STOCK MARKET PRICE PREDICTION USING REGRESSION ANALYSIS

*Project report submitted*

*in partial fulfilment of the requirement for the degree of*

**Bachelor of Technology**

(2015-2019)

By

**Aditya Agrawal - Roll:1505615**

**Abhijoy Sarkar - Roll:1505089**

**Harshit Sanganeria - Roll:1505663**



DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

Deemed to be University

**April 2018**

**Table of Contents**

1: Introduction............................................................................................................ 4

2: Literature Review………………………………………………………………...5

3: Present Investigations…………………………………………………................ 6

3.1: Simple Linear Regression…………………………………………....... 7

3.2: Multiple Linear Regression………………………………………….....7

3.3: Polynomial Linear Regression……………………………………........ 8

3.4: Random Forest Regression……………………………………............. 8

3.5: Support Vector Regression……………………………………............. 8

4: Results and Discussions………………………………………………................. 9

4.1: Simple Linear Regression………………………………………........... 11

4.2: Multiple Linear Regression…………………………………….............13

4.3: Polynomial Linear Regression………………………………................ 15

4.4: Random Forest Regression………………………………………......... 17

4.5: Support Vector Regression………………………………………......... 19

5: Summary and Conclusions…………………………………………….……....... 20

6: References…………………………………………………………………......... 21

7: Acknowledgement................................................................................................. 22

**CERTIFICATE**

This is to certify that ABHIJOY SARKAR (Roll No. 1505089), ADITYA AGARWAL (Roll No. 1505615) and HARSHIT SANGANERIA (Roll No. 1505663) have successfully completed the project titled **“Stock Market Prediction using Regression Analysis”** under my supervision and guidance in the partial fulfillment of requirements of Sixth Semester, **Bachelor of Technology (Computer Science and Engineering)** of KIIT University, Bhubaneswar.

**Prof. Abhaya Kumar Sahoo**

**Department of Computer Engineering**

**KIIT University**

**Bhubaneswar**

**Declaration**

We declare that this written submission of this report represents my ideas in our own words and where other ideas or words have been included. We have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We 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**

**Introduction**

The stock corporation organizes the equity stake of its owners. It represents the outstanding assets of the company that would be due to stockholders after discharge of all senior claims such as secured and unsecured debt. Stockholders’ equity cannot be withdrawn from the company in a way that is intended to be detrimental to the company’s creditors. The stock of a corporation is partitioned into shares the total of which has a certain declared per value, which is a nominal accounting value used to represent the equity on the balance sheet of the corporation. In other word, shares of a stock may be issued without associated per value. A stock market is the aggregation of buyers and sellers of stocks (also called shares).

Machine Learning is the science of getting computers to learn by analyzing data, fitting models on training data and predicting the model results’ using test data. In our project, the problem we are trying to solve is that we want to predict the stock market prices for a set of data. Since the time capital markets have existed, investors have strived to gain edges in predicting stock prices. Use of machine-learning techniques and quantitative analysis to make stock market predictions have become very popular. In this paper, we present a study to understand the trend in stock prices and their volatility using machine learning regression models. We developed a mathematical model which combine the best practices to calculate the cross-validation score which depict the nature of a company’s stock market. In using our model, we establish a hypothesis function which takes as input the data we need to get the desired value and output of the function is the value of the prediction.

**2**

**Literature Review**

A considerable number of studies have inferred that predicting stock market returns is a difficult task. The non-linear and nonstationary features of the stock market make it a complicated system. According to Park and Irwin, participants in different stock markets use technical analysis substantial support.

Technical analysis can be understood as a set of rules or charting that tends to anticipate future price shifts based on the study of certain information, such as, for example, open price, closed price and volume traded. The first major study that considered technical analysis as a study subject was “Can Stock Market Forecasters Forecast?”, written by Alfred Cowles 3rd and published in Econometrica, July 1933.

The prediction of financial assets returns is a subject that encompasses many knowledge areas such as financial econometrics, investment analysis, corporate finance, and most recently, behavioral finance.

By analyzing a paper on “Stock Market Prediction using Ridge Regression and Random Forest Regression”, we got to know about the efficiency of Random Forest Regression in analyzing stock market prices and direction of movement. We got to know about decision trees and how they split the data into parts and construct models on the same and combine all the decision tree models to get an optimum model.

This report is based upon a comparative study of all regression models to identify the pros and cons of each model and highlight the differences. This survey report utilizes concepts of regression models to fit the appropriate model to the dataset.

