**Investigating the Effectiveness of Machine Learning Algorithm on the Forecasting of Tehran Stock Exchange Index**

**Setila Rostami1\*, Marziyeh Bayat2, Darush Javid3, Mansur Esmaeilpuor 4**

*1 . Master Of Accounting, Hamedan Branch, Islamic Azad University, Hamedan, Iran.*

*2 . Accounting Dept, Hamedan Branch, Islamic Azad University, Hamedan, Iran.*

*3 . Accounting Dept, Hamedan Branch, Islamic Azad University, Hamedan, Iran.*

*4. Computer Engineering Dept, Hamedan Branch, Islamic Azad University, Hamedan, Iran.*

*Corresponding Author email:* [*setilarostami@yahoo.com*](mailto:setilarostami@yahoo.com)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Received** | 20 January, 2015 | **Accepted** | 18 March, 2015 | **Published** | 11 February, 2015 |

K E Y W O R D S***:*** Forecast, index, machine learning algorithm, Stock Exchange

**ABSTRACT:** Investing in stocks of the Stock Exchange is one of the lucrative options in the capital market. Stocks, on the one hand lead to the widespread participation of people in ownership, and on the other hand will achieve government’s anti-inflationary goals by attracting liquidity and guiding them in productive and beneficial economic activities. Since the forecast is mainly used to reduce risk and increase profits, the accuracy in prediction is a very important issue. Therefore, the aim of the present study is to evaluate the efficiency of machine learning algorithms in the forecasting of Tehran’s Stock Exchange index. This study, in terms of its purposes, is an applied research study, with a correlational design and uses library research to gather data from Tehran’s Stock Exchange organization through the Rahavard Novin software. The population and statistical samples of the study, is the Tehran Stock Exchange organization during 1385 to 1393. We forecasted Tehran Stock Exchange index using three models of decision- maker tree, Rough Set and logistic regression that are the subset are machine learning algorithm, and using Rosetta and Weka software. Subsequently, we compared the values predicted by these models with the actual values using the paired sample t-test in SPSS software; and finally we compared the superior performance of models using the ANOVA test. Since the hypothesis is confirmed in the study, the machine learning algorithm can be used as a reliable method to predict the Tehran Stock Exchange index.

**Introduction**

As we know, capital and labor force are the main pillars of production. The supply of these factors and their optimal diagnosis is essential for economic growth. This allocation requires the presence of markets and the optimal performance of market forces. With regards to the capital, the stock market can do this important task. The most important task of the stock market is to attract outspread capitals and direct them towards investment activities through an optimal allocation process. Investors, with the motivation to receive income, enter the field of investment from two channels, from the profits of the company whose shares they have purchased, and also from selling these shares again. The fluctuation of shares in all stock markets is a natural and normal issue; however, with a prediction of the price and index of stocks, a desirable combination of them can be chosen, and fluctuations can be reduced, and in this way the information individuals have can be increased. It seems that the increase in the information in the market will lead to its better performance. The prediction of what might happen in the future and planning on that basis are very important. It is clear that the characteristic of uncertainty is an undesirable issue; however, this characteristic in unavoidable for investors who have selected the stock market as a place to invest. Therefore, normally all efforts from the investor is to reduce uncertainty, and making predictions in the Stock market is a tool to reduce uncertainty. Forecasting of important Stock market indices can be helpful in increasing the information and making it transparent. Forecasts of the stock market or the capital market indices have always been the center of investigations. This great attention in recent years caused the development of models used in forecasting.

So far, in previous research studies, in order to predict the indices different models have been used such as the Autoregressive integrated moving average (ARIMA) model, autoregressive conditionally heteroscedastic model (ARCH), and the Artificial Neural Network (ANN). The Artificial Neural Network model (data mining model) compared to regression methods such as ARIMA and ARCH has shown better performance. Therefore, the present study intends to gain a more accurate conclusion by machine learning algorithm using algorithmic models of Rough Set, Decision Tree (decision tree) and logistic regression to estimate the future stock index. This article starts with a summarization of the theoretical foundations of stocks index, forecasts, and the models of decision tree, Rough Set and logistic regression. Then background literature will be reviewed, subsequently, the main discussion of the paper, i.e. the design of the model and their comparison with each other will be presented. The final section presents the findings and recommendations of the research.

