STOCK MARKET PREDICTION USING MACHINE LEARNING ALGORITHMS

CREATIVE AND INNOVATIVE PROJECT

*S*ubmitted by

SANJANA R 2017115584

SNEHASHREE R 2017115595

MONICA JAWAHAR 2017115619

Submitted to the Faculty of

INFORMATION SCIENCE AND TECHNOLOGY

For completion of Creative and Innovative Project for Third Years

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING, GUINDY

CHENNAI - 600025

MARCH 2020

ANNA UNIVERSITY CHENNAI

CHENNAI – 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled STOCK MARKET PREDICTION USING MACHINE LEARNING ALGORITHM is the bonafide work of SANJANA R(2017115584), SNEHASHREE R(20171195), MONICA JAWAHAR (2017115619) who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE: CHENNAI
DATE:

DR. S. BAMA, ASSISTANT PROFESSOR
MS. TINA ESTHER TRUEMAN, TEACHING FELLOW
PROJECT GUIDES
DEPARTMENT OF IST, CEG
ANNA UNIVERSITY
CHENNAI 600025

COUNTERSIGNED

DR. SASWATI MUKHERJEE
HEAD OF THE DEPARTMENT
DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY
COLLEGE OF ENGINEERING, GUINDY
ANNA UNIVERSITY
CHENNAI 600025

ABSTRACT

In Stock Market Prediction, the aim is to predict the future values of the financial stocks of a company. The recent trend in stock market prediction technologies is the use of machine learning which makes predictions based on the values of the current stock market indices by training on their previous values. Machine learning itself employs different models to make prediction easier and authentic. This project focuses on use of Regression and SVM based Machine learning to predict stock values. Factors considered are open, close, low, high, last, total trade quantity and turnover. The successful prediction of stock will be the great asset for the stock market institutions and will provide real-life solutions to the problems that stock investors face.

ACKNOWLEDGEMENT

First and foremost, we would like to express our deep sense of gratitude to our guide Dr. SASWATI MUKHERJEE, Professor and Head, Department of Information Science and Technology, Anna University, for her excellent guidance, counsel, continuous support and patience. She has helped us to come up with this topic and guided us in the development of this project. She gave us the moral support and freedom to finish our mini project in a successful manner. We also would like to thank her for her kind support and for providing necessary facilities to carry out the work.

We are thankful to the project committee members **Dr. S. BAMA** Assistant Professor, **Ms. TINA ESTHER TRUEMAN** Teaching Fellow, Department of Information Science and Technology, Anna University, Chennai, for their valuable guidance and technical support.

We express our heartiest thanks to all other teaching and non teaching staff those who have helped us in one way or other for the successful completion of the project. Last but not the least we would like to thank our parents, friends for their indirect contribution to the successful completion of my project.

SANJANA R

SNEHA SHREE R

MONICA JAWAHAR

Table of contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML	Machine Learning
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

The Stock market is basically an aggregation of various buyers and sellers of stock. The prediction is expected to be robust, accurate and efficient. The system must work according to the real life scenarios and should be suited to real world settings. Machine learning involves artificial intelligence which empowers the system to learn and improve from past experiences without being programmed time and again. The stock price movement over a long period of time usually develops a linear curve. People tend to buy those stocks whose prices are expected to rise in the near future. The uncertainty in the stock market refrain people from investing in stocks. Thus, there is a need to accurately predict the stock market which can be used in a real-life scenario. The methods used to predict the stock market includes a time series forecasting along with technical analysis, machine learning, modelling and predicting the variable stock market. The datasets of the stock market prediction model include details like the closing price, opening price, the data and various other variables that are needed to predict the object variable which is the price in a given day. By using Linear, Logistic Regression and SVM algorithms to predict the stock value.

