Study of Machine learning Algorithms for Stock Market Prediction

Ashwini Pathak
MPS in Analytics
Northeastern University
Boston, USA

Sakshi Pathak
B.Tech. Information Technology
SGSITS
Indore 452001, India

***Abstract*: Stock market prediction is a very important aspect in the financial market. It is important to predict the stock market successfully in order to achieve maximum profit. This paper will focus on applying machine learning algorithms like Random Forest, Support Vector Machine, KNN and Logistic Regression on datasets. We evaluate the algorithms by finding performance metrics like accuracy, recall, precision and f-score. Our objective is to identify the best possible algorithm for predicting future stock market performances. The successful prediction of the stock market will have a very positive impact on the stock market institutions and the investors also.**

Keywords: KNN, Logistic Regression, Machine Learning, Random Forest, Stock Market, Support Vector Machine

1. INTRODUCTION

Stock market consists of various buyers and sellers of stock. Stock market prediction means determining the future scope of market. A system is essential to be built which will work with maximum accuracy and it should consider all important factors that could influence the result. Various researches have already been done to predict stock market prices. The research is done over business and computer science domain. Sometime the stock market does well even when the economy is falling because there are various reasons for the profit or loss of a share. Predicting the performance of a stock market is tough as it takes into account various factors. The main aim is to identify the sentiments of investors. It is usually difficult as there must be rigorous analysis of national and international events. It is very important for an investor to know the current price and get a very close estimation of the future price.

There are some mechanisms for stock price prediction that comes under technical analysis[1]:

1. Statistical method
   Statistical methods were widely used before the advent of machine learning. The popular techniques are ARIMA, ESN and Regression. The main features of statistical approach is linearity and stationarity. An analysis of statistical approaches like Linear Discriminant Analysis(LDA), regression algorithms and Quadratic Discriminant Analysis(QDA) is done in [2]. An analysis of widely used technique called ARIMA model is done in [3]. An approach to use time series as input variables is Auto-Regressive Moving Average (ARMA).ARMA model combines Auto Regressive models. ARIMA can reduce non stationary series to a stationary series and is also an extension to ARMA models.
2. Pattern Recognition
   This method focuses on pattern detection. It studies data rigorously and identifies a pattern. Traders find buy and sell signals in Open-High-Low-Close Candlestick charts [4]. A study is done on pattern of stock prices that can help in predicting the future of a stock in [5]. An analysis of pattern is done in [6] by studying charts to develop predictions of stock market. A comparison of market price and its history to chart patterns for predicting future stock prediction is done in [7].
3. Machine learning
   Machine learning is used in many sectors. One of the most popular being stock market prediction itself. Machine learning algorithms are either supervised or unsupervised. In Supervised learning, labelled input data is trained and algorithm is applied. Classification and regression are types of supervised learning. It has a higher controlled environment. Unsupervised learning has unlabelled data but has lower controlled environment. It analyses pattern, correlation or cluster.
4. Sentiment analysis
   Sentiment analysis is an approach that is used in relation to the latest trends [8]. It observes the trends by analysing news and social trends like tweet activity. A study is done on using segment signals from text to improve efficiency of models to analyze trends in stock market in [9].

2. RELATED WORK

There has been several research work on implementing machine learning algorithm for predicting stock market. A study is done by implementing machine learning algorithms on Karachi Stock Exchange (KSE) in [10]. It compared Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP), Radial Basis Function (RBF) and Support Vector Machine (SVM). MLP performs best as compared to others. A comparison of four techniques Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and Naive-Bayes is done in [11]. A study used unsupervised learning as a precursor for supervised tasks [12]. A study compared various machine learning techniques like Random Forest, AdaBoost, Kernel Factory, Neural Networks, Logistic Regression, Support Vector Machine, KNN, on dataset of European companies [13]. An application of various machine learning algorithms (SVM, Naïve Bayes, Random Forest) was done and it was found that random forest gives the highest F-score [14]. A research applied RNN, LSTM, Gated Recurrent Unit (GRU) on google stock dataset and found that LSTM outperforms other algorithms [15]. An application of LSTM to predict Nifty prices is done in [16]. A proposal of an algorithm of multi layer feedforward networks on Chinese Stock dataset is done in [17]. A study applied random forest on Shenzhen Growth Enterprise Market (China) in [18]. It aims to predict stock price and also interval of growth rate. They found that this method is better than some existing methods in terms of accuracy. A proposal of a model that consists of LSTM and GRU is done in [19]. They applied it on S&P dataset. The result was better than some existing neural network approach. A research compared SVM (supervised) and K-means clustering (unsupervised) on S&P 500 dataset [20]. They perform Principal Component Analysis (PCA) for dimensionality reduction. They found that both algorithms give similar performance. The accuracy of SVM is 89.1% and accuracy of K-means is 85.6%. An algorithm is proposed in [21] which combined the Hierarchical Agglomerative Clustering (HAC) and reverse K-means clustering to predict the stock market. It compared HRK model with HAC, K-means, reverse K-means, SVM. The study found that the proposed system is better than SVM in terms of accuracy. An analysis of a model by AprioriALL algorithm (association rule learning) and K-means clustering is done in [22]. It converted data into charts and clustered using K-means to analyze patterns. A paper proposed a clustering method on the Stock Exchange of Thailand (SET) and found that the proposed method is better than other methods of stock market prediction [23].

