Taiyō
Machine Learning Assignment
Time-Series Analysis

By Remita Austin

PROBLEM STATEMENT

Use any daily time series from Investing.com or similar source with a strong sample of covariates. Target commodities price like: Oil, Natural Gas, Resin, or Metal Prices. Please make sure to get an extensive list of feature space, think through structural other external factors.

Option 1. Feature Importance. Dynamic Time Warping and/or XGBoost/Shapley Value hybrid model approach to quantify which factors influence the target positively or negative. Keep the analysis focused on the feature selection and feature importance aspects.

DATASET

The dataset that has been used for the time series analysis has been obtained from Investing.com. It consists of historical data of Crude Oil WTI Prices from 1 January 2022 to 14 June 2022. The dataset contains data fields of Date, Price, Open, High, Low, Vol. and Change %.

For the analysis and prediction, we would use only relevant fields of Date and Price.

Link to Dataset: https://www.investing.com/commodities/crude-oil

# IMPLEMENTATION

1. **Data Pre-processing**

   The Crude Oil WTI dataset was imported from Investing.com and loaded. We checked if the dataset contained missing/null values. Then string datatype of Date field was converted into datetime datatype. Similarly Vol. and Change % fields were converted from string into float datatype.
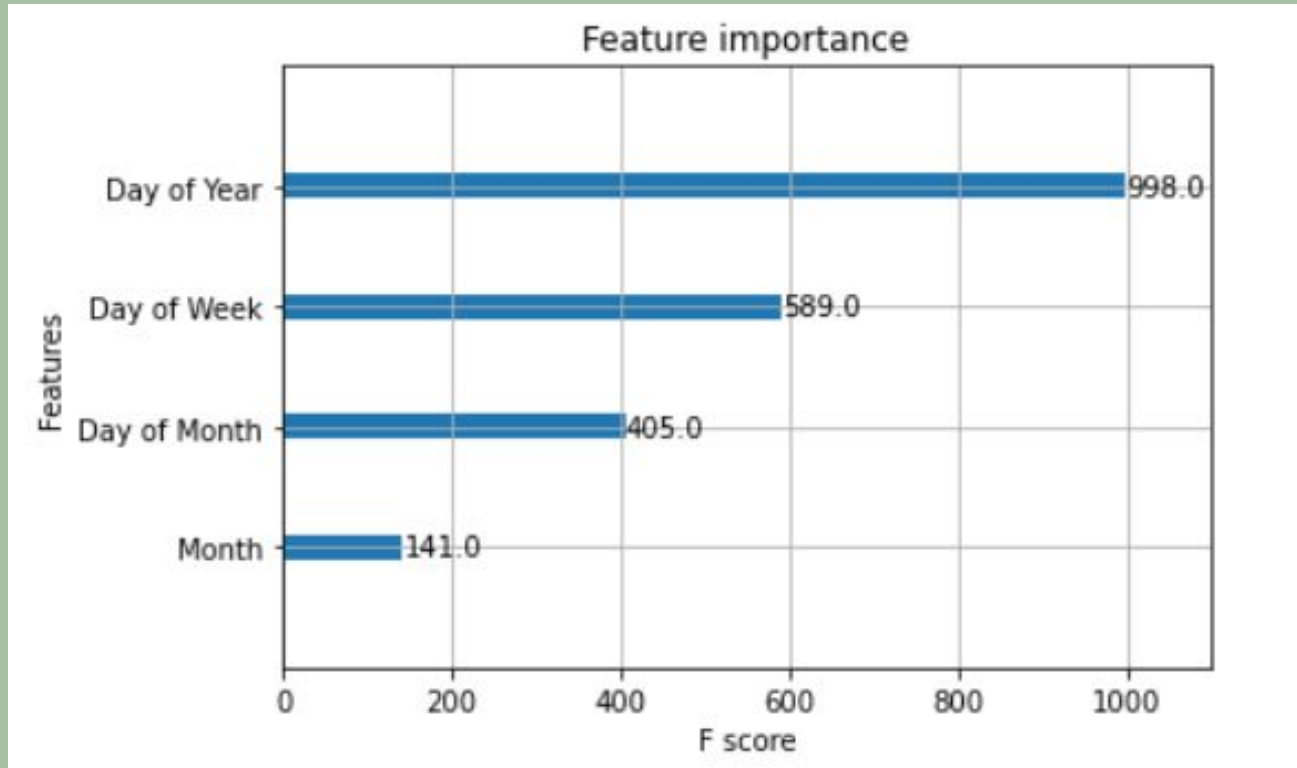
2. **Exploratory Analysis**

