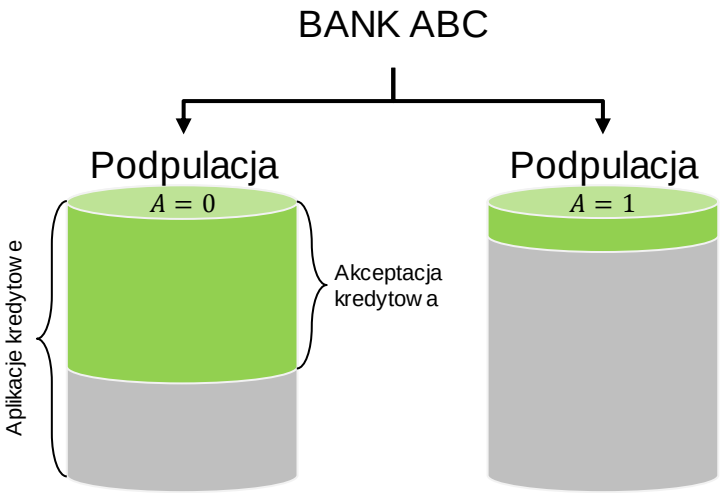
SGH
Warsaw School
of Economics

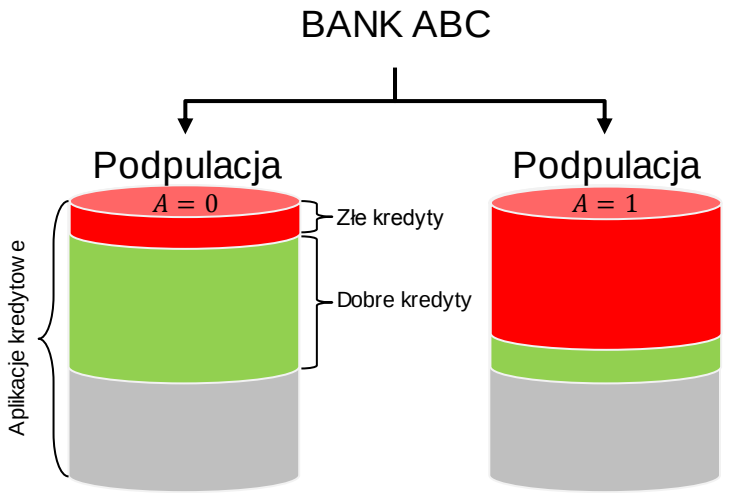
Społeczna ewaluacja modeli oceny zdolności kredytowej

Daniel Kaszyński
9 stycznia 2023 r.

Scenariusz wprowadzający



SCENARIUSZ 1



SCENARIUSZ 2

Czy takie działanie Banku ABC nazwiemy sprawiedliwym?

Dyskryminacja grup społecznych

Odmienny wpływ (ang. Disparate impact)

Odmienny wpływ to przykład dyskryminacji, dyskryminacja gdy osoby są traktowane jednakowo w ramach określonego zestawu zasad/procedur (np. model oceny zdolności kredytowej), ale te zasady dają przewagę jednej z grup. Określone zasady są stosowane w równym stopniu wobec wszystkich, ale mają one nieproporcjonalny, niekorzystny wpływ na wnioskodawców z chronionej grupy.

Przykład w modelowaniu zdolności kredytowej: istotnie różny odsetek akceptacji kredytów dla kobiet i mężczyzn.

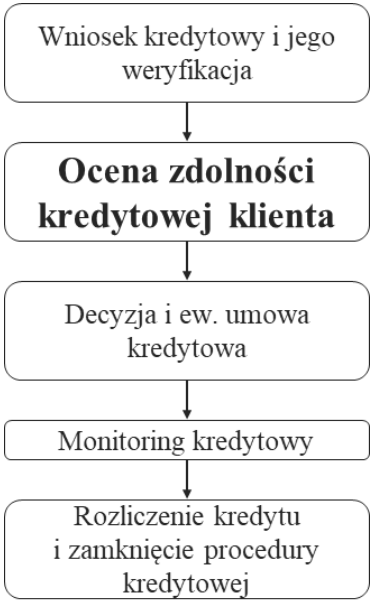
Odmienne traktowanie (ang. Disparate treatment)

Odmienne traktowanie to dyskryminacja, gdy równo wykwalifikowane osoby są traktowane odmiennie, ze względu na ich zmienną chronioną (płeć, kolor skóry, religia). Procedura oceny zdolności kredytowej działa odmiennie dla poszczególnych grup społecznych (np. kobiety-mężczyźni).

Przykład w modelowaniu zdolności kredytowej: jakość prognozy modelu skoringowego jest różna dla kobiet i mężczyzn (ale podobny odsetek akceptacji) – tzw. „lenistwo" (ang. leziness).

Źródło: Hunter N., 2018, https://fairlendingdiversity.com/disparate-treatment-vs-disparate-impact/

Zdolność kredytowa i metoda jej oceny



Schemat 1: Uproszczony diagram procesu kredytowego
Źródło: Opracowanie własne w oparciu o Klimontowicz 2017

Zdolności kredytowa, patrz art. 70. Prawo bankowe

> *Przez zdolność kredytową rozumie się zdolność do spłaty zaciągniętego kredytu wraz z odsetkami w terminach określonych w umowie.*

Model oceny zdolności kredytowej, patrz Schreiner 2003

> *Model oceny zdolności kredytowej to narzędzie, które oszacowuje prawdopodobieństwa niewykonania zobowiązania przez klienta na podstawie danych historycznych.*

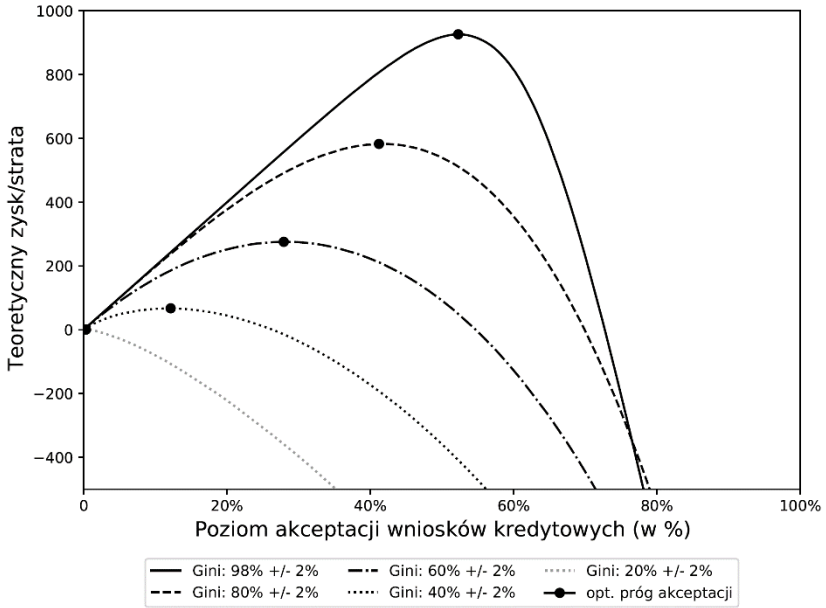
Ewaluacja – ocena działania - modeli oceny zdolności kredytowej obecnie jest prowadzona głównie w wymiarze technicznych.

Coraz częściej uwzględnia się aspekty społeczno-ekonomiczne w ewaluacji modeli oceny zdolności kredytowej.

Istota jakości modeli oceny zdolności kredytowych

- Poprawa jakości modelu oceny zdolności kredytowej oznacza lepsze rozróżnianie klientów „dobrych” i „złych”.
- Lepsze rozróżnienie klientów umożliwia podniesienie poziomu akceptacji wniosków kredytowych (więcej udzielanych kredytów).
- Lepsza ocena zdolności kredytowej umożliwia osiągnięcie wyższego zysku operacyjnego banku: przychód z oprocentowania minus koszt utraconych kredytów.
- Modele oceny zdolności kredytowej wykorzystywane w procesie podejmowania decyzji kredytowej mają bezpośrednie przełożenie na rentowności banku.

