Data Mining Project: FAANG Stock Prediction

Team: Data Diggers
Kunal Dhanaitkar, Anil Allikepalli, Koushik Akkenapalli

Professor Shivanjali Khare

University of New Haven
CSCI 6674: Data Mining

Spring 2022

Table of Contents

1. EMAILS AND AFFILIATION OF THE AUTHORS:

1) Anil Kumar Reddy Allikepalli (aalli2@unh.newhaven.edu)

   Department of Computer Science, Tagliatela College of Engineering, University of New Haven,West Haven, Connecticut.

2) Kunal Dhanaitkar (kdhan1@unh.newhaven.edu)

   Department of Computer Science, Tagliatela College of Engineering, University of New Haven,West Haven, Connecticut.

3) Koushik Akkenapally (kakk1@unh.newhaven.edu)

   Department of Computer Science, Tagliatela College of Engineering, University of New Haven,West Haven, Connecticut.

2. ABSTRACT

Stock price prediction is one of the most extensively studied and challenging glitches, which is acting so many academicians and industries experts from many fields comprising of economics, and business, arithmetic, and computational science. Predicting the stock market is not a simple task, mainly as a magnitude of the close to random-walk behavior of a stock time series. A good stock price prediction model will help investors, management, and decision makers in making correct and effective decisions.

In FAANG Stock Market Prediction, the aim is to predict the future value of the financial stocks of the FAANG companies namely Facebook, Apple, Amazon, Netflix, and Google. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. The paper focuses on the use of ARIMA, RNN and LSTM models based on Machine learning to predict stock values. Factors considered are open, close, low, high and volume.

3. INTRODUCTION

Stock prediction is widely used in the financial industry to help market participants determine future value. An accurate forecast of the future stock prices will be beneficial to the internal users, such as the management of the company, for planning out the capital activities. The successful prediction will also be valuable to external users, such as individual investors and institutional investors, for gaining significant profit from the market. Therefore, analysts worldwide have spent innumerable time and effort in using various theories and mechanisms to increase stocks' predictability.

The datasets we have selected are from the past few years and are from an authenticated website. We'll be working with datasets from Facebook, Apple, Amazon, Netflix, and Google (FAANG) big-tech companies. Since these five companies are the technological epitome, we would like to study their stocks in pre, during, and post COVID-19 pandemic times to see how their stocks were impacted by the pandemic that took the world by storm. We would also like to study on which company we can invest in the coming years.

4. LITERATURE AND RELATED WORK

Stock predictions are always prevailing in both industry and academia. Researchers and scholars have strived to use various methods and tools to conduct an accurate forecast. Researchers proposed a novel study on forecasting stocks of FAANG companies using HMM. The test data for analysis is split into two segments: relatively stable (I: February 2019 to June 2019) and highly vulnerable (II: February 2020 to June 2020) periods. The above-mentioned distribution in the test data analyzes the holistic performance of HMM in forecasting FAANG stocks. Using MAPE, the prediction efficiency of the FAANG HMM models is calculated to be nearly 99% for I and 97% for II. These findings reiterate the significance of HMM in stock forecasting: despite turbulent times, HMM could closely predict future stock values which are otherwise highly volatile and unpredictable. Additionally, the unique recovery of the FAANG companies justifies the value of these firms and their durability in global markets. Using more efficient algorithms apart from those considered in this study for enhancing the accuracy of the forecasted prices considering the global market's vulnerability due to any adverse circumstances that might arise in the future will be further contributions to this research work.

The previous paper did not consider integrating text mining and time series data analysis on a single stock and implementing a wide range of models to conduct the research. In our project, we are going to have exploration.

5. DATA

Our sourced data for predictive modeling is from Yahoo Finance. The datasets we have selected are from the past few years and are from an authenticated website. We'll be working with datasets from the Facebook, Apple, Amazon, Netflix, and Google (FAANG) big-tech companies.

Google's Dataset from 2018 to 2021

	High	Low	Open	Close	Volume	Adj Close
Date						
2018-01-02	1066.939941	1045.229980	1048.339966	1065.000000	1237600	1065.000000
2018-01-03	1086.290039	1063.209961	1064.310059	1082.479980	1430200	1082.479980
2018-01-04	1093.569946	1084.001953	1088.000000	1086.400024	1004600	1086.400024
2018-01-05	1104.250000	1092.000000	1094.000000	1102.229980	1279100	1102.229980
2018-01-08	1111.270020	1101.619995	1102.229980	1106.939941	1047600	1106.939941

High is the numerical data type that records the highest price at which Google traded during the trading day. Low is the numerical data type that describes the minimum price of Google in a period. Open is the numerical data type that records the price at which Google first trades upon the opening of an exchange on a trading day. Close is the numerical data type that represents the last price at which Google trades during a regular trading session. Volume is the numerical data type that explains the amount of Google stock that is traded during a daily trading period. Adj Close is the numerical data type that amends Google's closing price to reflect that Google stock's value after accounting for any corporate actions.

6. THE PROPOSED METHOD

The have used Regression in our project because the data in our dataset is numerical. Since our dataset is a collection of big 5 tech companies, i.e., FAANG and due to time- constraints, we have chosen to implement different solving techniques for different datasets, so that we can have better exposure how data mining works. We will be using proven statistical approach by implementing time series forecasting and data modeling using RNN, LSTM, Moving Averages and ARIMA. For Facebook and Amazon, we have used the RNN approach, for Netflix and Apple, we used Moving Averages and ARIMA, and for Google, we have used the LSTM approach.

