Stock Price Prediction

Reshan Faraz
Phd19006
reshanf@
iiitd.ac.in

Arpit Saxena
MT20058
arpit20058@
iiitd.ac.in

Waquar Shamsi
MT20073
waquar20073@
iiitd.ac.in

Abstract

Stock Market Prediction is used to predict the future value any stock using the previous trends to learn how the stock price varies along the time. Using accurate stock prediction techniques one can determine a loose estimate of what could be the possible value of the stock in near future.The successful prediction of a stock's future price will maximize investor's gains. Though the price of any stock is subject to multiple factors and not just the previous trends but still, previous trends do provide a rough estimate for near future. We use RandomForests, KNN Regression and Support Vector Regression(SVR), ARIMA, SARIMA, LSTM and Prophet Algorithm to compare their results. Also used RandomForest, KNN and LSTM in combination with news sentiment analysis to get a better prediction for the stock prices, as the news certainly plays major role in stock prices.

Keywords: SVR (Support Vector Regression), Linear Regression, KNN Regression, ARIMA, SARIMA, Long Short Term Memory (LSTM), Sentiment Analysis, Prophet Algorithm.

1 Introduction

Stock Price prediction is an important issue in finance and to the companies.As billions of dollars are involved in the stock market, companies want to invest in the stock at the peak time if the financial market is predicted by some company then it will bring wealth to the organization.Since there has been no consensus on the validity of Efficient Market Hypothesis (EMH) which states the market is efficient and there is no space for prediction, researchers have strived for proving the predictability of the financial market [Lawrence, 1997] Forecast Stock Market Prices', University of Manitoba.]
With the advancement of machine learning, fast computing and vast information of data many researchers, engineers and mathematicians find this as interesting and challenging. There are several techniques used to predict the stock prices like Linear Regression (LR), Neural Networks (NNs),Genetic Algorithms (GAs), Support Vector Machines (SVMs) and many more. Also not just the previous trends of stock prices but also several other factors do influence the stock prices which makes it much more unpredictable. Though of these challenge can be solved using Machine Learning Algorithms too. Such as we can use sentiment analysis to determine the sentiments of news (as investors are often influenced by the news about the stocks) to determine whether the news may have negative or positive impact on the prices for the stocks. Also if they do affect the price, by what magnitude will the prices be affected.

2 Importance of the Project

Accurate Stock Prediction is very sought after. Various banks also use some kind of algorithms to get a estimate of future stock prices to decide where to invest to get maximum output. Even though even best of the stock prediction algorithms have failed to achieve 100% accuracy still even getting a approximate idea may help to earn profits. If it becomes possible to accurately predict stock prices then who will lose the money for so that some other person can earn as its a zero-sum game.

3 Literature Review

In paper [1], "Stock market predication using a linear regression [D.Bhuriya, 2017], authors present Linear Regression to predict the stock prices. In the paper the dataset that was used was collected from the TCS Stock Database as a collection of comma-separated values where each row consisted of a stock on a specific day along

with data on the volume, shares out, closing price, and other features for that day in time. Regression predicts a numerical value. Regression performs operations on a dataset where the target values have been defined already. And the result can be extended by adding new information. The relations which regression establishes between predictor and target values can make a pattern. This pattern can be used on other datasets which their target values are not known. In the paper they have talked about Linearregression, polynomial and RBF regression approach

In paper [2], "Stock Price Prediction" [N P Samarth, 2019], authors present Random Forest for performing regression to predict the stock prices. The authors predicted stock price for Walt Disney, with stock information for the years 2013-17. Random Forest is a variant of ensemble method. The ensemble methods make use of more than one learning model and then integrate the models to get a better prediction. Random Forest is a type of supervised learning algorithms. It creates large number of decision trees which are un-correlated with each other.

In paper [3], "Stock Price Prediction Using K-Nearest Neighbor (kNN) Algorithm" [Khalid Alkhatib, 2013], authors present k-nearest neighbours algorithm to predict the stock prices. The authors used k value of 5 and predict the closing price of the stock for each day using the features available. kNN is considered a lazy learning algorithm and does not build a model to make predictions, instead finds the closest k records of the training data to the test instance. Thus prediction takes longer time as the distances have to be computed to find the nearest neighbours.

