



Analiza i budowa modelu predykcyjnego w oparciu o dane finansowe oraz wskaźniki finansowe

Wojciech Klusek, Aleksander Kuś



Przetwarzanie danych



Wstępne przetwarzanie danych

- Wszystkie zmienne kategoryczne zostały przekształcone przy użyciu metody One Hot Encoding, co pozwala na przetwarzanie tych zmiennych przez algorytmy uczenia maszynowego.
- Wszystkie brakujące wartości (NaN) w danych zostały zastąpione medianą odpowiednich kolumn, co pomaga w utrzymaniu integralności danych bez wprowadzania znaczących zakłóceń.
- Wartości nieskończone (Inf oraz -Inf) zostały zastąpione największą i najmniejszą wartością w danej kolumnie, odpowiednio, aby uniknąć problemów z obliczeniami.
- Zastosowano metodę identyfikacji i usuwania wartości odstających w kolumnach numerycznych. Metoda ta definiuje outliery jako wartości leżące poza przedziałem $[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$, gdzie Q1 to pierwszy kwartył, Q3 to trzeci kwartył, a IQR to rozstęp międzykwartyłowy.



WOE, IV

- WOE to technika statystyczna używana do oceny siły predykcyjnej poszczególnych zmiennych w modelach uczenia maszynowego. W naszym przypadku, zmienne numeryczne zostały podzielone na różne zakresy, a dla każdego z tych zakresów obliczono WOE, który mierzy, jak mocno dany zakres różni się od pozostałych pod względem predykcji wyniku.
- IV to miara używana do określenia siły predykcyjnej całej zmiennej, bazująca na obliczonym WOE. Wyższa wartość IV oznacza, że zmienna ma większe znaczenie predykcyjne.
- Usunięte zostały kolumny, które miały niską wartość IV, co oznacza, że ich zdolność do przewidywania wyniku jest ograniczona. Pozostawienie jedynie zmiennych o wysokiej wartości IV pozwala na zbudowanie bardziej efektywnego i skoncentrowanego modelu.