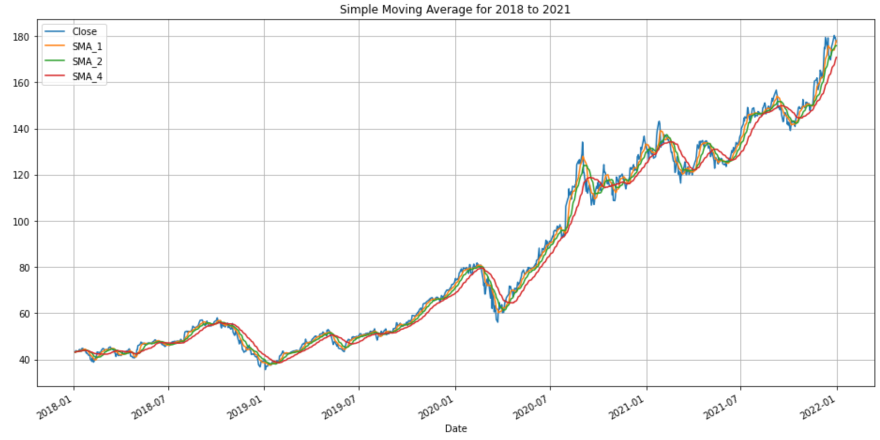
7. MODEL

7.1 MOVING AVERAGE

7.1.1 SIMPLE MOVING AVERAGE

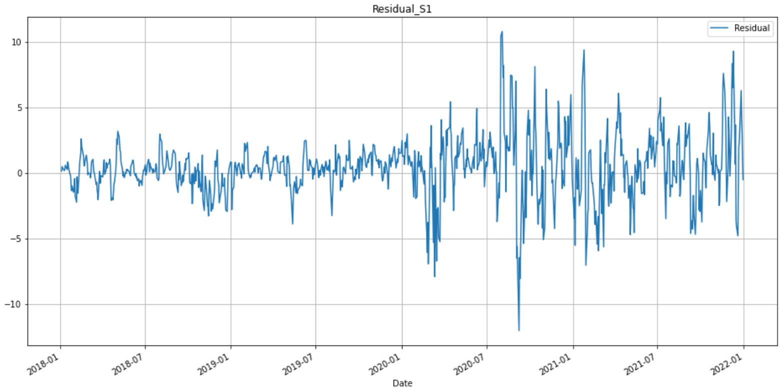
In a simple moving average model, the next value(s) in a time series is predicted based on the average of a fixed finite number of the previous values. It is an equally weighted mean of the previous data. We planned to fit the simple moving average models in our time series by setting the period as seven days, fourteen days, and twenty-eight days.

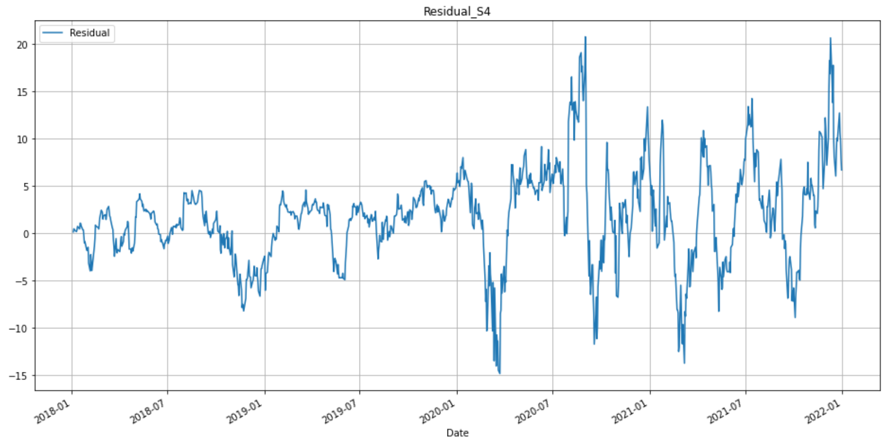
When we plotted the actual time series and three simple moving average models together, we can manually obverse that the simple moving average model with one week is the most accurate among these three models and the simple moving average model with approximately one month period has the worst performance. Therefore, for this time series, using fewer previous values can result in more accurate prediction.



To testify our manual observation, we created the residual plots for three simple moving average models, and we also calculated some model evaluation metrics, such as Mean Square Error (MSE), and Mean Absolute Error (MAE), Root Mean Square Error (RMSE).

Both the residual plots and model evaluation metrics results aligned with the previous observation we have, which is the simple moving average model with one-week period is the best-performed one among the three models. The residual of the simple moving average model with one-week period has the least fluctuation, and it also has the smallest MSE, MAE, and RMSE.



