Intelligent Stock Data Prediction using Predictive Data Mining Techniques

Pankaj kumar
(pankaj.thapar1416@gmail.com)
Master of Engineering (Software Engineering),
CSED Department, Thapar University, Punjab, India

Dr. Anju Bala
(anjubala@thapar.edu)
Assistant Professor, CSED Department, Thapar University, Punjab, India

Abstract—Cloud computing is the one of the admired paradigms of current era, which facilitates the users with on demand services and pay as you use services. It has tremendous applications in almost every sphere such as education, gaming, social networking, transportation, medical, business, stock market, pattern matching, etc. Stock market is such an industry where lots of data is generated and benefits are reaped on the basis of accurate prediction. So prediction is a vital part of stock market. Therefore, an attempt is being made to predict the stock market based on the given data set of stock market along with some features; using the techniques available for predictive data mining. Machine learning is one of the upcoming trends of data mining; hence few machine learning algorithms have been used such as Decision tree, Linear model, Random forest and further their results have been compared using the classification evaluation parameters such as H, AUC, ROC, TPR, FPR, etc. Random forest have been consider as the most effective model as it yield the highest accuracy of 54.12% whereas decision tree and linear model gives the accuracy of 51.87% and 52.83% respectively.

Keywords: cloud computing, data mining, machine learning

1. Introduction

Prediction system of stock market is very crucial and essentially important because it deals with the huge amount of money and in today's growing and forward time money is first priority. The predicted value directly affects the stock price and no one take risk to drop down in stock market index. So the due to money involvement and the reputation of the shares stock market needs to be a perfect or more accurate prediction about their upcoming market trends. In this paper various machine learning algorithms have been applied to the stock data set and the objective is to predict the stock market. Based on the accuracy, the comparison of the models used is shown in the paper. Also the ROC [10] graph, AUC plot, H-measure, smoothed score distributions are shown graphically.

In the proposed problem, three machine learning models have been used. These models are: Decision tree model [12], Linear model [13] and Random forest model [14]. We make two division of data first is training data and second is testing data. Firstly these models train the data then after completion of training test the data and find all the evaluation parameters of all models. Also find the accuracy of these models. These resulted graphs based on the evaluation parameters used to compare the models and finally we got the more accurate model who gives the close to or superior results.

2. Related Work

Most of the prediction problem solution of stock was introduced in many researches. Lee [1] uses Taiwan stock exchange data to p papers and publications. M. Credict the stock behavior and forecast the mid-term forecast price trend with the use of feature selection [7] and forecasting the stock in ARIMA based network forecasting system [8]. They use the linear model with formula:

$$x_t = r + \sum_{i=1}^{p} \varphi_i x_{t-i} \sum_{j=1}^{q} \theta_j e_{t-j} + e_t$$

They use 7 hidden layers in the neural network to train the data set using features of TSEWSI. They compare the ARIMA in two different manners one is without differentiate and second is differentiate and show the result. Binoy B. Nair, Dharini N.Mohana, Mohandas V.P. [3] automates the decision tree with neuro_fuzzy system with the technical analysis to extract the features and decision tree to select them. They validate their results on different four major stock exchanges with the help of confusion matrix. One of the other research paper on stock prediction [1] using the support vector machine and while feature selecting they use F_score and FSSFS method. F_score mathematical formula is:

$$F(i) \equiv \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_--1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

Then apply cross validation [11] after just classifier then compare the result with prediction accuracy of Back Propagation Neural Network and Support Vector Machine. In their methodology the SVM is the superior with back propagation neural network.

3. Methodology and Experimental setup

While performing any experiment it is necessary to have a dataset and a proper methodology as to how to work on that dataset so that a proper prediction could be made in lieu of future decisions to be made. In our experiment of prediction of stock market, we have a stock dataset with 21 features, 57772 data entries and target at 22nd position. It is a binary classification problem [15] with output values as 0 and 1; so therefore we have to apply classification models of machine learning [5]. Before using any model with the dataset, we must ensure that our data is pre-processed; it means that dataset should be in .csv (comma separated values) format, there should be no null values or any noisy data in the dataset.

