# Efficacy of Predicting Models in Bear and Bull Stock Run

Submitted in partial fulfillment of the requirements
of the degree of

BACHELOR OF ENGINEERING

in

INFORMATION TECHNOLOGY

by


Group No: 23

| Roll No. | Name |
| --- | --- |
| 67 | Manav Bahl |
| 68 | Divesh Bhatia |
| 69 | Dev Rajdev |

Supervisor:

DR. ARUN KULKARNI
(Professor, Department of Information Technology, TSEC)


THADOMAL SHAHANI
TSEC
ENGINEERING COLLEGE

Information Technology Department

Thadomal Shahani Engineering College)
University of Mumbai
2018-2019

# CERTIFICATE

This is to certify that the project entitled **"Efficacy Of Predicting Models In Bear and Bull Stock Run"** is a bonafied work of

| Roll No. | Name |
|---|---|
| 67 | Manav Bahl |
| 68 | Divesh Bhatia |
| 69 | Dev Rajdev |

Submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"BACHELOR OF ENGINEERING"** in **"INFORMATION TECHNOLOGY"**.

(Name and sign)
Project Guide

(Name and sign)
Head of Department

(Name and sign)
Principal

# Project Report Approval for B.E

Project report entitled ***Efficacy Of Predicting Models In Bear and Bull Stock Run*** by

| Roll No. | (Name) |
|---|---|
| 67 | Manav Bahl |
| 68 | Divesh Bhatia |
| 69 | Dev Rajdev |

is approved for the degree of ***"BACHELOR OF ENGINEERING" in "INFORMATION TECHNOLOGY"***.

Examiners

1.--------------------------------------------

2.--------------------------------------------

Date:

Place:

# Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1) ______________________

(Signature)

Manav Bahl, Roll No:67

2) ______________________

(Signature)

Divesh Bhatia, Roll No:68

3) ______________________

(Signature)

Dev Rajdev, Roll No:69

Date:

# Acknowledgement

It is our pleasure to present this project on Efficacy of Predicting Models in Bear and Bull Stock Run Using Data Science and Machine Learning in Stock Markets. First, we would like to express deep gratitude and sincere thanks to the I.T Department and our project guide Prof. Kumkum Saxena for their encouragement and support throughout the project, also the honest feedback and constructive criticism provided helped a lot to improve and eliminate flaws. We would like to give our sincere thanks to the information technology department staff, who led us to an excellent and timely completion of the project with great technical support and lab support. We hope our work will be taken as a step towards improving the current stock prediction systems.

Thank You.

Manav Bahl(67)

Divesh Bhatia(68)

Dev Rajdev(69)

# Abstract

This report aims to understand the models used for data training and testing to build a model for predicting stock prices. The concentration lies in studying various models like Linear Regression, Decision Tree Regression, Long Short Term Memory (LSTM), Support Vector Machine (SVM). These models' study helps us predict the stock prices using computation and different machine learning concepts. It is essential to allot clean datasets into the machine that the machine can interpret and calculate. The stock market price prediction also is supported by the sentiment analysis model. This model uses web-scrapers and web crawlers to extract live data from various credible websites. Since the stock price is affected by real-time news, market condition, weather conditions, and deviation in the gold price, sentiment analysis helps bridge the gap and increase the accuracy of the predicting models. The model is robust and functions effectively on all stocks, interprets the future run, and inclines which trend the stock price might progress in the future. In comparison with different machine learning models, we will determine the most accurate model for predicting stock prices

# Table of Content

# List of Figures

# List of Tables

# Chapter1

# Introduction

## 1.1 Introduction

The stock market is a very volatile in-deterministic system with many factors influencing the direction of the trend on varying scales and multiple layers. The efficient MarketHypothesis states that the market is unbeatable. This approach makes predicting the uptrend or downtrend a challenging task. This section describes the restrictions of the traditional method in stock marketing research and lists the advantages of using data science, machine learning, and sentiment analysis. Stock market prediction aims to determine the future movement of the stock value of a financial exchange. The accurate prediction of share price movement will result in more profit investors can make. Predicting how the stock exchange will move is one of the foremost challenging issues because of many factors involved in stock prediction, like interest rates, politics, and economic growth, making the stock exchange volatile and complex to predict accurately. The prediction of the stock exchange is a difficult task as market movements are always subject to uncertainties. Stock market prediction methods are divided into two main categories: technical and fundamental analysis. Technical analysis focuses on analyzing historical stock prices to predict future stock values (i.e., it focuses on the price). On the opposite hand, fundamental analysis relies totally on analyzing unstructured textual information like financial news and earning reports.

## 1.2 Aim & Objective

The project aims to predict the future stock price of various stocks by studying machine learning models like Linear Regression, Decision Tree Regression, Long Short Term Memory (LSTM), and Support Vector Machine (SVM). To understand the efficacy of these models, we would draw a comparative analysis of the accuracy achieved from the models. The project also consists of sentiment analysis which helps to add environmental factors to narrow the accuracy to the maximum. The objective of stock market price prediction is to determine the future movement of the stock value of a financial exchange. The accurate prediction of share price movement will lead to more profit investors can make. Predicting how the stock market will move is one ofthe most challenging issues due to many factors involved in the stock prediction, such as interest rates, politics, and economic growth that make the stock market volatile and very hard to predict accurately. The prediction of shares offers enormous chances for profit and is 12 a significant motivation for research in this area; knowledge of stock movements by a fraction of a second can lead to high profits. The project aims to combine multiple techniques into a much more robust prediction model that can handle various scenarios in which investment can be beneficial. This project helpsto analyze the stocks using various models and to understand why the predictions show a bear or bull run using the sentiment approach, which functions in real-time.

## 1.3 Scope

The project's scope is to have a detailed understanding of trends that the stocks function with and analyze the most efficient way to predict those values. The project has a future scope and can be very effective in finance and stockbroking organizations. The performance and accuracy of different models can also be compared to determine the accuracy. The project also serves as another user interface (UI) for visualizing results from the research apart from Jupyter notebooks with lots of tables and graphs.

## 1.4 Difference between the traditional and modern approach of stock market prediction:

| Traditional Approach | Modern Approach |
|---|---|
| Relying on uptrends decided by market dictators | Studying the market trends and by data analysis |
| Investing in stocks was termed dangerous because knowledge was not available. | Information on stock data and market trend learning is available online at ease |
| The focus was mainly on a company's past performance and credibility. | Newsfeeds regarding the stock market highly affect the market trend and form a downhill movement in case of negative news. |
| This approach is based on predicting future prices by applying time-series analysis on previous trends. | The models can be trained for individual stocks with an adjusted bias for most reflective features. |

# Chapter 2

## Review Of Literature

- Manoj S Hegde [1] highlights Stock market analysis and prediction tools prevalent for several years now with various techniques and models to efficiently predict stock markets. The paper presents the design and implementation of a novel method to predict stock market trends. The approach is an ensemble model that considers historical stock data, tweets, and news affecting the stock prices of various companies and provides recommendations on which stocks to invest in for a particular duration. The model uses Long-Short Term Memory(LSTM) to learn and predict the future stock trend. Finally, the predicted stock within a period is converted into a graphical image; When tested with actual stock prices over a week, it was found that the model was able to achieve extremely high accuracy in predicting the stock trends.

- The work proposed in “A Comparison of Extreme Learning Machine and Support Vector Machine Classifiers" by Mihai Bucurica [2] helps understand the previously existing trading rules. It produces results better than the research proposed before. This research uses multiple proven market strategies to stimulate a real-time autonomous trader. This research focuses on short-term gains, which are excellent for hands-off trading. Their model implements three commonly used models: the Linear SVM Model, Polynomial SVM Model, and Radial basis Function (RBF) SVM Model, and compares the accuracy and efficiency of short and long-term predictions. The paper is beneficial and provides excellent insights into the different SVM Models.

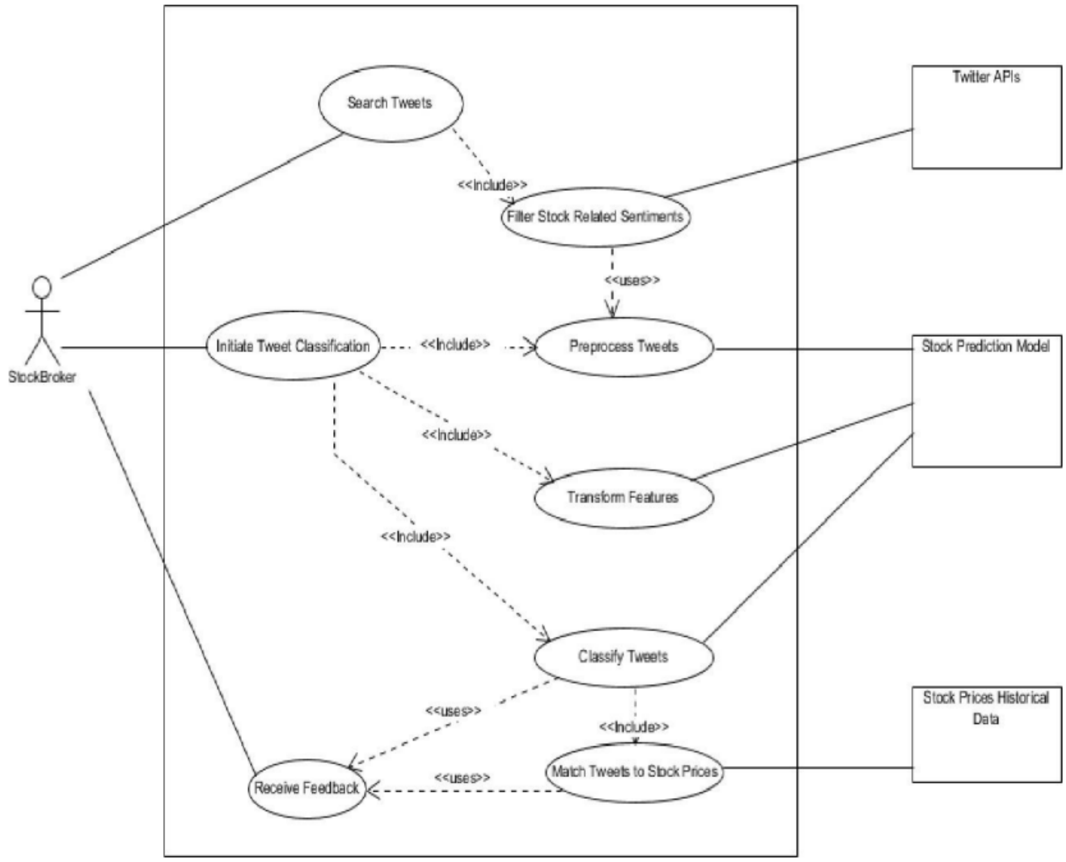
- Dinesh Bhuriya "Stock Market Prediction Using A Linear Regression" [3] investigates the success of machine learning models and event-driven Linear regression models. It is a severe challenge for investors and corporate stockholders to forecast the daily behavior of the stock market, which helps them invest with more confidence by taking risks and fluctuations into consideration. In this paper, by applying linear regression for the forecasting behavior of the TCS data set, the author proposed the method is suitable for comparing the other regression technique method, and the stockholders can invest confidentially based on that

- In “Sentimental Analysis Using Support Vector Machine,” [4] Aditya Gupta states that the addition of sentiment analysis adds valuable information to the prediction model and helps increase the accuracy of the model. Thus, this research takes news feeds into consideration to add credibility to sentiment analysis. They propose increasing the size of the corpus (training data) with each test. This is done by adding non-polarizing words found in the test data not present in the corpus. Thus, making the training data more efficient with each successive testing

# Chapter 3

# Proposed System

## 3.1 Data Flow Diagram




## 3.2 Technologies Used

1. Sklearn
2. Visualization
3. Python
4. Jupyter Notebook
5. Google Colab
6. SQL
7. Web Crawlers and Web Scrapers

# Chapter 4

# Design and Methodology

This project aims to build an application that outputs accurate recommendations in a quantifiable manner. For this purpose, five modules will be implemented, which are as follows:

● Long Short-Term Memory Networks (LSTM)
● Decision Tree Model
● Linear Regression Model
● Support Vector Machine(SVM)
● Sentiment Analysis

## 4.1 Datasets

For this comparative study, we have used fifteen different datasets from five different domains and compared the algorithms to find the algorithm that performs best on all three datasets.. These companies are chosen because they are one of the biggest contributors in their field in India. For instance, in India, Reliance is the biggest contributor in the telecommunications field. Similarly, Cipla is the biggest contributor in pharmaceuticals and HDFC is one of the biggest contributors in the finance field. Thus, these company datasets will be suitable for our comparative study as they have strong fundamentals and show a good uptrend. Moreover, HDFC also plays an important role in bank NIFTY. Furthermore, these datasets have been acquired from the historical stock prices section of the Bombay Stock Exchange(BSE) website. In the historical stock prices section, one can find datasets for any company listed in the Bombay Stock Exchange, for any date range required.