1.1 STOCK MARKET PREDICTION

- Stock market forecasting is the act of demanding to conclude the future price of a company stock or other financial instrument traded on an exchange. The successful forecast of a stock's future value might give up important profit.
- The efficient-market hypothesis suggests that stock prices reveal all currently existing information and any price changes that are not based on newly exposed information thus are intrinsically unpredictable.
- Stock price prediction is possible by MACHINE LEARNING ALGORITHMS.
- Machine learning involves artificial intelligence which empowers the system to learn and improve from past experiences without being programmed time and again.
- The future prices of the stock is predicted by Support Vector Machines(SVM) and Regression models.

1.2 PROBLEM STATEMENT

- Problem statement of Stock market is very vast and difficult to understand. It is considered too uncertain to be predictable due to huge fluctuation of the market.
- Stock market prediction task is interesting as well as divides researchers and academics into two groups, those who believe that we can devise

mechanisms to predict the market and those who believe that the market is efficient and whenever new information comes up the market absorbs it by correcting itself, thus there is no space for prediction.

- Investing in a good stock but at a bad time cam have disastrous result, while investing in a stock at the right time can bear profits.
- Financial investors of today are facing this problem of trading as they do not properly understand as to which stocks to buy or which stocks to sell in order to get optimum result.
- So, the purposed project will reduce the problem with suitable accuracy faced in such real time scenario.

1.3 PROCESS INVOLVED

- With the raw data from the dataset and the feature is extracted manually.
- After this process, the data obtained is trained and the data is tested simultaneously.
- Finally, from the trained data result is obtained from the testing phase.
- The output of the desired model is obtained which is the predicted value of the particular model.

1.4 OBJECTIVE

The aims of the project are as follows:

- To identify the factors that affect the share market.

- To generate the patterns from large set of data of stock market for prediction.
- To predict an approximate value of share price.
- To provide analysis for users through web application.

1.5 PROBLEMS/ISSUES FACED

- There was a minor problem in reading the file with dataset.
- In Logistic regression model, this been a major challenge in predicting the stock value. Also, the plot graph representation is also challenging one when undergoing this process there's some blunder in testing part result.
- Later on, result is rectified properly.
- In SVM model, there is mishap occurs while reading from the dataset.

CHAPTER 2

LITERATURE SURVEY

2. 1 SURVEY OF STOCK MARKET PREDICTION

The stock market prediction has become an increasingly important issue in the present time. One of the methods employed is technical analysis, but such methods do not always yield accurate results. So it is important to develop methods for a more accurate prediction. The technique that was employed in this instance was a regression. Since financial stock marks generate enormous amounts of data at any given time a great volume of data needs to undergo analysis before a prediction can be made. One of the noteworthy techniques that were mentioned was linear regression. The way linear regression models work is that they are often fitted using the least squares approach, but they may alternatively be also be fitted in other ways, such as by diminishing the "lack of fit" in some other norm, or by diminishing a handicapped version of the least squares loss function.

2.2 REVIEW ON MACHINE LEARNING TECHNIQUES FOR STOCK MARKET PREDICTION

Machine learning techniques intend to consequently learn and perceive patterns in huge information. There are large numbers of well-known machine learning algorithms that can be utilized to categorize an issue given a set of

peculiarities. In this area some of these algorithms that are especially utilized as a part of classifying stock market information into "up" or "down" periods given a set of inputs originated through macro-economic information and technical analysis has been presented.

2.2.1 APPROACH FOR SVM

(Lin et al,2013) propose a SVM based stock market forecast system .This system chooses a decent feature subset, assesses stock indicator and control over fitting on stock market inclination anticipation. They tried this methodology on Taiwan stock market datasets and found that this system performs well than the traditional stock market forecast system.

2.2.2 APPROACH FOR LOGISTIC REGRESSION

Li et al. [2010] used LR as a comparative method in order to build a better model for predicting stock returns effectively and efficiently. A 30 times holdout method was used in the assessment, along with the two commonly used methods in the top 10 data mining algorithms (the support vector machine and k nearest neighbour) and the two baseline benchmark methods from the statistical area (MDA and LR).

