Machine Learning: Python-based model of stock convictions to expected returns within given investment horizons - Raport

Maryla Sosna, Ewelina Księżniak, Angelika Haczykowska
https://github.com/MarylaSosna/umwf_projekt

1 Dane

W projekcie wykorzystano udostępniony zbiór danych zawierający podstawowe informacje o spółkach - nazwa spółki, identyfikator, score'y, daty, kategoria 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,
- Score - wynik obliczony dla spółki.

Dane zaciągnięte z wykorzystaniem API Yahoo Finance, połączono 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,
- Oil - cena ropy naftowej w danym okresie,
- Gold - cena złota w danym okresie,
- USD to Yuan - kurs USD/Yuan,
- Category (industry) - sektor, w którym działa spółka

Cenę ropy naftowej wybrano, ponieważ jest to główny surowiec energetyczny na świecie, zatem jego cena ma niewątpliwy wpływ na działanie spółek nie tylko z branży energetycznej.

Kolejna zmienna objaśniająca – cena złota, które mimo, że jest zaliczane do grona surowców, traktowane jest jako alternatywa dla rynków akcji, czy obligacji oraz jako bezpieczna przystań. Co więcej, jest to jedyny instrument finansowy, którego płynność nie została zachwiana przez wieki.

Jako kolejną zmienną objaśniającą przyjęto kurs dolara do chińskiego yuana, jako miarę cen i nastrojów kształtujących się na rynkach.

Ostatnią ze zmiennych jest sektor, w którym dana spółka działa. Zmienną tą wybrano wychodząc z założenia, że nie wszystkie sektory działają w tych samych trybach i na tych samych

zasadach – podczas kryzysów różnego typu, spółki działające w innych sektorach zapewne będą sobie dawać radę w inny sposób, zatem ich spodziewane wyniki finansowe też będą się różniły.

Wszystkie zmienne, poza industry, mają charakter liczbowy ciągły. W celach optymalizacji wyników uczenia maszynowego zmienną industry zamieniono na kategorie dodając następujące etykiety:

0 - Commercial Services,

1 - Communication,

2 - Consumer Durables,

3 - Consumer Non-Durables,

4 - Consumer Services,

5 - Distribution Services,

6 - Electronic Technology,

7 - Energy Minerals,

8 - Finance,

9 - Health Services,

10 - Health Technology,

11 - Industrial Services,

12 - Miscellaneous,

13 - Non-Energy Minerals,

14 - Process Industries,

15 - Producer Manufacturing,

16 - Retail Trade,

17 - Technology Services,

18 - Transportation,

19 - Utilities.

W konstruowanych w dalszej części pracy modelach za zmienną objaśnianą przyjęto wyniki ROR w ujęciu miesięcznym, kwartalnym, półrocznym i rocznym.

Do obliczenia dziennych logarytmicznych stóp zwrotu wykorzystano następujący wzór:

