Advances in mathematical finance & applications, 3 (1), (2018), 97-105

Published by IA University of Arak, Iran
Homepage: www.amfa.iau-arak.ac.ir

Forecasting Stock Trend by Data Mining Algorithm

***Sadegh Ehteshami*[a,*], *Mohsen Hamidian*[b], *Zohreh Hajiha*[c], *Serveh Shokrollahi*[b]**

[a]*Department of Accounting, Kish International Branch, Islamic Azad University-South Tehran, Tehran, Iran*
[b]*Department of Accounting and Economic, South Tehran Branch, Islamic Azad University, Tehran, Iran*
[c]*Young Researcher and Elite club, East Tehran Branch, Islamic Azad University, Tehran, Iran*

ARTICLE INFO

Article history:
Received 13 July 2017
Accepted 2 September 2017

Keywords:
Stock trend forecasting,
Random forest algorithm,
Decision tree algorithm.

ABSTRACT

Stock trend forecasting is a one of the main factors in choosing the best investment, hence prediction and comparison of different firms' stock trend is one method for improving investment process. Stockholders need information for forecasting firm's stock trend in order to make decision about firms' stock trading. In this study stock trend, forecasting performs by data mining algorithm. It should mention that this research has two hypotheses. It aimed at being practical and it is correlation methodology. The research performed in deductive reasoning. Hypotheses analyzed based on collected data from 180 firms listed in Tehran stock exchange during 2009-2015. Results indicated that algorithms are able to forecast negative stock return. However, random forest algorithm is more powerful than decision tree algorithm. In addition, stock return from last three years and selling growth are the main variables of negative stock return forecasting.

1 Introduction

One main component of capital market is Tehran stock exchange. Tehran stock exchange is a formal market in which firms' stock, government bond, and competent private enterprises bond trade based on certain regulations [1]. National economy and even global economy have effect on exchange market. One of the main standards for exchange decision making is stock return. Stock return has its own information content and many investors use them for their analysis. Forecasting price and stock return are financial issues that many researchers attempted to investigate them. Forecasting is not simple task because many variables have effect on stock return [6]. So previous decision need to evaluate for making decision about creating value for stockholder.

Stock return is main factor in choosing the best investment so comparing and forecasting different firms' stock return is one method of improving investment process. Stockholders need information for forecasting firm's stock trend in order to make decision about firms' stock trading [5]. Developing new technologies and applying them in different knowledge have become methods of accounting and financial management. Technology changes and using it in defend sciences made accountants to use new technologies along with enhancing efficiency. One of the main ways of increasing accuracy of firms' stock return forecasting is using new methods of data mining aiming at forecasting. Many re-

* Corresponding author Tel.:+98*9124154235*
E-mail address: *mrmahbobi@gmail.com*

searches have made investigations so far on providing patterns for identifying firms' return. Accordingly, most of them inclined to develop theories of free intelligent dynamic systems based on experimental studies. Decision tree and random forest belong to these dynamic systems, which transfer rule and knowledge latent in data to algorithm structure and the network [10].

2 Literature Review

Delen et al [9] studied firms' bankruptcy forecasting in Tehran stock exchange by decision tree and random forest. Research results indicated that both models (decision tree and random forest) are able to forecast bankruptcy with different accuracy. Therefore, the area under ROC curve in random forest model was more than decision tree model and it has better performance. Delen et al. [9] evaluated performance of business unit by financial ratio and decision tree by applying four algorithms and studied financial ratios forcasting power for performance assessment indices (stock exchange revenue return and assets' return). Results of stock exchange revenue return include gross profit margin, leverage ratio and growth ratio to selling. Considering assets return, main ratio includes earning before tax to stockholder's revenue, gross profit margin, and debt ratio to working asset.

Barzegari Khanghah et al. [3] forecasted stock return by financial ratios. Results illustrated that profitability ratios have more important role in stock return forecasting and also assets return ratio and stock holders revenue return have more power in explaining changes of stock return. He studied anticipation of future stock market return by ARIMA model, neural network and wavelet noise reduction. Results showed that reducing the noise improves performance of index return anticipation. In other word, wavelet neural network model (sign al noise reduction) has better performance than ARIMA models and neural network. In addition, neural network models are better predictors than ARIMA models. In addition, he compared stock return ratio forecasting in GARCH models by Bayesian and ML methods. Their results indicated that there is no significant different between mentioned methods although Baysian method is better performance in forecasting heavy tail functions.

