Data Mining Project: Stock Market Prediction

Team: BruteForce
Naga Anvesh Kunuguntla, Karthik Kannamreddy,
Surya Teja Chinigepalli

Professor Shivanjali Khare

University of New Haven
CSCI 6401-01: Data Mining

Fall 2022

Table of Contents

EMAILS AND AFFILIATION OF THE AUTHORS

1. Naga Anvesh Kunuguntla (nkunu1@unh.newhaven.edu)

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

2. Karthik Kannamreddy (kkann2@unh.newhaven.edu)

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

3. Surya Teja Chinigepalli(schin15@unh.newhaven.edu)

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

ABSTRACT

One of the most researched and difficult problems is stock price prediction, which is being worked on by numerous academics and industry specialists from a variety of disciplines, including economics, business, mathematics, and computational science. A stock time series' near resemblance to a random walk makes it difficult to accurately predict the stock market. Making wise judgments will be made easier for investors, managers, and decision-makers with the aid of a competent stock price prediction model.

In FAANG Stock Market Prediction, which focuses on Facebook, Apple, Amazon, Netflix, and Google, the goal is to forecast the value of the financial stocks of these firms in the future. The use of machine learning, which produces forecasts based on the values of current stock market indices by training on their prior values, is a new trend in stock market prediction technology. Multiple models are used by machine learning itself to facilitate and authenticate prediction. The research focuses on the application of machine learning-based ARIMA, RNN, and LSTM models to forecast stock prices. Open, close, low, high, and volume are all factors.

INTRODUCTION

In the financial sector, stock forecasting is frequently used to assist market players in estimating future worth. The internal users, including the management of the firm, will benefit from an accurate estimate of future stock prices for organizing the capital operations. External users, such as individual investors and institutional investors, would also benefit from the successful forecast since they may utilize it to earn significantly from the market. As a result, experts from all over the world have invested a tremendous amount of time and energy into employing various ideas and procedures to make stocks more predictable.

The datasets we chose are from a verified website and date back a few years. Datasets from the big-tech "FAANG" companies—Facebook, Apple, Amazon, Netflix, and Google—will be used by us. Since these five firms represent the pinnacle of technology, we would like to examine their stock prices before, during, and after the COVID-19 epidemic to see how the pandemic that shook the world affected their prices. We also want to research potential companies for future investments.

LITERATURE AND RELATED WORK

1. Title: Stock market prediction using Hidden Markov Model

Author Names: Poonam Somani, Shreyas Talele, Suraj Sawant

Affiliation: 2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference.

Publication date: 12/21/2014 Publisher's

name: IEEE publication

https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/7065011/authors#authors

Literature Review:

Training of data is used for developing model of HMM which is done using Baum Welch algorithm. This developed model is used for testing of data using Maximum a posteriori(MAP) approach. The model then selects one best probablity value using Viterbi algorithm. The next day's closing value is given as output. For initializing HMM parameters, prior probability π and transmission probability A were made uniform across all the states. K- means algorithm initializes mean, variance and weights of Gaussian Mixture Components. Using the stock market database, this algorithm predicts next days closing price. For training of above HMM from given sequences of observations is done using the Baum-Welch algorithm which uses Expectation-Maximization (EM) to arrive at the optimal parameters for the HMM. We have taken stocks of three Banks. Training for some period is done using EM algorithm.

This data is converted into Observation vector which is further used for processing of stock prediction. Authors used Mean Absolute Percentage Error as a metric to evaluate the performanceof the algorithm. MAPE is the average absolute error between the actual stock values and the predicted stock values in percentage.

2. Title: Risk Assessment and Analysis of FAANG stocks

Author Name: Aryan Kasera

Affiliation: International Journal of Emerging Technologies and Innovative Research Publicationdate: July 2020

Publisher's name: International Research Journal Publication (IJ Publication)

https://www.jetir.org/papers/JETIR2007413.pdf

Literature Review:

In order to generate data and evaluate the return on investment for these stocks over the previous 20 years, the research mostly employs statistical techniques with Python. The analysis yields interesting findings. The first business to reach a trillion dollars, Apple, is not the best investment, according to stock market patterns for the top IT companies globally. We then note that Amazon demonstrates to give a high yield of profits with relatively minimal investment risk. While Amazon and Microsoft have a relatively low likelihood of falling over 40% in 220 trading days, Apple provides a pretty high risk of doing so. Additionally, Amazon shows the Value at Riskto be positive, indicating that there is actually relatively little risk associated with investing in the company. Although Microsoft is comparatively low risk, its returns are not as great as those produced by Amazon, and its growth pace is slower than that of Apple, Amazon, and Google.

The research's findings have given a thorough understanding of the potential for investing in the top global tech companies. The study's key finding is that, over the course of five years, investments in Amazon have shown to be the most profitable for investors. Additionally, given thehigh risk and poor return that Facebook offers, we conclude that it is not a profitable investment inthe current market.

