Impact of News Sentiment on the Stock Market

Vishal Kapadia (vkapadia@usc.edu), Dhruvil Trivedi (dhruvila@usc.edu), Deep Amin (deepprak@usc.edu), Kartik Patel (kartikbh@usc.edu), Saurav Borse (borse@usc.edu)

Department of Data Science, Viterbi School of Engineering,
University of Southern California, Los Angeles, California 90089, USA

(Dated: March 22, 2021)

Abstract

The goal is to find the data and features that actually affect the stock prices of the oil companies directly or indirectly. There are many elements that can affect the stock prices but our goal is to find which features effect the most out of many different features that may or may not affect the stock prices.

1. Introduction:

Here we are starting with the data preprocessing for each data set that we are going to use. We have also attempted our hand in the feature selection. Our main goal is to find the features that are most relevant to the stock prices for the give oil and energy company. Then we are training the dataset to match the actual values of the stock prices using the predicted values of the stock prices. We have attempted this task on training dataset, validating dataset and testing dataset. From which we will find the most important features that effect the stock prices among the data that we have taken.

2. Model Selection:

We have selected total of 2 models to predict the prices of the stock. One is linear regression and the other is Random Forest. We have taken the linear regression to check for the linearity of the data and the Random Forest for analyzing the non-linearity in the data.

3. Data Pre-Processing:

There were 13 datasets primarily provided for the challenge which belonged various categories such as news sentiments, mobility index, oil and energy data and various financial data of given companies. We show all our results mainly for the analysis of various datasets of the Exxon Mobil Corporation.

The most basic processing done in every dataset used is that we transformed the data in such a manner that there was only 1 entry for a single date as it was the target variable i.e. the stock's closing price of the company. For most of the datasets, there were multiple entries for a single date which differed by certain column's categorical labels in them. Hence, those labels were transformed in such a manner that the new dataset had those categories as their column names and the values of the respective dataset as the feature values for those labels (columns) in the new dataset. Apart from that, the common process done for all was that the date range was taken from minimum to maximum date of the respective dataset. The new dataset was taken for this range and a new column showing a dummy variable was entered depicting whether or not the data was present or not at that date.

1. **Sentiment Data**
   - The sentiment data was considered from the global financial news dataset which gave the average sentiment for various number of news articles across multiple publishers for the given keyword. The data was very simple and didn't need specific processing. However, we added dummy variables and handled the values within the date range when there was no news regarding that specific company.
2. **Mobility Data**
   - The mobility section consisted of Apple mobility index which gave the data regarding the percentage change in number of people in a certain region using a specific means of transportation (Walking, Transit or Driving). This data is important as it can provide the information on the demand of fuel and energy resources in specific region. Hence for the processing, each categorical column was done in such a manner that for all the counties, there were 3 columns each for each model of transportation in that region. The values were stored based on the respective mapping of the columns and categories from the original dataset. We removed certain columns which had no values as there was not data for those categories in the given dataset.
   - Another dataset was Google mobility index which provided the information on the percentage change of number of people visiting a specific type of location such as parks in the specific region. This dataset is considered specially for influence of COVID-19 on the stock market. The values in the dataset are with respect to the baseline considered as average population of each day in the month of January 2020 before the adverse effects of COVID-19 began.
3. **Oil, Energy and Gas data**
   - The first dataset was the JODI Oil dataset. This was a monthly dataset which provided the information on the average use of barrels of oil for certain region. The basic process for label categories is same where

the country names are the columns, and the values are considered as per the unit (1000 barrels). The values were 1000 barrels per day as an average need for each day in specific month. Thus, we converted the date range into daily format from minimum to maximum date and inserted that average value for each day in specific region for all days of that month.

4. **Financial Data**

The financial data bucket consisted of many relevant datasets that provided various information for many companies.

- The main financial data is the stock market data. Hence, we filtered out the data for the Exxon company and it consisted of the closing price of every specific day. However, for making the data in a continuous date range, the days when the market was closed need to be handled. We cannot take any random or statistical value as it would imply incorrect behavior of the data. Hence, we added the dummy variable for those missing days and added the previous days closing value in the price depicting that the price did not change at all in those days.