Johnny et al. [4] studied relationship between profit and its components with stock return direction. Results of the study illustrates that firms with high profit quality, obtain positive return while firms with low profit quality attain negative return. Yu and wenjuan [11] used decision tree aiming at determining the most effective financial ratios to profit growth. They used c5 model as a technics of decision tree. Results of the study indicated that the model has high accuracy (95%) in forecasting firm's growth. Bheenick and Brooks (2015) investigated whether stock trade volume can help to forecast direction of return in Australian stock market. Results of the study illustrated that trade volume have power of forecasting for firms with high trade volume and for special industries in Australian market. However, for small firms, volume of the trade was not able to forecast stock return as well.

3 Research Methodology

This study aimed at being practical and it belongs to semi-experimental researches due to using historic information for testing hypotheses. Deductive reasoning and ex post facto were used as method of the study and decision tree and random forest algorithms were applied for forecasting stock negative return, so that required information and data were collected by library method and using Rah-Avard Novin software and also studying fundamental financial statements of companies listed in Tehran stock exchange during 2009-2015. In addition, financial statements data of exchange informational website collected. Statistical population included 180 companies listed as active firms in Tehran stock

exchange from 2009 to 2015 based on which we observed 1260 companies. In this study, we used screening method in order to select the best sample of the study as a good representative of the statistical population. Accordingly, 5 following standards were determined as standard for selecting firms as a sample of the study out of all other companies which eliminated from the study:

1- The firm is active in stock exchange before 2009 to the end of 2015.
2- Listed firm was not supposed to be holding, insurance, leasing, bank, and financial and investment enterprise due to its activities which are completely different from commercial and manufacturing companies.
3- Financial year ends at 19 March and it has no financial year change during years of the study.
4- The company has no trading pause longer than three months.
5- Required data need to be accessible in variables definition part.

In order to final analysis, firstly, central indices including mean, median and dispersion indices were used which included skewness and kurtosis as descriptive statistical indices. Then, each data were entered in R software and divided into three categories of teaching, validation and test. Next, decision tree and random forest used for forecasting stock negative return. Finally when forecasting model explored, accuracy of the model on test data was obtained and results were compared and importance degree of variables were identified by both methods.

4 Research Results

4.1 Descriptive statistic and durability of research variables

Table 1: Descriptive statistic of research variables

variables	numbers	mean	median	Standard deviation	skewness	kurtosis	minimum	maximum	Statistical value of Levin, Lin and Chu	probability value
I	1261	0.34	0	0.47	0.68	1.46	0	1	-17.99	0.000
FL	1261	0.61	0.62	0.21	-0.23	2.61	0.11	1.09	-8.42	0.000
OL	1261	0.99	1	0.08	-12.8	166.7	0	1	-2.36	0.009
WCTA	1261	0.1	0.11	0.21	-0.12	2.85	-0.48	0.6	-12.05	0.000
ICR	1261	1.33	1.22	2.22	0.34	5.47	-7.38	12	-57.62	0.000
CR	1261	0.18	0.17	0.54	0.05	5.2	-2.49	2.39	-12.23	0.000
CFL	1261	0.01	0.01	0.13	0.09	3.07	-0.32	0.36	-30.07	0.000
ROA	1261	0.08	0.08	0.11	-0.28	3.56	-0.29	0.34	-13.16	0.000

