(IJIEST) 2019, Vol. No. 5, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

EMPLOYABILITY OF A SMART MODEL BASED ON MACHINE LEARNING TOOLS AND TECHNIQUES IN THE ENHANCED FORECASTING OF STOCK PRICES

Suhasini Singh
Christ (deemed to be) University

ABSTRACT

Stock price prediction is a vital part of the monetary market. Forecasting the securities exchange effectively is essential to accomplish the greatest benefit. This paper focuses on applying AI algorithms like Random Forest, SVM, KNN and linear regression on datasets. We assess the algorithm by finding execution measurements like precision, Review, accuracy and f-score. We plan to distinguish the ideal calculation for predicting future financial exchange exhibitions. The fruitful forecast of the financial exchange will certainly affect the securities exchange structures and financial partners.

INTRODUCTION

The securities exchange comprises different buyers and sellers of stock. Predicting the stock market's future scope is called "prediction." A framework is crucial for being fabricated that will work with the most stunning precision, and it ought to consider exceptionally significant elements that could impact the outcome. Different investigations have proactively been finished to anticipate financial exchange costs. The analysis is done in the business and software engineering space. Sometimes, the securities exchange does well when the economy is falling because there are different purposes behind the benefit or loss of an offer. A stock market's performance can only be predicted by taking into account a variety of factors. The primary objective is to ascertain investor sentiments. Due to the need to conduct a thorough analysis of both national and international events, it is typically challenging. A financial supporter must know the ongoing cost and get an extremely close estimation of things.

There are a few mechanisms for stock cost expectation that go under specialized analysis [1]:

1. Measurable technique

Measurable techniques were generally utilized before the appearance of AI. The well-known procedures are ARIMA, ESN and Relapse. The primary elements of the factual methodology are linearity and stationarity. An analysis of factual methodologies like LDA, linear regression and Quadratic Discriminant Analysis (QDA) is done [2]. An examination of a generally utilized procedure called the ARIMA model is done [3]. A way to deal with involving time series as information factors is Auto-Backward Moving Normal (ARMA). ARMA model joins Auto-Backward models. ARIMA can decrease non-fixed series to fixed series and is likewise an augmentation to ARMA models.

2. Design Declaration

This strategy focuses on design discovery. It diligently examines the data and discovers a pattern. Merchants find trade signals in Open-High-Low-Close bar graphs [4]. A review is done on stock cost designs that can assist with foreseeing a stock's future [5]. The example is examined in [6] by concentrating on diagrams to facilitate securities exchange forecasts. Examining the market value and its set of experiences to outline designs for foreseeing future stock prediction is made in [7].

3. Machine learning

There are various applications for machine learning. Prediction of the stock market is one of the most liked. AI algorithms are either issued or solo. In machine learning, marked input data is trained, and the calculation is applied. Arrangement and Relapse are directed learning. It has a deep, controlled environment. Unsupervised learning uses data that has yet to be labelled, but the environment is less controlled. It examines examples, associations or groups.

