Data Analysis on the Bangladesh Share market using Machine-Learning

by

Syed Hassan Sadman
17101238
Hasib Al Rahat
17101017
Atikur Rahman Hamim
17101089
A.R.M. Emtiaj Al Kafi
17301203
Nasik Ali Khan
20301459

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

A.R.M. Emtiaj Al Kafi
17301203

Syed Hassan Sadman
17101238

Hasib Al Rahat
17101017

Atikur Rahman Hamim
17101089

Nasik Ali Khan
20301459

Approval

The thesis/project titled "Data Analysis on the Bangladesh Share market using Machine-Learning" submitted by

1. Syed Hassan Sadman (17101238)
2. Hasib Al Rahat (17101017)
3. Atikur Rahman Hamim (17101089)
4. A.R.M. Emtiaj Al Kafi (17301203)
5. Nasik Ali Khan (20301459)

Of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 2, 2021.

Examining Committee:

Supervisor:
(Member)

Tanvir Rahman
Lecturer
Computer Science Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Associate Professor
Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Mahbubul Alam Majumdar, PhD
Professor and Dean, School of Data and Sciences
Department of Computer Science and Engineering
Brac University

Ethics Statement

Hereby, we, the members, consciously assure that for the manuscript /insert title/ the following is fulfilled:

1) This material is the authors' own original work, which has not been previously published elsewhere.

2) The paper is not currently being considered for publication elsewhere.

3) The paper reflects the authors' own research and analysis in a truthful and complete manner.

4) The paper properly credits the meaningful contributions of co-authors and co-researchers.

5) The results are appropriately placed in the context of prior and existing research.

6) All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference.

7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

The violation of the Ethical Statement rules may result in severe consequences.
We agree with the above statements and declare that this submission follows the policies of BRAC University.

Abstract

The stock market plays an important role in the growth of industries by supplying funding. Thousands of Bangladeshis use the stock market as a means of employment. The stock markets in Bangladesh have been declining lately, impacting millions of individuals. What if stock markets could allow investors to know which stock is more reliable or less dependable. We will be using the Linear Regression algorithms along with the K-Nearest Neighbors algorithm, the Support Vector Machine (SVM), Lasso Regression and Multi-Linear Regression to predict the stocks.

Keywords: Data Mining; Machine Learning; Stock Market; Prediction; Linear Regression Analysis; K-Nearest Neighbour; Support Vector Machine; Lasso; Analysis

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our Advisor Mr. Tanvir Rahman sir for his kind support and advice in our work. He helped us whenever we needed help.
And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

β	Beta
ϵ	Epsilon
μi	Mu
υ	Upsilon

Chapter 1

Introduction

1.1 Data Analysis on the Bangladesh Share market using Machine-Learning

The most mysterious and difficult task in the world is predicting anything. Good forecast does good things and bad forecasting loses enormously. The prediction of stock markets is one of the most challenging tasks for all involved. Prediction is quite impossible with 100% accuracy. Good forecast means good average calculation prediction. If someone's forecast is better, he/she is a good analyst. It was our common goal from the beginning of the world to make our lives easier and more comfortable.In society, wealth brings comfort and luxury and so much has been done about ways of forecasting stock markets. Several methods, methods and ways with different results have been suggested and used. The stock market prediction is based on market statistics from previous years to forecast the future stock. Stock Market refers to the share and stock market of a country. Companies and governments may raise huge amounts of money through the stock market. There are two stock exchanges active in Bangladesh, one being the Dhaka stoke Exchange(DSE). In 1976 DSE started trading with 0.138 billion BDT paid- up capital. At the time, it had 0.147 billion BDT market capitalisation. For investors the correct forecasted movement of the stock companies can be of great benefit as BDT 46.2 billion has been capitalized in 2020. DSE includes 578 companies with a list of banks, institutions, treasury bonds, textiles, pharmaceuticals, insurance and reciprocal funds. [6] As long as stock markets existed, both Fortune Hunters and academics tried to predict them. It is interesting and valuable to predict stock markets and stocks, because financial and economic advantages can be gained by numerously different factors; stocks are well-known as difficult to predict. Researchers cannot agree that it is therefore very interesting if stock markets are predictable or not to examine if they can be predicted. The stock market is a capital market part and parcel. Any stock exchange contribution generally leads to economic growth by increasing funds to finance industry and other companies..

1.2 Problem Statement

Bangladesh is a predominantly agricultural region. The economy and industry of Bangladesh does not yet look good enough. The growth of the Gross Domestic

Product (GDP) and per capita income was below estimating. This problem stems from the illiteracy of the majority of people and their lack of awareness about what is to be achieved in the stock market and how it actually operates. Not all stock markets are always running the way they should. In most of the companies with lower dividend rates, the majority of the people believe that they are unfair. The businesses give low dividends for investors, so they are unable to make investments on the stock exchange. Since most people are illiterate, they live by hand to mouth. It is naive to expect that they will have knowledge of the stock market and the essence of the stock market. Even knowledgeable people don't know about the stock market. Moreover, commercial banks are not willing to disburse much loan for investing in the stock market and even they give a limited amount of loan that is not enough to form a substantial amount for capital required in industrial sectors. While the number of re-registered or public limited companies is not as much as we'd like, the number of shareholding companies still needs to be increased. Merely having very few publicly traded stocks, the stock market is still struggling to gain the interest of the broader public. Some clever traders or market manipulators trick investors and benefit more from the stock market. They exploit the artificially restricted supply of goods to market the same product at a higher price. Increase the price and purchase an ever-increasing number of shares. Through their speculative strength, they cause the stock market to rise or fall, robbing the individual investors of their gains. The stock market is not properly regulated. It was found that the rules and regulations were not strictly followed, so the management is not possible to do so. Also, It is unable to make timely business decisions which hinders usual operations as well as investor income. For being too influential, stock exchanges are not doing well and creating problems for the economy. Compared to other nations, our country is not politically peaceful. It causes tension and problems. Politics has a direct effect on the daily operations of a company. Investment within and outside the country may be removed. The bad cycle goes to the stock market and the investors are reluctant to invest anymore even withdraw capital from the stock market which creates a huge problem in the stock market and the companies. Our information flow is not free. What kinds of companies are there, which one is profitable, which one is not working properly and which ones will prosper in the upcoming economic situation etc. There is no burial in this land. Because of lack of information technology, the public's ability to accurately analyze the stock market is compromised.

1.3 Aims and Objectives

The aim of this research is just to study all of the other selected shares in different sectors along with aluminum, consumer goods, software, investment finance, banking, motor vehicles, pharmaceuticals, and general engineering with regard to the Indian stock market between 1994 and 2000. Several error concepts are used in order to decide how the value of stock prices is influenced by several variables. The study aims to determine whether safety markets in Bangladesh represent completely at any moment "all the publicly available information that can be gleaned again from the timeline in historical prices. Thus, the analysis relies on a weak form of efficient market hypothesis. To evaluate empirically whether a rough set hypothesis (or the weak EMH form of stock prices) is a plausible explanation for

share price behavior under Indian conditions. To decide whether historical prices have given the best price prediction for the future. The question is whether prices represent a continuous pattern over a short time or whether they are random. To estimate the effect of a certain share's price on the overall market by analyzing the behavior of other shares in the same market and shares in other industries. The key factors deciding the stock price behavior are those in the regression equation.

1.4 Work Plan

In our thesis, we are using past 15 years of data sets from the Bangladesh stock market to forecast the stock price movements for the future. An example of data set, which will be using for our analysis, is mentioned below. We have considered many status factors for stock exchanges, including the three companies from the 2016-2020 period. Now a question comes up on whether these businesses sufficiently consider the proper and necessary priority of data of this nature before investing in the stock market. "Yeah, probably or maybe not" Our research provides insights into how the business can be run for a profitable and healthy investment. We hope this research will help them make an accurate report of the state of the stock exchange market.