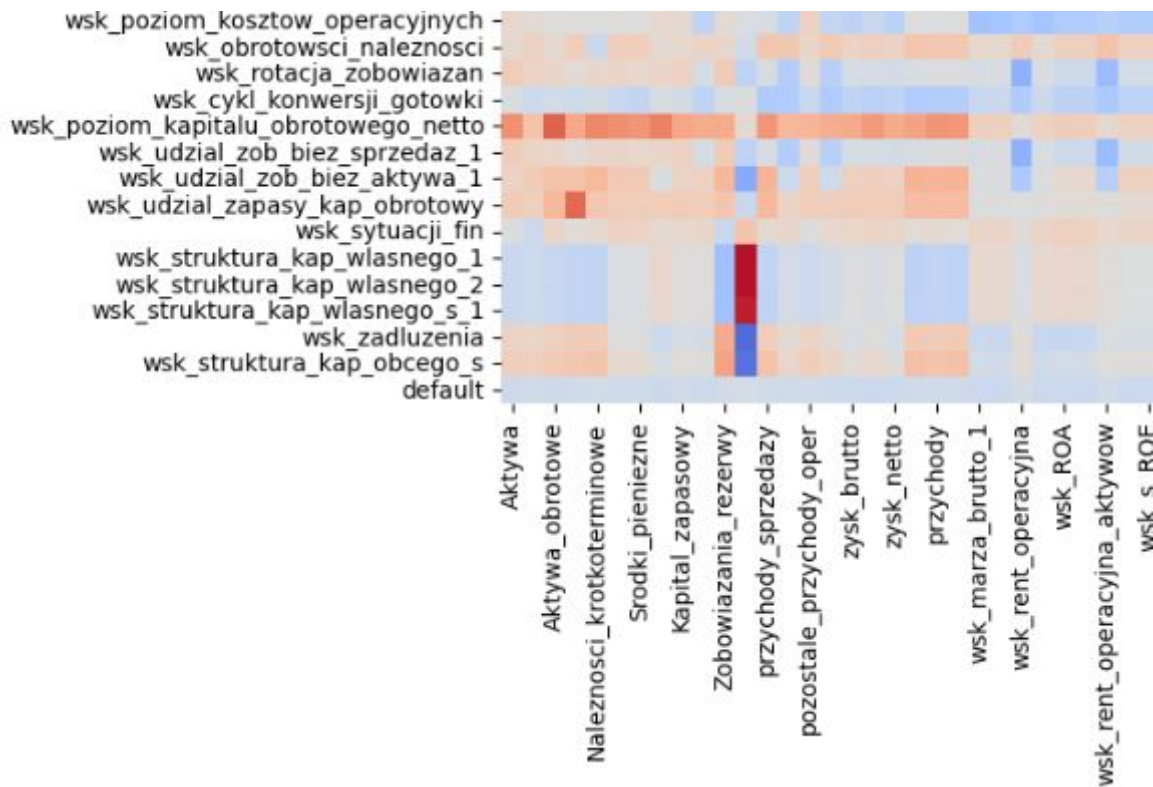


Test ANOVA

- ANOVA analizuje różnice w średnich wartościach między grupami i określa, czy te różnice są statystycznie znaczące.
- Test ANOVA został wykorzystany do oceny, czy różne zmienne w naszych danych mają statystycznie istotny wpływ na przewidywane zjawisko (w tym przypadku 'default').
- Test ANOVA pomógł nam zidentyfikować i usunąć zmienne, które miały niewielki lub żaden wpływ na przewidywany wynik, co pozwoliło na uproszczenie modelu i skoncentrowanie się na bardziej istotnych zmiennych.

Korelacja zmiennych

- Zbadaliśmy korelację pomiędzy zmiennymi w naszym zestawie danych, aby zidentyfikować pary zmiennych, które są silnie skorelowane.
- W przypadku identyfikacji par zmiennych o wysokiej korelacji (powyżej 95%), podjęliśmy decyzję o usunięciu zmiennej o mniejszej IV z pary.