In our case, the datasets range from March 31, 2021, to February 1, 2022. These datasets are up to date so as to include the impact of SARS-CoV-2(Covid-19) on the stock prices of the company, and range for a year, making up of 219 instances or rows, to give us enough data for training as well as testing the algorithms. These datasets contain 5 feature variables, which are the Date, Open Price, High Price, Low Price, Close Price, Volume of Shares. Below are the snapshots of the first five rows of all three datasets.

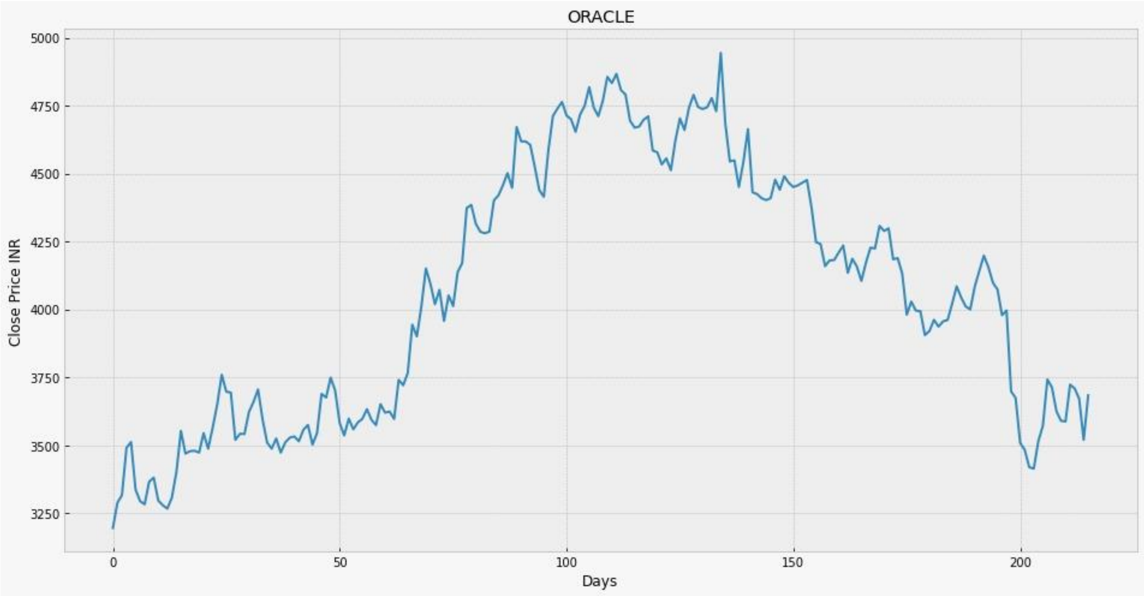
| 3 | Date | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|---|
| 4 | 04-05-2021 15:30 | 2024.95 | 2025 | 1962.1 | 1992.6 | 6864856 |
| 5 | 04-06-2021 15:30 | 2004 | 2004.95 | 1969 | 1984.3 | 6465241 |
| 6 | 04-07-2021 15:30 | 2000 | 2046.9 | 1993.3 | 2002.85 | 11198918 |
| 7 | 04-08-2021 15:30 | 2011 | 2022 | 1993 | 2005.35 | 7092878 |
| 8 | 04-09-2021 15:30 | 1998.45 | 2006.35 | 1980 | 1982.05 | 6478482 |
| 9 | 04-12-2021 15:30 | 1959 | 1961.4 | 1900.25 | 1911.15 | 9646031 |
| 10 | 4/13/2021 15:30:00 | 1924 | 1940.6 | 1917.85 | 1931.8 | 8958261 |
| 11 | 4/15/2021 15:30:00 | 1926.3 | 1961 | 1913 | 1944.3 | 9102492 |
| 12 | 4/16/2021 15:30:00 | 1936.6 | 1949.9 | 1926.45 | 1932.1 | 7225679 |
| 13 | 4/19/2021 15:30:00 | 1904 | 1916.4 | 1890 | 1901.7 | 8527967 |
| 14 | 4/20/2021 15:30:00 | 1910.5 | 1919 | 1890.45 | 1901.15 | 7939490 |
| 15 | 4/22/2021 15:30:00 | 1892.25 | 1914.45 | 1876.7 | 1906.4 | 6687573 |
| 16 | 4/23/2021 15:30:00 | 1906 | 1918.9 | 1895.35 | 1904.35 | 5459016 |
| 17 | 4/26/2021 15:30:00 | 1920 | 1962 | 1911.5 | 1937.85 | 9620785 |
| 18 | 4/27/2021 15:30:00 | 1940 | 1997.2 | 1938.25 | 1988.65 | 9226547 |
| 19 | 4/28/2021 15:30:00 | 1997.85 | 2008 | 1980.15 | 1997.3 | 7902002 |
| 20 | 4/29/2021 15:30:00 | 2022.9 | 2044.5 | 2007.3 | 2024.05 | 8035915 |
| 21 | 4/30/2021 15:30:00 | 2008.5 | 2036 | 1987.55 | 1994.5 | 9150974 |
| 22 | 05-03-2021 15:30 | 1966 | 1979 | 1943.1 | 1959.05 | 10909942 |
| 23 | 05-04-2021 15:30 | 1950 | 1967.8 | 1911 | 1916.6 | 10083693 |
| 24 | 05-05-2021 15:30 | 1923.35 | 1938.5 | 1908.05 | 1920.1 | 5719649 |
| 25 | 05-06-2021 15:30 | 1921.85 | 1935 | 1906.6 | 1931 | 6749281 |
| 26 | 05-07-2021 15:30 | 1937.75 | 1955.65 | 1926 | 1931.75 | 5671163 |
| 27 | 05-10-2021 15:30 | 1939 | 1946.8 | 1920.95 | 1926.2 | 6433879 |
| 28 | 05-11-2021 15:30 | 1915 | 1938.55 | 1910 | 1933.15 | 6220217 |

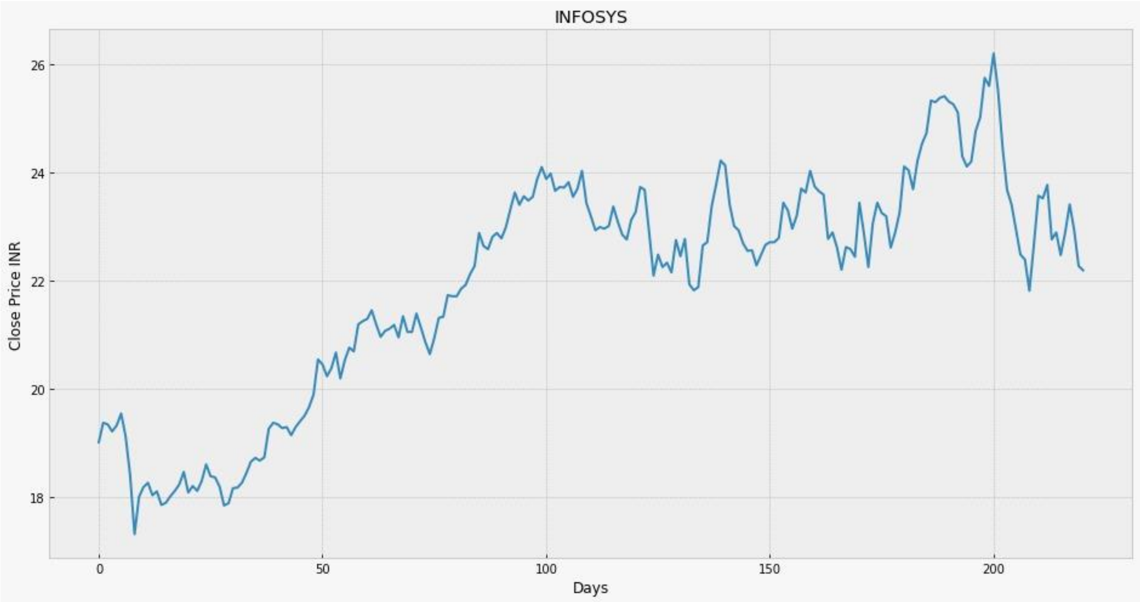
Stock Value Dataset of Reliance

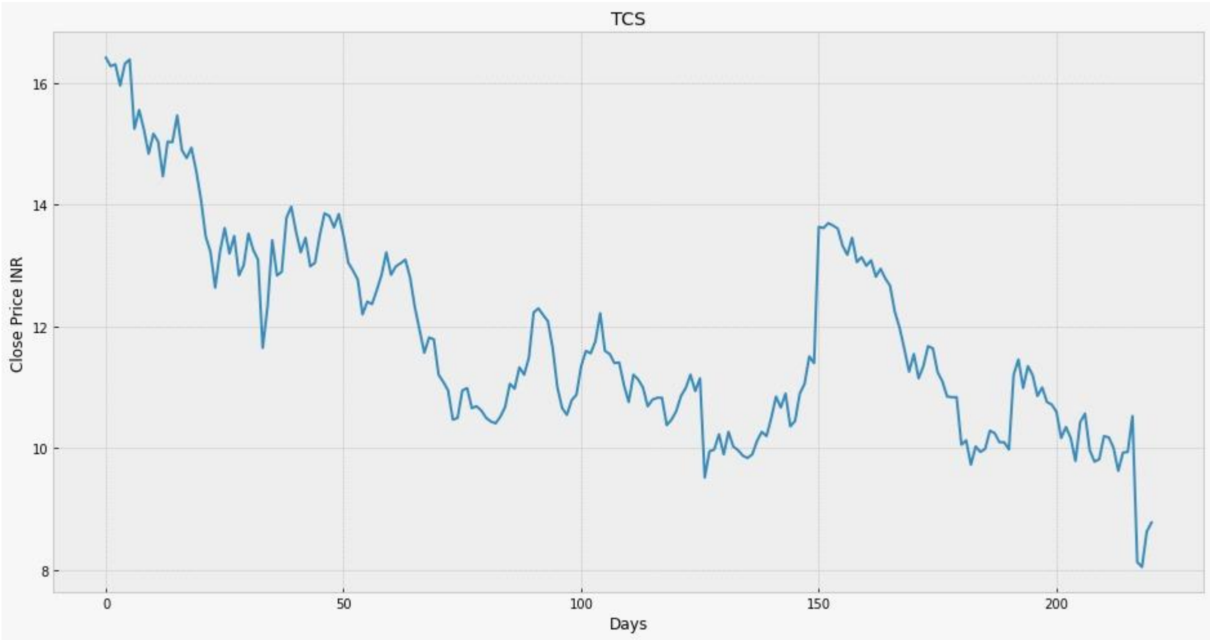
| 3 | Date | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|---|
| 4 | 04-05-2021 15:30 | 1480 | 1485 | 1431 | 1449.6 | 8003293 |
| 5 | 04-06-2021 15:30 | 1460 | 1462.65 | 1432.65 | 1440.25 | 7537867 |
| 6 | 04-07-2021 15:30 | 1439.3 | 1456.7 | 1421.55 | 1447.2 | 12544090 |
| 7 | 04-08-2021 15:30 | 1453 | 1460.9 | 1430.5 | 1432.8 | 8806796 |
| 8 | 04-09-2021 15:30 | 1426 | 1432.8 | 1415.1 | 1421.75 | 14078908 |
| 9 | 04-12-2021 15:30 | 1393 | 1399 | 1353 | 1367.05 | 11274564 |
| 10 | 4/13/2021 15:30:00 | 1368 | 1406.45 | 1361 | 1400.35 | 9300341 |
| 11 | 4/15/2021 15:30:00 | 1405 | 1436.7 | 1391 | 1430.1 | 17222492 |
| 12 | 4/16/2021 15:30:00 | 1434.95 | 1445 | 1423.5 | 1428.65 | 7803263 |
| 13 | 4/19/2021 15:30:00 | 1390 | 1417.7 | 1372.3 | 1412.4 | 12034621 |
| 14 | 4/20/2021 15:30:00 | 1425 | 1426.4 | 1383.95 | 1391.4 | 11593135 |
| 15 | 4/22/2021 15:30:00 | 1380 | 1426.8 | 1371.05 | 1422.5 | 19242656 |
| 16 | 4/23/2021 15:30:00 | 1409 | 1434.6 | 1400.2 | 1414.15 | 11356764 |
| 17 | 4/26/2021 15:30:00 | 1413 | 1429 | 1402.75 | 1404.8 | 15085476 |
| 18 | 4/27/2021 15:30:00 | 1407.25 | 1442 | 1404.8 | 1438.7 | 10296453 |
| 19 | 4/28/2021 15:30:00 | 1436.25 | 1479 | 1431 | 1476.8 | 12051970 |
| 20 | 4/29/2021 15:30:00 | 1486.2 | 1503.65 | 1461 | 1472.5 | 12039276 |
| 21 | 4/30/2021 15:30:00 | 1445 | 1453.8 | 1407.5 | 1412.3 | 17616451 |
| 22 | 05-03-2021 15:30 | 1393 | 1421.9 | 1377.3 | 1414.45 | 11236850 |
| 23 | 05-04-2021 15:30 | 1409.95 | 1423 | 1383.3 | 1388.35 | 10743164 |
| 24 | 05-05-2021 15:30 | 1401 | 1409.6 | 1381.7 | 1402.6 | 7210806 |
| 25 | 05-06-2021 15:30 | 1407.6 | 1410.8 | 1395 | 1400.9 | 5738522 |
| 26 | 05-07-2021 15:30 | 1412.95 | 1424.95 | 1410.25 | 1414.75 | 6024167 |
| 27 | 05-10-2021 15:30 | 1427 | 1430 | 1412.8 | 1419.85 | 5530025 |
| 28 | 05-11-2021 15:30 | 1396 | 1424.2 | 1395.05 | 1403.55 | 7259517 |

Stock Value Dataset of HDFC Bank

For our study, out of all the 5 features, we will be using only the Date and the Closing Price. The closing price is the raw price or cash value of the last transacted price in security before the market officially closes for normal trading. Even in the era of 24-hour trading, there is a closing price for any stock or other security, and it is the final price at which it trades during regular market hours on any given day. Until trade resumes on the next trading day, the closing price is considered the most reliable value of a stock or other security. Investors often use the closing price to compare a stock's output since the previous day, and closing prices are often used to create line graphs representing average price movements over time. For a better understanding of Close Price, the line graphs of the closing prices are provided below.

