Angelika Haczykowska, Ewelina Księżniak, Maryla Sosna

Functional requirements

Cały projekt dostępny jest w serwisie GitHub:
https://github.com/MarylaSosna/umwf_projekt

1. Cel projektu

Celem projektu jest stworzenie czterech różnych modeli, spośród których wybrany zostanie ten najlepiej przewidujący oczekiwane wartości score'ów na podstawie cotygodniowych cen zamknięcia przypisanych wszystkim badanym spółkom, z horyzontem inwestycyjnym miesięcznym, kwartalnym, półrocznym i rocznym.

2. Dane

2.1. Dane wejściowe (do poprawy opis)

W projekcie wykorzystano udostępniony zbiór danych zawierający podstawowe informacje o spółkach (takie jak identyfikator) oraz dodatkowe dane pochodzące z serwisu Yahoo Finance, które pobrano za pomocą udostępnionego API.

Oryginalny zbiór danych przekształcono i wybrano najbardziej istotne zmienne:

- Date - data
- Comp - skrócona nazwa spółki
- CompID - identyfikator spółki
- Score - wynik obliczony dla spółki

Dane zaciągnięte z wykorzystaniem API Yahoo Finance, zmapowano z utworzoną w wyniku przekształcenia ramką danych na podstawie daty i zmiennej Comp, lub wyłącznie z uwzględnieniem daty (w przypadku zmiennych, gdy nazwa spółki nie miała znaczenia np. Cena Ropy).

Zdecydowano się na analizę następujących dodatkowych zmiennych:

- Close - cena zamknięcia
- Dividends - wielkość wypłaconych dywidend
- Stock Splits
- totalRevenue - całkowita wartość przychodów w danym okresie
- totalDebt - całkowita wartość zadłużenia w danym okresie
- fullTimeEmployees - liczba zatrudnionych pracowników
- Oil - cena ropy naftowej w danym okresie
- Gold - cena złota w danym okresie
- USD to Yuan - kurs USD/Yuan
- industry - sektor, w którym działa spółka

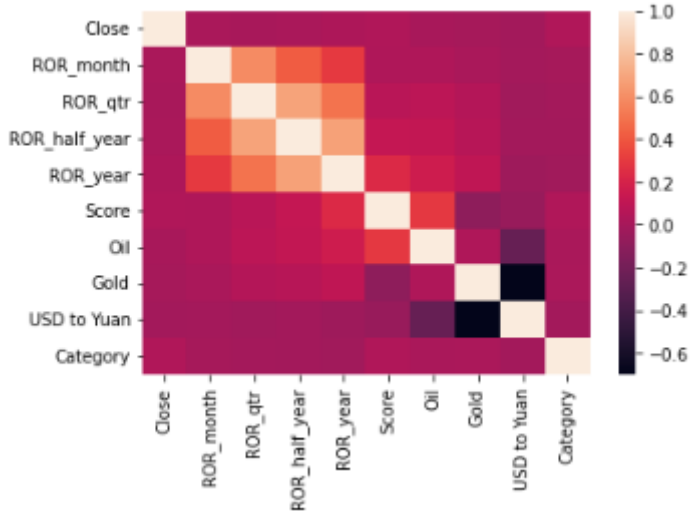
Wszystkie zmienne, poza industry, mają charakter liczbowy ciągły. W celach optymalizacji wyników uczenia maszynowego zmienną industry zdekodowano do poziomu binarnego (utworzono tzw. dummies variables).

W konstruowanych w dalszej części pracy modelach za zmienną objaśnianą przyjęto wynik Score.

2.2. Wizualizacja zmiennych

2.2.1. Korelogram

Zamieszczony poniżej wykres przedstawia korelogram wszystkich zmiennych.

