International Journal of Engineering

Journal Homepage: www.ije.ir

The Predictability of Tree-based Machine Learning Algorithms in the Big Data Context

F. Qolipour, M. Ghasemzadeh*, N. Mohammad-Karimi

Computer Engineering Department, Yazd University, Yazd, Iran

PAPER INFO

Paper history:
Received 22 August 2019
Accepted in revised form 22 October 2020

Keywords:
Stock Market
Big Data
Prediction
Machine Learning
Tree-based Algorithms
Ensemble Algorithms

ABSTRACT

This research work is concerned with the predictability of ensemble and singular tree-based machine learning algorithms during the recession and prosperity of the two companies listed in the Tehran Stock Exchange in the context of big data. In this regard, the main issue is that economic managers and the academic community require predicting models with more accuracy and reduced execution time; moreover, the prediction of the companies recession in the stock market is highly significant. Machine learning algorithms must be able to appropriately predict the stock return sign during the market downturn and boom days. Addressing the stated challenge will upgrade the quality of stock purchases and, subsequently, will increase profitability. In this article, the proposed solution relies on the utilization of tree-based machine learning algorithms in the context of big data. The proposed solution exploits the decision tree algorithm, which is a traditional and singular tree-based learning algorithm. Furthermore, two modern and ensemble tree-based learning algorithms, random forest and gradient boosted tree, has been utilized for predicting the stock return sign during recession and prosperity. The mentioned cases were implemented by applying the machine learning tools in python programming language and PYSPARK library that is used explicitly for the big data context. The utilized research data of the current study are the shares information of two companies of the Tehran Stock Exchange. The obtained results reveal that the applied ensemble learning algorithms have performed better than the singular learning algorithms. Additionally, adding 23 technical features to the initial data and subsequent applying of the PCA feature reduction method have demonstrated the best performance among other modes. In the meantime, it has been concluded that the initial data do not possess the proper resolution or generalizability, either during prosperity or recession.

***doi**: 10.5829/ije.2021.34.01a.10*

1. INTRODUCTION[1]

In recent years, machine learning methods are widely used in different areas [1,2]; also, the stock return predictability issue has been frequently investigated in financial studies, and in this regard, forecasting the stock returns for investment purposes is very imprtant [3]. Nowadays, the rapid increase in processing speed, low data storage costs, the big data availability, as well as a wide range of open-source software, have revolutionized application of machine learning techniques. However, the stated novel research field is not bound to computer science or software engineering. Currently, machine learning tools are also utilized for resolving financial issues [4].

Since large companies are more involved in economic activities and have a more transactions, they generate more massive data. As CPU speed increases, larger data packs could be analyzed; consequently, the data analyses improve the investor predictions and concurrently reduce the shareholders' uncertainties and company costs. The big data field is expanding to the modern economy context and may assist financial market participants to make more informed choices about the companies in which they intend to invest. Additionally, these data affect the price, stock value, and investment decisions of the mentioned companies [5]. Due to this, the stock return

*Corresponding Author Institutional Email: m.ghasemzadeh@yazd.ac.ir (M. Ghasemzadeh)

Please cite this article as: F. Qolipour, M. Ghasemzadeh, N. Mohammad-Karimi, The Predictability of Tree-based Machine Learning Algorithms in the Big Data Context, International Journal of Engineering, Transactions A: Basics Vol. 34, No. 01, (2021) 82-89

sign that is affected by the stock price is also influenced by them.

Capital markets include recession and prosperity. During prosperity, the stock return is positive in most companies, and vice versa. In this regard, if a company can keep its stock return positive while the market is in the state of recession, it is preferable for purchasing. Due to the mentioned reason, forecasting stock returns during the recession is highly significant. Machine learning algorithms must be able to appropriately predict stock returns in the times of market downturn and boom days. The primary motivation of conducting the present research is to study the predictability of machine learning algorithms in the boom and bust cycles of the stock market and to compare them in the context of big data. The selected data have been employed in this field for the first time. These data include 23 technical features, in addition to the 10 basic features of the Tehran stock market. In the current study, it has been sought to examine the impact of adding 23 technical features to the mentioned data in this case, in addition to the predictability of machine learning algorithms in boom and bust cycles in the big data field and its corresponding tools. Moreover, the obtained results of the original data and the new data of this field have been compared.