3. Title: Forecasting FAANG Stocks using Hidden Markov Model

Author Name: Aishwarya Jadhav, Jui Kale, Chinmayi Rane, Ankit Datta, Amol Deshpande,Dayanand Ambawade

Affiliation: 2021 6th International Conference for Convergence in Technology (I2CT) Publicationdate: 04/02/2021

Publisher's name: IEEE publication

https://ieeexplore-ieee- org.unhproxy01.newhaven.edu/stamp/stamp.jsp?tp=&arnumber=9418216

Literature Review:

The report suggests a fresh research method for predicting FAANG company stock prices using HMM. The test data is divided into two sections for analysis: a time that is largely steady (I: February 2019 to June 2019) and a period that is extremely fragile (II: February 2020 to June 2020). The distribution described above in the test data examines the overall effectiveness of the HMM inpredicting FAANG stocks. The FAANG HMM models' prediction accuracy is estimated to be 97%for model I and close to 99% for model II using MAPE. These results underline the importance of HMM in stock forecasting: despite turbulence, HMM was able to accurately estimate future stock prices, which are often very unexpected and volatile. Furthermore, the distinctive recovery of the FAANG enterprises supports the worth of these companies and their longevity in international markets. Additional contributions to this research work will come from the use of more effective algorithms in addition to those taken into account in this study for improving the accuracy of the predicted prices while taking into account the vulnerability of the global market due to any unfavorable circumstances that may arise in the future.

4. Title: Stock market prediction using different neural network classification architectures

Author Name: K. Schierholt; C.H. Dagli

Affiliation: IEEE/IAFE 1996 Conference on Computational Intelligence for Financial Engineering (CIFEr)

date: 08/06/2002 Publisher's

name: IEEE

https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/501826

Literature Review:

Based on prior stock exchange data and market swings, the stock forecast will be created. The authors of the current article used FAANG data spanning 10 years. The open price, close price, percentage of change, and other factors were utilised. The FAANG businesses' stock price is predicted using several neural network classification architectural models that are based on the ALBERT model. As a result, it would be possible to rate the situation as a buy, sell, or hold.

5. Title: Stock market prediction using Hidden Markov Models

Author Name: Aditya Gupta, Bhuwan Dhingra

Affiliation: 2012 Students Conference on Engineering and Systems date: 03/12/2012

Publisher's name: IEEE

https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/6199099

Literature Review:

Author used a continuous Hidden Markov Model (CHMM) to model the stock data as a time series. The suggested algorithm was tested on four different stock indices - TATA steel, AppleInc., IBM Corporation and Dell Inc. Details of the training and testing periods. Author presented an HMM based MAP estimator for stock prediction. The model uses a latency of d days to predictthe stock value for the (d+1)st day. A MAP decision is made over all the possible values of the stock using a previously trained continuous-HMM. We assume four underlying hidden states which emit the visible observations (fractional change, fractional high, fractional low). Emission probabilities conditioned on a given state are modeled as Gaussian Mixture Models (GMMs). Themodel can be easily extended to predict stock values for more than one day in the future, however the accuracy of such predictions would decrease as we look further into the future.

THE PROPOSED METHOD

Because the data in our dataset is numerical, regression has been employed in our study. Due to time restrictions and the fact that our dataset consists of the "big five" tech firms (FAANG), we have opted to use various solution strategies on various datasets in order to gain a better understanding of how data mining functions. By incorporating time series forecasting and data modeling utilizing RNN, LSTM, Moving Averages, and ARIMA, we will be applying a tried-and-true statistical methodology. We used the RNN approach for Facebook and Amazon, Moving Averages and ARIMA for Netflix and Apple, and the LSTM approach for Google.

