Predicting Stock Market Prices Using Machine Learning Models

Richard Maina
University of Nebraska
richardmaina3@gmail.com

Duc Phan
University of Nebraska

***Abstract*— The stock market is one of the most unpredictable markets known to man. This loose network of economic transactions is highly volatile due to the multitude of factors that go into determining prices on the stock market. These factors range across physical, psychological, rational and irrational behaviour just to name a few. Our hope is that by using data like the latest announcements about an organization, their quarterly revenue results and other data we can use machine learning techniques to uncover patterns and techniques to accurately predict future stock market predictions.**

Keywords—prediction, stock market.

I. Introduction

The idea of accurately predicting future stock market performance has been around as long as the stock market itself has been around. However, advancements in technology mean that we have been able to track and categorize more of the factors that affect stock market performance.

We plan to apply machine learning techniques to all the gathered data and statistics to see if we can accurately predict stock market performance. To do this we will be using linear regression, K-Nearest Neighbors (KNN), Auto Auto-Regressive Integrated Moving Averages (Auto ARIMA), and Long Short Term Memory (LSTM). LSTM is widely regarded as the most effective method for the time series data we will be using but using a range of models will allow us to better gauge and understand the performance of machine learning on stock market prediction.

II. PROBLEM DEFINITION

Given the multitude of factors that affect the stock market predicting its performance is particularly difficult. We plan to use machine learning techniques to provide a technical analysis of the stock market to identify trends in its performance. This kind of problem can be classified as a time series forecasting problem. This is because we use already gathered data to predict the future performance of time related data.

The ability to predict stock market performance is highly important because it provides investors with additional information to help them make decisions on any future trades. Investment managers would also rely on such information to make accurate recommendations to their clients. On a more personal scale we are interested in this because we actively invest in the American stock market. This way we can glean insight on future trades and how we expect them to perform. We chose this problem because predicting stock market performance and trends would be extremely helpful to us as investors.

III. DATASET

For our dataset we are using Historical prices for Tata Global beverages Limited from the National Stock Exchange of India. The dataset consists of 7 features and has 1235 samples. One of the features is a date value and all the rest are numerical real values. We did not have to recode any of the features as they were already in the form we needed them to be in.The dependent variable is the closing price as that will be used to calculate profit or loss. This field consists of real valued numbers. The link to the dataset can be found in the references.

IV. DATA PREPROCESSING

Fortunately for us the data set lacked high dimensionality and was presented to us in a form that required little to no preprocessing. The main preprocessing, we performed was to sort the data in ascending order of date. However, some of the methods we implemented required special preprocessing. For KNN and LSTM we scaled the feature set.

V. EXPLORATORY DATA ANALYSIS

To begin our exploratory data analysis we began by plotting the head of the data and followed that up with a describe. This gave us some insight on the type of data we are working with. Additionally, we plotted histograms of the data to visualize the data and better enable us to make a decision on whether or not to drop any of the features as well as our approach to normalize the data. Finally we plotted the target variable to understand how it varied in our data.

After performing this we decided that we would need to normalize our data but that each feature was important and there was no reason to reduce the dimensionality. (Refer to appendices 5 & 6)

VI. EVALUATION METRIC

To quantitatively provide an accurate technical analysis of the data we selected the closing price as the target variable. Seeing as it our target field we are using the root mean squared error (RMSE) of predicted closing scores as our metric for performance.

VII. METHODS

Linear Regression

This is a model we have implanted in class and we chose to apply it because it creates linear boundaries. We are hoping to exploit this quality in our prediction of stock market prices.

KNN

This is another method we have encountered in class. We chose this method because of its ability to find similarity between new data points and old data points. We hope to leverage this ability and use KNN to find trends in past data and use it to predict the new data. Really the only feature we can alter int this model is the number of neighbors.

Auto ARIMA

ARIMA is a very popular statistical method for time series forecasting. ARIMA models take into account the past values to predict the future values. There are three important parameters in ARIMA:

- p (past values used for forecasting the next value)
- q (past forecast errors used to predict the future values)
- d (order of differencing)

Parameter tuning for ARIMA consumes a lot of time. So we will use auto ARIMA which automatically selects the best combination of (p,q,d) that provides the least error. We chose this model because it is a state of the art model in predicting time series data which is perfect for prediction of future stock performance

LSTM

LSTMs are widely used for sequence prediction problems and have proven to be extremely effective. The reason they work so well is because LSTM is able to store past information that is important, and forget the information that is not. LSTM has three gates:

- The input gate: The input gate adds information to the cell state
- The forget gate: It removes the information that is no longer required by the model
- The output gate: Output Gate at LSTM selects the information to be shown as output

We chose this model because it is a state of the art tool as well. It will be a good basis for us to compare our other models against.

