ISSN No:-2456-2165

Machine Learning Algorithms for Stock Market Prediction

Ketan Juare
school of engineering Ajeenkya
DY Patil University,Pune, India

Akanksha Kulkarni
School of Engineering Ajeenkya
DY Patil University,Pune, India – 412105

Abstract:- Stock market forecasts are a very important aspect of the financial market. It is important to successfully predict the stock market in order to get maximum profit. This document focuses on applying machine learning algorithms such as Random Forest, Support Vector Machine, KNN, and Logistic Regression to datasets. We evaluate the algorithms by determining performance metrics such as accuracy, recall, precision, and F-Score. Our goal is to identify the best possible algorithm to predict the future development of the stock market. Successful stock market forecasts will have a very positive effect on stock exchange institutes and investors.

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

I. INTRODUCTION

Stock market consists of assorted patrons and sellers of stock. exchange prediction suggests that crucial the longer term scope of market. A system is crucial to be engineered which is able to work with most accuracy and it ought to think about all important factors that might influence the result. numerous analysis's have already been done to predict stock market prices. The research is completed over business and computing domain. someday the stock market will well even once the economy is falling as a result of there are various reasons for the profit or loss of a share. Predicting the performance of a stock market is hard as it takes under consideration numerous factors. the most aim is to spot the feelings of capitalists. it's typically tough as there should be rigorous Analysis of national and international events. it's vital for an investor beneath stand to grasp the present value and obtain a awfully shut estimation of the longer term price. There are some mechanisms for stock price prediction that comes under technical analysis:

- **Statistical method Statistical** ways were wide used before the appearance of machine learning. the favored techniques are ARIMA, ESN and Regression. the most options of statistical approach is one-dimensionality and stationarity. AN analysis of applied mathematics approaches like Linear Discriminant Analysis(LDA), regression algorithms and Quadratic Discriminant Analysis(QDA) is completed in. An analysis of wide used technique referred to as ARIMA model is done in [3]. An approach to use statistic 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.
- **Pattern Recognition** This method focuses on pattern recognition. Study the data carefully and identify apattern. Traders can find buy and sell signals on the Open, High, Low, and Close Candlestick charts. A study is conducted on the pattern of stock prices that can help predict the future of a stock in .In, pattern analysis is performed by studying charts to develop stock market predictions. In 3, a comparison of the market price and its history is performed to draw patterns to predict future actions.
- **Machine learning Machine** Learning Machine learning is used in many industries. One of the most popular is the stock market forecast itself. Machine learning algorithms are either monitored or unmonitored. Supervised learning trains the labeled input data and applies the algorithm. Classification and regression are types of supervised learning. It has a more controlled environment. Unsupervised learning has unlabeled data but a less controlled environment. Analyze patterns, correlation or clusters.
- **Sentiment analysis** Sentiment analysis is an approach used in relation to the latest trends [8]. See trends by analyzing news and social trends like tweet activity. A study is conducted on the use of text segment signals to improve the efficiency of models for analyzing trends in the stock market in.

II. UNDERSTANDING THE STOCK MARKET

In a nutshell, stock markets offer a secure and controlled setting wherever market participants will interact in shares and different eligible monetary instruments confidently with zero- to low operational seven risk. in operation underneath the outlined rules as expressed by the regulator, the stock markets act as primary markets and as secondary markets. As a primary market, the stock market permits firms to issue and sell their shares to the common public for the primary time through the method of initial public offerings (IPO). This activity helps companies raise necessary capital from investors. It basically implies that an organization divides itself into variety of shares (say, twenty million shares) and sells a vicinity of these shares (say, five million shares) to the common public at a value (say, $10 per share). To facilitate this process, an organization desires a marketplace wherever these shares will be sold. This marketplace is provided by the stock market. If everything goes as per the plans, the corporate will with success sell the 5 million shares at a price of $10 per share and collect $50 million worth of funds. Investors will get the company shares that they will expect to carry for his or her most popular duration, in anticipation of rising share value and any potential financial gain within the kind of dividend payments. The exchange acts as a supporter for this capital

