



Kolegium Analiz Ekonomicznych

Praca doktorska

Dziedzina: nauki społeczne
Dyscyplina: ekonomia i finanse

mgr Daniel Kaszyński

STRONNICZOŚĆ ALGORYTMICZNA W OCENIE ZDOLNOŚCI KREDYTOWEJ

Promotor:
prof. dr hab. Bogumił Kamiński

Warszawa, listopad 2024

Streszczenie

Celem rozprawy jest opracowanie metod wspomagania decyzji regulatorów rynku kredytów konsumenckich w zakresie tworzenia oraz banków w zakresie wdrażania przepisów równościowych w obszarze modeli oceny zdolności kredytowej konsumentów. Opracowane metody stanowią metodykę dla banków w zakresie konstrukcji modeli oceny zdolności kredytowej redukujących problem ich stronniczości algorytmicznej oraz metodykę dla regulatorów rynku kredytów konsumenckich w zakresie oceny konsekwencji społecznych wprowadzania regulacji równościowych.

Modele oceny zdolności kredytowej są wykorzystywane przez banki do automatyzacji procesu podejmowania decyzji kredytowych związanych z aplikacjami konsumentów. Decyzje kredytowe podejmowane przy wykorzystaniu modeli oceny zdolności kredytowej, z racji wpływu na sytuację życiową osób fizycznych, np. sytuację finansową lub mieszkaniową, mogą być dyskryminujące dla określonej grupy ludzi; modele oceny zdolności kredytowej stosowane przez banki mogą charakteryzować się problemem stronniczości algorytmicznej. *Stronniczość algorytmiczna* występuje wtedy, gdy decyzja podjęta za pomocą modelu systematycznie stawia w niekorzystnej, gorszej sytuacji wybraną grupę społeczną określoną przez cechę prawnie chronioną (w zależności od regulacji, np. płeć, pochodzenie etniczne lub wyznanie).

Regulatorzy rynków kredytów konsumenckich, mając na względzie możliwe działanie modelu decyzyjnego, które prowadzi do dyskryminacji lub naruszenia norm społecznych w zakresie równości – stronniczość algorytmiczna – wprowadzają przepisy równościowe zakazujące tej dyskryminacji. *Perspektywa decyzyjna regulatora* jest związana z projektowaniem regulacji równościowej, np. wybór miary wykorzystywanej do identyfikacji problemu stronniczości algorytmicznej.

Banki wdrażają zapisy regulacji równościowych przez konstrukcję modeli oceny zdolności kredytowej, uwzględniających konieczność redukcji ich stronniczości algorytmicznej. Wybór optymalnego ekonomicznie podejścia do redukcji problemu stronniczości algorytmicznej, pod warunkiem ograniczeń zadanych przez regulacje równościowe, jest *problemem decyzyjnym banku*.

Model oceny zdolności kredytowej, opracowany przez bank, zgodnie z regulacjami równościowymi zdefiniowanymi przez regulatora, jest następnie wykorzystywany do oceny zdolności kredytowej kredytobiorców. Skutki wcześniejszych decyzji regulatora rynku

kredytów konsumenckich oraz banku są odzwierciedlone w procesie kredytowym wspieranym przez model oceny zdolności kredytowej, a kredytobiorcy (społeczeństwo) są finalnym biorcą decyzji kredytowych. Wraz ze zmianą stosowanego podejścia, np. wprowadzenie nowych lub zmiana obowiązujących regulacji równościowych, powstają *konsekwencje społeczne* decyzji regulatora i banku.

Cel badawczy zweryfikowano sześcioma hipotezami, z których każda dotyczy innego aspektu badanego problemu. Po pierwsze, zbadano czy metoda stosowana obecnie przez banki, tzw. *sprawiedliwości przez nieświadomość*, polegająca na odrzuceniu informacji o wartości zmiennej prawnie chronionej z modelu oceny zdolności kredytowej jest efektywna ekonomicznie (np. zmienna płeć nie jest wykorzystywana ani na etapie budowy, ani na etapie wykorzystania modelu oceny zdolności kredytowej). Po drugie, przebadano czy w zależności od przyjętej miary stronniczości algorytmicznej, inne metody redukcji stronniczości algorytmicznej będą efektywne ekonomicznie. Po trzecie, oceniono możliwość opracowania procedury redukującej stronniczość algorytmiczną modelu oceny zdolności kredytowej maksymalizującej zysk banku. Po czwarte, weryfikacji poddano możliwość opracowania metodyki oceny długookresowych konsekwencji wprowadzania regulacji równościowych związanych z modelami oceny zdolności kredytowej. Po piąte, przebadano zasadność zastosowania okresu dostosowawczego po wprowadzeniu regulacji równościowej, wynikającego z możliwej krótkookresowej niestabilności rozkładu wydawanych decyzji kredytowych banku. Po szóste, oceniono w jaki sposób wprowadzenie regulacji równościowych prowadzi do kosztu ekonomicznego związanego ze zmianą populacji osób otrzymujących pozytywne decyzje kredytowe.

Wyniki badania potwierdzają, że obecnie wykorzystywane przez banki podejście *sprawiedliwości przez nieświadomość* nie jest efektywne i istnieją lepsze, z perspektywy rachunku ekonomicznego banku oraz zakazu dyskryminacji, metody redukcji stronniczości algorytmicznej. W rozprawie wykazano, że w zależności od dokonanego przez regulatora wyboru dotyczącego miary stronniczości algorytmicznej, inne metody redukcji stronniczości algorytmicznej będą, z perspektywy banku, efektywniejsze. Mając na względzie różne podejścia do redukcji stronniczości algorytmicznej, bank może wybrać to, które maksymalizuje zysk. W rozprawie opracowano i wykorzystano metodykę oceny długookresowych konsekwencji wprowadzania regulacji równościowych związanych z modelami oceny zdolności kredytowej. Przy wykorzystaniu tej metodyki wykazano, że wprowadzenie regulacji równościowych może powodować krótkookresową niestabilność rozkładu decyzji kredytowych. Dodatkowym wynikiem rozprawy jest wskazanie, że wprowadzenie regulacji równościowych prowadzi do kosztu ekonomicznego związanego ze zmianą populacji osób otrzymujących pozytywne decyzje kredytowe, wynikającego ze zmiany jakości portfela kredytowego lub zmiany poziomów akceptacji wniosków kredytowych.

Abstract

The dissertation's objective is to develop methods to support the decisions of consumer credit regulatory entities in creating, and banks in implementing equality regulations in consumer credit scoring models. The developed methods constitute a methodology for banks to construct credit scoring models that reduce algorithmic bias and a methodology for consumer credit regulatory entities to assess the consequences of introducing equality regulations.

Banks use credit scoring models to automate consumer credit-granting decisions. Impacting individuals' life situations, such as financial or housing conditions, these regulations can be discriminatory toward specific groups of people. Specifically, banks' credit scoring models may exhibit algorithmic bias, which occurs when a model's decision or prediction systematically disadvantages a selected social group defined by a legally protected characteristic (depending on regulations, such as gender, ethnic origin, or religion).

Consumer credit regulatory entities, aware of the risk that decision models entail, leading to discrimination or violating social norms regarding equality – algorithmic bias – introduce equality regulations prohibiting such discrimination. The *regulator's decision-making perspective* involves designing equality regulations, including the choice of the metric used to identify algorithmic bias.

Banks implement equality regulations by constructing credit scoring models that take into account the necessity of reducing algorithmic bias. The selection of the optimal approach to reduce algorithmic bias, given the constraints imposed by equality regulations, constitutes the *bank's decision-making problem*.

The credit scoring model, developed by the bank following the equality regulations defined by the regulator, is then used to assess borrowers' creditworthiness. The credit scoring model reflects the effects of prior decisions made by the consumer credit regulatory entity and the bank. Borrowers (society) are the ultimate recipients of these credit decisions. *Social consequences* of the regulators' and banks' decisions emerge with changes in credit scoring – either by introducing new or modifying existing equality regulations.

The research objective has been divided into six hypotheses concerning different aspects of the research problem. Firstly, the *fairness through unawareness* method currently used by banks is verified in terms of its economic efficiency. The method excludes legally

protected attributes from the credit scoring model (e.g., gender is used neither at the model construction stage nor during the utilization of the model). Secondly, it was examined whether different methods of reducing algorithmic bias would be economically efficient depending on the chosen measure of algorithmic bias. Thirdly, the possibility of developing a procedure that reduces algorithmic bias in the credit scoring model while maximizing the bank's profit is assessed. Fourthly, the development of a methodology for assessing the long-term consequences of introducing equality regulations related to credit scoring models is investigated. Fifthly, the justification for applying an adjustment period after introducing the equality regulations was examined, considering the potential short-term instability in the distribution of credit decisions made by the bank. Sixthly, the economic cost associated with the changes in the population of individuals receiving positive credit decisions due to the introduction of equality regulations is assessed.

The dissertation results confirm that *fairness through unawareness* approach currently used by banks is not economically efficient and that, considering the perspective of the bank's economics, there are more optimal methods for reducing algorithmic bias. The dissertation demonstrates that depending on the regulator's choice of algorithmic bias metric, different methods of reducing algorithmic bias are more efficient from the bank's perspective. Given various approaches to reducing algorithmic bias, the bank can select the one that maximizes its profit. The dissertation presents the developed and applied methodology for assessing the long-term consequences of introducing equality regulations related to credit scoring models. Based on this methodology, it is proved that the introduction of equality regulations may cause short-term instability in the distribution of credit decisions. An additional finding of the dissertation is that the introduction of equality regulations leads to economic costs associated with changes in the population of individuals receiving positive credit decisions, affecting the quality of the credit portfolio or altering levels of credit application acceptance.

Spis treści

Wykaz skrótów aktów prawnych	9
Wprowadzenie	11
1 Kredyt i proces kredytowy	31
1.1 Historia kredytu	31
1.2 Proces kredytowy	42
1.3 Dane wykorzystywane w procesie kredytowym	44
2 Ocena zdolności kredytowej konsumenta	49
2.1 Ocena zdolności kredytowej	50
2.2 Ocena jakości modeli oceny zdolności kredytowej	55
2.3 Ekonomiczne aspekty oceny zdolności kredytowej	60
3 Stronniczość algorytmiczna	67
3.1 Pojęcie stronniczości algorytmicznej	67
3.2 Prawo dotyczące równości i zakazu dyskryminacji	76
3.3 Stronniczość algorytmiczna w modelach oceny zdolności kredytowej	87
3.4 Metody redukcji stronniczości algorytmicznej	95
4 Bezpośredni wpływ redukcji stronniczości algorytmicznej	103
4.1 Schemat modelu statycznego	104
4.2 Scenariusz danych syntetycznych	107
4.3 Scenariusz danych rzeczywistych	127
5 Wielookresowy wpływ redukcji stronniczości algorytmicznej	137
5.1 Schemat modelu dynamicznego	138
5.2 Metoda bazowa: Unmitigated	139
5.3 Metoda bazowa: Fairness through Unawareness	145
5.4 Triada redukcji stronniczości algorytmicznej	150
Podsumowanie	156
Aneks: Specyfikacja danych	164

Bibliografia	170
Netografia	181
Indeks rzeczowy	183
Spis tablic	187
Spis rysunków	189

Wykaz skrótów aktów prawnych

- AI Act** – Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828
- Basel** – Basel Committee on Banking Supervision, 2023. IRB approach: minimum requirements to use IRB approach. Bank for International Settlements
- BSA** – Pub. L. 91–508, 84 Stat. 1114; 31 U.S.C. § 5311–5330, Bank Secrecy Act
- CCPA** – Pub. L. 90–321, 82 Stat. 146; 15 U.S.C. § 1601, Consumer Credit Protection Act
- CPPA** – Cal. Civ. Code §§ 1798.100 et seq., California Consumer Privacy Act of 2018
- decyzja 2006/01/30** – Decyzja Rady z dnia 30 stycznia 2006 r. w sprawie zasad, priorytetów i warunków zawartych w partnerstwie europejskim z Serbią i Czarnogorą, w tym z Kosowem w rozumieniu rezolucji nr 1244 Rady Bezpieczeństwa Organizacji Narodów Zjednoczonych z dnia 10 czerwca 1999 r. oraz uchylająca decyzję 2004/520/WE
- dyrektywa 2000/43/WE** – Dyrektywa Rady 2000/43/WE z dnia 29 czerwca 2000 r. wprowadzająca w życie zasadę równego traktowania osób bez względu na pochodzenie rasowe lub etniczne
- dyrektywa 2004/113/WE** – Dyrektywa Rady 2004/113/WE z dnia 13 grudnia 2004 r. wprowadzająca w życie zasadę równego traktowania mężczyzn i kobiet w zakresie dostępu do towarów i usług oraz dostarczania towarów i usług
- EBA** – European Banking Authority, 2016. Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013
- ECOA** – Pub. L. 93–495, 88 Stat. 1521; 15 U.S.C. § 1691, Equal Credit Opportunity Act

- ECOA-B** – 12 CFR Part 1002, Regulacja B, implementująca Equal Credit Opportunity Act
- EDPS** – European Data Protection Supervisor (EDPS), 2021. Opinion 11/2021 on the Proposal for Directive on customer credits
- EUNB** – Europejski Urząd Nadzoru Bankowego, 2010. Wytyczne CEBS w sprawie zarządzania ryzykiem koncentracji wynikające z procesu badania i oceny nadzorczej
- FCBA** – Pub. L. 93–495, 88 Stat. 1500; 15 U.S.C. § 1666, Fair Credit Billing Act
- FCRA** – Pub. L. 91–508, 84 Stat. 1127; 15 U.S.C. § 1681, Fair Credit Reporting Act
- FHA** – Pub. L. 90–284, 82 Stat. 73; 42 U.S.C. § 3601, Fair Housing Act
- FHAA** – Pub. L. 100–430, 102 Stat. 1619; 42 U.S.C. § 3601 et seq., Fair Housing Amendments Act
- IFRS 9** – International Financial Reporting Standard 9, 2021
- KNF 2015** – Rekomendacja W Komisji Nadzoru Finansowego dotycząca zarządzania ryzykiem modeli w bankach
- KNF 2019** – Rekomendacja S Komisji Nadzoru Finansowego dotycząca dobrych praktyk w zakresie zarządzania ekspozycjami kredytowymi zabezpieczonymi hipotecznie
- RODO** – Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE
- ustawa 2006/07/21** – Ustawa z dnia 21 lipca 2006 r. o nadzorze nad rynkiem finansowym
- ustawa 2010/12/03** – Ustawa z dnia 3 grudnia 2010 r. o wdrożeniu niektórych przepisów Unii Europejskiej w zakresie równego traktowania
- ustawa 2011/05/12** – Ustawa z dnia 12 maja 2011 r. o kredycie konsumenckim
- ustawa 2017/03/23** – Ustawa z dnia 23 marca 2017 r. o kredycie hipotecznym oraz o nadzorze nad pośrednikami kredytu hipotecznego i agentami
- ustawa 2022/10/06** – Ustawa z dnia 6 października 2022 r. o zmianie ustaw w celu przeciwdziałania lichwie
- ustawa 1964/04/23** – Ustawa z dnia 23 kwietnia 1964 r. – Kodeks cywilny
- ustawa Prawo Bankowe** – Ustawa z dnia 29 sierpnia 1997 r. – Prawo bankowe
-

Wprowadzenie

Celem rozprawy jest opracowanie metod wspomagania decyzji regulatorów rynku kredytów konsumenckich w zakresie tworzenia oraz banków w zakresie wdrażania przepisów równościowych w obszarze modeli oceny zdolności kredytowej konsumentów¹. Wypracowane metody stanowią metodykę dla banków w zakresie konstrukcji modeli oceny zdolności kredytowej redukujących problem stronniczości algorytmicznej tych modeli, a także metodykę dla regulatorów rynku kredytów konsumenckich w zakresie oceny konsekwencji społecznych wprowadzania regulacji równościowych.

Wprowadzanie przepisów równościowych jest związane z występującym w modelach oceny zdolności kredytowej problemem *stronniczości algorytmicznej*, tj. takim działaniem modelu decyzyjnego, które prowadzi do dyskryminacji lub naruszenia norm społecznych w zakresie równości ludzi². Decyzję o prawnym zakresie definicji dyskryminacji podejmuje podmiot regulujący rynek kredytów konsumenckich – *regulator*. Problem stronniczości algorytmicznej występuje wtedy, gdy wyniki działania modelu decyzyjnego prowadzą do szkodliwych konsekwencji dla niektórych osób lub grupy społecznej bez uzasadnionego powodu takiej nierówności (Friis i Riley 2023; Kordzadeh i Ghasemaghaei 2022).

Modele oceny zdolności kredytowej wykorzystywane są do automatyzacji procesu podejmowania decyzji kredytowych związanych z aplikacjami kredytowymi konsumentów (ustawa Prawo Bankowe, art. 105a). Z racji ich oddziaływania na sytuację życiową osób fizycznych, np. sytuację finansową, mieszkaniową oraz możliwe występowanie dyskryminacyjnego działania, modele oceny zdolności kredytowej są narażone na zagrożenie stronniczości algorytmicznej (Kozodoi i Lessmann 2022; Hurlin, Pérignon i Saurin 2022; Bono, Croxson i Giles 2021; Hurley i Adebayo 2016)

Definicja stronniczości algorytmicznej jest związana z negatywnymi konsekwencjami działania modeli decyzyjnych na ludzi; w literaturze przedmiotu wskazano na liczne sposoby pomiaru występowania problemu stronniczości algorytmicznej, tzw. *miary stronniczości algorytmicznej* (Barocas, Hardt i Narayanan 2023; Kim, Chen

¹ Zgodnie z art. 22¹ ustawy 1964/04/23, konsumentem jest „[...] osoba fizyczna dokonująca z przedsiębiorcą czynności prawnej niezwiązanej bezpośrednio z jej działalnością gospodarczą lub zawodową”.

² „Stronniczość algorytmiczna ma miejsce, gdy algorytmy podejmują decyzje, które systematycznie stawiają w niekorzystnej sytuacji określone grupy ludzi.” [tł. wł.], (Friis i Riley 2023).

„(Stronniczość algorytmiczna ma dop. aut.) miejsce, gdy wyniki algorytmu przynoszą korzyść lub niekorzyść pewnym osobom lub grupom bardziej niż innym, bez uzasadnionego powodu takich nierównych skutków.” [tł. wł.], (Kordzadeh i Ghasemaghaei 2022).

i Talwalkar 2020). Decyzja regulatora związana z wyborem miary wykorzystywanej do identyfikacji problemu stronniczości algorytmicznej będzie się wiązać z:

- kształtem projektowanej regulacji prawnej, np. wybór definicji dyskryminacji jako *dyskryminacji bezpośredniej* lub *dyskryminacji pośredniej*;
- rachunkiem ekonomicznym banku związanym z realizowanym procesem kredytowym;
- konsekwencjami społecznymi wynikającymi ze zmiany rozkładu wydawanych decyzji kredytów, tj. przed i po wprowadzeniu regulacji równościowej.

Decyzja związana z projektowaniem regulacji równościowych dotyczy *perspektywy decyzyjnej regulatora* rynku kredytów konsumenckich.

Standardowym i obecnie najczęściej wykorzystywanym w praktyce podejściem do zapewniania, że model oceny zdolności kredytowej nie dyskryminuje wybranej grupy społecznej zdefiniowanej na podstawie *cechy prawnie chronionej*³ jest odrzucenie informacji o tej cesze z procedury budowy modelu (Barocas, Hardt i Narayanan 2023; Dwork i in. 2012). Oznacza to, że ani na etapie transformacji zbioru treningowego, ani na etapie trenowania modelu, ani w prognozie z modelu i jej wykorzystaniu informacja o zmiennej prawnie chronionej nie jest wykorzystywana; podejście to nazywane jest podejściem *sprawiedliwości przez nieświadomość* i odnosi się do definicji dyskryminacji bezpośredniej, związanej z odmiennym traktowaniem osób fizycznych ze względu na atrybut prawnie chroniony. Budowa modeli oceny zdolności kredytowej zgodnie z podejściem sprawiedliwości przez nieświadomość jest, w większości jurysdykcji na świecie, obecnie prawnie wymagana (EDPS; ECOA-B), co ma swoje odzwierciedlenie w krajowej i międzynarodowej praktyce banków.

Współcześnie rozważa się również alternatywne podejścia do rozwiązywania, redukcji, problemu stronniczości algorytmicznej w modelach oceny zdolności kredytowej. Zmiana wykorzystywanego podejścia ze wskazanej wcześniej sprawiedliwości przez nieświadomość może się wiązać ze zmianą jakości prognostycznej konstruowanych modeli oceny zdolności kredytowej, a tym samym inną jakością samego procesu decyzji kredytowej, co w konsekwencji prowadzić będzie do innego wyniku finansowego uzyskiwanego przez bank (Barocas, Hardt i Narayanan 2023; Kozodoi i Lessmann 2022). Wybór optymalnego podejścia do redukcji problemu stronniczości algorytmicznej, pod warunkiem zadanych regulacji równościowych, jest *problemem decyzyjnym banku*.

Model oceny zdolności kredytowej, opracowany przez bank zgodnie z regulacjami równościowymi zdefiniowanymi przez regulatora, jest następnie wykorzystywany do oceny

³ Cecha prawnie chroniona to cecha osób fizycznych, na podstawie której normatywnie zabronionym jest różnicowanie sytuacji tych osób (RPO 2020). Dyskryminacją jest różnicowanie sytuacji podmiotów znajdujących się w podobnej sytuacji oparte na ich cechach prawnie chronionych. Do katalogu cech prawnie chronionych zalicza się m.in.: płeć, rasę, pochodzenie etniczne, narodowość, religię, wyznanie, światopogląd, niepełnosprawność, wiek lub orientacja seksualną (ustawa 2010/12/03). W rozprawie przyjmuje się również zamiennie określenia: *zmienna prawnie chroniona* lub *atrybut prawnie chroniony*.

zdolności kredytowej kredytobiorców. Skutki wcześniejszych decyzji regulatora rynku kredytów konsumenckich oraz banku są odzwierciedlone w procesie kredytowym wspieranym przez model oceny zdolności kredytowej; kredytobiorcy (społeczeństwo) są finalnym biorcą decyzji kredytowych. Wraz ze zmianą stosowanego podejścia, np. wprowadzenie nowych lub zmiana obowiązujących regulacji równościowych, powstają *konsekwencje społeczne* decyzji regulatora i banku.

Geneza celu badawczego

Decyzje kredytowe podejmowane przez banki związane z konsumentami są wspierane przez modele oceny zdolności kredytowej od ok. lat 60. XX w. (Yu i Dunn 2016; FICO 1998). Przed okresem wykorzystania modeli oceny zdolności kredytowych banki podejmowały decyzje kredytowe opierając się na *ocenie eksperckiej*: analitycy kredytowi (eksperci) wykorzystując wiedzę, doświadczenie oraz znajomość osobistą klienta dokonywali takiej oceny i podejmowali decyzję o udzieleniu kredytu (Marques, García i Sánchez 2013).

Metody oceny zdolności kredytowej

Ekspercka ocena zdolności kredytowej była oceną subiektywną, zależną od indywidualnej wiedzy i doświadczenia danego analityka kredytowego. Była narażona na nadużycia związane z indywidualnymi preferencjami lub animozjami danego analityka kredytowego (Uzoamaka, Gerald i Jude 2016). Tak realizowany proces oceny zdolności kredytowej był:

- mało wydajny, tj. czas od złożenia aplikacji kredytowej do decyzji kredytowej był długi, liczony nawet w tygodniach oraz względnie kosztowny (obsługiwany przez człowieka);
- nisko skalowany, tj. ograniczona liczba wydawanych decyzji kredytowych wynikająca z dostępności i obciążenia czasowego analityków kredytowych;
- zależny od indywidualnej wiedzy i doświadczenia analityka kredytowego, w tym podatny na indywidualne animozje lub uprzedzenia analityka kredytowego.

Wraz z rozwojem technologii informatycznych, zaawansowaniem metod analitycznych oraz występującymi zachowaniami dyskryminacyjnymi wybranych grup społecznych, podmioty regulujące rynek dostępu do kredytów konsumenckich zaczęły prawnie wymagać automatyzacji procesu oceny zdolności kredytowej konsumentów. Ze względu na wzrost wydajności, skalowalności oraz obiektywności procesu oceny zdolności kredytowej realizowanego przez modele oceny zdolności kredytowej, banki były skłonne wprowadzać automatyzację procesu oceny zdolności kredytowej przez wykorzystanie do tego modeli tej

oceny. W konsekwencji prowadziło to do upowszechnienia się wykorzystania modeli oceny zdolności kredytowej i zastąpienie nimi wcześniej prowadzonej oceny eksperckiej (Yu i Dunn 2016; Marques, García i Sánchez 2013; Abdou 2009).

Obecnie wykorzystanie *modeli oceny zdolności kredytowej* w procesie decyzji kredytowej konsumenta jest prawnie wymagane (ustawa Prawo Bankowe, art. 105a). Modele te są klasyfikatorami, tj. zmienna celu przyjmuje dyskretne wartości. W rozprawie przyjęto, że zmienną celu modelu oceny zdolności kredytowej będzie informacja o spłacie, bądź nie, zobowiązania kredytowego (Kaszyński, Kamiński i Szapiro 2020); zmienną celu jest zatem zmienna binarna przyjmująca odmienne wartości dla spłaconych kredytów oraz stanów, w których klient stracił *zdolność kredytową*⁴ i nie spłacił zaciągniętego zobowiązania. Modele uczenia nadzorowanego są to modele, w których uczenie (nazywane inaczej *trenowaniem*) odbywa się przy znanych wartościach zmiennej celu.

Uczenie modelu oceny zdolności kredytowej, oparte na o binarnej zmiennej celu, odbywa się przy wykorzystaniu historycznych danych, tj. historycznie udzielonych kredytów, obejmujących (Przanowski 2014):

- zmienne aplikacyjne (np. poziom miesięcznego wynagrodzenia i wydatków) lub zmienne behawioralne (np. terminowość spłat dotychczasowych zobowiązań) stanowiące zbiór zmiennych objaśniających modelu;
- informację o historycznych spłatach kredytów stanowiącą modelowaną zmienną celu.

Bank dysponujący danymi o historycznych kohortach swoich klientów (tj. kredytobiorcy, którzy w przeszłości otrzymali pozytywną decyzję kredytową), jest w stanie oszacować model klasyfikacyjny, w którym uczenie odbywa się w sposób nadzorowany, a wskazane dane historyczne są wykorzystywane do jego trenowania.

Adekwatność ekonomiczna i sprawiedliwość podejmowanych decyzji kredytowych, w reżimie oceny eksperckiej, wynikały z osądów i ewaluacji procesu oceny zdolności kredytowej realizowanego przez analityka kredytowego. Wraz z przejściem na wystandaryzowane, automatyczne modele oceny zdolności kredytowej, ocena adekwatności ekonomicznej i sprawiedliwości ich działania jest symetrycznie przeniesiona na poziom działania wykorzystywanego modelu. Coraz powszechniejsze wykorzystanie modeli oceny zdolności kredytowej przyczyniło się równocześnie do intensyfikacji potrzeb związanych z analizą wymiaru społecznego tych modeli, tj. sprawiedliwością ich działania rozumianą jako równością w dostępie do kredytu (Hurlin, Pérignon i Saurin 2022).

⁴ „Przez zdolność kredytową rozumie się zdolność do spłaty zaciągniętego kredytu wraz z odsetkami w terminach określonych w umowie. Kredytobiorca jest obowiązany przedłożyć na żądanie banku dokumenty i informacje niezbędne do dokonania oceny tej zdolności.” (ustawa Prawo Bankowe, art. 70).

Stronniczość algorytmiczna

Indywidualnie ludzie mogą posiadać odmienne opinie na temat definicji sprawiedliwości, w tym równości działania algorytmów decyzyjnych; w rozprawie stosuje się wykładnię sprawiedliwości definiowaną przez regulatora, tj. prawne zapisy dot. równości. W pracy badanymi będą konsekwencje podejmowanych przez regulatorów decyzji w zakresie definicji równości oraz wpływu tych decyzji na adekwatność stosowania podejść do redukcji stronniczości algorytmicznej, które mogą wykorzystać banki w przypadku modeli oceny zdolności kredytowej. Definicja *sprawiedliwości działania algorytmów* decyzyjnych wymaga odpowiedzi na następujące pytania:

- W jaki sposób można stwierdzić, że algorytm, model decyzyjny jest sprawiedliwy?
- Względem jakich cech, atrybutów można stwierdzić, że algorytm/model decyzyjny jest sprawiedliwy?

W ramach rozprawy wykorzystano wyniki przedstawiane w literaturze przedmiotu, zakresem obejmujące wskazane pytania.

Dyskryminacyjny, nierówny charakter prognoz generowanych przez modele decyzyjne, w tym modele oceny zdolności kredytowej, jest nazywany *stronniczością algorytmiczną* lub *skrzywieniem algorytmicznym* (Warchoń 2023). Stronniczość algorytmiczna jest negatywną cechą modeli decyzyjnych odpowiadającą definicji dyskryminacji występującej w decyzjach podejmowanych opartych na tych modelach; stronniczość algorytmiczna może być dalej kategoryzowana na: *odmienny wpływ* na osoby o różnej wartości cechy chronionej (np. kobiety i mężczyźni) lub *odmienne traktowanie* danej grupy społecznej, definiowanej przez cechę prawnie chronioną. Poniżej omówiono dwa wymiary dyskryminacji: odmienny wpływ oraz odmienne traktowanie, które mogą występować w modelach oceny zdolności kredytowej⁵.

Odmienny wpływ (ang. *disparate impact*) modelu oceny zdolności kredytowej jest przypadkiem tzw. dyskryminacji pośredniej. Pojęcie *dyskryminacji pośredniej* funkcjonuje w prawie i odnosi się do sytuacji „[...] w której dla osoby fizycznej ze względu na płeć, rasę, pochodzenie etniczne, narodowość, religię, wyznanie, światopogląd, niepełnosprawność, wiek lub orientację seksualną na skutek pozornie neutralnego postanowienia, zastosowanego kryterium lub podjętego działania występują lub mogłyby wystąpić niekorzystne dysproporcje lub szczególnie niekorzystna dla niej sytuacja [...]” (ustawa 2010/12/03). Odmienny wpływ to przykład dyskryminacji, w której osoby są traktowane jednakowo w ramach określonego zestawu zasad lub procedur (tj. model oceny zdolności kredytowej), ale te zasady lub procedury dają przewagę jednej z grup; model jest stosowany w równym stopniu wobec wszystkich, ale wpływ jego działania jest

⁵ W rozprawie pominięto inne wymiary definicji dyskryminacji, tj. *molestowanie* lub *molestowanie seksualne* (ustawa 2010/12/03).

nieproporcjonalnie niekorzystny dla wnioskodawców z grupy nieuprzywilejowanej. Model stroniczy algorytmicznie w rozumieniu odmiennego wpływu będzie charakteryzował się statystycznie różną, dla grup społecznych wyróżnionych przez wartości cechy prawnie chronionej, wartością wskaźnika zatwierdzeń kredytów. Regulator zastrzegł jednak, że odmienny wpływ może być dopuszczalny, gdy: „[...] postanowienie, kryterium lub działanie jest obiektywnie uzasadnione ze względu na zgodny z prawem cel, który ma być osiągnięty, a środki służące osiągnięciu tego celu są właściwe i konieczne” (ustawa 2010/12/03).

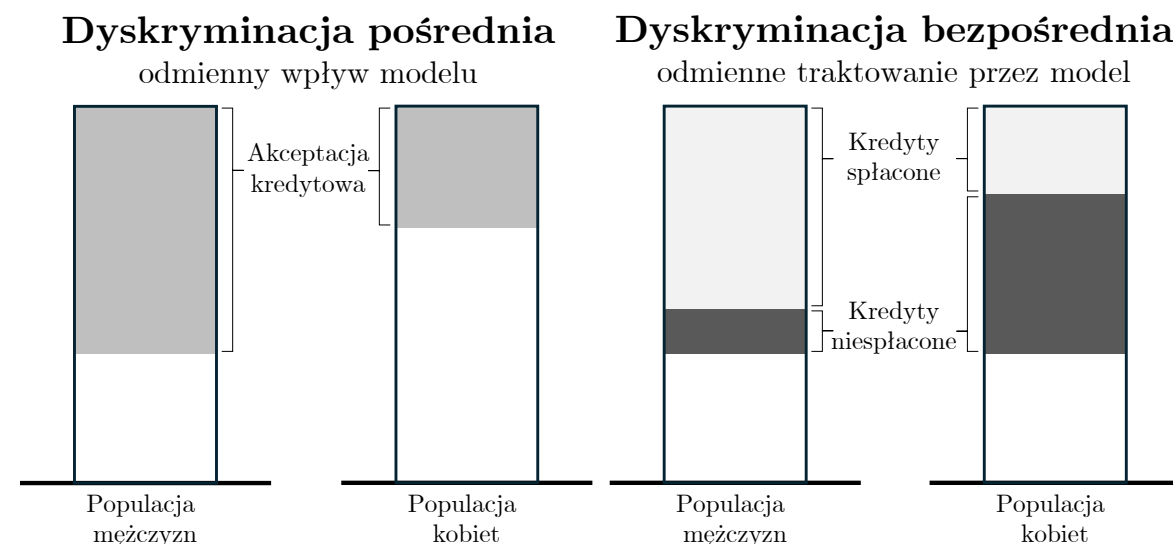
Przyjmijmy przykładowo, że w danym okresie, np. styczeń 2022 r., bank przyjął określoną liczbę aplikacji kredytowych n (np. $n = 500$), które były w ilości $n_{X_a=0}$ od mężczyzn (np. $n_{X_a=0} = 400$) oraz $n_{X_a=1}$ od kobiet (np. $n_{X_a=1} = 100$)⁶. Przyjmijmy, również, że bank, opierając się na wynikach modelu oceny zdolności kredytowej, udzielił kredyt dla: $n_{X_a=0}^+$ mężczyzn (np. $n_{X_a=0}^+ = 200$) oraz $n_{X_a=1}^+$ kobiet (np. $n_{X_a=1}^+ = 20$). Wtedy odmienny wpływ modelu oceny zdolności kredytowej odnosić się będzie do różnicy w proporcjach *wskaźnika zatwierdzeń* kredytów dla mężczyzn i kobiet, tj. $\frac{n_{X_a=0}^+}{n_{X_a=0}}$ oraz $\frac{n_{X_a=1}^+}{n_{X_a=1}}$ (w przykładzie: $\frac{200}{400} = 50\%$ dla mężczyzn oraz $\frac{20}{100} = 20\%$ dla kobiet).

Odmienne traktowanie (ang. disparate treatment) przez model oceny zdolności kredytowej jest przypadkiem tzw. dyskryminacji bezpośredniej. Pojęcie to odnosi się do sytuacji „[...] w której osoba fizyczna ze względu na płeć, rasę, pochodzenie etniczne, narodowość, religię, wyznanie, światopogląd, niepełnosprawność, wiek lub orientację seksualną jest traktowana mniej korzystnie niż jest, była lub byłaby traktowana inna osoba w porównywalnej sytuacji” (ustawa 2010/12/03). W przypadku odmiennego traktowania, model może posiadać taki sam wpływ na grupy społeczne, różnicowane przez wartość cechy prawnie chronionej – obie grupy społeczne posiadają taki sam poziom wskaźnika zatwierdzeń kredytów – ale sama procedura traktuje jedną z grup społecznych w sposób mniej korzystny (grupa nieuprzywilejowana). Jako przykład literatury podaje się zjawisko tzw. *lenistwa modelu* (ang. model laziness), tj. sytuacji, w której model dla jednej grupy (tzw. grupy uprzywilejowanej) posiada wysoką jakość prognostyczną (tj. niewielki błąd klasyfikacji), a dla drugiej grupy bardzo niską jakość prognostyczną (tj. wysoki błąd klasyfikacji). W skrajnym scenariuszu można wskazać jako przykład model, który np. dla mężczyzn wykorzystuje model idealny (tj. o zerowym błędzie klasyfikacji), a dla kobiet model losowy (Makhlouf, Zhioua i Palamidessi 2021).

Przykładowe scenariusze występowania problemu stroniczości algorytmicznej, wyrażone jako dyskryminacja pośrednia (odmienny wpływ) oraz dyskryminacja bezpośrednia (odmienne traktowanie) w kontekście oceny zdolności kredytowej, zostały przedstawione na rysunku 1.

Wskazane wymiary dyskryminacji – odmienny wpływ oraz odmienne traktowanie –

⁶ Przyjęto, że wartość zmiennej $X_a = 0$ oznacza mężczyznę, a $X_a = 1$ oznacza kobietę.



Rysunek 1: Przykład dyskryminacji pośredniej i bezpośredniej w ocenie zdolności kredytowej

Uwaga: Na wykresie lewym przedstawiono scenariusz banku, który w danym okresie posiadał *różny* wskaźnik zatwierdzeń kredytów między mężczyznami (lewa kolumna) i kobietami (prawa kolumna); scenariusz ten przedstawia *odmienny wpływ* modelu oceny zdolności kredytowej (niekorzystnie niższy wskaźnik zatwierdzeń dla kobiet). Na wykresie prawym przedstawiono scenariusz banku, który w danym okresie posiadał *równy* wskaźnik zatwierdzeń kredytów dla mężczyzn i kobiet, przy czym dla kobiet model charakteryzował się wyższym błędem klasyfikacji (tj. dla populacji udzielonych kredytów, więcej było tych, które nie zostały spłacone).

Źródło: opracowanie własne.

w pracy będą omawiane w kontekście atrybutów, cech konsumentów, które regulator chroni, tzw. *cech prawnie chronionych*. Prawna ochrona zmiennych odbywa się na poziomie regulacji antydyskryminacyjnych związanych z kredytami dla osób fizycznych; regulacje te nie wskazują jednak wprost na wymiar dyskryminacji, który należy stosować. Przykładowe przepisy prawne, krajowe i zagraniczne, zawierające zapisy regulacyjne przeciwdziałające dyskryminacji, zamieszczono w:⁷

⁷ Obok przepisów bezpośrednio lub pośrednio odnoszących się do obszaru modeli oceny zdolności kredytowej, warto wskazać również inne akty prawne, które wskazują na ogólne ramy zasad równościowych. Są to:

- Dyrektywa 2004/113/WE Parlamentu Europejskiego i Rady z dnia 13 grudnia 2004 r. w sprawie wprowadzenia w życie zasady równego traktowania mężczyzn i kobiet w zakresie dostępu do towarów i usług oraz dostarczania towarów i usług.
- Dyrektywa Rady 79/7/EWG z dnia 19 grudnia 1978 r. w sprawie stopniowego wprowadzania w życie zasady równego traktowania kobiet i mężczyzn w dziedzinie zabezpieczenia społecznego.
- Pub. L. No. 91–508, 84 Stat. 1114; 15 U.S.C. § 1681 et seq., Fair Credit Reporting Act.
- Międzynarodowy Pakt Praw Obywatelskich i Politycznych, przyjęty przez Zgromadzenie Ogólne ONZ w rezolucji 2200A (XXI) z dnia 16 grudnia 1966 r., część II, art. 2 i 3.
- Pub. L. No. 88–352, 78 Stat. 241; Civil Rights Act of 1964.
- Powszechna Deklaracja Praw Człowieka, przyjęta przez Zgromadzenie Ogólne ONZ w rezolucji 217 A (III) z dnia 10 grudnia 1948 r., art. 2.

- Rozporządzeniu Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (RODO). Pomimo że samo RODO nie zawiera przepisów odnoszących się do działania modeli oceny zdolności kredytowej, to jest ono wykorzystywane pośrednio przez kolejne przepisy i opinie prawne jako wykładnia w zakresie definicji cech prawnie chronionych (w szczególności art. 9 RODO dotyczący przetwarzania szczególnych kategorii danych osobowych) dla obszaru modeli oceny zdolności kredytowej (EDPS).
- Ustawie z dnia 3 grudnia 2010 r. o wdrożeniu niektórych przepisów Unii Europejskiej w zakresie równego traktowania (ustawa 2010/12/03), która zawiera zapisy związane z definicjami dyskryminacji i jej możliwymi formami oraz uwzględnia kwestie procesowe ułatwiające postępowanie o naruszenie zasady równego traktowania.
- Cal. Civ. Code §§ 1798.100 et seq., California Consumer Privacy Act of 2018 (CCPA); kalifornijska ustawa o ochronie prywatności konsumentów zawiera regulacje obejmujące kwestie zarządzania danymi osobowymi konsumentów gromadzonymi przez firmy oraz prawa do dysponowaniem tymi danymi, tzw. prawo do wiedzy, prawo do usunięcia danych, prawo do rezygnacji z udostępniania danych, prawo do niedyskryminacji.
- Pub. L. 93–495, 88 Stat. 1521; 15 U.S.C. § 1691, Equal Credit Opportunity Act (ECOA); ustawa jest rozszerzeniem ustawy Consumer Credit Protection Act CCPA w zakresie zapewnienia równości szans w dostępie do produktów kredytowych, zawierającym zapisy zakazujące praktyk dyskryminujących na podstawie rasy, koloru skóry, pochodzenia narodowego, religii, płci, stanu cywilnego, wieku lub informacji o tym czy kredytobiorca korzystał z programów pomocy publicznej. Ustawa ECOA uwzględnia wymogi w zakresie równego traktowania konsumentów przez instytucje finansowe.
- Pub. L. 90–284, 82 Stat. 73; 42 U.S.C. § 3601, Fair Housing Act (FHA). Celem ustawy jest zapobieganie dyskryminacji na rynku mieszkaniowym. Zakaz dyskryminacji określony w FHA obejmuje: rasę, kolor skóry, religię oraz narodowość. Zapisy prawne FHA obejmują czynności prawne związane z wynajmem i sprzedażą, udzielaniem kredytów hipotecznych oraz oceną nieruchomości.

Podsumowanie wybranych przepisów prawa w zakresie atrybutów wskazanych jako cechy prawnie chronione przedstawiono w tabeli 1.

W literaturze przedmiotu wskazano na liczne przykłady danych bankowych, w których cechy prawnie chronione są zmiennymi statystycznie istotnymi w oszacowanych modelach zdolności kredytowej: płeć (Andreeva i Matuszyk 2019; Agarwal i in. 2018; Henderson

Tabela 1: Zakres cech prawnie chronionych zgodnie z wybranymi przepisami prawa

Cecha konsumenta	RODO	Regulacja			
		ustawa 2010/12/03	CCPA	FHA	ECOA
pleć	+	+	+	+	+
rasa / pochodzenie etniczne	+	+	+	+	+
wyznanie	+	+	+	+	+
kolor skóry	-	-	+	+	+
narodowość	-	+	-	+	+
niesprawność	+	+	-	+	-
tożsamość płciowa	+	-	+	+	-
dane dot. zdrowia	+	-	+	-	-
poglądy polityczne	+	+	-	-	-
przynależność do związków zaw.	+	-	+	-	-
wiek	-	+	-	-	+
adres email	-	-	+	-	-
dane genetyczne	-	-	+	-	-
korzystanie z pomocy publicznej	-	-	-	-	+
lokalizacja	-	-	+	-	-
skład gospodarstwa domowego	-	-	-	+	-
stan cywilny	-	-	-	-	+

Uwaga: + oznacza cechę prawnie chronioną w ramach wskazanej regulacji; - oznacza, że dana cecha nie jest wskazana w regulacji jako zmienna prawnie chroniona. Skrótowe nazwy aktów prawnych przedstawiono w wykazie skrótów aktów prawnych.

Źródło: opracowanie własne.

i in. 2015; Coin 2013; Do i Pakey 2013; D’espallier, Guérin i Mersland 2011; Muravyev, Talavera i Schäfer 2009), pochodzenie etniczne (Butler, Mayer i Weston 2023; Asiedu Freeman i Nti-Addae 2012; Havard 2010) lub wiek konsumenta (Dumitrescu i in. 2022; Okesola i in. 2017; Lee i in. 2002).

Na problem *dyskryminacji pośredniej* modeli oceny zdolności kredytowej zwrócono uwagę m.in. podczas przesłuchania przed Izbą Reprezentantów Stanów Zjednoczonych (U.S. House of Representatives 2010, s. 57): „[...] wykorzystanie ocen kredytowych ma tendencję do niekorzystnego wpływu na mniejszości etniczne, kobiety oraz inne grupy, których wyniki kredytowe są często niższe niż wyniki białych pożyczkobiorców. Może to być spowodowane ograniczonym dostępem tych grup do głównych źródeł kredytu, co skutkuje ich zależnością od źródeł kredytu negatywnie wpływających na wyniki kredytowe, takich jak firmy pożyczkowe i pożyczkodawcy oferujący chwilówki.” [tł. wł.].

Sprawiedliwość przez nieświadomość

Problem dyskryminacji może występować, gdy w modelu oceny zdolności kredytowej wykorzystywana jest cecha prawnie chroniona jako jedna ze zmiennych egzogenicznych, bez szczególnego uwzględnienia tej zmiennej; dalej takie podejście będzie nazywane *podejściem*

nieograniczonym (ang. *unmitigated*). W przypadku podejścia nieograniczonego cecha prawnie chroniona, jeżeli jest statystycznie istotna w oszacowanym modelu, uzyskuje parametr wrażliwości, który bezpośrednio można interpretować jako poziom dyskryminacji: dwie osoby różniące się między sobą wartością cechy prawnie chronionej *ceteris paribus*, uzyskują dwa różne oszacowania z modelu oceny zdolności kredytowej. Tym samym model ma odmienny wpływ na osoby różniące się wartością zmiennej prawnie chronionej (dyskryminacja pośrednia).

Banki i regulatorzy, mając świadomość wskazanego problemu z nieograniczonym podejściem bezpośrednio wykorzystującym zmienną chronioną w celu redukcji problemu stronniczości algorytmicznej modelu oceny zdolności kredytowej, stosują podejście związane z odrzuceniem informacji o wartości atrybutów chronionych⁸. Podejście to, *sprawiedliwości przez nieświadomość* (ang. *fairness through unawareness*, *fairness through blindness*), jest obecnie najczęściej wykorzystywaną metodą *redukcji stronniczości algorytmicznej*, a jego zasadność wykazuje się na podstawie tego, że model nie może dyskryminować danej grupy społecznej, jeżeli nie posiada informacji o wartości zmiennej chronionej ani zmiennych istotnie z nią skorelowanych (Chen 2018).

Metoda sprawiedliwości przez nieświadomość jedynie powierzchownie rozwiązuje problem stronniczości algorytmicznej (Knight 2019; Andreeva i Matuszyk 2019). Dane socjodemograficzne charakteryzują się korelacjami i interakcjami między zmiennymi; nieuwzględnienie tych dwóch efektów w modelu, w którym istotna zmienna predyktywna zostanie wyłączona, np. wyłączenie zmiennej płeć ze zbioru treningowego, prowadzić będzie do powstania tzw. *problemu zmiennej pominiętej*.

W przypadku modeli parametrycznych, bezpośrednim efektem problemu zmiennej pominiętej będzie obciążenie estymatorów pozostałych parametrów modelu: pozostałe parametry zmiennych będą systematycznie zawyżane lub zaniżane względem ich prawdziwej wartości (tj. średnia z szeregu estymacji będzie systematycznie różnić się od prawdziwej wartości parametru z populacji). Efekt ten może być na tyle silny, że parametry poszczególnych zmiennych mogą zmienić swój znak, co, w przypadku modeli oceny zdolności kredytowej, prowadzić będzie do nieuzasadnianych ekonomicznie decyzji kredytowych, tj. przydzielanie wyższej oceny zdolności kredytowej osobom o niższych zarobkach lub osobom o wyższym zadłużeniu – przyjmując, że ekonomicznie zasadnym byłaby odwrotna zależność (Wilms i in. 2021).

Dodatkowym efektem wykluczenia ze zbioru zmiennych predyktywnej zmiennej egzogenicznej jest pogorszenie jakości prognostycznej modelu oceny zdolności kredytowej (Kozodoi i Lessmann 2022). Bank, wykluczając zmienną, która jest statystycznie istotna

⁸ „[...] wyraźne zakazanie wykorzystywania jakichkolwiek specjalnych kategorii danych osobowych na mocy art. 9 RODO [...]” [tł. wł.] (EDPS).

„Kredytodawca nie może zadawać pytań o rasę, kolor skóry, religię, pochodzenie narodowe lub płeć wnioskodawcy lub jakiegokolwiek innej osoby w związku z transakcją kredytową [...]” [tł. wł.], (ECOA-B).

w oszacowanym modelu oceny zdolności kredytowej, narażony jest na spadek jakości generowanych decyzji kredytowych. Tym samym, bank będzie udzielać więcej kredytów klientom o niższej zdolności kredytowej lub mniej kredytów klientom o wysokiej zdolności kredytowej.

Przykładem sytuacji, w której wykluczenie zmiennej chronionej nie jest efektywną strategią jest budowa modelu oceny zdolności kredytowej (np. model oparty na danych aplikacyjnych), w którym obok wielu innych zmiennych egzogenicznych występują zmienne: **pochodzenie etniczne** oraz **kod pocztowy** (Chen 2018). Zmienna **pochodzenie etniczne** jest zmienną prawnie chronioną (nie tylko w polskim i europejskim ustawodawstwie, ale również np. w krajach Ameryki Północnej), stąd banki wykluczają ją z zakresu zmiennych, które są wykorzystywane w trakcie budowy i prognozy z modelu, tj. przedstawione wcześniej podejście sprawiedliwości przez nieświadomość. Wykluczając wyłącznie zmienną prawnie chronioną – **pochodzenie etniczne** – a pozostawiając inne zmienne (np. **kod pocztowy**) bank naraża się na powstanie problemu zmiennej pominiętej: często pod danym kodem pocztowym zamieszkuje jeden typ grupy etnicznej. Wykluczając zmienną **pochodzenie etniczne**, nowo oszacowany model będzie wciąż kodował zmienność tej zmiennej chronionej w pozostałych zmiennych modelu, np. w zmiennej **kod pocztowy**.

Miary stronniczości algorytmicznej

W kontekście modeli oceny zdolności kredytowej, badając jakość prognostyczną modelu wykorzystuje się *macierz pomyłek* (więcej o ocenie modeli oceny zdolności kredytowej w podrozdziale 2.2). Standardowa macierz pomyłek w wierszach posiada realizację zmiennej celu (tj. informację 0 – 1 o braku lub spłacie kredytu), a w kolumnach prognozę wartości zmiennej celu (tj. prognozę wartości 0 – 1)⁹. Do wyznaczenia miar stronniczości algorytmicznej wykorzystuje się ogólne podejście *Fair Aware Confussion Table* – *FACT*, w ramach którego dla grup uprzywilejowanych i nieuprzywilejowanych (różnicowanych przez wartość atrybutu prawnie chronionego) definiuje się osobne macierze pomyłek, a wartości miar stronniczości algorytmicznej wyznaczane są jako funkcje poszczególnych elementów tych macierzy pomyłek (Kim, Chen i Talwalkar 2020).

Jednym ze sposobów identyfikacji występowania stronniczości algorytmicznej związanej z definicją dyskryminacji pośredniej jest tzw. *warunek parytetu demograficznego* (ang. *demographic parity*, *statistical parity* lub *independence*). Warunek parytetu demograficznego związany jest z *kryterium niezależności*, tj. założeniem, że poszczególne kategorie zmiennej chronionej nie powinny różnicować częstotliwości występowania poszczególnych prognoz z modelu (pozytywna lub negatywna decyzja kredytowa). Spełnienie warunku parytetu demograficznego oznacza równą częstotliwość uzyskiwania pozytywnej decyzji kredytowej w każdej z grup cechy prawnie chronionej.

⁹ Konwencję zapisu macierzy pomyłek przyjęto za Fawcett (2006). W literaturze macierz pomyłek definiowana jest również jako transpozycja przyjętej macierzy, bez utraty poprawności dalszego wywodu.

Innym sposobem identyfikacji stronniczości algorytmicznej związanej z definicją dyskryminacji bezpośredniej (odmienne traktowanie) jest *warunek wyrównanych szans* (ang. equalized odds condition lub separation). Warunek wyrównanych szans odwołuje się do tzw. *kryterium separacji*, tj. założenia że zmienna celu (tj. informacja o tym czy klient spłacił zobowiązanie, czy nie)¹⁰ stanowi jeden z możliwych podziałów (separacji) populacji na warstwy o równych prawdopodobieństwach uzyskania akceptacji wniosku kredytowego, bez względu na wartość zmiennej prawnie chronionej.

Ostatnim z omawianych sposobów identyfikacji stronniczości algorytmicznej, odnoszącym się również do definicji dyskryminacji bezpośredniej (odmienne traktowanie), jest *warunek parytetu kalibracji* (ang. sufficiency). Warunek parytetu kalibracji, nazywany również kryterium dostateczności, jest związany z cechą modelu oceny zdolności kredytowej zakładającą, że model ten posiada taką samą *jakość kalibracji*¹¹ dla każdej z grup definiowanych przez cechę prawnie chronioną (Barocas, Hardt i Narayanan 2023). W wersji słabszej warunku parytetu kalibracji przyjmuje się równość *wartość predykcyjna dodatnia* (ang. positive predictive value, PPV).

Przypisanie prawnych definicji dyskryminacji do poszczególnych miar stronniczości algorytmicznej, tj. w sytuacji, gdy dana definicja dyskryminacji związana jest z działaniem modelu decyzyjnego, zaprezentowano w tabeli 2.

Tabela 2: Mapowanie definicji dyskryminacji na miary stronniczości algorytmicznej

Prawna definicja dyskryminacji	Miara stronniczości algorytmicznej
dyskryminacja pośrednia	warunek parytetu demograficznego
dyskryminacja bezpośrednia	warunek wyrównanych szans warunek parytetu kalibracji

Uwaga: Przypisanie prawnych definicji dyskryminacji ustawa 2010/12/03) do miar stronniczości algorytmicznej.

Źródło: opracowanie własne.

Współistnienie kilku sposobów pomiaru występowania problemu stronniczości algorytmicznej, związanych z dyskryminacją pośrednią lub bezpośrednią, stanowi

¹⁰ Z racji na fakt, że informacja ta jest znana wyłącznie dla historycznie udzielonych kredytów, technicznie kryterium separacji ewaluowane może być na podstawie znanych wartości zmiennej celu na zbiorze testowym w ramach prognozy ex post.

¹¹ Jakość kalibracji modelu klasyfikacyjnego opisuje poprawność oszacowań prawdopodobieństwa przynależności do poszczególnych klas zmiennej celu. W kontekście oceny zdolności kredytowej jest to oszacowanie prawdopodobieństwa, że klient spłaci zobowiązanie kredytowe. *Jakość kalibracji* jest pojęciem węższym od pojęcia *jakości prognostycznej* modelu, które dotyczy *mocy dyskryminacyjnej* modelu, tj. możliwości odróżniania klientów spłacających kredyt, od tych, którzy kredytu nie spłacą (więcej o jakości kalibracji w podrozdziale 2.2.2). Możliwym jest, że model posiada wysoką moc dyskryminacyjną – tj. opierając się na prognozach modelu możliwym jest odróżnianie klientów spłacających kredyty – przy jednocześnie niskiej jakości kalibracji (tj. model niepoprawnie przypisuje prawdopodobieństwa spłaty kredytów).

perspektywę decyzyjną regulatora rynku kredytów konsumenckich w zakresie tworzenia przepisów równościowych w obszarze modeli oceny zdolności kredytowej konsumentów: wybór definicji i miary stronniczości algorytmicznej.

Sprawiedliwość przez świadomość

Mając na względzie wcześniej wskazane wady podejścia sprawiedliwości przez nieświadomość rozpoczęto prowadzenie badań w zakresie innych, alternatywnych rozwiązań wykorzystywanych do redukcji problemu stronniczości algorytmicznej (Kozodoi i Lessmann 2022; Hurlin, Pérignon i Saurin 2022). Metodami porównawczymi, skrajnymi względem siebie są opisane wcześniej podejście sprawiedliwości przez nieświadomość (tj. całkowite odrzucenie informacji o zmiennej prawnie chronionej) lub podejście nieograniczone (tj. pełne włączenie zmiennej chronionej do budowanego modelu oceny zdolności kredytowej). Pożądanymi cechami konstruowanych metod redukcji stronniczości algorytmicznej są stosunkowo niewielka (idealnie zerowa) utrata jakości prognostycznej modelu oraz możliwość redukcji stronniczości algorytmicznej.

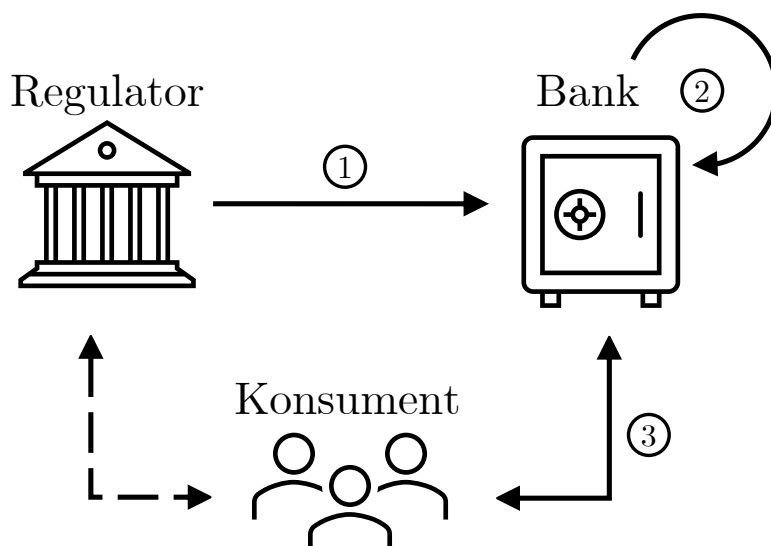
Z perspektywy zdefiniowanej regulacji równościowej oraz związanej z nią miary stronniczości algorytmicznej, bank stoi przed decyzją w zakresie wyboru optymalnej metody redukcji stronniczości algorytmicznej, w której kryterium optymalizacji jest zysk banku. W literaturze przedmiotu wyróżniono trzy klasy podejść związanych z redukcją stronniczości algorytmicznej: w zależności od umiejscowienia w procesie budowy modelu, w którym dana metoda jest wykorzystywana. Metody te dzieli się na:

- *metody przedtreningowe* (ang. pre-training), czyli metody redukcji stronniczości algorytmicznej związane ze wstępnym etapem przetwarzania danych zbioru uczącego; motywacją jest wskazanie, że przyczyną dyskryminacji są dane zbioru treningowego (Friedler i in. 2019);
- *metody treningowe* (ang. in-processing), czyli metody redukcji stronniczości algorytmicznej związane z modyfikacją algorytmu uczącego, np. przez uwzględnienie dodatkowych ograniczeń w procesie estymacji parametrów; obecnie najbardziej popularne podejście do zapewnienia sprawiedliwości (Friedler i in. 2019);
- *metody potreningowe* (ang. post-processing) związane z wprowadzeniem sprawiedliwości przez modyfikację wyników działania modelu (Friedler i in. 2019); jedna z metod potreningowych modyfikuje rozkład prognoz punktowych modelu oceny zdolności kredytowej (Kozodoi i Lessmann 2022).

Wybór optymalnego podejścia do redukcji problemu stronniczości algorytmicznej, pod warunkiem zadanych regulacji równościowych, jest *problemem decyzyjnym banku*.

Wpływ stronniczości algorytmicznej na ocenę zdolności kredytowej

Wyróżnić można co najmniej trzy obszary wpływu problemu stronniczości algorytmicznej na ocenę zdolności kredytowej. Zależności między wskazanymi obszarami przedstawiono na rysunku 2.



Rysunek 2: Relacje między regulatorem, bankiem i konsumentem

Uwaga: Strzałka 1 oznacza regulacje równościowe tworzone przez regulatora, będące wymaganiami nakładanymi na banki. Strzałka 2 oznacza proces tworzenia i wdrażania modelu oceny zdolności kredytowej przez bank zgodnie z obowiązującymi regulacjami. Strzałka 3 oznacza oddziaływanie modelu oceny zdolności kredytowej na konsumentów (tj. ocena aplikacji kredytowych). Strzałka przerywana oznacza cel działania regulatora, tj. ochrona konsumentów przez tworzone regulacje.

Źródło: opracowanie własne.

Działanie podmiotu regulującego rynek kredytów konsumenckich jest związane ze sprawowaniem ochrony konsumenta na rynku usług finansowych (strzałka przerywana na rysunku 2). *Perspektywa decyzyjna regulatora* jest związana z decydem projektującym regulacje równościowe, tj. wymagania nakładane na banki, obejmujące modele oceny zdolności kredytowej (strzałka 1 na rysunku 2). Proces analizy decyzji w tym zakresie dotyczy wpływu projektowanej regulacji (np. definicja miary stronniczości algorytmicznej, przedział wartości dopuszczalnych tej miary, okres przejściowy wdrożenia regulacji, inne wytyczne w zakresie wykorzystania zmiennej chronionej) na populację kredytobiorców. Regulator wprowadzając regulacje równościowe dąży do zrozumienia konsekwencji dla społeczeństwa i dla systemu bankowego projektowanej regulacji.

Podmiotem bezpośrednio odpowiedzialnym za wdrożenie regulacji tworzonych przez regulatora rynku jest bank, którego obowiązkiem jest wdrożenie regulacji równościowych w ramach wykorzystywanego modelu oceny zdolności kredytowej (strzałka 2 na rysunku 2). *Perspektywa decyzyjna banku* jest związana z analizą wpływu przyjętej i obowiązującej regulacji równościowej na rentowność procesu kredytowego, w tym poziom współczynnika zatwierdzeń kredytów, jakość portfela kredytowego oraz poziom zysku.

Decyzją w tym zakresie będzie wybór podejścia do redukcji stronniczości algorytmicznej, które optymalnie (uwzględniając wpływ na rachunek ekonomiczny) obniży poziom występowania stronniczości algorytmicznej wyrażony miarą tej stronniczości określoną przez regulatora.

Perspektywa społecznego odbiorcy decyzji jest związana z populacją konsumentów, którzy składając aplikacje kredytowe są oceniani przez model oceny zdolności kredytowej; wraz ze zmianą wykorzystywanego modelu, np. ze względu na wprowadzenie regulacji równościowych i dostosowaniem modelu oceny zdolności kredytowej, powstają konsekwencje społeczne (strzałka 3 na rysunku 2). Perspektywa ta wiąże się również z długookresowymi konsekwencjami społecznymi, które wystąpią ze względu na wprowadzenie określonych regulacji równościowych, a które będą zaimplementowane w modelu oceny zdolności kredytowej.

Hipotezy badawcze

W rozprawie zostanie zbadanych sześć następujących hipotez badawczych:

Hipoteza 1: O nieefektywności ekonomicznej metody sprawiedliwości przez nieświadomość. Metoda redukcji stronniczości algorytmicznej: sprawiedliwość przez nieświadomość nie jest ekonomicznie efektywna. Istnieją efektywniejsze metody redukcji stronniczości algorytmicznej klasy sprawiedliwości przez świadomość, które pozwalają na redukcję stronniczości algorytmicznej przy jednoczesnej mniejszej utracie zysku z realizowanego procesu kredytowego w porównaniu do podejścia sprawiedliwości przez nieświadomość.

Hipoteza 2: O skuteczności redukcji stronniczości algorytmicznej w zależności od miary. W zależności od przyjętej miary stronniczości algorytmicznej, inne metody redukcji stronniczości algorytmicznej będą efektywne ekonomicznie. Projektując regulacje równościowe, przeciwdziałające dyskryminacji w procesie oceny zdolności kredytowej konsumentów, regulator stoi przed wyborem miary opisującej występowanie problemu stronniczości algorytmicznej. W zależności od dokonanego przez regulatora wyboru, inne metody redukcji stronniczości algorytmicznej będą efektywne ekonomicznie z perspektywy działalności kredytowej realizowanej przez bank.

Hipoteza 3: O procedurze redukcji stronniczości algorytmicznej maksymalizującej zysk banku.

Możliwym jest opracowanie procedury redukującej stronniczość algorytmiczną modelu oceny zdolności kredytowej maksymalizującej zysk banku. W zakresie podejść do redukcji stronniczości algorytmicznej (metody przedtreningowe, metody treningowe oraz metody potreningowe), możliwym jest opracowanie metodyki pomiaru efektywności stosowania metod redukcji stronniczości algorytmicznej; opierając się na tej metodyce, możliwym jest opracowanie procedury redukcji stronniczości algorytmicznej modelu oceny zdolności

kredytowej, która maksymalizuje zysk banku z realizowanego procesu kredytowego.

Hipoteza 4: O metodyce oceny długookresowych konsekwencji regulacji równościowych.

Możliwym jest opracowanie metodyki oceny długookresowych konsekwencji wprowadzania regulacji równościowych związanych z modelami oceny zdolności kredytowej przy wykorzystaniu techniki symulacyjnej. W przypadku funkcjonujących obecnie banków wykorzystywaną metodą redukcji stronniczości algorytmicznej jest sprawiedliwość przez nieświadomość. Oznacza to, że podejście do oceny konsekwencji wprowadzenia regulacji równościowej (przy zdefiniowanej mierze tej stronniczości algorytmicznej), uwzględnić powinno porównanie z metodą dotychczas wykorzystywaną; w pracy weryfikacji podlegać będzie możliwość opracowania metodyki oceny wpływu przejścia z metody sprawiedliwości przez nieświadomość na metodę sprawiedliwości przez świadomość.

Hipoteza 5: O okresie dostosowawczym występującym po wprowadzeniu regulacji.

Wprowadzenie regulacji równościowych może powodować krótkookresową niestabilność rozkładu decyzji kredytowych. Projektując i wdrażając równościowe rozwiązania regulacyjne związane z modelami oceny zdolności kredytowej, regulator rynku kredytów konsumenckich powinien uwzględnić okresową niestabilność w okresie dostosowawczym.

Hipoteza 6: O koszcie ekonomicznym regulacji równościowych.

Wprowadzenie regulacji równościowych prowadzi do kosztu ekonomicznego związanego ze zmianą populacji osób otrzymujących pozytywne decyzje kredytowe. Koszty ekonomiczne związane są z:

- jakością portfela kredytowego, tj. więcej niespłaconych kredytów;
- zmianą poziomu akcji kredytowej, tj. inny współczynnik akceptacji wniosków kredytowych;
- kombinacją zmiany jakości portfela kredytowego i poziomu akcji kredytowej (tj. dwa powyższe czynniki).

Zmiany wynikające z wprowadzenia przepisów równościowych również będą się wiązać z konsekwencjami społecznymi wynikającymi ze zmiany populacji osób otrzymujących pozytywne decyzje kredytowe.

Metody badawcze

Modele oceny zdolności kredytowej konsumentów różnią się istotnie od sposobów oceny aplikacji kredytowych przedkładanych przez klientów korporacyjnych: zakres informacji wykorzystywanych przez banki w ocenie zdolności kredytowej, sposoby uwzględniania zabezpieczeń w kredytach bądź rodzaje dostępnych kredytów, w przypadku konsumenta oraz klienta korporacyjnego są istotnie różne (Zawadzka 1996). Uwzględniając fakt, że

problem badawczy rozprawy dotyczy konsumentów, w dalszej części będzie rozpatrywana ocena zdolności kredytowej konsumentów¹².

W rozprawie wykorzystano krytyczny przegląd literatury oraz podejście symulacyjne jako główne metody badawcze. Celem przeprowadzenia krytycznego przeglądu literatury było:

- opracowanie specyfikacji danych syntetycznych wykorzystanych w eksperymentach numerycznych;
- wybór architektury modeli klasyfikacyjnych, na podstawie których zostaną opracowane modele oceny zdolności kredytowej;
- specyfikacja miar stronniczości algorytmicznej;
- wybór metod redukcji stronniczości algorytmicznej, które będą w dalszej kolejności wykorzystane w eksperymentach numerycznych.

Na podstawie przeprowadzonego krytycznego przeglądu literatury przygotowano symulacje polegające na implementacji modeli oceny zdolności kredytowej wraz z wybranymi metodami redukcji stronniczości algorytmicznej, przygotowaniu danych (syntetycznych oraz rzeczywistych), na podstawie których zrealizowano symulacje Monte Carlo oraz utworzeniu modeli oceny skutków wprowadzania regulacji równościowych mierzących wpływ w ujęciu *statycznym* (model statyczny jest opisany w rozdziale 4) oraz *dynamicznym* (model dynamiczny jest opisany w rozdziale 5).

Celem utworzenia *modelu statycznego* była ocena jakości działania modelu oceny zdolności kredytowej przy wykorzystaniu różnych sposobów redukcji stronniczości algorytmicznej, np. podejście nieograniczone (wykorzystujące zmienną prawnie chronioną), podejście sprawiedliwości przez nieświadomość (wyłączenie zmiennej chronionej ze zbioru uczącego), modele przedtreningowej, treningowej oraz potreningowej redukcji stronniczości algorytmicznej. Model analizy statycznej był inspirowany podejściem przedstawionym przez Kozodoi i Lessmann (2022), w którym poszczególne podejścia redukcji stronniczości algorytmicznej zostały zweryfikowane na podstawie pełnego zbioru danych, tj. z wyłączeniem problemu braku możliwości obserwacji zmiennej celu dla kredytów odrzuconych. Podejście przedstawione w literaturze zostało rozwinięte zarówno w zakresie oszacowania przedziałów ufności uzyskiwanych wyników, jak również przez opracowanie drugiego modelu, tzw. modelu dynamicznego.

Celem utworzenia *modelu dynamicznego* była analiza długoterminowych konsekwencji wprowadzania regulacji równościowych. Podejścia prezentowane w literaturze przedmiotu przez Kozodoi i Lessmann (2022) przedstawiają poziom zysku generowanego przez model

¹² Ze względu na podobne potrzeby kredytowe, jak również wykorzystywane produkty kredytowe, praktyką w wielu bankach jest traktowanie jednoosobowej działalności gospodarczej lub małych firm jako konsumenta.

oceny zdolności kredytowej z perspektywy banku, przy różnych podejściach do redukcji stronniczości algorytmicznej. Model dynamiczny umożliwi odtworzenie: procesu akceptacji wniosków kredytowych, obserwacji spłacalności kredytów u kohort klientów, którym udzielono kredytu oraz budowy kolejnych instancji modelu oceny zdolności kredytowej, opartego na danych klientów, którzy wcześniej uzyskali pozytywną decyzję kredytową (tj. bank nie obserwuje zachowania klientów, którzy nie otrzymali kredytu). Model o takiej konstrukcji umożliwi ocenę konsekwencji ekonomicznych i społecznych zastosowania podejść do redukcji stronniczości algorytmiczną modeli oceny zdolności kredytowej konsumenta.

