

BRAC University

CSE422: Artificial Intelligence, Lab Project Report

**Simple AI Machine Learning Project**

[Per day Profit/Loss estimation of Stock Market]

**Submitted To:**

Mr. Mehran Hossain (MRH) && Ms. Zarin Tahia Hossain (ZTH)

C. Lecturer, CSE Department, BRAC University

**Submitted By:**

Group 1 - CSE422 Section 6 Lab

**Group Member’s Information:**

|  |  |  |  |
| --- | --- | --- | --- |
| S. No. | Name | Student ID | G-suite Mail |
| 1 | Ashok Lamichhane | 21201785 | ashok.lamichhane@g.bracu.ac.bd |
| 2 | Mukund Prasad Singh | 20201202 | mukund.prasad.singh@g.bracu.ac.bd |
| 3 | Udoy Saha | 23341134 | udoy.saha@g.bracu.ac.bd |
| 4 | Noor E Jannat Nafia | 21301631 | noor.ejannat.nafia@g.bracu.ac.bd |

Table of Contents

**Contents Page Number**

1. Introduction……………………………………………..…………………………………3
2. Dataset Description……………………………………….……………………………….3
3. Dataset Preprocessing……………………………………………………………….…….4
4. Feature Scaling………………….……………………………………...…………….……4
5. Dataset Splitting……………………………………………………………..……….……4
6. Model Training & Testing……………………………………………..…………….……5

* Linear Regression …………………………………….…………………………..5
* Support Vector Machine (SVM)…………………………………………………..5
* Decision Tree…………………………………………………….………………..6
* Random Forest……………………………………..……………………………..6
* KNN……………………………………………………………………...………..6

1. Comparison Analysis……………………………………………………………………...7
2. Conclusion………………………………………………………………………………...7
3. **Introduction**

In order to gain useful insights, this research entails doing a thorough examination of the stock market using historical price data, financial indicators, and market patterns. The initiative tries to find patterns, connections, and prospective investment possibilities using data visualization and statistical approaches. The study will give a greater knowledge of market dynamics and assist in guiding strategic investment decisions by looking at the performance of various stocks and taking into account external factors like economic data and news mood. In this model we are actually evaluating the actual loss or profit when we buy stock at opening price and sell at closing price on any particular day.

1. **Dataset Description**

We found out this data set from the website called: Kaggle.com

Used dataset: <https://drive.google.com/file/d/1vfBbadKNOCKoH4u_pIzUWn7_nz7DBoXo/view?usp=sharing>

Features: This dataset of stock market have 7 Features:

* Company Name
* Date
* Closing Price
* Volume
* Opening Price
* High Price
* Low price

Type: It is a regression type of problem because in this project we need to predict stock price with profit and loss. There is no category, all the necessary features are numerical, so it is a Regression problem.

Number of Data points: 25160

Feature type: All features in our dataset are of quantitative type. The qualitative features are also logically converted into quantitative data.

Correlation: We related all the features but especially we used Opening Price and Closing Price for finding profit and loss. The Opening and Closing price will be predicted. High and Low prices are highly correlated with those. Since those columns add no extra information, those are dropped.

Our dataset has 10 companies' data and it is balanced by an almost equal number of data in which we can see the companies data at a glance by giving the company name as input.

1. **Data Preprocessing**

Faults:

* Null Values in some features

Solution: NULL values are replaced by Mean values imputed according to each column. For doing this first we find out the mean values and we replace them in the place of NULL values. We used a fit function for that.

* Dollar Symbol before values

Solution: Using panda's function we removed the dollar symbol and converted all values to appropriate float values.

* The values of Date column needs to be formatted to Pandas 'datetime' format

Solution: We used pandas function and lambda function to convert all values to appropriate values. We did it for a date and also for others.

1. **Feature Scaling**

In this project we used features like open price, close price, volume, date high price and low price. Among them all are fixed and have the same ratio values in floating point values but time and volume don’t have. So, we used two types of feature scaling for scaling the. We used MinMaxScaler, StandardScaler for scaling the date and volume of stocks.

1. **Dataset Splitting**

Training Set → 70%

Testing Set → 30%

1. **Model Training & Testing**
2. Linear Regression

Incorporating the fundamental ideas of linear regression, our effort seeks to create a reliable framework for forecasting. We want to create a model that reliably predicts outcomes by investigating the connections between important variables. We will thoroughly examine the data to ascertain the direction and intensity of associations in order to build a solid linear regression model. This model will be a useful tool for generating informed judgments and projections as well as for gaining insightful understanding of the underlying dynamics of the data. Our research emphasizes the value of linear regression as a fundamental method for data analysis and forecasting, providing a direct route to better comprehension and foresight in challenging situations. We used following equation form:

y = mx + c in which m is slope and c is y-displacement

1. Support Vector Machine (SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. Non-Linear SVM can be used to classify data when it cannot be separated into two classes by a straight line (in the case of 2D). By using kernel functions, nonlinear SVMs can handle nonlinearly separable data. The original input data is transformed by these kernel functions into a higher-dimensional feature space, where the data points can be linearly separated.

1. Random Forest

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Random Forest works in two-phase first is to create the random forest by combining N decision trees, and second is to make predictions for each tree created in the first phase.

1. Decision Tree

A decision tree is one of the important supervised learning algorithms for dealing with classification and regression problems. The structure of the decision tree algorithm is basically compared to an actual tree. The branches in the tree represent the decisions while the leaf represents the result of the decision. It can handle both categorical and continuous data. The main goal of the decision tree is Build a decision tree to classify examples as positive or negative instances of a concept using supervised learning from a training set. A decision tree is a tree where each non-leaf node has associated with it an attribute (feature) each leaf node has associated with it a classification (+ or -) each arc has associated with it one of the possible values of the attribute at the node from which the arc is directed .

1. KNN

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. The KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. First of all, we have to choose the number K of neighbors, then Calculate the Euclidean distance between K neighbors, take the K closest neighbors based on the Euclidean distance.Count the number of data points in each category among these k neighbors. And finally we have to assign the new data points to the category with the highest number of neighbors.

.

**7. Model Selection & Comparison Analysis**

As we did with different five models we got different types of results so comparing them.

* By Model Accuracy

In the linear regression model we got almost 83% accuracy whereas from the SVR model we got more accuracy 93%. When we use Decision Tree, Random Forest, KNN model we got accuracy 99.3%, 99.4% & 98.9% respectively. It seems among our five models Decision Tree and Random forest have better performance.

* By Profit Got

We calculated the profit using our five different models by just calculating/using the open price and close price. The results seem like 50.33% for Linear Regression, 53.24% for SVR, 47.81% for Decision Tree, 49.60% for Random Forest, 45.96% for KNN. It seems SVR predicts better profit than other models.

**8. Conclusion**

In Conclusion we can see that Stock market has both loss and profits at a certain time period but our dataset has very good companies and their prices so we have profit in our model’s predictions. By evaluating all the five models (Linear Regression, SVR, Decision Tree, Random Forest, KNN) we can say that our five models are working properly fine but decision tree and random forest have more accuracy than the others. When we visualize our data with regards to profit KNN predicts high profits but in reality a good model is KNN as it has less profit and we know stock markets don’t always have profit.