Machine Learning Final Evaluation

CSE343/ECE343

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Motivation

Everyone knows how stock markets a lot of times work heavily on emotion, from this project our main goal was to analyse if that is actually true. We focus the study to only one stock, Apple, and finally try to find a correlation between the general sentiment of the company on Twitter and it’s stock price.

Literature Review

Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. *Journal of computational science:* They found a positive correlation between the sentiment of the tweets and DJIA values, and how simple text processing techniques can be used to track the sentiment of tweets, it also said how socio-cultural events also cause a lot of mood shifts which may or may not be predicted correctly by a machine learning model.

Literature Review

Mittal, A. and Goel, A., 2012. Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011):* They built upon the paper cited above, to neural networks and concluded that they perform quite well in terms of predicting DJIA values when trained on a feature set of DJIA values.

Dataset description and Generation

Tweets on Company: The original dataset has tweets from all other major tech companies from which we only focus on Apple in this study. This dataset contains tweets, retweets, epoch timestamps, comments, likes and the name of the user who has made the tweet. (source: Kaggle)

Historical Stock Data: Contains the open, close, high, low, volume traded for the day.

Final Dataset: for each day, contains the twitter sentiment (Discussed Later), twitter sentiment for previous day, open of current day, close of previous day, high of previous day, low of previous day, and volume of previous day, and we try to predict the close of current day.

Data preprocessing

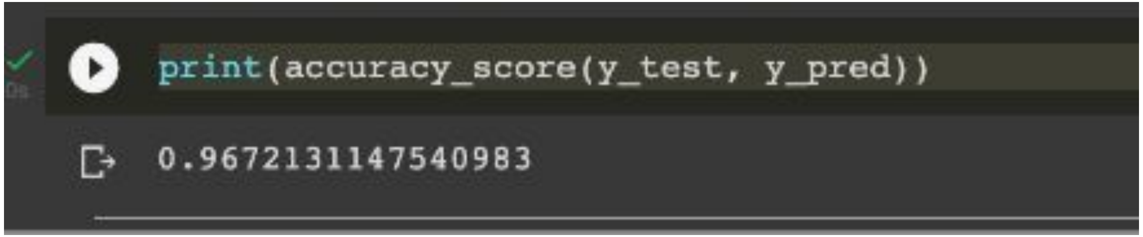
Preprocessing done on Company Tweet Data:

1) Conversion from epoch timestamps to human readable date formats.
2) Converting the format of date from the format obtained in dataset to the format which will be used by us (YYYY-MM-DD).
3) Creation of an engagement parameter which is sum of likes,comments and retweets.
4) Taking top 20 tweets based on engagement.
5) Cleaning of tweet text to allow sentiment analysis and then calculate polarity of tweets.
6) Clustering by dates from 2015 to 2016 and taking weighted average of polarity and engagement.

The problem with our mid evaluation results

In the mid evaluation, our idea was correct but we were taking the wrong values into consideration. We were considering the high value, close value and the open value. But using these values as parameters while training would be absolutely useless as any change in them reflects the overall sentiment. So, knowing these values beforehand is in fact equivalent to knowing the sentiment of the day.

Basically, we had already shown the model the data we were using for validation. That resulted in the unexpected high accuracy.



```
print(accuracy_score(y_test, y_pred))
```

