RESEARCH ARTICLE OPEN ACCESS

Forecasting the direction of stock market index movement using three data mining techniques: the case of Tehran Stock Exchange

[1]Sadegh Bafandeh Imandoust and [2]Mohammad Bolandraftar
[1]*Economic Department, Payame Noor University, Tehran, Iran*
[2]*Department of Contracts, Bandar Abbas Oil Refining Company, Bandar Abbas, Iran*

Abstract
Prediction of stock price return is a highly complicated and very difficult task because there are many factors such that may influence stock prices. An accurate prediction of movement direction of stock index is crucial for investors to make effective market trading strategies. However, because of the high nonlinearity of the stock market, it is difficult to reveal the inside law by the traditional forecast methods. In response to such difficulty, data mining techniques have been introduced and applied for this financial prediction. This study attempted to develop three models and compared their performances in predicting the direction of movement in daily Tehran Stock Exchange (TSE) index. The models are based on three classification techniques, Decision Tree, Random Forest and Naïve Bayesian Classifier. Ten microeconomic variables and three macroeconomic variables were chosen as inputs of the proposed models. Experimental results show that performance of Decision Tree model (80.08%) was found better than Random Forest (78.81%) and Naïve Bayesian Classifier (73.84%).
Keywords: Predicting direction of stock market index movement- Decision Tree- Random Forest- Naïve Bayesian Classifier

I. Introduction

Prediction of stock price return is a highly complicated and very difficult task because there are too many factors such as political events, economic conditions, traders' expectations and other environmental factors that may influence stock prices. In addition, stock price series are generally quite noisy, dynamic, nonlinear, complicated, nonparametric, and chaotic by nature [1-4]. The noisy characteristic refers to the unavailability of complete information from the past behavior of financial markets to fully capture the dependency between future and past prices [5-9].

Most of the studies have focused on the accurate forecasting of the value of stock price. However, different investors adopt different trading strategies; therefore, the forecasting model based on minimizing the error between the actual values and the forecasts may not be suitable for them. Instead, accurate prediction of movement direction of stock index is crucial for them to make effective market trading strategies. Specifically, investors could effectively hedge against potential market risk and speculators as well as arbitrageurs could have opportunity of making profit by trading stock index whenever they could obtain the accurate prediction of stock price direction. That is why there have been a number of studies looking at direction or trend of movement of various kinds of financial instruments [10-14].

In recent years, there have been a growing number of studies looking at the direction or trend of movements of financial markets. Although there exist some articles addressing the issue of forecasting financial time series such as stock market index, most of the empirical findings are associated with the developed financial markets (UK, USA, and Japan). However, few researches exist in the literature to predict direction of stock market index movement in emerging markets [15-17].

Because of the high nonlinearity of the stock market, it is difficult to reveal the inside law by the traditional forecast methods [18]. The difficulty of prediction lies in the complexities of modeling human behavior [19]. In response to such difficulty, data mining (or machine learning) techniques have been introduced and applied for this financial prediction. Recent studies reveal that nonlinear models are able to simulate the volatile stock markets well and produce better predictive results than traditional linear models in stock market tendency exploration [20]. With the development of artificial intelligence (AI) techniques investors are hoping that the market mysteries can be unraveled because these methods have great capability in pattern recognition problems such as classification and prediction.

In the present study, three classification methods, derived from the field of machine learning, are used to predict the direction of movement in the daily TSE index using. The employed methods are Random Forest, Decision Tree, and Naïve Bayesian Classifier. The remainder of the paper is organized as follows: Section 2 reviews the literature. Section 3 provides a brief description of Random Forest, Decision Tree, and Naïve Bayesian Classifier. Section 4 presents the

research design and methodology. Section 5 describes finding results from the comparative analysis. Finally, in the last section concluding remarks are given.

II. Literature Review

Data mining techniques have been introduced for prediction of movement sign of stock market index since the results of Leung et al. and Chen et al. [21], where LDA, Logit and Probit and Neural network were proposed and compared with parametric models, GMM-Kalman filter.

Kumar & Thenmozhi [22] compare the predictive ability of Random Forest and SVM with ANN, Discriminant Analysis and Logit model to predict Indian stock index movement based on economic variable indicators. Empirical experimentation suggests that the SVM outperforms the other classification methods in terms of predicting the S&P CNX NIFTY index direction and Random Forest method outperforms ANN, Discriminant Analysis and Logit model used in this study.

Afolabi and Olatoyosi [21] use fuzzy logics, neuro-fuzzy networks and Kohonen's self organizing plan for forecasting stock price. The finding results demonstrate that the deviation in Kohonen's self organizing plan is less than that in other techniques.

Chen and Han [24] propose an original and universal method by using SVM with financial statement analysis for prediction of stocks. They applied SVM to construct the prediction model and selected Gaussian radial basis function (RBF) as the kernel function. The experimental results demonstrate that their method improves the accuracy rate.

Abbasi and Abouec [20] investigate the current trend of stock price of the Iran Khodro Corporation at Tehran Stock Exchange by utilizing an Adaptive Neuro-Fuzzy Inference System (ANFIS). The findings of the research demonstrate that the trend of stock price can be forecast with a low level of error.

Jandaghi et al. [16] use ARIMA and Fuzzy-Neural networks to predict stock price of SAIPA auto-making company. The finding results show the preference of nonlinear Neural-Fuzzy model to classic linear model and verify the capabilities of Fuzzy-neural networks in this prediction.

Kara et al. [18] attempt to develop two models and compare their performances in predicting the direction of movement in the daily Istanbul Stock Exchange (ISE) National 100 Index. The models are based on two classification techniques, ANN and SVM. They selected ten technical indicators as inputs of the proposed models. Experimental results show that average performance of ANN model (75.74%) was found significantly better than SVM model (71.52%).

Other methods that have been used to predict the stock market include KNN and Bayesian belief networks.

