A

Project Report on

"Stock Market Analysis Using Machine Learning"

Submitted by

Sautrik Chaudhuri(213040032)

Somya Sahana(213040038)

Water Resources Engineering

Department of Civil Engineering

Indian Institute of Technology,

Bombay

(2022)

Table of Contents:

1. Introduction

One of the many buzzwords in machine learning nowadays is FINTECH. Financial analysts and data scientists all over the world have tried to make sense of the market to find a way to increase their return on investments. However due to various problems, like temporal variability and enormous scale of the system it has proved to be incredibly overwhelming for them even with state of the art data analytic tools.

This project deals with classification of the stock performance using OHLC (O-open, H-high, L-low and C-close) data of a stock comprising 10,000 data points. Conventional machine learning algorithms like Logistic Regression, Random Forest Classifier, Decision Tree Classifier and KNN Classifier have been used to predict the target variable 'y'. Results show a 53% accuracy on the testing data and 54% accuracy for the training data. This report summarizes the methodology, model description, outcomes and references.

2. Research

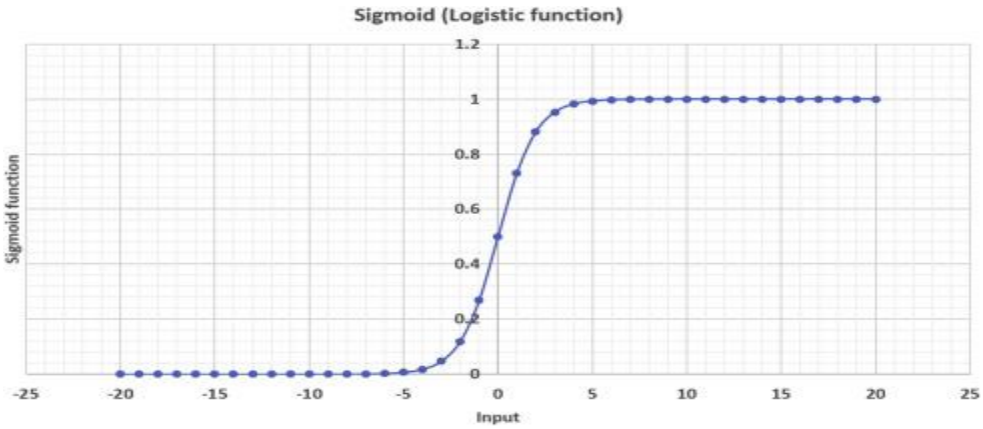
OHLC represents the opening, high, low and close in a stock for a time interval. Often the closing price is considered to be the most important by the traders. The chart type can be increasingly useful when depicting the momentum. When the open and close remain far apart, it indicates a strong momentum and when they are close, it shows indecision or weak momentum.

Classification algorithms can be applied on the OHLC data to derive important relationships. Some of the most important classification algorithms used in stocks trading and predictions include:

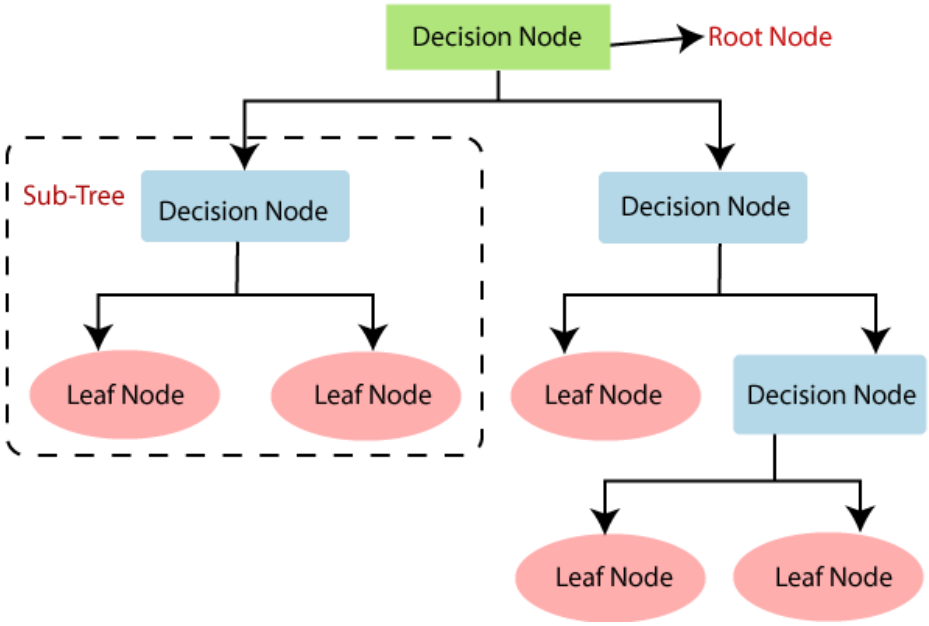
i) Logistic Regression: It is a supervised learning technique used for predicting a categorical dependent variable using a set of independent variables. It is used in classification algorithms as it gives probabilistic values which lie between 0 and 1. Instead of fitting a regression line the algorithm maps the independent variables

using a ‘S’ shaped sigmoid function. Hence this algorithm has been used in the analysis.

