A project report

on

STOCK PREDICTION USING THE MACHINE LEARNING AND TIME SERIES FORECASTING

Submitted By:

Pavan Kumar Kanjarla

Manideep Madavaram

Mounica Jakkam

Contents

List of Figures

Chapter 1. Introduction

1.1 Introduction

In recent the years the trend of investing in the stock Market has increased exponentially due to its high range of returns for the invested capital amount when compared with the returns in other investment sectors. Most of the stocks are volatile in nature and we can't predict the future of the stock. The stock market prediction is difficult due to the dynamic and volatile nature of the different stock values in the market. This dynamic and volatile nature of the stock results in the sudden profits and it may also results in the unexpected losses for the customers invested in the different stocks in the market.

Therefore the need of the proper stock forecasting models to forecast the trend of the stock or to predict the future price of the stock is enhanced. Therefore, to predict the trends in the stock market automatically, many Artificial Intelligence (AI) techniques have been investigated [2] [3] [4]. Some of the first research in prediction of stock prices dates back to 1994, in which a comparative study [5] with machine learning regression models was performed. Since then, many researchers were investing resources to devise strategies for forecasting the price of the stock. Through the use of these advanced technologies the two main approaches had been made in the field of stock prediction.

The first approach is completely focussed on the long term stock forecasting for the long term investments using the fundamental analysis. The other approach is the technical analysis of the stock using the advanced technologies to know the importance of different factors affecting the stock value in the market and predicting the future value of the stock by considering the variation in all the factors affecting the stock market. The main disadvantage in this technical approach is that we cannot predict the value of the stock for the long time periods. Through the use of advanced technologies like Artificial Intelligence and Machine learning the stock value of a company can be forecasted.

In this project the stock value of the AAPL Company is forecasted and the future value of the stock is predicted. This project mainly focusses on the analysing the stock value through the use of different machine learning algorithms and the deep learning neural networks and also to find out the best algorithms for the stock prediction. This project also focusses on extracting a best model for the stock forecasting through the use of these advanced technologies.

Initially in the process of stock prediction different analysis will be performed on the stock trend to know the exact variation in the stock value by comparing with the variation in the other factors that are affecting the stock value.

In generally the stock forecasting can be classified as

- Long Term Forecasting (Beyond 2 years)
- Medium Term forecasting (Between 1 and 2 years)
- Short Term Forecasting (For a period of days, weeks, months)

In this project the future stock value of the AAPL Company is predicted by using the following technical methods and the different machine learning algorithms.

1) Time Series Analysis
2) Regression Analysis
3) Machine Learning

4) Neural Networks (Deep Learning)

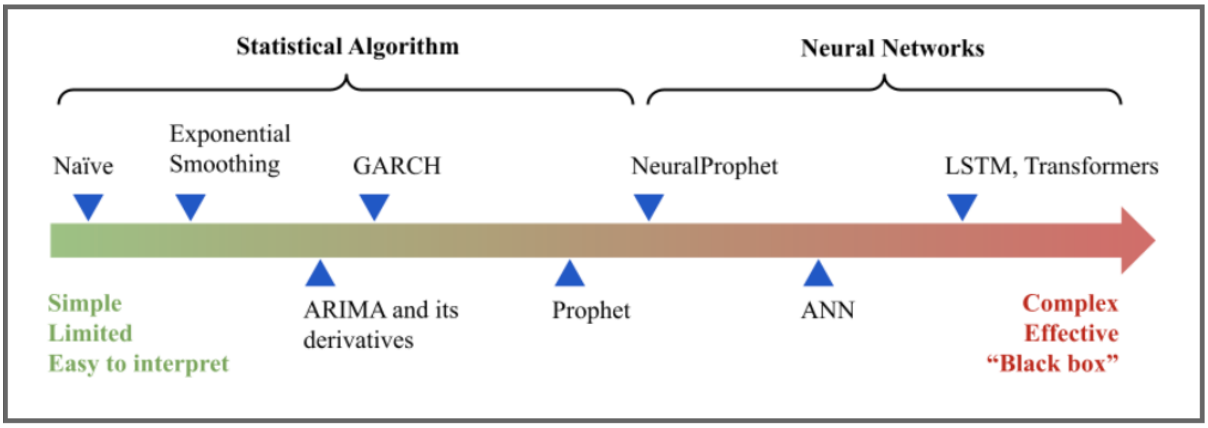

Fig 1.1 Visualization of different stock prediction algorithms

1.2 Time Series Analysis

The time series analysis is a linear classification of the selected variable in proportion with time. In the time series analysis the analysis is performed on the stock trend with reference to the time. The analysis is performed on the past historical stock data of the company to forecast the upcoming results of the stock data. In generally the time series analysis can be performed on the univariate and multi-variate data based on the requirements. If stocks of the different companies need to be analysed then it is a multi-variate analysis. In our case the Univariate time series analysis has been performed on the past historical stock data of the company and the future price and trend in the stock is predicted.

In the time series analysis the change in stock value over a certain time period is assessed and compared with the change in other variables that are affecting the stock value of the company. In this manner a basic fundamental relationship is developed between the various factors of the stock market. But this fundamental hypothesis is not accurate for the stock forecasting. Therefore necessary technical relations needs to be developed in order to support the initial fundamental hypothesis for the accurate forecasting of the stock. The technical analysis on the stock can also reject the fundamental hypothesis if it contradicts the results of the technical analysis.

Time series forecasting uses information from the historical values, and associated past trends of the stock in the stock market to predict future activity of the stock. Most often, this relates to trend analysis, cyclical fluctuation analysis, and issues of seasonality. In the time series forecasting of the stock value the result is volatile as it doesn't provide any technical hypothesis to support the forecasted output.

1.3 Regression Analysis

Regression analysis is a statistical tool for investigating the relationship between a dependent or response variable and one or more independent variables [6]. It is also a reliable method in identifying the relation between two or more variables of the data by providing the necessary mathematical equations. These mathematical equations supports the hypothesis of relation proposed by the regression analysis.

The regression analysis on the stock data identifies the relation between the variation of different aspects of the company on the stock value of the company. By performing the regression analysis on the stock data the exact stock value of the company is predicted using the different mathematical equations. The coefficients of different variables in these mathematical equations explains the significance of variation of other business factors that contribute to the variation in the stock value of the company. These estimated coefficients are used to predict the stock value of the company. In this manner the regression analysis is useful in the prediction of the future stock price.

1.4 Machine Learning

Machine learning is a field of computer science that deals with the learning ability of the intelligent systems (machines). Unlike traditional time series methods, these techniques can handle the nonlinear, chaotic, noisy, and complex data of the stock market, leading to more effective predictions (Chen & Hao, 2017). Mathematics plays a key role in dealing with the learning ability of the intelligent systems. The two main important categories in the machine learning are

1) Supervised Learning

2) Unsupervised Learning.

The supervised learning deals with the labelled data whereas unsupervised learning deals with the lumped data. Initially the machine learning models are trained on the past historical data to generate the mathematical equations for the output based ion the learning ability of the system. In the supervised learning the system is trained on the past historical data associated with labels which helps in developing a mathematical equation for the output based on the training data.

In the case of unsupervised learning the machine is trained on the data without the respective labels. Therefore the machine learning algorithm develops an output equation based on the internal structure of the input and analyses the different patterns in the data. In the stock market the variation in the stock value of a company is analysed and the significance of different factors affecting the stock value of the company is studied based on the past historical stock data of the company.

Different machine learning models are deployed and trained on the data for the efficient and effective analysis in the data. The machine learning models are completely data driven models. Therefore necessary precautions must be taken in pre-processing and filtering out the data. Inclusion of irrelevant information in the data effects the model learning ability and results in the inefficient performance if the system.

In generally the machine learning models has two objectives mainly. The first is to classify the data and the other is to make the predictions on the data based on these previous calculations. A machine learning algorithm for stock trading may inform the trader of future potential predictions. In this project different machine learning models have been choosen and these models are trained on the past historical stock data of the company to predict the future stock value of the company. The performance of these models have been analysed by calculating different performance metrics of the model and over all efficient model for the stock prediction is evaluated.

1.5 Deep Learning and Neural Networks

Deep learning is an advancement of the machine learning which mainly focuses on the development and performance of the Neural Networks. The working activity of these neural networks is similar to the working of different interconnected neurons present in the human brain. Deep learning models use a cascade of multi layered non-linear processing units called as neurons, which can perform feature extraction and transformation automatically. The complete network of these neuron is known as the Artificial Neural Network (ANN).

ANN is an interconnected group of nodes which simulate the structure of neurons present in the human brain. These neurons are organized in the form of consecutive layers, where output of the current layer of neurons is passed to the successive layer as the input. If these interconnection of neurons and layers are only in one direction then the network is termed as the Feed Forward Neural Network. In feed forward neural networks the output of each layer is analysed by the next layer and the corresponding output is generated based on the activation functions in each layers of the network.

In neural networks each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network. The number of nodes on the input and output layer of neural network is equal to the number of inputs and outputs of the system. The variation in the number of hidden layers and the activation functions in the network causes the variation in the output of the network. In generally the weights are assigned to the input signals the neural network processes the input information using the activation functions and the optimisers of the network.

Neural networks rely on training data to learn and improve their accuracy over time. However, once these learning algorithms are fine-tuned for accuracy, they are powerful tools in computer science and artificial intelligence, allowing us to classify and cluster data at a high velocity. Tasks in speech recognition or image recognition can take minutes versus hours when compared to the manual identification by human experts. The backpropagation algorithm in the neural networks helps in re-adjustment of weights of the input signals based on the previous output of the system. This backpropagation algorithm helps in fine tuning of the network for the efficient processing of the input information and to give out an efficient output.