ISSN No:-2456-2165

raising method and receives a fee for its services from the corporate and its monetary partners. Following the first-time share provision mercantilism exercise referred to as the listing process, the stock exchange conjointly is the commercialism platform that facilitates regular shopping for and commerce of the listed shares. This constitutes the secondary market. The stock exchange earns a fee for each trade that happens on its platform throughout the secondary market activity. The stock exchange shoulders the responsibility of making certain price transparency, liquidity, price discovery and truthful dealings in such commercialism activities. As most major stock markets across the world currently operate electronically, the exchange maintains trading systems that with efficiency manage the buy and sell orders from varied market participants. They perform the worth matching operate to facilitate trade execution at a price fair to each consumers and sellers. A listed company furtherly |might also|may additionally} offer new, additional shares through different offerings at a later stage, like through rights offering or through innings offers. they will even buy or delist their shares. The exchange facilitates such transactions. The stock exchange typically creates and maintains varied market-level and sector-specific indicators, just like the S&P five hundred index or information system one hundred index, which give a live to trace the movement of the market. different ways embody the random generator and random Momentum Index. The stock exchanges conjointly maintain all company news, announcements, and monetary reporting, which might be typically accessed on their official websites. A stock exchange also supports various other corporate- level, transaction-related activities. For instance, profitable firms may reward investors by paying dividends that usually comes from a vicinity of the company's earnings. The exchange maintains all such info and will support its process to a definite extent

Keywords: *Stock Market, Analysis, Data Visualization, Python, Machine Learning, Trends, Trend Prediction, Investment*

III. TECHNICAL INDICATORS

Technical Indicators are the properties associated with any stock based onits tick prices.

A. Simple Moving Average (SMA)

Simple Moving Average is calculated exclusively based on close price of the stock which we are trying to analyze. For instance, if we need to calculate SMA of 'x' intervals we need to get close prices of previous 20 intervals and divide it by 'x'. Hence the first available SMA value will correspond to xth interval. To calculate I add all close prices starting from the current interval looking back for n number of intervals. In this case n stands for number of intervals for which we need to calculate SMA.

B. Exponential Moving Average (EMA)

Exponential Moving Average is also calculated based on close price of the stock which we are analyzing. If we need to calculate EMA of 'x' intervals, we first need to calculate SMA (2.1.2.1) till xth interval and for every subsequent interval we calculate EMA based on the following formula.

$$EMA_{Today} = (Value_{Today} \times (\frac{Smoothing}{1 + Days})) + EMA_{Yesterday} \times (1 - (\frac{Smoothing}{1 + Days}))$$

Hence, the first available EMA value will be corresponding to xth interval. EMA moves hand in hand with price of stock. Higher the number of intervals we choose to calculate EMA, higher the stability of EMA. Hence, we chose 2, 5, 10, 20, 50 and 100 intervals for calculating EMA.

C. Relative Strength Index (RSI)

Relative Strength Index is calculated based on SMA and close price of the stock for the given interval. We must get familiar with the following terms to better understand the calculation of RSI: Gain, Loss. If the close price of the stock at a given interval is greater than it's open price then the stock resulted in Gain, vice versa it resulted in the Loss. Here are the formulae to calculate the RS and RSI. RS = Average Gain/ Average Loss RSI indicates the strength of the current trend. If Higher value of interval is chosen , then we get stable RSI values. We need to find out the threshold value. If RSI falls below its threshold, it is an indication of sellers taking over buyers. If RSI value rises over its threshold, it indicates that buyers are taking over sellers and stock prices will go high.

IV. RELATED WORK

There has been many analysis work on implementing machine learning algorithmic program for predicting stock market. A study is finished by implementing machine learning algorithms on metropolis exchange (KSE) in [10]. It compared Single Layer Perceptron (SLP), Multi-Layer Perceptron (MLP), Radial Basis perform (RBF) and Support Vector Machine (SVM). MLP performs best as compared to others. A comparison of 4 techniques Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and Naive-Bayes is done in . A study used unsupervised learning as a precursor for supervised tasks [12]. A study compared varied machine learning techniques like Random Forest, AdaBoost, Kernel Factory, Neural Networks, provision Regression, Support Vector Machine, KNN, on dataset of European firms An application of varied machine learning rules (SVM, Naïve Bayes, Random Forest) was done and it absolutely was found that random forest provides the very best Fscore. a research applied RNN, LSTM, Gated continual Unit (GRU) on google stock dataset and located that LSTM outperforms alternative algorithms. AN application of LSTM to predict cracking costs is completed in. A proposal of an algorithm of multi