   We need to observe and analyze the data to see what we are working with. For this, the matplotlib library was used and the data was visualized with respect to Date field.

Price of Crude Oil across Timeline
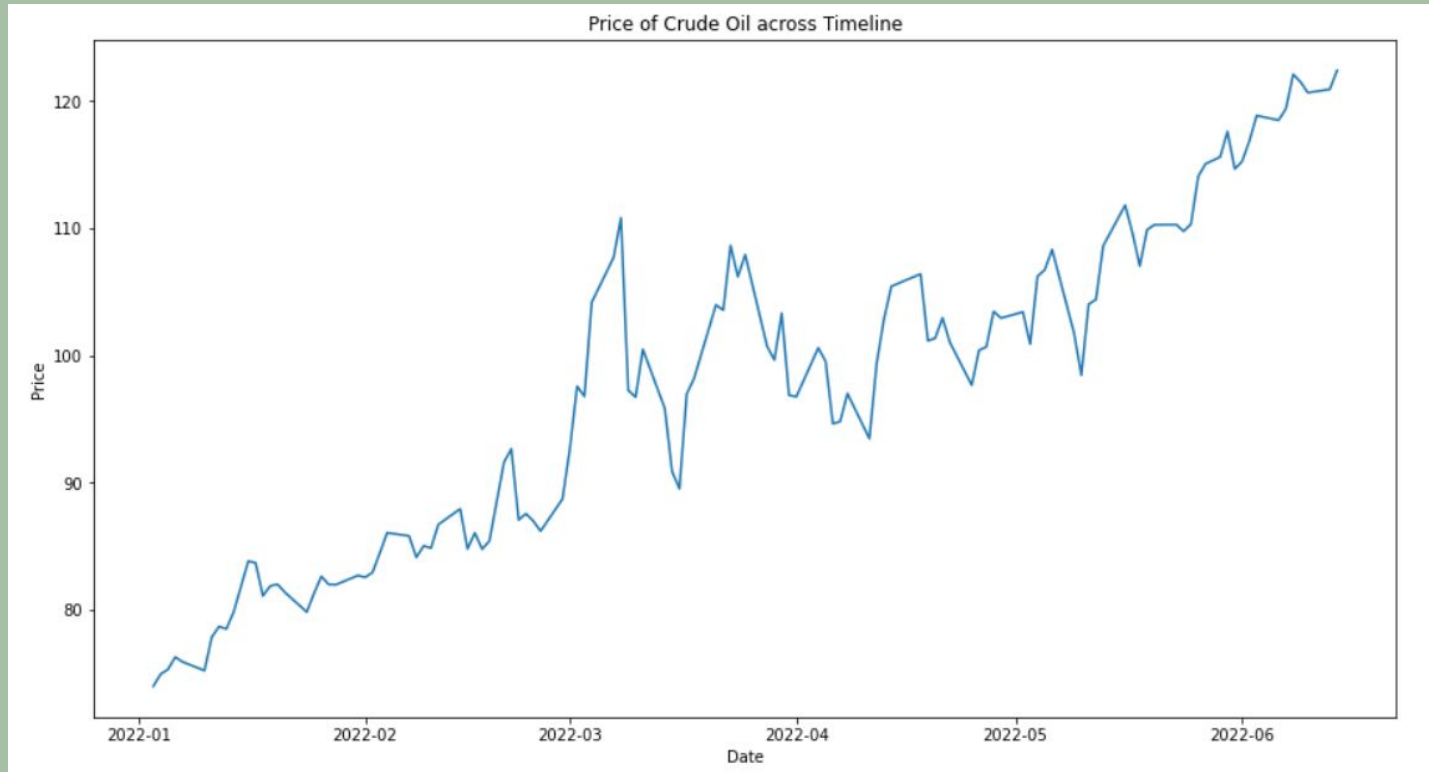
From the graph, we can infer that the price of crude oil has significantly increased over the past six months.