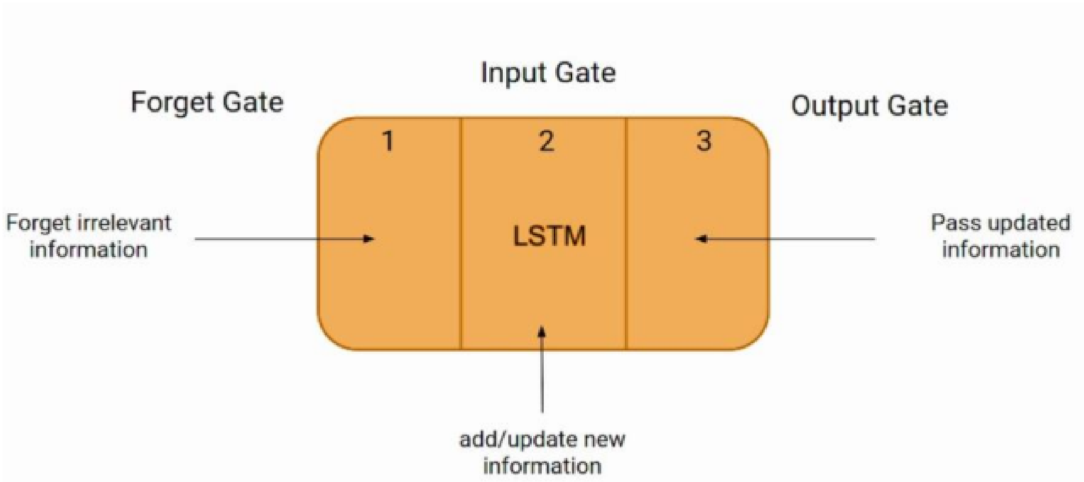




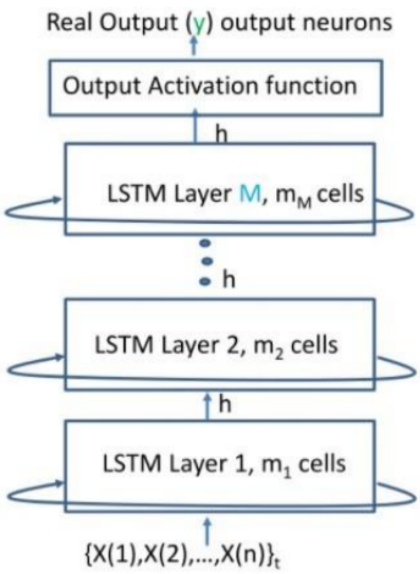

## 4.2 Model Implementation

### 4.2.1 Long Short-Term Memory (LSTM)




Long Short-Term Memory (LSTM) [5] is an artificial Recurrent Neural Network (RNN) architecture used in deep learning. LSTM networks are well-suited for classifying, processing, and making predictions based on time series data. There can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem that can be encountered when training traditional RNNs. Relative insensitivity to gap length is advantageous for LSTM over RNNs, hidden Markov models, and other sequence learning methods in numerous applications. One hundred epochs trained the model for each stock, and the testing sample is performed on each stock. As for long-term time series prediction, the LSTM brings advantages of selecting the essential and relevant information hence enhancing the predictive accuracy.

Typically, the deeper neural network can explain more complicated problems than a single-larger neural network. However, another exciting finding we discovered is that the 17 stacked-LSTM does not significantly outperform the LSTM in stock price prediction. Instead, the performance LSTM even beat the stacked-LSTM in certain instances. It has proved that the more complex representative does not necessarily improve the predictive power. It is possibly due to the more complicated neural network representation causing an overfitting issue; the more significant number of parameters in the stacked-LSTM memory model does not generalize well in the unseen data.




Long Short-Term Memory models are compelling time-series models. They can predict an arbitrary number of steps into the future. An LSTM module (or cell) has five essential components to model both long-term and short-term data.

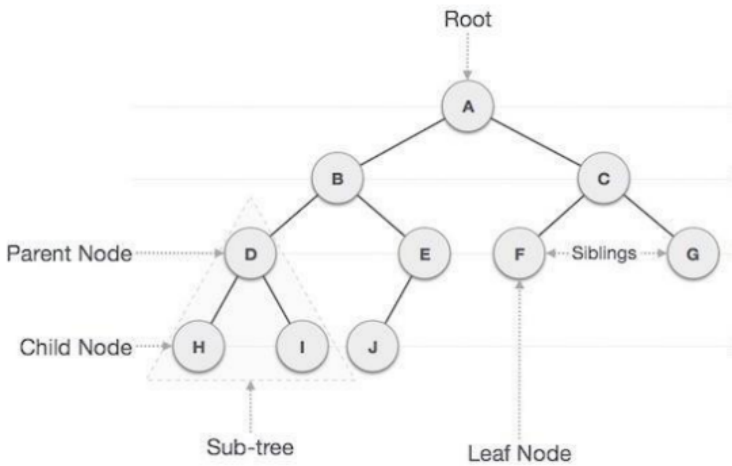
-Cell state (ct) - This represents the internal memory of the cell, which stores both short-term memory and long-term memories.

- Hidden state (ht) - This is output state information calculated w.r.t. current input, previously hidden state, and current cell input, which you eventually used to predict the future stock market prices. Additionally, the hidden state can only retrieve the short or long-term or both types of memory stored in the cell state to make the following prediction.

- Input gate (it) - Decides how much information from current input flows to the cell state
- Forget gate (ft) - Decides how much information from the current input and the previous cell state flows into the current cell state. 18
- Output gate (ot) - Decides how much information from the current cell state flows into the hidden state so that if needed, LSTM can only pick the long-term memories or short-term memories and long-term memorie

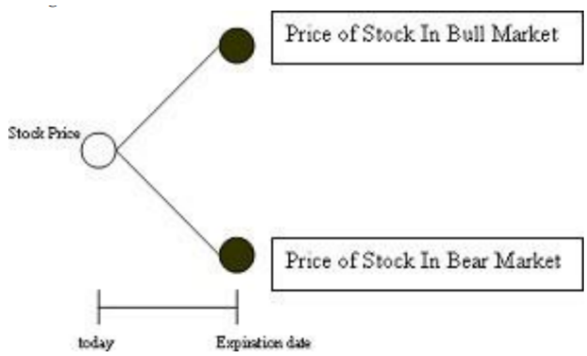
### 4.2.2 Decision Tree Model

Decision Tree [6]: A decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the trial, and each leaf node (terminal node) holds a class label. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the attribute's value, as shown in the above figure. This process is then repeated for the subtree rooted at the new node.




The goal of a decision tree is to encapsulate the training data in the smallest possible tree. The rationale for minimizing the tree size is the logical rule that the simplest possible explanation for a set of phenomena is preferred over other reasons. Also, small trees produce decisions faster than large trees, and they are much easier to look at and understand. There are various methods and techniques to control the depth of the tree. There are several steps involved in the building of a decision tree. Splitting: The process of partitioning the dataset into subsets.

Splits are formed on a particular variable and in a specific location. For each split, two determinations are made the predictor variable used for the split, called the splitting variable, and the set of values for the predictor variable (which are split between the left child node and the right child node), called the split point. For example, the split is based on a particular criterion, Gini (for classification) or sums of squares (for regression) from the entire data set. The leaf node, also called a terminal node, contains a small subset of the observations. Splitting continues until a leaf node is constructed. Pruning: The shortening of branches of the tree. Pruning reduces the size of the tree by turning some branch nodes into leaf nodes and removing the leaf nodes under the original branch.



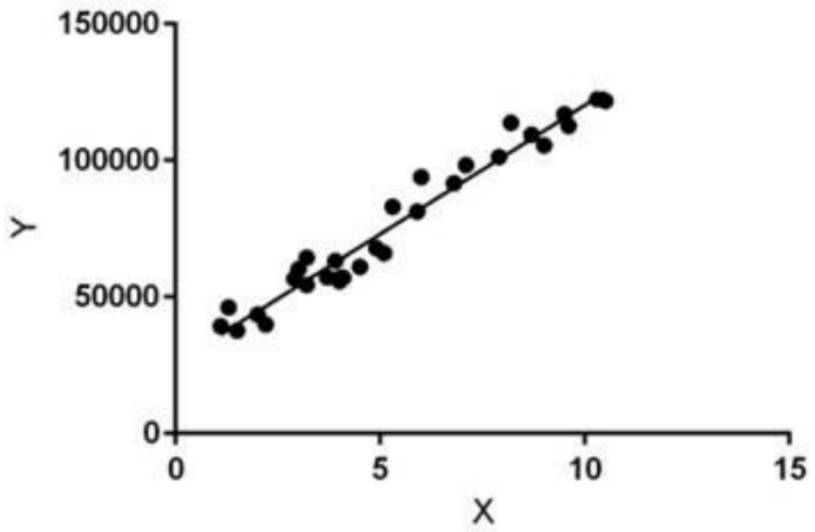

Pruning is helpful because classification trees may fit the training data well but may do a poor job of classifying new values. Lower branches may be strongly affected by outliers. Pruning enables you to find the next largest tree and minimize the problem. A simpler tree often avoids overfitting. Tree Selection: The process of finding the smallest tree that fits the data. Usually, this is the tree that yields the lowest cross-validated error. Using a decision tree for prediction is an alternative method to linear regression

### 4.2.3 Linear Regression Model

Linear Regression [7] is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models target a prediction value based on independent variables. It is mainly used for finding out the relationship between variables and forecasting. Different regression models differ based on the relationship between dependent and independent variables they are considering and the number of independent 20 variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

Hence, the name is Linear Regression. As shown in figure 7, X (input) is the work experience, and Y (output) is the person's salary. The regression line is the best fit line for our model.




Hypothesis function for Linear Regression :

$Y = \theta 1 + \theta 2.x$ (Eq.1)

While training the model we are given :

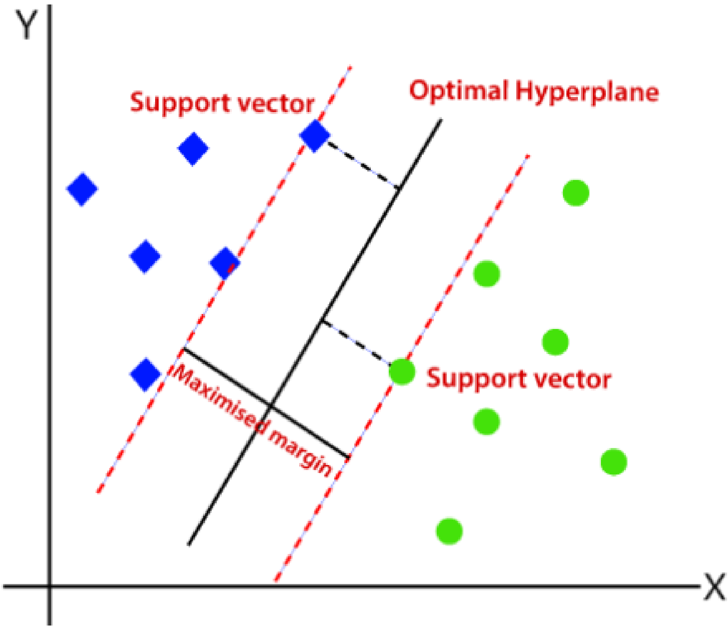
x: input training data (univariate – one input variable(parameter))y: labels to data (supervised learning) When training the model – it fits the best line to predict the value of y for a given value of x.The model gets the best regression fit line by finding the best θ1 and θ2 values.