Neural networks can be classified into different types, which are used for different purposes. Some of them are

1) Artificial Neural Networks (ANN)
2) Convolutional Neural Networks (CNN)
3) Recurrent Neural Networks (RNN)
4) Graph Neural Networks (GNN)

Chapter 2. Time Series Forecasting of the Stock Data

2.1 Introduction

Time series analysis is an approach to analyze time series data to extract meaningful characteristics of data and generate other useful insights applied in business situation. Generally, time-series data is a sequence of observations stored in time order. Time-series data often stands out when tracking business metrics, monitoring industrial processes and etc. The time series tracks the movement of the data points of the variable over a specified period of time with data points recorded at regular intervals. Time series forecasting is used to predict future values based on previously observed values and one of the best tools for trend analysis and future prediction. The stock market is very unpredictable, and there are many factors effecting the stock market in different ways either directly or indirectly.

In generally there are different time series forecasting models which predicts the trend of the stock market. The different time series forecasting models include

1) Moving Average Model
2) Exponential Smoothing Model
3) Autoregression Model

2.2 ARIMA Model

The Auto Regressive Integrated Moving Average (ARIMA) model is used to analyse and predict the future prices of the stock value based on the previous data. Initially the ARIMA model converts the Non stationary time series into the stationary series. It is one of the most widely used time series forecasting models for the prediction of the linear time series data.

In a multiple regression model, we forecast the variable using a linear combination of predictors. In an autoregression model, we forecast the variable using a linear combination of past values of the variable. The term autoregression indicates that it is a regression of the variable against itself. Autoregressive models are remarkably flexible at handling a wide range of different time series patterns.

Rather than using past values of the variable in a regression, a moving average model uses past forecast errors (residuals) in a regression like model.

An ARIMA model is characterized by 3 terms: p, d, q

Where,

p is the order of the AR term (Auto Regressive)

q is the order of the MA term (Moving Average)

d is the number of differencing required to make the time series stationary.

If the time series for the forecasting is not stationary then the first order difference of the time series need to be performed to make it stationary. A moving average model is used for forecasting future values, while moving average smoothing is used for estimating the trend-cycle of past values.

2.3 Exponential Smoothing Method

Exponential smoothing is a simple and pragmatic approach to forecasting, whereby the forecast is constructed from an exponentially weighted average of past observations. The largest weight is given to the present observation, less weight to the immediately preceding observation, even less weight to the observation before that, and so on (exponential decay of influence of past data).

Exponential smoothing is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. Exponential smoothing forecasting methods are similar in that a prediction is a weighted sum of past observations, but the model explicitly uses an exponentially decreasing weight for past observations.

There are three main types of exponential smoothing time series forecasting methods.

1) Single exponential smoothing
2) Double exponential smoothing
3) Triple exponential smoothing

The simplest of the exponentially smoothing methods is naturally called simple exponential smoothing (SES). This method is suitable for forecasting data with no clear trend or seasonal pattern.

2.4 Holt's Method

Holt's two-parameter model, also known as linear exponential smoothing, is a popular smoothing model for forecasting data with trend. Holt's model has three separate equations that work together to generate a final forecast. The first is a basic smoothing equation that directly adjusts the last smoothed value for last period's trend. The trend itself is updated over time through the second equation, where the trend is expressed as the difference between the last two smoothed values. Finally, the third equation is used to generate the final forecast. Holt's model uses two parameters, one for the overall smoothing and the other for the trend smoothing equation. The method is also called double exponential smoothing or trend-enhanced exponential smoothing.

2.5 TBATS Method

The TBATS refers to the Trigonometric seasonality Box cox transformation Arima Errors Trend Seasonal Components. TBATS is a forecasting method to model time series data. The main aim of this is to forecast time series with complex seasonal patterns using exponential smoothing. In TBATS forecasting different models are developed initially based on the trends and the seasonality patterns in the data. The various transformations needed to be done are also taken into account in this TBATS forecasting method. The final model for the TBATS is selected based on the AIC (Akaike information criterion) value.

Different models are developed in the TBATS forecasting method based on the different factors associated with the data, as shown below

- With Box-Cox transformation and without it Box-Cox transformation.
- With considering Trend and without Trend.
- With Trend Damping and without Trend Damping.
- With ARIMA (p,q) and without ARMA (p,q) process used to model residuals.

- Non-seasonal model.
- various amounts of harmonics used to model seasonal effects

The auto ARIMA model in the TBATS forecasting configure the modelling of the residuals based on the P and Q values of the ARIMA model.

In this project using all the above time series forecasting models the stock data of the AAPL company for the past five years is analysed and the present year 2022 stock value is forecasted. This forecasted value is compared with the true value of the stock and the final performance metrics of the model were calculated. Therefore the training data consists of the stock value of the company for the past five years and the testing data consists of the stock values of the present year to test the efficacy of the model.

Chapter 3. Regression Based Prediction of the Data

3.1 Introduction

Regression analysis provides a single mathematical relationship for two or more independent variables in the stock data for technical and quantitative analysis in financial markets. This mathematical regression equation gives out the significance of variation in different variables affecting the stock value of the company.

In generally the variables in the past historical stock data of a company are related linearly and the model trained on this data will be biased for this linear relation. Therefore the summary statistics of these model reflects the highest R squared value and the highest adjusted R squared value. The assumptions of the linear regression model are completely biased. This in-turn causes the significant statistical error in the regression model.

3.2 Prophet Model

Prophet is a procedure for forecasting time series data based additive regression model where non-linear trends are fit with yearly, weekly and daily seasonality. Prophet also includes the effect of missing values (Holiday effects) in the data. Prophet is robust to missing data and shifts in trends and typically handles outliers well.

In recent rimes this Prophet package is released by the Facebook and this version of the model is completely automatics in forecasting the trends and future values of the seasonality including the effects of missing values in the data.

The mathematical representation of the prophet model is given by

$$Y(t) = g(t) + s(t) + h(t) + e_t$$

Here all the components of the above equation are with reference to the time (t)

g(t) represents the trend of the pattern (linear or logistic)

s(t) represents the seasonality of the pattern

h(t) represents the effects of holidays (missing values of the stock)

e(t) represents the error.

In generally Prophet is framing the forecasting problem as a curve-fitting exercise rather than looking explicitly at the time-based dependence of each observation within a time series.

3.3 Assumptions of the Regression Model

Multiple regression is a statistical model to understand the relation between the multiple predictor variables and the final response variable. In this regression model to develop a relationship between the prediction variable and the predictor variable few assumptions will be made. These assumptions are a vital part of assessing the model.

In the regression analysis to explain the significance of the predictor variables the model should not violate this regression assumptions. The various assumptions in the linear regression model include:

1. There is a linear relationship between the predictor variables and the prediction (outcome) variable

2. The predictor variables are independent.
3. Residuals follows the normal distribution.
4. There exists a constant variance among the residuals (Equal Variance
5. Residual errors are independent of each other.

3.4 Verification of Assumptions of the Regression Model

To verify the assumptions of the regression model different characteristics of the model are plotted and analysed.

- The scatter plot of the predictor variable and the predictor variable gives the linearity relationship among the variables.
- The plot between the square root of the standardized residuals and the fitted values explains the constant variance assumption of the regression model.
- Durbin Watson test is conducted on the model to explain the hypothesis of the relation among the predictor variables in the regression model.
- Histogram of the residuals is plotted to verify the normality assumption of the residuals.

3.4.1 Verification of assumptions of the Prophet model.

i) Residual Plot:

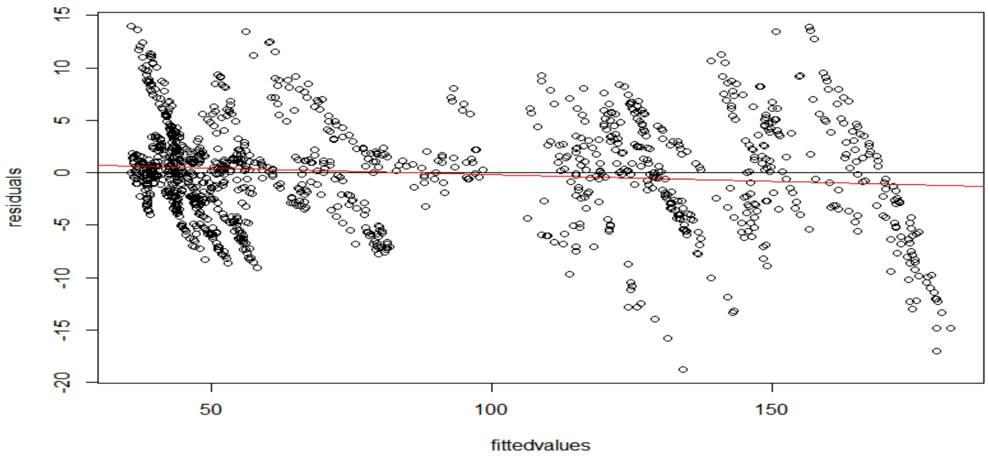

Fig 3.1 Residues vs Fitted values plot of the Prophet model

From the above plot of residuals and fitted values for the prophet model the mean of the residuals is almost equal to zero and the linearity assumption of the regression analysis is verified.

ii) Normal plot:

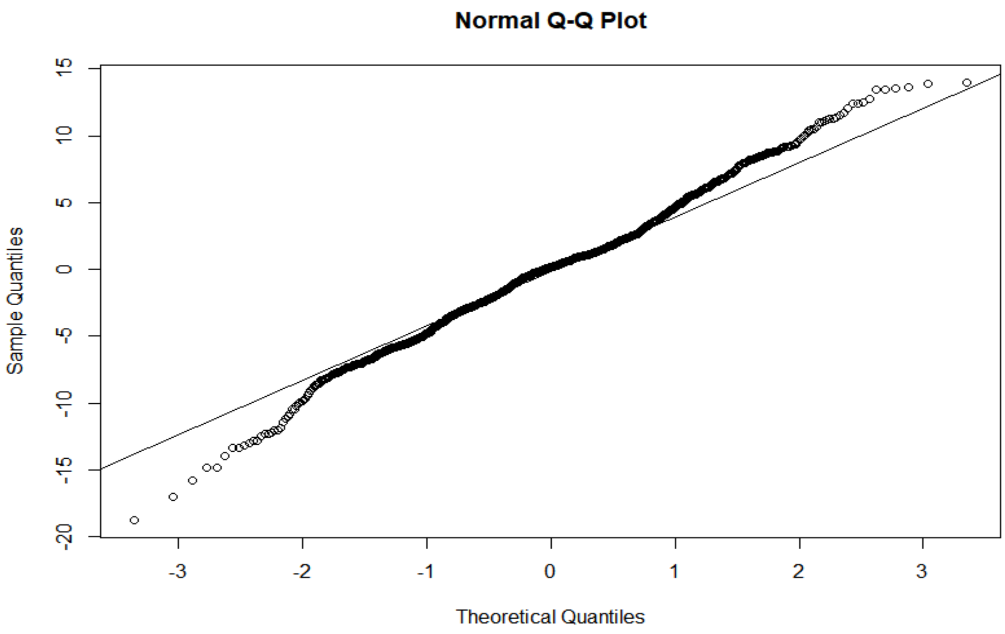

Fig 3.2 Normal Q-Q plot of the Prophet model

From the above normal plot the errors the variance of the errors is approximately equal and the regression assumption of Homoscedasity (Equal Variance) is verified.