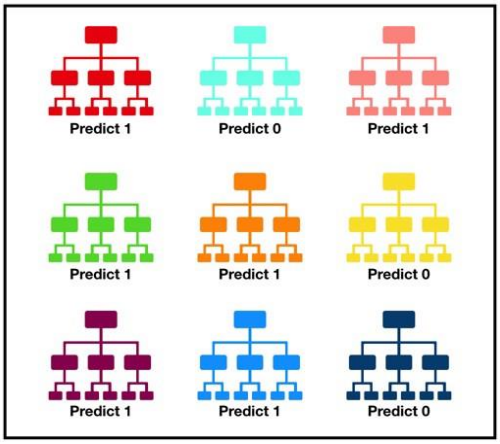
ii) Decision Tree Classifier: Decision Tree is a supervised learning algorithm that consists of a tree structure classifier, where the internal nodes represent the



feature variables, branches- the decision rules and each leaf node- the outcome. It simply answers the question ‘yes’ or ‘no’ to further split the tree into subtrees. It provides a graphical representation of all possible solutions to a problem based on given conditions.

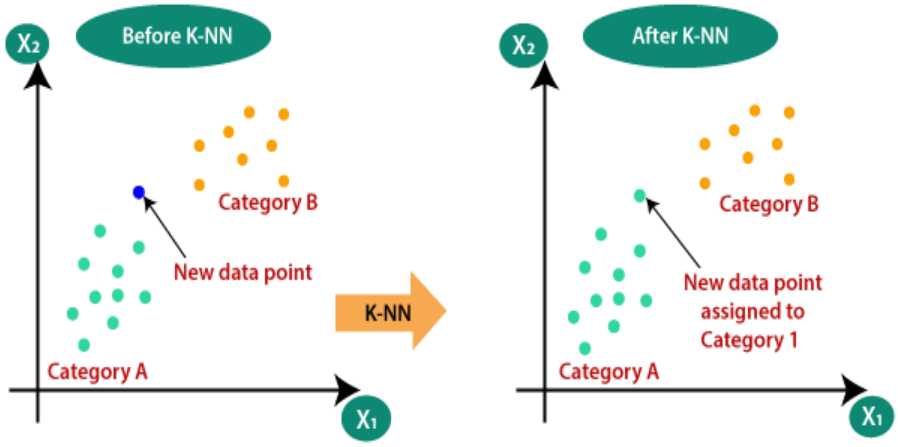


iii) Random Forest Classifier: Random Forest is an extension of the decision tree algorithm, where decisions from a number of decision trees on various subsets of the dataset are averaged to make a prediction. It takes less training time as compared to other algorithms even for large datasets generally with high accuracy.



Tally: Six 1s and Three 0s
Prediction: 1

iv) K-Nearest Neighbor(KNN): It is a supervised learning technique, which classifies new data based on the maximum similarities with the available categories (in our case 0 and 1). It is a nonparametric algorithm meaning it doesn't take any assumptions on underlying data. It is quite robust on noisy training data.



Momentum of stocks: It is an important characteristic used to define the rate of acceleration of a stock's price. Strong momentum indicates an upward trend and is often used by trading strategists to predict stock prices. This analysis uses momentum as a feature variable for identifying ‘y’.
Momentum(For A particular stock) = Close Price(Today) - Close Price(N-day ago)

3. Methodology

We performed the project on python interface. After importing the data, preprocessing has been done where we checked for missing values and outliers

in our data. Next we did basic EDA to explore relationships among the feature variables and the target variable. We also checked the correlation among each of the variables and we found that the target variable did not have any significant correlation with the features. Following this moving average was done over the data set to smoothen the noise. Different classification algorithms were applied where it was found that the model did not work well. So, we found the momentum for the stock and reevaluated the accuracy scores of the models where we found better results.

4. Model Description

Feature variables of the dataset include ‘x1’, ‘x2’, ‘x3’ and ‘x4’ which represent open, high, low and close respectively and the target variable is ‘y’. After preprocessing we performed a 90-10 training and testing split of the dataset. We used Grid search algorithm to find the best model algorithm with its best hyperparameters. Later we test the model with its testing dataset. We checked the model performance using roc-auc score, f1 score, recall, precision and confusion matrix.

5. Inference

From the above analysis we arrive at the following conclusions:

i) We trained our model with different classification algorithms like logistic regression, random forest classifier, decision tree classifier and KNN classifier. Accuracy scores from these models showed that logistic regression was more efficient as compared to the rest.

ii) The F1 score for y as 1 is 0.57 and 0 is 0.50. Here we can conclude that our model shows better performance while identifying the true positives as compared to the true negatives.

iii) The accuracy of our model is moderate **(0.54)**. OHLC being a random statistics, it is hard to obtain a higher accuracy score.

iv) From the roc-auc score which is slightly higher than 0.5 indicates that our model trained moderately.

6. References

1. www.sciencedirect.com. (n.d.). Logistic Regression - an overview | ScienceDirect Topics. [online] Available at: https://www.sciencedirect.com/topics/computer-science/logistic-regression#:~:text=Logistic%20regression%20is%20a%20process

2. broadcast, F.B.F.T.R.D. Momentum. [online] Investopedia. Available at: https://www.investopedia.com/terms/m/momentum.asp#:~:text=Momentum%20trading%20is%20a%20strategy.

3. Bruni, R. (2017). Stock Market Index Data and indicators for Day Trading as a Binary Classification problem. Data in Brief, 10, pp.569–575

4. Zhong, X. and Enke, D. (2017). A comprehensive cluster and classification mining procedure for daily stock market return forecasting. Neurocomputing, 267, pp.152–168.

5. Yıldırım, S. (2020). K-Nearest Neighbors (kNN) — Explained. [online] Medium. Available at: https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3.