VIII. SUMMARY OF RESULTS

Among four models Linear Regression, K-Nearest Neighbors, Auto ARIMA, and Long Short Term Memory, LSTM performed the best for our dataset with the lowest RMSE value.

Linear Regression

For Linear Regression, the RMSE value is 120.97, which is a very high error and indicative that the model did not do a good job of predicting the data. However, this was somewhat expected as linear regression is a somewhat simple model especially when it comes to time series data as it tends to fixate on the date causing it to overfit on that feature. Instead of taking into account the previous values from the point of prediction, the model will consider the value from the same date a month ago, or the same date/month a year ago. (Refer to Appendix 1)

KNN

For kNN, the RMSE value is 115.47, this is a slight improvement over linear regression but still shows that the model was not very effective. This was not a surprise because just like linear regression kNN was not expected to perform well on this time series data. Majority of regression algorithms as a whole would not perform well on this. For kNN specifically this could be caused by kNNs reliance on the dates. I make this assumption because kNN identified a drop in January 2018 since that has been the pattern for the past years.

Auto ARIMA

For auto ARIMA, the RMSE value is 45.48, which is a significant improvement from the two previous techniques. This improvement was largely expected as ARIMA is a popular model for time series forecasting because of its ability to utilize past values to predict future values. The model was able to understand the pattern and predict an increasing trend in the series. Although the predictions using this technique are far better than that of the previously implemented machine learning models, these predictions are still not close to the real values.

LSTM

For the LSTM model, the RMSE value is amazingly low of 8.92 and the plot captures the up and down value in the test set correctly. This was expected because LSTM is widely used in sequence prediction. It's ability to store information it fees is important while forgetting what it deems unimportant makes it a highly effective classification tool. This ability to recognize important patterns for our dataset, learn from them while ignoring the patterns that are not important are the reason LSTM performed so well on our data.

IX. CONCLUSION AND FUTURE WORK

It appears that LSTM was the best model overall. This was expected as it is a state of the art model whit it comes to handling time series data. Hopefully we get chance to work with such technology in the future..

X. REFERENCES

- Dataset: https://www.quandl.com/data/NSE/TATAGLOBAL-Tata-Global-Beverages-Limited
-

