Problem stronniczości algorytmicznej: metody wspomagania decyzji

Daniel Kaszyński
Szkoła Główna Handlowa w Warszawie, Kolegium Analiz Ekonomicznych

1. Cel badawczy

Celem badania jest opracowanie metod wspomagania dwóch typów decydentów w obszarze przepisów równościowych dotyczących modeli oceny zdolności kredytowej konsumentów: **(i) regulatora rynku kredytów konsumenckich** - w zakresie projektowania regulacji równościowych oraz oceny ich konsekwencji społecznych i ekonomicznych, oraz **(ii) banków** - w zakresie konstrukcji modeli oceny zdolności kredytowej redukujących problem stronniczości algorytmicznej w sposób ekonomicznie efektywny.

Punktem wyjścia badania jest obserwacja, że obecnie dominującym podejściem banków do zapewnienia braku dyskryminacji jest *sprawiedliwość przez nieświadomość* (*fairness through unawareness*), czyli usunięcie cechy prawnie chronionej (np. płeć) ze zbioru zmiennych modelu. Podejście to, mimo że prawnie ugruntowane w aktach takich jak ECOA-B czy opinia EDPS 11/2021, jest źródłem klasycznego problemu zmiennej pominiętej i - jak wykazuje literatura - może być zarówno ekonomicznie nieefektywne, jak i jedynie powierzchownie rozwiązujące problem dyskryminacji.

2. Główne hipotezy badawcze

Sformułowano sześć hipotez podzielonych na dwa bloki - statyczny (H1-H3, dotyczący jednorazowej decyzji modelowej banku) oraz dynamiczny (H4-H6, dotyczący wieloletnich konsekwencji regulacji):

H1. Metoda sprawiedliwości przez nieświadomość nie jest bezwarunkowo ekonomicznie efektywna; istnieją metody klasy *sprawiedliwości przez świadomość* (*fairness through awareness*) generujące wyższy zysk banku przy spełnieniu tego samego warunku równościowego.

H2. Skuteczność (w sensie ekonomicznym) poszczególnych metod redukcji stronniczości algorytmicznej zależy od miary stronniczości narzuconej przez regulatora - wybór miary determinuje optymalne podejście banku.

H3. Możliwe jest opracowanie procedury redukcji stronniczości algorytmicznej maksymalizującej zysk banku, ujętej w jednolitych ramach pomiaru efektywności metod przed-, śród- i potreningowych.

H4. Możliwe jest opracowanie metodyki oceny długookresowych konsekwencji regulacji równościowych z wykorzystaniem techniki symulacji dynamicznej.

H5. Wprowadzenie regulacji równościowych powoduje krótkookresową niestabilność rozkładu wydawanych decyzji kredytowych, co uzasadnia projektowanie okresu dostosowawczego.

H6. Wprowadzenie regulacji równościowych prowadzi do kosztu ekonomicznego wynikającego ze zmiany populacji osób otrzymujących pozytywne decyzje kredytowe - manifestującego się przez zmianę jakości portfela kredytowego, zmianę współczynnika akceptacji wniosków lub kombinację obu efektów.

3. Metody weryfikacji hipotez

Wykorzystano dwa autorskie modele symulacyjne oraz krytyczny przegląd literatury jako główne metody badawcze.

Model statyczny (weryfikacja H1-H3) odtwarza jednorazowy proces budowy modelu oceny zdolności kredytowej dla zadanej metody redukcji stronniczości oraz raportuje miary jakości prognostycznej (współczynnik Giniego), wybrane miary stronniczości algorytmicznej (parytet demograficzny, wyrównane szanse, parytet kalibracji) oraz zysk banku zgodny z macierzą kosztów zaproponowaną przez Kozodoi i Lessmann (2022). Stanowi rozwinięcie podejścia tych autorów o oszacowanie przedziałów ufności metodą Monte Carlo oraz o autorski *rysunek rozbieżności prognoz* służący do oceny skali zmian decyzyjnych względem modelu bazowego.

Model dynamiczny (weryfikacja H4-H6) odtwarza wieloletni proces kredytowy banku: w każdej iteracji bank trenuje model na danych klientów, którzy w poprzedniej iteracji uzyskali pozytywną decyzję kredytową (uwzględniając problem braku obserwacji zmiennej celu dla odrzuconych aplikacji), wydaje decyzje, obserwuje spłacalność i przekazuje wynik do kolejnej iteracji. Stanowi to autorski wkład w stosunku do literatury przedmiotu, w której modele tego typu były analizowane wyłącznie w ujęciu jednookresowym.

