Modele Trustworthy AI w skoringu kredytowym

Daniel Kaszyński, Małgorzata Wrzosek
Szkoła Główna Handlowa w Warszawie

Plan prezentacji

1. Skoring kredytowy
2. Metody uczenia maszynowego
3. Eksperyment numeryczny
4. Zagadnienia sprawiedliwości w modelach decyzyjnych

Ocena zdolności kredytowej

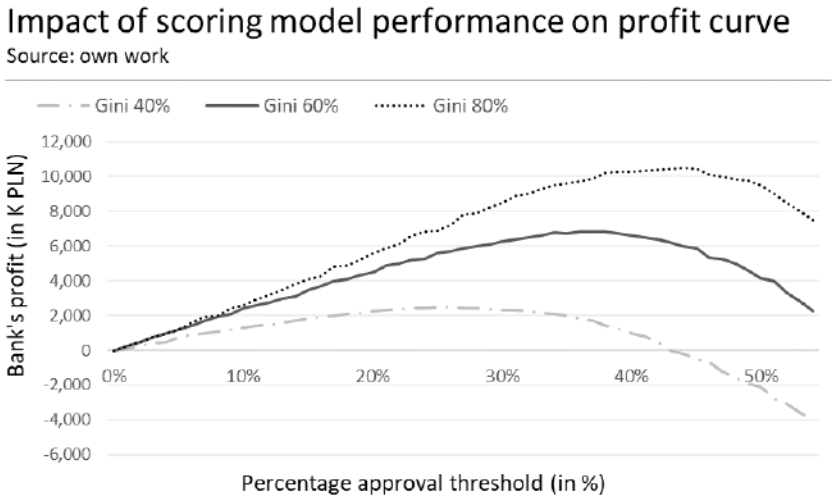
- **Wymogi regulacyjne:** „Bank uzależnia przyznanie kredytu od zdolności kredytowej kredytobiorcy. Przez zdolność kredytową rozumie się zdolność do spłaty zaciągniętego kredytu wraz z odsetkami w terminach określonych w umowie. Kredytobiorca jest obowiązany przedłożyć na żądanie banku dokumenty i informacje niezbędne do dokonania oceny tej zdolności" [Prawo Bankowe, 1997]

- **Modele skoringowe w bankowości:** modele skoringowe są wykorzystywane nie tylko na potrzeby akceptacji wniosków kredytowych, ale również użytkowane są na potrzeby księgowości, tj. standard MSSF 9 [MSSF 9, 2016], lub wyznaczania wymogów kapitałowych (wymogi komitetu Bazylejskiego).

- **Generyczność metod skoringowych (horyzontalny aspekt):** Skoring kredytowy jest wykorzystywany również tam, gdzie występuje tzw. kredyt kupiecki, tj odroczona płatność za produkty/usługi.

Ocena zdolności kredytowej

- **Długa historia wykorzystania metod skoringowych (wertykalny aspekt)**: pierwsze wspomnienia w literaturze (1820 r.), pierwsze komercyjne rozwiązanie firmy FICO (1950 r.) [Kaszynski et al., 2020]
- **Korzyść z dobrych modeli skoringowych** – z perspektywy instytucji (banku), lepsze modele skoringowe oznaczają: a) możliwość głębszej akceptacji kredytów, oraz b) uzyskiwać wyższe zyski z działalności gospodarczej.

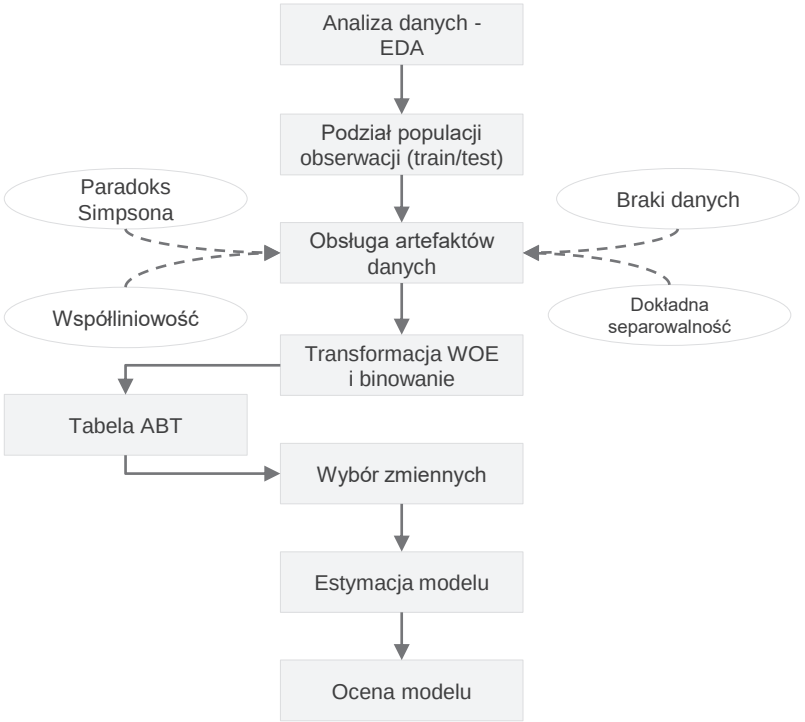


Klasyczny skoringu kredytowy

Jak **obecnie wygląda** przygotowanie modelu skoringowego?

Klasyczny skoring (regresja logistyczna) zakłada wieloetapowy proces analityki danych:

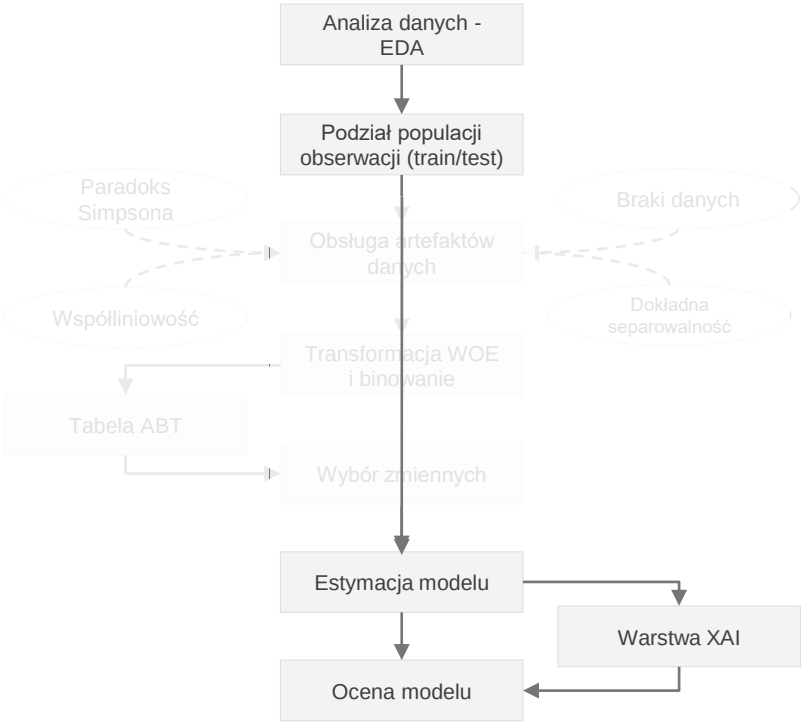
- obsługa artefaktów danych,
- transformację WOE,
- binowanie zmiennych,
- po wybór zmiennych i estymację modelu.