θ1: intercept θ2: coefficient of x Once we find the best θ1 and θ2 values, we get the best fit line. Linear Regression is usually very dependent on the NumPy values to calculate the values of θ1 and θ2.

### 4.2.4 Support Vector Machine Model (SVM)

A Support Vector Machine (SVM) [8] is a supervised machine learning model that produces significant accuracy with less computation power. In the SVM algorithm, we plot each data item as a point in n dimensional space (where n is the number of features you have), with the value of each feature being the value of a particular coordinate. We perform classification by finding the hyper-plane that differentiates the two classes. The approach of using SVMs to solve regression problems is called Support Vector Regression (SVR).

Hyperplane: There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but it is required to determine the best decision boundary that helps classify the data points. This best boundary is known as the hyperplane of SVM. The dimensions of the hyperplane depend on the features present in the dataset, which means if there are two features (as shown in the Figure 8), then the hyperplane will be a straight line, and if there are three features, then the hyperplane will be a 2- dimension plane. It is advisable to create a hyperplane with a maximum margin, which means the maximum distance between the data points.

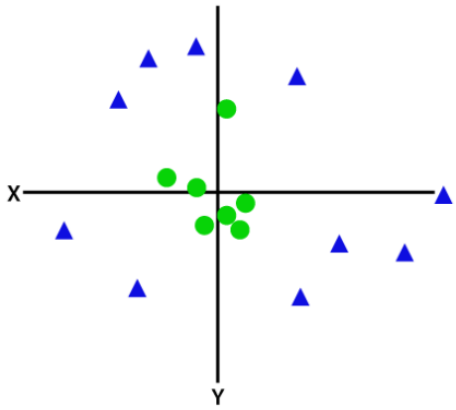



Support Vectors: The data points or vectors that are the closest to the hyperplane and affect the hyperplane position are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector. SVM can be of 2 types:

1. Linear
2. Non- Linear

Linear SVM: Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier. The working of the

Linear SVM algorithm can be understood by using an example. Suppose a dataset has two tags (green and blue), and the dataset has two features x1 and x2. A classifier is needed that can classify the pair (x1, x2) of coordinates in either green or blue. So, as it is a 2-D space, just using a straight line can easily separate these two classes. But there can be multiple lines that can separate these classes.




Hence, the SVM algorithm helps find the best line or decision boundary; this best boundary or region is called a hyperplane. The SVM algorithm finds the closest point of the lines from both classes. These points are called support vectors. The distance between the vectors and the hyperplane is called the margin. Moreover, the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane. Advantages of using Linear SVM:

1. Training an SVM with a Linear Kernel is Faster than with any other Kernel.
2. When training an SVM with a Linear Kernel, only the optimization of the C Regularization parameter is required.

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data. If a dataset cannot be classified using a straight line, such data is termed a non-linear data classifier used as a Non-linear SVM classifier. The working of the Non-Linear SVM algorithm can be understood by using an example. If data is linearly arranged, it can be separated by using a straight line, but for non-linear data one cannot draw a single straight line. So to separate these data points, add one more dimension. For linear data, two dimensions, x and y, were used, so for non-linear data, add a third dimension, z. Adding a third dimension makes the separation of data points muchmore accessible and accurate. Therefore, the data have been plotted from

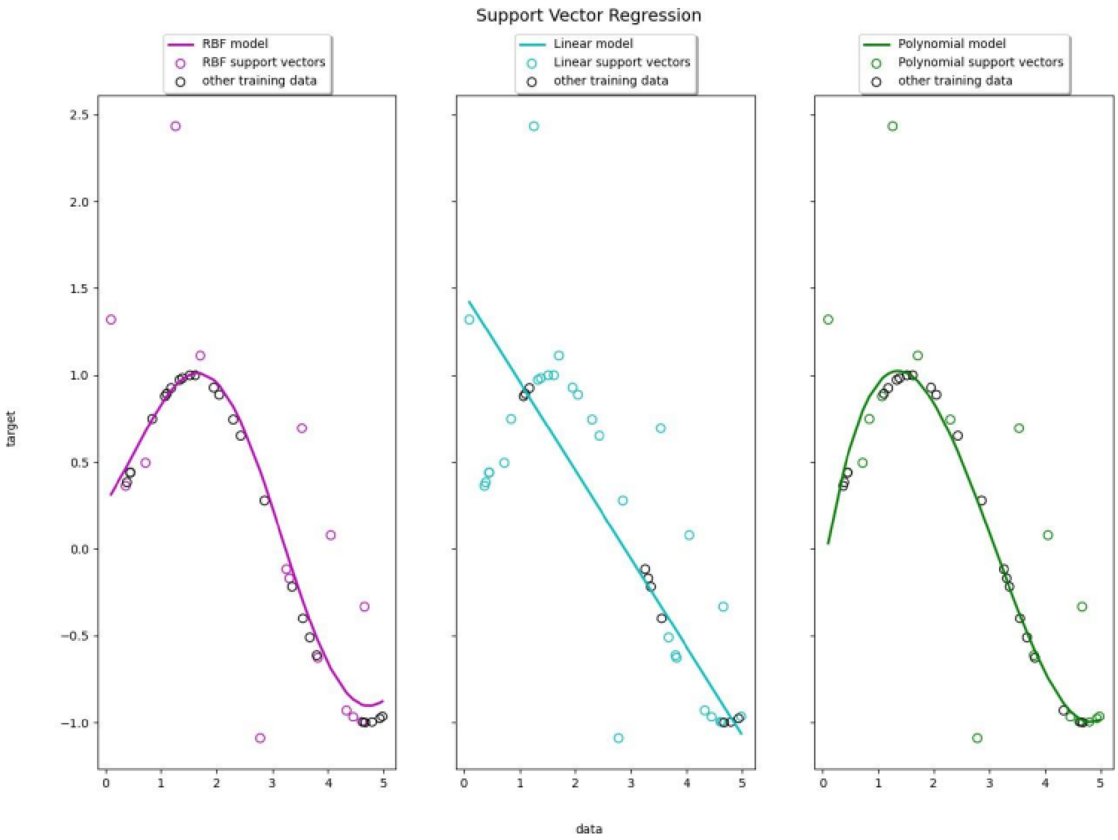
2-D space to 3-D space.Now one can easily classify the data by drawing the best hyperplane between them.

Radial basis Function(RBF) kernel [9]: RBF Kernel is similar to K-Nearest NeighborAlgorithm. It has the advantages of KNN and overcomes the space complexity problem. RBF Kernel Support Vector Machines need to store the support vectors during training and not the entire dataset. The RBF Kernel Support Vector Machines is implemented in the scikit-learn library and has two hyperparameters associated with it, 'C' for SVM and 'γ' for the RBF Kernel. C- It is a hypermeter in SVM to control error. If a low value of C, then lesserror and a higher value of C means more significant error. That does not mean that low error means we have a good model. It depends upon on datasets how much the error dataset consists of—these conditions for ideal datasets. Gamma is also a hyperparameter that we have to set before training the model. Gamma decides how much curvature we want in a decision boundary. Gamma high means more curvature. Gamma low means less curvature

Polynomial Kernel: It represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models. The polynomial kernel looks at the given features of input samples to determine their similarity and combinations. Selecting the optimal degree of a polynomial kernel is critical to ensure good generalization of the resulting support vector machine model. The degree parameter controls the flexibility of the decision boundary. Higher degree kernels yield a more flexible decision boundary. Using degree=1 is the same as using a 'linear' kernel.

Advantages of SVM:

● Good for smaller cleaner datasets
● It is more effective in high dimensional spaces.
● It is relatively memory efficientDisadvantages of SVM:
● The algorithm is not suitable for large data sets.
● It does not perform very well when the data set has more noise i.e. target classes are overlapping.
● As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.
● Picking the right kernel can be computationally intensive


Support Vector Regression
RBF model
RBF support vectors
other training data
Linear model
Linear support vectors
other training data
Polynomial model
Polynomial support vectors
other training data
target
data
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
0
1
2
3
4
5

Difference between Linear SVM and Non-Linear SVM:

| LINEAR SVM | NON-LINEAR SVM |
|---|---|
| 1. It can be easily separated with a linear line. | 1. It cannot be easily separated with a linear line. |
| 2. Data is classified with the help of hyperplane. | 2. We use Kernels to make non-separable data into separable data. |
| 3. Data can be easily classified by drawing a straight line. | 3. We map data into high dimensional space to classify. |

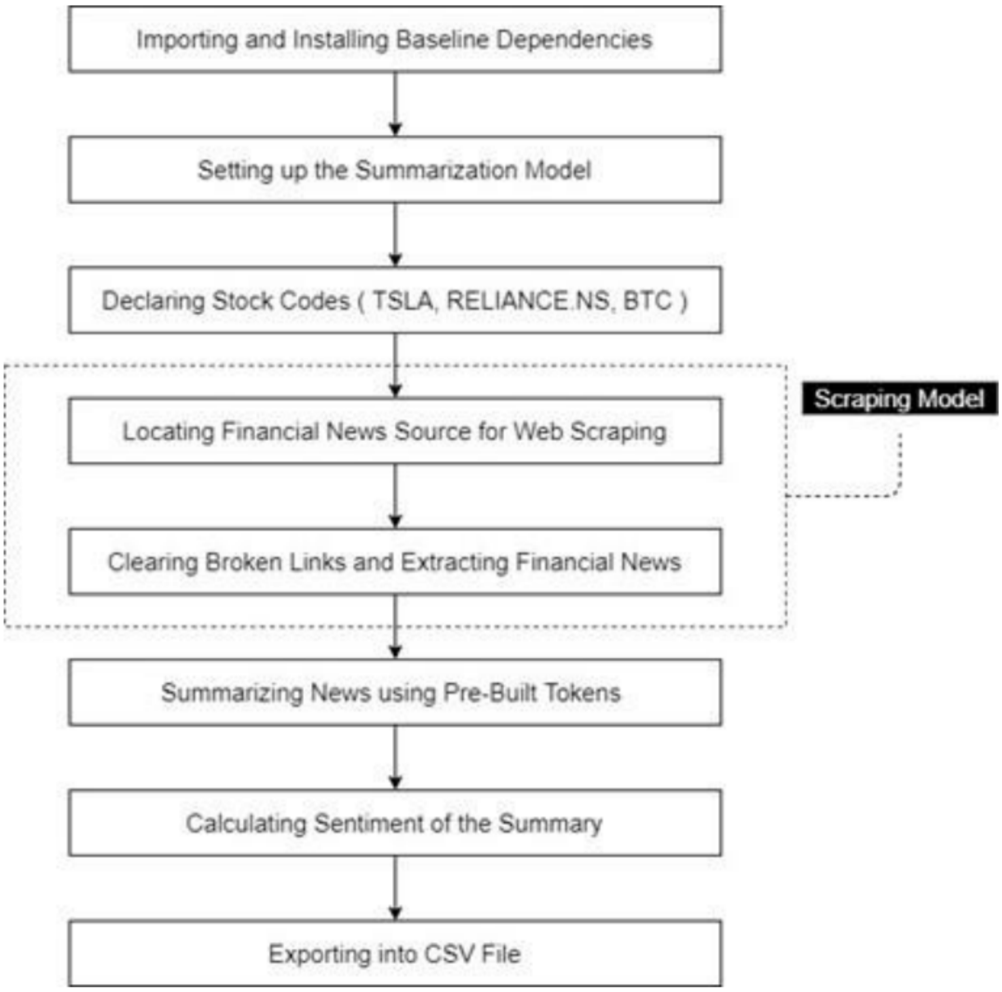
### 4.2.5 Sentiment Analysis Model

Sentiment analysis [10] is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Sentiment Analysis plays a crucial role in Stock value prediction as the highs and lows of stock depend on previous trends and are subjected to surrounding news, weather conditions, and highly targeted insights in the respective domain. The sentiment analysis works based on various API's and Web Scrapers.The working of a sentiment analysis model is as follows

The Scraping Model is used to access the internet and analyze and extract financial information from websites such as Google News and Yahoo Finance. It implements a throughout the search and returns the website links, which then have to be cleaned andtested to check if they are functional. The scraping tool also searches the website for particularly mentioned stock codes and excludes news with political importance.