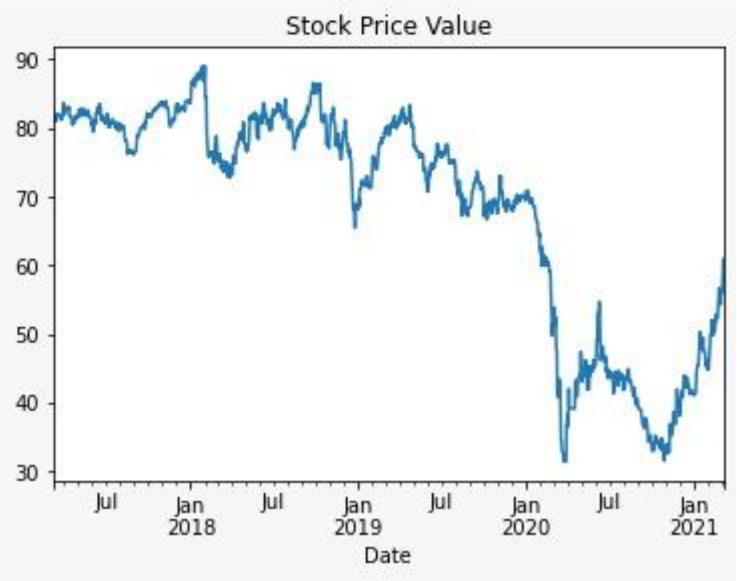

Figure 0: Stock price value vs Date

- Apart from that, the next dataset was the financial statement data provided many insights for the categories such as the Net profit margin, gross profit margin, interest expense, revenue Income Tax expense and so on. However, these data are quarterly and some of them are in percentage values. So, we cannot take average for the same for all days. Furthermore, this data is such that is made public only at the given time period. Hence, it is influencing the stock prices before the information is public is highly unlikely. Adding the averages or any other values in the null entries for the days between two data points will create a bias and will let the model learn that at that point of time, there is some influence on the stock which is not the case.
- The next dataset was Energy Future Prices which consist of future values of the crude oil products which provided information related to the product id and its description. This data is vital as it provides information about the future values of the crude oil and refined products. Hence, processing of the data was done in such a way that each category of the product id Contracts was assigned as column and its values was set accordingly. Therefore, values were mapped to the product id contracts from the original dataset.
- The next dataset is the stock market indicators and the indices. This dataset is fairly simple and had categories for the United States' stock market indices and indicators showing their values. The indices

were Dow Jones utility average, S&P 500, Dow Jones Composite average, Dow Jones industrial average and Dow Jones transportation average prices. For the missing dates in between, it is handled in the same manner as the stock data is processed. The previous day's values are added in the missing date, but the dummy variable is kept 0 indicating that the data was not actually present there.

- Furthermore, the next dataset is the interest rates dataset. Here, it indicates annual interest rate percentage for certain categories. These categories are the indicators as Long term interest rates, short term interest rates, long term interest rates forecast and short term interest rates forecast. These indicators are given for many different regions across the globe such as clusters of European areas, India, United States, Russia and China. This is also a quarterly dataset and hence cannot be handled directly for the missing dates. Apart from that, another problem is that that though the data is quarterly, the unit index of measurement is percentage of interest rates annually. However, the interest rates changing can potentially indicate the trends in the payment of the debt by the certain company.
- The next dataset that we worked on is commodity set dataset. In this data set we can see that there are total of 2 distributers of the oils to the major oil companies. One is NYMEX and other one is BRENT. We have separated both the companies with their respected prices on the date taken as index. We have taken the missing dates as null for processing of the data. This dataset was Daily dataset so we have only filled the missing values of the dates for processing purposes.

4. **Model Evaluation:**

We are using Root Mean Squared Error(RMSE) for evaluating our results for the predicted prices. We are prediction price for training dataset, Validation dataset and Testing dataset

1. JODI Oil dataset: The Visualization of the Predicted Values for training dataset plotted on Actual Values is as follows for JODI dataset.