Współczesny skoring kredytowy

Jak **może wyglądać** proces przygotowania modelu skoringowego?

We współczesnym skoringu, proces analityczny może zostać istotnie skrócony. Koniecznym jest jednak uwzględnienie Warstwy XAI – wytłumaczalności modelu analitycznego.



Argumenty za i przeciw metodom ML

✓ Lepsza elastyczność (nieliniowość), i większe dopasowanie do danych (mniejsze błędy prognoz)

✓ Część artefaktów związanych z danymi jest automatycznie obsługiwanych (np. braki danych *obserwacje odstające,* lub *współliniowość*)

✓ *Feature engineering* jest przeprowadzany automatycznie (tj. nieliniowe zależności, interakcje między zmiennymi)

✗ Tracimy jednak prostą interpretowalność i audytowalność modeli

✗ Często dużo większa złożoność obliczeniowa

✗ Możliwe problemy z przeuczeniem modeli (potencjalnie utrata stabilności)

Argumenty za i przeciw metodom ML

✓ Lepsza elastyczność (nieliniowość), i większe dopasowanie do danych (mniejsze błędy prognoz)

✓ Część artefaktów związanych z danymi jest automatycznie obsługiwanych (np. braki danych *obserwacje odstające,* lub *współliniowość*)

✓ *Feature engineering* jest przeprowadzany automatycznie (tj. nieliniowe zależności, interakcje między zmiennymi)

✗ Tracimy jednak prostą interpretowalność i audytowalność modeli

✗ Często dużo większa złożoność obliczeniowa

✗ Możliwe problemy z przeuczeniem modeli (potencjalnie utrata stabilności)

#ZWIADoAI

Przewodnia i nadzorcza rola człowieka

W tym prawa podstawowe, przewodnia i nadzorcza rola człowieka

Techniczna solidność i bezpieczeństwo

W tym odporność na atak i bezpieczeństwo, plan rezerwowy i ogólne bezpieczeństwo, dokładność, wiarygodność i odtwarzalność

Ochrona prywatności i zarządzanie danymi

W tym poszanowanie prywatności, jakość i integralność danych oraz dostęp do danych

Przejrzystość

W tym identyfikowalność, wytłumaczalność i komunikacja

Różnorodność, niedyskryminacja i sprawiedliwość

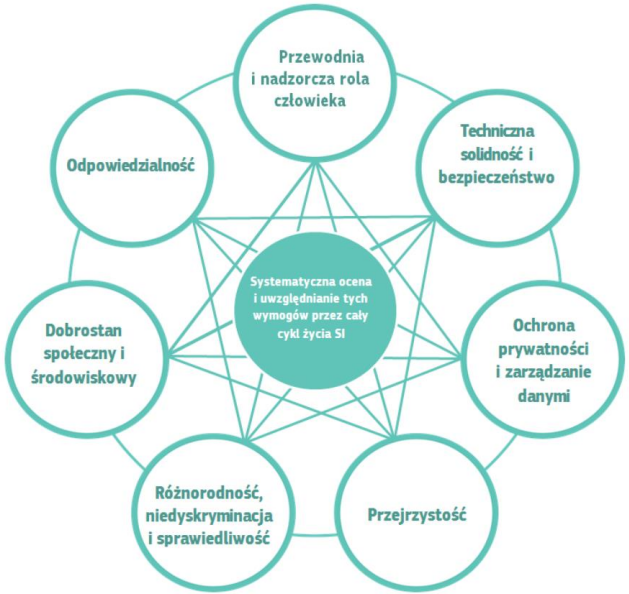
W tym unikanie niesprawiedliwej stronniczości, dostępność i zasada „projektowanie dla wszystkich" oraz udział zainteresowanych stron

Dobrostan społeczny i środowiskowy

W tym zrównoważony charakter i przyjazne podejście wobec środowiska, skutki społeczne, społeczeństwo i demokracja

Odpowiedzialność

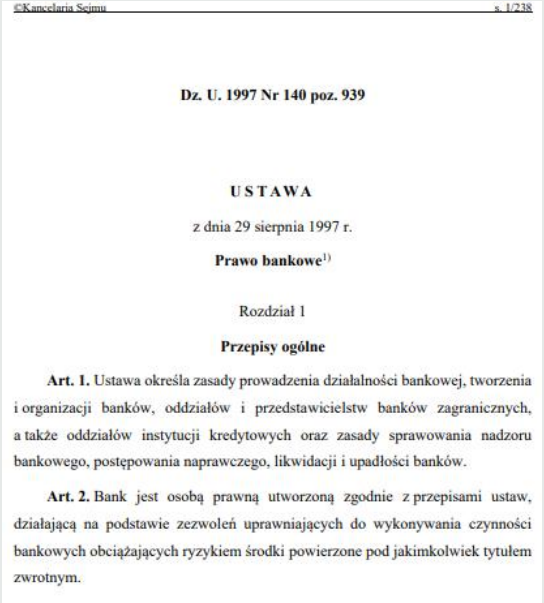
W tym możliwość kontrolowania, minimalizacja i zgłaszanie negatywnych skutków, kompromisy i dochodzenie roszczeń



*Źródło: **Komisja Europejska**, 2019, Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji*

Metody XAI

Dlaczego potrzebny jest XAI?



©Kancelaria Sejmu s. 1/238

Dz. U. 1997 Nr 140 poz. 939

USTAWA

z dnia 29 sierpnia 1997 r.

Prawo bankowe[1)]

Rozdział 1

Przepisy ogólne

Art. 1. Ustawa określa zasady prowadzenia działalności bankowej, tworzenia i organizacji banków, oddziałów i przedstawicielstw banków zagranicznych, a także oddziałów instytucji kredytowych oraz zasady sprawowania nadzoru bankowego, postępowania naprawczego, likwidacji i upadłości banków.

Art. 2. Bank jest osobą prawną utworzoną zgodnie z przepisami ustaw, działającą na podstawie zezwoleń uprawniających do wykonywania czynności bankowych obciążających ryzykiem środki powierzone pod jakimkolwiek tytułem zwrotnym.

Art. 70. 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.

Art. 70a. 1. 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**.
2. 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.