III. Theoretical background

3.1 Random Forest

Random forest (RF) is a popular and very efficient algorithm, based on model aggregation ideas, for both classification and regression problems, introduced by Breiman [20]. A RF is in fact a special type of simple regression trees ensemble, which gives a prediction based on the majority voting (the case of classification) or averaging (the case of regression) predictions made by each tree in the ensemble using some input data [14].

The RF is an effective prediction tool in data mining. It employs the Bagging method to produce a randomly sampled set of training data for each of the trees. This method also semi-randomly selects splitting features; a random subset of a given size is produced from the space of possible splitting features. The best splitting is feature deterministically selected from that subset. A pseudo to classify a test instance, the random forest classifies the instance by simply combining all results from each of the trees in the forest. The method used to combine the results can be as simple as predicting the class obtained from the highest number of trees.

The principle of random forests is to combine many binary decision trees built using several bootstrap samples coming from the learning sample L and choosing randomly at each node a subset of explanatory variables X. More precisely, with respect to the well-known CART model building strategy performing a growing step followed by a pruning one, two differences can be noted. First, at each node, a given number of input variables are randomly chosen and the best split is calculated only within this subset. Second, no pruning step is performed, so all the trees of the forest are maximal trees.

For each observation, each individual tree votes for one class and the forest predicts the class that has the plurality of votes. The user has to specify the number of randomly selected variables to be searched through for the best split at each node. The largest tree possible is grown and is not pruned. The root node of each tree in the forest contains a bootstrap sample from the original data as the training set. The observations that are not in the training set, roughly 1/3 of the original data set, are referred to as out-of-bag (OOB) observations. One can arrive at OOB predictions as follows: for a case in the original data, predict the outcome by plurality vote involving only those trees that did not contain the case in their corresponding bootstrap sample. By contrasting these OOB predictions with the training set outcomes, one can arrive at an estimate of the prediction error rate, which is referred to as the OOB error rate. The RF

construction allows one to define several measures of variable importance.

3.2 Decision Tree

Decision tree (DT) algorithm is a data mining induction technique which recursively partitions a dataset of records using depth-first greedy approach or breadth-first approach until all the data items belong to a particular class.

A DT is a mapping from observations about an item to conclusion about its target value as a predictive model in data mining and machine learning. Generally, for such tree models, other descriptive names are classification tree (discrete target) or regression tree (continuous target). The general idea of a decision tree is splitting the data recursively into subsets so that each subset contains more or less homogeneous states of target predictable attribute. At each branch in the tree, all available input attributes are calculated again for their own impact on the predictable attribute.

In a classification problem, which the target variable is categorical; all variables in a dataset are assigned to the root node. The data is then divided into two child nodes, based on a splitting criterion that splits data characterized by a question. A splitting criterion at each node depends on the single variable value selected from the dataset.

Depending on the answer to the question, whether yes or no, data is split into left or right nodes. The splitting of parent nodes continues until the resulting child nodes are pure or until the numbers of cases inside the node reach a predefined number. Thus the tree is constructed by examining all possible splits at each node until maximum depth is reached or no gain in purity is observed with further splitting. Nodes that are pure or homogeneous, which could not be split further, are called terminal or leaf nodes, and they are assigned to a class.

DT classification technique is performed in two phases: tree building and tree pruning. Tree building is done in top-down manner. It is during this phase that the tree is recursively partitioned till all the data items belong to the same class label. Tree pruning is done in a bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training dataset).

Although other methodologies such as neural networks and rule based classifiers are the other options for classification, DT has the advantages of interpretation and understanding for the decision makers to compare with their domain knowledge for validation and justify their decisions.

Figure 1 is an illustration of the structure of DT built by some credit database, where x, y, z, u in inner nodes of the tree are predictive attributes and "good" and "bad" are the classifications of target attribute in the credit database.

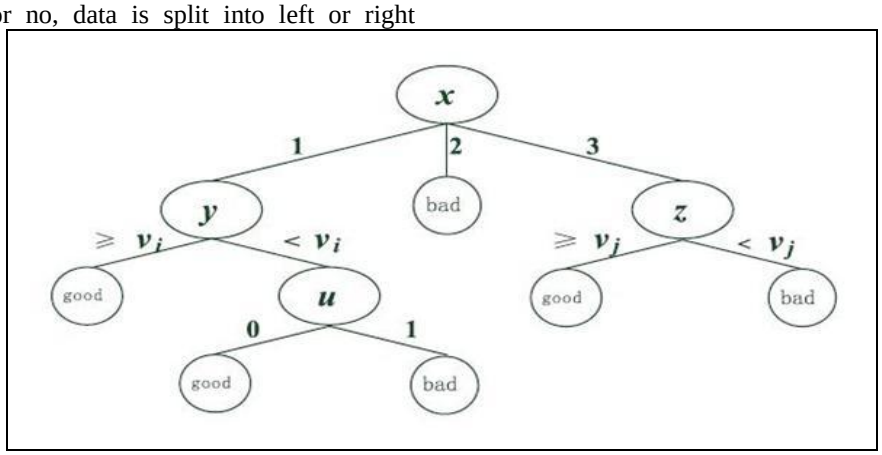


Fig. 1. A structure of decision tree

3.3 Naïve Bayesian Classifier

A Naive Bayesian Classifier (NBC) is well known in the machine learning community. It is one kind of Bayesian classifier, which is now recognized as a simple and effective probability classification method, and works based on applying Bayes' theorem with strong (naive) independence assumptions.

The NBC is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naïve Bayes can often outperform more sophisticated classification methods.