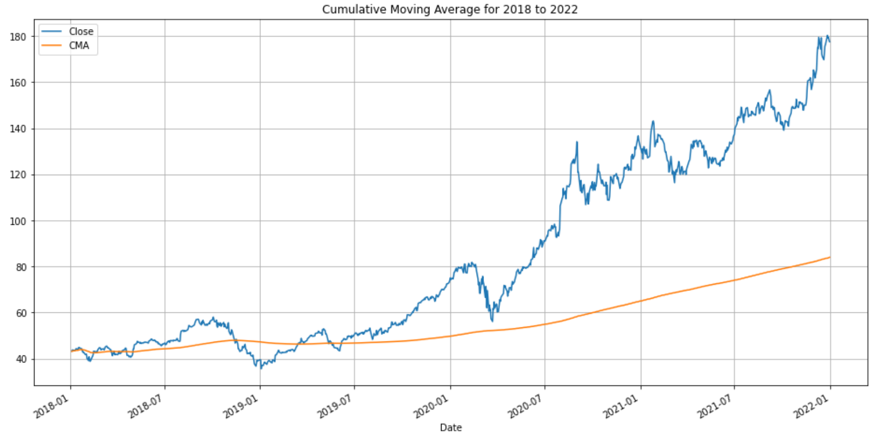


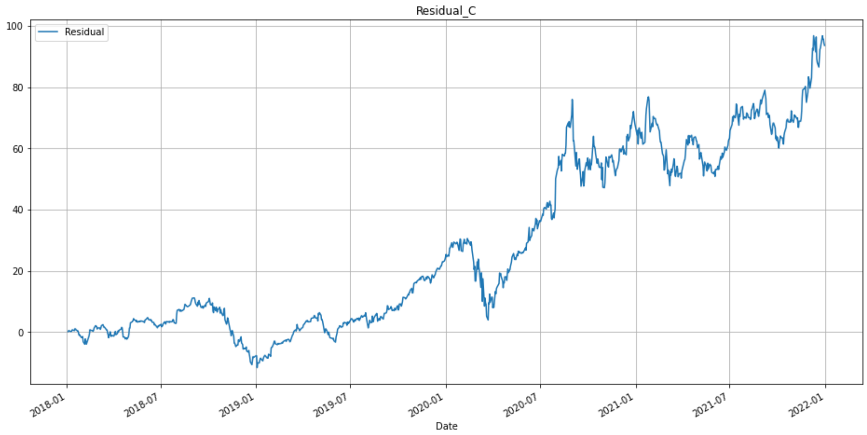
Mean Square Error (MSE): 2.8166306485739674
Mean Absolute Error (MAE): 1.1645611768425583
Root Mean Square Error (RMSE): 1.6782820527473823

7.1.2 CUMULATIVE MOVING AVERAGE

The cumulative moving average model shares the similarity with the simple moving average model. As the simple moving average, the cumulative moving average is also an equally weighted mean of the previous data. What makes the cumulative moving average model different from the simple one is that the cumulative moving average considers all prior observations. In contrast, the simple moving average will drop the oldest observation as the new one gets added to the calculation. Based on the previous experience and the mathematic equation behind the model, we realized that CMA is not a decent model to analyze the trend and smooth the time series because it takes the average of all the previous data until the current data point.

The cumulative moving average plot shows that our judgment is correct since this model has the worst performance as far and it cannot capture the variation in time series data at all. The residual plot also demonstrates that the fluctuation range is huge, which means that the prediction is extremely inaccurate.

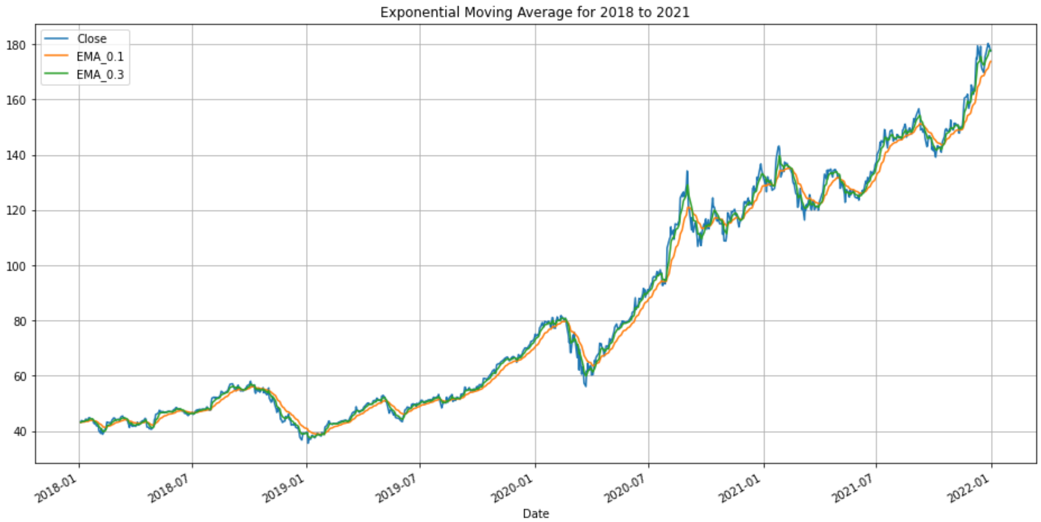


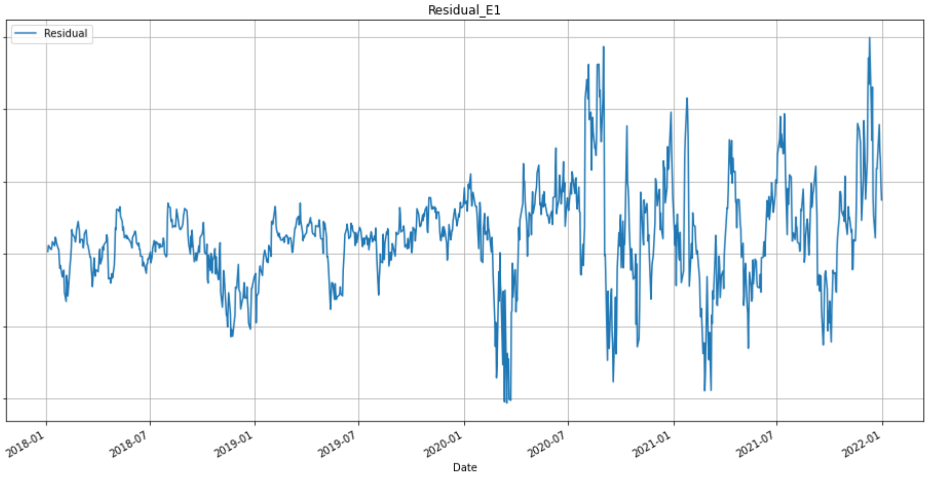


7.1.3 EXPONENTIAL MOVING AVERAGE

The intuition behind the exponential moving average model gives more weight to more recent data, which is different from the simple average moving of treating all previous data with equal weight. This improvement is comprehensible and ingenious because usually, older observations have less impact on current data. Therefore, compared with the simple moving average, the exponential moving average can capture the time series' movement more accurately and faster because it is more responsive to the latest time series changes. The exponential moving average characteristics make it a preferable model to use when it comes to moving average models.