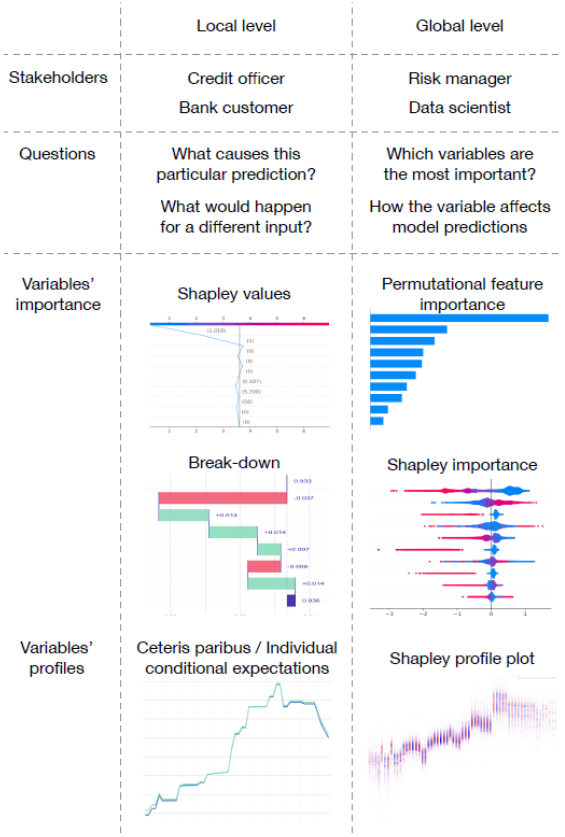
Art. 105a. 1. […] mogą w celu oceny zdolności kredytowej i analizy ryzyka kredytowego podejmować decyzje, **opierając się wyłącznie na zautomatyzowanym przetwarzaniu**, w tym profilowaniu, danych osobowych – również stanowiących tajemnicę bankową […]

Metody XAI

Jaki problem chcemy rozwiązać?

Wyróżnia się następujące klasyfikacje metod XAI:

- **Metody lokalne/globalne** – metody lokalne umożliwiają wytłumaczenie wyniku modelu dla pojedynczej obserwacji
- **Metody specyficzne/agnostyczne** – modele agnostyczne nie zakładają żadnych specyficznych własności modeli – mogą być wykorzystywane dla wszystkich modeli prognostycznych
- **Metody wewnętrzne/post-hoc** – metody wewnętrzne związane są z postacią modelu (np. wrażliwość modelu regresji logistycznej jest jawna)



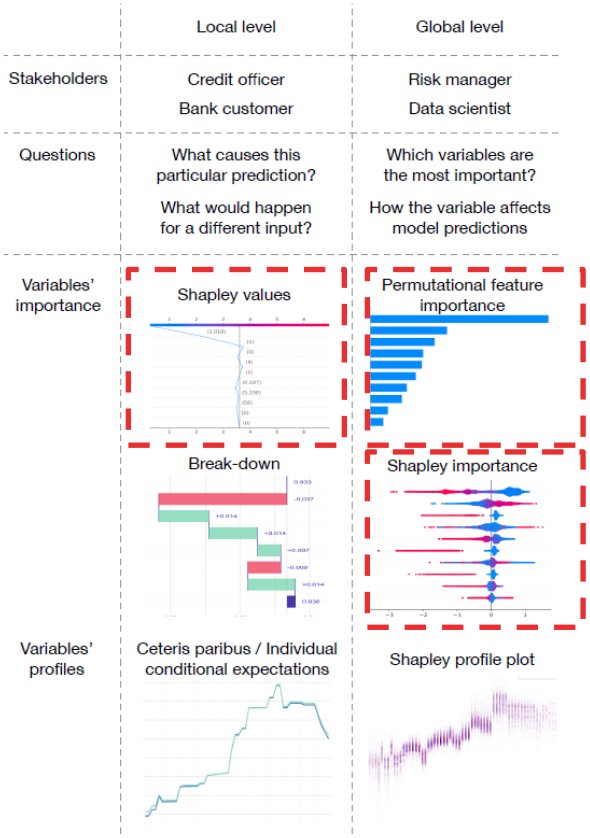
Źródło: D. Kaszyński, B. Kamiński, T. Szapiro, "Credit Scoring in Context of Interpretable Machine Learning", SGH Publishing House, 2020

Metody XAI

Jaki problem chcemy rozwiązać?

Wyróżnia się następujące klasyfikacje metod XAI:

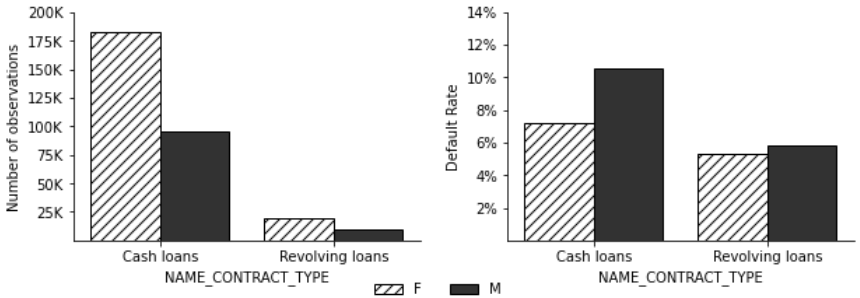
- **Metody lokalne/globalne** – metody lokalne umożliwiają wytłumaczenie wyniku modelu dla pojedynczej obserwacji
- **Metody specyficzne/agnostyczne** – modele agnostyczne nie zakładają żadnych specyficznych własności modeli – mogą być wykorzystywane dla wszystkich modeli prognostycznych
- **Metody wewnętrzne/post-hoc** – metody wewnętrzne związane są z postacią modelu (np. wrażliwość modelu regresji logistycznej jest jawna)



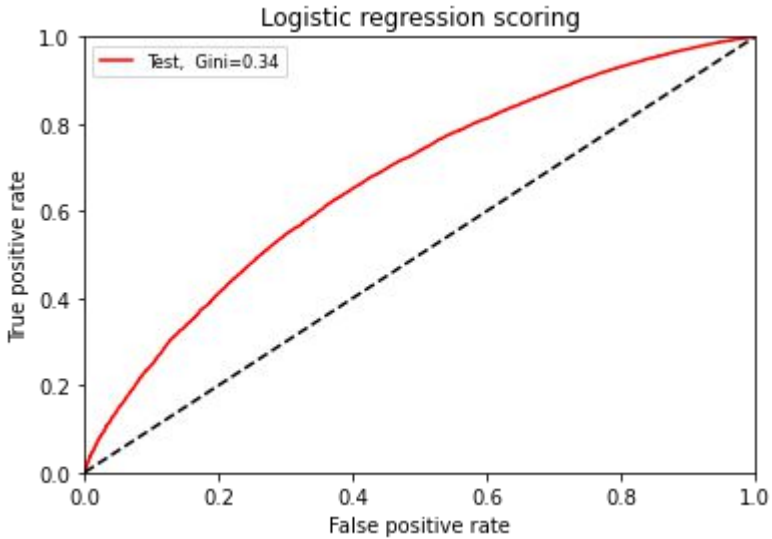
Źródło: D. Kaszyński, B. Kamiński, T. Szapiro, "Credit Scoring in Context of Interpretable Machine Learning", SGH Publishing House, 2020