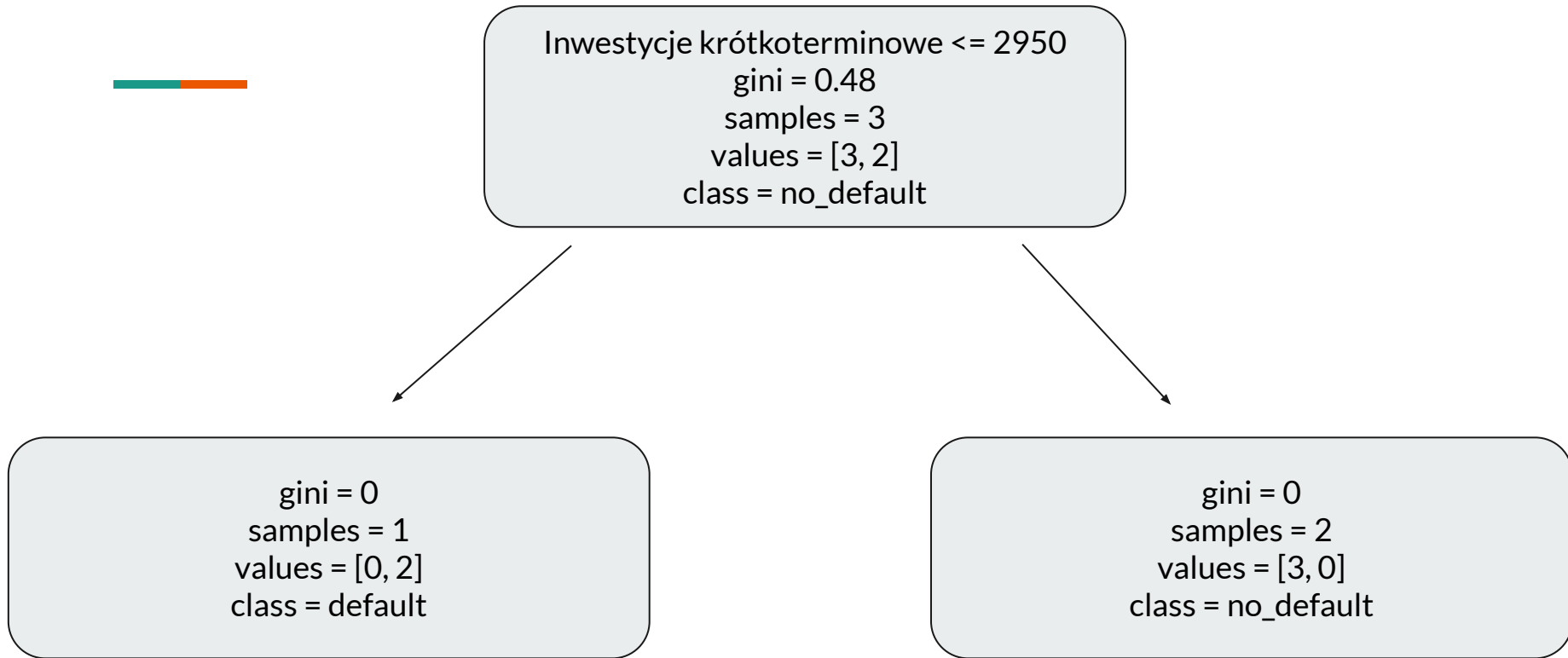


Model interpretowalny




Las losowy

- Las losowy to algorytm uczenia maszynowego, który wykorzystuje zespół drzew decyzyjnych do generowania prognoz. Każde drzewo w lesie dokonuje indywidualnej predykcji, a wynik lasu losowego jest średnią lub najczęściej występującą predykcją wszystkich drzew.
- W lasach losowych każde drzewo jest trenowane na nieco innym zestawie danych.
- Jeden z głównych atutów lasu losowego to jego zdolność do redukcji przesadnego dopasowania, które często występuje w przypadku pojedynczych drzew decyzyjnych.
- Model ten umożliwia także ocenę, które cechy mają największy wpływ na prognozowane wyniki, co jest przydatne w analizie danych i interpretacji modelu.