We developed two exponential moving average models with different smoothing factor alpha, 0.1 and 0.3, respectively. From the plot, we can discover that the exponential moving average model with the smoothing factor alpha 0.3 has a more accurate prediction than the model with alpha 0.1. The same conclusion also came from the residual plots since the exponential moving average models with the smoothing factor alpha 0.3 has less error than the model with alpha 0.1.

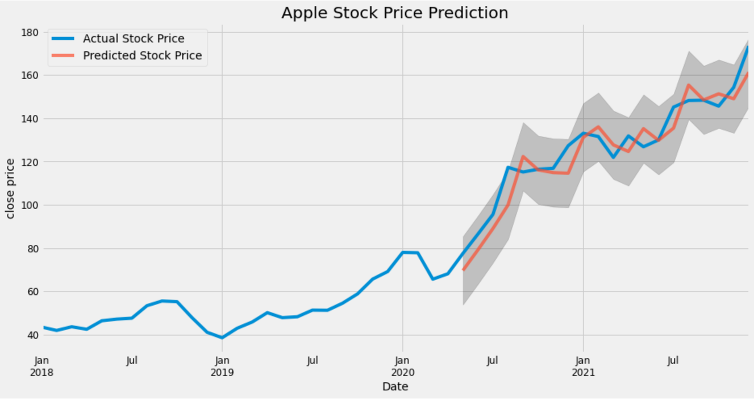




7.2 ARIMA

The Autoregressive Integrated Moving Average (ARIMA) model is widely used in time series forecasting. The ARIMA model is robust in analyzing time series because it can capture a suite of standard structures in time series data. ARIMA (p, d, q) is the standard notation for the ARIMA model where the p, d, and q will be replaced with integer values to specify the detailed ARIMA model being used. The parameter p stands for the lag order, which means the number of lag observations in the model. The parameter d represents the degree of differencing, which means the number of times that the observations are differenced. Lastly, the parameter q illustrates the order of moving average, which means the size of the moving average window. ARIMA model is advantageous in time series analysis because it considers the past values of the series and previous error terms when forecasting.

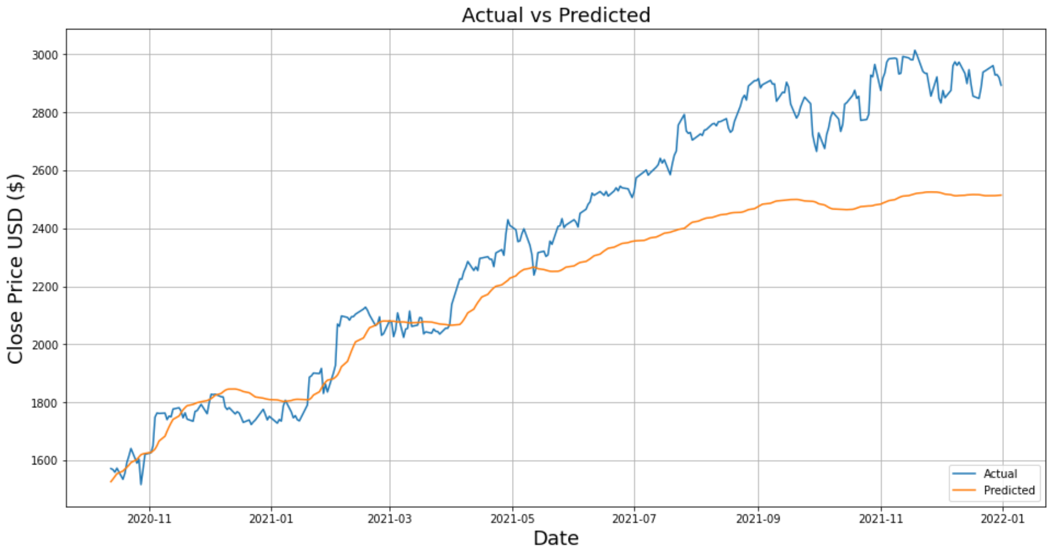
We used the auto-ARIMA in Python to find the best model without seasonality. Auto ARIMA is exceedingly productive in implementing an ARIMA model. It can help us make the time series stationary, determine d value, create ACF and PACF plots, and select the p and q values. Auto ARIMA will also return the best model that it can find based on the lowest AIC under the provided constraints. According to the result we got after running the code, the ARIMA (0,1,0) model has the best performance given the conditions. We also tried to set the seasonality as true to fit in the auto ARIMA model, and the best model we generated is still ARIMA (0,1,0).