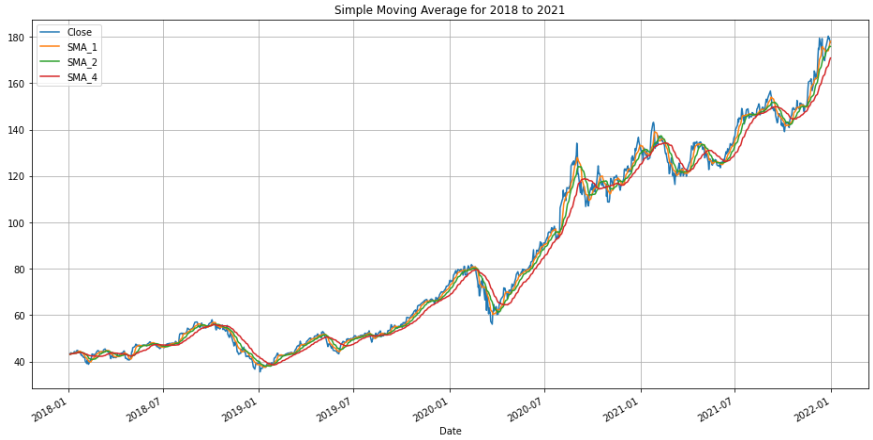
MODEL

1. MOVING AVERAGE

1.1 SIMPLE MOVING AVERAGE

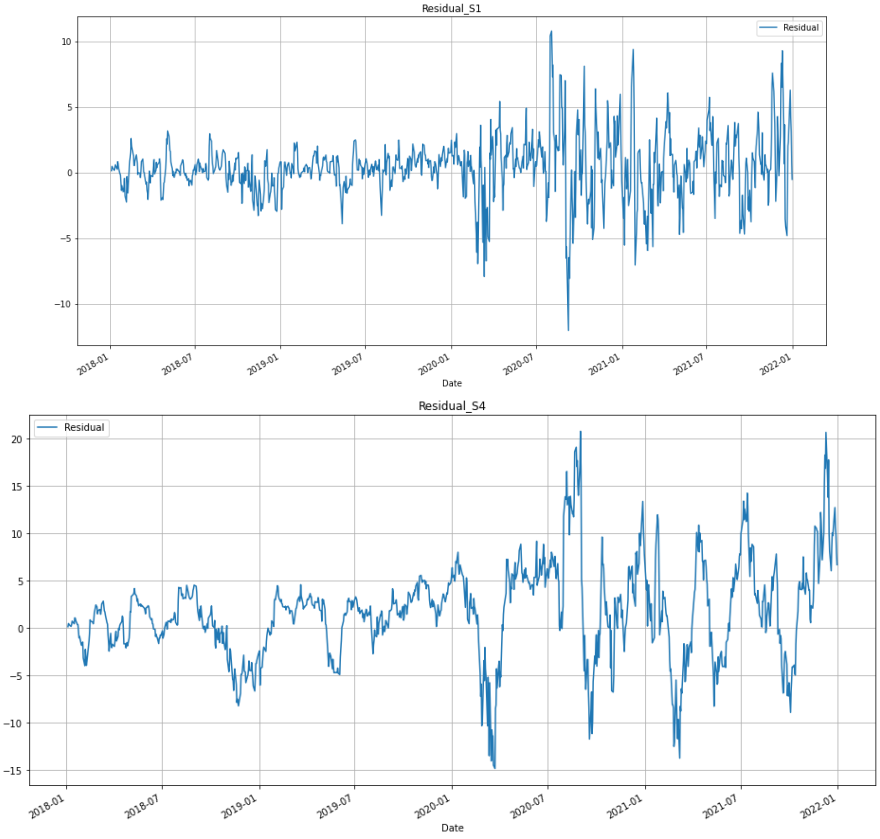
In a basic moving average model, the average of a specified finite number of the prior values is used to forecast the future value or values in a time series. It is a mean of the prior data that is equally weighted. We intended to use the periods of seven, fourteen, and twenty-eight days to fit simple moving average models to our time series.

The simple moving average model with a one-week period is the most accurate among these three models, and the simple moving average model with a roughly one-month period performs the worst, according to a manual analysis of the actual time series and the three simple moving average models plotted together. Therefore, employing fewer preceding values for this time series might lead to more accurate prediction.



We generated residual plots for three simple moving average models to support our manual observations, and we also computed various model assessment metrics, such as Mean Square Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).

The simple moving average model with a one-week period is the best-performing one among the three models, according to both the residual plots and model assessment metrics findings. The residual of the one-week simple moving average model has the smallest MSE, MAE, and RMSE as well as the least variation.

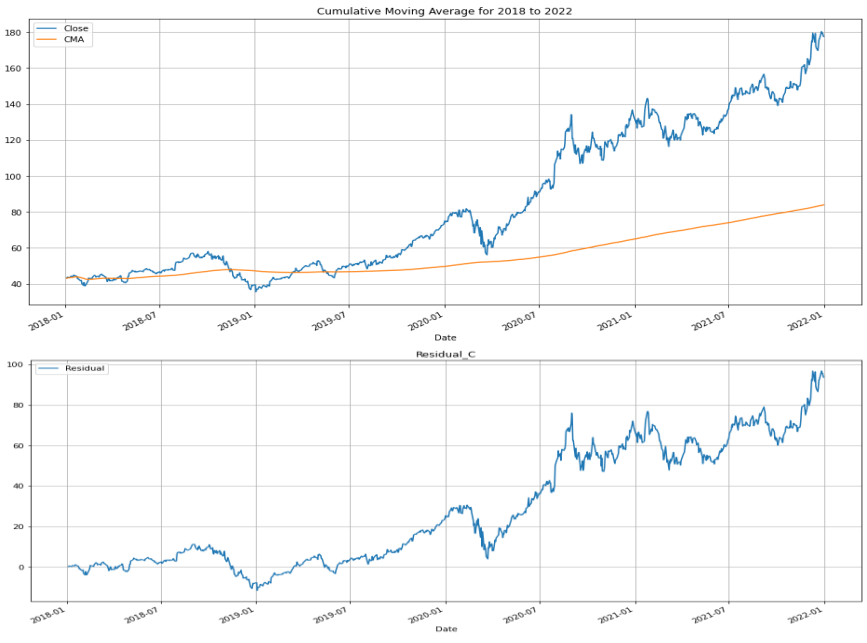


Mean Square Error (MSE): 45.81663062743927489
Mean Absolute Error (MAE): 2.162983749832723874
Root Mean Square Error (RMSE): 3.45678909897983