CHAPTER 3

METHODOLOGIES

Regression analysis is used to determine the magnitude of relationships between variables as well as to model relationships between variables and for predictions based on the models. 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. Thus, the data needed for regression are two part, first section for defining model and the other for testing model. In this section we choose linear regression for our analysis. First, we divide the data into two parts of training and testing. Then we use the training section for starting analysis and defining the model.

3.1 LINEAR REGRESSION

- **Linear regression** is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).
- Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable.
- However, a number of non-linear techniques could be used to obtain a more accurate regression if the relationship between variables is not linear in parameters.

- If the goal is prediction, linear regression can be used to fit a predictive model to an observed data set of y and X values.
- After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted.

3.2 SUPPORT VECTOR MACHINE(SVM)

- The main task of the support machine algorithm is to identify an N-dimensional space that distinguishably categorizes the data points.
- Here, N stands for a number of features. Between two classes of data points, there can be multiple possible hyperplanes that can be chosen.
- The objective of this algorithm is to find a plane that has maximum margin. Maximizing margin refers to the distance between data points of both classes.
- The benefit associated with maximizing the margin is that it provides is that it provides some reinforcement so that future data points can be more easily classified.
- Decision boundaries that help classify data points are called hyperplanes.
- Based on the position of the data points relative to the hyperplane they are attributed to different classes.
- The dimension of the hyperplane relies on the number of attributes, if the number of attributes is two then the hyperplane is just a line, if the number of attributes is three then the hyperplane is two dimensional.

- Here by using features like Date, Open, Close, High, Low and Turnover SVM is concluded with input and output values.

3.3 LOGISTIC REGRESSION

- Logistic regression (LR), which is helpful for predicting the presence or absence of a characteristic or outcome based on values of a set of predictor variables, is a multivariate analysis model.
- Logistic regression has the advantage of being less affected than discriminant analysis when the normality of the variable cannot be assumed. It has the capacity to analyse a mix of all types of predictors.
- Logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these.
- In instances where the independent variables are categorical, or a mix of continuous and categorical, logistic regression is preferred.
- Regression uses a more complex cost function, this cost function can be defined as the '**Sigmoid function**' or also known as the 'logistic function'.
- Since the probability of an event must lie between 0 and 1, it is unrealistic to model probabilities with linear regression techniques, because the linear regression model allows the dependent variable to take values greater than 1 or less than 0.
- The logistic regression model is a type of generalized linear model that extends the linear regression model by linking the range of real numbers to the 0-1 range.

CHAPTER 4

SYSTEM ARCHITECTURE

Kaggle is an online community for data analysis and predictive modelling. It allows the users to use their datasets so that they can build models and work with various data science engineers to solve various real-life data science challenges. The dataset used in the proposed project has been downloaded from Kaggle. However, this data set is present in what we call raw format. The data set is a collection of stock market information about a companies .

4.1 DATASET DESCRIPTION

Dataset is taken from the Kaggle website , Sample data obtained from Tata company under Beverages. Though it includes stocks from different industries has some distinguishing characteristics such as sensitivity, predictability, scalability, to name a few. We have used all the indices for future calculations. A sample of data obtained is shown in the table. Date, Open, High, Low, Last, Close, Total Trade Quantity and Turnover.

DATE	OPEN	HIGH	LOW	LAST	CLOSE	TTQTY	TURNOVER IN LAKHS
08-10-18	208.00	222.25	206.85	216.00	215	4642146	10062.83
05-10-18	217	218.5	205.5	210.25	209.05	3519515	7407.06
04-10-18	223.3	227	216	217	218	1728786	3815.79
01-10-18	234	234.5	221.8	230	230	1534749	3486.50
03-10-18	230	237	225	226	227.6	1708590	3960.27
27-9-18	234	236	231	233	233	5082859	11859.95
26-9-18	240	240	232.5	235	234.5	2240909	5248.6
25-9-18	233	236	232	236	236	2349368	5503.90
24-9-18	233.50	239.20	230	234	233	3423509	7999.55

Table 4.1 Sample data

4.2 DATA NORMALISATION

There are a couple important details to note about the way the data must be pre- processed in order to be fit into regression models. Firstly, dates are normally represented as strings of the format “YYYY-MM-DD” when it comes to database storage. This format must be converted to a single integer in order to be used as a column in the feature matrix. This is done by using the date’s ordinal value.