Wyższa jakość prognoz modelu oceny zdolności kredytowej, przekłada się na wyższy zysk.



Schemat 2: Krzywa zysku/straty z akcji kredytowej w zależności od jakości modelu oceny zdolności kredytowej

Źródło: Opracowanie własne

Przyjęta notacja

- $X_{m\times n}$ to macierz zmiennych objaśniających, gdzie m to liczba obserwacji, a n to liczba zmiennych objaśniających model.
- $A_{m\times 1} \subset X_{m\times n}$ to pojedynczą zmienną chronioną (np. płeć – jedna z kolumn macierzy X).
- $Y_{m\times 1}$ to wektor binarnych wartości zmiennej objaśnianej.
- $s(.)$ to funkcja oceny skoringowej, oszacowana w oparciu o historyczne wartości X i Y.
- $s(X_{i,.})$ to wartość funkcji oceny skoringowej dla i-tej obserwacji.

W oparciu o powyższe definicje, można zdefiniować $\hat{\ }_{m\times 1}$ jako wektor binarny zawierający klasyfikację (tj. prognoza z modelu) obserwacji na podstawie macierzy X. Definiowany jest jako:

$$\hat{\ }_i = \begin{cases} 1 & s(X_{i,.}) > \tau \\ 0 & s(X_{i,.}) \leq \tau \end{cases}$$

gdzie: τ to próg odcięcia klasyfikacji.

	TARGET	CONSTANT	GENDER	AGE	INCOME	SPENDINGS
0	0.0	1.0	0.0	-0.247904	0.950364	2.344213
1	1.0	1.0	0.0	0.062824	-1.954827	-0.916302
2	0.0	1.0	0.0	0.516308	0.102586	1.886797
3	0.0	1.0	1.0	-2.120281	-1.709788	-0.294365
4	0.0	1.0	1.0	0.980328	-0.266985	-1.794711
...	...	...	...	...	...	...
999995	0.0	1.0	0.0	-1.129343	-0.332387	0.581598
999996	0.0	1.0	0.0	-0.664906	1.212868	-1.673147
999997	0.0	1.0	0.0	-2.347681	0.225940	-0.593212
999998	0.0	1.0	0.0	0.725457	0.430236	1.864633
999999	1.0	1.0	1.0	0.814198	-0.587235	-1.392114

1000000 rows × 6 columns

Schemat 4:Przykład wysymulowanych danych wykorzystywanych w procesie oceny zdolności kredytowej

Źródło: Opracowanie własne

Przyjęta notacja

	TARGET	Y_PROB	CONSTANT	GENDER	AGE	INCOME	SPENDING
61402	0.0	0.024498	1.0	0.0	1.842158	0.272055	-0.153557
60366	0.0	0.239996	1.0	1.0	0.715060	-0.851210	0.579654
34966	0.0	0.084533	1.0	1.0	-0.389781	1.415681	0.196077
42763	0.0	0.274303	1.0	1.0	0.335925	-0.395488	-0.250881
45034	1.0	0.096161	1.0	0.0	0.704650	-0.455414	0.330789
...	...	...	...	...	...	...	...
61343	0.0	0.259004	1.0	0.0	-0.086876	-0.828243	-0.127760
10863	1.0	0.687423	1.0	1.0	-0.585117	-1.303460	-0.541361
63603	0.0	0.050254	1.0	1.0	2.040028	-0.225813	0.832249
4119	0.0	0.270870	1.0	0.0	-0.103258	-0.384307	-1.357185
71045	1.0	0.148212	1.0	1.0	-1.135105	1.393522	0.158147

20000 rows × 7 columns

- Wysokie prawdopodobieństwo wejścia w stan *default*
- Niskie prawdopodobieństwo wejścia w stan *default*

Schemat 4: Prawdopodobieństwo wejścia w stan default vs. zmienna objaśniana
Źródło: Opracowanie własne

Miary zdolności predykcyjnej modeli oceny zdolności kredytowej

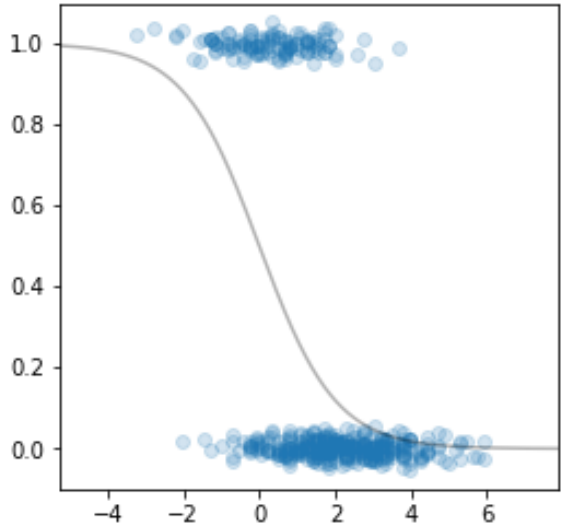
Ewaluacja techniczna (1/5)

Modele oceny zdolności kredytowej są to modele klasyfikacji binarnej.
Klasyfikacja odbywa się przy założeniu wykorzystania progu τ:

$$s[Y|X] > \tau \Rightarrow \hat{} = 1$$
$$s[Y|X] \leq \tau \Rightarrow \hat{} = 0$$

gdzie:

- $s[.]$ to model oceny zdolności kredytowej (klasyfikator) zwracający ocenę punktową – tzw. score,
- y to binarna zmienna celu (np. klient spłacający / niespłacający kredyt),
- X to zbiór atrybutów modelu,
- τ to próg decyzji kredytowej (tzw. prób odcięcia klasyfikacji).

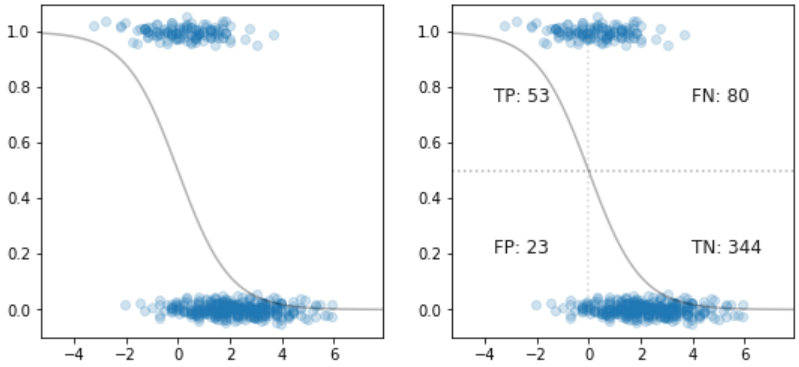


Schemat 2: Krzywa logistyczna dopasowana do wylosowanej próby
Źródło: Opracowanie własne

Ewaluacja techniczna (2/5)

- Przy założeniu progu odcięcia na określonym poziomie (np. 0.5) uzyskuje się prognozę z modelu, tj. $\hat{Y}$.
- Zestawiając w kolumnach prognozę z modelu $\hat{Y}$ oraz w wierszach realizację zmiennej objaśnianej y można uzyskać tzw. macierz pomyłek:

		Realizacja	
		$Y = 1$	$Y = 0$
Prognoza	$\hat{Y} = 1$	True Positive TP	False Positive FP
	$\hat{Y} = 0$	False Negative FN	True Negative TN