The summary process is essential as it helps to understand the article in less than 55 characters and makes it accurate for the model to detect the sentiment with precision. The model then tabulates the scores and their respective analysis in a CSV file, making it easier for the users to understand the market standings of the particular stocks.

The sentiment analysis works on a broader perspective like National (BSE) as well as international stocks (NSE). The web crawlers and scrapers are trained to extract information on the particular stock code mentioned in the model. This makes the Sentiment Analysis very effective and makes the result more precise


Importing and Installing Baseline Dependencies
Setting up the Summarization Model
Declaring Stock Codes ( TSLA, RELIANCE.NS, BTC )
Locating Financial News Source for Web Scraping
Clearing Broken Links and Extracting Financial News
Scraping Model
Summarizing News using Pre-Built Tokens
Calculating Sentiment of the Summary
Exporting into CSV File

# Chapter 5

# Implementation

## 5.1 Implementation of Linear Regression Model

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

θ1: intercept

θ2: coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. Linear Regression is usually very dependent on the NumPy values to calculate the values of θ1 and θ2. In this model, 75 % of the data is used for training the model and 25% of data is used for testing. The prediction is implemented for 25 days prior to comparing the prediction to the actual value of the stock. Linear Prediction Model is used to understand the point-to-point plotting and to interpret a result based on previous trends. Linear Regression library is imported from sklearn.linear_model

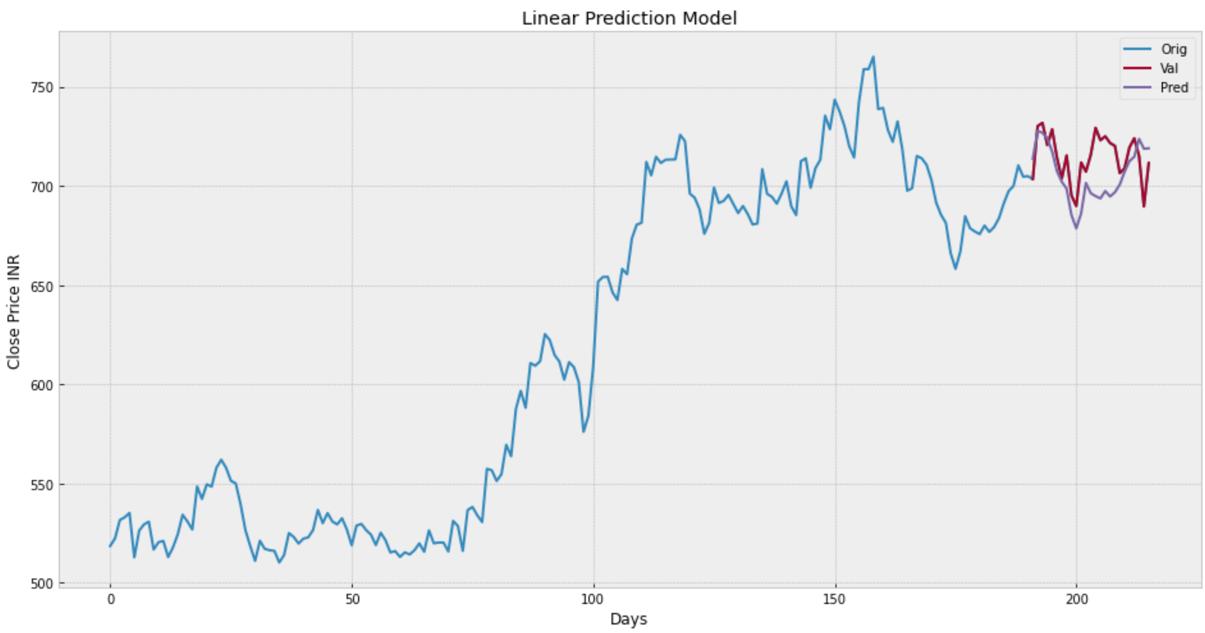



*Figure: Linear Regression Model Prediction for Airtel Stock*

## 5.2 Implementation of Decision Tree

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the attribute's value. This process is then repeated for the subtree rooted at the new node. Decision Tree Prediction Model dependent on nodes, splitting, branching and pruning of stock values In this model, 75 % of the data is used for training the model and 25% of data is used for testing. The model provides results in a comparative analysis of trends of the particular stock in the past 25 days Decision Tree Regressor library is imported from sklearn.

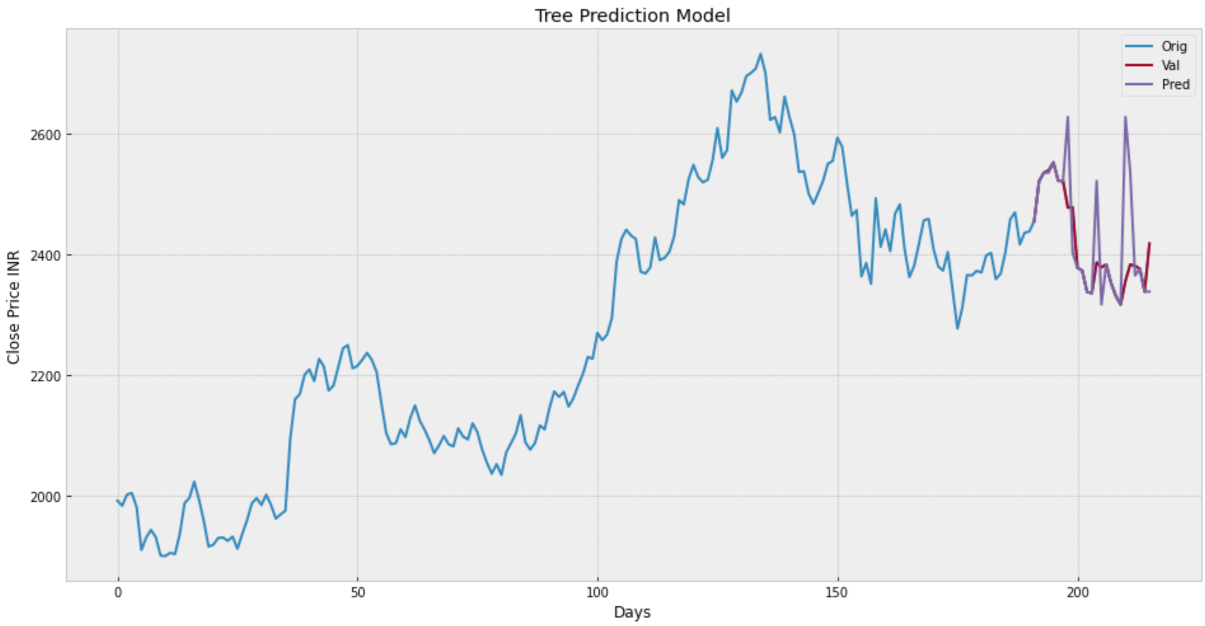
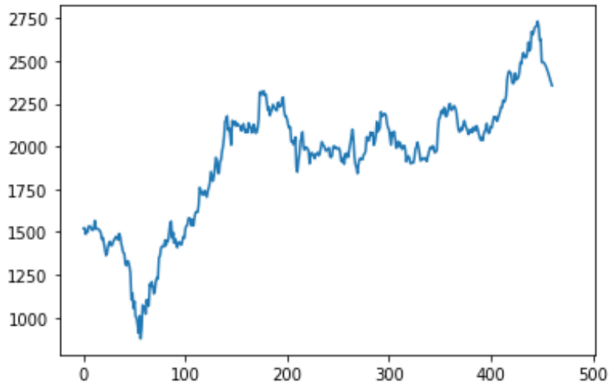



*Fig : Decision Tree Model Prediction for Reliance Stock*

## 5.3 Implementation of LSTM Model

LSTM Model is efficient for long term and time series data dependencies in data. This is achieved because the recurring module of the model has a combination of four layers interacting with each other. As for long-term time series prediction, the LSTM brings advantages of selecting the essential and relevant information hence enhancing the predictive accuracy. It is possibly due to the more complicated neural network representation causing an overfitting issue; the more significant number of parameters in the stacked-LSTM memory model does not generalize well in the unseen data.

In this model, 75 % of the data is used for training the model and 25% of data is used for testing The model provides results in a comparative analysis of trends of the particular stock for the 10 days LSTM Model library is imported from tensorflow.keras




## 5.4 Implementation of Sentiment Analysis

Sentiment Analysis determines the market response of the specific stock. We used a web scraper to extract information from Google Finance. The model summarizes the article and understands the emotion of the article The result is in the form of Positive, Neutral, Negative with a confidence score Implementation: We have run Sentiment Analysis on 'Reliance' stock and have deduced a result

```
In [55]: sentiment(summaries['RELIANCE'])

Out[55]: [{'label': 'NEGATIVE', 'score': 0.9359803795814514},
          {'label': 'POSITIVE', 'score': 0.9979088306427002},
          {'label': 'NEGATIVE', 'score': 0.9994615912437439},
          {'label': 'POSITIVE', 'score': 0.9393962025642395},
          {'label': 'NEGATIVE', 'score': 0.9864209890365601},
          {'label': 'POSITIVE', 'score': 0.9348860383033752},
          {'label': 'POSITIVE', 'score': 0.9986051917076111},
          {'label': 'NEGATIVE', 'score': 0.9723230004310608},
          {'label': 'NEGATIVE', 'score': 0.9953311085700989},
          {'label': 'NEGATIVE', 'score': 0.7673423886299133}]
```

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Ticker | Summary | Label | Confidence | URL |
| 2 | RELIANCE | Spread over 17.5 acres in Maker Maxity complex at Bandra Kurla. Reliance's new mall aims to fill a gap in Mumbai | NEGATIVE | 0.93598038 | https://finance.yahoo.com/news/india-jio-world-mall-gives-150017763.html |
| 3 | RELIANCE | We are aware of the issue and are working to resolve it. | POSITIVE | 0.997908831 | https://finance.yahoo.com/news/arbitration-panel-rejects-futures-plea-041247301.html |
| 4 | RELIANCE | Ex-billionaire Anil Ambani told a UK court his net worth was zero. ICIJ says it has obtained records of 18 offshore firms | NEGATIVE | 0.999461591 | https://finance.yahoo.com/news/anil-ambani-might-not-penniless-115340691.html |
| 5 | RELIANCE | Sharma cuts Reliance to neutral from buy, cites 'rich' valuations. | POSITIVE | 0.939396203 | https://finance.yahoo.com/news/reliance-industries-rating-cut-top-060532461.html |
| 6 | RELIANCE | Reliance International set up in UAE. India's Reliance plans to sell 20% stake in oil-to-chemical business to Aramco. | NEGATIVE | 0.986420989 | https://finance.yahoo.com/news/reliance-sets-subsidiary-uae-trading-123810949.html |
| 7 | RELIANCE | 7-Eleven stores to open in Mumbai on Oct. 9. Moderna drug to be given to people under 30 years old | POSITIVE | 0.934886038 | https://finance.yahoo.com/news/reliance-retail-launch-7-eleven-030327468.html |
| 8 | RELIANCE | Oil-to-retail giant is investing $10 billion in alternative energy. REC and Sterling & Wilson will help Reliance New Energy Solar expand globally | POSITIVE | 0.998605192 | https://finance.yahoo.com/news/billionaire-ambani-reliance-buys-rec-090930410.html |
| 9 | RELIANCE | Reliance Steel & Aluminum Co. was in 27 hedge funds' portfolios at the end of the second quarter. | NEGATIVE | 0.972323 | https://finance.yahoo.com/news/buy-reliance-steel-aluminum-co-175331293.html |
| 10 | RELIANCE | Third-quarter net income was $1.3 million, or $0.16 per diluted 800-273-3217 800-273-3217 800-273-3217 800-273-3217 800-273-3217. | NEGATIVE | 0.995331109 | https://finance.yahoo.com/news/first-reliance-bancshares-reports-third-123000322.html |
| 11 | RELIANCE | NexWafe is developing and producing monocrystalline silicon wafers grown directly from inexpensive raw materials. | NEGATIVE | 0.767342389 | https://finance.yahoo.com/news/correcting-replacing-reliance-energy-solar-170235913.html |

