Measuring Accuracy of Stock Price Prediction Using Machine Learning Based Classifiers

Ranjeet Kaur¹, Dr. Yogesh Kumar Sharma², Devershi Pallavi Bhatt³

¹Research Scholar, Department of Computer Science & Engineering, Sri Jagdish Prasad Jhabarmal Tibrewala University, Jhunjhunu, India
²Associate Professor & Head, Department of Computer Science & Engineering, Shri Jagdish Prasad Jhabarmal Tibrewala University, Jhunjhunu, India
³Associate Professor & Head, Department of Computer Applications, Manipal University Jaipur, India

E-mail: er.ranjeetsandhu@gmail.com

Abstract. Stock market prediction means to decide the future development of the stock estimation of a budgetary trade. The exact forecast of offer value development will prompt more benefit speculators can make. The investigation dependent on the past gathered enormous information with the utilization of the AI strategies is appropriate for different fields. The basic aim is to generate the analysis for driving good information which will be useful for the purpose of decision making. The quality of the decisions will definitely be enhanced. There are various machine learning techniques lies with different accuracies. The selection of the best technique such that the highest level of accuracy can be achieved. In the current research there are three techniques with different variant are tested for showing the relative accuracy of the specific technique. The all these techniques are based on supervised learning will requires training for the better accuracy. In current research paper all the techniques with different variants are trained with the different sizes of the training sets. These training sizes are 70:30, 50:50 and 30:70. The best variant is the 70:30 for the KNN. The given variant shows the highest accuracy in terms of the prediction.

1. Introduction
The decision for the purchase of the stock such that it can generate the profit is the major part of the business related to sales and purchase of stock. There are various people who are having such decision power for interpreting the stock with relation to the different factors which affect the stock prices. There are various parameters like global political scenario, local political scenario, favorable economic policy etc. which affect the stock prices. The interpretation of these parameters will help in knowing the future prospects for the stock whether it will be down or will be up. There are various instruments in the stock purchase or sales for example, trading, delivery, future sell, future buy etc. These instruments require some intelligence so that the prediction for the stock price can be done for maximizing the profit from the stock [1].

The involvement of the machine learning algorithms for the purpose of prediction of the stock price is the major area of thrusts for various researchers. There are various researchers who are emphasizing the prediction with various new machine learning techniques in the both supervised and unsupervised learning. There are abundance number of tools related to the machine learning are available which are having higher accuracy for the prediction based on the various parameters which affect the most. The large data related the previous prices of the stock prices with the relative values of different parameters

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
which affect the stock prices are collected. The parameters are either has unsupervised relation or may be having supervised relation is used for knowing the relative effect on the stock prices. The prediction algorithm based on the learning from the exiting data help in the predicting the stock prices with higher accuracy [2].

There are two categories of the machine learning techniques those are used for the purpose of prediction.

- Supervised learning
- Unsupervised leaning

### 1.1 Supervised learning:

It is the machine learning technique where the learning is done in the supervision of the analyst, which will provide various set of attributes for the learning process. Take a example of basket full of different types of fruits, the machine learning algorithm will be required which based on the values of the attributes identifies the objects for the training purpose. These attributes can be shape, color, smell etc. Once the learning for the object recognition is completed machine will be having full understanding of the shape recognition [3][4]. This generated learning will be applied to the testing set for the classification purpose. The algorithm based on the identification attributes will identify the objects with higher accuracy.

Supervised learning process is of two types [3].

- Classification
- Regression

**Classification:** It is the process where the objects are classified based on the label of the object. These labels are red, blue or may be disease or no disease etc [3].

**Regression:** it is the technique where the prediction variables are in the form of real values for example dollar, weight. The rate of the change of the value of one attribute will affect the change in another attribute value will be interpreted [5].

There are various algorithms which are used in the supervised learning process either based on classification or may be based on the regression.

- Regression
- Logistic regression
- Classification
- Naïve Bayes classification
- Decision tree
- Support vector machine

There are various advantages and dis-advantages of the supervised learning process.

