Stock Price Trend Prediction Using Multiple Linear Regression

Shruti Shakhla[1], Bhavya Shah[1], Niket Shah[1], Vyom Unadkat[1]
Pratik Kanani[2]

[1](Student, Information Technology Engineering,
Dwarkadas J. Sanghvi College of Engineering, Mumbai University, India)
[2](Assistant Professor, Information Technology Engineering,
Dwarkadas J. Sanghvi College of Engineering, Mumbai University, India)
Corresponding Author: Shruti Shakhla

Abstract : *Stock markets are major institutions where trades are made daily having their worth in millions and billions on a daily basis. Some consider as a fast way to fill up their pockets whereas some have a traditional approach to invest in a company and reap long term benefits. A well done research about the company and its performance can help to gain better financial profits. Industry experts and major conglomerates invest heavily in research and development and with the technological developments, minimising the error probability and estimating the future performance has become achievable. Machine Learning algorithms have proved to fetch benevolent results in predicting stock prices. In this paper, we have studied and documented the performance of APPLE INC.'s stock price using Multiple Linear Regression and gauged its performance using Root Mean Squared Error. The results are promising but can be improved by taking into consideration more parameters.*
Keywords -*Stock Market, Finance, Linear Regression, Machine Learning*

I. INTRODUCTION

Making money hand over fist is the real incentive which draws investors towards the stock market. The stock market is a platform where individuals can buy and sell shares of publicly traded companies. It is a system with individuals of opposite views, ideas and approach. One investor will sell a particular stock whereas someone else will buy it. One of them would profit from the trade and hence, it is very important to be well versed while dealing with stocks to maintain a profitable account.[1]

There are many factors affecting the price of the stock be it supply and demand, media, risk, statements made by the market experts, political moves, natural disasters and many more. All of these and an array of other factors together when compiled, have a direct impact on the stock price. Investors need to study extensively before investing in a particular stock. The stock's actual price and its fair value need to be taken into consideration to make crucial decisions. The actual price is the price at which it is traded in the market. The fair value can be estimated by the company's assets keeping the liabilities aside. It can also be derived by the net value of the company's future earnings.

The indispensable surge towards prediction models has compelled financial analysts in adopting machine learning models and computational intelligence. Technical analysis of stock markets is made easier by implementing machine learning models such as Regression, Support Vector Regression and Deep Networks with the help of various tools and libraries.[2]

Here we use a linear regression model to learn the trends of the market and then we predict the price of the stock, to givea brief idea about how the stock is going to perform.

II. RELATED WORK

R. Yamini and Dr. C. Dhaya highlighted the major factors that are to be considered while selecting the best suited Machine Learning Prediction Model for stock analysis. They used multiple prediction models to compare and analyze the performance of each model in order to find the most accurate model that must be implemented for stock analysis.[3]

Another paper written by Han Lok Siew and Md Jan Nordin demonstrated the importance of using ordinal or standardized values from a dataset made of heterogeneous types of data, in regression prediction techniques. The paper hypothesized the factors to be considered for a prediction model, which conjugated the basis of performing fundamental analysis of stocks. The variables were extracted from balance sheets, income statements and other cash flow statements provided in the annual reports of the particular stock tickers.[4]

K.R. Sekar, Manav Srinivasan, K.S.Ravichandran and J.Sethuram wrote a paper on the prediction of gold price in the stock market based on several independent yet influential variables. The prediction model used multiple linear regression algorithm to predict the price of gold in the market. Their model took a dataset consisting of historical gold prices, along with other variables of many years on a monthly basis to feed into their model which would be used for prediction later on.[5]

III. PROBLEM DEFINITION

Nature of the stock market is extremely volatile. The major goal is to minimize the uncertainty of the returns by meticulously predicting (as accurately as possible) the future stock prices and thereby identifying their fluctuations in advance to mitigate financial risks. For the following reasons we have developed, our own prediction model which takes into account multiple factors in order to predict the stock prices.
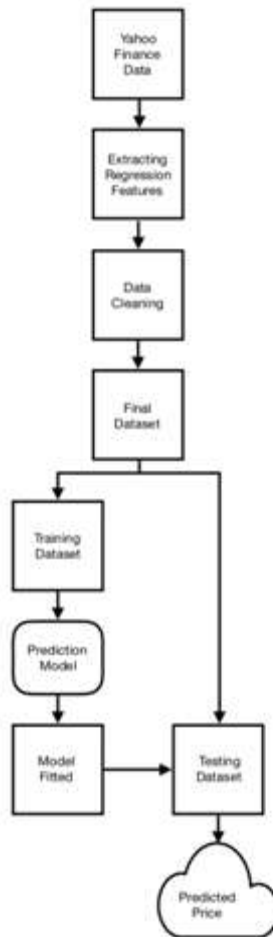
IV. DATA METHODOLOGY

Figure 1: Methodology

The stocks considered for our prediction model are AAPL(Apple), which comes under the market index NASDAQ(NDAQ). Our prediction model will be trained on the historical data gathered using the Yahoo Finance Developer API.[6][7][8]

The Yahoo API provides various methods which facilitates data acquisition. Using this API, we focus on gathering the following two data parameters:

- Stock's Day Open Price
- Stock's Day High Price; for both, AAPL and NDAQ.