| Ticker | Summary | Label | Confidenc | URL |
|---|---|---|---|---|
| MTNL | TikTok user posts video in which unidentified flight attenc | NEGATIVE | 0.999143 | https://www.businessinsider.in/thelife/news/flight-attendant-reportedly- |
| MTNL | TCS, L&T Infotech among companies to report earnings thi | POSITIVE | 0.971559 | https://www.business-standard.com/article/markets/it-shares-in-demand |
| MTNL | Union says sale of mobile towers to corporates is beginnir | NEGATIVE | 0.996904 | https://www.newsclick.in/BSNL-Employees-Union-Protest-Monetisation-T |
| MTNL | Stocks to watch out for this week: Prasad. | NEGATIVE | 0.822156 | https://economictimes.indiatimes.com/markets/stocks/news/biggest-gain |
| MTNL | Over 92,000 employees on BSNL's rolls, over 13,500 in MTI | NEGATIVE | 0.948525 | https://www.newsclick.in/VRS-BSNL-MTNL-Non-retiring-Employees-Appre |
| MTNL | Againt fertilizer waste dumping | POSITIVE | 0.748121 | https://www.businessinsider.in/stock-market/news/shares-of-gujarat-stat |
| MTNL | Stocks to watch out for this week: Prasad. | NEGATIVE | 0.822156 | https://economictimes.indiatimes.com/markets/stocks/news/day-trading- |
| MTNL | We are aware of the issue and are working to resolve it. | POSITIVE | 0.997909 | https://finance.yahoo.com/news/india-telecoms-industry-report-2021-141 |
| MTNL | Finance ministry has decided to shut down BSNL and MTN | NEGATIVE | 0.99606 | https://www.newsclick.in/Finance-Ministry-Plans-Close-Down-BSNL-MTNL |
| MTNL | With a spate of Union Budget proposals, buzz has suddenl | NEGATIVE | 0.881286 | https://www.sentinelassam.com/editorial/budget-buzz-on-the-street-577 |
| | | | 0.918382 | |

| Ticker | Summary | Label | Confidence | URL |
|---|---|---|---|---|
| GLENMARK | We are aware of the issue and are working to resolve it. | POSITIVE | 0.997908831 | https://finance.yahoo.com/news/sanotize-inks-ag |
| GLENMARK | 46% of the market's growth will originate from North America. | POSITIVE | 0.992512584 | https://finance.yahoo.com/news/pruritus-drugs-n |
| GLENMARK | Says China Plus One is a real thing: CEO | POSITIVE | 0.99923563 | https://www.businessinsider.in/stock-market/nev |
| GLENMARK | Another 248,000 Americans filed new claims last week, more than the exp | POSITIVE | 0.987415731 | https://finance.yahoo.com/news/aerosol-delivery |
| GLENMARK | We are aware of the issue and are working to resolve it. | POSITIVE | 0.997908831 | https://finance.yahoo.com/news/glenmark-sanoti |
| GLENMARK | Global topical drug delivery market is projected to reach at a market value | NEGATIVE | 0.914623439 | https://finance.yahoo.com/news/topical-drug-del |
| GLENMARK | We are aware of the issue and are working to resolve it. | POSITIVE | 0.997908831 | https://finance.yahoo.com/news/vaneck-india-gr |
| GLENMARK | We are aware of the issue and are working on it. | POSITIVE | 0.748121142 | https://www.businessinsider.in/stock-market/nev |
| GLENMARK | Employee is working hard to get metformin hydrochloride tablets market | POSITIVE | 0.997908831 | https://finance.yahoo.com/news/metformin-hydr |
| GLENMARK | Global drugs market is expected to grow from $24.68 billion in 2021 to $26 | POSITIVE | 0.98520875 | https://finance.yahoo.com/news/gynecology-drug |
| | | | 0.96187526 | |