4.3 FEATURES

Stock market close price is an important piece of information that is very useful for every short-term trader. The close prices are very important, especially for swing traders and position traders. It also has implications for practical day trading in many day trading systems. The stock market close price level provides very important information about the general mood of investors. It tells a lot about the thinking of big investors that allocate large amount of money into the stock market for their asset management purposes.

In total there 8 columns including Date, Open, High, Low, Last, Close, Total Trade Quantity, Turnover(Lacs). Now explore each specific column:

Date — Our dataset date ranges from Oct’18 to Aug’18.

Open — It represents the price at which the stock started trading on a particular date.

Close — It represents the price at which the stock closed on a particular date.

High — It represents the maximum price stock encounter on a particular date.

Low — It represents the minimum price stock encounter on a particular date.

Total Trade Quantity — The total no. of shares/stocks bought or sold on a particular date.

Turnover (Lacs) — The total sales generated in a single day.

4.4 SYSTEM DESIGN

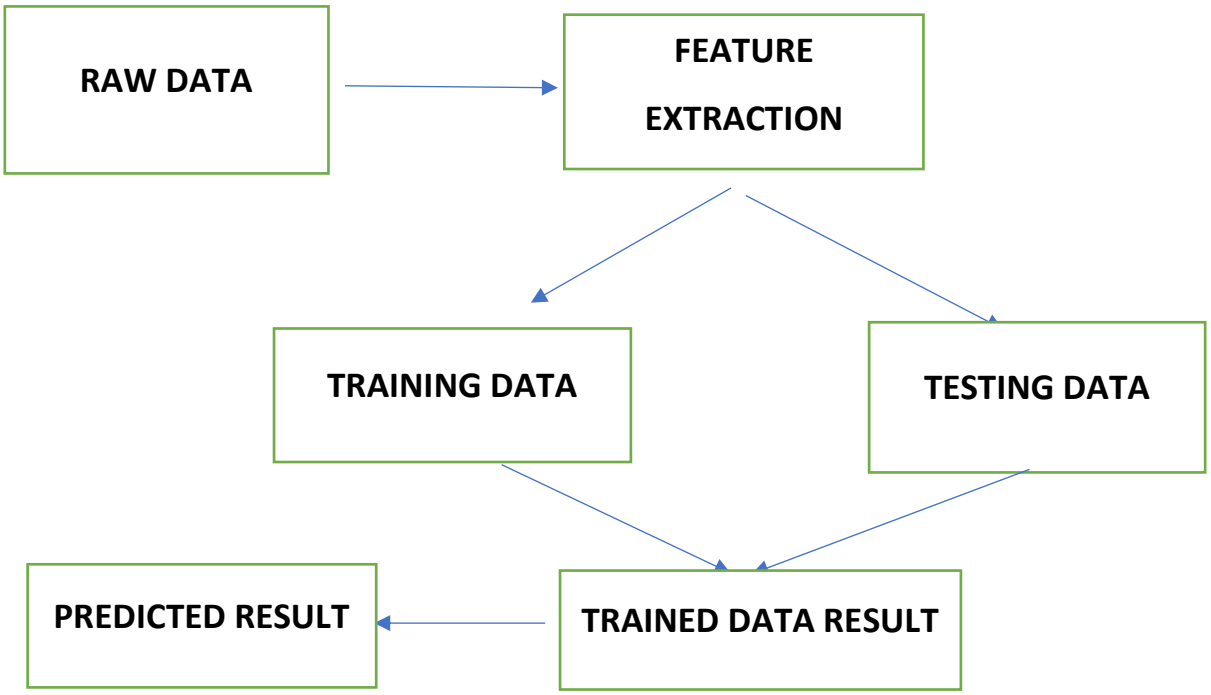


Figure 4.1: ARCHITECTURE OF THE SYSTEM

4.1 STEPS INVOLVED IN DESIGNING THE SYSTEM:

1. The first step is the conversion of this raw data into processed data. This is done using feature extraction, since in the raw data collected there are

multiple attributes but only a few of those attributes are useful for the purpose of prediction.