In simple terms, a NBC assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Given feature variables F_1, F_2,…, F*n* and a class variable C. The Bayes' theorem states

$$P(C \mid F_1, F_2, ..., F_n) = \frac{P(C)P(F_1, F_2, ..., F_n \mid C)}{P(F_1, F_2, ..., F_n)} \quad (1)$$

Assuming that each feature is conditionally independent of every other feature for i : one obtains

$$P(C \mid F_1, F_2, \ldots, F_n) = \frac{P(C)\prod_{i=1}^{n} P(F_i \mid C)}{P(F_1, F_2, \ldots, F_n)} \quad (2)$$

The denominator serves as a scaling factor and can be omitted in the final classification rule.

$$\arg\max_{c}^{a} P(C = c)\prod_{i=1}^{n} P(F_i = f_i \mid C = c) \quad (3)$$

IV. Research design

The research data used in this study is the direction of daily closing price movement in the TSE Index. The entire data set covers the period from April 17, 2007 to March 18, 2012. The total number of cases is 1184 trading days. The number of days with increasing direction is 654, the number of days with decreasing direction is 523, and the number of days with constant direction is 7 days. In other words, in 55.24% of the all days market has an increasing direction, in 44.17% of the all days it has a decreasing direction, and in 0.59% of the all trading days the direction is constant. The historical data and the number of days for each year are given in Table 1. The research historical data was obtained from Tehran Securities Exchange Technology Management Co. and the official website of the Tehran Stock Exchange.

80% of the data was used for training the models, and the remaining 20% was used to test them and compare their performance. That is, 947 of all the cases are training data and 237 cases are used to test the models.

The direction of daily change in the stock price index is categorized as ‘‘Positive’’, ‘‘Negative’’ and “No Change”. If the TSE Index at time t is higher than that at time t-1, the direction is ‘‘Positive’’, if the TSE Index at time t is lower than that at time t-1, the direction is ‘‘Negative’’, and if the TSE Index at time t is equal to that at time t-1, the direction is ‘‘No Change’’.

Year	Increase	Decrease	Constant (No Change)	Total
2007	100	74	1	175
2008	99	139	2	240
2009	136	103	0	238
2010	159	81	1	242
2011	130	104	3	237
2012	30	22	0	52
Sum	**654**	**523**	**7**	**1184**

Table 1. The number of cases in the entire data set

In the study technical and fundamental variables have been employed, and the input variables to models are divided into three parts. In the first part, the input variables to models are technical indicators, and the predictive power of the data mining methods, used in the study, are evaluated and compared. In the second part, fundamental inputs are applied to the models, and in the last part, fundamental and technical variables are applied to machine learning techniques simultaneously. Figure 2 depicts an overview of the research steps.

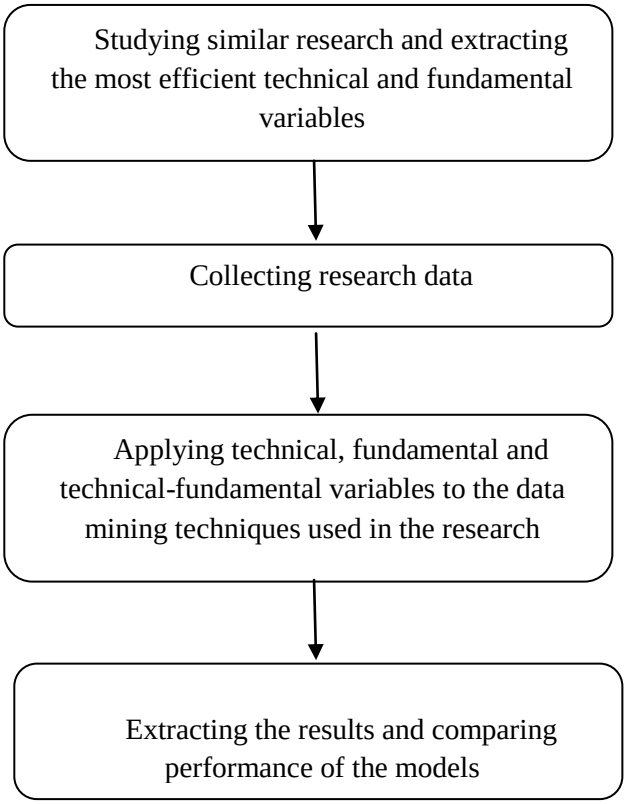


Fig 2. Research steps

Ten technical indicators for each case were used as input variables. A variety of technical indicators are available. Some technical indicators are effective under trending markets and others perform better under no trending or cyclical markets. Table 2 summarizes the selected technical indicators.

Name of indicators
Simple 10-day moving average
Weighted 10-day moving average
Momentum
Stochastic K%
Stochastic D%
RSI (Relative Strength Index)
MACD (moving average convergence divergence)
Larry William's R%
A/D (Accumulation/Distribution) Oscillator
CCI (Commodity Channel Index)

Table 2. Selected technical indicators and their formulas

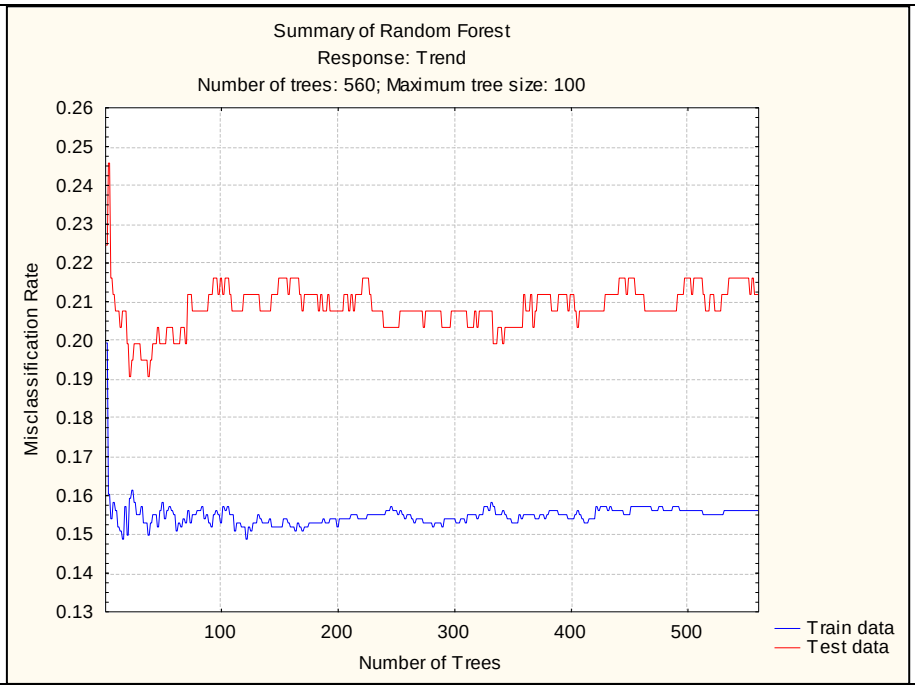
Since oil, gold and USD/IRR play prominent roles in the Iranian economy, these three variables were considered as fundamental indicators.