ISSN No:-2456-2165

layer feedforward networks on Chinese Stock dataset is done in. A study applied random forest on Shenzhen Growth Enterprise Market (China) in. It aims to predict stock worth and additionally interval of growth rate. They found that this methodology is best than some existing ways in terms of accuracy. A proposal of a model that consists of LSTM and GRU is completed in. 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 clump (unsupervised) on S&P five hundred dataset. They perform Principal element ANalysis (PCA) for spatiality reduction. They found that each rules offer similar performance. The accuracy of SVM is89.1% and accuracy of K-means is 85.6%. An algorithm is projected within which combined the ranked agglomerated clump (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 projected system is best than SVM in terms of accuracy. AN analysis of a model by AprioriALL rule (association rule learning) and K-means clump is completed in . It born-again information into charts and Clustered victimisation K-means to research patterns. A paper proposed a clustering methodology on the stock market of Asian country (SET) and located that the proposed method is better than alternative ways of stock market prediction.

A. Dataset

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

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

Table 1: Description of dataset

B. Data Pre-Processing

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

- **Handling missing data:** The dataset is in raw format. The dataset has to be regenerate into a format that may be analyses. so thereare some steps that are performed before building the model: 1. Handling missing data
- **One Hot Encoding:** It converts categorical data to quantitative variable as any data within the style of string or object doesn't facilitate in analysing data. opening move is to convert the columns to ‘category’ data type. Second step is to use label coding so as to convert it into numerical values which will be valuable for analysis. Third step is to convert the column into binary worth (either zero or 1).
- **Data Normalization:** it's usually potential that if data isn't normalized, the column with high values are given additional importance in prediction. so as to tackle that, we have a tendency to scale the data.

V. CLASSIFIERS

Classifiers are given coaching information, it constructs a model. Then it's provided testing data and also the accuracy of model is calculated. The classifiers employed in this paper are : A. Random Forest Classifier: it's a supervised algorithmic program and a kind of ensemble learning program. it's a really versatile algorithm capable of playacting regression furthermore as classification. it's engineered on decision trees. It primarily builds multiple decision tree and merges them for manufacturing result. during this algorithm, solely a set of options is taken into consideration. it's same hyper parameters as a choice tree. blessings of Random Forest are that it works terribly effectively on large dataset. It will work for each regression and classification problems. It adds additional randomness to the model that makes it a far better model. The disadvantage of this model is that it makes use of huge range of trees that creates it slow. Algorithm:

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

ISSN No:-2456-2165

- **SVM (Support Vector Machine):** It is a supervised learning formula that classifies cases by a separator. It works by mapping data to a high dimensional feature house so finds a separator. It finds n-dimensional space that categorizes data points. This algorithm finds the simplest plane. This plane should have a most margin.

- The boundary that classifies data points is named hyperplanes. the info points are classifies on the premise of position with relation to hyperplanes. Kernel parameter, gamma parameter and regularization parameter are standardization parameters of SVM. Linear kernel predicts new input by real number between input and support vector. Mapping data to a higher dimensional house is named kernelling. Kernel perform will be linear, polynomial, RBF and Sigmoid. Regularization parameter is that the C parameter with default price of 10. Less regularization suggests that wrong classification. tiny value of gamma means ineffectual to search out the region of data. One will improve the model by increasing the importance of classification of every data. blessings of SVM are that it's an honest formula for estimation in high dimensional space and it is terribly memory efficient. Disadvantages of SVM are that it can suffer from over-fitting which it works fine on small datasets.

- **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-**

✓ Choose k.
✓ 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, xi is the ith element of the instance x, yi is the ith element of the instance y, n is the total number of features in the data set.
✓ k number of data points are chosen near unknown data.
✓ The unknown data will belong to the majority cases in chosen k neighbours.