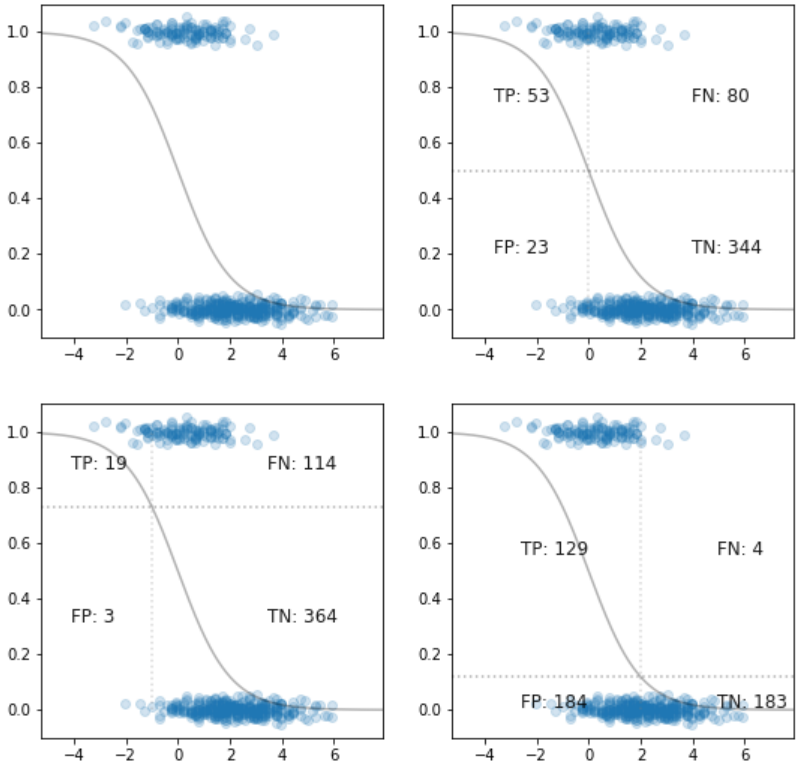


Schemat 3: Przykład klasyfikacji $\tau = 0.5$

Źródło: Opracowanie własne

Ewaluacja techniczna (3/5)

- Sterując odpowiednio parametrem $\tau \in \{0.5, 0.75, 0.175\}$ uzyskiwane są różne wartości parametrów w macierzy pomyłek – różne błędy.
- Przy progu konserwatywnym $\tau = 0.175$ (tj. prognozującym więcej obserwacji na $\hat{} = 1$), wartość False Negative jest zdecydowanie mniejsza niż w pozostałych przypadkach, przy czym rośnie błąd False Positive. Odwrotna sytuacja zachodzi przy progu liberalnym $\tau = 0.75$.
- Sterowaniem parametrem τ umożliwia generowanie różnych macierzy pomyłek-wartości błędów FN i FP.
- Optymalny dobór wartości τ opiera się na *koszcie* związanym z popełnieniem błędów FN i FP, oraz zysku z poprawnej klasyfikacji TP i TN.



Schemat 4: Przykłady klasyfikacji: $\tau \in \{0.5, 0.75, 0.175\}$

Źródło: Opracowanie własne

Ewaluacja techniczna (4/5)

- W oparciu o macierz pomyłek, można wyznaczyć standardowe miary, m.in.:

$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{TP+TN+FP+FN}$, *accuracy*

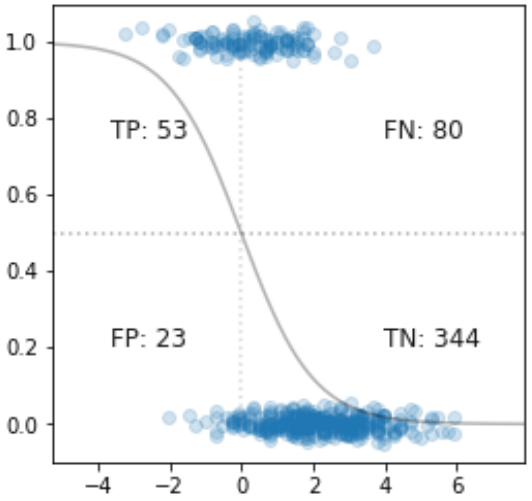
$\mathrm{TPR} = \frac{TP}{TP+FN}$, *sensitivity*

$\mathrm{TNR} = \frac{\mathrm{TN}}{TN+FP} = 1 - FP\mathrm{R}$, *specificity*

$\mathrm{FPR} = \frac{\mathrm{FP}}{N}$, *fall-out*

$\mathrm{PPV} = \frac{TP}{TP+FP}$, *precision, positive predictive value*

$F_1 = \frac{2TP}{2TP+FP+FN}$, *średnia harmoniczna PPV i TPR*

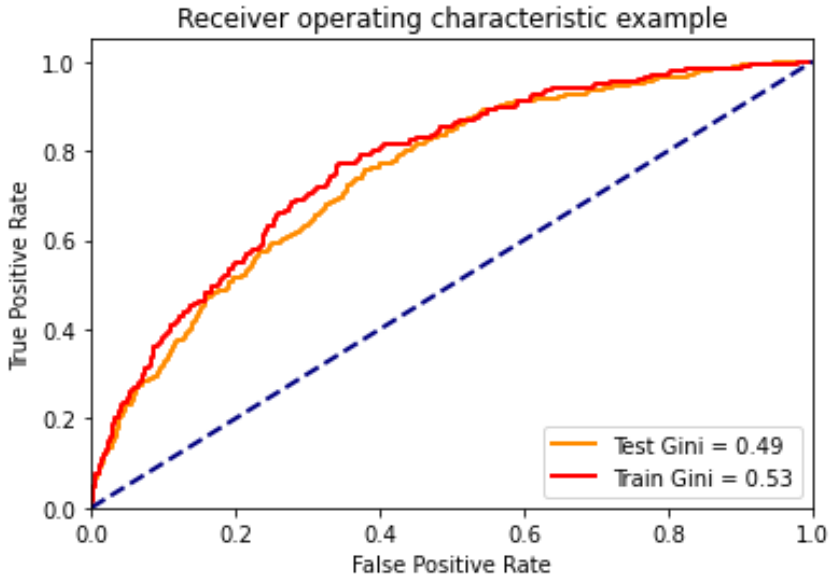


Schemat 2: Krzywa logistyczna dopasowana do wylosowanej próby
Źródło: Opracowanie własne

Ewaluacja techniczna (5/5)

- Moc dyskryminacyjną modelu można oceniać poprzez tzw. krzywą ROC, oraz pole pod krzywą ROC – wartość AUC (ang. Area Under the ROC Curve).
- Popularną miarą mocy dyskryminacyjnej modelu klasyfikacji binarnej jest również indeks Gini, określony na wartościach [0,1]:

$$Gini = 2 \times AUC - 1$$



Schemat 2: Krzywa ROC dla oszacowanego modelu oceny zdolności kredytowej

Źródło: Opracowanie własne

Społeczne aspekty ewaluacji modeli oceny zdolności kredytowej

Motywacja regulacyjna

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.

Bazowe miary stronniczości algorytmicznej

Warunek Parytetu Demograficznego

Demographic parity

$$P[s(X|A=1) > \tau] = P[s(X|A=0) > \tau]$$

Warunek Parytetu Demograficznego, nazywany również kryterium niezależności wymaga, aby ocena skoringowa $s(X)$ była niezależna względem wartości zmiennej chronionej A. W praktyce wykorzystuje się przybliżoną formę:

$$|P[s(X|A=1) > \tau] - P[s(X|A=0) > \tau]| < \xi$$

Z perspektywy kontekstu oceny zdolności kredytowej, parytet demograficzny wymaga, aby bank, każdej grupie osób różnicowanych przez zmienną A, udzielał kredytów w takiej samej frakcji. Kryterium niezależności odnosi się do równości wpływu modelu.

Warunek wyrównanych szans

Equalized odds

$$\begin{cases} P[s(X|Y=0, A=1) > \tau] = P[s(X|Y=0, A=0)] \\ P[s(X|Y=1, A=1) \leq \tau] = P[s(X|Y=0, A=0)] \end{cases}$$

Warunek wyrównanych szans, nazywany również kryterium separacji wymaga, aby dla każdej z grup wartości zmiennej chronionej A, wartości FPR (False Positive Rate), oraz FNR (False Negative Rate) były takie same. W praktyce korzysta się z miary separacji:

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

Kryterium separacji odnosi się do błędu klasyfikacji między grupami o różnych wartościach zmiennej chronionej A; kryterium separacji odnosi się do równości traktowania modelu .