Company name	Date	Open	High	Close	Low	Volume
Square Pharma	01/01/2020	8.1	8.1	8	8.1	24327
	01/01/2019	12	12.1	11.8	11.9	208373
	01/01/2018	21.9	22.2	21.5	21.6	2715640
	01/01/2017	20	20.08	19.64	19.73	3136275
	01/01/2016	18.75	18.93	18.84	18.66	382216.5
Marico	01/01/2020	1585	1585	1562.1	1566.8	16295
	01/01/2019	950.2	993.5	950.2	969.9	38
	01/01/2018	800	810	790	791.1	453
	01/01/2017	620.5	662.4	612	662.4	6930
	01/01/2016	1.33	1042.5	1033	1035.7	1409
First Bank	01/01/2020	8.90	8.90	8.72	8.81	690287.4
	01/01/2019	9.00	9.09	8.84	9.00	1491458.1
	01/01/2018	11.57	11.72	11.49	11.57	1071271.322
	01/01/2017	9.78	9.84	9.21	9.42	11964621.21
	01/01/2016	5.20	5.20	4.99	4.99	4.990876892

Table 1.1: Data Set of Different Companies

Chapter 2

Related Work

Some similar work in this subject of other include work of Wasiat Khan and Mustansar Ali Ghazanfar, who published "Stock market forecast using machine learning classifiers and social media, news". They wrote that the accurate forecasting of stocks is of tremendous interest to investors but stock markets are influenced by volatile elements such as microblogs and news which make it difficult to anticipate stock market index on the basis of historical data alone. The high volatility of stocks stresses that the function of external factors in stock planning must be assessed correctly. Stock markets can be forecasted with algorithms for machine learning of social media information and news, as this data can alter the behavior of investors. They employed social media and financial news algorithms in this paper, to Discover the influence of this data on the accuracy of stock market predictions over the following 10 days. The selection and spam tweets on the data sets are conducted to improve the performance and quality of forecasts. We also experiment with such hard-to-predict stock markets and those that are more affected by social media and financial news. We are comparing Results for a consistent classifier of different methods. Finally, to achieve the highest forecast Precision, deep learning and certain classifiers are utilized together. The findings of their experiments suggest that 80.53% and 75.16% of highest forecast accuracies have been attained by social media and financial news correspondingly. they also show that the stock markets of New York and Red Hat are difficult to anticipate, and that social media influences New York and IBM stocks more, while London and Microsoft stocks are affected by Financial news. It has been discovered that the random forest classification is consistent and accurate in 83.22% of its ensemble is realized. In another paper, Bo Qian and Khaled Rasheed did for Applied Intelligence, titled "Stock market prediction with multiple classifiers". They wrote that the Prediction of the stock market is interesting and complex. According to the effective market theory, stock prices should follow a random pattern, and hence with more than 50% accuracy ought not be predictable. The Dow Jones Industrial Average Index was analyzed to reveal that not all periods are equally random. They chose a period with high predictability using the Hurst exponent. Auto-mutual information and fake closest neighbor algorithms were used to hone parameters for creating training patterns heuristically. These created patterns were then used to train certain inductive machine-learning classifiers- including an artificial neural network, a decision tree, and k-nearest neighbor. We were able to attain prediction accuracy of up to 65% by combining these models in the right way. Kim and Han created a model

for forecasting stock price index using a mixture of artificial neural networks (ANN) and genetic algorithms (GAs) with feature discretization. The technical indicators as well as the direction of change in the daily Korea stock price index were employed in their research (KOSPI). They picked features and formulas from data including samples of 2928 trading days spanning the years January 1989 to December 1998. They also used feature discretization optimization, which is a method comparable to dimensionality reduction. The strengths of their work are that they presented GA to optimize the ANN. First, the hidden layer's input characteristics and processing components are 12 and are not changeable. Another disadvantage is that the authors only concentrated on two elements in the optimization phase throughout the ANN learning phase. They still feel GA has a lot of promise when it comes to feature discretization optimization. Qiu and Song also offered a solution based on an improved artificial neural network model for predicting the direction of the Japanese stock market. The authors use genetic algorithms in conjunction with artificial neural network-based models to create a hybrid GA-ANN model.
To forecast stock trends, Lee combined the support vector machine (SVM) with a hybrid feature selection approach. This study used a subset of the NASDAQ Index from the Taiwan Economic Journal Database (TEJD) in 2008. A hybrid method was used for feature selection, with supported sequential forward search (SSFS) serving as the wrapper. Another benefit of this work is that it included a detailed approach for parameter tuning that included performance under various parameter values. The feature selection model's clear structure also serves as a heuristic for the first stage of model structuring. One of the drawbacks was that the performance of SVM was only compared to that of the back-propagation neural network (BPNN) and not to that of other machine learning methods.
.

Chapter 3

Prediction Modeling using Various Algorithms

We will be using the Linear Regression algorithms along with the K-Nearest Neighbors algorithm, the Support Vector Machine (SVM) and Lasso Regression and Multi Linear Regression to predict the stocks by analyzing the datasets.

3.1 Linear Regression

In preparation of our model we have used a linear regression algorithm. Linear regression means that a link between two factors relating to the observed data on a linear equation can be demonstrated. One variable is seen as an explanatory and the other as a dependent. In the case of suspicion that the relation between the factors will continue later, the linear regression strategy can be used. A Linear Recovery Eq. (1) is as follows

$$\mathrm{Yi} = \beta 0 + \beta 1 \mathrm{Xi} + \varepsilon \mathrm{i} \tag{3.1}$$

In this case, Yi is the result (projected output) of the variable for the test unit; Xi is the indie variable used for the ith test unit prediction. $\beta 0 + \beta 1$Xi is the linear Yi-Xi relationship $\beta 0$ is the mean Y when X =0 and $\beta 1$ are either the slope or the mean change if X increases by 1. The μi is the term of the error. We used our model as our independent variable the closing column for data. So, Close I = Xi in our model. Unknown are the $\beta 0$ and $\beta 1$ parameters. They are calculated with the least quadratic method. [3] The least square method (2) is as follows:

$$\mathrm{b} = (\mathrm{n}(\Sigma \mathrm{X}) - (\Sigma \mathrm{X}) \cdot (\Sigma \mathrm{Y})) / (\mathrm{n}\,(\Sigma \mathrm{X}\text{^}2) - (\Sigma \mathrm{X})\text{^}2) \tag{3.2}$$

Considering the least square the best-fitted line is taken and uses that as a function for the prediction. There are several techniques to prepare a linear regression.

Simple Linear Regression: If we have a single input, we can use statistics to estimate the coefficients with a simple linear regression. This requires us to calculate statistical data properties, such as deviations, standard differences, correlations and covariance. To traverse and calculate statistics, all data must be available.

Ordinary Least Squares: We can use Ordinary Least Squares to evaluate the values in the coefficients if we have more than one input. This process aims at minimizing the sum of squared residues for ordinary squares. This way, the distance from each data point to the return line is calculated, placed and summarized together with the regression line through the data. This is to minimize the volume of ordinary squares. Data are treated as a matrix in this approach and linear algebra operations are used to determine ideal coefficient values. This requires all data to be made available and enough memory to fit the data and run the op matrix.

Gradient Descent: When one or more inputs are available, we can optimize the values of the coefficients by minimizing the model error on your training data iteratively. This procedure is called the Gradient Descent and works at random values for each coefficient. For each pair of input and output values, the sum of the squared errors is determined. A learning rate shall be used as a scale factor and the coefficients shall be updated to minimize the error. The process is repeated until there is a minimum amount of error or no further improvement. We must select an alpha parameter when using this method that determine the size of the improvement process to be performed in every iteration. A linear regression model is often taught to descend gradients because it is relatively simple to understand. Practically, when the number of rows or columns that may not fit in the storage is highly useful. It should be used.

Regularization: Linear model training extensions known as regularization methods Both these instructions decrease the squared mistake in the training data model, while also reducing the model's complexity (like the total of all the coefficients in the model number or absolute size) Examples of regularization procedures for linear regression:

Lasso Regression: Here, regular lower places are changed, so that the absolute number of coefficients can be minimized (called L1 regularization).

Ridge Regression: Ordinary lower squares are modified here to also reduce the absolute squared sum of the coefficients (called L2 regularization).

While doing a predictive analysis usually linear regression analysis is used to ensure the solutions are providing significant outcomes. The regression model shows the relationship between the one dependent and independent variable (that may be greater than one in number). Scrutinizing, if the variables can predict a dependent variable accurately and how they can predict or affect the result of the dependent variable. The equation of the regression model consists of y: the dependent variable, x: the independent variable, c: the constant and b: regression coefficient, and would formulate as y = x*b + C. [3] The dependent and independent variables in the regression are often termed as Exogenous Variable, Regressors and predictor Variable for the independent variable whereas the dependent ones are termed as Endogenous Variable, Criterion Variable, Regress and and Outcome Variable. The regression model and analysis is generally used for three major factors, being:

- To know the strength of the predictors: how strong of an impact will an indepen-

dent variable have on a dependent variable such as how is social media increasing the usage of mobile phones.

- To forecast an effect: how or to what extent the dependent variable changes if the independent variable is changed such as if social media apps are banned, how much would it decrease in the number of people using mobile phones, and

- To forecast trend: to understand future trends and make predictions such as what will be the reaction if the TikTok ban was lifted. The analysis can be generated from the various kinds of linear regression that are listed below:

1. Simple Linear Regression: Where a single input of the independent variable (interval or ratio or dichotomous) is provided to generate the relationship with one dependent variable (interval or ratio).