Table 1: Continue

0.000	-10.63	1.71	-11.3	5.67	-1.09	1.59	-1.58	-1.79	1261	**ETL**
0.000	-8.12	1	0	4.4	1.84	0.37	0	0.16	1261	**NEG**
0.000	-41.02	46.2	-27.5	8.19	1.03	8.74	0	0.16	1261	**RET**
0.000	-28.23	1027.56	-77.8	8.43	2.17	149.56	3	82.38	1261	**RET_3**
0.000	-295.54	8.46	5.27	2.66	0.46	0.66	6.55	6.64	1261	**CCC**
0.000	-100.96	2.59	-3.11	5.35	-0.04	0.75	-0.12	-0.13	1261	**CROA**
0.000	-15.96	1.08	-3.28	5.18	-1	0.64	-0.34	-0.42	1261	**TR**
0.000	-34.76	1.75	-0.73	5.98	1.11	0.31	0	0.02	1261	**CTR**
0.000	-522.94	3.88	-2.07	4.18	0.54	0.76	0.87	0.96	1261	**INV**
0.000	-47.99	0.98	-0.8	3.25	0.37	0.32	-0.02	0.01	1261	**CINV**
0.000	-27.06	30.85	0.02	3.6	1.01	6.62	7.38	10.47	1261	**SD**
0.000	-20.19	18.48	6.57	3.78	0.5	1.75	11.5	11.57	1261	**SD_Sale**
0.000	-6.18	16.18	3.48	3.37	0.41	1.81	10.23	10.39	1261	**SD_NI**
0.000	-124.33	16.4	5.09	3.61	0.48	1.7	10.61	10.77	1261	**SD_CF**
0.000	-8.38	18.48	6.57	3.78	0.5	1.75	11.5	11.57	1261	**ACC/TA**
0.000	-46.45	5.94	-7.49	9.17	-0.61	1.06	-0.92	-1.05	1261	**BtoM**
0.000	-61.26	0.04	-0.04	10.79	0.26	0.01	0	0	1261	**EtoP**
0.000	-102.56	0.53	-0.25	4.01	0.11	0.12	0.12	0.12	1261	**EQUITY**
0.000	-25.49	0.19	-0.24	5.75	-0.75	0.06	-0.02	-0.03	1261	**SG**
0.000	-97.19	0.92	-0.72	3.12	0.12	0.29	0.12	0.13	1261	**CEtoP**
0.000	-13.90	1	0	1.77	0.88	0.46	0	0.3	1261	**SD**

Since dependent variable of the study is a qualitative variable, obtaining central and dispersion indices does not provide good information from variable nature. For these variables we need to draw many tables in order to achieve comprehensible observation from data dispersion. Based on Table 1 and considering skewness of variable lower than -2 and higher than 2, dispersion of variable is not completely normal. For example, growth and stock return of the company have no normal dispersion. In addition, mean value of interest coverage ratio is very higher than median value and this proved that skewness is toward right direction. However, mean and median of assets return are close together and by consideration of its skewness, which is close to zero, it can be concluded that this variable is sym-

metric. In addition, high standard deviation of interest coverage ratio proves that data are highly dispersed which indicates that there are outlier observations in data.

Table 2: Numbers and percentage of stock return variable

percentage	numbers	Stock return
33.9	427	Negative return
66.1	833	Positive return
100	1260	Total

As it is indicated in Table 2, number of stock negative return is 40 percent and 60 percent of stock return is nonnegative. Therefore, numbers of negativity of stock return is less than positivity of stock during 2009 to 2015. Results of durability test for each variable are indicated in Table 1. Based on Levin, Lin and Chu test, as probability value of all variables is less than 5%, all variables are stationary. Stationary means that mean and variance of research variables were fixed during mentioned time and variables covariance was fixed in different years.

4.2 Data Grouping

In order to use data in decision tree and random forest algorithms, whole sample were divided into two groups of teaching and experiment's groups. The more is teaching samples; the better is performed process of teaching. So 882 cases of whole sample (70 percent) were selected as teaching samples and 378 numbers (30 percent) were selected as experiment group. It must be noted that this division was performed systematically. So 30 percent data were considered as experiment data. Remained ones were used for teaching decision tree algorithm.