4 DataSet Used

We used the Nifty50 data set available at

`https://www.kaggle.com/rohanrao/nifty50-stock-market-data`

It consists of records of top 50 Nifty stocks from 1st January, 2000 to 31st July, 2020. There are 50 csv files each for one of the stock and each row in the csv file corresponds to one day. The features available in the dataset are:

- Date
- Previous Close : Previous day's close price
- Open: Open price of day
- High: Highest price in day
- Low: Lowest price in day
- Close: Close price of day

Many of the features were not required and hence were dropped before using them to train the model.

For obtaining the news data for sentiment analysis, used beautiful soup python library to scrap data from Economic Times' archive website. Fetched all the headlines from the archive and stored it along with the date. Also calculated the polarity and subjectivity and stored it in the dataframe along each of the news headlines.

5 Data Analysis

We found NULL values in the columns $Trade$, $Deliverable\ Volume$, $\%Deliverable$ but we dropped these columns and have not used them in our training and prediction as they were not necessary for making the predictions.

6 Tasks Completed

6.1 Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called linear regression. The linear equation assigns one scale factor to each input value or column, called a coefficient and one additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

6.2 k-Nearest Neighbors

KNN is simple supervised machine learning algorithm. KNN assumes that similar data points appear close to each other and thus it uses distance functions to find the k nearest samples near the test instance and then give prediction according to values of the k-nearest samples.

6.3 Random Forest Regresion

Random Forest uses many decision trees to make the predictions. It uses bagging and tries to make the decision trees un-correlated. Its an ensemble learning method.

6.4 SVR (Support Vector Regression)

SVR uses the same principle as SVM, but for regression problems.(SVR) is characterized by the use of kernels, sparse solution, and VC control of the margin and the number of support vectors. SVR has been proven to be an effective tool in real-value function estimation.

6.5 Prophet Forecasting

Prophet, or "Facebook Prophet," is an open-source library for univariate (one variable) time series forecasting developed by Facebook.Prophet is open source software released by Facebook's Core Data Science team.we used version 0.7.It is robust to missing data and shifts in the trend, and typically handles outliers well. Prophet procedure is an additive regression model with piecewise linear or logistic growth curve trend,A yearly seasonal component modeled using Fourier series,A weekly seasonal component using dummy variables and even a user-provided list of important holidays.

6.6 ARIMA

ARIMA stands for AutoRegressive Integrated Moving Average. It is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.
Autoregression (AR) refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
Integrated (I) represents the differencing of raw observations to allow for the time series to become stationary, i.e., data values are replaced by the difference between the data values and the previous values.
Moving average (MA) incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations. Each component acts as a parameter with p, d, q respectively.

6.7 SARIMA

SARIMA stands for Seasonal AutoRegressive Integrated Moving Average,SARIMA model are ARIMA models with seasonal component with formula SARIMA(p,d,q)x(P,D,Q,s).
p and seasonal P: indicate number of autoregressive terms (lags of the stationarized series)
d and seasonal D: indicate differencing that must be done to stationarize series
q and seasonal Q: indicate number of moving average terms (lags of the forecast errors)
s: indicates seasonal length in the data

6.8 LSTM

LSTM (Long short-term memory) is an Artificial Recurrent Neural Network (RNN) architecture. The most important feature of LSTM is that it has feedback connections using which it becomes perfect for sequence data. Stock prediction is thus suitable application area for LSTMs as the prices of stocks depends on the sequence of past stock prices.

6.9 KNN,RandomForests,LSTM with News Sentiment Analysis

To achieve much higher accuracy predictions other factors influencing stock prices must be used, one of the most influential source is economic news. Economic news helps investors to decide where should they invest thus, a negative news may lead to fall in prices for the stock. In this project used Beautiful python library to scrape the news from Economic Times website `https://economictimes.indiatimes.com/archivelist`. Used the archives to scrape the news for the days in past. Found the sentiments of all news on a particular day and the took the sum of their polarity to finally determine whether that day has more positive news or negative and by what magnitude is it positive or negative. Increased or Decreased the original predicted values by the percentage equal to the magnitude of positivity or negativity of news for that day.