**Advantages**

- The supervised learning algorithm allows the collection of the data on the continuous basis from the internal and external environment. Finally, it generates the output from the collected data.
- There are various researchers who have given the optimization criteria for the purpose of analysis.
  - The optimization criteria will help in identifying the optimum classification of the results.
- There are various real-world problems can be resolved using supervised learning mechanism [6].

**Disadvantages**

- There are various dis-advantages of the supervised learning mechanism.
- There are large data items which are difficult to classify.
- There requires large time for the purpose of training for the learning. There requires large time because of large data sets [7].
1.2 Unsupervised Learning

It is the machine learning algorithm where there is no description of the rules from the outside, rather it is the machine to learn by itself about the different aspects of the data provided for the purpose of classification. There are various attributes which are identified at the run time by the system for the purpose of analysis. Take an example of the dataset having various images of the dogs. There is no sub category of the animals. Now rather than having multiple classes there are single class of the animal and it is on the basis of the algorithm to interpret the shape and various other attributes which will build the classes into the large category of dogs. There are various sub categories based on the attribute’s values [8].

There are two categories of the algorithms in the unsupervised learning process.

- **Clustering:** it is the first category of the algorithms where the data will be clustered into the groups. There are various grouping-based classes for example the persons having higher shopping, having medium level shopping, having lower level shopping [9].
- **Association:** there are various association principles that are applied such that two entities are having association may be derived association or may be enforced relationship. In either is the case the relationship will be established for understanding of the data [10].

There are various unsupervised learning techniques in the category of the clustering.

- Exclusive (partitioning)
- Agglomerative
- Overlapping
- Probabilistic
- There are various cluster types
- Hierarchical clustering
- K-means clustering
- K-NN (k nearest neighbors)
- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis

The supervised learning techniques are better technique compared to the unsupervised because supervised learning technique provide results of higher accuracy.

2. Techniques

2.1 Logistic Regression

Logistic regression is the classification technique based on the supervised learning technique where the classification can be performed when some threshold function will be brought into the picture. So, threshold value will be highly important for the logistic regression for the classification of the data into multiple classes. The decision for the threshold value will be highly dependent on the researcher wish and experience. The main aim of the threshold value to be set is based on the optimization of the precision and the recall. These values should be kept at 1, but that may not be possible in some of the cases. The aim is to minimize the number of false positive and false negatives. The logistic regression is performed with the sigmoid function as shown in figure 1.

\[ G(z) = \frac{1}{1+e^{-z}} \]

\[ \text{eq. I} \]
2.2. Linear Regression
Dissimilar to the logistic regression the linear regression uses single feature for the purpose of classification. In the linear regression there are two variables which are related to each other in the linear way as shown in figure 2.

This table shows the two variables named as x and y. There is a relationship between these two variables such that there is relationship between the change of x with respect to y.

| x  | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|----|---|---|---|---|---|---|---|---|---|---|
| y  | 1 | 3 | 2 | 5 | 7 | 8 | 8 | 9 | 10| 12 |

2.3 KNN
It is the k nearest Neighbour is the machine learning algorithm where the data based on the feature’s values will be classified into the multiple sets. The KNN uses the mean value as the centroid value for identifying the distance of each value of the attribute from the centroid value. All the distances of each entry will be kept in the ascending order of distance. The least distance entries are considered as highly related and higher distance values are considered as the lest related. The k number of entries are picked in one class and rest are kept in other class. The KNN is the measuring technique in the supervised learning mechanism as shown in figure 3.
KNN accepts that the information is in a component space. All the more precisely, the information focuses are in a measurement space. The information can be scalars or potentially even multidimensional vectors. Since the points are in feature space, they have a notion of distance – This need not necessarily be Euclidean distance although it is the one commonly used. Each of the training data consists of a set of vectors and class label associated with each vector. In the simplest case, it will be either + or – (for positive or negative classes). But KNN, can work equally well with arbitrary number of classes. We are also given a single number "k". This number decides how many neighbors (where neighbors is defined based on the distance metric) influence the classification. This is usually a odd number if the number of classes is 2. If k=1, then the algorithm is simply called the nearest neighbor algorithm [8].