2. LITERATURE REVIEW

Machine learning is a scientific study that innovates various algorithms to improve its performance on a particular task gradually. Due to the remarkable ability to extract valid information from datasets and the most optimized pattern recognitions, numerous recent papers have focused on the application of machine learning techniques in financial subjects. These methods encompass basic statistical models such as logistic regression as well as artificial intelligence methods such as decision trees, support vector machines, and artificial neural networks. In contrast to traditional machine learning models, ensemble models (i.e., a combination of several models) are machine learning-based approaches in which several basic algorithms are utilized for solving a particular problem, and it has been proven that they demonstrate a higher performance in predicting financial time series in comparison to singular learning models.

In ensemble learning algorithms, bagging and boosting are among the most popular techniques in the machine learning field. Bagging (bootstrap aggregating) that has been developed by Bryman [6] is one of the most straightforward and most intuitive approaches in the ensemble, which in addition to superb performance, it also reduces the variance and prevents the occurrence of overfitting. The begging algorithm is obtained from the Bootstrap technique, which produces subsets of training data by repeating the training data set. Each subset is utilized for fitting a separate basic learner, and the final prediction results are gathered through the majority voting method. Boosting is another ensemble technique that is according to the research of Freund and Shapir [7]. In contrast to the bagging technique, this method creates various learners by applying a sequential weighting algorithm to training samples. Any sample that has not been classified by the previous learner will gain more weight in the next round of training. Consequently, unclassified training samples will usually occur in the subsequent Bootstrap sample, and the bias can be effectively reduced. The ultimate model of the Boosting algorithm is a combination of all basic learners that are weighted through their corresponding predictive performance [4]. The random forest algorithm exploits the bagging method, and the gradient boosted tree algorithm uses the boosting method.

Two types of ensemble classifiers have been organized (i.e., homogeneous and heterogeneous ensemble classifiers) through utilizing the majority voting and the bagging method by Tsai et al. [8]. In this regard, the financial ratio and macroeconomic characteristics in the Taiwan stock market have been considered to examine the performance of stock return forecasts. The result has indicated that ensemble classifiers perform more beneficial than singular classifiers from the aspect of forecast accuracy and return on investment (ROI) [8].

Similarly, a comparative study has been conducted by Ballings et al. [9] in which ensemble learning algorithms, including random forests, and the AdaBoost, have been compared to singular learning models, including neural networks, linear regression, support vector, and k-nearest neighbors algorithms. Afterward, the one-year stock price direction of European companies have been predicted. The AUC results illustrate that the random forest is the superior algorithm among the examined algorithms [9].

Random forest and XGBoost algorithms have been applied for the classification problem by Basak et al. [10]. Moreover, according to the prevailing price of the past few days, it has been predicted that stock prices would rise or fall. Eventually, experimental results have displayed that the performance of the forecasting process for different types of companies has improved in comparison to the available companies [10].

Regarding the perspective that macroeconomic indicators can solely predict the accurate one-month ahead price of major US stock indices, four ensemble models of random forest quantile regression, quantile regression neural network, bagging regression and boosting regression have been created by Wong et al. [11]. The results have demonstrated that the forecasting performances of these ensemble learning methods are superior to traditional time series models. Additionally, this study proposes a hybrid approach of long short term

memory, and then, it proves that macroeconomic features are pioneers [11].