iii) Histogram of the Residuals:

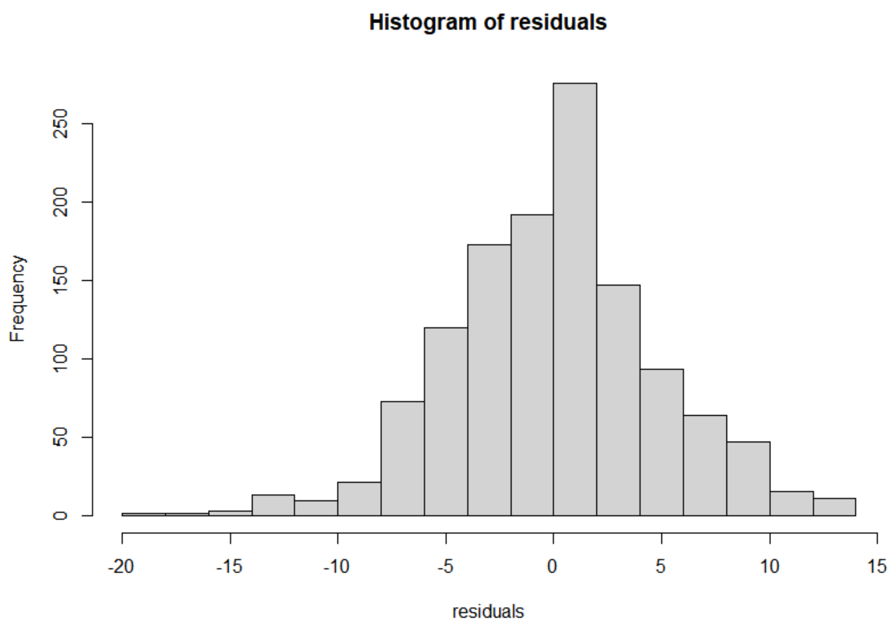

Fig 3.3 Histogram of the residuals of the Prophet model

From the above histogram the errors are normally distributed. Therefore we can conclude that all the assumptions of regression model are verified.

3.5 Stock Prediction using the Prophet Model

In this project an autoregressive model (PROPHET) is implemented for the regression analysis of the stock data of the AAPL Company. The stock value of a company is forecasted for the next on year (365 days) and time series visualization is provided to observe the predicted trend in the stock data. The performance metrics of the model is calculated.

Chapter 4. Stock Prediction using the Machine Learning

4.1 Introduction

4.2 Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes. A decision node has two or more branches each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node.

Decision tree is a supervised learning algorithm, and it works for both the categorical and continuous variables. Decision tree is a hierarchical model of supervised learning where the decision is made in a sequence of recursive splits in a smaller number of steps. As a supervised machine learning model, a decision tree learns to map data to outputs in the training phase of model building.

During training, the model is fitted with any historical data that is relevant to the problem domain and the true value we want the model to learn to predict. The model learns any relationships between the data and the target variable. After the training phase, the decision tree produces a tree, calculating the best questions and the order of questions to ask in order to make the most accurate estimates possible. In the process of testing the model the same data format should be provided to the model in order to make a prediction. The prediction will be an estimate based on the train data.

The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria is different for classification and regression trees. Decision trees regression normally use mean squared error (MSE) to decide to split a node in two or more sub-nodes. For each subset, it will calculate the MSE separately. The tree chooses the value with results in smallest MSE value.

In this project a decision tree regressor is implemented to predict the stock value of the company and to evaluate the performance metrics of the model to evaluate the best regression model for stock prediction based on the performance metrics of the model.

4.3 Random Forest

The random forest is a model made up of many decision trees. The random forest model uses two key concepts that gives it the name random:

1) Random sampling of training data points when building trees.
2) Random subsets of features considered when splitting nodes

The generalization error for forests converges as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Random forest comes under the category of supervised learning and is widely used as the classification and the regression model. The most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Hyper-parameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster. Random forest is one of the most accurate general-purpose learning techniques available in the machine learning.

In this project the random forest regression model is implemented and trained on the past historical stock data of the AAPL Company. Then the stock values for the present running year 2022 is predicted. The performance of the model is evaluated by comparing the predicted stock values with the actual stock values. The performance metrics for the model is calculated.

Chapter 5. Stock Prediction using the Deep Learning

5.1 Introduction

Deep learning is part of a broader family of ML methods based on learning data representations, as opposed to task specific algorithms. Deep learning models use a cascade of multi layered non-linear processing units called as neurons, which can perform feature extraction and transformation automatically. The network of such neurons is called an Artificial Neural Network.

Deep learning mainly deals with the learning and the performance of the neural networks. Deep learning attempts to mimic the human brain far from matching its ability enabling systems to cluster data and make predictions with incredible accuracy. Deep learning drives many artificial intelligence (AI) applications and services that improve automation, performing analytical and physical tasks without human intervention. Deep learning technology lies behind everyday products and services as well as emerging technologies (such as self-driving cars).

Deep learning algorithms eliminates some of data pre-processing that is typically involved with machine learning. These algorithms can ingest and process unstructured data, like text and images, and it automates feature extraction, removing some of the dependency on human experts. Deep learning algorithms determine the significance of different features that are most important to distinguish between then other. In machine learning, this hierarchy of features is established manually by a human expert.

Then, through the processes of gradient descent and backpropagation, the deep learning algorithm adjusts and fits itself for accuracy, allowing it to make predictions about a new photo of an animal with increased precision.

Deep neural networks consist of multiple layers of interconnected nodes, each building upon the previous layer to refine and optimize the prediction or categorization. This progression of computations through the network is called forward propagation. The input and output layers of a deep neural network are called visible layers. The input layer is where the deep learning model ingests the data for processing, and the output layer is where the final prediction or classification is made.

5.2 Artificial Neural Networks

Artificial Neural networks (ANN) or neural networks are computational algorithms. Artificial Neural Networks are intended to simulate the behavior of biological systems composed of “neurons”. ANNs are computational models inspired by an animal’s central nervous systems. It is capable of machine learning as well as pattern recognition. These presented as systems of interconnected “neurons” which can compute values from inputs.

A neural network is an oriented graph. It consists of nodes which in the biological analogy represent neurons, connected by arcs. It corresponds to dendrites and synapses. Each arc associated with a weight at each node. An Artificial Neural Network is an information processing technique. It works like the way human brain processes information. ANN includes many connected processing units that work together to process information. They also generate meaningful results from it.

Artificial Neural network is typically organized in layers. Layers are being made up of many interconnected 'nodes' which contain an 'activation function'. A neural network may contain the following 3 layers

- Input Layer
- Hidden Layer
- Output Layer

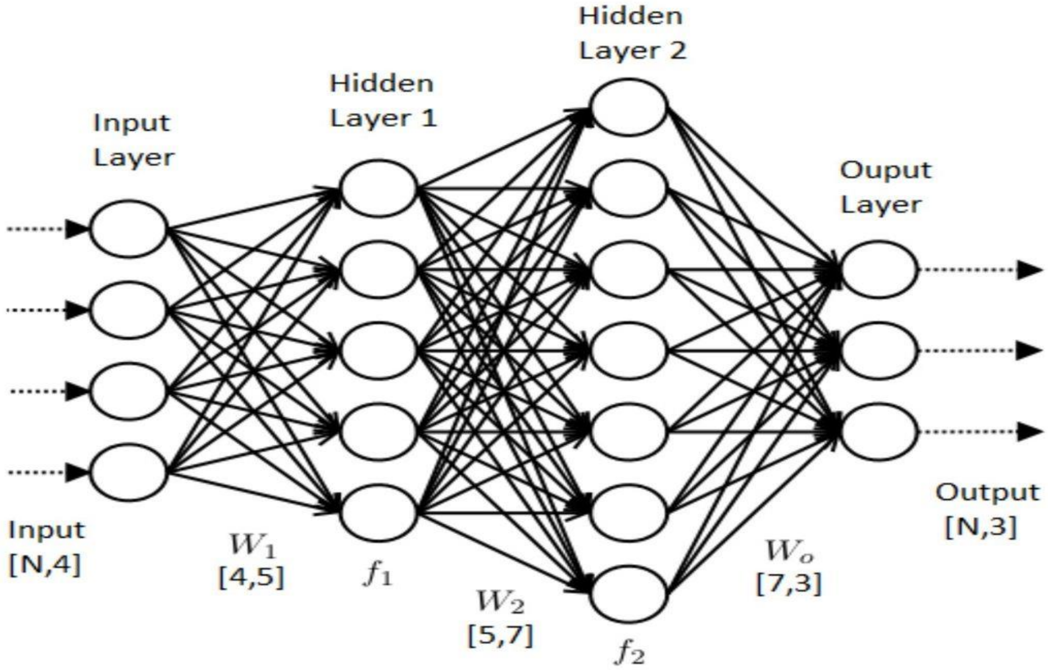

Fig 5.1 Representation of the Artificial Neural Network (ANN)

The purpose of the input layer is to receive as input the values of the explanatory attributes for each observation. Usually, the number of input nodes in an input layer is equal to the number of explanatory variables. 'Input layer' presents the patterns to the network, which communicates to one or more 'hidden layers'.