2. The first step is feature extraction, where the key attributes are extracted from the whole list of attributes available in the raw dataset. Feature extraction starts from an initial state of measured data and builds derived values or features.
3. These features are intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps.
4. Feature extraction is a dimensionality reduction process, where the initial set of raw variables is diminished to progressively reasonable features for ease of management, while still precisely and totally depicting the first informational collection.
5. The feature extraction process is followed by a classification process wherein the data that was obtained after feature extraction is split into two different and distinct segments.
6. Classification is the issue of recognizing to which set of categories a new observation belongs. The training data set is used to train the model whereas the test data is used to predict the accuracy of the model.
7. The splitting is done in a way that training data maintain a higher proportion than the test data.
8. Finally, the proposed system for analysing the stock is predicted.

CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1 OVERVIEW

The proposed system is a machine learning model, a programming language used to predict the stock market is python. Here, Accuracy plays the major role in stock market prediction. Although many algorithms are available for this purpose, selecting the most accurate one continues to be the fundamental task for getting best results. The algorithm used here such as Linear regression, Logistic Regression, SVM. This involves training the algorithms, executing them, getting the results, comparing various performance parameters of these algorithms and finally obtaining the most accurate one.

5.2 PERFORMANCE ANALYSIS OF ALGORITHMS

In this project, there is a comparison between machine learning algorithms and stock market dataset. By performing experiments with various algorithms on stock market dataset and observed the mean square error to predict accuracy using those algorithms. We used Python libraries such as pandas, numpy to load the dataset and to perform mathematical calculations respectively and we used sklearn to model different machine learning algorithms. The whole dataset with about 0.25 used to test the model and about 0.75 is used to train the model. Meanwhile, the algorithms for each model is predicted appropriately.

5.3 TOOLS AND LANGUAGES USED

PROGRAMMING LANGUAGES

- Python

TOOL

- Python IDE
- Anaconda
- Jupiter Notebook

5.4 EXPERIMENTAL RESULTS

The proposed system is trained and tested over the dataset taken from TATAGLOBAL BEVERAGES. It is spilt into training and testing sets respectively and yields the following results upon passing through the different models:

5.4.1 LINEAR REGRESSION MODEL RESULTS

In this model, the first (dependent variable) is continuous, the second variables (independent variable (independent variable) can be continuous and this leads to a linear line which is the nature of regression.

It establishes a relationship between X and Y, it makes a straight line which is best fit after computation (which is the regression line).

By the equation it is obtained:

Y = mx+c is used to obtain a straight line.

By considering low (X axis) and high(Y axis), the graph is obtained.

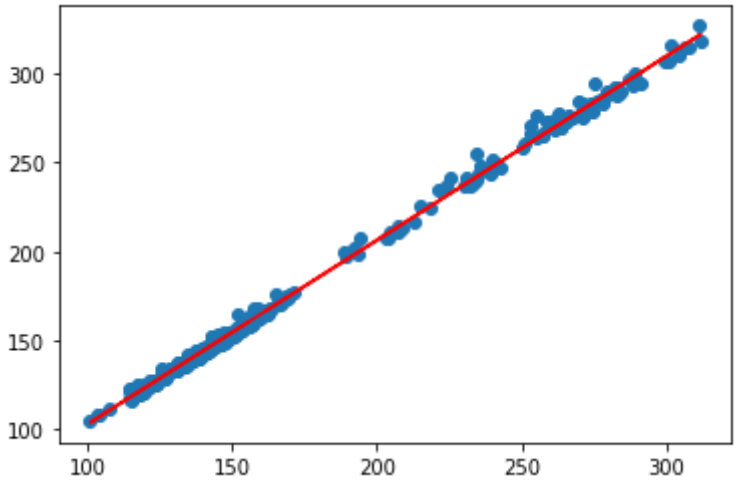


Fig 5.1: Plot between Low and High data using Regression.