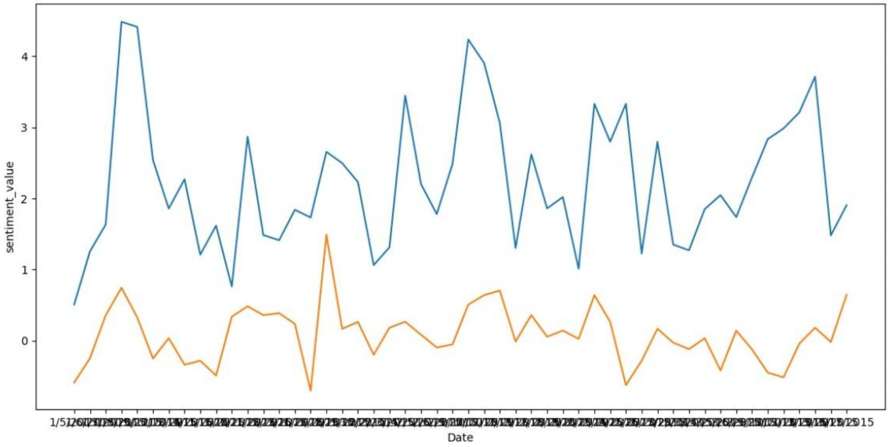
```
0.9672131147540983
```

Methodology

- Scraped the tweets of Apple, and cleaned them to get the final data set.
- Ran a sentiment analysis - Used TextBlob as the model for each tweet.
- The average of all sentiments of a day is the sentiment of a day.
- Plot the sentiment, change in stock price vs day to check even there is a relation between the two that can be learnt.
- Either solve this as a classification problem or a regression problem, use different models, Linear Regression, Decision Trees, Random Forests, Support Vector Machines, and finally LSTM Neural Networks.

Does sentiment of Tweets influence the Stock Prices

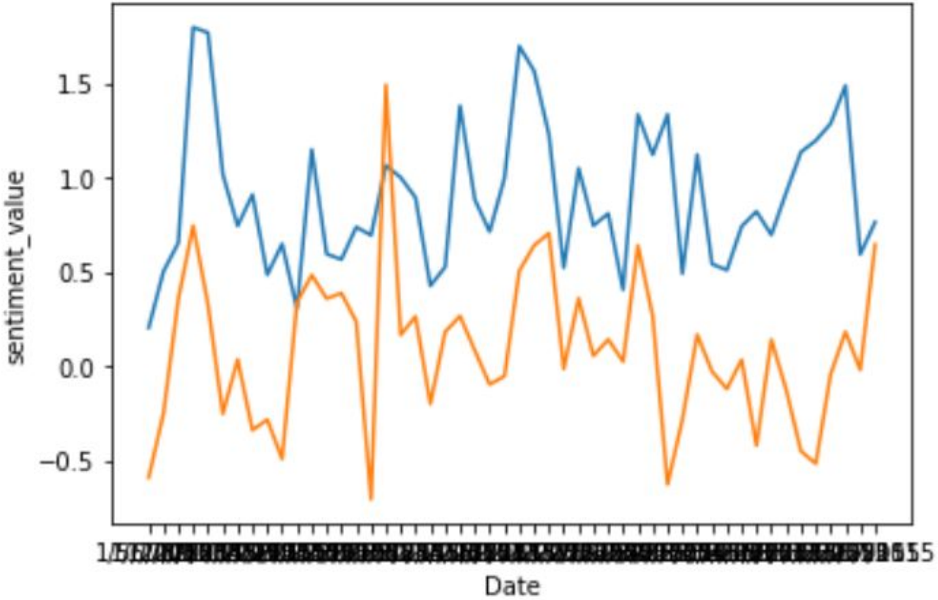
This study becomes pointless incase there is no correlation between the two, however as seen in the plot, where the blue line represents the change in stock value from previous day, and the orange line plot represents the current sentiment. Can be seen how higher sentiment value increases the stock price.



The correlation between the two was coming out to be 0.63 which indicates they are positively correlated.