1.2 CUMULATIVE MOVING AVERAGE

The basic moving average model and the cumulative moving average model are comparable. The cumulative moving average is the same equally weighted mean of the prior data as the simple moving average. Because it takes into account all previous observations, the cumulative moving average model differs from the basic one. In contrast, as each new observation is added to the computation, the simple moving average will discard the oldest data. We concluded that the CMA is a poor model to examine the trend and smooth the time series since it takes the average of all the data up until the present data point. This conclusion was based on prior experience and the mathematical equation behind the model.

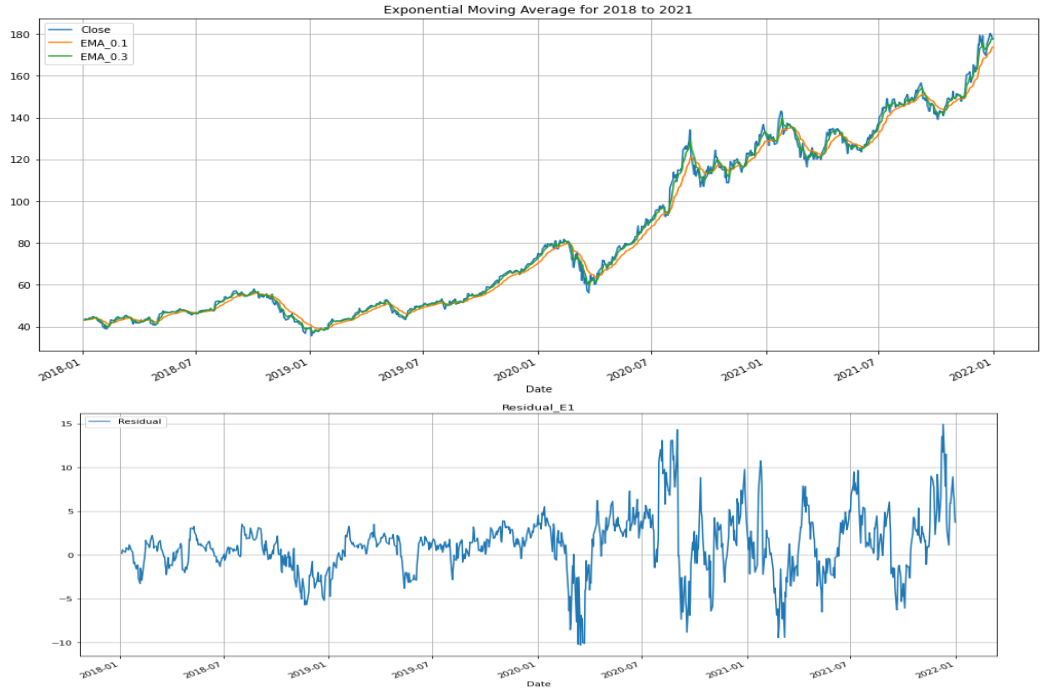
The cumulative moving average plot demonstrates that we made the right decision because this model has so far had the lowest performance and is completely unable to capture the variety in time series data. The residual plot also shows that the forecast is quite poor since the fluctuation range is very wide.



1.3 EXPONENTIAL MOVING AVERAGE

In contrast to the ordinary moving average model, which treats all prior data equally, the exponential moving average model assigns greater weight to more recent data. Because earlier observations often have less of an influence on current data, this solution makes sense and is clever. Because it is more sensitive to the most recent time series changes, the exponential moving average may therefore capture the movement of the time series more precisely and quickly than the simple moving average. When it comes to moving average models, the exponential moving average features make it a more advantageous model to utilize.