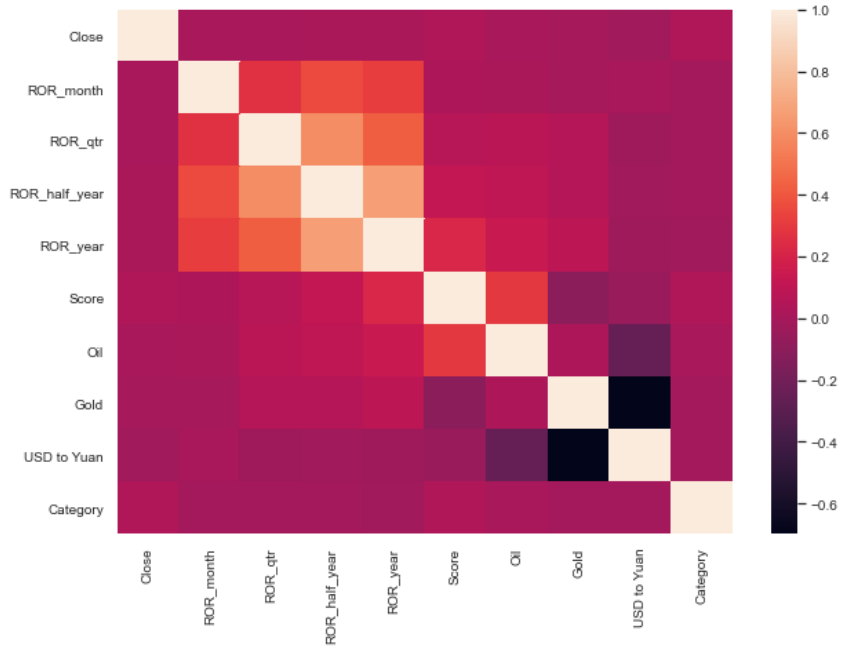
$$r_t = ln\frac{C_t}{C_{t-1}} \tag{1}$$

- C_t - cena zamknięcia w bieżącym okresie,
- C_{t-1} - cena zamknięcia w okresie bazowym.

2 Wizualizacja zmiennych

2.1 Korelogram

Zamieszczony poniżej wykres przedstawia korelogram wszystkich zmiennych.

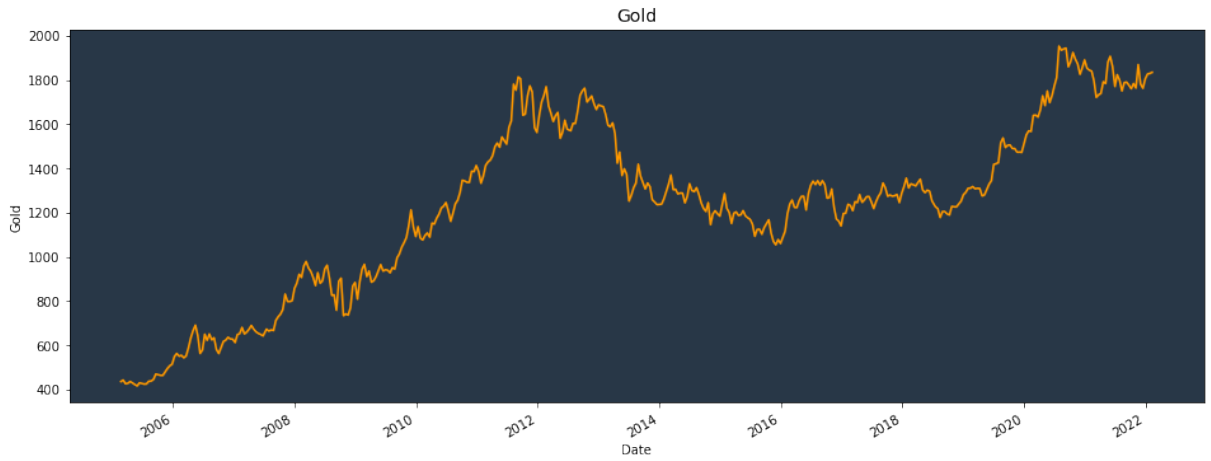


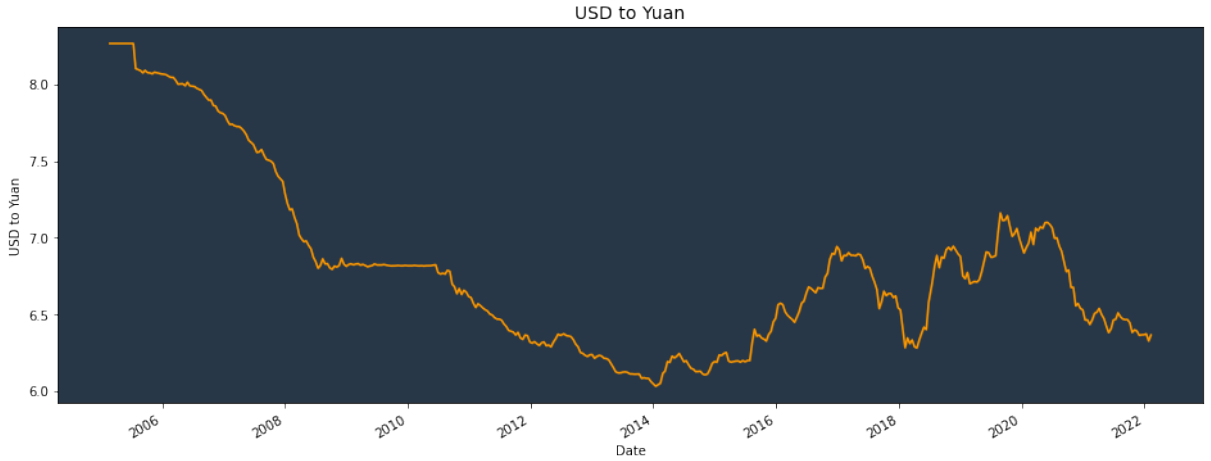
Na podstawie korelogramu można stwierdzić, że:

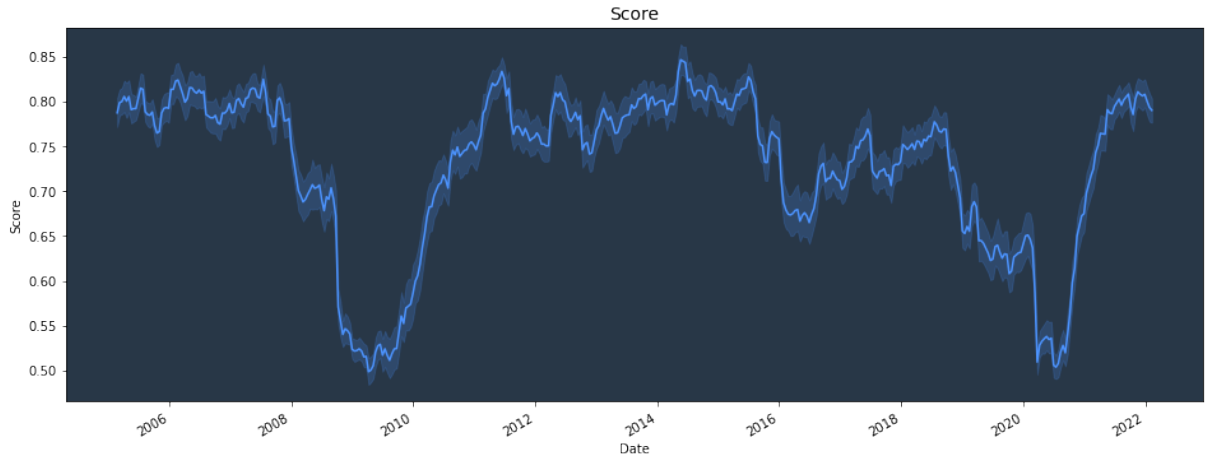
- większość zmiennych objaśniających nie jest ze sobą wzajemnie skorelowanych,
- silna ujemna korelacja jest widoczna pomiędzy zmiennymi: Gold (cena złota), a USD to Yuan (kurs USD/Yuan), co może świadczyć o niższej zdolności prognostycznej drugiej cechy.

2.2 Inspekcja zmiennych objaśniających ciągłych

Zamieszczone poniżej wykresy przedstawiają, jak kształtują się wielkości poszczególnych cech ciągłych na przestrzeni lat. Żółte wykresy zastosowano dla zmiennych pobranych za pomocą YahooFinace, natomiast niebieski dla zmiennej, która znajdowała się w wejściowym zbiorze.