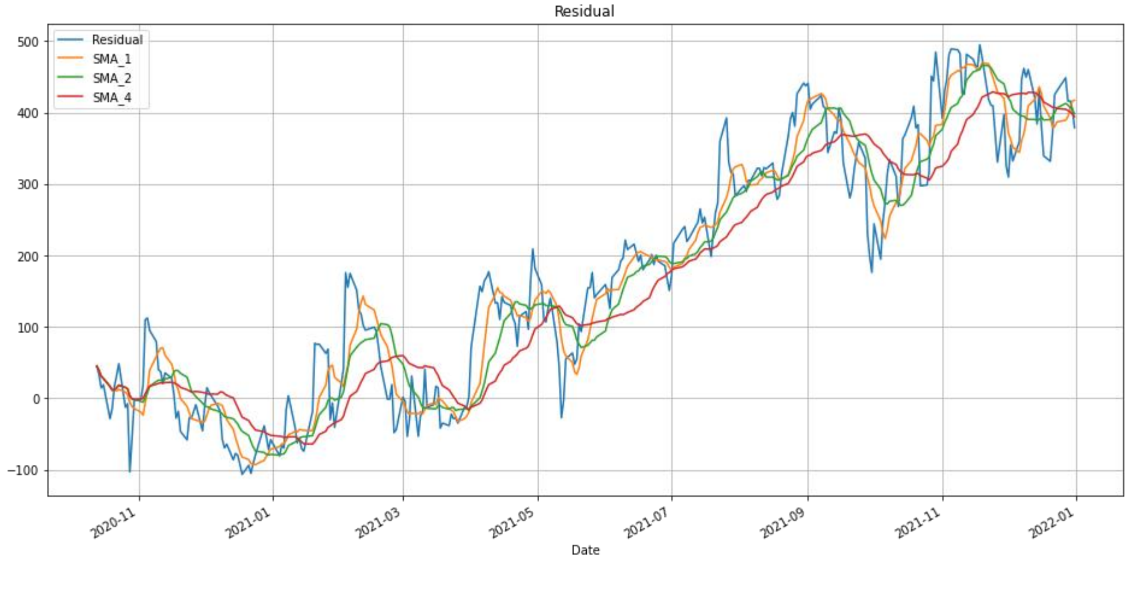


7.3 LSTM

The Long Short-Term Memory network (LSTM) is a recurrent neural network used in deep learning. Because of the feedback connections, it can be used for forecasting the time series data. Apart from learning the long sequences of data, LSTM can learn to make a one-shot multi-step forecast. In an LSTM model, five essential components allow it to model both long-term and short-term data. The five elements include cell state, hidden state, input gate, forget gate, and output gate.

After building the model and plotting the prediction with the actual time series, we can spot that thought the LSTM model can capture the trend and variation of the time series, it still has observable errors, which is also illustrated in the residual plot. Long short-term memory (LSTM) data modeling technique is used for Google's dataset.





7.4 RNN

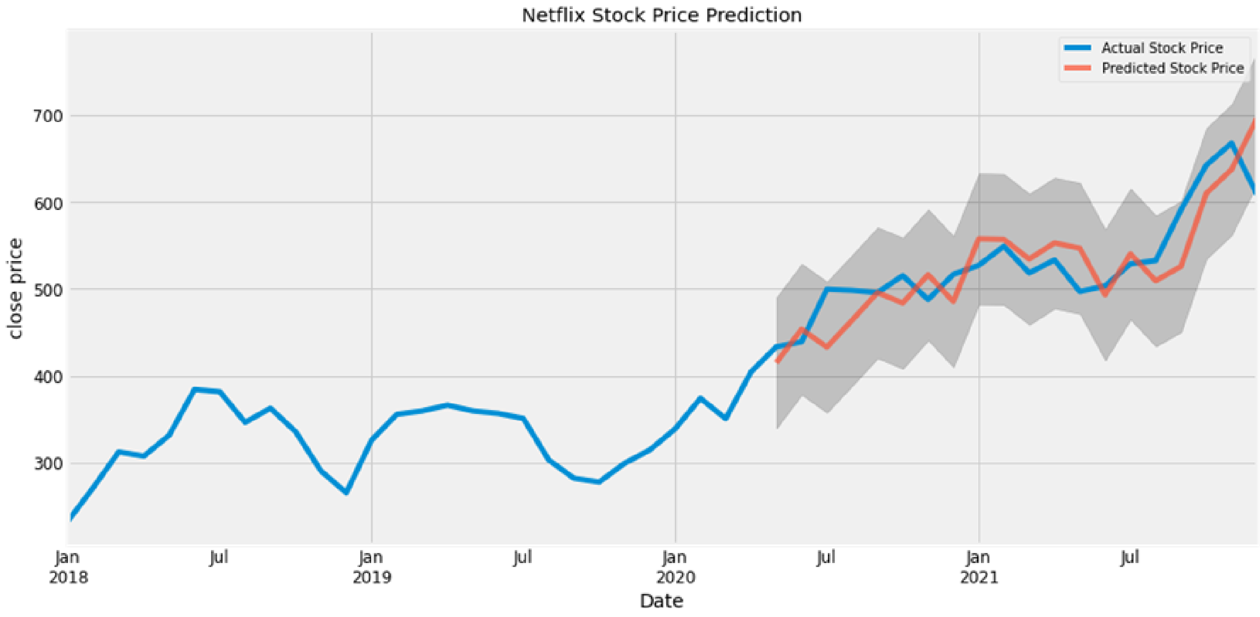
Recurrent neural networks (RNN) are the state-of-the-art algorithm for sequential data and are used by Apple's Siri and Google's voice search. It is the first algorithm that remembers its input, due to an internal memory, which makes it perfectly suited for machine learning problems that involve sequential data.