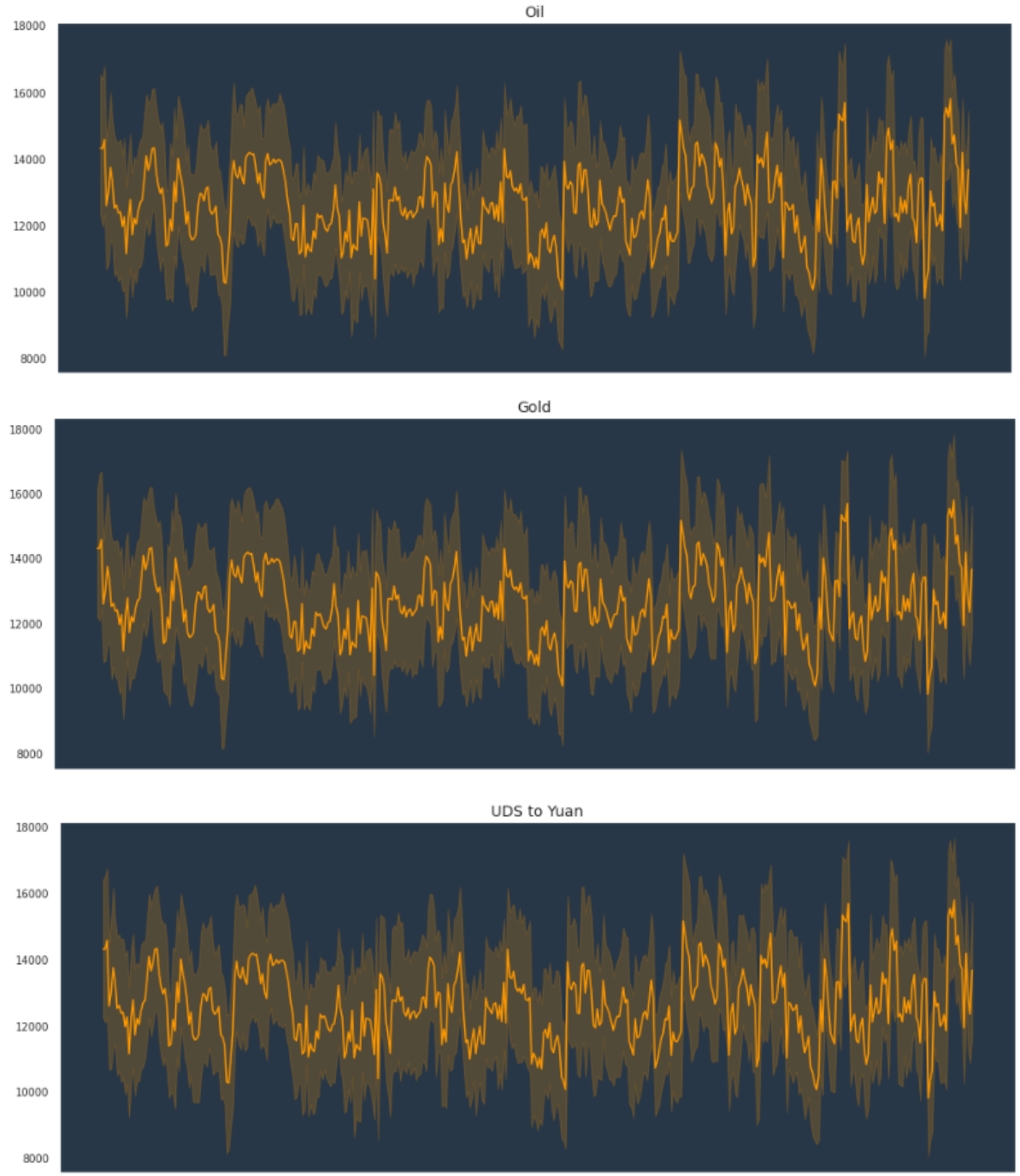


Na podstawie korelogramu można stwierdzić, że:

- większość zmiennych objaśniających nie jest ze sobą wzajemnie skorelowanych,
- silna ujemna korelacja jest widoczna pomiędzy zmiennymi: Gold (cena złota), a USD to Yuan (kurs USD/Yuan), co może świadczyć o niższej zdolności prognostycznej drugiej cechy.