2. Multiple Linear Regression: Where more than two inputs of the independent variable (interval or ratio or dichotomous) are provided to generate the relationship with one dependent variable (interval or ratio).

3. Ordinal Regression: Where more than one input of the independent variable (ordinal or dichotomous) is provided to generate the relationship with one dependent variable (ordinal).

4. Logistic Regression: Where more than two inputs of the independent variable (interval or ratio or dichotomous) are provided to generate the relationship with one dependent variable (dichotomous).

5. Multinomial Regression: Where more than one input of the independent variable (interval or ratio or dichotomous) is provided to generate the relationship with one dependent variable (nominal).

6. Discriminant Regression: Where more than one input of the independent variable (interval or ratio) is provided to generate the relationship with one dependent variable (nominal). While generating the model for the regression analysis we must consider the model that is fitting, as adding any more independent variable will only increase the level of variance and as a result the standard deviation. On the other hand, if too many models are fitted, overfitting can take place, as if too many large numbers of the variables are used in the analysis, some of them will be significant statistically.

3.2 K-Nearest Neighbor (KNN):

K-nearest neighbor's technique is an easy to implement machine learning algorithm (Aha et al . 1991). The question of stock prediction can be transformed into a similarity classification. Data from the historical inventory and test data was mapped into a set of vectors. Each vector represents N dimensions for each stock function. Then a similarity metric is determined, such as Euclidean distance. A definition of KNN is given in this section. KNN is called a lazy learning mechanism that doesn't

previously construct a model or function but produces the closest k records of a training data set with the highest similarity (i.e. query record) to the tests. [2] The selected k-records then take a majority vote to decide the class name, then allocate it to the query record.

The stock market closing price forecast shall be determined using KNN as follows:

a) Define the number of closest neighbors, k. b) determine the distance between the question record and the training samples. c) Sort all training records by distance values. d) Use a plurality vote for the class labels of k's closest neighbors and assign it as a query record prediction weight.

K-NN or K-Nearest Neighbor usually collects the data from a training database to make future predictions. KNN algorithm ensures similarity with distance, nearness and proximity. Based on the supervised learning technology, K-Nearest Neighbor is one of the simplest machine learning algorithms. K-NN algorithms presume that the new case/data is comparable to the available cases and place the new instance in the most comparable category to the existing categories. The KNN algorithm can compete with the most precise models because it provides very precise forecasts. Therefore, for applications with great accuracy, but without a human-readable model, you can use the KNN method.

The quality of the forecasts varies with the distance. The KNN method is therefore acceptable for use with enough domain knowledge. This understanding helps to pick a suitable measure. The KNN algorithm is a sort of lazy learning in which the prediction calculation is delayed until it is classified. Although this procedure raises computational costs compared with other algorithms, KNN remains the best option for situations in which forecasts are not frequently asked for, but accuracy is critical.

The KNN algorithm process flow would start as:

1. Insert the data (new or from a training dataset)

2. For the pre-determined neighbours, initiate the value for K

3. Then calculate the distance from the existing example to the newly input example

4. Then adding the distance and the index of the example in an order

5. Then sorting the order from ascending to descending by the distances measured

6. After which picking the first K entry made and labelling them

7. If it is classified as regression then return the mode for the entries of K or if mean then return the mean for the entries of K

However, while selecting the value for K, a few things should be kept in mind, reducing K to 1, predictions now will not be so stable and vice versa; increasing the

value will increase stability. K is selected of an odd number when the majority from the labels are picked. For the algorithm to be made no other model needs to be built or no other assumptions were to be made either and as from its name being lazy, it shows that the algorithm is quite simpler and easier to implement. The ability to use this for regression, classification and search makes it more diverse for the data to be implemented.

Therefore, being the supervised simple learning algorithm, KNN brings versatility to the algorithms from the ability to solve data of both regression, classification and search and with the right value of K chosen will simplify the process even further with more accurate results.

Where v is a value of A, Sv is the subset of instances of S where A takes the value v and S is the number of instances With the help of this node evaluation technique we can proceed recursively through the subset we create until leaf nodes have been reached throughout and all subsets are pure with zero entropy. This is how a decision tree algorithm works.

3.3 Support Vector Machine (SVM):

One of the best binary classifiers is support for Vector Machines. They create a decision limit, such that most points in one category are on the one side, while most of the points in the other category are on the other. Please take an n-dimensional vector x = (X 1 .... X n). A linear boundary (hyperplane) can be described as,

$$\beta 0 + \beta 1X1 + \cdots + \beta nXn = \beta 0 + \Sigma ni = 1\beta iXi = 0 \tag{3.3}$$

Then elements in one category are such that the sum is larger than 0, while the sum of elements in the other type is less than 0. In our grouping,

$$y \subseteq \{-1, 1\} \tag{3.4}$$

We can rewrite the hyperplane equation using inner products,

$$y = \beta 0 + \Sigma \alpha iyix(i) * x \tag{3.5}$$

Where * represents the operator of the internal product. It should be noted that the label weighs the internal product. The ideal hyperplane is such that the distance from the airplane to some point is maximized. This is called the margin. The maximum hyperplane margin (MMH) separates the data best. Since it is not a perfect distinction, however, we may add Ś 1 ... β n error variables and keep their sum below any budget B. The fundamental element is that only the closest points to the boundary matter are irrelevant for hyperplane selection. These points are known as support vectors and the hyperplane is known as the Sustainable Vector Classifier.

(SVC) since each support vector is put in the same or different classes. The SVC definition is similar to the common Ordinary Least Squares (OLS) linear regression, but both are optimizing various quantities. The OLS defines the residue, the distance of and data point to the correct line and reduces the sum of the residue squares. On the other hand, the SVC only investigates the support vectors.

And uses the internal product to optimize the hyperplane distance from the support vector. In addition, their labels weigh internal products in SVC, while OLS weights the residual square. SVC and OLS are therefore two different approaches to this problem. [1] SVCs are constrained because they are linear constraints. SVMs correct that by using non-linear kernel functions to map and linearly classify the inputs into a higher dimension space. In the original space a linear classification in the higher-dimensional space is not linear. The SVM replaces the internal product with a more general Kernel function that allows for a higher dimension mapping of the input. So, in an SVM,

$$y = \beta\_0 + \Sigma\alpha\_\text{iy}\_(\text{I})\text{K}(\text{x(i)}, \text{x}) \tag{3.6}$$

SVM (Support Vector Machine) is a supervised machine learning technique that may be used to solve classification and regression problems. It is, however, mostly employed to solve categorization difficulties. Each data item is plotted 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 certain coordinate in the SVM algorithm. Then we accomplish classification by locating the hyper-plane that clearly distinguishes the two classes.

Support Vectors are just individual observation coordinates. The SVM is a boundary that best separates both classes.

Consider machine learning algorithms as an inventory full with axes, swords, blades, bows, daggers, etc. You have different tools, but you need learn to utilize them at the proper moment. Consider "Regeneration," an analogy, as a blade that can efficiently cut and dictate the data but cannot manage very complicated data. 'Support Vector Machines' is like a sharp tool - it works on less datasets, but in creating models it may be much stronger, much more powerful on the complex ones.

Pros and Cons of SVM:

- **Pros:**

o SVM works perfectly with a clear margin of separation.

o SVM is effective in high dimensional spaces.

o SVM is effective in cases where number of samples are smaller than the number of dimensions.

o In the decision function (called support vector), SVM employs a subset of training points so it is also memory efficient.

• **Cons:**

o When we have a large data set SVM doesn't perform well because the required training time is higher.

o When the data set has more noise SVM doesn't perform well.

However, for text classification, SVMs are quite excellent. SVMs look nice at the best linear separator. The technique in the kernel makes SVM algorithms for non-linear algorithms. Selecting a kernel is the key for a decent SVM, and it is not very easy to select the correct kernel function.

3.4 Lasso Regression:

To know about 'Lasso Regression', we need to have a clear concept on regression, regularization.
Regression Regression is a technique of statistics used to determine the relationship between one variable and one or more independent variables. In plain words, an analysis of regression will tell you how different factors depend on your results.

There are many factors such as training, experiences, skills, the role of the job, the company etc. The dependent variable - salary can be predicted by means of regression analysis. y = mx+c

The dependent variable measures the independent variable in the above equation.

In terms of mathematics,

Y is the value dependent, X is the value independent, m is the pitch of the line, c is the constant value.

In machine learning or in the statistical world, the same terms are referred to as slightly different.

In machine learning or the statistical realm, the same terms are named somewhat different. Y is the forecast value, X is the characteristic value, m is the coefficient or the weight, c is the partial value. [5]

Regularization The problem of overfitting is resolved by regularization. Overfitting is responsible for the low precision of the model. It occurs when both data and noises of the training set are heard by the model. In a training set, noises are random data and do not represent the actual data characteristics.