2.4. Support Vector Machine

It is another supervised learning technique. This is used for the classification and regression challenge.

**Manhattan distance**

\[ \sum_{i=1}^{n} |x_i - y_i| \]

…………………(1)

**Chebyshev**

\[ \max_{i} |x_i - y_i| \]

…………………(2)

**Euclidean**

\[ \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \]

…………………(3)
It is the classification algorithm. Where each item based on the feature values is put into the 3-dimensional space. N is the number of features. Each value of the feature being the value of particular coordinate. This way the various classes will be set up based on isolation of the information along the hyper plane as shown in figure 4 [6].

![SVM Plane](image)

**Figure 4. SVM Plane [2]**

3. Datasets
Different parameters are considered for the prediction of the stock market. The attributes of dataset are either related directly or indirectly. The dataset is taken from kaggle.com.

3.1 Training Sets and Testing Sets
There are various proportions of the training set and testing set are considered for the testing purpose. This will help in denoting the relative comparison for the accuracy at the different proportion of the training and testing set. In the following table 1, we declare three types of data variance. The first ration is for training data and second for testing set proportions.

| Sr. no. | Training and Testing set proportions |
|---------|-------------------------------------|
| 1       | 70:30                               |
| 2       | 50:50                               |
| 3       | 30:70                               |

3.2 Experimental Setup
Here, we are taking the parameters for the purpose of analysis. We are taking KNN with three different parameters Manhattan, Euclidean and Chi Square. Next parameter is Support vector machine (SVM), in this parameter we are taking four types RBF kernel, Poly Kernal, Sigmoid and Linear kernel. After that we are taking linear regression and Logistic regression. The last parameter is Linear regression with multiple variables. These are all mentioned cases that will be considered for the analysis purpose. These techniques are classified based on the results of accuracy and various other parameters as shown in table 2.
Table 2. Classifier Techniques for Measuring Accuracy

| S. no | Case                        |
|-------|-----------------------------|
| 1     | KNN with Manhattan          |
| 2     | KNN with Euclidean          |
| 3     | KNN with Chi Square         |
| 4     | SVM with RBF kernel         |
| 5     | SVM with Poly Kernel        |
| 6     | SVM with sigmoid            |
| 7     | SVM with Linear kernel      |
| 8     | Linear Regression           |
| 9     | Logistic Regression         |
| 10    | Linear Regression with multiple variables |

4. Comparison Results

4.1. Comparison of accuracy between the linear regression with single variable or multiple variables
This shows the comparison for the accuracy and RMSE for the linear regression with single variable and the multiple variables. First is representing the single variable linear regression and the second is showing multiple variables based linear regression. The linear regression with the multiple variables is having higher accuracy compared to the single variable. The RMSE for the multiple variables is lower compared to the single variable. The whole test is performed on the three types of partition of the datasets. One is the 70:30, second is the 30:70, and the third is the 50:50. The highest efficient partition is the 70:30. Where 70% is the training set and 30% is the testing set as shown in figure 5.

![Figure 5. Comparison of Accuracy and RMSE](image)

4.2. Comparison of different parameters for the SVM with different kernels
Here, we are analyzing the performance of SVM with numerous kernels. There are four different kernels being taken for the purpose of accuracy analysis based on different parameters. we are calculating accuracy, prediction, sensitivity, specificity and FPR. The best performance is for the linear kernel with the dataset partition of 70:30 as shown in figure 6.
4.3. Comparison of different parameters for KNN with different distances

The comparison for the KNN on different distances is compared with different partitions of the datasets. There are three partitions that are taken for the purpose of accuracy comparison. Here, we are calculating accuracy, prediction, sensitivity, specificity and FPR. The results show that, the highest performance is measure for KNN with chi square distance. The most accuracy comes under the case of 70:30 data set as shown in figure 7.

![Figure 7. Comparison of Parameters for different Distances on KNN](image)

The whole analysis shows that there is vast different between the level of the accuracy for the different techniques when selected with the specific variants. There are various variants for example KNN with the different types of distance evaluation methodologies. The selected of the KNN with the specific distance will provide the ability to optimize the performance.