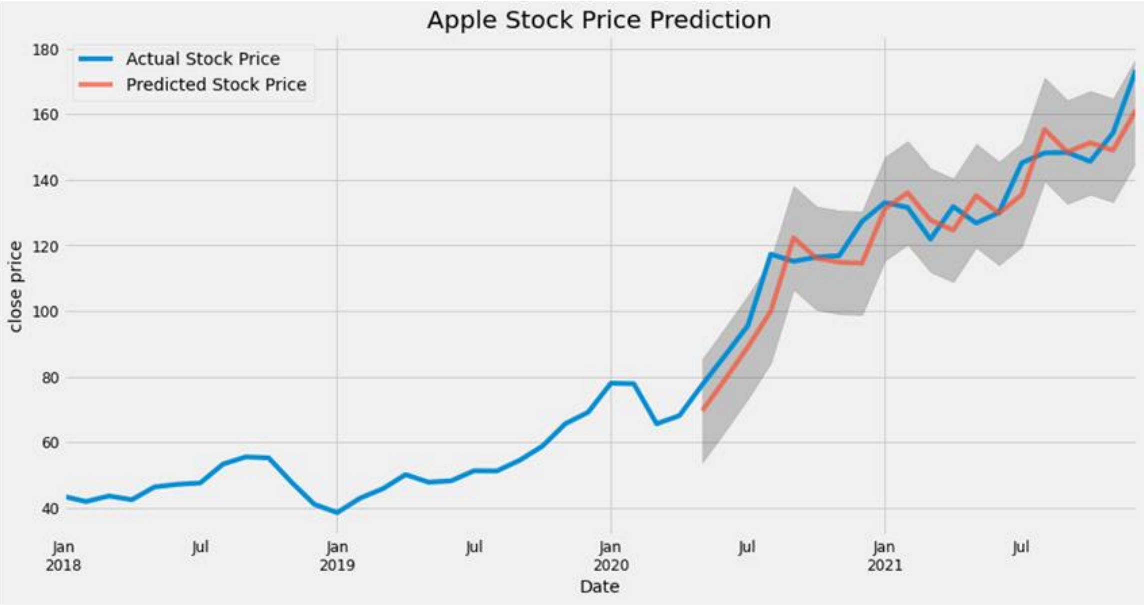
We have used Recurrent Neural Networks (RNN) to model the data for Amazon and Facebook. We preprocessed the data which generated Mean Square Error (MSE), Mean Absolute Value (MAE) and Root Mean Square Error (RMSE).

Mean Square Error (MSE): 127882.65888394567
Mean Absolute Error (MAE): 279.21269820908367
Root Mean Square Error (RMSE): 357.60684960434645

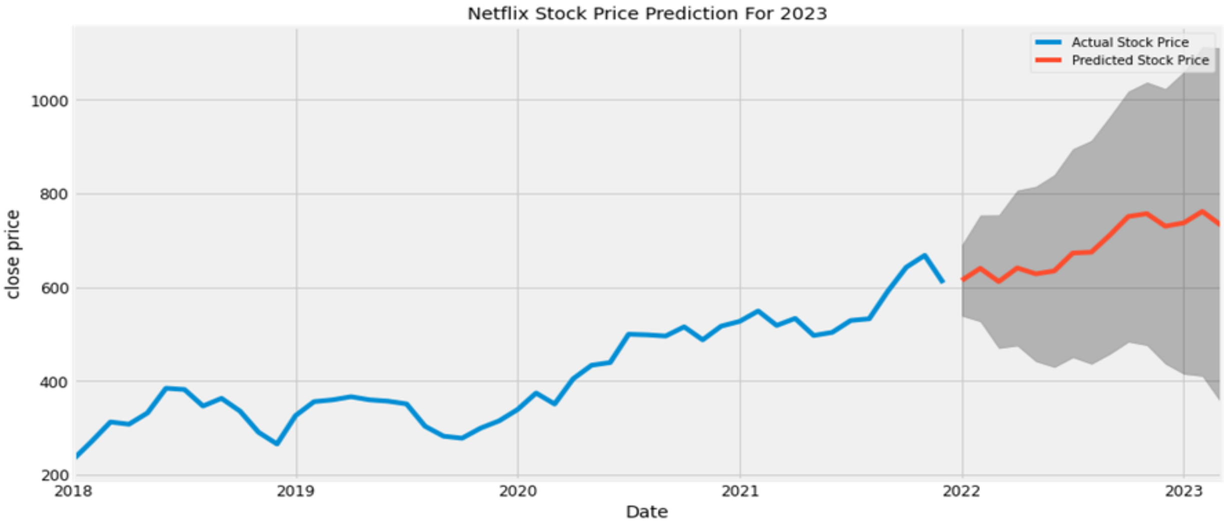
8. THE EXPERIMENTAL RESULTS

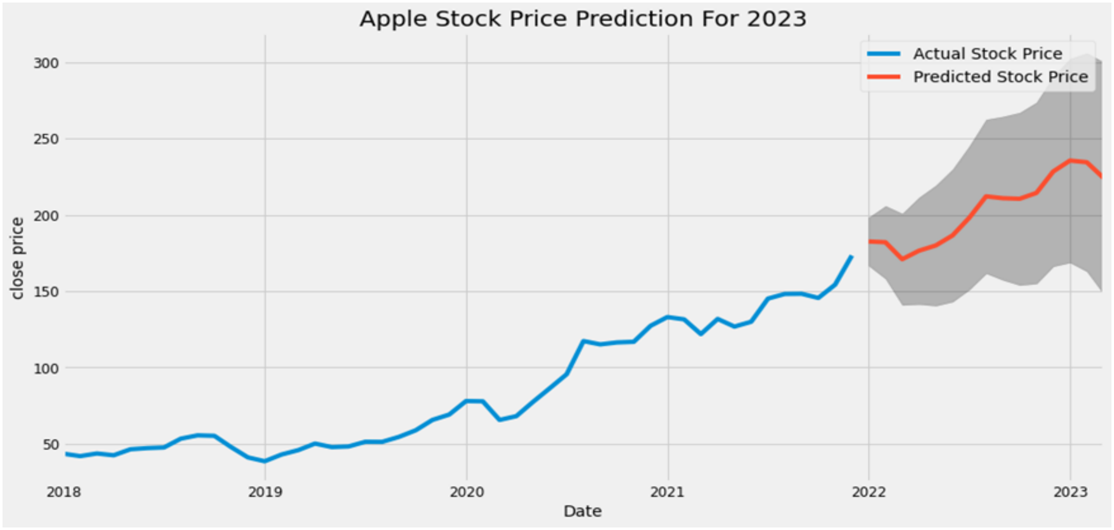
The line plot is showing the observed values compared to the rolling forecast predictions. Overall, our forecasts align with the true values very well, showing an upward trend in stock prices. The first half of the data was used to train the model and the second half for the prediction.