Symulacje Monte Carlo przeprowadzono, opierając się na modelu statycznym i modelu dynamicznym, w celu uzyskania oszacowanie średniej i jej przedziałów ufności. W rozprawie posłużono się zarówno danymi syntetycznymi, jak i danymi rzeczywistymi. Wykorzystanie danych syntetycznych w ocenie procedur związanych z tworzeniem modeli oceny zdolności kredytowej jest inspirowane istniejącą literaturą przedmiotu: Przanowski (2014); Kennedy, Delany i Namee (2011); za zasadnością rozszerzenia analiz prowadzonych na podstawie danych rzeczywistych o dane syntetyczne wskazują: brak historycznych obserwacji bankowych (np. pandemia Covid-19), powolna dynamika rzeczywistych procesów bankowych lub brak publicznie dostępnych danych kredytowych.

Struktura rozprawy

Rozprawa składa się ze wstępu, 5 rozdziałów, podsumowania i aneksu zawierającego specyfikację danych wykorzystanych w modelu statycznym oraz dynamicznym.

W rozdziale 1 wprowadzono podstawowe pojęcia wykorzystywane w rozprawie: kredyt oraz proces kredytowy, które przedstawiono w kontekście zmian historycznych, wskazując na ewolucję procedur i zasad związanych z udzielaniem kredytów konsumentom. Przedstawiono również przebieg procesu kredytowego wraz z danymi wykorzystywanymi przez bank w ramach tego procesu.

W rozdziale 2 opisano ewolucję historyczną i obecnie wykorzystywane praktyki banków w zakresie oceny zdolności kredytowej, tj. przejście z oceny eksperckiej na modele oceny zdolności kredytowej. W zakresie podejść współcześnie wykorzystywanych, tj. modeli oceny zdolności kredytowej, zaprezentowano podstawowe podejścia do oceny jakości działania takich modeli, tj. ocenę mocy dyskryminacyjnej oraz ocenę jakości kalibracji. Rozdział zakończono uzasadnieniem istotności poprawnego działania modeli oceny zdolności kredytowej, wskazując przy tym zarówno na aspekty makroekonomiczne jak również mikroekonomiczne: zysk banku oraz poziom ryzyka udzielonych kredytów.

Rozdział 3 poświęcono problematyce stronniczości algorytmicznej. Uwzględniono analizę prawnych wytycznych dotyczących modeli decyzyjnych, jak również aspektów stronniczości algorytmicznej w modelach oceny zdolności kredytowej. W podrozdziale 3.1

omówiono pojęcie stronniczości algorytmicznej, wskazując na odmienny wpływ modelu (dyskryminacja pośrednia) lub odmienne traktowanie przez model decyzyjny (dyskryminacja bezpośrednia); uwagę skierowano również na główne źródła skrzywienia algorytmicznego oraz nawiązano do historycznych teorii ekonomicznych traktujących o sprawiedliwości podejmowania decyzji. W podrozdziale 3.2 podano współcześnie funkcjonujące przepisy prawne zabraniające dyskryminacji, w tym dyskryminacji w modelach oceny zdolności kredytowej. W podrozdziale 3.3 ujęto sposoby identyfikacji stronniczości algorytmicznej w modelach decyzyjnych. W podrozdziale 3.4 przedstawiono podejścia do redukcji występującego problemu stronniczości algorytmicznej.

W rozdziale 4 przedstawiono autorski model statyczny umożliwiający badanie wpływu zastosowania poszczególnych metod redukcji stronniczości algorytmicznej na wynik ekonomiczny realizowanego procesu kredytowego. Przeanalizowano ten wpływ opierając się na danych symulacyjnych i danych rzeczywistych. Celem zastosowania danych symulacyjnych była duża dowolność parametryzacji procesu generującego dane, np. uwzględnienie współliniowości zmiennych oraz dokładna znajomość prawdopodobieństwa spłat kredytów.

W rozdziale 5 skupiono uwagę na autorskim modelu dynamicznym umożliwiającym odtworzenie procesu kredytowego w horyzoncie wielu okresów. W modelu dynamicznym przyjmuje się, że populacja kredytobiorców, która otrzymała kredyt, będzie w kolejnych iteracjach symulacji wykorzystywana do trenowania następnych instancji modelu oceny zdolności kredytowej. Wykorzystując model dynamiczny możliwym jest obserwowanie zmienności w czasie średnich wartości atrybutów klientów, którzy otrzymują pozytywne decyzje kredytowe. Celem zastosowania takiego mechanizmu symulacyjnego jest umożliwienie zbadania wpływu zastosowania metody redukcji stronniczości algorytmicznej, przy założeniu, że bank zachowuje ciągłość działania, tj. przechodzi z dotychczas wykorzystywanej metody na metodę redukującą stronniczość algorytmiczną.

Rozdział 1

Kredyt i proces kredytowy

Celem niniejszego rozdziału jest omówienie pojęć kredytu oraz procesu kredytowego. Dodatkowo w ramach rozdziału przedstawiono rodzaje danych wykorzystywanych przez banki w ramach procesu kredytowego. W celu prezentacji genezy pojęcia procesu kredytowego, z uwzględnieniem wewnętrznych praktyk bankowych i zewnętrznych uwarunkowań wymogów regulacyjnych, pojęcie kredytu osadzono w kontekście historycznym.

Mając na względzie wskazany cel, przyjęto następującą strukturę dalszej części rozdziału: w podrozdziale 1.1 przedstawiono historyczny zarys instytucji kredytu; w podrozdziale 1.2. zaprezentowano ramy procesu kredytowego, wraz z oceną zdolności kredytowej; w podrozdziale 1.3 wskazano na źródła danych, które są wykorzystywane w procesie kredytowym do oceny zdolności kredytowej. Wskazane pojęcia kredytu, procesu kredytowego oraz oceny zdolności kredytowej będą dalej wykorzystywane w kolejnych rozdziałach rozprawy.

1.1 Historia kredytu

Pod pojęciem *kredytu* współcześnie rozumie się kwotę pieniężną, która została dostarczona przez bank na określony czas i zgodnie z warunkami zdefiniowanymi w pisemnej umowie – tzw. *umowie kredytowej* (Iwanicz-Drozdowska i in. 2017)¹³. Kredyt ma na celu zaspokojenie potrzeb finansowych kredytobiorcy, np. zakup nieruchomości (tzw. *kredyt hipoteczny*), wydatki bieżące (tzw. *kredyt konsumpcyjny*) lub

¹³ Istotnymi cechami kredytu są: ściśle pieniężny charakter oraz to, że kredyty mogą być udzielane wyłącznie przez banki. Instytucja kredytu w Polsce jest regulowana przez ustawę Prawo Bankowe. Należy jednocześnie wskazać, że pojęcie *pożyczki* jest pojęciem różnym względem kredytu: pożyczką jest czynność związana z udzieleniem przez jedną ze stron (osoba fizyczna lub instytucja) określonej sumy pieniężnej lub określonego przedmiotu pożyczkobiorcy. Istotnym warunkiem jest jednak fakt, że przedmiot pożyczki jest własnością pożyczkodawcy (Wójcicka 2015a). Instytucję pożyczki reguluje ustawa 1964/04/23, która precyzuje przedmiot pożyczki następująco: „Przez umowę pożyczki dający pożyczkę zobowiązuje się przenieść na własność biorącego określoną ilość pieniędzy albo rzeczy oznaczonych tylko co do gatunku, a biorący zobowiązuje się zwrócić tę samą ilość”.

konsolidacja bieżących kredytów (tzw. *kredyt konsolidacyjny*). Współcześnie udzielanie kredytów klientom indywidualnym i niedetalicznym leży u podstaw funkcjonowania banków uniwersalnych (Baesens, Roesch i Scheule 2016; Zawadzka 1996).

Słowo kredyt historycznie pochodzi od łacińskiego *creditum* ‘rzecz powierzona komuś innemu, zaufanie do kogoś’. Ludzkość od zarania wykorzystywała instytucję kredytu jako sposobu finansowania wydatków konsumpcyjnych lub inwestycji (Online Etymology Dictionary 2022). Sposób udzielania kredytów, łącznie z warunkami umowy kredytowej regulowanej prawnie lub społecznie, ulegały na przestrzeni wieków przekształceniom i zmianom. Poniżej przedstawiono zestawienie wybranych okresów historycznych wskazujących na stan zaawansowania instytucji kredytu.

1.1.1 Starożytność, Sumerowie, 3500 lat p.n.e.

Pierwsze odnotowane doniesienia o wykorzystaniu kredytu w celach finansowania przedsięwzięć agrarnych (w szczególności zakupu nasion) lub finansowanie wypraw eksploracyjnych, pochodzą z okresów cywilizacji sumeryjskiej zamieszkującej ówczesny obszar południowej Mezopotamii¹⁴ (Goetzmann 2017). Historyczne doniesienia wskazują, że poziom odsetek od takich kredytów sięgał około 1,5–2,0% w skali miesiąca: koszt kredytu był wyznaczany na poziomie ok. 1 shekel (tj. 8,41 g srebra) za każdą minę (tj. 505 g srebra) wartości pożyczonego kapitału.

Władcy sumeryjscy okresowo wydawali tzw. *dekrety czystej karty* (ang. *clean slate decrees*), które znosiły kredyty agrarne, tym samym uwalniając kredytobiorców od konieczności spłaty zobowiązań. Takie działania były prowadzone najczęściej w pierwszych latach panowania lub w czasach konfliktów zbrojnych lub nieurodzaju. Dodatkowo celem tych dekretów było zapobieganie narastaniu długów do stopnia, w którym mogłyby one zagrozić bezpieczeństwu państwa, jeśli chłopci z powodu długów straciliby swoją ziemię (Hudson i Mierop 2002).

1.1.2 Starożytność, Babilon, 1800 lat p.n.e.

Nowe państwo Babilon, powstałe w wyniku napływu i podboju ludności Amorytów, rozbudowało instytucję kredytu o pierwsze spisane prawo obejmujące zasady udzielania kredytów. Prawo to obecnie nazwane jest *prawem Hammurabiego*.

Zapisy prawa Hammurabiego wskazywały na maksymalne dopuszczalne odsetki od udzielanych kredytów: do 33,3% rocznie za kredyty na ziarno lub 20,0% na kredyty udzielane na zakup srebra do celów handlowych (Desjardins 2018). Obok powyższych zapisów, które wyznaczały techniczne normy dopuszczalnego oprocentowania kredytów,

¹⁴ Obecnie Irak; cywilizacja ta zajmowała tereny między rzekami Tygrysem i Eufratem, które cyklicznie wylewając, nawadniały i użyźniały glebę, co w efekcie prowadziło do wysoko efektywnego rolnictwa, które sprzyjało rozwojowi instytucji kredytu.

prawo to obejmowało również pierwsze zasady *ładu procesu udzielania kredytów*: kredyty musiały być udzielane przy urzędnikach publicznych i spisywane w formie kontraktów. Wskazując przykład cywilizacji babilońskiej można uwypuklić współlistnienie dwóch wymiarów opisujących uwarunkowania udzielania kredytów: *ilościowe*, m.in. oprocentowanie, horyzont czasu pożyczki, liczba rat w roku oraz *jakościowe*, m.in. sposób spisania umowy, kluczowe punkty umowne, wymogi poświadczeń urzędniczych lub uwarunkowania udzielania kredytów (Hermann i Johns 1910)¹⁵.

1.1.3 Starożytność, Grecja, 1500 lat p.n.e.

Mieszkańcy starożytnej Grecji dokonywali transakcji finansowych wykorzystując do tego instytucje tworzone wokół greckich świątyń, co umożliwiało transfer środków bez konieczności ich fizycznego transportowania i składowania. Doniesienia historyczne wskazują, że środki deponowane w tych świątyniach mogły być wydatkowane (np. na zakup nasion) bez konieczności fizycznego wykorzystywania monet. Świątynie Artemidy, Hery oraz Apolla, oprócz pełnienia funkcji miejsc sakralnych świadczyły usługi deponowania środków finansowych oraz innych bardziej złożonych usług finansowych, w tym wymianę obcych walut, weryfikację autentyczności monet, ale również udzielanie pożyczek (Cohen 2011).

1.1.4 Starożytność, Cesarstwo Rzymskie, 50 lat p.n.e.

Ekspansja obszarowa Republiki Rzymskiej przyniosła równoczesne innowacje w obszarze udzielania kredytów oraz transferów pieniężnych. Przesył dużych sum pieniężnych między poszczególnymi dominiami Republiki Rzymu był możliwy wyłącznie dzięki współlistniającym siatkom instytucji publicznych, kredytowych. W momencie ekspansji Rzymu na tereny Grecji, Północnej Afryki, Azji i dzisiejszej Hiszpanii, logistyka transferu pieniędzy pomiędzy poszczególnymi obszarami (m.in. w celu finansowania przedsięwzięć na nowo zdobytych obszarach) wymagała istotnych innowacji¹⁶. Kredytowanie i finansowanie daleko położonych obszarów, z którym wiązał się przesył pieniędzy był możliwy za sprawą tzw. *permutatio*, tj. transferu środków za pomocą *transakcji papierowej* (Harris 2008; Barlow 1978). Nowo zdobyte terytoria często podlegały opodatkowaniu zbieranemu przez prywatne podmioty, niejednokrotnie

¹⁵ „Ich akt zgody został spisany w świątyni przez notariusza i potwierdzony przysięgą – na boga i króla.” [tł. wł.]

¹⁶ Dla przykładu, zapiski historyczne wskazują na transakcję zakupu nieruchomości przez Cyncerona na wzgórzu Palatyn za kwotę 3,5 miliona srebrnych sestercjuszów (srebrna moneta o wadze 1,13 g); zapłata takiej kwoty w fizycznych momentach oznaczałaby konieczność przewiezienia prawie 4 ton monet; wskazana transakcja zakupu nieruchomości nie była największą w tym okresie (Harris 2008). Z technicznego punktu widzenia, zakup tej nieruchomości wiązałoby się z dużymi kosztami transakcyjnymi (m.in. zgromadzenia takiej sumy monet, pakowanie monet, przewóz i ochrona transportu) zarówno dla kupującego nieruchomość, jak również dla sprzedającego.

poszczególne osoby. Jednostki te mogły kumulować znaczne środki finansowe pochodzące z opodatkowania; za sprawą istnienia rozbudowanej sieci filii takich podmiotów w różnych dominium, jak również w samym Rzymie, podmioty zbierające podatek mogły świadczyć usługi przesyłania środków finansowych. Transakcje papierowe odbywały się poprzez wpłatę określonych środków np. w filii w Rzymie z możliwością rozdysponowania tych środków w odległym dominium Republiki Rzymskiej (Desjardins 2018; Barlow 1978)¹⁷.

1.1.5 Średniowiecze, Europa, 800 rok n.e.

W okresie po upadku Zachodniego Cesarstwa Rzymskiego Kościół chrześcijański zabronił praktykowania *lichwy*, czyli pobierania oprocentowania za pożyczane środki. Pod rządami Karola Wielkiego (768–814 lata n.e.) chrześcijanie nie mogli naliczać odsetek od udzielanych pożyczek. Zakazy te spowolniły, lecz nie zatrzymały rozwoju gospodarczego i prowadzonych inwestycji. Przykład wskazanego zakazu z późniejszego okresu można odnaleźć w Sumie Teologicznej św. Tomasza z Akwinu „Pożyczający pieniądze przenosi posiadanie ich na tego, kto od niego pożycza; ten przeto, któremu pieniądze pożyczono, posiada je na własne ryzyko i obowiązany jest oddać je w całości: stąd wniosek, że nie wolno wierzycielowi żądać więcej, niż pożyczył.” (Suma Teologiczna, II, II, q. 78, a. 1.)¹⁸.

Historyczne doniesienia wskazują, że pod koniec średniowiecza naliczanie oprocentowania za udzielanie kredytów stało się akceptowalne, a lichwą zaczęto określać wyłącznie kredyty obarczone wyższym niż średnio stosowane poziomy oprocentowania. Społeczeństwo korzystało z instytucji kredytu, pożyczając pieniądze od wyznawców innych religii. Chrześcijanie mogli konkurować z wyznawcami innych wiar w nisko oprocentowanym pożyczaniu pieniędzy; jednocześnie z racji na chrześcijański zakaz lichwy, wyznawcy innych wiar uzyskali monopol na pożyczki wysokiego ryzyka, które wymagały wysokich oprocentowań je rekompensujących (Wilson 2016).

Kończący się okres średniowiecza pozostawił wiele dużych domów bankowych, np. bankowy ród Medyceuszy lub działający od 1472 r. do czasów współczesnych Banca Monte dei Paschi di Siena.

¹⁷ Podobne systemy, tj. nieformalne systemy transferu wartości znane są również współcześnie jako *bankowość podziemna* lub *bankowość równoległa* i są spotykane np. w działalności siatek terrorystycznych lub w krajach o niesko rozwiniętym sektorze bankowym (Filipkowski 2006).

¹⁸ Św. Tomasz wyróżniał dwa rodzaje pożyczek (Bednarski 1970): pożyczanie trwałych dóbr gospodarczych (np. ziemia, budynki, naczynia) oraz pożyczanie dóbr, które ulegają zniszczeniu wraz z użyciem (np. pokarm, pieniądze). Pożyczka trwałych dóbr gospodarczych pozbawiała właściciela pewnych korzyści, np. zużycie przedmiotu, brak możliwości wykorzystania gospodarczego; wierzyciel ma praw, zgodnie z koncepcją św. Tomasza, żądać *wynagrodzenia* za pożyczanie trwałych dóbr gospodarczych, ponieważ ich natura, w razie pożyczania, naraża go na utratę pewnych korzyści. W przypadku pożyczania dóbr, które ulegają zniszczeniu, właściciel, jeżeli przedmiot zostanie zwrócony w takiej samej ilości i jakości, nie ponosi straty; wartość pożyczonych rzeczy pozostaje niezmienną, a więc wierzyciel nie powinien otrzymać z tego tytułu *wynagrodzenia*.

1.1.6 Nowożytność, Europa, 1500 r. n.e.

W okresie nowożytności nastąpił istotny wzrost zapotrzebowania na kapitał: kolejne misje i wyprawy morskie wymagały istotnych nakładów finansowych. W celach dywersyfikacji ryzyka ukształtowano instytucję *spółek handlowych*, które organizowały finansowanie takich wypraw. Powstanie spółek handlowych ustrukturyzowało relacje łączące wielu handlarzy: mogli oni występować o kredyt na wyprawę, występując łącznie jako spółka.

Jednym z pierwszych współczesnych państw europejskich, które ustanowiło prawne limity na poziom oprocentowania od kredytów była Anglia za czasów Henryka VIII (1509–1547 r. n.e.): wartość ta wynosiła 10% rocznie; wyższe oprocentowanie było traktowane jako lichwa (Munro 2011; Desjardins 2018). W późniejszych okresach, w celu wspierania handlu, limit oprocentowania od kredytów ulegał redukcji: do 8% w 1624 r., później do 6% w 1651 r. aż do 5% w 1713 r. Wartość limitu oprocentowania została uchylona wraz ze zniesieniem przepisów dotyczących lichwy w 1854 r. (Munro 2011).

1.1.7 Epoka XIX w., kredyt konsumencki

Dynamiczny wzrost liczby udzielanych kredytów z okresu nowożytności był kontynuowany również w XIX w. Coraz częściej pojawiały się argumenty przeciw narzucaniu odgórnych limitów na poziom odsetek od kredytów, które miały powodować hamowanie wzrostu gospodarczego (Bentham 1816). Wskazywano jednocześnie, że cele i potrzeby kredytowe są heterogeniczne: w przypadku bardziej ryzykownych przedsięwzięć, kompensacja ryzyka odbywa się za sprawą zwiększenia poziomu odsetek od kapitału.

W XIX w. powstało wiele instrumentów i procesów kredytowych, które stanowiły bazę rozwiązań instytucjonalnych, które znane są we współczesnym procesie kredytowym; w XIX w. angielskie grupy zrzeszających się handlarzy zaczęły wymieniać się informacjami o klientach, którzy mieli trudności lub zatory finansowe. Mechanizmy te stanowiły pierwowzór dzisiejszych *biur informacji kredytowej*, które dysponowały szczegółowymi informacjami o spłatach i zadłużeniach poszczególnych osób w różnych bankach. Przepływ informacji o kredytobiorcach między poszczególnymi kredytodawcami – bankami lub handlarzami oferującymi towar na termin – sprzyjał obniżaniu poziomów oprocentowań kredytów oraz dalszemu zwiększeniu *akcji kredytowej*: liczby udzielanych kredytów. Raportowanie informacji o kredytobiorcach przynosiło na tyle dobre skutki, że w 1826 r. powołano organizację The Manchester Guardian Society, publikującą co miesiąc informacje o osobach, które nie były w stanie spłacać swoich zobowiązań. Podobny charakter miały również The Mercantile Agency, która rozpoczęła swoją działalność w 1841 r. w Nowym Jorku, oraz The Retail Credit Company, otwarta w Atlancie i działająca od 1899 r. do czasów współczesnych – jedno z najstarszych biur

kredytowych na świecie (Norris 1978).

Wraz z rozwojem sektora motoryzacyjnego (przede wszystkim firma Ford), pojawiała się potrzeba na udzielanie kredytów konsumenckich dla osób, których nie było stać na zakup własnego samochodu. Popyt na kredyt był tak duży, że spółka General Motors powołała podmiot General Motors Acceptance Corporation, który oferował klientom plany finansowania samochodów, przy założeniu 35% płatności wstępnej. Kredyt konsumencki na zakup samochodu był na tyle popularny, że ok. 70% wszystkich kupowanych samochodów była kredytowana (Farber 2002).

1.1.8 Okres powojenny, Stany Zjednoczone

W latach 50 XX w. wprowadzono pierwsze *karty kredytowe*, które stały się akceptowalnym środkiem płatniczym w rosnącej liczbie podmiotów gospodarczych: początkowo były to restauracje. Firmy takie jak BankAmericard lub Mastercard rozpoczęły emisję uniwersalnych kart kredytowych, które stawały się powszechnym środkiem płatniczym (Durkin 2000).

Biura informacji kredytowej zbierały i przechowywały informacje o spłacalności kart kredytowych, które dodatkowo wzbogacano danymi pochodzącymi z zewnętrznych rejestrów: lokalnych gazet, raportów policji lub notyfikacji sądowych. Powszechnie zaczęto wykorzystywać *raporty kredytowe* przygotowywane przez biura informacji kredytowej: w latach 60 XX w. amerykańskie biura informacji kredytowej emitowały średnio ponad 50 milionów raportów kredytowych w każdym roku.

W latach 70 XX w. wprowadzono również pierwsze, obowiązujące do dziś, regulacje prawne dotyczące udzielania i obsługi kredytów. Definiują one obecnie wymogi prawne związane z udzielaniem kredytów konsumenckich. Są to:

- Pub. L. 90–284, 82 Stat. 73; 42 U.S.C. § 3601, Fair Housing Act (FHA). Ustawa mająca na celu zapobieganie dyskryminacji na rynku mieszkaniowym. Zakaz dyskryminacji obejmował cechy: rasę, kolor skóry, religię oraz narodowość. Prawo obejmowało czynności prawne związane z wynajmem i sprzedażą, udzielaniem kredytów hipotecznych oraz oceną nieruchomości.
- Pub. L. 90–321, 82 Stat. 146; 15 U.S.C. § 1601, Consumer Credit Protection Act (CCPA). Ustawa mająca na celu ochronę konsumentów przed nieuczciwymi praktykami kredytowymi. Ustawa CCPA zawiera wiele przepisów w zakresie kredytów konsumenckich:
 - Truth in Lending Act – wymogi od kredytodawców w zakresie ujawniania kluczowych warunków kredytu i kosztów konsumentom przed zawarciem umowy kredytowej;

- Fair Credit Reporting Act – wymogi w zakresie zbierania, dostępu i używania informacji kredytowych konsumentów przez biura informacji kredytowej;
 - Equal Credit Opportunity Act – wymogi zakazujące dyskryminacji podstawie rasy, koloru skóry, religii, narodowości, płci, wieku, stanu cywilnego lub z powodu korzystania z programów pomocy publicznej w procesie kredytowym;
 - Fair Debt Collection Practices Act – regulacje dot. praktyk agencji windykacyjnych w celu ochrony konsumentów przed nadużyciami.
- Pub. L. 91–508, 84 Stat. 1114; 31 U.S.C. § 5311-5330, Bank Secrecy Act (BSA). Celem ustawy BSA była walka z procederami prania pieniędzy, finansowania terroryzmu oraz innymi nielegalnymi aktywnościami finansowymi. Ustawa BSA nakłada na instytucje finansowe wymogi w zakresie monitorowania, utrzymywania historii rekordów transakcji oraz zgłaszania podejrzanych transakcji.
 - Pub. L. 91–508, 84 Stat. 1127; 15 U.S.C. § 1681, Fair Credit Reporting Act (FCRA). Ustawa jest rozszerzeniem ustawy CCPA, a jej celem jest dążenie do dokładności i prywatności danych zbieranych przez biura informacji kredytowej. Ustawa FCRA wprowadziła wymogi dla biur informacji kredytowej, użytkowników tych danych (np. banki, pracodawców) oraz zewnętrznych dostawców danych. Zapisy ustawy FCRA wskazują na wymogi w zakresie informacji osobistych i finansowych konsumentów, które mogą być wykorzystywane w procesach kredytowych; m.in. w ustawie określono zakres informacji, które mogą być gromadzone i wykorzystywane przez biura informacji kredytowych oraz prawa dostępu do informacji przez konsumentów (tj. jeden bezpłatny raport kredytowy co 12 miesięcy oraz po odrzuceniu wniosku kredytowego z powodu informacji zawartych w raporcie kredytowym).
 - Pub. L. 93–495, 88 Stat. 1500; 15 U.S.C. § 1666, Fair Credit Billing Act (FCBA). Ustawa FCBA jest rozwinięciem ustawy CCPA, w szczególności zasady Truth in Lending Act, a jej wprowadzenie miało na celu ochronę konsumentów przed nieuczciwymi praktykami rozliczeniowymi w odniesieniu do otwartych kont kredytowych, takich jak karty kredytowe i linie kredytowe. Zapisy ustawy FCBA zawierają definicję mechanizmów zapewnienia konsumentom środków do kwestionowania i rozstrzygania błędów w rozliczeniach na ich kontach kredytowych. Dodatkowo zawarto w niej zapisy wskazujące na prawa konsumentów do kwestionowania błędnych opłat, takich jak nieautoryzowane transakcje lub błędne kwoty transakcji, opłat za towary albo usługi.
 - Pub. L. 93–495, 88 Stat. 1521; 15 U.S.C. § 1691, Equal Credit Opportunity Act (ECOA). Ustawa jest rozszerzeniem ustawy CCPA w zakresie zapewnienia równości

szans w dostępie do produktów kredytowych, zakazując jednocześnie praktyk dyskryminujących na podstawie rasy, koloru skóry, pochodzenia narodowego, religii, płci, stanu cywilnego, wieku lub informacji o tym czy kredytobiorca korzystał z programów pomocy publicznej. W zapisach ustawy ECOA określono wymogi w zakresie równego traktowania konsumentów przez instytucje finansowe.

- 12 CFR Part 1002, Regulacja B, implementująca Equal Credit Opportunity Act (ECOA–B). Rozporządzenie wykonawcze ECOA–B jest implementacją przepisów określonych w ustawie ECOA i zawiera zasady, które mają na celu ochronę konsumentów przed zachowaniami dyskryminującymi w procesie kredytowym. W ramach Regulacji B określono w jaki sposób banki powinny postępować przy rozpatrywaniu wniosków kredytowych, aby zapewnić równy dostęp do kredytu.
- Pub. L. 100–430, 102 Stat. 1619; 42 U.S.C. § 3601 et seq., Fair Housing Amendments Act (FHAA). Zapisy ustawy FHAA rozszerzają ustawę FHA w zakresie zakazu dyskryminacji osób niepełnosprawnych oraz rodzin i dzieci. Dodatkowo ustawa FHAA zawiera zapisy wzmacniające mechanizmy egzekwowania prawa, np. większe uprawnienia lub bardziej rygorystyczne mechanizmy.

Decyzje kredytowe do lat 60 XX w. były podejmowane na podstawie oceny eksperckiej: analitycy kredytowi (eksperci) wykorzystując wiedzę, doświadczenie oraz znajomość osobistą klienta dokonywali takiej oceny i podejmowali decyzję o udzieleniu kredytu (Marques, García i Sánchez 2013). Ekspercka ocena zdolności kredytowej była oceną subiektywną, zależną od indywidualnej wiedzy i doświadczenia danego analityka kredytowego. Ocena taka była narażona na nadużycia związane z indywidualnymi preferencjami lub animozjami danego analityka kredytowego (Uzoamaka, Gerald i Jude 2016).

W latach 60 XX w. banki zaczęły wykorzystywać w procesie kredytowym – do wydawania decyzji kredytowych – modele oceny zdolności kredytowej (Yu i Dunn 2016; FICO 1998). Jedną z innowacji tamtego okresu, opracowaną przez spółkę analityczną Fair, Isaac and Company, był wystandaryzowany, analityczny model oceny zdolności kredytowej, tzw. FICO score.

1.1.9 Czasy współczesne, Polska

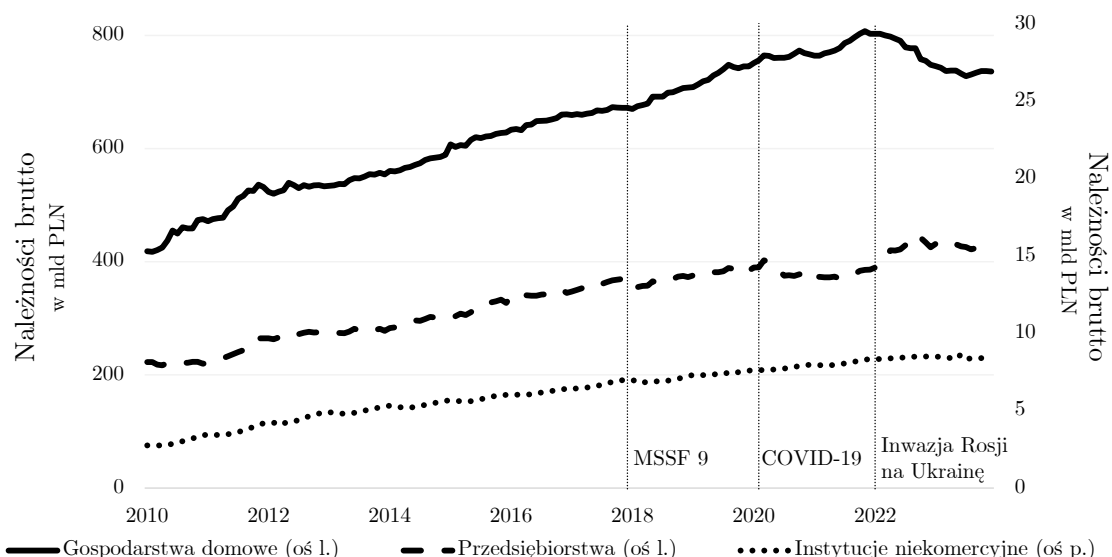
Rynek kredytów oraz proces kredytowy w Polsce jest regulowany prawnie przez następujące akty prawne:

- Ustawa z dnia 29 sierpnia 1997 r. – Prawo bankowe jest dokumentem prawnym regulującym rynek bankowy w Polsce; jej zapisy określają zasady funkcjonowania banków, w tym: tworzenia, funkcjonowania, nadzoru i likwidacji. Celem jej

wprowadzenia było zapewnienie stabilności i bezpieczeństwa sektora finansowego oraz ochrona interesów konsumentów.

- Ustawa z dnia 12 maja 2011 r. o kredycie konsumenckim (ustawa 2011/05/12) zawiera zasady udzielania kredytów konsumenckich w Polsce, w tym: warunki przy zawieraniu umów kredytowych, prawa konsumenta (odstąpienie od umowy kredytowej lub wcześniejsza spłata kredytu).
- Ustawa z dnia 21 lipca 2006 r. o nadzorze nad rynkiem finansowym (ustawa 2006/07/21) zawiera zasady sprawowania nadzoru nad rynkiem finansowym (w tym bankowym) w Polsce. Celem jej wprowadzenia było zapewnienie stabilnego i prawidłowego funkcjonowania rynku finansowego oraz przejrzystości i ochrony interesów uczestników tego rynku. Zapisy ustawy powołują Komisję Nadzoru Finansowego na organ nadzorujący rynek finansowy.
- Ustawa z dnia 23 marca 2017 r. o kredycie hipotecznym oraz o nadzorze nad pośrednikami kredytu hipotecznego i agentami (ustawa 2017/03/23) zawiera zasady prawne związane z kredytami hipotecznymi (prawa i obowiązki stron), a także zasady nadzoru nad pośrednikami i agentami oferującymi kredyty hipoteczne.
- Rekomendacja S Komisji Nadzoru Finansowego dotycząca dobrych praktyk w zakresie zarządzania ekspozycjami kredytowymi zabezpieczonymi hipotecznie zawiera zbiór dobrych praktyk w zakresie kredytów zabezpieczonych hipotecznie. Zawarto w niej zasady procesu przyznawania kredytów hipotecznych, m.in. maksymalny 25-letni okres spłacalności kredytu do wyznaczania oceny zdolności kredytowej lub konieczność zestawienia ryzyka kredytów walutowych w porównaniu z kredytami złotówkowymi.
- Ustawa z dnia 6 października 2022 r. o zmianie ustaw w celu przeciwdziałania lichwie (ustawa 2022/10/06), (w szczególności nadużyć firm pożyczkowych) zawiera zapisy regulujące rynek pożyczek konsumenckich. Określono w niej maksymalne koszty pozaodsetkowe dla pożyczek, redukcję kosztów pozaodsetkowych dla kredytów konsumenckich, sposoby zapobiegania *rolowaniu kredytów* (zaciąganiu kolejnych kredytów w celu spłacania poprzednich), konieczność każdorazowej oceny zdolności kredytowej oraz większy nadzór KNF nad instytucjami pożyczkowymi.

Rozwój instytucji kredytu, postrzegany przez pryzmat poziomu akcji kredytowej banków, również nastąpił we współczesnej Polsce. Zgodnie z danymi Komisji Nadzoru Finansowego na koniec 2023 r. gospodarstwa domowe posiadały należności na kwotę 725 mld PLN w produktach kredytach, a przedsiębiorstwa 413 mld PLN (KNF 2023). Zestawienie poziomu udzielonych kredytów przez banki w Polsce, w podziale na podmioty kredytowane, przedstawiono na rysunku 3.



Rysunek 3: Wielkość rynku kredytów sektora niefinansowego w Polsce

Uwaga: Ze względu na duże zróżnicowanie skali sumy udzielonych kredytów, oś lewa odnosi się do gospodarstw domowych oraz przedsiębiorstw, a oś prawa do instytucji niekomercyjnych. Na rysunku wyróżniono trzy momenty w czasie: 1 stycznia 2018 r. rozpoczęcie obowiązywania Międzynarodowego Standardu Sprawozdawczości Finansowej 9 (MSSF 9), 1 marca 2020 r. umowna data wybuch pandemii COVID-19 w Polsce oraz 24 lutego 2022 r. data inwazji Rosji na Ukrainę („specjalna operacja wojskowa”). Wszystkie wskazane wydarzenia miały wpływ na raportowany poziom kredytów dla przedsiębiorstw.

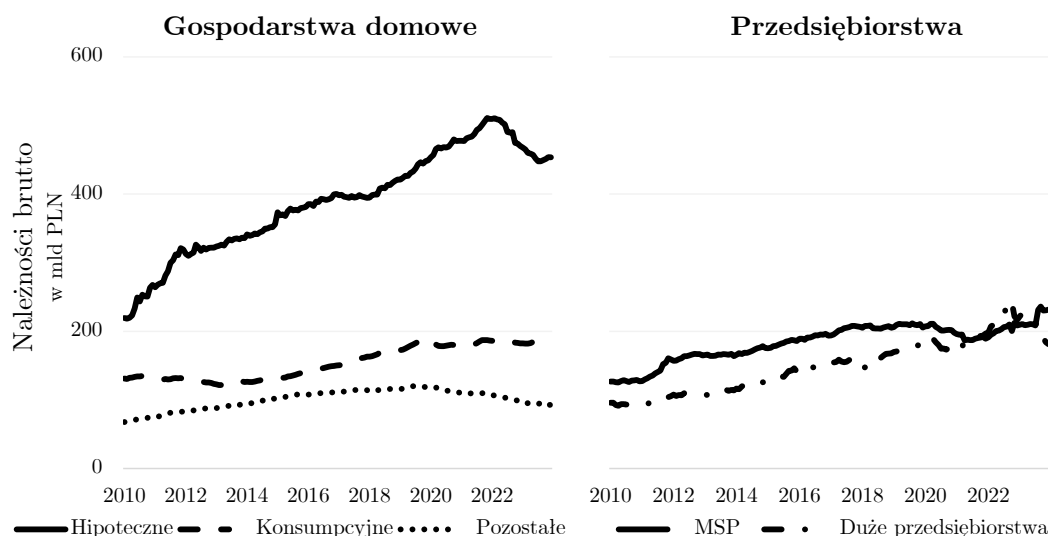
Źródło: opracowanie własne na podstawie danych KNF (2023).

Klientom indywidualnym – konsumentom – udzielane są obecnie następujące rodzaje kredytów (BIK 2021):

- *hipoteczny* – nazywany również *kredytem mieszkaniowym* jest długoterminowym kredytem udzielanym na zakup nieruchomości (tj. mieszkania, domu lub działki budowlanej) lub kredytem udzielanym na budowę; zabezpieczeniem takiego kredytu jest wpis hipoteki nieruchomości na rzecz banku do księgi wieczystej;
- *konsumpcyjny* – udzielany na bieżące wydatki, np. zakup sprzętu, remont mieszkania, zakup samochodu; dzieli się go na kredyty gotówkowe (udzielane bez zabezpieczenia na dowolny cel), kredyty w rachunku bieżącym, karty kredytowe¹⁹, kredyty ratalne, kredyty samochodowe;
- *inwestycyjny* – przeznaczony na realizację przedsięwzięcia skutkującego powiększeniem majątku kredytobiorcy (np. zakup papierów wartościowych);
- *konsolidacyjny* – zaciągany w celu spłaty innych zobowiązań finansowych, najczęściej stanowi zamianę kilku mniejszych zobowiązań finansowych na jedno o większym kapitale do spłaty i standardowo dłuższym okresie spłaty.

¹⁹ Karta kredytowa jest instrumentem płatniczym umożliwiającym zdalny dostęp do środków finansowych (KNF 2023); karta płatnicza umożliwia dokonywanie tzw. *płatności bezgotówkowych* (tj. zapłata za towary i usługi) oraz transakcji bankomatowych. Wraz z kartą kredytową jest powiązany dostępny dla klienta limit kredytowy (Wójcicka 2015b).

Głównymi typami kredytów udzielanych przez banki dla klientów indywidualnych są kredyty hipoteczne i kredyty konsumpcyjne, a dla przedsiębiorstw są to kredyty operacyjne i inwestycyjne. Poziomy udzielonych kredytów różnym typom klientów oraz linie produktowe oferowane przez banki przedstawiono na rysunku 4.



Rysunek 4: Wielkość rynku kredytów w Polsce w podziale na klienta

Uwaga: Kredyty dla klientów indywidualnych (lewy wykres) zostały zaprezentowane w rozbiciu na trzy klasy: kredyty hipoteczne, konsumpcyjne oraz pozostałe. Szczyt poziomu udzielonych kredytów w 2022 r. oraz późniejszy spadek wielkości udzielonych kredytów na hipotecznych związany był z wysokim poziomem stóp procentowych w tamtym okresie.

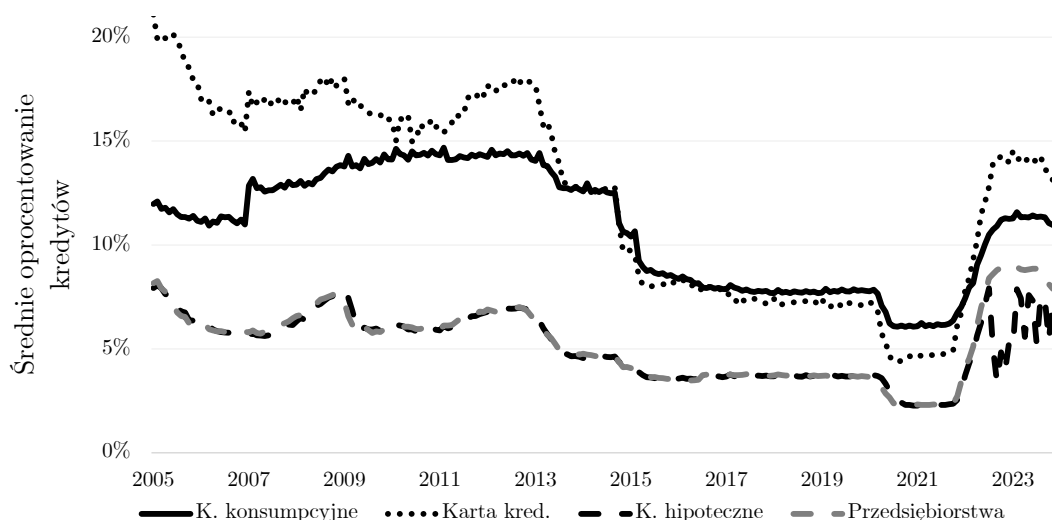
Źródło: opracowanie własne na podstawie danych KNF (2023).

Z perspektywy zarządzania ryzykiem, kredytami najbezpieczniejszymi są kredyty hipoteczne, dla których kwota jest zabezpieczona zakupowaną nieruchomością, a skłonność do spłaty zobowiązania hipotecznego jest z reguły największa (Przanowski 2014). Prawidłowość tą można zaobserwować w rzeczywistych różnicach oprocentowania kredytów hipotecznych i konsumpcyjnych. Dla zobrazowania różnic między poszczególnymi kategoriami kredytów, jak również klientów, na rysunku 5 przedstawiono zestawienie historycznych stóp oprocentowania kredytów dla gospodarstw domowych oraz przedsiębiorstw w rozbiciu na poszczególne kategorie kredytów.

Maksymalne oprocentowanie kredytu konsumpcyjnego, zgodnie z ustawą o kredycie konsumenckim z 12 maja 2011 r. nie może być większe niż dwukrotność odsetek ustawowych, tj. suma stopy referencyjnej NBP i stawki marży na poziomie 3,5% (ustawa 2011/05/12)²⁰.

Instytucja kredytu ewoluowała; zmiany te dotyczyły współistniejących dwóch wymiarów opisujących uwarunkowania udzielania kredytów: *ilościowe*, m.in. oprocentowanie, horyzont czasu pożyczki, liczba rat w roku, oraz *jakościowe*,

²⁰ Dla przykładu: 7 października 2023 r. stopa referencyjna NBP wynosi 5,75%, co daje maksymalne oprocentowanie kredytu konsumpcyjnego na poziomie $2 \times (5,75\% + 3,50\%) = 18,50\%$. Wartość ta między 2021 r. a 2023 r. uległa wzrostowi: 7 października 2021 r. stopa referencyjna NBP wynosiła 0,50%, co przekłada się na maksymalne oprocentowanie na poziomie $2 \times (0,50\% + 3,50\%) = 8,00\%$.



Rysunek 5: Średni poziom stóp procentowych kredytów w Polsce

Uwaga: Dla gospodarstw domowych średnie historyczne poziomy oprocentowanie kredytów konsumpcyjnych są na poziomach zdecydowanie wyższych niż kredytów hipotecznych – fakt ten jest bezpośrednio związany z różnicą w poziomie ryzyka, które związane jest z każdym z tych rodzajów kredytów. W przypadku przedsiębiorstwa, poziomy oprocentowania są zbliżone do kredytów hipotecznych udzielanych gospodarstwom domowym.

Źródło: opracowanie własne na podstawie danych NBP (2024).

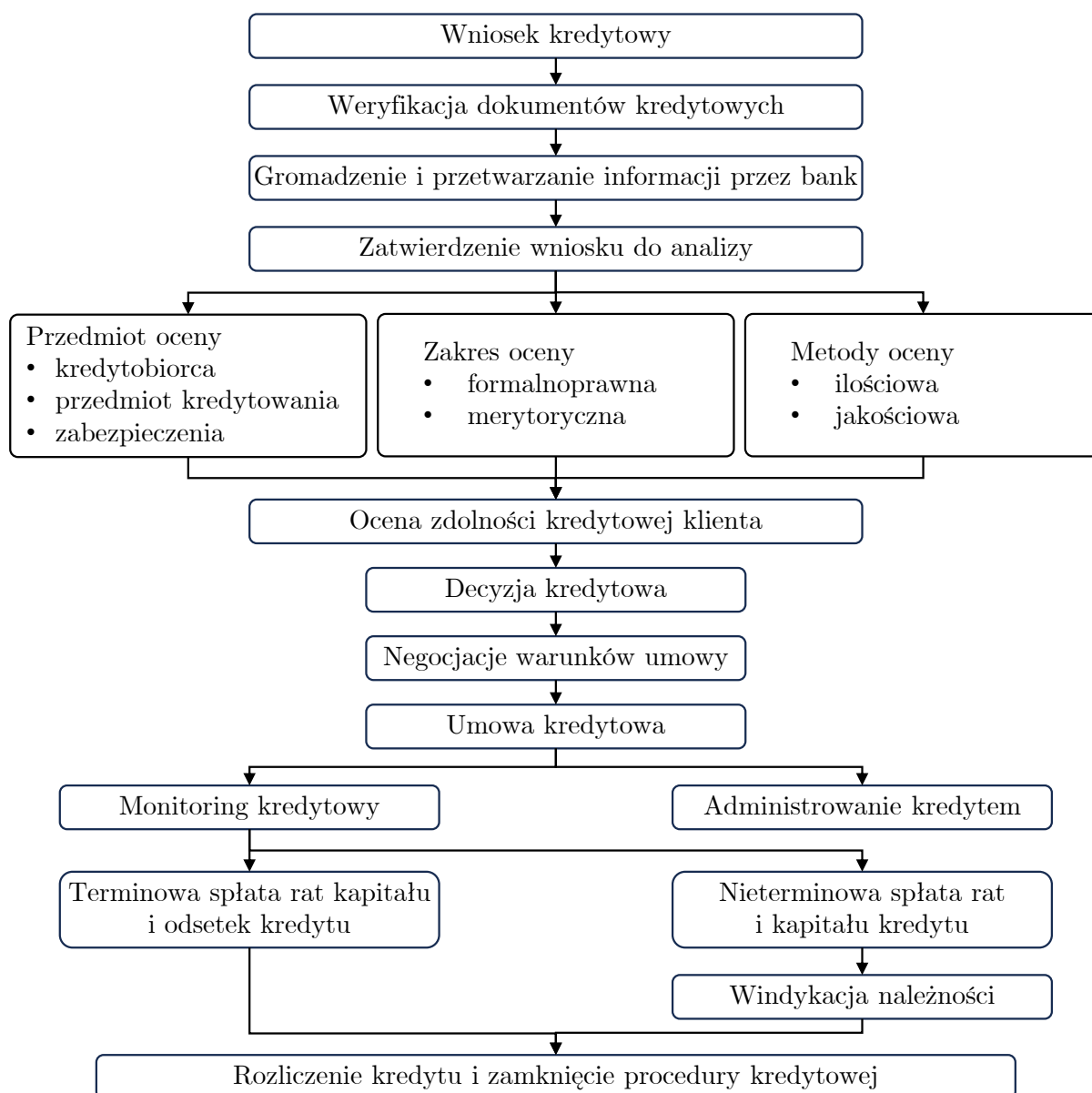
m.in. sposób spisania umowy, kluczowe punkty umowne, wymogi poświadczeń urzędniczych lub uwarunkowania udzielania kredytów. Standaryzacja i popularyzacja wśród konsumentów instytucji kredytu z czasem prowadziła do ukształtowania się *procesu* udzielania i obsługi kredytów – tzw. *procesu kredytowego*.

1.2 Proces kredytowy

Proces udzielania oraz ewentualnej późniejszej obsługi kredytu nazywa się *procesem kredytowym* (Klimontowicz 2017). Na przestrzeni lat zmieniał się i ewoluował wraz z instytucją kredytu, z procedury opartej na ręcznej pracy analityka–eksperta, do obecnie procesu podlegającego dużej automatyzacji. Główną jego funkcją jest standaryzacja obsługi kredytobiorców: od momentu pierwszej interakcji, tj. od złożenia wniosku kredytowego do zamknięcia kredytu (spłaty zobowiązania albo windykacji należności banku).

Standardowy proces kredytowy funkcjonujący w polskich bankach, przy założeniu, że jest on inicjowany przez potencjalnego kredytobiorcę przedstawiono na rysunku 6. Proces ten dotyczy aplikacji kredytobiorcy (proces kredytowy jest inicjowany aplikacją kredytową), tj. klient samodzielnie rozpoczął ścieżkę starania się o kredyt²¹.

²¹ Wyróżnia się dwa typy sprzedaży kredytów: dla nowych klientów oraz do klientów bieżących. Podejścia te różnicowane są przede wszystkim w zakresie poziomu dostępnej informacji o kliencie: w przypadku kredytów dla nowych klientów, bank dysponuje mniejszą liczbą informacji umożliwiającą ocenę zdolności kredytowej takiego klienta. W przypadku klientów bieżących, bank posiada dodatkowe informacje behawioralne danego klienta: np. poziom wpływów na rachunek ROR, jakość spłaty wcześniejszych



Rysunek 6: Etapy procesu kredytowego w banku

Źródło: opracowanie własne na podstawie Klimontowicz (2017).

Ogólne ramy etapów procesu kredytowego w banku mogą być takie same dla różnych typów klientów, tj. konsumentów oraz klientów korporacyjnych; występują jednak istotne różnice między szczegółowymi zasadami realizacji tych etapów ze względu na typ klienta. Zakres informacji wykorzystywanych przez banki, sposoby uwzględniania zabezpieczeń, bądź rodzaje dostępnych kredytów, w przypadku konsumenta oraz klienta korporacyjnego są istotnie różne (Zawadzka 1996). Przykładowo: dla ekspozycji detalicznych uwzględniane są cechy konsumenta, np. stan majątkowy, zawód, wykształcenie, uzyskiwany dochód lub stan cywilny; dla ekspozycji niedetalicznych uwzględnia się m.in. sytuację finansową spółki, strukturę kapitałową lub kwalifikację kierownictwa firmy.

zobowiązań, poziom średnich miesięcznych wydatków.

W przypadku gdy kredyt jest oferowany obecnemu klientowi, ma miejsce *behavioralna ocena zdolności kredytowej*, a cały proces kredytowy jest uproszczony: ocena zdolności kredytowej może być prowadzona jeszcze przed przyjęciem formalnego wniosku kredytowego i rozmową z klientem (Kaszyński, Kamiński i Szapiro 2020).

Spośród etapów procesu kredytowego w dalszej części rozprawy będzie rozważany etap oceny zdolności kredytowej. Z racji na temat rozprawy dotyczący problemu stroniczości algorytmicznej w modelach oceny zdolności kredytowej, w dalszej części tekstu skupiono uwagę na procesie kredytowym dotyczącym konsumentów. Jako segment analizowanych klientów kredytowych wybrano najbardziej liczny, tj. gospodarstwa domowe, dalej przedstawiane również jako *klient indywidualny*. Z perspektywy produktowej w rozprawie przede wszystkim będą rozpatrywane kredyty konsumpcyjne, które stanowią dużą składową ekspozycji banków oraz posiadają stosunkowo wyższe poziomy oprocentowania. Jak wskazał Przanowski (2014), kredyty ratalne (np. hipoteczne) są kredytami niskomarżowymi, wykorzystywanymi przez banki w głównej mierze do akwizycji nowych klientów i budowy z nimi stabilnej relacji; kredyty konsumpcyjne (na potrzeby rozprawy przyjęto, że są to kredyty gotówkowe) są kredytami, na których bank wypracowuje znaczną część zysku z działalności kredytowej²².

1.3 Dane wykorzystywane w procesie kredytowym

W poprzedniej części rozdziału zaprezentowano historię zmian związanych z instytucją kredytu do czasów współczesnych wraz z ich uwarunkowaniami regulacyjnymi i praktykami banków. Dodatkowo przedstawiono ramy współczesnego procesu kredytowego wraz z umiejscowieniem oceny zdolności kredytowej w tym procesie. W niniejszym podrozdziale przedstawiono zakres danych, które współczesne banki wykorzystują w procesie kredytowym; należy wskazać, że ich zakres jest zróżnicowany ze względu na segment klienta oraz cel kredytowy (Kaszyński, Kamiński i Szapiro 2020). W zależności od relacji klienta wnioskującego o kredyt z bankiem, tj. czy jest to kredyt dla nowego, czy dotychczasowego klienta, wyróżnia się odpowiednio dane aplikacyjne oraz dane behawioralne. *Dane aplikacyjne* (pozyskane z aplikacji–wniosku kredytowego) oraz *dane behawioralne* (dane związane z obserwacją zachowania klienta w relacji z bankiem) stanowią wewnętrzne źródła danych banku o kliencie indywidualnym. Dodatkowo bank może pozyskiwać dane ze źródeł zewnętrznych, tj. dane pozyskiwane z biur informacji kredytowych lub dane pochodzące z tzw. *alternatywnych* źródeł.

²² „Kredyt ratalny [przyp. wł.: np. kredyt hipoteczny] jako akwizycja jest oferowany klientom, by pozyskać ich dla banku i związane jest to najczęściej z niskimi marżami dla banku. Klient musi odnieść wrażenie, że mu się opłaca korzystać z usług naszego banku. Dopiero na kredycie gotówkowym bank jest w stanie zarobić na tyle dużo, by pokryć stratę z kredytu ratального.” (Przanowski 2014).

1.3.1 Dane aplikacyjne

Standardowy proces kredytowy, w przypadku nowych klientów jest uruchamiany przez tzw. *wniosek kredytowy*, który klient indywidualny składa w banku. Wniosek ten obejmuje dane związane z celem wnioskowanego kredytu i kredytobiorcą; załącznikami do takiego wniosku niejednokrotnie są informacje poświadczające poziom dochodów klienta. Dane pozyskiwane w ramach tego procesu zawarte są w aplikacji o kredyt, stąd dane te nazywa się *danymi aplikacyjnymi*, i obejmują swoim zakresem m.in. (Kaszyński, Kamiński i Szapiro 2020):

- dane osobowe kredytobiorcy, tj. podstawowe informacje dotyczące kredytobiorcy, w tym dane osobowe, dane o stanie cywilnym, liczbie członków rodziny na utrzymaniu;
- deklarację poziomu dochodów i wydatków, obejmującą informacje o dochodach kredytobiorcy (w tym czy są to dochody stałe, czy zmienne) i wydatkach, które ponosi;
- cel i wielkość wnioskowanego kredytu, tj. podstawowe informacje o celu i wielkości wnioskowanego kredytu;
- informacje o innych zobowiązaniach kredytobiorcy, tj. dane dotyczące zobowiązań finansowych kredytobiorcy w stosunku do innych banków.

Dane aplikacyjne stanowią podstawowe źródło informacji o nowym konsumencie.

1.3.2 Dane behawioralne

W przypadku klientów, którzy posiadają bądź posiadali zobowiązanie w banku, bank ten posiada dodatkowe informacje o współpracy z tym klientem. Dane te odnoszą się do przeszłego zachowania klienta (stąd nazwa – dane behawioralne) związanego z płatnościami za zobowiązania, informacjami o wpływach i wydatkach. Dane behawioralne, z racji ich wewnętrznej genezy, należą do danych wewnętrznych banku i obejmują m.in. dane o:

- spłacie poprzednich zobowiązań, tj. informacja o występowaniu i wielkości przeszłych kredytów, informacja o opóźnieniach w spłatach historycznych zobowiązań;
- poziomie obecnego zadłużenia, tj. poziom obecnego zadłużenia klienta w banku, w tym wskaźnik poziomu wartości zadłużenia do wartości (ang. loan-to-value, LTV).

Dane aplikacyjne i dane behawioralne stanowią wewnętrzny zasób danych, którym dysponuje bank; pochodzą one z aplikacji kredytowej oraz obserwacji historycznego zachowania klienta. Zestaw tych danych może być wzbogacany o dodatkowe zmienne pochodzące z zewnętrznych źródeł, np. dane biur informacji kredytowej.

1.3.3 Dane biur informacji kredytowej

Biura informacji kredytowej powstały w XIX w. (podrozdział 1.1.7). Informacje o klientach, którymi dysponują poszczególne banki mogą być niepełne i fragmentaryczne; w przypadku klientów indywidualnych, którzy posiadają aktywne rachunki kredytowe w różnych bankach, występuje potrzeba i zasadność wymiany informacji między bankami zainteresowanymi. Celem funkcjonowania biur informacji kredytowej jest gromadzenie i udostępnianie danych o spłaconych i aktywnych zobowiązaniach kredytowych innym bankom.

Potrzeba istnienia biur informacji kredytowej wiąże się z występującą na rynku kredytów *asymetrią informacji*, tj. sytuacją w której kredytobiorca-konsument posiada większą wiedzę o swojej zdolności kredytowej niż bank udzielający kredytu (Varian 2007). Działanie biur informacji kredytowych zmniejsza asymetrię informacyjną, jaka występuje między kredytobiorcą a kredytodawcą. Z perspektywy klienta, informacje o istniejących w innych bankach zobowiązaniach kredytowych obniżają bieżącą zdolność kredytową tego klienta; może mieć on zatem przesłanki do utajniania informacji o już zaciągniętych zobowiązaniach kredytowych.

Dane udostępniane przez biura informacji kredytowej mogą być wykorzystywane w procesie kredytowym w ocenie zdolności kredytowej, ponieważ uwzględniają informacje raportowane przez inne banki. Dodatkowo umożliwiają ograniczania wyłudzeń kredytów, jak również monitorowania poziomów zadłużenia konsumenta. Zakres danych, które są udostępniane przez biura informacji kredytowej obejmuje, m.in.:

- wskaźniki bieżącego regulowania zobowiązań, tj. informacja o tym, czy dany klient obecnie posiada jakiegokolwiek opóźnienia w innych bankach;
- historię kredytową konsumenta, w tym dane dotyczące spłat przeszłych kredytów w innych bankach;
- historię składanych przez danego klienta zapytań kredytowych, w tym dane wskazujące czy w ostatnim okresie konsument składał wnioski kredytowe i jakie uzyskiwały one decyzje kredytowe;
- ocenę zdolności kredytowej wykonaną przez biuro informacji kredytowej; biura informacji kredytowej posiadają duże zasoby danych o klientach z różnych banków i mogą w oparciu o te zasoby danych budować własne modele oceny zdolności kredytowej.

1.3.4 Dane alternatywne

Współczesnym trendem badawczym w zakresie modeli oceny zdolności kredytowej jest analiza zasadności wykorzystania alternatywnych źródeł danych o konsumentach. Wskazany

trend obejmuje dwa wymiary: prawną możliwość wykorzystania tych danych oraz wpływ ich włączenia na jakość predykcyjną opracowywanych modeli oceny zdolności kredytowej.

Alternatywne źródła danych o konsumentach powinny obejmować informacje – atrybuty do tej pory niedostępne bankom, a jednocześnie charakteryzujące się wysokim poziomem mocy predykcyjnej (Klein 2019). Dodatkowo, wykorzystywanie alternatywnych źródeł danych powinno umożliwić ocenę zdolności kredytowej osób, które do tej pory były wykluczone (Berg i in. 2020; Klein 2019; Biuro Wykonawcze Prezydenta USA 2016; Brevoort, Grimm i Kambara 2015)²³. Trend ten jest na tyle silny, że może już obejmować ponad połowę wszystkich instytucji udzielających kredytów; zgodnie z danymi agencji Experian, wartość ta może nawet wynosić 65% (Experian 2019).

Dane alternatywne nie są bezpośrednio związane z zachowaniami kredytowymi klientów (podrozdział 1.3.2). Dane z alternatywnych źródeł są uzyskiwane za pomocą, np. sieci społecznościowych, aktywności konsumenta w sieci, informacji o płatnościach za usługi nie bankowe lub rachunków za wynajem nieruchomości.

W literaturze przedmiotu wskazano na różne źródła danych alternatywnych, które mogą zostać wykorzystane w procesie kredytowym, w szczególności na etapie oceny zdolności kredytowej (ale również na etapie monitoringu kredytu). Zakres tych danych obejmuje m.in.:

- aktywność klientów w Internecie – dane obejmujące informacje o korzystaniu z sieci społecznościowych, usług sektora e-commerce, informacje z przeglądarki internetowej;
- rachunki za świadczone usługi – informacje o płatnościach rachunków telefonicznych, za media;
- dane transakcyjne – dane o transakcjach wykonywanych przez danego klienta (m.in. czy wykonuje dużo zakupów internetowych, z jakiego rodzaju usług korzysta);
- informacje z publicznych rejestrów – dane obejmujące m.in. rejestry sądowe, rejestry policyjne, dane o zajmowanych stanowiskach, posiadanych akcjach lub udziałach w spółkach, informacje o przyznanych dofinansowaniach lub grantach.

* * *

W niniejszym rozdziale przedstawiono historyczny zarys instytucji kredytu oraz wskazano ramy procesowe – proces kredytowy – wykorzystywane do udzielania kredytów konsumentom; opisano również źródła danych wykorzystywanych przez współczesne banki w procesie kredytowym do realizacji tzw. oceny zdolności kredytowej.

²³ Takie osoby nazywa się *kredytowo niewidzialnymi* (ang. credit invisible), (Brevoort, Grimm i Kambara 2015).

Instytucja kredytu zmieniała się w czasie; zmiany te dotyczyły kwestii *ilościowych* (np. poziom oprocentowania, horyzont czasu pożyczki) oraz zagadnień *jakościowych* (np. warunki sporządzania umowy kredytowej). Standaryzacja i popularyzacja instytucji kredytu z czasem prowadziła do ukształtowania się procesu kredytowego.

Jednym z etapów procesu kredytowego jest ocena zdolności kredytowej. Z racji na podjętą w rozprawie tematykę – stronniczości algorytmiczną w modelach oceny zdolności kredytowej – w dalszej części tekstu skupiono uwagę na kredytach udzielanych konsumentom.

Wprowadzone do wywodu pojęcia kredytu, procesu kredytowego oraz danych wykorzystywanych w procesie kredytowym posłużą w kolejnych rozdziałach do zdefiniowania procesu oceny zdolności kredytowej, który jest narażony na problem dyskryminacji, tj. stronniczość algorytmiczną.

Rozdział 2

Ocena zdolności kredytowej konsumenta

Celem niniejszego rozdziału jest przedstawienie oraz omówienie oceny zdolności kredytowej oraz modelu oceny zdolności kredytowej. Oba te pojęcia z perspektywy rozprawy są pojęciami podstawowymi: zagadnienie stronniczości algorytmicznej, w kolejnych rozdziałach będzie rozważane w kontekście modeli decyzyjnych wykorzystywanych do oceny zdolności kredytowej.

W przypadku udzielania kredytów banki są prawnie zobligowane do przeprowadzenia oceny zdolności kredytowej (ustawa 2011/05/12; ustawa Prawo Bankowe). W przypadku kredytów udzielanych konsumentom banki „mogą w celu oceny zdolności kredytowej i analizy ryzyka kredytowego podejmować decyzje, opierając się wyłącznie na zautomatyzowanym przetwarzaniu” (ustawa Prawo Bankowe, art. 105a), które to *zautomatyzowane przetwarzanie* w dalszej części rozprawy będzie nazywane modelem oceny zdolności kredytowej.

Niniejszy rozdział posiada następującą strukturę: w podrozdziale 2.1 wprowadzono pojęcie oceny zdolności kredytowej, wskazując na ewolucję tego procesu: przejście z eksperckiej oceny realizowanej przez analityka kredytowego, na proces zautomatyzowany wykorzystujący oszacowany model oceny zdolności kredytowej. W podrozdziale 2.2 przedstawiono sposoby oceny jakości działania modeli oceny zdolności kredytowej, tj. procedurę walidacji takiego modelu; w rozprawie rozważanym wymiarem prowadzonej walidacji będzie ocena mocy dyskryminacyjnej modelu²⁴.

²⁴ W tym miejscu należy wskazać na zbieżność pojęć: „ocena mocy dyskryminacyjnej modelu”, a „stronniczość algorytmiczna” tj. dyskryminacja przy wykorzystaniu modelu oceny zdolności kredytowej. W rozprawie *ocena mocy dyskryminacyjnej modelu* rozumiana jest jako zdolność modelu do sortowania klientów względem ich poziomu zdolności kredytowej; nie jest zatem związana z powszechnie przyjmowaną definicją dyskryminacji, ale jest sposobem weryfikacji jakości prognostycznej modelu oceny zdolności kredytowej. Ocena mocy dyskryminacyjnej modelu *nie odnosi* się do pojęcia dyskryminacji w rozumieniu mniej korzystnego traktowania danej osoby ze względu na wartość cechy prawnie chronionej (tak w rozprawie będzie rozumiane pojęcie stronniczości algorytmicznej).

2.1 Ocena zdolności kredytowej

W ramach prowadzonej działalności kredytowej bank ocenia wnioski–aplikacje kredytowe przedłożone przez konsumentów; ocena ta obejmuje: *ocenę zdolności prawnej* i *ocenę zdolności kredytowej* (Zawadzka 1996)²⁵.

Ocena zdolności prawnej również jest nazywana *oceną zdolności formalnoprawnej* i dotyczy oceny konsumenta w zakresie możliwości podejmowania przedmiotowych czynności prawnych, tj. do zawierania umowy kredytowej. Przeprowadzenie oceny zdolności formalnoprawnej umożliwia odpowiedź na pytanie czy bank *może* udzielić kredytu danemu konsumentowi (Zawadzka 1996). W dalszej części rozprawy przyjęto, że konsumenci składający wnioski kredytowe mogą zawierać umowy kredytowe.

Po pozytywnej weryfikacji oceny zdolności prawnej bank przystępuje do badania zdolności kredytowej konsumenta i próby odpowiedzi na pytanie, czy bank *powinien* dany kredyt udzielić. Zdolność kredytowa rozumiana jest jako (ustawa Prawo Bankowe, art. 70):

„Art. 70. [Zdolność kredytowa]

1. Bank uzależnia przyznanie kredytu od zdolności kredytowej kredytobiorcy. Przez zdolność kredytową rozumie się zdolność do spłaty zaciągniętego kredytu wraz z odsetkami w terminach określonych w umowie. Kredytobiorca jest obowiązany przedłożyć na żądanie banku dokumenty i informacje niezbędne do dokonania oceny tej zdolności.”

Zdolność kredytowa konsumenta jest badana na potrzeby procesu kredytowego; w kolejnych podrozdziałach przedstawiono podejścia do oceny zdolności kredytowej: historyczne podejście eksperckiej oceny zdolności kredytowej oraz współcześnie wykorzystywanych modeli oceny zdolności kredytowej.

2.1.1 Eksperska ocena zdolności kredytowej

Zanim pierwsze modele oceny zdolności kredytowej zostały opracowane, tj. do lat 60 XX w., banki podejmowały decyzję kredytową opierając się na ocenie eksperckiej. Eksperska ocena zdolności kredytowej polegała na tym, że analitycy bankowi pracujący na lokalnym rynku, często znający osobiście kredytobiorcę, określali możliwości spłaty wnioskowanego zobowiązania kredytowego (Marques, García i Sánchez 2013).

W efekcie kryzysu ekonomicznego, który miał miejsce w 1837 r.²⁶, powstała agencja handlowa The Mercantile Agency, której zadaniem było gromadzenie informacji

²⁵ W literaturze, obok pojęcia *oceny zdolności kredytowej* wyróżnia się pojęcie *oceny wiarygodności kredytowej* (Zawadzka 1996); w rozprawie oba pojęcia będą traktowane jako synonimy.

²⁶ Obecnie kryzys ten jest nazywany „Paniką w Stanach Zjednoczonych z roku 1837”. Wśród 850 banków, które istniały w tamtym okresie w Stanach Zjednoczonych ponad 340 zostało zamkniętych, a ponad 60 podlegała częściowemu wstrzymaniu działania operacyjnego (Morawski 2003).

o majątkach dłużników bankowych (podrozdział 1.1.7); dane zbierane w bazach tego podmiotu były wykorzystywane przez banki do oceny zdolności kredytowej w ramach eksperckiej oceny zdolności kredytowej. Przy takiej ocenie analityk bankowy bazował wyłącznie na swojej wiedzy (Thomas, Crook i Edelman 2017). Nierzadko w tak prowadzonej ocenie kwestie pochodzenia etnicznego lub wyznania odgrywały istotną rolę (Cohen 2012).

Wskazane agencje handlowe stawały się biurami informacji kredytowej, które wspierały banki w realizacji oceny eksperckiej, uzupełniając informację o konsumencie o zewnętrzne, względem zasobów banku, źródła danych. Pierwsze takie biura działające w Stanach Zjednoczonych Ameryki Północnej były lokalnymi jednostkami, których funkcjonowanie opierało się na prowadzeniu analiz publicznie dostępnych dokumentów i informacji, miejscowych gazet lub rejestrów, w poszukiwaniu informacji, m.in. o zawiadomieniach o aresztowaniach, małżeństwach, zgonach, działalności politycznej, awansach. Bazując na tych informacjach analityk kredytowy wiedział np. czy potencjalny klient nie posiada historii kryminalnej lub czy informacje, które podał na etapie rozmowy są prawdziwe. Eksperska ocena zdolności kredytowej wykorzystująca informacje z biur informacji kredytowych posiadała jednak ograniczenia związane z poziomem szczegółowości zbieranych danych, np. brak było informacji o skali zaciąganych w innych instytucjach zobowiązań finansowych lub poziomem–istotnością przewinień lub przestępstw danego konsumenta (Thomas, Crook i Edelman 2017).

Eksperska ocena zdolności kredytowej była oceną subiektywną, zależną od indywidualnej wiedzy i doświadczenia danego analityka kredytowego. Była narażona na nadużycia związane z indywidualnymi preferencjami lub animozjami danego analityka kredytowego (Uzoamaka, Gerald i Jude 2016). Tak realizowany proces oceny zdolności kredytowej był (Thomas, Crook i Edelman 2017):

- mało wydajny, tj. czas od złożenia aplikacji kredytowej do decyzji kredytowej był długi, liczony nawet w tygodniach oraz względnie kosztowny (obsługiwany przez człowieka);
- nisko skalowany, tj. ograniczona liczba wydawanych decyzji kredytowych wynikająca z dostępności i obciążenia czasowego analityków kredytowych;
- zależny od indywidualnej wiedzy i doświadczenia analityka kredytowego, w tym podatny na indywidualne animozje lub uprzedzenia analityka kredytowego; z racji na brak wypracowanych standardów jakość decyzji kredytowej była zależna w dużej mierze od analityka kredytowego.