***Theoretical Foundations***

Forecast: In a general definition, the prediction of the future conditions and events is called forecast and how this is done, is called forecasting. (Afsar, 1384)

Index: Index, in general, means figure, representative or indicator. In terms of applications, the word Index (INDEX) is a quantity that represents several homogeneous variables. Index is a tool for measuring and comparing the phenomena that have certain nature and properties on the basis of which the changes in certain variables can be investigated during one period. (Pars Khebre Brokerage Company).

Machine learning algorithm: It means the design and development of algorithms on the basis of which computers or other machines gain learning ability. Its purpose is to achieve machines thatare able to extract knowledge (learning) from the environment.

Stock Exchange: The formal and structured capital market where buying and selling of shares of companies and governmental or private institutions’ Stock Exchange are done under the terms, rules and specific regulations (Lunni, 1386).

***Review of Literature***

***Local Studies***

Monajemi et al., (1388), in their study entitled "The prediction of stock market prices in Stock Exchange using the neuro -Fuzzy network using genetic algorithms, and its comparison with artificial neural network" showed that in terms of performance evaluation criteria, the prediction of the stock price of the next day through the hybrid model of neuro- fuzzy network and genetic algorithm is more accurate than neural network. In other words, prediction of the stock price using neuro- fuzzy network and genetic algorithm reduces the stock price estimation error in relation to the neural network technique.

Moshiri and Morovat (2006) forecasted the total index of stock output by linear and nonlinear models. Using the daily and weekly data of indices in the period 1377 to 1382, and different forecast methods such GARCH, ARIMA, and neural network models, they predicted the total index. The result suggested that the neural network model had fewer errors than the other two models. However, the statistical test of significance showed that the difference is not significant. In other words, the accuracy of prediction models is not statistically different.

Adel Azar et al. (1385) in a study predicted the stock index by three approaches of classical methods, artificial intelligence approach and hybrid approach. The findings of this research suggest that the neuro-fuzzy networks are superior over ARIMA method, and have the unique features of quick convergence and high accuracy and are appropriate for the prediction of stock price.

Sinai, Mortazavi, and Teimoori Asldar (1384) predicted the stock price index at Tehran Stock Exchange by artificial neural network, and presented some evidence on the chaotic behavior of stock prices on the Stock Exchange. They selected two sets of data as the input for the neural network, and selected several interruptions of Index and macroeconomic factors as the independent variables. In this study, the linear ARIMA model was used to predict the price index in the next weeks. Results from the study show that the neural network outperformed the linear ARIMA model to predict the price index.

Abbaspoor (1381) conducted a study to predict the stock price "Iran Khodro" company in the Tehran stock market using artificial neural network and used the daily data between 1379 to 1380. Based on the findings of the research, the variables affecting the stock price of "Iran Khodro" company include currency exchange rate, oil prices, the P / E ratio (price to earnings) and the volume of the stock exchanges. The results of the study show the superiority of the results of the price forecast by the artificial neural network compared to the Box - Jenkins.

***Foreign Studies***

In another study, Yakup Kara et al. (2011) attempted to predict the direction of stock price index in Istanbul through neural network models and the Support Vector Machine (SVM), and used the daily data from 1997 to 2007, along with 10 technical indices as input variables of the model. Neuro- Fuzzy Network managed to forecast 75.74%; and the Support Vector Machine (SVM) model 71.52%, and the better performance of Neuro- Fuzzy Network in comparison with the Support Vector Machine model was confirmed. Further, the best predictive performance is related to 2001.

Ming-Chi Lee (2009), predicted the NASDAQ index with a hybrid model of Support Vector Regression (SVR) and compared it with the neural network. In this research, the Support Vector Regression (SVR) was combined with the function of F-score and Supported Sequential Forward Search (FSSFS) and was used by 29 technical indices as a set of complete features to change the index . The data of research was gathered from 2001 to 2007, 80% of which was used for the teaching of the model and 20% for the testing. The results showed the superiority of the hybrid model of Support Vector Regression (SVR) compared to the neural network.

Kelly Logan (2007), forecasted the amount of money in the economy of America by the Least-angle Regression (LARS) and Bayesian methods. She used the variables such as long-term interest rates, short-term interest rates, unemployment rates, the deposit amount, and costs for monetary services between 1960 and 2009 on a monthly.