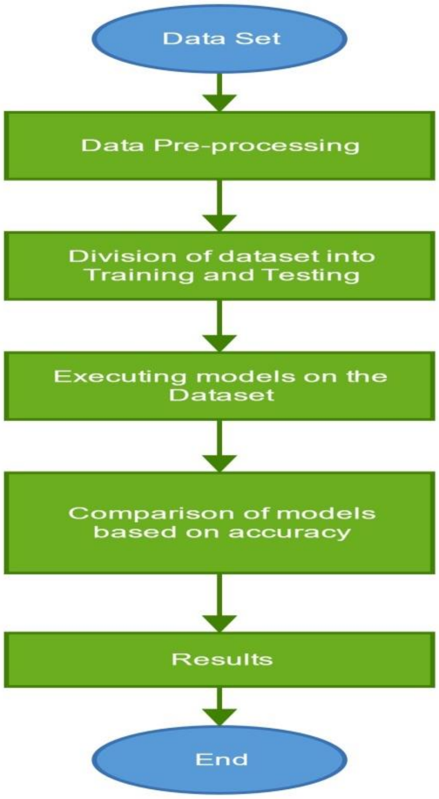
Data Set
Data Pre-processing
Division of dataset into Training and Testing
Executing models on the Dataset
Comparison of models based on accuracy
Results
End

Figure 1. Approach to solve Classification problem

After data pre-processing [9], next step is to divide the dataset into two parts: - training data subset and testing data subset (normally it is in 70:30 ratio but it could be changed as per the experimenter's requirements or as per the performance of the model). Once the dataset is divided into training and testing data subsets, classification models are executed on the dataset and results are generated in the form of evaluation parameters such as H, Gini, AUC, F-measure, Sensitivity, Specificity, TPR, FPR, Error Rate, Recall, Precision, Accuracy and Time. Any of the above mentioned parameters can be chosen to compare the results, we have chosen accuracy parameter to compare the classification models and choose the best one giving the most accurate results. Confusion Error matrix show the true positive rate and false positive rate. Table 1, shown the confusion error matrix is shown below:

	Predicted Class	
Actual Class	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Table 1. Confusion Matrix

Various other parameters results in the form of graph are also shown in the paper ahead. Final results will help the experimenter to predict the stock market so that correct decisions could be made. For the experiment, we use the R-studio for implement all the models in R language and all the results and graph are build through the R. For linear model use ‘nnet’, ‘hmeasure’ packages and for decision tree we use the ‘nnet’, ‘hmeasure’, ‘caret’ packages and for random forest we use ‘randomForest’ and ‘hmeasure’ packages. R gives the best graphically representation of data and result set.

4. Results and discussion

The problem of stock market classification problem is done through the approach discuss in this paper. Finally we have evaluation parameters to compare the result of binary classification problem. Now we analyze three models results and after the approach we get some graphs. Every model gives the ROC [10] plot graph, H-measure, Area under curve (AUC) graph and smoothed score distribution graph. After studying the graph ROC the random forest model shows the closest value to 1 rather than the other models. If the ROC [10] value close to 1 than it excellent and the accuracy is more.