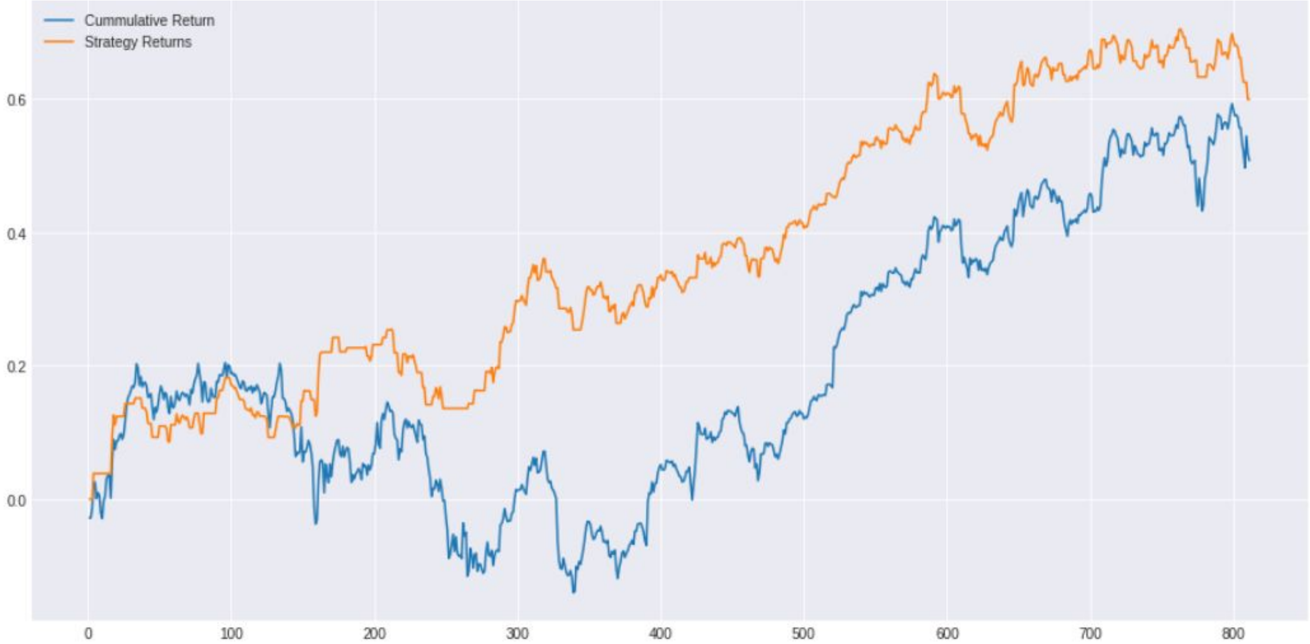
Results

We noticed that the sentiment value of some day had a direct correlation with the change in stock prize on that day as shown in the graph below.



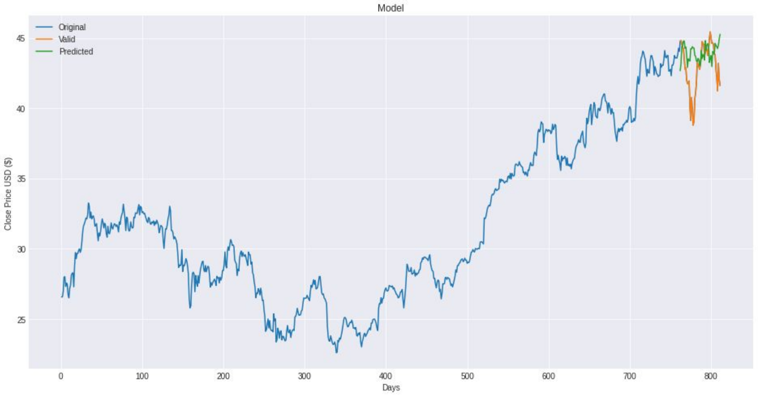
Results

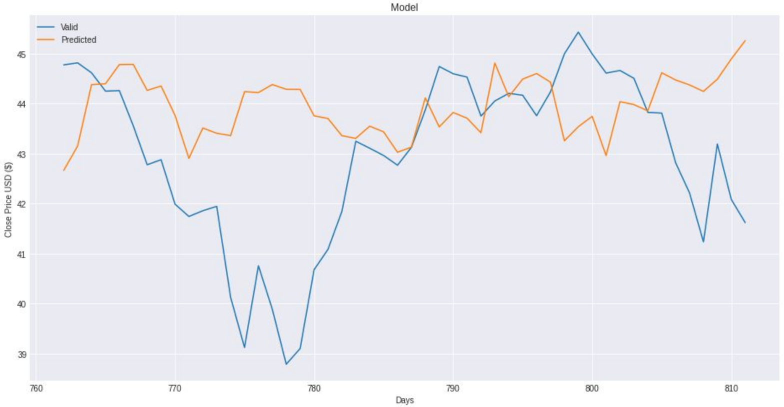
We used SVM to classify the data and we obtained an accuracy of 0.4723926380368098 which is not a good enough accuracy



Results

We try to run **linear regression** to predict the value of the stock for next fifty days. The graph was as follows and we can see that the graph for the predicted value is very different from the graph for actual value. The root mean squared error in this case is 2.1414 , so we conclude that this is not a good model and we have to do better than this.