Y C0 + C1X1 + C2X2 + . . . + CpXp

Y stands for a dependent variable, X stands for independent variables and C for various variables in the following linear regression equation is the coefficient estimate.

The fitting of the model involves a loss function called squares. In order to reduce the loss function to a minimum, the equation coefficients are chosen. False coefficients are chosen when the training set contains many irrelevant data. For model forecasts in the future, this will not go well. In cases like this, we can regularize or decrease these misunderstood coefficients to zero with regularization. The Lasso regression is one of the popular techniques for improving the performance of the model.

The regression in Lasso is like a linear regression, but uses a shrinkage technique where the determination coefficients are shrunk to nil. As noted in the dataset, linear regression gives you regression coefficients. With lasso regression, these coefficients can be shrunk or regulated to prevent overfitting and make them work better for different datasets. This regression type is used if the data set shows high multi-collinearity or if the elimination and selection of the variable is automated.

The choice of a model depends on the dataset and the problem statement. The dataset and how the features interact with each other are essential to understand. The regression of Lasso penalizes less important characteristics in your dataset and zero, eliminating them. It therefore offers you the benefit of selecting the feature and creating a simple model.

Therefore, lasso regression can be used when the data set has high dimensionality and high correlation.

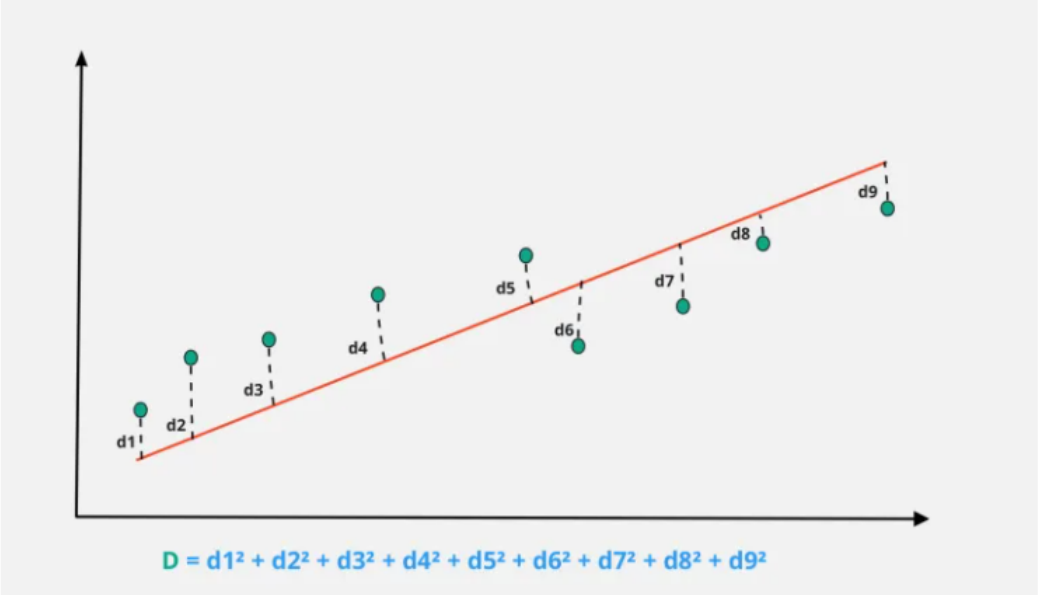


Figure 3.1: Statistics of Lasso Regression

The distance between the real information points and the model line in the pre-

ceding graph is represented by d1, d2, d3, etc.

The least square shall be the sum of the squares between the points of the plotted curve. The best model is chosen to minimize the least-squares in linear regression. We add a penalizing factor to the lowest squares while regressing. The model is therefore chosen to minimize the underlying loss function.

D = least-squares + lambda * summation (absolute values of the magnitude of the coefficients)

All estimated parameters include the lasso regression penalty. Any value between zero and infinity can be Lambda. This value determines how to regularize the system aggressively. It is often selected with cross-validation. Lasso penalizes the sum of the coefficients' absolute values. The coefficients decrease and ultimately get zero as the lambda value increases. This eliminates insignificant variables of our model from lasso regression. Our regularized model may be slightly biased compared to the linear regression, but for future predictions less.

3.5 Multi Linear Regression:

Like the usual regression model, multilinear regression models are also used to understand the relationship between dependent and independent variables, uniquely multilinear regression model uses two or more independent variables with one deponent variable to find the relationship between the variables. This enables to find out the strength of the relationship between the variables, the effect the independent variable will have on the dependent variable and to predict the trends for the future with a specific value.

This regression model forms a formula with y: the predicted dependent variable, B0: the y-intercept, B1X1: the regression coefficient of the first independent variable, BnXn: regression coefficient for the last independent variable and e: the model error, turning into y = B0 + B1X1 + BnXn and the estimates or values found from the model are known as coefficients. This is used to calculate the specified t-static for the entire model along with the p-value that is associated with the model and the regression coefficients so that the error value is at minimum. [4]

It is assumed that in the case of the multilinear regression model, the variance is often similar or constant as the size of the error does not change much, however, the accuracy is maintained as the observations are made independently, making no hidden relationship between the variables existent as the data usually follow a normal distribution and the linearity makes it more clear to understand the trend as it is a straight line rather than a curve or scattered. It is also pre-assumed that the independent variables do not have a high correlation and the best way to test this assumption would be the factor method. The residuals in the case of multilinear regression tend to be approximately rectangular in shape. Mostly multilinear regression model uses statistical software, but it can be completed by hand as well and the best-fitted line is plotted at the centre of the analysis for the model, the line is fitted through multi-dimensional points of data to represent the best outcome.

For example, a motor vehicle company can use multilinear regression to get a better understanding of their customer base, with the independent variables being the size of the car, model of the car, year of launch, etc. with the dependent variable being price of the car.

In the case of understanding the correlation between the independent and dependent variables, the multiple linear regression plays a greater impact while examining the data the relative influence of the variables on one another can be easier to predict with strong positive correlation to negative correlation. On the other hand, the ability to which the model helps to determine outliers makes it more efficient due to the generalizable results. The model is also able to deal with a large dataset and provide an effective output easily, due to the ability to provide a fair prediction based on the significance level of the dataset.

Multiple regressions go a step further and two or more independent variables will take place instead of one. Multiple applications for machine learning based on regression can (at least) be grouped into three different problem domains:

1) Quantifying relationships: We know that the strength (expressed in a number between zero and one) of the relationship between the two variables can be found by correlation with the direction (positive or negative). But multiple regressions go further and quantify the relationship in fact.

2) Prediction: Everything about prediction is implemented machine learning. Since regression can quantify relationships, this ability can be converted into the solution of prediction problems. This means we should be able to predict, using a regression equation, what salary levels should be for this individual if the level of education and the years of experience are known to us.

3) Forecasting: The third problem domain is predicting time series analysis for multiple regression. Two powerful predictive techniques that employ principles of multiple regression include vector autoregression and panel data regression.

Chapter 4

Data and Preliminary Analysis

4.1 Analyzing the Data using the Algorithms

For the data analysis of Bangladesh stock market, we have written a code which will predict the increase and decrease of the price of a particular company's stock. From the data set of various companies, we get to know about their opening price, closing price, high price, low piece and volume. Based on this information we have applied liner regression, K-nearest neighbor, SVM, LASSO regression and Multi linear regression. Here is some scenario of Square Pharmaceuticals:

Figure 4.1: Data Summary of Square Pharmaceuticals

From the above data summary, we can see the mean value, standard value, min value, max value of volume, open price, close price, high price and low price.

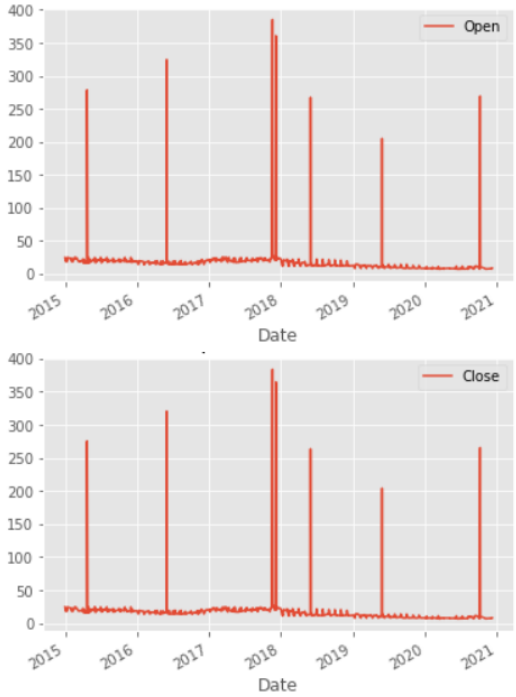


Figure 4.2: Opening and Closing Price of Square Pharmaceuticals

Based on opening and closing prices of Square we can see a graph like above. Here we used the data from 2015 to 2021. Through the above graph we can see the various fluctuation of prices. Among the years, in 2018 the opening and closing prices were so high and it's near about to 400.