XI. APPENDICIES

APPENDIX 1: Logistic Regression Classification graph

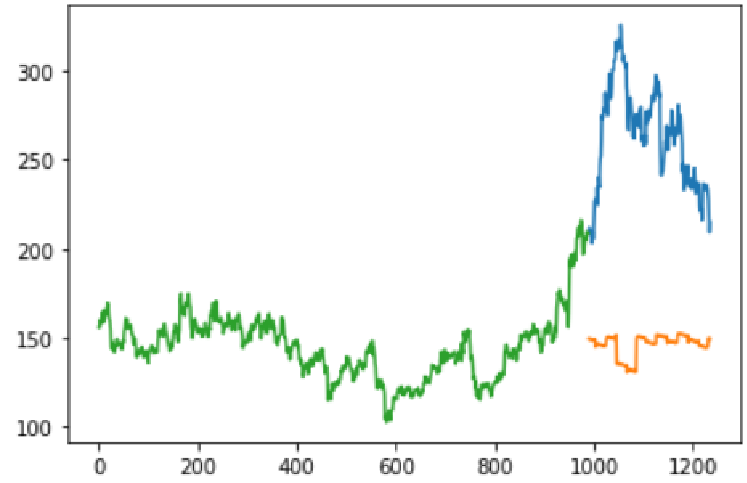


Figure 1: Graph of logistic regression prediction

APPENDIX 2: kNN Classification graph

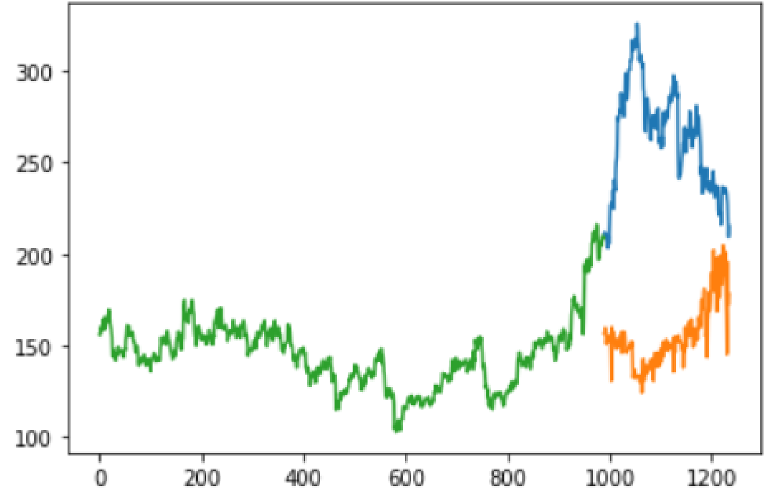


Figure 2: Graph of kNN prediction

APPENDIX 3: Auto- ARIMA Classification graph

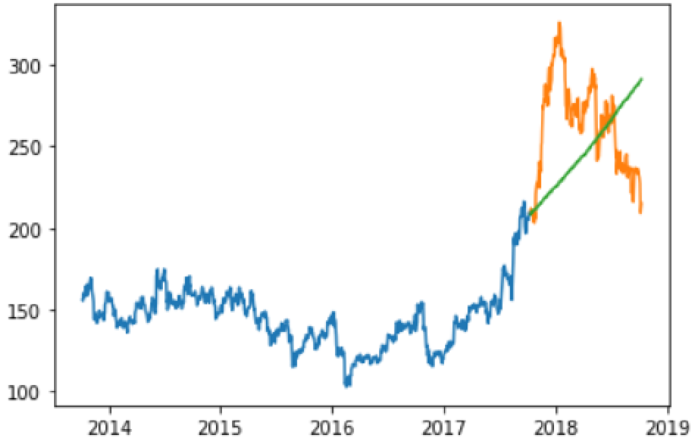