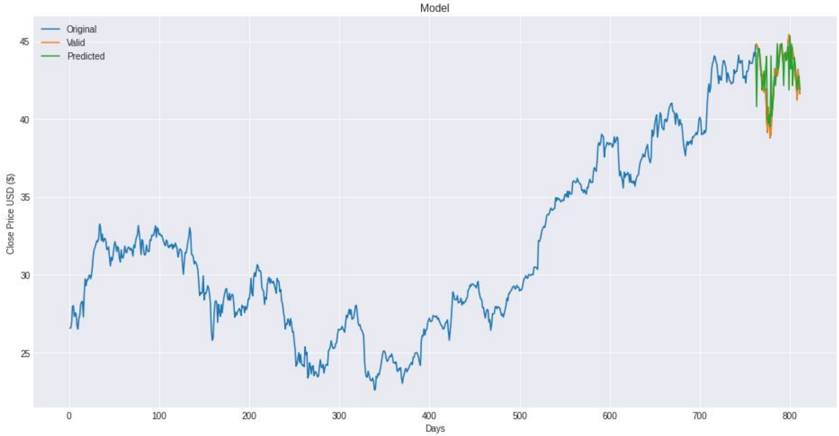


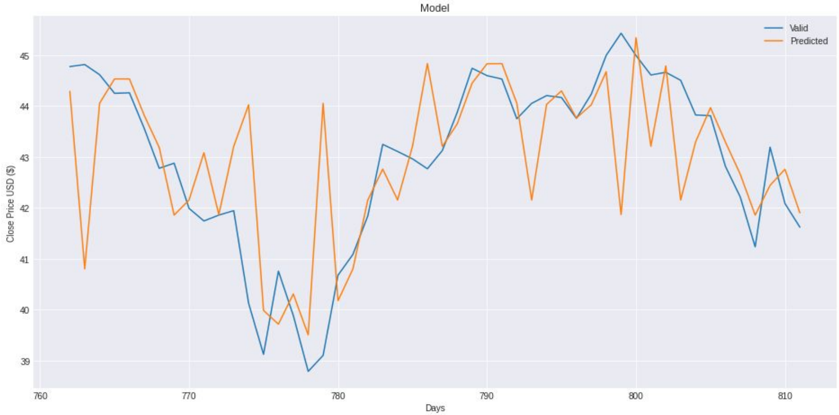


Results

Then, we used Decision Trees for the problem and we got an RMSE value of 1.2647

This is better as compared to linear regression on the data as shown by the figures.

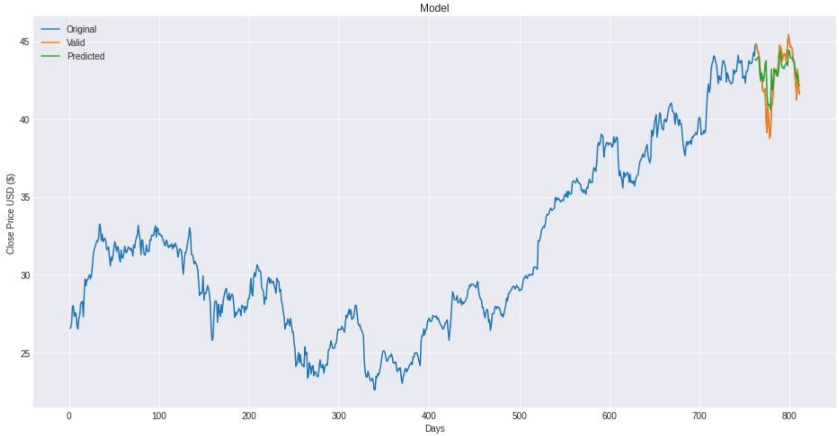


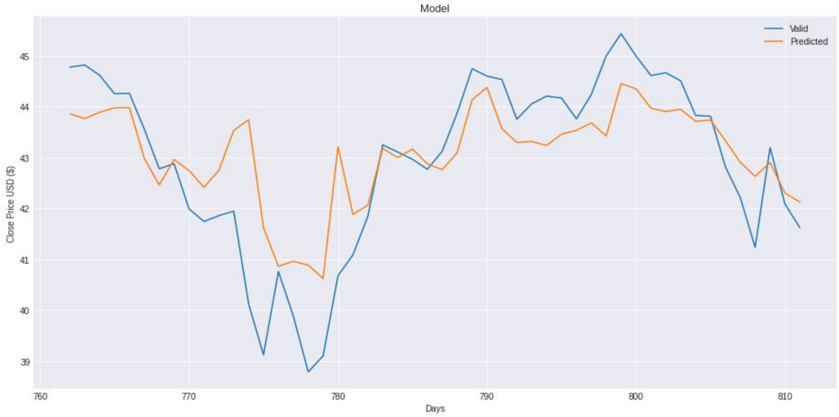


Results

Then, we used Random forest for the problem and we got an RMSE value of 1.0505. So, the RMSE value goes down on switching to random forest.