Wraz z rozwojem technologii informatycznych, zaawansowaniem metod analitycznych oraz występującymi zachowaniami dyskryminacyjnymi wybranych grup społecznych, podmioty regulujące rynek dostępu do kredytów konsumenckich zaczęły prawnie wymagać automatyzacji procesu oceny zdolności kredytowej konsumentów. Ze względu na wzrost wydajności, skalowalności oraz obiektywności procesu oceny zdolności kredytowej

realizowanego przez modele oceny zdolności kredytowej, banki były skłonne wprowadzać automatyzację procesu oceny zdolności kredytowej wykorzystując do tego modele. W konsekwencji prowadziło to do upowszechnienia się wykorzystania *modeli oceny zdolności kredytowej* i zastąpienie nimi wcześniej prowadzonej oceny eksperckiej (Yu i Dunn 2016; Marques, García i Sánchez 2013; Abdou 2009; ustawa Prawo Bankowe).

2.1.2 Modele oceny zdolności kredytowej

Zmiany w podejściu do oceny zdolności kredytowej zaczęły zachodzić w latach 50. i 60. XX w. wraz z opracowaniem pierwszego powszechnie stosowanego *modelu* przewidywania prawdopodobieństwa niewypłacalności pożyczkobiorców. Cechami konsumentów, które były wykorzystywane przez model był np. fakt posiadania telefonu, długość zamieszkania pod tym samym adresem, długość stażu pracy w jednym zakładzie, bądź sam wiek wnioskodawcy. Opierając się na takim podejściu do realizacji oceny zdolności kredytowej założona została firma Fair, Isaac and Company, która oferowała usługi w postaci karty skoringowej, zawierającej informację, w jaki sposób wcześniej wskazane cechy wnioskodawcy związane były z prawdopodobieństwem spłaty zobowiązania kredytowego (The Economist 2019).

W 1989 r. trzy biura informacji kredytowej: Equifax, Experian i TransUnion, opracowały pierwszą metodykę wyznaczania *punktowej* oceny zdolności kredytowej. Punktowa ocena zdolności kredytowej wskazywała na wartość z zakresu między 300 (niska zdolność kredytowa), a 850 (wysoka zdolność kredytowa), i stała się standardem wśród amerykańskich pożyczkodawców. Pierwszy model punktowej oceny zdolności kredytowej uwzględniał pięć podstawowych cech klienta związanych z informacjami finansowymi, wraz z odpowiednim rozkładem wag²⁷.

Wraz z dynamicznym wzrostem zapotrzebowania na kredyty konsumpcyjne związane z pojawieniem się m.in. kart kredytowych lub masową produkcją samochodów (podrozdział 1.1.8), banki zaczęły wykorzystywać coraz częściej automatycznie wyznaczone wartości punktowe (stąd angielska nazwa: *credit scoring*), do podejmowania decyzji kredytowych (Thomas, Crook i Edelman 2017). Banki stosujące modele oceny zdolności kredytowej odnotowywały wzrost wydajności, skalowalności, obiektywności realizowanego procesu, przy jednoczesnej poprawie jakości portfeli kredytowych (Thomas, Crook i Edelman 2017)²⁸.

²⁷ Dokładna formuła nie jest upubliczniona, ale wykorzystywanymi zmiennymi były: historia płatności (ok. 35% wagi), poziom obecnych zobowiązań kredytowych klienta (ok. 30%), długości historii kredytowej (ok. 15%), poziom ekspozycji w otwartych liniach kredytowych lub kredytach hipotecznych (ok. 10%) oraz wnioski o nowe ekspozycje kredytowe (ok. 10%).

²⁸ W początkowej fazie wprowadzenia automatycznych modeli oceny zdolności kredytowej zaobserwowano poprawę wskaźników jakości portfeli kredytowych: „(dop. aut.: stosując automatyczne modele oceny zdolności kredytowej można było) wyeliminować około 6% strat banków bez utraty wolumenu dobrych transakcji; można by wyeliminować około 24% strat poświęcając 3% procent wolumenu dobrych

W konsekwencji wykorzystanie przez banki automatycznych modeli oceny zdolności kredytowej upowszechniło się i zastąpiło wcześniej realizowaną ocenę ekspercką (Yu i Dunn 2016; Marques, García i Sánchez 2013; Abdou 2009; FICO 1998)²⁹.

Obecnie wykorzystanie modeli oceny zdolności kredytowej jest prawnie wymagane w procesie podejmowania decyzji kredytowych związanych z aplikacjami kredytowymi konsumentów (ustawa Prawo Bankowe, art. 105a). Z racji na oddziaływanie tych modeli na sytuację życiową osób fizycznych, np. sytuację finansową, mieszkaniową oraz możliwe występowanie dyskryminacyjnego działania, modele oceny zdolności kredytowej są narażone na zagrożenie stronniczości algorytmicznej (Kozodoi i Lessmann 2022; Hurlin, Pérignon i Saurin 2022; Bono, Croxson i Giles 2021; Hurley i Adebayo 2016).

Modele oceny zdolności kredytowej są konstruowane jako modele klasyfikacyjne uczenia nadzorowanego, w których zmienna celu przyjmuje dyskretne wartości. W rozprawie przyjęto, że zmienną celu modelu oceny zdolności kredytowej będzie informacja o spłacie, bądź nie, zobowiązania kredytowego (Kaszyński, Kamiński i Szapiro 2020; Przanowski 2014); zmienną celu jest zatem zmienna binarna przyjmująca odmienne wartości dla spłaconych kredytów oraz stanów, w których klient stracił *zdolność kredytową* i nie spłacił zaciągniętego zobowiązania. Modele uczenia nadzorowanego są to modele, w których uczenie modelu (nazywane inaczej trenowaniem) odbywa się przy znanych wartościach zmiennej celu.

Obecnie większość banków w ramach oceny zdolności kredytowej wykorzystuje model oparty na regresji logistycznej. Binarna zmienna celu tworzona jest na podstawie historycznych kohort klientów, którzy spłacali zobowiązania kredytowe³⁰.

Standardowo w *ocenie zdolności kredytowej* dla ekspozycji detalicznych uwzględniane są cechy kredytobiorcy, tj. stan majątkowy, zawód, wykształcenie, uzyskiwany dochód lub stan cywilny; dla ekspozycji niedetalicznych dotychczasowa i perspektywiczna sytuacja finansowa, w tym m.in. struktura kapitałowa, poziomy aktywów obrotowych, wskaźniki

kredytów; a około 50% strat wynikających z ryzyka kredytowego można było uniknąć kosztem zaledwie 7% wolumenu dobrych kredytów. Inne badanie przeprowadzone dla dealera samochodowego wykazało, że około 20% strat kredytowych można było wyeliminować kosztem zaledwie 1% wolumenu dobrych kredytów.” [tł. wł.] (Myers i Forgy 1963).

²⁹ Rozwój modeli oceny zdolności kredytowej zainicjował również prace nad rozszerzaniem systemu oceny na inne obszary życia społecznego, tzw. modelu oceny *kredytu społecznego*. Jest to zunifikowana ewidencja osób fizycznych i przedsiębiorstw, w której jednostki są bieżąco monitorowane–śledzone, i w oparciu o ich zachowania społeczne uzyskują ocenę; obecnie modele tej klasy obejmują swoim zasięgiem 325 mln obywateli Chin, co stanowi ok. 25% ogółu mieszkańców tego regionu (Chandler 2019). Wykorzystanie tego typu systemów kredytu społecznego jest obecnie zakazane w Europie (AI Act, art. 5 ust. 1).

³⁰ W literaturze przedmiotu wskazuje się na tzw. *dobrego* oraz na *złego* klienta: pojęcia te odwołują się odpowiednio do klientów, którzy historycznie spłacili oraz nie spłacili zobowiązania kredytowego (Przanowski 2014). Kryterium spłaty kredytu nie jest jedynym: banki w ramach oceny klientów wykorzystują np. miarę liczby dni opóźnienia w spłacie (ang. days-post-due, DPD): dobrzy klienci są definiowani jako tacy, u których nie wystąpiło opóźnienie w spłacie przewyższające 30 lub 90 DPD. Regulatorzy rynku kredytów, zarówno na szczeblu Polski jak i Europejskim, przedstawiają swoje rekomendacje i zalecenia w zakresie wykorzystywanych definicji dobrego i złego klienta; przykładem takiego działania mogą być wytyczne Europejskiego Nadzoru Bankowego z 2013 r. wprowadzające tzw. nową definicję default (EBA).

finansowe lub kwalifikacje kierownictwa firmy (Zawadzka 1996). Zakres stosowania poszczególnych cech, zarówno dla portfeli detalicznych, jak i niedetalicznych jest zróżnicowany i zależy od stosowanej przez bank polityki w tym zakresie³¹.

W zakresie metod i technik ekonometrycznych wykorzystywanych do budowy modeli oceny zdolności kredytowej w literaturze przedmiotu wyróżnia się m.in. następujące klasy rozwiązań (Louzada, Ara i Fernandes 2016; Kaszyński, Kamiński i Szapiro 2020): regresja logistyczna, regresja liniowa, drzewa klasyfikacyjne, analiza dyskryminacyjna, metoda wektorów nośnych, metody baggingowe i boostingowe, sieci neuronowe, sieci bayesowskie, algorytm genetyczny, modele logiki rozmytej.

Współczesne metody uczenia maszynowego, w tym m.in. sieci neuronowe, metody baggingowe i boostingowe, drzewa klasyfikacyjne, sieci bayesowskie, wykorzystywane w ocenie zdolności kredytowej charakteryzują się (Zhou i in. 2021):

- elastyczną specyfikacją modelu związaną z faktem, że w większości metody uczenia maszynowego są podejściami nieparametrycznymi, co w rezultacie skutkuje relatywnie lepszym dopasowaniem do danych, a w konsekwencji niższymi błędami klasyfikacji;
- automatyzacją rutynowych czynności modelowania ekonometrycznego, które wykonywane były przez analityków tworzących modele, w tym np. dyskretyzacja zmiennych, tworzenie interakcji między zmiennymi, uzupełnianie wartości brakujących³².

Poziom obecnego wykorzystania wyżej wskazanych klas modeli w polskich bankach nie jest znany; w artykule MamStartup (2023) przywołano następującą statystykę: „Obecnie ponad 80% decyzji kredytowych podejmowanych jest z wykorzystywaniem algorytmów machine learning (dop. aut.: uczenia maszynowego). Na podstawie olbrzymiej ilości informacji pochodzących z różnych źródeł jesteśmy (dop. aut.: bank PKO BP) w stanie tworzyć coraz bardziej zaawansowane modele do oceny ryzyka, prognozowania różnych zjawisk i zachowań klienta.” (MamStartup 2023). Należy jednak wskazać, że bank jest zobowiązany do złożenia, w formie pisemnej wyjaśnienia dotyczącego dokonanej przez siebie oceny zdolności kredytowej wnioskującego; wyjaśnienie to powinno obejmować informacje „[...] na temat czynników, w tym danych osobowych wnioskującego, które miały wpływ na dokonaną ocenę zdolności kredytowej” (ustawa Prawo Bankowe, art.

³¹ W szczególności ma to znaczenie przy różnych strategiach banków w zakresie docelowej struktury ekspozycji; część banków specjalizuje się w klientach korporacyjnych, inne zaś prowadzą zbilansowaną politykę struktury klienckiej. W przypadku klientów niedetalicznych, podmioty te są z reguły niejednorodne, różnią się wielkością, rodzajem działalności, strukturą kapitałową i własnościową, główny region działania – takie uwarunkowania skutkują różnicami zakresu stosowanych cech w modelach oceny kredytowej poszczególnych banków.

³² Klasyczne podejścia w tym zakresie zakładały budowę tzw. ABT (ang. analytical base table), tj. wielokolumnowych macierzy obserwacji zawierających transformacje na danych wejściowych, np. binaryzacja lub interakcje między zmiennymi (Kaszyński, Kamiński i Szapiro 2020; Przanowski 2014).

70a). W przypadku modeli uczenia maszynowego występuje konieczność wykorzystania dodatkowej warstwy tzw. explainable artificial intelligence (XAI), aby bank mógł taką informację wydać (Kaszyński, Kamiński i Szapiro 2020).

W kolejnym podrozdziale opisano główne miary wykorzystywane do oceny działania modelu analitycznego z uwzględnieniem specyfiki modeli oceny zdolności kredytowej, tj. faktu, że są to klasyfikatory binarne.

2.2 Ocena jakości modeli oceny zdolności kredytowej

Bez względu na to, czy opracowywany przez bank model oceny zdolności kredytowej jest zbudowany na podstawie podejścia klasycznego (regresja logistyczna) czy współczesnych metodach uczenia maszynowego (np. metody baggingowe lub boostingowe), każdy model oceny zdolności kredytowej jest poddawany ocenie jakości predykcyjnej. Wymogi tej oceny są zdefiniowane zarówno regulacyjnie, jak również wewnętrznie przez bank. W zależności od konkretnego momentu przeprowadzania analizy, tj. czy analiza jest prowadzona na powstającym modelu, czy modelu już funkcjonującym w środowisku produkcyjnym banku, jej zakres jest zróżnicowany (Kaszyński, Kamiński i Szapiro 2020).

2.2.1 Walidacja modeli oceny zdolności kredytowej

Ocena zdolności predykcyjnej modeli oceny zdolności kredytowej jest elementem szerszego procesu walidacji i przeglądu tych modeli (Kaszyński, Kamiński i Szapiro 2020). Celem przeprowadzania okresowej *walidacji* modelu oceny zdolności kredytowej jest weryfikacja, czy model spełnia założenia przyjęte na etapie budowy tego modelu, czy jakość prognostyczna tego modelu jest wystarczająca oraz czy procesy związane z rozwojem i utrzymaniem tego modelu są prowadzone w sposób odpowiedni. *Przeglądem* natomiast nazywa się zakresowo węższy proces oceny działania modelu oceny zdolności kredytowej, wykonywanej z reguły częściej niż walidacja.

Wskazuje się, że walidacja modelu powinna zakładać przeprowadzenie analizy co najmniej w zakresie (Kaszyński, Kamiński i Szapiro 2020):

- oceny jakości danych wejściowych – dane wykorzystywane w modelu oceny zdolności kredytowej muszą obejmować odpowiedni zakres informacji i mieć odpowiednią jakość;
- ilościowej oceny jakości prognostycznej modelu – model oceniany jest z perspektywy ilościowych miar ewaluacji zdolności predykcyjnej oceny zdolności kredytowej, tj. moc dyskryminacyjna oraz jakość kalibracji;
- jakościowej oceny modelu – jakość modelu oceniana z perspektywy jakościowych aspektów wykorzystywania modelu, np. jakość dokumentacji modelu, odpowiednio zdefiniowane role i odpowiedzialności związane z modelem;

- oceny wpływu zmian wprowadzonych w modelu – w przypadku zmian w modelu koniecznym jest ocena ich wpływu na jakość modelu, np. zmiana specyfikacji modelu z modeli o architekturze drzew decyzyjnych, na modele sztucznych sieci neuronowych.

Walidacja modelu oceny zdolności kredytowej jest działaniem obligatoryjnym w przypadku nowo powstałych modeli; w przypadku modeli już działających w środowisku docelowym wymaga się przeprowadzenia cyklicznej walidacji, tzw. *monitoring* modelu oceny zdolności kredytowej (KNF 2015).

Kluczowym wymaganiem z perspektywy modeli oceny zdolności kredytowych jest wymaganie, aby walidacja była przeprowadzona w sposób niezależny (Kaszyński, Kamiński i Szapiro 2020). W przypadku polskich banków wymogi regulacyjne Rekomendacji W Komisji Nadzoru Finansowego zakładają, że departament walidacji powinien być niezależnym od departamentu budowy modeli i zarządzania ryzykiem. W rozprawie przyjmuje się, że walidacja modelu będzie prowadzona wyłącznie w oparciu o ilościową ocenę jakości prognostycznej modelu, tj. ocenę mocy dyskryminacyjnej oraz ocenę jakości kalibracji. Z powodów ograniczeń technicznych – brak niezależności autora rozprawy – nie przeprowadzono pełnej walidacji modeli oceny zdolności kredytowej przedstawianych w rozprawie w rozdziałach 4 i 5.

2.2.2 Ocena mocy dyskryminacyjnej modelu

Niech $PD_{i,t}$ oznacza nieobserwowalne, rzeczywiste prawdopodobieństwo niewykonania zobowiązania przez konsumenta w okresie t , np. w ramach kalkulacji wymogów kapitałowych w reżimie podejścia Internal Ratings-Based approach, IRB (Basel), parametr $PD_{i,t}$ jest modelowany w oknie jednego roku; niech $s_{i,t}$ oznacza wynik punktowej oceny zdolności kredytowej przyznany i -temu klientowi w okresie t . Niech $\widehat{PD}_{i,t}$ oznacza oszacowane przez model oceny zdolności kredytowej prawdopodobieństwo wejścia w stan niewykonania zobowiązania i -tego klienta (bankructwo klienta, zły klient) w okresie do t (Kaszyński, Kamiński i Szapiro 2020).

Ocena mocy dyskryminacyjnej modelu jest związana z oceną możliwości rozróżnienia dobrych i złych klientów przez model. Model w pełni poprawnie (idealnie) rozróżnia te grupy, jeśli dla $PD_{1,t} > PD_{2,t} > \dots > PD_{n,t}$ uzyskane z modelu punkty scoringowe spełniają zależności $s_{1,t} < s_{2,t} < \dots < s_{n,t}$ ³³.

Jak wskazano powyżej, wartości $PD_{i,t}$ nie są obserwowalne; obserwowanym jest fakt wejścia w stan upadłości konsumenckiej. Warto podkreślić, że ocena mocy dyskryminacyjnej modelu nie bierze pod uwagę rozkładu $PD_{i,t}$, tj. badanie mocy dyskryminacyjnej nie daje odpowiedzi na pytanie, czy model posiada dobrą kalibrację prawdopodobieństwa, czyli czy

³³ Funkcjonują różne konwencje przejścia między oceną punktową $s_{i,t}$, a oceną prawdopodobieństwem bankructwa i upadłości konsumenckiej $\widehat{PD}_{i,t}$; na potrzeby niniejszego wywodu przyjęto, że czym wyższe wartości $s_{i,t}$, tym niższe jest $\widehat{PD}_{i,t}$.

prognoza prawdopodobieństwa wejścia w stan upadłości konsumenckiej spełnia warunek $\widehat{PD}_{i,t} = PD_{i,t}$.

Sposobem weryfikacji oceny mocy dyskryminacyjnej modelu jest *macierz pomyłek*, która zestawia prognozy spłacalności kredytu wraz z ich realizacjami (Fawcett 2006). W przypadku, gdy model oceny zdolności kredytowej zwraca wartości ciągłe (prognozę prawdopodobieństwa spłaty kredytu lub punktową ocenę zdolności kredytowej), w celu uzyskania binarnej zmiennej prognozowanej, wykorzystywany jest próg odcięcia decyzji kredytowej (np. punktowa ocena zdolności kredytowej $s_{i,t} > \tau$ oznacza pozytywną decyzję kredytową). Oznacza to, że próg odcięcia decyzji kredytowej jest zmienną kontrolowaną przez bank, mającą przełożenie na politykę prowadzonej akcji kredytowej (tj. niższe wartości progu odcięcia decyzji kredytowej τ będą oznaczać więcej zaakceptowanych kredytów).

Technicznie macierz pomyłek jest macierzą o wymiarach $N \times N$, gdzie N oznacza liczbę możliwych stanów prognozowanej zmiennej; w przypadku modeli oceny zdolności kredytowej konsumenta rozważanych w rozprawie prognozowana zmienna jest binarna (informacja o spłaconym, bądź nie, kredycie), a więc $N = 2$. W przypadku wykonywanej prognozy zmiennej binarnej może zajść jedna z czterech sytuacji:

- Prognoza spłaty kredytu jest zgodna z realizacją, tj. klient spłacił kredyt. Sytuacji tej odpowiadają obserwacje *prawdziwie pozytywnych* prognoz (ang. true positive – TP)³⁴. W przypadku obserwacji, dla których prognozy były prawdziwie pozytywne, bank na podstawie prognozy z modelu oceny zdolności kredytowej wskazującej na spłatę kredytu wyda pozytywną decyzję kredytową dla konsumenta, który w rzeczywistości spłacił kredyt; z perspektywy banku obserwacje te oznaczają zysk banku wynikający z marży kredytowej.
- Prognoza spłaty kredytu nie jest zgodna z realizacją, tj. klient nie spłacił kredytu. Sytuacji tej odpowiadają obserwacje *fałszywie pozytywnych* prognoz (ang. false positive – FP). W przypadku scenariusza fałszywie pozytywnych prognoz, bank w oparciu o model oceny zdolności kredytowej wydał pozytywną decyzję kredytową klientowi, który nie spłacił zobowiązania kredytowego (nie miał, bądź stracił zdolność kredytową); przypadek ten oznacza konieczność utworzenia odpisu na kredyt, tj. realną stratę kredytową banku oraz konieczność uruchomienia procesów windykacyjnych; sytuacja ta jest również negatywna z perspektywy konsumenta: konsument nieposiadający zdolności kredytowej otrzymał kredyt, którego nie był w stanie spłacić.
- Prognoza braku spłaty kredytu jest zgodna z realizacją, tj. klient nie spłaciłby kredytu. Sytuacji tej odpowiadają obserwacje o *prawdziwie negatywnych* prognoz (ang. true negative – TN). Obserwacje otrzymujące prognozy prawdziwie negatywne odnoszą się

³⁴ W rozprawie przyjęto anglojęzyczne skróty miar pochodzących z macierzy pomyłek.

do klientów, którzy nie posiadają zdolności kredytowej, i otrzymali od banku odmowę udzielenia kredytu; scenariusz ten odnosi się do sytuacji, w której bank uniknął straty.

- Prognoza braku spłaty kredytu nie jest zgodna z realizacją, tj. klient byłby w stanie spłacić kredyt. Sytuacji tej odpowiadają obserwacje *fałszywie negatywnych* prognoz (ang. false negative – FN). W przypadku obserwacji o prognozach fałszywie negatywnych, bank nie udzielił kredytu klientom, którzy posiadali zdolność do jego spłaty; scenariusz prognoz fałszywie negatywnych bank ponosi koszt utraconych korzyści, tj. w przypadku udzielenia takiego kredytu bank uzyskałby marżę.

Poglądową macierz pomyłek przedstawiono w tabeli 3.

Tabela 3: Poglądowa struktura macierzy pomyłek.

		Prognoza modelu	
		Pozytywna d. kredytowa	Negatywna d. kredytowa
Realizacja	Spłata kredytu	true positive	false negative (błąd II-rodzaju)
	Brak spłaty kredytu	false positive (błąd I-rodzaju)	true negative

Uwaga: Bez utraty ogólności, przyjęto konwencje definicji kolumn i wierszy zgodnie z Fawcett (2006).

Źródło: opracowanie własne.

Na podstawie uzyskanej macierzy pomyłek, której poszczególne komórki oznaczają *liczby obserwacji* danego wariantu prognozy modelu i realizacji, możliwym jest wyznaczenie miar jakości modelu. Są to m.in.:

- *czułość* (ang. sensitivity, true positive rate, TPR), zdolność klasyfikatora do prawidłowego identyfikowania klasy pozytywnej, $TPR = \frac{TP}{TP+FN}$;
- *swoistość* (ang. specificity, true negative rate, TNR), zdolność klasyfikatora do prawidłowego identyfikowania klasy negatywnej, $TNR = \frac{TN}{TN+FP}$;
- *wskaźnik fałszywie dodatnich* (ang. fall-out, false positive rate, FPR), czym bliższy zeru tym udział fałszywie pozytywnych obserwacji FPR również powinien być jak najbliższy wartości 0. Wskaźnik ten wskazuje, jaki jest udział fałszywie pozytywnych przypadków wśród wszystkich przypadków w rzeczywistości negatywnych $FPR = \frac{FP}{TN+FP}$;
- *wskaźnik fałszywie negatywnych* (ang. miss rate, false negative rate, FNR) częstość błędnej klasyfikacji do klasy negatywnej, $FNR = \frac{FN}{FN+TP}$;
- *dokładność* (ang. accuracy, ACC), wskazująca na sprawność klasyfikatora związaną z częstością poprawności klasyfikacji obserwacji o zmiennej klasyfikowanej zarówno pozytywnej jak i negatywnej, $ACC = \frac{TP+TN}{TP+TN+FP+FN}$;

- *precyzja* (ang. precision, positive predictive value, PPV) wskazująca na częstotliwość prawidłowo prognozowanych pozytywnych wartości do sumy wszystkich pozytywnych prognoz, $PPV = \frac{TP}{TP+FP}$.

Powyższe miary umożliwiające ocenę mocy dyskryminacyjnej modelu są wyznaczane dla pojedynczego poziomu progu odcięcia decyzji kredytowej τ : zmieniając poziom tego progu, zmianie podlegać będzie również rozkład wartości w macierzy pomyłek. Istnieją również metody oceny działania modelu uwzględniające wszystkie możliwe wartości progów odcięcia decyzji kredytowej τ .

Jednym z takich podejść jest *krzywa ROC* (ang. Receiver Operating Characteristic) wskazująca na kompromis między wartościami czułości klasyfikatora oraz wartościami wskaźnika fałszywie dodatnich obserwacji, dla poszczególnych progów odcięcia (Kaszyński, Kamiński i Szapiro 2020). Krzywa ta wskazuje na jakość opracowywanego modelu na zbiorze testowym, agregując wszystkie możliwe progi odcięcia klasyfikacji. Nachylenie krzywej wskazuje na stosunek czułości (true positive rate) do swoistości (false positive rate) klasyfikatora przy różnych progach decyzyjnych, a ewentualna wypukłość wskazuje na klasyfikator gorszy niż przypadek klasyfikatora losowego (może oznaczać np. błąd popełniony na etapie przygotowywania danych).

Inną miarą oceny mocy dyskryminacyjnej modelu oceny zdolności kredytowej jest *pole powierzchni pod krzywą ROC* (ang. Area Under the ROC Curve, AUC). Wartości te mieszczą się w przedziale od 0 do 1, przy czym wartości poniżej 0,5 wskazują na model o mocy dyskryminacyjnej niższej niż model losowo przyporządkowujący prognozę. Miarą stosowaną w praktyce modeli oceny zdolności kredytowej jest miara wyznaczana w oparciu o wartość AUC, tj. indeks Giniego, który jest wyznaczany jako $2 \times AUC - 1$ (Schechtman i Schechtman 2019). Indeks Giniego przyjmuje wartości w przedziale od -1 do 1 ; model charakteryzujący się mocą dyskryminacyjną mierzoną indeksem Giniego równym 1 jest modelem idealnie różnicującym dobrych i złych klientów; wartość 0 indeksu Giniego odnosi się do modelu, który losowo sortuje klientów dobrych i złych; model o wartości indeksu Giniego poniżej wartości 0 może oznaczać model, w którym zmienne w zbiorze treningowym i testowym mają odmiennie kodowaną zmienną celu – taki model posiada moc dyskryminacji niższą niż losowe sortowanie (tj. niepoprawny kierunek sortowania klientów).

2.2.3 Ocena jakości kalibracji modelu

W przypadku, gdy model posiada moc dyskryminacyjną, tj. na podstawie prognoz oceny punktowej można posortować klientów na spłacających kredyty, i tych którzy nie spłacają, kolejnym krokiem walidacji jest ocena jakości kalibracji prawdopodobieństwa. Kalibracja modeli będących klasyfikatorami binarnymi – np. model oceny zdolności kredytowej – polega na dostosowaniu wynikowych ocen punktowych modelu do

rzeczywistych prawdopodobieństw wystąpienia zdarzenia braku spłaty kredytu. Mechanizm kalibracji polega na przekształceniu ocen punktowych modelu w wartości prawdopodobieństwa spłaty kredytu (opierając się na częstości tej spłaty obserwowanej).

Oszacowany model oceny zdolności kredytowej należy poddać testowi statystycznemu, który weryfikuje czy oszacowane prawdopodobieństwo spłaty kredytu (lub braku tej spłaty), jest równe obserwowanej częstości. Technicznie, ocena jakości kalibracji modelu może być wykonana np. za pomocą statystyki Brier (Brier 1950) lub testu Hosmera–Lemeshowa (Hosmer i Lemeshow 2013). Dla przykładu: test Hosmera–Lemeshowa polega na podziale na G grup posortowanego rozkładu ocen punktowych z modelu oceny zdolności kredytowej. Grupowanie może odbywać się na podstawie kwantyli rozkładu oceny punktowej z modelu lub dla zadanego stałego kroku oszacowanego prawdopodobieństwa. Statystyka testowa jest zadana przez:

$$H = \sum_{g=1}^G \frac{(O_{1g} - E_{1g})^2}{N_g \pi_g (1 - \pi_g)} \stackrel{H_0}{\sim} \chi^2(G - 2)$$

gdzie: O_{1g} to liczba obserwacji rzeczywistych, dla których zmienna celu przyjęła wartość równą 1, E_{1g} to oczekiwana (z modelu) liczba obserwacji dla których prognoza wartości zmiennej celu z modelu wyniosła 1, N_g to liczba obserwacji w przedziale g , π_g to empiryczna częstotliwość zdarzenia default w grupie g .

W przypadku, gdy statystyka testu wskazuje na odrzucenie H_0 , koniecznym jest przeprowadzenie kalibracji modelu. Istnieje kilka technik, które można wykorzystać do kalibracji modelu klasyfikatora binarnego, tj. kalibracja funkcjami sklejanymi (Lucena 2018), regresja izotoniczna (Niculescu–Mizil i Caruana 2005) lub kalibracja Platta (Platt 1999).

2.3 Ekonomiczne aspekty oceny zdolności kredytowej

Ocena zdolności kredytowej stanowi istotny element działalności operacyjnej banku związanej z realizacją procesu kredytowego. Warto również wskazać, że jakość realizowanego procesu kredytowego, w tym ocena zdolności kredytowej przekładają się na sytuację makroekonomiczną gospodarek. W niniejszym podrozdziale przedstawiono makroekonomiczny i mikroekonomiczny wpływ oceny zdolności kredytowej.

2.3.1 Makroekonomiczny wpływ oceny zdolności kredytowej

Z perspektywy gospodarki przykładem oddziaływania regulacji prawnych na procesy oceny zdolności kredytowej, a w konsekwencji na gospodarkę danego kraju, była sytuacja Meksyku w 2012 r. Pomimo wysokiej dynamiki wzrostowej populacji mieszkańców (zgodnie z danymi za 2012 r. ok. 115 milionów ludzi – wzrost ok. 1,4% r/r), zadłużenie prywatne w tym kraju wynosiło tylko ok. 20% PKB. Wartość ta była jedną z niższych wskaźników na świecie³⁵. Tylko jedna trzecia wszystkich meksykańskich firm miała dostęp do pożyczek banków komercyjnych; wśród mniejszych firm odsetek ten był jeszcze niższy (The Economist 2012). Zgodnie z raportem The Global Competitiveness Report 2010–2011 wydanym przez Światowe Forum Ekonomiczne, dostęp do kredytów był trzecią największą barierą rozwojową Meksyku, zaraz po nieefektywności biurokracji rządowej oraz problemie korupcji (Schwab 2010).

Sytuacja Meksyku po części jest związana z panującym tam bardzo rygorystycznym systemem oceny kredytowej obsługiwany przez dwie prywatne agencje będące własnością banków. Zamiast standardowej oceny zdolności kredytowej, związanej z przydzieleniem punktacji kredytowej, potencjalni kredytobiorcy są tam klasyfikowani do jednej z dwóch grup: zdolnych do spłaty lub nie. Taki system oceny zdolności kredytowej nie różnicuje dalej klientów, bez względu na stosowaną strategię banku, apetyt instytucji na ryzyko kredytowe, bądź przyczynę odrzucenia aplikacji kredytowej. W praktyce oznacza to, że brak opłacenia jednego rachunku telefonicznego (np. na kwotę 500 pesos – w przeliczeniu ok. 100 zł) może skutkować odmową udzielenia kredytu np. na samochód, co przyniosłoby ok. 10 tys. pesos przychodu dla państwa w formie podatku VAT – ok. 2 tys. PLN (The Economist 2012).

2.3.2 Mikroekonomiczny wpływ oceny zdolności kredytowej

Z perspektywy przedsiębiorstwa – banku udzielającego kredytu – ocena zdolności kredytowej jest mechanizmem filtracji konsumentów, którzy mają relatywnie niższe prawdopodobieństwo spłaty. Kredyty obciążone wysokim prawdopodobieństwem straty posiadają stosunkowo wyższy koszt ryzyka, tj. przewidywalna strata banku związana z brakiem spłaty kredytu jest wyższa. Koszt ryzyka jest jedną ze składowych kosztów banku związanych z prowadzoną działalnością gospodarczą.

W zakresie czynników, które banki biorą pod uwagę przy ocenie ryzyka kredytowego transakcji są:

- rodzaj kredytu, np. udzielanie finansowania instytucjom publicznym jest względnie bezpiecznym kredytowaniem;

³⁵ Dla porównania: dług prywatny w krajach Unii Europejskiej zgodnie z danymi za 2012 r. wynosił odpowiednio od: 62,5% na Litwie, 72,9% w Rumunii lub 73,1% na Słowacji, do 317,4% w Luksemburgu, 306,4% w Irlandii lub 299,2% na Cyprze; Polska odnotowała wartość 74,7%.

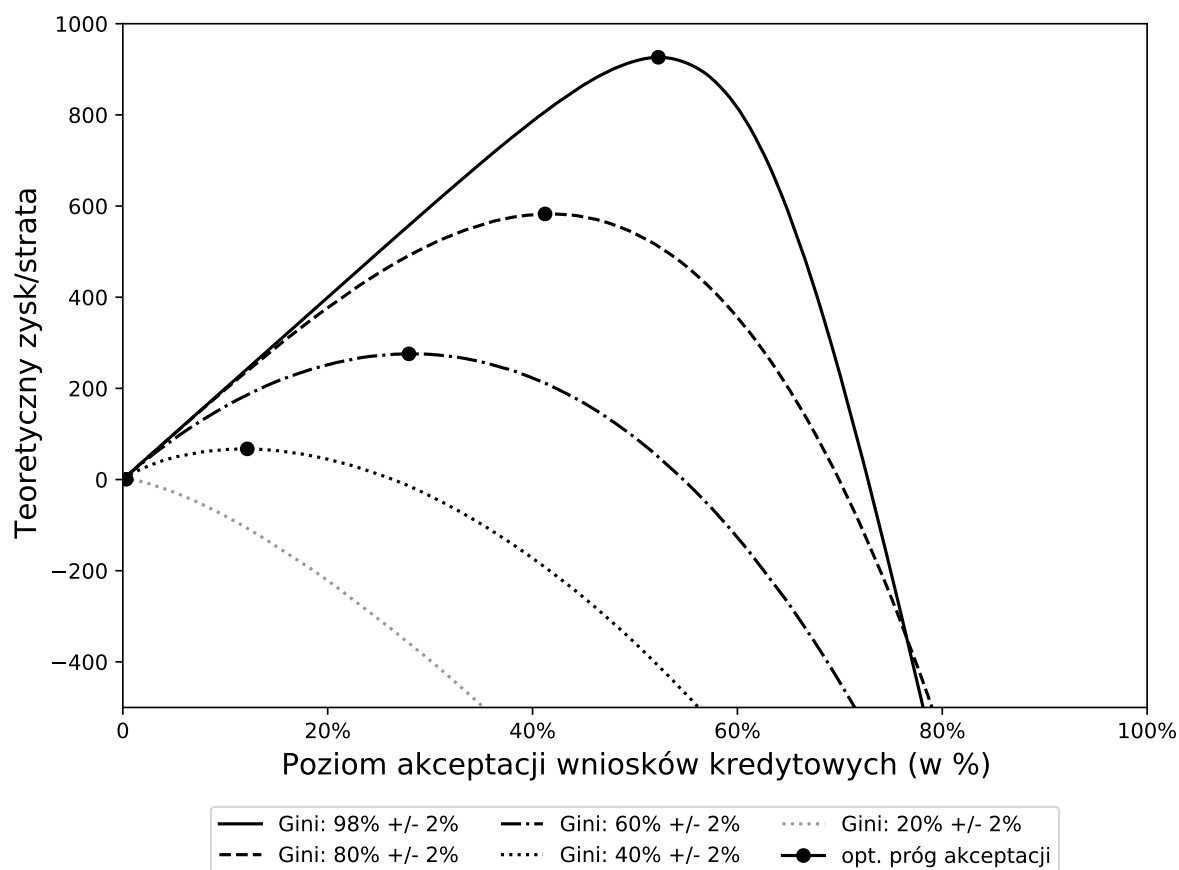
- wysokość udzielanego finansowania, tj. przy dużych zobowiązaniach kredytowych powstaje tzw. ryzyko koncentracji³⁶;
- przeznaczenie kredytu, np. finansowanie bieżących potrzeb konsumpcyjnych jest obarczone większym ryzykiem kredytowym;
- terminu zapadalności kredytu, tj. im krótszy termin spłaty, tym mniej ryzykowne ekspozycje (Matuszyk 1996; Zawadzka 1996).

Proces akceptacji wniosków kredytowych oparty na modelu oceny zdolności kredytowej charakteryzujący się wyższą mocą rozróżniania klientów z niskim i wysokim prawdopodobieństwem spłaty umożliwia obniżenie kosztu ryzyka banku przy tym samym poziomie akceptacji wniosków kredytowych (Myers i Forgy 1963). Model taki ma przełożenie na wynik finansowy banku – bank mogąc lepiej odróżniać klientów *dobrych* i *złych* – obniży koszt ryzyka, a tym samym osiągnie wyższy zysk finansowy.

Przykład poziomu zyskowności banku w zależności od mocy dyskryminacyjnej działania modelu oceny zdolności kredytowej, mierzonej indeksem Giniego (podrozdział 2.2.2), został przedstawiony na rysunku 7. W ramach przykładu zestawiono wartości średnie dla pięciu modeli oceny zdolności kredytowej; poszczególne krzywe przedstawiają średnie zależności poziomu akceptacji wniosków kredytowych od teoretycznego zysku banku. Ciągła krzywa (opisana jako Gini: 98% + / – 2%) przedstawia zależność poziomu akceptacji wniosków kredytowych oraz teoretycznego zysku dla modelu o najwyższej mocy dyskryminacyjnej; krzywa przerywana jasna (opisana jako Gini: 20% + / – 2%) przedstawia model o najniższej mocy dyskryminacyjnej.

Wraz z rosnącą mocą dyskryminacyjną modelu oceny zdolności kredytowej bank dla zadanego poziomu akceptacji wniosków kredytowych posiada wyższy poziom teoretycznego zysku. Bank posiadający model oceny zdolności kredytowej o wyższym poziomie mocy dyskryminacyjnej może podnosić poziom akceptacji wniosków kredytowych – dodatkowo zwiększając tym samym poziom zysku generowanego przez akcję kredytową. Relacja ta ma kształt odwróconej paraboli: za punktem optymalnego progu akceptacji (zaznaczony czarnym punktem), bank zwiększając poziom akceptacji wniosków kredytowych, będzie uzyskiwał gorsze wyniki ekonomiczne.

³⁶ Ryzyko koncentracji wiąże się z ekspozycjami, które mogą powstać w jednej lub w różnych kategoriach ryzyka przy uwzględnieniu całej instytucji, które mogą: generować straty na tyle istotne, by narazić kondycję finansową instytucji lub wpływać na zmianę bieżącego profilu ryzyka instytucji (EUNB 2010).



Rysunek 7: Poglądowa zależność zysku banku i poziomu akceptacji kredytowej w zależności od jakości modelu oceny zdolności kredytowej.

Uwaga: W ramach przykładu wykorzystano symulację spłacalności kredytów konsumentów; prawdopodobieństwo spłaty kredytu przez konsumenta jest zadane rozkładem logistycznym $p_i(x) = \frac{1}{1+\exp(\beta_0+\beta_1 x_{i,1})}$ gdzie parametry $\beta_0 = -2$, $\beta_1 = 7$, a cecha klienta $x_{i,1}$ jest zmienną losową ze standardowego rozkładu normalnego. Przyjęto że strata wynikająca z udzielenia kredytu, który nie zostanie spłacony to -1 j.p., a przychód wynikając z udzielenia kredytu spłaconego to $0,2$ j.p. Przyjęto również wielkość populacji kredytobiorców równą $10\,000$. Miarą jakości prognostycznej modelu oceny zdolności kredytowej jest indeks Giniego (podrozdział 2.2.2), który jest tym wyższy, im wyższą moc dyskryminacyjną model oceny zdolności kredytowej posiada.

Źródło: opracowanie własne.

2.3.3 Ryzyko kredytowe konsumenta – stan dla Polski

Wyniki oszacowania zdolności kredytowej konsumentów, którzy otrzymali kredyt, są wykorzystywane przez bank w ramach procesu oceny ryzyka kredytowego³⁷. Bank udzielający kredytów konsumentom bierze pod uwagę zarówno marżę z udzielanych kredytów (przychód banku), jak również ryzyko poniesienia straty wynikające z braku spłaty kredytu (koszt banku). Dla każdego inicjowanego kredytu, bank wyznacza poziom marży operacyjnej ponad stopę bazową (np. 3 miesięczna stopa WIBOR); zmiana rynkowych stóp procentowych oznacza dla banku zmianę obserwowanego popytu na kredyt, tj. wraz ze wzrostem stóp procentowych maleje popyt na kredyty (rysunek 4), jak również zmianę ryzyka kredytów które już zostały udzielone. Z racji na wrażliwość banku na zewnętrzne czynniki (np. stopy procentowe, cykl gospodarczy), banki są zobowiązane do zarządzania tzw. *ryzykiem kredytowym*.

W ramach procesu zarządzania ryzykiem kredytowym, bank przypisuje każdą z ekspozycji kredytowych do jednej z trzech *faz*, tj. koszyków ekspozycji, o podobnej jakości kredytowej. W Polsce, na datę 31 grudnia 2023 r., kwotę należności konsumentów przypisanych na wskazane koszyki można odpowiednio rozłożyć na 621 mld PLN w należnościach bez utraty wartości, tj. Faza 1 ekspozycji zgodnie ze standardem MSSF 9 (IFRS), oraz 70 mld PLN na należności bez utraty wartości, ale z zidentyfikowanym pogorszeniem jakości kredytowej (Faza 2 zgodnie z MSSF 9). Pozostała kwota 35 mld PLN to ekspozycje z rozpoznaną utratą wartości – tj. ekspozycje w stanie niewykonania zobowiązania – również określane mianem ekspozycji w stanie *default*³⁸. W tabeli 4 zaprezentowano zestawienie udzielonych w Polsce kredytów w rozbiciu na segment klienta oraz jakość kredytową ekspozycji; stan zgodny z danymi KNF (2023).

Z perspektywy całkowitej kwoty udzielonych kredytów gospodarstwom domowym, aż 14% tej wartości stanowią kredyty z pogorszoną jakością kredytową, tj. należące do Fazy 2 lub Fazy 3. Wartość ta dla przedsiębiorstw jest nawet większa i wynosi ok. 20%.

Na ekspozycje kredytowe znajdujące się w Fazie 2 wymaga się podwyższonego poziomu odpisów księgowych z tytułu utraty wartości kredytowej: odpisy te kalkuluje się

³⁷ Słowo *ryzyko* pochodzi od słowa *risk* z języka arabskiego, gdzie oznacza ‘los lub dopust boży’, oraz z języka greckiego, w którym oznacza ‘ostrą skałę lub rafę’, tj. coś niebezpiecznego, na co powinni żeglarze zwracać uwagę (Kaczmarek 2006). Pojęcie ryzyka zostało wyodrębnione z pojęcia niepewności, która mogła mieć charakter mierzalny lub niemierzalny (Knight 1921); ryzykiem nazywa się niepewność mierzalną, dla której możliwe jest określenie prawdopodobieństwa. Niepewnością niemierzalną uznaje się niepewność sensu stricte (Knight 1921). Ryzyko kredytowe jest związane z ryzykiem straty wynikającej ze zmiany zdolności kredytowej kredytobiorcy, która to zmiana w efekcie może prowadzić do niewypełnienia zobowiązania zawartego w umowie kredytowej przez kredytobiorcę (Iwanicz-Drozdowska i in. 2017).

³⁸ Stanem *default* określa się stan braku możliwości wykonania zobowiązania (tj. stan niewykonania zobowiązania). Definicja stanu niewykonania zobowiązania (stanu *default*) wynika m.in. z art. 178 Rozporządzenia Parlamentu Europejskiego i Rady (UE) nr 575/2013. Zgodnie z definicją EBA, przyjmuje się, że niewykonanie zobowiązania przez kredytobiorcę zachodzi gdy spełniona jest jedna z przesłanek: bank określi wywiązanie się kredytobiorcy w pełni ze swoich zobowiązań jako mało prawdopodobne – tj. bez konieczności realizacji zabezpieczenia oraz opóźnienie w realizacji zobowiązań przez kredytobiorcę wobec banku przekracza 90 dni (dla nieruchomości mieszkalnych wartość ta może wynosić 180 dni).

Tabela 4: Zestawienie udzielonych kredytów w Polsce, stan na koniec 2023 r.

Segment klienta	Bez utraty wart. Faza 1	Bez utraty wart. Faza 2	Z utratą wart. Faza 3	Procent opóźnionych
Gosp. domowe	621	70	35	14%
Mieszkaniowe	414	41	11	11%
Konsumpcyjne	155	20	15	18%
Pozostałe	51	9	8	25%
Przedsiębiorstwa	331	57	25	20%
MŚP	176	36	19	24%
Duże przedsiębiorstwa	154	20	6	15%

Uwaga: Kwoty w tabeli w mld PLN. Faza 1 oznacza ekspozycje kredytowe, których ryzyko kredytowe nie wzrosło istotnie od momentu początkowego ujęcia; są to kredyty na bieżąco obsługiwane. Faza 2 oznacza ekspozycje kredytowe, których ryzyko kredytowe wzrosło istotnie od momentu początkowego ujęcia; dla tych kredytów występują przesłanki ilościowe (np. opóźnienie w spłacie poniżej 90 dni) lub jakościowe pogorszenia jakości kredytowej. Faza 3 oznacza ekspozycje kredytowe w stanie niewywiązania się z zobowiązania kredytowego.

Źródło: opracowanie własne na podstawie danych KNF (2023).

w horyzoncie trwania całego kredytu. Dla ekspozycji zaklasyfikowanych do Fazy 1 odpisy aktualizujące wyznacza się wyłącznie w horyzoncie najbliższych 12 miesięcy. Poprawa procesów oceny zdolności kredytowej, poprzez lepszą identyfikację kredytobiorców o obniżonej zdolności kredytowej, prowadzi do zmniejszenia przyszłych przepływów ekspozycji z Fazy 1 do Fazy 2 (oraz dalej Fazy 3, czyli utraty wartości), co w efekcie przekłada się na mniejsze odpisy księgowe z tytułu utraty wartości.

* * *

W niniejszym rozdziale wprowadzono pojęcie procesu oceny zdolności kredytowej konsumenta. Wskazano na ewolucję tego procesu, tj. przejście z oceny eksperckiej na ocenę wykorzystującą modele decyzyjne: tzw. modele oceny zdolności kredytowej. Ewolucja ta, oprócz zmian w zakresie narzędzi szacowania prawdopodobieństwa spłaty zobowiązania przez konsumenta, zawiera również wskazania prawne, które regulują kwestie udzielania kredytów konsumentom. Regulacje te, pomimo że częściowo są to dokumenty jeszcze z lat 60–70 XX w., zawierały już klauzule przeciwdziałające dyskryminacji ze względu na atrybuty prawnie chronione.

W rozdziale 2 opisano również podejścia do weryfikacji jakości działania modelu oceny zdolności kredytowej, w szczególności w zakresie ilościowej oceny jakości prognostycznej. Zakończono go przykładami makroekonomicznymi i mikroekonomicznymi, wskazującymi na zasadność budowy modeli oceny zdolności kredytowej oraz istotność parametrów opisujących ich jakość prognostyczną (podsumowując: większa jakość prognostyczna tych modeli przekłada się na wyższą zyskowność dla banku). Modele oceny zdolności kredytowej,

wskazujące na oszacowania prawdopodobieństwa spłaty kredytu zgodnie z warunkami umownymi, są wykorzystywane przez banki również w procesie zarządzania ryzykiem kredytowym (proces biznesowy) oraz tworzeniem odpisów aktualizujących wartość zgodnie z Międzynarodowym Standardem Sprawozdawczości Finansowej 9.

Modele oceny zdolności kredytowej konsumenta, opisane w niniejszym rozdziale, będą w dalszym toku rozprawy analizowane w kontekście występowania problemu stroniczości algorytmicznej.

Rozdział 3

Stronniczość algorytmiczna

Celem niniejszego rozdziału jest wyjaśnienie pojęcia stronniczości algorytmicznej oraz wskazanie ram prawnych określających zasady równości oraz zakazu dyskryminacji. Omówiono w nim zagadnienia związane ze stronniczością algorytmiczną w modelach decyzyjnych, wskazując przy tym na kryteria oceny występowania i metody redukcji problemu stronniczości algorytmicznej.

W podrozdziale 3.1 przedstawiono definicję stronniczości algorytmicznej, w podrozdziale 3.2 wskazano na zapisy prawne odnoszące się do zasad antydyskryminacyjnych, w tym definicje dyskryminacji pośredniej i dyskryminacji bezpośredniej. W podrozdziale 3.3 zaprezentowano zagadnienie stronniczości algorytmicznej w modelach oceny zdolności kredytowej; fragment ten zawiera podsumowanie sposobów pomiaru stronniczości algorytmicznej odnoszących się do pojęć dyskryminacji pośredniej i dyskryminacji bezpośredniej oraz metod redukcji stronniczości algorytmicznej w modelach oceny zdolności kredytowej. Następnie wykazano główne przepisy prawne odnoszące się do kwestii równości oraz zakazu dyskryminacji. Podrozdział 3.4 poświęcono kwestiom modelowania statystycznego, umożliwiającym pomiar i redukcję stronniczości algorytmicznej.

3.1 Pojęcie stronniczości algorytmicznej

Stronniczość algorytmiczna, nazywana również *skrzywieniem algorytmicznym*, to takie działanie modelu decyzyjnego, które prowadzi do dyskryminacji lub naruszenia norm społecznych w zakresie równości ludzi. Friis i Riley (2023) wskazują, że „stronniczość algorytmiczna ma miejsce, gdy algorytmy podejmują decyzje, które systematycznie stawiają w niekorzystnej sytuacji określone grupy ludzi.” [tł. wł]. Podobną definicję przedstawiają Kordzadeh i Ghasemaghaei (2022) „(dop. aut.: stronniczość algorytmiczna ma) miejsce, gdy wyniki algorytmu przynoszą korzyść lub niekorzyść pewnym osobom lub grupom bardziej niż innym, bez uzasadnionego powodu takich nierównych skutków.” [tł. wł].

Stronniczość algorytmiczna jest zatem związana z dyskryminacyjnym, nierównym charakterem prognoz generowanych przez modele decyzyjne, w tym modele oceny zdolności kredytowej (Warchoń 2023). Ze względu na uwarunkowania prawne związane z zakazem dyskryminacji, stronniczość algorytmiczna jest negatywną cechą modeli decyzyjnych odpowiadającą definicji dyskryminacji występującej w decyzjach podejmowanych na podstawie tych modeli.

W celu zobrazowania zagadnienia stronniczości algorytmicznej w modelach oceny zdolności kredytowej, poniżej przedstawiono przykładowy scenariusz dyskryminacyjnego działania banku.

3.1.1 Odmienny wpływ modelu

Założmy, że bank w danym okresie przyjął określoną liczbę aplikacji kredytowych $n = n_{X_a=0} + n_{X_a=1}$, gdzie $n_{X_a=0}$ oznacza liczbę wniosków kredytowych złożonych przez mężczyzn, a $n_{X_a=1}$ liczbę wniosków kredytowych złożonych przez kobiety. Przyjmijmy również, że grupa osób składających aplikacje kredytowej jest grupą reprezentatywną dla rozważanej populacji oraz że liczba aplikacji kredytowych n jest na tyle duża, że zagadnienia związane z dyskretnym charakterem miar wyznaczanych na podstawie $n_{X_a=0}$ i $n_{X_a=1}$ mogą być pominięte. Przyjmijmy, również, że bank, opierając się na wynikach modelu oceny zdolności kredytowej, udzielił kredyt dla: $n_{X_a=0}^+$ mężczyzn oraz $n_{X_a=1}^+$ kobiet. W przypadku, gdy wskaźnik zatwierdzeń kredytów, dla mężczyzn $\frac{n_{X_a=0}^+}{n_{X_a=0}}$ oraz dla kobiet $\frac{n_{X_a=1}^+}{n_{X_a=1}}$ są od siebie statystycznie istotnie różne, model oceny zdolności kredytowej posiadać będzie odmienny wpływ na mężczyzn i kobiety – jest to przypadek *dyskryminacji pośredniej* wyrażany jako różnica w proporcjach *wskaźnika zatwierdzeń* kredytów dla mężczyzn i kobiet (Kozodoi i Lessmann 2022).

Argumentem wskazującym na dyskryminacyjny charakter takiego modelu jest definicja dyskryminacji pośredniej (dyrektywa 2004/113/WE, art. 2), tj. „pośrednia dyskryminacja: występuje w sytuacji, gdy z pozoru neutralny przepis, kryterium lub praktyka stawiałyby osoby danej płci w szczególnie niekorzystnym położeniu w porównaniu do osób innej płci, chyba że dany przepis, kryterium lub praktyka są obiektywnie uzasadnione zgodnym z prawem celem, a środki służące osiągnięciu tego celu są właściwe i niezbędne”. Pomimo, że definicja ta zakłada zapis o „[...] obiektywnie uzasadnione zgodnym z prawem celem”, w przypadku oceny zdolności kredytowej istnieją kazusy sądowe, wskazujące na brak zasadności zastosowania tego zapisu (Louis v. Saferent Sols. 2023).

Proces sądowego dowodzenia odmiennego wpływu modelu oceny zdolności kredytowej wykorzystywanego przez bank zawiera co najmniej dwa etapy (Watkins, McKenna i Chen 2022).

Na początku jest ustanawiany przypadek uprawdopodobniający – *prima facie* – zaistnienie dyskryminacji pośredniej, tj. odmiennego wpływu. Watkins, McKenna i Chen

(2022) przedstawili możliwą procedurę uprawdopodobniającą zajście odmiennego wpływu modelu oceny zdolności kredytowej na grupę definiowaną przez zmienną prawnie chronioną. Procedura ta zakłada dwa etapy:

- *Wybór odpowiedniej populacji porównawczej.* W celu identyfikacji odmiennego wpływu modelu decyzyjnego na osoby o odmiennych wartościach cechy prawnie chronionej, koniecznym jest zdefiniowanie populacji porównawczej. W przypadku modeli oceny zdolności kredytowej, np. dla cechy **płeć**, należy określić populację porównawczą mężczyzn i kobiet aplikujących o kredyt w danym banku. Wybór zbyt szerokiej (np. populacja całego kraju), lub zbyt wąskiej (np. mieszkańcy osiedla, na którym znajduje się oddział banku) może utrudnić bądź uniemożliwić określenie przypadku *prima facie* (Alexander v. Edgewood Mgmt. Corp. 2019). Wybór określonej populacji porównawczej, zaakceptowanej przez sąd zależny będzie od historii ustawodawczej oraz szczegółów danego przypadku pozwu sądowego; dobór populacji porównawczej jest złożony i zależny od kontekstu danej sprawy (Watkins, McKenna i Chen 2022).
- *Wybór odpowiedniego testu statystycznego.* Opierając się na populacji porównawczej, możliwym jest zweryfikowanie czy próba obserwacji z określoną decyzją kredytową (tj. przyznanie kredytu bądź odmowa) posiada statystycznie istotnie różny rozkład wydanych decyzji kredytowych od populacji porównawczej. Powszechnie używane statystyki testowe w tym zakresie to test χ^2 lub *test Fishera* (Fisher 1970)³⁹. Spotykaną również praktyką oceniania odmiennego wpływu modelu jest zastosowanie *zasady czterech–piątych* (ang. four–fifths rule), która wskazuje, że stosunek wskaźnika zatwierdzeń kredytów u kobiet do wskaźnika zatwierdzeń kredytów u mężczyzn nie może być mniejszy niż $\frac{4}{5} = 80\%$. Dalej, stosując symetryczną zasadę dla grupy uprzywilejowanej, otrzymuje się wartość $\frac{5}{4} = 125\%$ (Saleiro i in. 2019). Watkins, McKenna i Chen (2022) wskazują jednak, że w rzeczywistości zasada ta nie powinna być traktowana jako jedyne kryterium oceny odmiennego wpływu modelu, ze względu na tzw. epistemiczne wkroczenie (ang. epistemic trespassing)⁴⁰.

Drugim etapem dowodowym jest uzasadnienie braku „obiektywnie uzasadnionego, zgodnego z prawem celu gospodarczego”. Wystąpienie przypadku *prima facie*, np. odrzucenie hipotezy zerowej o takim samym wpływie modelu oceny zdolności kredytowej dla testu

³⁹ W przypadku, gdy wskazania testów się różnią, sąd jest ostateczną instancją podejmującą decyzję, ze względu na fakt, że „Statystyki [...] mają nieskończoną różnorodność [...] ich użyteczność zależy od wszystkich otaczających faktów i okoliczności.” [tł. wł.], (Hazelwood School District v. United States 1976)

⁴⁰ Watkins, McKenna i Chen (2022) definiują *epistemiczne wykroczenie* jako przedwczesne wnioskowanie w jednym obszarze (np. modele oceny zdolności kredytowej) w oparciu o wnioski lub rozwiązania stosowane w innym obszarze (np. modele decyzyjne wspierające proces zatrudnienia). Autorzy wskazują, że zasada czterech–piątych została początkowo wprowadzona jako regulacja obszaru równości w dostępie do rynku pracy (tj. zatrudnienie); jednocześnie Watkins, McKenna i Chen (2022) wskazują, że zasada ta nie powinna zostać bezpośrednio przeniesiona na inne obszary, np. modele oceny zdolności kredytowej, ze względu na różnice kontekstowe obu tych obszarów.

Fishera, nie indukuje dowodu, że model posiada odmienny wpływ. W przypadku modeli oceny zdolności kredytowej organy regulacyjne mogą wnioskować o wykazanie, że wybrany model oceny zdolności kredytowej jest najmniej dyskryminujący z klasy wszystkich modeli zapewniających wystarczającą jakość prognostyczną (Elston v. Talladega Cty. Bd. of Educ 1993; Watkins, McKenna i Chen 2022).

W przypadku, gdy dowiedziony zostanie przypadek *prima facie* oraz nie zostanie obiektywnie uzasadniony cel gospodarczy, a zostaną zidentyfikowane możliwe sposoby redukcji poziomu odmiennego wpływu modelu, bank jest zobligowany, aby je udokumentować i wdrożyć (FDIC 2019).

Poniżej przedstawiono statystykę *testu Fishera*, ze względu na jego późniejsze wykorzystanie w pracy. Przyjmijmy, że bank wydawał decyzje kredytowe dla osób o różnych wartościach zmiennej prawnie chronionej – np. zmiennej binarnej *płeć*. Niech rozkład tych decyzji – tablica częstości – będzie jak w tabeli 5.

Tabela 5: Test Fishera – macierz rozkładu decyzji kredytowych, ze względu na zmienną prawnie chronioną.

	$X_a = 0$ (mężczyzna)	$X_a = 1$ (kobieta)	Σ
Pozytywna d. kredytowa	a	b	$a + b$
Negatywna d. kredytowa	c	d	$c + d$
Σ	$a + c$	$b + d$	n

Źródło: opracowanie własne.

Niech rozkład poszczególnych *liczności* wydawanych decyzji kredytowych i wartości zmiennej prawnie chronionej *płeć* będzie zadany przez parametry a , b , c i d ; dodatkowo niech całkowita liczba wydanych decyzji kredytowych oznaczona będzie przez n , a sumy wierszy lub kolumn stanowią sumy brzegowe.

Prawdopodobieństwo p zaobserwowania dokładnie takiego rozkładu parametrów a , b , c i d jest równe:

$$p = \frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{n}{a+b}}$$

Hipoteza zerowa H_0 testu przyjmuje, że rozkład brzegowy decyzji kredytowej dla mężczyzn i kobiet jest taki sam⁴¹.

⁴¹ Do wyznaczenia wartości p-value, w badaniu przedstawionym w rozprawie w rozdziale 4 wykorzystano implementację testu Fishera w języku Python `scipy.stats.fisher_exact`.

3.1.2 Odmienne traktowanie przez model

Przyjmijmy, że wskaźnik zatwierdzeń kredytów jest zbliżony dla mężczyzn i kobiet. W takim przypadku, model posiada taki sam wpływ na grupy społeczne, różnicowane przez wartość cechy prawnie chronionej – obie grupy społeczne posiadają taki sam poziom wskaźnika akceptacji kredytowej. Przyjmijmy jednak, że sama procedura traktuje jedną z grup społecznych w sposób mniej korzystny (tzw. grupa nieuprzywilejowana). Jako przykład literatury podaje się zjawisko tzw. *lenistwa modelu* (ang. model laziness), tj. sytuacji, w której model dla jednej grupy (tzw. *grupy uprzywilejowanej*) posiada wysoką jakość prognostyczną (tj. niewielki błąd klasyfikacji), a dla drugiej bardzo niską jakość prognostyczną (tj. wysoki błąd klasyfikacji). W skrajnym scenariuszu można wskazać jako przykład model, który np. dla mężczyzn wykorzystuje model idealny (tj. o zerowym błędzie klasyfikacji), a dla kobiet model losowy (Makhlouf, Zhioua i Palamidessi 2021).

We wskazanym przykładzie populacja kobiet jest narażona na niską jakość działania modelu oceny zdolności kredytowej: kobiety, które nie posiadają zdolności kredytowej, mają wyższe prawdopodobieństwo błędnego uzyskania wyższej oceny zdolności kredytowej i tym samym uzyskania pozytywnej decyzji kredytowej. W literaturze przedmiotu wskazano, że to właśnie błędne działania modelu, w szczególności udzielanie finansowania osobie, która nie ma zdolności do spłaty tego zobowiązania jest niebezpiecznie, zarówno z perspektywy banku (konieczność uruchomienia procesu windykacyjnego), jak również konsumenta (pogorszenie sytuacji finansowej w związku z utratą kredytu).

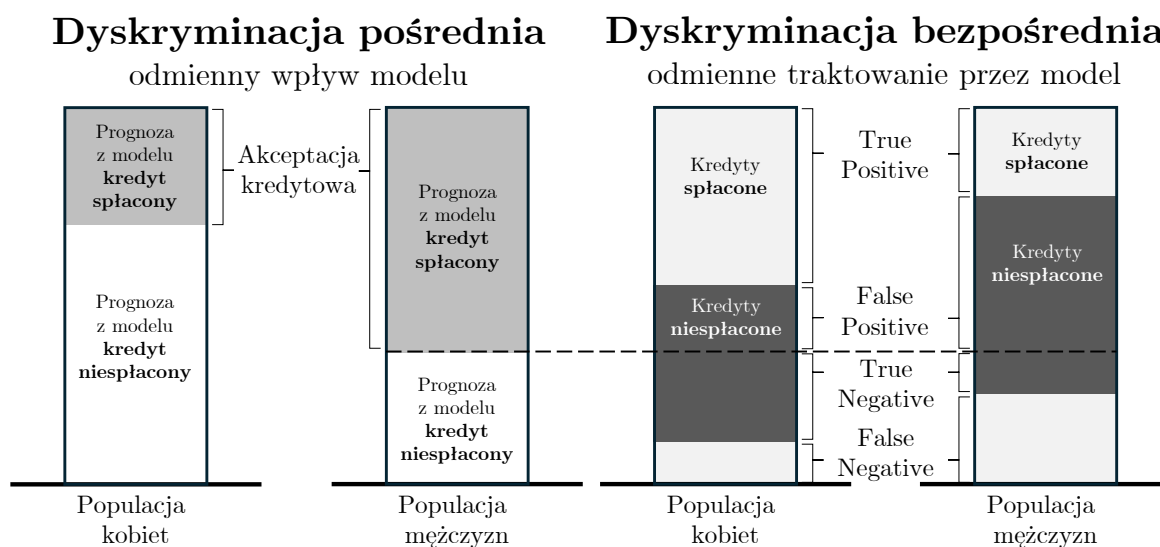
Taki przypadek odnosi się do tzw. dyskryminacji bezpośredniej, tj. sytuacji „[...] w której osoba fizyczna ze względu na płeć, rasę, pochodzenie etniczne, narodowość, religię, wyznanie, światopogląd, niepełnosprawność, wiek lub orientację seksualną jest traktowana mniej korzystnie niż jest, była lub byłaby traktowana inna osoba w porównywalnej sytuacji” (ustawa 2010/12/03). Przypadek odmiennego traktowania odnosi się do sprawiedliwości i równości w kontekście *równości szans* (Segall 2013)⁴².

Przykładowe scenariusze występowania problemu stronniczości algorytmicznej, wyrażone jako dyskryminacja pośrednia (odmienny wpływ) oraz dyskryminacja

⁴² Segall (2013) wskazuje na następujące rozróżnienie pojęć równości szans, m.in.:

- *Formalna* równość szans (ang. formal equality of opportunity) występuje, gdy nie ma formalnych (tj. prawnych ani innych instytucjonalnych) barier dla jednostek w dążeniu do celów życiowych (np. kariera, wykształcenie, dobra i usługi). Historycznymi przykładami jawnego braku formalnej równości szans, był apartheid lub utrwalony system kastowy. Literatura wskazuje jednak, że historycznie utrwalone nierówności społeczno-ekonomiczne skutkują w praktyce nierównym dostępem do dóbr i usług, pomimo, że nie ma formalnych barier tego dostępu.
- *Rzeczywista* równość szans (ang. substantive equality of opportunity) występuje, gdy formalna równość szans jest uzupełniona o neutralizację nierówności w pochodzeniu: „Aby osiągnąć naprawdę równy punkt wyjścia w wyścigu o karierę, szanse muszą zostać w pewnym sensie wyrównane. [...] Wymóg sprawiedliwego pochodzenia można realizować wcześniej (na przykład poprzez podatek od spadków i dofinansowaną edukację publiczną) lub później w życiu (na przykład poprzez działania afirmatywne).” [tł. wł.], (Segall 2013).

bezpośrednia (odmienne traktowanie) w kontekście oceny zdolności kredytowej, zostały przedstawione na rysunku 8. Wykres po lewej stronie rysunku przedstawia scenariusz dyskryminacji pośredniej, w którym kobiety posiadają niekorzystnie mniejszy wskaźnik zatwierdzeń kredytów od mężczyzn (tj. poziom akceptacji kredytowej jest dla kobiet niższy). Wykres po prawej stronie rysunku przedstawia przykład dyskryminacji bezpośredniej, w którym pomimo, że obje płcie posiadają taki sam poziom akceptacji wniosków kredytowych (linia przerywana), to mężczyźni posiadają wyższy odsetek kredytów niespłaconych. Obserwacje, dla których udzielony został kredyt (tj. model prognozował spłatę kredytu), ale w rzeczywistości nie został spłacony, oznaczane były wcześniej jako obserwacje *fałszywie pozytywne* (podrozdział 2.2.2).



Rysunek 8: Przykład odmiennego wpływu i odmiennego traktowania przez model oceny zdolności kredytowej

Uwaga: Na rysunek lewym przedstawiono scenariusz odmiennego wpływu modelu (dyskryminacja pośrednia) rozumiany jako odmienny poziom wskaźnika zatwierdzeń kredytów dla kobiet i mężczyzn; rysunek prawy przedstawia scenariusz *odmiennego traktowania* przez model oceny zdolności kredytowej. Na wykresie prawym przyjęto taki sam poziom wskaźnika zatwierdzeń kredytów (przerywana linia); dla zadanych decyzji kredytowych wyznaczyć można parametry macierzy pomyłek, zgodnie z podrozdziałem 2.2.2.

Źródło: opracowanie własne.

- *Radykalna równość szans* (ang. radical equality of opportunity) integruje w swojej definicji zarówno formalną, jak i rzeczywistą równość szans, dodając do tego trzy aspekty: walutę, przedmiot i zakres (Segall 2013). Przez walutę rozumie się wyrównanie nierówności związanych z naturalnymi predyspozycjami, talentem lub stanem psychofizycznym danej osoby; przedmiotem, w rozumieniu radykalnej równości szans, jest wyrównanie dostępu do fundamentalnych wartości ludzi, tj. zdrowia (wskazuje się tutaj stan zdrowia, a nie dostęp do usług medycznych), dochodu oraz szerzej dobrobytu; Segall (2013) wskazuje, że zakres radykalnej równości szans obejmuje zarówno przestrzeń publiczną, jak również przestrzeń prywatną.

Decyzję o prawnym zakresie definicji dyskryminacji podejmuje podmiot regulujący: *regulator*. Problem stronniczości algorytmicznej występuje wtedy, gdy wyniki działania modelu decyzyjnego prowadzą do szkodliwych konsekwencji dla niektórych osób lub grupy społecznej – *grupy nieuprzywilejowanej* – bez uzasadnionego powodu takiej nierówności (Friis i Riley 2023; Kordzadeh i Ghasemaghaei 2022).