7 Results

7.1 Linear Regression

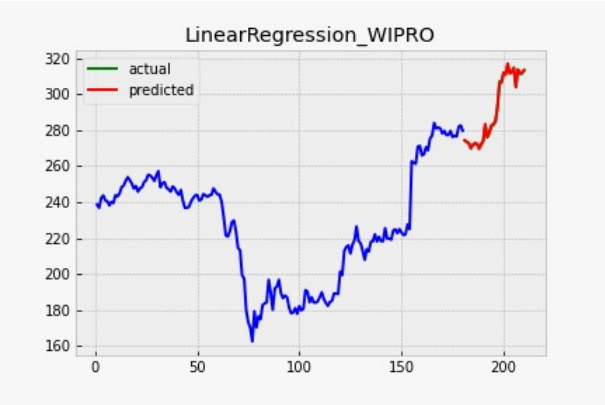

Figure 1: Wipro Stock Price Prediction, MSE: 0.394019802093781

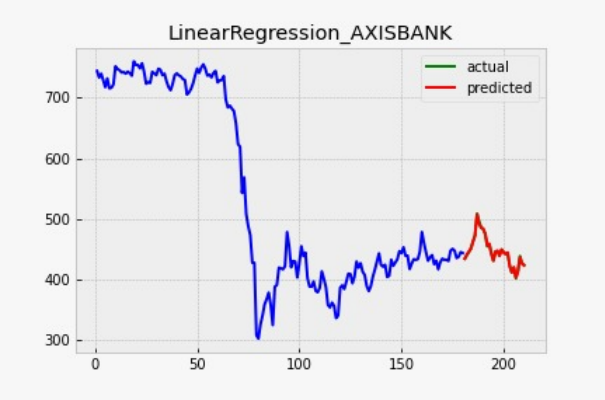

Figure 2: Axis Bank Stock Price Prediction, MSE: 2.095532259485699

7.2 k-Nearest Neighbors

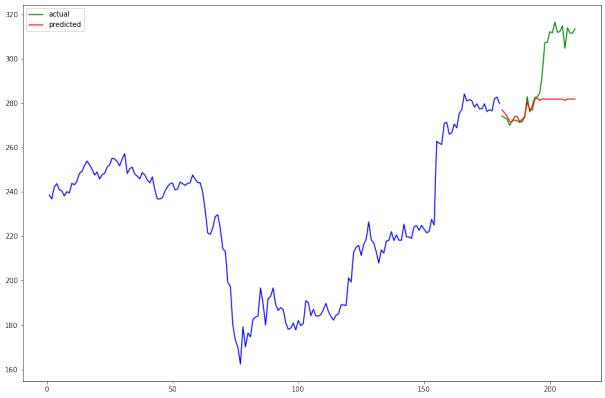

Figure 3: Wipro Stock Price Prediction

We can observe that it predicts very poorly for the stock Wipro.
MSE = 392.37358999999856
SCORE = -0.21528576624113516

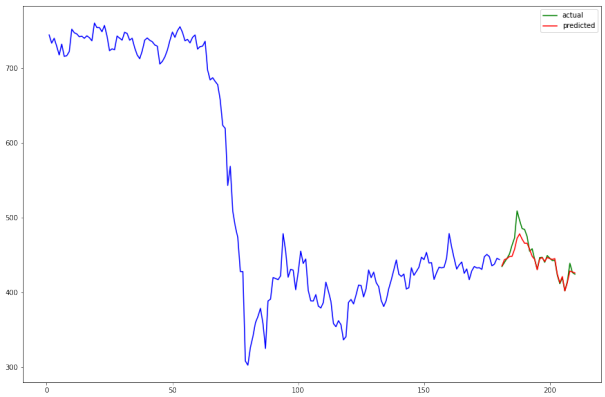

Figure 4: Axis Bank Stock Price Prediction

We can observe that it predicts very well for the stock AXISBANK.