Wyniki modelu interpretowalnego



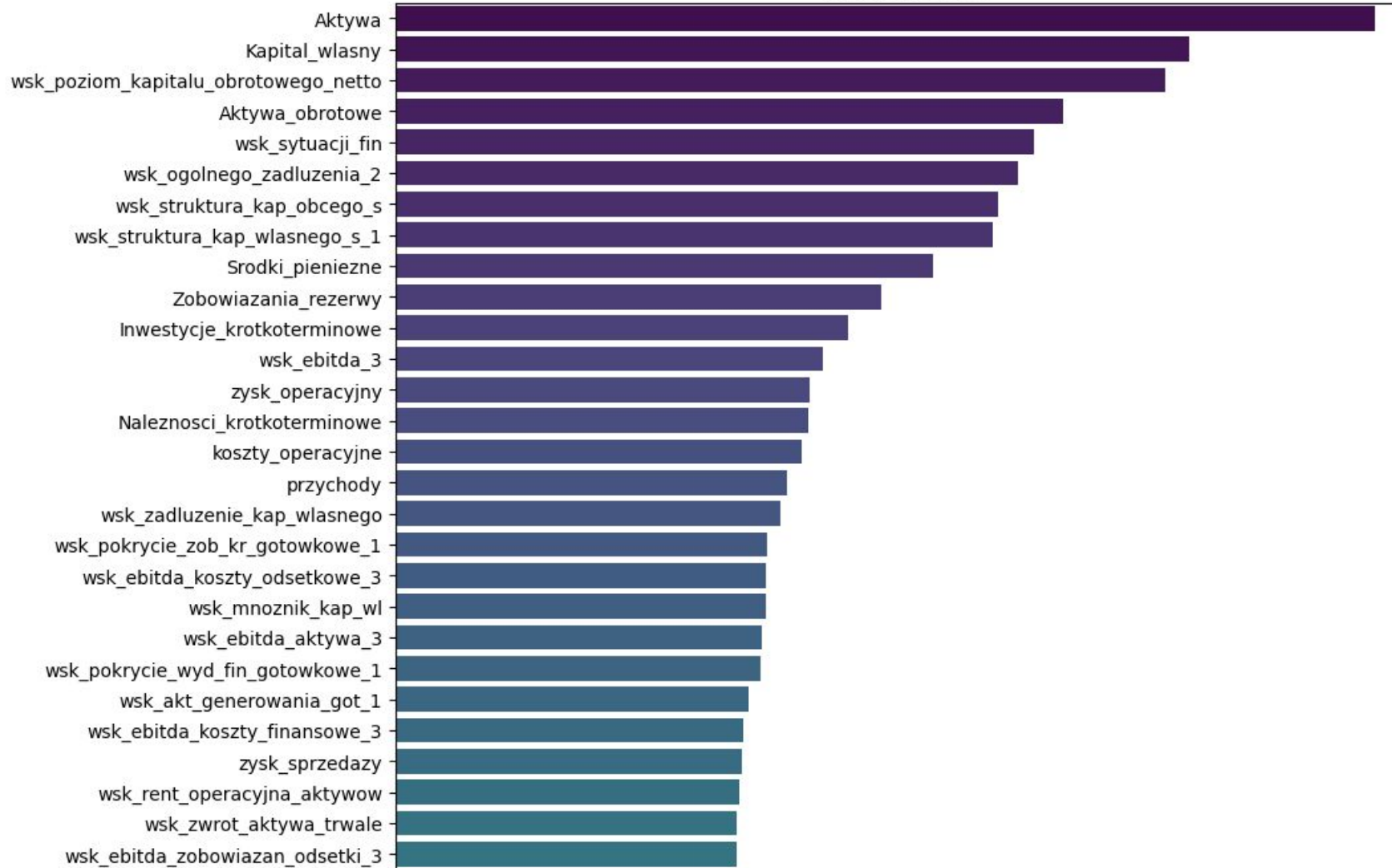
Accuracy: 0.95

AUC: 0.73

Gini: 0.45

	precision	recall	f1-score	support
0	0.95	1.00	0.97	568
1	1.00	0.03	0.06	32
accuracy			0.95	600
macro avg	0.97	0.52	0.52	600
weighted avg	0.95	0.95	0.92	600

Random forest





Model black-box



XGBoost

- XGBoost (eXtreme Gradient Boosting) to algorytm uczenia maszynowego, który korzysta z techniki wzmacniania gradientowego do optymalizacji modeli predykcyjnych.
- XGBoost buduje model w sposób iteracyjny, dodając kolejne drzewa, gdzie każde kolejne drzewo próbuje skorygować błędy popełnione przez poprzednie drzewa. Algorytm ten minimalizuje funkcję straty, stosując algorytm gradientowego spadku.
- XGBoost jest znany z szybkości działania i wydajności, co wynika z optymalizacji wykorzystywanych algorytmów i struktur danych, jak również z możliwości równoległego przetwarzania.




PCA

- PCA to technika statystyczna używana do uproszczenia złożoności danych, którą wykorzystaliśmy w modelu black-box. Pozwala ona na redukcję liczby zmiennych w zestawie danych, zachowując przy tym jak najwięcej informacji.
- PCA przekształca zestaw oryginalnych zmiennych w nowy zestaw zmiennych, które są ze sobą niepowiązane (tzw. składowe główne). Te nowe zmienne są kombinacją oryginalnych zmiennych i są wybierane tak, aby najbardziej różnić się od siebie pod względem posiadanych informacji.
- Jedną z konsekwencji użycia PCA jest utrata interpretowalności modelu. Z uwagi na to, że składowe główne są kombinacją oryginalnych zmiennych, trudno jest bezpośrednio zrozumieć, jak poszczególne oryginalne zmienne wpływają na wyniki modelu.