Further, since the stock prices do not change over the weekend, we wish to eliminate them. Yahoo API provides a feature which helps acquire only weekday's stock prices. After acquiring the requisite data, it is significant to follow it up with data-preprocessing. To aid this process, the data gathered is converted to Panda's Data-frame and indexed. Furthermore, it is noticed that there are various NAN (Not A Number) values in the Data-frame. A NAN value in a Panda's Data-frame is indicative of a missing value. These missing values make

the Data-frame inconsistent and prevent application of any predictive models on them. There are various ways to deal with these NAN values and we chose to use the panda's function fillna() on the Data-frame to forward-fill and backward-fill the values. Forward-fill uses last valid observation to fill missing values up to the next valid observation whereas backward-fill uses the next valid observation to fill in the gap. After eliminating all redundancies and inconsistencies in the data, the predictive model can now be trained using it.[9]

V. PROPOSED MODEL

Multiple linear Regression [10] is a highly established statistical technique used in stock market analysis. It allows the analyser to consider multiple variables which affect the quantity to be predicted. The quantity to be predicted is usually referred to as the independent variable. The various factors which demonstrate a strong correlation with the independent variable are referred to as the dependent variables. The measure by which the dependent variable changes due to a unit change in the independent variable is known as the regression coefficient of that independent variable.

This model mathematically calculates a linear (straight line) relation between the dependent variable and every other independent variable (Apple's daily high and Market's daily high). The general mathematical equation representing a multiple linear regression is:

$Y_i = B_0 + B_1X_{i1} + B_2X_{i2} + \ldots + B_nX_{in} + E$[11]

where Y_i : dependent variable
$X_{i1}, X_{i2}, \ldots, X_{in}$: n-independent variables
$B_1, B_2, \ldots, B_n$: regression coefficients
B_0 : y-intercept
E : random error in the

model

One of the popular use of this model is considering Least Squares Regression. Least Squares Regression aims at calculating a best fit line by minimising the residual sum of squares of the deviations of the predicted values from the corresponding data points. Squaring the deviations removes the complexity introduced in the model due to positive and negative values. Based on above information we have formulated our own prediction model to carefully predict a stock's daily high price.
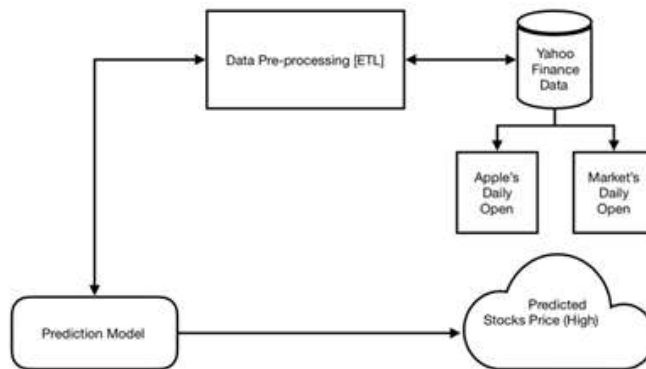
Figure 2: Stock Prediction Model

The Prediction Model using Multiple Linear Regression Method has been built using Python Programming. We aim to predict a stock's daily high using historical data. The data used is the stock's open and the market's open. The model used is a Multi-Linear Regression model which is one of the most extensively used model in stock market analysis.

Considering the dynamic nature of the stock market, the opening price of a stock on a particular day is highly dependent on a plethora of factors. Hence considering the opening price of a stock as an independent variable for our prediction model will give us an amalgamation of all those factors. The market's opening price gives us a broader understanding of the trends of the stock, especially in case of a drastic fluctuation in the entire market.

The stock considered is AAPL, which comes under the market index Nasdaq (NDAQ). Historical data is gathered using Yahoo Finance [6].