MSE = 100.68490333333332
Score = 0.8340696682558897

7.3 Random Forest Regression

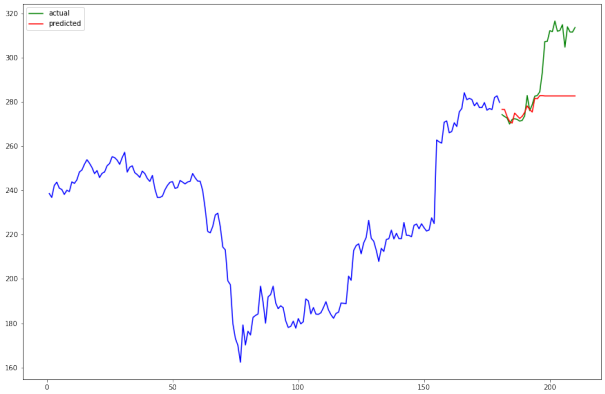

Figure 5: Wipro Stock Price Prediction

Again, random forest trees also perform poorly for stock WIRPO
MSE 370.2952646416647
Score -0.1469032980163396

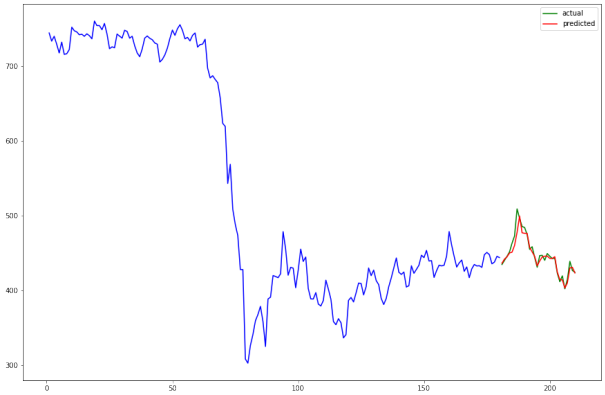

Figure 6: Axis Bank Stock Price Prediction

Again, random forest trees also perform very well for stock AXISBANK
MSE 58.96308539286477
Score 0.9028278918091259

7.4 SVR (Support Vector Regression)

The three model of SVR are linear ,polynomial and RBF. Fig. 7 represent the graphical view of all model and how they are different from each other

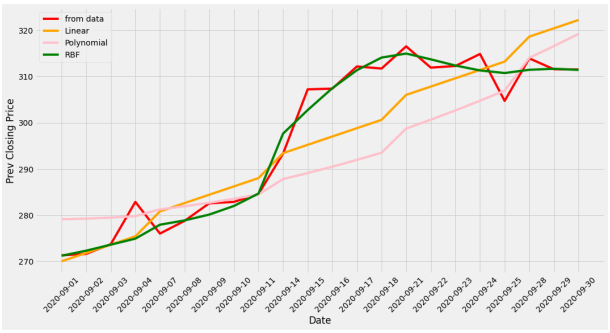

Figure 7: Wipro Stock Price Prediction

Model	Score
Linear	0.8147
Polynomial	0.6136
RBF	0.9699

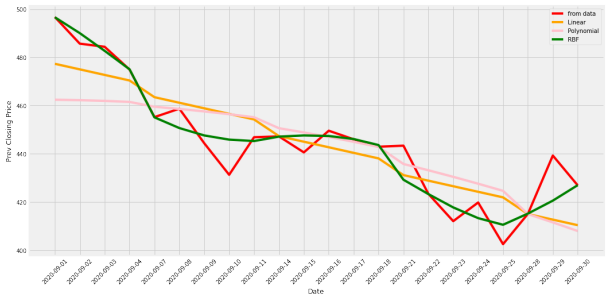

Figure 8: Axis Bank Stock Price Prediction

Model	Score
Linear	0.72056
Polynomial	0.5574
RBF	0.9158

7.5 Prophet Forecasting

We forecast for two company Wipro and Axis Bank and find out that it outperform both of them from previous methods.We perform grid search to find out the hyperparameter which gives least RMSE(Root mean Sqaure Error).Follwing are the parameters used for grid search
$changepoint_prior_scale$: $[0.001, 0.01, 0.1, 0.5]$
$seasonality_prior_scale$: $[0.01, 0.1, 1.0, 10.0]$ we took the past 5 year data starting from $2015-01-01$ to $2020-09-30$ and trying to predict the stock price for whole next year(365 days).
We perform two different approach:-
1-we predict the stock price for each date.
2-we predict the stock price monthly(Monthly Data Prediction). Following is the result from grid search.
We can observe the prediction for Axis Bank stock price as