3.1.3 Źródła stronniczości algorytmicznej

Wyróżnia się kilka głównych przyczyn stronniczości algorytmicznej modeli oceny zdolności kredytowej (Zhou i in. 2021):

1. Dane historyczne – dane, na jakich trenowane są modele uczenia nadzorowanego mogą posiadać silnie niezbalansowaną próbę względem dyskryminowanej grupy społecznej (np. niska liczba kobiet, którym historycznie udzielono kredytów). W takim przypadku model może charakteryzować się stronniczością algorytmiczną dla populacji o mniejszej liczbie obserwacji.
2. Proces zbierania danych – w przypadku gdy w trakcie procesu zbierania danych zachodzą systematyczne błędy lub selekcja obserwacji, tzw. *błąd przeżywalności* (ang. survivorship bias), trenowane modele będą charakteryzować się stronniczością algorytmiczną. Zagadnienie to jest o tyle istotne, że coraz częściej banki wykorzystują alternatywne źródła danych, aby poszerzyć zakres zbieranych atrybutów o klientach. Literatura przedmiotu przedstawia wyzwania związane z wykorzystaniem alternatywnych źródeł danych w kontekście oceny zdolności kredytowej (Berg i in. 2020). Alternatywne zmienne, np. te związane z tzw. *śladem cyfrowym* (ang. digital footprint) posiadają wysoką moc predykcyjną, lecz w większości są istotnie skorelowane z cechami prawnie chronionymi (Klein 2019).
3. Niepoprawne przypisanie etykiet obserwacji – niepoprawne przypisanie etykiet zmiennej celu może być przyczyną występowania stronniczości algorytmicznej (Corbett–Davies i in. 2023). Wartości zmiennej celu, w przypadku oceny zdolności kredytowej są etykietami kredytów spłaconych lub nie; etykiety te są wyznaczone oparciu o kryteria ilościowe (np. próg liczby dni opóźnienia – tzw. Days Past Due, DPD na poziomie 180 dni) lub jakościowych (np. informacja czy klient jest na liście obserwowanych klientów). Wybór kryteriów ilościowych oraz jakościowych jest podyktowany zarówno wewnętrzną polityką instytucji finansowej, jak również zewnętrznymi regulacjami prawnymi⁴³.
4. Występowanie wysokiej wartości korelacji pozostałych zmiennych egzogenicznych z cechami prawnie chronionymi – w przypadku automatycznych procesów analizy,

⁴³ Regulacje prawne i wewnętrzna polityka instytucji mogą się zmieniać, co technicznie prowadzi do zagadnień spójności definicji zmiennej celu na przestrzeni czasu; jako przykład można wskazać wprowadzoną w 2016 r. zmianę regulacyjną dotyczącą nowej definicji default (EBA)

budowy i utrzymania modeli decyzyjnych, zachodzi ryzyko włączenia do modelu zmiennej wysoce skorelowanej ze zmienną chronioną (np. *pochodzenie etniczne* oraz *kod pocztowy*). Wyłączenie atrybutu chronionego (nazywane w rozprawie podejściem *sprawiedliwości przez nieświadomość*), przy jednoczesnym zachowaniu wysoce skorelowanych innych zmiennych egzogenicznych, prowadzi do algorytmicznej stronniczości, tj. włączone zmienne egzogeniczne przybliżają wartość atrybutu chronionego (Hooker 2021).

5. Brak bezpośredniej możliwości interpretacji wartości parametrów modelu – w większości modele uczenia maszynowego są modelami nieparametrycznymi, nieposiadającymi bezpośredniej interpretacji parametrów, w przeciwieństwie do modeli parametrycznych – np. model regresji logistycznej. Brak możliwości prostej weryfikacji parametrów modelu decyzyjnego utrudnia analizę występowania obciążenia, a w efekcie uniemożliwia korektę lub reakcję ekspercką analityka tworzącego model.

Występuje co najmniej kilka przyczyn stronniczości algorytmicznej, przy czym większość z nich (w szczególności w kontekście bankowym) dotyczy danych, na podstawie których trenowane są później modele. Ważnym aspektem badania przyczyn stronniczości algorytmicznej jest definicja dyskryminacji, która zostanie wykorzystana do pomiaru tej stronniczości. W dalszej części rozdziału, zanim przedstawione zostaną przepisy prawne wskazujące na pojęcie dyskryminacji, przedstawiono klasyczne rozważania dotyczące sprawiedliwości w ekonomii.

3.1.4 Teoria ekonomii a sprawiedliwość decyzji

Kończąc wątek wprowadzenia pojęcia stronniczości algorytmicznej warto wskazać, że kwestiami sprawiedliwości decyzji, ludzkiej lub wygenerowanej przez model, ekonomiści zajmowali się od połowy XX w. w ramach tzw. nurtu ekonomii behawioralnej (Kahneman, Knetsch i Thaler 1986). Jednym z rozszerzeń teorii ekonomii było uwzględnienie w modelach preferencji ludzi do traktowania i bycia traktowanymi w sposób sprawiedliwy. Problem związany z uwzględnieniem w formalnym modelu m.in. pojęcia sprawiedliwości jest złożony ze względu na fakt, że bardzo często są one niedoprecyzowane, subiektywne lub nawet sprzeczne, mają charakter anegdotyczny lub nawet obyczajowy (Biggar 2010; Fehr, Kirchsteiger i Riedl 1993; Bonbright, Danielsen i Kamerschen 1988; Baumol, Koehn i Willig 1987).

Klasyczne modele mikroekonomiczne maksymalizacji zysku przedsiębiorstwa zakładają istnienie agentów przestrzegających norm prawnych; nie muszą oni jednak działać w sposób sprawiedliwy lub etyczny (Kahneman, Knetsch i Thaler 1986). W praktyce jednak ludzie postępują w sposób uwzględniający aspekty sprawiedliwości i przestrzeganie norm etycznych niejednokrotnie ponad efektywność. Z perspektywy różnych społeczeństw,

np. zjawisko problemu free-rider czyli korzystanie z wspólnych dóbr bez ponoszenia kosztu, występuje z różnym natężeniem i zróżnicowanym wpływem na gospodarkę jako całość (Delton i in. 2012)⁴⁴.

Czy to przez normy społeczne i związane z nimi wywieranie wpływu na podejmowanie przez ludzi decyzji, czy przez rozwój świadomości i bogacenie się społeczeństw poziom oceny istotności przestrzegania zasad sprawiedliwości się zmienia, co determinuje skutki ekonomiczne dla podmiotów i wpływa na poziom całej gospodarki (Biggar 2010). Wyniki empirycznych badań wskazują, że jednostki i podmioty gospodarcze uwzględniają w swoim procesie decyzyjnym kwestie sprawiedliwości, nawet w przypadkach, gdy nie ma zagrożenia karą regulacyjnymi lub utraty przyszłego przychodu, np. wynikającego z ryzyka reputacji (Kahneman, Knetsch i Thaler 1986).

W ramach rozważań przedstawicieli ekonomii wyróżnia się co najmniej dwie doktryny związane z definicją sprawiedliwości (Konow 2003), tj. *sprawiedliwość proceduralna* (ang. procedural fairness, natural justice) oraz *sprawiedliwość wyników* (ang. fairness in outcomes, end-state fairness). Definicje te są odpowiednio zbieżne z definicją dyskryminacji bezpośredniej (odmienne traktowanie) oraz dyskryminacji pośredniej (odmienny wpływ)⁴⁵.

Inne wątki rozważań w zakresie pojęcia sprawiedliwości są związane z powstawaniem i występowaniem monopolu oraz towarzyszącej im tzw. *zbędnej straty społecznej* (ang. deadweight loss). Rozważania nad sprawiedliwością miały zatem wymiar wspierania procesów regulacji rynków monopolistycznych, przede wszystkim przez wycenę produktów na poziomie krańcowego kosztu lub przez stosowanie różnych podejść do różnicowania cen (Biggar 2010).

Jednymi z pierwszych podejść do formalizacji pojęcia sprawiedliwości było zestawienie go ze stanem tzw. *braku zazdrości* (ang. absence of envy), (Pazner i Schmeidler 1978; Varian 1973; Foley 1966). W ramach tej definicji, podział dóbr pozbawiony zazdrości jest taki, w którym żaden z jego uczestników nie preferuje części która przypadła innemu uczestnikowi⁴⁶. Wskazuje się jednak na wątpliwość, czy sprawiedliwość przez brak zazdrości,

⁴⁴ Kahneman, Knetsch i Thaler (1986) przedstawili przykład konfliktu między produktywnością, a oceną sprawiedliwości danych rozwiązań: porównano 3 metody alokacji rzadkiego dobra. Pierwszą z nich była aukcja: dobro było sprzedawane dla klientów oferujących najwyższą cenę; druga to loteria: dobro trafiało do osób, które zostały wylosowane; ostatnia strategia to kolejka, w ramach której dobro trafiało do klientów, którzy pierwsi zgłosili chęć zakupu. Z perspektywy poziomu przychodu ze sprzedaży dobra, najbardziej efektywną strategią jest aukcja, a najmniej loteria i kolejka. W przypadku jednak oceny sprawiedliwości poszczególnych strategii alokacji dobra, respondenci wskazywali właśnie loterię i kolejkę jako najbardziej sprawiedliwe podejścia.

⁴⁵ Inną doktryną jest tzw. *sprawiedliwość dystrybutywna* (ang. distributive equity, distributive justice), która jest związana z analizą rozkładu dochodu w gospodarce (Biggar 2010). W tym wątku rozważań ekonomicznych analizuje się m.in. rozkład możliwości siły nabywczej, uwzględniając przede wszystkim najbiedniejszych członków społeczności (Muir 2001).

⁴⁶ Warto wskazać również, że sprawiedliwość jako wartość, którą można ocenić (np. czy dana sytuacja, zachowanie były sprawiedliwe, czy czujemy, że *dotknęła nas niesprawiedliwość*) nie jest wyłącznie odbierana przez ludzi. Przeprowadzono wiele eksperymentów, w ramach których wskazano, zwierzęta

jest dalej sprawiedliwością – przykładem są sytuacje, w których warunek sprawiedliwości poprzez brak zazdrości jest spełniony, ale większość ludzi uważa, że taki podział dóbr jest niesprawiedliwy (Konow 2003).

W niniejszym podrozdziale przedstawiono definicję stronniczości algorytmicznej, wraz z odniesieniem jej do definicji dyskryminacji pośredniej i bezpośredniej – pojęć występujących w regulacjach antydyskryminacyjnych. Ponadto wskazano na potencjalne źródła powstawania stronniczości algorytmicznej. Pojęcie stronniczości algorytmicznej będzie w dalszej części rozprawy rozpatrywane w kontekście modeli oceny zdolności kredytowej. Z racji silnego ugruntowania pojęcia dyskryminacji w prawie, zarówno międzynarodowym, jak i krajowym, w kolejnym podrozdziale zestawiono podsumowanie treści wybranych aktów prawnych dotyczących dyskryminacji.

3.2 Prawo dotyczące równości i zakazu dyskryminacji

Polskie banki realizujące ocenę zdolności kredytowej konsumenta są zobligowane do przestrzegania norm i regulacji prawnych (Blicharz i in. 2019).

Opierając się na syntezie prawodawstwa krajowego, europejskiego i międzynarodowego, przedstawionej w niniejszym podrozdziale, w kontekście rozważań dotyczących sprawiedliwości i zapisów antydyskryminacyjnych związanych z modelami decyzyjnymi, w tym również modelami wykorzystywanymi do oceny zdolności kredytowej, można wyróżnić:⁴⁷

- zapisy przeciwdziałające *dyskryminacji w dostępie*, nakładające obowiązek równości w dostępie do usług (np. równe koszty transakcyjne, równy dostęp do kredytów);
- zapisy przeciwdziałające *dyskryminacji w ocenie*, nakładające obowiązek do równego traktowania ludzi ze względu na cechy chronione (np. modele oceny zdolności kredytowej nie powinny dyskryminować ludzi m.in. ze względu na płeć, pochodzenie etniczne).

W dalszej części rozprawy przedstawiono zestawienie wybranych aktów prawnych, które odnoszą się do zakazu dyskryminacji.

potrafią ocenić zachowania niesprawiedliwe i niejednokrotnie zareagować w sposób jednoznaczny agresją lub zachowaniem wskazującym na podirytowanie, zdenerwowanie (Brosnan i De Waal 2003). Wyniki te mogą sugerować, że kryteria sprawiedliwości są uniwersalne lub ponad gatunkowe (Brosnan i De Waal 2012).

⁴⁷ Pojęcia te odpowiednio związane są z pojęciami dyskryminacji pośredniej i dyskryminacji bezpośredniej.

3.2.1 Prawo krajowe

Konstytucja RP

Konstytucja Rzeczypospolitej Polskiej, uchwalona 2 kwietnia 1997 r. (wraz ze zmianami – ostatnia miała miejsce w 2009 r.) jest najwyższym aktem prawnym Polski, tzw. ustawą zasadniczą (Stelmach 1993). Sprawiedliwość i równość płci, definiowana z perspektywy prawnej, jest przedstawiona w ustawie zasadniczej, w art. 32 wskazującym, że:

Art. 32. [Zasada równości; zakaz dyskryminacji]

1. Wszyscy są wobec prawa równi. Wszyscy mają prawo do równego traktowania przez władze publiczne.
2. Nikt nie może być dyskryminowany w życiu politycznym, społecznym lub gospodarczym z jakiejkolwiek przyczyny. Wszyscy są wobec prawa równi. Wszyscy mają prawo do równego traktowania przez władze publiczne.

Konstytucja wskazuje na równość wszystkich wobec prawa oraz na prawo do równego traktowania. O ile art. 32 nie precyzuje cech, które stanowią atrybuty chronione prawnie przed dyskryminacją (tj. pozostawiając otwarty katalog), o tyle już art. 33 ustawy zasadniczej zawęża perspektywę do płci, wskazując, że:

Art. 33. [Zasada równości kobiet i mężczyzn]

1. Kobieta i mężczyzna w Rzeczypospolitej Polskiej mają równe prawa w życiu rodzinnym, politycznym, społecznym i gospodarczym.
2. Kobieta i mężczyzna mają w szczególności równe prawo do kształcenia, zatrudnienia i awansów, do jednakowego wynagradzania za pracę jednakowej wartości, do zabezpieczenia społecznego oraz do zajmowania stanowisk, pełnienia funkcji oraz uzyskiwania godności publicznych i odznaczeń.

Przytoczone fragmenty Konstytucji RP stanowią najistotniejsze artykuły z perspektywy zasad antydyskryminacyjnych (Łojkowska i Pawłęga 2014). Odnoszą się one bezpośrednio do płci jako atrybutu, który nie powinien różnicować ludzi względem prawa w: „życiu rodzinnym, politycznym, społecznym i gospodarczym”. Konstytucja RP jest podstawowym, krajowym aktem prawnym, który ze swojej definicji porządkuje i ustanawia najważniejsze prawa i obowiązki obywateli, w tym zakazuje jakichkolwiek praktyk dyskryminacyjnych (Burek i Klaus 2013).

Rzecznik Praw Obywatelskich

Instytucja Rzecznika Praw Obywatelskich RP wskazuje na następujące cechy ludzi, które mogą być zagrożone dyskryminacją (RPO RP 2021): płeć, rasę, pochodzenie etniczne, narodowość, religię, wyznanie, światopogląd, niepełnosprawność, wiek lub

orientację seksualną. Katalog zmiennych chronionych prawnie jest zatem dłuższy w porównaniu do Konstytucji RP, i obejmuje zmienne, które mogą nie być łatwo obserwowalne (np. wyznanie, orientacja seksualna) lub jednoznacznie zdefiniowane (np. światopogląd).

Zagadnienie związane z brakiem obserwowalności tych cech jest istotne z perspektywy procesu i modelu oceny zdolności kredytowej, ponieważ nie są łatwo identyfikowalne na etapie zbierania danych kredytobiorcy oraz prawnie muszą być chronione przed dyskryminacją. Co więcej, zmienne te mogą być skorelowane z innymi cechami behawioralnymi (np. wydatki miesięczne i wyznanie, liczba istniejących innych kredytów danego kredytobiorcy i światopogląd). W przypadku występowania takiej korelacji konstruowany model oceny zdolności kredytowej, którego specyfikacja zakłada wyłączenie tych zmiennych chronionych, przez mechanizm korelacji z pozostałymi zmiennymi będzie dalej uwzględniać zmienność atrybutów prawnie chronionych⁴⁸.

Istotnym wątkiem w tym zakresie są również interakcje, jakie występują między współistniejącymi regulacjami. Przykładem może być np. ochrona przed dyskryminacją osób niepełnosprawnych, która, w wyniku współistnienia i współzależności z innymi ograniczeniami prawnymi, może prowadzić do pośredniej dyskryminacji. Sytuacja taka zachodziła dla obecnie uchylonego art. 80 Kodeksu Cywilnego, który brzmiał:

Art. 80.

Jeżeli osoba niemogąca czytać ma złożyć oświadczenie woli na piśmie, oświadczenie powinno być złożone w formie aktu notarialnego.

Ratio legis tego przepisu była ochrona praw i interesów zarówno osób niemogących czytać (np. niewidomych lub niedowidzących) jak również ich kontrahentów. Odczytujący zapisy pisemnej umowy, mógłby zmienić treść takiego oświadczenia, ukrywając tym samym szkodliwe zapisy umowne; inną sytuacją jest również przypadek, w którym osoba niemogąca czytać, mogłaby intencjonalnie podważyć zapis umowny, wskazując przy tym, że jego treść została zniekształcona podczas czytania jego treści. Zapis ten oznacza, że ochronie podlegają zarówno osoby niemogące czytać, jak również ich kontrahenci (Niezbecka i Kidyba 2009). W przypadku usług bankowych ustawa Prawo Bankowe zakłada wymóg formy pisemnej do korzystania z rachunku bankowego (art. 52), kredytu (art. 69).

Oznaczało to w praktyce, że osoba niewidoma zmuszona była do przedstawiania aktów notarialnych w przypadku chęci korzystania z większości popularnych usług bankowych.

⁴⁸ Przykładem sytuacji, w której wykluczenie zmiennej chronionej nie jest efektywną strategią jest budowa modelu oceny zdolności kredytowej (np. model oparty na danych aplikacyjnych), w którym obok szeregu innych zmiennych egzogenicznych występują zmienne: **pochodzenie etniczne** oraz **kod pocztowy** (Chen 2018). Zmienna **pochodzenie etniczne** jest zmienną prawnie chronioną, i jest skorelowana ze zmienną **kod pocztowy** (George 2018; LaFrance 2014). Wykluczając wyłącznie zmienną prawnie chronioną – **pochodzenie etniczne** – a pozostawiając inne zmienne skorelowane z wyłączonej zmienną (np. **kod pocztowy**) bank naraża się na powstanie *problemu zmiennej pominiętej* (ang. omitted variable problem), który skutkuje powstaniem obciążenia na innych zmiennych (Clarke 2005).

Koszt związany ze stawką taksy notarialnej pozostaje w relacji do wartości przedmiotu czynności; koszty prawne dla osób niewidomych stanowiłyby istotne ograniczenie w korzystaniu z usług bankowych, co oznaczałoby występowanie dyskryminacji pośredniej.

Ustawa antydyskryminacyjna – ustawa 2010/12/03

Głównym celem ustawy antydyskryminacyjnej – ustawa 2010/12/03 – jest przeciwdziałanie dyskryminacji i ochrona przed nią, a także precyzowanie, kiedy występuje przestępstwo nierównego traktowania, a kiedy takie zachowanie jest wręcz dopuszczalne: np. ubezpieczyciel może udzielić zniżki na ubezpieczenie samochodu kobiecie, jeśli z badań wynika, że kobiety powodują mniej wypadków drogowych (Małocha-Krupa i in. 2013). W ustawie tej zdefiniowano pojęcie dyskryminacji, określając jej możliwe formy, oraz wskazując na kwestie procesowe ułatwiające postępowanie o naruszenie zasady równego traktowania. Ustawa 2010/12/03 wskazuje następujące definicje, przytaczane w rozprawie:

Art. 3. Ilekroć w ustawie jest mowa o:

- 1) dyskryminacji bezpośredniej – rozumie się przez to sytuację, w której osoba fizyczna ze względu na płeć, rasę, pochodzenie etniczne, narodowość, religię, wyznanie, światopogląd, niepełnosprawność, wiek lub orientację seksualną jest traktowana mniej korzystnie niż jest, była lub byłaby traktowana inna osoba w porównywalnej sytuacji;
- 2) dyskryminacji pośredniej – rozumie się przez to sytuację, w której dla osoby fizycznej ze względu na płeć, rasę, pochodzenie etniczne, narodowość, religię, wyznanie, światopogląd, niepełnosprawność, wiek lub orientację seksualną na skutek pozornie neutralnego postanowienia, zastosowanego kryterium lub podjętego działania występują lub mogłyby wystąpić niekorzystne dysproporcje lub szczególnie niekorzystna dla niej sytuacja, chyba że postanowienie, kryterium lub działanie jest obiektywnie uzasadnione ze względu na zgodny z prawem cel, który ma być osiągnięty, a środki służące osiągnięciu tego celu są właściwe i konieczne⁴⁹;
- 5) nierównym traktowaniu – rozumie się przez to traktowanie osób fizycznych w sposób będący jednym lub kilkoma z następujących zachowań: dyskryminacją bezpośrednią, dyskryminacją pośrednią, molestowaniem, molestowaniem seksualnym, a także mniej korzystnym traktowaniem osoby fizycznej wynikającym z odrzucenia molestowania lub molestowania seksualnego lub podporządkowania się molestowaniu lub molestowaniu seksualnemu, oraz zachęcanie do takich zachowań i nakazywanie tych zachowań.

⁴⁹ W rozprawie nie podjęto tematu problemu molestowania lub molestowania seksualnego.

3.2.2 Prawo Unii Europejskiej

Równość wobec prawa i brak dyskryminacji zapisano w prawie Unii Europejskiej, w zapisach znajdujących się w Traktacie o Unii Europejskiej:⁵⁰

TYTUŁ III [RÓWNOŚĆ]

Artykuł 20 [Równość wobec prawa]

Wszyscy są równi wobec prawa.

Artykuł 21 [Niedyskryminacja]

1. Zakazana jest wszelka dyskryminacja w szczególności ze względu na płeć, rasę, kolor skóry, pochodzenie etniczne lub społeczne, cechy genetyczne, język, religię lub przekonania, poglądy polityczne lub wszelkie inne poglądy, przynależność do mniejszości narodowej, majątek, urodzenie, niepełnosprawność, wiek lub orientację seksualną.

Dyrektywa 2000/43/WE

Jednym ze wcześniejszych dokumentów regulujących kwestie sprawiedliwości jest dyrektywa Rady 2000/43/WE z dnia 29 czerwca 2000 r. wprowadzająca w życie zasadę równego traktowania osób bez względu na pochodzenie rasowe lub etniczne (dyrektywa 2000/43/WE), wskazująca że:

Unia Europejska odrzuca wszystkie teorie zmierzające do określenia istnienia odrębnych ras ludzkich. Użycie terminu „pochodzenie rasowe” w niniejszej dyrektywie nie oznacza wcale uznania tych teorii.

Dyrektywa ta wskazuje na pojęcie dyskryminacji rasowej, przyjmując następującą kategoryzację⁵¹:

Artykuł 2 [Pojęcie dyskryminacji]

1. Do celów niniejszej dyrektywy zasada równego traktowania oznacza brak jakichkolwiek form bezpośredniej lub pośredniej dyskryminacji ze względu na pochodzenie rasowe lub etniczne.
2. Do celów ust. 1:
 - a) dyskryminacja bezpośrednia ma miejsce, gdy ze względu na pochodzenie rasowe lub etniczne osoba traktowana jest mniej przychylnie niż traktuje się,

⁵⁰ Pełen zakres ochrony wskazany w Traktacie o Unii to: Równość wobec prawa, Niedyskryminacja, Różnorodność kulturowa, religijna i językowa, Równość kobiet i mężczyzn, Prawa dziecka, Prawa osób w podeszłym wieku, oraz Integracja osób niepełnosprawnych.

⁵¹ Warto zwrócić uwagę, że zaprezentowana klasyfikacja jest przykładem ogólnej definicji dyskryminacji, wyróżniającej: dyskryminację bezpośrednią, dyskryminację pośrednią, molestowanie i molestowanie seksualne. W przypadku dyrektywy 2000/43/WE cechą prawnie chronioną jest *pochodzenie rasowe lub etniczne*.

traktowano lub traktowano by inną osobę w podobnej sytuacji;

b) dyskryminacja pośrednia ma miejsce, gdy pozornie neutralny przepis, kryterium lub praktyka mogą doprowadzić do szczególnie niekorzystnej sytuacji dla osób danego pochodzenia rasowego lub etnicznego w stosunku do innych osób, chyba że taki przepis, kryterium lub praktyka są obiektywnie uzasadnione legalnym celem, a środki mające służyć osiągnięciu tego celu są odpowiednie i konieczne.

Dyrektywa 2004/113/WE

Kluczowym dokumentem regulacyjnym w zakresie sprawiedliwości traktowania ze względu na płeć na obszarze Unii Europejskiej jest dyrektywa Rady 2004/113/WE z dnia 13 grudnia 2004 r. wprowadzająca w życie zasadę równego traktowania mężczyzn i kobiet w zakresie dostępu do towarów i usług oraz dostarczania towarów i usług (dyrektywa 2004/113/WE). Celem dyrektywy jest utworzenie ram legislacyjnych do przeciwdziałania dyskryminacji wynikającej z płci w zakresie dostępu do towarów i usług.

Pojęcie dyskryminacji, jest różnicowane przez zapisy dyrektywy na następujące podtypy:⁵²

- *Dyskryminacja bezpośrednia* – jest związana z sytuacją, w której dana osoba jest traktowana w sposób mniej korzystny ze względu na płeć, niż gdyby była traktowana inna osoba w porównywalnej sytuacji; przykładem dyskryminacji bezpośredniej jest bezpośrednio uwzględnienie płci jako cechy dyskryminującej np. w rekrutacji zawodowej lub w procesie oceny zdolności kredytowej.
- *Dyskryminacja pośrednia* – związana jest neutralnym przepisem, kryterium lub praktyką, które stawiają reprezentanta danej płci w szczególnie niekorzystnym położeniu względem osób o innej płci⁵³; dyskryminacją pośrednią jest np. wymaganie 10 letniego stażu pracy w ogłoszeniach o prace – osoby młodsze pomimo posiadania kwalifikacji będą wykluczone z tego procesu.
- *Molestowanie* – wystąpienie niepożądanego zachowania związanego z płcią danej osoby, a której celem lub efektem jest naruszenie godności osoby oraz stworzenie onieśmielającej, upokarzającej, poniżającej, wrogiej lub agresywnej atmosfery.
- *Molestowanie seksualne* – wystąpienie niepożądanego zachowania (w tym zachowanie fizyczne o charakterze seksualnym), werbalne lub niewerbalne związane z płcią danej osoby, a której celem lub efektem jest naruszenie godności osoby oraz stworzenie onieśmielającej, upokarzającej, poniżającej, wrogiej lub agresywnej atmosfery.

⁵² Wskazane podtypy nie wykluczają się wzajemnie; np. zgodnie z zapisami dyrektywy 2004/113/WE, molestowanie seksualne może być szczególnym przypadkiem dyskryminacji bezpośredniej.

⁵³ Wyjątkiem tej sytuacji jest przepis, kryterium lub praktyka, której środki osiągnięcia obiektywnego celu są zgodne z prawem, są właściwe i niezbędne.

Zagadnienie wykorzystania zmiennej płeć w modelach ubezpieczeniowych i związanych usług finansowych jest bezpośrednio poruszony w art. 5 ust. 1 dyrektywy, który wskazuje, że:

1. Państwa Członkowskie zapewniają, że we wszystkich nowych umowach zawartych najpóźniej po 21 grudnia 2007 r. użycie płci jako czynnika w kalkulowaniu składek i świadczeń do celów ubezpieczenia i związanych usług finansowych nie powoduje różnic w składkach odszkodowaniach poszczególnych osób.

Dodatkowo, w tej dyrektywie wskazano jednak zapis w art. 5 ust. 2, który umożliwił wcześniejsze wykorzystanie płci gdy jej użycie „jest czynnikiem decydującym w ocenie ryzyka opartego na odpowiednich i dokładnych danych aktuarialnych i statystycznych”.

Wytyczne w zakresie etyki z 2019 r.

W kwietniu 2019 r., niezależna grupa ekspertów wysokiego szczebla do spraw sztucznej inteligencji, powołana przez Komisję Europejską, opublikowała Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji. Celem przygotowania wskazanych wytycznych jest promowanie rozwiązań godnej zaufania sztucznej inteligencji jako mechanizmów analitycznych posiadających trzy cechy, tj.:

- 1) Zgodność z prawem – zapewnienie pokrycia i zgodności z obowiązującymi przepisami prawa.
- 2) Przestrzeganie zasad i wartości etycznych – zapewnienie zgodności z zasadami i normami etycznymi.
- 3) Solidność techniczna i społeczna – zapewnienie, że rozwiązania wykorzystujące systemy sztucznej inteligencji nie wywołują żadnych niezamierzonych szkód, działają w sposób bezpieczny pewny i niezawodny.

Komisja Europejska wskazuje, że wykorzystanie systemów sztucznej inteligencji w celach wypracowania przewagi konkurencyjnej powinno wiązać się dążeniem do maksymalizacji korzyści ze stosowania takich systemów, przy jednoczesnym ograniczaniu zagrożeń wynikających z ich stosowania. Zapisy dotyczące sprawiedliwości są obecne w każdej z cech, które Komisja Europejska przedstawia w przytoczonych wytycznych. Jeżeli od modelu sztucznej inteligencji wymaga się zgodności z prawem, to zapisy dotyczące przeciwdziałania dyskryminacji, które obecnie funkcjonują i zostały częściowo opisane w niniejszej pracy, są w tym punkcie również obecne. W ramach wytycznych wskazuje się, że sztuczna inteligencja powinna być opracowywana w taki sposób, aby przestrzegane były zasady i normy etyczne.

Definicja sprawiedliwości wskazana w wytycznych Komisji Europejskiej odnosi się do dwóch wymiarów: materialnego oraz proceduralnego⁵⁴.

Wymiar *materialny* odnosi się do zapewnienia sprawiedliwego i równego podziału kosztów i korzyści wynikających z zastosowania algorytmów sztucznej inteligencji oraz zapewnienia braku stronniczości, dyskryminacji oraz stygmatyzacji. Definicja ta odnosi się również do zapewnienia, że system sztucznej inteligencji będzie: „[...] wspierać równe szanse w zakresie dostępu do edukacji, towarów, usług i technologii”.

Wymiar *proceduralny* wskazuje na możliwość podważenia decyzji i rekomendacji wskazywanych przez systemy sztucznej inteligencji oraz skutecznego dochodzenia roszczeń wynikających z tych decyzji. W tym zakresie odnosi się również do identyfikacji podmiotu odpowiedzialnego za decyzje podejmowane przez modele decyzyjne oraz konieczności zapewnienia, aby procesy decyzyjne mogły być wytłumaczalne.

Wytyczne Komisji Europejskiej odnoszą się do zasad zachowania różnorodności, niedyskryminacji i sprawiedliwości, poprzez wskazanie, że kwestie włączenia społecznego i różnorodności powinny być uwzględniane przez cały cykl życia modelu. Wskazują również na konieczność unikania niesprawiedliwej stronniczości, które mogą wynikać z: uwzględnienia obciążonej historycznej próby obserwacji, oraz niekompletności i złego zarządzania modelem. Dotyczą one bezpośrednio możliwych skutków takiej stronniczości – tj. powstania bezpośredniej i pośredniej dyskryminacji wobec określonych grup lub osób.

3.2.3 Organizacja Narodów Zjednoczonych

W Preambule do Karty Narodów Zjednoczonych z 1945 r. wpisano wezwanie do równouprawnienia mężczyzn i kobiet, a wśród celów ONZ wskazano działanie na rzecz praw człowieka:

My, Ludy Narodów Zjednoczonych, zdecydowane

- uchronić przyszłe pokolenia od klęsk wojny, która dwukrotnie za naszego życia wyrządziła ludzkości niewypowiedziane cierpienia,
- przywrócić wiarę w podstawowe prawa człowieka, w godność i wartość człowieka, w równouprawnienie mężczyzn i kobiet, w równość narodów dużych i małych,

[...]

Organizacja Narodów Zjednoczonych w ramach trzecich obrad wskazała przepisy zakazujące dyskryminację w dostępie do praw człowieka w Powszechnej deklaracji praw człowieka z 1948 r. (Preambuła, art. 2, art. 7), a więc pierwszym dokumencie

⁵⁴ Wytyczne wskazują, że definicja sprawiedliwości posiada „wiele różnych interpretacji”, co zgodnie z przedstawionymi w rozprawie argumentami, wskazuje, że pojęcie sprawiedliwości nie jest jednorodne oraz umożliwia niejednakową interpretację.

międzynarodowym zawierającym ogólny katalog praw człowieka. W momencie jego ustanawiania, był on niewiążący prawnie – dziś większość postanowień deklaracji jest traktowana jako wiążące prawo zwyczajowe.

Na zapisach Powszechnej deklaracji praw człowieka wzorowano się przyjmując na forum ONZ w 1966 r. dwie kolejne umowy międzynarodowe: Międzynarodowy pakt praw obywatelskich i politycznych oraz Międzynarodowy pakt praw gospodarczych, społecznych i kulturalnych. Umowy te w zamierzeniu miały przekształcić postanowienia Powszechnej deklaracji w obowiązujące normy prawne. Ogólne klauzule antydyskryminacyjne zawarto w art. 2 obu paktów.

3.2.4 Organizacja Narodów Zjednoczonych do spraw Oświaty, Nauki i Kultury

W ramach Organizacji Narodów Zjednoczonych do spraw Oświaty, Nauki i Kultury (ang. United Nations Educational, Scientific and Cultural Organization – UNESCO), której Polska jest członkiem, opracowano konwencje związane z prawem zabraniającym dyskryminacji:

- Konwencja w sprawie zwalczania dyskryminacji w dziedzinie oświaty, sporządzona w Paryżu dnia 15 grudnia 1960 r. (Dz.U. 1964 nr 40 poz. 268).
- Międzynarodowa konwencja w sprawie likwidacji wszelkich form dyskryminacji rasowej, otwarta do podpisu w Nowym Jorku dnia 7 marca 1966 r. (Dz.U. 1969 nr 25 poz. 187).
- Konwencja w sprawie likwidacji wszelkich form dyskryminacji kobiet przyjęta przez Zgromadzenie Ogólne Narodów Zjednoczonych dnia 18 grudnia 1979 r. (Dz.U. 1982 nr 10 poz. 71).
- Konwencja o prawach osób niepełnosprawnych, sporządzona w Nowym Jorku dnia 13 grudnia 2006 r. (Dz.U. 2012 poz. 1169).

W umowach zawierających ogólne katalogi praw człowieka występują otwarte katalogi cech chronionych, tj.: płeć, rasa, kolor skóry, język, religia/wyznanie, poglądy polityczne lub inne przekonania, narodowość, pochodzenie społeczne, pochodzenie narodowe i etniczne, niepełnosprawność, stan cywilny, sytuacja majątkowa, urodzenie, przynależność do mniejszości narodowych.

Biuro urzędu prezydenta USA

W 2016 r. biuro urzędu prezydenta USA opublikowało raport „Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights”, w którym wskazano na wpływ postępu technologicznego na życie gospodarcze, społeczne i obywatelskie (Biuro

Wykonawcze Prezydenta USA 2016). Zapisano w nim, że odpowiednie wykorzystanie rozwiązań zaawansowanych metod analitycznych i big data umożliwi przezwycięzenie długotrwałych uprzedzeń i dyskryminacji.

W raporcie przedstawiono bezpośrednio przykład oceny zdolności kredytowej, jako przykład obszaru szczególnie eksponowanego na zagrożenia sprawiedliwości, jak również potencjału wykorzystania rozwiązań Big Data. Zgodnie z danymi Consumer Financial Protection Bureau (CFPB) ok. 11% Amerykanów jest *niewidzialna kredytowo* (ang. credit invisible), tj. nie posiada wystarczającej liczby danych, w szczególności danych behawioralnych o przeszłych spłatach kredytowych, żeby można było ocenić ich zdolność kredytową (Brevoort, Grimm i Kambara 2015). W CFPB wskazano, ten problem dotyka przede wszystkim Amerykanów, zamieszkujących biedniejsze obszary, tj. ok. 30% z nich nie posiada odpowiednich danych w systemach biur kredytowych, żeby mogła uzyskać pozytywną ocenę zdolności kredytowych. Dodatkowo, zgodnie z CFPB Afroamerykanie i Latynosi są dużo bardziej narażeni na problem niewidzialności kredytowej (15% z nich jest na to narażona) w porównaniu z osobami o białym kolorze skóry (9%).

3.2.5 Cechy prawnie chronione i cechy wrażliwe

W ramach wskazanej wcześniej prezentacji prawa równościowego oraz regulacji antydyskryminacyjnych wykorzystywano zamiennie pojęcia cech prawnie chronionych⁵⁵ oraz cech wrażliwych⁵⁶.

Cecha prawnie chroniona to cecha osób fizycznych, na podstawie której normatywnie zabronionym jest różnicowanie sytuacji tych osób (RPO 2020). Pojęcie *cechy wrażliwej* dotyczy informacji o osobie fizycznej, która to informacja podlega szczególnej ochronie; pojęcie cech wrażliwych zostało wyłonione w kontekście kategorii danych osobowych, które wymagają szczególnej ochrony.

Pojęcie cechy prawnie chronionej jest pojęciem szerszym definicyjnie od pojęcia cech wrażliwych. Na grupę cech wrażliwych wskazuje Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (RODO). W RODO wymienia się następujące kategorie danych podlegających szczególnej ochronie: pochodzenie rasowe lub etniczne, poglądy polityczne, przekonania religijne lub światopoglądowe, przynależność do związków zawodowych, dane genetyczne, dane biometryczne (wykorzystywane w celu jednoznacznego zidentyfikowania osoby fizycznej), dane dotyczące zdrowia, seksualność lub orientacja seksualna.

⁵⁵ Analogiczne pojęcie funkcjonuje również w innych jurysdykcjach: w Stanach Zjednoczonych są to tzw. protected group lub protected class, w Kanadzie są nimi prohibited grounds.

⁵⁶ Podobne rozróżnienie istnieje w innych jurysdykcjach, gdzie wyróżnia się tzw. sensitive features.

Wskazuje się również brak możliwości przetwarzania danych osobowych wrażliwych, o ile nie jest spełniona jedna z przesłanek wskazanych w art. 9 ust. 3 RODO, m.in.:

- Osoba, której dane dotyczą, wyraziła wyraźną zgodę na przetwarzanie tych danych osobowych w jednym lub kilku konkretnych celach, chyba że prawo Unii lub prawo państwa członkowskiego UE przewidują, iż osoba, której dane dotyczą, nie może uchylić zakazu.
- Przetwarzanie jest niezbędne do wypełnienia obowiązków i wykonywania szczególnych praw przez administratora lub osobę, której dane dotyczą, w dziedzinie prawa pracy, zabezpieczenia społecznego i ochrony socjalnej, o ile jest to dozwolone prawem Unii lub prawem państwa członkowskiego UE lub porozumieniem zbiorowym na mocy prawa państwa członkowskiego UE, przewidującymi odpowiednie zabezpieczenia praw podstawowych i interesów osoby, której dane dotyczą.
- Przetwarzanie dotyczy danych osobowych w sposób oczywisty upublicznionych przez osobę, której dane dotyczą.

Regulacje antydyskryminacyjne funkcjonują w prawie, krajowym i międzynarodowym od wielu lat jako akty normatywne. Na potwierdzenie tej tezy można wskazać Kartę Narodów Zjednoczonych z 1945 r., która pomimo że jest *umową międzynarodową*, jest traktowana przez kraje sygnatariusze jako *prawo*. Przepisy prawne stanowiące ogólne ramy (tj. nieodnoszące się bezpośrednio do konkretnego kontekstu gospodarczego) odnajdują swoje odzwierciedlenie w przepisach i rekomendacjach szczegółowych, m.in. ustawa z dnia 3 grudnia 2010 r. o wdrożeniu niektórych przepisów Unii Europejskiej w zakresie równego traktowania, Equal Credit Opportunity Act, Fair Housing Amendments Act lub Opinion 11/2021 on the Proposal for Directive on customer credits. Przykładem takiej transpozycji prawa ogólnego na regulacje domenowe (modele oceny zdolności kredytowej) może być RODO; pomimo że samo RODO nie odnosi się do działania modeli oceny zdolności kredytowej, to jest ono wykorzystywane pośrednio przez kolejne przepisy i opinie prawne jako wykładnia w zakresie definicji cech prawnie chronionych (art. 9 RODO) dla obszaru modeli oceny zdolności kredytowej (EDPS).

W kolejnym podrozdziale uwagę przeniesiono na kwestie metodyczne związane identyfikacją oraz redukcją problemu stroniczości algorytmicznej w modelach oceny zdolności kredytowej.

3.3 Stronniczość algorytmiczna w modelach oceny zdolności kredytowej

W niniejszym podrozdziale przybliżono zagadnienia stronniczości algorytmicznej w modelach oceny zdolności kredytowej. *Sprawiedliwość* w kontekście modeli oceny zdolności kredytowej jest rozumiana jako brak występowania problemu stronniczości algorytmicznej względem poszczególnych grup osób definiowanych na podstawie cech prawnie chronionych.

3.3.1 Obciążenie estymatorów a stronniczość algorytmiczna

Jednym z podstawowych sposobów oceny parametrów oszacowanego modelu ekonometrycznego jest ocena statystycznej obciążoności estymatora modelu (Taylor 2020). Jeżeli estymator danego parametru jest obciążony, to wartość oczekiwana szacowanego parametru nie jest równa wartości rzeczywistej tego parametru, tj. $E[\hat{\beta}_i] \neq \beta_i$, gdzie β_i jest parametrem oszacowanego, parametrycznego modelu ekonometrycznego. Obciążenie estymatorów modelu może się pojawić w przypadku m.in. gdy występuje współliniowość zmiennych egzogenicznych, występuje problem zmiennej pominiętej lub bardziej ogólnie: estymowany model ma błędną specyfikację.

Model oceny zdolności kredytowej, charakteryzujący się występowaniem problemu statystycznej obciążoności estymatorów, będzie zawyżał lub zaniżał rzeczywisty wpływ zmiennych na ocenę zdolności kredytowej. Problem ten może być na tyle silny, że odwrócony zostanie znak stojący przy oszacowanym parametrze (np. w przypadku regresji logistycznej oszacowany parametr zmiennej *dochód* może mieć ujemny znak co będzie interpretowane jako: „czym większy dochód, tym mniejsza zdolność do spłaty kredytu”). Zagadnienie statystycznego obciążenia estymatorów jest zagadnieniem technicznym, dotyczącym parametrów modelu. Wystąpienie tego rodzaju obciążenia powoduje, że prognozy modelu⁵⁷ mogą być niewiarygodne (np. model systematycznie zaniża lub zawyża prawdopodobieństwo spłaty zadłużenia), a interpretacja parametrów tego modelu będzie niepoprawna (Greene 2000).

W przypadku gdy model oceny zdolności kredytowej przechodzi pomyślnie weryfikację obciążenia estymatorów (tj. brak obciążenia), w modelu może występować problem stronniczości algorytmicznej. Zagadnienie stronniczości algorytmicznej jest zagadnieniem społeczno-ekonomicznym, które może pojawić się wyłącznie przy odpowiednim kontekście społecznym lub prawnym, np. historyczne dane o udzielonych kredytach w niektórych krajach nie będą obejmować kobiet, które nie mogły aplikować o kredyty konsumenckie (Morsy 2020). Stronniczość algorytmiczna jest zatem utożsamiana z pojęciem

⁵⁷ W kontekście oceny zdolności kredytowej jest to klasyfikacja na klienta, który spłaci oraz tego, który nie spłaci kredytu.

dyskryminacji: w zależności od konkretnego kontekstu regulacyjnego można ocenić dane działanie modelu oceny zdolności kredytowej jako dyskryminujące.

3.3.2 Kryteria oceny sprawiedliwości

Poniżej przedstawiono trzy główne kryteria oceny sprawiedliwości, które można wykorzystać w kontekście oceny zdolności kredytowej. Z każdym ze wskazanych kryteriów oceny sprawiedliwości jest związana definicja sprawiedliwości oraz miara stronniczości algorytmicznej. Miary te mają indywidualne interpretacje ekonomiczne, które w zestawieniu z kontekstem przedmiotowym – modele oceny zdolności kredytowej – determinują swoją użyteczność praktycznego zastosowania (Kozodoi i Lessmann 2022).

Przyjęta notacja

W celu zdefiniowania podstawowych miar stronniczości algorytmicznej posłużono się następującą notacją inspirowaną pracą Kozodoi i Lessmann (2022). Niech X oznacza rzeczywistą zmienną losową k -wymiarową opisującą atrybuty związane z daną aplikacją kredytową, a $x \in \mathbb{R}^k$ oznacza realizację zmiennej losowej X . Atrybutami tymi mogą być np. cechy opisujące konkretną aplikację kredytową otrzymaną przez bank: wiek, płeć lub wielkość dochodów kredytobiorcy. Niech Y oznacza zmienną losową wskazującą na przyszły stan kredytobiorcy, a $y \in \{0, 1\}$ oznacza realizację zmiennej losowej Y . Przyjmujemy, że przyszły stan kredytobiorcy może przyjmować dwa poziomy: dobry ($y = 1$) lub zły ($y = 0$). Zgodnie z powyższym przestrzeń probabilistyczna rozważana w dalszej części rozprawy opisana jest przez rozkład łączny (X, Y) .

Bank podejmuje decyzję kredytową opierając się na modelu oceny zdolności kredytowej zdefiniowanym jako funkcja $s : \mathbb{R}^k \rightarrow [0; 1]$:

$$s(x) = \mathbb{P}[Y = 1|x],$$

której wartości interpretuje się jako prawdopodobieństwo, że stan kredytobiorcy będzie dobry pod warunkiem, że atrybuty aplikacji kredytowej przyjmują wartość x .

Wartość $s(x)$ można interpretować następująco: im jest większa, tym bardziej prawdopodobny jest fakt, że dany kredytobiorca spłaci swoje zobowiązania kredytowe. W dalszej części wywodu pomija się sposób budowy i estymacji funkcji $s(x)$; przyjmuje się, że taka funkcja została oszacowana i spełnia wskazane wcześniej własności⁵⁸.

Bank wykorzystuje model oceny zdolności kredytowej do podejmowania decyzji kredytowej. W tym celu można przyjąć, że wszystkie wartości prognozy modelu oceny zdolności kredytowej powyżej progu $\tau \in [0, 1]$, implikują akceptację wniosku kredytowego.

⁵⁸ Kaszyński, Kamiński i Szapiro (2020) przedstawili opis metod budowy i estymacji modeli oceny zdolności kredytowej.

Niech $\hat{y}(x) \in 0, 1$ oznacza decyzję podjętą przez bank dla wniosku o atrybutach x . Wtedy zachodzi zależność:

$$\hat{y}(x) = \begin{cases} 1, & s(x) \geq \tau \\ 0, & s(x) < \tau \end{cases},$$

gdzie wartość $\hat{y}(x) = 1$ oznacza pozytywną decyzję kredytową, a $\hat{y}(x) = 0$ oznacza negatywną decyzję kredytową. Dalej przez $\hat{Y} = \hat{y}(X)$ będziemy oznaczali binarną zmienną losową opisującą rozkład decyzji kredytowych podejmowanych przez bank.

Niech X_a oznacza binarną zmienną losową, a $x_a \in \{0, 1\}$ jest realizacją tej zmiennej. Zmienna ta jest jednym z atrybutów X opisujących wniosek kredytowy i oznacza wartość cechy prawnie chronionej. Atrybut ten oznaczany jest symbolem a . W dalszej części tekstu, dla ustalenia uwagi, przyjęto, że cechą tą jest płeć kredytobiorcy, gdzie $x_a = 1$ oznacza kobietę, a $x_a = 0$ mężczyznę⁵⁹. Wartość takiej zmiennej, zgodnie z zapisami regulacyjnymi, nie powinna wpływać na fakt, czy aplikant otrzyma pozytywną lub negatywną decyzję kredytową, tj. dla dowolnej obserwacji, zmiana wartości zmiennej chronionej nie powinna zmieniać prognozy z modelu⁶⁰.

Badając jakość prognostyczną modelu wykorzystuje się *macierz pomyłek* (więcej o ocenie modeli oceny zdolności kredytowej w podrozdziale 2.2). Do wyznaczenia miar stronniczości algorytmicznej wykorzystuje się ogólne podejście *Fair Aware Confussion Table – FACT*, w ramach którego dla grup uprzywilejowanych i nieuprzywilejowanych (różnicowanych przez wartość zmiennej prawnie chronionej) definiuje się osobne macierze pomyłek, a wartości miar stronniczości algorytmicznej wyznaczone są jako funkcje poszczególnych elementów tych macierzy (Kim, Chen i Talwalkar 2020).

Poniżej zaprezentowano matematyczne definicje wcześniej wykorzystywanych pojęć związanych z odmiennym wpływem (dyskryminacja pośrednia) oraz odmiennym traktowaniem (dyskryminacja bezpośrednia) przez model oceny zdolności kredytowej.

⁵⁹ W pracy, dla zapewnienia przejrzystości wyводу, rozważana jest 2-wartościowa cecha prawnie chroniona; podejście jest to analogiczne do publikacji naukowych traktujących o problematyce stronniczości algorytmicznej, patrz: Barocas, Hardt i Narayanan (2023); Kozodoi i Lessmann (2022); Andreeva i Matuszyk (2019); Chouldechova (2017); Pleiss i in. (2017) lub Kleinberg, Mullainathan i Raghavan (2016).

⁶⁰ ECOA–B, 1976 (oficjalna interpretacja zasady: „Kredytodawca nie może dyskryminować wnioskodawcy na podstawie zakazu w jakimkolwiek aspekcie transakcji kredytowej”): „[...] niezależnie od tego, czy praktyka kredytowa jest specyficznie zakazana gdzie indziej w regulacji, traktowanie wnioskodawców w różny sposób na zakazanej podstawie narusza prawo, ponieważ narusza ogólną zasadę. Różne traktowanie na zakazanej podstawie jest nielegalne, niezależnie od tego, czy wynika z świadomego zamiaru dyskryminacji.” [tł. wł.]

Dyrektywa 2004/113/WE, art. 5 ust. 1: „Państwa Członkowskie zapewniają, że we wszystkich nowych umowach zawartych najpóźniej po 21 grudnia 2007 r. użycie płci jako czynnika w kalkulowaniu składek i świadczeń do celów ubezpieczenia i związanych usług finansowych nie powoduje różnic w składkach i odszkodowaniach poszczególnych osób.”

Warunek parytetu demograficznego

Pierwszym z przedstawionych sposobów pomiaru stronniczości algorytmicznej, związanym z definicją dyskryminacji pośredniej (tj. odmienny wpływ), jest tzw. *parytet demograficzny* (ang. demographic parity, statistical parity lub independence). Warunek parytetu demograficznego jest związany z kryterium niezależności, tj. założeniem, że poszczególne kategorie zmiennej chronionej nie powinny różnicować częstotliwości występowania poszczególnych prognoz z modelu (tj. pozytywna lub negatywna decyzja kredytowa). Formalnie warunek ten przyjmuje równe prawdopodobieństwo (częstotliwość) uzyskania pozytywnego wyniku klasyfikacji (tj. $\hat{Y} = 1$), dla wskazanego progu odcięcia τ w każdej z grup cechy prawnie chronionej. Niech model oceny zdolności kredytowej spełnia założenia parytetu demograficznego, wtedy:

$$P \left[\hat{Y} = 1 | X_a = 0 \right] = P \left[\hat{Y} = 1 | X_a = 1 \right]$$

Powyższe równanie wskazuje na brak wpływu wartości zmiennej X_a na oszacowanie prawdopodobieństwa zdarzenia $\hat{Y} = 1$. Wykorzystując powyższą definicję kryterium niezależności można wskazać miarę stronniczości algorytmicznej będącą iloczynem wskazanych wyżej prawdopodobieństw, *Disparate Impact* DI (ang. disparate impact):

$$DI = \frac{P \left[\hat{Y} = 1 | X_a = 1 \right]}{P \left[\hat{Y} = 1 | X_a = 0 \right]} \quad (3.1)$$

Z perspektywy oceny zdolności kredytowej, parytet demograficzny wymaga, aby bank, każdej grupie osób różnicowanych przez zmienną a , udzielał kredytów w takiej samej frakcji (tj. taki sam poziom *wskaźnika zatwierdzeń kredytów*). Kryterium niezależności odnosi się do dyskryminacji rozumianej jako brak odmiennego wpływu modelu na poszczególne grupy społeczne.

Sposoby identyfikacji odmiennego wpływu modelu – brak niezależności – opisano w podrozdziale 3.1.1 (tj. test χ^2 , test Fishera lub zasada czterech–piątych).

Warunek wyrównanych szans

Drugi ze sposobów identyfikacji występowania problemu stronniczości algorytmicznej jest związany z definicją dyskryminacji bezpośredniej (tj. odmienne traktowanie) i odnosi się do *warunku wyrównanych szans* (ang. equalized odds condition, lub separation). Warunek wyrównanych szans odwołuje się do tzw. kryterium separacji, tj. założenia że zmienna celu Y stanowi jeden z możliwych podziałów (separacji) populacji na warstwy o równych prawdopodobieństwach uzyskania akceptacji wniosku kredytowego, bez względu na wartość zmiennej prawnie chronionej, tj.:

$$\begin{aligned} P \left[\hat{Y} = 1 | Y = 1, X_a = 0 \right] &= P \left[\hat{Y} = 1 | Y = 1, X_a = 1 \right] \\ P \left[\hat{Y} = 1 | Y = 0, X_a = 0 \right] &= P \left[\hat{Y} = 1 | Y = 0, X_a = 1 \right] \end{aligned}$$

Warunek wyrównanych szans zakłada parytet błędów prognozy modelu oceny zdolności kredytowej dla poszczególnych grup różnicowanych wartością zmiennej chronionej⁶¹; dla poszczególnych grupy różnicowanych wartością cechy prawnie chronionej prawdopodobieństwo niepoprawnej klasyfikacji, zarówno false positive rate⁶² (FPR) jak i false negative rate⁶³ (FNR) są takie same. W praktyce korzysta się z miary separacji **Separation** (Kozodoi i Lessmann 2022):

$$SP = \frac{|FPR_{X_a=1} - FPR_{X_a=0} + FNR_{X_a=1} - FNR_{X_a=0}|}{2}$$

Można zwrócić uwagę, że dla parametrów $FPR_{X_a=1} = 1$, $FPR_{X_a=0} = 0$, $FNR_{X_a=1} = 0$, oraz $FNR_{X_a=0} = 1$, wartość miary SP będzie wskazywać na model bez dyskryminacji (tj. miara $SP = 0$), a w rzeczywistości model będzie odmiennie traktował klientów $X_a = 1$, oraz $X_a = 0$; w rozprawie zdefiniowano miarę kryterium separacji nie posiadającą takiej własności:

$$SP = \frac{|FPR_{X_a=1} - FPR_{X_a=0}| + |FNR_{X_a=1} - FNR_{X_a=0}|}{2}$$

Warunek parytetu kalibracji

Ostatnim z omawianych w rozprawie sposobów identyfikacji stronniczości algorytmicznej, odnoszącym się również do definicji dyskryminacji bezpośredniej (odmienne traktowanie), jest *kryterium dostateczności* (ang. sufficiency). Kryterium dostateczności związane jest z tzw. *parytetem kalibracji*: taką cechą modelu oceny zdolności kredytowej, że model ten posiada taką samą *jakość kalibracji*⁶⁴ dla każdej z grup definiowanych przez cechę prawnie chronioną (Barocas, Hardt i Narayanan 2023).

W założeniach *parytetu kalibracji* jest to, że model posiada taką samą jakość kalibracji⁶⁵ modelu dla każdej z grup definiowanych przez cechę prawnie chronioną (Barocas, Hardt

⁶¹ Pierwsze równanie warunku wyrównanych szans można zastąpić $TPR = 1 - FNR$, co jest równoważne $P[\hat{Y} = 1 | Y = 1] = 1 - P[\hat{Y} = 0 | Y = 1]$.

⁶² Odsetek fałszywie pozytywnych, tzw. błąd pierwszego rodzaju, definiowany jako: $FPR = P[\hat{Y} = 1 | Y = 0]$.

⁶³ Odsetek fałszywie negatywnych, tzw. błąd drugiego rodzaju, definiowany jako: $FNR = P[\hat{Y} = 0 | Y = 1]$.

⁶⁴ Jakość kalibracji modelu klasyfikacyjnego opisuje poprawność oszacowań prawdopodobieństwa przynależności do poszczególnych klas zmiennej celu. W kontekście oceny zdolności kredytowej jest to oszacowanie prawdopodobieństwa, że klient spłaci zobowiązanie kredytowe. *Jakość kalibracji* jest pojęciem węższym od pojęcia *jakości prognostycznej* modelu, które to obejmuje również pojęcie *mocy dyskryminacyjnej* modelu, tj. możliwości odróżniania klientów spłacających kredyt od tych, którzy kredytu nie spłacą. Możliwym jest, że model posiada wysoką moc dyskryminacyjną – tj. na podstawie prognozy modelu możliwym jest odróżnianie klientów spłacających kredyty – przy jednoczesnej niskiej jakości kalibracji (tj. model niepoprawnie przypisuje prawdopodobieństwa spłaty kredytów).

⁶⁵ Niech $P[Y = 1 | s(X)] = s(X)$ oznacza model o poprawnej kalibracji prawdopodobieństwa. Wtedy parytet kalibracji można zdefiniować jako: $P[Y = 1 | s(X), X_a = 0] = P[Y = 1 | s(X), X_a = 1]$. Więcej o ocenie jakości kalibracji modelu oceny zdolności kredytowej opisano w podrozdziale 2.2.3.

i Narayanan 2023). W wersji słabszej parytetu kalibracji przyjmuje się równość wartości predykcyjnej dodatniej, PPV (ang. positive predictive value):

$$P\left[Y = 1|\hat{Y} = 1, X_a = 0\right] = P\left[Y = 1|\hat{Y} = 1, X_a = 1\right]$$

Miarą stronniczości algorytmicznej wynikającą z kryterium dostateczności może być bezwzględna różnica między wartościami PPV dla poszczególnych grup wartości zmiennych wrażliwych; tj. miara Sufficiency:

$$SF = |PPV_{X_a=0} - PPV_{X_a=1}|$$

Wskazane miary stronniczości algorytmicznej pokrywają zakres definicji pośredniej (kryterium niezależności) jak i bezpośredniej (kryterium separacji i dostateczności), które w ramach projektowanych regulacji równościowych regulator może wykorzystać do identyfikacji występowania problemu stronniczości algorytmicznej w modelach oceny zdolności kredytowej.

Kompromisy między poszczególnymi miarami

Wskazane wcześniej miary stronniczości algorytmicznej umożliwiają jej pomiar, przyjmując inne definicje *sprawiedliwości* działania modelu decyzyjnego. W tabeli 3 (rozdział 2) przedstawiono macierz pomyłek klasyfikatora binarnego. Macierz tą osadźmy w kontekście stronniczości algorytmicznej; podejmiemy próbę odpowiedzi na pytanie, jak poszczególne miary tej stronniczości związane są z różnymi perspektywami jej oceny. Kwestie kompromisu między miarami stronniczości algorytmicznej są szerzej omówione przez Barocas, Hardt i Narayanan (2023). W ramach przeprowadzonej analizy, inspirowanej pracą Biecek (2024), zostaną wykorzystane trzy perspektywy oceny sprawiedliwości procesu kredytowego: zewnętrzny obserwator, klient banku oraz bank (tabela 6).

Tabela 6: Macierz pomyłek – perspektywa obserwatora, banku i klienta

		Perspektywa banku	
		Pozytywna d. kredytowa	Negatywna d. kredytowa
Perspektywa klienta	Y = 1 (spłata)	true positive	false negative (błąd II-rodzaju)
	Y = 0 (brak spłaty)	false positive (błąd I-rodzaju)	true negative

Uwaga: Bez utraty ogólności, przyjęto konwencje definicji kolumn i wierszy zgodnie z Fawcett (2006). Kolorem czerwonym zaznaczono perspektywę banku; kolorem niebieskim przedstawiono perspektywę klienta banku.

Źródło: opracowanie własne.

Pierwsza perspektywa związana jest z *niezależnym obserwatorem* procesu kredytowego realizowanego przez bank. Dotyczy ona wcześniej zdefiniowanego warunku parytetu demograficznego; niezależny obserwator procesu kredytowego, odnotowuje ile aplikacji kredytowych zostało zaakceptowanych przez bank. Z tej perspektywy dyskryminacją pośrednią (odmienny wpływ) będzie sytuacja, gdy zaobserwowany zostanie różny współczynnik zatwierdzeń kredytów dla poszczególnych klas zmiennej chronionej (np. dla kobiet i dla mężczyzn). Perspektywa niezależnego obserwatora jest związana z kryterium niezależności, tj.:

$$\hat{Y} \perp\!\!\!\perp X_a$$

Oznacza to, że decyzja kredytowa $\hat{Y}$, wynikająca z prognozy modelu oceny zdolności kredytowej, powinna być niezależna od wartości zmiennej prawnie chronionej X_a .

Druga perspektywa związana jest z *klientem banku*, który składając aplikację kredytową oczekuje, że jego zdolność kredytowa zostanie poprawnie oceniona: jeżeli jego zdolność kredytowa jest wystarczająca to powinien otrzymać kredyt (w przeciwnym przypadku powinien otrzymać odmowę kredytu). Z perspektywy klienta banku (przyjmując, że jest to klient, który wie, że spłaci kredyt), dyskryminacją będzie sytuacja, w której bank częściej będzie się mylić i będzie podejmować błędną decyzję kredytową dla osób z grupy nieuprzywilejowanej. Osoby, dla których bank będzie się średnio częściej mylił (tj. będzie odmawiał kredytu, pomimo, że są to rzetelni klienci, którzy spłaciliby kredyt) będą *odmiennie traktowane* od osób grupy uprzywilejowanej, dla której błąd ten będzie mniejszy. Perspektywa ta, dotycząca dyskryminacji bezpośredniej modelu oceny zdolności kredytowej – warunku wyrównanych szans (np. „będąc dobrym klientem, czy zostanie poprawnie oceniony”). Sprawiedliwym, dla klienta banku, będzie scenariusz w którym:

$$\hat{Y}|Y \perp\!\!\!\perp X_a$$

Oznacza to, że rozkład warunkowy podejmowanych decyzji $\hat{Y}$ (tj. pozytywna lub negatywna decyzja kredytowa), względem prawdziwego stanu klienta Y (klient spłaci lub nie spłaci kredytu) jest niezależna względem wartości cechy prawnie chronionej.

Trzecią perspektywą analizy procesu kredytowego będzie perspektywa *banku*. Bank wydając decyzję kredytową wie, że nie zna wszystkich aspektów życia klienta, które mają wpływ na jego ocenę zdolności kredytowej (np. faktu, czy dany klient prowadzi zdrowy styl życia: śmierć lub rozwód klienta mogą być stosowane jako przesłanki utraty zdolności kredytowej). Wydając określoną decyzję kredytową, bank antycypuje, że z pewnym prawdopodobieństwem wydana decyzja może być niepoprawna (tj. przyznano kredyt osobie, która nie posiadała zdolności kredytowej). Względem wydawanych decyzji kredytowych $\hat{Y}$, bank odnotowuje, jaka część z tych decyzji została wydana poprawnie, a jaka nie. Perspektywa banku wiąże się z warunkiem parytetu kalibracji modelu

opisanym wcześniej oraz kryterium dostateczności. Sprawiedliwym z perspektywy banku jest scenariusz, w którym częstotliwość mylnie wydawanych decyzji kredytowych przez bank nie zależy od wartości zmiennej chronionej.

$$Y|\hat{Y} \perp\!\!\!\perp X_a$$

Przyjmijmy, że cecha prawnie chroniona *nie jest niezależna* względem faktu spłaty lub nie kredytu⁶⁶. Mając wskazane powyższe perspektywy: zewnętrznego obserwatora, klienta banku i banku, oraz przyjmując również, że występuje związek między cechą prawnie chronioną, a faktem, czy ktoś spłaci kredyt lub nie, tj.:

$$Y \not\perp\!\!\!\perp X_a$$

można wskazać, że naraz nie jest możliwe spełnienie dowolnych dwóch z trzech wcześniej wskazanych perspektyw (Biecek 2024).

Jeżeli perspektywy sprawiedliwości *niezależnego obserwatora* ($\hat{Y} \perp\!\!\!\perp X_a$) i perspektywa *klienta banku* ($\hat{Y}|Y \perp\!\!\!\perp X_a$) będą naraz spełnione, to fakt spłaty kredytu jest niezależny względem cechy prawnie chronionej ($Y \perp\!\!\!\perp X_a$) *lub* decyzja kredytowa jest niezależna względem faktu spłaty kredytu, a model oceny zdolności kredytowej nie prognozuje spłaty kredytu ($\hat{Y} \perp\!\!\!\perp Y$):

$$\hat{Y} \perp\!\!\!\perp X_a \quad \wedge \quad \hat{Y}|Y \perp\!\!\!\perp X_a \quad \implies \quad Y \perp\!\!\!\perp X_a \quad \vee \quad \hat{Y} \perp\!\!\!\perp Y$$

Jeżeli perspektywy sprawiedliwości *niezależnego obserwatora* ($\hat{Y} \perp\!\!\!\perp X_a$) i perspektywa *banku* ($Y|\hat{Y} \perp\!\!\!\perp X_a$) będą naraz spełnione, to fakt spłaty kredytu jest niezależny względem cechy prawnie chronionej ($Y \perp\!\!\!\perp X_a$):

$$\hat{Y} \perp\!\!\!\perp X_a \quad \wedge \quad Y|\hat{Y} \perp\!\!\!\perp X_a \quad \implies \quad Y \perp\!\!\!\perp X_a$$

Jeżeli perspektywy sprawiedliwości *klienta banku* ($\hat{Y}|Y \perp\!\!\!\perp X_a$) i perspektywa *banku* ($Y|\hat{Y} \perp\!\!\!\perp X_a$) będą naraz spełnione, to fakt spłaty kredytu jest niezależny względem cechy prawnie chronionej ($Y \perp\!\!\!\perp X_a$) *oraz* ocena zdolności kredytowej jest niezależna względem wartości cechy prawnie chronionej ($\hat{Y} \perp\!\!\!\perp X_a$):

$$\hat{Y}|Y \perp\!\!\!\perp X_a \quad \wedge \quad Y|\hat{Y} \perp\!\!\!\perp X_a \quad \implies \quad Y \perp\!\!\!\perp X_a \quad \wedge \quad \hat{Y} \perp\!\!\!\perp X_a$$

W każdym z przedstawionych porównań kryteriów sprawiedliwości wskazano na brak zależności między spłacalnością kredytu (czyli nieobserwowalna zdolność kredytowa klienta) oraz zmienną prawnie chronioną lub jak, w przypadku kryterium separacji, model oceny

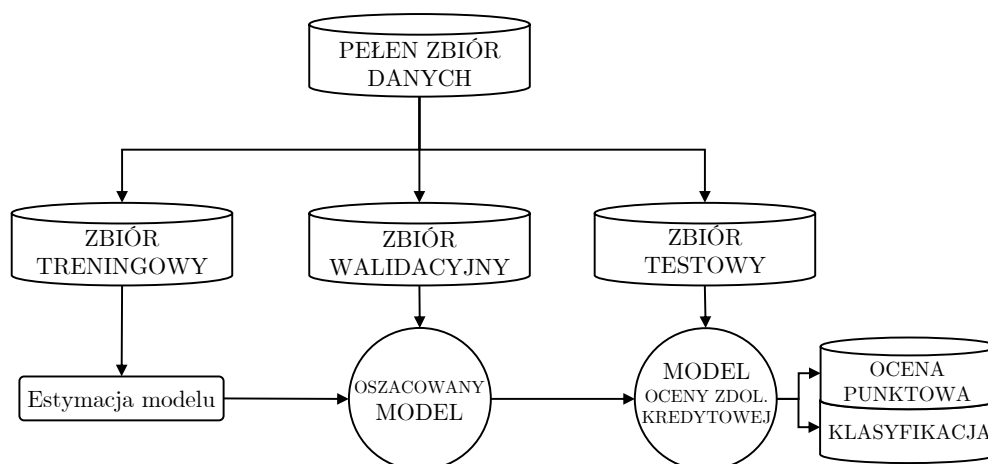
⁶⁶ W podrozdziale 3.4.1 wskazano odwołania do źródeł literatury wskazujące, że cechy prawnie chronione są zmiennymi statystycznie istotnymi w modelowaniu oceny zdolności kredytowej.

zdolności kredytowej nie prognozuje spłacalności kredytu. Oznacza to, że jeżeli $Y \not\subseteq X_a$, czyli spłacalność kredytu (realizacja nieobserwowalnej zdolności kredytowej klienta) jest zależna od zmiennej prawnie chronionej, to nie jest możliwe spełnienie dwóch z trzech wskazanych kryteriów sprawiedliwości modeli decyzyjnych.

3.4 Metody redukcji stronniczości algorytmicznej

W poprzednim podrozdziale przedstawiono metody identyfikacji występowania problemu stronniczości algorytmicznej. W przypadku, gdy taka stronniczość występuje – wybór miary identyfikacji stronniczości algorytmicznej jest *problemem decyzyjnym regulatora* rynku kredytów konsumenckich – banki, aby spełniać założenia regulacji przeciwdziałających dyskryminacji, muszą wykorzystać metody redukcji problemu stronniczości algorytmicznej. W ramach niniejszego podrozdziału przedstawiono metody redukcji stronniczości algorytmicznej, które banki mogą wykorzystać – wybór ten jest *problemem decyzyjnym banku*.

Poglądowy schemat relacji między zbiorami danych, modelem oraz procesem jego budowy przedstawiono na rysunku 9. Rysunek ten, w dalszej części tekstu będzie rozwinięty o klasy metod wykorzystywanych do redukcji problemu stronniczości algorytmicznej.



Rysunek 9: Proces budowy i wykorzystania modelu oceny zdolności kredytowej

Źródło: opracowanie własne.

3.4.1 Sprawiedliwość przez nieświadomość

Gdy bank, opracowując model oceny zdolności kredytowej, włącza cechę prawnie chronioną jako jedną ze zmiennych egzogenicznych (bez szczególnego uwzględnienia tej zmiennej) może wystąpić problem stronniczości algorytmicznej. Włączenie zmiennej prawnie chronionej jako kolejnej zmiennej egzogenicznej będzie nazywane *podjęciem nieograniczonym* (ang. *unmitigated*). W przypadku tego podejścia, jeżeli szacowany jest

model parametryczny, a cecha prawnie chroniona jest statystycznie istotna⁶⁷ parametr wrażliwości przy tej zmiennej wskazuje na występowanie dyskryminacji pośredniej: dwie osoby różniące się między sobą wyłącznie wartością cechy prawnie chronionej, uzyskują dwa różne oszacowania z modelu oceny zdolności kredytowej.