“The Stock Market Trend Analysis” by Harold gave us an idea on how the direction of movement varies in the stock market and how the trend changes with other factors like seasonal influence, profits, etc.

**3**

**Present Investigation**

Dataset Gathered from: Yahoo Finance

Environment used: Conda environment with Python 3.6

Libraries: Numpy, Pandas, Sklearn, Matplotlib, statsmodel, SciPy

To predict the trends in stock prices in the market, we used regression analysis. Regression analysis is used to predict a continuous dependent variable from several independent variable(s). If the dependent variable is dichotomous, we should go for logistic regression. Since we are predicting the price which is a continuously varying dependent variable, our source of investigation relies on using regression models suitable for continuous data. The independent variables we use can be either continuous or discrete. Discrete variables, such as seasonal influence in which a stock market price trends changes abruptly during a season can be categorized as independent dummy variables. Usually, regression analysis is used with naturally occurring variables, rather than experimental variables.

We choose to predict the open price for a given day as we shall have some correspondence with the High, Low, Close, Adj Close and Volume.

**3.1**

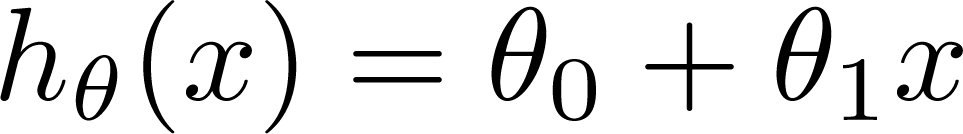
**Regression Models used:**

**3.1.1**

**Simple Linear Regression**

Simple Linear Regression is a regression model which uses two variables. One variable, the independent variable (named x) is used to predict the value of a dependent variable (named y).

Hypothesis function for the same is given by:

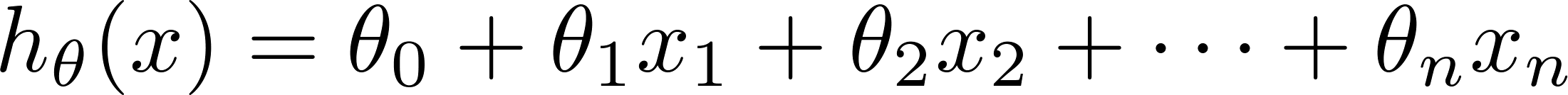


**3.1.2**

**Multiple Linear Regression**

Multiple linear regression is like simple linear regression, instead it uses more than one independent variable(s).

Hypothesis function for the same is given by:



**3.1.3**

**Polynomial Linear Regression**

This is a more general form of a linear regression in which we have independent variables of more than one degree, while the constants have degree one.

Hypothesis function for the same is given by:

**3.1.4**

**Support Vector Regression**

Support Vector Machines can also be used as a regression method, maintaining all the features that characterize the algorithm. SVR uses the same principles as SVM for classification, with only a few minor differences. First, because output is a continuous variable it becomes very difficult to predict the information at hand, which has infinite possibilities. In case of regression, a margin of tolerance is set in approximation to the SVM which would have already requested from the problem.

Hypothesis function for the same is given by:

**3.1.5**

**Random Forest Regression**

Ensemble technique called Bagging is like Random Forests. The idea behind this technique is to decor relate the several trees. They are basically an ensemble learning method for classification. It uses the concept of decision trees. Random Forest builds multiple decision trees to get a more accurate and stable prediction. Decision trees split the dataset based on some value, or average value(s).

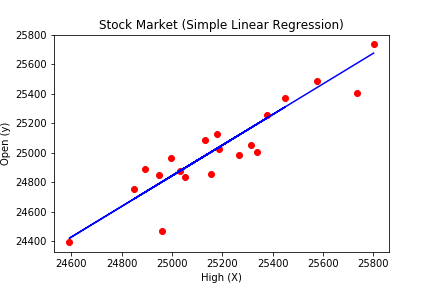
**4**

**Results and Conclusions**

**4.1**

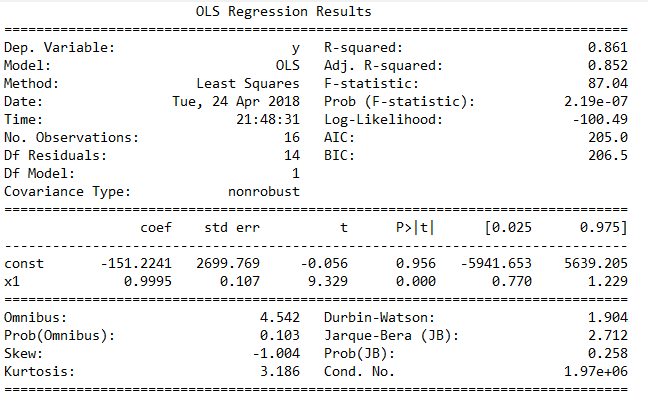
**Simple Linear Regression**

**Graphs:**

****

**Figure 4.1** Simple Linear Regression plot

**Ordinary Least Squares Summary**

****

**Regression Model Score:**



**Cost Function Value:**

108.05238

|  |  |  |  |
| --- | --- | --- | --- |
|  | **X testing set** | **Y testing set** | **Y predicted values** |