The hidden layers process the output of the input layer in the network with reference to the weights assigned to the input signal. The input signal is processed based on the activation function and the optimizers in the hidden layers. These hidden layers processes the input information and gives out the outpour to the output layers of the network.

The main function of the output layer is to give out the output of the neural network. The number of nodes in the output layers is equal to the number of output (Prediction) variable.

5.3 Recurrent Neural Networks

Recurrent neural networks (RNNs) are identified by their feedback loops. These learning algorithms are primarily leveraged when using time-series data to make predictions about future outcomes, such as stock market predictions or sales forecasting. RNNs are a powerful and robust type of neural network, and belong to the most promising algorithms in use because it is the only one with an internal memory.

Because of their internal memory, RNN's can remember important things about the input they received, which allows them to be very precise in predicting what's coming next. This is why they're the preferred algorithm for sequential data like time series, speech, text, financial data, audio, video, weather and much more. Recurrent neural networks can form a much deeper understanding of a sequence and its context compared to other algorithms.

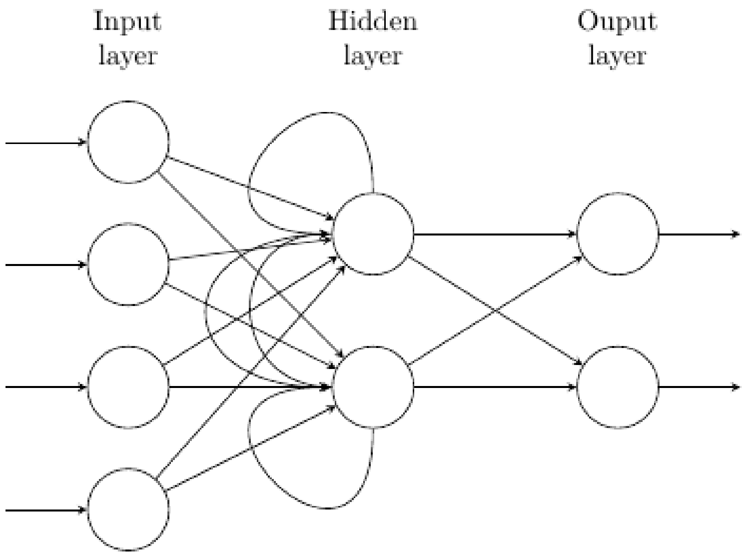

Fig 5.2 Representation of the Recurrent Neural Network (RNN)

In a RNN the information cycles through a loop. When it makes a decision, it considers the current input and also what it has learned from the inputs it received previously.

A usual RNN has a short-term memory. In combination with a LSTM they also have a long-term memory.

5.4 Long Short term Forecasting Memory

Long short-term memory networks (LSTMs) are an extension for recurrent neural networks, which basically extends the memory. Therefore it is well suited to learn from important experiences that have very long time lags in between. The units of an LSTM are used as building units for the layers of a RNN, often called an LSTM network.

LSTMs enable RNNs to remember inputs over a long period of time. This is because LSTMs contain information in a memory, much like the memory of a computer. The LSTM can read, write and delete information from its memory.

This memory can be seen as a gated cell, with gated meaning the cell decides whether or not to store or delete information (i.e., if it opens the gates or not), based on the importance it assigns to the information. The assigning of importance happens through weights, which are also learned by the algorithm.

In an LSTM you have three gates: input, forget and output gate. These gates determine whether or not to let new input in (input gate), delete the information because it isn't important (forget gate), or let it impact the output at the current timestep (output gate). Below is an illustration of a RNN with its three gates:

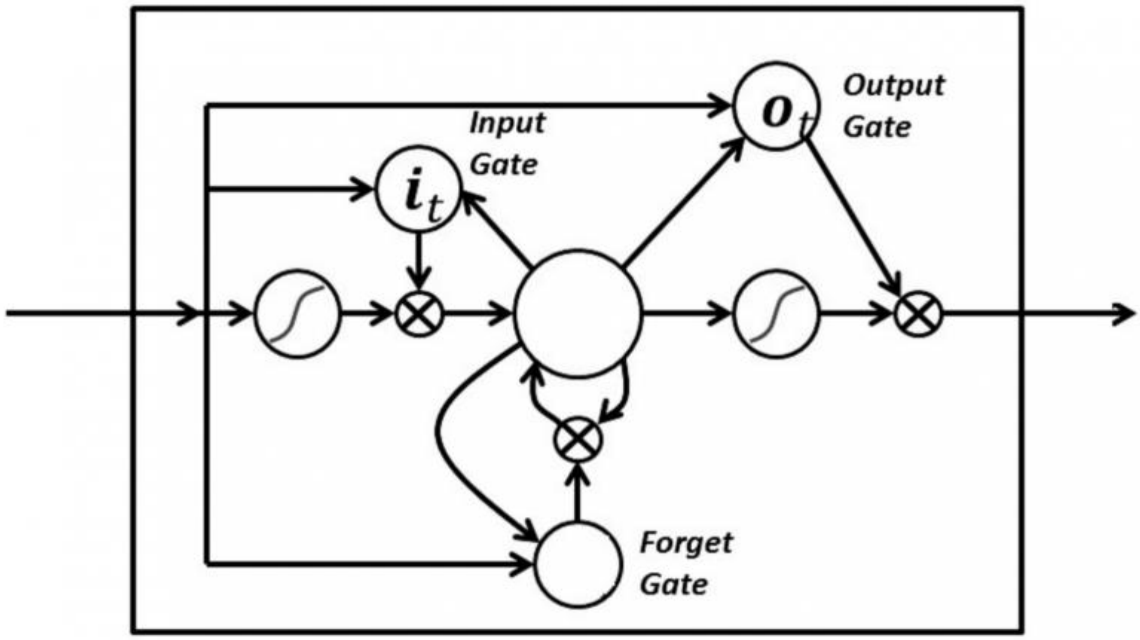

Fig 5.3 Representation of the Long Short Term Forecasting Model (LSTM)

The gates in an LSTM are analog in the form of sigmoids, meaning they range from zero to one. The fact that they are analog enables them to do backpropagation.

The problematic issues of vanishing gradients is solved through LSTM because it keeps the gradients steep enough, which keeps the training relatively short and the accuracy high.

Chapter 6. Results and Conclusion

6.1 Results

6.1.1 Stock Prediction Using the Prophet model:

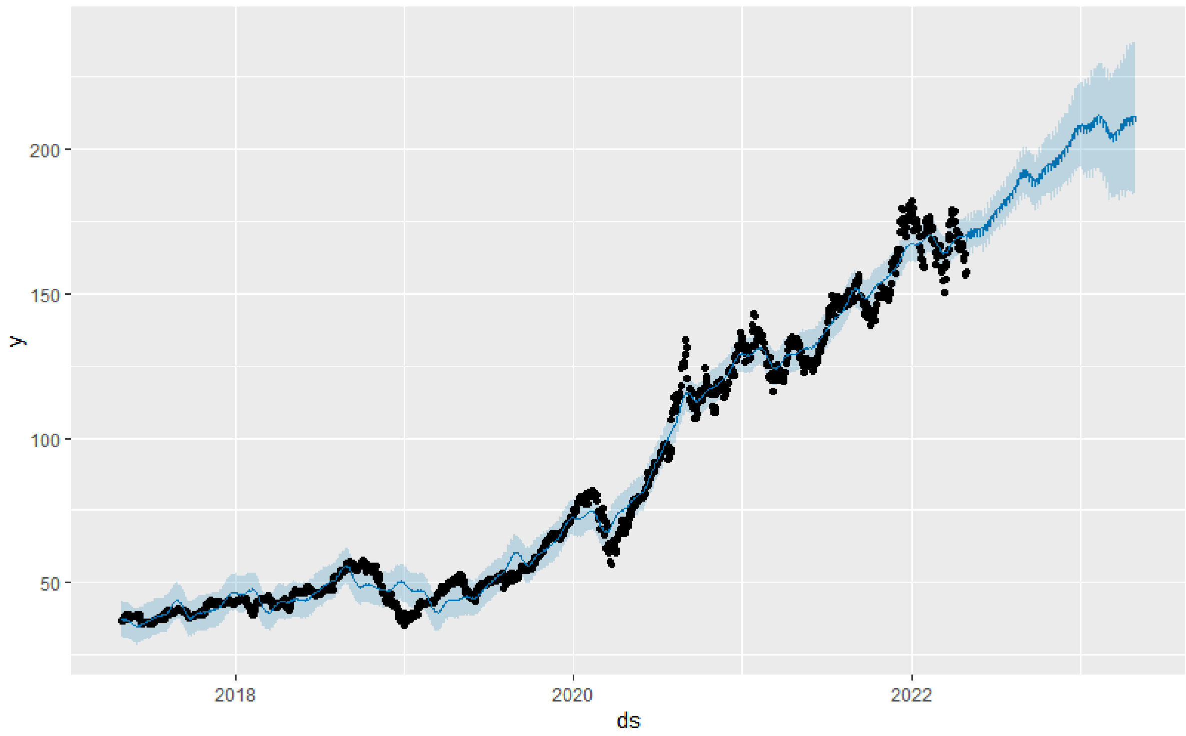

Fig 6.1 Forecasted time series plot of the stock value of the company

The above time series plot refers to the forecasted value of the stock using the Prophet model. The blue region around the prediction line represents the confidence interval of the forecasted stock value of the company.