Warunek Parytetu Prognostycznego

Predictive parity

$$P[y=1|s(X) > \tau, A=0] = P[y=1|s(X) > \tau, A=1]$$

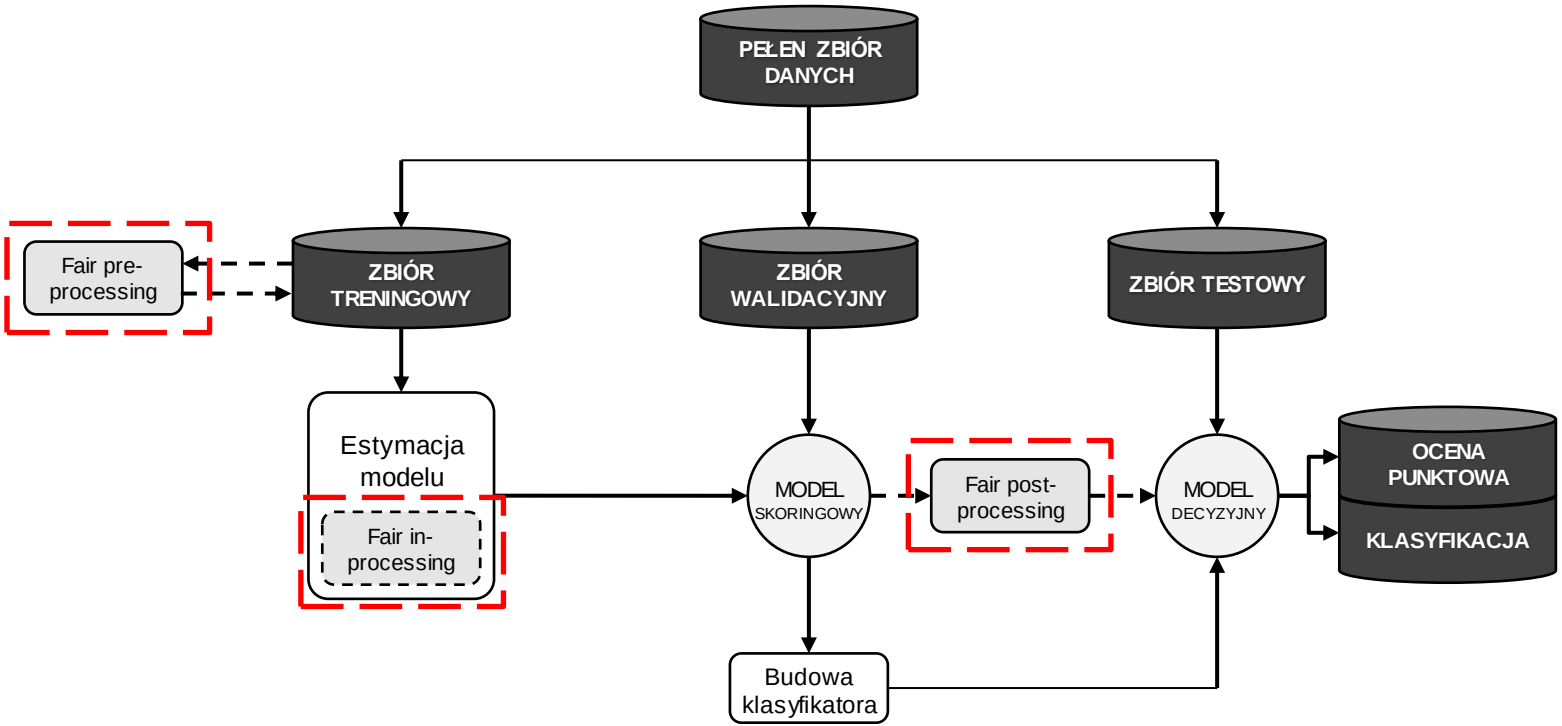
Warunek Parytetu Predyktywności, nazywany również kryterium dostateczności wymaga, aby PPV (Positive Predictive Value) dla poszczególnych grup wartości zmiennej chronionej A była taka sama. W praktyce korzysta się z miary dostateczności:

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

Kryterium dostateczności odnosi się do równości traktowania modelu.

Źródło: Barocas, S., Hardt, M., & Narayanan, A., 2019., Hardt, M., Price, E., & Srebro, N., 2016, Kozodoi, N., Jacob, J., Lessmann, S., 2022.

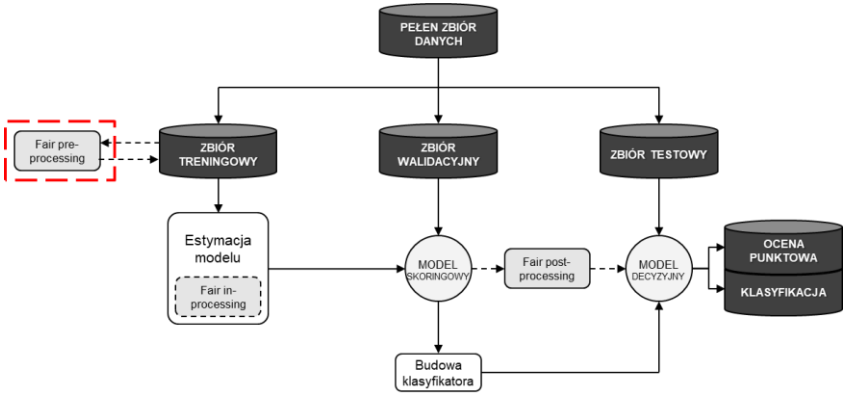
Metody redukcji stronniczości algorytmicznej (1/4)



Źródło: Kozodoi, N., Jacob, J., Lessmann, S., 2022.

Metody redukcji stronniczości algorytmicznej (2/4)

Metody *fair pre-processing:*

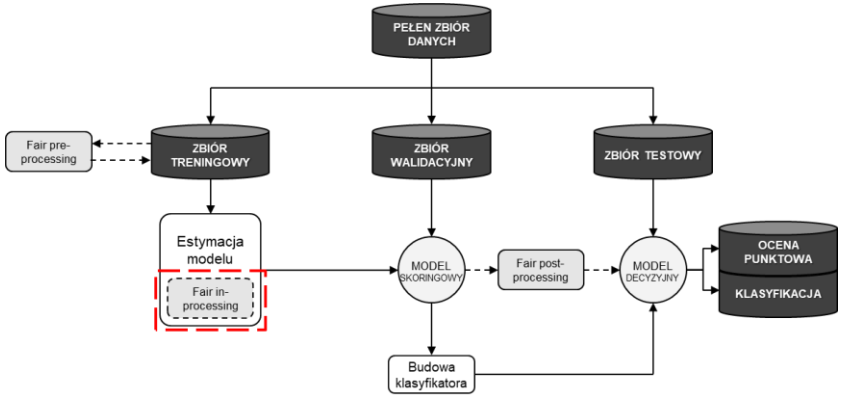


- związane ze wstępnym etapem przetwarzania danych zbioru uczącego; motywacją jest wskazanie, że przyczyną dyskryminacji są dane zbioru treningowego, patrz Friedler et al. 2019 – modyfikując zbiór uczący, można zredukować poziom dyskryminacji;
- może być to wynikowe względem dyskryminacji w danych historycznych, 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 Calmon et al. 2017;
- innym podejściem jest algorytm opisany w Feldman et al. 2015, 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, patrz Barocas et al. 2019.