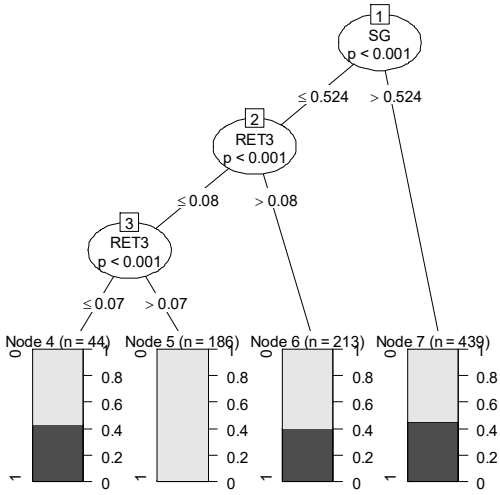
4.3 Decision Tree Forecasting

Table 3: Accuracy of decision tree forecasting for all data

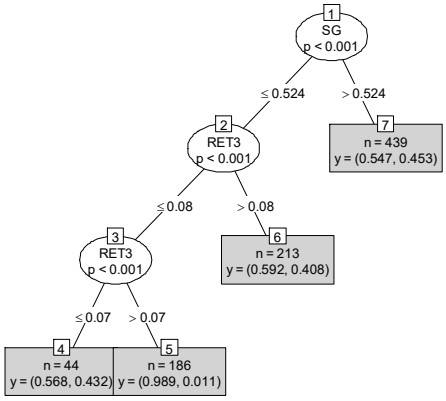
Observations		Forecasting		
		Negative return	Positive return	Accuracy
Teaching	Negative return	165	18	92.16%
	Positive return	83	616	88.12%
	All observations	66.53%	97.16%	88.55%
Test	Negative return	206	38	84.43%
	Positive return	55	79	58.95%
	All observations	69.05%	67.52%	75.40%
Final forecasting	Negative return	371	56	86.88%
	Positive return	138	695	83.43%
	All observations	72.88%	92.54%	84.55%

As it is shown in Table 3, decision tree algorithm was able to accurately forecast 206 negative returns out of 244 negative return of total experiment data. In addition, it could accurately forecast 79 positive returns from 134 positive returns of total experimental data. However, 45 positive returns were mistakenly identified as negative return and 38 negative returns were wrongly identified as positive re-

turn. So considering both teaching and experiment groups, decision tree algorithm could accurately predict 84.43 percent of negative return and 58.95 percent of positive return. Decision tree algorithm 'power of forecasting is, therefore, 75.40%.



(a) not pruned decision tree



(b) pruned decision tree

Fig. 1: results of decision tree in forecasting stock negative return

In Fig. 1, decision tree of stock negative return was obtained. In part a, decision tree was drawn with all leaves and branches while in part b decision tree was drawn with pruned leaves. As it is clear from the figure decision tree identified selling growth and stock return of last three years as the main decision making variables. Decision tree choose variable with more entropy of data as the root and other variables are placed in other nodes based on their entropy level. Evidently, in this study, selling growth has the highest entropy.

4.4 Random Forest Forecasting

Based on obtained results, random forest forecasting is carried out with optimal trees and important variables. There are 100 trees and variables with less importance degree are eliminated from model formula so that which obtain best forecasting for stock return. Random forest algorithm chooses the best model for stock return forecasting by test and error and by adding and omitting variables in the formula. Results of forecasting for experiment data are shown in Table 4. As it is indicated in Table 4, random forest algorithm was able to accurately forecast 226 negative return out of 244 negative returns of total experiment data. In addition, it could accurately forecast 98 positive returns from 134 positive returns of total experimental data.

However, 36 positive returns were mistakenly identified as negative return and 18 negative returns were wrongly identified as positive return. So considering both teaching and experiment groups, random forest algorithm could accurately predict 92.92 percent of negative return and 73.13 percent of positive return. Thus, random forest algorithm 'power of forecasting is 85.71%.

Table 4: Accuracy of random forest forecasting for all data

Observations		Forecasting		
		Negative return	**Positive return**	**Accuracy**
Teaching	Negative return	169	14	92.35%
	Positive return	76	623	89.13%
	Total observations	98.68%	97.80%	89.80%
Experiment	Negative return	226	18	92.62%
	Positive return	36	98	73.13%
	Total observations	61.9%	88.1%	85.71%
Final forecasting	Negative return	395	32	92.51%
	Positive return	112	721	86.55%
	Total observations	77.91%	95.75%	88.50%

5 Results and Discussion

Decision algorithm could accurately predict 84.43 percent of negative return and 58.95 percent of positive return. Since decision tree algorithm 'power of forecasting is 75.40%, it is concluded that decision tree has high power of stock return forecasting. This algorithm had also high error in identifying positive and negative return. By entropy calculation, decision tree algorithm recognized selling growth as the main variable and replaced this variable at the root of the tree. Therefore, it made decisions mostly based on selling growth and stock return of last three years. Random forest algorithm could accurately predict 92.92 percent of negative return and 73.13 percent of positive return. Considering that random forest algorithm 'power of forecasting is 85.71%, it is concluded that this algorithm has high power of stock return forecasting. In identifying positive return, random forest algorithm had more error due to closeness of some observations of variables. Since negative return and positive return's grouping was based on the sign, this error can be ignored. Because returns close to zero are forecasted as positive returns which is probably due to closeness of their group centre to negative returns' centre.