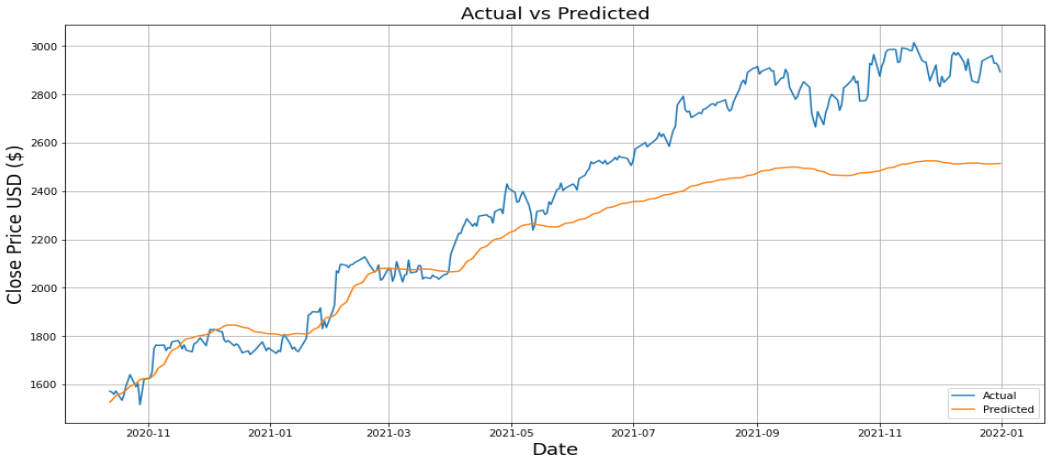
With two alternative smoothing factors, alpha, of 0.1 and 0.3, respectively, we created two exponential moving average models. According to the graphic, the exponential moving average model with smoothing factor alpha 0.3 makes predictions that are more precise than the model with smoothing factor alpha 0.1. The residual plots also led to the same result since they showed that the model with alpha 0.1 had more error than the model with alpha 0.3 for exponential moving averages.

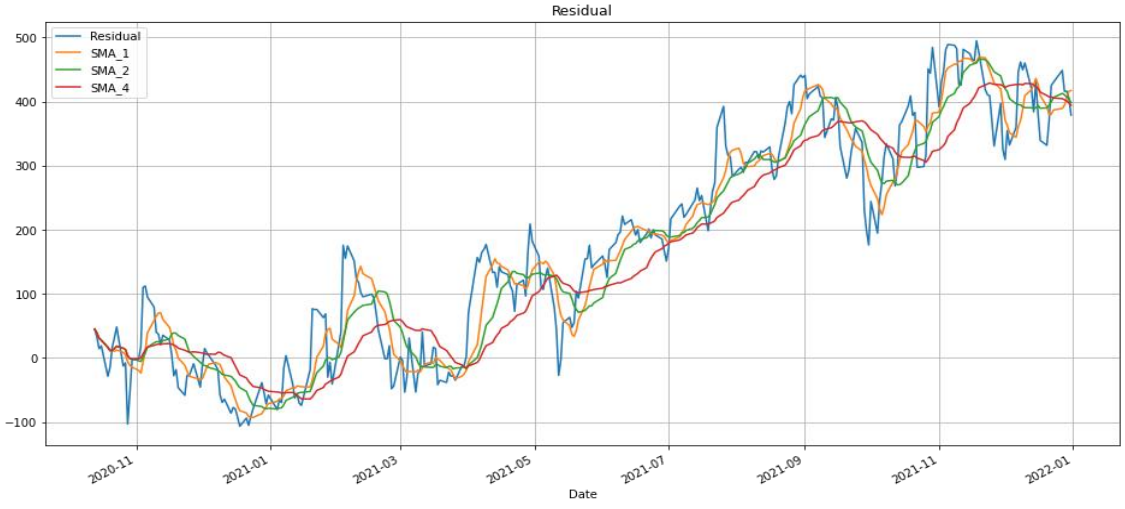


2.LSTM

Deep learning uses the Long Short-Term Memory network (LSTM), a recurrent neural network. It may be used to forecast time series data because of the feedback links. LSTM can learn to create a one-shot multi-step forecast in addition to learning the lengthy data sequences. An LSTM model can represent both long-term and short-term data thanks to five key elements. Cell state, hidden state, input gate, forget gate, and output gate are the five components.

Though the LSTM model can capture the trend and variance of the time series, it still contains obvious mistakes, which are also displayed in the residual plot, after developing the model and comparing the forecast with the actual time series. For the Google dataset, the long short-term memory (LSTM) data modeling approach is employed.