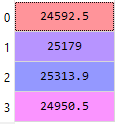
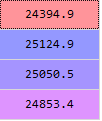
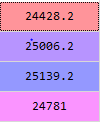


Table 4.1: Simple Linear Regression Prediction Results

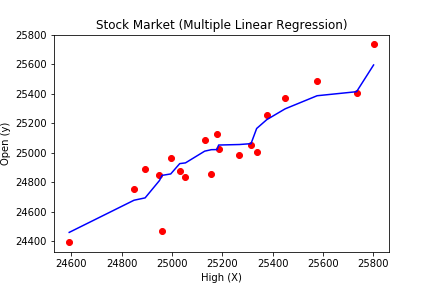
**Conclusion:**

The Simple Linear Regression model gave us an accuracy of 91% on the test data. This model was particularly used with Backward Elimination to find the optimum set of features for the prediction. The Significance level for the analysis was set to 5%, to identify the required optimum features using ordinary least square methods.

**4.2**

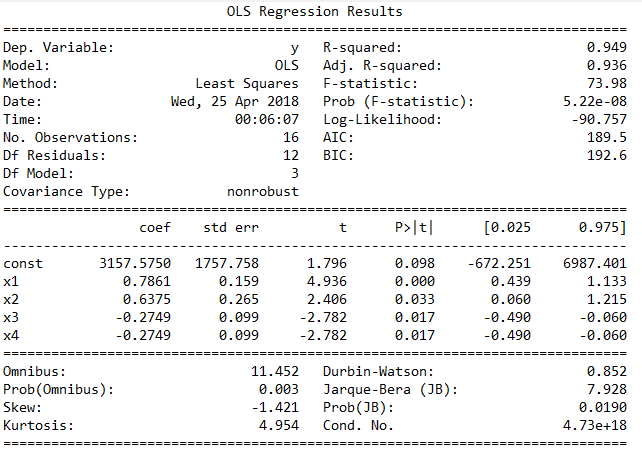
**Multiple Linear Regression**

**Graphs:**

****

**Figure 4.2** Multiple Linear Regression plot

**Ordinary Least Squares Summary:**

****

**Regression Model Score:**



**Cost Function Value:**

107.20564

|  |  |  |  |
| --- | --- | --- | --- |
|  | **X testing set** | **Y testing set** | **Y predicted values** |

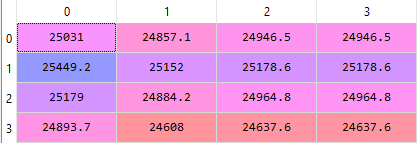
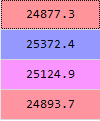
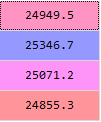


Table 4.2: Multiple Linear Regression Results

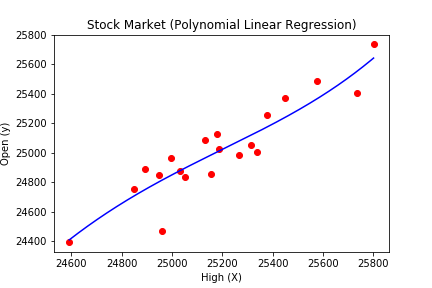
**Conclusion:**

The Multiple Linear Regression model gave us an accuracy of 93% on the test data. This model was particularly used with Backward Elimination to find the optimum set of features for the prediction. The Significance level for the analysis was set to 5%, to identify the required optimum features using ordinary least square methods.