Metody redukcji stronniczości algorytmicznej (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, patrz Friedler et al. 2019;

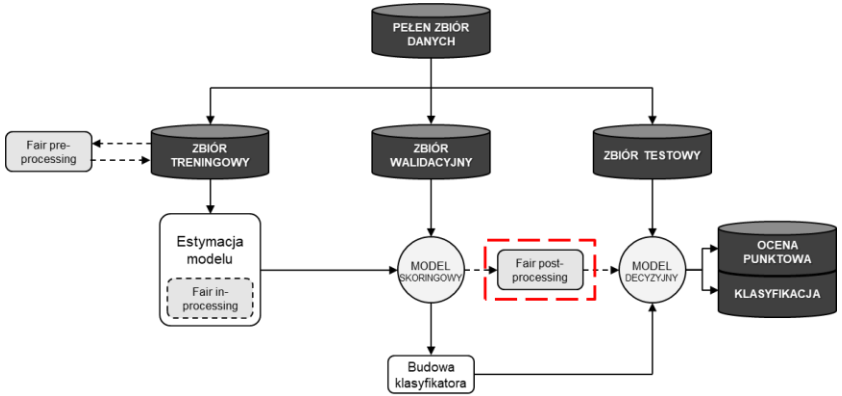


- w Kamishima et al. 2012 zaproponowano podejście regresji logistycznej, z dodatkowym wyrazem regularyzacyjnym; w Zafar et al., 2017 wskazano, że standardowe ograniczenia sprawiedliwości są niewypukłe, i trudne do spełnienia – w artykule wprowadzono mechanizm wypukłej relaksacji dla celów optymalizacyjnych;
- w Calders i Verwer 2010 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 redukcji stronniczości algorytmicznej (4/4)

Metody *fair post-processing:*

- związane z wprowadzeniem sprawiedliwości poprzez modyfikację wyników działania modelu (klasyfikatora) patrz Friedler et al. 2019;
- standardowa procedura zakłada modyfikację rozkładu oszacowanych scorów lub etykiet Kozodoi et al. 2022;



- w Kamiran et al. 2010, 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, patrz Barocas et al. 2019.

Eksperyment numeryczny

Eksperymenty numeryczne

Eksperyment 1

Perspektywa krótkoterminowa,
tj. wpływ modeli FairAware na bieżącą wartość monitorowanych parametrów (np. jakość modelu, dyskryminacji podpopulacji).

Eksperyment 2

Perspektywa długoterminowa,
tj. wpływ modeli FairAware na długookresową strukturę populacji osób, których aplikacje kredytowe są akceptowane..

Eksperyment 1

Perspektywa krótkoterminowa

Metoda	Model bazowy	GINI *train*	GINI *test*	STP *train*	STP *test*	TPR *train*	TPR *test*	FPR *train*	FPR *test*
Unmitigated	Logistic Regression	0.61	0.60	0.03	0.03	0.07	0.08	0.03	0.03

Eksperyment 1

Perspektywa krótkoterminowa

Metoda	Model bazowy	GINI *train*	GINI *test*	STP *train*	STP *test*	TPR *train*	TPR *test*	FPR *train*	FPR *test*
Unmitigated	Logistic Regression	0.61	0.60	0.03	0.03	0.07	0.08	0.03	0.03

True Positive Rate – czułość. Wskazana miara jest miarą względną.

Analizowany przykład: dla kobiet czułość klasyfikatora wynosi 8% wartości TPR dla mężczyzn

Indeks Gini – miara jakości modelu.

Porównanie na zbiorze treningowym i testowym daje pogląd na wielkość **przetrenowania modelu.**

Statistical Parity – parytet statystyczny. Wskazana miara jest miarą względną.

Analizowany przykład: na 100 mężczyzn otrzymujących akceptację kredytową, przypada 0.03×100=3 kobiet.

False Positive Rate.

Eksperyment 1

Perspektywa krótkoterminowa

Metoda	Model bazowy	GINI *train*	GINI *test*	STP *train*	STP *test*	TPR *train*	TPR *test*	FPR *train*	FPR *test*
Unmitigated	Logistic Regression	0.61	0.60	0.03	0.03	0.07	0.08	0.03	0.03
Fairness Through Unawareness	Logistic Regression	0.52	0.52	0.39	0.41	1.76	1.83	0.24	0.27

Spadek jakości modelu

Poprawa jakości miar dyskryminacji.

Poprawa nie jest pełna, ponieważ pozostałe zmienne są skorelowane ze zmienną chronioną

Eksperyment 1

Perspektywa krótkoterminowa

Metoda	Model bazowy	GINI *train*	GINI *test*	STP *train*	STP *test*	TPR *train*	TPR *test*	FPR *train*	FPR *test*
Unmitigated	Logistic Regression	0.61	0.60	0.03	0.03	0.07	0.08	0.03	0.03
Fairness Through Unawareness	Logistic Regression	0.52	0.52	0.39	0.41	1.76	1.83	0.24	0.27
Reweighing	Logistic Regression	0.50	0.50	0.65	0.71	2.07	2.13	0.60	0.75

Podejścia pre-processing

Dalszy spadek jakości modelu + poprawa miar dyskryminacji

Eksperyment 1

Perspektywa krótkoterminowa

Metoda	Model bazowy	GINI *train*	GINI *test*	STP *train*	STP *test*	TPR *train*	TPR *test*	FPR *train*	FPR *test*
Unmitigated	Logistic Regression	0.61	0.60	0.03	0.03	0.07	0.08	0.03	0.03
Fairness Through Unawareness	Logistic Regression	0.52	0.52	0.39	0.41	1.76	1.83	0.24	0.27
Reweighing	Logistic Regression	0.50	0.50	0.65	0.71	2.07	2.13	0.60	0.75
Exponentiated Gradient Reduction *Demographic Parity*	Logistic Regression	0.08	0.07	1.71	1.86	2.01	2.17	0.89	1.08
Grid Search Reduction *Demographic Parity*	Logistic Regression	0.28	0.28	7.18	7.32	9.79	9.99	1.84	2.28

Niedopuszczalnie niski poziom jakości modelu decyzyjnego!

Eksperyment 1

Perspektywa krótkoterminowa

Metoda	Model bazowy	GINI *train*	GINI *test*	STP *train*	STP *test*	TPR *train*	TPR *test*	FPR *train*	FPR *test*
Unmitigated	Logistic Regression	0.61	0.60	0.03	0.03	0.07	0.08	0.03	0.03
	Decision Tree	0.85	0.68	0.09	0.09	0.26	0.19	0.12	0.12
Fairness Through Unawareness	Logistic Regression	0.52	0.52	0.39	0.41	1.76	1.83	0.24	0.27
	Decision Tree	0.79	0.63	0.58	0.88	1.27	1.50	3.00	1.77
Reweighing	Logistic Regression	0.50	0.50	0.65	0.71	2.07	2.13	0.60	0.75
	Decision Tree	0.81	0.65	0.59	0.65	1.26	1.15	2.64	1.12
Exponentiated Gradient Reduction *Demographic Parity*	Logistic Regression	0.08	0.07	1.71	1.86	2.01	2.17	0.89	1.08
	Decision Tree	0.42	0.26	0.95	1.05	2.04	1.79	10.18	2.08
Grid Search Reduction *Demographic Parity*	Logistic Regression	0.28	0.28	7.18	7.32	9.79	9.99	1.84	2.28
	Decision Tree	0.67	0.54	0.97	1.07	2.09	1.87	9.16	2.13

