Journal of Financial Management Strategy
Alzahra University
Vol. 12, No. 46, Fall 2024

pp. 33-36

Research Paper

Comparison of Intelligence Accuracy of Data Mining Algorithms to Estimate Stocks Prices[1]

Hossein Kianizadeh[2], Ali Baghani[3], Mohsen Hamidian[4]

Received: 2023/05/09 Accepted: 2024/08/09

INTRODUCTION

The volume of capital market information is expanding dramatically, making it impossible to effectively use this data without data mining algorithms and big data models. Past studies have indicated the potential for stock price prediction using machine learning models; however, the prediction accuracy of these models has not been thoroughly evaluated. The aim of this research is to compare the predictive accuracy of five commonly used data mining algorithms: neural networks, logistic regression, k-nearest neighbors, support vector machines, and cross-validation.

Among the 385 active companies listed on the Tehran Stock Exchange, 72 companies were selected using a systematic elimination method, and the accuracy of the aforementioned models in predicting stock prices was assessed using daily data from these selected stocks for the years 2008 to 2019. To evaluate the models' accuracy, three indices-R^2, MSE, and RMSE-were used. An analysis of variance (ANOVA) employing the F statistic was conducted to assess the models' fit, while the t-test was used to compare the models in pairs.

The main objective of this research is to compare the predictive accuracy of data mining algorithms for stock price forecasting. To achieve this, widely-used machine

1. doi: 10.22051/jfm.2024.40333.2685

2. Ph.D. Student, Department of Financial Management, Kish International Branch, Islamic Azad University, Kish Island, Iran. Email:kianizadeh@gmail.com.

3. Assistant Professor, Department of Financial Management, Kish International Branch, Islamic Azad University, Kish Island, Iran. Corresponding Author. Email:ali.baghani.85@gmail.com.

4. Associate Professor, Department of Financial Management, Kish International Branch, Islamic Azad University, Kish Island, Iran. Email:hamidian_2002@yahoo.com.

learning models, including support vector machines, neural networks, k-nearest neighbors, logistic regression, and cross-validation, were first identified through a review of the literature. Then, using statistical methods, the accuracy of these models in estimating stock prices in the Tehran Stock Exchange was calculated.

The primary innovation of this research lies in not only evaluating the potential of stock price prediction using data mining algorithms-something that has been done independently in previous studies-but also in assessing the accuracy of these models, which is novel research within the context of the Iranian capital market.

MATERIALS AND METHODS

The method of data collection in this research is the library method. The data were extracted daily from the Rahavard Novin software and the tsetmc.com information database. Additionally, the gold and currency statistics database was used to obtain the free exchange rate, the price of gold per ounce, and the price of oil.

The statistical population of this research includes companies listed on the main board of the Tehran Stock Exchange during the period from 2008 to 2019. Given the large size of the population and certain inconsistencies among its members, the systematic elimination method was employed to select the sample (Sarmed, 2019). The criteria used for sample selection are as follows:

- The company must have been listed on the Tehran Stock Exchange before 2008.
- The company's financial year must end at the end of Esfand (to increase comparability).
- The company must not have changed its financial year during the period from 2008 to 2019.
- The company must not have experienced a trading suspension of more than six months during any financial year.
- More than 600,000 of its shares must be traded each year.

By applying these criteria, 72 companies active on the main board of the stock exchange were selected as the statistical sample.

RESULTS AND DISCUSSION

By classifying the aforementioned indicators, it was concluded that the support vector machine exhibited the highest accuracy in predicting stock prices, followed by neural networks, logistic regression, cross-validation, and k-nearest neighbors.

The following table presents the descriptive statistical indicators of the five models used in the research. According to these indicators, the average R^2 for the support vector machine method has the highest value, equal to 0.94, with a standard deviation of 0.01. It is followed by the artificial neural network model, logistic regression, cross-validation, and lastly, k-nearest neighbors. Other descriptive statistics also support the reliability and validity of the data.

Table 1. Descriptive statistics of the models used in the research

kurtosis	Skewness	Median	max	min	Standard Deviation	Mean	Descriptive statistics indicators	
-1/14	0/24	0/96	0/99	0/91	0/01	0/94	R^2	support vector machine
-1/16	-0/24	97/28	131/93	58/33	22/17	96/24	MSE	
-1/15	-0/26	57	85	23	18/82	55/49	RMSE	
-1/21	-0/01	0/89	0/95	0/83	0/02	0/59	R^2	logistic regression
-0/94	-0/13	122/53	150/85	87/06	17/9	121/07	MSE	
-0/93	-0/14	86/5	113	53	16/89	85/46	RMSE	
-1/19	-0/02	0/87	0/92	0/83	0/02	0/87	R^2	cross-validation
-0/98	-0/02	152/42	183/55	119/97	17/9	152/19	MSE	
-0/98	-0/05	115/5	145	84	16/99	115/13	RMSE	
-1/15	0/01	0/82	0/88	0/82	0/02	0/85	R^2	k-nearest neighbors
-1/05	-0/17	181/74	219/1	137/4	23/84	181/33	MSE	
-1/05	-0/16	143/5	179	101	22/69	142/79	RMSE	
-1/21	-0/01	0/91	0/96	0/88	0/02	0/91	R^2	neural network
-1/28	0/01	98/15	131/17	66/49	19/51	98/13	MSE	
-1/27	0/01	64	95	33	18/58	63/78	RMSE	

CONCLUSION

Based on the analysis conducted in this research, it is evident that machine learning models can be used to estimate stock prices in the capital market, and the accuracy of stock price estimation varies across different models. To validate this hypothesis, the mean difference test (variance analysis) was employed. According to the results presented in the table, the significance level of the R^2 index is 0.0440, which is less than 0.05. This leads to the acceptance of hypothesis H1, which states that "There is a significant difference in accuracy between the support vector machine, cross-validation, logistic regression, neural networks, and k-nearest neighbor's models." Consequently, the null hypothesis, which posits no significant difference, is rejected.

According to the research findings, the support vector machine model demonstrates the highest accuracy in stock price estimation, ranking first with $R^2 = 0.944$. This is followed by the artificial neural network model with $R^2 = 0.908$, logistic regression with $R^2 = 0.888$, cross-validation with $R^2 = 0.868$, and k-nearest neighbors with $R^2 = 0.848$. This ranking is further confirmed by the MSE and RMSE

methods. As shown in the table, the support vector machine method, being the most accurate, has the lowest values for these two indices. As model accuracy decreases, the values of these indices increase, following the same ranking observed in the R^2 results.

The results of this research are consistent with previous empirical findings, which demonstrated that the models used have the capability to predict stock prices, as highlighted in the literature review. However, since previous studies did not involve a comparative analysis of the models used in this research, it is not possible to directly compare or contradict their findings. The main innovation of this research lies in comparing a range of machine learning models, which was not done in prior studies.

Keywords: Stock exchange, Forecast, Stock prices, Intelligent algorithms, Machine learning, Data mining.
JEL Classification: C02, C13, C53, C55, C58, G17.