V. Experimental results

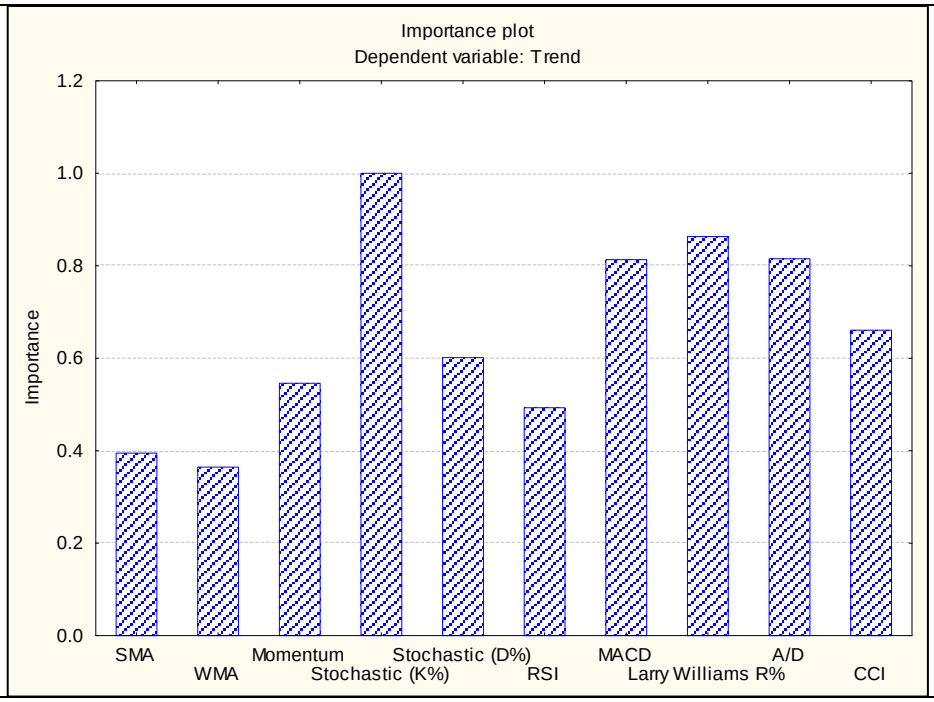
5.1 Random Forest

In this part, the results of applying technical, fundamental and technical-fundamental indicators to RF are considered.

Figure 3 shows the finding results of applying technical indicators to RF. As we can see, Stochastic (K%) is the most important variable, and 560 trees lead to the best result. Also, in this part, the RF model can predict stock market movement direction with the accuracy of 78.81%.



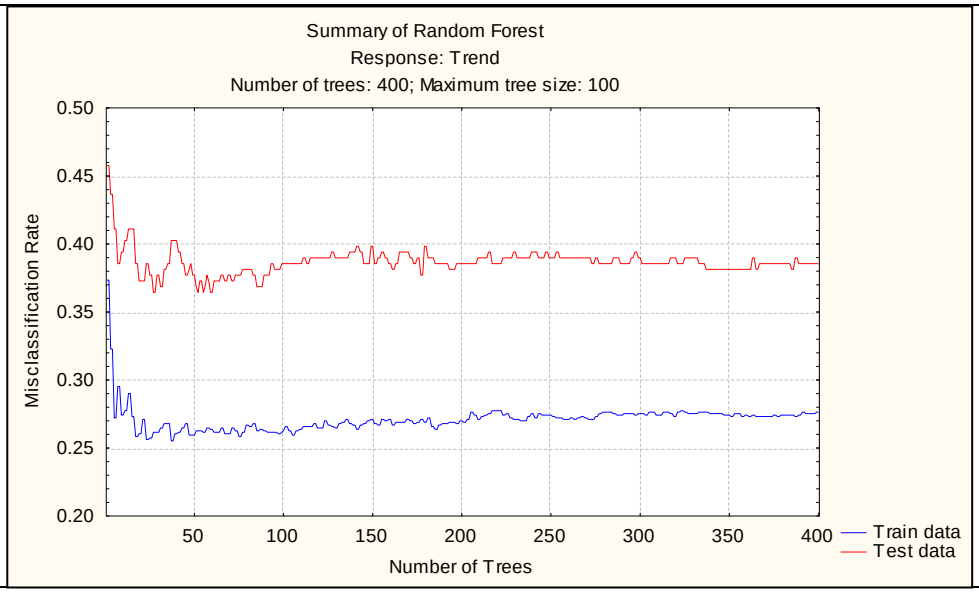
(a)