6.1.2 Stock prediction using the ARIMA model:

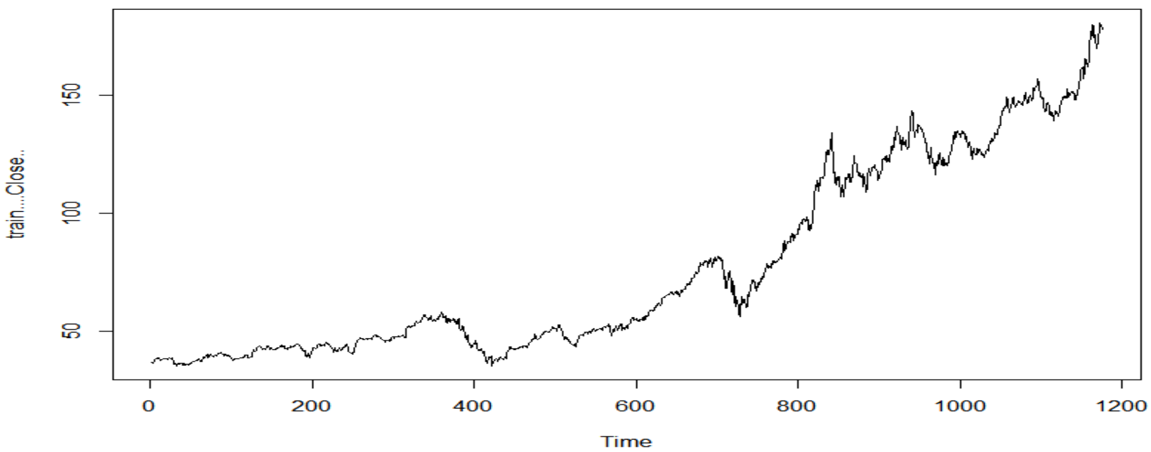

Fig 6.2 Time series plot of the training data of the ARIMA model

The above time series plot represents the training data of the ARIMA model. The above time series is not stationary. In order to implement an ARIM model the input time series must be necessarily stationary. Therefore we need to convert the above time series into stationary time series, the first order logarithmic difference of the data is calculated to convert the data into the stationary time series.

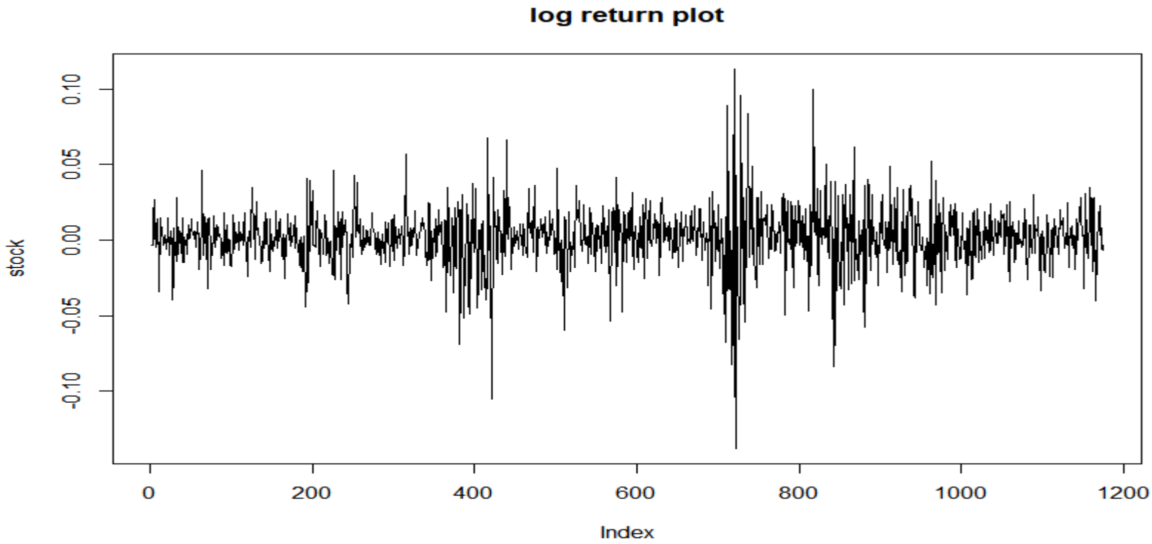

Fig 6.3 Stationary time series plot of the first order logarithmic difference data

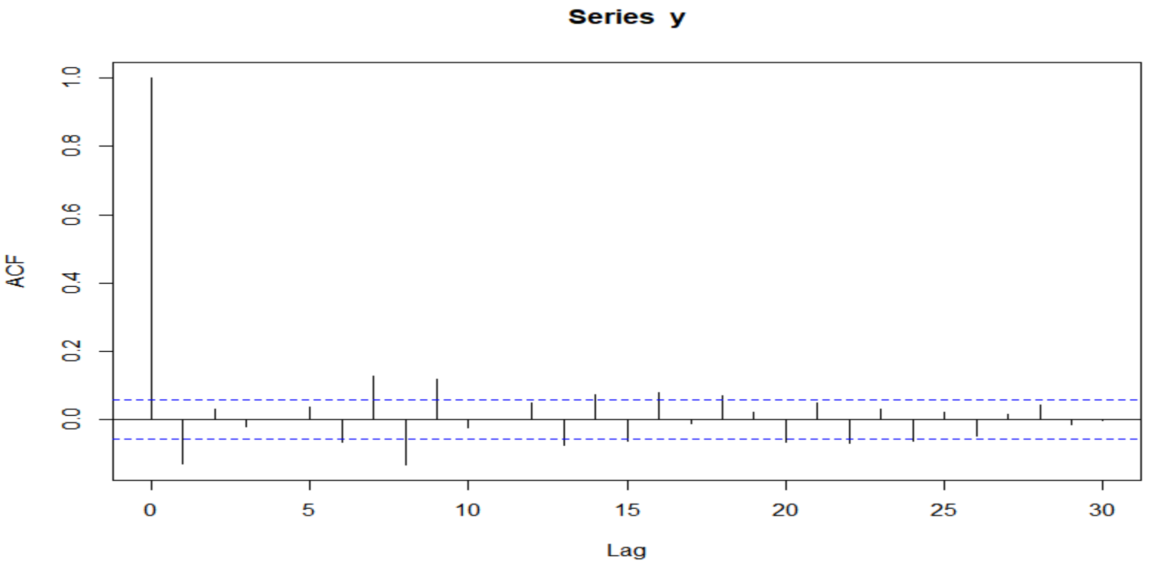

Fig 6.4 ACF plot of the stationary time series data.

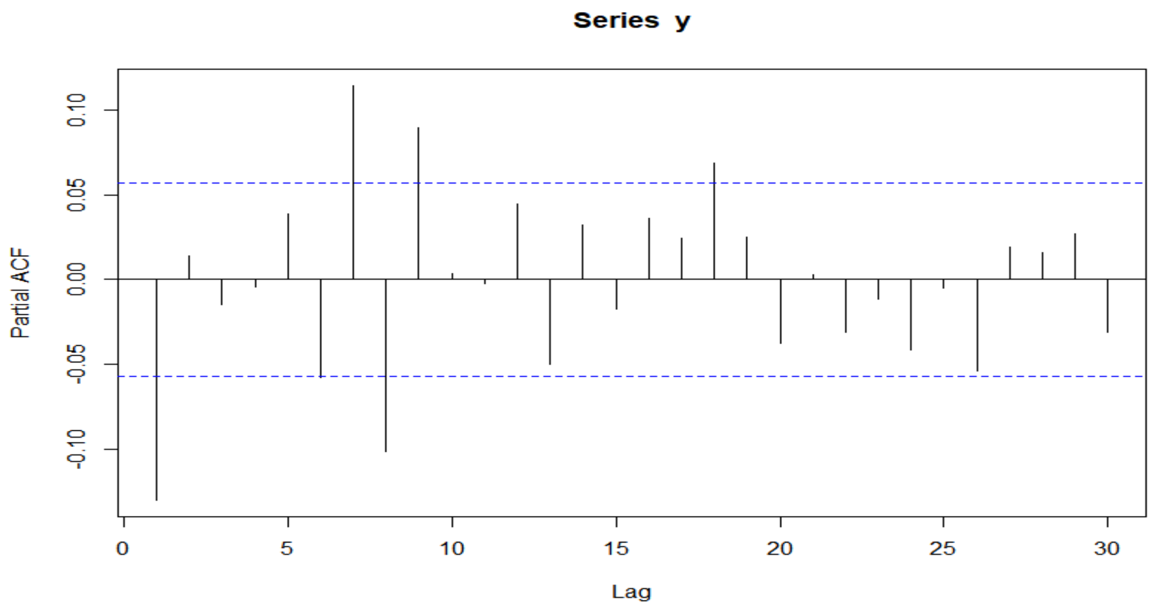

Fig 6.5 PACF plot of the stationary time series data

From the above ACF and the PACF plots we can conclude that the time series is stationary and we can apply the ARIMA model on the data to forecast the stock data.

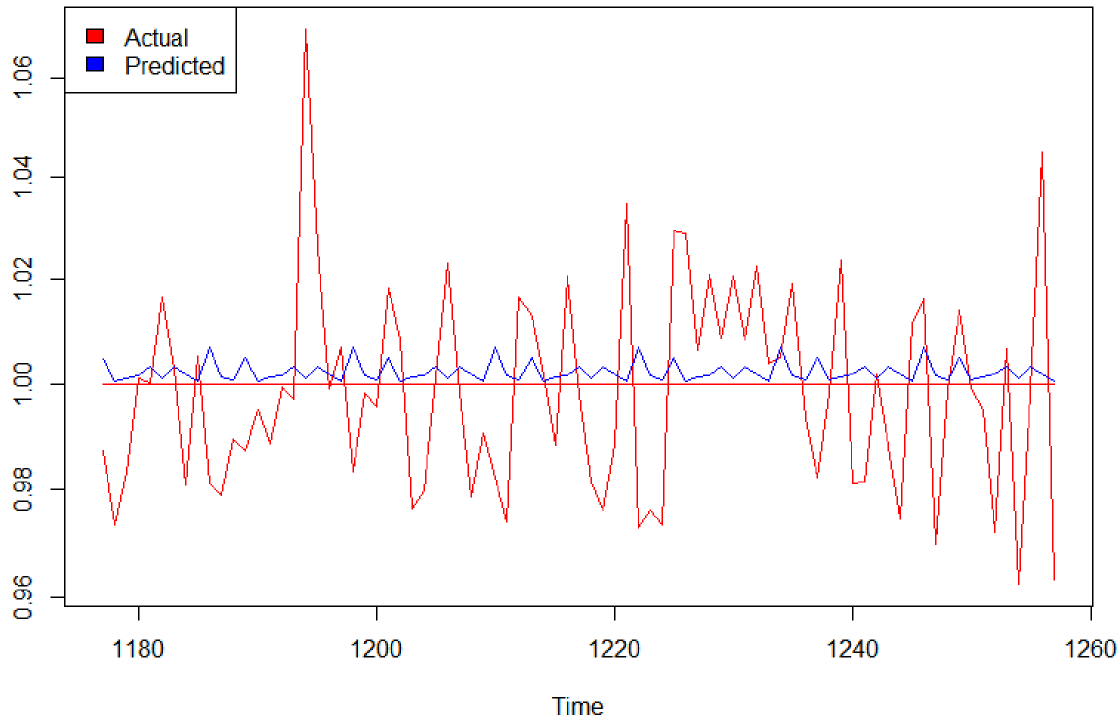

Fig 6.6 Time series plot of the predicted value of the stock using the ARIMA model

The above plot represents the time series plot of the predicted and the true value of the stock forecasted by using the ARIMA model.