3.RNN

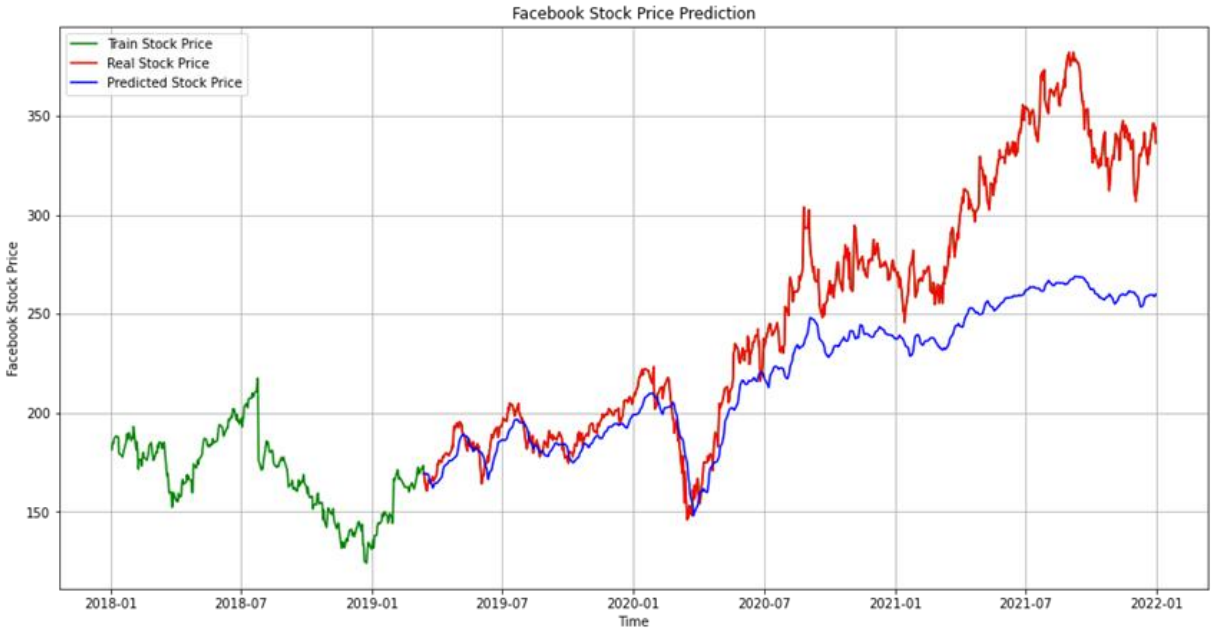
Recurrent neural networks (RNNs) are the most advanced algorithm for sequential data and are the foundation of Google voice search and Apple's Siri. Due to its internal memory, it is the first algorithm to recall its input, making it ideal for machine learning issues involving sequential data.

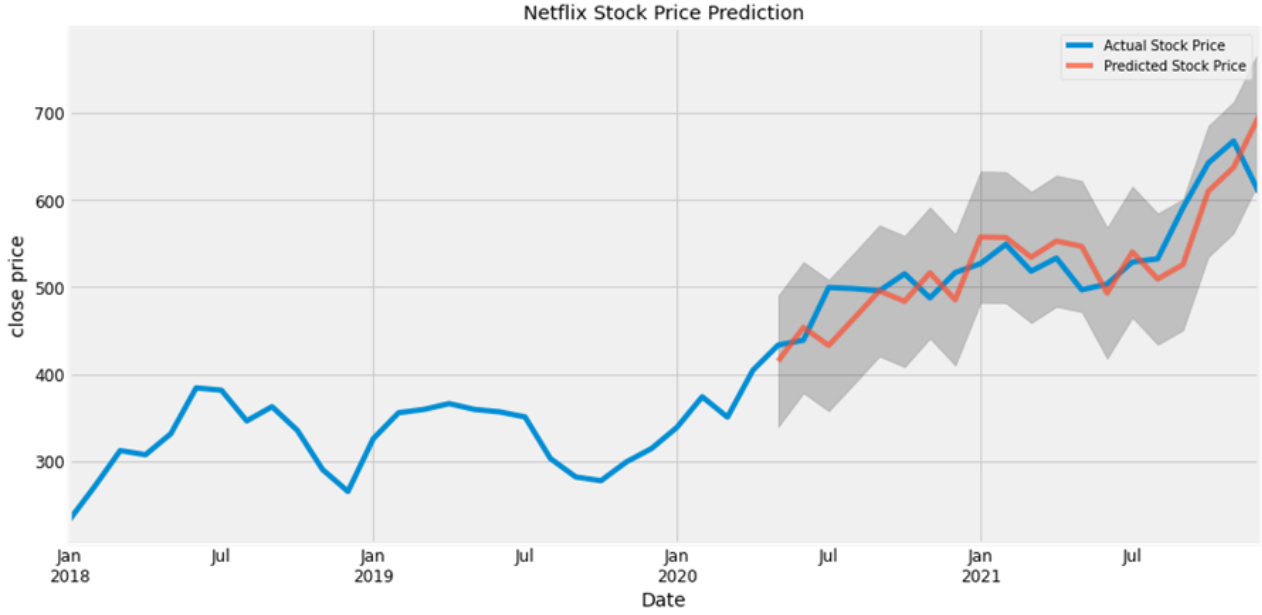
Recurrent neural networks (RNN) were utilized to model the data for Facebook and Amazon. In order to obtain Mean Square Error (MSE), Mean Absolute Value (MAE), and Root Mean Square Error, we preprocessed the data (RMSE).