VI. PROPOSED METHOD

The model predicts the price movement in tn considering all available historical data, i.e. H. from tn−1,tn−2,...t1, where tn represents prediction transaction data. All available data is trained using a supervised machine learning algorithm. Sentiment is extracted from social media and news data. The sentiments extracted later are integrated with the historical price to create the predictive model. Researchers reported conflicting opinions about the impact of sentiment on the stock market. Few studies14 reported that sentiment extracted from social media has no effect on stock price movement, while reporting that sentiment has a strong or weak effect on stock price movement. Two different models have been developed to predict the development of the stock market. The first model forecasts the development of the stock market for the next day (Daily Prediction Model) taking into account all daily available data as input. The second model forecasts the development of the stock market for the next month (monthly forecast model), taking into account the monthly available data. The first contribution of the proposed work is that few features have been derived from the available historical data through the use of statistics. One of the statistical parameters considered is the ratio between the trend of a day and the volume of shares traded on the same day15. The traded volume function in the historical data reflects the shares bought and sold on a daily basis.

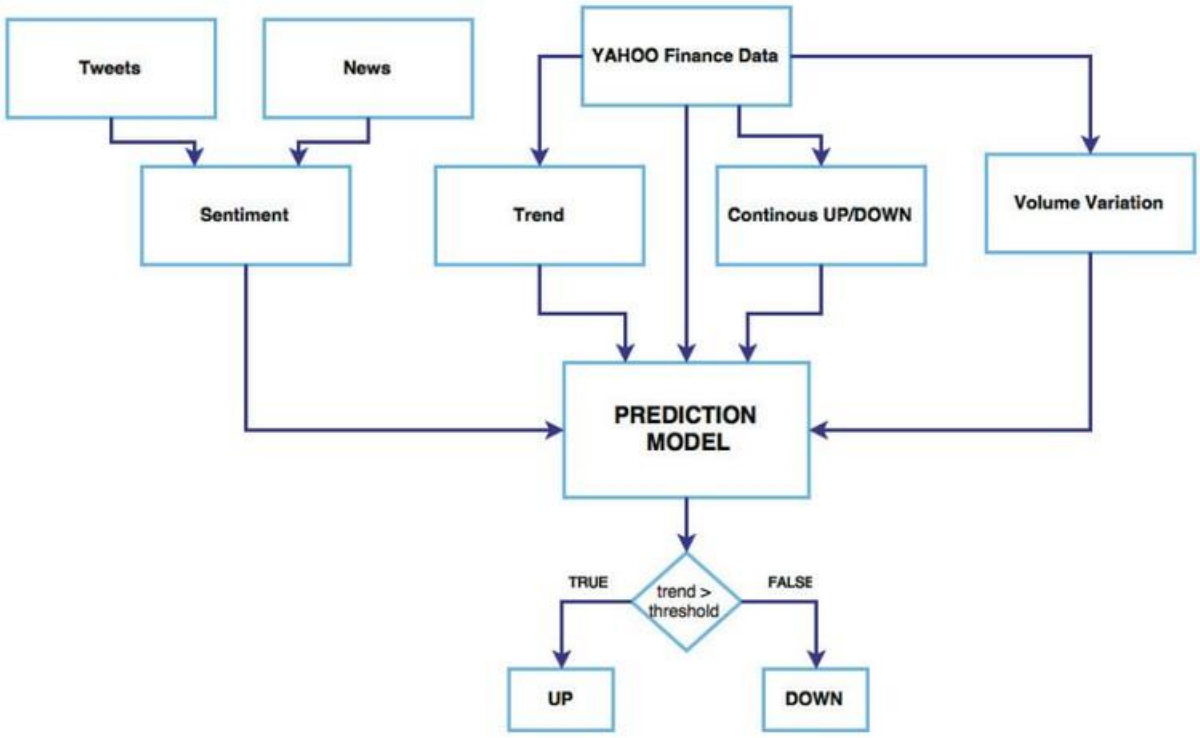

Fig. 1: Prediction Model for Daily Prediction Model.

VII. RESULTS

The data set consists of characteristics: "Open" is the starting price at which a stock trades in a day, and "Close" is the ending price at which a stock trades in a day. We create a new class tag that will have binary values (0 or 1). We formulate the idea that if the opening value is less than the closing value, we assign it 1 value. If the opening value is greater than the closing value, we assign it a value of 0. The data is trained using a model, and then the test data is passed through the trained model. We get a confusion matrix. The confusion matrix represents the values True Positive, False Negative, False Positive, True Positive. True positive is the number of correct predictions that a value belongs to the same class. True negative is the number of correct predictions that a value belongs to the same class. False positives are the number of incorrect predictions that a value belongs to one class when it belongs to another class.