6.1.3 Stock prediction using the Exponential Smoothing model:

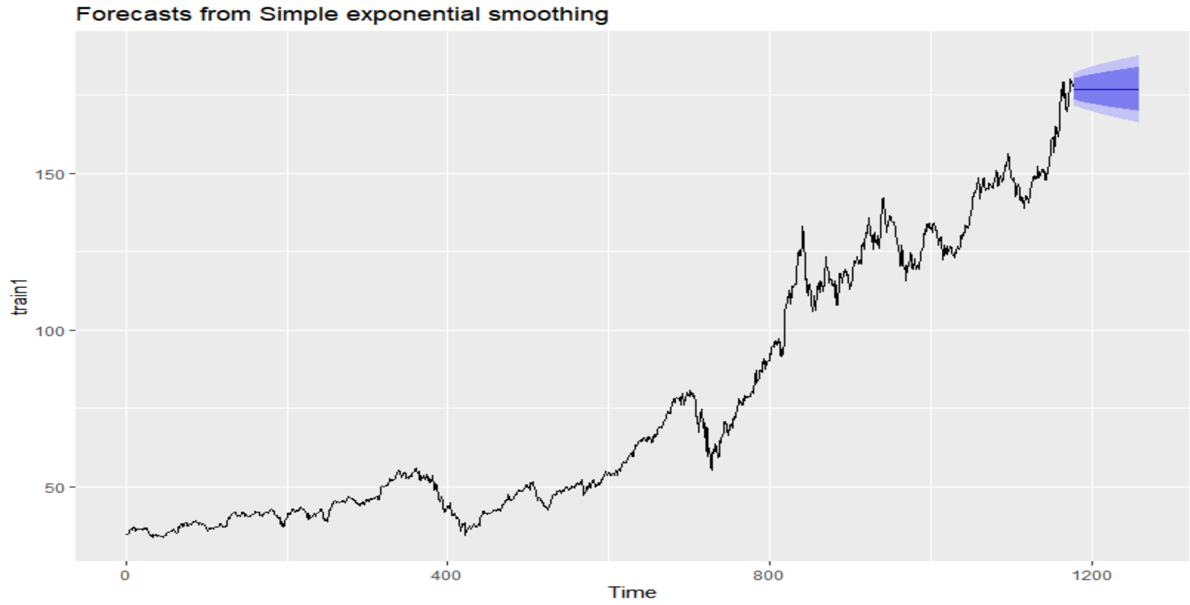

Fig 6.7 Time series plot of the forecasted stock data using the Exponential Smoothing model

The above time series plot represents the forecasted stock value of the company. The dense blue region around the stock trend represents the interval of 95% confidence interval and the light blue region represents the interval of 80% confidence interval for the prediction of stock data.

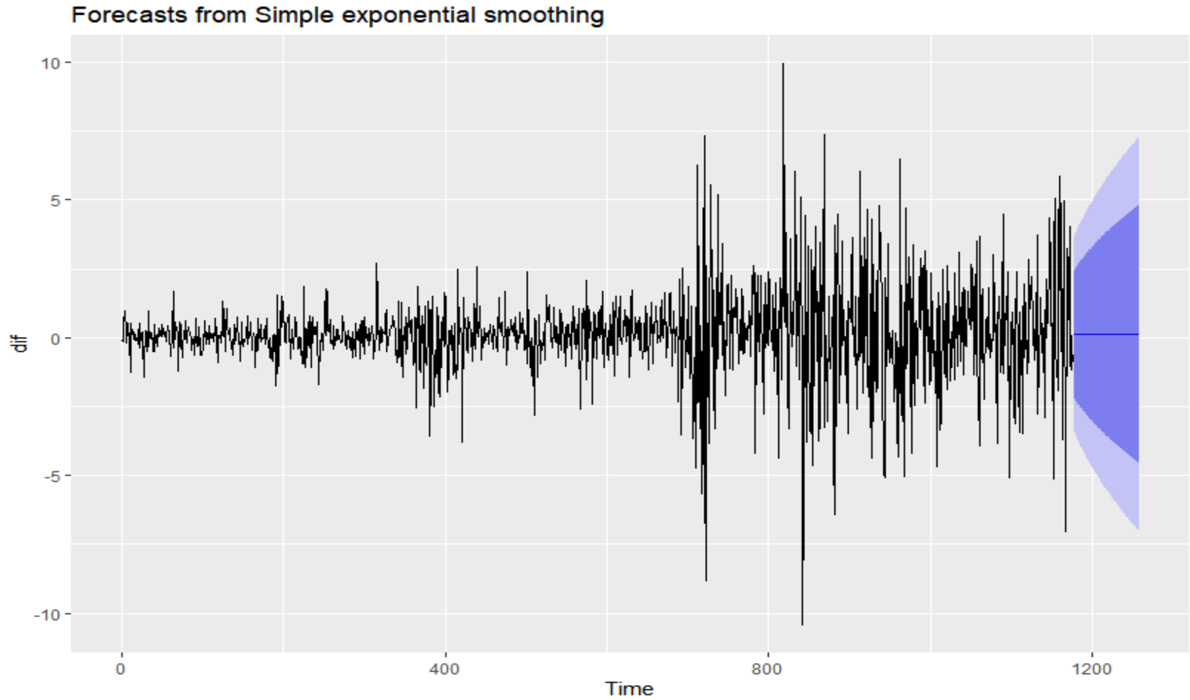

Fig 6.8 Time series plot of the first order difference value of the forecasted stock data using the exponential smoothing model

The above time series plot represents the first order difference value forecasted stock of the company. The dense blue region around the stock trend represents the interval of 95%

confidence interval and the light blue region represents the interval of 80% confidence interval for the prediction of stock data.

6.1.4 Stock prediction using the Holt's model:

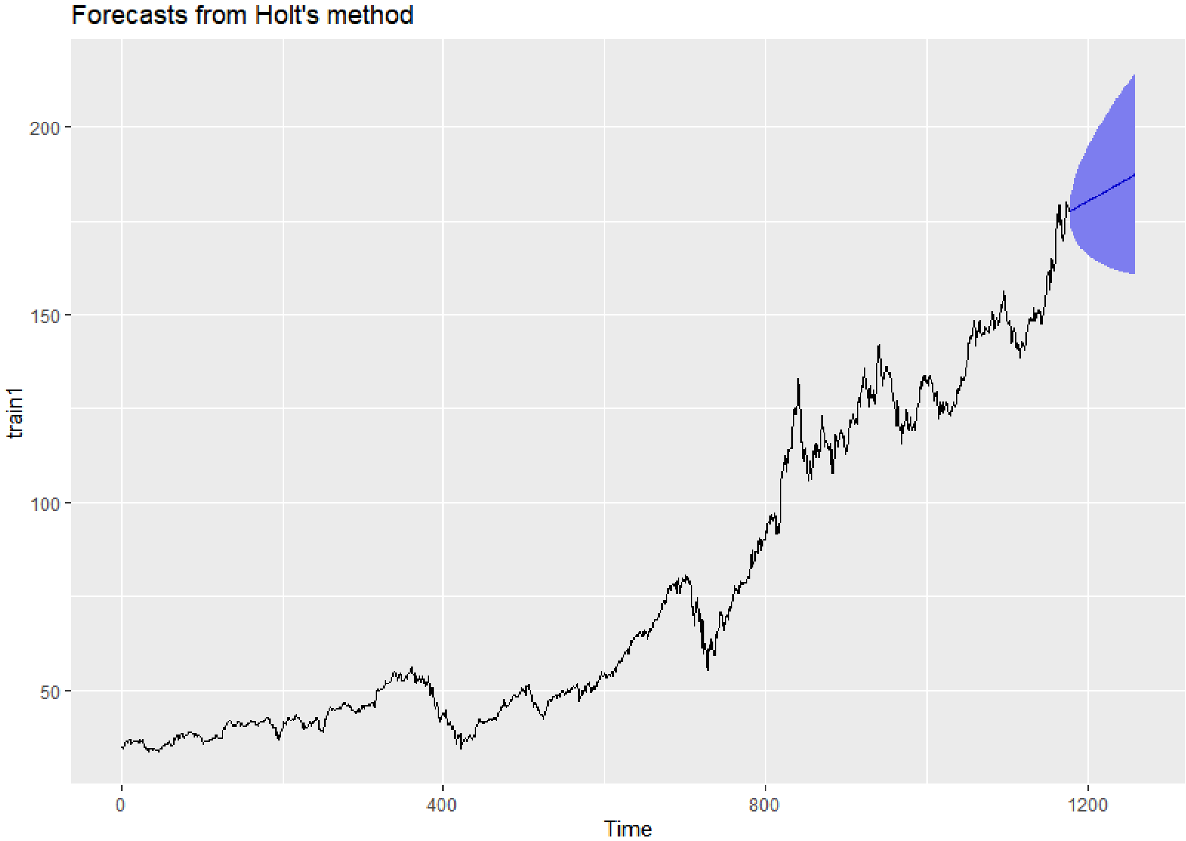

Fig 6.9 Time series plot of the forecasted stock data using the Holt's trend model

The above time series plot represents the forecasted stock value of the company using the holt's trend method. The dense blue region around the stock trend represents the interval of 95% confidence interval and the light blue region represents the interval of 80% confidence interval for the prediction of stock data.