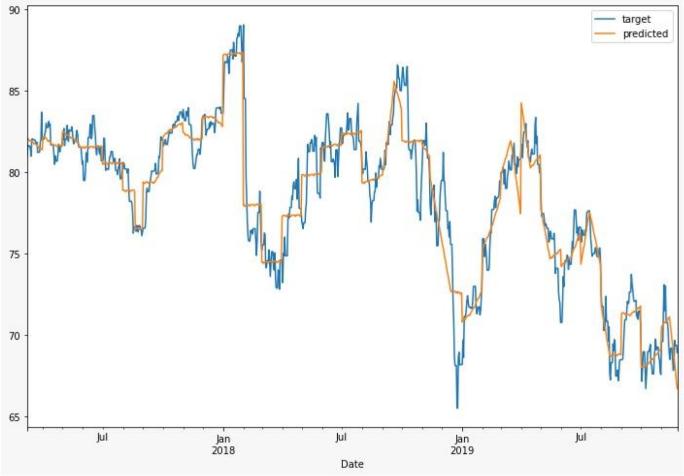

Figure 1: Predicted Values for training dataset plotted on Actual Values for Linear Regression

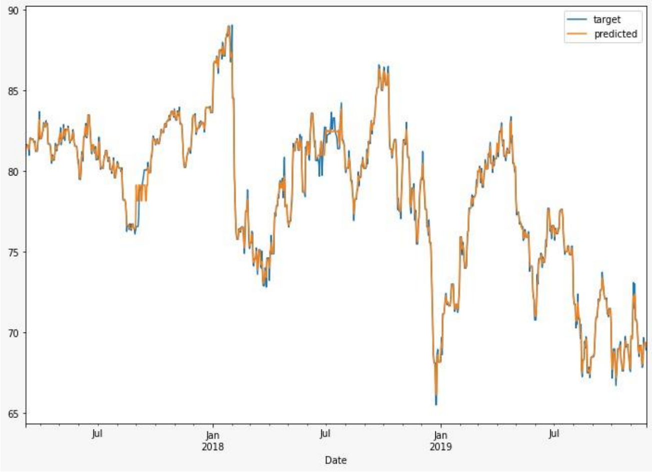

Figure 1: Predicted Values for training dataset plotted on Actual Values for Random Forest

2. Sentiment dataset: The Visualization of the Predicted Values for training dataset plotted on Actual Values is as follows for Sentiment dataset.

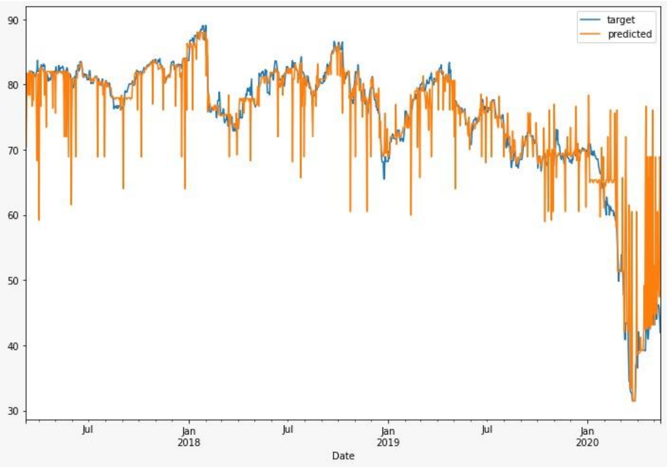

Figure 3: Predicted Values for training dataset plotted on Actual Values for Random Forest

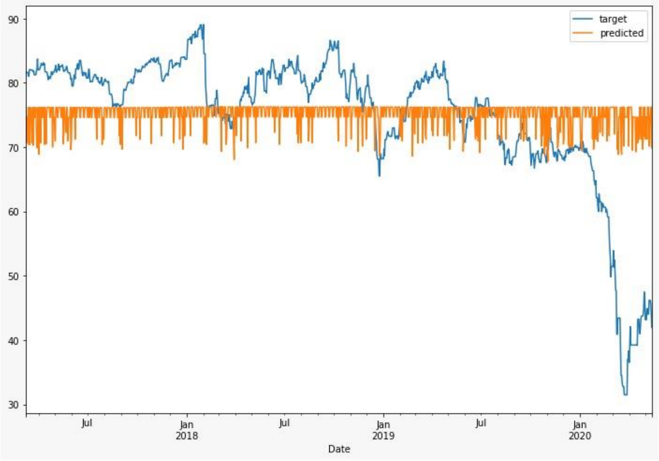

Figure 4: Predicted Values for training dataset plotted on Actual Values for Linear Regression

3. Apple Mobility Dataset: The Visualization of the Predicted Values for training dataset plotted on Actual Values is as follows for Apple Mobility Dataset.