M.Tsang et al. (2007) investigated the effectiveness of neural network model (NN) on the prediction of Stocks prices in Hong Kong. This system was applied on the events of two Banking Joint Stock Companies in Hong Kong and Shanghai. The system indicated an overall success rate of over 70 percent. This study suggested the superiority of the forecast based on the Least-angle Regression (LARS) model.

zhi yank zhank (2006) attempted to predict the trend of the stock price in Shanghai using Support Vector Machine (SVM). He extracted the daily index price in Shanghai stock market from 2003 to 2005. Further, the recommendations of the nearly 400 capital market analysts and their prediction were used as input variables. The results of his study indicated that the Support Vector Machine (SVM) has a high predictability; and a combination of the Support Vector Machine (SVM) with smart models has even better results compared to the Support Vector Machine (SVM).

**Method data analysis**

***Machine learning***

Machine Learning is one of the most important branches of artificial intelligence research which is currently going through a period of growth and evolution. It means the design and development of algorithms based on which computers or other machines gain the ability of learning. Its purpose is to achieve machines that are able to extract knowledge (learning) from the environment. Learning machines have been used to accelerate and automate this process. Research in the field of machine learning is focused on the production of systems capable of extracting the concepts and their relationships in an environment based on observing some examples of them (at least one sample per meaning), and use this knowledge to identify other phenomena in the future. These machines, in general, use the induction technique for their learning. Obtaining knowledge is one of the most important applications of the learning machine in the sense that the act of learning extracts basic information from the environment and uses it for the analysis of future events. Also, another application of the learning machine is to extract large amounts of data. Learning machines are used in intelligent systems to increase knowledge and change it, increase efficiency and automatic error correction.

***Decision Tree***

Decision trees are one of the most powerful, well-known, and common tools used for classification and prediction, which is a subset of machine learning algorithms. Decision tree is a data structure that can be used to split a large collection of records to smaller sets of records. Decision-making trees use a series of questions and very simple decision rules to do this. With each successful division, the elements that are in each set are more similar to each other. As an overview of a decision tree, it can be considered a hierarchical structure in which the intermediate nodes are used to test a feature. The branches are indicators of test output; the leaves indicate class tag or the distribution of class tags. The number of sub-trees of a node determines its grade. Leaf nodes are rated as zero. In the decision trees each node of the tree does the act of categorization based on the values ​​of one the features, and the final decisions are made in the leaves.

***Rough Theory***

Finding an equivalent term in Persian for the term ROUGH SETS is difficult. In the dictionary, ROUGH equivalents are coarse, rude, approximate, turbulent and uneven, among which the word ‘approximate’ is more like the concept of the founder of the theory. But none of these words have the exact meaning of the Latin word; hence, in this study the term "Rough Set" is used. Rough Set Theory is founded in early 1980 by professor Pawlak. This theory deals with the analysis of data tables. In this theory the data tables can be obtained by measurement, or expert and specialists. The main aim of Rough sets is to obtain approximate concepts of the acquired data. This theory is a powerful mathematical tool for reasoning in cases of ambiguity and uncertainty which can provide a method for eliminating and reducing irrelevant knowledge that is more than the needs of databases. This process is done by eliminating redundant data on the basis of education (main task of the system) without loss of essential data of the database. As a result of data reduction, a set of abridged and meaningful rules would result that makes the decision-making process much easier. In fact, we can say that the Rough Sets model, by reducing the data space and selecting important terms, perform a shift from a space of the raw data and terminologies to a semantics space (meaning). Thus, due to the explosive growth of data volumes, the Rough sets can be very effective in decision support systems. Rough set theory has many similarities with fuzzy set theory, intuition theory, Boolean logic methods and discriminant analysis; however, the rough set theory is considered an independent theory. Rough set theory is a smart mathematical tool that deals with the collections and the relationships between them. Rough theory is built on the basis of the information is of concern to any member of the international community. This method attempts to suggest a way to convert data into knowledge and it is a useful method to discover hidden patterns of data. The main advantage of Rough set theory is that it doesn’t need the additional information of the data such as probability in statistics and membership grid in the fuzzy theory.

**Methods and Population**

The present study, is applied in terms of purpose, and the design of the research study is descriptive and a correlational survey which attempts to extend the quantitative data obtained from the sample to the population. The population and its statistical samples include Tehran Stock Exchange indices in a period of 9 year from 1385 to 1393.

In general, data collection procedure can be divided into two categories of library research and field method. To gather the required information, the website of the Central Bank and the software Rah Avard Novin has been utilized. For the implementation of models, the Rosetta and Weka software, and to run the statistical tests, SPSS software was used.