Figure 4.3: Density Plot of Opening and Closing of Square Pharmaceuticals

Here we plot the density of opening and closing prices. In X-axis we took prices and in Y-axis we took the density. We can see that price between 0 to 45 the density is so high. From price 0 the curve was growing but after the price 45, the curve was going downward.

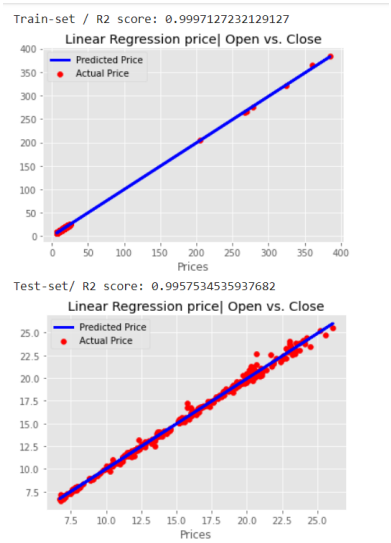


Figure 4.4: Train Set and Test Set of Square Pharmaceuticals

After applying linear regression, we got the train-set score and test-set score. Here we can see that our predicted price line is very static, a blue colored straight line. But the problem is, sometimes the prices are high but the prediction price is below the actual price and also sometimes the prices are low but the prediction price is above the actual price. So, it's not giving the accurate prediction of prices.

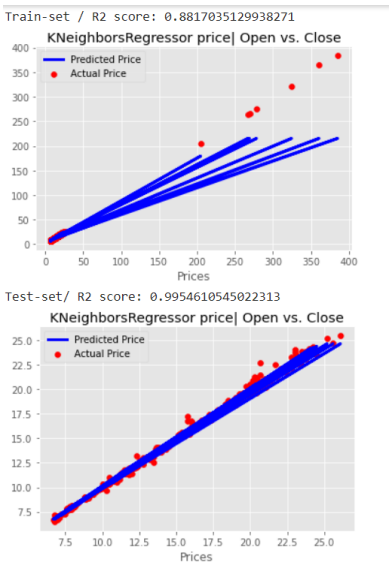


Figure 4.5: Train Set and Test Set of Square Pharmaceuticals using K Nearest Neighbours

Upper graph: Depicts observations of train data set. Lower graph: Depicts observations of Test data set. Where each of the 10 lines (predicted Price) lies under a specific area observes and calculates the nearest value of these areas. However, there is a slight difference in both results.

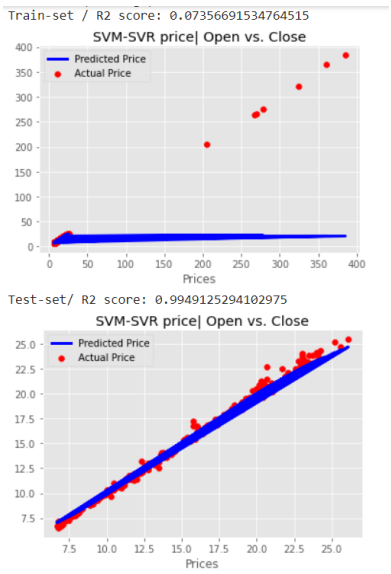


Figure 4.6: Train Set and Test Set of Square Pharmaceuticals using Support Vector Machine

From the above graph we can see the huge difference of prediction score and also the differences of prediction price lines. In linear and k-nearest neighbor regression we saw that our prediction price line was very static. But here in SVM our prediction price lines of train-set can't predict accurately. Which is a very huge problem.

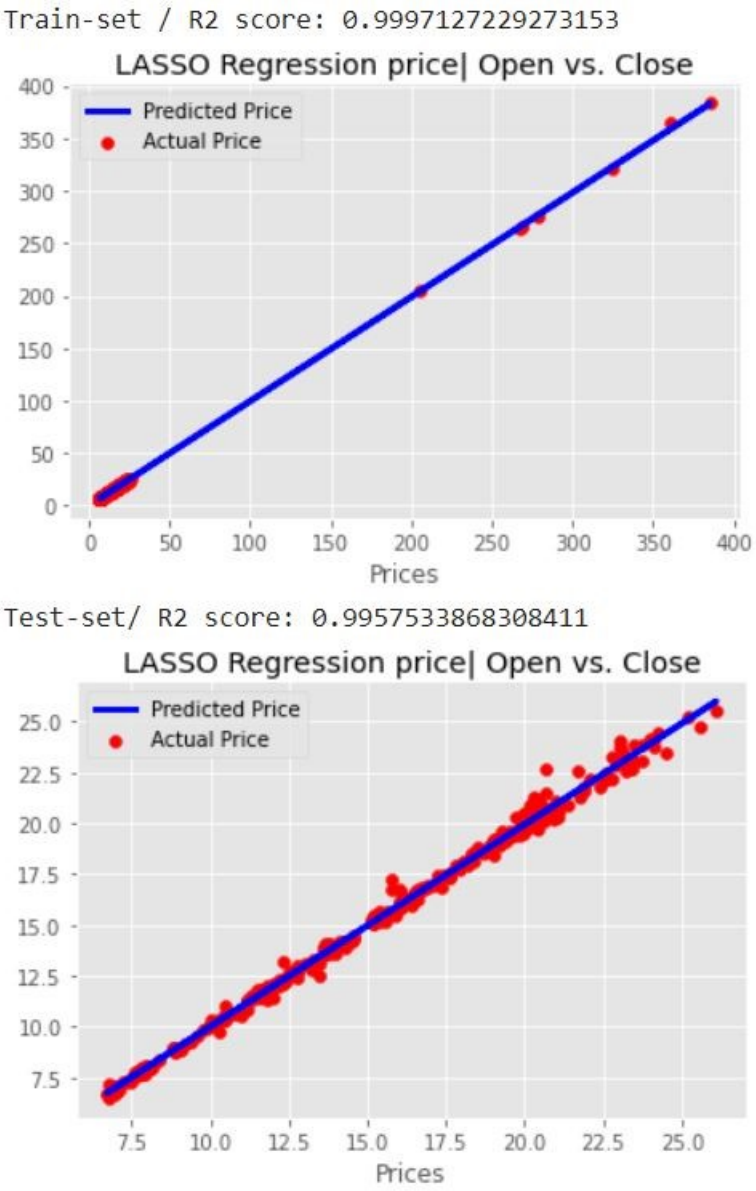


Figure 4.7: Train Set and Test Set of Square Pharmaceuticals using Lasso Regression

From the above graph we can see that it's kind of similar to linear regression but lasso regression is used for more accurate prediction. The score of train and test set of lasso regression is more accurate or closer to 1 than the train and test set score of linear regression. We can say it gives us better prediction of prices than linear regression. From the above graph we can see that it's kind of similar to linear regression but lasso regression is used for more accurate prediction. The score of train and test set of lasso regression is more accurate or closer to 1 than the train and test set score of linear regression. We can say it gives us better prediction of prices than linear regression.

OLS Regression Results

Dep. Variable:	Close	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	2.600e+07
Date:	Mon, 10 May 2021	Prob (F-statistic):	0.00
Time:	05:01:54	Log-Likelihood:	-20285.
No. Observations:	6734	AIC:	4.058e+04
Df Residuals:	6731	BIC:	4.060e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
0	-0.2777	0.008	-32.860	0.000	-0.294	-0.261
1	0.6752	0.007	98.897	0.000	0.662	0.689
2	0.6002	0.006	92.395	0.000	0.587	0.613

Omnibus:	8604.220	Durbin-Watson:	1.784
Prob(Omnibus):	0.000	Jarque-Bera (JB):	22641698.029
Skew:	6.012	Prob(JB):	0.00
Kurtosis:	286.814	Cond. No.	159.

Figure 4.8: Train Set and Test Set of Square Pharmaceuticals using Multi Linear Regression

Here is some scenario of Summit Alliance Port Limited:

Figure 4.9: Data Summary of Summit Alliance Port Limited

From the above data summary, we can see the mean value, standard value, min value, max value of volume, open price, close price, high price and low price.