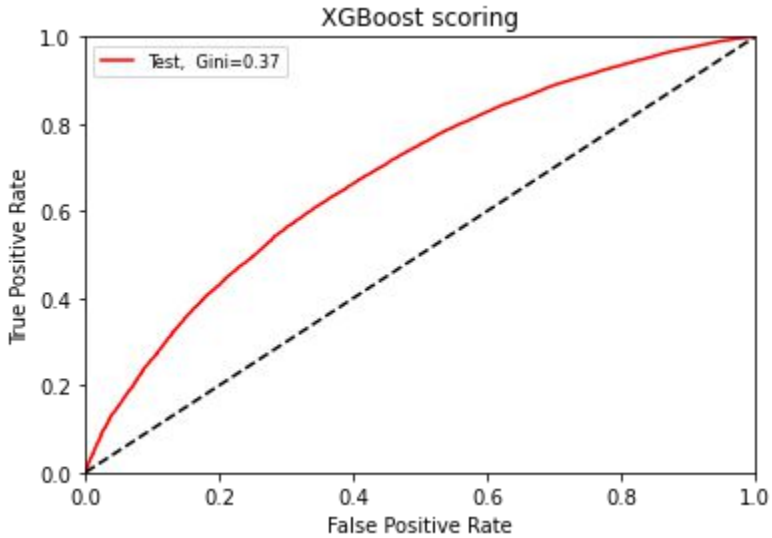
Eksperyment numeryczny

- **Dane** - Home Credit Default Risk, https://www.kaggle.com/c/home-credit-default-risk
- **Model klasyczny** – Regresja Logistyczna
- **Model Machine Learning** – XGBoost



Eksperyment numeryczny





Metody XAI – Shapley values

- Wartości Shapleya wyprowadzone zostały z **teorii gier** (interakcji i akcji w grach kooperacyjnych).
- Interpretacja wartości Shapleya dla zmiennej j jest następująca: wartość zmiennej j wpływa na poziomie ϕ_j na wartość predykcji w przypadku danej obserwacji, w porównaniu do średniej predykcji na zbiorze danym.

Algorithm 1: Shapley values $\phi_j(x)$ for a single predictor (x).

for $1, \ldots,$ m **do**

Pick a random observation w from the data set;

Pick a random permutation $O \in \pi(n)$, where $\pi(n)$ is set of ordered permutations of the feature indexes;

Create two new instances b_1 and b_2 based on x and w:

- reorder features in x and w according to indices in O: $x' = (x_1, x_2, \ldots, x_j, \ldots, x_n)$ and $w' = (w_1, w_2, \ldots, w_j, \ldots, w_n)$
- create $b_1 = (x_1, x_2, \ldots, x_j, w_{j+1}, \ldots, w_n)$ and $b_2 = (x_1, x_2, \ldots, w_j, w_{j+1}, \ldots, w_n)$

Instance b_1 is created by taking feature values of x' for $i = 1, \ldots,$j and features values of w' for $i = j+1, \ldots, n$. Analogously, instance b_2 takes feature values of x' for $i = 1, \ldots, j-1$ and features values of w' for $i = j, \ldots, n$;

Calculate $\phi_j^m(x)$ of the mth iteration from the formula:

$\phi_j(x) = \hat{f}(b_1) - \hat{f}(b_2)$

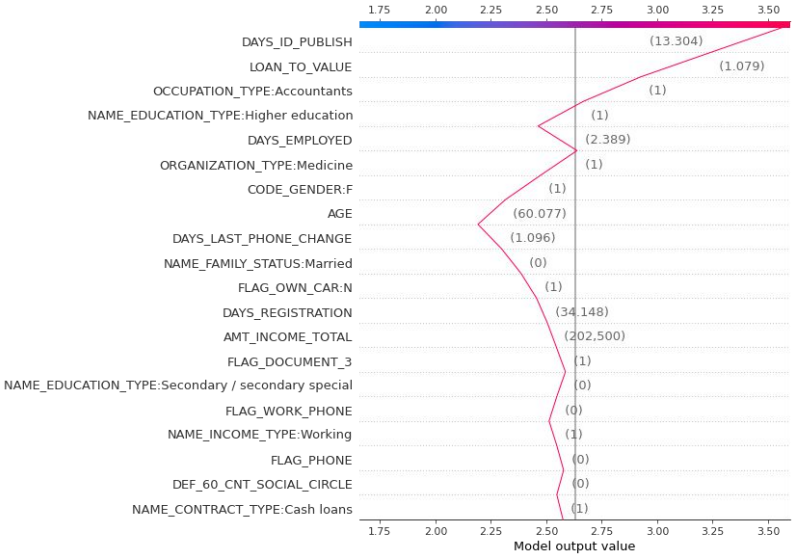
end

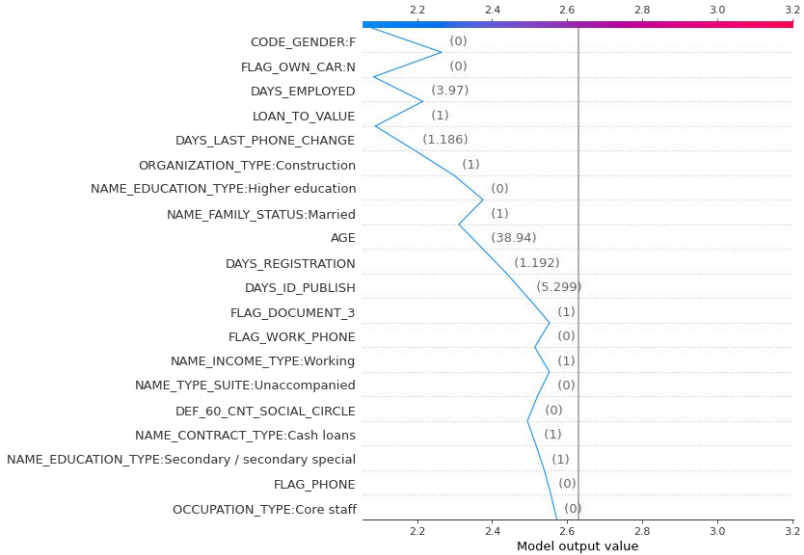
Shapley value is the average $\phi_j(x) = \frac{1}{m}\sum_{i=1}^{m}\phi_j^m(x)$.

Source: own work

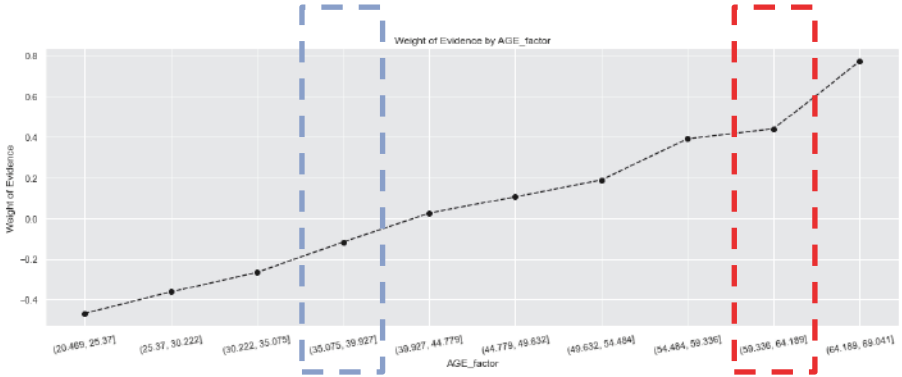
Źródło: D. Kaszyński, B. Kamiński, T. Szapiro, "Credit Scoring in Context of Interpretable Machine Learning", SGH Publishing House, 2020