False negative is the number of incorrect predictions that a value belongs to a different class even though it belongs to the same class. We then calculated performance metrics represented by accuracy, recall, precision and F-score. Precision = (TP+TN) / (TP+TN+FN+FP) Recovery = TP/ (TP+FN) Precision = TP/ (TP+FP) F-Score = 2*(Precision*Recall) / (Precision + recovery) Table 2 shows the collected values for accuracy, recall, precision and f-score when the four algorithms (Random Forest, SVM, KNN, Logistic Regression) are applied to the data set. Figure 1. shows the precision comparison of the four algorithms.

Figure 2 shows the recovery comparison of the four algorithms. Fig. 3. shows the precision comparison of the four algorithms. Figure 4. shows the comparison of the f-scores of the four algorithms

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

Table 2: Result of Experiment

ISSN No:-2456-2165

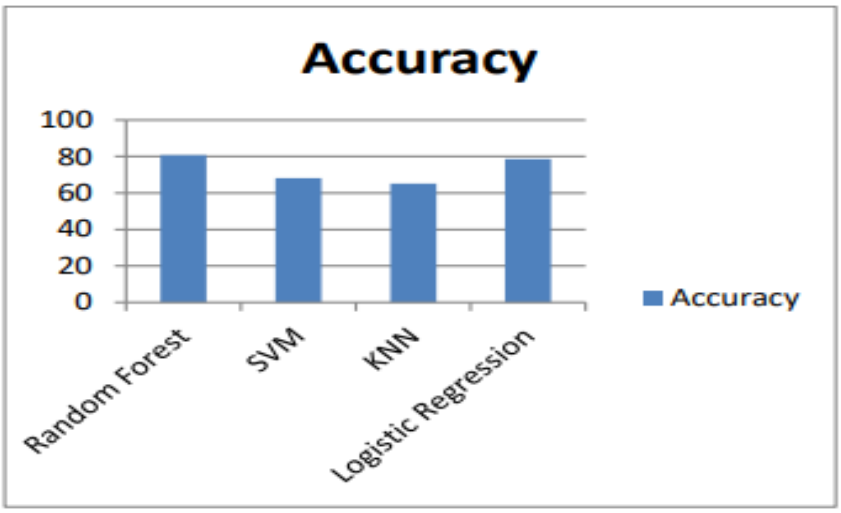

Fig. 1: Accuracy of four algorithms

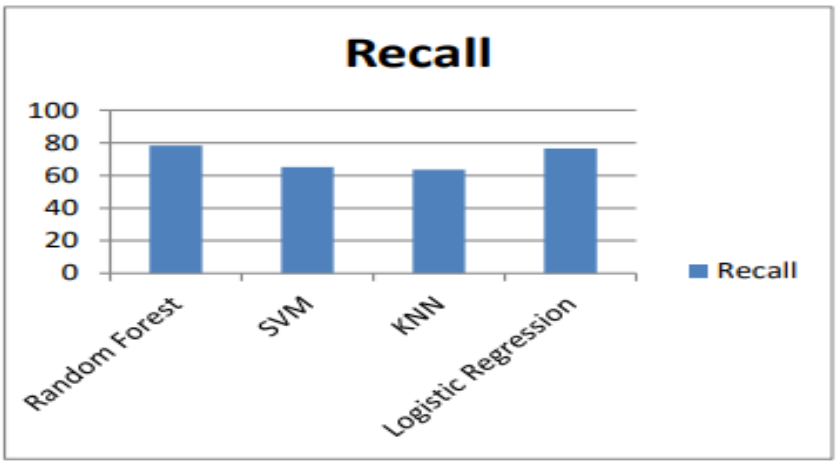

Fig. 2: Recall of four algorithms

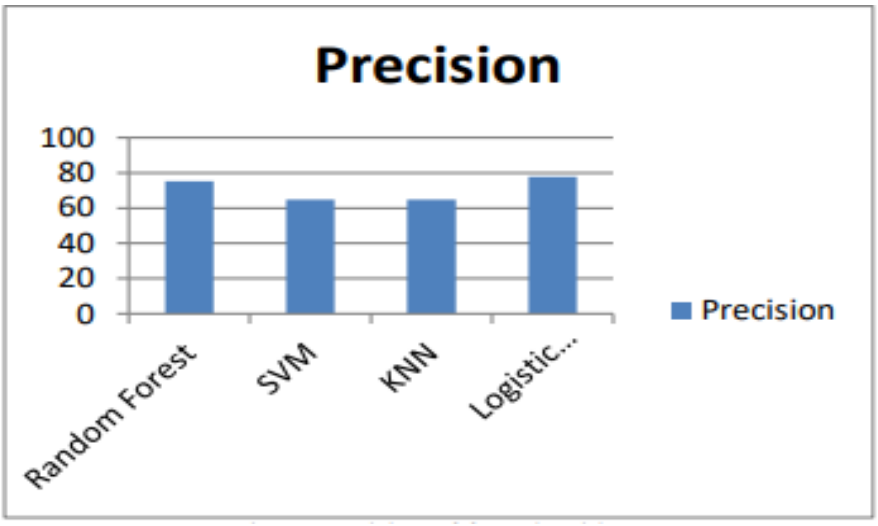

Fig. 3: Precision of four algorithms

ISSN No:-2456-2165

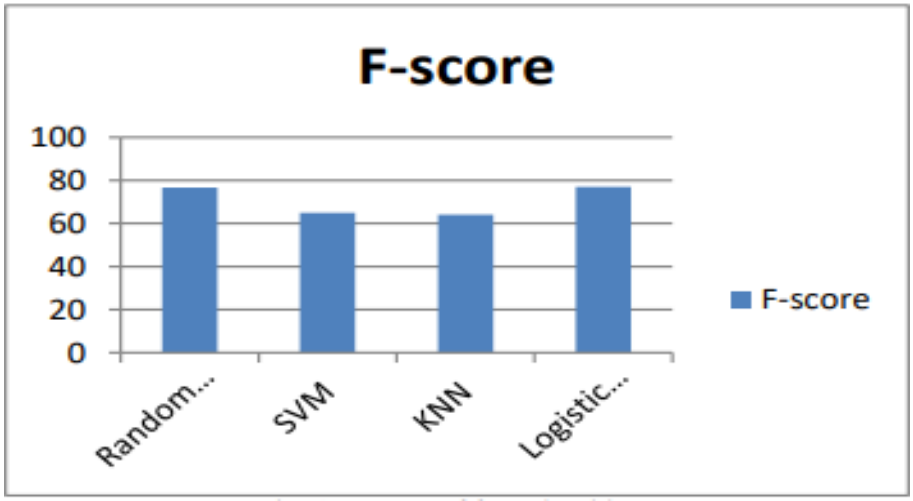

Fig. 4: F-Score of our algorithms

The observations made from the performance of the algorithms are: Random Forest gives the highest accuracy rate for predicti on.

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

VIII. CONCLUSION