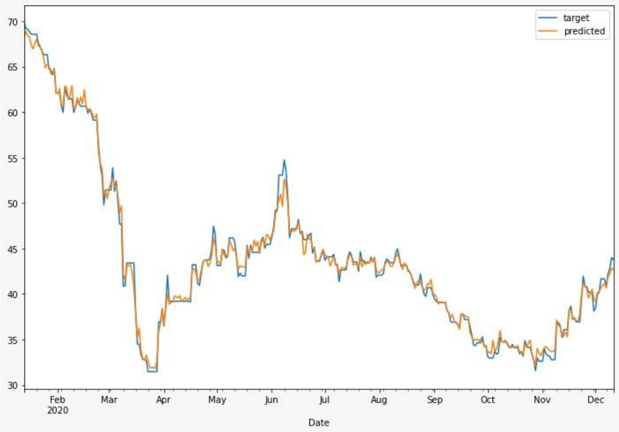

Figure 5: Predicted Values for training dataset plotted on Actual Values for Random Forest

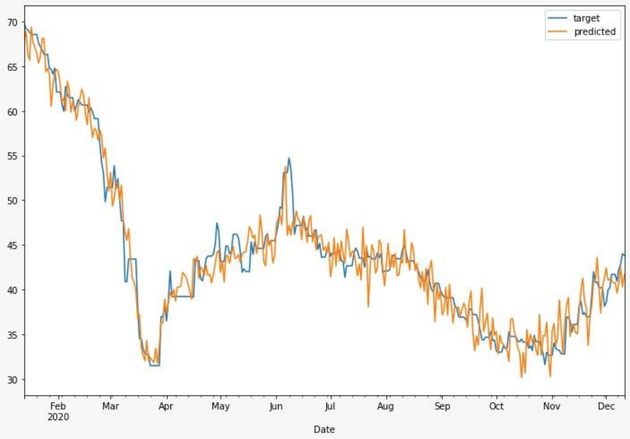

Figure 6: Predicted Values for training dataset plotted on Actual Values for Linear Regression

4. Commodity Dataset: The Visualization of the Predicted Values for training dataset plotted on Actual Values is as follows for Commodity Dataset.

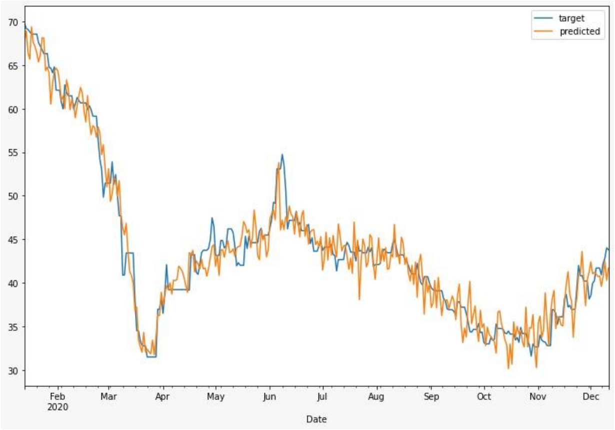

Figure 7: Predicted Values for training dataset plotted on Actual Values for Linear Regression

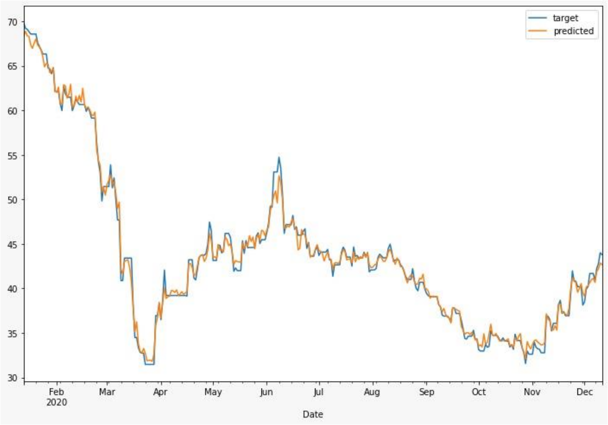

Figure 8: Predicted Values for training dataset plotted on Actual Values for Random Forest

5. Stock Indicator: The Visualization of the Predicted Values for training dataset plotted on Actual Values is as follows for Stock Indicator Dataset.