Metody XAI – Shapley values

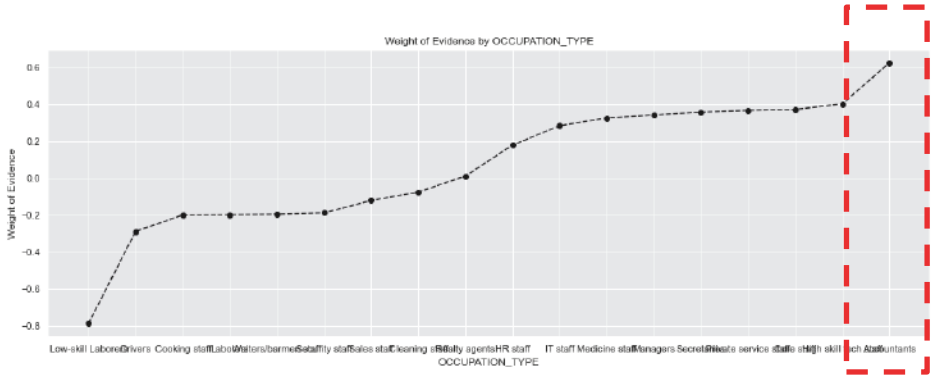




Metody XAI – Shapley values

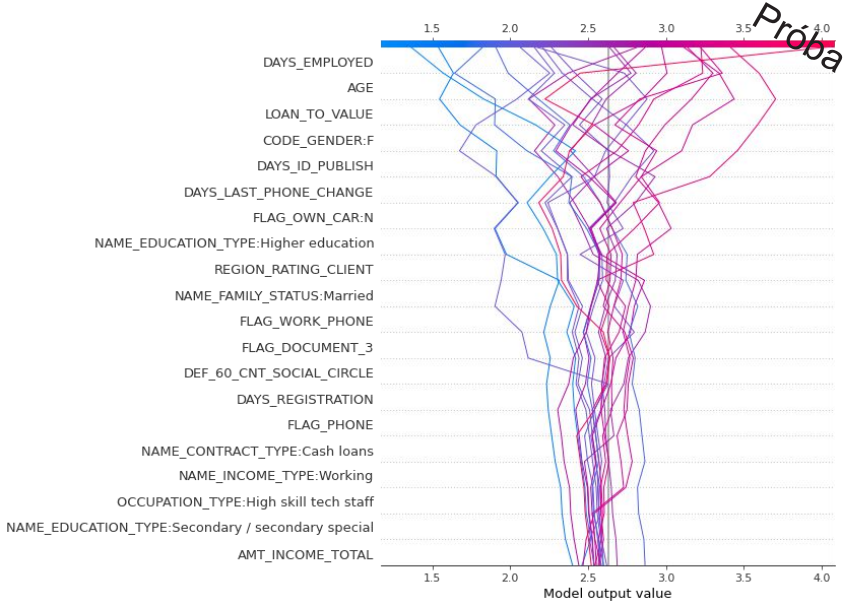


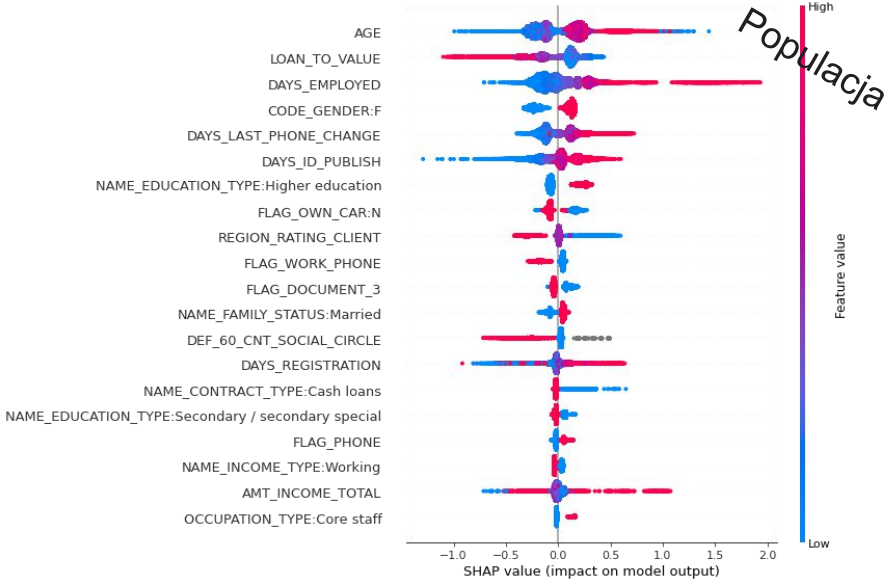
38 60



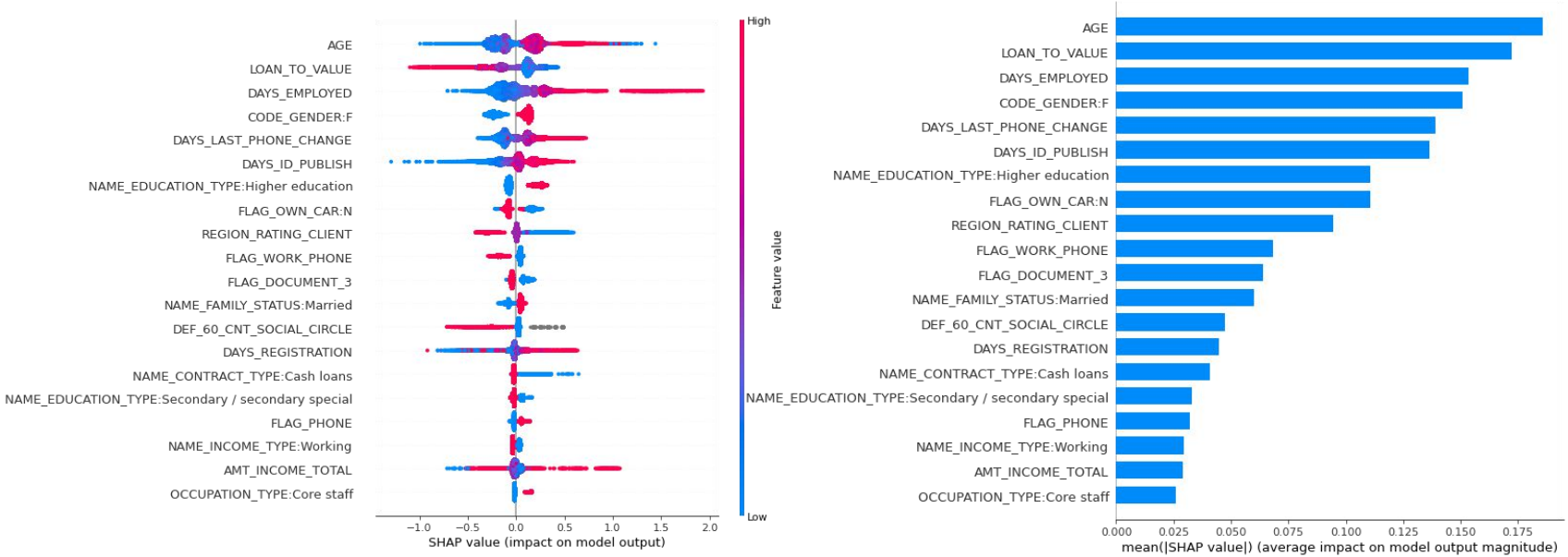
Accountants

Metody XAI – Shapley importance

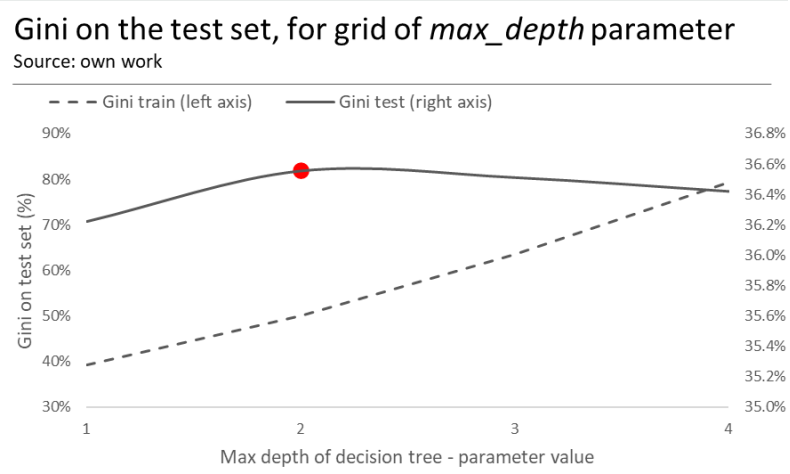




Metody XAI – Permutational feature importance



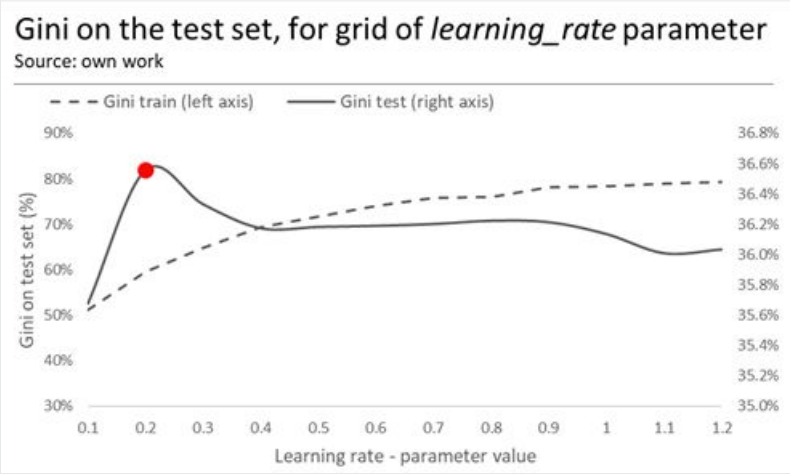
Parametryzacja XGBoost



Parametr modelu *max_depth*

- XGBoost jest oparty o drzewa decyzyjne – parametr wskazuje na głębokość drzew decyzyjnych
- Model bardzo szybko przeucza się wraz ze wzrostem parametru *max_depth* – optymalna wartość to 2

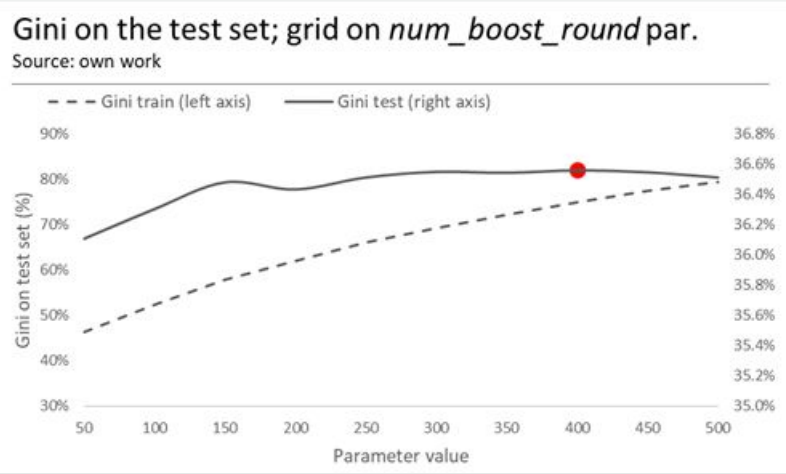
Parametryzacja XGBoost



Parametr modelu *learning_rate*

- Tempo uczenia metody boostingowej
- Optymalna wartość to 0.2

Parametryzacja XGBoost



Parametr modelu *num_boost_round*

- Parametr wskazujący na liczbę rund boostingowych.
- Optymalna wartość to 400, przy czym wskazujemy, że parametr ten nie zmienia się istotnie.

Dane wrażliwe (RODO), a cechy chronione (KPP)

RODO	KPP
art. 9 ust. 1: Zabrania się przetwarzania danych osobowych ujawniających **pochodzenie rasowe lub etniczne, poglądy polityczne, przekonania religijne lub światopoglądowe**, przynależność do związków zawodowych oraz przetwarzania danych genetycznych, danych biometrycznych w celu jednoznacznego zidentyfikowania osoby fizycznej lub danych dotyczących zdrowia, **seksualności lub orientacji seksualnej tej osoby.