3. DATASET

The dataset is downloaded from kaggle. The dataset represent data of National Stock Exchange of India for the years 2016 and 2017. The description of dataset is given in Table1.

Table 1. Description of dataset

Feature	Description
Symbol	Symbol of the listed company
Series	Series of the equity(EQ, BE, BL, BT, GC, IL)
Open	Starting price at which a stock is traded in a day
High	Highest price of equity symbol in a day
Low	Lowest price of share in a day
Close	Final price at which a stock is traded in a day
Last	Last traded price of the equity symbol in a day
Prevclose	The previous day closing price of equity symbol in a day
TOTTRDQTY	Total traded quantity of equity symbol on the date
TOTTRDVAL	Total traded volume of equity symbol on the date

4. DATA PRE-PROCESSING

The dataset is in raw format. The dataset needs to be converted into a format that can be analysed. Therefore there are some steps that are performed before building the model:

1. *Handling missing data*
2. One Hot Encoding: It converts categorical data to quantitative variable as any data in the form of string or object does not help in analysing data. First step is to convert the columns to ‘category’ data type. Second step is to apply label encoding in order to convert it into numerical values which will be valuable for analysis. Third step is to convert the column into binary value (either 0 or 1).
3. Data Normalization: It is often possible that if data is not normalized, the column with high values will be given more importance in prediction. In order to tackle that, we scale the data.

5. CLASSIFIERS

Classifiers are given training data, it constructs a model. Then it is supplied testing data and the accuracy of model is calculated. The classifiers used in this paper are :

A. Random Forest Classifier:

It is a supervised algorithm and a type of ensemble learning program. It is a very versatile algorithm capable of performing regression as well as classification. It is built on decision trees. It basically builds multiple decision tree and merges them for producing result. In this algorithm, only a subset of features is taken into consideration. It has same hyperparameters as a decision tree. Advantages of Random Forest are that it works very effectively on large dataset. It can work for both regression and classification problems. It adds more randomness to the model which makes it a better model. The disadvantage of this model is that it makes use of large number of trees that makes it slow.

Algorithm:

1. Randomly select m features.
2. For a node, find the best split.
3. Split the node using best split.
4. Repeat the first 3 steps.
5. Build the forest by repeating these 4 steps.

B. *SVM (Support Vector Machine):*

It is a supervised learning algorithm which classifies cases by a separator. It works by mapping data to a high dimensional feature space and then finds a separator. It finds n-dimensional space that categorizes data points. This algorithm finds the best plane. This plane must have a maximum margin. The boundary that classifies data points is called hyperplanes. The data points are classifies on the basis of position with respect to hyperplanes. Kernel parameter, gamma parameter and regularization parameter are tuning parameters of SVM. Linear kernel predicts new input by dot product between input and support vector. Mapping data to a higher dimensional space is called kernelling. Kernel function can be linear, polynomial, RBF and Sigmoid. Regularization parameter is the C parameter with default value of 10. Less regularization means wrong classification. Small value of gamma means not able to find the region of data. One can improve the model by increasing the importance of classification of each data. Advantages of SVM are that it is a good algorithm for estimation in high dimensional space and it is very memory efficient. Disadvantages of SVM are that it can suffer from over-fitting and that it works very well on small datasets.