We successfully implemented machine learning algorithms on the dataset to predict the stock price. We apply data preprocessing and feature selection to the data set. We apply four algorithms: ANN, SVM, Random Forest, Logistic Regression on the data set. We analyzed the difference in algorithms by calculating the performance metrics (Accuracy, Retrieval, Precision, F-Score). We also find the advantages and disadvantages of algorithms. We conclude that Random Forest is the best of the four algorithms with an accuracy rate of 80.7%. The future scope of this document would be toadd more parameters affecting the stock market forecast. Adding a larger number of parameters ensures a better estimate. The new work may also incorporate the concept of sentiment analysis, where we will consider public comment, news and social influence. This will improve investors' understanding and give a better prediction.

REFERENCES

[1.] Yahoo finance, https://finance.yahoo.com/ for the history of stocks
[2.] 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
[3.] Velay M, Fabrice D. Stock Chart Pattern recognition with Deep Learning. arXiv, 2018
[4.] 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
[5.] 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
[6.] 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.
[7.] Milosevic N. Equity Forecast: Predicting Long Term Stock Price Movement Using Machine Learning. arXiv, 2016.
[8.] 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.
[9.] Roondiwala M, Patel H, Varma S. Predicting Stock Prices Using Lstm. International Journal of Science and Research (IJSR), 2017, vol. 6, pp. 1754–1756.
[10.] 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.
[11.] Treading view, https://in.tradingview.com/
[12.] Dhan, https://dhan.co/#.