6.1.4 Stock prediction using the Random Forest

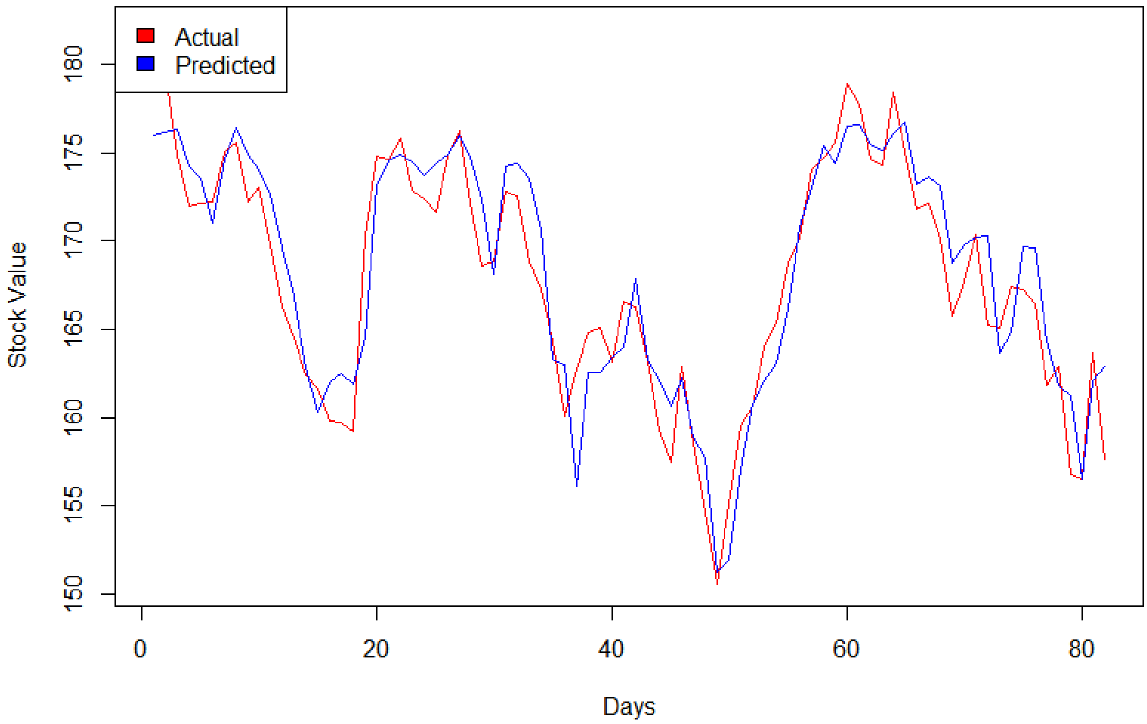

Fig 6.10 Time series plot of the forecasted stock data using the Random Forest model

In the above plot the time series visualization of the true value of the stock and the predicted value of the stock is provided. The prediction of the random forest algorithm is approximately accurate in the stock prediction.

6.1.5 Stock prediction using the Decision tree

6.1.5.1 Loose one out Cross Validation (LOOCV) on Decision tree

```
CART

1177 samples
   4 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 1176, 1176, 1176, 1176, 1176, 1176, ...
Resampling results across tuning parameters:

  cp          RMSE      Rsquared    MAE
  0.04219079  11.79525  0.91527048   8.674269
  0.05340969  15.08423  0.86361094  11.884093
  0.86424994  38.98045  0.09935914  35.766687

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cp = 0.04219079.
```

Fig 6.11 Cross validation metrics of the Decision Tree model

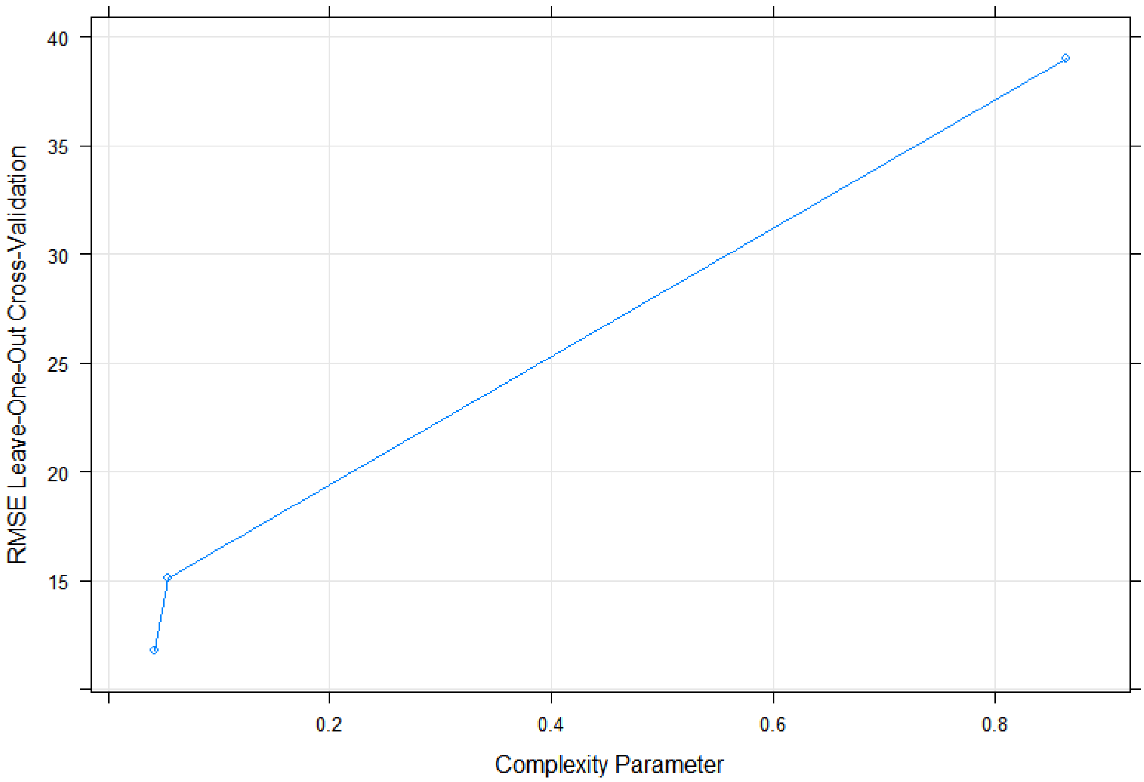

Fig 6.12 RMSE vs complexity parameter plot of the cross validated decision tree model

6.1.6 Stock Prediction using the Artificial Neural Network

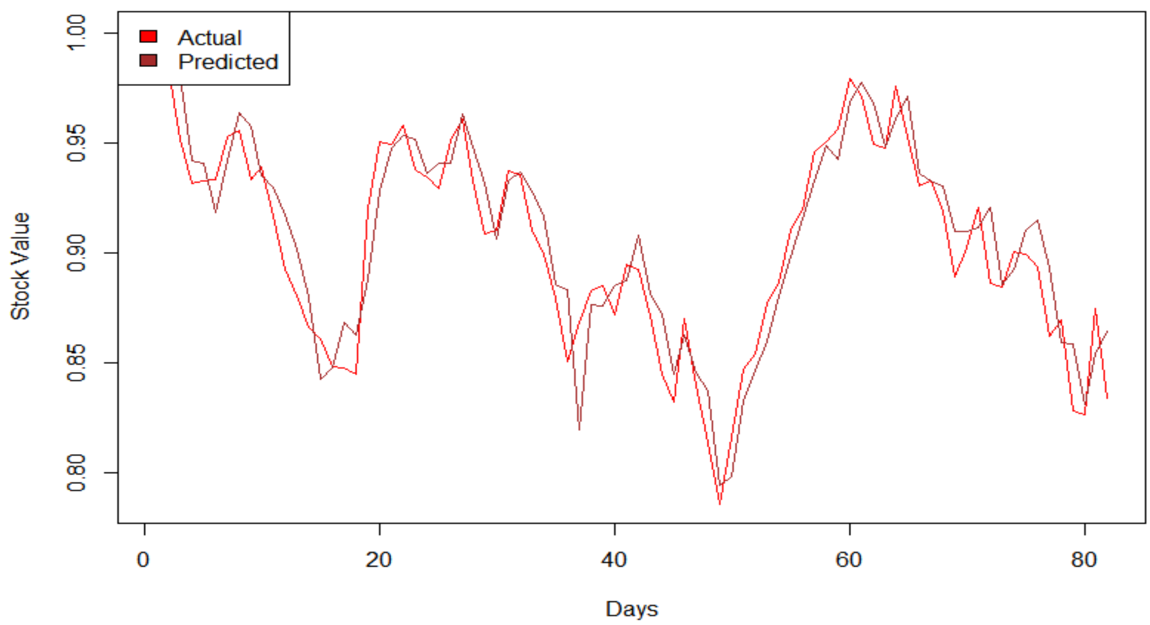

Fig 6.13 Time series plot of the Actual and predicted stock value using the Neural Networks