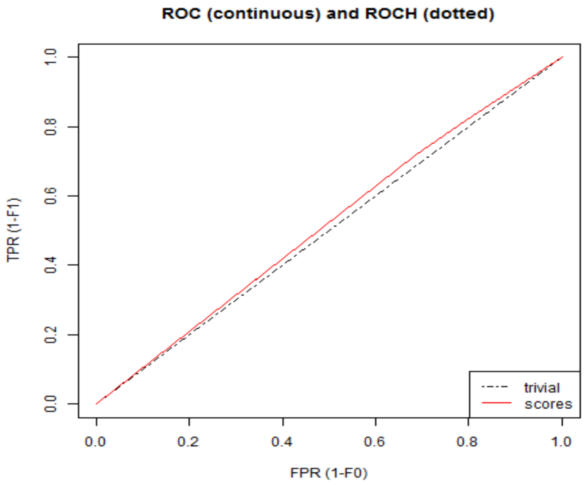

Figure 1. Decision tree ROC plot graph

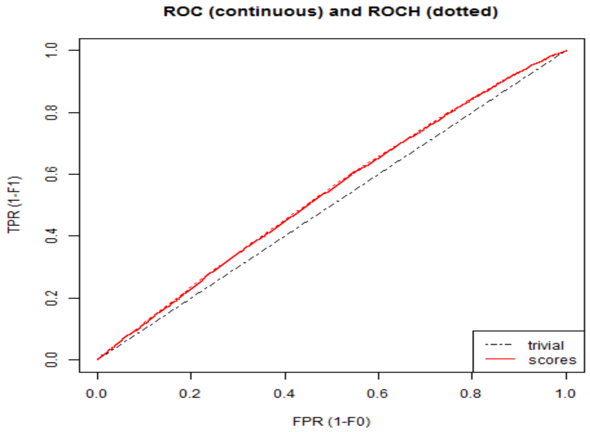

Figure 5. Linear Model ROC plot graph

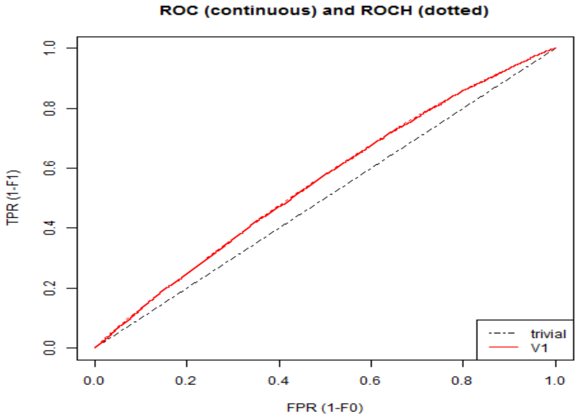

Figure 9. Random Forest ROC plot graph

In the binary classification AUC [6] is the popular evaluation metric and it shows the accuracy in terms of true positive rate. If the AUC value is close to 1 that means the true positive rate is increase or more and if it is near to 0 than the true positive rate decrease or less. In our experiment the AUC value of random forest model (0.554) is higher than the linear model (0.538) and decision tree model (0.517). So the random forest shows the high true positive ratio that means more accuracy.