Eksperyment 1

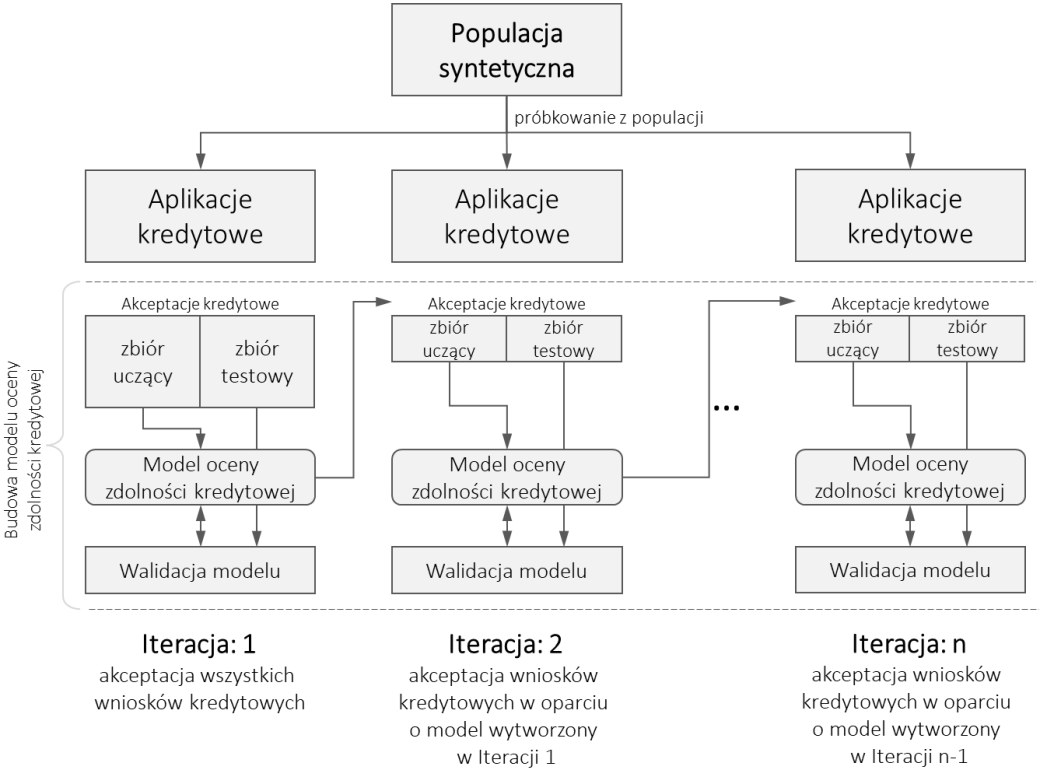
Perspektywa krótkoterminowa

Metoda	Model bazowy	GINI *train*	GINI *test*	STP *train*	STP *test*	TPR *train*	TPR *test*	FPR *train*	FPR *test*
Unmitigated	Logistic Regression	0.61	0.60	0.03	0.03	0.07	0.08	0.03	0.03
	Decision Tree	0.85	0.68	0.09	0.09	0.26	0.19	0.12	0.12
	Random Forest	0.90	0.77	0.04	0.04	0.16	0.10	0.03	0.04
	Gradient Boosting	1.00	0.74	0.29	0.14	0.95	0.29		0.19
Fairness Through Unawareness	Logistic Regression	0.52	0.52	0.39	0.41	1.76	1.83	0.24	0.27
	Decision Tree	0.79	0.63	0.58	0.88	1.27	1.50	3.00	1.77
	Random Forest	0.85	0.68	0.52	0.84	1.52	1.81	2.32	2.90
	Gradient Boosting	1.00	0.65	0.31	0.95	1.02	1.48		2.46
Reweighing	Logistic Regression	0.50	0.50	0.65	0.71	2.07	2.13	0.60	0.75
	Decision Tree	0.81	0.65	0.59	0.65	1.26	1.15	2.64	1.12
	Random Forest	0.88	0.73	0.39	0.41	1.01	0.81	3.04	1.25
	Gradient Boosting	1.00	0.73	0.31	0.24	1.01	0.45		0.42
Exponentiated Gradient Reduction *Demographic Parity*	Logistic Regression	0.08	0.07	1.71	1.86	2.01	2.17	0.89	1.08
	Decision Tree	0.42	0.26	0.95	1.05	2.04	1.79	10.18	2.08
	Random Forest	0.28	0.18	0.56	0.60	1.48	1.22	3.49	2.06
	Gradient Boosting	0.84	0.21	0.52	0.56	1.71	1.19		0.72
Grid Search Reduction *Demographic Parity*	Logistic Regression	0.28	0.28	7.18	7.32	9.79	9.99	1.84	2.28
	Decision Tree	0.67	0.54	0.97	1.07	2.09	1.87	9.16	2.13
	Random Forest	0.75	0.60	0.61	0.66	1.60	1.34	6.00	2.19
	Gradient Boosting	0.37	0.22	0.39	0.46	1.21	0.80		1.06
Adversarial Debiasing	AdversarialDebiasing	0.81	0.80	0.17	0.17	0.42	0.41	0.28	0.29

Schemat eksperymentu symulacyjnego (1/2)

Perspektywa długoterminowa

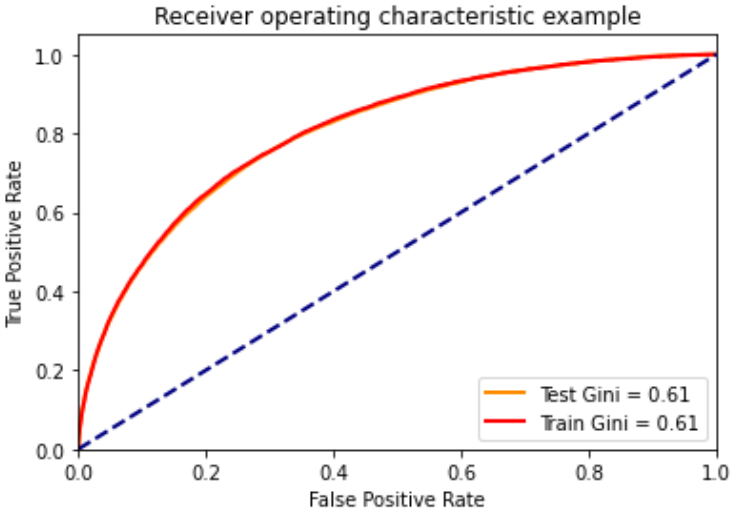
- wygenerowano populację agentów z łącznego rozkładu (populacja generyczna),
- z populacji generycznej losowane są próby agentów, którzy ubiegają się o kredyt (tj. aplikacje kredytowe),
- w pierwszej iteracji akceptowane są wszystkie wnioski kredytowe – bank pozyskuje bazę klientów, na której może zbudować model oceny zdolności kredytowej,
- w kolejnych iteracjach akceptowane są wyłącznie wnioski, dla których szacowane prawdopodobieństwo wejścia w stan niewykonania zobowiązania (tzw. PD – *Probability of Default*) jest mniejsze niż ustalona wartość progowa.

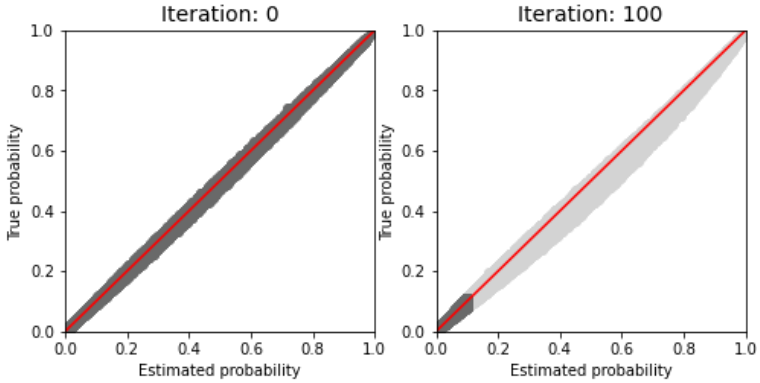


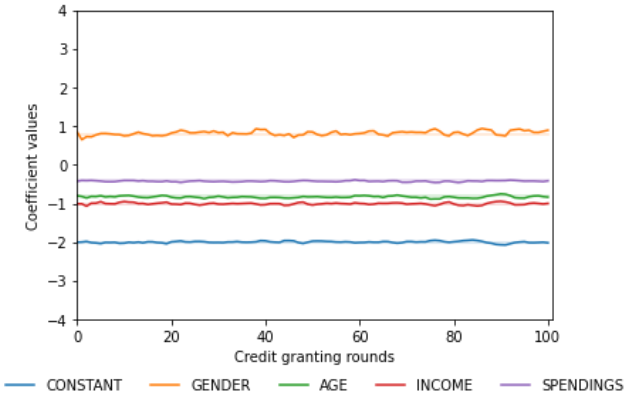
Scenariusz 1

Model „słabszy”

- oszacowany model zgodny z procesem generowania danych (tj. regresja logistyczna)
- brak korelacji między zmiennymi,
- wszystkie zmienne włączone do modelu,