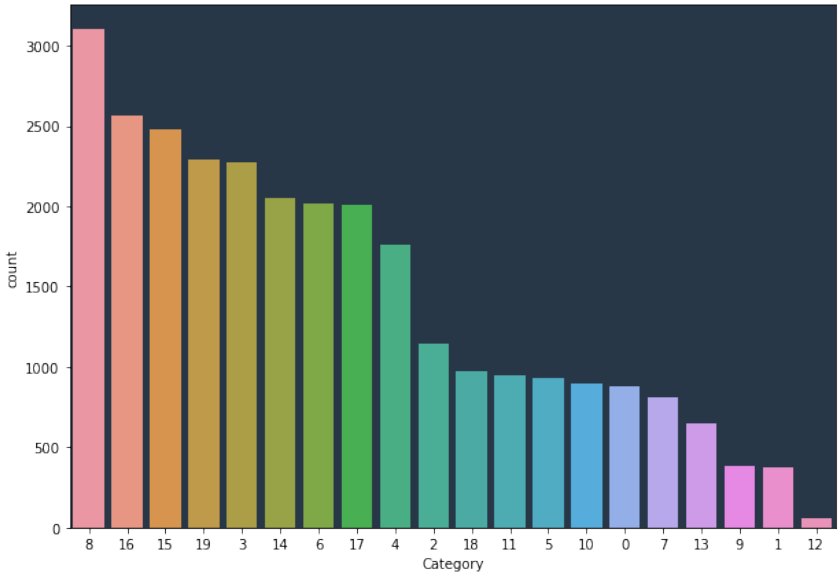




Na podstawie wykresów można zauważyć, że poziom każdej z analizowanych cech jest zmienny w czasie (nie posiada cech szeregu stacjonarnego). Zidentyfikowano swoiste załamania Score'ów i cen ropy (szczególnie w okolicach roku 2009). Cena złota od 2018 roku systematycznie rośnie. Kurs dolara do yuana malał w okresie 2004-2014 i ponownie maleje od 2020 roku.

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 (branża finansowa), 16 (sektor sprzedaży detalicznej), 15 (branża wytwórcza).
- Najmniej spółek działa w sektorach: 12 (różnorodne), 9 (usługi zdrowotne), 1 (komunikacja).

3 Stworzone modele

3.1 Regresja liniowa

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

Szacunek	Horyzont			
	roczny	półroczny	kwartalny	miesięczny
Mean absolute error	0,17	0,13	0,1	0,06
Mean squared error	0,06	0,04	0,03	0,01
Median absolute error	0,13	0,1	0,07	0,04
Explain variance score	0,09	0,03	0,01	0
R^2	0,0861	0,0322	0,0121	0,001

Na podstawie analizy wyników osiągniętych za pomocą modelu regresji liniowej, można dojść do następujących wniosków:

- MAE oraz MSE osiągają niskie wartości. MAE określa o ile średnio różniły się prognozy od wartości rzeczywistych.
- Miara Explain variance score mówi na ile dobrze stworzony model wyjaśnia zmienność w zbiorze danych. Im wynik bliższy jest jedności, tym lepiej. W przypadku stworzonego modelu regresji liniowej, miara ta utrzymuje się na niskim poniżej 10 % poziomie, co oznacza, że model nie wyjaśnia dobrze zmienności w zbiorze.
- Współczynnik R^2 , który mówi nam, jak dobrze model będzie radził sobie z nieznanymi próbkami jest niski i nie przekracza 10 % (idealną jest wartość 1), co oznacza, że skonstruowany model regresji liniowej nie będzie dobrze radzić sobie z predykcją na nieznanych danych.
- W każdym z analizowanych horyzontów czasowych współczynnik R^2 jest niski i waha się w okolicach od 0,01 % do 8,6 %. Najniższy poziom miernika R^2 został osiągnięty dla horyzontu kwartalnego, a najwyższy - przyjmując horyzont roczny.

3.2 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 czterech horyzontów czasowych: rocznego, półrocznego, kwartalnego oraz miesięcznego. 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.

3.2.1 Las losowy

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

Szacunek	Horyzont			
	roczny	półroczny	kwartalny	miesięczny
Mean absolute error	0.15	0.12	0.09	0.05
Mean squared error	0.04	0.03	0.02	0.01
Median absolute error	0.11	0.09	0.07	0.04
Explain variance score	0.34	0.19	0.29	0.28
R^2	0.34	0.19	0.29	0.28

- 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.

3.2.2 XGBoost

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

Szacunek	Horyzont			
	roczny	półroczny	kwartalny	miesięczny
Mean absolute error	0.13	0.12	0.08	0.05
Mean squared error	0.03	0.03	0.02	0.01
Median absolute error	0.1	0.09	0.06	0.04
Explain variance score	0.45	0.14	0.35	-0.24
R^2	0.45	0.14	0.35	-0.24

- 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.