After undergoing training and testing process, the model returns the needed predicted value by evaluating the low and high value to produce Turnover (predicted value).

```
Enter the low value:
206.87

Enter the High value:
300.98

Accuracy value for Low and Turn:
0.33586467511617535

Accuracy value for High and Turn:
0.36712327234463626

The RMS value is
7277.6696730622525
```

Figure 5.2: The predicted value of Linear Regression.

The above image shows the predicted value for linear regression.

5.4.2 SVM MODEL RESULTS

A Support Vector Machine(SVM) is a discriminative classifier that formally defined by the separating hyperplane. The SVM involves in plotting of data as point in the space of n dimnsions. After training data, the algorithm will obtain the output which is optimal hyperplane.

By considering Low(X axis) and High(Y axis) the graph is obtained.

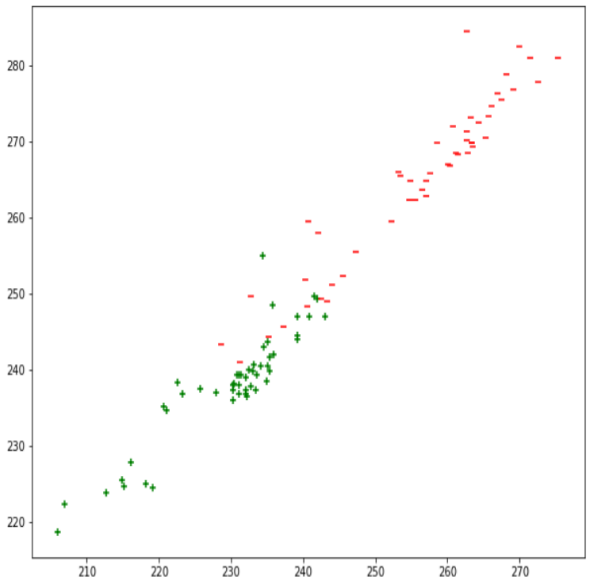


Figure 5.3: Plot between low and high data using SVM.

The above graph shows SVM decision marking boundary. SVM algorithm draws a boundary over the data set called as the **hyper-plane,** which separates the data into two classes as shown in the Fig5.2.

After testing and training process, the model returns predicted value which is obtained is turnover value.