**	art. 21 ust. 1: Zakazana jest wszelka dyskryminacja 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.**

Tabela 1: Zakres definicji danych wrażliwych (RODO) i cech chronionych (Karta Podstawowych Praw Unii Europejskiej).
Źródło: Niklas J., 2019. Problem dyskryminacji automatycznej – uwagi na tle ogólnego rozporządzenia o ochronie danych osobowych, Europejski Przegląd Sądowy

Dyrektywa Rady 2004/113/WE

Pojęcie dyskryminacji, jest różnicowane przez zapisy Dyrektywy na 4 podtypy, tj.:

- dyskryminacja bezpośrednia, związana z sytuacją, w której dana osoba jest traktowana w sposób **mniej korzystny ze względy na płeć**, niż gdyby była traktowana inna osoba w porównywalnej sytuacji,
- […]

Ocena zdolności kredytowej, która bierze pod uwagę **płeć, lub wiek, jest traktowana jako przejaw dyskryminacji bezpośredniej**. Oznacza to, że automatyczny model decyzyjny opracowany przez bank w celu oceny zdolności kredytowej (a w dalszej konsekwencji wydawania decyzji kredytowej), nie może dyskryminować / oceniać zdolności kredytowej ze względu na cechy wrażliwe.

Literatura naukowa obszaru oceny zdolności kredytowej niejednokrotnie wskazuje na przykłady, w których zmiennymi statystycznie istotnymi są **płeć**: [1], [2], [3], [4], [5], lub **wiek kredytobiorcy**: [6], [7], [8].