**4.3**

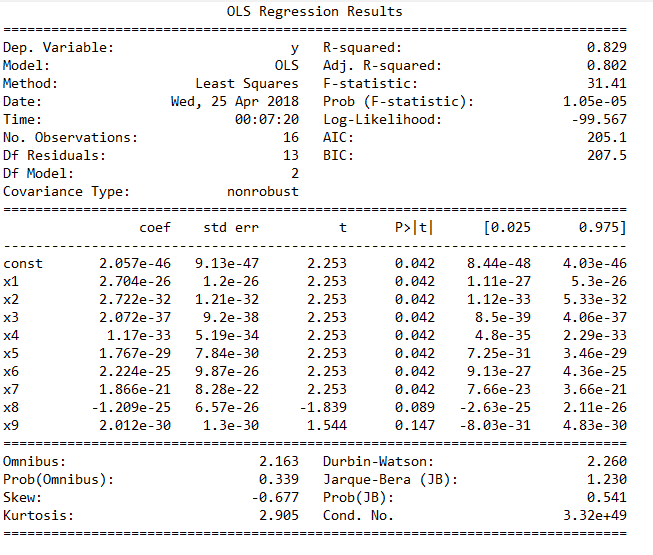
**Polynomial Regression (n=8)**

**Graphs:**

****

**Figure 4.3** Polynomial Linear Regression plot

**Ordinary Least Square Summary:**

****

**Regression Model Score:**



**Cost Function Value:**

153.71241

|  |  |  |  |
| --- | --- | --- | --- |
|  | **X testing set** | **Y testing set** | **Y predicted values** |

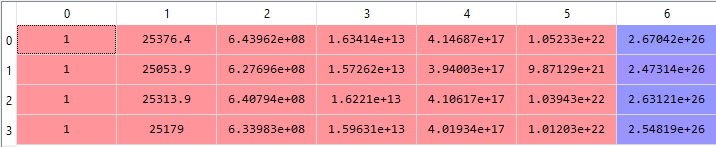
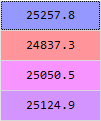
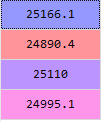


Table 3: Polynomial Linear Regressi

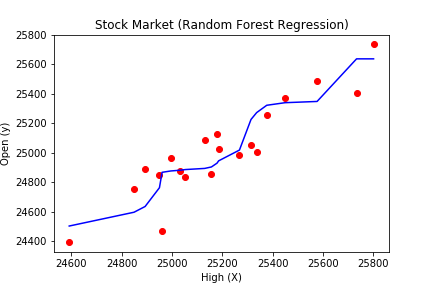
**Conclusion:**

The Polynomial Linear Regression model gave us an accuracy of 94% on the test data. This model was particularly used with Backward Elimination to find the optimum set of features for the prediction. The Significance level for the analysis was set to 5%, to identify the required optimum features using ordinary least square methods.

**4.4**

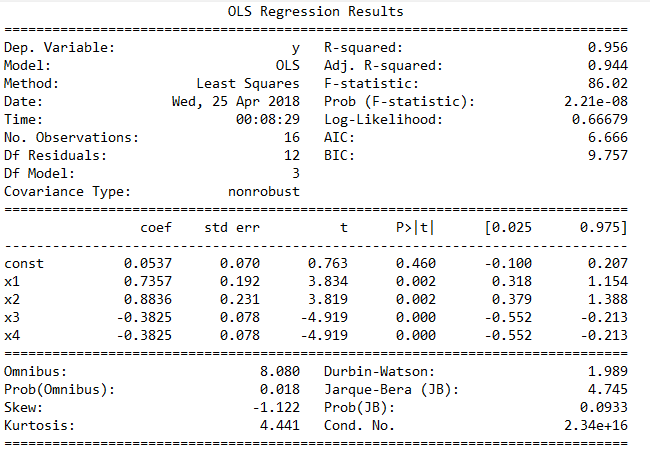
**Random Forest Regression**

**Graphs:**

****

**Figure 4.4** Random Forest Regression plot

**Ordinary Least Squares Summary**

****

**Regression Model Score:**

0.8231536056523545

**Cost Function Value:**

160.715384

|  |  |  |  |
| --- | --- | --- | --- |
|  | **X testing set** | **Y testing set** | **Y predicted values** |

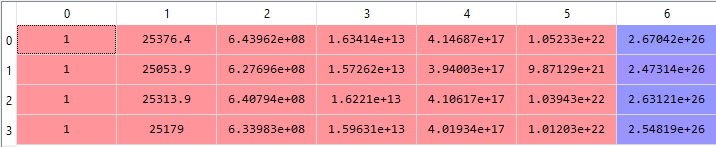
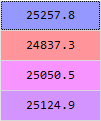
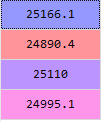


Table 3: Ra

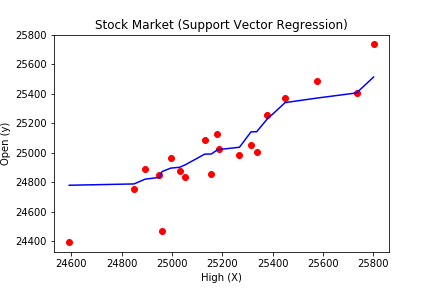
**Conclusion:**

The Random Forest Linear Regression model gave us an accuracy of 82.31% on the test data. This model was particularly used with Backward Elimination to find the optimum set of features for the prediction. The Significance level for the analysis was set to 5%, to identify the required optimum features using ordinary least square methods.