5. Conclusion

The stock prediction will be highly advantageous process for various people who are working in the stock market. The level of the risks attached to the stock purchase and sale will be reduced if some automated tool for the prediction is used. These automated tools use the data analytics based on the machine learning algorithms for the purpose of prediction. There are various machine learning
algorithms which are having different levels of the accuracy. The current system is based on the prediction of the stock prices based on the traditional data. There are various factors which effect the system at the varying level so these attributes values are collected along with the stock prices. The prediction mechanism will be used for the purpose of enhancing the accuracy of the decision for the stock purchase and sales. In the current research paper, there are different techniques with different variants on different sizes of training and testing sets are used which help in having varying level of accuracy of the prediction. The supervised learning based on KNN is the best suitable technique at the training and testing set proportion of 70:30.

6. Future Work
Mining is the most important aspect as far as system of decision making is concerned. There are various applications which are used by different businesses for the enhancement of the quality of the decision. Various classification techniques are there which can be used for the purpose of prediction. In future the accuracy can be upgraded further by hybridization of classifiers.

7. References
[1]. L. Sayavong, Z. Wu and S. Chalita, "Research on Stock Price Prediction Method Based on Convolutional Neural Network," *2019 International Conference on Virtual Reality and Intelligent Systems* (ICVRIS), Jishou, China, 2019, pp. 173-176, doi: 10.1109/ICVRIS.2019.00050.
[2]. Y. Wang and Y. Wang, "Using social media mining technology to assist in price prediction of stock market," *2016 IEEE International Conference on Big Data Analysis* (ICBDA), Hangzhou, 2016, pp. 1-4, doi: 10.1109/ICBDA.2016.7509794.
[3]. Amit Kumar Sirohi, Pradeep Kumar Mahato and V. Attar, "Multiple Kernel Learning for stock price direction prediction," *2014 International Conference on Advances in Engineering & Technology Research* (ICAETR - 2014), Unnao, 2014, pp. 1-4, doi: 10.1109/ICAETR.2014.7012901.
[4]. L. Zhao and L. Wang, "Price Trend Prediction of Stock Market Using Outlier Data Mining Algorithm," *2015 IEEE Fifth International Conference on Big Data and Cloud Computing, Dalian*, 2015, pp. 93-98, doi: 10.1109/BDCloud.2015.19.
[5]. Z. Zhang, Y. Shen, G. Zhang, Y. Song and Y. Zhu, "Short-term prediction for opening price of stock market based on self-adapting variant PSO-Elman neural network," *2017 8th IEEE International Conference on Software Engineering and Service Science* (ICSESS), Beijing, 2017, pp. 225-228, doi: 10.1109/ICSESS.2017.8342901.
[6]. A. Ariyo, A. O. Adewumi and C. K. Ayo, "Stock Price Prediction Using the ARIMA Model," *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, Cambridge*, 2014, pp. 106-112, doi: 10.1109/UKSim.2014.67.
[7]. H. Liu and B. Song, "Stock Price Trend Prediction Model Based on Deep Residual Network and Stock Price Graph," *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, Hangzhou, China, 2018, pp. 328-331, doi: 10.1109/ISCID.2018.10176.
[8]. M. Yu and J. Wu, "CEAM: A Novel Approach Using Cycle Embeddings with Attention Mechanism for Stock Price Prediction," *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Kyoto, Japan, 2019, pp. 1-4, doi: 10.1109/BIGCOMP.2019.8679218.
[9]. M. Billah, S. Waheed and A. Hanifa, "Stock market prediction using an improved training algorithm of neural network," *2016 2nd International Conference on Electrical, Computer &
[10]. D. Wei, "Prediction of Stock Price Based on LSTM Neural Network," 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Dublin, Ireland, 2019, pp. 544-547, doi: 10.1109/AIAM48774.2019.00113.

[11]. W. Yang, "Stock Price Prediction Based on Fuzzy Logic," 2007 International Conference on Machine Learning and Cybernetics, Hong Kong, 2007, pp. 1309-1314, doi: 10.1109/ICMLC.2007.4370347.