3.2.3 NN

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

Szacunek	Horyzont			
	roczny	półroczny	kwartalny	miesięczny
Mean absolute error	0.15	0.12	0.09	0.06
Mean squared error	0.05	0.03	0.03	0.01
Median absolute error	0.12	0.09	0.07	0.04
Explain variance score	0.3	0.19	0.1	0.1
R^2	0.28	0.19	0.1	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%

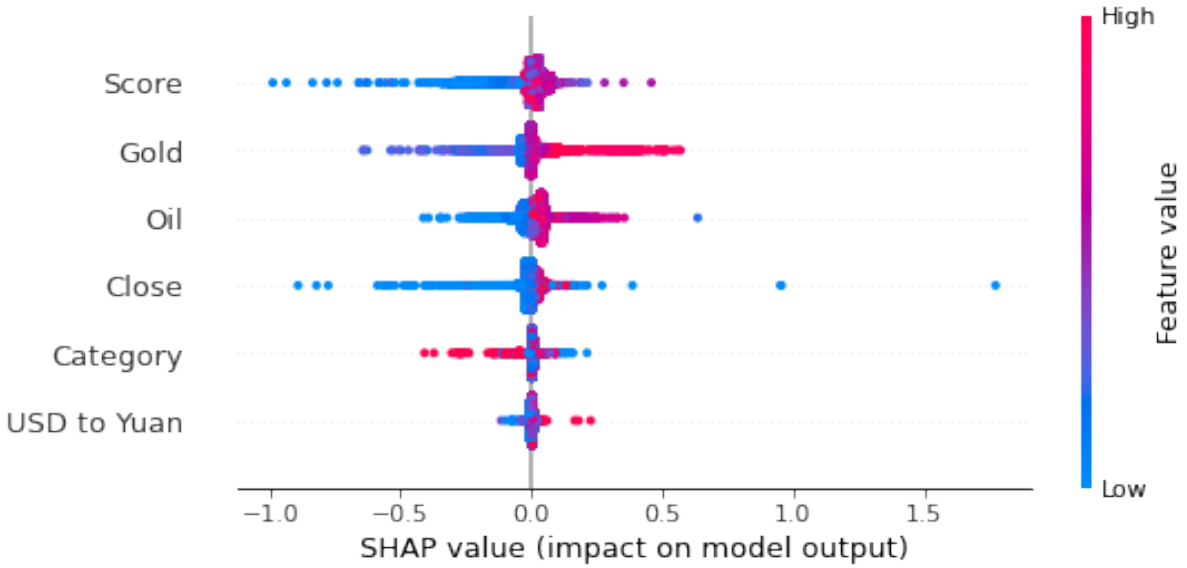
- Model jest najbardziej wrażliwy na zmiany w przyjętym horyzoncie czasowym. Różnica pomiędzy R2 dla horyzontu kwartalnego i rocznego wynosi 18

3.3 Współczynnik Shapley'a

Współczynnik Shapley'a pozwala określić, w jaki sposób poziom określonych cech użytych w zbiorze treningowym wpływa na otrzymane predykcje, w porównaniu z predykcjami, które otrzymano by, gdyby zastosowano ich bazową wartość.

Poniżej przedstawione zostały wykresy wartości zmiennej Shapley'a dla zmiennych prognozowanych, czyli stóp zwrotu w horyzontach rocznym, półrocznym, kwartalnym i miesięcznym.