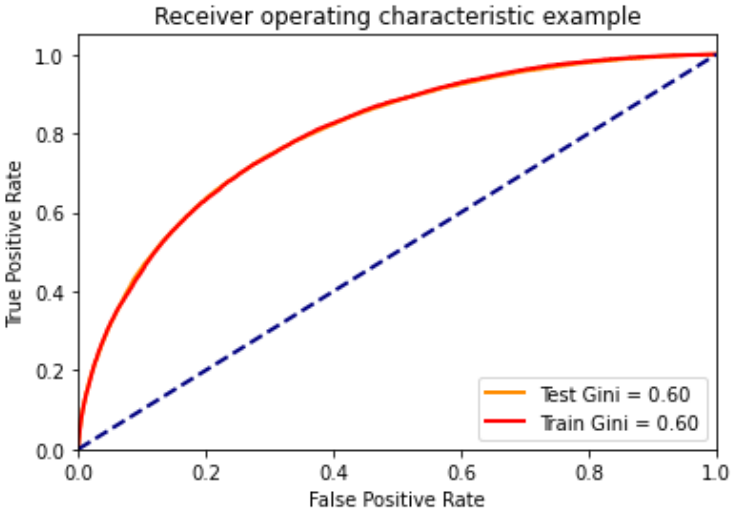


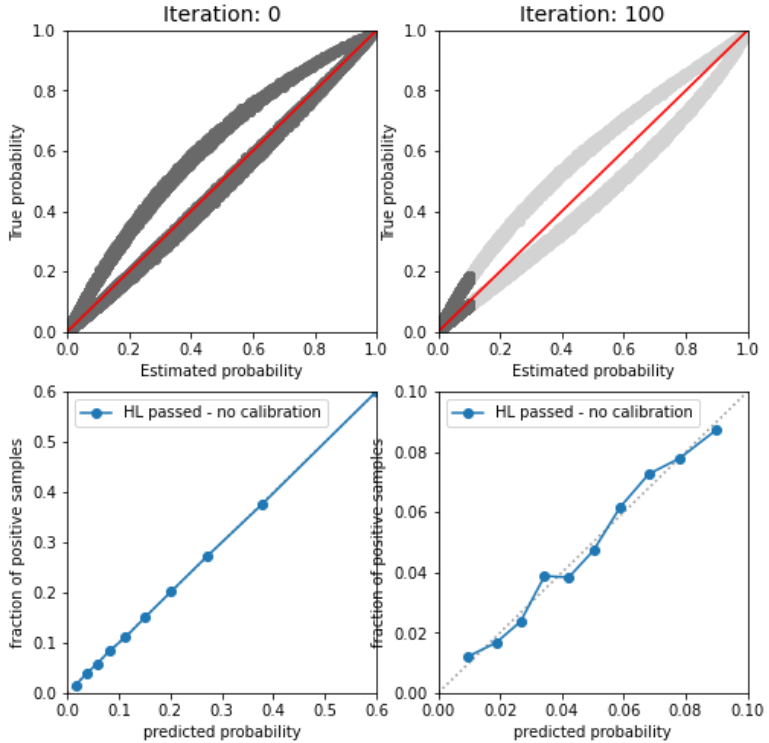


Scenariusz 2

Model „słabszy” – *fairness through unawareness*

- oszacowany model zgodny z procesem generowania danych (tj. regresja logistyczna)
- brak korelacji między zmiennymi,
- wyłączenie zmiennej „płeć” z modelu,

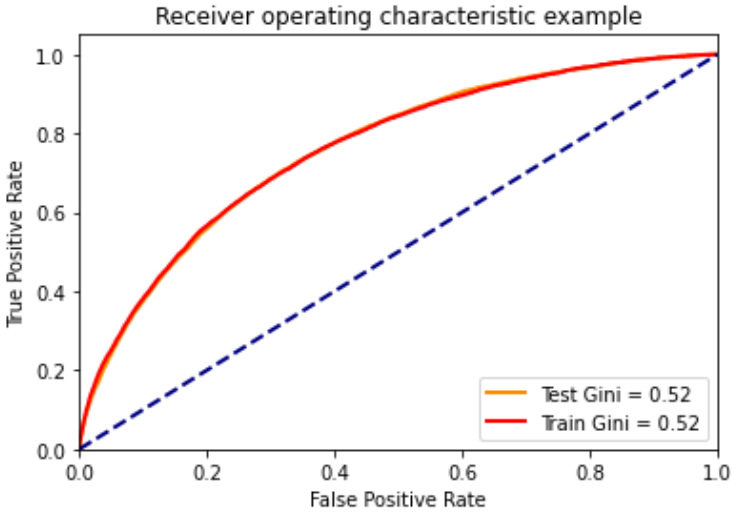


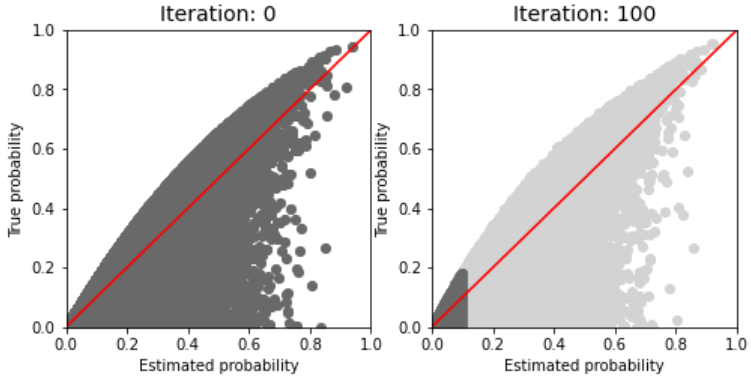


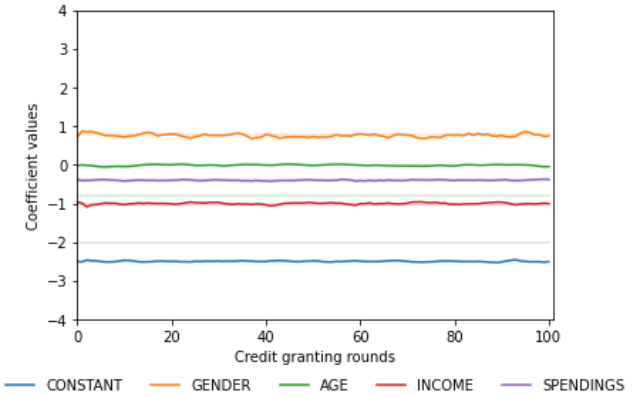
Scenariusz 3

Model o złej specyfikacji

- oszacowany model niezgodny z procesem generowania danych (tj. regresja logistyczna)
- brak korelacji między zmiennymi objaśniającymi,
- wszystkie zmienne włączone do modelu,



