Kamila Wachulec
Michał Tałałaj

17 listopad 2025r.

Analiza modeli predykcyjnych

Porównanie skuteczności modeli predykcyjnych

- Cel: Budowa narzędzia, które rozdziela klientów na ryzykownych i rzetelnych
- Podejście: Porównanie Regresji Logistycznej vs metody XGBoost
- Odpowiedź: Jak modele zachowają się dla konkretnych danych finansowych

Przygotowanie danych

Model Regresji Logistycznej

1. Pobieramy dane, dzielimy je na zbiór treningowy i testowy
2. Dane kategoryczne za pomocą OHE stają się numeryczne, rodzaj działalności gospodarczej zamieniany jest na miarę WOE
3. Braki danych są uzupełniane medianą oraz tworzone są kolumny wskaźnikowe dla braków
4. Nadmiernie skorelowane kolumny są usuwane
4. Zmienne są skalowane do modelu regresji

Model XGBoost

1. Pobieramy dane, dzielimy je na zbiór treningowy, walidacyjny i testowy
2. Dane kategoryczne za pomocą OHE stają się numeryczne, rodzaj działalności gospodarczej zamieniany jest na miarę WOE
3. Dane numeryczne są uzupełniane, pozbawiane nieskończoności i oznaczane
4. Zapisanie nowych danych, zwracając uwagę na brak wycieków danych

Podejście modeli

Model Regresji Logistycznej

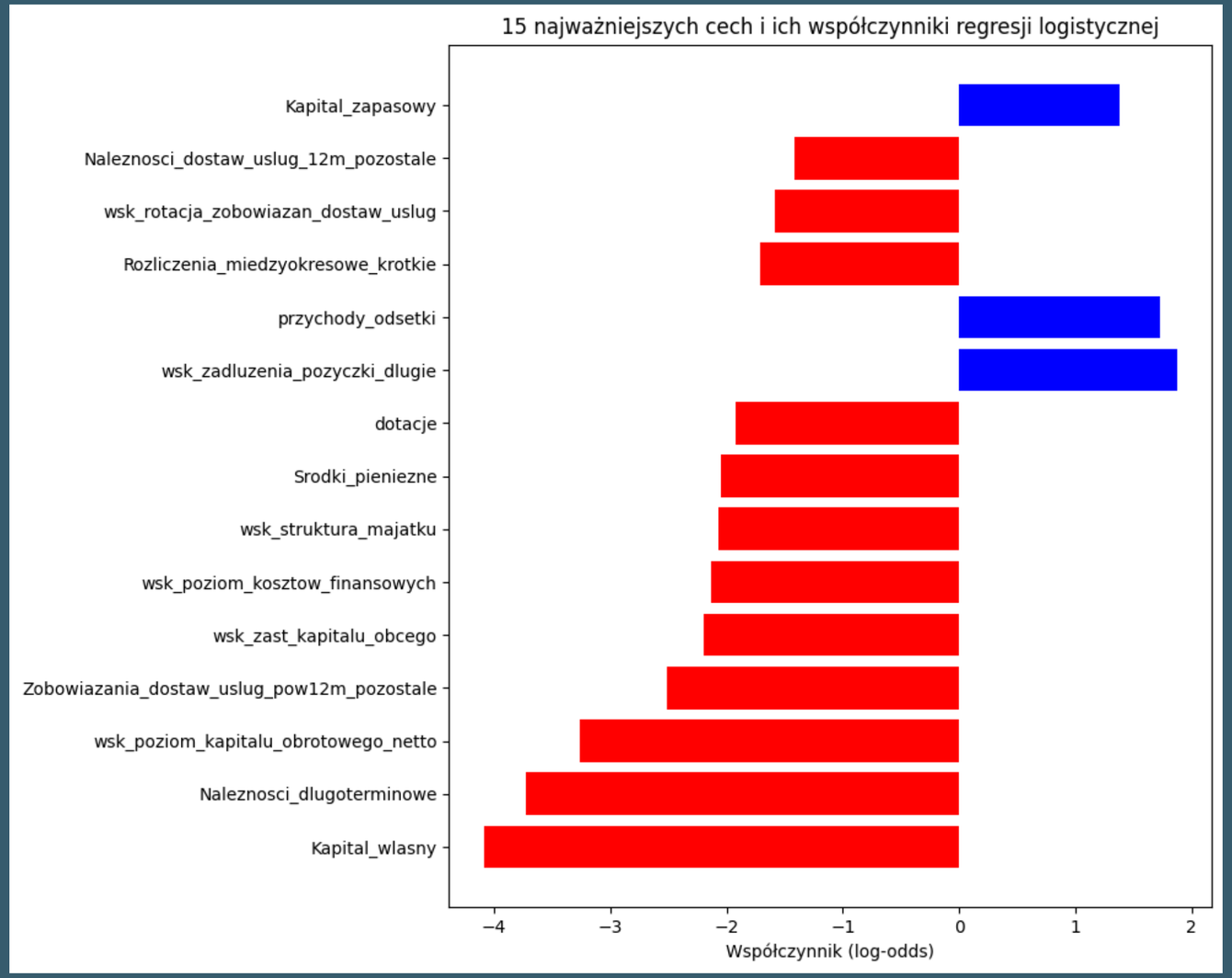
- Na podstawie zestawu cech wejściowych przypisuje każdej z cech wagę, który mówi, jak mocno dana cecha wpływa na prawdopodobieństwo niespłaty kredytu.
- Nie widzi złożonych interakcji między danymi i zakłada, że każdy z czynników działa osobno.

Model XGBoost

- Sekwencja pytań Tak/Nie w określonej kolejności, które tworzy to rozbudowane drzewo decyzyjne.
- Potrafi połączyć fakty, których regresja nie widzi, czyli zależności między zmiennymi.
- Model buduje się iteracyjnie, a każde kolejne drzewo w modelu skupia się na naprawianiu błędów popełnionych przez poprzednie, co zwiększa precyzję w trudnych przypadkach.