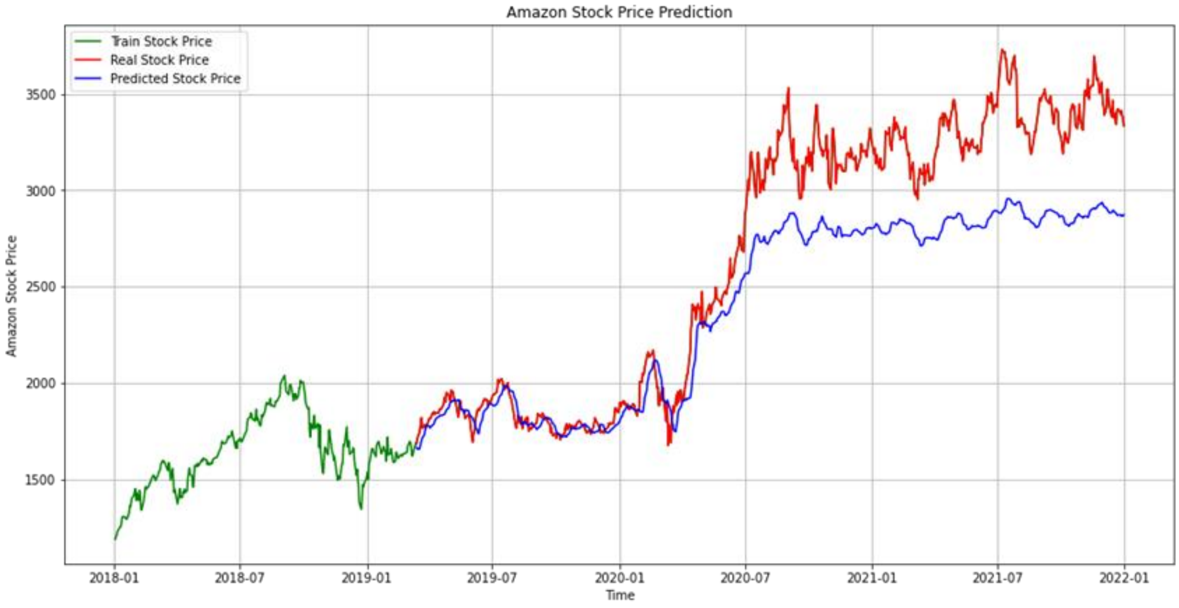


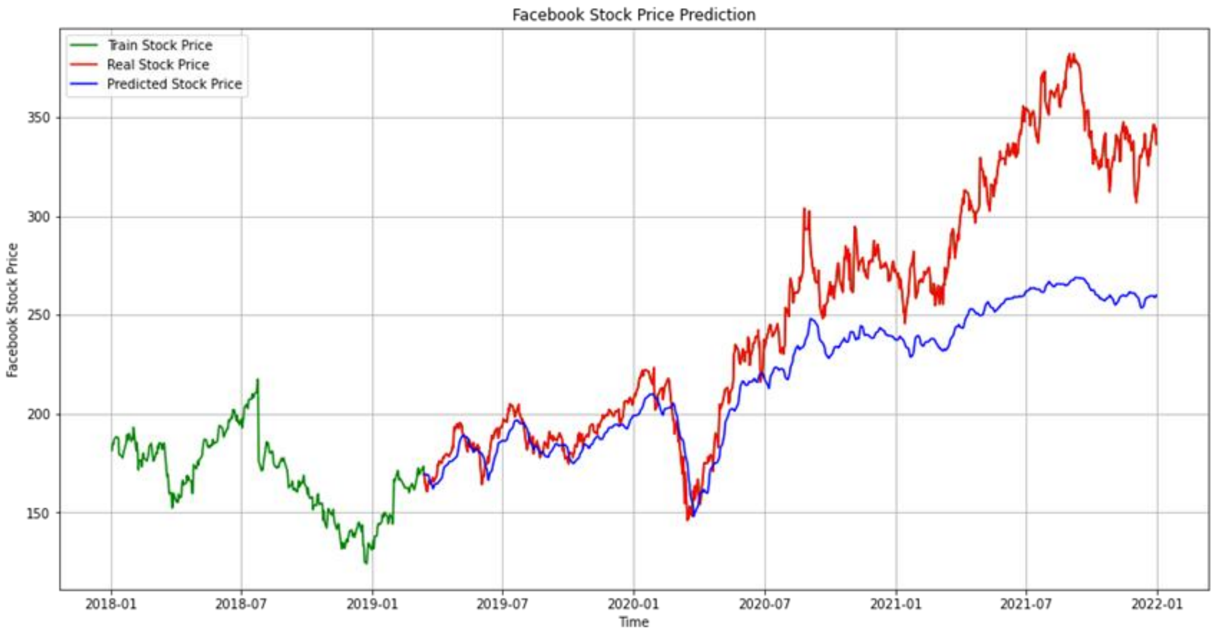
We were also able to predict the stock prices for 2023 using the ARIMA model for Apple and Netflix.