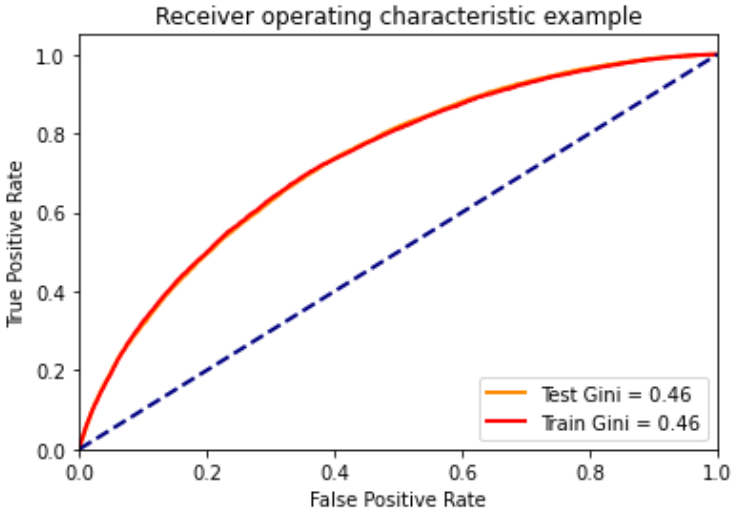


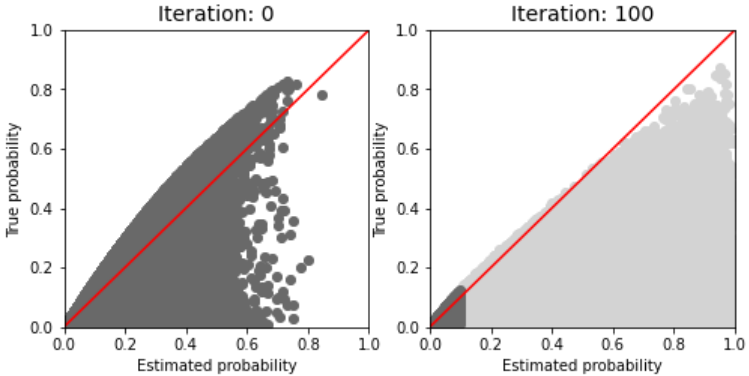


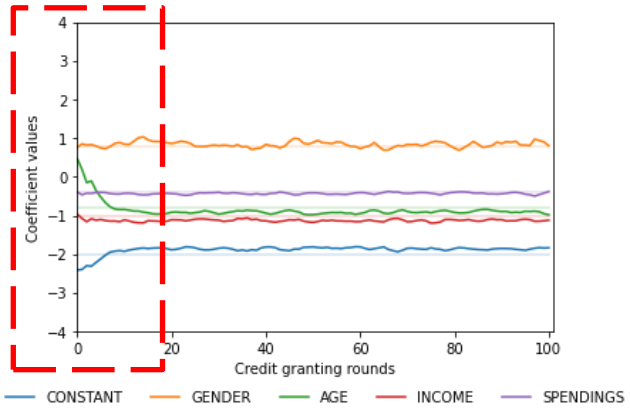
Scenariusz 4

Model o złej specyfikacji + silnej współliniowości

- oszacowany model niezgodny z procesem generowania danych (tj. regresja logistyczna)
- silna korelacja między zmiennymi objaśniającymi,
- wszystkie zmienne włączone do modelu,



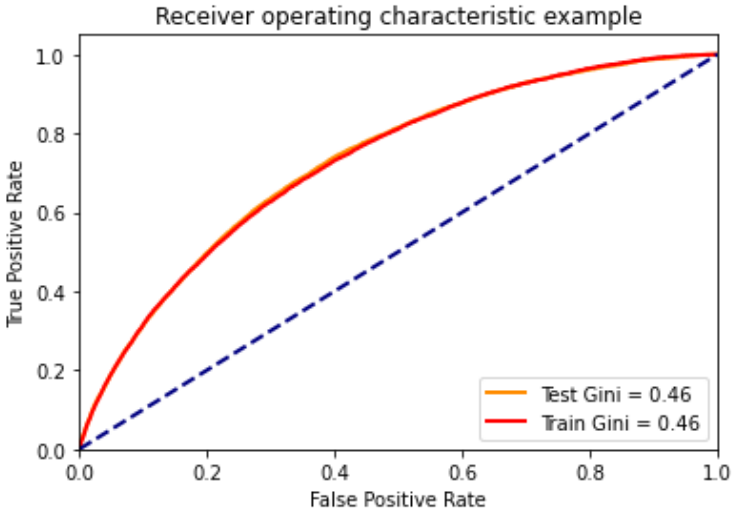


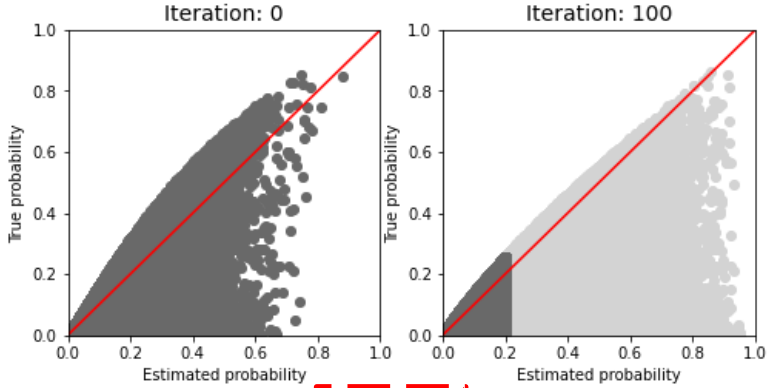


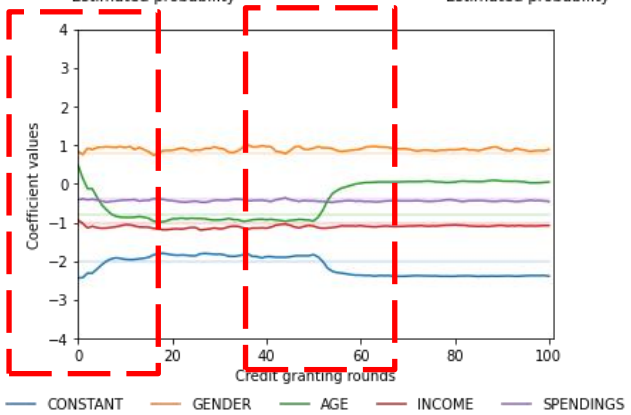
Scenariusz 5

Model o złej specyfikacji + silnej współliniowości + zmianie progu

- oszacowany model niezgodny z procesem generowania danych (tj. regresja logistyczna)
- silna korelacja między zmiennymi objaśniającymi,
- wszystkie zmienne włączone do modelu,







Bibliografia (1/2)

- Klimontowicz, M., 2017. Bankowość dla praktyków. Europejski Certyfikat Bankowca, EFCB 3E.
- Schreiner, M., 2003. Scoring: the next breakthrough in microcredit. Occasional paper, 7.
- European Commission , 2019. Ethics guidelines for trustworthy AI.
- Przanowski, K., 2014. Rola danych symulacyjnych w badaniach Credit Scoring, Monografia „Statystyka w służbie biznesu i nauk społecznych", Wydawnictwo Wyższej Szkoły Menedżerskiej w Warszawie.
- Zhou, N., Zhang, Z., Nair, V. N., Singhal, H., Chen, J. & Sudjianto, A., 2021. Bias, fairness, and accountability with ai and ml algorithms, arXiv preprint arXiv:2105.06558
- Kleinberg, J., Mullainathan, S., Raghavan, M., 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:1609.05807v2.
- Barocas, S., Hardt, M., & Narayanan, A., 2019. Fairness and Machine Learning. fairmlbook.org.
- Hardt, M., Price, E., & Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3315–3323).
- Kozodoi, N., Jacob, J., Lessmann, S., 2022. Fairness in Credit Scroing: Assessment, Implementation and Profit Implications, arXiv preprint arXiv: 2103.01907v4.

Bibliografia (2/2)

- Hunter N., 2018, Statutory Protections for Individual Rights, in: The Law of Emergencies (Second Edition), Butterworth-Heinemann
- Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P. and Roth, D., 2019, January. A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the conference on fairness, accountability, and transparency (pp. 329-338).
- Kamishima, T., Akaho, S., Asoh, H. and Sakuma, J., 2012, Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 35-50). Springer, Berlin, Heidelberg.
- Zafar, M.B., Valera, I., Rogriguez, M.G. and Gummadi, K.P., 2017, Fairness constraints: Mechanisms for fair classification. In Artificial Intelligence and Statistics (pp. 962-970). PMLR.
- Calders, T. and Verwer, S., 2010. Three naive bayes approaches for discrimination-free classification. Data mining and knowledge discovery, 21(2), pp.277-292.
- Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K.N. and Varshney, K.R., 2017, December. Optimized pre-processing for discrimination prevention. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 3995-4004).
- Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S., 2015, August. Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268).
- Kamiran, F., Calders, T. and Pechenizkiy, M., 2010, December. Discrimination aware decision tree learning. In 2010 IEEE International Conference on Data Mining (pp. 869-874). IEEE.