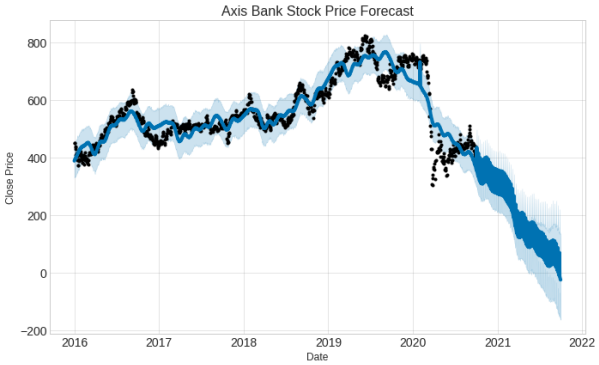

Figure 9: Axis Bank Stock Price Prediction

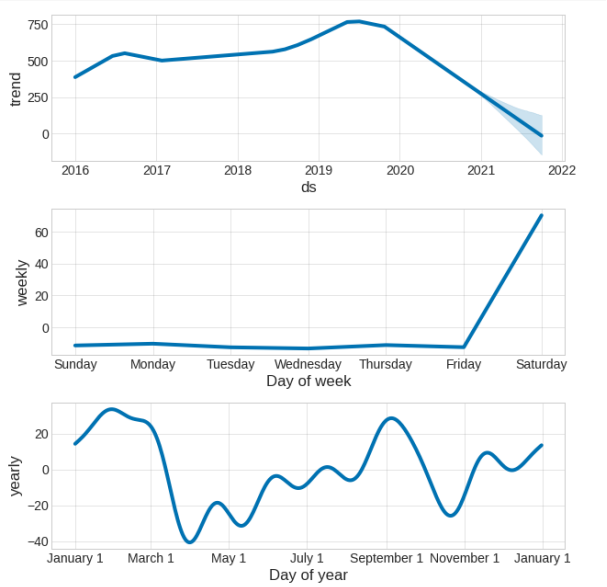

Figure 10: Axis Bank Stock Price Prediction

Following is the Monthly Data Prediction.