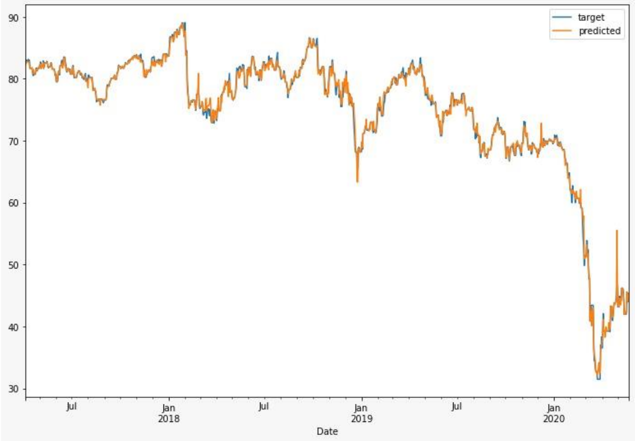

Figure 9: Predicted Values for training dataset plotted on Actual Values for Random Forest

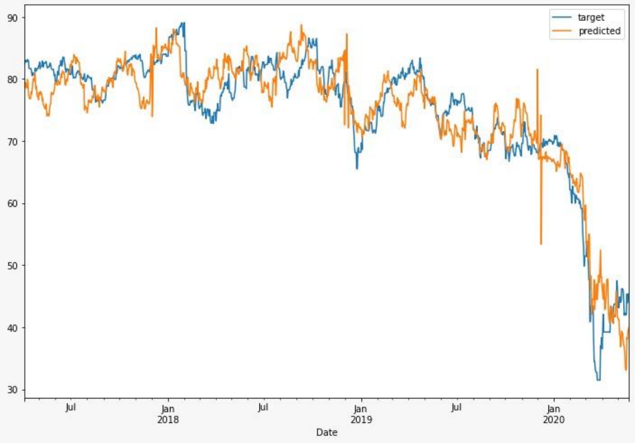

Figure 10: Predicted Values for training dataset plotted on Actual Values for Linear Regression

As mentioned earlier in the introduction section our main goal is to find the important features that effect the stock prices of the given oil and energy company. We are using permutation_importance() function of sklearn to find those features for each dataset.

permutation_importance(model,train_X,train_y,n_repeats = 30, random_state = 43)

Here is the result table on all the datasets for which we have found out the important results of.

RMSE Value	Linear Regression			Random Forest		
	Training	Validation	Testing	Training	Validation	Testing
JODI Dataset	1.517	910.095	3374.238	0.309	16.585	26.106
Sentiment Dataset	10.508	32.464	31.916	10.165	32.918	32.330
Apple Mobility Dataset	2.054	3.145	5.059	0.691	3.550	11.422
Commodity Dataset	6.849	20.327	23.682	1.520	19.079	26.163
Stock Indicator	4.062	16.595	24.204	0.635	30.432	27.189

Table 1: Results on Different datasets

5. Feature Importance

Here, we describe that for each individual dataset, which specific features of the processed data proved to be the most crucial factors in predicting the stock price thus displaying the relationship between the stock price and the respective feature.

1. JODI Dataset:

In the JODI oil dataset, for the linear model training, the important features that came up were the countries name which had major oil fields in reality. This shows that the linear model can be used in order to determine a general trend of where the oil supplies and transactions would take place. However, for the non-linear model, the important features that popped up were such countries where the population density was relatively high. This shows that when working on tasks which require the information of specific time period or seasonality, the non-linear model suggests better and gives countries with high density as important features instead of a general notion.

2. News Sentiment Dataset

For new sentiment dataset, the linear model shows that important features in such a manner that the presence of news (the dummy variable) is more important than the actual sentiment of the news. But for the non-linear model, the sentiment value is more important feature compared to others. On the contrary, the error for the linear model is lesser compared to that of the non-linear model. This shows that the linear model though shows better results, is averaging out the data vector and this is not able to catch the importance of the sentiments that actually matter.

3. Apple Mobility Index Dataset

The apple mobility index proved to be a very important dataset in impacting the stock prices. Here, the linear model and non-linear model gave on =e of the best insights regarding the importance of various features. The features were the columns showing a region name and the mode of transportation. The most important features were the ones where the mode of transportation was driving and transit. This is to be expected because they are the actual consumers of the fuel resources. Hence, the apple mobility index is the best dataset that explains the behavior of the stock prices and thus acts as a great influencer. Especially the features providing the values for the driving and transit modes of transportation in various regions.

