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

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.

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:

Mean absolute error	0,09
Mean squared error	0,01
Median absolute error	0,07
Explain variance score	0,15
R^2	0,15

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).