Independent variables in this study are the price of gold, oil, dollar, copper, silver, inflation and construction certificate; and the dependent variable is the predicted changes of Stock Index.

**Descriptive Statistics**

Table 1. Results Kolmogorov - Smirnov

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | Forecast Rough set | Forecast decision tree | Forecast logistic regression |
| N | | 3240 | 3240 | 3240 |
| Normal parameters | Mean | 0.0006 | 0.0071 | 0.8509 |
| Std. Deviation | 0.024 | 0.083 | 0.778 |
| Test Statistic k-s | | 1.289 | 1.197 | 1.263 |
| Significant quantities | | 0.071 | 0.114 | 0.082 |

According to the table 1, greater than 0.05 significance level for all models is the accuracy. The value logistic regression model, and Rough set of normally distributed random tree addressing. So it can be for review and comparison of mean precision, independent sample t-test and ANOVA parametric use.

**Evaluation of Models**

**The results of the decision tree**

Table 2. The classification Table of the variable “index changes " using Random Tree model

|  |  |  |  |
| --- | --- | --- | --- |
| Percent correct | predicted | | |
| Chang=Dec | Chang=No\_Chg | Chang=INC |
| 100 | 0 | 0 | 1213 | Chang= INC | observed |
| 99/08 | 0 | 1080 | 10 | Chang = No\_Chg |
| 98/71 | 924 | 12 | 0 | Chang= Dec |
| %99/32 | Total | | | | |

Chart 1 . The dot chart for the prediction accuracy of different levels of variable " index changes " based on Random Tree model.

**The Results of Rough Sets**

Table 3. The classification of the variable “index changes" using Rough Sets theory

|  |  |  |  |
| --- | --- | --- | --- |
| Data collection | accuracy | Data collection | accuracy |
| First quarter 85 | 97.70 | Third quarter 89 | 100 |
| Second quarter 85 | 100 | Fourth quarter 89 | 100 |
| Third quarter 85 | 98.88 | First quarter 90 | 100 |
| Fourth quarter 85 | 96.62 | Second quarter 90 | 98.92 |
| First quarter 86 | 98.86 | Third quarter 90 | 100 |
| Second quarter 86 | 100 | Fourth quarter 90 | 100 |
| Third quarter 86 | 100 | First quarter 91 | 98.86 |
| Fourth quarter 86 | 100 | Second quarter 91 | 98.92 |
| First quarter 87 | 100 | Third quarter 91 | 100 |
| Second quarter 87 | 98.92 | Fourth quarter 91 | 98.87 |
| Third quarter 87 | 100 | First quarter 92 | 98.86 |
| Fourth quarter 87 | 96.62 | Second quarter 92 | 100 |
| First quarter 88 | 100 | Third quarter 92 | 100 |
| Second quarter 88 | 98.92 | Fourth quarter 92 | 98.87 |
| Third quarter 88 | 98.88 | First quarter 93 | 100 |
| Fourth quarter 88 | 100 | Second quarter 93 | 100 |
| First quarter 89 | 100 | Third quarter 93 | 98.88 |
| Second quarter 89 | 100 | Fourth quarter 93 | 98.88 |

Chart 2.The dot chart for the prediction accuracy of different levels of variable "index changes" based on the Rough Sets theory

**Inferential Statistics**

**Paired sample t-test for the decision tree model**

Hypothesis 1: The decision tree model can forecast the Tehran Stock Exchange index.

Decision tree model is able to forecast Tehran Stock Exchange index: H0

Decision tree model is unable to forecast Tehran Stock Exchange index: H1

The results of the paired sample t-test for comparing the actual and predicted values ​​of the decision tree model are shown in the table 4

Table 4 .Paired sample t-test results for the decision tree model

|  |  | Levene test | | T-test for comparison of means | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | F statistic | Significant | T-statistics | Degrees of freedom | Significant bilateral | The average deviation | The standard deviation | %95 for SD | |
|  |  | lower limit | upper limit |
| Index Change | Assuming equality of variance | 0/004 | 0/095 | 0/352 | 6477 | 0/725 | 0/007 | 0/020 | -0/032 | 0/047 |
| Assuming Un equality of variance |  |  | 0/352 | 6477 | 0/725 | 0/007 | 0/020 | -0/032 | 0/047 |

According to the results of the above table 4, a significant level for Levene test (Sig. = 0.951) was found which is greater than 0.05, therefore, the equality hypothesis of variance of two populations (predicted values ​​and the actual values) is confirmed. So in order to examine the comparison between the mean score of the two populations, we should refer to the relevant results and the equality of variances. The results of the paired sample t-test, in case the equality of variance is assumed, shows that the level of significance for t-test (two-tailed sig. = 0.725) is greater than 0.05, therefore, the mean of the two populations will be accepted, so it can be said that the decision tree model has had a high accuracy in the prediction of index changes variable.