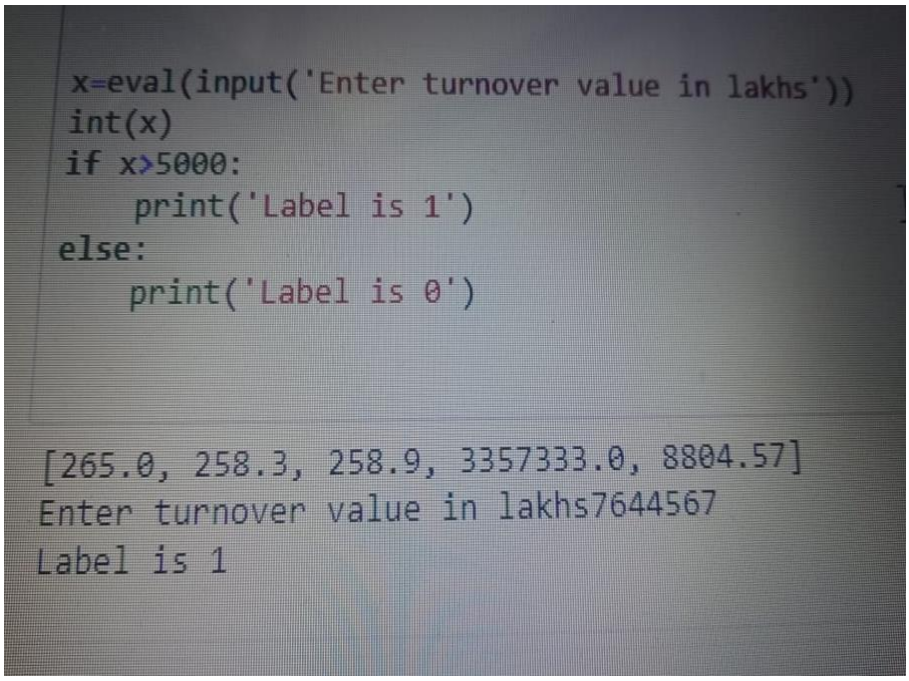


Figure 5.4: The predicted value of SVM

The above image shows predicted value using SVM.

5.4.3 LOGISTIC REGRESSION MODEL RESULTS

This model Is used to find the probability of how much chance there is for cases such that the event is either success or failure. Logistic Regression(LR) can be implemented when the dependent variable is binary in nature, that is it can have at most two values.

Logistic regression is quite similar to Linear regression. The only difference arises here is the sigmoid/logistic function is used.

LOGISTIC FUNCTION: $y=1/1+e^{-x}$ is used in regression.

The Scattered plot graph is obtained by taking Low(x axis) and High(Y axis).

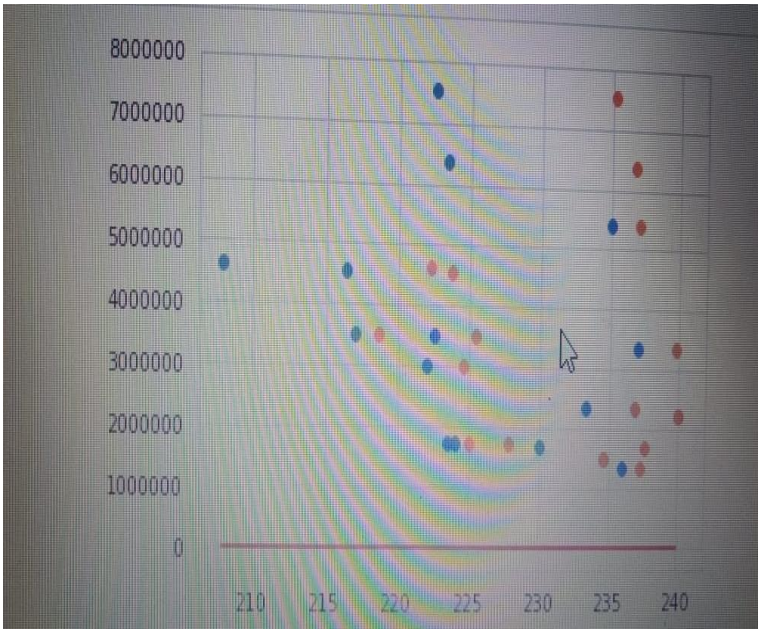
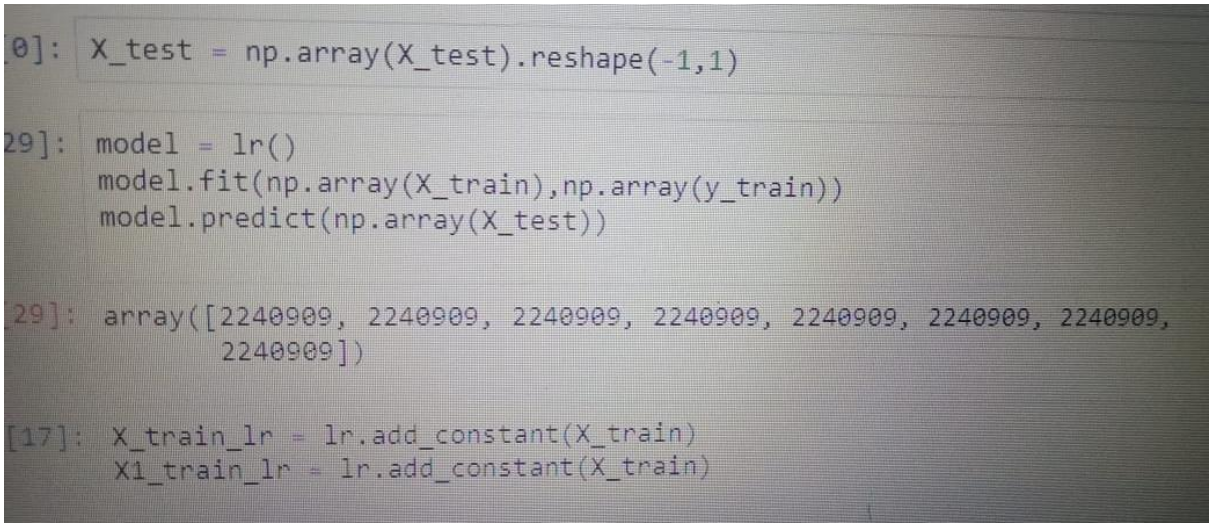