Random forest algorithm is more powerful in forecasting stock negative return rather than decision tree algorithm, because percentage of correct forecasting made by random forest algorithm was approximately 10 percent higher than decision tree algorithm. Since applied random forest created 100 decision tree and it selects the best tree as forecasting results by OOB method, so it has higher power in forecasting. Generally, random forest is generalized algorithm of decision tree and it is considered as a powerful algorithm among double groups variables. Therefore, it is more powerful than decision tree. Variable importance is different and importance of last three years' stock returns indices are higher than performance standards. Other variables are important equally whole SD, NEG and Ol had

leas importance degree. Since importance degree of different input variables groups (leverage criteria, performance, working, instability, and quality) is different in forecasting stock negative return, variables including selling growth and stock return of last three years have considerable power of forecasting rather than other variables, so that if these variables are used in forecasting, decision making power does not become very different so eliminating these variables from model formula reduces decision making power considerably. On the other hand, operative leverage variables, company's loss and selling drop have less forecasting power so elimination of these variables from the model formula is increased.

6 Conclusion

An important issue that researchers and scholars in decision-making and forecasting fields have challenge with is choosing effective variables on decision output and forecasting. So if stock return is can be predicted by good variables and some models can be providing, in fact, more insured condition is provided in capital market which help investment development in financial markets. Regarding percentage of correct forecasting of applied algorithms in stock return forecasting, it can be concluded that decision tree and random forest algorithms have high power of stock negative return forecasting. However, random forest also has more power than decision tree.

References

[1] Pour Heydar A., Hemati D., *Study effect of debt contract, political costs, award plans, and property of earn ing management in companies listed in Tehran stock exchange*, Accounting and Auditing Studies, 2004, **36**, P.23-45.

[2] Hosseini, S. M., Rashidi Z., *Forecasting bankruptcy of companies listed in Tehran stock Exchange by deci sion tree and logestic regression*, Financial Accounting Researches, 2017, **5**(3), P.105-130

[3] Barzegari Khanghah J., Jamali Z., *Stock return forecasting by financial ratios: investigating recent researches,* Accounting Quarterly Journal, 2017, **6**(1), P.71-92

[4] Johnny R.J., Khodadadi V., *Study relationship between earning and its component and stock return by focus of earning quality in companies listed in Tehran stock exchange*, Financial Accounting Journal, 2015, **3**(9), P.84-113

[5] Mehrani S.M., Pishvayi F., Khalatbari H., *Assessing earning management in different levels of co seratism and institutional investors by Bedford law*, Journal of Accounting and Auditing, 2016, 5, P.234-257.

[6] Abder Rovf M.D., *The Corporate Social responsibility Disclosure*, Business and Economics Research Journal, 2007, **2**(3), P.19-32.

[7] Ardia D., Hoogerheide L.F*., Bayesian Estimation of the GARCH (1,1) Model with Student-t Innovations in RMPRA* working paper, 2016, URL http://mpra.ub. uni-muenchen.de/17414/.

[8] Bheenick, E. B., Brooks, R. D., *Does Volume Help in Predicting Stock Return? An Analysis of the Australian Market*, Research in International Business and Finance, 2015, **24**, P.146-157.

[9] Delen, D., Kuzey, C., Uyar, A., *Measuring firm performance using financial ratios: A Decision tree approach*. Expert System with Application, 2017, **40**(10), P.3970-3983.

[10] Mac Lyons, M., *Dolphin deaths, organizational legitimacy and potential employees' reactions to as sured environmental disclosure*s. Accounting Forum, 2005, **34** (1), P.1–19.

[11] Yu G., Wenjuan, G., *Decision tree method in financial analysis of listed logistics company. In* 2010 International conference on intelligent computation technology and automation, 2015.