**Paired sample t-test for the Rough Sets**

Hypothesis 2: Rough Sets model is able to predict Tehran Stock Exchange's index.

Rough Sets model is able to predict Tehran Stock Exchange index: H0

Rough Sets model is unable to predict Tehran Stock Exchange index: H1

The results of the paired sample t-test for the comparison of actual and predicted values of the Rough Sets model are ​​shown in the table 5

Table 5 . Paired sample t-test for the Rough Sets

|  |  | Levene test | | T-test for comparison of means | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | F statistic | Significant | T-statistics | Degrees of freedom | Significant bilateral | The average deviation | The standard deviation | %95 for SD | |
|  |  | lower limit | upper limit |
| Index Change | Assuming equality of variance | 0.001 | 0.99 | 0/019 | 6476 | 0.992 | -0.00069 | 0/020 | -0.039 | 0.039 |
| Assuming Un equality of variance |  |  | 0/019 | 6476 | 0.992 | -0.00069 | 0/020 | -0.039 | 0.039 |

According to the results of the table 5, the significant level for Levene test (sig. = 0.99) is greater than 0.05, therefore, the hypothesis of the equality of variance of the two populations (actual and predicted values) is supported. Hence, in order to examine the comparison of the two populations’ means, attention should be paid to the results of the variance equality. The results of the paired sample t-test, in case the equality of variance is assumed, shows that the level of significance for t-test (two-tailed sig. = 0.992) is greater than the alpha level 0.05; therefore, the mean equality of the two populations will be accepted, so it can be said that the Rough Sets model has had a high accuracy in the prediction of index changes variable, and that our hypothesis is supported.

**Assortment of the Results of Statistical Analysis and Correctness or Incorrectness of the Hypotheses**

In this study, two main hypothesis were proposed, and their results after the analysis and statistical tests are as follows:

Table 6 . Summary of the Hypotheses Results

|  |  |  |
| --- | --- | --- |
| Hypothesis | Title of the hypothesis | Result |
| Hypothesis 1 | Decision tree model is able to forecast Tehran Stock Exchange index: H0 | H0 is confirmed.  H1 is rejected. |
| Decision tree model is unable to forecast Tehran Stock Exchange index: H1 |
| Hypothesis 2 | Rough Sets model is able to predict Tehran Stock Exchange index :H0 | H0 is confirmed.  H1 is rejected. |
| Rough Sets model is unable to predict Tehran Stock Exchange index : H1 |

**Conclusions and Suggestions for Further Research**

In this paper, the efficiency of machine learning algorithms to predict Tehran Stock Exchange index was investigated. According to research findings, some suggestions can be presented as follows:

**Suggestions arising from the research**

The forecast of Index using the Nearest Neighbor method and its comparison with models of decision tree, Rough Sets model, and logistic regression.

Taking longer periods of time into consideration and involving the independent variables such as subsidies and currency of various countries.

**Practical suggestions for future research**

The application of machine learning algorithm to predict the commercial issues in optimizing the results

Investigation and comparison of machine learning algorithm with other mathematical methods for industry profitability

The investigation of the results of effective factors in forecasting index under the influence of unavoidable circumstances such as international sanctions

**References**

Adel A, Afsar A. 2006. "The comparison of classical and artificial intelligence methods in predicting the stock price index and designing a hybrid model", Journal of Humanities teacher, Issue (4).

Afsr A. 2006. Forecast modeling of Stock price using the neuro-fuzzy network and hybrid methods. MA. thesis. Industrial Management. Tarbiat Modares University.

Allahyari E. 2008. Examining the weak form of capital markets effectiveness on Tehran Stock Exchange. Stock Exchange Quarterly. No. 4

Fallahpour, Saeed; Golarzi, Gholam Hossain; Fatooreh Chian, Nasser. (2013). Forecast of the trend of stock prices using the Support Vector Machine based on the genetic algorithm in Tehran Stock Exchange. *Financial Research*. Volume 15, Issue 2. pp. 269-288.