Dane. Weryfikację przeprowadzono dwutorowo: na zbiorze syntetycznym z parametryzowaną nieliniowością, współliniowością i niezbilansowaniem względem

zmiennej chronionej (n = 2 000 000 obserwacji, regresja logistyczna z członem kwadratowym) oraz na rzeczywistym zbiorze German Credit Data. Symulacje Monte Carlo umożliwiają oszacowanie wartości średnich i przedziałów ufności wyników.

4. Związki ze światowymi nurtami badań

Badanie wpisuje się w trzy aktywnie rozwijane nurty międzynarodowej literatury naukowej.

Algorithmic fairness w finansach. Praca rozwija dorobek Kozodoi i Lessmann (2022), Hurlin, Pérignon i Saurin (2022), Bono, Croxson i Giles (2021) oraz Hurley i Adebayo (2016), korzystając z taksonomii metod redukcji stronniczości (przed-, śród-, potreningowe) zaproponowanej przez Friedler i in. (2019) oraz z formalizmu Fair Aware Confusion Table (FACT) Kim, Chen i Talwalkar (2020). Podstawą teoretyczną pojęć dyskryminacji pośredniej i bezpośredniej oraz miar parytetu demograficznego, wyrównanych szans i parytetu kalibracji jest monografia Barocas, Hardt i Narayanan (2023). Wkład badania w stosunku do wymienionych prac dotyczy w szczególności: (i) rozszerzenia analizy o przedziały ufności wyników, (ii) wprowadzenia ujęcia wielookresowego oraz (iii) jednolitego porównania trzech klas metod redukcji stronniczości względem ekonomicznego kryterium zysku banku.

Empiryczne dowody dyskryminacji w modelach kredytowych. Praca odnosi się do udokumentowanej w literaturze statystycznej istotności cech prawnie chronionych w modelach kredytowych - płci (Andreeva i Matuszyk 2019; Agarwal i in. 2018), pochodzenia etnicznego (Butler, Mayer i Weston 2023; Asiedu, Freeman i Nti-Addae 2012) oraz wieku konsumenta (Dumitrescu i in. 2022).

Regulacja AI i systemów decyzyjnych wysokiego ryzyka. Praca odnosi się do rozporządzenia 2024/1689 (AI Act), które klasyfikuje systemy oceny zdolności kredytowej jako systemy wysokiego ryzyka, oraz do regulacji branżowych: ECOA i ECOA-B, FHA, RODO, ustawy 2010/12/03 oraz opinii EDPS 11/2021. Wprowadza do tej dyskusji ekonomiczny rachunek konsekwencji wyboru miary stronniczości - element pomijany w dotychczasowej debacie prawno-regulacyjnej.

5. Uzyskane rezultaty

Wszystkie sześć hipotez zostało zweryfikowane pozytywnie. Najważniejsze ustalenia przedstawiono poniżej.

Dla scenariusza syntetycznego sprawiedliwość przez nieświadomość prowadzi do istotnego spadku zysku banku przy jednoczesnym *braku* spełnienia warunku braku dyskryminacji pośredniej lub bezpośredniej - tj. nie tylko jest droga, ale i nieskuteczna w swoim deklarowanym celu. Dla danych German Credit jedynie w przypadku warunku wyrównanych szans podejście to było uzasadnione ekonomicznie; dla warunku parytetu demograficznego oraz warunku parytetu kalibracji istniały podejścia efektywniejsze (H1).

Wybór miary stronniczości algorytmicznej przez regulatora determinuje, która klasa metod (przed-, śród- czy potreningowa) jest ekonomicznie optymalna dla banku - w przebadanych scenariuszach każda z klas była optymalna w co najmniej jednej konfiguracji miary i zbioru danych (H2). Opracowana procedura statyczna pozwala bankowi wskazać podejście maksymalizujące zysk dla zadanej miary stronniczości i zbioru danych (H3).