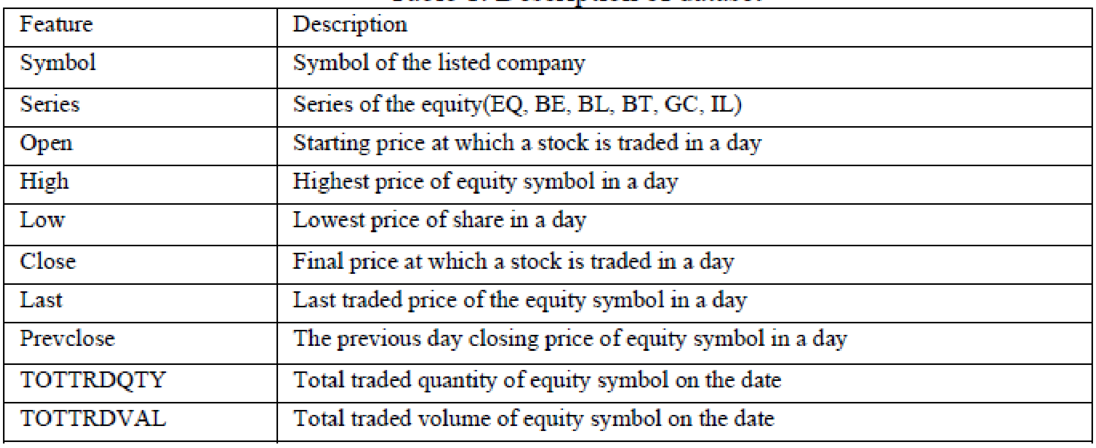
4. Sentiment analysis

It is a methodology for the most recent patterns [8]. It notices the patterns by investigating news and social patterns like tweet action. Segment signals from text have been used in a study to make it easier for models to analyse stock market trends [9].

DATASET

Kaggle is used to download the dataset. The dataset addresses information from the Public Stock Trade of India for 2016 and 2017. Table 1 describes the dataset.

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

DATA PRE-PROCESSING

The dataset is in raw form. The dataset should be changed over into an organization that can be examined. Thus, a few stages are performed before building the model:

1. Taking care of missing information

2. One Hot Encoding: It switches clear-cut information over completely to quantitative factors, as any information as a string or item doesn't assist with examining information. The initial step is to switch the sections over completely to a 'classification' information type. The subsequent step is to apply mark encoding to change it into mathematical qualities, which will be significant for examination. The column must be converted into a binary value (0 or 1) in the third step.

3. Normalization of Data: It is often conceivable that if the information isn't standardized, the section with high qualities will be more significant in the forecast. That's what to handle; we scale the information.

CLASSIFIERS

Classifiers are given preparation information and develop a model. Then, testing information is provided, and the model's exactness is determined. The classifiers utilized in this paper are :

A. Classifier Based on Random Forests:

It is a managed calculation and a gathering learning program. It is an exceptionally flexible calculation fit for performing Relapse and characterization. It is based on choice trees. It constructs numerous choice trees and consolidates them to deliver results. In this calculation, just a subset of elements is thought about. It has a similar hyperparameter as a choice tree. Random Forest's advantage is that it works well with a large dataset. It can work for both relapse and arrangement issues. It adds more irregularity to the model, improving it. The hindrance of this model is that it utilizes many trees, making it slow.

B. Support Vector Machine

It is a managed learning calculation that orders cases by a separator. It works by planning information to a high-layered, including space and tracking down a separator. It finds n-layered space that classifies pieces of information. This calculation is viewed as the best plane. There must be a maximum margin on this plane. Hyperplanes are the boundaries used to classify data points. The information focuses are ordered in light of the position concerning hyperplanes. SVM's tuning parameters are the kernel, gamma, and regularization parameters. The straight portion predicts new contribution by the spot item between the info and backing vector. Planning information to a higher layered space is called kernelling. Portion capability can be straight, polynomial, RBF and Sigmoid. The regularization boundary is the C boundary with a default worth of 10. Less regularization implies wrong grouping. A small gamma value indicates the inability to locate the data region. The model can be improved by making each data classification more important. SVM's advantages include its ability to estimate in high-dimensional space and its high memory efficiency. The burdens of SVM are that it can experience the ill effects of over-fitting and functions admirably on little datasets.

C. KNN

It is a calculation for grouping comparable cases. It produces results just when they are mentioned. In this manner, it is known as a sedentary student since there is no learning stage. The benefit of KNN is that it is the least difficult calculation, as it needs to figure out the worth of k and the Euclidean distance. It is, at times, quicker than different calculations as a result of its sluggish learning highlight. It functions admirably for multiclass issues. Because it does not go through the learning phase, the KNN algorithm may not be able to be applied to a wide range of problems. It is slower for an enormous dataset as it should work out to sort every one of the good ways from the obscure thing. Information standardization is important for the KNN calculation to obtain the best outcome.