Ustawa Prawo Bankowe wskazuje, że „[...] banki i inne instytucje ustawowo upoważnione do udzielania kredytów na wniosek osoby fizycznej, prawnej lub jednostki organizacyjnej niemającej osobowości prawnej, o ile posiada zdolność prawną, ubiegającej się o kredyt przekazują, w formie pisemnej, wyjaśnienie dotyczące dokonanej przez siebie oceny zdolności kredytowej wnioskującego” (ustawa Prawo Bankowe, art. 70a). Punkt ten jest rozszerzony o dopisek, że wyjaśnienie, o którym mowa w ust. 1, obejmuje informacje na temat czynników, w tym danych osobowych wnioskującego, które miały wpływ na dokonaną ocenę zdolności kredytowej. Oznaczałoby to, że gdyby bank jako jednostka zobowiązana do przedstawienia osobom ubiegającym się o kredyt informacji (w nawiązaniu do wcześniej złożonego wniosku o taką informację), o czynnikach które miały wpływ na dokonaną ocenę zdolności kredytowej, musiałby wskazać wpływ czynników związanych z atrybutami prawnie chronionymi (np. płeć, pochodzenie etniczne). Dla wskazanego przykładu banku, kobieta, która posiada wartości pozostałych atrybutów taką samą jak przykładowy mężczyzna, mogłaby otrzymać informację, że został jej odmówiony kredyt ze względu na jej płeć.

Banki i regulatorzy mając świadomość wskazanego problemu z nieograniczonym podejściem; w związku z tym, w celu redukcji problemu stronniczości algorytmicznej modelu oceny zdolności kredytowej banki wykorzystują podejście związane z odrzuceniem informacji o wartości atrybutów chronionych. Podejście to, *sprawiedliwości przez nieświadomość* (ang. fairness through unawareness, fairness through blindness), jest obecnie najczęściej wykorzystywaną metodą *redukcji stronniczości algorytmicznej*. Uzasadnia się wykorzystanie tego podejścia przez wskazanie, że model nie może dyskryminować danej grupy społecznej jeżeli nie posiada informacji o wartości zmiennej chronionej (Chen 2018).

Metoda *sprawiedliwości przez nieświadomość* jedynie powierzchownie rozwiązuje problem stronniczości algorytmicznej (Knight 2019; Andreeva i Matuszyk 2019). Dane socjodemograficzne charakteryzują się korelacjami i interakcjami między zmiennymi; nieuwzględnienie tych dwóch efektów w modelu, w którym istotna zmienna predyktywna zostanie wyłączona, np. wyłączenie zmiennej płeć ze zbioru zmiennych wykorzystywanych w uczeniu modelu, prowadzić będzie do powstania tzw. *problemu zmiennej pominiętej*.

⁶⁷ Istnieją silne przesłanki wskazujące, że parametr zmiennej prawnie chronionej będzie statystycznie istotny; w literaturze występują liczne danych bankowych, w których to cechy prawnie chronione są zmiennymi statystycznie istotnymi: płeć (Andreeva i Matuszyk 2019; Agarwal i in. 2018; Henderson i in. 2015; Coin 2013; Do i Pakey 2013; D’espallier, Guérin i Mersland 2011; Muravyev, Talavera i Schäfer 2009), pochodzenie etniczne (Butler, Mayer i Weston 2023; Asiedu Freeman i Nti-Addae 2012; Havard 2010), lub wiek konsumenta (Dumitrescu i in. 2022; Okesola i in. 2017; Lee i in. 2002).

Efektom wyłączenia predyktywnej zmiennej objaśniającej z modelu (tj. cechy prawnie chronionej), skorelowanej z pozostałymi zmiennymi objaśniającymi będzie obciążenie estymatorów pozostałych parametrów modelu. Efekt ten może być na tyle silny, że parametry poszczególnych zmiennych mogą zmienić swój znak, co w przypadku modeli oceny zdolności kredytowej, prowadzi będzie do nieuzasadnianych ekonomicznie decyzji kredytowych, tj. przydzielanie wyższej oceny zdolności kredytowej osobom o niższych zarobkach (podrozdział 3.3.1).

Dodatkowo, wykluczenie predyktywnej zmiennej prowadzi będzie do pogorszenia jakości prognostycznej modelu oceny zdolności kredytowej (Kozodoi i Lessmann 2022); bank narażony będzie na spadek jakości podejmowanych decyzji kredytowych (m.in. udzielanie większej liczby kredytów klientom o niższej zdolności kredytowej lub mniej kredytów klientom o wysokiej zdolności kredytowej).

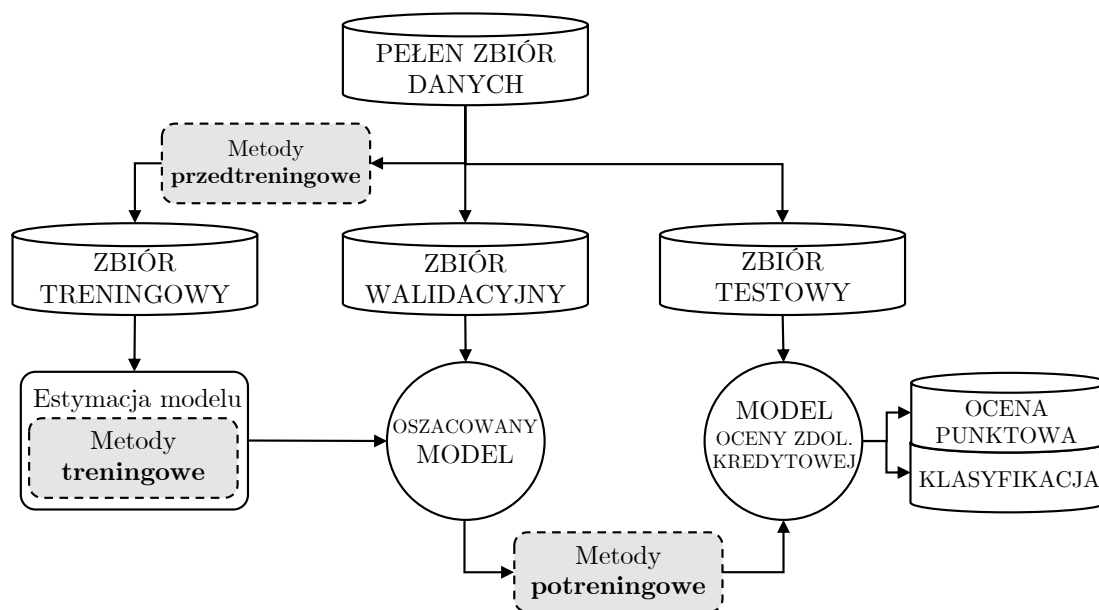
3.4.2 Podejścia: sprawiedliwość przez świadomość

Alternatywne względem podejścia *sprawiedliwości przez nieświadomość* (ang. bias-unaware) są podejścia *sprawiedliwości przez świadomość* (ang. bias-aware), które w ramach budowy modelu decyzyjnego wykorzystują pełen zbiór zmiennych (zarówno zmienne prawnie chronione, jak i niechronione), i na późniejszych etapach, wprowadzane są odpowiednie korekty.

W literaturze przedmiotu wyróżniono trzy klasy rozwiązań związanych z redukcją stroniczości algorytmicznej, w zależności od umiejscowienia w procesie budowy modelu, w którym dane podejście jest wykorzystywane. Podejścia dzieli się na:

- *Metody przedtreningowe* (ang. pre-processing) związane ze wstępnym etapem przetwarzania danych zbioru uczącego; motywacją jest wskazanie, że przyczyną dyskryminacji są dane zbioru treningowego, a modyfikując zbiór uczący, można zredukować poziom dyskryminacji (Friedler i in. 2019).
- *Metody treningowe* (ang. in-processing) redukcji stroniczości algorytmicznej związane z modyfikacją algorytmu uczącego, np. przez dodanie dodatkowych ograniczeń w procesie estymacji parametrów (Friedler i in. 2019).
- *Metody potreningowe* (ang. post-processing) redukcji stroniczości algorytmicznej związane z modyfikacją ocen punktowych generowanych w prognozie z modelu klasyfikacyjnego (Friedler i in. 2019); standardowa procedura zakłada modyfikację rozkładu oszacowanych ocen punktowych, a w konsekwencji zmianę przypisywanych przez model prognoz zmiennej celu (Kozodoi i Lessmann 2022).

Umiejscowienia metody redukcji stroniczości algorytmicznej w procesie budowy i wykorzystania modelu decyzyjnego przedstawiono na rysunku 10.



Rysunek 10: Umiejscowienie metody redukcji stronniczości algorytmicznej w procesie budowy i wykorzystania modelu decyzyjnego

Źródło: opracowanie własne na podstawie Kozodoi i Lessmann (2022).

W literaturze przedmiotu istnieje wiele alternatywnych podejść do redukcji stronniczości algorytmicznej. Metody realizują ten cel na różne proceduralnie sposoby, co w efekcie może wiązać się z różnymi konsekwencjami i wpływem na zysk generowany z akcji kredytowej oraz docelowy poziom stronniczości algorytmicznej modelu oceny zdolności kredytowej. W niniejszej rozprawie opracowano metodykę badania wpływu zastosowania poszczególnych metod redukcji stronniczości algorytmicznej w modelach oceny zdolności kredytowej.

Metody przedtreningowe

Metody przedtreningowe są związane ze wstępnym etapem przetwarzania danych zbioru uczącego – oddziałujące na rozkładach zmiennych na historycznych danych. Motywacją wykorzystania tej klasy podejść jest wskazanie, że przyczyną dyskryminacji są dane zbioru treningowego (Friedler i in. 2019) – modyfikując zbiór uczący, można zredukować poziom dyskryminacji.

Wyróżnia się następujące metody w ramach podejścia przedtreningowego:

- Metoda **Reweighing**⁶⁸ – metoda przedtreningowa, która zakłada wyznaczenie odpowiednio dobranych wag dla poszczególnych obserwacji, aby poziom zróżnicowanego wpływu (disparate impact) został zredukowany (Kamiran i Calders 2012).
- Metoda **Disparate Impact Remover**⁶⁹ – technika przedtreningowa, która modyfikuje wartości atrybutów poszczególnych obserwacji, aby zredukować poziom zróżnicowanego wpływu zachowując kolejność rang w grupach – Disparate Impact (Feldman i in. 2015).

⁶⁸ Przykładowa implementacja: `aif360.sklearn.preprocessing.Reweighing`

⁶⁹ Przykładowa implementacja: `aif360.algorithms.preprocessing.DisparateImpactRemover`

Główną zaletą metod klasy przedtreningowych jest fakt wczesnego włączania sprawiedliwego traktowania w procesie budowy modelu (Barocas, Hardt i Narayanan 2023). W przeciwieństwie do kolejnych metod, sam proces uczenia modelu i wykonywania prognoz na modelu oceny zdolności kredytowej pozostają bez zmian; modyfikacji podlega zbiór treningowy, na który uzyskiwane są oszacowania parametrów modeli.

Metody treningowe

Metody wykorzystywane w trakcie procesu estymacji parametrów modelu ekonometrycznego, tj. metody treningowe, związane są z modyfikacją algorytmu uczącego, np. przez dodanie ograniczeń w procesie estymacji parametrów modelu (Friedler i in. 2019).

W literaturze przedmiotu wskazuje się na przykładowe metody treningowe umożliwiające redukcję stroniczości algorytmicznej modelu decyzyjnego. Są to:

- Metoda **Grid Search Reduction**⁷⁰ – redukcja wyjściowego problemu klasyfikacji do sekwencji problemów klasyfikacyjnych wrażliwych na koszty (związany z niespełnieniem kryterium sprawiedliwości, podrozdział 3.3.2), zwracając deterministyczny klasyfikator z najniższym błędem empirycznym, podlegający ograniczeniom na wartość odmiennego wpływu – Disparate Impact (Agarwal, Dudík i Wu 2019).
- Metoda **Adversarial Debiasing**⁷¹ – model oparty na Generatywnych Sieciach Współzawodniczących; jedna z sieci szacowana jest mając na celu dokładność przewidywania zmiennej celu przy jednoczesnym zmniejszaniu zdolności drugiej sieci do określenia wartości zmiennej prawnie chronionej; wykonywanie prognozy z modelu nie uwzględnia wykorzystania informacji o zmiennych prawnie chronionych (Zhang, Lemoine i Mitchell 2018).
- Metoda **Prejudice Remover**⁷² – model regresji logistycznej z dodanym komponentem regularyzacyjnym związanym z poziomem wartości odmiennego wpływu modelu (Kamishima i in. 2012).

Główną zaletą metod treningowych jest fakt, że metody tej klasy nie oddziałują na dane ani zbiór treningowy, ani rozkład oceny punktowej prognozy z modelu. Oznacza to, że model może być traktowany jako odrębny moduł w całym procesie, w którym jednym z parametrów jest określenie miary sprawiedliwości.

⁷⁰ Przykładowa implementacja: `aif360.sklearn.inprocessing.GridSearchReduction`

⁷¹ Przykładowa implementacja: `aif360.sklearn.inprocessing.AdversarialDebiasing`

⁷² Przykładowa implementacja: `aif360.sklearn.inprocessing.PrejudiceRemover`

Metody potreningowe

Metody potreningowe związane są z redukcją stronniczości algorytmicznej przez modyfikację wyników oceny punktowej prognozy modelu (Friedler i in. 2019).

Przykładową metodą potreningową jest metoda `Reject Option Classifier`⁷³, która umożliwia zapewnienie redukcji odmiennego wpływu modelu przez zmianę prognozy modelu; w modelu oceny zdolności kredytowej byłaby to zmiana decyzji kredytowych. Zmiana ta odbywa się dla poszczególnych obserwacji, dając korzystne wyniki grupom nieuprzywilejowanym i niekorzystne wyniki grupom uprzywilejowanym w przedziale ufności wokół granicy decyzji (Kamiran i in. 2012). Metoda, wybierając obserwacje wokół granicy decyzyjnej oddziałuje, na obserwacje o najwyższej niepewności prognozy. Zaletą tej klasy podejść jest możliwość aplikacji do dowolnego rozkładu prognoz z modelu (tj. dowolnego modelu), przy czym bardzo często te podejścia charakteryzują się istotnym spadkiem mocy dyskryminacyjnej (Barocas, Hardt i Narayanan 2023).

* * *

Historia krajowych i międzynarodowych regulacji antydyskryminacyjnych sięga dziesięcioleci; pierwsze ramy prawne i umowy międzynarodowe zakazujące dyskryminacji m.in. ze względu na płeć lub pochodzenie etniczne były tworzone w latach 40. XX wieku. Przepisy te odnosiły się kierunkowo do zakazu dyskryminacji, przy czym nie wskazywały szczegółowych definicji lub interpretacji związanych z pojęciem dyskryminacji⁷⁴.

Regulacje zakazujące dyskryminacji zaczęły obejmować działalność banków, w szczególności proces udzielania kredytów dla konsumentów. Początkowo zapisy regulacyjne wprowadzane były w Stanach Zjednoczonych Ameryki Północnej (m.in. *Fair Housing Act* z 1968 r. lub *Equal Credit Opportunity Act* z 1974 r.), a później w Unii Europejskiej (m.in. dyrektywa 2004/113/WE). Wraz z ewolucją przepisów prawa, zmianom podlegał również proces oceny zdolności kredytowej. Początkowo był on realizowany przez ekspertów i analityków kredytowych, a z czasem, wraz z popularyzacją wykorzystania modeli analitycznych, regulatorzy rynku kredytów konsumenckich zaczęli również prawnie wymagać stosowania *modeli* w ocenie zdolności kredytowej (ustawa Prawo Bankowe, art. 105a).

Współistnienie zarówno *kontekstu prawnego* zakazującego dyskryminacji ze względu na określoną zmienną prawnie chronioną oraz rozpowszechnienie *wykorzystania modeli*

⁷³ Przykładowa implementacja: `aif360.sklearn.postprocessing.RejectOptionClassifier`

⁷⁴ Dla przykładu: Karta Narodów Zjednoczonych z 1945 r. wskazuje na zapis: „przywrócić wiarę w podstawowe prawa człowieka, w godność i wartość człowieka, w równouprawnienie mężczyzn i kobiet [...]”. W treści tej umowy nie są wskazane zapisy szczegółowe wskazujące na interpretację równouprawnienia, tj. równości w ramach określonego systemu prawnego. Zapisy ustawy z dnia 3 grudnia 2010 r. o wdrożeniu niektórych przepisów Unii Europejskiej w zakresie równego traktowania wyszczególniały pojęcia dyskryminacji pośredniej i dyskryminacji bezpośredniej.

decyzyjnych (w tym modeli oceny zdolności kredytowej) wprowadziło do dyskursu nowe pojęcie, tj. stronniczość algorytmiczną. Jak wskazano w początkowej części niniejszego rozdziału, stronniczość algorytmiczna „[...] ma miejsce, gdy algorytmy podejmują decyzje, które systematycznie stawiają w niekorzystnej sytuacji określone grupy ludzi.” (Friis i Riley 2023). Oznacza to, że dyskryminacja w zakresie równości ludzi („niekorzystna sytuacja grupy ludzi”) spowodowana przez decyzje podjęte przez algorytm stanowi przesłankę o jego stronniczości.

Nie istnieje jeden sposób na identyfikację czy model jest stronniczy algorytmicznie, czy nie; w podrozdziale 3.3.2 przedstawiono trzy alternatywne sposoby pomiaru problemu stronniczości algorytmicznej. Wskazano również, że w przypadku, gdy zmienna prawnie chroniona jest zmienną *istotną* z perspektywy zdolności kredytowej, to nie jest możliwe spełnienie dwóch z trzech przedstawionych kryteriów sprawiedliwości modeli decyzyjnych. Oznacza to, że jeżeli zmienna prawnie chroniona nie jest niezależna od faktu czy dany kredytobiorca spłaci kredyt, to co najmniej dwie miary identyfikujące stronniczość algorytmiczną będą wskazywać na występowanie tego problemu.

Wskazany fakt determinuje implikacje dla sposobu tworzenia regulacji zakazujących stronniczość algorytmiczną w modelach decyzyjnych, w tym w modelach oceny zdolności kredytowej. Aktualnie wprowadzane regulacje prawne dotyczące modeli sztucznej inteligencji, m.in. rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń, wskazują na zapisy zakazujące dyskryminacji ludzi przez działanie modeli decyzyjnych (stronniczość algorytmiczna). W przypadku przytoczonego rozporządzenia, modele oceny zdolności kredytowej są tam wskazane jako modele wysokiego ryzyka, które to powinny być regularnie weryfikowane pod kątem występującego problemu stronniczości algorytmicznej. Regulacja ta nie podaje jednak sposobu identyfikacji stronniczości algorytmicznej.

Regulator rynku kredytów konsumenckich tworzący regulacje obejmujące modele oceny zdolności kredytowej stoi przed wyborem miary umożliwiającej identyfikację stronniczości algorytmicznej; ten problem decyzyjny w rozprawie nazwano *perspektywą decyzyjną regulatora* rynku kredytów konsumenckich.

W przypadku, gdy taka stronniczość występuje banki, aby spełnić założenia regulacji zakazujące dyskryminacji, wykorzystują metody redukcji problemu stronniczości algorytmicznej. Wybór odpowiedniego podejścia jest uwarunkowany zarówno skutecznością redukcji stronniczości algorytmicznej, jak również rachunkiem ekonomicznym z zrealizowanego procesu kredytowego; ten problem decyzyjny w rozprawie nazwano *problemem decyzyjnym banku*.

W kolejnym rozdziale przedstawiono metodykę oceny wykorzystania metod redukcji stronniczości algorytmicznej w kontekście: wymogów zakazujących dyskryminacji i konsekwencji ekonomicznych dla banku.

Rozdział 4

Bezpośredni wpływ redukcji stronniczości algorytmicznej

Celem niniejszego rozdziału jest opracowanie procedury redukcji stronniczości algorytmicznej oraz jej zastosowanie na przykładowych danych – syntetycznych i rzeczywistych. Jako metodę badawczą wykorzystano symulacje, które umożliwiają kontrolę środowiska prowadzonego badania i porównywania własności poszczególnych metod redukcji stronniczości algorytmicznej.

Motywacją do przeprowadzenia badania bezpośredniego wpływu redukcji stronniczości algorytmicznej był *problem decyzyjny banku*. Decyzje związane z modelami oceny zdolności kredytowej podejmowane przez bank w zakresie ograniczania stronniczości algorytmicznej są oceniane przy uwzględnieniu wartości ekonomicznej generowanej przez dany model oceny zdolności kredytowej pod warunkiem ograniczeń związanych z obowiązującymi wymogami regulacji antydyskryminacyjnych.

W ramach badania wykorzystano autorski model symulacyjny – model statyczny – umożliwiający ocenę wybranych podejść do redukcji stronniczości algorytmicznej w zakresie:

- skuteczności redukcji problemu stronniczości algorytmicznej w kontekście modeli oceny zdolności kredytowych;
- skutków zastosowania tych metod w zakresie realizowanej oceny zdolności kredytowej, w tym rachunku ekonomicznego banku.

W rozdziale wykorzystano definicje miar stronniczości algorytmicznej przedstawione w podrozdziale 3.3.2 oraz podejścia do redukcji tego problemu zaprezentowane w podrozdziale 3.4.2.

Struktura niniejszego rozdziału jest następująca: w podrozdziale 4.1 opisano procedurę symulacyjną wykorzystaną w modelu statycznej oceny zastosowanie metod redukcji stronniczości algorytmicznej. W podrozdziale 4.2 przedstawiono opis wyników symulacji przeprowadzonych na podstawie danych syntetycznych, w których zmienną chronioną jest **płeć** (specyfikacja symulatora danych syntetycznych została zamieszczona w Aneksie: Specyfikacja danych). W podrozdziale 4.3 zamieszczono wyniki symulacji wykorzystania

modelu statycznego na zbiorze danych rzeczywistych: **German Credit Data**⁷⁵, w których zmienną chronioną jest **narodowość**.

4.1 Schemat modelu statycznego

Autorski model statyczny jest modelem symulacji, w którym dla zadanego zbioru danych jest możliwa statystyczna ocena bezpośredniego wpływu zastosowania metod redukcji stroniczości algorytmicznej w modelach oceny zdolności kredytowej.

Nazwa badania „bezpośredni wpływ redukcji stroniczości algorytmicznej” wiąże się z założeniem, że w badaniu nie jest brane pod uwagę, jaki model oceny zdolności kredytowej jest stosowany przez bank *przed* rozpoczęciem wykorzystania metod redukcji stroniczości algorytmicznej. Dla przykładu, oceniając metodę **Reweighing** opisaną w podrozdziale 3.4.2, w modelu statycznym nie jest uwzględniana informacja o modelu oceny zdolności kredytowej, który bank wykorzystywał przed włączeniem tej metody. Założenie to jest uchylone w rozdziale 5, w którym model dynamiczny umożliwia badanie wpływu włączenia modelu redukcji stroniczości algorytmicznej w miejsce modelu do tej pory wykorzystywanego.

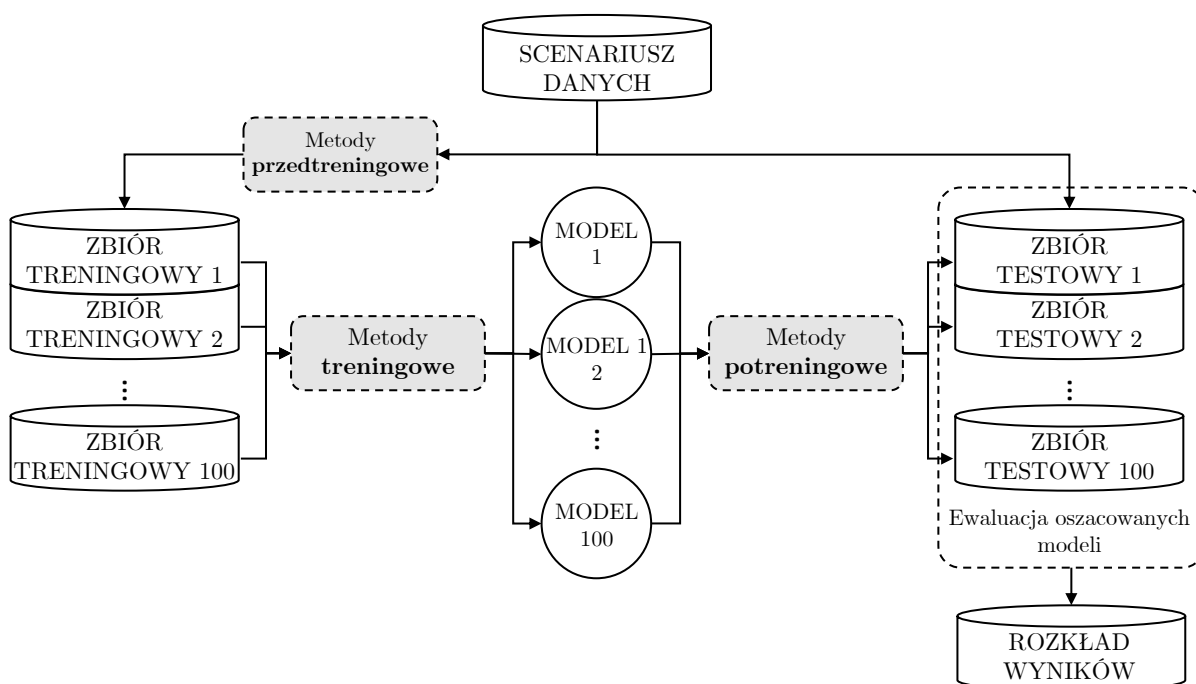
Badanie z wykorzystaniem modelu statycznego przeprowadzono opierając się na danych syntetycznych: **Scenariusz Syntetyczny**, jak również danych rzeczywistych: **Scenariusz German Credit Data**. Opis danych wykorzystanych w poszczególnych scenariuszach został zamieszczony w aneksie. Ogólny schemat realizacji symulacji w ramach podejścia oceny bezpośrednich konsekwencji (model statyczny) zaprezentowano na rysunku 11. Realizacja badania przy wykorzystaniu modelu statycznego zakłada następujące etapy:

1. *Wygenerowanie zbiorów uczących i testowych.* Pierwszym krokiem badania jest wygenerowanie – z danego scenariusza danych – zestawu zbiorów treningowych i testowych. W rozprawie liczba par tych zbiorów wynosiła 100 (tj. 100 zbiorów treningowych oraz 100 zbiorów testowych). Z racji na wykorzystanie modelu regresji (tj. regresja logistyczna) w przypadku zmiennych kategorycznych jest wykorzystywane kodowanie binarne (tzw. dummy coding).
2. *Dobór typów modeli ekonometrycznych.* Drugim krokiem badania jest określenie zakresu typów modeli ekonometrycznych, na podstawie których będą szacowane modele oceny zdolności kredytowej. W rozprawie wykorzystano dla scenariusza danych syntetycznych model regresji logistycznej, drzewa klasyfikacyjne, lasy losowe oraz metodę gradient boosting.
3. *Dobór metod redukcji stroniczości algorytmicznej.* Trzecim krokiem jest dobór zestawu metod, które umożliwiają redukcję stroniczości algorytmicznej. Warto wskazać, że

⁷⁵ Dane upublicznione przez Hofmann (1994).

podejścia **Adversarial Debiasing** i **Prejudice Remover** posiadają wewnętrzne modele ekonometryczne, tj. metoda **Adversarial Debiasing** bazuje na sieciach neuronowych, a metoda **Prejudice Remover** polega na obciążeniu regresji logistycznej (tj. regularyzowana regresja logistyczna). W rozprawie w badaniu przy wykorzystaniu modelu statycznego w scenariuszy danych syntetycznych wykorzystano podejścia do redukcji stroniczości algorytmicznej opisane w podrozdziale 3.4.

4. *Przeliczenie kombinacji: dane $\times$ model ekonometryczny $\times$ metoda redukcji stroniczości algorytmicznej.* W kroku czwartym dla kombinacji: wyznaczonego w pkt. 1 zestawu zbiorów danych, określonego w pkt. 2 zakresu typów modeli ekonometrycznych oraz wybranego w pkt. 3 zestawu metod redukcji stroniczości algorytmicznej, są wyznaczane wartości miar stroniczości algorytmicznej (opisane w podrozdziale 3.3.2) oraz parametry oceny jakości procesu kredytowego (tj. moc dyskryminacyjna lub zysk).
5. *Zebranie i podsumowanie wyników.* W kroku piątym na podstawie zestawu danych z wygenerowanego zgodnie z pkt. 4, jest tworzone zestawienie wyników w postaci tabel i rysunków, które zawiera: dane o redukcji stroniczości algorytmicznej oraz informację o jakości progностycznej modelu oceny zdolności kredytowej.



Rysunek 11: Schemat realizacji eksperymentów numerycznych oceny bezpośrednich konsekwencji
Źródło: opracowanie własne.

W zakresie metod redukcji stroniczości algorytmicznej wykorzystano przykłady każdej z klas metod opisanych w rozdziale 3.4, tj.: metody przedtreningowe, treningowe oraz potreningowe. Jako dwie metody porównawcze przyjęto: podejście bez ograniczania stroniczości algorytmicznej – metodę **Unmitigated** (estymacja modelu z uwzględnieniem

zmiennej chronionej) oraz metodę sprawiedliwość przez nieświadomość – metody *Fairness through Unawareness* (wyłączenie zmiennej chronionej ze zbioru zmiennych wykorzystanych do estymacji modelu). W zakresie metod przedtreningowych wykorzystano metody *Reweighting*, *Disparate Impact Remover*. Metody treningowe były reprezentowane przez metody *Grid Search Reduction*, *Adversarial Debiasing* oraz *Prejudice Remover*. Zakres metod potreningowych obejmował metodę *Reject Option Classifier*.

Symulacje dla danych syntetycznych – *Scenariusz Syntetyczny* – zrealizowano przy wykorzystaniu 4 typów modeli ekonometrycznych: regresja logistyczna – LR⁷⁶, drzewa klasyfikacyjne – DT⁷⁷, lasy losowe – RF⁷⁸, gradient boosting – GB⁷⁹. Dobór wartości hiperparametrów modeli DT, RF oraz GB był wykonany na niezależnym zbiorze walidacyjnym.

Choć w literaturze przedmiotu dla danych – *Scenariusz German Credit Data* – wykorzystywane są również innego typu modele, z racji na istotnie mniejszy zbiór danych oraz zagadnienia stabilności symulacji, dla danych rzeczywistych przedstawiono wyniki wyłącznie dla modelu regresji logistycznej.

Podstawowym zagadnieniem związanym z zastosowaniem metod ograniczania stronniczości algorytmicznej jest wpływ ich wykorzystania na jakość prognostyczną modeli decyzyjnych oraz ocena poziomu redukcji stronniczości algorytmicznej. Jakość ta, szczególnie w kontekście oceny zdolności kredytowej, jest powiązana z wartością ekonomiczną, jaką dany model generuje dla banku. Modele oceny zdolności kredytowej o wysokiej jakości prognostycznej, tj. charakteryzujące się stosunkowo niskim poziomem błędów klasyfikacji, będą umożliwiały podejmowanie lepszych, z perspektywy rachunku ekonomicznego, decyzji kredytowych przez bank (podrozdział 2.3.2). W przypadku modeli o niskiej jakości prognostycznej, włącznie ich do stosowania skutkować będzie generowaniem złych decyzji kredytowych, które w rezultacie wpłyną na wzrost poziomu szkodowości portfeli kredytowych banku. Z perspektywy oceny jakości prognostycznej modelu przygotowano 2 miary wykorzystane w badaniu, tj.: indeks Giniego oraz syntetyczną miarę *Zysk*, przy czym główną miarą będzie syntetyczna miara *Zysk*⁸⁰. Wszystkie raportowane w rozprawie wyniki eksperymentów numerycznych są wyznaczane w oparciu o zbiór testowy.

Drugim wymiarem badania jest zaburzenie struktury populacyjnej kredytobiorców,

⁷⁶ Przykładowa implementacja w języku Python: `sklearn.linear_model.LogisticRegression`

⁷⁷ Przykładowa implementacja w języku Python: `sklearn.tree.DecisionTreeClassifier`

⁷⁸ Przykładowa implementacja w języku Python: `sklearn.ensemble.RandomForestClassifier`

⁷⁹ Przykładowa implementacja w języku Python: `sklearn.ensemble.GradientBoostingClassifier`

⁸⁰ Zdecydowano się na wykorzystanie syntetycznej miary *Zysk*, ponieważ odnosi się ona do teoretycznej macierzy kosztów związanych z wydawanymi decyzjami kredytowymi (Kozodoi i Lessmann 2022). W przypadku indeksu Giniego, mierzy on jakość separacji klientów (spłacających od niespłacających) dla wszystkich progów odcięcia decyzji kredytowej. W praktyce banki wykorzystują jeden próg pozytywnej decyzji kredytowej, związany z oszacowaniem prawdopodobieństwa spłaty zobowiązania kredytowego.

która odnosi się do wpływu zastosowanego modelu oceny zdolności kredytowej na rozkład wyników procesu kredytowego. Jest to perspektywa zmiany rozkładu akceptowanych wniosków kredytowych poprzez zastosowanie metody ograniczania stronniczości algorytmicznej. Wdrożenie nowego modelu oceny zdolności kredytowej, w tym modyfikacja polegająca na włączeniu metody ograniczania stronniczości algorytmicznej, może spowodować, że profile klienckie, które do tej pory były odrzucane w procesie kredytowym, zaczną otrzymywać pozytywne oceny kredytowej. Analogicznie, wnioski kredytowe typowych kredytobiorców banku mogą zacząć być odrzucane. Zjawisko to będzie dalej nazywane zaburzeniem rozkładu akceptowanych wniosków kredytowych.

4.2 Scenariusz danych syntetycznych

Scenariusz Syntetyczny oparty jest na danych aplikacji kredytowych wygenerowanych przy wykorzystaniu generatora danych, którego specyfikacja została przedstawiona w Aneksie. Zbiór przygotowanych danych zawiera cztery zmienne egzogeniczne opisujące klienta: **wiek**, **płeć**, **dochód** i **wydatki**; zbiór danych zawiera również zmienną celu, **target**, stanowiącą informację o tym czy klient spłacił kredyt. W badaniu przyjęto, że zmienną prawnie chronioną ze względu na którą są wyznaczone miary stronniczości i względem której stosowane są metody redukcji stronniczości algorytmicznej jest binarna zmienna **płeć**⁸¹. Zbiór danych liczy 2 mln obserwacji, przy czym w jednej symulacji (zbiory treningowe/testowe od 1 do 100 na rysunku 11) losowanych bez zwracania jest 20 tys. obserwacji.

Wykorzystanie danych symulacyjnych miało na celu zobrazowanie i wyeksponowanie wpływu występowania typowych zjawisk w danych wykorzystywanych do uczenia modeli oceny zdolności kredytowej. Zakres tych zjawisk obejmował:

- występowanie współliniowości zmiennych egzogenicznych, tj. zmienne objaśniające są ze sobą skorelowane⁸²;
- występowanie nieliniowości wpływu zmiennych egzogenicznych – zmienna **wiek** wpływa na zmienną celu w sposób nieliniowy (tj. przyjęto zależność względem zmiennej celu jako kwadrat wartości zmiennej **wiek** z ekstremum paraboli w środku rozkładu);
- niebilansowanie próby obserwacji względem zmiennej **płeć**, tj. w populacji aplikacji kredytowych występuje więcej mężczyzn.

⁸¹ Zgodnie z ustawą 2010/12/03 zmienna **wiek** jest również zmienną prawnie chronioną. W celu uproszczenia przeprowadzonego badania i prezentowanych wyników, w przedstawionym w rozprawie przykładzie posłużono się binarną zmienną **płeć** jako zmienną prawnie chronioną.

⁸² W procedurze symulacyjnej wykorzystano kopule łączące rozkłady, z których losowano poszczególne zmienne egzogeniczne (Houssou i in. 2022).

Wykorzystanie danych syntetycznych w ocenie procedur związanych z tworzeniem modeli oceny zdolności kredytowej jest inspirowane istniejącą literaturą przedmiotu: Przanowski (2014); Kennedy, Delany i Namee (2011). Argumenty wskazujące na zasadność wykorzystania danych syntetycznych związane są z:

- Brakiem historycznych obserwacji – część zjawisk zachodzących w otoczeniu banków są zjawiskami typu „Czarny Łabędź” – przykładem może być pandemia Covid-19. Na wczesnym etapie pandemii, nie był znany wpływ wystąpienia pandemii lub wprowadzonych „wakacji kredytowych”, na ryzyko kredytowe ponoszone przez banki⁸³.
- Powolną dynamiką rzeczywistych procesów – aby zaobserwować pełen cykl życia portfela kredytów potrzeba jest przynajmniej kilku-kilkunastu lat (tj. od aplikacji kredytowej, po ewentualną windykację). Zmiany wartości parametrów procesu generującego dane dodatkowo generują szum do tej obserwacji⁸⁴. Model symulacji dynamicznej opisany w rozdziale 5 umożliwia badanie zjawisk, które w systemie bankowym zachodziłyby w skali kilkudziesięciu lat.
- Utrudnionym dostępem do danych publicznych – w przypadku danych bankowych, formalne uzyskanie zgodny na dostęp oraz wykorzystanie danych np. o spłacalności kredytów przez konsumentów może być utrudnione lub niemożliwe. Częstą praktyką banków jest odmowa lub udostępnienie danych, które zostały poddane „zaszumieniu” – dane, które są pozbawione interpretacji oraz odniesienia do prawdziwego procesu generującego dane. W rozprawie posłużono się danymi publicznie dostępnymi – **Scenariusz German Credit Data** – przy czym jest to jeden z niewielu publicznie dostępnych zbiorów danych bankowych.
- Brakiem możliwości uogólnienia tezy w oparciu o wybrany zbiór danych – teoretyczny dowód wyższości jednej metody nad inną (np. metody **Reweight** nad metodą **Disparate Impact Remover**), korzystając z pojedynczego zbioru danych (np. wybranego zbioru danych rzeczywistych) nie jest możliwy; w celu empirycznego wykazania lepszych własności metody, należałoby wykorzystać wszystkie możliwe zbiory danych rzeczywistych, które jak wskazano wyżej, w przypadku banków, są trudno dostępne.

Dane syntetyczne nie stanowią alternatywy do danych rzeczywistych; stanowią jednak „[...] coś pośredniego. Mianowicie możliwość rozważenia zbiorów najbardziej typowych spotykanych w rzeczywistości oraz ich wielu uogólnień. [...] O tego typu danych będzie

⁸³ Pojęcie zjawiska „Czarnego Łabędzia” wprowadził Nassim Taleb, który definiuje je jako „nieprzewidywalne i nieregularne zdarzenia o ogromnej skali i potężnych konsekwencjach, nieoczekiwane dla każdego obserwatora jeśli nie tylko go zaskoczą, ale zaszkodzą” (Taleb 2014).

⁸⁴ Kiedyś informacja o posiadaniu telefonu komórkowego była zmienną istotną z perspektywy oceny zdolności kredytowej; dziś, ze względu na zmiany społeczne, gospodarcze i technologiczne, zmienna ta nie niesie żadnej istotnej informacji z perspektywy oceny zdolności kredytowej.

można pisać w sposób bardziej precyzyjny podając ich parametry, własności rozkładów itp. Będzie zatem możliwe przeprowadzanie dowodów nie tylko na zasadzie, która technika modelowa jest lepsza, ale także przy jakich parametrach” (Przanowski 2014).

4.2.1 Badanie jakości prognostycznej modeli

W ramach oceny bezpośredniego wpływu stosowania metod ograniczania stroniczości algorytmicznej na wartość jakości prognostycznej modeli oceny zdolności kredytowej przedstawiono wyniki badania poszczególnych kryteriów oceny sprawiedliwości opisanych w podrozdziale 3.3.2, tj.: warunku parytetu demograficznego, warunku wyrównanych szans oraz warunku parytetu kalibracji.

Warunek parytetu demograficznego

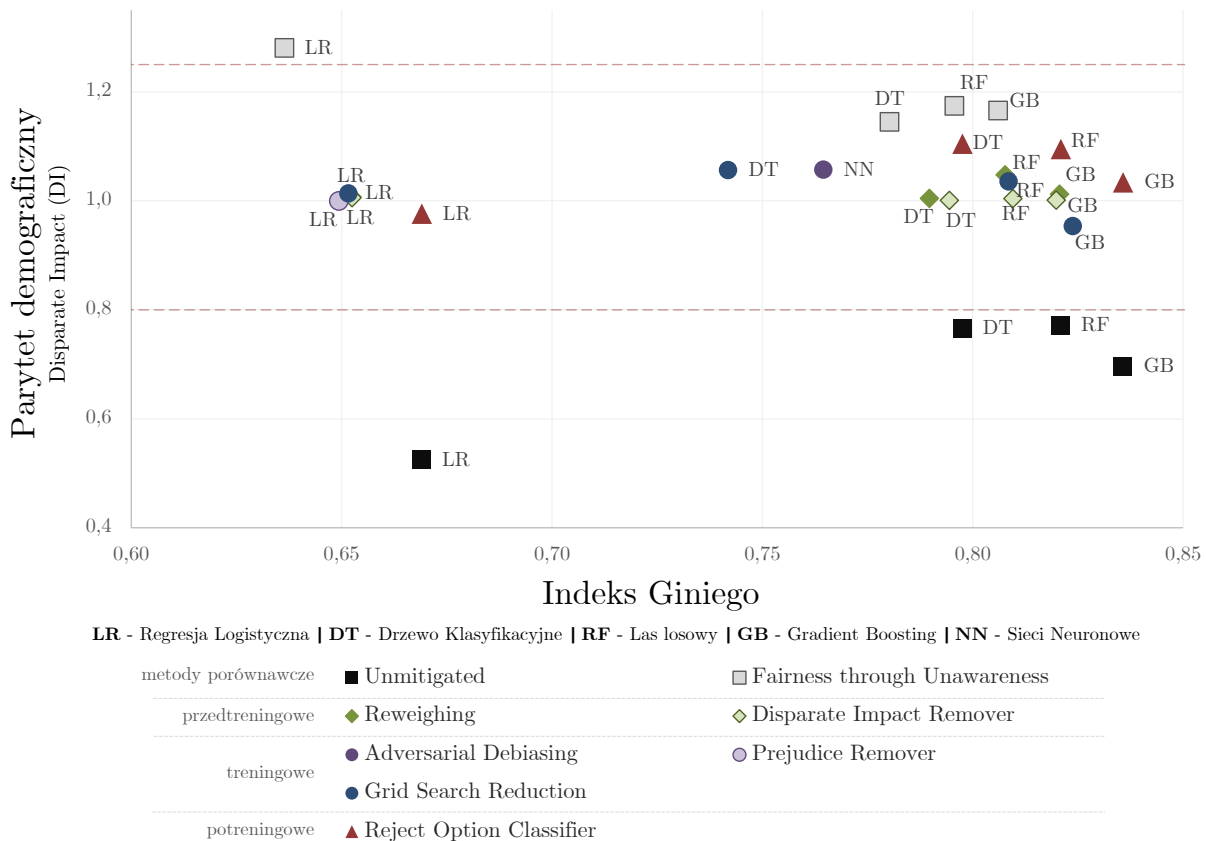
Ocenę wpływu zastosowania metod redukcji stroniczości algorytmicznej na jakość prognostyczną modeli oceny zdolności kredytowej przeprowadzono na podstawie parametrów jakości modelu (tj. współczynnik Giniego oraz syntetyczna miara *Zysk*) oraz miary *Disparate Impact*, związanej z warunkiem parytetu demograficznego, przedstawionej w podrozdziale 3.3.2.

Wpływ zastosowania poszczególnych metod redukcji stroniczości algorytmicznej na indeks Giniego (opisany w podrozdziale 2.2.2) przedstawiono na rysunku 12. Wszystkie raportowane w rozprawie wyniki są wyznaczone na podstawie zbioru testowego.

Indeks Giniego jest miarą stosowaną do oceny mocy dyskryminacyjnej modeli klasyfikacyjnych⁸⁵ – należy jednak wskazać, że jest to miara wyłącznie częściowo użyteczna z perspektywy prowadzonego badania: indeks Giniego jest bowiem miarą zbiorczą, prezentującą działanie modelu dla wszystkich poziomów progów odcięcia decyzji kredytowej. Technicznie, wartość indeksu Giniego wyznaczana jest w oparciu o pole powierzchni pod krzywą ROC. W kontekście oceny zdolności kredytowej, badaniu poddawać należy zdefiniowany, specyficzny próg pozytywnej decyzji kredytowej, najczęściej określany z perspektywy dopuszczalnego poziomu prawdopodobieństwa spłaty kredytu.

W celu zobrazowania rzeczywistego wpływu modelu na ekonomiczny rachunek banku, skonstruowano syntetyczną miarę *Zysk*, która wyznaczana jest na podstawie predykcji klas oraz macierzy kosztu przedstawionych w tabeli 7.

⁸⁵ Moc dyskryminacyjna modelu klasyfikacyjnego rozumiana, zgodnie z rozdziałem 2, jako zdolność modelu do sortowania klientów względem ich poziomu zdolności kredytowej (np. od najniższej do najwyższej).



Rysunek 12: Średnie wartości indeksu Giniego, a warunek parytetu demograficznego,
Scenariusz Syntetyczny

Uwaga: Średnie wartości dla 100 powtórzeń symulacji. Przerywanymi, czerwonymi liniami poziomymi zaznaczono krytyczne poziomy 80% oraz 125% wartości miary **Disparate Impact**; obserwacje wartości dla modeli znajdujących się między tymi poziomami spełnią kryterium 4/5 opisane w podrozdziale 3.1.1. Model idealny posiada wartość miary **Disparate Impact** na poziomie równym 1 oraz posiada jak najwyższą wartość indeksu Giniego (prawa strona wykresu).

Źródło: opracowanie własne.

Tabela 7: Macierz kosztu dla syntetycznej miary **Zysk**

	Pozytywna d. kredytowa	Negatywna d. kredytowa
$Y = 1$ (spłata)	$C = +0,2664$	0
$Y = 0$ (brak spłaty)	$B = -0,2750$	0

Źródło: opracowanie własne.

Kozodoi i Lessmann (2022) zdefiniowali syntetyczną miarę odnoszącą się do poziomu zysku/straty, wynikającej z macierzy kosztów. Parametry macierzy kosztu, zdefiniowane przez autorów, to odpowiednio: $C = +0,2664$, $B = -0,275$; w przeciwieństwie jednak do podejścia przedstawionego przez Kozodoi i Lessmann (2022) w rozprawie dla parametru kosztu związanego z fałszywie negatywną oceną (tj. prognoza z modelu wskazywała złego klienta, a klient w rzeczywistości spłaciłby zobowiązanie) przypisano wartość równą 0;

Kozodoi i Lessmann (2022) przyjęli koszt utraconych przychodów na poziomie równym $-0,2664$.

Podejście zastosowane w rozprawie miało na celu uniknięcie podwójnej kalkulacji kosztów wynikających z nieprzyznania kredytu dla dobrego klienta: dla zobrazowania tego przypadku, rozważmy scenariusz graniczny, w którym bank w zdefiniowanym horyzoncie czasu otrzymał wyłącznie jedną aplikację kredytową. Przyjęto również, że ten bank odrzucił wniosek tego klienta, przy czym klient ten spłaciłby kredyt, gdyby go otrzymał. Macierz kosztów dla takiego przypadku, zgodnie z Kozodoi i Lessmann (2022), zaprezentowano w tabeli 8.

Tabela 8: Macierz kosztu dla fałszywie negatywnego przypadku

	Pozytywna d. kredytowa	Negatywna d. kredytowa
$Y = 1$ (spłata)	0	$-0,2664$
$Y = 0$ (brak spłaty)	0	0

Uwaga: W przypadku odrzucenia jednego wniosku kredytowego klienta, który w rzeczywistości spłaciłby kredyt, macierz kosztu zgodnie z Kozodoi i Lessmann (2022) zawierałaby wyłącznie wartość $-0,2664$ kosztu utraconych przychodów.

Źródło: opracowanie własne.

Następnie rozważono sytuację, gdy dany bank zwiększa wartość progu odcięcia, w taki sposób, że wniosek klienta zostałby rozpatrzony pozytywnie. Macierz kosztów dla takiego przypadku, zgodnie z Kozodoi i Lessmann (2022), zaprezentowano w tabeli 9.

Tabela 9: Macierz kosztu dla prawdziwie pozytywnego przypadku

	Pozytywna d. kredytowa	Negatywna d. kredytowa
$Y = 1$ (spłata)	$+0,2664$	0
$Y = 0$ (brak spłaty)	0	0

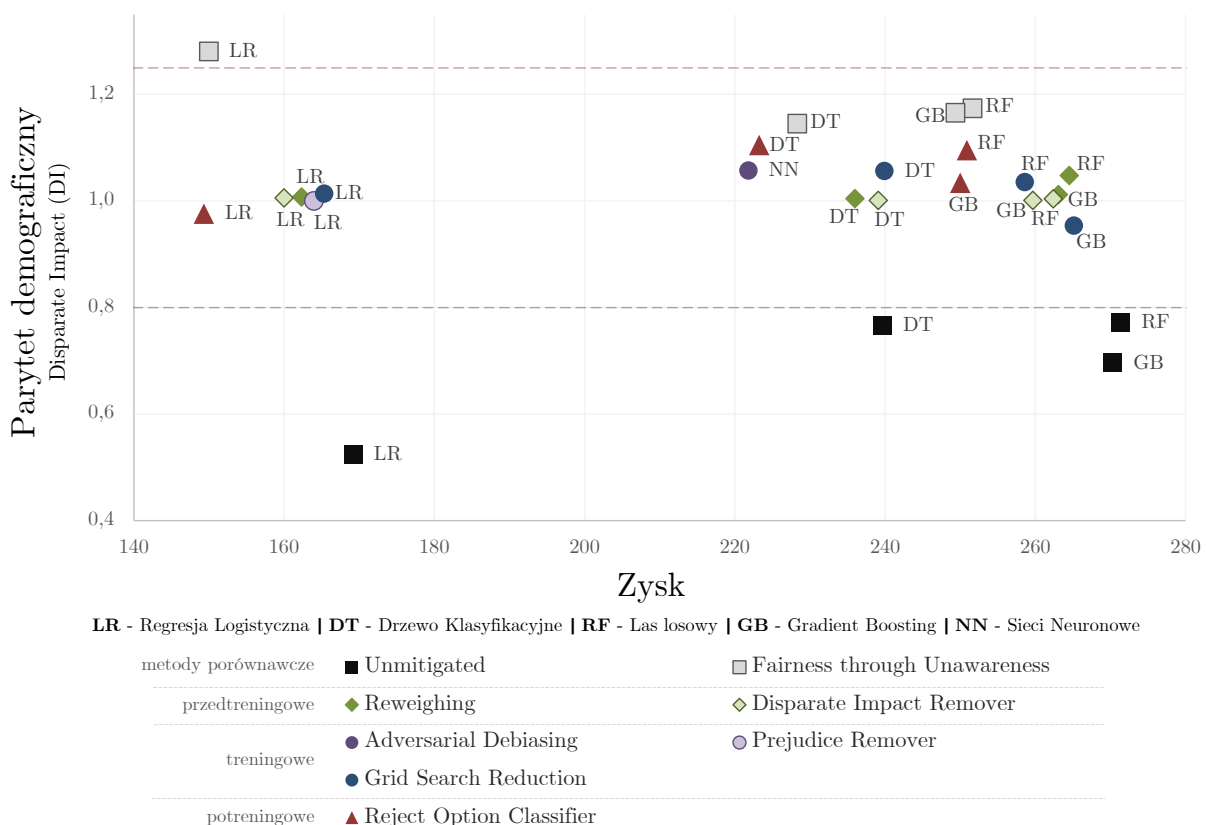
Uwaga: W przypadku przyjęcia jednego wniosku kredytowego klienta, który w rzeczywistości spłaciłby kredyt, macierz kosztu zgodnie z Kozodoi i Lessmann (2022) zawierałaby wyłącznie wartość $+0,2664$ zysku ze sprzedanego kredytu.

Źródło: opracowanie własne.

Różnica między tymi dwoma scenariuszami jest równa $+0,2664 - (-0,2664) = 0,5328$. Z perspektywy jednak rzeczywistego procesu kredytowego, bank zarabia wyłącznie $+0,2664$ na udzieleniu kredytu wnioskodawcy, który spłaci zobowiązanie kredytowe

(przypadek prawdziwie pozytywny). W rozprawie przyjęto podejście *rachunkowości finansowej*, która wskazuje, że rejestrowane powinny być wyłącznie rzeczywiste wydatki i przychody (Horngren, Datar i Rajan 2014).

Ocenę wpływu zastosowania metod ograniczania stronniczości algorytmicznej na poziom na ekonomiczną wartość modelu, tj. poziom zysku uzyskiwany przy zastosowaniu danego modelu oceny zdolności kredytowej dla poszczególnych podejść do redukcji stronniczości algorytmicznej, przedstawiono na rysunku 13. Wyniki te są w dużej mierze zbieżne z wynikami osiągniętymi dla badania z wykorzystaniem indeksu Giniego – największe różnice obserwowane są dla modeli potreningowego (którego mechanizm działania wykorzystuje jeden określony próg odcięcia decyzji). Rekomenduje się jednak wykorzystanie miar typu Zysk, ponieważ część z metod redukcji stronniczości algorytmicznej (np. metoda *Reject Option Classifier*) nie modyfikuje rozkładu wartości punktowych (tzw. scorów) z modeli oceny zdolności kredytowej, a wyłącznie progi odcięcia decyzji.



Rysunek 13: Miara Zysk a warunek parytetu demograficznego, **Scenariusz Syntetyczny**

Uwaga: Średnie wartości dla 100 powtórzeń symulacji. Przerywanymi, czerwonymi liniami poziomymi zaznaczono krytyczne poziomy 80% oraz 125% wartości miary Disparate Impact; obserwacje wartości dla modeli znajdujących się między tymi poziomami spełniają kryterium 4/5 opisane w podrozdziale 3.1.1. Model idealny posiada wartość miary Disparate Impact na poziomie równym 1 oraz posiada jak najwyższą wartość miary Zysk (prawa strona rysunku).

Źródło: opracowanie własne.

Jedną z głównych obserwacji jest fakt, że modele uczenia maszynowego wskazują na statystycznie lepsze wyniki w zakresie jakości prognostycznej modeli: modele DT, RF i GB uzyskują znacząco lepsze wartości jakości prognostycznej od modelu LR.

Średnie wartości miary Zysk oraz miary związanej z warunkiem parytetu demograficznego – **Disparate Impact** – uzyskanych dla różnych kombinacji metod ograniczania stronniczości algorytmicznej i modeli bazowych, wraz ze wskazaniem na ich wartości błędów standardowych (ang. standard error – SE) przedstawiono w tabeli 10. Wartości błędu standardowego wskazują na niepewność uzyskanych wyników: czym większe wartości SE, tym wyższy poziom niepewności uzyskanych wyników.

W zakresie metod porównawczych przedstawiono metodę **Unmitigated**, która jawnie włącza zmienną chronioną do zbioru zmiennych wykorzystywanych w procesie trenowania modelu oraz metodę **Fairness through Unawareness**, która zmienną chronioną wyłącza z procesu trenowania modelu. Uzyskane wyniki dla metody **Unmitigated** wskazują na wysoką jakość prognostyczną modelu oceny zdolności kredytowej: każdy z modeli bazowych uzyskał najwyższy wynik miary Zysk w swojej grupie modeli. Wyniki uzyskane dla tej metody nie spełniają jednak kryterium 4/5 warunku parytetu demograficznego: wartość **Disparate Impact** były poniżej progu 80%. Można jednocześnie wskazać, że wartości te dla metod bardziej zaawansowanych, tj. DT, RF oraz GB, są wartościami bardziej zbliżone do tego progu. Model LR dla metody **Unmitigated** uzyskał najgorszą wartość **Disparate Impact** z pośród wszystkich przebadanych wariantów metod ograniczania stronniczości algorytmicznej i estymatorów bazowych.

Wyłączenie zmiennej chronionej ze zbioru treningowego – metoda **Fairness through Unawareness** – prowadzi do powstania stronniczości algorytmicznej będącej odbiciem względem sytuacji początkowej, tj. włączając zmienną chronioną w metodzie **Unmitigated** kobiety uzyskiwały statystycznie mniej pozytywnych decyzji kredytowych. Sytuacja ta ma miejsce wyłącznie w przypadku modelu regresji logistycznej LR; w przypadku pozostałych modeli bazowych, wartości **Disparate Impact** mają wartości w dopuszczalnym przedziale 4/5.

Stosując metodę **Fairness through Unawareness** obserwuje się również statystycznie gorsze wyniki w zakresie jakości prognostycznej (miara Zysk), względem metody **Unmitigated**. Każdy z modeli bazowych posiada niższy wynik względem wyników metody **Unmitigated**, co jest wynikiem wyłączenia statystycznie istotnej zmiennej z modelu.

Drugą z klas badanych metod są metody przedtreningowe, dla których redukcja stronniczości algorytmicznej odbywa się poprzez ingerowanie w zbiór danych treningowych, na którym szacowany jest w dalszej kolejności model bazowy. W przedstawionym badaniu wykorzystano dwa podejścia: **Reweighing** nakładająca odpowiednio dobrane wagi na poszczególne obserwacje oraz **Disparate Impact Remover** zmieniająca wartości poszczególnych atrybutów, w wyniku czego redukowana jest stronniczość algorytmiczna.

Tabela 10: Wartości Zysk oraz Disparate Impact dla Scenariusz Syntetyczny

Podejście	Metoda redukcji stroniczości algorytmicznej	Model bazowy	Zysk	SE _{Zy}	DI	SE _{DI}	$\chi^2\%$	F%
porównawcze	Unmitigated	LR	169	8,3	0,52	0,05	0,00	0,00
	Unmitigated	DT	240	8,7	0,77	0,08	0,10	0,10
	Unmitigated	RF	271	8,4	0,77	0,05	0,00	0,00
	Unmitigated	GB	270	7,9	0,70	0,04	0,00	0,00
	Fairness t. Unawareness	LR	150	8,4	1,28	0,09	0,15	0,14
	Fairness t. Unawareness	DT	228	9,7	1,15	0,06	0,57	0,56
	Fairness t. Unawareness	RF	252	8,8	1,17	0,05	0,22	0,21
	Fairness t. Unawareness	GB	249	8,5	1,17	0,06	0,33	0,32
przedtreningowe	Reweighing	LR	162	8,3	1,01	0,07	0,99	0,99
	Reweighing	DT	236	9,2	1,00	0,08	0,91	0,91
	Reweighing	RF	265	8,6	1,05	0,05	0,93	0,93
	Reweighing	GB	263	8,4	1,01	0,05	0,99	0,99
	Disparate Impact Remover	LR	160	9,5	1,01	0,04	1,00	1,00
	Disparate Impact Remover	DT	239	9,1	1,00	0,03	1,00	1,00
	Disparate Impact Remover	RF	262	9,4	1,00	0,03	1,00	1,00
	Disparate Impact Remover	GB	260	8,8	1,00	0,03	1,00	1,00
treningowe	Adversarial Debiasing	NN	222	9,5	1,06	0,07	0,94	0,94
	Prejudice Remover	LR	164	11,0	1,00	0,13	0,81	0,81
	Grid Search Reduction	LR	165	9,2	1,01	0,13	0,75	0,75
	Grid Search Reduction	DT	240	12,3	1,06	0,15	0,52	0,52
	Grid Search Reduction	RF	259	8,9	1,04	0,08	0,86	0,85
	Grid Search Reduction	GB	265	8,6	0,95	0,06	0,91	0,90
potreningowe	Reject Option Classifier	LR	149	8,6	0,98	0,09	0,98	0,98
	Reject Option Classifier	DT	223	10,6	1,11	0,11	0,66	0,65
	Reject Option Classifier	RF	251	9,4	1,10	0,07	0,76	0,75
	Reject Option Classifier	GB	250	8,6	1,03	0,06	0,97	0,97

Uwaga: Kolumny $\chi^2\%$ oraz F% odnoszą się do frakcji symulacji, dla których odpowiednio test χ^2 (Tallarida i Murray 2012) oraz test Fishera (Fisher 1970) (podrozdział 3.3.2), przy 1% poziomie istotności nie dał podstaw do odrzucenia H_0 , która to hipoteza stanowiła o braku różnic między populacją obserwacji uprzywilejowanych i nieuprzywilejowanych. Oznacza to, że im mniejsze wartości w tych kolumnach, tym częściej model statystyczny wykazywał różnice w poziomie akceptacji wniosków kredytowych między kobietami i mężczyznami. Wartości 1,00 wskazuje na metody, dla których we wszystkich ze 100 symulacjach nie było podstaw do odrzucenia H_0 (tj. nie występowały statystyczne różnice między poziomem akceptacji kredytowych dla poszczególnych populacji).

Źródło: opracowanie własne.

Metody Reweighing oraz Disparate Impact Remover generują zbliżone wyniki. Warto wskazać, że wyniki dla tych podejść nie różnią się znacząco od wartości jakości prognostycznej dla metody Unmitigated, lecz spełniają one kryterium braku stroniczości algorytmicznej zgodnie z podrozdziałem 3.3.2 (wartości Disparate Impact oscylują około wartości 1). Zatem, metody te posiadają dobre własności w zakresie jakości

prognostycznej modelu oceny zdolności kredytowej, redukują w pełni problem stronniczości algorytmicznej oraz, ze względu na sposób działania, posiadają prostą interpretację.

Kolejną z grup badanych metod były metody treningowe. Sposób działania poszczególnych metod jest istotnie zróżnicowany. W przypadku metody **Adversarial Debiasing**, redukcja stronniczości algorytmicznej odbywa się poprzez zastosowanie wewnętrznej metody – sztucznych sieci neuronowych NN – w architekturze modelu generatywnych sieci współzawodniczących. Metoda **Prejudice Remover** związana jest z estymacją modelu regresji logistycznej przy warunkach ograniczających związanych ze współczynnikiem **Disparate Impact** (tj. regularyzacja regresji logistycznej). Metoda **Grid Search Reduction**, która polegała na estymacji szeregu modelu z warunkami ograniczającymi.

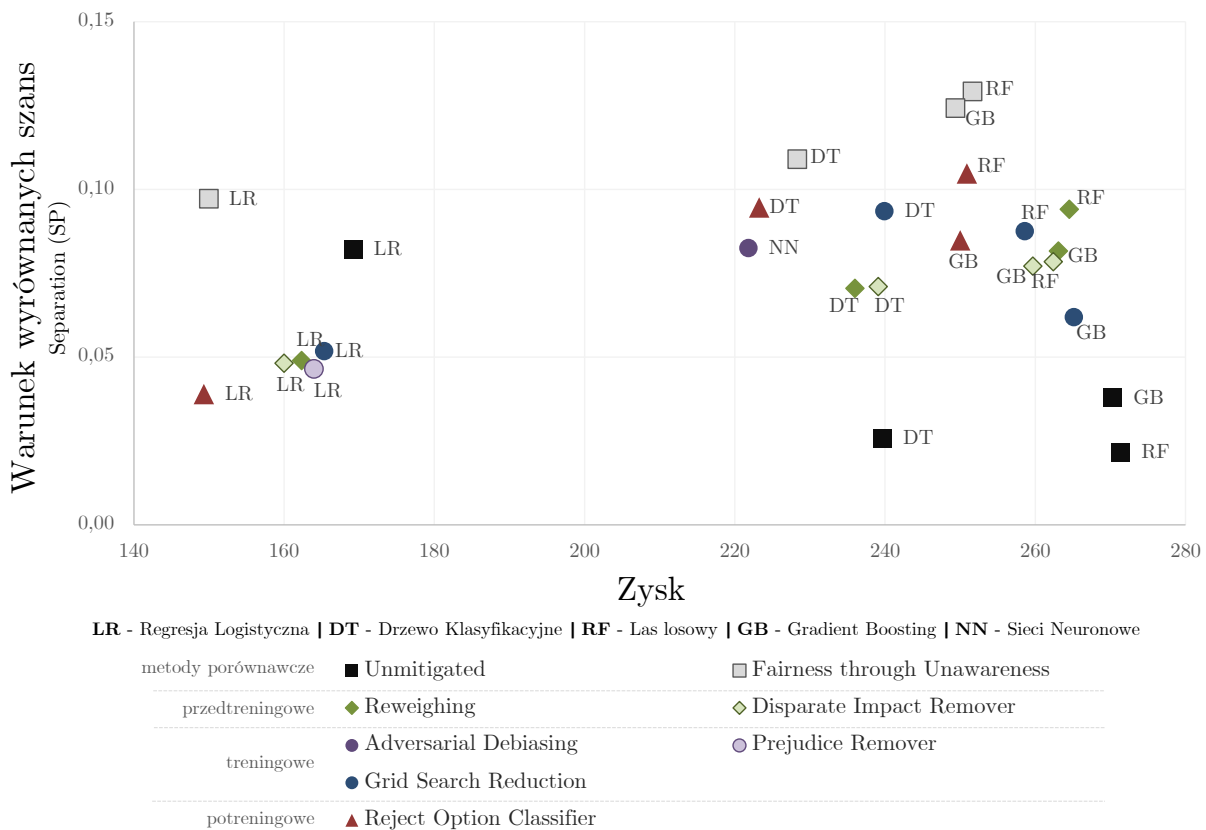
Uzyskane wyniki dla wszystkich metod wskazują na dobre własności prognostyczne modeli oceny zdolności kredytowej. Wszystkie oszacowane metody uzyskują wartość **Disparate Impact** na poziomach w przedziale wartości krytycznych 80% – 125%.

Ostatnią z klas wykorzystanych metod są metody potreningowe, w których uwzględnienie stronniczości algorytmicznej odbywa się na wartościach punktowych. Metodą, która stanowi przykład takiego podejścia jest **Reject Option Classifier**, której mechanizm działania zakłada, że obserwacje znajdujące się w klasie nieuprzywilejowanej (tj. z perspektywy zmiennej chronionej płeć, są to kobiety), których ocena prawdopodobieństwa znajduje się poniżej progu decyzji (maksymalna różnica jest parametrem tej metody), uzyskują pozytywną decyzję kredytową; analogicznie, mężczyźni, którzy uzyskaliby akceptację kredytową, ale znajdują się blisko progu akceptacji, uzyskują odrzucenie wniosku kredytowego. Mechanizm działania tej metody zakłada zmianę decyzji kredytowej wokół progu akceptacji kredytowej.

W przypadku metody **Reject Option Classifier**, jakość prognostyczna modelu kształtuje się na poziomie zbliżonym do modelu porównawczego **Fairness through Unawareness**, przy czym dla wszystkich estymatorów bazowych spełniony jest warunek sprawiedliwości (tj. wartości wskaźnika **Disparate Impact** mieszczą się w zakresie 80% – 125%).

Warunek wyrównanych szans

Drugim z kryteriów jest warunek wyrównanych szans, który odnosi się do zbiorczej miary prawdopodobieństwa popełnienia błędu prognozy w poszczególnych grupach zmiennej chronionej. Czym wyższa jest wartość kryterium separacji (kryterium związane z warunkiem wyrównanych szans), tym większa jest różnica w błędach prognozy dla poszczególnych grup zmiennej chronionej. Zbiorcze wyniki jakości prognostycznej mierzonej miarą **Zysk** oraz wartości kryterium separacji przedstawiono na rysunku 14.



Rysunek 14: Miara Zysk a warunek wyrównanych szans, **Scenariusz Syntetyczny**

Uwaga: Średnie wartości dla 100 powtórzeń symulacji. Wartości mniejsze kryterium separacji, związanego z warunkiem wyrównanych szans wskazują na mniejsze rozbieżności między błędami klasyfikacji dla poszczególnych grup zmiennej chronionej (mężczyźni i kobiety). Model idealny posiada wartość miary **Separation** na poziomie równym 0 oraz posiada jak najwyższą wartość miary **Zysk** (prawa strona rysunku).

Źródło: opracowanie własne.

Opierając się na przedstawionych wynikach, można wskazać następujące prawidłowości:

- metoda **Fairness through Unawareness** charakteryzowała się najwyższymi wartościami kryterium separacji;
- dla podejścia **Unmitigated** zaawansowane modele: DT, RF oraz GB charakteryzowały się wysokimi poziomami jakości progностycznej, przy równocześnie stosunkowo niskich wartościach kryterium separacji;
- metody ograniczania stronniczości algorytmicznej oparte o model regresji logistycznej LR charakteryzowały się stosunkowo niskimi wartościami kryterium separacji, przy równocześnie niskich wartościach jakości progностycznej.

Obecnie wykorzystywane przez banki podejście – **Fairness through Unawareness** – charakteryzuje się najwyższymi wartościami kryterium separacji, co powoduje stosunkowo wysokie różnice w poziomach błędów klasyfikacji między grupami zmiennej prawnie

chronionej. Średnie wartości miary Zysk oraz miary kryterium separacji przedstawiono w tabeli 11.