Figure 3: Graph of Auto ARIMA prediction

APPENDIX 4: LSTM Classification graph

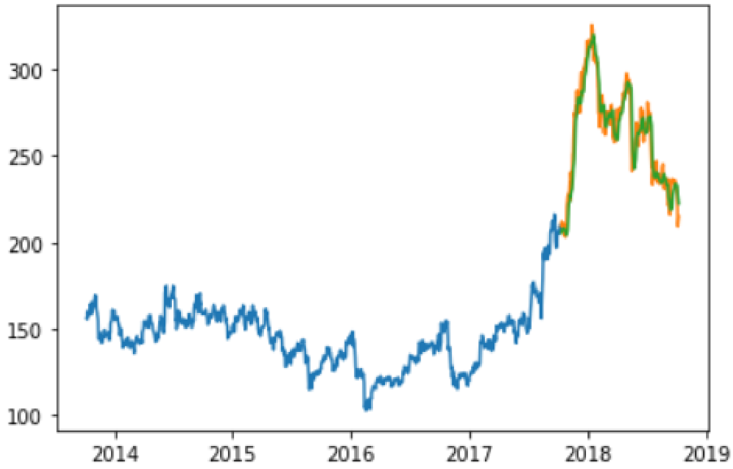
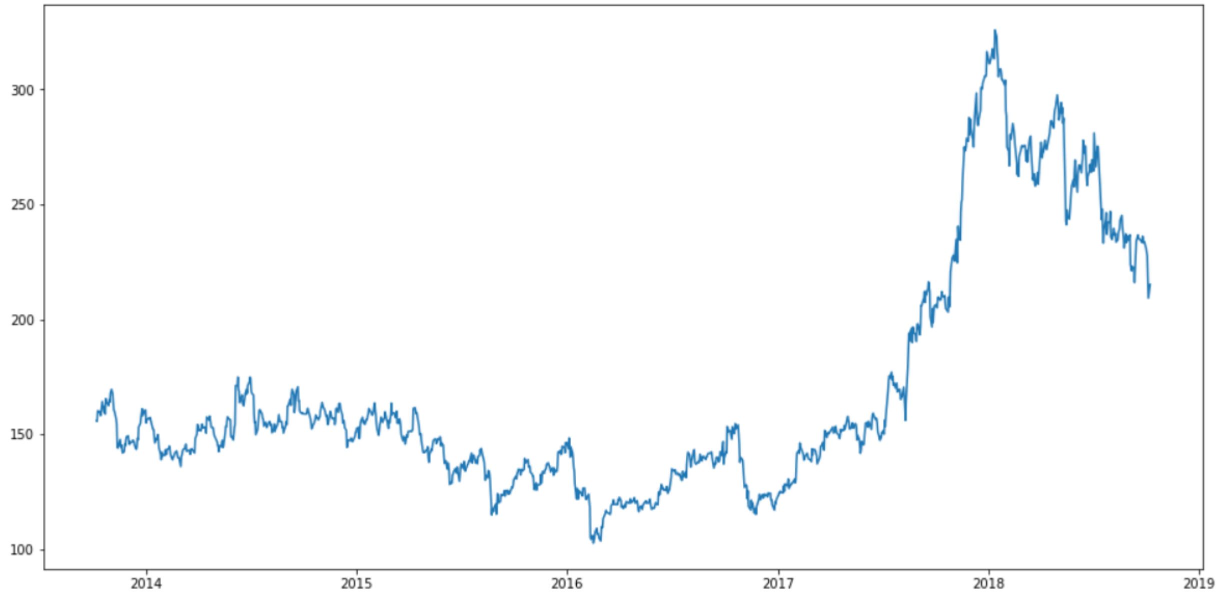
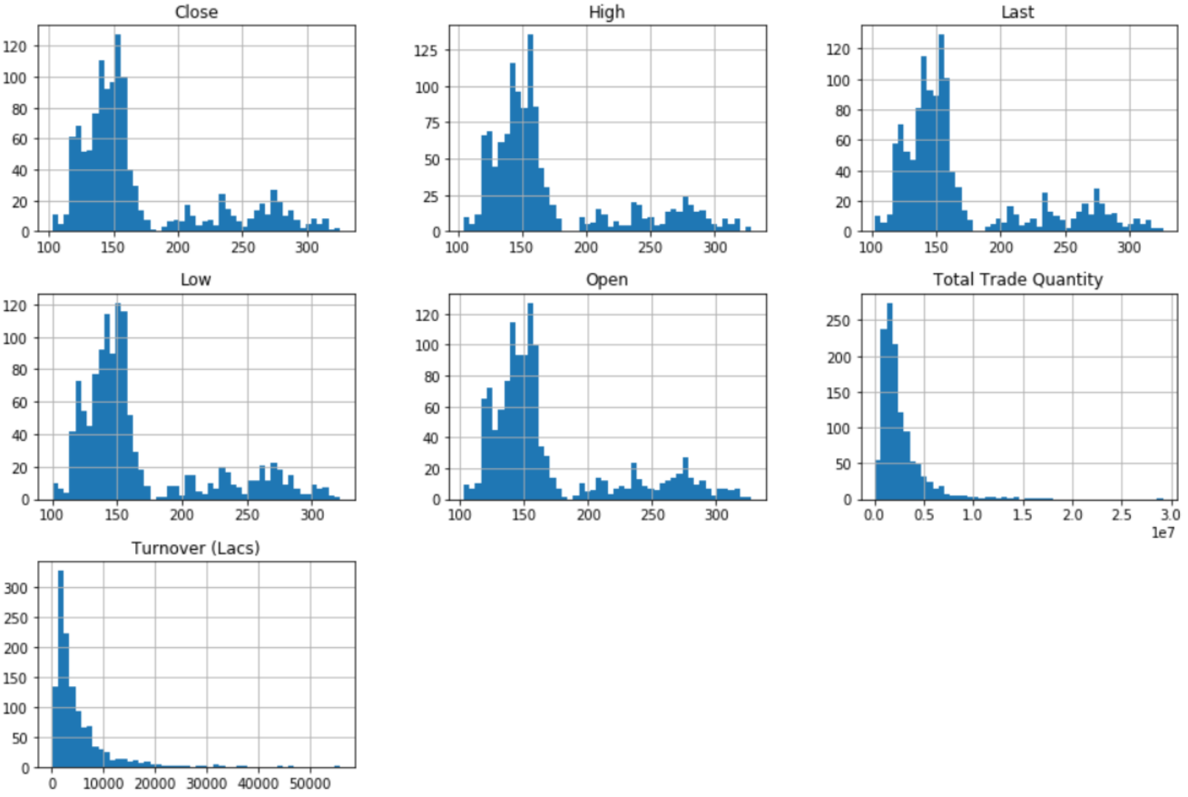


Figure 2: Graph of LSTM prediction

APPENDIX 5: EDA



APPENDIX 6: EDA


Close
High
Last
Low
Open
Total Trade Quantity
Turnover (Lacs)
1e7