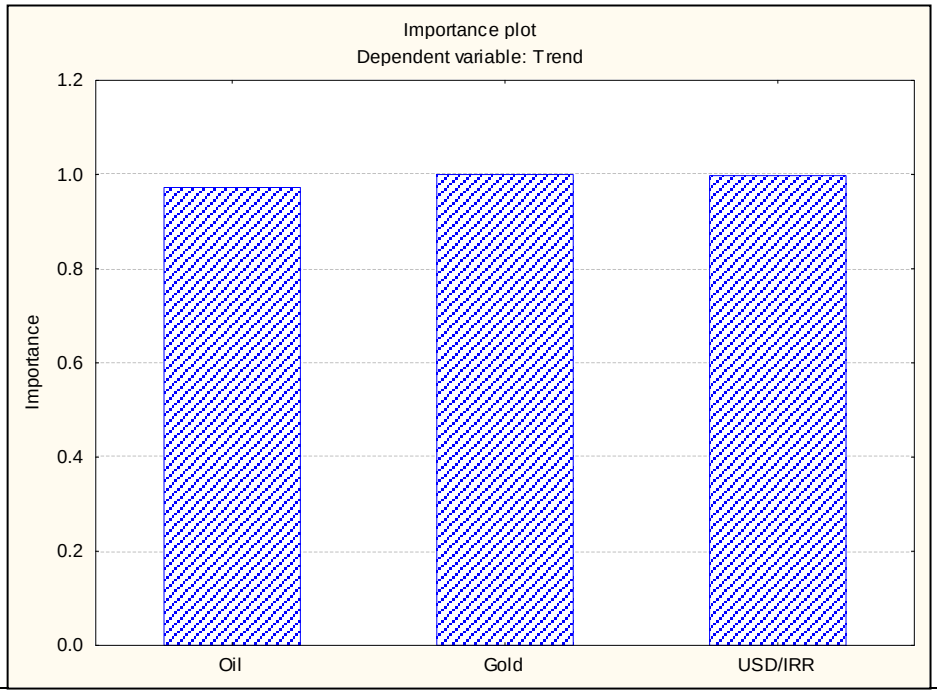
(b)

Fig. 3. The results of applying technical indicators to RF (a) misclassification rate graph (b) importance plot

In the second part, fundamental indicators are employed in RF. The RF model, in this part, can predict TSE market movement direction with the accuracy of 61.86%. As it is demonstrated in Figure 4, when the number of trees is 400, the optimized answer is obtained. Additionally, the variable Gold is as important as the variable USD/IRR.



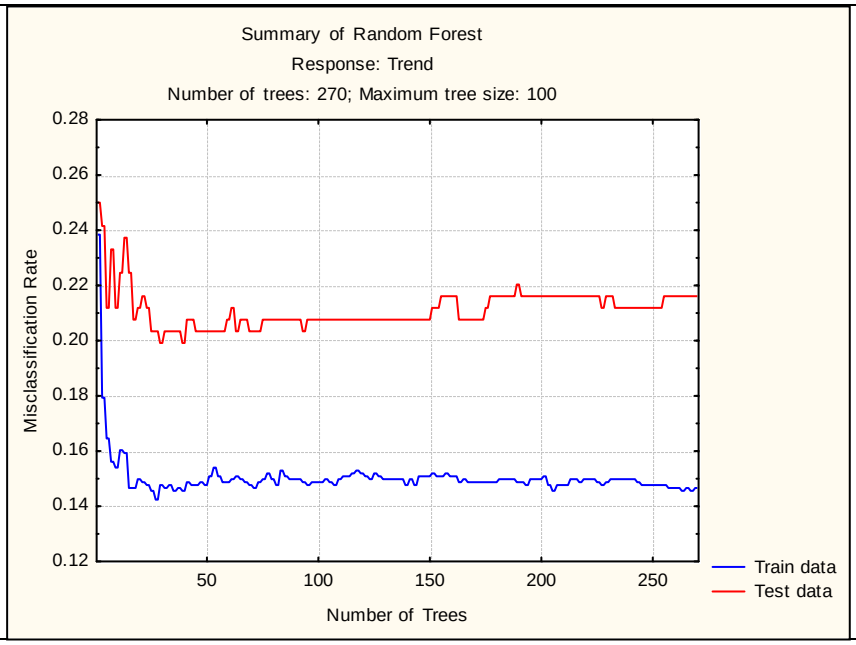
(a)



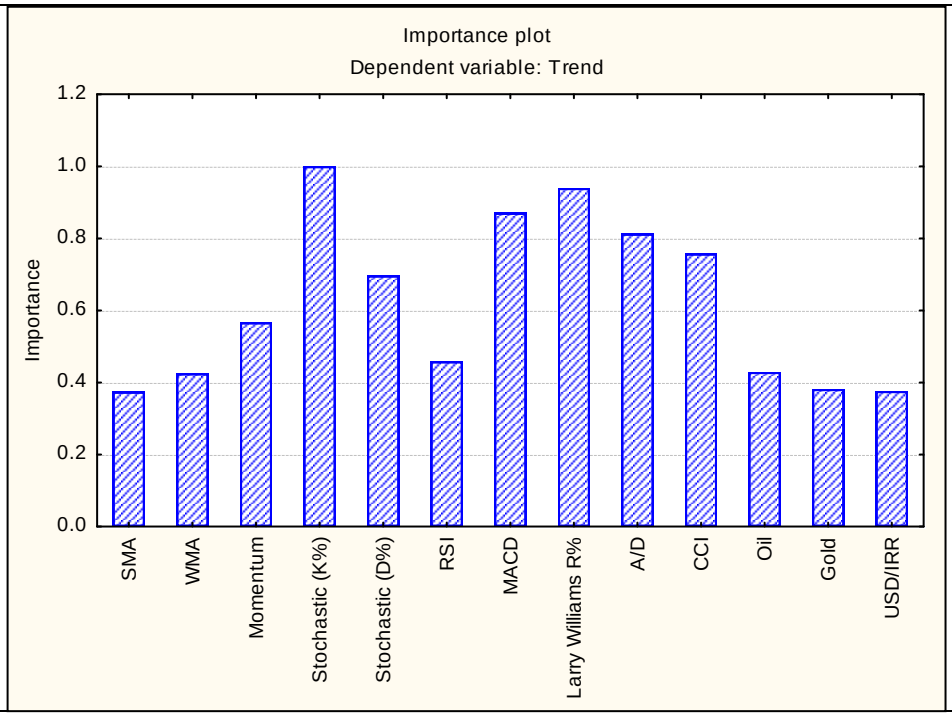
(b)

Fig. 4. The results of applying fundamental indicators to RF (a) misclassification rate graph (b) importance plot

Finally, in the third part 13 technical-fundamental indicators were exploited in RF. The results, which are displayed in Figure 5, indicate that the optimal number of trees is 270 and Stochastic (K%) is the most notable variable. Furthermore, the RF model can forecast the market movement direction with the accuracy of 78.39%.