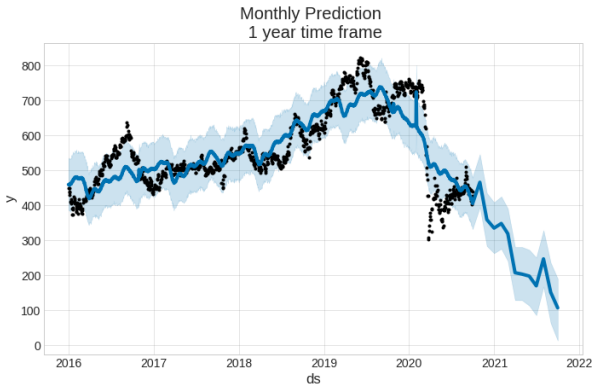

Figure 11: Axis Bank Monthly Stock Prediction

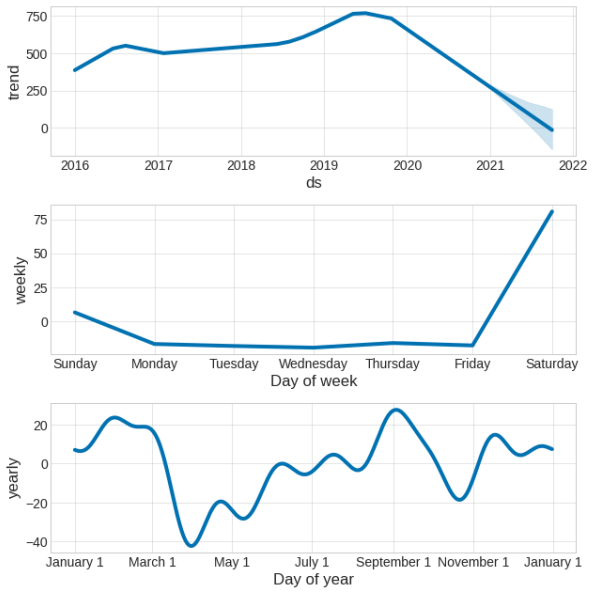

Figure 12: Axis Bank Monthly Stock Prediction

The MSE (Mean Square Error) is 50.43808496225144.
We can observe the prediction for WIPRO stock price as

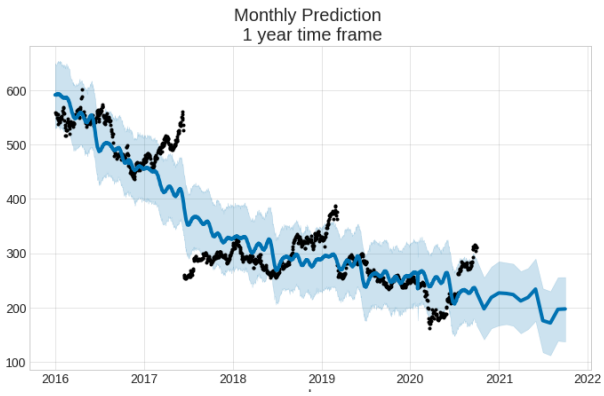

Figure 13: WIPRO Monthly Stock Prediction

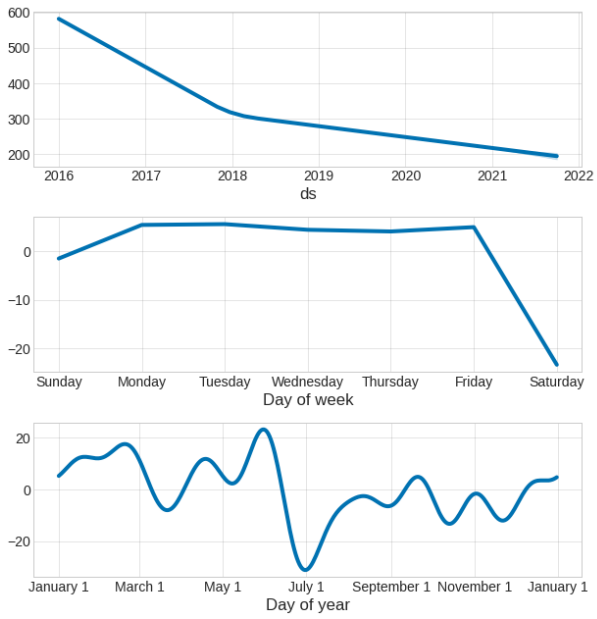

Figure 14: WIPRO Bank Monthly Stock Prediction, MSE=35.19616710299317

7.6 ARIMA

This model was trained on all the data except for the last 120 entries which were used as testing data