Analiza współczynników – regresja

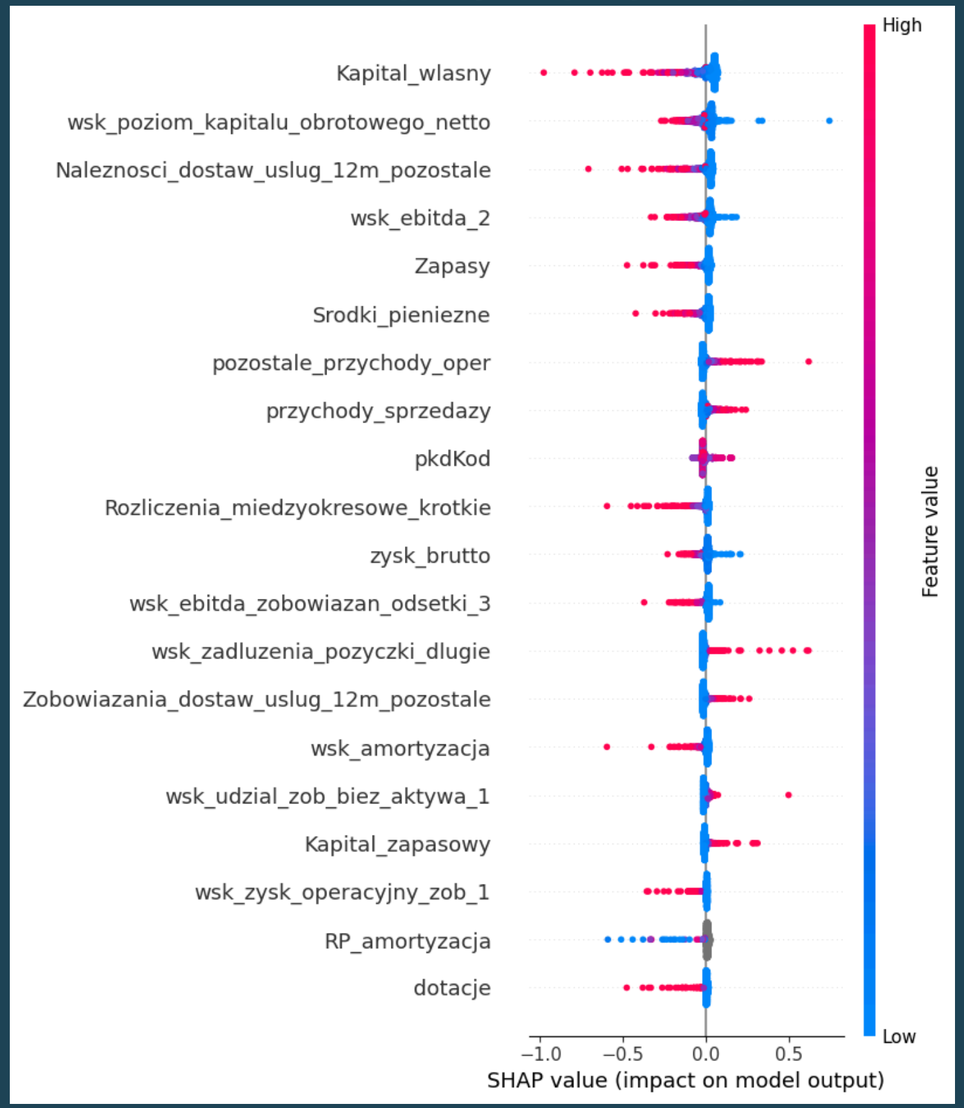
- Wartość bezwzględna współczynników pokazuje “siłę” z jaką każda z cech wpływa na predykcję modelu.
- Ujemne wartości zwiększają prawdopodobieństwo spłaty, a dodatnie szansę niespłaty.



Sterowanie ryzykiem

Wpływ cech na decyzję.

Model Regresji Logistycznej



SHAP > 0

zmienna zwiększa ryzyko niespłaty

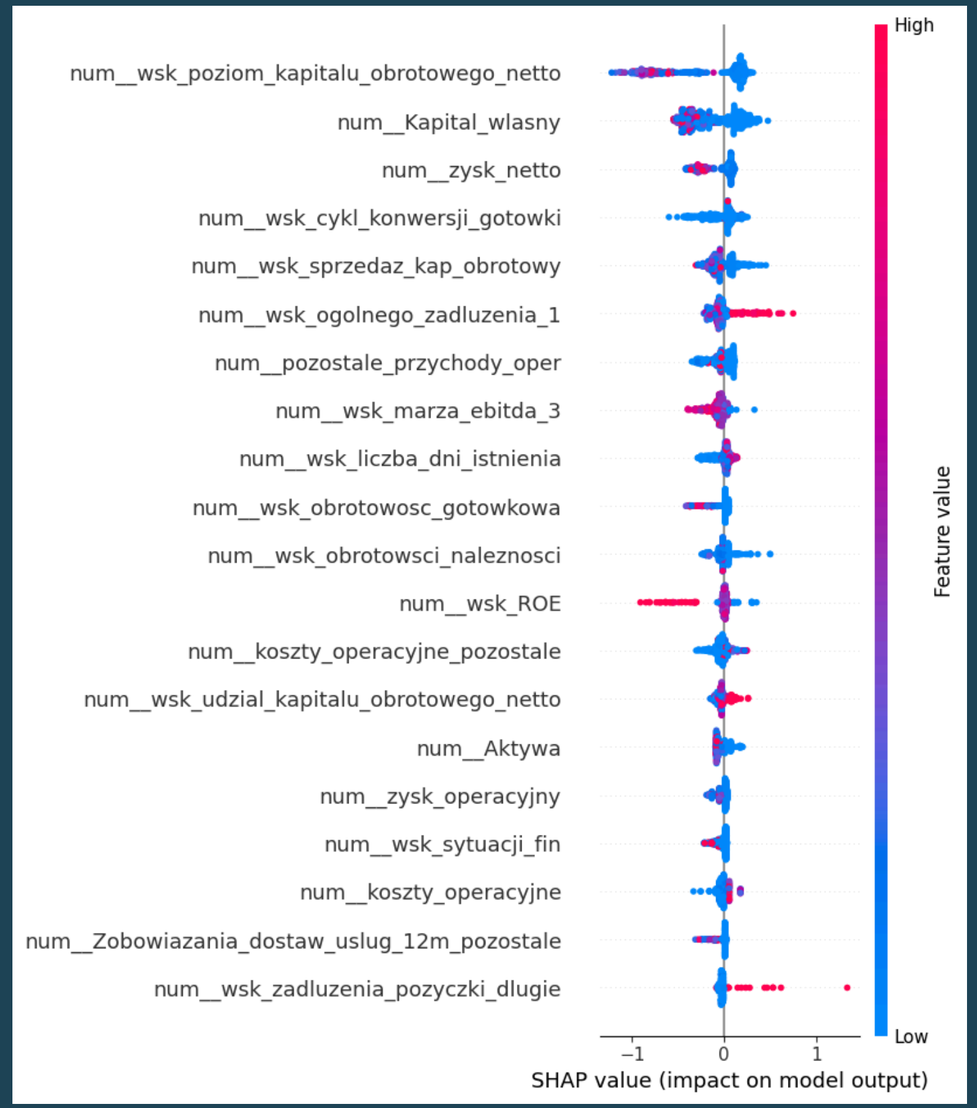
SHAP < 0

zmienna zmniejsza ryzyko niespłaty

Obserwacje blisko zera:

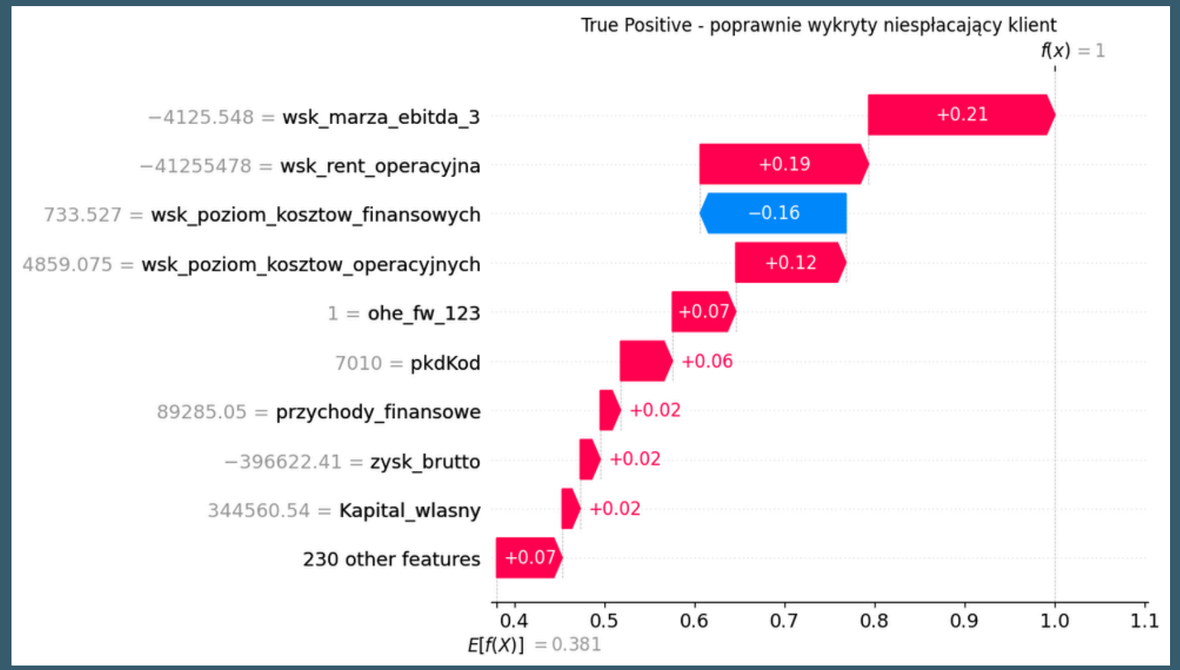
cecha nie ma istotnego wpływu na wynik dla wielu firm

Model XGBoost



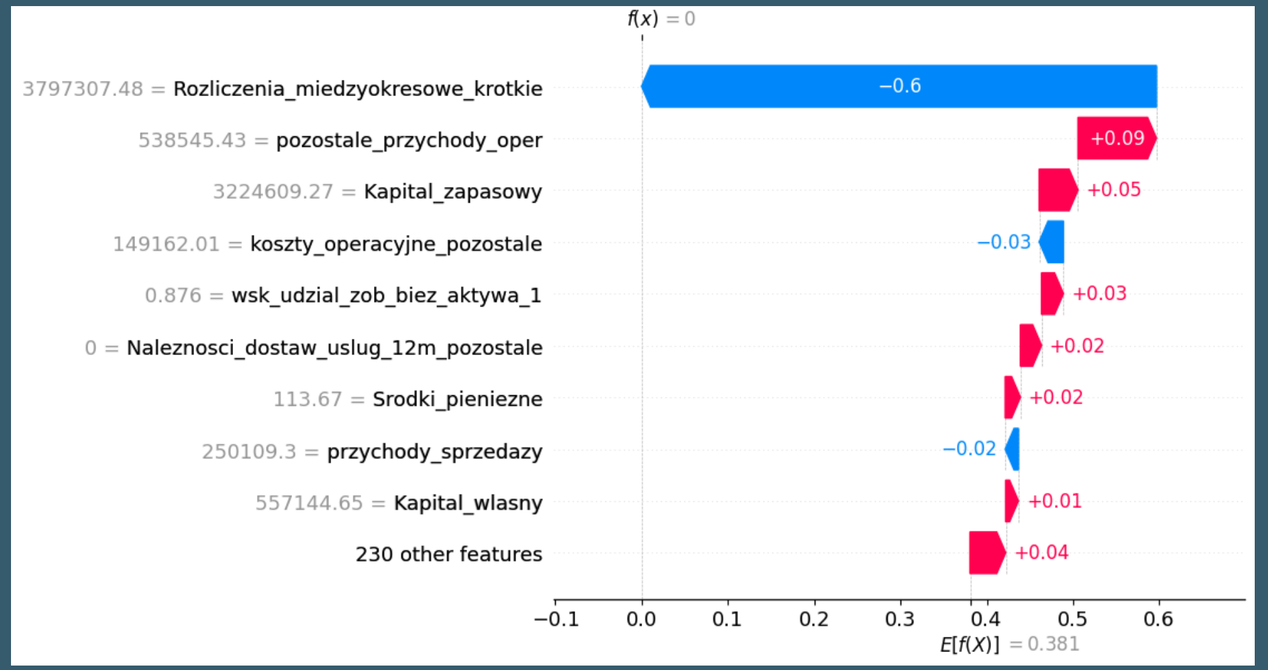
Case studies - regresja

Cechy TP: poprawne założenie niespłaty kredytu



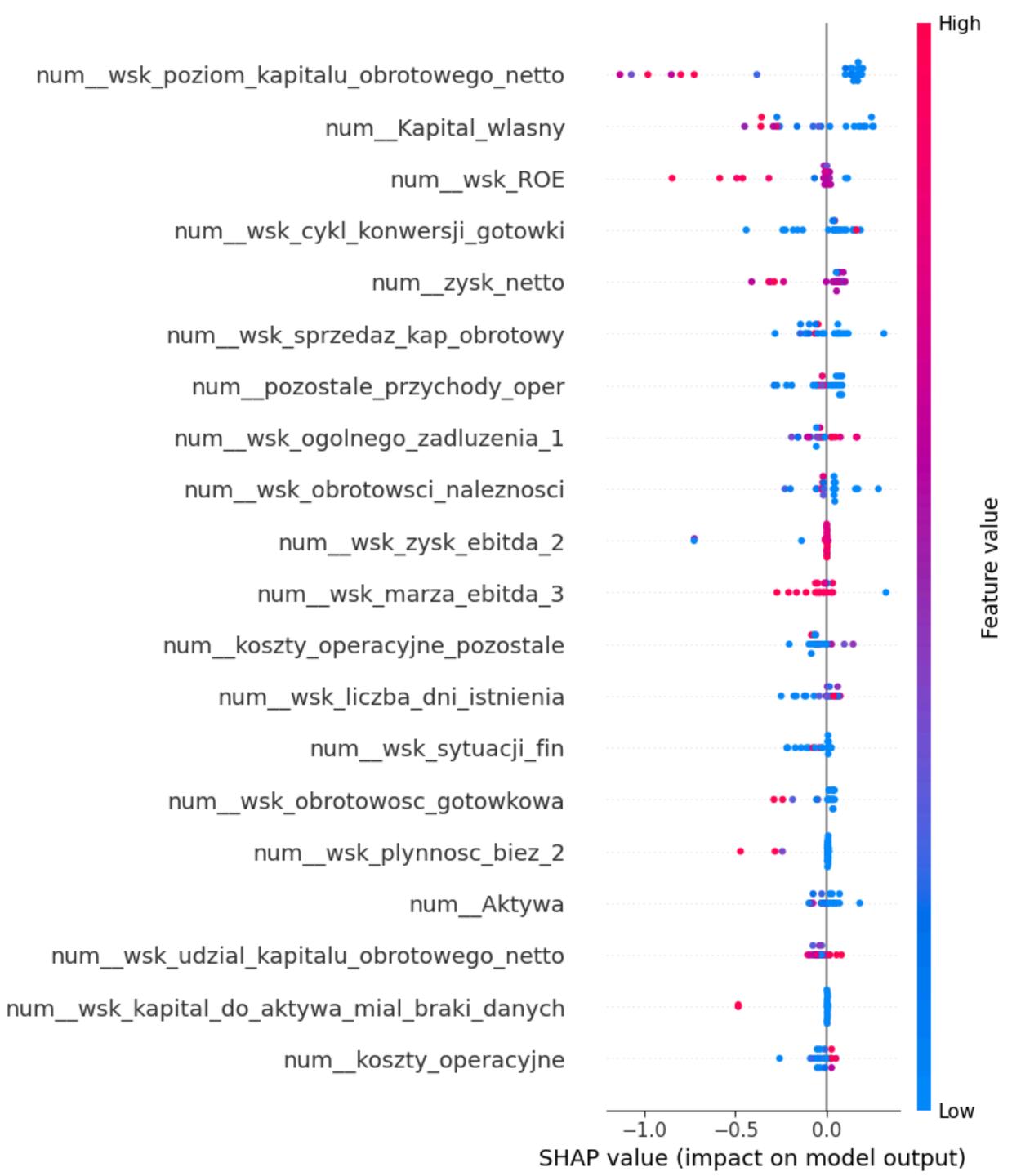
Firma generuje wielokrotnie wyższe straty niż posiadany kapitał własny. Struktura kosztów wskazuje na zaawansowaną niewypłacalność. Ryzyko upadłości jest ekstremalnie wysokie.