2.2.2. Inspekcja zmiennych objaśniających ciągłych

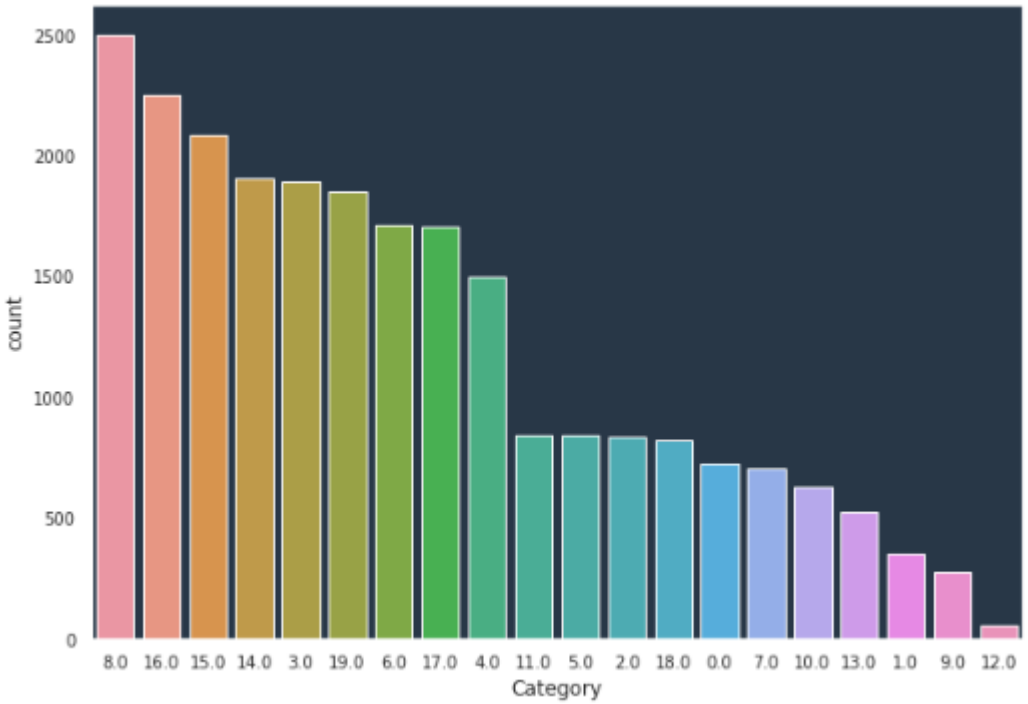
Zamieszczone poniżej wykresy przedstawiają, jak kształtują się wielkości poszczególnych cech ciągłych na przestrzeni lat.



Na podstawie wykresów można zauważyć, że poziom z każdej z analizowanych cech jest stabilny w czasie (posiada cechy szeregu stacjonarnego). Nie zidentyfikowano wartości odstających.

2.2.3. Inspekcja zmiennych objaśniających kategorycznych

Zamieszczony poniżej wykres przedstawia, ile z analizowanych spółek funkcjonuje w poszczególnych sektorach.



- Najwięcej z analizowanych spółek działa w sektorach: 8, 16, 15
- Najmniej spółek działa w sektorach: 12, 9, 1

3. Stworzone modele

3.1. Regresja liniowa (dopisać wnioski dla poprawionych wyników)

Zbiór testowy stanowił 30% wejściowych danych. Korzystano z pakietu sklearn. Wytrenowany model uzyskał następujące metryki na zbiorze testowym:

Szacunek	Horyzont roczny	Horyzont półroczny	Horyzont kwartalny
Mean absolute error	0,16	0,12	0,08
Mean squared error	0,05	0,03	0,01
Median absolute error	0,12	0,09	0,06
Explain variance score	0,18	0,2	0,34
R^2	0,18	0,19	0,34

MAE oraz MSE osiągają niskie wartości. MAE określa o ile średnio różniły się prognozy od wartości rzeczywistych.

Explain variance score mówi na ile dobrze stworzony model wyjaśnia zmienność w zbiorze danych. Im wynik bliższy jest jedności, tym lepiej.

Współczynnik R^2, który mówi nam, jak dobrze model będzie radził sobie z nieznanymi

próbkami nie jest, niestety, zbyt wysoki (idealną jest wartość 1).

3.1 Modelowanie zaawansowane

W sekcji omówiono wyniki badania uzyskane z wykorzystaniem zaawansowanych modeli uczenia maszynowego. W projekcie dokonano modelowania za pomocą algorytmów:

- lasu losowego,
- XGboost,
- sztucznej sieci neuronowej.

Modele zaimplementowano dla trzech horyzontów czasowych: rocznego, półrocznego oraz kwartalnego. Dla modeli lasu losowego oraz XGboost dokonano hiperparametryzacji z wykorzystaniem algorytmu grid search.