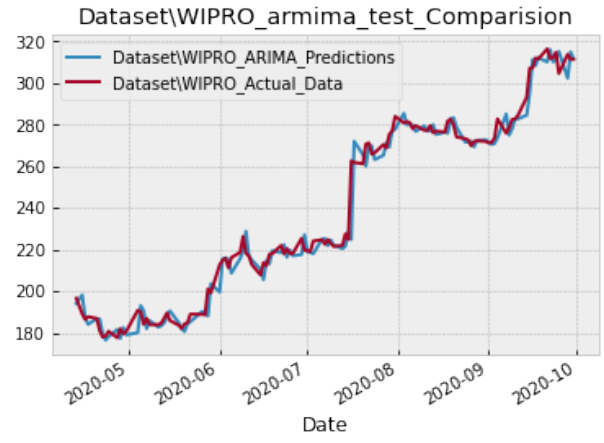

Figure 15: Wipro Stock Price Prediction, MSE: 34.887186486702966

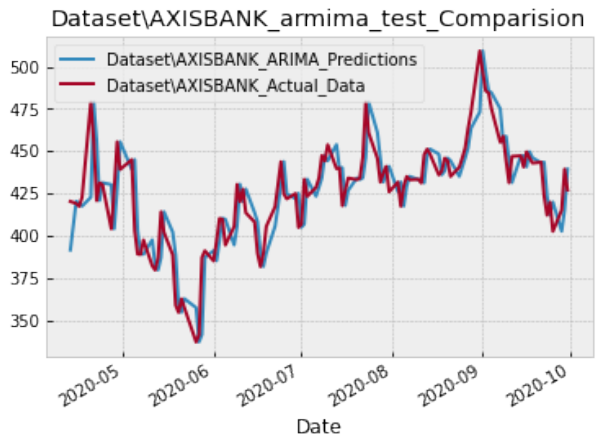

Figure 16: Axis Bank Stock Price Prediction, MSE: 242.14351669335326

7.7 SARIMA

This model was trained on all the data except for the last 120 entries which were used as testing data.

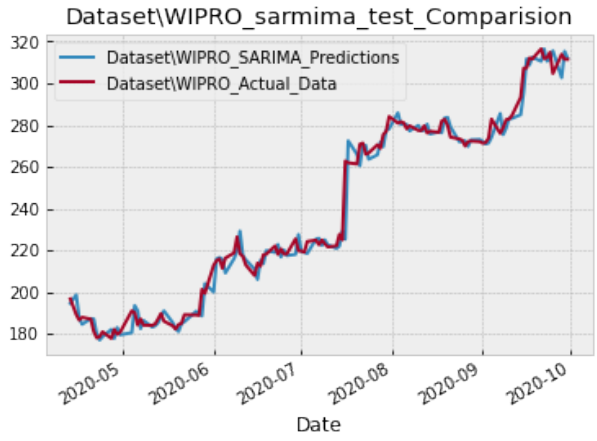

Figure 17: Wipro Stock Price Prediction, MSE: 34.33071791122736, s=12

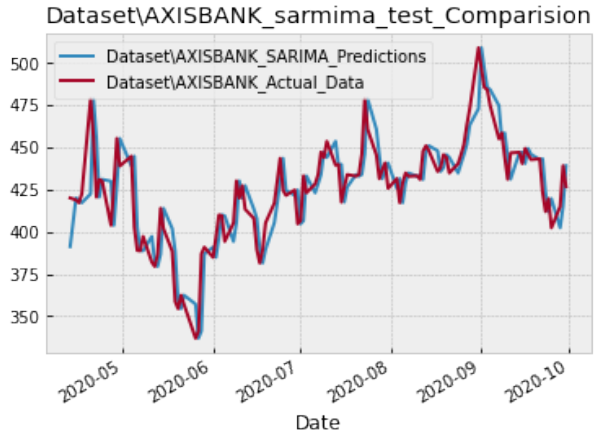

Figure 18: Axis Bank Stock Price Prediction, MSE: 242.1835416666665, s=12

7.8 LSTM

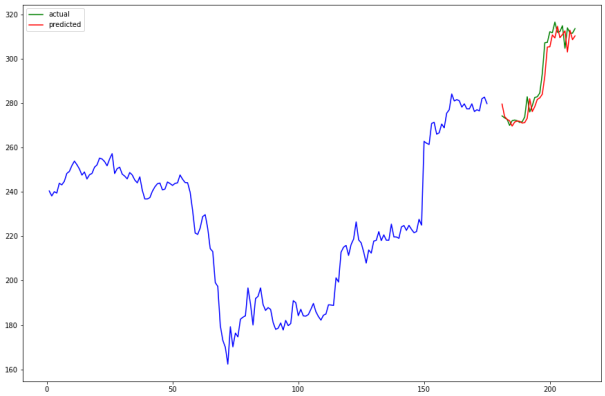

Figure 19: Wipro Stock Price Prediction

LSTM seems to be very accurate in predictions, MSE dropped significantly.
MSE 37.243806389843456
Score 0.8846459880589376