## 5.5 Implementation of Predicting Models on Stock Domains

A. Glenmark

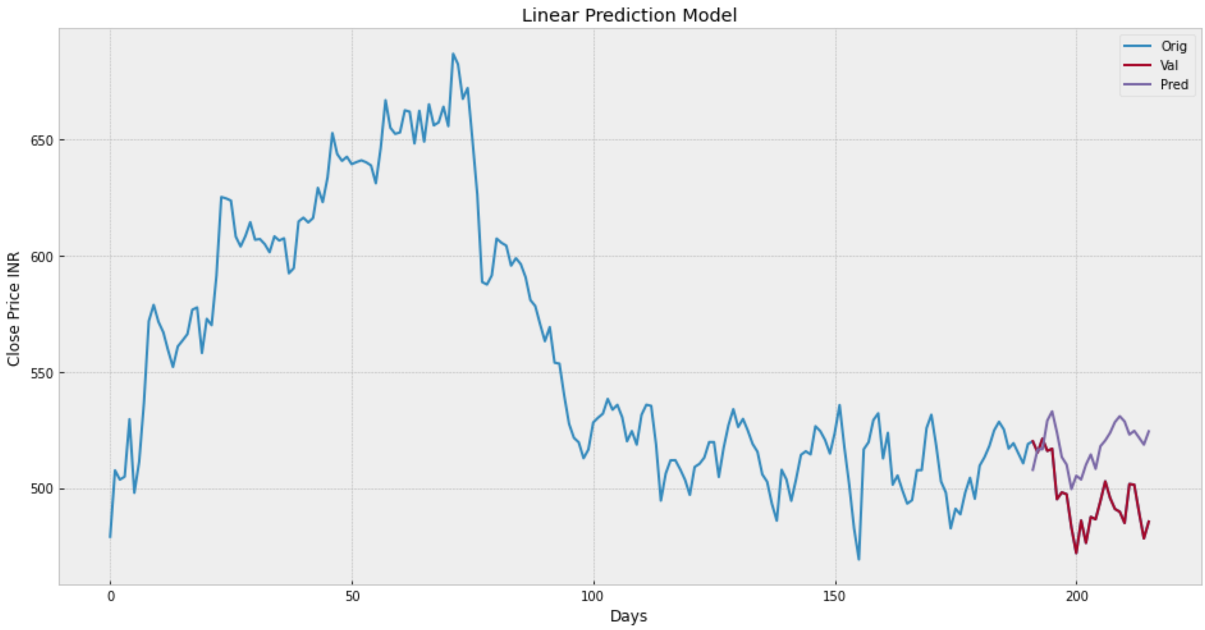



*Fig __ : Linear Prediction Model for Glenmark*

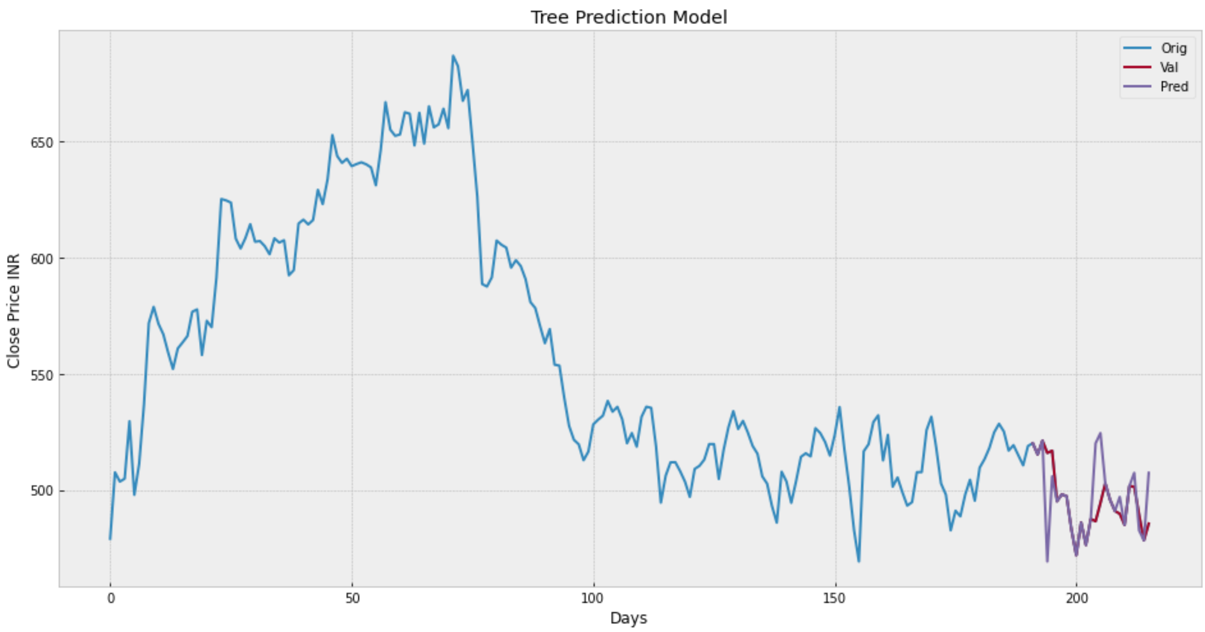



*Fig __ : Tree Prediction Model for Glenmark*

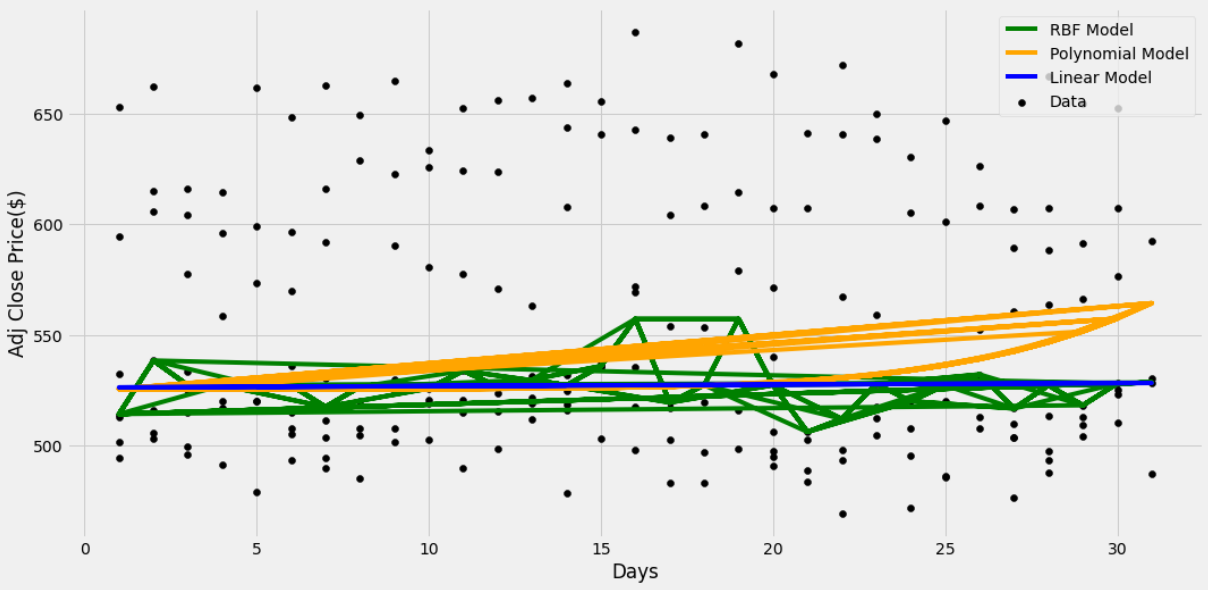



*Fig __ : SVM Prediction Model for Glenmark*

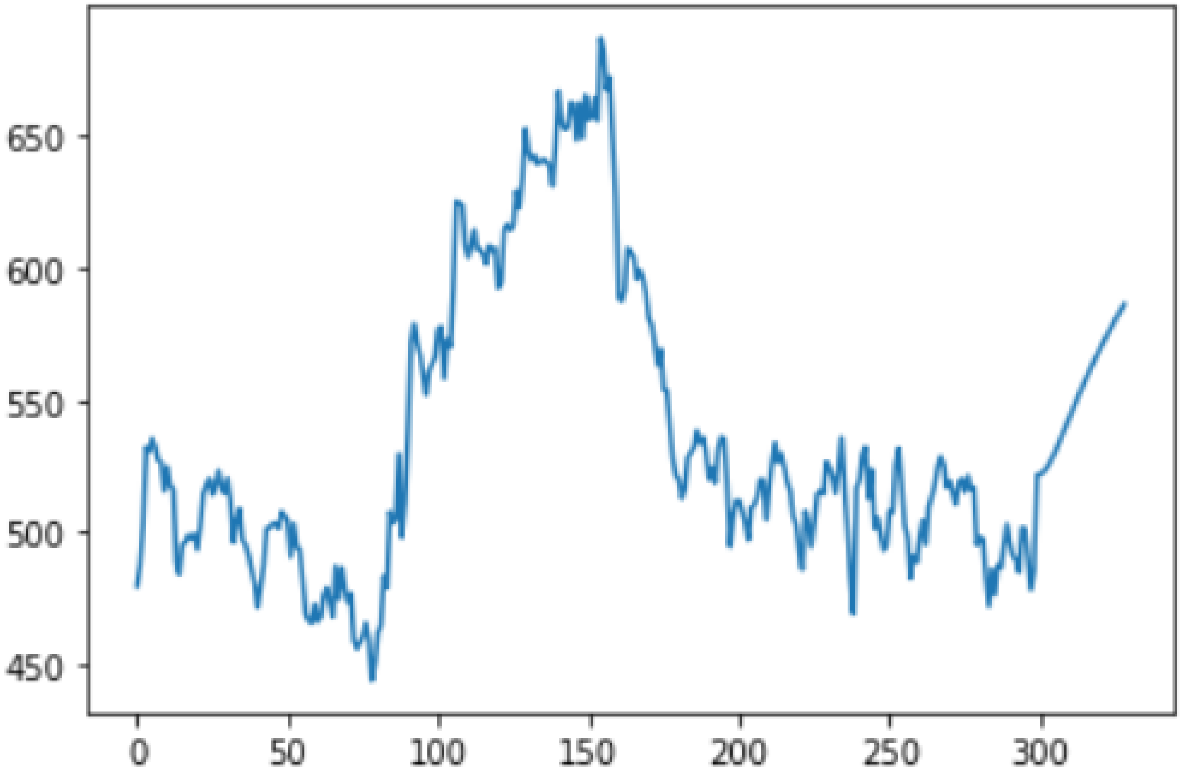



*Fig __ : LSTM Prediction Model for Glenmark*

B. ICICI

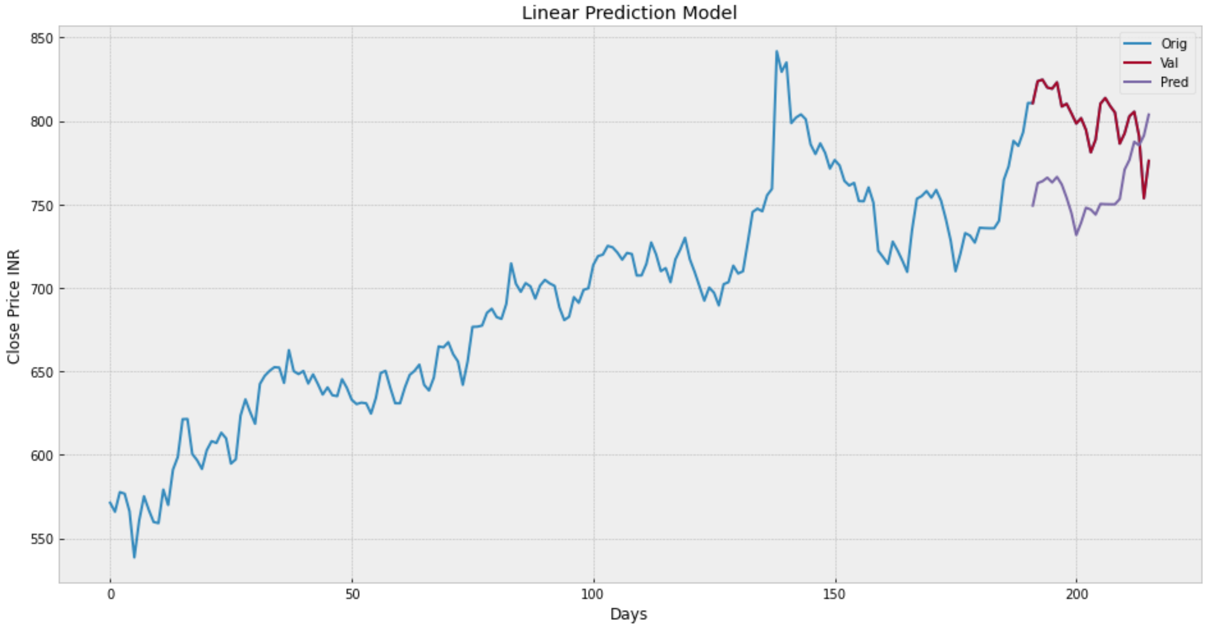



*Fig __ : Linear Prediction Model for ICICI*

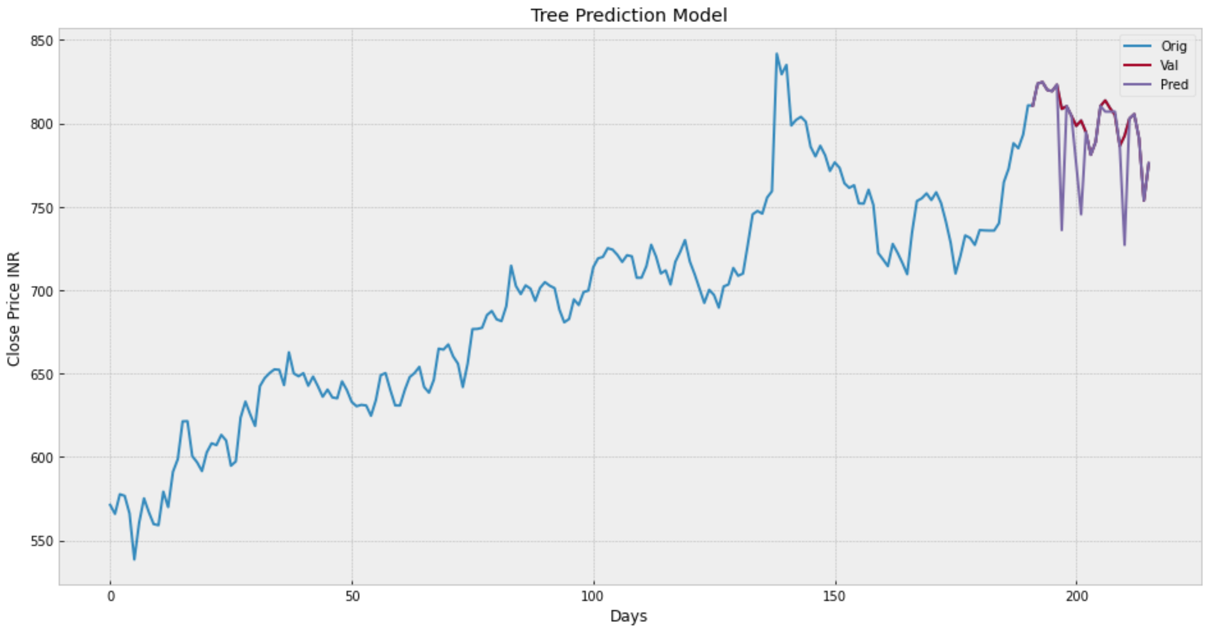



*Fig __ : Tree Prediction Model for ICICI*

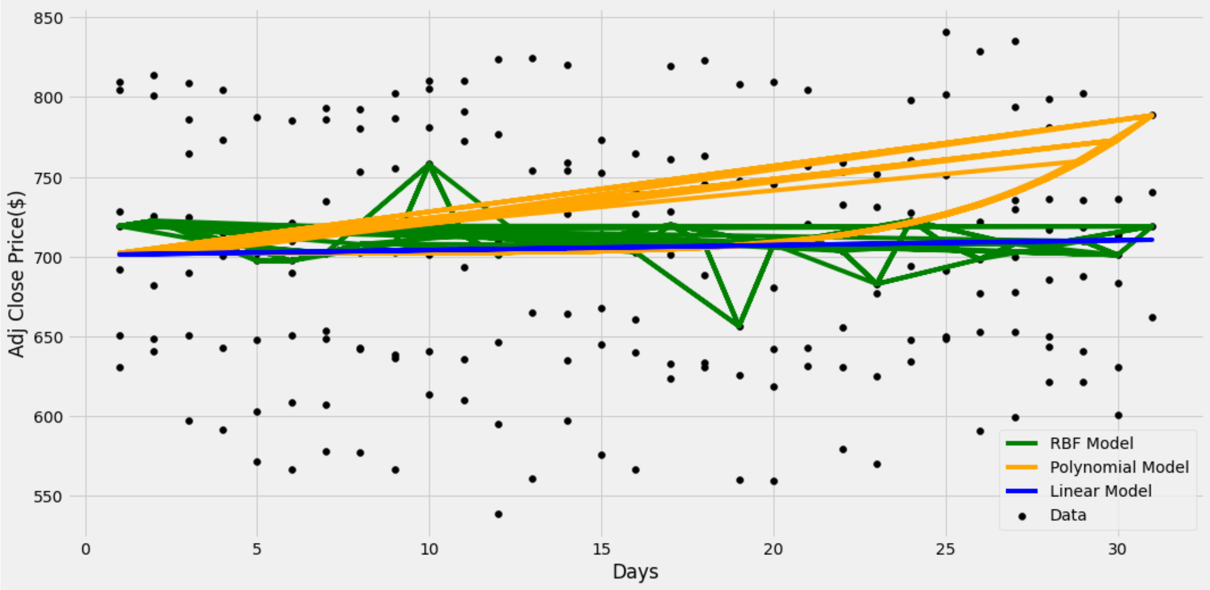



*Fig __ : SVM Prediction Model for ICICI*

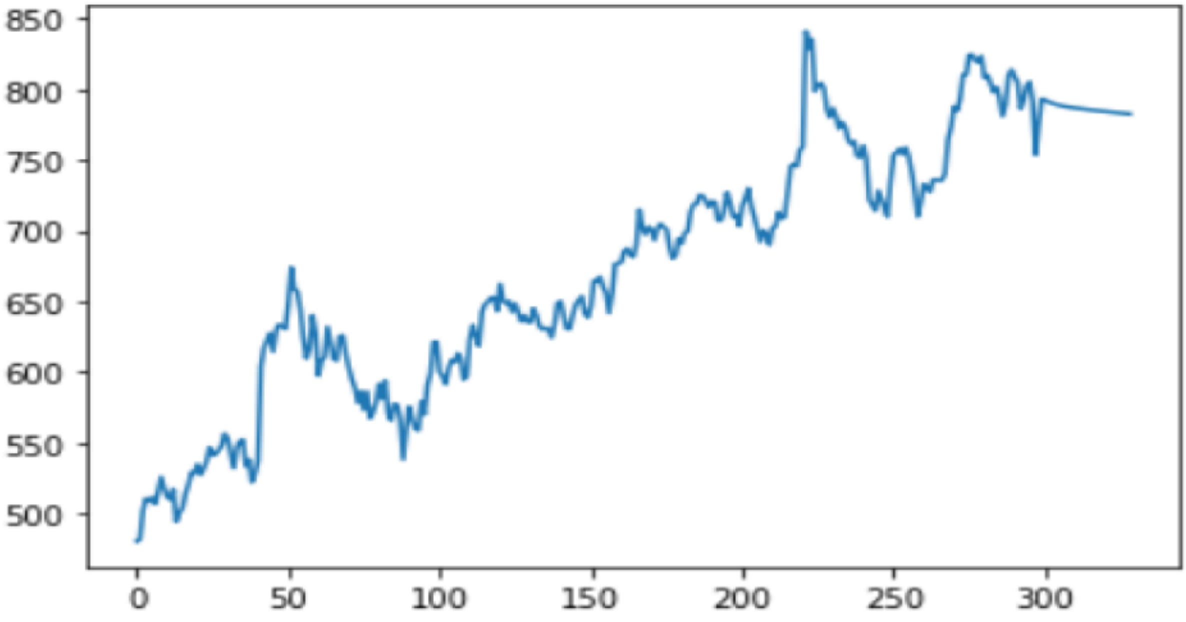



*Fig __ : LSTM Prediction Model for ICICI*

## C. Reliance

## 5.6 Sentiment Analysis

## 5.7 Comparative Analysis

A. Pharmacticual Industry

1. Cipla

| Name of Model | Current Value | Predicted Value | Sentiment | Confidence |
|---|---|---|---|---|
| Linear Regression | 921.85 | 927 | Positive | 0.895834863 |
| Decision Tree | 921.85 | 924 | Positive | 0.895834863 |
| LSTM | 921.85 | 948 | Positive | 0.895834863 |
| Linear SVM | 921.85 | 959.4590 | Positive | 0.895834863 |
| Polynomial SVM | 921.85 | 1252.5093 | Positive | 0.895834863 |
| RBF | 921.85 | 926.3960 | Positive | 0.895834863 |

2. Glenmark

| Name of Model | Current Value | Predicted Value | Sentiment | Confidence |
|---|---|---|---|---|
| Linear Regression | 485.55 | 524 | Positive | 0.96187526 |
| Decision Tree | 485.55 | 501 | Positive | 0.96187526 |
| LSTM | 485.55 | 520 | Positive | 0.96187526 |
| Linear SVM | 485.55 | 540.5999 | Positive | 0.96187526 |
| Polynomial SVM | 485.55 | 411.9913 | Positive | 0.96187526 |
| RBF | 485.55 | 527.1616 | Positive | 0.96187526 |

3. Pfizer

| Name of Model | Current Value | Predicted Value | Sentiment | Confidence |
|---|---|---|---|---|
| Linear Regression | 4385.8 | 4972 | Neutral | 0.897523654 |
| Decision Tree | 4385.8 | 5002 | Neutral | 0.897523654 |
| LSTM | 4385.8 | 4810 | Neutral | 0.897523654 |
| Linear SVM | 4385.8 | 4210.5833 | Neutral | 0.897523654 |
| Polynomial SVM | 4385.8 | 4816.8421 | Neutral | 0.897523654 |
| RBF | 4385.8 | 4286.9119 | Neutral | 0.897523654 |