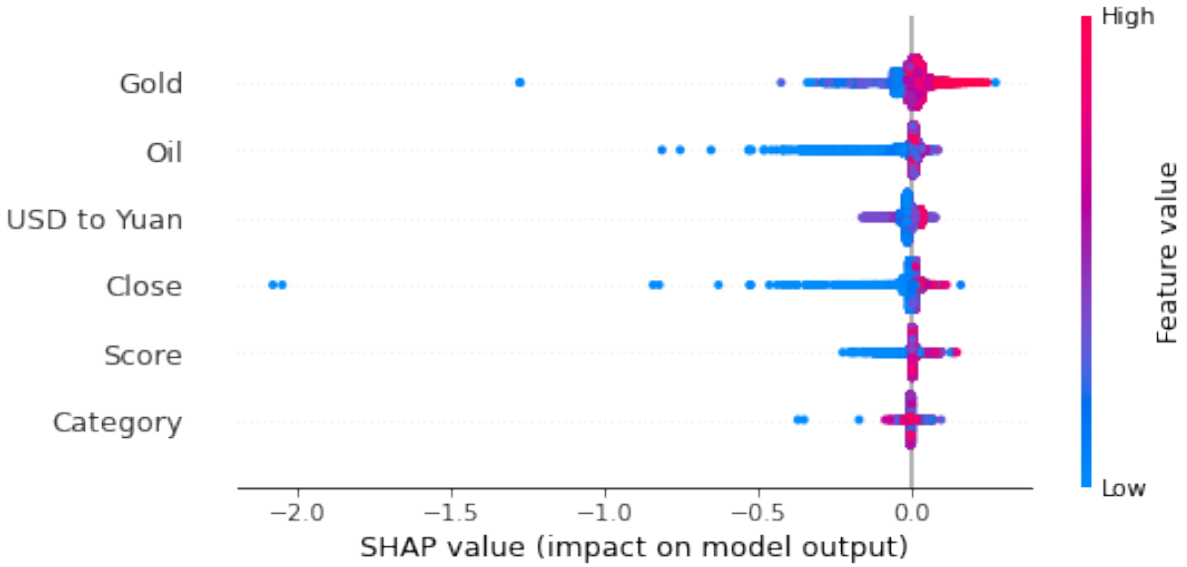
Horyzont roczny:



Z powyższego diagramu można dowiedzieć się, że:

- im niższa jest wartość Score'a, tym mniejsza roczna stopa zwrotu (ROR_year),
- wyższe ceny złota (Gold), jak i ropy (Oil) prowadzą do wyższych wartości rocznych stóp zwrotu, natomiast niższe prowadzą do niższych wartości rocznej stopy zwrotu.

Horyzont półroczny:

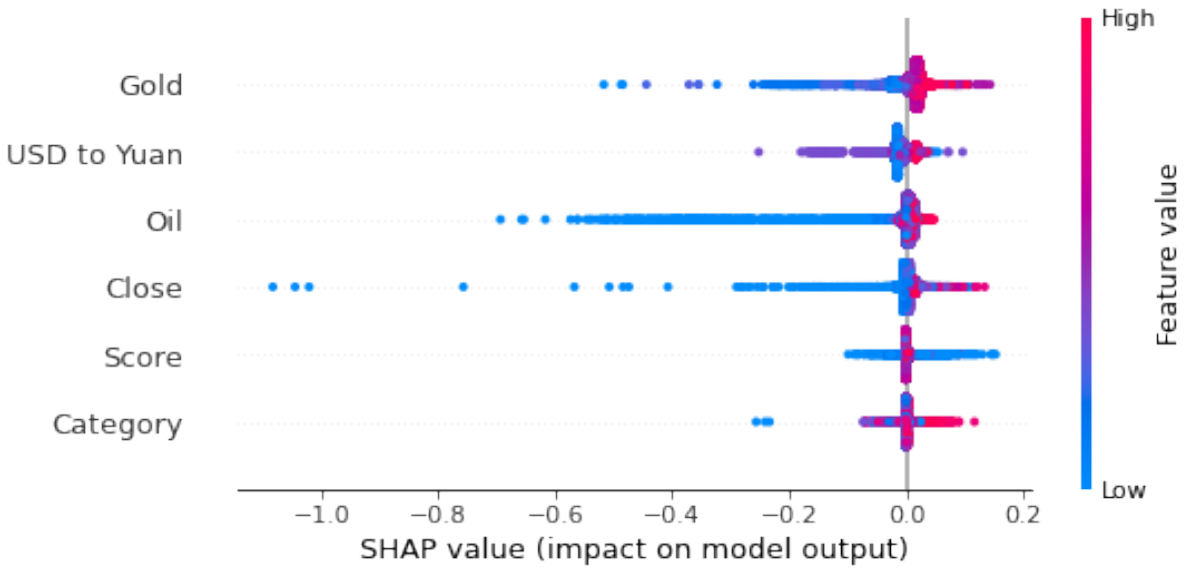


Dla horyzontu półrocznego diagram współczynnika Shapley'a pozwala wysunąć następujące wnioski:

- niższe ceny ropy (Oil) zmniejszają wartość półrocznej stopy zwrotu,

- podobnie ceny zamknięcia (Close) – mniejsza wartość prowadzi do mniejszej wartości półrocznej stopy zwrotu.

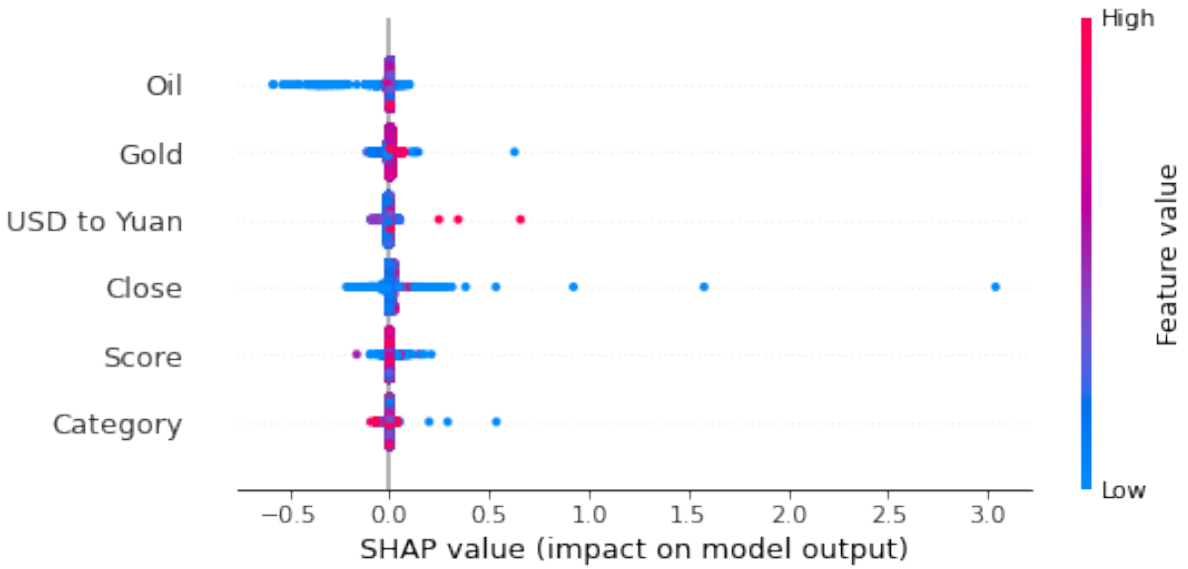
Horyzont kwartalny:



Wykres współczynnik Sharpley'a dla stóp zwrotu w horyzoncie kwartalnym świadczy o:

- niższe ceny ropy (Oil) zmniejszają wartość kwartalnej stopy zwrotu,
- podobnie jest z ceną zamknięcia (Close) – niższe wartości łączą się z niższym RORem kwartalnym,
- w przypadku Score'a zarówno wyższa jak i niższa wartość tej zmiennej wpływa na zmniejszenie oczekiwanej kwartalnej stopy zwrotu.

Horyzont miesięczny:



Wyraźny wpływ zmiennych objaśniających na objaśnianą w horyzoncie miesięcznym widać jedynie w przypadku cen ropy (zmienna Oil) – niższa cena zmniejsza wartość spodziewanej stopy zwrotu (ROR_month). W przypadku pozostałych zmiennych nie można stwierdzić jednoznacznych wpływów. Zaobserwować można jedynie poszczególne wartości odstające dla cen zamknięcia (Close) – wysokie wartości sprawiają, że oczekiwana miesięczna stopa zwrotu jest niższa.