Dla modelu lasu losowego optymalizowano parametry:

- maksymalna głębokość drzewa (4 lub 8)
- liczba drzew (150, 200 lub 400)

Dla każdego z analizowanych przypadków (zakres horyzontu czasowego) optymalnymi wartościami parametrów okazały się być odpowiednio: 8 (głębokość drzewa) oraz 400 (liczba drzew).

Dla modelu xgBoost optymalizowano parametry:

- maksymalna głębokość drzewa (4 lub 8)
- liczba drzew (150, 200 lub 400)
- learning rate (0.01 lub 0.015)

Dla każdego z analizowanych przypadków (zakres horyzontu czasowego) optymalnymi wartościami parametrów okazały się być odpowiednio: 8 (głębokość drzewa), 400 (liczba drzew) oraz 0.015 (learing rate).

Trzecią techniką modelowania było utworzenie sztucznej sieci neuronowej, zaimplementowaną z użyciem gotowych rozwiązań z biblioteki Keras. Utworzono sieć neuronową zawierającą 25 neuronów w warstwie ukrytej. Zdecydowano się użyć funkcji aktywacji typu relu. Model uczono 150 cyklach treningowych (liczba epok). Sieć zawierają warstwę wyjściową składającą się z 1 neuronu wskazującego na prognozowany wynik ROR.

3.2 Wnioski - las losowy

Zamieszczona poniżej tabela zawiera informację o wynikach uzyskanych dla modelu lasu losowego w horyzoncie: rocznym, półrocznym oraz kwartalnym.

Szacunek	Horyzont roczny	Horyzont półroczny	Horyzont kwartalny
Mean absolute error	0.14	0.11	0.07
Mean squared error	0.04	0.02	0.01
Median absolute error	0.11	0.08	0.05

Explain variance score	0.38	0.35	0.43
R^2	0.38	0.35	0.43

- Zastosowanie algorytmu lasu losowego pozwoliło zwiększyć R^2 do ok. 40%. Mimo zwiększenia trafności modelu, uzyskane wyniki wciąż nie są zadowalające (są odległe od 100%).
- Model zachowuje się podobnie w każdym z analizowanych horyzontów czasowych. Nie ma istotnych różnic w wynikach uzyskanych dla poszczególnych horyzontów.

3.3 Wnioski - XGBoost

Zamieszczona poniżej tabela zawiera informację o wynikach uzyskanych dla modelu XGBoost w horyzoncie: rocznym, półrocznym oraz kwartalnym.

Szacunek	Horyzont roczny	Horyzont półroczny	Horyzont kwartalny
Mean absolute error	0.13	0.1	0.07
Mean squared error	0.03	0.02	0.01
Median absolute error	0.1	0.08	0.05
Explain variance score	0.47	0.44	0.52
R^2	0.47	0.44	0.52

- Zastosowanie algorytmu XGboost pozwoliło zwiększyć R^2 do ok. 50% - tzn. zastosowanie boostingu pozwoliło uzyskać trafność lepszą o ok. 10% aniżeli w przypadku modelu lasu losowego. Mimo zwiększenia trafności modelu, uzyskane wyniki wciąż nie są zadowalające (są odległe od 100%).
- Model zachowuje się podobnie w każdym z analizowanych horyzontów czasowych. Nie ma istotnych różnic w wynikach uzyskanych dla poszczególnych horyzontów.

3.4 Wnioski - NN

Zamieszczona poniżej tabela zawiera informację o wynikach uzyskanych dla modelu sztucznej sieci neuronowej w horyzoncie: rocznym, półrocznym oraz kwartalnym.

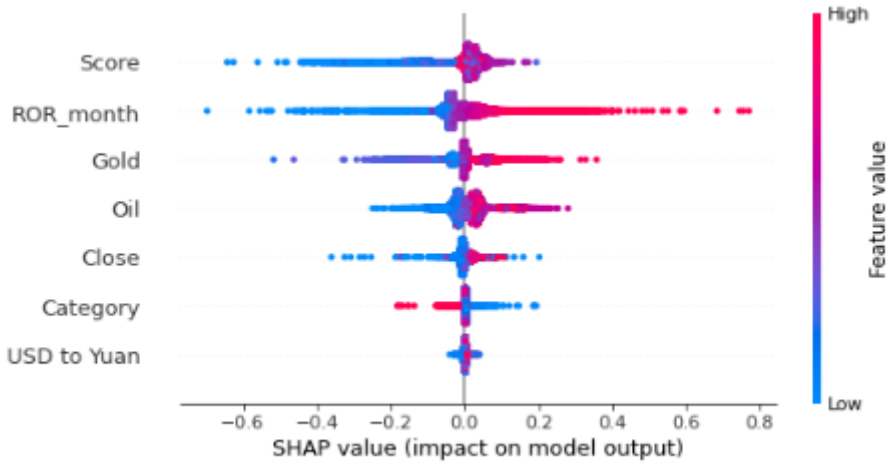
Szacunek	Horyzont roczny	Horyzont półroczny	Horyzont kwartalny
Mean absolute error	0.15	0.12	0.1
Mean squared error	0.04	0.03	0.02