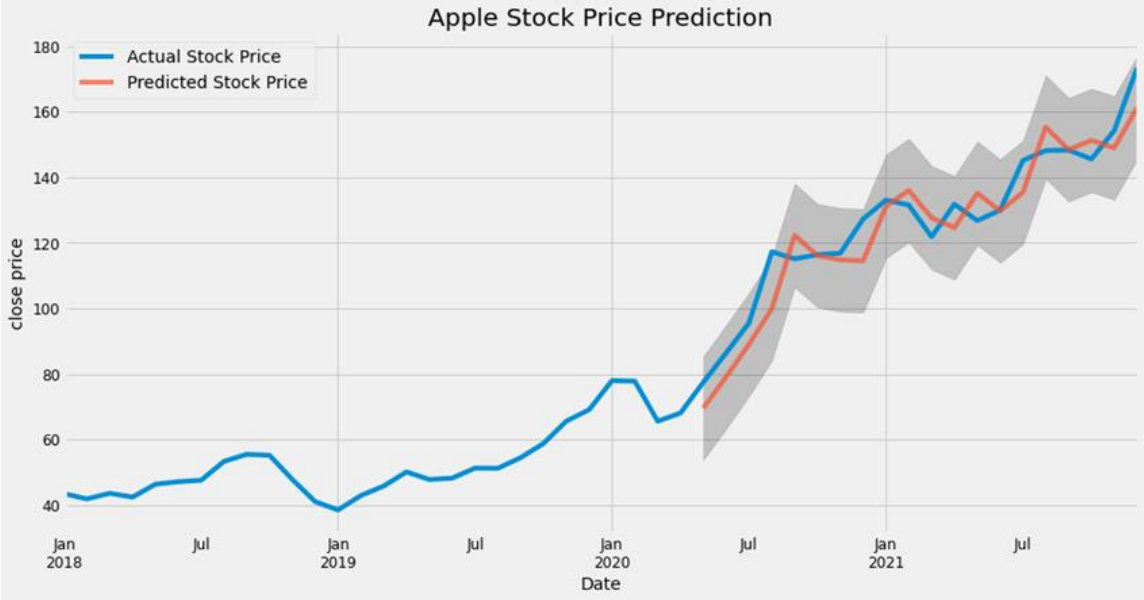
Mean Square Error (MSE): 127.782378427339
Mean Absolute Error (MAE): 25.813687812367
Root Mean Square Error (RMSE): 30.918273897837

THE EXPERIMENTAL RESULTS

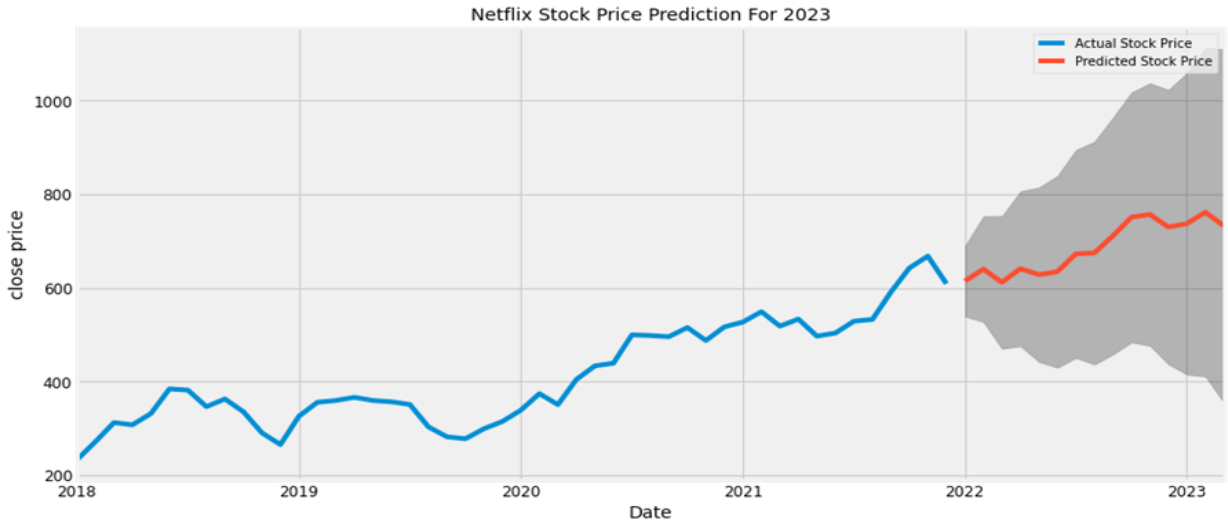
The line plot displays the observed values in relation to the predictions from the rolling forecast. Overall, our predictions closely match the real numbers, indicating a rising trend in stock prices. The model was trained using the first half of the data, and the prediction was made using the second half.