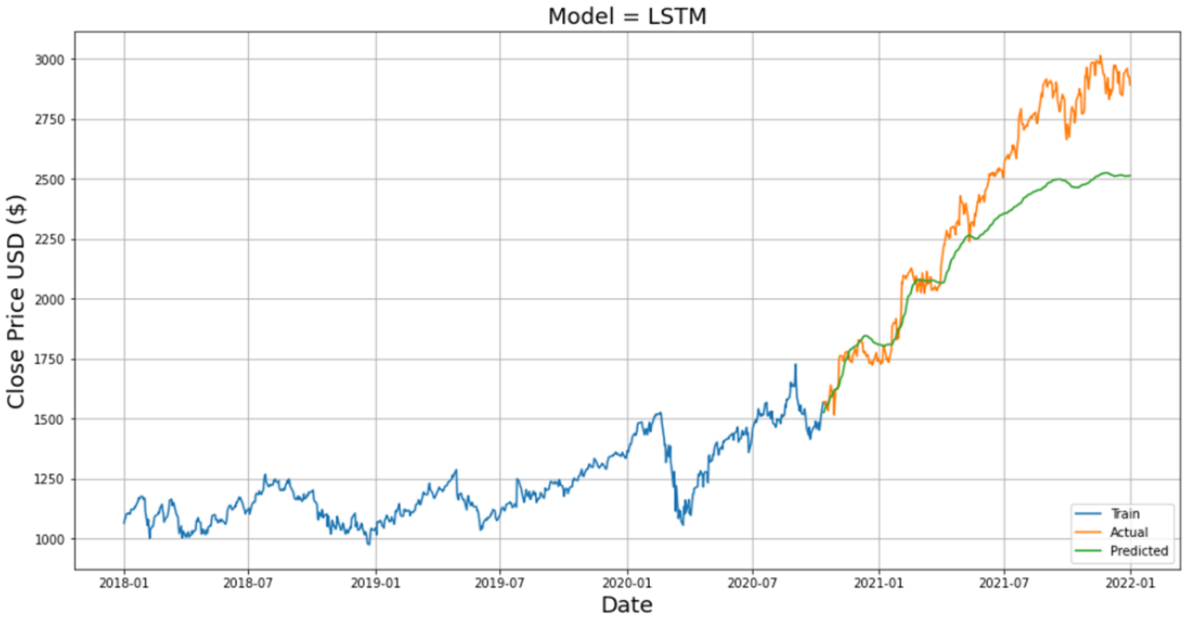


The datasets of Facebook and Amazon were trained with the Recurrent Neural Network data model. The predictions can be found below:





Long short-term memory (LSTM) data modeling technique is used for Google’s dataset.



9. DISCUSSION

The stock prediction is attractive to people ranging from brokerage analysts to ordinary investors with little financial background. Numerous tools and techniques have been exercised in the forecasting process. Our project aims to use text mining to conduct sentiment analysis among people and then use predictive modeling to select the best model that can predict FAANG's future price accurately. We hope our project can be insightful to people who are interested in stock prediction.

Market sentiment is critical in stock analysis, but it does not have the capability to accurately predict the future because most people make the wrong decisions in the equity market. Only a small portion of people can make the correct decisions and bet the market. It is well known that statistically, over time, 80 percent of the investors lose, 10 percent break even and 10 percent make money consistently. Therefore, the market sentiment is not very reliable since truth always rests with the minority.

10. CONCLUSION AND FUTURE WORK

The project applies the data mining technology of neural networks to stock price forecasts. Determining the Stock market forecasts has always been challenging work for business analysts. We attempted to make use of huge textual data to predict the stock market indices. This data was found on the official website of yahoo finance, which provides historical data for companies, and organizations. Our models clearly capture close price seasonality. As we forecast further out into the future, it is natural for us to become less confident in our values. This is reflected by the confidence intervals generated by our model, which grow larger as we move further out into the future. We conclude that the ARIMA model is more effective than the other two models providing higher returns, with a higher number of trades with profits.

We plan to apply other Techniques and compare the results for our future work. We will be using News feeds and Experts Advice from websites and taking them as datasets for Advanced predictions.

11. APPENDIX FOR LINK TO GitHub REPOSITORY

https://github.com/aanilkumarreddy/DataDiggers

12. REFERENCES

1. Analysis and Risk Management of FAAMG Stocks, Aryan Kasera, 2020 JETIR July 2020, Volume 7, Issue 7.

2. Rise of Facebook, Amazon, Apple, Netflix and Google during COVID-19 Pandemic, Shivraj Pisal California State University - San Bernardino

3. Text Mining of Stocktwits Data for Predicting Stock Prices, Mukul Jaggi, Priyanka Mandal, Shreya Narang, Usman Naseem, Matloob Khushi, School of Computer Science, The University of Sydney, Sydney

4. Jason Brownlee, A Gentle Introduction to Autocorrelation and Partial Autocorrelation, posted on February 6, 2017, https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/

5. Amanda Iglesias Moreno, Moving averages with Python, posted on July 8, 2020, https://towardsdatascience.com/moving-averages-in-python-16170e20f6c

6. Jason Brownlee, How to Create an ARIMA Model for Time Series Forecasting in Python, posted on January 9, 2017, https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

7. 14. Fares Sayah, Stock Market Analysis and Prediction using LSTM, https://www.kaggle.com/faressayah/stock-market-analysis-prediction-using-lstm

8. Eugene F. Fama, “Random Walks in Stock Market Prices”, Financial Analysts Journal (1995), 51:1, 75-80, DOI: 10.2469/faj.v51.n1.1861