C. *KNN (K-nearest neighbour):*

It is an algorithm for classifying similar cases. It produces results only when they are requested. Therefore, it is called lazy learner because there is no learning phase. Advantages of KNN are that it is one of the simplest algorithms as it has to compute the value of k and the Euclidean distance only. It is sometimes faster than other algorithms because of its lazy learning feature. It works well for multiclass problem. Disadvantages of KNN are that the algorithm may not generalize well as it does not go through the learning phase. It is slower for a large dataset as it will have to calculate sort all the distances from the unknown item. Data normalization is necessary for KNN algorithm in order to get best result.

Algorithm for KNN-

1. Choose k.
2. Calculate the Euclidean distance of all cases from unknown case.

   The Euclidean distance (also called the least distance) between sample x and y is :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

   Where,
   x_i is the ith element of the instance x,
   y_i is the ith element of the instance y,
   n is the total number of features in the data set.

3. k number of data points are chosen near unknown data.
4. The unknown data will belong to the majority cases in chosen k neighbours.

D. *Logistic Regression*

This algorithm is used when response is binary (either 1 or 0). It is used for both binary and multiclass classification. Logistic Regression provides most accurate results among all but requires finding the best possible feature to fit. In this model, the relationship between Z and probability of event is given in [24] as,

$$p_i = \frac{e^{zi}}{1+e^{zi}} = \frac{1}{1+e^{-zi}}$$

$$z_i = log(p_i/1 - p_i)$$

Where, p_i is the probability that ith case occurs
Z_i is unobserved continuous variable for ith case
Z value is odd ratio expressed in [24] as:

$$z_i = c + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

x_{ij} is the jth predictor of ith case
B_j is the jth coefficient
P is the number of predictors

6. RESULTS

The dataset consists of features: “Open” that is starting price at which a stock is traded in a day and “Close” is the final price at which a stock is traded in a day. We create a new class label that will have binary values(either 0 or 1). We formulate the idea that if the Open value is less than the close value then we assign it 1 value. If Open value is greater than Close value, we assign it a value of 0.

The data is trained using a model and then the test data is run through the trained model. We obtain a confusion matrix. Confusion matrix represents the values of True positive, false negative, false positive, true positive. True positive is the number of correct prediction that a value belongs to same class. True negative is the number of correct prediction that a value does belong to same class. False positive is the number of incorrect prediction that a value belongs to a class when it belongs to some other class. False negative is the number of incorrect prediction that a value belongs to some other class when it belongs to the same class. Then we calculate performance metrics represented by accuracy, recall, precision and f-score.

Accuracy = (TP+TN) / (TP+TN+FN+FP)

Recall = TP/ (TP+FN)

Precision = TP/ (TP+FP)

F-Score = 2*(precision*recall) / (precision + recall)

Table 2. shows the acquired values of accuracy, recall, precision and f-score when the four algorithms(Random Forest, SVM, KNN, Logistic Regression) are implemented on the dataset. Fig. 1. shows the comparison of accuracy of the four algorithms. Fig. 2. shows the comparison of recall of the four algorithms. Fig. 3. shows the comparison of precision of the four algorithms. Fig. 4. shows the comparison of f-score of the four algorithms.

Table 2. Result of Experiment

Algorithm	Accuracy	Recall	Precision	F-score
Random Forest	80.7	78.3	75.2	76.7
SVM	68.2	65.2	64.7	64.9
KNN	65.2	63.6	64.8	64.1
Logistic Regression	78.6	76.6	77.8	77.1