(a)



(b)

Fig. 5. The results of applying technical-fundamental indicators to RF (a) misclassification rate graph (b) importance plot

5.2 Decision Tree

In the case of technical input indicators, the accuracy of the optimal model is 80.08%. The importance plot of the DT model is given in Figure 6. As it is shown, the variable MACD is the most considerable variable in this part of analysis.

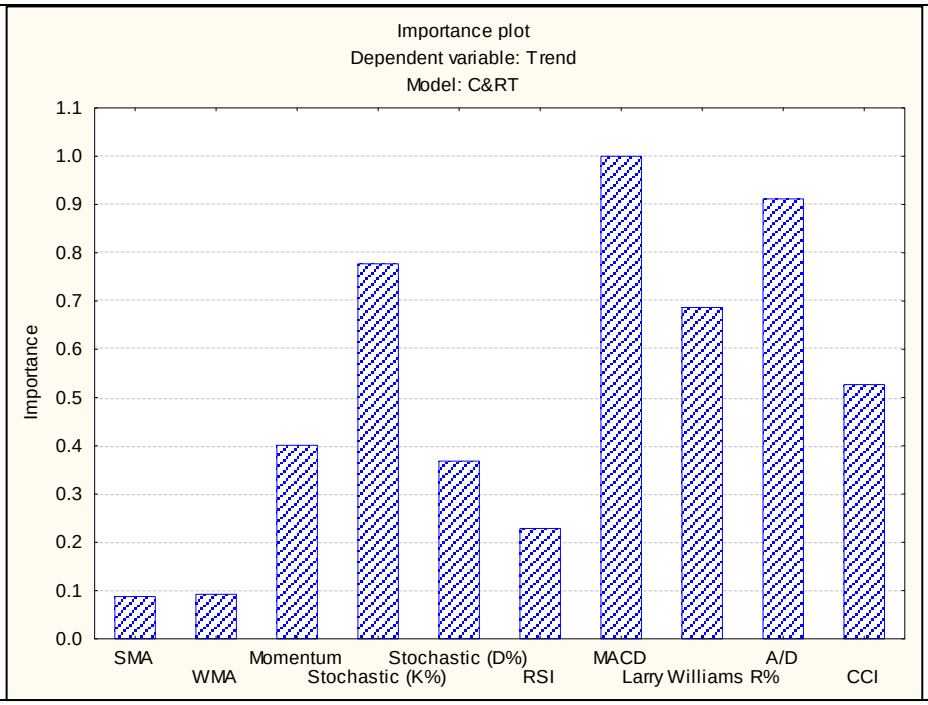


Fig. 6. Variable importance plot of technical indicators in the DT model

Figures 7 and 8 represent the importance plot of fundamental and technical-fundamental indicators to in DT respectively. In the case of fundamental variables oil price is the most key variable, whereas in the case of technical-fundamental variables MACD is the most important variable. Performances of the most accurate models in fundamental and technical-fundamental indicators are about 58.44% and 80.08% respectively.