Wyniki modelu black-box



Accuracy: 0.94

AUC: 0.73

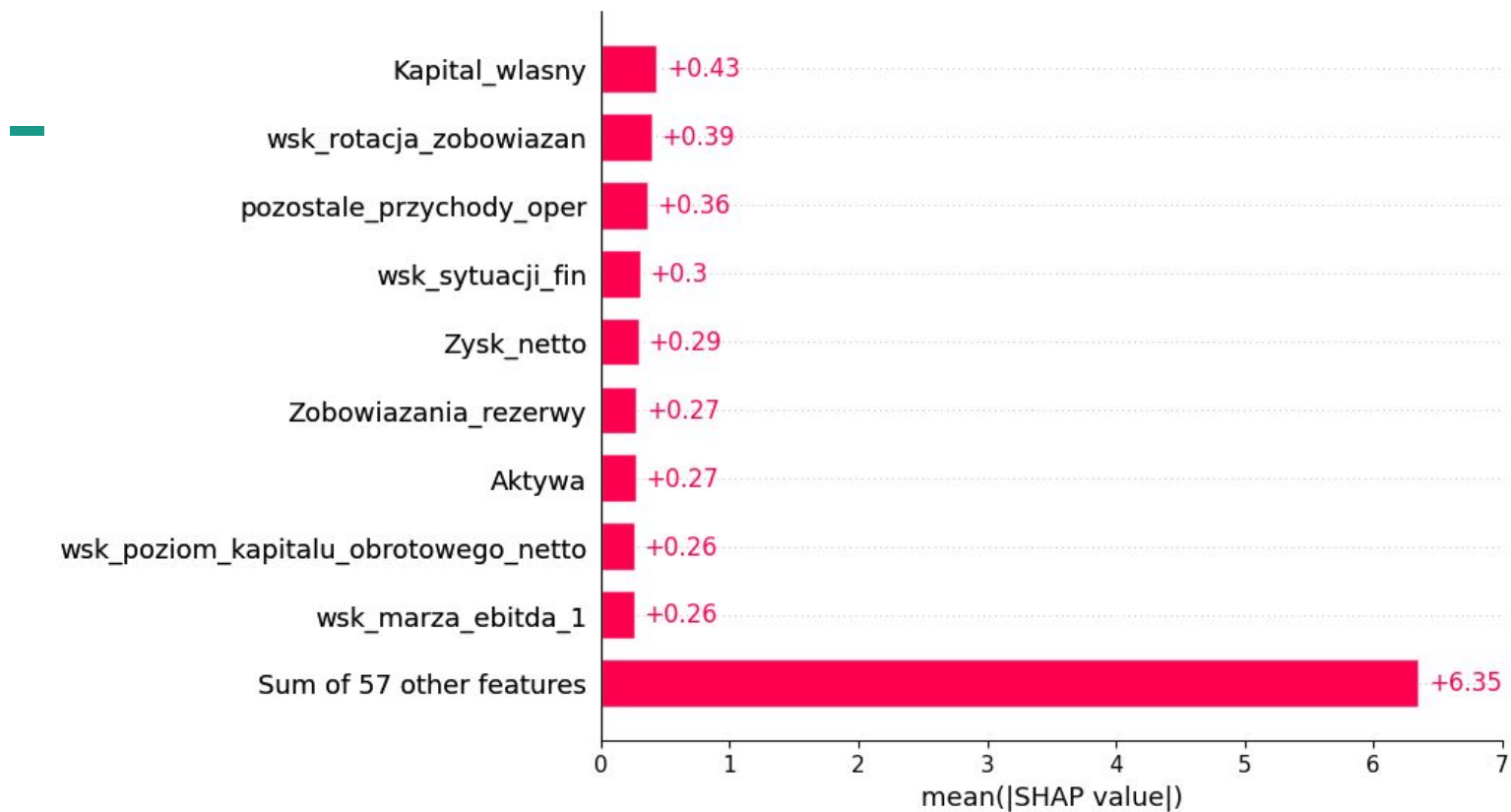
Gini: 0.47

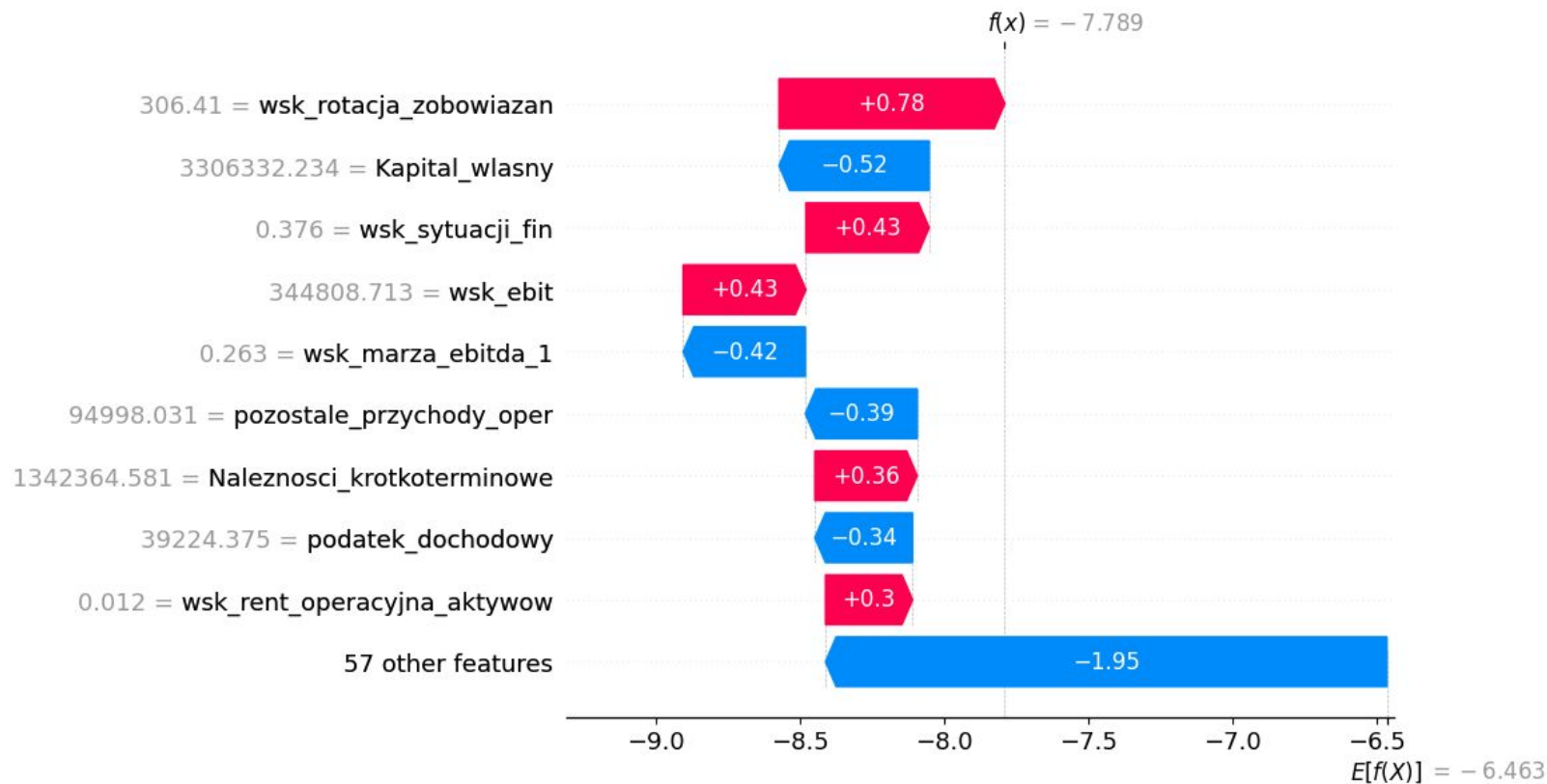
	precision	recall	f1-score	support
0	0.95	0.99	0.97	568
1	0.40	0.06	0.11	32
accuracy			0.94	600
macro avg	0.67	0.53	0.54	600
weighted avg	0.92	0.94	0.93	600



SHAP

- SHAP (SHapley Additive exPlanations) jest metodą stosowaną w uczeniu maszynowym do wyjaśniania wkładu poszczególnych cech w prognozowane wyniki modelu.
- Za jej pomocą możemy zinterpretować model i sprawdzić, jak poszczególne cechy wpływają na wynik klasyfikacji.
- Możemy sprawdzić globalny wpływ poszczególnych zmiennych na predykcje klasyfikatora lub zawęzić analizę do pojedynczej próbki.







Bibliografia

1. <https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html>
2. <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>
3. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
4. <https://xgboost.readthedocs.io/en/stable/>