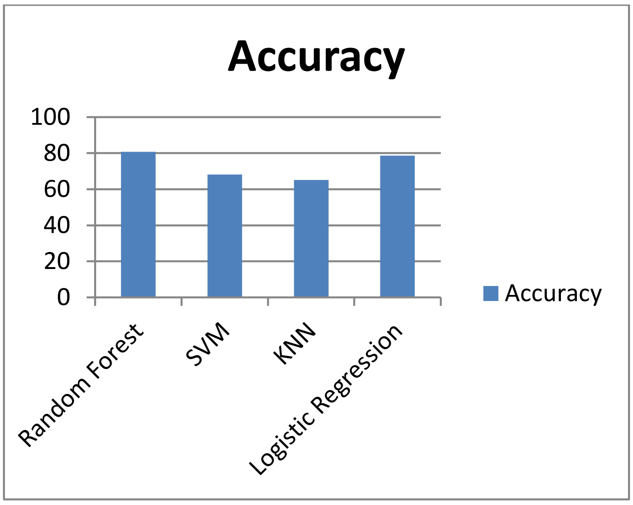

Fig. 1. Accuracy of four algorithms

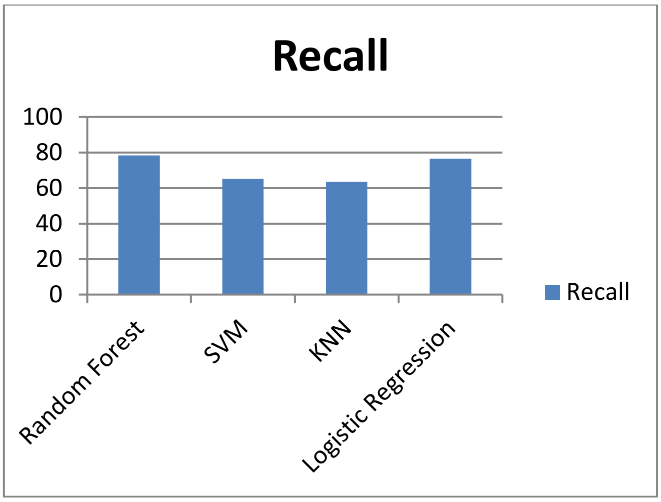

Fig. 2. Recall of four algorithms

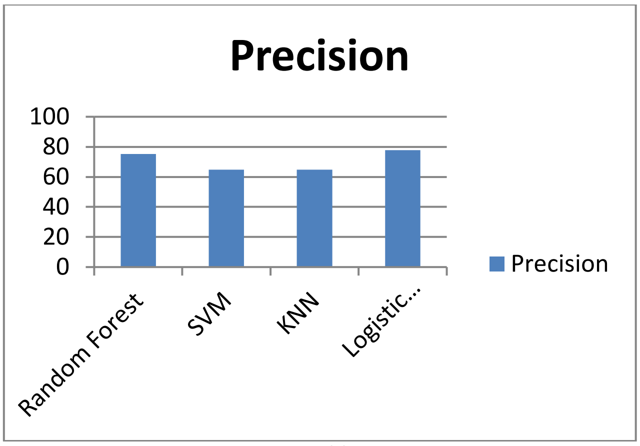

Fig. 3. Precision of four algorithms

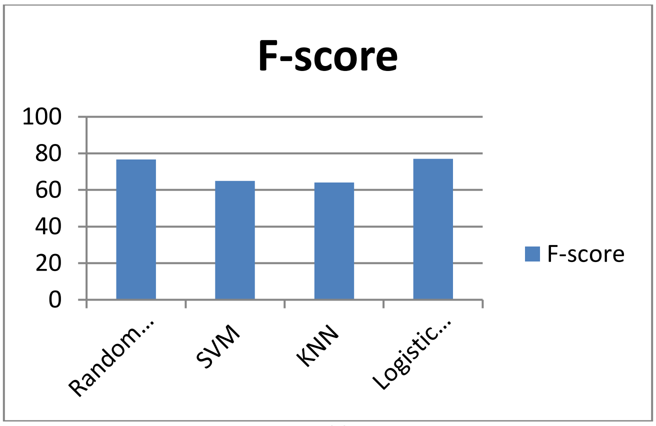

Fig. 4. F-score of four algorithms

The observations made from the performance of the algorithms are:

Random Forest gives the highest accuracy rate for prediction.

1. Random Forest reaches highest recall rate.
2. Logistic Regression reaches highest precision and f-score.
3. KNN is the worst algorithm among the four algorithms for prediction in terms of accuracy.
4. Time taken for building of KNN algorithm is higher than the others.

7. CONCLUSION

We have successfully implemented machine learning algorithms on the dataset for predicting the stock market price. We applied data pre processing and feature selection on the dataset. We applied four algorithms: KNN, SVM, Random Forest, Logistic Regression on the dataset. We analysed the difference of the algorithms by calculating the performance metrics (accuracy, Recall, precision, f-score). We also found the advantages and disadvantages of the algorithms. We conclude that Random Forest is the best algorithm out of the four with an accuracy rate of 80.7%. Future scope of this paper would be adding more parameters that effect the stock market prediction. Adding more number of parameters will ensure better estimation. The new work can also take in account the concept of sentiment analysis where we will consider the public comments, news and social influence. This will increase the understanding of investors and give better prediction.

8. REFERENCES

[1] Shah D, Isah H, Zulkernine F. Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. Int. J. Financial Stud., 2019, 7(2), pp. 1-22.