Cechy FN: błędne założenie spłaty kredytu



Symptomy złej kondycji firmy były zbyt słabe, aby skompensować dominującą wartość pozytywnej cechy. Firma została zaklasyfikowana jako bezpieczna, mimo że jej sytuacja w rzeczywistości nie była stabilna.

Analiza FN - XGBoost



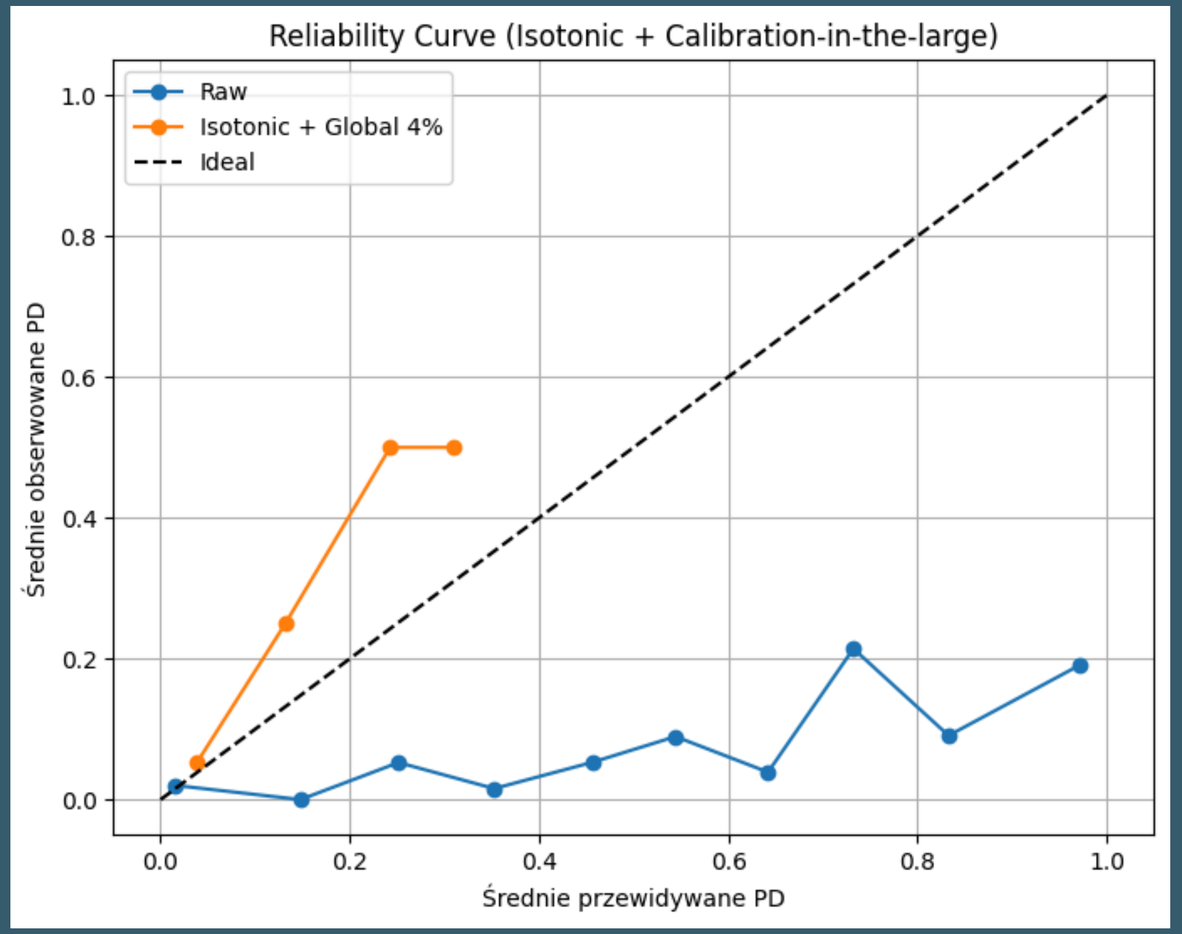
Błędne założenie spłaty kredytu

Poprzez wyizolowanie przypadków błędnego stwierdzenia, że klient spłaci kredyt, możemy wyznaczyć statystyczny wpływ konkretnych cech wyłącznie dla tej grupy.

Wykres ten wskazuje, dlaczego ten konkretny podzbiór klientów, mimo że w rzeczywistości ryzykowny, został uznany przez model za bezpieczny.

Dodatkowo wiemy teraz, na jakie dane należy uważać przy rozpatrywaniu sytuacji brzegowych.