B. Banking Industry

1. Housing Development Finance Corporation Limited (HDFC)

| Name of Model | Current Value | Predicted Value | Sentiment | Confidence |
|---|---|---|---|---|
| Linear Regression | 1517.8 | 1539 | Positive | 0.961340624 |
| Decision Tree | 1517.8 | 1522 | Positive | 0.961340624 |

| | | | | |
|---|---|---|---|---|
| LSTM | 1517.8 | 1490 | Positive | 0.961340624 |
| Linear SVM | 1517.8 | 1349.0462 | Positive | 0.961340624 |
| Polynomial SVM | 1517.8 | 1388.1945 | Positive | 0.961340624 |
| RBF | 1517.8 | 1507.2504 | Positive | 0.961340624 |

2. Industrial Credit and Investment Corporation of India (ICICI)

| **Name of Model** | **Current Value** | **Predicted Value** | **Sentiment** | **Confidence** |
|---|---|---|---|---|
| Linear Regression | 776.05 | 804 | Positive | 0.938085347 |
| Decision Tree | 776.05 | 775 | Positive | 0.938085347 |
| LSTM | 776.05 | 786 | Positive | 0.938085347 |
| Linear SVM | 776.05 | 707.6871 | Positive | 0.938085347 |
| Polynomial SVM | 776.05 | 860.5156 | Positive | 0.938085347 |
| RBF | 776.05 | 767.3937 | Positive | 0.938085347 |

3. State Bank of India (SBI)

| **Name of Model** | **Current Value** | **Predicted Value** | **Sentiment** | **Confidence** |
|---|---|---|---|---|
| Linear Regression | 524.8 | 500 | Positive | 0.780220318 |
| Decision Tree | 524.8 | 459 | Positive | 0.780220318 |
| LSTM | 524.8 | 480 | Positive | 0.780220318 |

| | | | | |
|---|---|---|---|---|
| Linear SVM | 524.8 | 429.2000 | Positive | 0.780220318 |
| Polynomial SVM | 524.8 | 415.8842 | Positive | 0.780220318 |
| RBF | 524.8 | 442.8431 | Positive | 0.780220318 |

C. Information Technology Industry

1. Tata Consultancy Services (TCS)

| **Name of Model** | **Current Value** | **Predicted Value** | **Sentiment** | **Confidence** |
|---|---|---|---|---|
| Linear Regression | 8.78 | 11.6 | Positive | 0.953021818 |
| Decision Tree | 8.78 | 8.9 | Positive | 0.953021818 |
| LSTM | 8.78 | 11.3 | Positive | 0.953021818 |
| Linear SVM | 8.78 | 11.2782 | Positive | 0.953021818 |
| Polynomial SVM | 8.78 | 15.2829 | Positive | 0.953021818 |
| RBF | 8.78 | 8.1699 | Positive | 0.953021818 |

2. Infosys

| **Name of Model** | **Current Value** | **Predicted Value** | **Sentiment** | **Confidence** |
|---|---|---|---|---|
| Linear Regression | 22.19 | 24 | Positive | 0.860656768 |
| Decision Tree | 22.19 | 22.3 | Positive | 0.860656768 |

| | | | | |
|---|---|---|---|---|
| LSTM | 22.19 | 23 | Positive | 0.860656768 |
| Linear SVM | 22.19 | 21.3739 | Positive | 0.860656768 |
| Polynomial SVM | 22.19 | 13.2396 | Positive | 0.860656768 |
| RBF | 22.19 | 22.7019 | Positive | 0.860656768 |

3. Oracle Finance (OFSS)

| Name of Model | Current Value | Predicted Value | Sentiment | Confidence |
|---|---|---|---|---|
| Linear Regression | 3684.4 | 4124 | Neutral | 0.874268109 |
| Decision Tree | 3684.4 | 3742 | Neutral | 0.874268109 |
| LSTM | 3684.4 | 3850 | Neutral | 0.874268109 |
| Linear SVM | 3684.4 | 3411.6843 | Neutral | 0.874268109 |
| Polynomial SVM | 3684.4 | 4082.7042 | Neutral | 0.874268109 |
| RBF | 3684.4 | 3783.1223 | Neutral | 0.874268109 |

D. Automobile Industry

1. 10.Tata Motors

| Name of Model | Current Value | Predicted Value | Sentiment | Confidence |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Linear Regression | 504 | 500 | Positive | 0.917204666 |
| Decision Tree | 504 | 507 | Positive | 0.917204666 |
| LSTM | 504 | 498 | Positive | 0.917204666 |
| Linear SVM | 504 | 446.9214 | Positive | 0.917204666 |
| Polynomial SVM | 504 | 400.5352 | Positive | 0.917204666 |
| RBF | 504 | 442.9382 | Positive | 0.917204666 |

2. TVS Motor Company Limited

| **Name of Model** | **Current Value** | **Predicted Value** | **Sentiment** | **Confidence** |
|---|---|---|---|---|
| Linear Regression | 663.3 | 624 | Positive | 0.891178614 |
| Decision Tree | 663.3 | 659 | Positive | 0.891178614 |
| LSTM | 663.3 | 634 | Positive | 0.891178614 |
| Linear SVM | 663.3 | 585.0153 | Positive | 0.891178614 |
| Polynomial SVM | 663.3 | 523.7746 | Positive | 0.891178614 |
| RBF | 663.3 | 612.0200 | Positive | 0.891178614 |

3. Bajaj Motors

| **Name of Model** | **Current Value** | **Predicted Value** | **Sentiment** | **Confidence** |
|---|---|---|---|---|
| Linear Regression | 3589.2 | 3524 | Positive | 0.926721853 |

| | | | | |
|---|---|---|---|---|
| Decision Tree | 3589.2 | 3600 | Positive | 0.926721853 |
| LSTM | 3589.2 | 3603 | Positive | 0.926721853 |
| Linear SVM | 3589.2 | 3705.0909 | Positive | 0.926721853 |
| Polynomial SVM | 3589.2 | 3906.4202 | Positive | 0.926721853 |
| RBF | 3589.2 | 3772.2087 | Positive | 0.926721853 |

E. Telecommunication Industry

1. Mahanagar Telephone Nigam Limited (MTNL)

| **Name of Model** | **Current Value** | **Predicted Value** | **Sentiment** | **Confidence** |
|---|---|---|---|---|
| Linear Regression | 25.45 | 28.1 | Negative | 0.9183819 |
| Decision Tree | 25.45 | 24.7 | Negative | 0.9183819 |
| LSTM | 25.45 | 24 | Negative | 0.9183819 |
| Linear SVM | 25.45 | 13.8714 | Negative | 0.9183819 |
| Polynomial SVM | 25.45 | 33.8046 | Negative | 0.9183819 |
| RBF | 25.45 | 19.7923 | Negative | 0.9183819 |

2. Reliance

| **Name of Model** | **Current Value** | **Predicted Value** | **Sentiment** | **Confidence** |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Linear Regression | 2417.95 | 2421 | Negative | 0.975703311 |
| Decision Tree | 2417.95 | 2428 | Negative | 0.975703311 |
| LSTM | 2417.95 | 2250 | Negative | 0.975703311 |
| Linear SVM | 2417.95 | 2305.0722 | Negative | 0.975703311 |
| Polynomial SVM | 2417.95 | 1565.9586 | Negative | 0.975703311 |
| RBF | 2417.95 | 2394.9978 | Negative | 0.975703311 |

3. Airtel

| Name of Model | Current Value | Predicted Value | Sentiment | Confidence |
|---|---|---|---|---|
| Linear Regression | 711.55 | 712.36 | Positive | 0.787999946 |
| Decision Tree | 711.55 | 713 | Positive | 0.787999946 |
| LSTM | 711.55 | 750 | Positive | 0.787999946 |
| Linear SVM | 711.55 | 646.0702 | Positive | 0.787999946 |
| Polynomial SVM | 711.55 | 635.2810 | Positive | 0.787999946 |
| RBF | 711.55 | 700.7368 | Positive | 0.787999946 |

# Chapter 6

# Conclusion and Future Work

## 6.2 Challenges and Solution

Once we selected our topic Efficacy of Predicting Models in Bear and Bull Stock Run, we searched the already implemented projects under this domain. We wanted to identify the efficacy and usability of Machine Learning, the challenging task of predicting the stock market prices. The existing studies mainly considered non-Indian companies and outdated data. Our goal was to find a reliable algorithm that provides an approximate trend in future stock prices of some of the top Indian companies.

The challenges we faced during the project making process are briefly explained below, along with our approach to dealing with them.

We could easily obtain recent data from the Bombay Stock Exchange (BSE) website,but it contained many features such as Open Price, High Price, Low Price, Close Price, Weighted Average Price, Number of Shares, Number of Trades, etc. Since the stock market relies on many factors, majorly on human emotion and news, we had to determine which of these features impact our prediction accuracy. Also, we had to think about how many days in the future we should predict with minimal error.

Solution: We surveyed several existing papers and learnt that the only feature from the dataset that affects the Closing Price is the previous Closing Price. Since we had only one variable to train our models on, it was difficult to obtain accuracy in linear regression models. So we scaled our dataset using MinMaxScaler. We experimented with the future number of days and realized that all our models do best when predicting prices of 60 days in the future.

There are many Machine Learning algorithms which have been implemented over the years. Deciding on which algorithms to be used in the prediction process was also a challenge. We had to research and study each algorithm to determine the best algorithm for our time-series financial datasets.

Solution: We decided that we will be including LSTM and ARIMA as they are known to work well with time-series datasets. For the remaining, we applied seven existing algorithms and picked the three that performed comparatively more decently, i.e. SVM, KNN and Random Forest. We noticed that Decision Tree was prone to overfitting so we discarded it. Linear Regression was far too simple for a complicated problem statements like this, so we discarded that as well.

We noticed that the companies we had chosen as datasets had undergone a stock split which was leading to inaccuracy in the forecasting. We had to adjust the split to have a well-defined graph to train our models.

Solution: To deal with splits, we had two options: To drop all the data before the split or fix it by dividing the data before the split by the ratio of the split. We did not want to lose the majority of our data as the primary purpose of this study was research on long-term data to the present data. So we adjusted the split and our models improved significantly.

## 6.2 Conclusion

Efficacy of Predicting Models in Bear and Bull Stock Run works on the BSE stocks and uses the stock prices to generate datasets. The project has provided a fair idea about the complexity of the market, and the environmental factors that have a drastic effect on the future predictions. We noted the factors affecting the sentiment of the market and used the models to derive variation on these predictions to deduce the most effective model. The overall result generated from these datasets have been noted and documented.

# References


[1] M. S. Hegde, G. Krishna, and R. Srinath, “An Ensemble Stock Predictor and Recommender System,” in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.

[2] M. Bucurica, R. Dogaru, and I. Dogaru, “A comparison of Extreme Learning Machine and Support Vector Machine classifiers,” in 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), 2015.

[3] D. Bhuriya, G. Kaushal, A. Sharma, and U. Singh, “Stock market prediction using a linear regression,” in 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 2017.

[4] N. Zainuddin and A. Selamat, “Sentiment analysis using Support Vector Machine,” in 2014 International Conference on Computer, Communications, and Control Technology (I4CT), 2014.

[5] Researchgate.net. [Online]. Available: https://www.researchgate.net/publication/46280096_A_Decision_Tree-_Rough_Set_Hybrid_System_for_Stock_Market_Trend_Prediction. [Accessed: 29-Apr-2021].

[6] A. Wilson, “Stock Prediction Using Linear Regression - Analytics Vidhya - Medium,” Analytics Vidhya, 06-Dec-2020. [Online]. Available: https://medium.com/analytics-vidhya/stock-prediction-using-linear-regression-cd1d8351f536. [Accessed: 29-Apr-2021].

[7] J. Briggs, “Sentiment analysis for stock price prediction in python,” Towards Data Science, 04-Dec-2020. [Online]. Available: https://towardsdatascience.com/sentiment-analysis-for-stock-price-prediction-in-python-bed40c65d178. [Accessed: 29-Apr-2021].

[8] “Support Vector Machine (SVM) Algorithm - Javatpoint,” Javatpoint.com. [Online]. Available: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm. [Accessed: 29-Apr-2021].

[9] A. Yadav, “SUPPORT VECTOR MACHINES(SVM) - Towards Data Science,” Towards Data Science, 20-Oct-2018. [Online]. Available: https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589. [Accessed: 29-Apr-2021].

[10] S. Sreenivasa, “Radial basis function (RBF) kernel: The go-to kernel,” Towards Data Science, 12-Oct-2020. [Online]. Available: https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a. [Accessed: 29-Apr-2021].