[2] Zhong X, Enke D. Forecasting daily stock market return using dimensionality reduction. Expert Systems with Applications, 2017, vol. 67, pp. 126–139.

[3] Hiransha M, Gopalakrishnan E A, Menon V K, Soman K P. NSE stock market prediction using deep-learning models. Procedia Computer Science, 2018, vol. 132, pp. 1351–1362.

[4] Velay M, Fabrice D. Stock Chart Pattern recognition with Deep Learning. arXiv, 2018.

[5] Parracho P, Neves R, Horta N. Trading in Financial Markets Using Pattern Recognition Optimized by Genetic Algorithms. 12th Annual Conference Companion on Genetic and Evolutionary Computation, 2010, pp. 2105-2106.

[6] Nesbitt K V, Barrass S. Finding trading patterns in stock market data. IEEE Computer Graphics and Applications, 2004, 24(5), pp. 45–55.

[7] Leigh W, Modani N, Purvis R, Roberts T. Stock market trading rule discovery using technical charting heuristics. Expert Systems with Applications, 2002, 23(2), pp. 155–159.

[8] Bollen J, Mao H, Zeng X. Twitter Mood Predicts the Stock Market. Journal of Computational Science, 2011, 2(1), pp. 1–8.

[9] Seng J L, Yang H F. The association between stock price volatility and financial news—A sentiment analysis approach. Kybernetes, 2017, 46(8), pp. 1341–1365.

[10] Usmani M, Adil S H, Raza K, Ali S S A. Stock market prediction using machine learning techniques. 3rd International

Conference on Computer and Information Sciences (ICCOINS), 2016, pp. 322-327.

[11] Patel J, Shah S, Thakkar P, Kotecha K. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques. Expert Systems with Applications, 2015, 42(1), pp. 259-268.

[12] Bhardwaj A, Narayan Y, Vanraj, Pawan, Maitreyee D. Sentiment analysis for Indian stock market prediction using Sensex and nifty. Procedia Computer Science, 2015, 70, pp. 85–91.

[13] Ballings M, Poel D V D, Hespeels N, Gryp R. Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 2015, 42(20), pp. 7046–56.

[14] Milosevic N. Equity Forecast: Predicting Long Term Stock Price Movement Using Machine Learning. arXiv, 2016.

[15] Luca D P, Honchar O. Recurrent Neural Networks Approach to the Financial Forecast of Google Assets. International Journal of Mathematics and Computers in simulation, 2017, vol. 11, pp. 7–13.

[16] Roondiwala M, Patel H, Varma S. Predicting Stock Prices Using Lstm. International Journal of Science and Research (IJSR), 2017, vol. 6, pp. 1754–1756.

[17] Yang B, Gong Z J, Yang W. Stock Market Index Prediction Using Deep Neural Network Ensemble. 36th Chinese Control Conference (CCC), 2017, pp. 26–28.

[18] Zhang J, Cui S, Xu Y, Li Q, Li T. A novel data-driven stock price trend prediction system. Expert Systems with Applications, 2018, 97(1), pp. 60–69.

[19] Hossain M A, Karim R, Thulasiram R K, Bruce N D B, Wang Y. Hybrid Deep Learning Model for Stock Price Prediction. IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 18–21.

[20] Powell N, Foo S Y, Weatherspoon M. Supervised and Unsupervised Methods for Stock Trend Forecasting. Paper presented at the 40th Southeastern Symposium on System Theory (SSST), 2008, pp. 203-205.

[21] Babu M S, Geethanjali N, Satyanarayana B. Clustering Approach to Stock Market Prediction. International Journal of Advanced Networking and Applications, 2012, vol. 3, pp. 1281-1291.

[22] Wu K P, Wu Y P, Lee H M. Stock Trend Prediction by Using K-Means and Aprioriall Algorithm for Sequential Chart Pattern Mining. Journal of Information Science and Engineering, 2014, vol. 30, pp. 669–686.

[23] Peachavanish R. Stock selection and trading based on cluster analysis of trend and momentum indicators. International MultiConference of Engineers and Computer Scientists, 2016, vol. 1, pp. 16–18.

[24] Zaidi M, Amirat A. Forecasting Stock Market Trends by Logistic Regression and Neural Networks Evidence from KSA Stock Market. Euro-Asian Journal of Economics and Finance, 2016, 4(2), pp. 50-58.