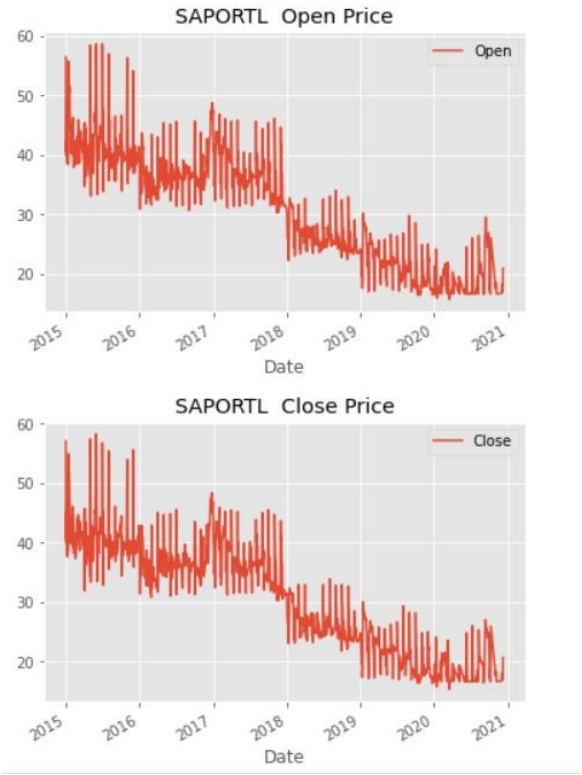


Figure 4.10: Opening and Closing Prices of Summit Alliance Port Limited

Based on opening and closing prices of SAPORTL we can see a graph like above. Here we used the data from 2015 to 2021. Through the above graph we can see the various fluctuation of prices. Year 2015 to 2016 the opening and closing prices were so high. But gradually it's going down.

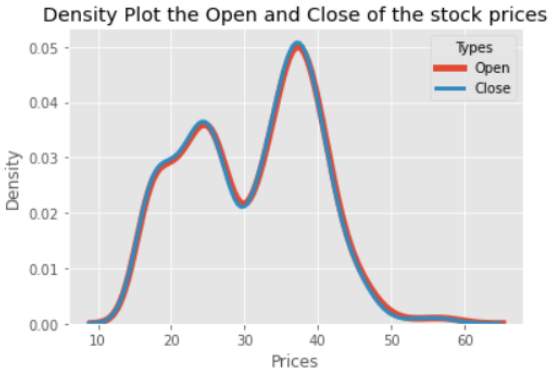


Figure 4.11: Density Plot of Opening and Closing of Summit Alliance Port Limited

Here we plot the density of opening and closing prices. In X-axis we took prices and in Y-axis we took the density. We can see that price between 30 to 40 the density is so high. From price 40, the curve was going downward.

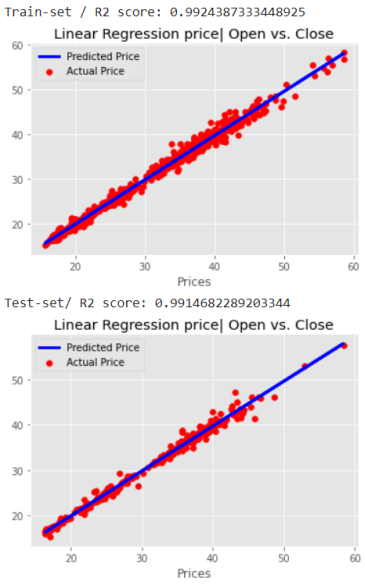


Figure 4.12: Train Set and Test Set of of Summit Alliance Port Limited using Linear Regression

After applying liner regression algorithm, we got the train-set score and test-set score. Here we can see that our predicted price line is very static, a blue colored straight line. But the problem is, sometimes the prices are high but the prediction price is below the actual price and also sometimes the prices are low but the prediction price is above the actual price. So it's not giving the accurate prediction of prices.

Figure 4.13: Train Set and Test Set of of Summit Alliance Port Limited using K Nearest Neighbour

Upper graph: Depicts observations of train data set. Lower graph: Depicts observations of Test data set. Where each of the 10 lines (predicted Price) lies under a specific area observes and calculates the nearest value of these areas. However, there is a slight difference in both results.

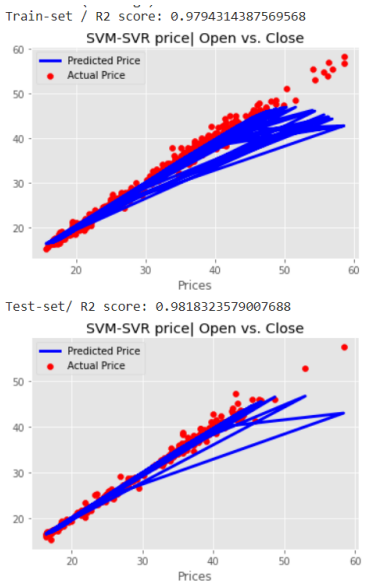


Figure 4.14: Train Set and Test Set of of Summit Alliance Port Limited using Support Vector Machine

From the above graph we can see the huge difference of prediction score and also the differences of prediction price lines. In linear and k-nearest neighbor regression we saw that our prediction price line was very static. But here in SVM our prediction price lines are kind of scattered. For that reason, we are not getting the accurate prediction of prices.

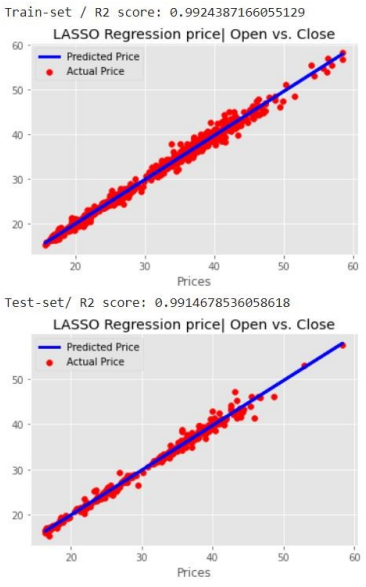


Figure 4.15: Train Set and Test Set of of Summit Alliance Port Limited using Lasso Regression

From the above graph we can see that it's kind of similar to linear regression but lasso regression is used for more accurate prediction. The score of train and test set of lasso regression is more accurate or closer to 1 than the train and test set score of linear regression. We can say it gives us better prediction of prices than linear regression.

OLS Regression Results

Dep. Variable:	Close	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	5.174e+06
Date:	Mon, 10 May 2021	Prob (F-statistic):	0.00
Time:	05:00:08	Log-Likelihood:	-284.04
No. Observations:	1348	AIC:	574.1
Df Residuals:	1345	BIC:	589.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
0	-0.5255	0.017	-30.096	0.000	-0.560	-0.491
1	0.7398	0.015	48.359	0.000	0.710	0.770
2	0.7830	0.015	52.851	0.000	0.754	0.812

Omnibus:	277.448	Durbin-Watson:	1.957
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2550.706
Skew:	0.679	Prob(JB):	0.00
Kurtosis:	9.601	Cond. No.	147.

Figure 4.16: Train Set and Test Set of of Summit Alliance Port Limited using Multi Linear Algorithm

Here is some scenario of First Bank:

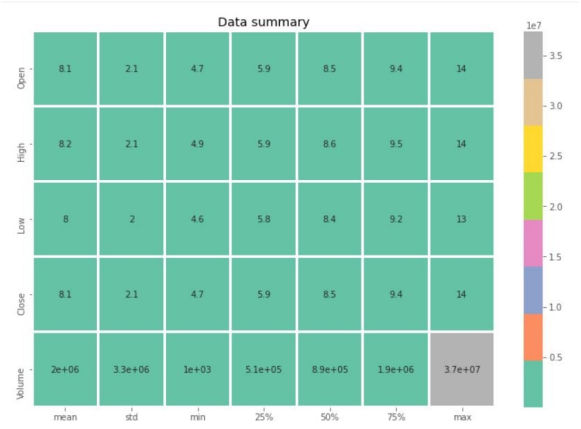


Figure 4.17: Data Summary of First Bank

From the above data summary, we can see the mean value, standard value, min value, max value of volume, open price, close price, high price and low price.

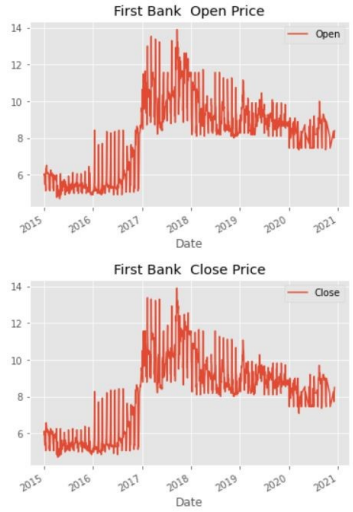


Figure 4.18: Opening and Closing Prices of First Bank

Based on opening and closing prices of First Bank we can see a graph like above. Here we used the data from 2015 to 2021. Through the above graph we can see the various fluctuation of prices. Year 2017 to 2018 the opening and closing prices were so high.