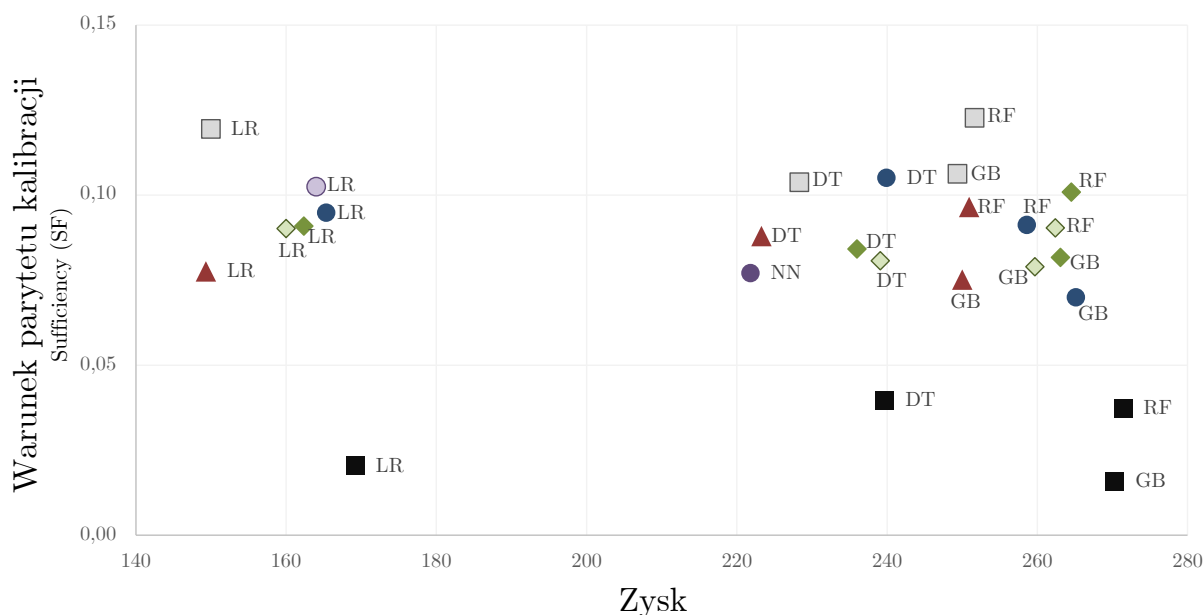
Tabela 11: Zestawienie wartości miary Zysk oraz kryterium separacji dla **Scenariusz Syntetyczny**

Podejście	Metoda redukcji stroniczości algorytmicznej	Model bazowy	Zysk	SE _{Zy}	SP	SE _{SP}
porównawcze	Unmitigated	LR	169	8,3	0,08	0,02
	Unmitigated	DT	240	8,7	0,03	0,02
	Unmitigated	RF	271	8,4	0,02	0,01
	Unmitigated	GB	270	7,9	0,04	0,02
	Fairness t. Unawareness	LR	150	8,4	0,10	0,01
	Fairness t. Unawareness	DT	228	9,7	0,11	0,01
	Fairness t. Unawareness	RF	252	8,8	0,13	0,01
	Fairness t. Unawareness	GB	249	8,5	0,12	0,01
przedtreningowe	Reweighing	LR	162	8,3	0,05	0,02
	Reweighing	DT	236	9,2	0,07	0,02
	Reweighing	RF	265	8,6	0,09	0,02
	Reweighing	GB	263	8,4	0,08	0,01
	Disparate Impact Remover	LR	160	9,5	0,05	0,01
	Disparate Impact Remover	DT	239	9,1	0,07	0,01
	Disparate Impact Remover	RF	262	9,4	0,08	0,01
	Disparate Impact Remover	GB	260	8,8	0,08	0,01
treningowe	Adversarial Debiasing	NN	222	9,5	0,08	0,02
	Prejudice Remover	LR	164	11,0	0,05	0,03
	Grid Search Reduction	LR	165	9,2	0,05	0,03
	Grid Search Reduction	DT	240	12,3	0,09	0,04
	Grid Search Reduction	RF	259	8,9	0,09	0,02
	Grid Search Reduction	GB	265	8,6	0,06	0,02
potreningowe	Reject Option Classifier	LR	149	8,6	0,04	0,02
	Reject Option Classifier	DT	223	10,6	0,09	0,03
	Reject Option Classifier	RF	251	9,4	0,10	0,02
	Reject Option Classifier	GB	250	8,6	0,08	0,02

Źródło: opracowanie własne

Warunek parytetu kalibracji

Ostatnią z omawianych w rozdziale 3 miar stroniczości algorytmicznej była miara związana z warunkiem parytetu kalibracji. Analogicznie jak w przypadku warunku wyrównanych szans, wyższe wartości miary kryterium dostateczności (związanego z warunkiem parytetu kalibracji) sugerują występowanie stroniczości algorytmicznej. Zbiornicze wyniki jakości prognostycznej mierzonej miarą Zysk oraz wartości kryterium dostateczności przedstawiono na rysunku 15.



LR - Regresja Logistyczna | DT - Drzewo Klasyfikacyjne | RF - Las losowy | GB - Gradient Boosting | NN - Sieci Neuronowe

metody porównawcze	■ Unmitigated	□ Fairness through Unawareness
przedtreningowe	◆ Reweighing	◆ Disparate Impact Remover
treningowe	● Adversarial Debiasing	● Prejudice Remover
	● Grid Search Reduction	
potreningowe	▲ Reject Option Classifier	

Rysunek 15: Miara Zysk a warunek parytetu kalibracji, Scenariusz Syntetyczny

Uwaga: Średnie wartości dla 100 powtórzeń symulacji. Wartości mniejsze kryterium dostateczności, związanego z warunkiem parytetu kalibracji wskazują na mniejsze rozbieżności między błędami dla poszczególnych grup zmiennej chronionej (mężczyźni i kobiety). Model idealny posiada wartość miary Sufficiency na poziomie równym 0 oraz posiada jak najwyższą wartość miary Zysk (prawa strona rysunku).

Źródło: opracowanie własne.

Warto wskazać, że metoda porównawcza **Unmitigated**, w której zmienna chroniona płeć jest włączona w trenowaniu modelu oceny zdolności kredytowej, charakteryzowała się stosunkowo wysokimi wartościami jakości prognostycznej, przy jednoczesnych, najniższych wartościach miary dostateczności (tj. miary wartości predykcyjnej dodatkowo dla poszczególnych grup wartości zmiennej chronionej są do siebie zbliżone). Najwyższymi wartościami kryterium dostateczności związanego z warunkiem parytetu kalibracji są modele sprawiedliwości przez nieświadomość **Fairness through Unawareness**, które są obecnie głównie wykorzystywane przez banki. Zestawienie kryterium dostateczności oraz poziomów miary Zysk przedstawiono w tabeli 12.

Tabela 12: Zestawienie wartości miary Zysk oraz kryterium dostateczności dla Scenariusz Syntetyczny

Podejście	Metoda redukcji stroniczości algorytmicznej	Model bazowy	Zysk	SE _{Zy}	SF	SE _{SF}
porównawcze	Unmitigated	LR	169	8,3	0,02	0,02
	Unmitigated	DT	240	8,7	0,04	0,02
	Unmitigated	RF	271	8,4	0,04	0,02
	Unmitigated	GB	270	7,9	0,02	0,01
	Fairness t. Unawareness	LR	150	8,4	0,12	0,02
	Fairness t. Unawareness	DT	228	9,7	0,10	0,02
	Fairness t. Unawareness	RF	252	8,8	0,12	0,02
	Fairness t. Unawareness	GB	249	8,5	0,11	0,02
przetreningowe	Reweighing	LR	162	8,3	0,09	0,03
	Reweighing	DT	236	9,2	0,08	0,02
	Reweighing	RF	265	8,6	0,10	0,02
	Reweighing	GB	263	8,4	0,08	0,02
	Disparate Impact Remover	LR	160	9,5	0,09	0,03
	Disparate Impact Remover	DT	239	9,1	0,08	0,02
	Disparate Impact Remover	RF	262	9,4	0,09	0,02
	Disparate Impact Remover	GB	260	8,8	0,08	0,02
treningowe	Adversarial Debiasing	NN	222	9,5	0,08	0,02
	Prejudice Remover	LR	164	11,0	0,10	0,04
	Grid Search Reduction	LR	165	9,2	0,09	0,03
	Grid Search Reduction	DT	240	12,3	0,11	0,03
	Grid Search Reduction	RF	259	8,9	0,09	0,02
	Grid Search Reduction	GB	265	8,6	0,07	0,02
potreningowe	Reject Option Classifier	LR	149	8,6	0,08	0,02
	Reject Option Classifier	DT	223	10,6	0,09	0,03
	Reject Option Classifier	RF	251	9,4	0,10	0,02
	Reject Option Classifier	GB	250	8,6	0,08	0,02

Źródło: opracowanie własne.

4.2.2 Zaburzenie struktury populacyjnej kredytobiorców

Drugim z badanych wymiarów, istotnym z perspektywy decydenta bankowego, jest wpływ nowego modelu na funkcjonujący proces oceny zdolności kredytowej. Obok badania zmiany jakości modelu (opisanej w podrozdziale 4.2.1), istotnym zagadnieniem decyzyjnym jest badanie potencjalnego wpływu zastosowania nowego modelu oceny zdolności kredytowej na profil akceptowanego klienta kredytowego. W przypadku, gdyby podjęta została decyzja o wdrożeniu metod redukcji stroniczości algorytmicznej, istotnym zagadnieniem jest weryfikacja poziomu wpływu wdrożenia nowego modelu na profile klientów, którzy uzyskują akceptację kredytową.

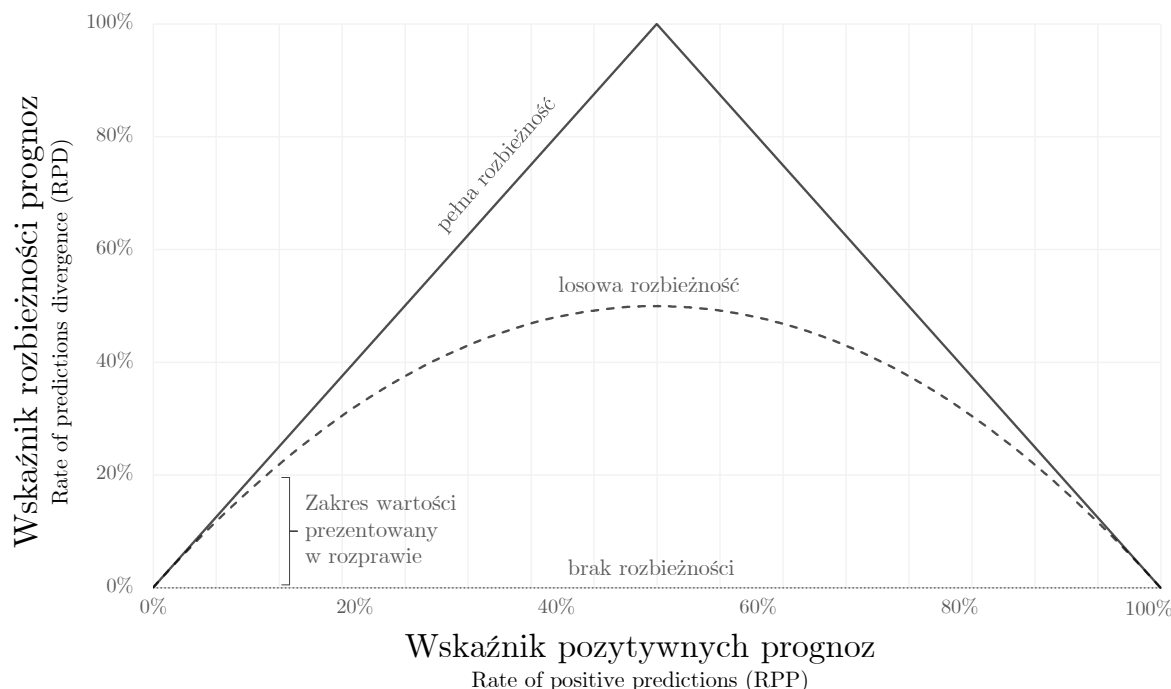
Rozbieżność decyzji kredytowej

Bank podejmując decyzję o wykorzystaniu metody redukcji stronniczości algorytmicznej⁸⁶, bierze pod uwagę zarówno rachunek ekonomiczny (tj. zysk) i zgodność z przepisami zakazującymi dyskryminacji. Jednym z elementów, pośrednio związanym z zyskiem, jest poziom rozbieżności jaką nowe podejście za sobą niesie – tj. w jaki sposób zmieni się rozkład wydawanych decyzji kredytowych. W niniejszej pracy przyjmuje się, że mniejszy poziom zaburzenia populacji wydawanych decyzji kredytowych, będzie bardziej przez bank preferowany.

W celu zobrazowania poziomu wpływu wdrożenia nowego modelu na populację kredytobiorców, poniżej przedstawiono autorskie podejście do wyznaczania rozbieżności w wydawanych decyzjach kredytowych nazwane *rysunkiem rozbieżności prognoz* (lub dalej również *rysunkiem rozbieżności* – ang. disturbance plot). Rysunek rozbieżności wskazuje na rozbieżność generowanych prognoz między dwoma modelami: nazwijmy te modele referencyjnym oraz porównawczym (lub porównywanym). Przyjmijmy, że model referencyjny będzie modelem wykorzystywanym w dotychczasowej praktyce oceny zdolności kredytowej (np. model charakteryzujący się stronniczością algorytmiczną). W przedstawionym w niniejszej rozprawie badaniu przyjmuje się, że modelem referencyjnym będzie model *Unmitigated* (tj. włączający explicite zmienną chronioną). Założenie to jest jednak założeniem bez utraty ogólności, tj. można jako model referencyjny przyjąć dowolny model generujący wartość oceny zdolności kredytowej. Modelem porównywanym jest model, który planowany jest do wdrożenia. Wartości rozbieżności prognoz będą mierzone dla modelu porównywanego, względem prognoz modelu referencyjnego. Przykład rozbieżności prognoz, bez odniesienia się do żadnej z metod ograniczania stronniczości algorytmicznej, przedstawiono na rysunku 16.

W ramach badania nie zaprezentowano wyników dla modelu potreningowego, podejście *Reject Option Classifier*, z racji na mechanizm działania tego podejścia, tj. dla danego rozkładu prawdopodobieństw spłaty (tj. wynik modelu oceny zdolności kredytowej, w którym nie zastosowano redukcji stronniczości algorytmicznej), w tym podejściu przesuwane są progi pozytywnej decyzji kredytowej. Oznacza to, że rozkład oceny punktowej (po kalibracji – prawdopodobieństwa) pozostaje bez zmian, ulega zmianie wyłącznie próg odcięcia decyzji kredytowej.

⁸⁶ Również szerzej: zmiana istniejącego modelu oceny zdolności kredytowej.



Rysunek 16: Rozbieżności prognoz, przykład bez wskazania żadnej z metod ograniczania stronicowości algorytmicznej

Uwaga: Oś x przedstawia poziom pozytywnych prognoz z modelu; oś ta może być interpretowana jako kolejne, coraz niższe wartości proggu odcięcia decyzji. Wartość 0% wskazuje na sytuację, w której żadna aplikacja kredytowa nie została zaakceptowana; wartość 100% wskazuje na sytuację, że wszystkie wnioski kredytowe zostały zaakceptowane.

Oś y przedstawia poziom rozbieżności prognoz. Wartość 0% oznacza, że prognozy obu modeli (tj. referencyjny oraz porównywany) są identyczne; wartości 100% oznacza, że modele mają dokładnie odmienne prognozy. Kształt krzywej rozbieżności dla:

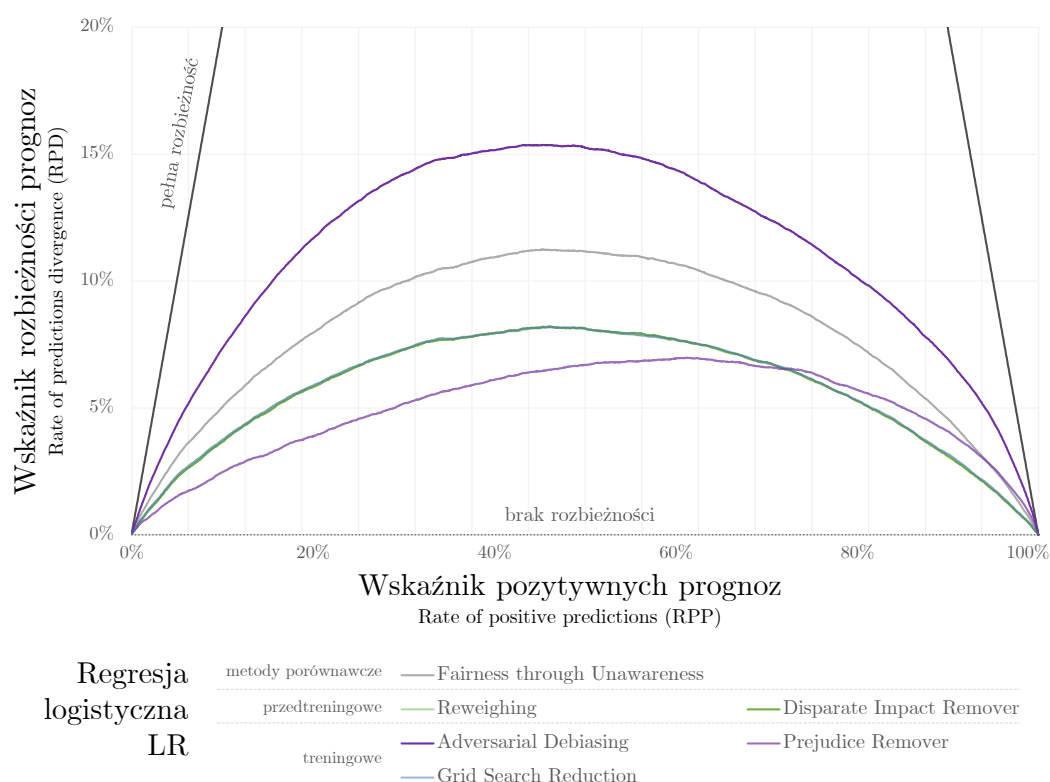
- Braku rozbieżności jest linią poziomą na wartości 0%, tj. dla każdego proggu odcięcia prognoza modelu referencyjnego i modelu porównywanego są tożsame.
- Pełnej rozbieżności jest trójkątem równoramiennym o wierzchołku o wartości 100%, tj. wraz ze wzrostem poziomu odcięcia prognozy, do mediany rozkładu oceny zdolności kredytowej poziom niezgodności rośnie, a później maleje (tj. w przypadku gdy akceptowane są wszystkie wnioski kredytowe, nawet jeżeli rozkład oceny zdolności kredytowej jest dokładnie przeciwny, prognoza z modeli jest taka sama).
- Losowej rozbieżności jest n-kształtna, co stanowi scenariusz mieszany pełnej oraz braku rozbieżności.

W rozprawie, z racji na rozkład wartości rozbieżności, krzywe rozbieżności będą prezentowane na rysunkach w przedziale 0 – 20%.

Źródło: opracowanie własne.

Krzywa rozbieżności – regresja logistyczna

Wpływ zastosowania metod ograniczania stroniczości algorytmicznej na poziom rozbieżności generowanych prognoz, gdzie modelem referencyjnym jest model uwzględniający zmienną chronioną (model *Unmitigated*), bazujący na regresji logistycznej – LR przedstawiono na rysunku 17. Z racji tego, że metoda *Adversarial Debiasing* oparta jest na architekturze sztucznych sieci neuronowych, dla tej metody nie przedstawiono modelu referencyjnego; postanowiono zestawzić wyniki tej metody z pozostałymi wynikami regresji logistycznej. Wszystkie rysunki rozbieżności prezentują rozbieżność wyznaczoną jako wartość średnia dla 100 symulacji – pojedyncza symulacja może charakteryzować się wysokim poziomem szumu losowego.



Rysunek 17: Średnie krzywe rozbieżności, Scenariusz Syntetyczny, model regresji logistycznej LR.

Źródło: opracowanie własne.

Z badania wykresu rozbieżności wynikają następujące, najważniejsze wnioski:

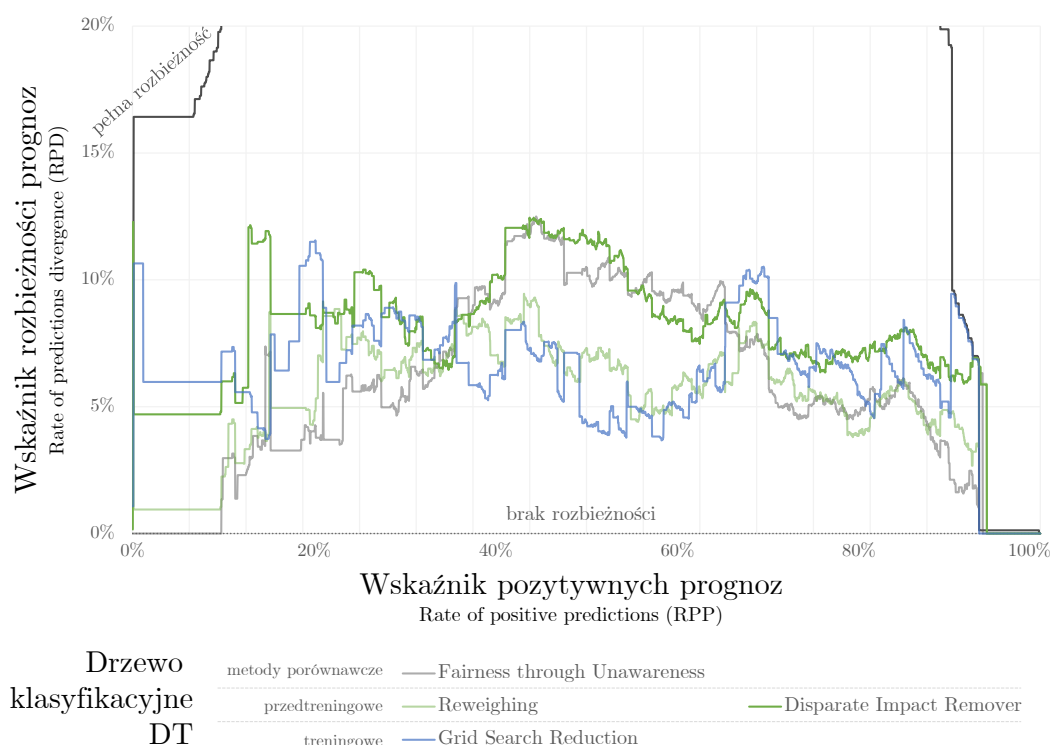
- W przypadku metod przedtreningowych, poziom rozbieżności względem modelu referencyjnego był umiarkowany; poziom tej rozbieżności był jednak niższy niż dla modelu *Fairness through Unawareness* – tj. modelu, w którym wyłączono zmienną chronioną; obie metody – *Reweighing* oraz *Disparate Impact Remover* – wskazują na wysoki poziom zbieżności generowanych prognoz (na rysunku 17 stykają się).
- W przypadku metod treningowych, rozbieżności są zróżnicowane; dla metody opartej na sztucznych sieciach neuronowych NN, poziom rozbieżności był istotnie różny–wyższy od

poziomu rozbieżności modelu **Fairness through Unawareness**. Z drugiej jednak strony, metoda **Prejudice Remover** jest metodą charakteryzującą się najniższym poziomem rozbieżności względem modelu referencyjnego w klasie metod treningowych.

- W przypadku metod potreningowych wskazać można na brak rozbieżności generowanych prognoz z modelu; stan ten wynika z faktu, że metoda **Reject Option Classifier** nie zmienia rang poszczególnych wniosków kredytowych (tj. n–ty najlepszy wniosek kredytowy w modelu referencyjnym jest również n–tym najlepszym wnioskiem kredytowym w modelu porównawczym) oraz nie zmienia wartości prognozowanego prawdopodobieństwa spłaty kredytu (tj. wnioski kredytowe, których ocena zdolności kredytowej jest „blisko” progu decyzji uzyskują odmienną prognozę). W dalszej części rozprawy pominięto opis rozbieżności dla metody **Reject Option Classifier**.

Krzywa rozbieżności – drzewa klasyfikacyjne

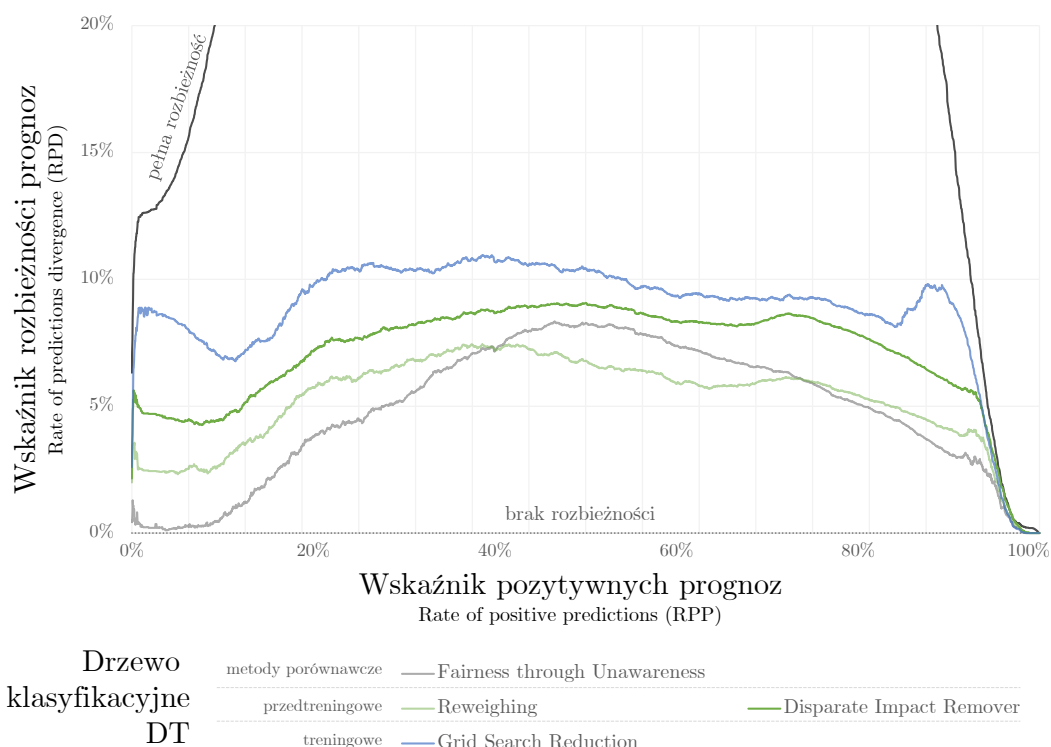
Poniżej przedstawiono wpływ zastosowania metod ograniczania stroniczości algorytmicznej na poziom rozbieżności generowanych prognoz. Modelem referencyjnym jest model uwzględniający zmienną chronioną (model **Unmitigated**), a oparty o model drzewa klasyfikacyjnego – DT. Krzywą rozbieżności dla *pojedynczego* modelu drzewa klasyfikacyjnego przedstawiono na rysunku 18



Rysunek 18: Pojedyncza krzywa rozbieżności, **Scenariusz Syntetyczny**, model drzewa klasyfikacyjnego DT.

Źródło: opracowanie własne.

W zakresie metod redukcji stroniczości algorytmicznej opartych na drzewach klasyfikacyjnych można wskazać, że wykres rozbieżności charakteryzuje się dużą zmiennością; niewielkie zmiany progu odcięcia mogą powodować skokowe zmiany generowanych prognoz, a co za tym idzie – rozbieżności między modelami referencyjnym i porównawczym. Uśrednione krzywe rozbieżności prognoz przedstawiono na rysunku 19.

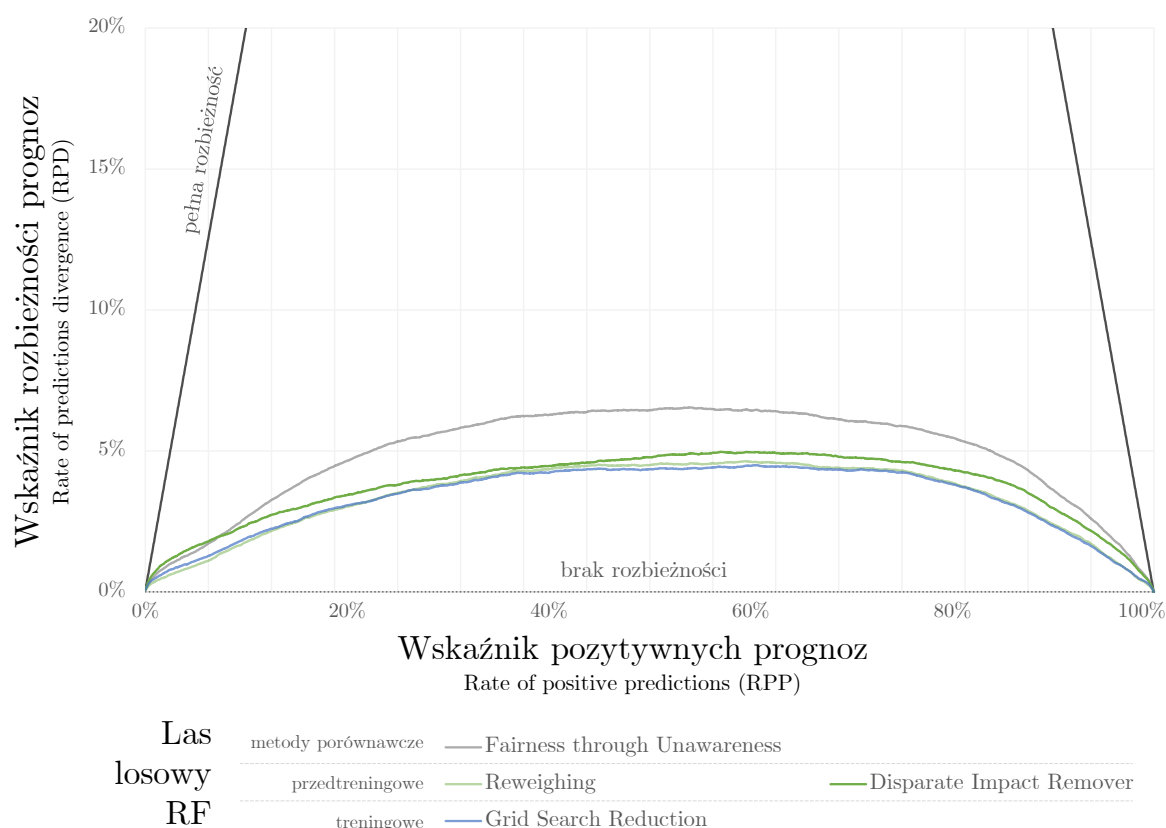


Rysunek 19: Średnie krzywe rozbieżności, Scenariusz Syntetyczny, model drzewa klasyfikacyjnego DT.

Źródło: opracowanie własne.

Krzywa rozbieżności – lasy losowe

W niniejszej podsekcji przedstawiono wpływ zastosowania metod ograniczania stroniczości algorytmicznej na poziom rozbieżności generowanych prognoz – modelem referencyjnym jest model uwzględniający zmienną chronioną (model *Unmitigated*), o architekturze modelu lasu losowego – RF. Wykres rozbieżności decyzji kredytowej, zakładając kolejne metody redukcji stroniczości algorytmicznej, przy wykorzystaniu lasów losowych – RF, zamieszczono na rysunku 20. W przypadku lasu losowego poziom rozbieżności nie jest znaczący. Spośród wszystkich przebadanych podejść najniższym poziomem rozbieżności charakteryzują się metody przedtreningowe.



Rysunek 20: Średnie krzywe rozbieżności, **Scenariusz Syntetyczny**, model lasu losowego RF.
Źródło: opracowanie własne.

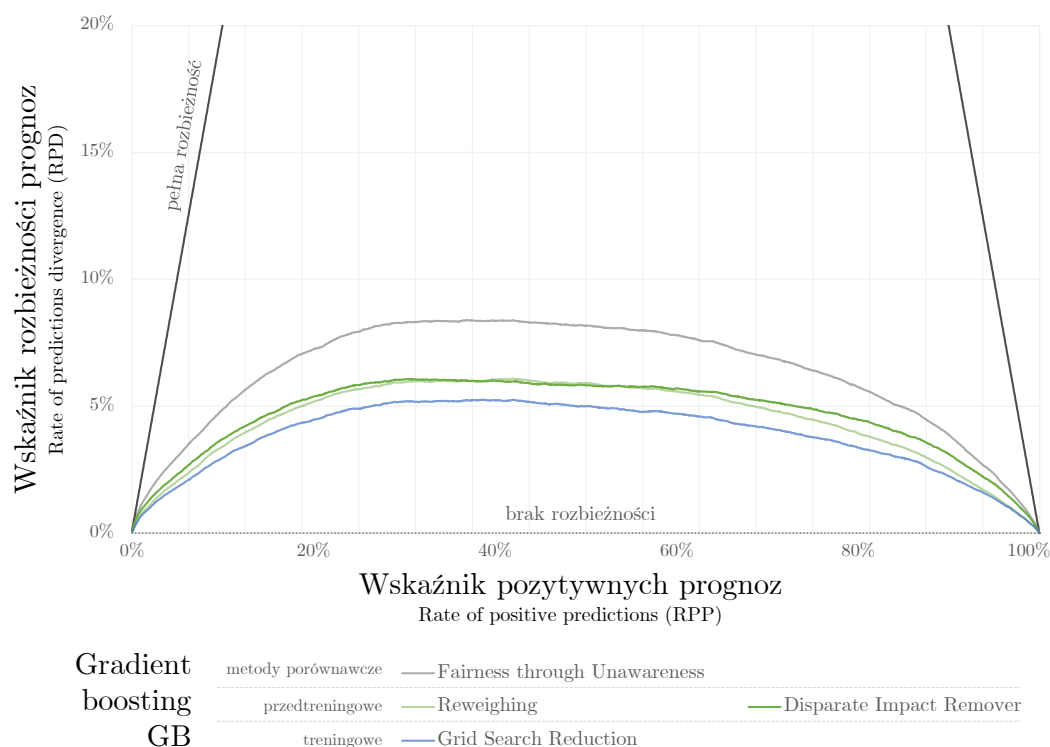
Krzywa rozbieżności – gradient boosting

W niniejszej podsekcji przedstawiono wpływ zastosowania metod ograniczania stroniczności algorytmicznej na poziom rozbieżności generowanych prognoz; modelem referencyjnym jest model uwzględniający zmienną chronioną (model **Unmitigated**) oparty o model gradient boosting – GB.

Wykres rozbieżności decyzji kredytowej, zakładając kolejne metody redukcji stroniczności algorytmicznej, przy wykorzystaniu metody gradient boosting – GB zamieszczono na rysunku 21.

Z badania wykresu rozbieżności wynikają następujące, najważniejsze wnioski:

- W przypadku metod przedtreningowych, poziom rozbieżności względem modelu referencyjnego był umiarkowany; poziom tej rozbieżności był jednak niższy niż dla modelu **Fairness through Unawareness** – tj. modelu, w którym wyłączono zmienną prawnie chronioną; obie z badanych metod – **Reweighing** oraz **Disparate Impact Remover** – wskazują na wysoki poziom zbieżności generowanych prognoz między sobą;
- W przypadku metody treningowej – **Grid Search Reduction**, rozbieżności między prognozami względem modelu referencyjnego są niewielkie i nie przekraczają 1,7%.



Rysunek 21: Średnie krzywe rozbieżności, **Scenariusz Syntetyczny**, model gradient boosting **GB**.

Źródło: opracowanie własne.

W ramach badania wpływu zastosowania modeli ograniczania stronniczości algorytmicznej przeprowadzono analizę zaburzenia struktury zaakceptowanych wniosków kredytowych, która odnosi się do wpływu modelu oceny zdolności kredytowej na rozkład wyników procesu kredytowego. Wskazane badanie miało na celu przedstawienie poziomu zaburzenia rozkładu akceptowanych wniosków kredytowych poprzez zastosowanie metody ograniczania stronniczości algorytmicznej. W celu przedstawienia skondensowanej informacji o wpływie zastosowania poszczególnych metod redukcji stronniczości algorytmicznej, przedstawiono średni poziom rozbieżności prognozy dla wszystkich poziomów akceptacji kredytowej, jak również wskazano statystyki opisowe rozkładu rozbieżności: 25%, 50% oraz 75% percentyl rozkładu wartości rozbieżności (tabela 13).

Tabela 13: Średni poziom rozbieżności prognozy dla wszystkich poziomów akceptacji kredytowej

Model bazowy	Metoda redukcji stroniczości algorytmicznej	Średnia rozbieżność	25%	50%	75%
LR	Fairness through Unawareness	7,8%	5,6%	8,8%	10,6%
	Reweighing	5,7%	4,1%	6,4%	7,7%
	Disparate Impact Remover	5,7%	4,0%	6,4%	7,7%
	Prejudice Remover	4,9%	3,5%	5,4%	6,5%
	Grid Search Reduction	5,7%	4,1%	6,4%	7,7%
DT	Fairness through Unawareness	4,8%	2,9%	5,2%	7,1%
	Reweighing	5,2%	4,1%	5,8%	6,5%
	Disparate Impact Remover	7,1%	5,9%	8,1%	8,5%
	Grid Search Reduction	8,8%	8,5%	9,3%	10,4%
RF	Fairness through Unawareness	4,8%	3,7%	5,6%	6,3%
	Reweighing	3,3%	2,4%	3,9%	4,4%
	Disparate Impact Remover	3,7%	3,0%	4,2%	4,7%
	Grid Search Reduction	3,3%	2,5%	3,8%	4,3%
GB	Fairness through Unawareness	6,3%	4,9%	7,0%	8,1%
	Reweighing	4,4%	3,4%	5,0%	5,8%
	Disparate Impact Remover	4,6%	3,8%	5,2%	5,8%
	Grid Search Reduction	3,8%	3,0%	4,2%	5,0%
NN	Adversarial Debiasing	11,0%	8,2%	12,2%	14,7%

Uwaga: Wartości przedstawione w kolumnach zostały uśrednione na podstawie 100 symulacji. Wartości większe wskazują na wyższy poziom rozbieżności prognozy modelu oceny zdolności kredytowej, względem modelu referencyjnego (tj. w badaniu wykorzystano model *Unmitigated*).

Źródło: opracowanie własne.

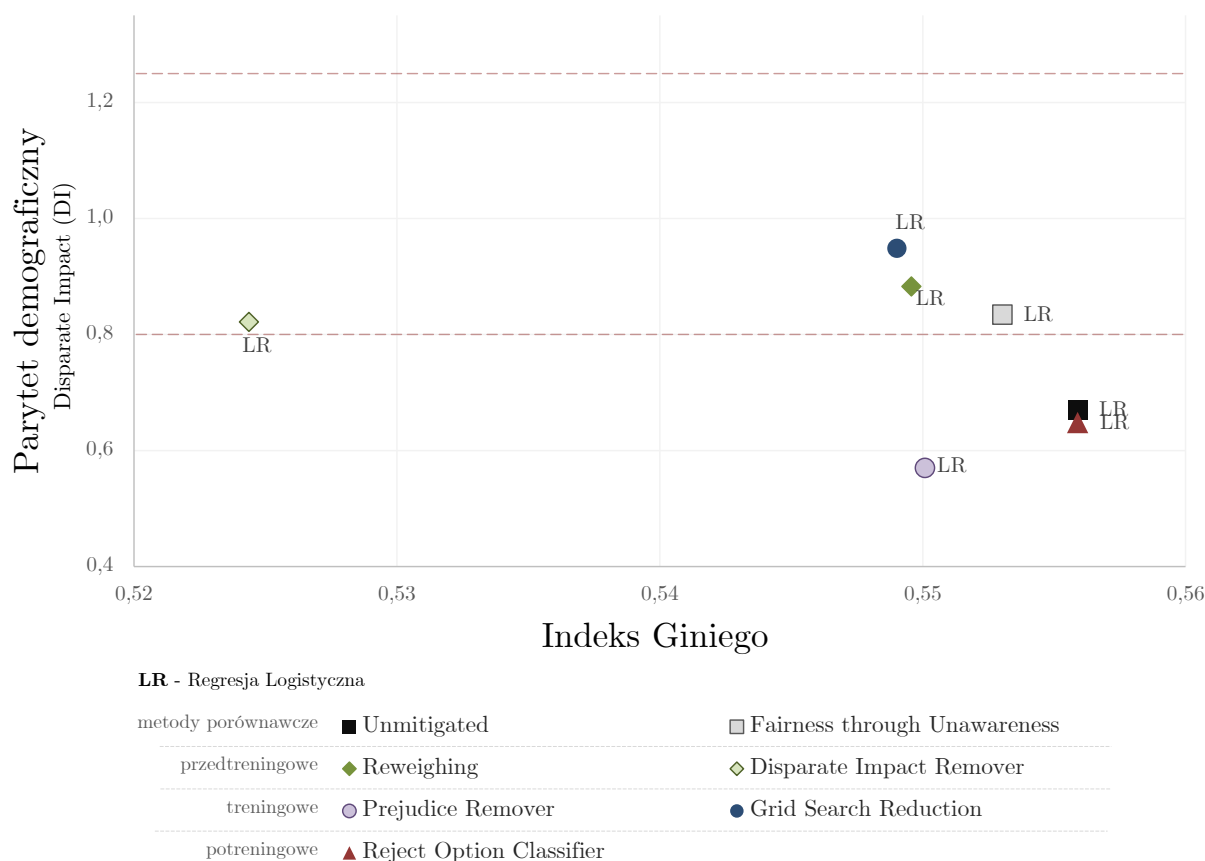
4.3 Scenariusz danych rzeczywistych

Scenariusz German Credit Data jest oparty na danych rzeczywistych⁸⁷. Zmienną chronioną, którą przyjęto w przeprowadzonym badaniu jest zmienna **narodowość**: jest to zmienna binarna, przyjmująca wartość 1 dla obcokrajowców. Celem badania przedstawionego w tym podrozdziale jest przedstawienie problemu stroniczości algorytmicznej na przykładzie danych rzeczywistych. Należy jednak wskazać, że charakteryzowały się one nieznaczną wielkością próby (1000 obserwacji); z racji na niewielką próbę oraz jednoczesne wykorzystanie metody symulacji obserwacji, omawianymi modelami referencyjnymi będą modele oparte na regresji logistycznej.

⁸⁷ „Dane kredytowe zawierające 20 zmiennych oraz zmienną celu: czy wnioskodawca należy do dobrej czy złej klasy ryzyka kredytowego dla 1000 osób ubiegających się o pożyczkę.” [tł. wł.], (Hofmann 1994).

Warunek parytetu demograficznego

Wpływ zastosowania poszczególnych metod redukcji stronniczości algorytmicznej na indeks Giniego (opisany w podrozdziale 2.2.2) przedstawiono na rysunku 22. Wszystkie raportowane w rozprawie wyniki są wyznaczane na podstawie zbioru testowego.



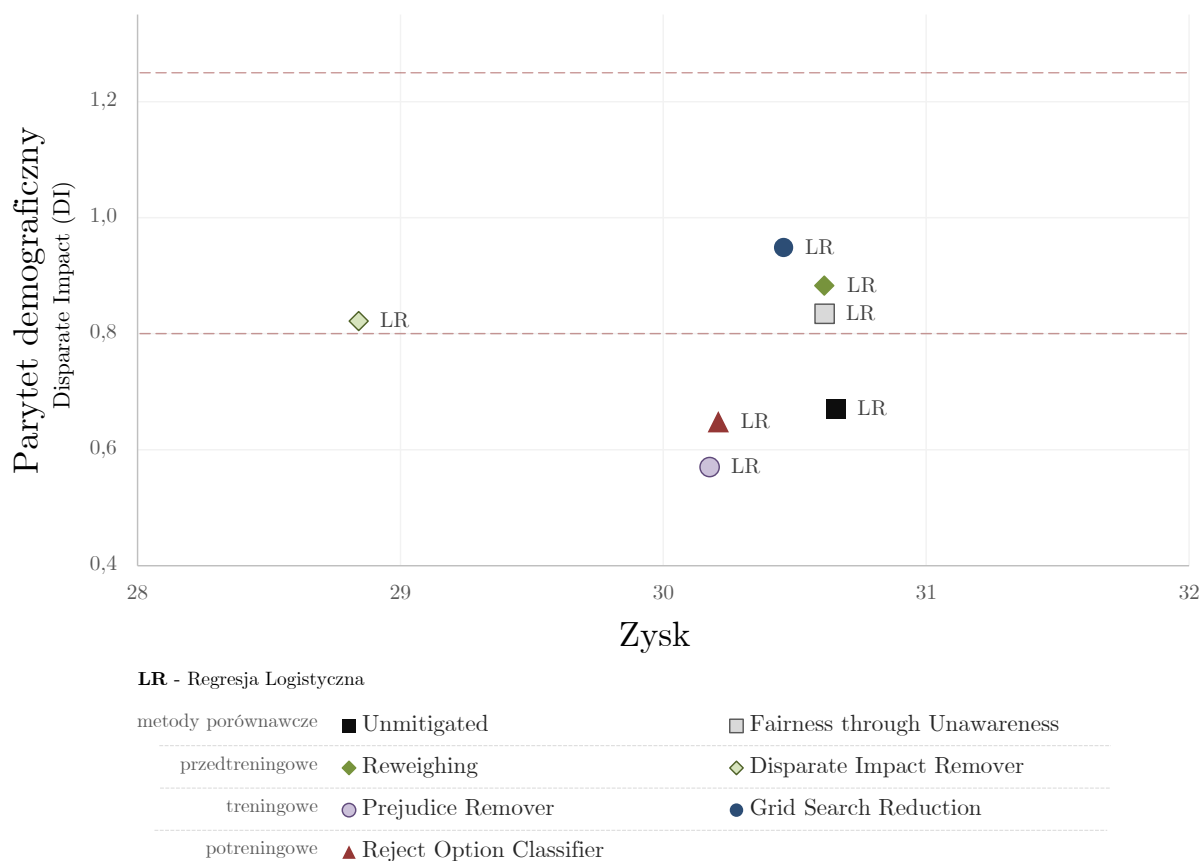
Rysunek 22: Wartość indeksu Giniego, a warunek parytetu demograficznego, **Scenariusz German Credit Data**

Uwaga: Średnie wartości dla 100 powtórzeń symulacji. Przerywanymi, czerwonymi liniami poziomymi zaznaczono krytyczne poziomy 80% oraz 125% wartości miary **Disparate Impact**; obserwacje wartości dla modeli znajdujących się między tymi poziomami spełnią kryterium 4/5 opisane w podrozdziale 3.1.1. Model idealny posiada wartość miary **Disparate Impact** na poziomie równym 1 i posiada jak najwyższą wartość indeksu Giniego (prawa strona rysunku).

Źródło: opracowanie własne.

Ocenę wpływu zastosowania metod ograniczania stronniczości algorytmicznej na ekonomiczną wartość modelu, tj. poziom zysku uzyskiwany przy zastosowaniu danego modelu oceny zdolności kredytowej dla poszczególnych podejść do redukcji stronniczości algorytmicznej, przedstawiono na rysunku 23.

Średnie wartości miary **Zysk** oraz miary związanej z warunkiem parytetu demograficznego – **Disparate Impact** – uzyskanych dla poszczególnych metod ograniczania stronniczości algorytmicznej, wraz ze wskazaniem na wartość błędu standardowego (ang. standard error – SE), zostały przedstawione w tabeli 14. Wartości błędu standardowego wskazują na niepewność uzyskanych wyników: czym większe



Rysunek 23: Miara Zysk a warunek parytetu demograficznego, **Scenariusz German Credit Data**
 Uwaga: Średnie wartości dla 100 powtórzeń symulacji. Przerywanymi, czerwonymi liniami poziomymi zaznaczono krytyczne poziomy 80% oraz 125% wartości miary Disparate Impact; obserwacje wartości dla modeli znajdujących się między tymi poziomami spełniają kryterium 4/5 opisane w podrozdziale 3.1.1. Model idealny posiada wartość miary Disparate Impact na poziomie równym 1 oraz posiada jak najwyższą wartość miary Zysk (prawa strona rysunku).
 Źródło: opracowanie własne.

wartości SE, tym wyższy poziom niepewności uzyskanych wyników.

Wyniki uzyskane dla danych rzeczywistych wskazują na niższe, niż w przypadku przyjętego procesu symulacyjnego znaczenie zmiennej chronionej: występuje znacznie mniejszy spadek zysku pomiędzy wykorzystaniem metod Unmitigated oraz Fairness through Unawareness. Model, który zbliżył się najbardziej do wartości 1,00 względem miary Disparate Impact, był model Grid Search Reduction. Istotnie najslabsze wyniki, otrzymano metodą Disparate Impact Remover – istotny spadek zysku, przy jednoczesnym porównywalnym poziomie miary Disparate Impact dla wyników uzyskanych metodą Fairness through Unawareness.

Warunek wyrównanych szans

Kolejnym z kryteriów poddanych badaniu bezpośredniego wpływu zastosowania metod redukcji stronniczości algorytmicznej był warunek wyrównanych szans. Jak wskazano wcześniej w rozprawie, czym wyższa jest wartość kryterium separacji (kryterium związane

Tabela 14: Wartości Zysk oraz Disparate Impact dla Scenariusz German Credit Data

Podejście	Metoda redukcji stronniczości algorytmicznej	Zysk	SE _{Zy}	DI	SE _{DI}	$\chi^2\%$	F%
porównawcze	Unmitigated	31	1,93	0,67	0,10	0,93	0,90
	Fairness t. Unawareness	31	1,85	0,83	0,23	1,00	0,99
przedtreningowe	Reweighing	31	1,84	0,88	0,29	0,99	0,99
	Disparate Impact Remover	29	3,72	0,82	0,19	1,00	1,00
treningowe	Prejudice Remover	30	2,03	0,57	0,08	0,64	0,51
	Grid Search Reduction	30	1,96	0,95	0,34	1,00	1,00
potreningowe	Reject Option Classifier	30	1,98	0,65	0,10	0,92	0,88

Uwaga: Kolumny $\chi^2\%$ oraz F% odnoszą się do frakcji symulacji, dla których odpowiednio test χ^2 (Tallarida i Murray 2012) oraz test Fishera (Fisher 1970), przy 1% poziomie istotności nie dał podstaw do odrzucenia H_0 , która to hipoteza stanowiła o braku różnic między populacją obserwacji uprzywilejowanych i nieuprzywilejowanych. Wartości 1,00 wskazuje na metody, dla których we wszystkich symulacjach nie było podstaw do odrzucenia H_0 (tj. nie występowały statystyczne różnice między poziomem akceptacji kredytowych dla poszczególnych populacji).

Źródło: opracowanie własne.

z warunkiem wyrównanych szans), tym większa jest różnica w błędach prognozy dla poszczególnych grup zmiennej chronionej. Zbiorcze wyniki jakości prognostycznej mierzonej miarą Zysk oraz wartości kryterium separacji przedstawiono na rysunku 24.

Obecnie wykorzystywane przez banki podejście – **Fairness through Unawareness** – charakteryzuje się niskimi wartościami kryterium separacji; średnie wartości miary Zysk oraz miary kryterium separacji przedstawiono w tabeli 15.

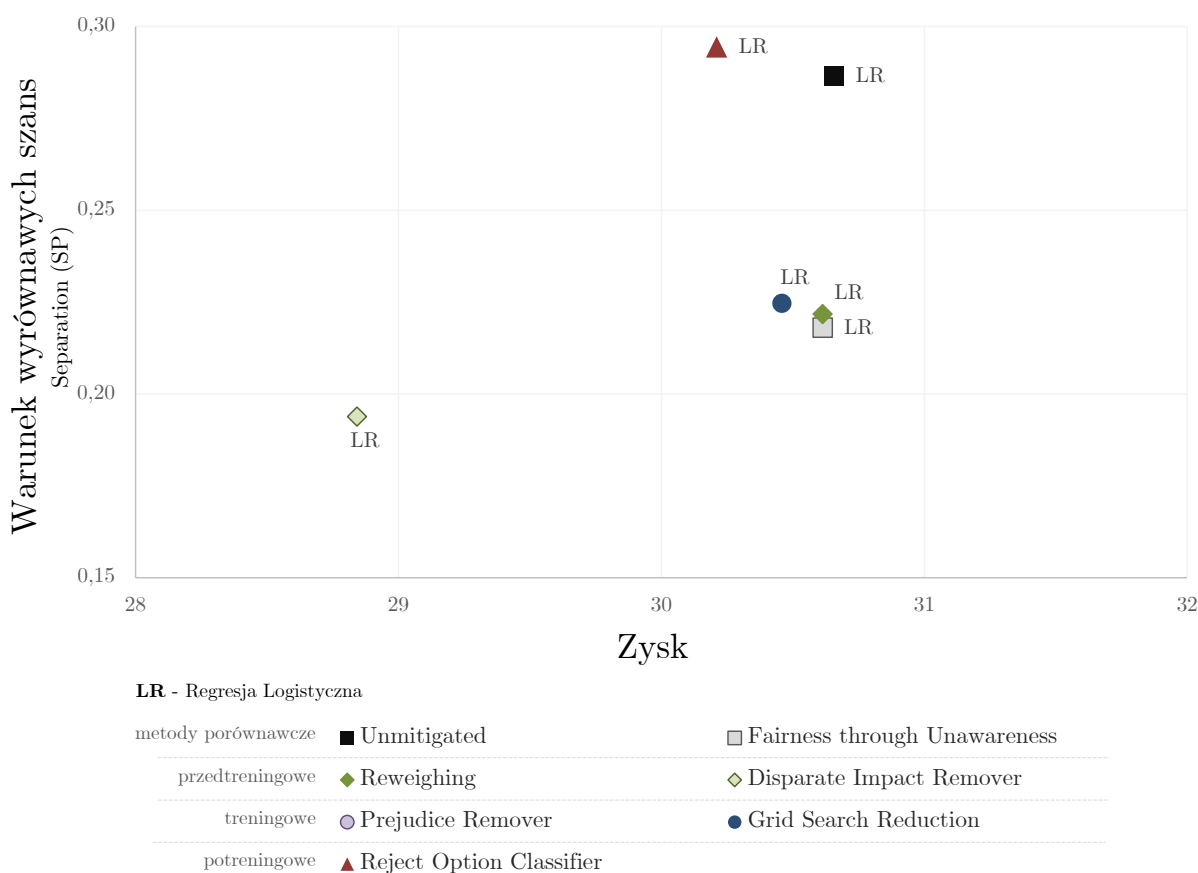
Tabela 15: Zestawienie wartości miary Zysk oraz kryterium separacji dla Scenariusz German Credit Data

Podejście	Metoda redukcji stronniczości algorytmicznej	Zysk	SE _{Zy}	SP	SE _{SP}
porównawcze	Unmitigated	31	1,93	0,29	0,17
	Fairness t. Unawareness	31	1,85	0,22	0,13
przedtreningowe	Reweighing	31	1,84	0,22	0,14
	Disparate Impact Remover	29	3,72	0,19	0,11
treningowe	Prejudice Remover	30	2,03	0,46	0,24
	Grid Search Reduction	30	1,96	0,22	0,13
potreningowe	Reject Option Classifier	30	1,98	0,29	0,17

Źródło: opracowanie własne.

Warunek parytetu kalibracji

Ostatnią z zaprezentowanych miar jest miara związana z warunkiem parytetu kalibracji. Analogicznie jak w przypadku warunku wyrównanych szans, wyższe wartości



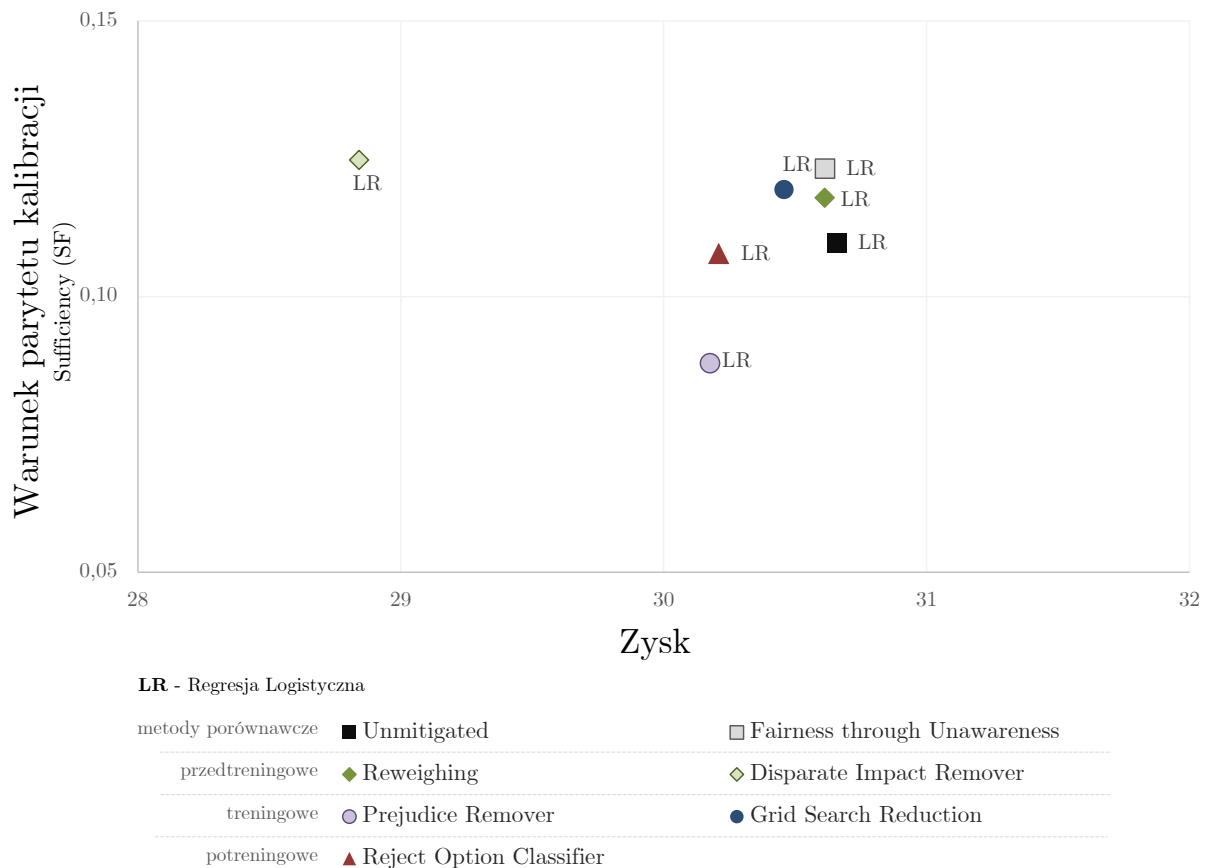
Rysunek 24: Miara **Zysk** a warunek wyrównanych szans, **Scenariusz German Credit Data**

Uwaga: Średnie wartości dla 100 powtórzeń symulacji. Wartości mniejsze kryterium separacji, związanego z warunkiem wyrównanych szans wskazują na mniejsze rozbieżności między błędami klasyfikacji dla poszczególnych grup zmiennej chronionej (mężczyźni i kobiety). Model idealny posiada wartość miary **Separation** na poziomie równym 0 oraz posiada jak najwyższą wartość miary **Zysk** (prawa strona rysunku).

Źródło: opracowanie własne.

miary kryterium dostateczności (związanego z warunkiem parytetu kalibracji) sugerują występowanie stronnictwa algorytmicznego. Zbiorcze wyniki jakości prognostycznej mierzonej miarą **Zysk** oraz wartości kryterium kalibracji przedstawiono na rysunku 25.

Warto wskazać, że metoda porównawcza **Unmitigated**, w której zmienna chroniona **płeć** jest włączona w trenowanie modelu oceny zdolności kredytowej, charakteryzowała się stosunkowo wysokimi wartościami jakości prognostycznej, przy jednoczesnych, najniższych wartościach miary dostateczności (tj. miary wartości predykcyjnej dodatniej dla poszczególnych grup wartości zmiennej chronionej są do siebie zbliżone). Najwyższymi wartościami kryterium dostateczności związanego z warunkiem parytetu kalibracji charakteryzują się modele sprawiedliwości przez nieświadomość **Fairness through Unawareness**, które są obecnie głównie wykorzystywane przez banki. Zestawienie kryterium dostateczności oraz poziomów miary **Zysk** przedstawiono w tabeli 16.



Rysunek 25: Miara Zysk a warunek parytetu kalibracji, Scenariusz German Credit Data

Uwaga: Średnie wartości dla 100 powtórzeń symulacji. Wartości mniejsze kryterium dostateczności, związanego z warunkiem parytetu kalibracji wskazują na mniejsze rozbieżności między błędami dla poszczególnych grup zmiennej chronionej (mężczyźni i kobiety). Model idealny posiada wartość miary Sufficiency na poziomie równym 0 oraz posiada jak najwyższą wartość miary Zysk (prawa strona rysunku).

Źródło: opracowanie własne.

Tabela 16: Zestawienie wartości miary Zysk oraz kryterium dostateczności dla Scenariusz German Credit Data

Podejście	Metoda redukcji stroniczości algorytmicznej	Zysk	SE _{Zy}	SF	SE _{SF}
porównawcze	Unmitigated	31	1,93	0,11	0,06
	Fairness t. Unawareness	31	1,85	0,12	0,05
przedtreningowe	Reweighing	31	1,84	0,12	0,05
	Disparate Impact Remover	29	3,72	0,12	0,05
treningowe	Prejudice Remover	30	2,03	0,09	0,05
	Grid Search Reduction	30	1,96	0,12	0,05
potreningowe	Reject Option Classifier	30	1,98	0,11	0,05

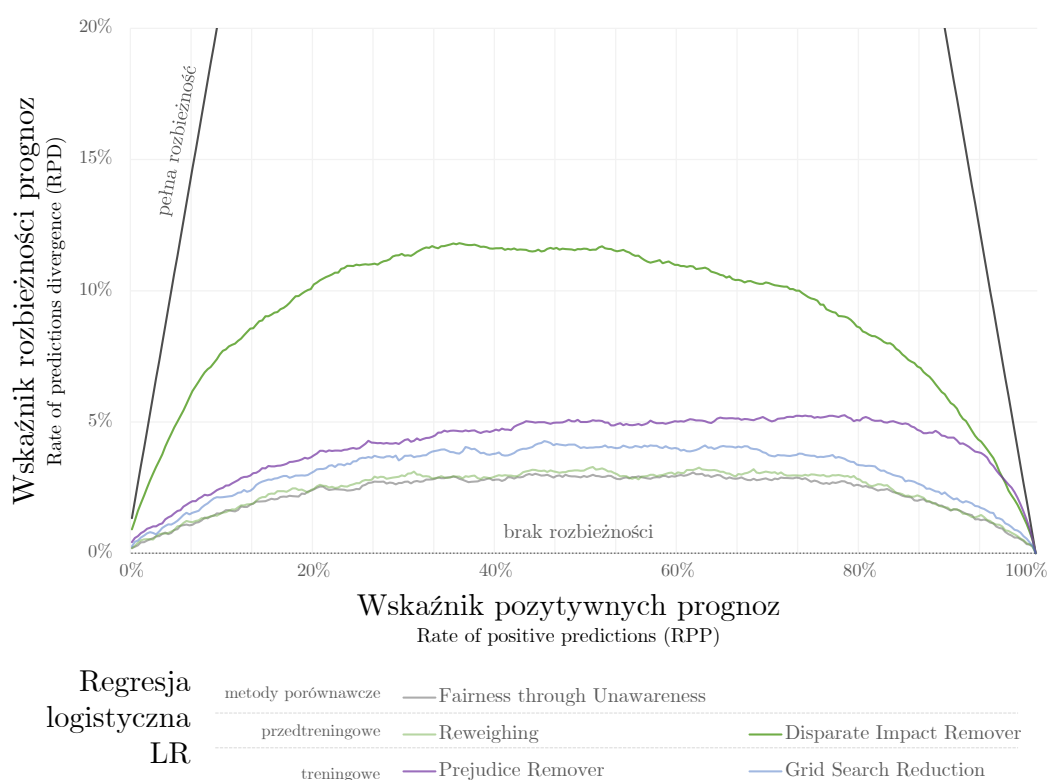
Źródło: opracowanie własne.

4.3.1 Zaburzenie struktury populacyjnej kredytobiorców

Analogicznie jak dla scenariusza danych syntetycznych, w przypadku scenariusza German Credit Data przeprowadzono analizę poziomu rozbieżności prognoz.

Krzywa rozbieżności – regresja logistyczna

Wpływ zastosowania metod ograniczania stronniczości algorytmicznej na poziom rozbieżności generowanych prognoz, gdzie modelem referencyjnym jest model uwzględniający zmienną chronioną (model *Unmitigated*) regresji logistycznej – LR przedstawiono na rysunku 26.



Rysunek 26: Średnie krzywe rozbieżności, **Scenariusz Syntetyczny**, model regresji logistycznej LR.

Źródło: opracowanie własne.

Poniżej zestawiono najważniejsze wnioski wynikające z badania wykresu rozbieżności:

- W przypadku metod przedtreningowych, poziom rozbieżności względem modelu referencyjnego był zróżnicowany; poziom tej rozbieżności był jednak wyższy niż dla modelu *Fairness through Unawareness* – tj. modelu, w którym wyłączono zmienną chronioną; metoda *Disparate Impact Remover* charakteryzowała się najwyższym poziomem rozbieżności generowanych prognoz (najwyżej leżąca krzywa na rysunku 26).
- W przypadku metod treningowych, rozbieżności są zróżnicowane; metoda *Prejudice*

Remover charakteryzuje się najniższym poziomem rozbieżności względem modelu referencyjnego w klasie metod treningowych.

W ramach badania wpływu zastosowania modeli ograniczania stronniczości algorytmicznej przeprowadzono analizę zaburzenia struktury zaakceptowanych wniosków kredytowych, która odnosi się do wpływu modelu oceny zdolności kredytowej na rozkład wyników procesu kredytowego. Wskazane badanie miało na celu przedstawienie poziomu zaburzenia rozkładu akceptowanych wniosków kredytowych poprzez zastosowanie metody ograniczania stronniczości algorytmicznej. W celu przekazania skondensowanej informacji o wpływie zastosowania poszczególnych metod redukcji stronniczości algorytmicznej, przedstawiono średni poziom rozbieżności prognozy dla wszystkich poziomów akceptacji kredytowej, jak również wskazano statystyki opisowe rozkładu rozbieżności: 25%, 50% oraz 75% percentyl rozkładu wartości rozbieżności (tabela 17).

Tabela 17: Średni poziom rozbieżności prognozy dla wszystkich poziomów akceptacji kredytowej

Model bazowy	Metoda redukcji stronniczości algorytmicznej	Średnia rozbieżność	25%	50%	75%
LR	Fairness through Unawareness	2,3%	1,9%	2,7%	2,9%
	Reweighing	2,4%	1,9%	2,9%	3,0%
	Disparate Impact Remover	9,0%	7,6%	10,3%	11,4%
	Prejudice Remover	4,1%	3,6%	4,7%	5,0%
	Grid Search Reduction	3,1%	2,4%	3,7%	4,0%

Uwaga: Wartości przedstawione w kolumnach zostały uśrednione na podstawie 100 symulacji. Wartości większe wskazują na wyższy poziom rozbieżności prognozy modelu oceny zdolności kredytowej, względem modelu referencyjnego (tj. w badaniu wykorzystano model **Unmitigated**).

Źródło: opracowanie własne.

* * *

Celem niniejszego rozdziału było opracowanie procedury redukcji stronniczości algorytmicznej oraz zastosowanie jej na przykładowych danych – syntetycznych i rzeczywistych. W ramach rozdziału 4 opracowano procedurę wspierającą banki w badaniu bezpośredniego wpływu redukcji stronniczości algorytmicznej, tzw. *problem decyzyjny banku*. Decyzje, które banki podejmują związane z modelami oceny zdolności kredytowej konsumentów dotyczą maksymalizacji zysku z realizacji procesu kredytowego, przy uwzględnieniu ograniczeń związanych z regulacjami zakazującymi dyskryminacji. Utworzona procedura umożliwia bankom weryfikację poszczególnych podejść redukcji stronniczości algorytmicznej w zakresie oceny jakości tej redukcji (tj. czy modele przestają być stronnicze algorytmicznie) oraz rachunku ekonomicznego związanego z procesem kredytowym.

W ramach zaprezentowanego w niniejszym rozdziale badania wykorzystano autorski model symulacyjny – model statyczny – umożliwiający ocenę wybranych podejść do redukcji stroniczości algorytmicznej w zakresie:

- skuteczności redukcji problemu stroniczości algorytmicznej w kontekście modeli oceny zdolności kredytowych;
- skutków zastosowania tych metod na realizowaną ocenę zdolności kredytowej, w tym rachunek ekonomiczny banku.

W rozdziale tym przedstawiono wyniki badania przeprowadzonego na podstawie danych symulacyjnych oraz danych rzeczywistych. Wyniki tego badania są związane z trzema hipotezami badawczymi rozprawy.

Hipoteza 1: O nieefektywności ekonomicznej metody sprawiedliwości przez nieświadomość. Na podstawie przeprowadzonych prac i wykonanego, z wykorzystaniem modelu statycznego badania hipotezę zweryfikowano pozytywnie: metoda sprawiedliwości przez nieświadomość, bezwarunkowo, nie jest metodą efektywną ekonomicznie. W przypadku scenariusza danych syntetycznych: **Scenariusz Syntetyczny** wykazano, że odrzucenie informacji o zmiennej prawnie chronionej może prowadzić do istotnego spadku zysku z generowanego procesu kredytowego, przy równoczesnym braku spełnienia warunków na brak dyskryminacji pośredniej lub bezpośredniej (tabele 10, 11, 12). Dla scenariusza danych rzeczywistych: **Scenariusz German Credit Data**, odrzucenie danych o zmiennej chronionej, jedynie w przypadku warunku wyrównanych szans, było ekonomicznie uzasadnione (tabela 15); zarówno dla warunku parytetu demograficznego oraz warunku jakości kalibracji istniały ekonomicznie efektywniejsze metody (tabele 14, 16).

Hipoteza 2: O skuteczności redukcji stroniczości algorytmicznej w zależności od miary. Zastosowanie metod redukcji stroniczości algorytmicznej związane jest z nałożeniem restrykcji na model; w przypadku zastosowania metod redukcji stroniczości algorytmicznej, jakość progностyczna (a tym samym zysk z procesu kredytowego) modelu oceny zdolności kredytowej nie będzie większa od modelu bez tych restrykcji (tj. modelu jawnie wykorzystującego zmienną prawnie chronioną). W badaniu wykazano, że w zależności od przyjętej definicji dyskryminacji: odmienny wpływ lub odmienne traktowanie, w danym kontekście danych, inne metody redukcji stroniczości algorytmicznej były ekonomicznie efektywne. W badanych scenariuszach danych modelami oceny zdolności kredytowej, które charakteryzowały się najwyższym poziomem zyskowności były modele reprezentujące każdą z klas rozwiązań związanych z redukcją stroniczości algorytmicznej: metody przedtreningowe, treningowe i potreningowe.

Hipoteza 3: O procedurze redukcji stronniczości algorytmicznej maksymalizującej zysk banku. Procedura określająca zysk banku, związany z wykorzystaniem modelu oceny zdolności kredytowej, została w ramach rozprawy opracowana i przedstawiona jako autorski model statyczny. Jednym z założeń tego podejścia jest określenie postaci macierzy kosztów banku. Dla rozważanych scenariuszy danych – dane syntetyczne oraz dane rzeczywiste – przeprowadzono symulację działania modelu statycznego, która wskazywała oszacowania stronniczości algorytmicznej poszczególnych metod, jak również poziomów zyskowności banku.

Dodatkowym rozwinięciem literatury przedmiotu w zakresie problemu stronniczości algorytmicznej modelu oceny zdolności kredytowej, tj. obok oceny rachunku ekonomicznego banku, jak i zgodności z przepisami zakazującymi dyskryminacji, było przedstawienie autorskiego podejścia do wyznaczania rozbieżności w wydawanych decyzjach kredytowych nazwane *rysunkiem rozbieżności prognoz* (lub *rysunkiem rozbieżności* – ang. disturbance plot), który wskazuje na rozbieżność generowanych prognoz między dwoma modelami: modelem referencyjnym oraz porównawczym (lub porównywanym).