Percentage of Change across Timeline
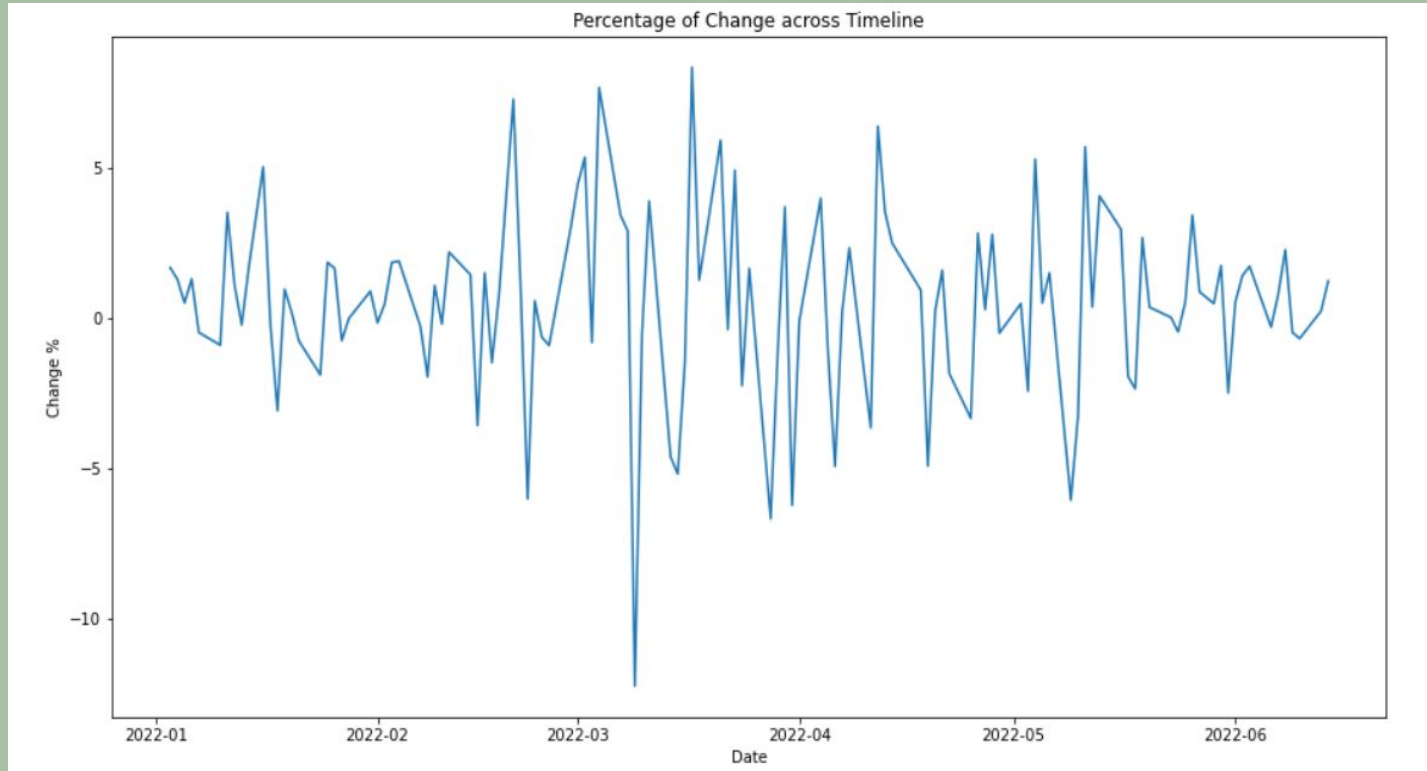
From the graph, we can see that there has been a steep decline in change % during the beginning of March and similarly a steep incline in change % by the middle of March. We can also infer that Change % is not consistent; it always keeps increasing and decreasing across the six months.