Art. 5 ust. 1 Dyrektywy Rady 2004/113/WE

Zagadnienie wykorzystania zmiennej płeć w modelach ubezpieczeniowych i związanych z usługami finansowymi (a więc również modele oceny zdolności kredytowej) jest bezpośrednio poruszony w art. 5 ust. 1 Dyrektywy, który wskazuje, że:

> *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****.*

W Dyrektywie tej, wskazano jednak zapis w art. 5 ust. 2, który umożliwiał wykorzystanie płci gdy jej użycie jest czynnikiem decydującym w ocenie ryzyka opartego na odpowiednich i dokładnych danych aktuarialnych i statystycznych. Oznacza to, że pierwotna intencja wskazywała na możliwość wykorzystania płci, w celach wyznaczania składek ubezpieczeniowych, oraz celach związanych z innymi usługami finansowymi.

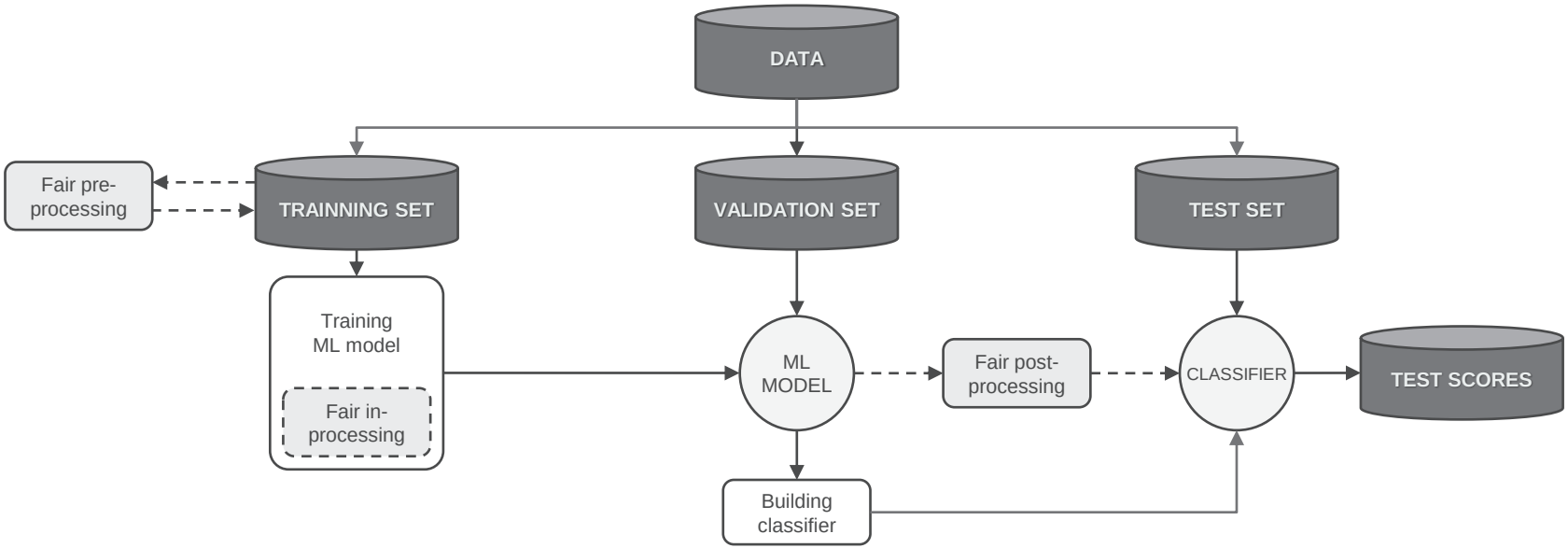
The use of gender in insurance pricing wydanym przez Europejski Trybunał Sprawiedliwości (2011)

Dokument [9] wskazuje, na **niemożność różnicowania składek ubezpieczeniowych oraz celów związanych z innymi usługami finansowymi ze względu na płeć**.

- wskazuje się, **płeć jest jedynie statystycznie związana z ryzykiem** – wiele innych czynników (ekonomiczne, środowisko i społeczne), wchodzące w interakcję z osobistymi nawykami i stylem życia poszczególnych osób kształtują poziom ryzyka.
- zmiany społeczne wynikające zanikającym tradycyjnym podziałem ról i wzorców mężczyzn i kobiet, powodują, że nie jest możliwe ustalenie wyraźnego związku między płcią, a czynnikami behawioralnymi.
- pomimo, że płeć jest czynnikiem łatwiej identyfikowalnym (niż np. skłonność osoby do ryzyka), to inne czynniki wpływają na poziom ryzyka; argument przemawiający za wygodą użycia płci, jako czynnika dyskryminującego, nie może stanowić odpowiedniego uzasadnienia, a wykorzystanie płci danej osoby jako pewnego rodzaju kryterium zastępczego dla innych cech.

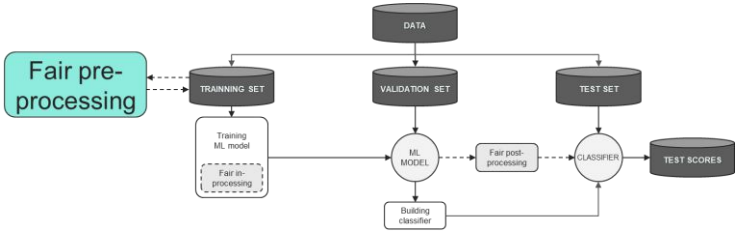
Dokument *The use of gender in insurance pricing* wskazuje jednak na ważny zapis, tj.: ***stosowanie płci jako czynnika dyskryminującego nie jest zabronione; tylko w przypadkach, gdy zastosowanie tego czynnika skutkuje zróżnicowaniem poziomem świadczeń, byłoby to uznane za nieważną praktykę***.

Metody zapewniania sprawiedliwości (1/4)



Metody zapewniania sprawiedliwości (2/4)

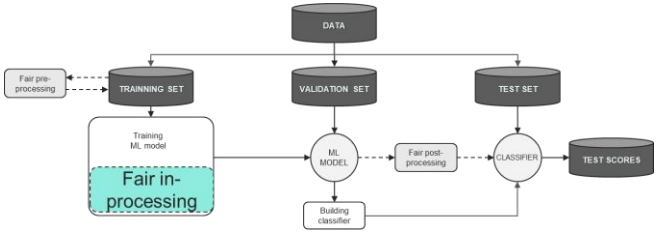
Metody *fair pre-processing:*



- związane ze wstępnym etapem przetwarzania danych zbioru uczącego; motywacją jest wskazanie, że przyczyną dyskryminacji są dane zbioru treningowego [11] – modyfikując zbiór uczący, można zredukować poziom dyskryminacji;
- może być to wynikowe względem dyskryminacji w danych historycznych, lub występowania niedostatecznej reprezentacji grupy mniejszościowej (tj. błędy w tych grupach są bardziej prawdopodobne ze względu na niektóre miary dokładności);
- zazwyczaj przeprowadza się dekorelację atrybutów ze zmienną chronioną [12];
- innym podejściem jest algorytm opisany w [13], który zakłada modyfikację każdego z atrybutów, aby rozkłady krańcowe atrybutów były równe (nie modyfikuje się etykiet treningowych);
- zaletą jest wczesne włączenie *sprawiedliwego* traktowania w procesie budowy modelu [14].