Jednym z założeń badania przeprowadzonego w niniejszym rozdziale – badania bezpośredniego wpływu redukcji stronniczości algorytmicznej – było założenie, że w badaniu nie jest brane pod uwagę, jaki model oceny zdolności kredytowej jest stosowany przez bank *przed* rozpoczęciem wykorzystania metod redukcji stronniczości algorytmicznej. Założenie to zostanie uchylone w rozdziale 5, w którym model dynamiczny umożliwia badanie wpływu włączenia modelu redukcji stronniczości algorytmicznej w miejsce modelu do tej pory wykorzystywanego.

Rozdział 5

Wielookresowy wpływ redukcji stronniczości algorytmicznej

Celem niniejszego rozdziału jest opracowanie procedury wspierającej ocenę skutków zastosowania określonej definicji dyskryminacji (miara stronniczości algorytmicznej) w regulacji – *perspektywa regulatora rynku kredytów konsumentów*. Przedstawiono w nim zastosowanie wskazanej procedury oparte na przykładowych danych. Jako metodę badawczą w niniejszym rozdziale wykorzystano również symulacje; konstruowany model jest modelem wielookresowym, tj. modelem dynamicznym.

Modele decyzyjne konstruowane są na podstawie zadanej próby obserwacji zmiennych objaśniających oraz zmiennej endogenicznej; decyzje oparte na wskazaniach tych modeli mają jednak konsekwencje długoterminowe (Liu i in. 2018). W przypadku modelowej oceny zdolności kredytowej terminowa spłata zobowiązania kredytowego oznacza, przy założeniu komercyjnych warunków kredytowania, zysk dla banku udzielającego finansowania przy równoczesnej poprawie ogólnej sytuacji kredytobiorcy. W przypadku odwrotnego scenariusza, tj. braku spłaty zobowiązania, bank może ponieść stratę, a sytuacja finansowa kredytobiorcy się pogarsza, np.: spada zdolność kredytowa, powstają koszty wynikające z niespłaconego kredytu.

Motywacją przeprowadzenia badania wielookresowego była ocena konsekwencji stosowania metod ograniczania stronniczości algorytmicznej – *perspektywa decyzyjna regulatora*. Wprowadzane regulacje obejmujące konieczność redukcji stronniczości algorytmicznej, wiązać się będzie ze zmianą procedur budowy modeli oceny zdolności kredytowej konsumentów oraz wpływem tej zmiany na sytuację ekonomiczną banku jak również rozkład populacji akceptowanych kredytów. W niniejszym rozdziale nawiązano do prac z obszaru *analizy wpływu regulacji* (Deighton–Smith i Jacobs 1997).

Wraz z włączeniem metod redukcji stronniczości algorytmicznej, odmienne wnioski kredytowe będą uzyskiwać pozytywną decyzję kredytową, co w kolejnych okresach powodować będzie zmianę zbiorów uczących wykorzystywanych do tworzenia kolejnych edycji modeli oceny zdolności kredytowej. W przypadku oceny zdolności kredytowej, bank

nie wnioskuję operując się na próbach klientów, których wnioski kredytowe zostały odrzucone. Powyższy kontekst powoduje, że modele oceny zdolności kredytowej zawsze są budowane na wybranej, niereprezentatywnej dla populacji ogólnej, uczącej próbie obserwowanych kredytobiorców, którzy wcześniej uzyskali akceptację.

Struktura niniejszego rozdziału jest następująca: podrozdział 5.1 zawiera opis procedury symulacyjnej zastosowanej w modelu dynamicznej oceny konsekwencji wprowadzenia regulacji nakazującej redukcję stronniczości algorytmicznej. W podrozdziale 5.2 oraz podrozdziale 5.3 przedstawiono zastosowanie opracowanej procedury przy założeniu, że modelem bazowym (tj. obecnie wykorzystywanym przez bank) jest odpowiednio model **Unmitigated**, tj. model, w którym zmienna prawnie chroniona jest jawnie włączona jako jedna ze zmiennych egzogenicznych lub model **Fairness through Unawareness**, tj. model, w którym zmienna prawnie chroniona jest wyłączona ze zbioru zmiennych egzogenicznych. W podrozdziale 5.4 przedstawiono autorską tezę–obserwację związaną z kompromisem stronniczości algorytmicznej, tzw. *triadę redukcji stronniczości algorytmicznej*. Rozdział kończy podsumowanie.

5.1 Schemat modelu dynamicznego

W ramach długoterminowej oceny konsekwencji stosowania metod redukcji stronniczości algorytmicznej w modelach oceny zdolności kredytowej zrealizowano 2 scenariusze symulacyjne; scenariusze różnią się wykorzystaniem modelu bazowego, tj. modelu wykorzystywanego przez bank przed wprowadzeniem regulacji przeciwdziałającej stronniczości algorytmicznej: model **Unmitigated** i model **Fairness through Unawareness**. Zakres realizacji tych scenariuszy obejmował dane syntetyczne opisane jako **Scenariusz Syntetyczny**; opis procesu generowania danych scenariusza został przedstawiony w Aneksie. Wykorzystanie danych syntetycznych w badaniu miało na celu zobrazowanie wpływu stosowania metod redukcji stronniczości algorytmicznej w warunkach kontrolowanych, tj. przy znanych rozkładach zmiennych objaśniających ich strukturze korelacyjnej oraz postaci funkcyjnej zmiennej objaśnianej.

Badanie z wykorzystaniem modelu dynamicznego przeprowadzono opierając się na danych syntetycznych: **Scenariusz Syntetyczny**. Opis wykorzystanych danych został zamieszczony w Aneksie. Ogólny schemat realizacji symulacji w ramach podejścia oceny wielookresowego wpływu redukcji stronniczości algorytmicznej (model dynamiczny) został zaprezentowany na rysunku 27. Realizacja badania przy wykorzystaniu modelu dynamicznego zakłada następujące etapy:

1. *Inicjacja modelu bazowego*. W pierwszym etapie symulacji, na wylosowanej populacji 20 000 obserwacji szacowany jest model oceny zdolności kredytowej. Procedura zakłada, że w pierwszym kroku iteracji (ze 100) wszystkie wnioski kredytowe są

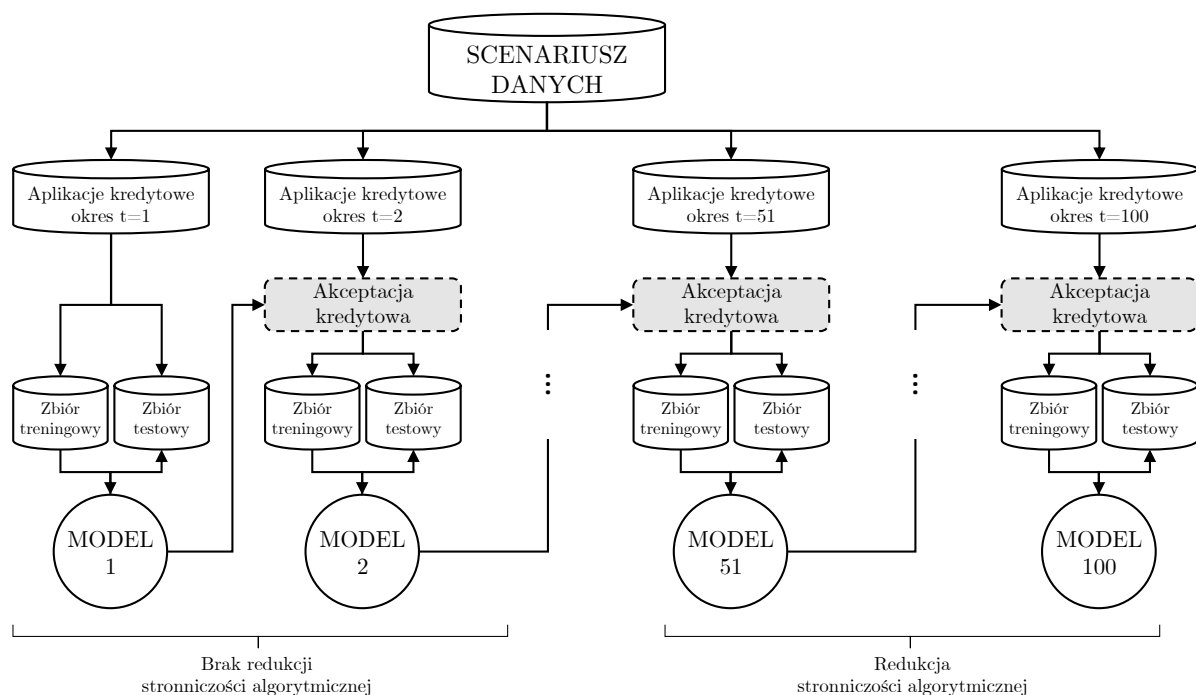
akceptowane i opierając się na danych tych kredytów, w kolejnych iteracjach możliwe będzie estymowanie kolejnych modeli (brak „Akceptacja kredytowa” dla okresu $t = 1$).

2. *Stabilizacja i oszacowanie modelu bazowego.* W drugim etapie symulacji, tj. dla każdej iteracji od 2 do 50 włącznie, dla wylosowanej populacji wniosków kredytowych wykorzystywany jest bazowy model oceny zdolności kredytowej (tj. metoda **Unmitigated**, lub **Fairness through Unawareness**, w zależności od scenariusza) przy zastosowanym progu akceptacji kredytowej na poziomie 75% (tj. prawdopodobieństwo spłaty kredytu co najmniej 75%). Pula wniosków pozytywnie ocenionych staje się pulą wniosków zaakceptowanych, a pozostałe wnioski zostają odrzucane. Opierając się na puli wniosków zaakceptowanych trenowany jest nowy model oceny zdolności kredytowej, który w kolejnej iteracji będzie wykorzystywany do oceny zdolności kredytowej wniosków.
3. *Redukcja stronniczości algorytmicznej.* W trzecim etapie symulacji, tj. dla każdej iteracji od 51 do 100 włącznie, wylosowaną populację wniosków kredytowych ocenia się przy pomocy modelu oceny zdolności kredytowej zawierającego mechanizm redukcji stronniczości algorytmicznej.
4. *Powtórzenie symulacji.* Etapy przedstawione w pkt. 1, 2 oraz 3 są powtarzane (w rozprawie sto razy) w celu uzyskania przedziałów ufności dla uzyskiwanych wyników.
5. *Zebranie i podsumowanie wyników.* Opierając się na zestawie danych wygenerowanych zgodnie z pkt. 4, tworzone jest zestawienie wyników.

W sekcjach 5.2 oraz 5.3 przedstawiono wyniki wykorzystania modelu dynamicznego, dla dwóch różnych metod bazowych: metody **Unmitigated** oraz **Fairness through Unawareness**. W rozprawie przyjmuje się, że metoda bazowa to metoda, którą bank wykorzystuje *przed* wprowadzeniem regulacji nakazującej redukcję stronniczości algorytmicznej związanej z dyskryminacją pośrednią (odmienny wpływ). Warto wskazać, że sekcja 5.3 przedstawia podejście bazowe **Fairness through Unawareness**, które jest obecnie najczęściej wykorzystywane przez banki, jak również jest prawnie obecnie wymagane (EDPS; ECOA-B).

5.2 Metoda bazowa: Unmitigated

Pierwszą symulację przeprowadzono przy wykorzystaniu metody bazowej oznaczanej wcześniej jako metody **Unmitigated** z wykorzystaniem metody regresji logistycznej – LR. Oznacza to, że dla pierwszych 50 iteracji symulacji jako metodę bazową stosowano metodę, która w zbiorze treningowym wykorzystuje wszystkie zmienne, łącznie ze zmienną



Rysunek 27: Schemat realizacji eksperymentów numerycznych oceny długoterminowej

Uwaga: Rysunek przedstawia schemat jednej symulacji modelu dynamicznego; w rozprawie prezentowane wyniki przedstawiają wartości uśrednione dla 100 symulacji.

Źródło: opracowanie własne.

chronioną – płeć. Dobór liczby iteracji symulacji (tj. liczba okresów) oraz liczby powtórzeń symulacji miał na celu stabilizację wyników symulacji (tj. stosunkowo niewielki przedział ufności oszacowanych średnich wartości badanych parametrów).

Jako metodę redukcji stronniczości algorytmicznej wykorzystano metodę **Disparate Impact Remover**; metoda ta, dla scenariusza danych syntetycznych w rozdziale 4, wygenerowała stabilne wyniki w zakresie redukcji stronniczości algorytmicznej: jako jedyna metoda dla wszystkich zrealizowanych symulacji, w oparciu o wszystkie modele ekonometryczne, na zbiorze testowym wygenerowała decyzje kredytowe, które zgodnie z testami statystycznymi χ^2 oraz Fishera nie generowały dyskryminacji pośredniej (tj. odmienny wpływ).

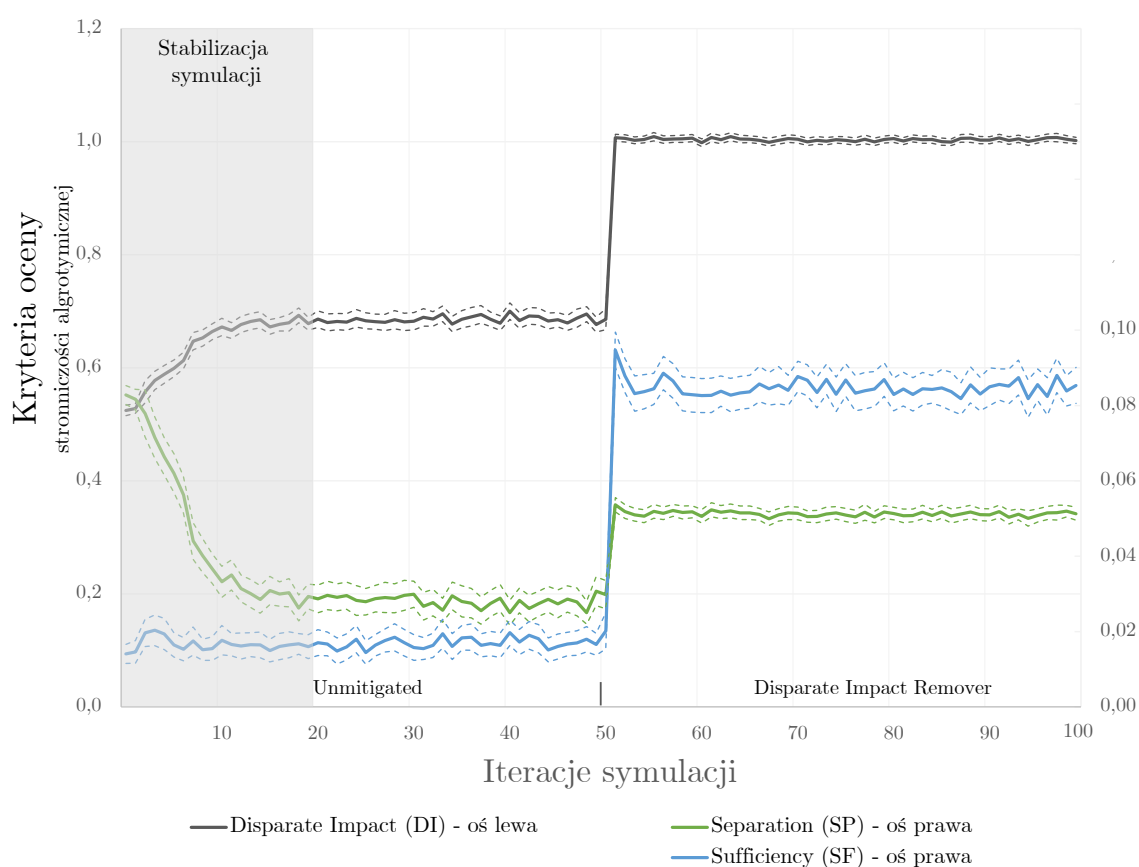
W kolejnych podrozdziałach przedstawiono długoterminowy wpływ stosowania metody redukcji stronniczości algorytmicznej na różne wymiary oceny tego wpływu: miary stronniczości algorytmicznej, rozkład wartości atrybutów populacji akceptowanych kredytów oraz wskaźniki ekonomiczne procesu kredytowego.

5.2.1 Miary stronniczości algorytmicznej

W pierwszej kolejności przedstawiono wpływ zastosowania metod redukcji stronniczości algorytmicznej, na przykładzie metody **Disparate Impact Remover**, na poszczególne miary stronniczości algorytmicznej opisane w podrozdziale 3.3.2, odpowiednio: warunek parytetu demograficznego, mierzony miarą **Disparate Impact** – DI, warunek wyrównanych

szans, mierzony miarą **Separation** – SP oraz warunek parytetu kalibracji, mierzony miarą **Sufficiency** – SF.

Wpływu zastosowania metody **Disparate Impact Remover** na poziom miary stronniczości algorytmicznej został przedstawiony na rysunku 28. Faza stabilizacji wyników wynika z nieliniowości procesu generującego dane, występowania współliniowości zmiennych objaśniających oraz filtrowania próby uczącej w kolejnych iteracjach przez akceptację części najlepszych wniosków kredytowych. Początkową fazę symulacji, tj. 20 pierwszych okresów, stanowi okres stabilizacji wyników. Po stabilizacji miara **Disparate Impact** zbiega do stanu stabilnego na poziomie ok. 69%.



Rysunek 28: Kryteria oceny stronniczości algorytmicznej, **Scenariusz Syntetyczny**, model dynamiczny

Uwaga: Liniami przerywanymi zaznaczono 99% przedział ufności wartości średniej (linia ciągła). Wartość średniej oraz odchylenia standardowego wyznaczono na podstawie 100 symulacji przy wykorzystaniu modelu dynamicznego.

Źródło: opracowanie własne.

Od okresu $t = 51$ włączono metodę redukcji stronniczości algorytmicznej **Disparate Impact Remover**, która to metoda w jednej iteracji spowodowała poprawę miary **Disparate Impact**. Kierunek zmiany wartości **Disparate Impact** jest poprawiający, ponieważ wartość miary **Disparate Impact** z wartości 69% (brak spełnienia warunku parytetu demograficznego) zbiega do wartości 100% (dokładny parytet demograficzny).

Wartość miar **Separation** oraz **Sufficiency** znacząco wzrósł względem poziomów

uzyskiwanych dla stosowania podejścia **Unmitigated**. Zaobserwowany mechanizm wzrostu miar stronniczości algorytmicznej, dotyczących dyskryminacji bezpośredniej ma charakter negatywny – stosowanie metody redukcji stronniczości algorytmicznej względem poprawy warunku parytetu demograficznego (mierzonego miarą **Disparate Impact**) spowodowało pogorszenie wartości pozostałych miar. Zaobserwowany przypadek jest kompromisem między poszczególnymi miarami stronniczości algorytmicznej (podrozdział 3.3.2) – poprawa wartości miary **Disparate Impact** powoduje pogorszenie miar **Separation** oraz **Sufficiency**.

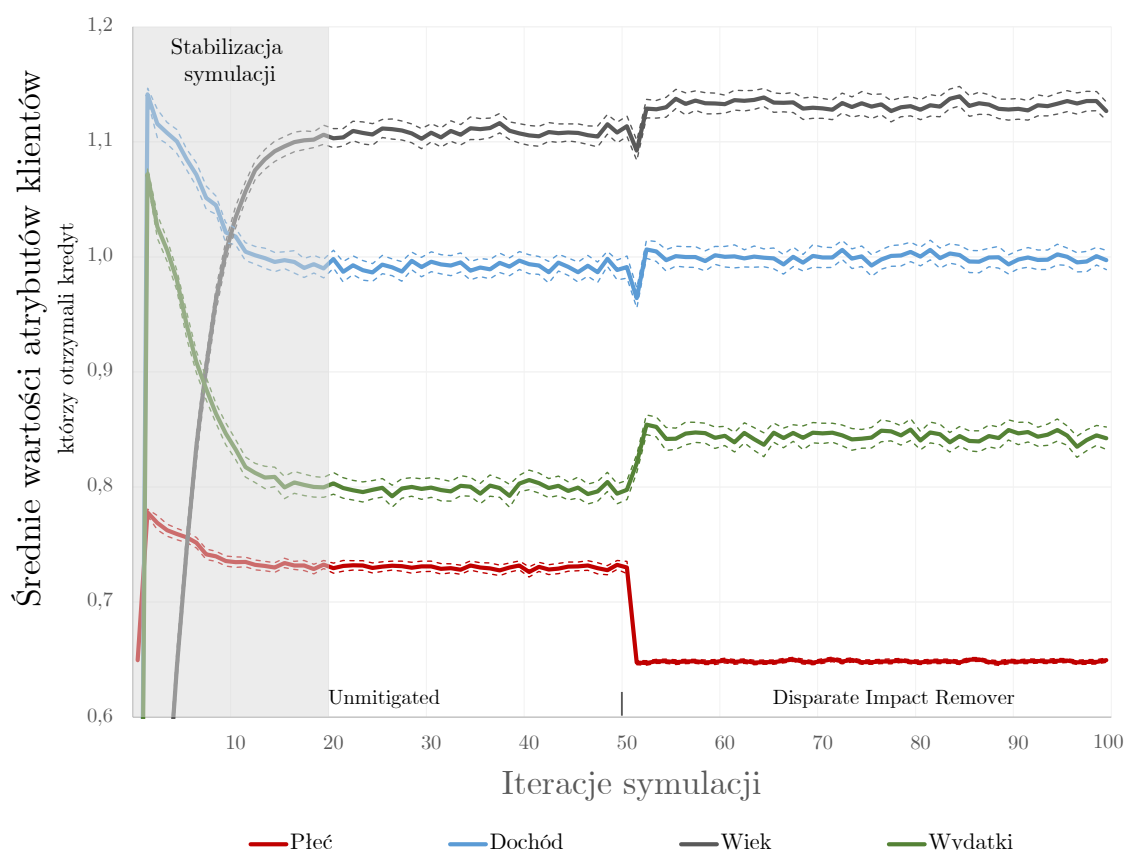
Dodatkowo, w przypadku miar stronniczości algorytmicznej odnoszących się do dyskryminacji bezpośredniej: miary **Separation** oraz **Sufficiency**, warto wskazać na krótkookresowy efekt niestabilności wynikający z wprowadzenia metody redukcji stronniczości algorytmicznej – dla iteracji 51 widać krótkookresowe wybiecie wartości tych miar. Kwestia ta jest istotna z *perspektywy regulatora rynku kredytów konsumenckich*: mechanizm wprowadzania regulacji zakazujących dyskryminacji w modelach oceny zdolności kredytowej, *powinien* uwzględniać fakt możliwej, krótkookresowej destabilizacji oszacowań modelu oceny zdolności kredytowej. Fakt ten mógłby zostać uwzględniony m.in. przez wprowadzenie regulacji stopniowo (tj. krokowe dojście do docelowych, wymaganych przez regulację wartości miar stronniczości algorytmicznej).

5.2.2 Atrybuty populacji akceptowanych kredytów

Kolejnym wymiarem przeprowadzonego badania był rozkład wartości parametrów populacji akceptowanych kredytów. Badanie to miało na celu odpowiedź na pytanie, w jaki sposób zmienia się rozkład klientów, którzy otrzymują akceptację wniosków kredytowych ze względu na zastosowanie metody redukcji stronniczości algorytmicznej. Ze względu na występowanie współliniowości zmiennych egzogenicznych, jak również nieliniowości procesu generującego dane, zastosowanie metody redukującej stronniczość algorytmiczną (zakładającej redukcję nierównego wpływu stosowania modelu na populację zmiennej chronionej) będzie wpływać na zmianę rozkładu parametrów wniosków kredytowych uzyskujących akceptację kredytową.

Średnie wartości atrybutów klientów, którzy otrzymali pozytywną decyzję kredytową przedstawiono na rysunku 29. Dla okresu $t = 1$, zgodnie ze schematem symulacji na rysunku 27, wszystkie wnioski są akceptowane; stąd średnie wartości parametrów wniosków akceptowanych są równe średnim wartościom tych parametrów dla całej populacji (Aneks: Specyfikacja danych).

Ponownie, na rysunku 29 początkowy okres – ok. 20 iteracji – wskazuje na zasadność wykorzystania okna stabilizacji wyników. Zastosowanie metody redukcji stronniczości algorytmicznej **Disparate Impact Remover** powoduje również zmianę innych parametrów akceptowanych parametrów: dla przykładu po zastosowaniu wskazanej metody akceptacji



Rysunek 29: Analiza dynamiczna wpływu na rozkład parametrów akceptowanych kredytów, metoda bazowa: **Unmitigated**

Uwaga: Liniami przerywanymi zaznaczono 99% przedział ufności wartości średniej (linia ciągła). Wartość średniej oraz odchylenia standardowego wyznaczono na podstawie 100 symulacji przy wykorzystaniu modelu dynamicznego.

Źródło: opracowanie własne.

zaczynają podlegać wnioski osób starszych (zmienna *Wiek* zwiększa swoją wartość), osób o wyższym poziomie dochodu (zmienna *Dochód*), o wyższych wydatkach (zmienna *Wydatki*).

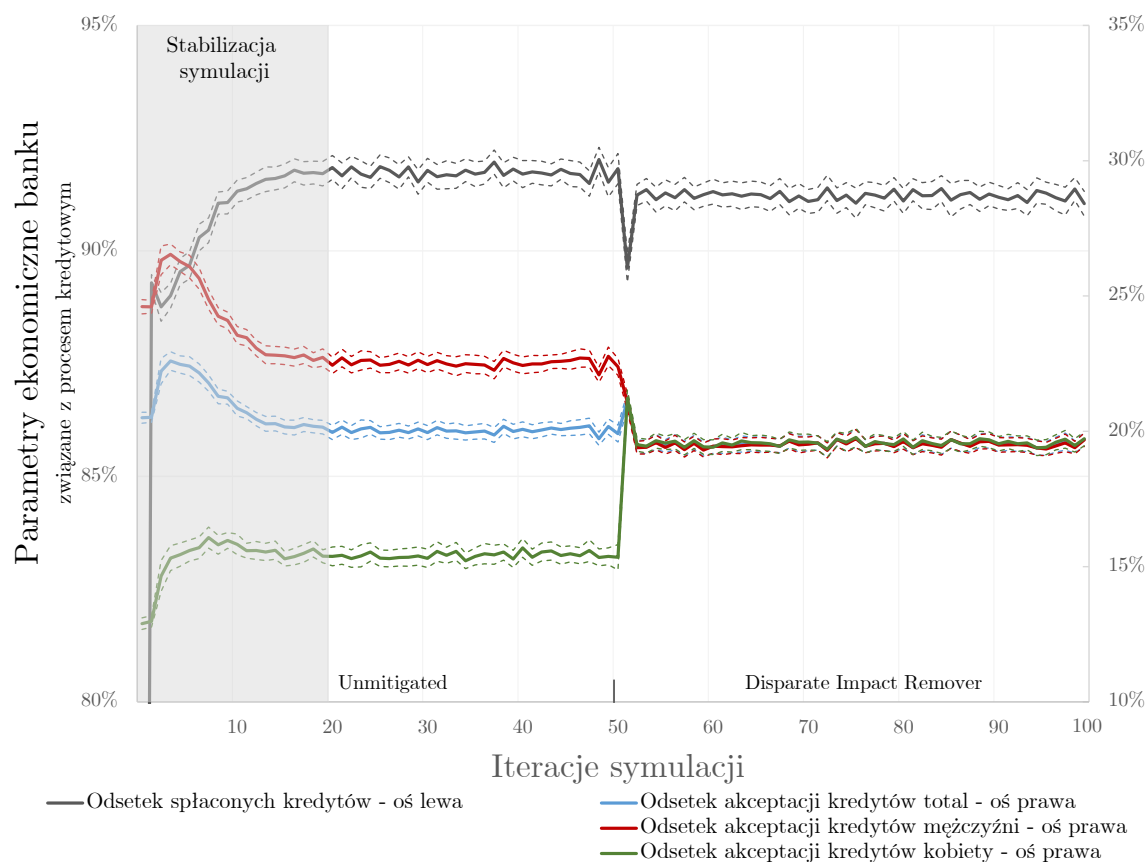
Dodatkowo warto wskazać na krótkookresowy efekt braku stabilności średnich wartości parametrów: średnie wartości niektórych zmiennych (np. zmienne *Wiek*, *Dochód*) dla iteracji 51 maleją, przed późniejszym, strukturalnym wzrostem wartości.

5.2.3 Parametry ekonomiczne procesu kredytowego

Ostatnim wymiarem badania jest ocena zmian parametrów ekonomicznych procesu kredytowego. W ramach ewaluacji długoterminowego wpływu zastosowania metod redukcji stronniczości algorytmicznej ocenie poddano: jakość portfela kredytowego, jakość prognostyczną modelu oceny zdolności kredytowej, wskaźnik zatwierdzeń oraz poziom zysku generowanego z zastosowania modelu oceny zdolności kredytowej.

Wpływ zastosowania metod redukcji stronniczości algorytmicznej na parametry ekonomiczne procesu kredytowego: jakość portfela kredytowego (mierzoną poziomem

splacania kredytów) oraz wolumenem udzielanych kredytów (mierzonym wskaźnikiem zatwierdzeń kredytów) przedstawiono na rysunku 30.



Rysunek 30: Analiza długoterminowa wpływu na jakość i wolumen portfela kredytowego, metoda bazowa: **Unmitigated**

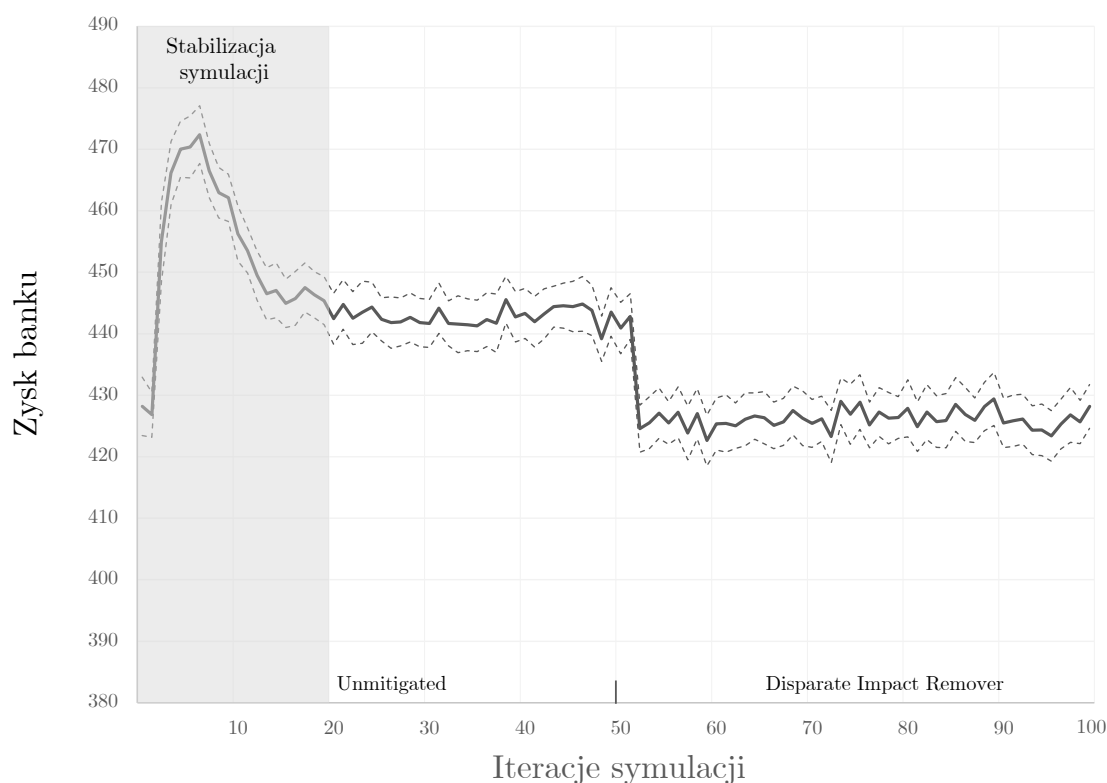
Uwaga: Liniami przerywanymi zaznaczono 99% przedział ufności wartości średniej (linia ciągła). Wartość średniej oraz odchylenia standardowego wyznaczono na podstawie 100 symulacji przy wykorzystaniu modelu dynamicznego.

Źródło: opracowanie własne.

Wskaźnik jakości portfela kredytowego wraz z wprowadzeniem metody redukcji stronniczości algorytmicznej **Disparate Impact Remover** ulega czasowej destabilizacji – dla iteracji 51 pojawia się istotne obniżenie jakości portfela, która następnie wraca na niższy (niż dla metody bazowej) poziom. Wraz z zastosowaniem metody redukcji stronniczości algorytmicznej, długookresowo poziom jakości portfela ulega pogorszeniu.

Zgodnie ze spodziewanym wpływem, zastosowanie podejścia do redukcji stronniczości algorytmicznej opartego o miarę **Disparate Impact** spowoduje wyrównanie wskaźnika akceptacji wniosków kredytowych w obu podpopulacjach (mężczyźni oraz kobiety) do jednego poziomu.

Ostatnim wymiarem łączącym wcześniejsze – jakość portfela i wskaźnik akceptacji – jest poziom zysku generowanego przez zastosowanie modelu oceny zdolności kredytowej. Miara ta uwzględnia równocześnie zmianę liczby udzielanych akceptacji wniosków kredytowych, jak również ich jakości.



Rysunek 31: Analiza długoterminowa wpływu na zysk, metoda bazowa: **Unmitigated**

Uwaga: Liniami przerywanymi zaznaczono 99% przedział ufności wartości średniej (linia ciągła).
Wartość średniej oraz odchylenia standardowego wyznaczono na podstawie 100 symulacji przy wykorzystaniu modelu dynamicznego.

Źródło: opracowanie własne.

Łączny poziom zysku z realizowanej akcji kredytowej po zastosowaniu metody **Disparate Impact Remover** powoduje spadek zysku. Oznacza to, że bank włączając mechanizm ograniczania stronniczości algorytmicznej powoduje spadek wartości ekonomicznej generowanej z takiego procesu. Wskazana zmiana poziomu zysku oznacza koszt, jaki bank (szerzej instytucja wykorzystująca model oceny zdolności kredytowej) musi ponieść.

5.3 Metoda bazowa: Fairness through Unawareness

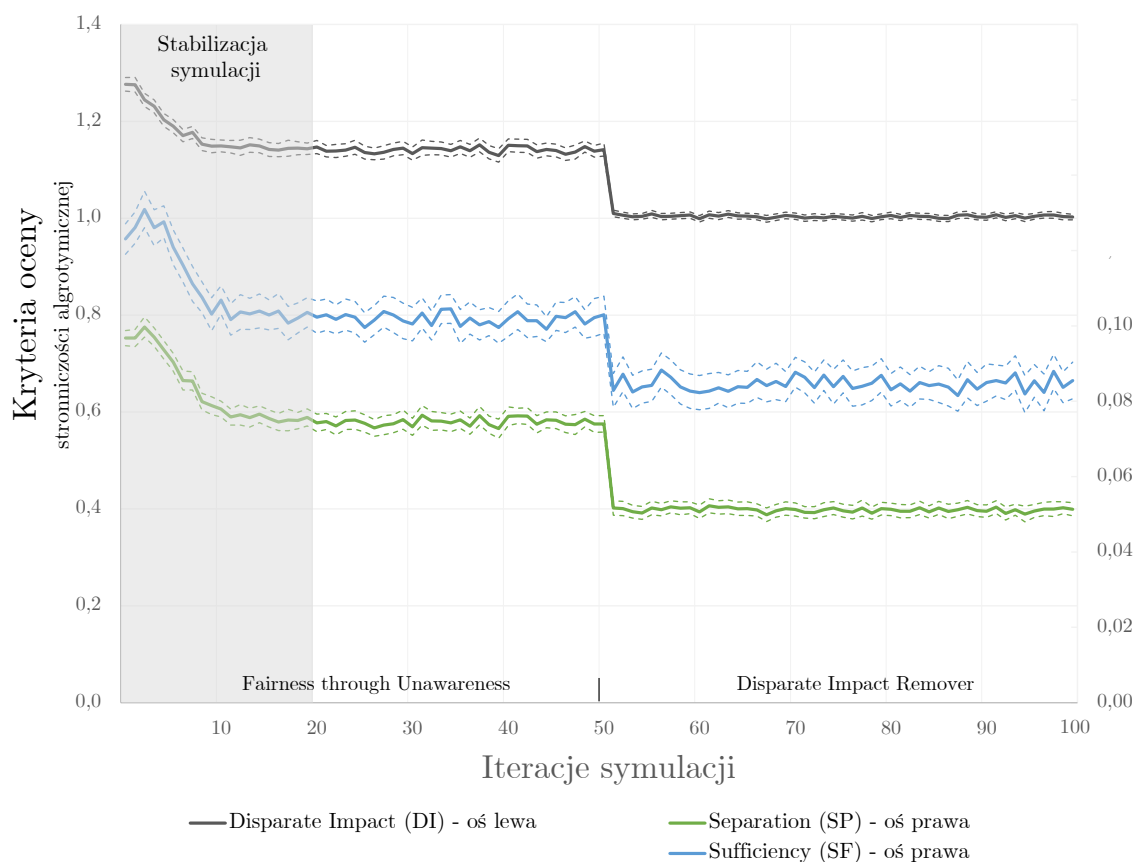
Drugą z symulacji przy wykorzystaniu modelu dynamicznego przeprowadzono przy wykorzystaniu metody bazowej oznaczanej wcześniej jako metody **Fairness through Unawareness** z wykorzystaniem metody regresji logistycznej – LR. Dla pierwszych 50 iteracji symulacji dla regresji logistycznej ze zbioru treningowego wykluczono zmienną chronioną płeć.

Opierając się na wynikach badania przedstawionego w niniejszym podrozdziale zauważono, że występuje zjawisko krokowej stabilizacji symulowanych parametrów: modelu oceny zdolności kredytowej, miar stronniczości algorytmicznej oraz populacji

otrzymującej kredyty. W zakresie wyżej wskazanych wyników, jedną z rekomendacji dla regulatorów w zakresie projektowania regulacji obejmujących modele oceny zdolności kredytowej jest rozważenie stosowania wydłużonych okresów dostosowawczych, umożliwiających np. stopniową realizację zakładanych progów miar stronniczości algorytmicznej.

5.3.1 Miary stronniczości algorytmicznej

Wpływ zastosowania metody **Disparate Impact Remover** na poziom miar stronniczości algorytmicznej, przy początkowo zastosowanej metodzie bazowej **Fairness through Unawareness** przedstawiono na rysunku 32.



Rysunek 32: Analiza długoterminowa wpływu na miarę **Disparate Impact**, metoda bazowej: **Fairness through Unawareness**

Uwaga: Liniami przerywanymi zaznaczono 99% przedział ufności wartości średniej (linia ciągła). Wartość średniej oraz odchylenia standardowego wyznaczono na podstawie 100 symulacji przy wykorzystaniu modelu dynamicznego.

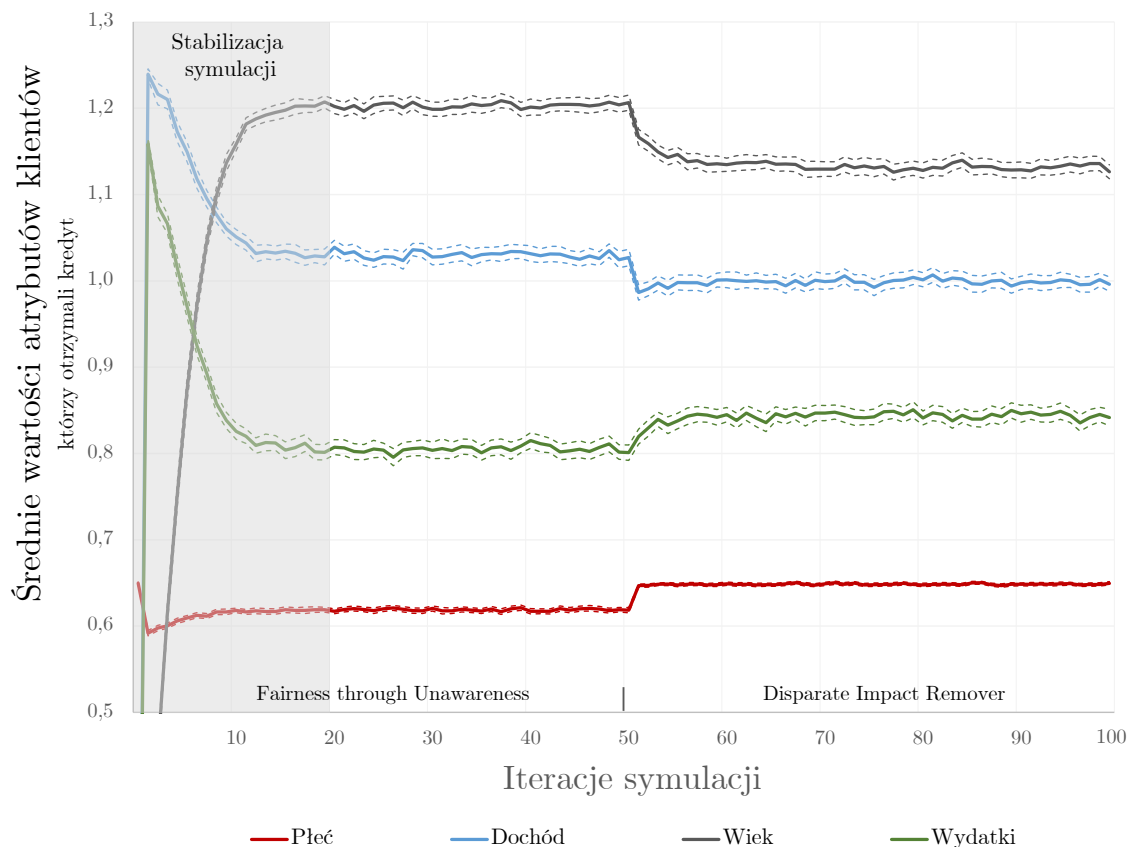
Źródło: opracowanie własne.

Jak wskazano w rozdziale 4, regresja logistyczna wykorzystująca podejście **Fairness through Unawareness**, dla danych **Scenariusza Syntetycznego** wykazuje miarę **Disparate Impact** wyższą niż 1,00. Oznacza to, że grupa nieuprzywilejowana (np. kobiety) otrzymują częściej kredyt niż grupa uprzywilejowana. Zastosowanie metody

Disparate Impact Remover sprowadza miarę Disparate Impact Remover do wartości 1,00. Zastosowanie metody redukcji stronniczości algorytmicznej Disparate Impact Remover powoduje również spadek wartości miary Separation – niższe wartości tej miary wskazują na mniejsze różnicowanie podgrup populacji ze względu na wielkość błędów prognozy FN i FP. Dodatkowo warto zauważyć, że osiągnięcie zgodności na wartości Disparate Impact powodować będzie równoczesną poprawę na wartościach Separation oraz Sufficiency. Oznacza to, że model regresji logistycznej w podejściu Fairness through Unawareness jest podejściem Pareto nieoptymalnym: można uzyskać poprawę wszystkich miar sprawiedliwości.

5.3.2 Atrybuty populacji akceptowanych kredytów

Drugim wymiarem przeprowadzonego badania była analiza rozkładu wartości parametrów populacji akceptowanych kredytów. Wartości średnie poszczególnych atrybutów akceptowanych wniosków kredytowych przedstawiono na rysunku 33.



Rysunek 33: Analiza dynamiczna wpływu na rozkład parametrów akceptowanych kredytów, metoda bazowej: **Fairness through Unawareness**

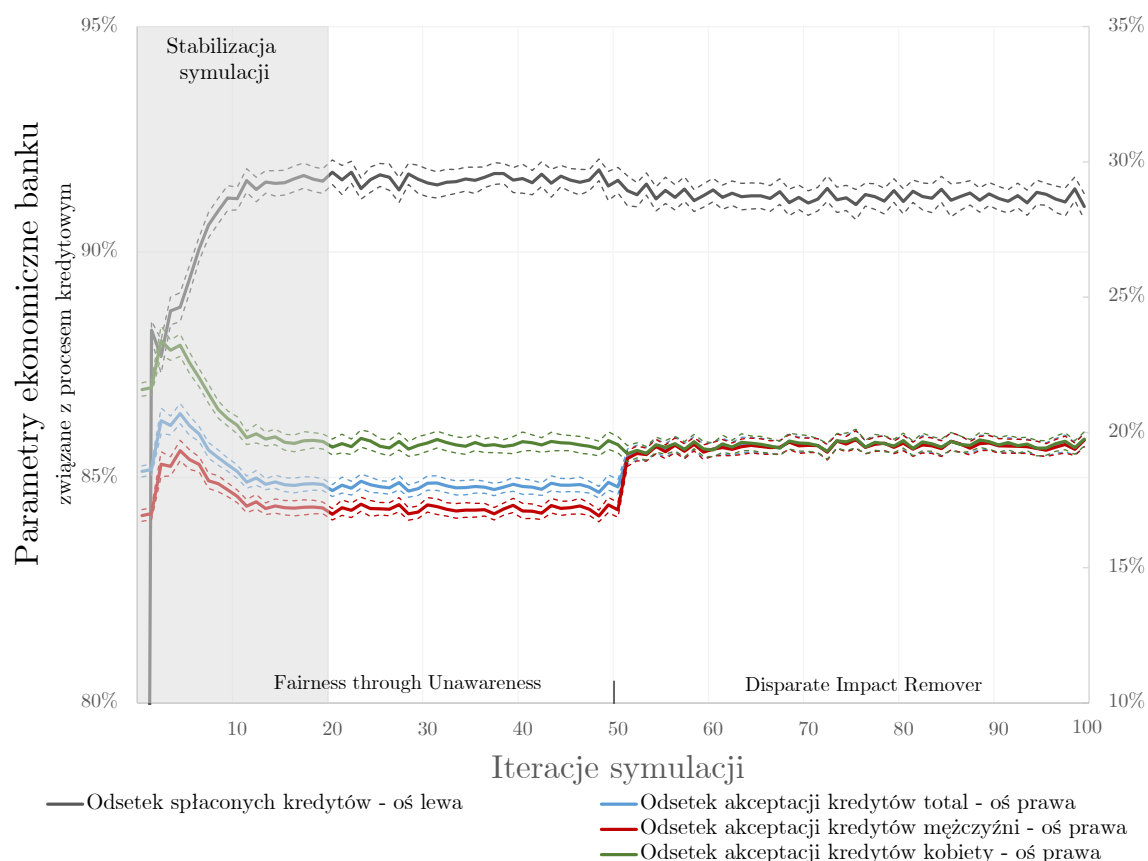
Uwaga: Liniami przerywanymi zaznaczono 99% przedział ufności wartości średniej (linia ciągła). Wartość średniej oraz odchylenia standardowego wyznaczono na podstawie 100 symulacji przy wykorzystaniu modelu dynamicznego.

Źródło: opracowanie własne.

Wykorzystanie podejścia **Disparate Impact Remover** w przypadku średnich wartości atrybutów klientów, których wnioski kredytowe zostały zaakceptowane, powoduje krokowe (tj. nie natychmiastowe) zmiany do poziomu ustalonego – prawidłowość tę widać np. przy zmiennej **Wiek** lub **Wydatki**. Zmiany wartości atrybutu **Wiek** następują stopniowo – ok. 5–8 iteracji między włączeniem mechanizmu redukcji stronniczości algorytmicznej a osiągnięciem poziomu ustalonego.

5.3.3 Parametry ekonomiczne procesu kredytowego

Ostatnim wymiarem badania jest ocena zmian wskaźników ekonomicznych procesu kredytowego względem uwzględnienia redukcji stronniczości algorytmicznej. Wpływ zastosowania podejścia **Disparate Impact Remover** na parametry ekonomiczne procesu kredytowego: spłacalność oraz wskaźnik zatwierdzeń kredytów, przedstawiono na rysunku 34.



Rysunek 34: Analiza długoterminowa wpływu na jakość portfela kredytowego, metoda bazowej: **Fairness through Unawareness**

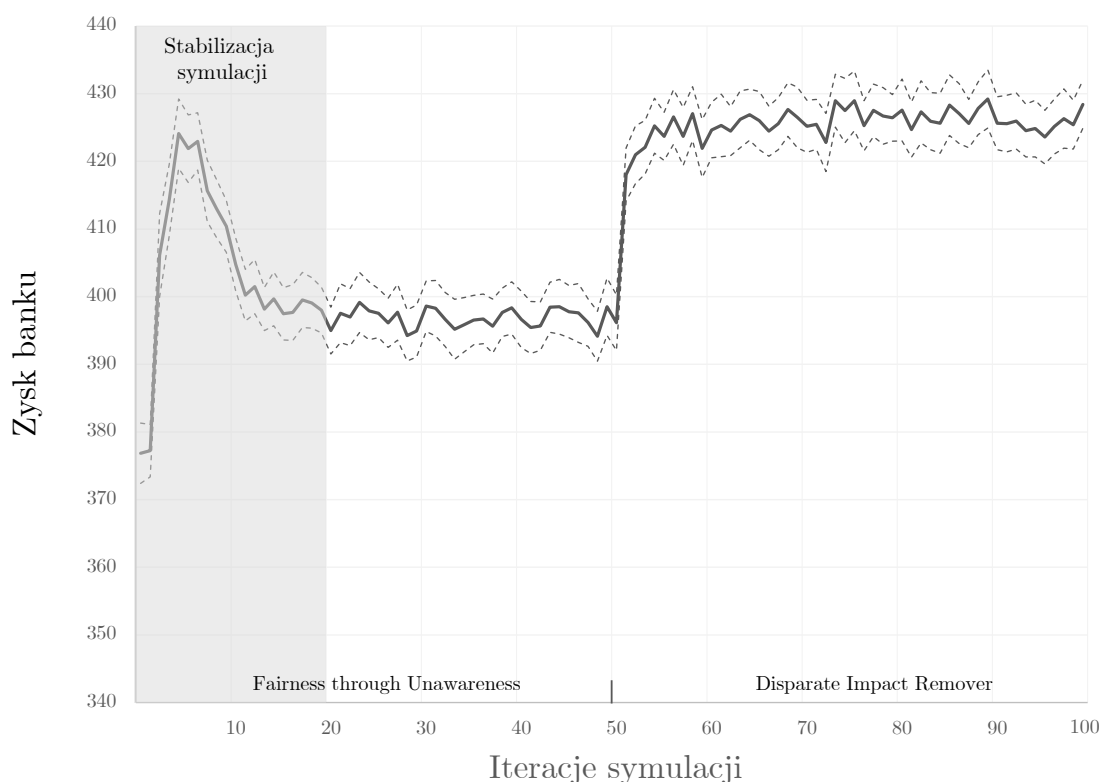
Uwaga: Liniami przerywanymi zaznaczono 99% przedział ufności wartości średniej (linia ciągła). Wartość średniej oraz odchylenia standardowego wyznaczono na podstawie 100 symulacji przy wykorzystaniu modelu dynamicznego.

Źródło: opracowanie własne.

Wskaźnik jakości portfela kredytowego wraz z wprowadzeniem metody redukcji

stronniczości algorytmicznej **Disparate Impact Remover** ulega obniżeniu. Oznacza to, że wdrożenie metody redukcji stronniczości algorytmicznej skutkuje obniżeniem jakości portfela kredytowego. Wdrożenie redukcji stronniczości algorytmicznej, w rozważanym scenariuszu symulacyjnym, powoduje wyrównanie wskaźnika zatwierdzeń kredytów.

Ostatnim wymiarem badania jest poziom zysku generowanego przez zastosowanie modelu oceny zdolności kredytowej. Miara ta uwzględnia zmianę liczby udzielanych akceptacji wniosków kredytowych, jak również ich jakości. Wpływ zastosowania metod redukcji stronniczości algorytmicznej na poziom zysku z realizowanego procesu kredytowego przedstawiono na rysunku 35.



Rysunek 35: Analiza długoterminowa wpływu na zysk, metoda bazowej: **Fairness through Unawareness**

Uwaga: Liniami przerywanymi zaznaczono 99% przedział ufności wartości średniej (linia ciągła). Wartość średniej oraz odchylenia standardowego wyznaczono na podstawie 100 symulacji przy wykorzystaniu modelu dynamicznego.

Źródło: opracowanie własne.

Warto zauważyć, że włączenie metod redukcji stronniczości algorytmicznej, jeżeli modelem bazowym jest model regresji logistycznej przy podejściu z **Fairness through Unawareness** powoduje wzrost wartości zysku generowanego z procesu oceny zdolności kredytowej. Wy tłumaczeniem tego stanu rzeczy jest fakt, że podejście bazowe nie uwzględnia informacji zawartej w zmiennej płeć – **Fainess through Unawarness** nie zawiera w zbiorze treningowym zmiennych chronionych. W przypadku metody **Disparate Impact Remover**, zmienna ta nie jest bezpośrednio w zbiorze uczącym wykorzystywana, niemniej jednak

zmienna ta wykorzystywana jest do transformacji pozostałych zmiennych objaśniających. W przypadku predyktywnych zmiennych chronionych, metoda **Disparate Impact Remover** umożliwia podniesienie zarówno zysku z realizowanej akcji kredytowej, jak również obniżenie wartości wszystkich miar stronniczości algorytmicznej (tj. w przypadku miary **Disparate Impact** sprowadzenie jej wartości do wartości 1,00).

5.4 Triada redukcji stronniczości algorytmicznej

We wcześniejszych podrozdziałach przedstawiono wpływ wprowadzenia metod redukcji stronniczości algorytmicznej, na przykładzie metody **Disparate Impact Remover**, na poszczególne wymiary prowadzonego badania. Na podstawie wyników dokonano następującej obserwacji: w modelach oceny zdolności kredytowej występują co najmniej trzy współzależne wymiary oceny takiego modelu. Ważnym tutaj jest wskazanie, że przez model oceny zdolności kredytowej rozumie się całokształt opracowanej procedury, uwzględniając przy tym m.in. sposób trenowania takiego modelu, uwzględnienie (lub jego brak) sposobu redukcji stronniczości algorytmicznej, próg akceptacji wniosków kredytowych lub transformacje na zmiennych objaśniających.

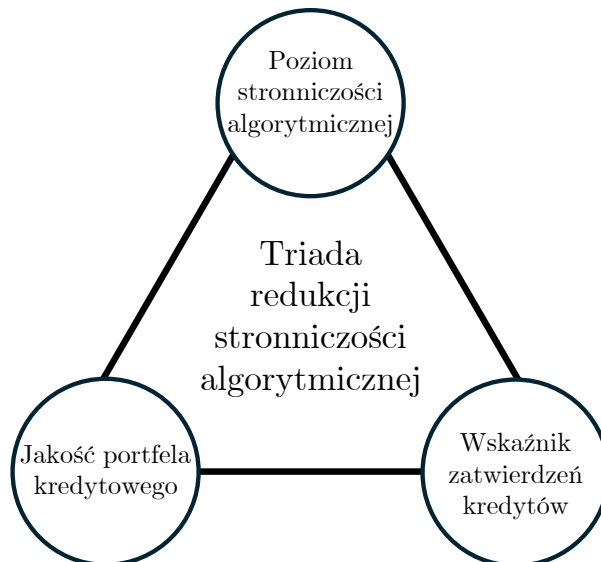
Pierwszym z wymiarów takiej oceny jest *jakość portfela kredytowego*. Opracowując model oceny zdolności kredytowej, który jest następnie wykorzystywany do oceny wniosków kredytowych, bank uzyskuje *określony* poziom szkodowości portfela kredytowego (mierzony np. kohortowym poziomem braku spłacalności kredytów). Parametrem, który jest wykorzystywany do zarządzania szkodowością takich portfeli jest próg akceptacji wniosków kredytowych. Jednak w zależności od *jakości* modelu oceny zdolności kredytowej dla danego progu odcięcia bank będzie mógł uzyskiwać różne poziomy szkodowości. Przykładowo, dla losowego modelu oceny zdolności kredytowej, zmiana progu akceptacji wniosków kredytowych będzie liniowo zwiększać/zmniejszać liczbę kredytów złych; w przypadku idealnego modelu oceny zdolności kredytowej, liczba złych kredytów będzie zerowa do momentu, w którym wszystkie dobre kredyty uzyskają akceptację kredytową – dopiero wtedy złe kredyty będą uzyskiwać akceptację.

Drugim z wymiarów jest *wskaźnik akceptacji wniosków kredytowych*, który wskazuje, jaki procent złożonych wniosków kredytowych uzyskuje pozytywną decyzję kredytową (np. jeżeli dla 1 000 złożonych wniosków kredytowych, tylko 100 uzyskało akceptację, to poziom wskaźnika akceptacji będzie wynosił $\frac{100}{1000} = 10\%$). Wskaźnik akceptacji wniosków jest funkcją progu akceptacji: czym bardziej restrykcyjna jest wartość progu, tym mniej wniosków kredytowych uzyskuje akceptację. Przy skrajnych podejściach, bank może prowadzić tak konserwatywną politykę kredytową, że żaden wniosek kredytowy nie uzyska akceptacji kredytowej. Przeciwna sytuacja występuje przy liberalnej polityce kredytowej, akceptującej wszystkie wnioski kredytowe.

Powyższe dwa wskaźniki modeli oceny zdolności kredytowej są ściśle związane z ekonomicznym wpływem stosowania modelu oceny zdolności kredytowej. Bank będzie osiągać wysoki poziom zysku z prowadzonej akcji kredytowej, kiedy będzie utrzymywać wysoki poziom akceptacji wniosków kredytowych, z których kredyty docelowo będą terminowo spłacane. Oba wskaźniki są ze sobą sprzężone: jeżeli bank udziela niewielu kredytów (tj. wskaźnik akceptacji jest niski), to kredyty uzyskiwać będą jednostki, o wysokim prawdopodobieństwie spłaty. Jeżeli bank zacznie zwiększać swoją tolerancję na ryzyko, zwiększając tym samym poziom wskaźnika akceptacji kredytów, to coraz więcej *złych* kredytów będzie uzyskiwać finansowanie.

Trzecim wymiarem, odnoszącym się do społecznej ewaluacji działania modelu jest *poziom stronniczości algorytmicznej*, jaką dany model posiada. Stanowi on warunek ograniczający na dwa wcześniejsze parametry: dla zadanego poziomu oczekiwanej sprawiedliwości istnieje zbiór możliwych wartości odpowiednio jakości portfela kredytowego oraz wskaźnika akceptacji – oba będące funkcją progu pozytywnej decyzji kredytowej. Dla zadanego poziomu ograniczenia stronniczości algorytmicznej (mierzony *Disparate Impact*) istnieje krzywa możliwych poziomów akceptacji wniosków kredytowych i wynikowej jakości portfela kredytowego.

Podsumowanie składowych triady redukcji stronniczości algorytmicznej przedstawiono na rysunku 36.



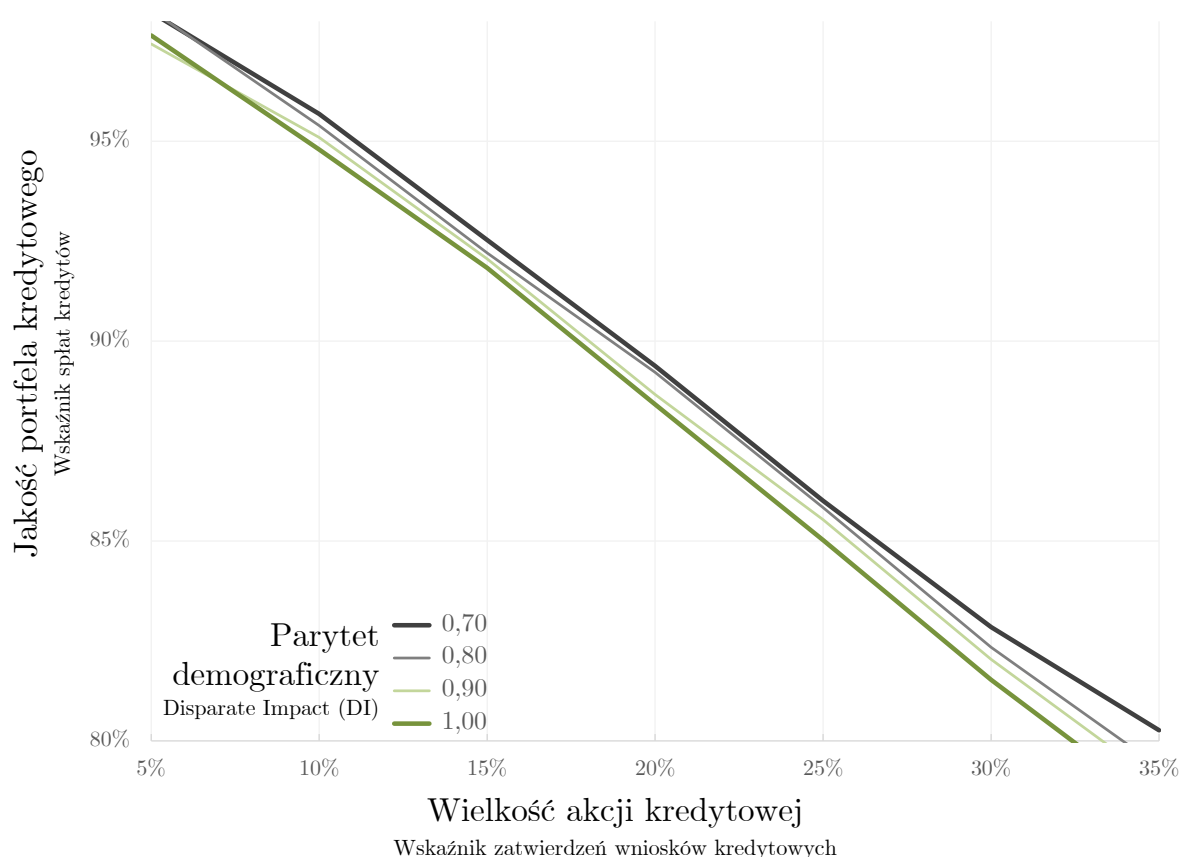
Rysunek 36: Triada redukcji stronniczości algorytmicznej

Uwaga: Bank, przed okresem obowiązywania regulacji związanych z koniecznością redukcji stronniczości algorytmicznej sterował swoją zyskownością przez próg decyzji kredytowej (problem doboru progu decyzji kredytowej); zmiana tego progu wpływała na zmianę wskaźnika zatwierdzeń kredytów oraz jakość portfela kredytowego. Wprowadzenie obowiązku redukcji stronniczości algorytmicznej – w badaniu w rozprawie wykorzystano definicję dyskryminacji pośredniej (odmienny wpływ) – powoduje, że we wskazanym zagadnieniu doboru progu decyzji kredytowej pojawia się dodatkowy wymiar: poziom stronniczości algorytmicznej.

Źródło: opracowanie własne.

Zależność wskaźnika zatwierdzeń kredytów od jakości portfela kredytowego, dla zadanych poziomów miary **Disparate Impact**, przedstawiono na rysunku 37. Dla wartości **Disparate Impact** na poziomie 0,70, średni poziom jakości portfela kredytowego i wskaźnika akceptacji/selekcji są na najwyższym poziomie. Stanowią one zbiorczo rozwiązanie optymalnym w sensie Pareto, tj. „alokacja zasobów generowanych przez rynek jest efektywna (optymalna w sensie Pareto), jeśli nie istnieje alternatywna, dopuszczalna alokacja zasobów, która może poprawić sytuację jakiejś jednostki bez pogorszenia sytuacji kogoś innego” (Stiglitz 1981).

Wprowadzenie regulacji, która nakładałaby na bank obowiązek wyrównania wskaźnika zatwierdzeń kredytów dla kobiet i mężczyzn (krzywa zielona o wartości **Disparate Impact** na poziomie 1,00) spowodowało przesunięcie całej krzywej zależności między wielkością akcji kredytowej a jakością portfela kredytowego w kierunku początku układu współrzędnych. Oznacza to, że realizacja założeń regulacyjnych, wymagających redukcji problemu stronniczości algorytmicznej zgodnie z definicją dyskryminacji pośredniej, spowodowałaby spadek jakości portfela kredytowego (wskaźnik spłat kredytów) dla zadanej wielkości akcji kredytowej.



Rysunek 37: Izolinie wartości **Disparate Impact** w zależności od wskaźnika zatwierdzeń kredytów oraz jakości portfela kredytowego

Źródło: opracowanie własne.

Celem niniejszego rozdziału było opracowanie procedury wspierającej ocenę skutków zastosowania określonej definicji dyskryminacji (miara stronniczości algorytmicznej) w regulacji oraz jej zastosowanie na przykładowych danych. W ramach rozdziału 5 opracowano procedurę wspierającą regulatorów rynku kredytów konsumenckich w badaniu pośredniego wpływu redukcji stronniczości algorytmicznej, tzw. *perspektywa regulatora rynku*. Decyzje, które regulatorzy podejmują są związane z kształtem tworzonej regulacji – m.in. definicja dyskryminacji (pośrednia lub bezpośrednia), miara stronniczości algorytmicznej lub warunki wprowadzanie regulacji (np. okres wprowadzania obowiązku regulacyjnego). Utworzona procedura umożliwi regulatorom weryfikację poszczególnych podejść redukcji stronniczości algorytmicznej, w zakresie oceny skutków zastosowania tej procedury na konsumentów (np. zmiana rozkładu parametrów populacji otrzymującej pozytywne decyzje kredytowe).

W ramach rozdziału przedstawiono zastosowanie wskazanej procedury na podstawie przykładowych danych. Jako metodę badawczą w niniejszym rozdziale wykorzystano również symulacje, przy czym konstruowanym modelem jest model wielookresowy, tj. model dynamiczny, umożliwiający obserwację skutków zmiany stosowanego przez bank podejścia, np. podejście obecnie stosowane przez większość banków komercyjnych zakładające odrzucenie wartości zmiennej chronionej z modelu oceny zdolności kredytowej – podejście **Fairness through Unawareness**.

Rozdział ten przedstawia wyniki badania przeprowadzonego na podstawie danych symulacyjnych. Wyniki badania przedstawionego w rozdziale są związane z trzema hipotezami badawczymi rozprawy.

Hipoteza 4: O metodyce oceny długookresowych konsekwencji regulacji równościowych. Procedura umożliwiająca ocenę długookresowych skutków decyzji regulacyjnych została przedstawiona i zastosowana na danych symulacyjnych w ramach rozprawy. Jednym z nieoczywistych wniosków wykorzystania modelu dynamicznego jest ten dotyczący krótkookresowych niestabilności średnich wartości parametrów związanych z procesem kredytowym. W rozprawie przedstawiono przykład, w którym krótkookresowe skutki wprowadzenia takiej regulacji były co do kierunku przeciwne względem kierunków zmian długookresowych; na rysunku 29 przedstawiono dynamikę zmian średnich wartości parametrów populacji otrzymujących kredyty: dla scenariusza danych syntetycznych, średni poziom dochodu (lub wieku) osób uzyskujących kredyt w efekcie redukcji stronniczości algorytmicznej krótkookresowo spadł; w późniejszych okresach wartość średniej ustabilizowała się na poziomach wyższych od wartości wyjściowej.

Hipoteza 5: O okresie dostosowawczym występującym po wprowadzeniu regulacji. W rozprawie opracowano dynamiczny model umożliwiający odpowiedzi na pytanie,

jaka będzie dynamika zmian kluczowych parametrów związanych z procesem kredytowym – zarówno z perspektywy banku, jak i konsumentów – w związku z koniecznością wprowadzania redukcji stronniczości algorytmicznej w ocenie zdolności kredytowej. Na podstawie uzyskanych wyników badania hipotezę zweryfikowano pozytywnie: po rozpoczęciu stosowania metody redukcji stronniczości algorytmicznej, może wystąpić okresowa niestabilność wyników lub krokowa stabilizacja wyników; oba zjawiska wskazują na zasadność wydłużonego okresu stabilizacji przedstawionych wartości. W zakresie wyżej wskazanych wyników, jedną z rekomendacji dla regulatorów w zakresie projektowania regulacji obejmujących modele oceny zdolności kredytowej jest rozważanie stosowania wydłużonych okresów dostosowawczych, umożliwiających np. stopniową realizację zakładanych progów miar stronniczości algorytmicznej.

Hipoteza 6: O koszcie ekonomicznym regulacji równościowych. W rozprawie wskazano, że w modelach oceny zdolności kredytowej występują co najmniej trzy współzależne wymiary oceny takiego modelu, tj.: jakość portfela kredytowego, wskaźnik zatwierdzeń wniosków kredytowych oraz poziom stronniczości algorytmicznej jaką dany model posiada. Dwa pierwsze wskaźniki, tj.: jakość portfela kredytowego oraz wskaźnik zatwierdzeń wniosków kredytowych są ściśle związane z ekonomicznym wpływem stosowania modelu oceny zdolności kredytowej. Poziom stronniczości algorytmicznej stanowi niejako warunek ograniczający na dwa wcześniejsze parametry: dla zadanego poziomu dyskryminacji pośredniej istnieje zbiór możliwych wartości odpowiednio jakości portfela kredytowego oraz wskaźnika akceptacji – oba będące funkcją progu pozytywnej decyzji kredytowej. Dla zadanego poziomu ograniczenia stronniczości algorytmicznej istnieje krzywa, Pareto efektywna, możliwych poziomów akceptacji wniosków kredytowych i wynikowej jakości portfela kredytowego.

W literaturze przedmiotu w zakresie problemu stronniczości algorytmicznej modelu oceny zdolności kredytowej wskazuje się na ocenę krótkookresowych skutków wykorzystania metod redukcji stronniczości algorytmicznej. Dotychczas przedstawiono w niej podejście, zbieżne z omówionym w rozprawie w rozdziale 4, w którym przeprowadzono analizę skutków bezpośrednich zastosowania metod redukcji stronniczości algorytmicznej. W niniejszym rozdziale przedstawiono wyniki badania przy rozważanym, wielookresowym horyzoncie oceny konsekwencji użycia metod redukcji stronniczości algorytmicznej.

W zakresie kontrybucji do obszaru stronniczości algorytmicznej można wskazać na wyniki badania związane z okresową niestabilnością lub krokową stabilizacją wyników działania modelu oceny zdolności kredytowej. Efekty te zostały zauważone i przedstawione w ramach niniejszej rozprawy, a skutkować mogą rekomendacjami dla regulatorów rynku kredytów konsumenckich w zakresie kształtu projektowanych regulacji obejmujących zakazanie dyskryminacji w modelach oceny zdolności kredytowej.