4. Commodity Prices Dataset

By performing the feature importance on Commodity for knowing the impact of features in commodity data towards the stock price value using Permutation_Importance it is observed that when the feature importance is created for linear model the commodity features by Brent is more important as compared to those by NYMEX. Similar, observation are observed will working with Random Forest Model. So, probably the commodity data by Brent is more important for Exxon Mobil Data. So, in future if more data from Brent Company is obtained it would really impact the Stock Price Value of Exxon Mobil Data.

5. Stock Indicators and Indices Dataset

When the Permutation_Importance is executed on Stock Indicator Data using Linear Regression Model, the importance of Composite Average was more as compared to Transportation Average and Utility Average this is because the data is non-linear and so this linear model is not able the catch the non-linear patterns from the data. When the feature importance is done on Random Forest Model it catches the non-linear patterns and hence it gives higher importance for Transportation Average. Hence, non-linear model would work efficiently in this case.

6. Interpretation

- The final interpretations from the datasets that were tested for being related to be responsible for creating biases in the performance of stock of a specific company. The table shows the results of tests conducted using 2 models on the same dataset. The two models are the Linear Regression model which helps understand that how much the linear behavior does the data depicts. The second model is the Random Forest regression model which is used to catch and analyze the non-linearities in the dataset. The reason to use these models because they are very easy to interpret and help understand the behavior of the data in an efficient manner. The data values in the given table shows the error or the deviation of the predictions from the actual values. The table can be interpreted in many forms.
- Firstly, for the JODI dataset, the linear model gives a lot of error, but random forest has very lesser error. This shows that the data is majorly non-linear in nature. Furthermore, it shows that the non-linear model (random forest model) though shows better results, has an error of approximately 26 USD in the predictions. Hence, the JODI dataset does influence the stock market by a certain factor but not much.
- For the news sentiment dataset, it shows that in a general view, the data is linear in nature as the linear model performs slightly better. However, it performs least efficiently as compared to other datasets. Thus, it can be said that the new sentiments are important influencers but not the most crucial factor.
- The apple mobility index is as of now proving to be the best bias that helps influence the stock price. It is also a linearly behaving data. The best part is that it can be observed that the training error for linear model is more than that of non-linear model which shows that the non-linear model is overfitting the data to a certain extent. This is the reason why the testing error is lesser for linear mode. It is about 5 USD. The average price for the testing dataset in linear model is 48 USD. Thus, it can be said that the error rate in predictions is nearly 10.42% on average in general view. This shows that using the apple mobility index, we can predict the stock prices more efficiently as they are the one that provide the actual information on the demand of the fuel and energy resources.
- The commodity dataset and the stock indicator dataset both display linear characteristics as well and also have a good fit for linear model. However, they as well are not very important by themselves.
- However, the commodity prices, stock indicators and indices and the JODI oil dataset prove to be nearly equally important datasets as the biases for the stock prices. But the news sentiment didn't prove to be as well. Although, the apple mobility index proved to be the best influencer.

7. Conclusion

- In a nut-shell, we preprocessed many datasets from the given data. However, based on our observations during the processing, there were certain problems in some of the datasets which we felt would create more ambiguity and errors if used to determine their individual influences. For an instance, energy futures and the financial statement ratios were quarterly data and thus were very sparse and there was no reliable means of handling the null values for al the dates lying between two such complete datapoints. Hence, we didn't take those into the account for the final submission. However, the 5 datasets namely the JODI oil dataset, the Apple Mobility dataset, the News Sentiment dataset, the Commodity prices dataset and the Stock indicators and indices proved to be good datasets where we were efficiently able to generate the data in its smallest granularity available (daily basis) and were able to handle the null values as well in an efficient manner.

- Here, the Apple mobility index, describing the indirect information on how the fuel and energy resources were consumed, proved to be the best supporter in predicting the stock price showing that the stock price for energy and oil based companies is highly dependent on such data. Although, contrasting to the general belief, the new sentiment though very important and a good factor, proved out to be least effective in having an influence on the stock price.

8. Acknowledgement:

- We heartily thank our mentor Dr. Marcin Abram for helping us through the journey which was this competition.
- We also thank RMDS Labs for giving this opportunity to work for this project.
- This competition has helped us a lot in improving our knowledge on the stock market datasets and what features and entities that might affect them, hence we would also like to thank WorldData.Ai for giving us this wonderful opportunity to work on a real-life stock market data.

Further Codes of the analysis can be found at https://github.com/DhruvilATrivedi/RMDS_Competition_Team-5ACES