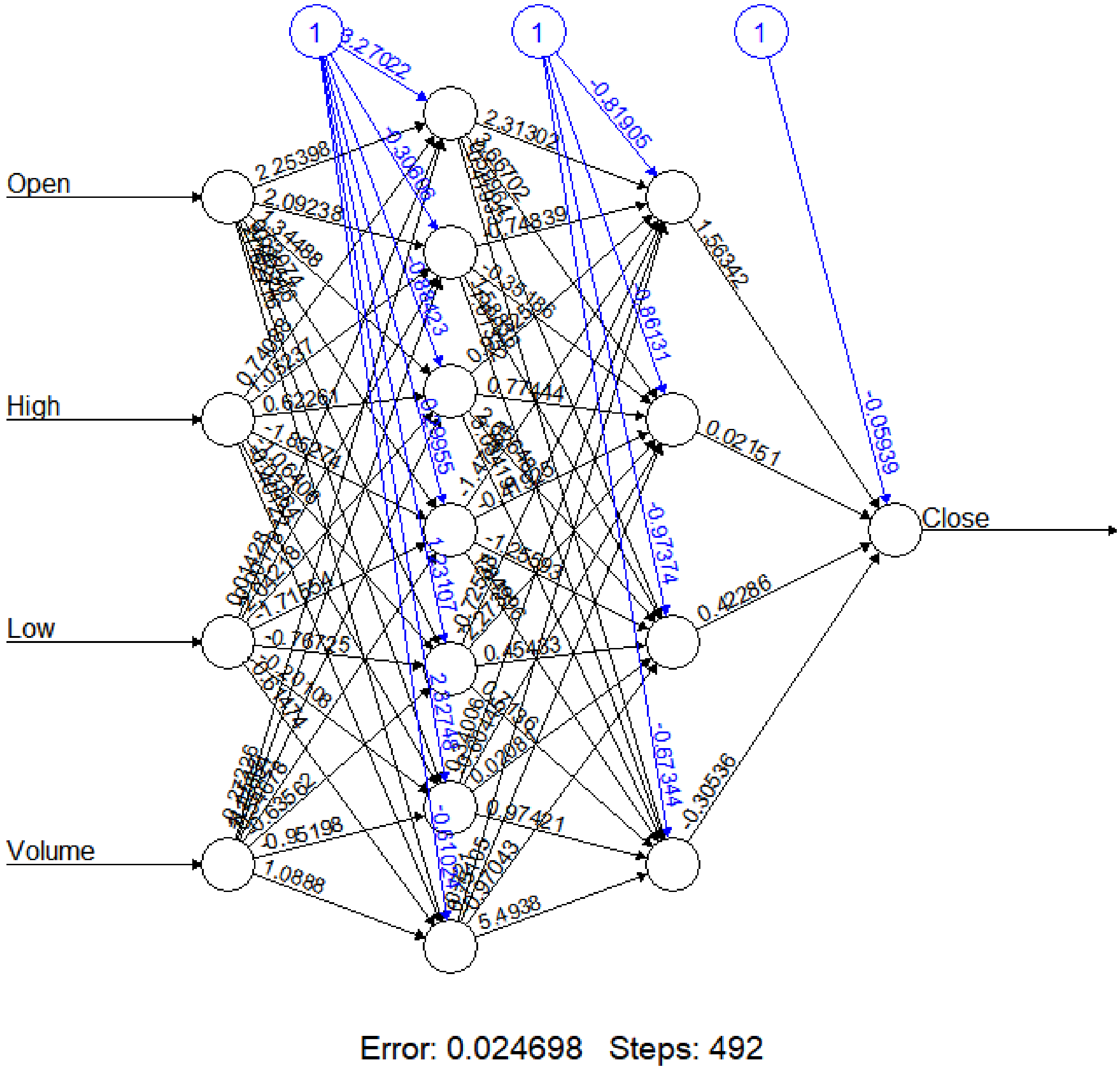

Fig 6.14 Developed Artificial Neural Network for the stock prediction

6.1.7 Stock Prediction using the LSTM

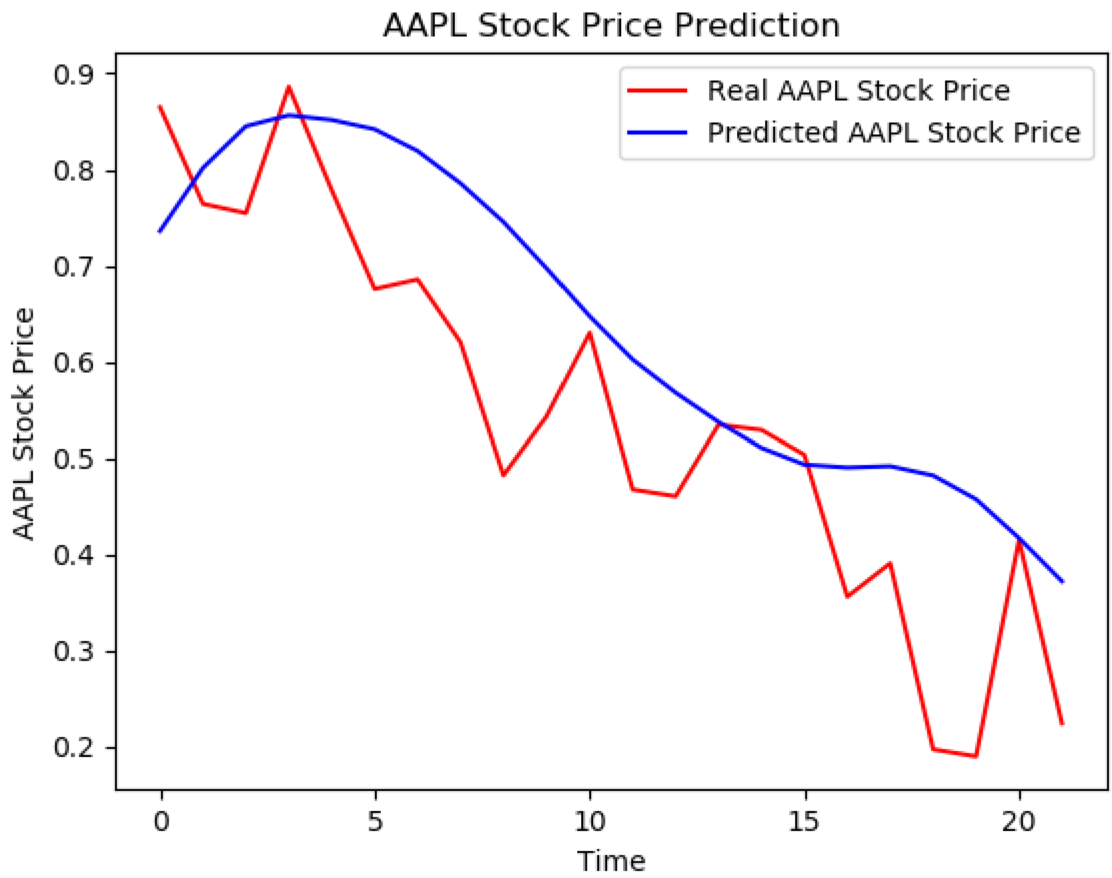

Fig 6.15 Time series plot of the predicted and true stock price using the LSTM model

6.2 Conclusion

Model	Prediction Class	Variables Involved	MAPE	Accuracy
Linear Regression	Regression	Open, Low, High	0.0786	79.21
ARIMA	Time series Forecasting	Close and Date	1.475	54.32
Exponential Smoothing	Time series Forecasting	Close and Date	0.05276	84.32
Holter's Method	Times series Forecasting	Close and Date	0.080937	74.38
Decision Tree	Machine Learning	Open, Low, High, Volume	0.10589	67.21
Random Forest	Machine Learning	Open, Low, High, Volume	0.0225	89.02
Artificial Neural Network	Neural Networks	Open, Low, High, Volume	0.01244	93.32
LSTM (RNN)	Neural Networks	Open, Low, High, Volume	0.00669	97.32
TBATS	Time series Forecasting	Close and Date	0.060511	82.26

- From all the above models created the LSTM network has the less error when compared with the remaining models.
- The time series forecasting models predicts the output based on the seasonality patter of the previous data but not on the other independent variables of the data.
- The neural networks uses the backpropagation technique to decrease the loss in the process of training of the network.

Chapter 7. References

1) HTTPS://TOWARDSDATASCIENCE.COM/IN-DEPTH-UNDERSTANDING-OF-NEURALPROPHET-THROUGH-A-COMPLETE-EXAMPLE-2474F675BC96
2) T. FINLEY AND T. JOACHIMS, "TRAINING STRUCTURAL SVMS WHEN EXACT INFERENCE IS INTRACTABLE," IN 25TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, HELSINKI, FINLAND, 2008.
3) Y. S. X. Z. Y. DING, "FORECASTING FINANCIAL CONDITION OF CHINESE LISTED COMPANIES BASED ON SUPPORT VECTOR MACHINE," EXPERT SYST. APPL, VOL. 34, NO. 4, PP. 3081-3089, 2008.
4) H. N. I. H. M. K. A. S. KHALID ALKHATIB, "STOCK PRICE PREDICTION USING KNEAREST NEIGHBOR (KNN) ALGORITHM".
5) A. N. REFENES, A. ZAPRANIS AND G. FRANCIS, "STOCK PERFORMANCE MODELING USING NEURAL NETWORKS: A COMPARATIVE STUDY WITH REGRESSION MODELS," NEURAL NETWORKS, VOL. 7, NO. 2, PP. 375-388, 1994.
6) DR. P. K. SAHOO, MR. KRISHNA CHARLAPALLY "STOCK PRICE PREDICTION USING REGRESSION ANALYSIS" IN INTERNATIONAL JOURNAL OF SCIENTIFIC & ENGINEERING RESEARCH, VOLUME 6, ISSUE 3, MARCH-2015.
7) CHEN Y., HAO Y.A FEATURE WEIGHTED SUPPORT VECTOR MACHINE AND K-NEAREST NEIGHBOR ALGORITHM FOR STOCK MARKET INDICES PREDICTION
8) Theappsolutions.com/blog/development/recurrent-neural-networks/
9) Medium.com/analytics-vidhya/predicting-stock-prices-using-facebooks-prophet-model-b1716c733ea6
10) https://finance.yahoo.com/quote/AAPL/history?p=AAPL
11) Analyticsvidhya.com/blog/2021/12/stock-price-prediction-using-lstm/
12) https://www.analyticsvidhya.com/blog/2021/07/stock-market-forecasting-using-time-series-analysis-with-arima-model/
13) (2000) HOLT'S FORECASTING MODEL. In: Swamidass P.M. (eds) Encyclopedia of Production and Manufacturing Management. Springer, Boston, MA. https://doi.org/10.1007/1-4020-0612-8_409.
14) Medium.com/analytics-vidhya/time-series-forecasting-using-tbats-model.