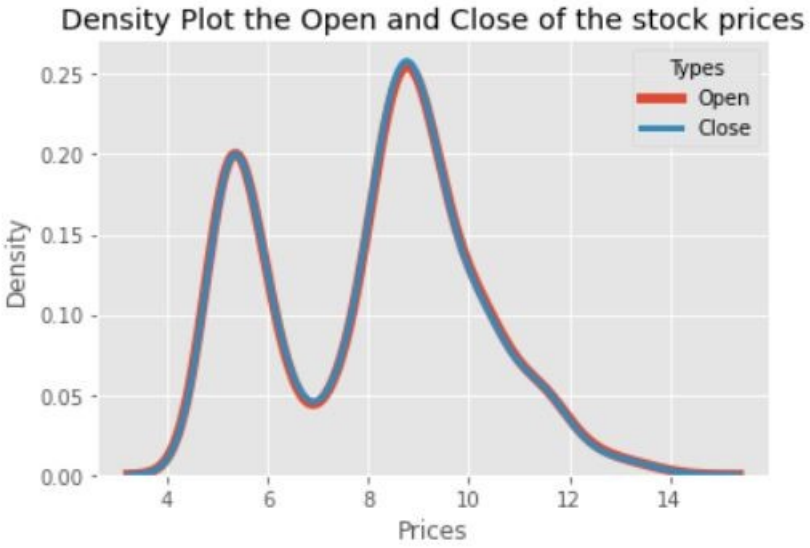


Figure 4.19: Density Plot of First Bank

Here we plot the density of opening and closing prices. In X-axis we took prices and in Y-axis we took the density. We can see that price between 8 to 10 the density is so high. From price 8 the density curve is growing but after the price 9, the curve was going downward.

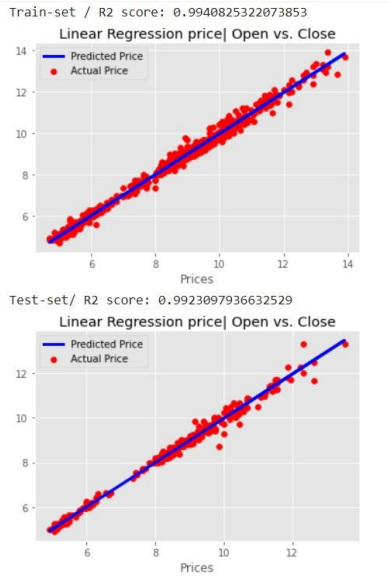


Figure 4.20: Train Set and Test Set of First Bank using Linear Regression

After applying linear regression algorithm, we got the train-set score and test-set score. Here we can see that our predicted price line is very static, a blue colored straight line. But the problem is, sometimes the prices are high but the prediction price is below the actual price and also sometimes the prices are low but the prediction price is above the actual price. So it's not giving the accurate prediction of prices.

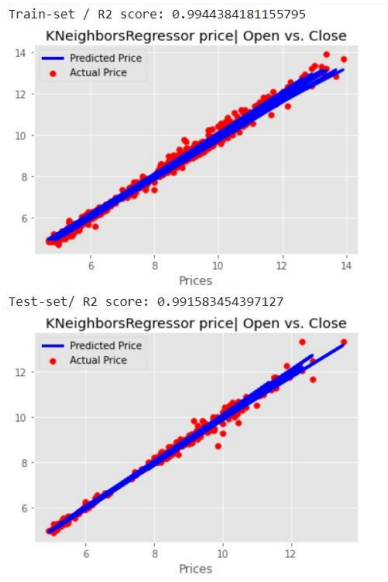


Figure 4.21: Train Set and Test Set of First Bank using K Nearest Neighbour

Upper graph: Depicts observations of train data set. Lower graph: Depicts observations of Test data set. Where each of the 10 lines (predicted Price) lies under a specific area observes and calculates the nearest value of these areas. However, there is a slight difference in both results.

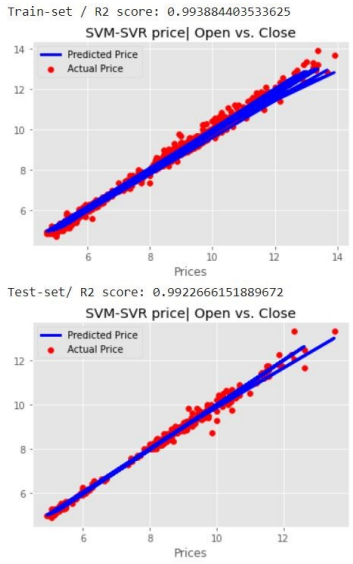


Figure 4.22: Train Set and Test Set of First Bank using Support Vector Machine

From the above graph we can see the huge difference of prediction score and also the differences of prediction price lines. In linear and k-nearest neighbor regression we saw that our prediction price line was very static. But here in SVM our prediction price lines are kind of curvy. For that reason, we are not getting accurate prediction of prices.

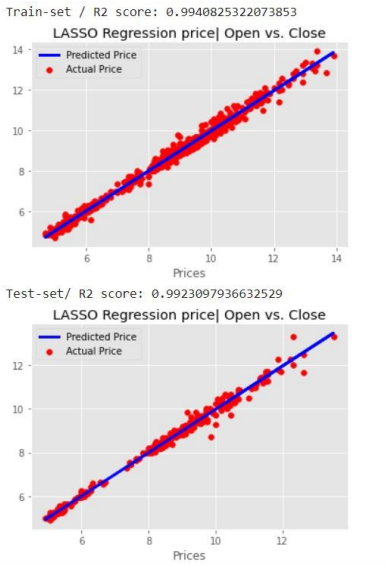


Figure 4.23: Train Set and Test Set of First Bank using Lasso Regression

From the above graph we can see that it's kind of similar to linear regression but lasso regression is used for more accurate prediction. The score of train and test set of lasso regression is more accurate or closer to 1 than the train and test set score of linear regression. We can say it gives us better prediction of prices than linear regression.

OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	6.646e+06
Date:	Mon, 10 May 2021	Prob (F-statistic):	0.00
Time:	04:58:35	Log-Likelihood:	1701.3
No. Observations:	1352	AIC:	-3397.
Df Residuals:	1349	BIC:	-3381.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
0	-0.4620	0.018	-25.325	0.000	-0.498	-0.426
1	0.7453	0.016	47.350	0.000	0.714	0.776
2	0.7156	0.016	43.414	0.000	0.683	0.748

Omnibus:	69.603	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	247.500
Skew:	-0.024	Prob(JB):	1.80e-54
Kurtosis:	5.096	Cond. No.	173.

Figure 4.24: Train Set and Test Set of First Bank using Multi Linear Regression

Here is some scenario of Marico Bangladesh:

Figure 4.25: Data Summary of Marico Bangladesh

From the above data summary, we can see the mean value, standard value, min value, max value of volume, open price, close price, high price and low price.

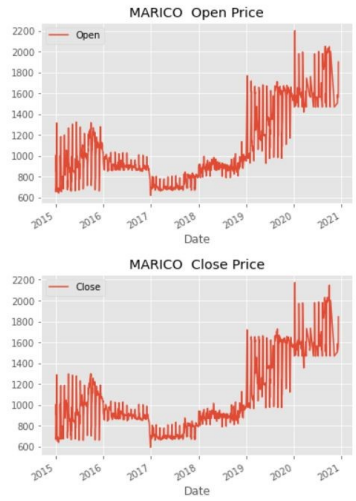


Figure 4.26: Opening and Closing Prices of Marico Bangladesh

Based on opening and closing prices of MARICO we can see a graph like above. Here we used the data from 2015 to 2021. Through the above graph we can see the various fluctuation of prices. Year 2020 to 2021 the opening and closing prices were so high.

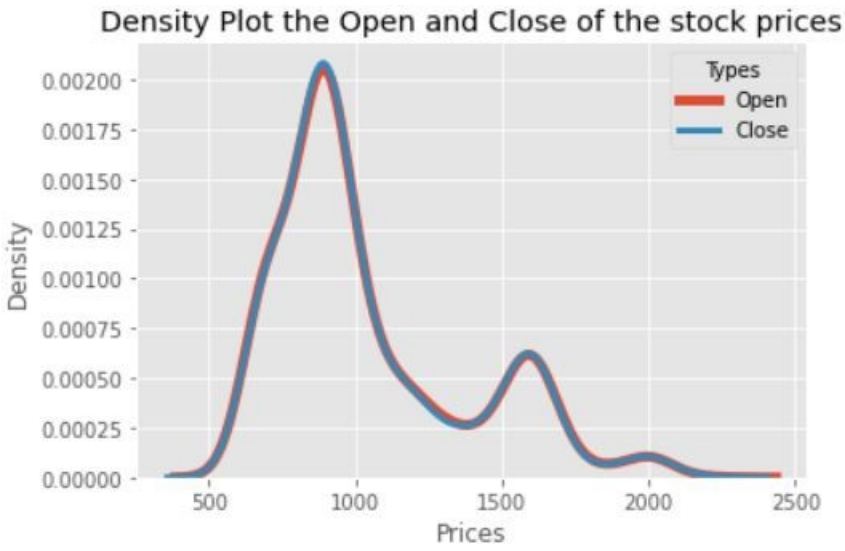


Figure 4.27: Density Plot of Marico Bangladesh

Here we plot the density of opening and closing prices. In X-axis we took prices and in Y-axis we took the density. We can see that price between 800 to 1000 the density is so high. From 500 the density curve is growing up when it reached approximately 900, the curve was going downward.