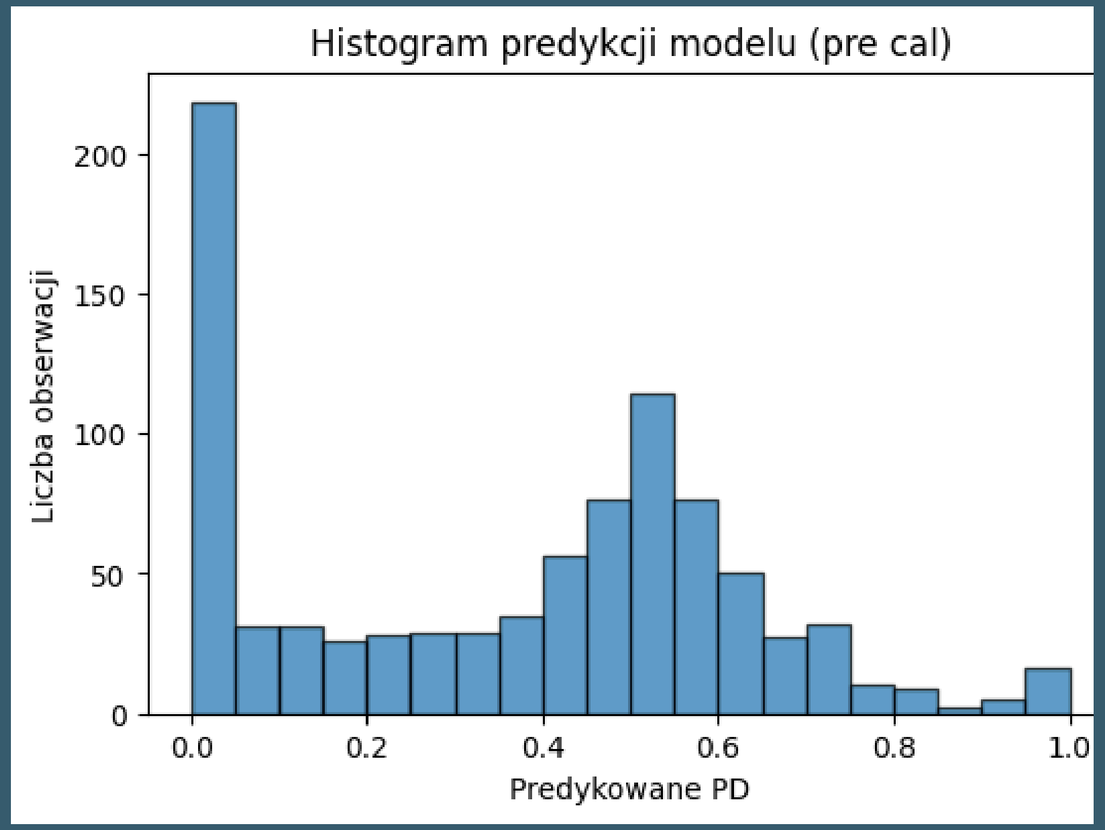
Kalibracja 4% - regresja

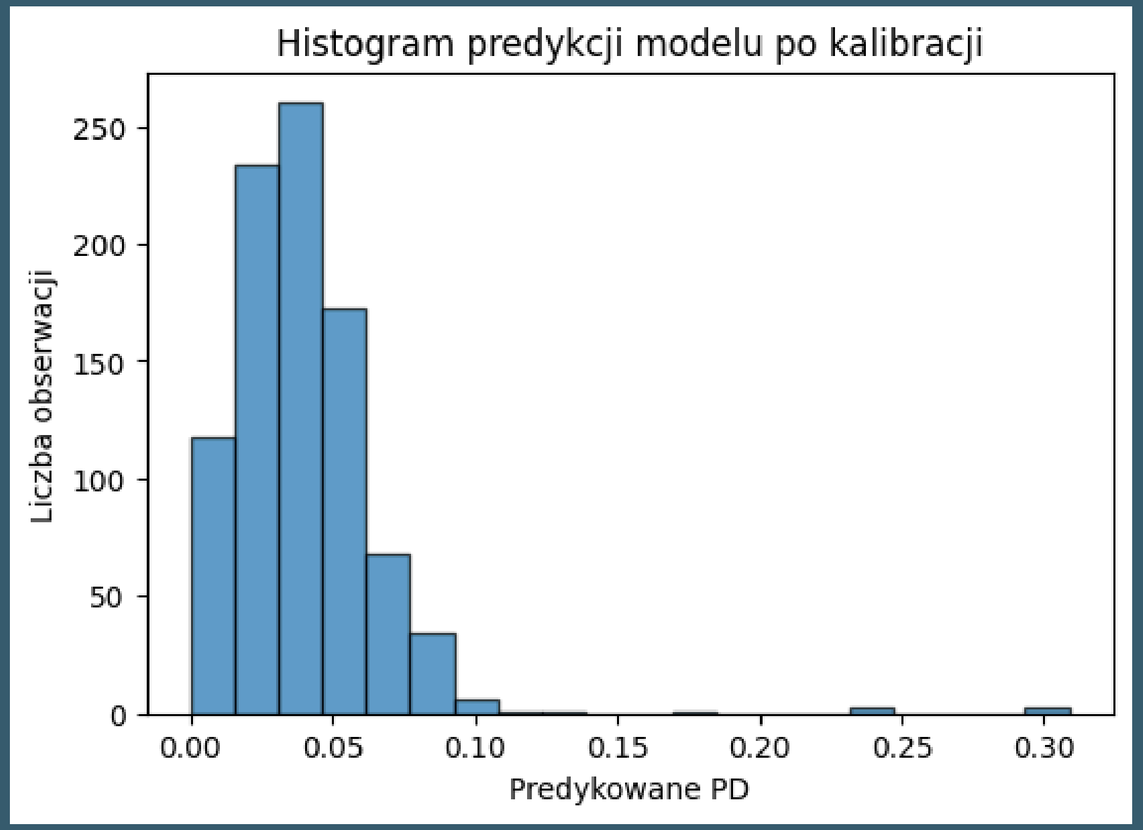


Model przed kalibracją wyraźnie zaniżał prawdopodobieństwo niespłaty

Kalibracja 4% - regresja

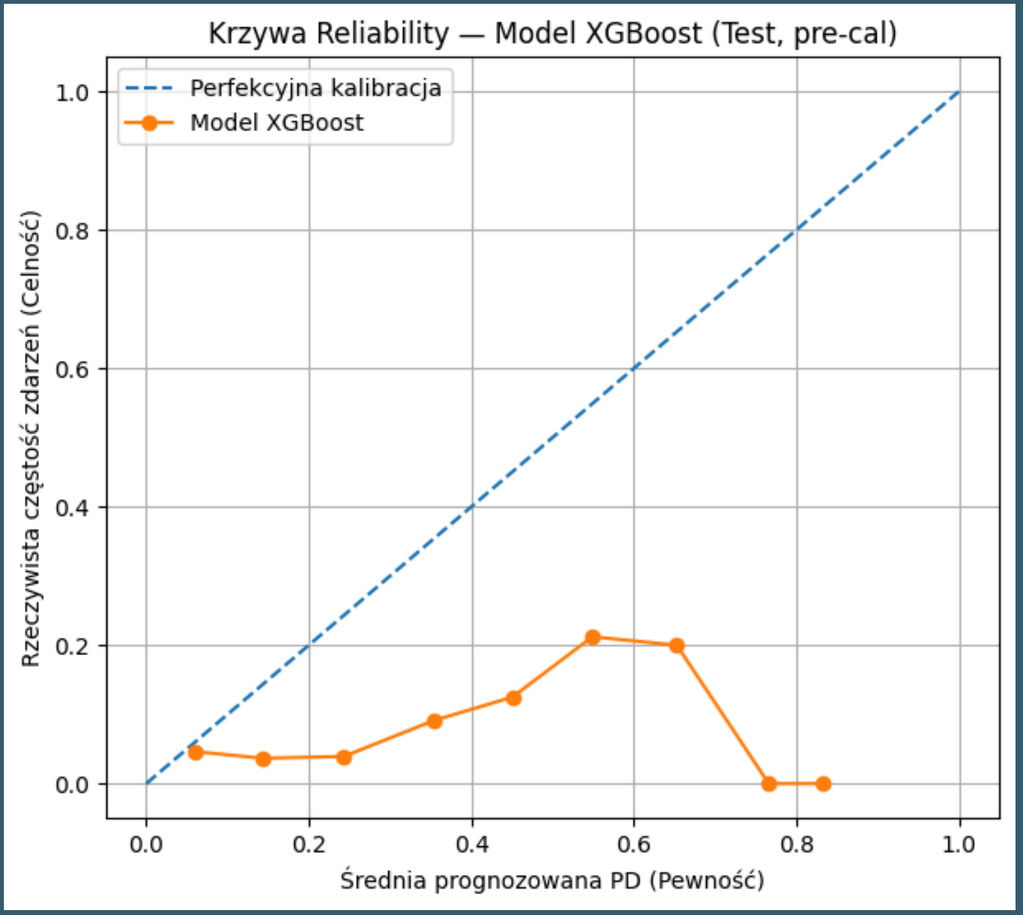