Median absolute error	0.12	0.09	0.08
Explain variance score	0.3	0.26	0.28
R^2	0.28	0.21	0.1

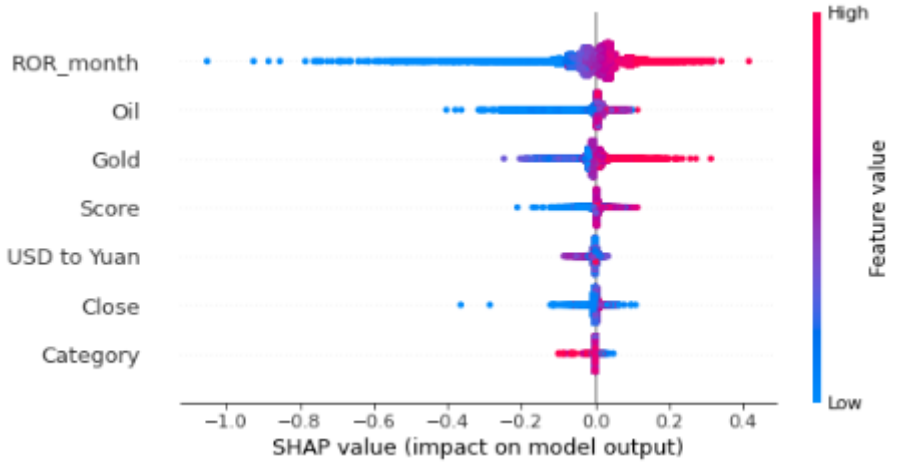
- Trafność modelu sztucznej sieci neuronowej jest najniższa spośród dwóch omówionych wcześniej modeli. R^2 w zależności od przyjętego horyzontu czasowego waha się w przedziale 10 - 28%
- Model jest najbardziej wrażliwy na zmiany w przyjętym horyzoncie czasowym. Różnica pomiędzy R2 dla horyzontu kwartalnego i rocznego wynosi 18%.

3.5. Współczynnik Shapleya

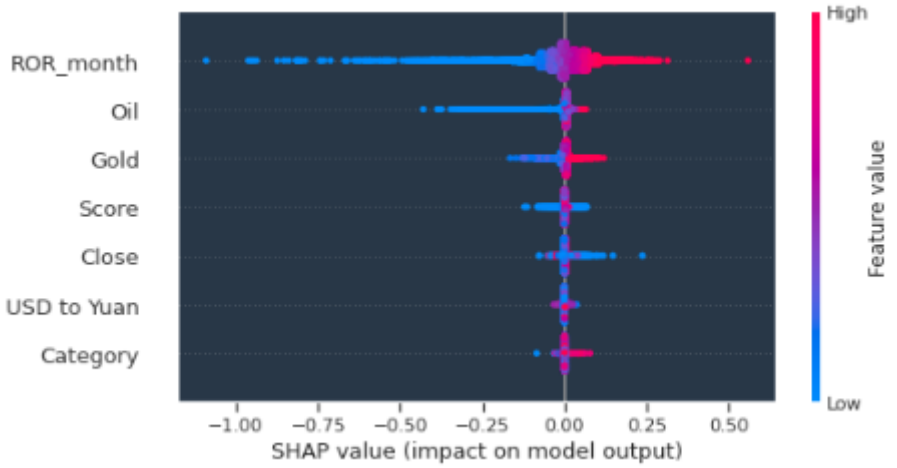
Horyzont roczny:



Horyzont półroczny:



Horyzont kwartalny:


ROR_month
Oil
Gold
Score
Close
USD to Yuan
Category
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
SHAP value (impact on model output)
High
Feature value
Low