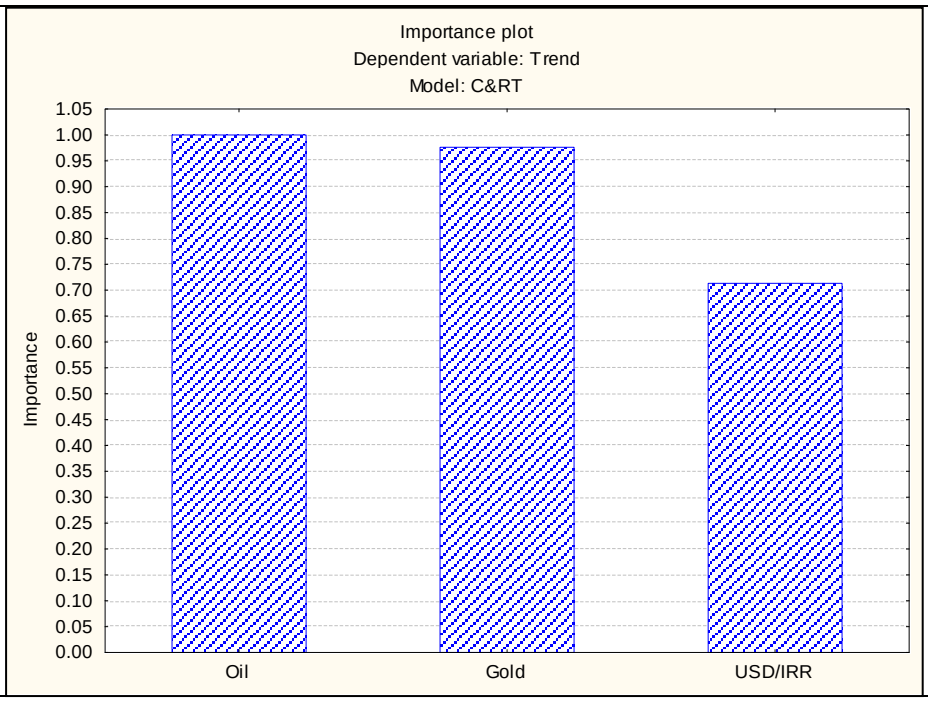


Fig. 7. Variable importance plot of fundamental indicators in the DT model

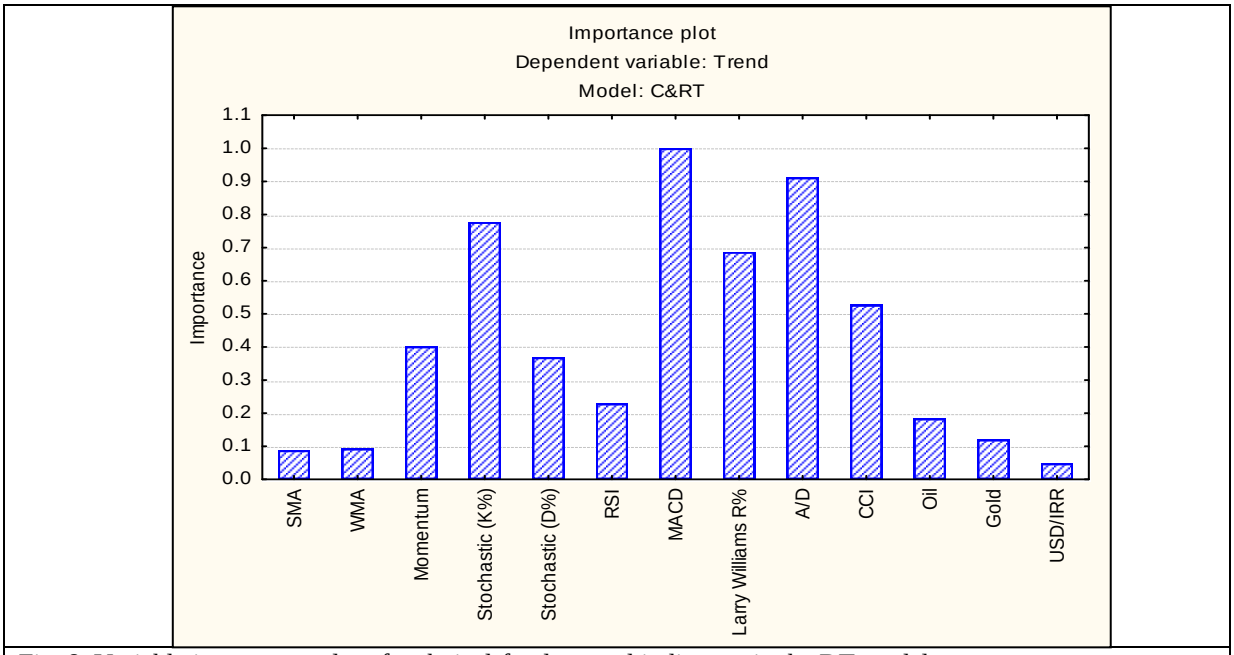


Fig. 8. Variable importance plot of technical-fundamental indicators in the DT model

5.3 Naïve Bayesian Classifier

First, ten technical indicators as input variables were given to NBC. Next, three fundamental variables used in the research were applied to the model. Finally, 13 fundamental-technical variables applied to the model. The finding results indicate that 10-variable NBC, 3-variable NBC, and 13-variable NBC are able to forecast the next day price movement with the accuracy of 73.84%, 54.43%, and 72.15% consequently.

Figure 9 shows the comparative results of using three data mining techniques used in the study. As it is displayed, when the input variables are technical or technical-fundamental, DT outperforms the other methods, but if fundamental variables are employed, RF is more accurate. Furthermore, there is dramatic difference among the performance of the models in the case of technical/technical-fundamental variables and fundamental variables. Additionally, DT with technical/technical-fundamental indicators is able to predict the TSE market movement direction more accurately than the other methods.

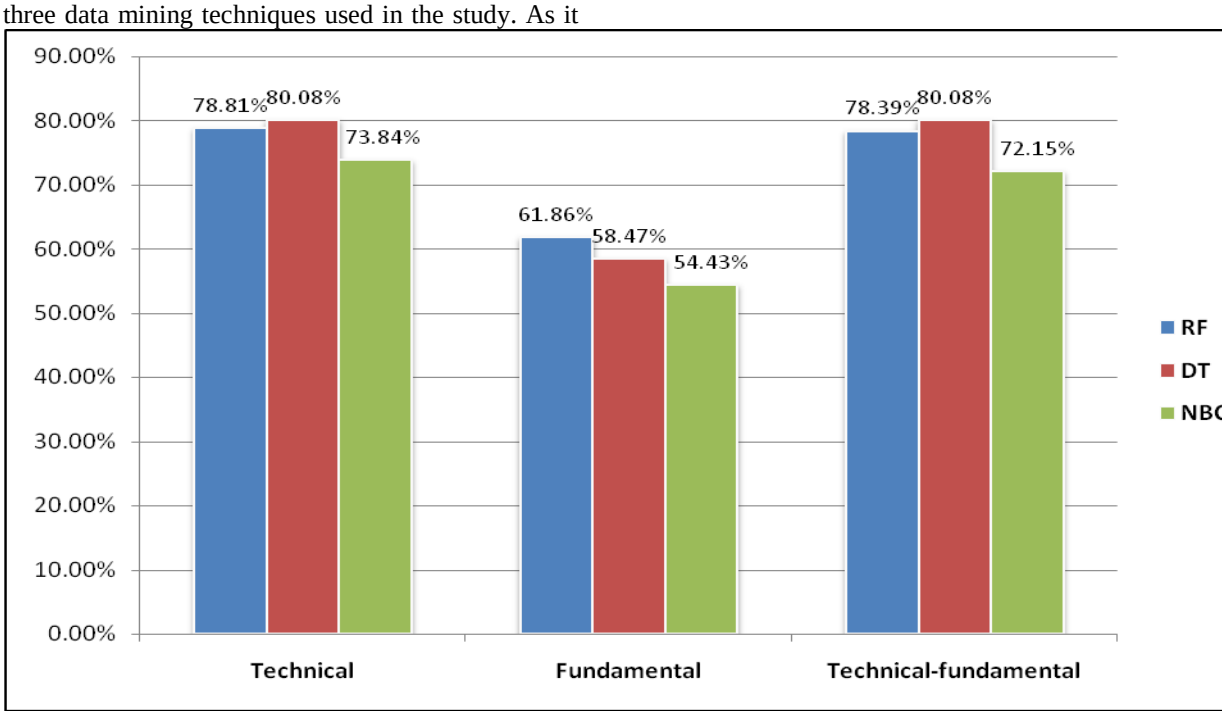


Fig. 9. Comparative results of using three data mining techniques

VI. Conclusions

Predicting stock market due to its nonlinear, dynamic and complex nature is very difficult. Predicting the direction of movements of the stock market index is important and of great interest because successful prediction may promise attractive benefits. It usually affects a financial trader's decision to buy or sell an instrument.