Metody zapewniania sprawiedliwości (3/4)

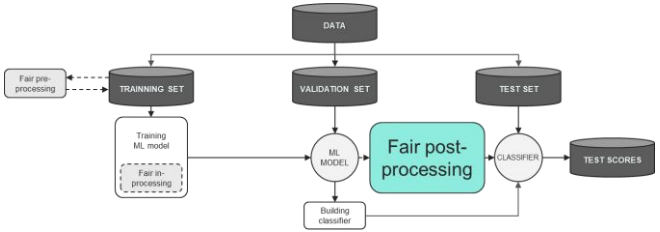
Metody *fair in-processing:*



- związane z modyfikacją algorytmu uczącego, np. poprzez dodanie dodatkowych ograniczeń w procesie estymacji parametrów – obecnie najbardziej popularne podejście do zapewnienia sprawiedliwości [11];
- w [15] zaproponowano podejście regresji logistycznej, z dodatkowym wyrazem regularyzacyjnym; w [16] wskazano, że standardowe ograniczenia sprawiedliwości są niewypukłe, i trudne do spełnienia – w artykule wprowadzono mechanizm wypukłej relaksacji dla celów optymalizacyjnych;
- W [17] zaproponowano algorytm pt. *Two Naive Bayes*, polegający na wytrenowaniu osobnych modeli dla poszczególnych wartości zmiennych chronionych – w kolejnym kroku iteracyjne łączy się prognozy z tych modeli, ze względu na zdefiniowane miary sprawiedliwości;

Metody zapewniania sprawiedliwości (4/4)

Metody *fair post-processing:*

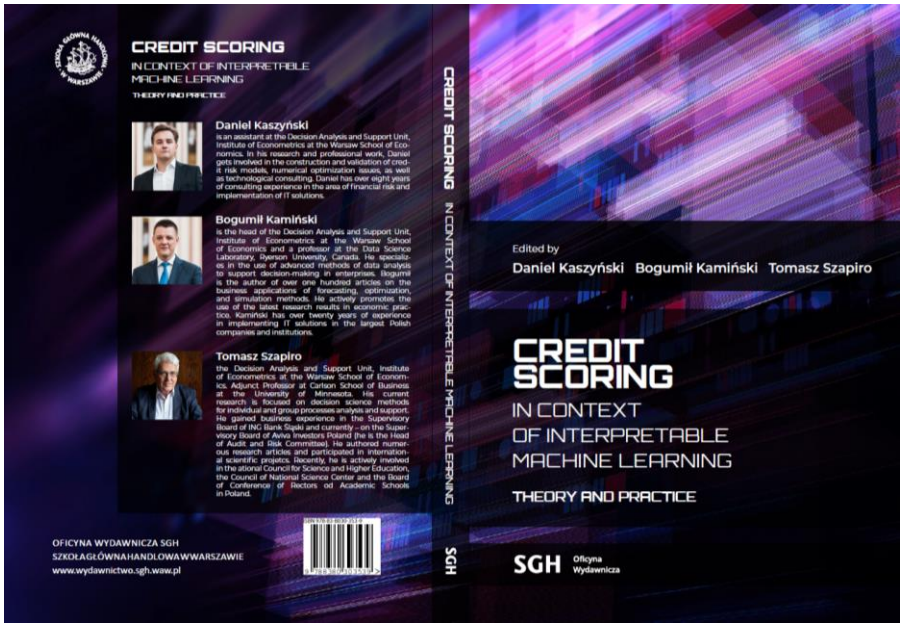


- związane z wprowadzeniem sprawiedliwości poprzez modyfikację wyników działania modelu (klasyfikatora) [11];
- standardowa procedura zakłada modyfikację rozkładu oszacowanych scorów lub etykiet [10];
- w [18] przedstawiono technikę modyfikacji etykiet w liściach drzew decyzyjnych po estymacji, w celu spełnienia ograniczeń miar związanych ze sprawiedliwością;
- 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 [14].

Bibliografia

- [Kaszynski et al., 2020]: Kaszyński, D., Kamiński, B. and Szapiro, T., 2020. Credit scoring in the context of interpretable machine learning.
- [Prawo Bankowe, 1997]: Ustawa z dnia 29 sierpnia 1997 r. – Prawo bankowe (Dz.U. z 2020 r. poz. 1896)
- [Przanowski, 2014]: Przanowski, K., 2014. Credit Scoring w erze Big Data. Warszawa: Oficyna Wydawnicza SGH.
- [MSSF 9, 2016]: Rozporządzenie Komisji (UE) 2016/2067 z dnia 22 listopada 2016 r. zmieniające rozporządzenie (WE) nr 1126/2008 przyjmujące określone międzynarodowe standardy rachunkowości zgodnie z rozporządzeniem (WE) nr 1606/2002 Parlamentu Europejskiego i Rady w odniesieniu do Międzynarodowego Standardu Sprawozdawczości Finansowej 9

#ZWIADoAI


CREDIT SCORING
IN CONTEXT OF INTERPRETABLE MACHINE LEARNING
THEORY AND PRACTICE
Daniel Kaszyński
is an assistant at the Decision Analysis and Support Unit, Institute of Econometrics at the Warsaw School of Economics. In his research and professional work, Daniel gets involved in the construction and validation of credit risk models, numerical optimization issues, as well as technological consulting. Daniel has over eight years of consulting experience in the area of financial risk and implementation of IT solutions.
Bogumił Kamiński
is the head of the Decision Analysis and Support Unit, Institute of Econometrics at the Warsaw School of Economics and a professor at the Data Science Laboratory, Ryerson University, Canada. He specializes in the use of advanced methods of data analysis to support decision-making in enterprises. Bogumił is the author of over one hundred articles on the business applications of forecasting, optimization, and simulation methods. He actively promotes the use of the latest research results in economic practice. Kamiński has over twenty years of experience in implementing IT solutions in the largest Polish companies and institutions.
Tomasz Szapiro
the Decision Analysis and Support Unit, Institute of Econometrics at the Warsaw School of Economics. Adjunct Professor at Carlson School of Business at the University of Minnesota. His current research is focused on decision science methods for individual and group processes analysis and support. He gained business experience in the Supervisory Board of ING Bank Śląski and currently - on the Supervisory Board of Aviva Investors Poland (he is the Head of Audit and Risk Committee). He authored numerous research articles and participated in international scientific projects. Recently, he is actively involved in the ational Council for Science and Higher Education, the Council of National Science Center and the Board of Conference of Rectors od Academic Schools in Poland.
OFICYNA WYDAWNICZA SGH
SZKOŁAGŁÓWNAHANDLOWAWWARSZAWIE
www.wydawnictwo.sgh.waw.pl
CREDIT SCORING IN CONTEXT OF INTERPRETABLE MACHINE LEARNING
SGH
Edited by
Daniel Kaszyński Bogumił Kamiński Tomasz Szapiro
CREDIT SCORING
IN CONTEXT OF INTERPRETABLE MACHINE LEARNING
THEORY AND PRACTICE
SGH Oficyna Wydawnicza