Wpływ kalibracji na wartości predykcji

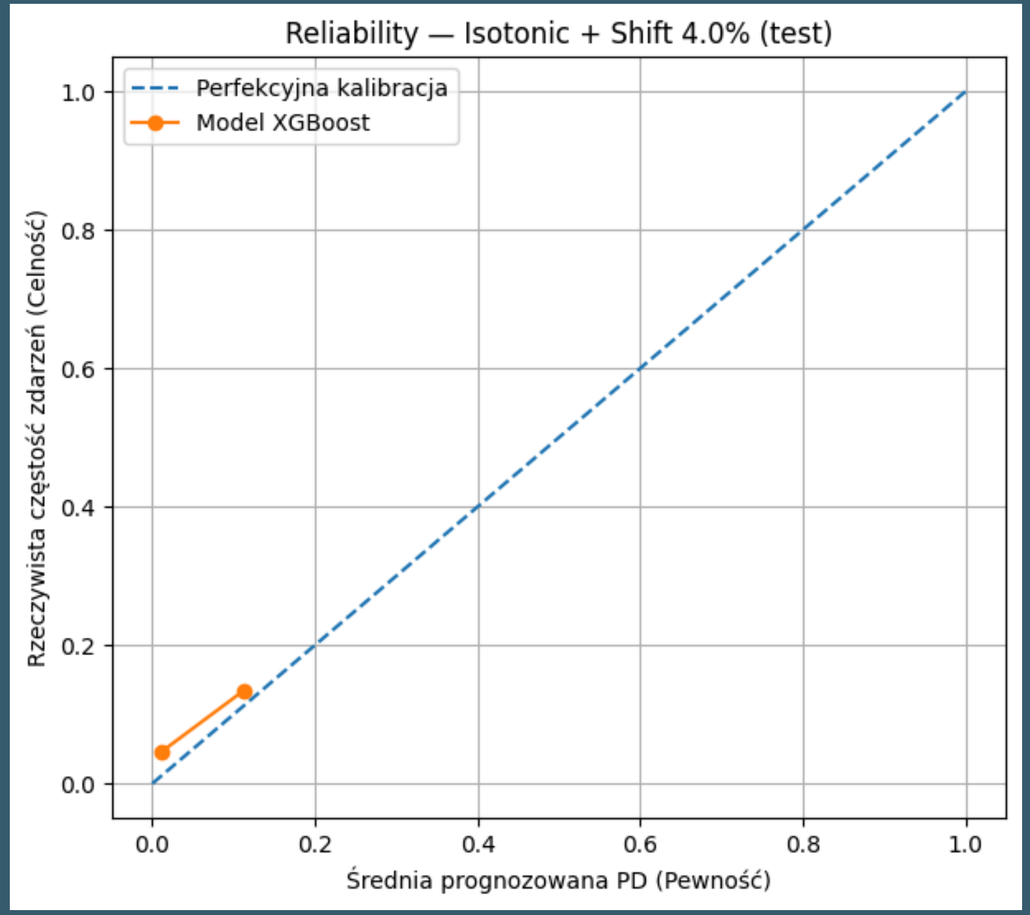




Kalibracja 4% - XGBoost

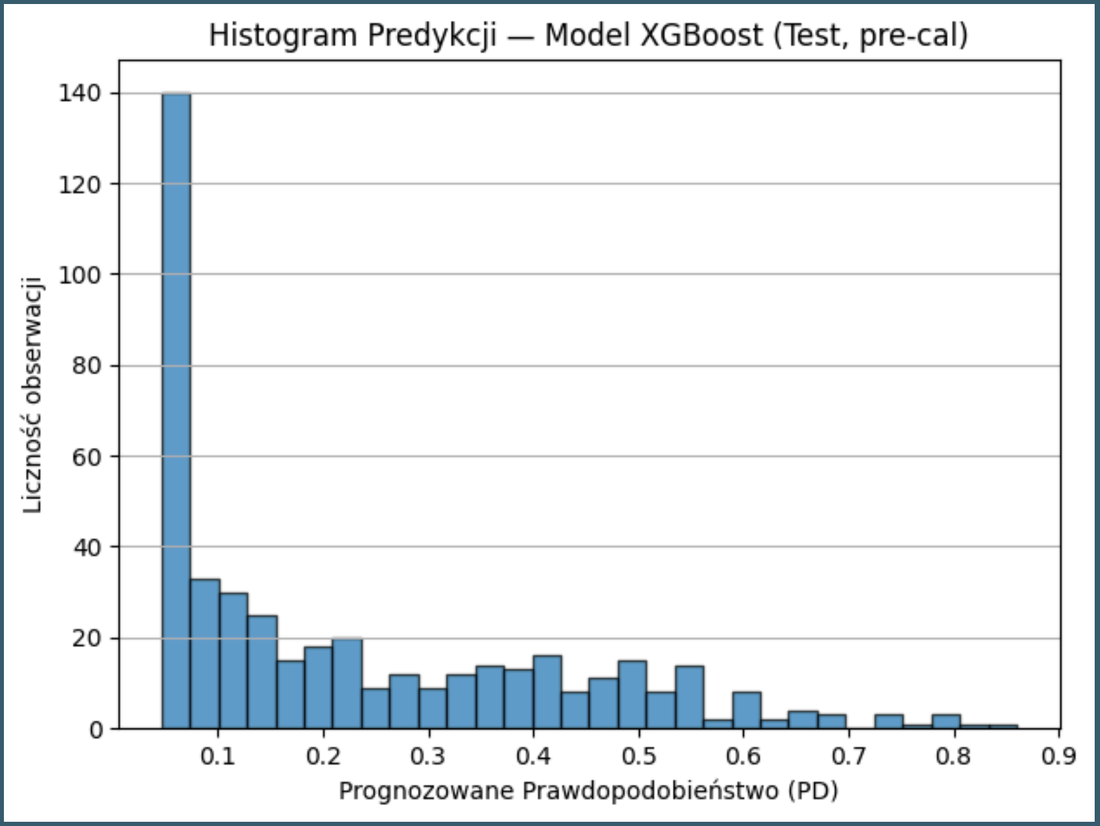
Widzimy zbyt wielką pewność modelu przed kalibracją i przesunięciem.