3. **Split the data into training and testing data**

   The dataset was split such that training data consisted of 89 rows and 1 column while testing data consisted of 30 rows and 1 column.

4. **Feature Selection**

   The feature_selection() function was used to create features from the Date field like Day of Week, Month, Year, Day of Year and Day of Month. Using XGBoost model, feature importance was computed and it was visualized in the form of a graph.

Feature importance

It is clear from the graph that the 'Day of Year' feature is the most important with an F-score of 998 while 'Month' has the least importance with an F-score of 141.

5.   Making Predictions

The price of crude oil from the month of May was predicted and the results were compared with the actual prices and visualized.
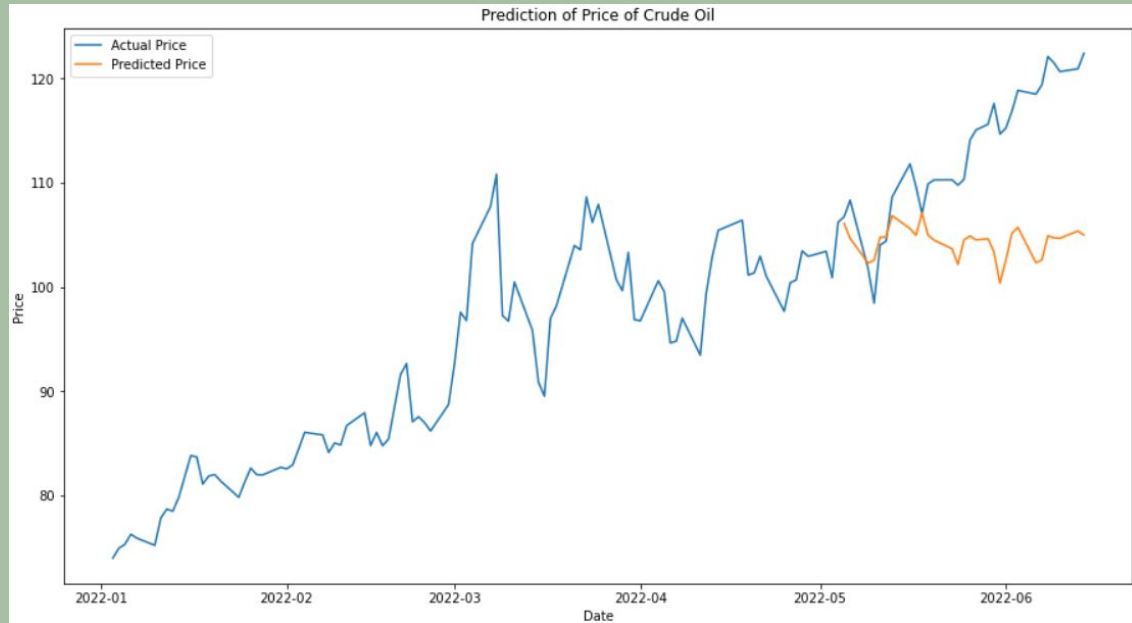
Here the price is predicted for the last one month, which shows to be in the range 100-110, while the actual price goes up to 120.

CONCLUSION

The Crude Oil WTI time series data were analyzed, and on performing feature selection using XGBoost, the most important feature was found to be 'Day of Year' while the least important was 'Month'. Then the price of crude oil was predicted and compared with the actual prices visually.