This is better as compared to all the models till now on the data as shown by the figures.

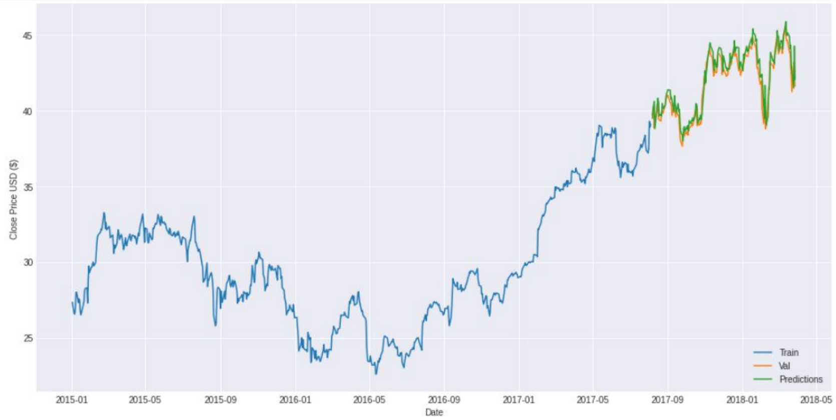


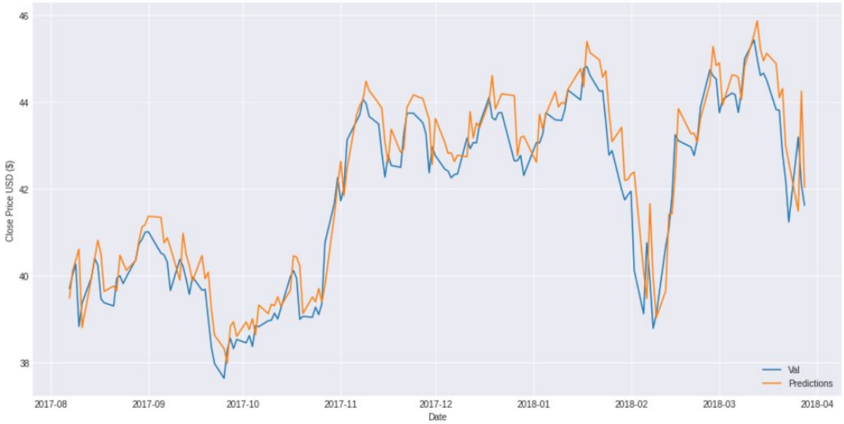


Results

We finally used the LSTM model (long short term memory), which greatly improved the results as shown in the figures

The RMSE value for this model was 0.3650404639802157 which is very small as compared to all the models till now on the data as shown by the figures.





Results

The overall results were as follows:

Model	Accuracy /RMSE
Support vector machine (SVM)	Accuracy = 47.23926380368098 %
Linear Regression	RMSE = 2.1414
Decision Trees	RMSE = 1.2647
Random Forests	RMSE = 1.0505
LSTM	RMSE = 0.3650404639802157

Conclusion

From the models presented, we get two things, Firstly this problem is better scene as a regression problem than a classification problem as seen from the errors. Also the same was discussed in the research papers too. Another interesting thing is that yes there is a correlation between the stock price and the sentiment but it is delayed by a few days, and hence LSTMs work so well giving a RMSE 0.36, which was the best model out of all covered and Random Forests having 71% accuracy.

Lastly, stock prediction is a very complex problem, a lot of unrelated events influence stock prices which are very hard to factor in models, even the best models in literature review, if solved using a classification approach, had a 75% accuracy, and thus this study comes very close to them.

Timeline

- We successfully got the data by trying various online datasets and scraping methods
- We sanitized the dataset according to our needs
- We took the various values for the stock for each day like open value, close value, high value, low value, volume, etc.
- We tried some basic techniques to predict the sentiment of market.

Individual member contribution

Abhik - Twitter Scraping, Literature Review, LSTM, Decision Trees, Random Forests.

Sohum - Twitter Scraping, Literature Review, Correlation between Sentiment and Stock Value, TextBlob usage, LSTM.

Ishaan - Dataset generation, Literature review, Linear Regression, Logistic Regressions, Support Vector Machines.

Thank You