VI. FINDINGS AND DISCUSSIONS

The model has been trained on data of 12 months.
- Training Dates: 30th September, 2016 – 30th September, 2017
- Test Dates: 1st October, 2017 – 31st October, 2017

Table 1: Table of regression coefficients and intercept

Intercept	AAPL Open Coefficient	NDAQ Open Coefficient
-0.25142712813	0.99658308	0.02304602

APPL$_{High}$ = -0.251427128131 + (0.99658308 * AAPL$_{Open}$) + (0.02304602 * NDAQ$_{Open}$)
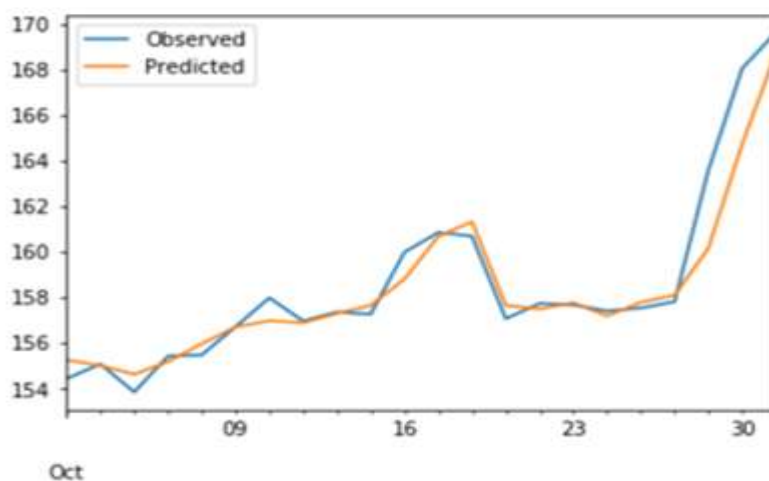
Figure 3: Predicted vs Observed values

This model produces a R^2 score of 0.91433103366658208 and Root Mean Square Error (RMSE) of 1.145040040250809. The R^2 score, also known as the coefficient of determination is the proportion of variance in the dependent variable that can be predicted from the independent variables. The Root Mean Square Error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed.

VII. CONCLUSION

Modelling of capital markets has authentically been done in partial equilibrium. Such machine learning models indeed provide us with a broader understanding of the capital market. Hence we have harnessed this opportunity to gain an insight about the market. In this paper we have incorporated the multiple linear regression technique, which is one of the most routinely used technique to make predictions in the capital market. Considering the open price of the stock as an independent variable eliminates the need to consider other factors to predict the high of the stock. The open of the market provides with an insight of the variations in stock prices of the entire market. Taking in account the above mentioned factors, the linear regression model produced a RMSE of 1.145040040250809.

We believe that the performance measures obtained here can be improved further with the use of genetic-based algorithms [12-18]. Taking into consideration a wide array of factors could widen the scope of the model.

REFERENCES

[1]. The Economic Times, "Definition of Stock Market" https://economictimes.indiatimes.com/definition/stock-market
[2]. Sebastien Heiz, "A simple deep learning model for stock price prediction using TensorFlow" https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877
[3]. R. Yamini Nivetha, Dr. C. Dhaya, "Developing a Prediction Model for Stock Analysis" 2017 International Conference on Technical Advancement in Computers and Communications.
[4]. Han Lock Siew, Md Jan Nordin, "Regression Techniques for the Prediction of Stock Price Trend".
[5]. K.R Sekar, "Gold price estimation using multivariable model", International Conference on Networks and Advances in Computational Technologies (NetACT) July 2017.

[6].   Yahoo Finance Developer Network https://developer.yahoo.com/yql/?guccounter=1Apple Inc. A technology company https://www.apple.com/Nasdaq Financial Company https://www.nasdaq.com/"Padas" Library documentation for the function "fillna()" https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.fillna.html
[7].   Lucas Nunno, "Stock Market Price Prediction Using Linear and Polynomial Regression Models", 2014.
[8].   yale.edu, "Multiple Linear regression", http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm
[9].   Khaled Assaleh , Hazim El-Baz , Saeed Al-Salkhadi, "Predicting Stock Prices Using Polynomial Classifiers: The Case of Dubai Financial Market ", Journal of Intelligent Learning Systems and Applications, 2011, 3, 82-89
[10].   Phua, P.K.H. Ming, D., Lin, W. (2000), Neural Network With Genetic Algorithms For Stocks Prediction, Fifth Conference of the Association of Asian-Pacific Operations Research Societies, 5th - 7th July, Singapore. J. Moody. Learning to trade via direct reinforcement. IEEE Transactions on Neural Networks
[11].   A.S. Weigend and N.A. Gershenfeld, "Time Series Prediction: Forecasting the Future and Understanding the Past," Addison-Wesley, Reading, MA, 1993.
[12].   G. Grudnitzky and L. Osburn, "Forecasting S&P and Gold Futures Prices: An Application of Neural Networks," Journal of Futures Markets, vol. 13, No. 6, pp. 631-643, September 1993.
[13].   P.J. Werbos, "Generalization of back propagation with application to a recurrent gas market model," Neural Networks, vol. 1, pp. 339-356, 1988.
[14].   Hiransha M , Gopalakrishnan E.A , Vijay Krishna Menon, Soman K.P, "NSE Stock Market Prediction Using Deep-Learning Models", International Conference on Computational Intelligence and Data Science (ICCIDS 2018)