Figure 5.5 shows the scattered plot graph using Regression.

After training and testing process, the model returns predicted value between Low and High value which gives turnover(predicted value).



```
0]: X_test = np.array(X_test).reshape(-1,1)

29]: model = lr()
     model.fit(np.array(X_train),np.array(y_train))
     model.predict(np.array(X_test))

29]: array([2240909, 2240909, 2240909, 2240909, 2240909, 2240909, 2240909,
            2240909])

17]: X_train_lr = lr.add_constant(X_train)
     X1_train_lr = lr.add_constant(X_train)
```

Figure 5.6 Predicted value of Logistic Regression

The above image shows the predicted value of stock market using Logistic Regression.

CHAPTER 6

CONCLUSION AND FUTUREWORK

6.1 CONCLUSION

By measuring the accuracy of the different algorithms, we found that the most suitable algorithm for predicting the market price of a stock based on various data points from the historical data is the linear regression algorithm. The algorithm will be a great asset for brokers and investors for investing money in the stock market since it is trained on a huge collection of historical data and has been chosen after being tested on a sample data.

6.2 FUTURE WORK

In the future, the stock market prediction system can be further improved by utilizing a much bigger dataset than the one being utilized currently. This would help to increase the accuracy of our prediction models. Furthermore, other models of Machine Learning could also be studied to check for the accuracy rate resulted by them.

REFERENCES

[1]https://www.pantechsolutions.net/stock-market-prediction-using-machine-learning.

[2] Ashish Sharma, Dinesh Bhuriya, Upendra Singh. "Survey of Stock Market Prediction Using Machine Learning Approach", ICECA 2017

[3] Dutta, Bandopadhyay, Sengupta. “Prediction of Stock performance in the Indian Stock Market Using Logistic Regression”, IEEE 2017

[4] Ishita Parmar, Navanshu Agarwal, Sheirsh Saxena, Ridam Arora, Shikhin Gupta, Himanshu Dhiman, Lokesh Chouhan,”Stock Market Prediction using Machine Learning”, Conference paper December 2018.

[5] V Kranthi Sai Reddy,” Stock Market Prediction using Machine Learning”, IRJET 2018.

[6] Zhen Hu, Jibe Zhu, and Ken Tse “Stocks Market Prediction Using Support Vector Machine”, 6th International Conference on Information Management, Innovation Management and Industrial Engineering, 2013.M

[7] M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "Stock market prediction using machine learning techniques," 2016 3rd International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, 2016, pp. 322-327.

[8] K. Raza, "Prediction of Stock Market performance by using machine learning techniques," 2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT), Karachi, 2017, pp.