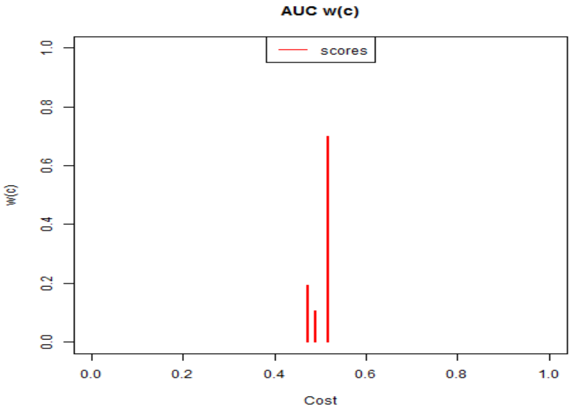

Figure 3. Decision tree AUC graph plot

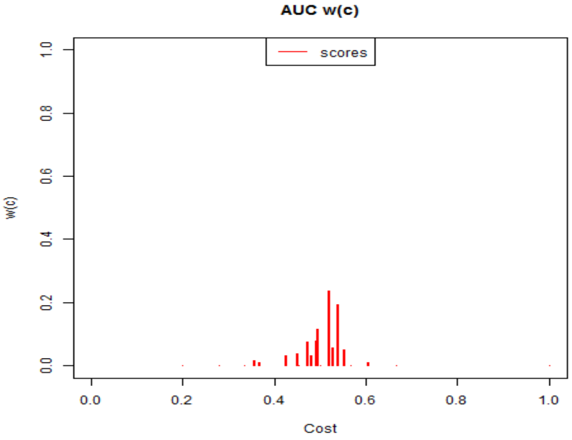

Figure 7. Linear Model AUC graph plot

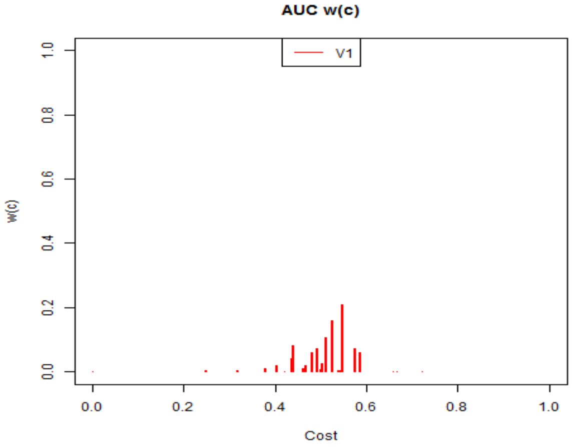

Figure 11. Random Forest AUC graph plot

The smoothed score distribution graph represents the score versus density graph and how the change in the density while score will be increased. The peak of the curve shows the mean of the graph. It is under the normal distribution and due to smoothing the noise part will be eliminated.

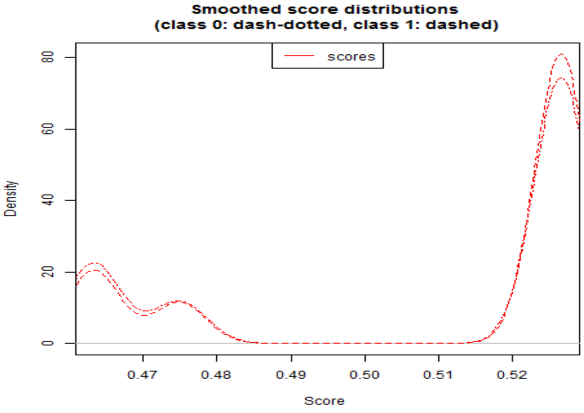

Figure 4. Decision tree density v/s score graph

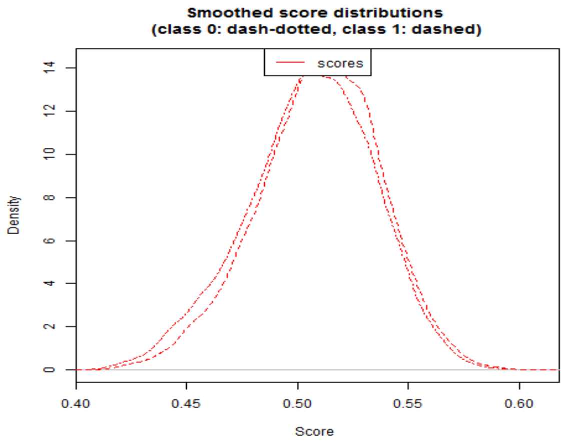

Figure 8. Linear Model density v/s score graph

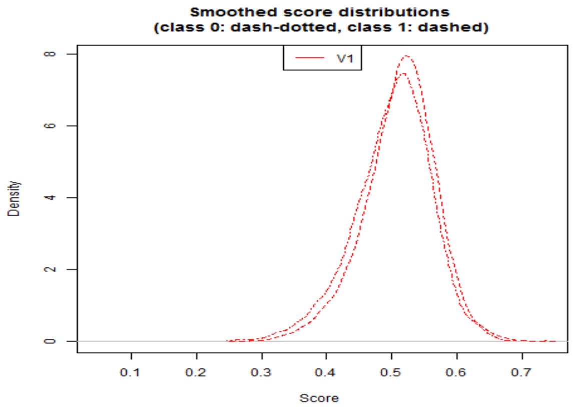

Figure 12. Random Forest density v/s score graph

Finally, after all analysis and result set of the prediction problem solving though our approach we found that the accuracy of the random forest is superior rather than the linear model and decision tree. In the table 2, it shows the True positive rate (TPR), false positive rate (FPR), Area under the curve (AUC) and accuracy of the all models. We can clearly see that AUC value of random forest is closest to 1 and other models AUC value smaller than the random forest model. Accuracy of random forest is 54.12 and other models decision tree and linear model accuracy is 51.87 and 52.83 respectively. So, random forest model gives the more accurate accuracy when we applied it on the binary classification problem stock data.

Evaluation matrix	Decision tree	Linear model	Random Forest
TPR	0.717	0.656	0.627
FPR	0.685	0.603	0.549
AUC	0.517	0.538	0.554
Accuracy	51.87	52.83	54.12

Table 2. Evaluation matrix for decision tree, Linear model and Random forest

5. Future work

In the future, we can merge the different models for more accuracy to predict the stock market. This model can be combined and there result efficiency could be merge using the process of ensembling. Ensembling methods are used to obtain the more accurate predictive results through using multiple learning algorithms. So it can be used in the future for solving these types of problems for more accuracy and efficient predicted values. We can also implement this approach in cloud environment with the help of some tools like hadoop or Amazon cloud.