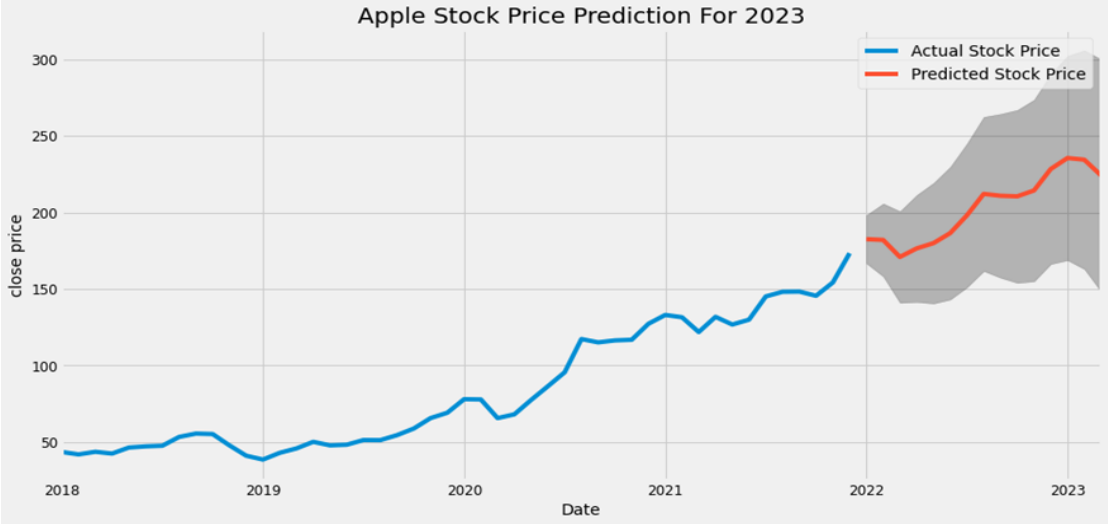

Facebook Stock Price Prediction
Train Stock Price
Real Stock Price
Predicted Stock Price
Facebook Stock Price
350
300
250
200
150
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
Time


Netflix Stock Price Prediction
Actual Stock Price
Predicted Stock Price
close price
700
600
500
400
300
Jan
2018
Jul
Jan
2019
Jul
Jan
2020
Jul
Jan
2021
Jul
Date



Using the RNN model, we were also able to forecast the stock values for Apple and Netflix for the year 2023.




Apple Stock Price Prediction For 2023
Actual Stock Price
Predicted Stock Price
close price
300
250
200
150
100
50
2018
2019
2020
2021
2022
2023
Date

DISCUSSION

The stock forecast appeals to a wide range of people, including brokerage analysts and novice investors. The practice of predicting has been carried out using a variety of instruments and methodologies. The goal of our research is to do sentiment analysis among individuals using text mining, and then apply predictive modeling to choose the model that will most accurately forecast the future price of FAANG. We believe that anyone interested in stock prediction will find our effort to be instructive.

Although market sentiment is important for stock analysis, it cannot reliably forecast the future because most investors make poor choices in the equity market. Only a small percentage of people are capable of making wise choices and betting the market. It is common knowledge that, statistically, over time, 80% of investors lose money, 10% achieve break-even, and 10% continually gain money. Since the minority is always right, market sentiment is not particularly dependable because of this.

CONCLUSION AND FUTURE WORK

The research uses neural networks for data mining to estimate stock prices. For business experts, determining stock market projections has always been a difficult task. We attempted to forecast stock market indexes using a vast amount of textual data. This information was discovered on the official Yahoo Finance page, which offers historical data for businesses and organizations. Our algorithms successfully simulate seasonality in closing prices. It is only normal for us to lose faith in our principles as we project deeper into the future. The confidence intervals produced by our model, which expand as we go further into the future, reflect this. We come to the conclusion that the ARIMA model is superior to the other two models because it offers higher returns and more profitable trades per trade.

For our upcoming study, we want to use various Techniques and compare the outcomes. For advanced predictions, we will use news feeds and expert commentary from websites as datasets.

APPENDIX FOR LINK TO GitHub REPOSITORY

https://github.com/anvesh-lp/BruteForce.git

REFERENCES

1. **Stock market prediction using Hidden Markov Model** | Publisher: IEEE | Poonam Somani; Shreyas Talele; Suraj Sawant |
   https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/7065011/authors#authors
2. **ANALYSIS AND RISK ASSESSMENT OF FAAMG STOCKS** | Aryan Kasera, Student, The Doon School.|
   https://www.jetir.org/papers/JETIR2007413.pdf
3. **Forecasting FAANG Stocks using Hidden Markov Model** | Publisher: IEEE |Author Name: Aishwarya Jadhav, Jui Kale, Chinmayi Rane, Ankit Datta, Amol Deshpande, Dayanand Ambawade |
   https://ieeexplore-ieee- org.unhproxy01.newhaven.edu/stamp/stamp.jsp?tp=&arnumber=9418216
4 **Stock market prediction using different neural network classification architectures** | Publisher: IEEE | Author Name: K. Schierholt; C.H. Dagli |
   https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/501826
5 **Stock market prediction using Hidden Markov Models** | Publisher: IEEE | Author Name: Aditya Gupta, Bhuwan Dhingra
   https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/6199099