7.9 KNN with News Sentiment Analysis

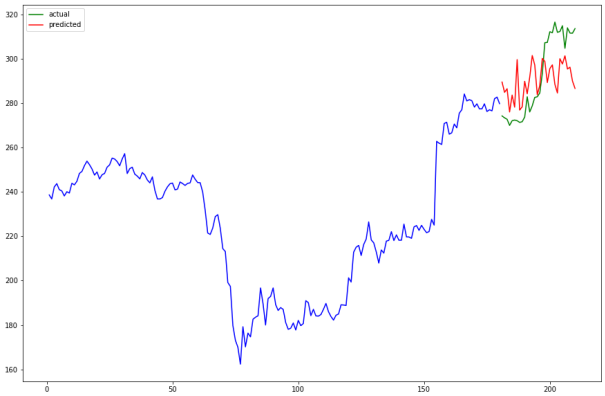

Figure 20: Wipro Stock Price Prediction

MSE 251.09891001940832
Score 0.22227963593255196
Here we can observe an improvement of 56.2626% over KNN without sentiment analysis.

7.10 RandomForests with News Sentiment Analysis

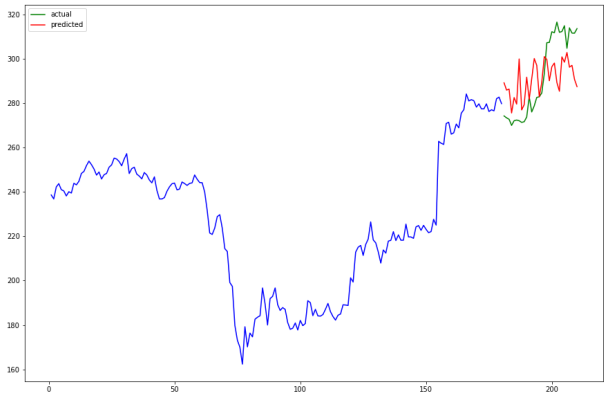

Figure 21: Wipro Stock Price Prediction

MSE 240.55499924577018
Score 0.25493694266850997
Here we can observe an improvement of 53.9337% over RandomForest without sentiment analysis.

7.11 LSTM with News Sentiment Analysis

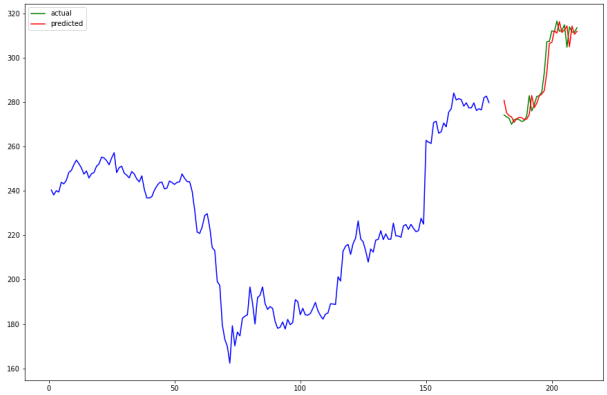

Figure 22: Wipro Stock Price Prediction

Again, random forest trees also perform poorly for stock WIRPO
MSE 24.82531760875757
Score 0.9231093633688786 Here we can observe an improvement of 50.0235% over LSTM without sentiment analysis.

8 Inferences:

We can conclude that predicting using only historical data is inferior compared to prediction using sentiment analysis on news along with historical data.
Also it can be concluded that LSTM is best for stock prediction among the algorithms tried in this report.

References

A.Sharma U.Singh D.Bhuriya, G.Kaushal. 2017. Stock market predication using a linear regression. 2:510–513.

Ismail Hmeidi Mohammed K. Ali Shatnawi Khalid Alkhatib, Hassan Najadat. 2013. Stock price prediction using k-nearest neighbor (knn) algorithm.

Ramon Lawrence. 1997. Using neural networks to forecast stock market prices. *University of Manitoba*, 333:2006–2013.

Hema N N P Samarth, Gowtham V Bhat. 2019. Stock price prediction. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*.

GOOGLE DRIVE LINK FOR CODE AND DATASET:

https://drive.google.com/drive/folders/1uVpQEoGsJJrIt5UipL08CozBa0559vXj?usp=sharing