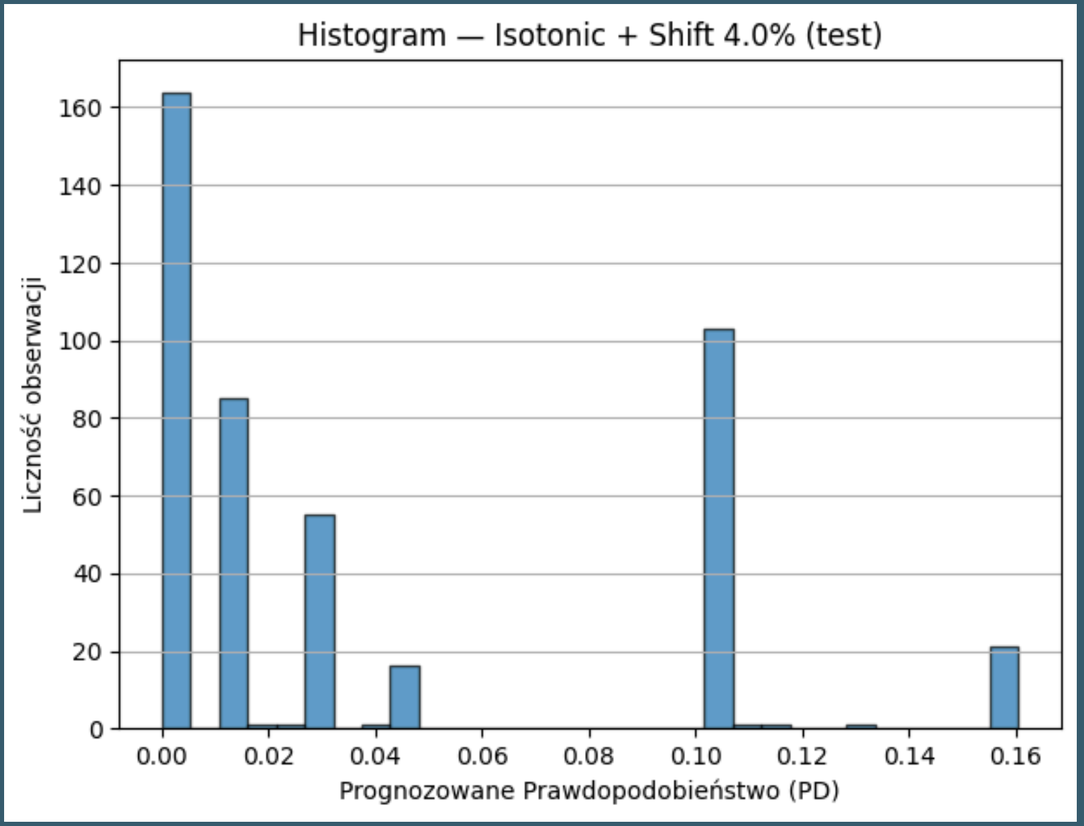




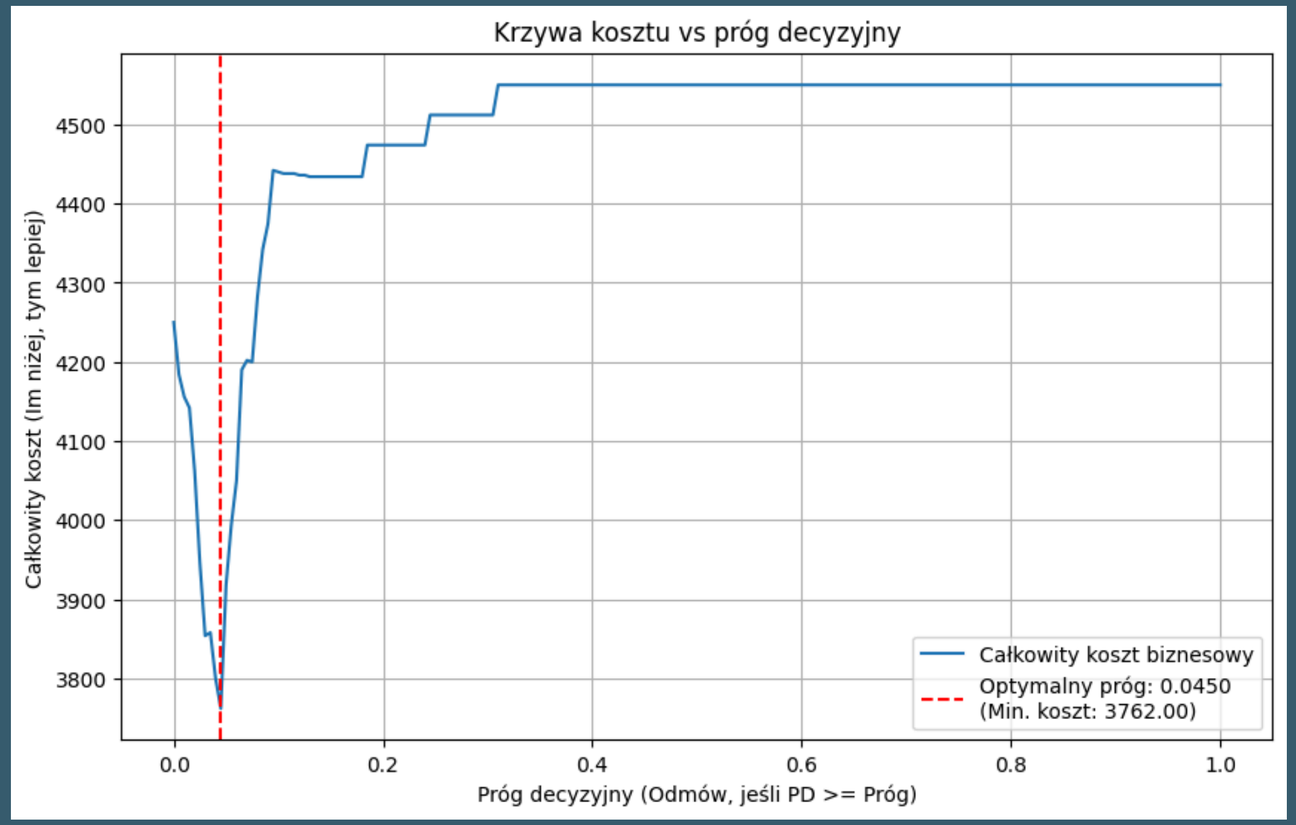
Kalibracja 4% - XGBoost

Widzimy jak kalibracja wprowadziła schody dla wartości.