W ujęciu dynamicznym wykazano krótkookresową niestabilność rozkładu decyzji kredytowych po wdrożeniu regulacji. W jednym z badanych scenariuszy kierunek krótkookresowych zmian średniego dochodu i wieku populacji akceptowanej był *przeciwny* do kierunku zmian długookresowych - wynik nieoczywisty z perspektywy analizy statycznej. Dla podejścia wyjściowego *Fairness through Unawareness* zaobserwowano dodatkowo zjawisko stabilizacji krokowej trwającej kilkanaście iteracji (H4, H5).

Wykazano istnienie krzywej Pareto-efektywnej pomiędzy współczynnikiem akceptacji wniosków a jakością portfela kredytowego przy zadanym poziomie ograniczenia stronniczości algorytmicznej - co stanowi formalne ujęcie kompromisu, przed którym staje bank po wprowadzeniu regulacji (H6).

Dodatkowymi rezultatami rozprawy są: autorska miara zaburzenia struktury populacji kredytobiorców oraz *rysunek rozbieżności prognoz* - narzędzia użyteczne w analizie decyzji banku o zmianie modelu oceny zdolności kredytowej.

6. Nierozwiązane problemy badawcze

Z przeprowadzonego badania wyłaniają się cztery kierunki dalszych prac.

Po pierwsze, długoterminowa konkurencyjność banków w warunkach regulacji równościowych: otwartym pytaniem pozostaje, czy w długim okresie banki o różnych funkcjach kosztów dochodzą do podobnych portfeli kredytowych, czy też regulacja prowadzi do trwałego rozejścia się strategii kredytowych.

Po drugie, wpływ własności zmiennych objaśniających (nieliniowość, interakcje, współliniowość) na poziom stronniczości algorytmicznej modelu. Punktem wyjścia może być analiza wariancji parametrów modelu liniowego pod wpływem współliniowości przedstawiona przez O'Briena (2007), rozszerzona o miary stronniczości algorytmicznej.

Po trzecie, weryfikacja zaproponowanych metodyk na innych zbiorach danych - zarówno rzeczywistych (poza German Credit Data), jak i syntetycznych z innych generatorów (np. Przanowski 2014). Pozwoliłoby to ocenić, na ile uzyskane wyniki są zależne od specyfiki danych.

Po czwarte, rozwój metod wspomagania decyzji banku o zmianie modelu oceny zdolności kredytowej - w szczególności podejść opartych na tabelach częstości, uwzględniających rozbieżność decyzji generowaną przez przejście na nowy model. Jest to istotne, ponieważ koszt wdrożenia zmiany w banku jest funkcją skali rozbieżności decyzji, a nie wyłącznie miar agregatywnych jakości modelu.

7. Obszary i możliwości zastosowań w praktyce gospodarczej

Wyniki badań są bezpośrednio aplikowalne w trzech obszarach praktyki gospodarczej.

Banki - proces budowy modelu kredytowego. Opracowana metodyka pozwala porównać metody przed-, śród- i potreningowe redukcji stronniczości algorytmicznej w ramach jednolitej miary efektywności (zysk banku przy ograniczeniu miarą stronniczości), wskazać optymalne podejście dla danego portfela oraz oszacować koszt zmiany modelu w stosunku do aktualnie wykorzystywanego (rysunek rozbieżności prognoz). Stanowi to alternatywę dla obecnie stosowanego, suboptymalnego podejścia *fairness through unawareness*.

Regulatorzy rynków kredytowych (KNF, EBA, organy nadzoru państw członkowskich UE). Model dynamiczny umożliwia oszacowanie *ex ante* długookresowych konsekwencji projektowanej regulacji - w tym wpływu na strukturę populacji kredytobiorców, jakość portfeli sektora bankowego oraz poziom akcji kredytowej. Wynik dotyczący krótkookresowej niestabilności i krokowej stabilizacji stanowi argument za wprowadzaniem okresu dostosowawczego w przepisach równościowych, analogicznie do okresów przejściowych funkcjonujących w pakiecie Basel III/IV.

Inne modele decyzyjne wysokiego ryzyka w rozumieniu AI Act. Metodyka jest przenośna na inne klasyfikatory binarne wykorzystywane w decyzjach mających wpływ na sytuację osób fizycznych: scoring ubezpieczeniowy, modele wykrywania nadużyć

finansowych, systemy oceny w procesach rekrutacyjnych. We wszystkich tych obszarach występuje analogiczny problem wyboru miary sprawiedliwości oraz kompromisu między jakością prognostyczną a wymogami równościowymi.