**4.5**

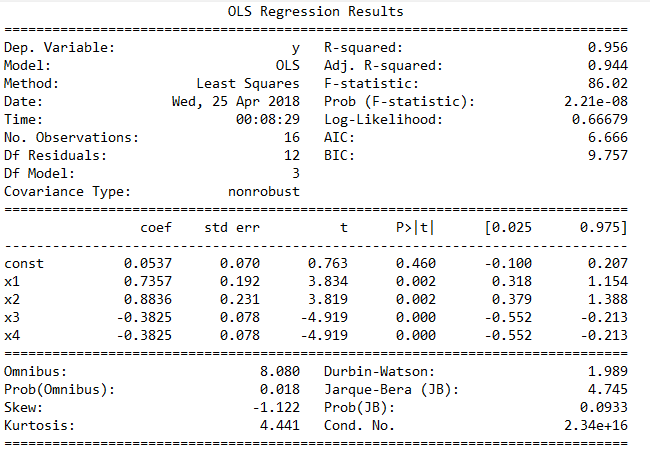
**Support Vector Regression**

**Graphs:**

****

**Figure 4.5** Random Forest Regression plot

**Ordinary Least Squares Summary:**

****

**Regression Score:**

0.8149234

**Cost Function Value:**

343.02835

|  |  |  |  |
| --- | --- | --- | --- |
|  | **X testing set** | **Y testing set** | **Y predicted values** |

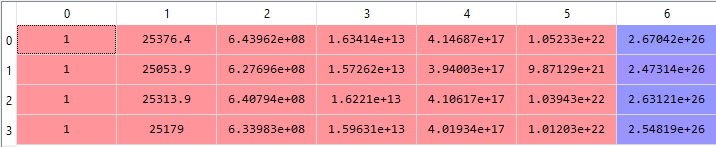
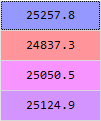
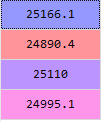


Table 3: Support Vector Regre

**Conclusion:**

The Support Vector Regression model gave us an accuracy of 81% on the test data. This model was particularly used with Backward Elimination to find the optimum set of features for the prediction. The Significance level for the analysis was set to 5%, to identify the required optimum features using ordinary least square methods.

**5**

**Summary and Results**

With the dataset gathered from Yahoo Finance, we performed regression analysis on the data using 5 regression models. Simple Linear Regression is a simple, basic algorithm in which we used ordinary least squares method to get the most optimum feature. In multiple linear regression, we used 4 variables to identify the output during which we had an accuracy of 93%. We analyzed the trends in dataset to find a polynomial nature in the data and we tried to fit a polynomial regression model where we got an accuracy of 94%.

To keep in mind, the various fluctuations and keeping in account of drastic changes in the stock market, we used support vector regression with a Gaussian kernel model.

Lastly, to survey all models, we performed a Random Forest Regression model to expect a more accurate and stable result. All in all, the highest accuracy we attained on a dataset was Polynomial Linear Regression.

**6**

**REFERENCES**

[1] Harold Jane, “Method of Regression Analysis”, [www.towardsinsight.com/regression-analysis](http://www.towardsinsight.com/regression-analysis)

[2] Shivank Choudhary, “Stock Market Prediction using Ridge and Random Forest Regression”,

pp. 5,2015

[3] Andrew Y. Ng," On Feature Selection: Learning with exponentially many irrelevant features as

training examples”, pp. 4-6, 1998

[4] Andrew Y. Ng,” Presenting overfitting on cross validation data”,

pp. 3-5, 1997

**7**

**Acknowledgement**

We hereby thank our mentor Prof. Abhaya Kumar Sahoo for the constant guidance and without whom, it would not had been possible if we had him by our side, helping us wade through all difficult times.

We also thank our friends who have been a constant support and provided us with all the help we needed during the successful completion of the project.

We would also like to thank our University for giving us the privilege of working as a team on a project of our choice.