Efektom przeprowadzonego badania, którego wyniki przedstawiono w rozdziale 5, było zauważenie relacji, którą w rozprawie nazwano *triadą redukcji stronniczości algorytmicznej*. Decyzje kredytowe w zakresie wskaźnika zatwierdzeń kredytów, a pośrednio również jakości kredytowej (tj. pośrednio, ponieważ mechanizm ten odbywa się przez dobór progu zatwierdzeń kredytów), w reżimie regulacji zakazującej dyskryminacji, będą musiały uwzględniać również poziom stronniczości algorytmicznej. W praktyce oznaczać to będzie, że banki, aby spełniać wymogi związane z uzyskaniem wartości określonej miary stronniczości algorytmicznej (w rozprawie zbadano miarę *Disparate Impact* związaną z definicją dyskryminacji pośredniej), będą musiały poświęcić jakość kredytową lub wskaźnik zatwierdzeń kredytów (lub daną kombinację tych dwóch wymiarów) – poziom dotkliwości wskazanego skutku zależny będzie od podejścia wykorzystywanego przez bank w dotychczasowej praktyce.

Podsumowanie

Celem rozprawy było opracowanie metod wspomagania decyzji regulatorów w zakresie tworzenia oraz banków w zakresie wdrażania przepisów równościowych w obszarze modeli oceny zdolności kredytowej konsumentów. Wypracowane metody stanowią metodykę dla banków w zakresie konstrukcji modeli oceny zdolności kredytowej redukujących problem stronniczości algorytmicznej oraz metodykę dla regulatorów rynku kredytów konsumenckich w zakresie oceny konsekwencji społecznych wprowadzania regulacji równościowych. W rozprawie rozpatrywano modele oceny zdolności kredytowej, które są wykorzystywane przez banki do automatyzacji procesu podejmowania decyzji kredytowych związanych z aplikacjami kredytowymi konsumentów. Ich użycie oddziałuje na sytuację życiową osób fizycznych – konsumentów – i może wiązać się z dyskryminacją wydawanych decyzji kredytowych, która może mieć charakter odmiennego wpływu modelu (dyskryminacja pośrednia) lub odmiennego traktowania przez model (dyskryminacja bezpośrednia). W przypadku dyskryminacji w wydawanych decyzjach kredytowych, wykorzystywany model oceny zdolności kredytowej jest stronniczy algorytmicznie, co jest niezgodne z wymogami regulatorów w zakresie tych modeli.

Na podstawie celu rozprawy sformułowano sześć hipotez badawczych:

Hipoteza 1: O nieefektywności ekonomicznej metody sprawiedliwości przez nieświadomość. Metoda redukcji stronniczości algorytmicznej: sprawiedliwość przez nieświadomość nie jest ekonomicznie efektywna.

Hipoteza 2: O skuteczności redukcji stronniczości algorytmicznej w zależności od miary. W zależności od przyjętej miary stronniczości algorytmicznej, inne metody redukcji stronniczości algorytmicznej będą efektywne ekonomicznie; istnieje potrzeba utworzenia procedury wspierającej banki w ocenie poszczególnych metod redukcji stronniczości algorytmicznej.

Hipoteza 3: O procedurze redukcji stronniczości algorytmicznej maksymalizującej zysk banku. Możliwym jest opracowanie procedury redukującej stronniczość algorytmiczną modelu oceny zdolności kredytowej maksymalizującej zysk banku.

Hipoteza 4: O metodyce oceny długookresowych konsekwencji regulacji równościowych. Możliwym jest opracowanie metodyki oceny długookresowych konsekwencji

wprowadzania regulacji równościowych związanych z modelami oceny zdolności kredytowej, przy wykorzystaniu techniki symulacyjnej.

Hipoteza 5: O okresie dostosowawczym występującym po wprowadzeniu regulacji. Wprowadzenie regulacji równościowych może powodować krótkookresową niestabilność rozkładu wydawanych decyzji kredytowych.

Hipoteza 6: O koszcie ekonomicznym regulacji równościowych. Wprowadzenie regulacji równościowych prowadzi do kosztu ekonomicznego związanego ze zmianą populacji osób otrzymujących pozytywne decyzje kredytowe.

Hipotezy badawcze poddano weryfikacji na podstawie autorskich modeli symulacyjnych opracowanych w ramach przygotowania rozprawy: w zakresie weryfikacji hipotez 1, 2 oraz 3 przygotowano model statyczny opisany w rozdziale 4, a do weryfikacji hipotez 4, 5 oraz 6 utworzono model dynamiczny opisany w rozdziale 5.

Hipoteza 1: O nieefektywności ekonomicznej metody sprawiedliwości przez nieświadomość. Obecnie banki wykorzystują do redukcji stronniczości algorytmicznej podejście sprawiedliwość przez nieświadomość, zakładające całkowite odrzucenie informacji o wartości zmiennej prawnie chronionej z procesu budowy modelu i generowania prognozy modelu. Na podstawie przeprowadzonych prac i wykonanego, z wykorzystaniem modelu statycznego badania, których wyniki opisano w rozdziale 4, hipotezę zweryfikowano pozytywnie: metoda sprawiedliwości przez nieświadomość, bezwarunkowo, nie jest metodą efektywną ekonomicznie. W przypadku scenariusza danych syntetycznych: **Scenariusz Syntetyczny** wykazano, że odrzucenie informacji o zmiennej prawnie chronionej może prowadzić do istotnego spadku zysku z generowanego procesu kredytowego, przy równoczesnym braku spełnienia warunków na brak dyskryminacji pośredniej lub bezpośredniej (tabele 10, 11 oraz 12). Dla scenariusza danych rzeczywistych: **Scenariusz German Credit Data**, odrzucenie danych o zmiennej chronionej, jedynie w przypadku warunku wyrównanych szans, było ekonomicznie uzasadnione (tabela 15); zarówno dla warunku parytetu demograficznego oraz warunku jakości kalibracji istniały ekonomicznie efektywniejsze metody (tabele 14, oraz 16). Na podstawie powyższych wyników pracy można wskazać, że metoda sprawiedliwości przez nieświadomość nie jest bezwarunkowo (bez względu na miarę stronniczości algorytmicznej oraz wykorzystane dane) efektywna ekonomicznie.

Hipoteza 2: O skuteczności redukcji stronniczości algorytmicznej w zależności od miary. Wprowadzenie regulacji, które wymagać będą od banków redukcji problemu stronniczości algorytmicznej będzie wiązać się z decyzją regulatorów w zakresie definicji tej stronniczości. W zależności od przyjętej definicji stronniczości algorytmicznej i związanej z nią miary, inne metody redukcji stronniczości algorytmicznej będą efektywne ekonomicznie. W rozdziale 4, przy wykorzystaniu dwóch scenariuszy danych,

zidentyfikowano najbardziej ekonomicznie efektywne podejścia do redukcji stronniczości algorytmicznej. Podejście, nazwane w rozprawie podejściem *Unmitigated*, polegające na jawnym włączeniu zmiennej chronionej do zbioru zmiennych wykorzystywanych w procesie uczenia modelu, było podejściem generującym najwyższy wynik ekonomiczny spośród wszystkich metod (rysunki 13, lub 23). Wynik ten jest zgodny z intuicją ekonomiczną: model oceny zdolności kredytowej bez zastosowania warunków ograniczających na działanie tego modelu (np. wykluczenie predyktywnej zmiennej), będzie posiadał najwyższą jakość prognostyczną; w kontekście oceny zdolności kredytowej przekłada się to na najwyższy poziom generowanego zysku. Zastosowanie metod redukcji stronniczości algorytmicznej jest związane z nałożeniem restrykcji na model; w takim przypadku jakość prognostyczna modelu oceny zdolności kredytowej (a tym samym zysk) będzie nie większy od modelu bez tych restrykcji. W zależności od przyjętej definicji dyskryminacji: odmienny wpływ lub odmienne traktowanie, w danym kontekście danych, inne metody redukcji stronniczości algorytmicznej będą ekonomicznie efektywne. W badanych scenariuszach danych modelami oceny zdolności kredytowej, które charakteryzowały się najwyższym poziomem zyskowności były modele reprezentujące każdą z klas rozwiązań związanych z redukcją stronniczości algorytmicznej: metody przedtreningowe, treningowe i potreningowe (opisane w podrozdziale 3.4.2).

Hipoteza 3: O procedurze redukcji stronniczości algorytmicznej maksymalizującej zysk banku. Przy określonej przez regulatora definicji i mierze stronniczości algorytmicznej oraz zadanim zbiorze danych bankowych (z którym jest związana specyfika klientów banku), możliwym jest określenie metody redukcji stronniczości algorytmicznej, która umożliwia maksymalizację zysku banku. Procedura określająca ten zysk banku została w ramach rozprawy opracowana i przedstawiona w rozdziale 4 jako autorski model statyczny. Jednym z założeń tego podejścia jest określenie postaci macierzy kosztów banku; w rozprawie przyjęto macierz kosztów wskazaną przez Kozodoi i Lessmann (2022). Dla rozważanych scenariuszy danych – dane syntetyczne oraz dane rzeczywiste – przeprowadzono symulację działania modelu statycznego, które wskazywało oszacowania (na zbiorach testowych) stronniczości algorytmicznej poszczególnych metod, jak również poziomów zyskowności banku. Dodatkowo, w ramach modelu były raportowane wartości indeksu Giniego, stanowiącej miarę jakości działania modelu klasyfikacyjnego dla wszystkich progów odcięcia prognozy modelu.

Hipoteza 4: O metodyce oceny długookresowych konsekwencji regulacji równościowych. W ramach rozprawy poruszono jedną z luk badawczych istniejących w literaturze przedmiotu: oceny skutków wprowadzania regulacji dotyczących redukcji stronniczości algorytmicznej w długim horyzoncie czasu. W ramach tej metodyki opracowano model dynamiczny (rozdział 5). Jako przykład zastosowania metodyki oceny długookresowych konsekwencji regulacji równościowych wskazano symulacje przedstawione w rozdziale 5. Jednym z nieoczywistych wniosków wykorzystania modelu dynamicznego był wniosek

dotyczący krótkookresowych niestabilności średnich wartości parametrów związanych z procesem kredytowym. W rozprawie przedstawiono przykład, w którym krótkookresowe skutki wprowadzenia takiej regulacji były co do kierunku przeciwne względem kierunków zmian długookresowych; na rysunku 29 przedstawiono dynamikę zmian średnich wartości parametrów populacji otrzymujących kredyty: w krótkim okresie, dla scenariusza danych syntetycznych, średni poziom dochodu (lub wieku) osób uzyskujących kredyt w efekcie redukcji stronniczości algorytmicznej krótkookresowo spadł; w późniejszych okresach wartość średniej ustabilizowała się na poziomach wyższych od wartości wyjściowej.

Hipoteza 5: O okresie dostosowawczym występującym po wprowadzeniu regulacji. W ramach rozprawy opracowano dynamiczny model umożliwiający odpowiedzi na pytanie, jaka będzie dynamika zmian kluczowych parametrów związanych z procesem kredytowym – zarówno z perspektywy banku, jak i konsumentów – w związku z koniecznością wprowadzania redukcji stronniczości algorytmicznej w ocenie zdolności kredytowej. Decyzje regulatora dotyczące, w szczególności wyboru miary stronniczości algorytmicznej, mogą powodować krótkookresową niestabilność działania procesu kredytowego. Na podstawie wyników przedstawionych w rozdziale 5, hipotezę zweryfikowano pozytywnie: po rozpoczęciu stosowania metody redukcji stronniczości algorytmicznej, może wystąpić okresowa niestabilność wyników lub krokowa stabilizacja wyników (w przeciwieństwie do sytuacji, w której stabilność działania wyników modelu osiągnąć byłąby w jednym okresie – natychmiast); oba zjawiska sugerują zasadność wprowadzenia okresu dostosowawczego działania modelu oceny zdolności kredytowej po wprowadzenia metody redukcji stronniczości algorytmicznej. W rozprawie przedstawiono wyniki symulacji zmian w parametrach umożliwiających monitorowanie procesu kredytowego, zarówno dla podejścia *Unmitigated*, jak również *Fairness through Unawareness*. Krótkotrwała niestabilność występowała w modelach, w których jako podejście wyjściowe przyjęto metodę *Unmitigated*. Na rysunku 28 przedstawiono niestabilność dla miary *Sufficiency* po wdrożeniu metody *Disparate Impact Remover*; na rysunku 29 przedstawiono lokalną (krótkoterminową) niestabilność w wartościach atrybutów populacji otrzymującej kredyty; na rysunku 30 przedstawiono niestabilności związane z kluczowymi parametrami ekonomicznymi procesu kredytowego, tj. jakością portfela kredytowego oraz odsetek akceptacji kredytów. W zakresie podejścia wyjściowego *Fairness through Unawareness* zauważono, że występuje inne zjawisko: stabilizacja symulowanych parametrów (modelu oceny zdolności kredytowej, miar stronniczości algorytmicznej oraz populacji otrzymującej kredyty) była krokowa i trwała od kilku, do kilkunastu iteracji. Na rysunku 33 przedstawiono stabilizację atrybutów populacji otrzymującej kredyty, a na rysunku 35 poziom zysku z procesu kredytowego – oba rysunki wskazują na wydłużony okres stabilizacji przedstawionych wartości. W zakresie wyżej wskazanych wyników, jedną z rekomendacji dla regulatorów w zakresie projektowania regulacji obejmujących modele oceny zdolności kredytowej jest rozważanie stosowania

wydłużonych okresów dostosowawczych, umożliwiających np. stopniową realizację zakładanych progów miar stronniczości algorytmicznej.

Hipoteza 6: O koszcie ekonomicznym regulacji równościowych. Wprowadzenie regulacji równościowych prowadzi do kosztu ekonomicznego związanego ze zmianą populacji osób otrzymujących pozytywne decyzje kredytowe. W podrozdziale 5.4 wskazano na obserwację: w modelach oceny zdolności kredytowej występują co najmniej trzy współzależne wymiary oceny takiego modelu, przy czym przez model oceny zdolności kredytowej rozumie się całokształt opracowanej procedury, uwzględniając przy tym m.in. sposób trenowania takiego modelu, uwzględnienie (lub jego brak) sposobu redukcji stronniczości algorytmicznej, próg akceptacji wniosków kredytowych lub transformacje na zmiennych objaśniających. Wskazane wymiary to: jakość portfela kredytowego (mierzona kohortowym poziomem spłacalności kredytów), wskaźnik zatwierdzeń wniosków kredytowych (mierzony liczbą udzielonych kredytów w stosunku do złożonych wniosków) oraz poziom stronniczości algorytmicznej jaką dany model posiada (w rozważaniach przyjęto miarę dyskryminacji pośredniej). Dwa pierwsze wskaźniki, tj.: jakość portfela kredytowego oraz wskaźnik zatwierdzeń wniosków kredytowych są ściśle związane z ekonomicznym wpływem stosowania modelu oceny zdolności kredytowej, tj. wysoki poziom zysku z prowadzonej akcji kredytowej będzie osiągany przy stosunkowo wysokim poziomie akceptacji wniosków kredytowych, z których kredyty docelowo będą terminowo spłacane. Poziom stronniczości algorytmicznej stanowi niejako warunek ograniczający na dwa wcześniejsze parametry: dla zadanego poziomu dyskryminacji pośredniej istnieje zbiór możliwych wartości odpowiednio jakości portfela kredytowego oraz wskaźnika akceptacji – oba będące funkcją progu pozytywnej decyzji kredytowej. Dla zadanego poziomu ograniczenia stronniczości algorytmicznej istnieje krzywa, Pareto efektywna, możliwych poziomów akceptacji wniosków kredytowych i wynikowej jakości portfela kredytowego. Dodatkowo w ramach weryfikacji hipotezy wskazano na autorską metodę oceny poziomu zaburzenia struktury populacji kredytobiorców, przedstawionej w podrozdziale 4.2.2.

Wszystkie hipotezy zostały pozytywnie zweryfikowane. Badanie przeprowadzone w ramach weryfikacji tych hipotez pozwoliło na wypracowanie metodyki wspomagania decyzji regulatorów w zakresie tworzenia oraz banków w zakresie wdrażania przepisów równościowych w obszarze modeli oceny zdolności kredytowej konsumentów – tym samym przedstawiony w rozprawie cel został osiągnięty.

Wyniki uzyskane w rozprawie nie dostarczają odpowiedzi na wszystkie pytania, jakie można sformułować w zakresie wdrażania metod ograniczających stronniczość algorytmiczną w modelach oceny zdolności kredytowej konsumenta. Tematyka rozprawy ma charakter interdyscyplinarny i dotyczy zarówno zagadnień prawnych, obszaru modelowania ekonometrycznego oraz jest związana z technicznymi aspektami implementacji metod redukcji stronniczości algorytmicznej. Poniżej przedstawiono proponowane kierunki rozwoju omawianego w rozprawie tematu.

Jedną z kontrybucji rozprawy jest wprowadzenie do dyskursu naukowego analizy efektów długoterminowych w obszarze stronniczości algorytmicznej modeli oceny zdolności kredytowej (podejście modelu dynamicznego opisane w rozdziale 5). Wcześniejsze prace, przytoczone w rozprawie, odnoszą się do zagadnień technicznych i metodycznych pomiaru oraz redukcji stronniczości algorytmicznej tych modeli. Luką badawczą jest zatem rozwinięcie analizy długoterminowych skutków wprowadzania regulacji równościowych i towarzyszących im rozwiązań modelowych w zakresie redukcji stronniczości algorytmicznej. Ważnymi zagadnieniami badawczymi są kwestie długoterminowej konkurencyjności banków w przypadku wprowadzenia regulacji mających na celu redukcję problemu stronniczości algorytmicznej; rozważania te byłyby rozwinięciem modeli przedstawionych przez Broecker (1990) o aspekty regulacji przeciwdziałających dyskryminacji.

Inną luką badawczą, niepodjętą w ramach niniejszej rozprawy, jest odpowiedź na pytanie, w jaki sposób cechy zmiennych i relacje między nimi (np. nieliniowość, interakcje lub korelacje między zmiennymi objaśniającymi) wpływają na poziom stronniczości algorytmicznej. W ramach rozprawy opracowano metodykę oceny skutków dla banków stosowania metod redukcji stronniczości algorytmicznej, których stosowanie uzasadnione jest w przypadku wprowadzenia regulacji nakładających obowiązek spełnienia określonych miar stronniczości algorytmicznej. Jako przykład identyfikacji tych zależności można podać ocenę wpływu występowania problemu współliniowości zmiennych objaśniających na oszacowania wariancji parametrów modelu liniowego przedstawioną przez O'Brien (2007).

Ważnym rozwinięciem przedstawionych w rozprawie wyników byłoby również rozważenie zaproponowanych metodyk oceny skutków wykorzystania metod redukcji stronniczości algorytmicznej (model statyczny opisany w rozdziale 4 oraz model dynamiczny opisany w rozdziale 5), o inne zbiory danych. Dane te mogłyby dotyczyć zarówno danych rzeczywistych, jak również innych modeli oraz generatorów danych symulacyjnych, jak np. te przedstawione przez Przanowski (2014).

Ponadto, tematyka rozprawy dotyczy zmiany funkcjonowania procesu podejmowania decyzji w banku, jakim jest proces oceny zdolności kredytowej. W rozdziale 4 przedstawiono autorską metodykę analizy wpływu wykorzystywanych metod redukcji stronniczości algorytmicznej – rysunek rozbieżności prognoz. Zaproponowane rozwiązanie umożliwia porównanie rozbieżności prognoz modelu klasyfikacyjnego z modelem dotychczas wykorzystywanym; w kontekście oceny zdolności kredytowej dotychczasowym modelem jest model nieuwzględniający mechanizmu redukcji problemu stronniczości algorytmicznej. Motywacją za opracowaniem rysunku rozbieżności prognoz było założenie, że czym większa rozbieżność prognoz spowodowana przez zmianę modelu, tym w mniejszym stopniu bank jest skłonny do podjęcia takiej zmiany; jednocześnie autor rozprawy dostrzega inne aspekty związane z podejmowaniem decyzji banku o zmianach

w modelach oceny zdolności kredytowej, tj. strukturalne zmiany procesu generującego dane lub wprowadzenie regulacji związanych z wymogiem redukcji stronniczości algorytmicznej. W opinii autora rozprawy istnieje luka badawcza związana z metodami wspomagania analizy decyzji banków w zakresie zmian modeli oceny zdolności kredytowej (np. podejść opartych o tabelę częstości); banki podejmując decyzje w zakresie zmiany modelu oceny zdolności kredytowej, np. ze względu na problem stronniczości algorytmicznej, powinny uwzględniać aspekt *poziomu* rozbieżności decyzji kredytowych, jakie dana zmiana spowoduje.

Aneks: Specyfikacja danych

W niniejszym aneksie przedstawiono scenariusze przeprowadzonych eksperymentów numerycznych.

Scenariusz Syntetyczny

Scenariusz danych syntetycznych zakładający występowanie zjawiska współliniowości zmiennych egzogenicznych oraz nieliniowości wpływu jednej ze zmiennych objaśniających (tj. zmienna **wiek**). Dodatkowo przyjmuje się niebilansowanie próby względem zmiennej **płeć** (odsetek mężczyzn wynosi 65%, a kobiet 35%). W scenariuszu założono, że wartości zmiennych ciągłych są już po wstępnym procesie standaryzacji (dla tych zmiennych w symulacji wykorzystano standardowy rozkład normalny).

Wielkość populacji: $n = 2\,000\,000$

Zbiór uczący / testowy: 80/20

Postać procesu generującego dane

$$P(y_i = 1 \mid x_i) = \left\{ 1 - \exp \left[- \left(\sum_{i=0,1,3,4} \beta_i x_i + \beta_2 x_2^2 \right) \right] \right\}^{-1}$$

Zmienne egzogeniczne

stała (x_0) $\sim$ const., $\beta_0 = +2,80$
płeć (x_1) $\sim$ Bern(0,65), $\beta_1 = -1,60$
wiek (x_2) $\sim$ $\mathcal{N}(0, 1)$, $\beta_2 = -1,60$
dochody (x_3) $\sim$ $\mathcal{N}(0, 1)$, $\beta_3 = -1,50$
wydatki (x_4) $\sim$ $\mathcal{N}(0, 1)$, $\beta_4 = -1,50$

Zakładana struktura korelacji Pearsona zmiennych egzogenicznych

$$\rho \sim \begin{bmatrix} 1,00 & 0,00 & 0,05 & -0,25 \\ 0,00 & 1,00 & 0,30 & 0,15 \\ 0,05 & 0,30 & 1,00 & 0,40 \\ -0,25 & 0,15 & 0,40 & 1,00 \end{bmatrix}$$

Scenariusz German Credit Data

Zbiór danych **German Credit Data** zawiera 1000 rekordów (wiersze), z których każdy reprezentuje jednego klienta banku starającego się o kredyt w Niemczech. Dane zawierają 20 zmiennych egzogenicznych (kolumny) opisujących poszczególne aspekty profilu kredytobiorcy oraz jedną zmienną docelową oznaczającą decyzję kredytową. Zmienna docelowa przyjmuje wartość binarną, gdzie „1” oznacza klienta o niskim ryzyku kredytowym, a „0” klienta o wysokim ryzyku.

W analizie przyjęto, że zmienną prawnie chronioną będzie narodowość – zmienna **foreign worker** przyjmuje wartości binarne „0” dla obywateli niemiec, a „1” dla osób o innej narodowości.

Wielkość populacji: $n = 1\,000$ **Zbiór uczący / testowy:** 80/20

Tabela 18: Scenariusz German Credit Data – zmienne egzogeniczne

Nazwa zmiennej	Typ	Grupa demograficzna	Opis	Informacje o zmiennych
Attribute1	Kategoryczna		Status istniejącego konta czekowego	A11: ... < 0 DM A12: 0 <= ... < 200 DM A13: ... >= 200 DM A14: brak konta czekowego
Attribute2	Całkowita		Czas trwania	Ilość miesięcy
Attribute3	Kategoryczna		Historia kredytowa	A30: brak pobranych kredytów / wszystkie kredyty należycie spłacone A31: wszystkie kredyty w banku spłacone należycie A32: istniejące kredyty spłacone należycie aż do teraz A33: opóźnienie w spłacie w przeszłości A34: konto krytyczne / inne istniejące kredyty

Nazwa zmiennej	Typ	Grupa demograficzna	Opis	Informacje o zmiennych
Attribute4	Kategoryczna		Cel	A40: samochód (nowy), A41: samochód (używany) A42: meble/sprzęt, A43: radio/telewizja A44: sprzęt AGD, A45: remont A46: edukacja, A47: (wakacje - nie istnieje?) A48: przekwalifikowanie, A49: biznes A410: inne
Attribute5	Całkowita		Kwota kredytu	Kwota liczbowa
Attribute6	Kategoryczna		Konto oszczędnościowe/obligacje	A61: ... < 100 DM A62: 100 <= ... < 500 DM A63: 500 <= ... < 1000 DM A64: ... >= 1000 DM A65: nieznany/brak konta oszczędnościowego
Attribute7	Kategoryczna	Inne	Obecne zatrudnienie od	A71: bezrobotni A72: ... < 1 rok A73: 1 <= ... < 4 lata A74: 4 <= ... < 7 lat A75: ... >= 7 lat
Attribute8	Całkowita		Stopa raty kredytu jako procent dochodu do dyspozycji	

Nazwa zmiennej	Typ	Grupa demograficzna	Opis	Informacje o zmiennych
Attribute9	Kategoryczna	Stan cywilny	Status osobisty i płeć	A91: mężczyzna: rozwiedziony/w separacji A92: kobieta: rozwiedziony/w separacji/zamężna A93: mężczyzna: kawaler A94: mężczyzna: żonaty/wdowiec A95: kobieta: niezamężna
Attribute10	Kategoryczna		Inni dłużnicy / poręczyciele	A101: brak A102: współwnioskodawca A103: poręczyciel
Attribute11	Całkowita		Obecne miejsce zamieszkania od	
Attribute12	Kategoryczna		Własność	A121: nieruchomość A122: jeśli nie A121: umowa oszczędnościowa spółdzielni budowlanej /ubezpieczenie na życie A123: jeśli nie A121/A122: samochód lub inny, nie w atrybucie 6 A124: nieznany / brak własności
Attribute13	Całkowita	Wiek	Wiek	lata
Attribute14	Kategoryczna		Inne plany ratalne	A141: bank A142: sklepy A143: brak

Nazwa zmiennej	Typ	Grupa demograficzna	Opis	Informacje o zmiennych
Attribute15	Kategoryczna	Inne	Mieszkanie	A151: wynajem A152: własne A153: darmowe
Attribute16	Całkowita		Liczba istniejących kredytów w tym banku	
Attribute17	Kategoryczna	Zawód	Praca	A171: bezrobotny/niewykwalifikowany (nierezydent) A172: niewykwalifikowany (rezydent) A173: pracownik wykwalifikowany / urzędnik A174: kadra zarządzająca / osoba samozatrudniona / wysoko wykwalifikowany pracownik / urzędnik
Attribute18	Całkowita		Liczba osób zobowiązanych do utrzymania	
Attribute19	Binarna		Telefon	A191: brak A192: tak, zarejestrowany na nazwisko klienta
Attribute20	Binarna	Inne	Pracownik zagraniczny	A201: tak A202: nie
Class	Binarna		1 = Dobry, 2 = Zły	

Źródło: opracowanie własne na podstawie Hofmann (1994).

Bibliografia

Akty prawne

- Rozporządzenie Parlamentu Europejskiego i rady (UE) nr 575/2013 z dnia 26 czerwca 2013 r. w sprawie wymogów ostrożnościowych dla instytucji kredytowych i firm inwestycyjnych, zmieniające rozporządzenie (UE) nr 648/2012 (Dz.U. L 176 z 26.6.2013)
- Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 czerwca 2024 r. w sprawie ustanowienia zharmonizowanych przepisów dotyczących sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (AI Act), (Dz.U. L 1689 z 12.7.2024)
- Basel Committee on Banking Supervision, 2023. *IRB approach: minimum requirements to use IRB approach*. Bank for International Settlements
- European Data Protection Supervisor (EDPS), 2021. *Opinion 11/2021 on the Proposal for Directive on customer credits*
- International Financial Reporting Standard 9, 2021. [ifrs.org/issued-standards/list-of-standards/ifrs-9-financial-instruments/#about](https://www.ifrs.org/issued-standards/list-of-standards/ifrs-9-financial-instruments/#about)
- European Banking Authority, 2016. *Guidelines on the application of the definition of default under Article 178 of Regulation (EU) No 575/2013*
- Europejski Urząd Nadzoru Bankowego, 2010. *Wytyczne CEBS w sprawie zarządzania ryzykiem koncentracji wynikające z procesu badania i oceny nadzorczej (GL31)*. EUNB
- Ustawa z dnia 6 października 2022 r. o zmianie ustaw w celu przeciwdziałania lichwie (Dz.U. 2022 poz. 2339)
- Rekomendacja S Komisji Nadzoru Finansowego dotycząca dobrych praktyk w zakresie zarządzania ekspozycjami kredytowymi zabezpieczonymi hipotecznie (Rekomendacja S)
- Ustawa z dnia 23 marca 2017 r. o kredycie hipotecznym oraz o nadzorze nad pośrednikami kredytu hipotecznego i agentami (Dz.U. 2017 poz. 819)
- Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (RODO), (Dz.U. L 119/1 z 4.5.2016)
- Komisja Nadzoru Finansowego, 2015. *Rekomendacja W dotycząca zarządzania ryzykiem modeli w bankach*

Ustawa z dnia 12 maja 2011 r. o kredycie konsumenckim (Dz.U. 2011 nr 126 poz. 715)

Ustawa z dnia 3 grudnia 2010 r. o wdrożeniu niektórych przepisów Unii Europejskiej w zakresie równego traktowania (Dz.U. 2010 nr 254 poz. 1700)

Ustawa z dnia 21 lipca 2006 r. o nadzorze nad rynkiem finansowym (Dz.U. 2006 nr 157 poz. 1119)

Decyzja Rady z dnia 30 stycznia 2006 r. w sprawie zasad, priorytetów i warunków zawartych w partnerstwie europejskim z Serbią i Czarnogórą, w tym z Kosowem w rozumieniu rezolucji nr 1244 Rady Bezpieczeństwa Organizacji Narodów Zjednoczonych z dnia 10 czerwca 1999 r. oraz uchylająca decyzję 2004/520/WE (Dz.U. L 35/32 z 7.2.2006)

Dyrektywa Rady 2004/113/WE z dnia 13 grudnia 2004 r. wprowadzająca w życie zasadę równego traktowania mężczyzn i kobiet w zakresie dostępu do towarów i usług oraz dostarczania towarów i usług (Dz.U. L 373/37 z 21.12.2004)

Dyrektywa Rady 2000/43/WE z dnia 29 czerwca 2000 r. wprowadzająca w życie zasadę równego traktowania osób bez względu na pochodzenie rasowe lub etniczne (Dz.U. L 180 z 19.7.2000)

Ustawa z dnia 23 kwietnia 1964 r. – Kodeks cywilny (Dz.U. 1964 nr 16 poz. 93)

Ustawa z dnia 29 sierpnia 1997 r. – Prawo bankowe (Dz.U. 1997 nr 140 poz. 939)

Cal. Civ. Code §§ 1798.100 et seq., California Consumer Privacy Act of 2018

Pub. L. 100–430, 102 Stat. 1619; 42 U.S.C. § 3601 et seq., Fair Housing Amendments Act (FHAA 1988)

12 CFR Part 1002, Regulacja B, implementująca Equal Credit Opportunity Act (ECOA–B 1976)

Pub. L. 93–495, 88 Stat. 1500; 15 U.S.C. § 1666, Fair Credit Billing Act (FCBA 1974)

Pub. L. 93–495, 88 Stat. 1521; 15 U.S.C. § 1691, Equal Credit Opportunity Act (ECOA 1974)

Pub. L. 91–508, 84 Stat. 1114; 31 U.S.C. § 5311–5330, Bank Secrecy Act (BSA 1970)

Pub. L. 91–508, 84 Stat. 1127; 15 U.S.C. § 1681, Fair Credit Reporting Act (FCRA 1970)

Pub. L. 90–284, 82 Stat. 73; 42 U.S.C. § 3601, Fair Housing Act (FHA 1968)

Pub. L. 90–321, 82 Stat. 146; 15 U.S.C. § 1601, Consumer Credit Protection Act (CCPA 1968)

Dokumenty

Wachoł, M., 2023. *Odpowiedź na interpelację nr 38607 w sprawie zagrożeń ze strony algorytmów AI*. Warszawa, Ministerstwo Sprawiedliwości

Rzecznik Praw Obywatelskich, 2021. *Czym Jest Dyskryminacja*. Biuletyn Rzecznika Praw Obywatelskich

Rzecznik Praw Obywatelskich (RPO), 2020. *Ochrona przed dyskryminacją w Polsce. Stan prawny i świadomość społeczna. Wnioski i rekomendacje RPO*. Biuletyn Rzecznika Praw

Obywatelskich

Experian, 2019. *State of alternative credit data*.

Biuro Wykonawcze Prezydenta USA, 2016. *Big data: A report on algorithmic systems, opportunity, and civil rights*.

Wójcicka, K., 2015. *Kredyt bankowy i pożyczka – czy znaczą to samo?*. Warszawa, Komisja Nadzoru Finansowego

Wójcicka, K., 2015. *Karty płatnicze w Polsce*. Warszawa, Komisja Nadzoru Finansowego

Łojkowska, M., i Pawłęga, M., 2014. *Rozdział 1: Polityka antydyskryminacyjna*.

W: Bregin, D., Pawłęga, M., Rutkowska, E., i Łojkowska, M., 2014. *Równe traktowanie w administracji publicznej*. Biuro Pełnomocnika Rządu do spraw Równego Traktowania

Małocha-Krupa, A., Hołojda, K., Krysiak, P. i Pietrzak, W., 2013. *Równościowy savoir-vivre w tekstach publicznych*. Biuro Pełnomocnika Rządu do spraw Równego Traktowania

U.S. House of Representatives, 2010. *Hearing before the Subcommittee on Capital Markets, Insurance, and Government Sponsored Enterprises of the Committee on Financial Services*. 111th Congress, 2nd Session, Washington, D.C.: U.S. Government Printing Office

Schwab, K., 2010. *The global competitiveness report 2010–2011*. World Economic Forum Geneva

Niezbecka, E., Kidyba, A., 2009. *Kodeks cywilny. Komentarz. t. 1*. WKP 2009

Fair, Isaac i Company Incorporated (Annual Report), 1998. *FICO Investor Relations*

Orzecznictwo

Louis v. Saferent Sols., Civil Action 22–CV–10800–AK (D. Mass. Jul. 26, 2023)

Alexander v. Edgewood Mgmt. Corp. 2019. Civil Case No. 15–1140 (D.D.C. Jun. 25)

Elston v. Talladega Cty. Bd. of Educ, 997 F.2d 1394 (11th Cir. 1993)

United States v. Hazelwood School Dist, 534 F.2d 805 (8th Cir. 1976)

Druki zwarte i ciągłe

Abdou, H.A., 2009. *Genetic programming for credit scoring: The case of Egyptian public sector banks*. Expert Systems with Applications 36(9), s. 11402–11417

Agarwal, S., He, J., Sing, T.F. i Zhang, J., 2018. *Gender gap in personal bankruptcy risks: Empirical evidence from Singapore*. Review of Finance 22(2), s. 813–847

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. i Wallach, H., 2018. *A reductions approach to fair classification*. International Conference on Machine Learning, s. 60–69

Agarwal, A., Dudík, M. i Wu, Z.S., 2019. *Fair regression: Quantitative definitions and reduction-based algorithms*. International Conference on Machine Learning, s. 120–129

- Allen, D.M., 1971. *Mean square error of prediction as a criterion for selecting variables*. Technometrics 13(3), s. 469–475
- Andreeva, G. i Matuszyk, A., 2019. *The law of equal opportunities or unintended consequences?: The effect of unisex risk assessment in consumer credit*. Journal of the Royal Statistical Society Series A: Statistics in Society 182(4), s. 1287–1311
- Andreeva, G., Ansell, J. i Crook, J., 2004. *Impact of anti-discrimination laws on credit scoring*. Journal of Financial Services Marketing 9, s. 22–33
- Asiedu, E., Freeman, J.A. i Nti-Addae, A., 2012. *Access to credit by small businesses: How relevant are race, ethnicity, and gender?* American Economic Review 102(3), s. 532–537
- Baesens, B., Roesch, D. i Scheule, H., 2016. *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. Hoboken, John Wiley & Sons
- Barlow, C. T., 1978. *Bankers, moneylenders, and interest rates in the Roman Republic*. Chapel Hill
- Barocas, S., Hardt, M. i Narayanan, A., 2023. *Fairness and machine learning: Limitations and opportunities*. Cambridge, MIT Press
- Baumol, W.J., Koehn, M.F., i Willig, R.D., 1987. *How Arbitrary is „Arbitrary” – or, Toward the Deserved Demise of Full Cost Allocation*. Public Utilities Fortnightly, s. 16–21
- Bednarski, F., W., 1970. *Suma teologiczna, II-II, q. 57–80: Sprawiedliwość*. Londyn
- Bentham, J., 1816. *A Defence of Usury: Shewing the Impolicy of the Present Legal Restraints on the Terms of Pecuniary Bargains, in a Series of Letters to a Friend*. Cambridge, Cambridge University Press
- Berg, T., Burg, V., Gombović, A. i Puri, M., 2020. *On the rise of fintechs: Credit scoring using digital footprints*. The Review of Financial Studies 33(7), s. 2845–2897
- Biecek, P., 2024. *Sprawiedliwość nie istnieje i można to udowodnić*. hAI Magazine CampusAI 1(1), s. 101–103
- Biggar, D., 2010. *Fairness in public-utility regulation: a theory*. Agenda: A Journal of Policy Analysis and Reform, s. 5–29
- Blicharz, G., Oreziak, B., Ratowski, E., Ratowski, K., Walczak, P. i Wielec, M., 2019. *Rynek finansowy: zapobieganie przyczynom przestępczości*. Warszawa, Wydawnictwo Instytutu Wymiaru Sprawiedliwości
- Bonbright, J.C., Danielsen, A.L., i Kamerschen, D.R., 1988. *Principles of Public Utility Rates*. Arlington, Public Utilities Reports
- Bono, T., Croxson, K. i Giles, A., 2021. *Algorithmic fairness in credit scoring*. Oxford Review of Economic Policy 37(3), s. 585–617
- Brevoort, K.P., Grimm, P. i Kambara, M., 2015. *Data Point: Credit invisibles*. Consumer Financial Protection Bureau Office of Research Reports Series 15(1), s. 1–37
- Brier, G.W., 1950. *Verification of forecasts expressed in terms of probability*. Monthly Weather Review 78(1), s. 1–3
- Broecker, T., 1990. *Credit-worthiness tests and interbank competition*. Econometrica:

- Journal of the Econometric Society 58(2), s. 429–452
- Brosnan, S.F. i De Waal, F.B., 2003. *Monkeys reject unequal pay*. Nature 425(6955), s. 297–299
- Brosnan, S.F. i De Waal, F.B., 2012. *Fairness in animals: Where to from here?* Social Justice Research 25, s. 336–351
- Burek, W. i Klaus, W., 2013. *Definiowanie dyskryminacji w prawie polskim w świetle prawa Unii Europejskiej oraz prawa międzynarodowego*. Problemy Współczesnego Prawa Międzynarodowego, Europejskiego i Porównawczego 11, s. 72–90
- Butler, A.W., Mayer, E.J. i Weston, J.P., 2023. *Racial disparities in the auto loan market*. The Review of Financial Studies 36(1), s. 1–41
- Chen, J., 2018. *Fair lending needs explainable models for responsible recommendation*. Proceedings of Workshop on Responsible Recommendation 18, s. 1–4
- Chouldechova, A., 2017. *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. Big Data 5(2), s. 153–163
- Clarke, K.A., 2005. *The phantom menace: Omitted variable bias in econometric research*. Conflict Management and Peace Science 22(4), s. 341–352
- Cohen, E., 2011. *Athenian Economy and Society: A Banking Perspective*. Princeton, Princeton University Press
- Cohen, M., 2012. *Cotton, Capital, and Ethnic Networks: Jewish Economic Growth in the Postbellum Gulf South*. American Jewish Archives Journal 64, s. 112–36
- Coin, D., 2013. *Are female entrepreneurs better payers than men?*. Bank of Italy Occasional Paper 186, s. 1–16
- Corbett–Davies, S., Gaebler, J.D., Nilforoshan, H., Shroff, R. i Goel, S., 2023. *The measure and mismeasure of fairness*. The Journal of Machine Learning Research 24(1), s. 14730–14846
- D’espallier, B., Guérin, I. i Mersland, R., 2011. *Women and repayment in microfinance: A global analysis*. World Development, 39(5), s. 758–772
- Danks, D. i London, A.J., 2017. *Algorithmic Bias in Autonomous Systems*. Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), s. 4691–4697
- Deighton–Smith, R. i Jacobs, S.H., 1997. *Regulatory impact analysis: best practices in OECD countries*. Paris, OECD
- Delton, A.W., Cosmides, L., Guemo, M., Robertson, T.E. i Tooby, J., 2012. *The psychosemantics of free riding: dissecting the architecture of a moral concept*. Journal of Personality and Social Psychology 102(6), s. 1252–1270
- Do, C. i Paley, I., 2013. *Does gender affect mortgage choice? Evidence from the US*. Feminist Economics 19(2), s. 33–68
- Dumitrescu, E., Hué, S., Hurlin, C. i Tokpavi, S., 2022. *Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects*. European Journal of

Operational Research 297(3), s. 1178–1192

- Durkin, T.A., 2000. *Credit cards: Use and consumer attitudes, 1970–2000*. Federal Reserve Bulletin 86, s. 623–634
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O. i Zemel, R., 2012. *Fairness through awareness*. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, s. 214–226
- Farber, D., 2002. *Sloan Rules: Alfred P. Sloan and the Triumph of General Motors*. Chicago, The University of Chicago Press
- Fawcett, T., 2006. *An introduction to ROC analysis*. Pattern Recognition Letters 27(8), s. 861–874
- Fehr, E., Kirchsteiger, G., i Riedl, A., 1993. *Does fairness prevent market clearing? An experimental investigation*. The Quarterly Journal of Economics 108(2), s. 437–459
- Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C. i Venkatasubramanian, S., 2015. *Certifying and removing disparate impact*. KDD '15: The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, s. 259–268
- Filipkowski, W., 2006. *Nieformalne systemy transferu wartości majątkowych*. Ruch Prawniczy, Ekonomiczny i Socjologiczny 2006(1), s. 123–139
- Fisher, R., A., 1970. *Statistical Methods for Research Workers*. Edinburgh, Hafner Publishing
- Foley, D.K., 1966. *Resource allocation and the public sector*. Ann Arbor, Yale University
- Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P. i Roth, D., 2019. *A comparative study of fairness-enhancing interventions in machine learning*. Proceedings of the Conference on Fairness, Accountability, and Transparency, s. 329–338
- Friis, S. i Riley, J., 2023. *Eliminating algorithmic bias is just the beginning of equitable AI*. Harvard Business Review 2023(9), s. 1–7
- Goetzmann, W., 2017. *Money changes everything: How finance made civilization possible*. New Haven, Princeton University Press
- Greene, W.H., 2000. *Econometric analysis 4th edition*. New Jersey, Prentice Hall
- Harris, W.V., 2008. *The nature of Roman money*. W: *The monetary systems of the Greeks and Romans* red. Harris, W.V., Oxford University Press, s. 174–207
- Havard, C.J., 2010. *On the take: The black box of credit scoring and mortgage discrimination*. Boston University Public Interest Law Journal 20(2), s. 241–287,
- Henderson, L., Herring, C., Horton, H.D. i Thomas, M., 2015. *Credit where credit is due? Race, gender, and discrimination in the credit scores of business startups*. The Review of Black Political Economy 42(4), s. 459–479
- Hermann, C., Johns, W., 1910. *Babylonian Law–The Code of Hammurabi*. Eleventh Edition of the Encyclopedia Britannica
- Hooker, S., 2021. *Moving beyond „algorithmic bias is a data problem”*. Patterns 2(4), s. 1–4

- Horngren, C., T., Datar, S., M., Rajan, M., V., 2014. *Cost Accounting – a managerial emphasis*. Harlow, Pearson
- Hosmer, D. W., Lemeshow, S., 2013. *Applied Logistic Regression*. New York, Wiley
- Houssou, R., Augustin, M.C., Rappos, E., Bonvin, V. i Robert–Nicoud, S., 2022. *Generation and Simulation of Synthetic Datasets with Copulas*. arXiv:2203.17250, s. 1–11
- Hudson, M., Mieroop, M., 2002. *Debt and Economic Renewal in the Ancient Near East*. Michigan, CDL Press
- Hurley, M. i Adebayo, J., 2016. *Credit scoring in the era of big data*. Yale Journal of Law & Technology 18, s. 148–216
- Hurlin, C., Pérignon, C. i Saurin, S., 2022. *The fairness of credit scoring models*. HEC Paris Research Paper 2021(1411), s. 1–69
- Iwanicz–Drozdowska, M., Jaworski, W.L., Szelągowska, A. i Zawadzka, Z., 2017. *Bankowość: instytucje, operacje, zarządzanie*. Warszawa, Poltext
- Kaczmarek, T. T., 2006. *Ryzyko i zarządzanie ryzykiem. Ujęcie interdyscyplinarne*. Warszawa, Difin
- Kahneman, D., Knetsch, J.L., i Thaler, R.H., 1986. *Fairness and the assumptions of economics*. Journal of Business 59(4), s. 285–300
- Kamiran, F. i Calders, T., 2012. *Data preprocessing techniques for classification without discrimination*. Knowledge and Information Systems 33(1), s. 1–33
- Kamiran, F., Karim, A. i Zhang, X., 2012. *Decision theory for discrimination-aware classification*. IEEE 12th International Conference on Data Mining, s. 924–929
- Kamishima, T., Akaho, S., Asoh, H. i Sakuma, J., 2012. *Fairness-aware classifier with prejudice remover regularizer*. Machine Learning and Knowledge Discovery in Databases: European Conference, s. 35–50
- Kaszyński, D., Kamiński, B., Szapiro, T., 2020. *Credit scoring in context of interpretable machine learning*. Warszawa, SGH Publishing House
- Kennedy, K., Delany, S.J., Namee, B.M., 2011. *A framework for generating data to simulate application scoring*. Credit Scoring and Credit Control XII: Conference Proceedings, Credit Research Centre, s. 1–21
- Kim, J.S., Chen, J. i Talwalkar, A., 2020. *FACT: A diagnostic for group fairness trade-offs*. International Conference on Machine Learning, s. 5264–5274
- Kleinberg, J., Mullainathan, S. i Raghavan, M., 2016. *Inherent trade-offs in the fair determination of risk scores*. 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), s. 1–23
- Klimontowicz, M., 2017. *Bankowość dla praktyków. Europejski Certyfikat Bankowca EFCB 3E*. Warszawa, Związek Banków Polskich
- Knight, F.H., 1921. *Risk, uncertainty and profit*. Iowa, Houghton Mifflin Company
- Konow, J., 2003. *Which is the fairest one of all? A positive analysis of justice theories*. Journal of Economic Literature 41(4), s. 1188–1239

- Kordzadeh, N. i Ghasemaghaei, M., 2022. *Algorithmic bias: review, synthesis, and future research directions*. European Journal of Information Systems 31(3), s. 388–409
- Kozodoi, N., Jacob, J. i Lessmann, S., 2022. *Fairness in credit scoring: Assessment, implementation and profit implications*. European Journal of Operational Research 297(3), s. 1083–1094
- Lee, T.S., Chiu, C.C., Lu, C.J. i Chen, I.F., 2002. *Credit scoring using the hybrid neural discriminant technique*. Expert Systems with Applications 23(3), s. 245–254
- Liu, L.T., Dean, S., Rolf, E., Simchowitz, M. i Hardt, M., 2018. *Delayed impact of fair machine learning*. International Conference on Machine Learning, s. 3150–3158
- Louzada, F., Ara, A. i Fernandes, G.B., 2016. *Classification methods applied to credit scoring: Systematic review and overall comparison*. Surveys in Operations Research and Management Science 21(2), s. 117–134
- Makhlouf, K., Zhioua, S. i Palamidessi, C., 2021. *On the applicability of machine learning fairness notions*. ACM SIGKDD Explorations Newsletter 23(1), s. 14–23
- Marques, A.I., García, V. i Sánchez, J.S., 2013. *A literature review on the application of evolutionary computing to credit scoring*. Journal of the Operational Research Society 64(9), s. 1384–1399
- Matuszyk, A., 2004. *Credit scoring: metoda zarządzania ryzykiem kredytowym*. Warszawa, CeDeWu
- Morawski, W., 2003. *Kronika kryzysów gospodarczych*. Warszawa, Wydawnictwo Trio
- Morsy, H., 2020. *Access to finance: Why aren't women leaning in?*. Finance & Development 57(1), s. 52–53
- Muir, J., 2001. *Fair pricing regulation*. Agenda: A Journal of Policy Analysis and Reform 8(1), s. 3–18
- Munro, J.H., 2011. *Usury, Calvinism, and Credit in Protestant England: from the Sixteenth Century to the Industrial Revolution*. University of Toronto Working Papers 439, s. 1–41
- Muravyev, A., Talavera, O. i Schäfer, D., 2009. *Entrepreneurs' gender and financial constraints: Evidence from international data*. Journal of Comparative Economics 37(2), s. 270–286
- Myers J.H., Forgy, E. W., 1963. *The development of numerical credit evaluation systems*. Journal of the American Statistical Association 58(303), s. 799–806
- Niculescu-Mizil, A. i Caruana, R., 2005. *Predicting good probabilities with supervised learning*. Proceedings of the 22nd International Conference on Machine learning, s. 625–632
- Norris, J., 1978. *R.G. Dun and Co. 1841–1900: The Development of Credit-Reporting in the Nineteenth Century*. Westport, Greenwood Press
- O'Brien, R.M., 2007. *A caution regarding rules of thumb for variance inflation factors*. Quality & quantity 41, s. 673–690
- Okesola, O.J., Okokpuije, K.O., Adewale, A.A., John, S.N. i Omoruyi, O., 2017, December.

- An improved bank credit scoring model: a naïve Bayesian approach*. 2017 International Conference on Computational Science and Computational Intelligence (CSCI), s. 228–233
- Pazner, E.A. i Schmeidler, D., 1978. *Egalitarian equivalent allocations: A new concept of economic equity*. The Quarterly Journal of Economics 92(4), s. 671–687
- Platt, J., 1999. *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods*. Advances in Large Margin Classifiers 10(3), s. 61–74
- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. i Weinberger, K.Q., 2017. *On fairness and calibration*. Advances in Neural Information Processing Systems 30 (NIPS 2017), s. 1–19
- Przanowski, K., 2014. *Credit Scoring w erze Big Data*. Warszawa, SGH Publishing House
- Schechtman, E. i Schechtman, G., 2019. *The relationship between Gini terminology and the ROC curve*. Metron 77(3), s.171–178
- Segall, S., 2013. *Equality and opportunity*. Oxford, Oxford University Press
- Stelmach, A., 1993. *Konstytucja i podstawowe akty prawne*. W: Współczesne instytucje polityczne Polski, Stelmach, A., Zyborowicz, S., Poznań, Instytut Nauk Politycznych i Dziennikarstwa
- Stiglitz, J.E., 1981. *Pareto optimality and competition*. The Journal of Finance 36(2), s. 235–251
- Taleb, N.N., 2014. *Antifragile: Things that gain from disorder*. Bucharest, Random House Trade Paperbacks
- Tallarida, R.J. i Murray, R.B., 2012. *Manual of pharmacologic calculations: with computer programs*. Philadelphia, Springer Science & Business Media
- Thomas, L., Crook, J. i Edelman, D., 2017. *Credit scoring and its applications*. Philadelphia, Society for industrial and Applied Mathematics
- Uzoamaka, O.E., Gerald, N. i Jude, E., 2016. *Women empowerment: Panacea for poverty reduction and economic development in Nigeria*. Journal of Policy and Development Studies 10(2), s. 31–41
- Varian, H.R., 1973. *Equity, envy, and efficiency*. Journal of Economic Theory 9(1), s. 63–91
- Varian, H. R., 2007. *Mikroekonomia Kurs średni ujęcie nowoczesne*. Warszawa, Wydawnictwo Naukowe PWN
- Watkins, E., A., McKenna, M., i Chen, J., *The four-fifths rule is not disparate impact*. FAccT'24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, s. 764–775
- Wilms, R., Mäthner, E., Winnen, L. i Lanwehr, R., 2021. *Omitted variable bias: A threat to estimating causal relationships*. Methods in Psychology 5, s. 1–10
- Wilson, P.H., 2016. *The Holy Roman Empire: a thousand years of Europe's history*. London, Penguin UK
- Yu, X. i Dunn, L.F., 2016. *The Impact of Credit Checks on Employment*. SSRN Electronic Journal 2016, s. 1–33
- Zawadzka, Z., 1996. *Zarządzanie ryzykiem w banku komercyjnym*. Warszawa, Poltext

- Zhang, B.H., Lemoine, B. i Mitchell, M., 2018. *Mitigating unwanted biases with adversarial learning*. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, s. 335–340
- Zhou, N., Zhang, Z., Nair, V.N., Singhal, H., Chen, J. i Sudjianto, A., 2021. *Bias, fairness, and accountability with AI and ML*. International Statistical Review 90(3), s. 468–480

Netografia

- Biuro Informacji Kredytowej, 2021. *Rodzaje kredytów*, bik.pl/poradnik-bik/rodzaje-kredytow [16.03.2024]
- Chandler, C., 2019. *How China's Surveillance State Reflects Black Mirror*, Fortune, fortune.com/2019/03/27/china-surveillance-state-credit-score/ [16.03.2024]
- Desjardin, J., 2018. *How Tech is Changing the Modern Credit Landscape?*, Visual Capitalist. visualcapitalist.com/tech-changing-modern-credit-landscape/ [24.04.2022]
- FDIC. 2019. *Policy statement on discrimination in lending*. fdic.gov/regulations/laws/rules/5000-3860.html [29.04.2024]
- George, A., 2018. *Thwarting bias in AI systems*. College of Engineering at Carnegie Mellon University. engineering.cmu.edu/news-events/news/2018/12/11-datta-proxies.html [30.04.2024]
- Hofmann, H., 1994. *Statlog (German Credit Data)*. UCI Machine Learning Repository. doi.org/10.24432/C5NC77 [13.02.2024]
- Klein, A., 2019. *Credit denial in the age of AI*, brookings.edu/research/credit-denial-in-the-age-of-ai/ [13.02.2023]
- Knight, W., 2019, *The apple card didn't „see” gender—and that's the problem*. wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem [10.06.2024]
- Komisja Nadzoru Finansowego, 2023. *Dane miesięczne sektora bankowego – grudzień 2023*, knf.gov.pl/?articleId=56224&p_id=18 [10.03.2024]
- LaFrance, A., 2014. *Big Data Can Guess Who You Are Based on Your Zip Code*. Atlantic Media Company. theatlantic.com/technology/archive/2014/10/big-data-can-guess-who-you-are-based-on-your-zip-code/381414/ [16.02.2023]
- Lucena, B., 2018. *Spline-based probability calibration*. doi.org/10.48550/arXiv.1809.07751 [12.10.2024]
- MamStartup, 2023. *PKO Bank Polski: Sztuczna inteligencja wspiera zarządzanie ryzykiem kredytowym*. mamstartup.pl/pko-bank-polski-sztuczna-inteligencja-wspiera-zarzadzanie-ryzykiem-kredytowym/ [10.07.2024]
- Narodowy Bank Polski, 2024. *Statystyka stóp procentowych*, nbp.pl/statystyka-i-sprawozdawczosc/statystyka-monetarna-i-finansowa/

statystyka-stop-procentowych/ [16.03.2024]

Online Etymology Dictionary, 2022. etymonline.com/search?q=credit [24.04.2022]

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K.T. i Ghani, R., 2019. *Aequitas: A Bias and Fairness Audit Toolkit*. arxiv.org/abs/1811.05577 [12.10.2024]

Taylor, C., 2020 *Unbiased and Biased Estimators*. ThoughtCo. thoughtco.com/what-is-an-unbiased-estimator-3126502 [01.02.2023]

The Economist, 2012. *Mexican banks - From tequila crisis to sunrise*, economist.com/finance-and-economics/2012/09/22/from-tequila-crisis-to-sunrise [01.01.2024]

The Economist, 2019. *A brief history – and future-of credit scores*, economist.com/international/2019/07/06/a-brief-history-and-future-of-credit-scores [01.01.2024]

Indeks rzeczowy

- absence of envy, 75
- ABT, 54
- ACC, 58
- accuracy, 58
- Adversarial Debiasing, 99
- akcja kredytowa, 35
- algorytm genetyczny, 54
- alternatywne dane, 46
- analiza
 - dyskryminacyjna, 54
 - wpływu regulacji, 137
- analytical base table, 54
- Area Under the ROC Curve, 59
- asymetria informacji, 46
- atrybut prawnie chroniony, 12, 21
- BankAmericard, 36
- bankowość
 - podziemna, 34
 - równoległa, 34
- behawioralna ocena zdolności kredytowej, 44
- bias-aware, 97
- bias-unaware, 97
- biuro informacji kredytowej, 35–37, 44, 46, 51
- brak zazdrości, 75
- błąd
 - drugiego rodzaju, 91
 - pierwszego rodzaju, 91
 - przeżywalności, 73
 - standardowy, 113, 128
- cecha prawnie chroniona, 12, 17, 20, 21, 23, 78, 85, 89
- clean slate decrees, 32
- credit
 - invisible, 47, 85
 - scoring, 52
- czułość, 58
- dane
 - alternatywne, 44
 - aplikacyjne, 44, 45
 - behawioralne, 44, 45
- Days Past Due, 73
- deadweight loss, 75
- decyzja kredytowa, 38
- dekret czystej karty, 32
- demographic parity, 21, 90
- DI, 90
- digital footprint, 73
- disparate
 - impact, 15, 90, 113, 128
 - treatment, 16
- Disparate Impact Remover, 98
- distributive
 - equity, 75
 - justice, 75
- disturbance plot, 120, 136
- dobry klient, 53
- dokładność, 58
- DPD, 73

drzewa klasyfikacyjne, 54
 dyskryminacja
 bezpośrednia, 12, 16, 22, 67, 71, 72, 76, 81, 89–91, 93, 142
 pośrednia, 12, 15, 19–21, 67, 68, 72, 76, 81, 89, 90, 93, 139
 w dostępie, 76
 w ocenie, 76

 end-state fairness, 75
 equalized odds condition, 22, 90

 FACT, 21, 89
 Fair Aware Confussion Table, 21, 89
 fairness
 in outcomes, 75
 through blindness, 20, 96
 through unawareness, 20, 96
 fall-out, 58
 false
 negative, 58
 negative rate, 58, 91
 positive, 57
 positive rate, 58, 59, 91
 faza
 1 MSSF 9, 65
 2 MSSF 9, 65
 3 MSSF 9, 65
 fałszywie
 negatywne, 58
 pozytywne, 57, 72
 FICO score, 38
 FN, 58
 FNR, 58
 formalna równość szans, 71
 four-fifths rule, 69
 FP, 57
 FPR, 58

 Grid Search Reduction, 99

 grupa
 nieuprzywilejowana, 21, 73, 89
 uprzywilejowana, 16, 21, 71, 89

 in-processing, 23, 97
 independence, 21, 90
 Internal Ratings-Based, 56
 IRB, 56

 jakość
 kalibracji, 22, 91
 prognostyczna modelu, 16, 21, 23, 71, 89

 karta kredytowa, 36, 40
 KNF, 39
 konsekwencje społeczne, 13
 konsument, 40
 kredyt, 31, 32
 hipoteczny, 31, 40
 inwestycyjny, 40
 konsolidacyjny, 32, 40
 konsumpcyjny, 31, 40
 mieszkaniowy, 40
 społeczny, 53
 kryterium
 dostateczności, 91, 117, 131
 niezależności, 21, 90, 93
 separacji, 22, 90, 115, 129
 krzywa ROC, 59

 lenistwo modelu, 16, 71
 lichwa, 34, 35
 loan-to-value, 45
 LTV, 45

 macierz pomyłek, 21, 57, 89
 Mastercard, 36
 metoda wektorów nośnych, 54
 metody
 baggingowe i boostingowe, 54

potreningowe, 23, 97, 100
 przedtreningowe, 23, 97, 98
 treningowe, 23, 97, 99
 miara stronniczości algorytmicznej, 11,
 21, 22, 89
 miss rate, 58
 moc
 dyskryminacyjna, 22, 91
 dyskryminacyjna modelu, 109
 model
 dynamiczny, 27
 laziness, 16, 71
 logiki rozmytej, 54
 oceny zdolności kredytowej, 14, 38,
 50, 52, 53, 68
 statyczny, 27, 103, 104, 135
 monitoring modelu, 56
 MSSF 9, 40
 natural justice, 75
 ocena
 ekspercka, 13, 14, 38, 50
 wiarygodności kredytowej, 50
 zdolności formalnoprawnej, 50
 zdolności kredytowej, 39, 50
 zdolności prawnej, 50
 odmienne traktowanie, 16, 22, 72, 89–91
 odmienny wpływ, 15–17, 20, 68, 69, 89,
 90, 139
 odsetek
 fałszywie negatywnych, 91
 fałszywie pozytywnych, 91
 omitted variable problem, 78
 parytet
 demograficzny, 90
 kalibracji, 91
 permutatio, 33
 perspektywa decyzyjna regulatora, 12,
 23, 24, 95, 101
 podejście nieograniczone, 19, 23, 95
 positive predictive value, 22, 59, 92
 post-processing, 23, 97
 pożyczka, 31
 PPV, 22, 59, 92
 prawdziwie
 negatywne, 57
 pozytywne, 57
 prawo Hammurabiego, 32
 pre-processing, 97
 pre-training, 23
 precision, 59
 precyzja, 59
 Prejudice Remover , 99
 prima facie, 68
 problem
 decyzyjny banku, 12, 23, 95, 101
 free-rider, 75
 współliniowości, 87
 zmiennej pominiętej, 20, 21, 78, 87,
 96
 procedural fairness, 75
 proces kredytowy, 37, 38
 prohibited grounds, 85
 protected
 class, 85
 group, 85
 próg odcięcia decyzji kredytowej, 59
 płatność bezgotówkowa, 40
 radykalna równość szans, 72
 raport kredytowy, 36
 Receiver Operating Characteristic, 59
 redukcja stronniczości algorytmicznej, 20,
 23, 96, 97, 135, 159
 regresja
 liniowa, 54

- logistyczna, 54
- Reject Option Classifier, 100
- Reweighing, 98
- ROC, 109
- rolowanie kredytu, 39
- rysunek
 - rozbieżności, 120, 136
 - rozbieżności prognoz, 120, 136, 162
- ryzyko
 - koncentracji, 62
 - kredytowe, 64
- rzeczywista równość szans, 71
- równość szans, 71
- sensitive features, 85
- sensitivity, 58
- separation, 22, 90
- sieci
 - bayesowskie, 54
 - neuronowe, 54
- skrzywienie algorytmiczne, 15, 67
- specificity, 58
- sprawiedliwość
 - dystributywna, 75
 - działania algorytmów, 15
 - proceduralna, 75
 - przez nieświadomość, 12, 19, 20, 23, 95, 96
 - przez świadomość, 23, 97
 - wyników, 75
- spółka handlowa, 35
- standard error, 113, 128
- statistical parity, 21, 90
- stronniczość algorytmiczna, 11, 15, 49, 67, 68, 87, 90, 95, 97, 135, 159
- sufficiency, 22, 91
- Sumerowie, 32
- survivorship bias, 73
- swoistość, 58
- śląd cyfrowy, 73
- test
 - Fishera, 69, 70, 90
 - Hosmera–Lemeshowa, 60
- TN, 57
- TNR, 58
- TP, 57
- TPR, 58
- transakcja papierowa, 33
- trenowanie modelu, 14
- true
 - negative, 57
 - negative rate, 58
 - positive, 57
 - positive rate, 58, 59
- umowa kredytowa, 31, 39, 64
- unmitigated, 20, 95
- wartości predykcyjnej dodatniej, 22
- warunek
 - parytetu demograficznego, 89, 93, 109, 113, 127, 128
 - parytetu kalibracji, 22, 91, 93, 117, 130
 - wyrównanych szans, 22, 90, 115, 129
- wniosek kredytowy, 45
- wskaźnik
 - fałszywie dodatnich, 58
 - fałszywie negatywnych, 58
 - zatwierdzeń, 16, 17, 68, 72
- wymiar
 - materialny, 83
 - proceduralny, 83
- XAI, 55
- zasada czterech–piątych, 69, 90
- zbędna strata społeczna, 75
- zdolność kredytowa, 14, 53, 64
- zmienna prawnie chroniona, 12, 20
- zły klient, 53, 56

Spis tabel

1	Zakres cech prawnie chronionych zgodnie z wybranymi przepisami prawa .	19
2	Mapowanie definicji dyskryminacji na miary stronniczości algorytmicznej .	22
3	Poglądowa struktura macierzy pomyłek.	58
4	Zestawienie udzielonych kredytów w Polsce, stan na koniec 2023 r.	65
5	Test Fishera – macierz rozkładu decyzji kredytowych, ze względu na zmienną prawnie chronioną.	70
6	Macierz pomyłek – perspektywa obserwatora, banku i klienta	92
7	Macierz kosztu dla syntetycznej miary Zysk	110
8	Macierz kosztu dla fałszywie negatywnego przypadku	111
9	Macierz kosztu dla prawdziwie pozytywnego przypadku	111
10	Wartości Zysk oraz Disparate Impact dla Scenariusz Syntetyczny . . .	114
11	Zestawienie wartości miary Zysk oraz kryterium separacji dla Scenariusz Syntetyczny	117
12	Zestawienie wartości miary Zysk oraz kryterium dostateczności dla Scenariusz Syntetyczny	119
13	Średni poziom rozbieżności prognozy dla wszystkich poziomów akceptacji kredytowej	127
14	Wartości Zysk oraz Disparate Impact dla Scenariusz German Credit Data	130
15	Zestawienie wartości miary Zysk oraz kryterium separacji dla Scenariusz German Credit Data	130
16	Zestawienie wartości miary Zysk oraz kryterium dostateczności dla Scenariusz German Credit Data	132
17	Średni poziom rozbieżności prognozy dla wszystkich poziomów akceptacji kredytowej	134
18	Scenariusz German Credit Data – zmienne egzogeniczne	166

Spis rysunków

1	Przykład dyskryminacji pośredniej i bezpośredniej w ocenie zdolności kredytowej	17
2	Relacje między regulatorem, bankiem i konsumentem	24
3	Wielkość rynku kredytów sektora niefinansowego w Polsce	40
4	Wielkość rynku kredytów w Polsce w podziale na klienta	41
5	Średni poziom stóp procentowych kredytów w Polsce	42
6	Etapy procesu kredytowego w banku	43
7	Poglądowa zależność zysku banku i poziomu akceptacji kredytowej w zależności od jakości modelu oceny zdolności kredytowej.	63
8	Przykład odmiennego wpływu i odmiennego traktowania przez model oceny zdolności kredytowej	72
9	Proces budowy i wykorzystania modelu oceny zdolności kredytowej	95
10	Umieszczenie metody redukcji stroniczości algorytmicznej w procesie budowy i wykorzystania modelu decyzyjnego	98
11	Schemat realizacji eksperymentów numerycznych oceny bezpośrednich konsekwencji	105
12	Średnie wartości indeksu Giniego, a warunek parytetu demograficznego, Scenariusz Syntetyczny	110
13	Miara Zysk a warunek parytetu demograficznego, Scenariusz Syntetyczny	112
14	Miara Zysk a warunek wyrównanych szans, Scenariusz Syntetyczny . . .	116
15	Miara Zysk a warunek parytetu kalibracji, Scenariusz Syntetyczny . . .	118
16	Rozbieżności prognoz, przykład bez wskazania żadnej z metod ograniczania stroniczości algorytmicznej	121
17	Średnie krzywe rozbieżności, Scenariusz Syntetyczny, model regresji logistycznej LR.	122
18	Pojedyncza krzywa rozbieżności, Scenariusz Syntetyczny, model drzewa klasyfikacyjnego DT.	123
19	Średnie krzywe rozbieżności, Scenariusz Syntetyczny, model drzewa klasyfikacyjnego DT.	124
20	Średnie krzywe rozbieżności, Scenariusz Syntetyczny, model lasu losowego RF.	125

21	Średnie krzywe rozbieżności, Scenariusz Syntetyczny, model gradient boosting GB.	126
22	Wartość indeksu Giniego, a warunek parytetu demograficznego, Scenariusz German Credit Data	128
23	Miara Zysk a warunek parytetu demograficznego, Scenariusz German Credit Data	129
24	Miara Zysk a warunek wyrównanych szans, Scenariusz German Credit Data	131
25	Miara Zysk a warunek parytetu kalibracji, Scenariusz German Credit Data	132
26	Średnie krzywe rozbieżności, Scenariusz Syntetyczny, model regresji logistycznej LR.	133
27	Schemat realizacji eksperymentów numerycznych oceny długoterminowej .	140
28	Kryteria oceny stronniczości algorytmicznej, Scenariusz Syntetyczny, model dynamiczny	141
29	Analiza dynamiczna wpływu na rozkład parametrów akceptowanych kredytów, metoda bazowa: Unmitigated	143
30	Analiza długoterminowa wpływu na jakość i wolumen portfela kredytowego, metoda bazowa: Unmitigated	144
31	Analiza długoterminowa wpływu na zysk, metoda bazowa: Unmitigated .	145
32	Analiza długoterminowa wpływu na miarę Disparate Impact, metoda bazowej: Fairness through Unawareness	146
33	Analiza dynamiczna wpływu na rozkład parametrów akceptowanych kredytów, metoda bazowej: Fairness through Unawareness	147
34	Analiza długoterminowa wpływu na jakość portfela kredytowego, metoda bazowej: Fairness through Unawareness	148
35	Analiza długoterminowa wpływu na zysk, metoda bazowej: Fairness through Unawareness	149
36	Triada redukcji stronniczości algorytmicznej	151
37	Izolinie wartości Disparate Impact w zależności od wskaźnik zatwierdzeń kredytów oraz jakości portfela kredytowego	152