6. Conclusion

This study of problem solving of binary classification data based on the machine learning models gives the best out of models which is experimentally used. The overall study and experiments shows the random forest is the more efficient to predict the binary classification due to its accuracy. Random forest accuracy is much better than the others model. We can apply that model on bulk data sets, its approx 57772 entry of data in comma separated format data sheet with targeted values in binary (0,1) format. Our experimental setup gives the random forest as the best model based on the accuracy measure on this problem.

7. References

[1]. M. C. Lee (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction *[online] 36 (2009) 10896–10904 Available: www.elsevier.com/locate/eswa*

[2]. E. P. George Box; Gwilym M. Jenkins; C. Reinsel Gregoly, Time Series Analysis, Prentic; Hall, 1994.

[3]. Binoy B. Nair, Dharini N.Mohana, Mohandas V.P., A Stock market trend prediction system using a hybrid decision tree-neuro-fuzzy system. 2010 International Conference on Advances in Recent Technologies in Communication and Computing. [online] Available: *http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5655295&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F5654667%2F5655294%2F05655295.pdf%3Farnumber%3D5655295*

[4]. E. W. Saad, D.V. Prokhorov, and, D.C. Wunsch, *"Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks", IEEE Transactions on Neural Networks, Vol. 9, No. 6, pp. 1456-1470, 1998.*

[5]. S.B.Kotsiantis,"*Supervised Machine Learning: A Review of classification Techniques*" in Emerging Artificial Intelligence Application in Computer Science, vol 160, Amsterdam, Nederland.

[6]. Andrew P. Bradley. (1997, july.). The use of area under the ROC curve in the evaluation of machine learning algorithms. Volume 30, Issue 7, Pages 1145–1159 *Available:http://www.sciencedirect.com/science/article/pii/S0031320396001422*

[7]. G. Forman. (2003, January.). An extensive empirical study of feature selection metrics for text classification. Volume 3, Pages 1289-1305 *Available: http://dl.acm.org/citation.cfm?id=944974*

[8]. Connor, J.T.; Martin, R.D.; Atlas, L.E. "Recurrent neural networks and robust time series prediction," IEEE *Transactions on Neural Networks,* Vol. *5,* Iss. 2, pp. 240-254, March 1994.

[9]. D. V. Zhora (2005). Data preprocessing for stock market forecasting using random subspace classifier network. IEEE international conference joint conference on neural network. Available: *http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1556304&newsearch=true&queryText=data%20preprocessing%20for%20stock%20market*

[10]. P. A. Emelia Akashah, S. K. Sugathan, Anthony TS. Ho. Receiver Operating Characteristic (ROC) Graph to Determine the Most Suitable Pairs Analysis Threshold Value. Advances in Electrical and Electronics Engineering - IAENG Special Edition of the World Congress on Engineering and Computer Science 2008. [Online]. Available: *http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5233162*

[11]. M. Das, O. Y. (2008) Hao. Efficient Cross Validation Over Skewed Noisy Data. IEEE International Conference on Systems, Man and Cybernetics (SMC 2008). Available: *http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4811368*

[12]. U. Johansson, H. Bostrom, T. Lofstrom. (2013). Conformal Prediction Using Decision Trees. IEEE 13th International Conference on Data Mining. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6729517

[13]. B. Xu, D. Han, C. Xu. (2012). Linear Fixed Weight Combination Prediction Model and Model Optimum Seeking Method. [Online]. Available:http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6246502

[14]. K. V. Kavitha, R. Saritha, S. S. V. Chandra. (2013). Computational Prediction of Continuous B-Cell Epitopes Using Random Forest Classifier. 4th ICCCNT. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6726820

[15]. P. Jeatrakul, K.W. Wong. (2009). Comparing the Performance of Different Neural Networks for Binary Classification Problems. Eighth International Symposium on Natural Language Processing. [Online]. Available:http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5340935