This study attempted to predict the direction of stock price movement in the Tehran Stock Exchange. Three data mining techniques including RF, DT, and NBC were constructed and their performances were compared on the daily data from 2007 to 2012. In order to integrity of the study, three types of input data have been used. These data include technical indicators, fundamental indicators and technical-fundamental indicators. The results demonstrate that technical and technical-fundamental variables are capable to predict the direction of the market movement with acceptable accuracy, which DT (with the accuracy of 80.08% in both technical and technical-fundamental variables) outperforms the other techniques. When fundamental indicators are employed, RF with the accuracy of 61.86% has a better performance than other methods. In general, the predictive power of machine learning methods in this case is not acceptable.

According to the results, predicting the direction of the Tehran Stock Exchange using data mining techniques is possible. Since technical variables led to much more accurate results, we can conclude fundamental analysis plays less important role than technical analysis in the process of decision-making of traders and stakeholders.

References

[1] Abbasi, E., & Abouec, A. (2008) "*Stock price forecast by using neuro-fuzzy inference system. Proceedings of World Academy of Science*", Engineering and Technology, 36: 320–323.

[2] Afolabi, M., & Olatoyosi, O. (2007) "*Predicting stock prices using a hybrid Kohonen self-organizing map (SOM)*", In 40th Annual Hawaii international conference on system sciences, 1–8.

[3] Antipov, E.A. & Pokryshevskaya, E.B. (2012) "*Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics*", Expert Systems with Applications, 39(2): 1772-1778.

[4] Anyanwu, M.N. & Shiva, S.G. (2009) "*Comparative Analysis of Serial Decision Tree Classification Algorithms*", International Journal of Computer Science and Security (IJCSS), 3: 230-240.

[5] Boyacioglu, M.A. & Avci, D. (2010) "*An Adaptive Network-Based Fuzzy Inference System (ANFIS) for the prediction of stock market return: The case of the Istanbul Stock Exchange*", Expert Systems with Applications, 37: 7908–7912.

[6] Bozkir, A.S. & Sezer, E.A. (2011) "*Predicting food demand in food courts by decision tree approaches*", Procedia Computer Science, 3: 759–763.

[7] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984) "*Classification and Regression Trees", Chapman & Hall, New York.*

[8] [8] Breiman, L. (2001) "*Random Forests*", Machine Learning, 45(1): 5-32.

[9] Chen, A.-S., Daouk, H. & Leung, M.T. (2003) "*Application of Neural Networks to an Emerging Financial Market: Forecasting and Trading the Taiwan Stock Index*", Journal Computers and Operations Research, 30(6): 901 – 923.

[10] Chen, R.C. & Han, S. (2007) "*Using SVM with Financial Statement Analysis for Prediction of Stocks*", Communications of the IIMA, 7(4): 63-72.

[11] Fernandez-Rodriguez, F., Sosvilla-Rivero, S. & Garcia-Artiles, M.D. (1999) "*Dancing with Bulls and Bears: Nearest Neighbor Forecast for the Nikkei Index*", Japan and the World Economy, 11: 395–413.

[12] [11] Genuer, R., Poggi, J.M. & Malot, T.C. (2010) "*Variable selection using random forests*", Pattern Recognition Letters, 31(14): 2225-2236.

[13] Hari, V. (2009) "*Empirical Investigation of CART and Decision Tree Extraction from Neural Networks*", Athens: Ohio University, MSc dissertation in Industrial and Systems Engineering.

[14] [13] Hill, T., Lewicki, P. (2007). STATISTICS: *Methods and Applications.* StatSoft, Tulsa, OK.

[15] Huang, W., Nakamori, Y. & Wang, S.Y. (2005) "*Forecasting stock market movement direction with support vector machine*", Computers & Operations Research, 32(10): 2513-2522.

[16] Hunt, E.B., Marin, J. & Stone, P.J. (1966) "*Experiments in induction*", Academic Press, New York.

[17] Jandaghi et al. (2010) *used ARIMA and Fuzzy-neural networks to predict stock price of SAIPA auto-making company*. The finding results show the preference of nonlinear neural-Fuzzy model to classic linear model and verify the capabilities of Fuzzy-neural networks in this prediction.

[18] Jie, L., Bo, S. (2011) "*Naive Bayesian classifier based on genetic simulated annealing algorithm*", Procedia Engineering, 23: 504 – 509.

[19] Kara, Y., Boyacioglu, M.A. & Baykan, Ö.K. (2011) "*Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange*", Expert Systems with Applications, 38, 5311–5319.

[20] Kumar, M. & Thenmozhi, M. (2005) "*Forecasting stock index movement: A comparison of Support Vector Machines and Random Forest*", In Proceedings of Ninth Indian institute of capital markets conference, Mumbai, India, [Online] Available: http://ssrn.com/abstract=876544.

[21] Leung, M.T., Daouk, H., & Chen, A.-S. (2000) "*Forecasting Stock Indices: A Comparison of Classification and Level Estimation Models*", International Journal of Forecasting, 16(2): 173–190.

[22] [21] Mehta, M., Agrawal, R. & Rissanen, J. (1996) "*SLIQ: A fast scalable classifier for data mining*", Proceedings of the Fifth International Conference on Extending Database Technology (EDBT).

[23] O'Connor, M., Remus, W., & Griggs, K. (1997) "*Going up-going down: How good are people at forecasting trends and changes in trends?*", Journal of Forecasting, 16: 165-176.