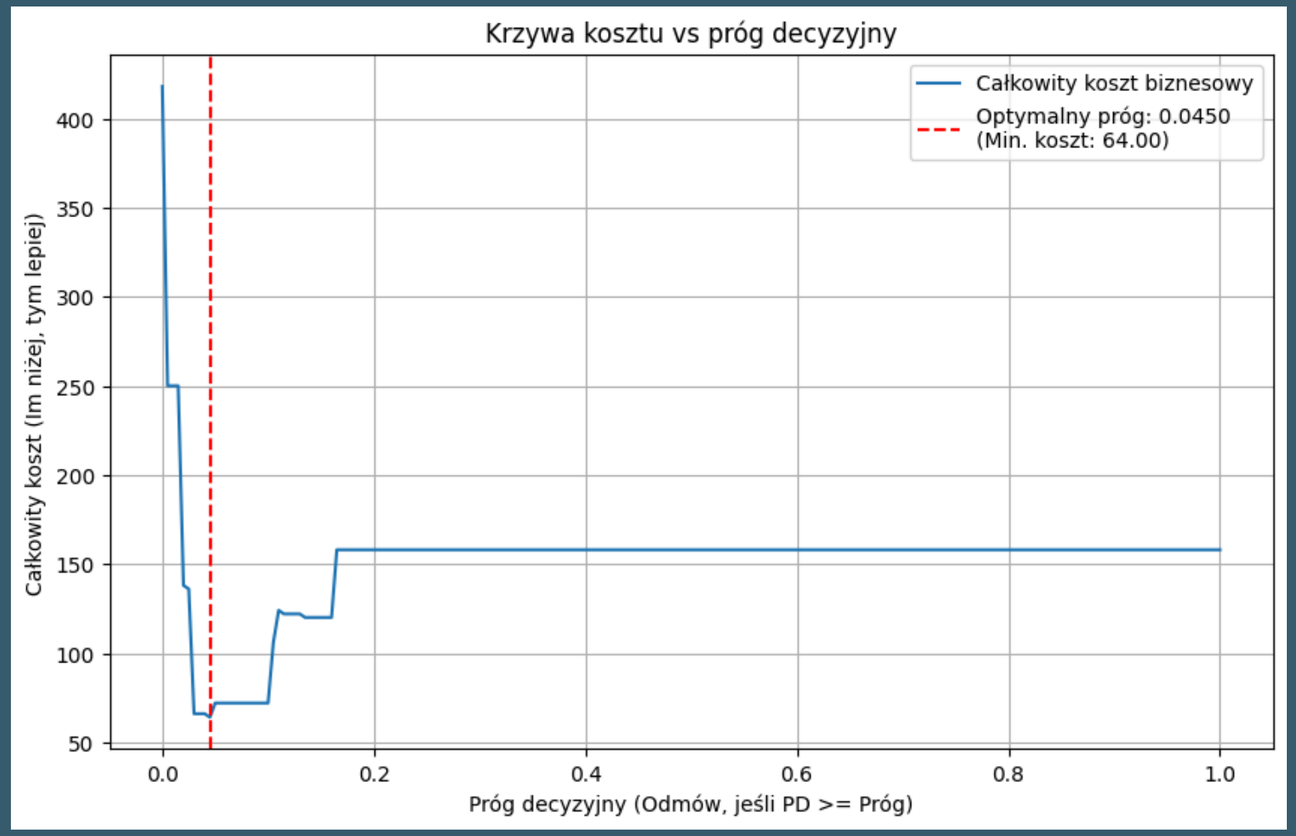
Krzywe kosztu i progi decyzyjne – LR



Optymalny próg: 0.0450
Stopa akceptacji: 57.44%

- **Zbyt niski próg – odrzucamy wielu dobrych klientów, co zmniejsza zyski.**
- **Zbyt wysoki próg – akceptujemy zbyt wielu ryzykownych klientów, co zwiększa straty.**

Krzywe kosztu i progi decyzyjne - XGBoost



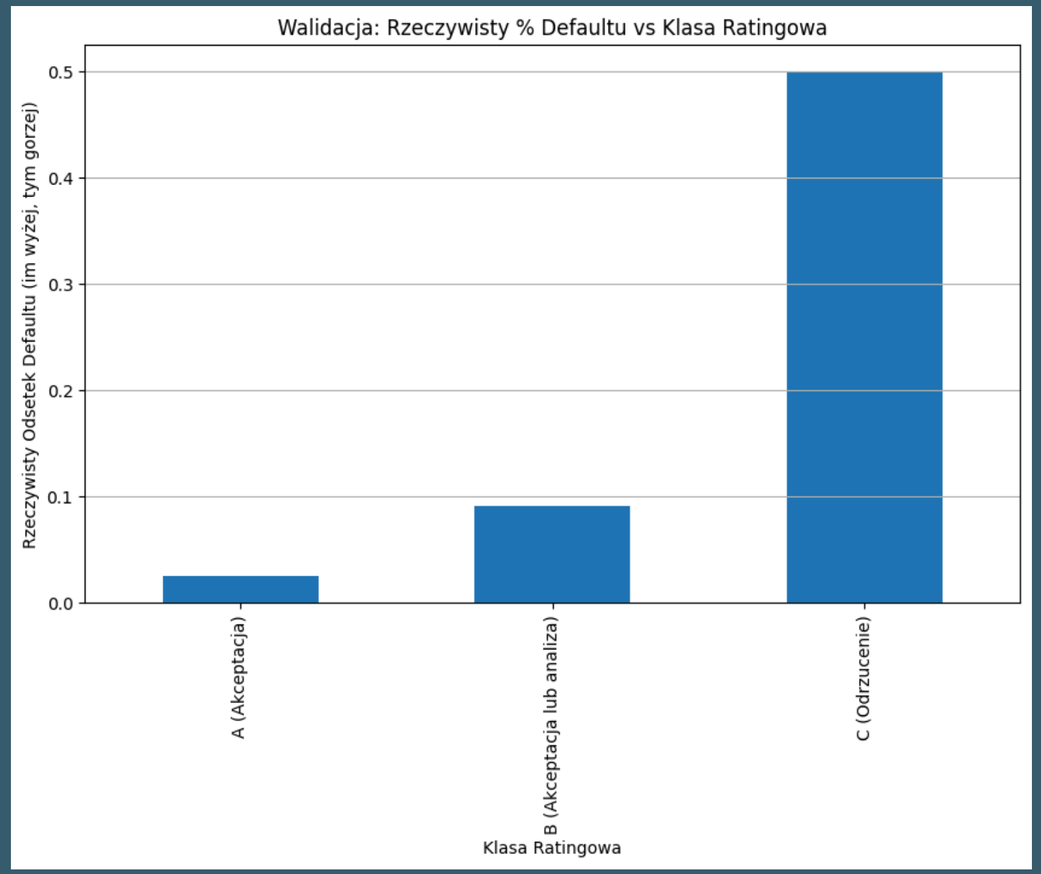
Optymalny próg: 0.0450 ← to samo co dla LR
Stopa akceptacji: 68.22%

Macierz kosztów decyzyjnych (proporcje):
TP: 0.0 FP: 1.0 FN: 18.0 TN = -1.0

Mamy na celu znalezienie progu przy którym zminimalizujemy koszt podejmowanych decyzji w oparciu o model.

Klasy ratingowe

Model Regresji Logistycznej



Model XGBoost