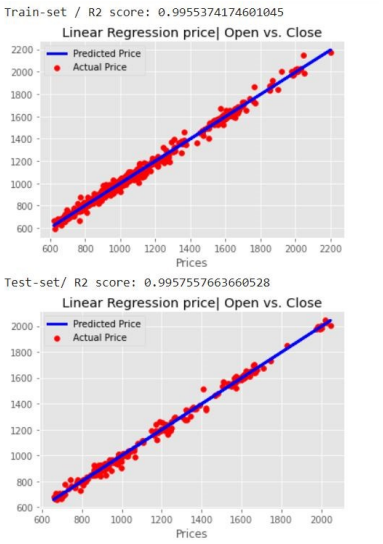


Figure 4.28: Train Set and Test Set of of Marico Bangladesh using Linear Regression

After applying liner regression algorithm, we got the train-set score and test-set score. Here we can see that our predicted price line is very static, a blue colored straight line. But the problem is, sometimes the prices are high but the prediction price is below the actual price and also sometimes the prices are low but the prediction price is above the actual price. So it's not giving the accurate prediction of prices.

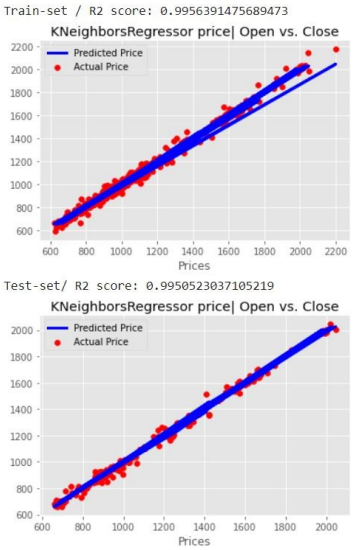


Figure 4.29: Train Set and Test Set of of Marico Bangladesh using K Nearest Neighbour

Upper graph: Depicts observations of train data set. Lower graph: Depicts observations of Test data set. Where each of the 10 lines (predicted Price) lies under a specific area observes and calculates the nearest value of these areas. However, there is a slight difference in both results.

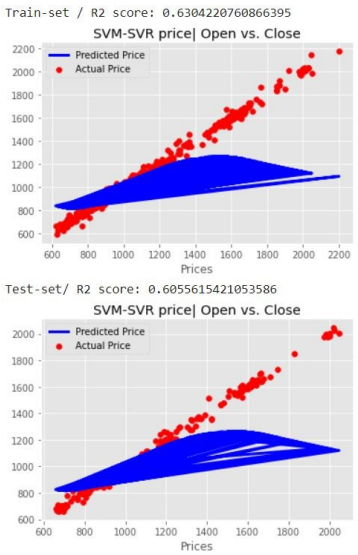


Figure 4.30: Train Set and Test Set of of Marico Bangladesh using Support Vector Machine

From the above graph we can see the huge difference of prediction score and also the differences of prediction price lines. In linear and k-nearest neighbor regression we saw that our prediction price line was very static. But here in SVM our prediction price lines are kind of curvy. For that reason, we are not getting accurate prediction of prices.

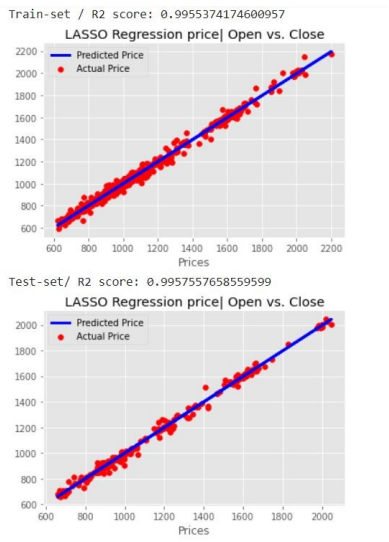
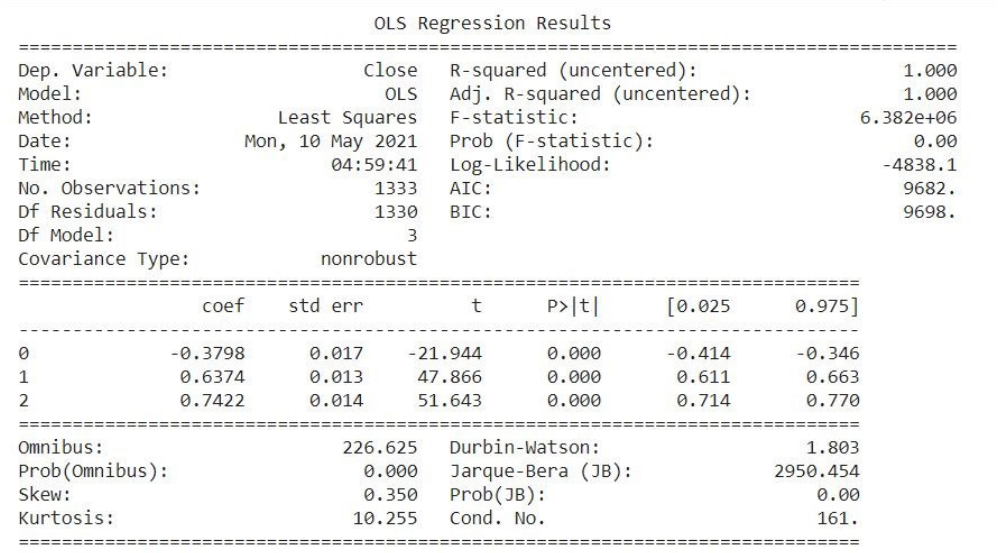


Figure 4.31: Train Set and Test Set of Marico Bangladesh using Lasso Regression

From the above graph we can see that it's kind of similar to linear regression but lasso regression is used for more accurate prediction. The score of train and test set of lasso regression is more accurate or closer to 1 than the train and test set score of linear regression. We can say it gives us better prediction of prices than linear regression.



OLS Regression Results

Dep. Variable:	Close	R-squared (uncentered):	1.000
Model:	OLS	Adj. R-squared (uncentered):	1.000
Method:	Least Squares	F-statistic:	6.382e+06
Date:	Mon, 10 May 2021	Prob (F-statistic):	0.00
Time:	04:59:41	Log-Likelihood:	-4838.1
No. Observations:	1333	AIC:	9682.
Df Residuals:	1330	BIC:	9698.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
0	-0.3798	0.017	-21.944	0.000	-0.414	-0.346
1	0.6374	0.013	47.866	0.000	0.611	0.663
2	0.7422	0.014	51.643	0.000	0.714	0.770

Omnibus:	226.625	Durbin-Watson:	1.803
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2950.454
Skew:	0.350	Prob(JB):	0.00
Kurtosis:	10.255	Cond. No.	161.

Figure 4.32: Train Set and Test Set of Marico Bangladesh using Multi Linear Regression

4.2 Future Plan

In the near future we want to make a user friendly website which will predict the data and told the user which shares are in the top and which shares are's in below. In the website there will be all the previous data from the different companies which will help him to predict it properly. Top most when a user want to invest his/her money is stock market and he/she visit our website they will be so benefited. Because our website help him/her out to understand the market and told him/her where to invest after predict the data from the market. In our website we will also add dailies opening price and closing price of every stocks. So user can be use on daily basis for checking the opening and closing price of the stock's.

Bibliography

[1] S. Vishwanathan and M. N. Murty, "Ssvm: A simple svm algorithm," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, IEEE, vol. 3, 2002, pp. 2393–2398.

[2] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[3] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012, vol. 329.

[4] G. K. Uyanık and N. Güler, "A study on multiple linear regression analysis," *Procedia-Social and Behavioral Sciences*, vol. 106, pp. 234–240, 2013.

[5] J. Ranstam and J. Cook, "Lasso regression," *Journal of British Surgery*, vol. 105, no. 10, pp. 1348–1348, 2018.

[6] M. Karimuzzaman, N. Islam, S. Afroz, and M. M. Hossain, "Predicting stock market price of bangladesh: A comparative study of linear classification models," *Annals of Data Science*, vol. 8, no. 1, pp. 21–38, 2021.