D. Linear regression

This calculation is utilized when the reaction is paired (either 1 or 0). It is utilized for both parallel and multiclass characterization. Linear regression gives the most reliable outcomes except for tracking the ideal fit component. In this model, the connection between Z and the likelihood of an events.

RESULTS

The dataset comprises elements: " Open" is the beginning cost at which a stock is exchanged a day, and "Close" is the last cost at which a stock is exchanged a day. A brand-new class label with binary values (either 0 or 1) is created. We form the possibility that if the Open worth is not exactly the nearby worth, we dole out it 1 worth. If the Open worth is more prominent than the Nearby, we relegate it to 0.

The information is prepared to utilize a model, and afterwards, the test information is gone through the prepared model. A confusion matrix is what we get. The disarray grid addresses the upsides of Genuine positive, misleading negative, bogus positive, and genuine positive. Genuine positive is the quantity of right forecasts that a worth has a place with a similar class. The genuine negative is the number of right forecasts that have a place with a similar class. Bogus positive is the number of erroneous expectations that worth has a place with a class when it has a place with another class. A misleading negative is the number of inaccurate forecasts that worth has a place with some other class when it has a place with The perceptions produced using the presentation of the calculations are:

Random forest gives the most elevated precision rate for expectation.

1. Irregular Backwoods arrive at the most elevated review rate.

2. Calculated Relapse arrives at the most elevated accuracy and f-score.

3. Regarding accuracy, KNN is the worst of the four prediction algorithms.

4. The time to fabricate the KNN calculation is higher than the others.

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

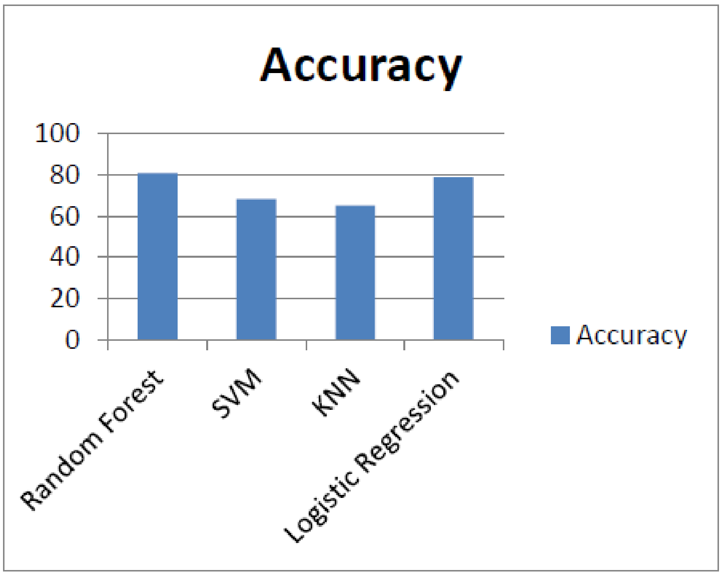


Fig. 1. Accuracy of four algorithms

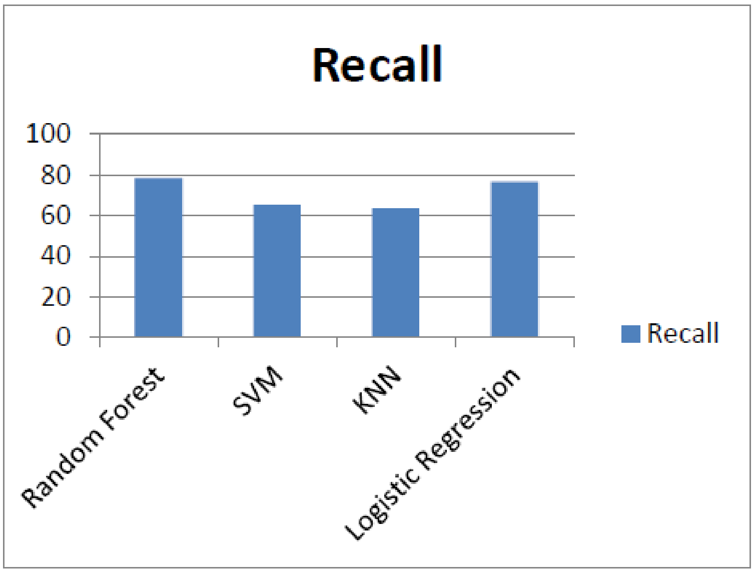


Fig. 2. Recall of four algorithms

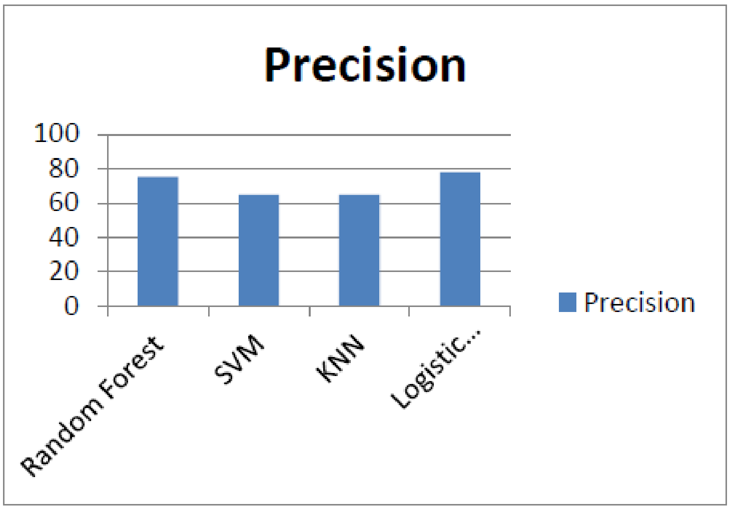


Fig. 3. Precision of four algorithms

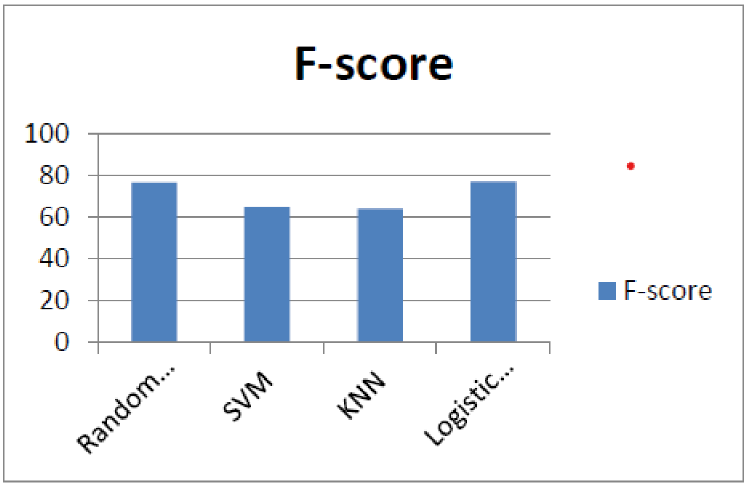


Fig. 4. F-score of four algorithms

CONCLUSION

We have effectively conducted AI calculations on the dataset to anticipate the securities exchange cost. We applied information pre-handling and element choice on the dataset. We applied four calculations: KNN, SVM, Arbitrary Woodland, and Calculated Relapse on the dataset. We dissected the distinction of the calculations by ascertaining the presentation measurements (exactness, Review, accuracy, f-score). We likewise tracked down the benefits and drawbacks of the calculations. Irregular Woods is the best calculation, with a precision pace of 80.7%. Adding additional parameters that influence stock market prediction would expand the scope of this paper in the future. Adding more number of boundaries will guarantee better assessment. The new work can also incorporate sentiment analysis, in which public comments, news, and social influence are considered. Investors will gain a deeper understanding and better predictions due to this.

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.