3. MATERIALS AND METHODS

3. 1. System Model and Hypotheses The structure of a machine learning model does not necessarily require the development of an entirely novel algorithm. Customization and utilization of investigated models can also lead to improved prediction results. Even the preprocessing of information before implementing the model is also part of the study innovation. In this research, through using 10 basic features of the Tehran Stock Exchange, 23 technical features for two active companies in the stock exchange, are extracted, and the new data with 10 basic features along with the mentioned 23 technical features are generated. Afterward, the recession and prosperity of these companies are separated, and the predictability of the stock return is compared by utilizing singular, and ensemble tree-based algorithms such as decision tree and random forest ensemble algorithms and gradient boosted tree in big data space. Furthermore, it has been intended to evaluate the impact of adding 23 technical features to the initial data and exploiting the PCA feature reduction technique on the performance of these algorithms. In this regard, the general process of the study can be observed in Figure 1.

3. 2. The Proposed Method In huge markets such as the stock market, which is daily encountered with a massive amount of data and the prompt reaction of shareholders is crucial, it is highly significant to be able to select the right decisions as soon as possible.

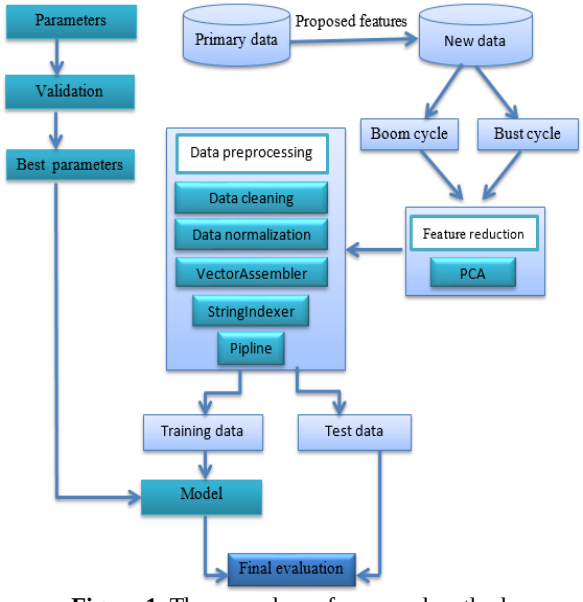

Figure 1. The procedure of proposed method

Accordingly, in forecasting the huge financial series with machine learning algorithms, achieving the minimum error rate in the minimum amount of time is critical. The initial data of this study is the stock information of two active companies in the Tehran Stock Exchange that each stock data contains ten basic features. These features include the date, initial price, highest price, lowest price, final price, volume, value, number of trades, and yesterday's stock price each day. Through technical analysis, 23 technical features were measured from the collected information of 10 basic features from companies' stocks over ten years. The new dataset is generated by extracting 23 technical features and adding to the original data along with ten basic characteristics.

Proposed features, which are based on the technical analysis are total price index, industry index, equal-weighted price index, industries index, total return, industry return, beta coefficient of industry return, beta coefficient of total index, moving average divergence convergence, three-day moving average, five-day moving average, moving average Ten-Day, 20-Day Moving Average, 30-Day Moving Average, Seven-Day Moving Average, Weighted Moving Average, Relative Strength Index, Bollinger Bands (Upper and Lower bands), First Days of Each Week, Latest Days of Stock Market, First Months of Each Year and Exchange rate index).

After collecting the suggested features, through the assistance of the total price index feature, a new feature was generated that was called the period. In this regard, it represents the market boom and bust cycles. If the stock return sign can be predicted during the recession as well as the prosperity, the great achievement will be acquired due to the particular importance that they possess recently in the capital market.

According to the period feature, the collected data have been converted into two categories of prosperity and recession periods with positive and negative signs, respectively. In addition to comparing the prosperity and recession period for each share, the impact of adding 23 technical features to the basic characteristics have been considered, as well. To achieve this aim, the performance of the decision tree, random forests, and gradient boosted tree algorithms in six created modes have been compared.

The six created modes are:

- prosperity period for each share and 10 basic features
- prosperity period for each share and new features (10 basic features + 23 technical features)
- prosperity period and feature reduction utilizing PCA feature extraction method
- recession period for each share and 10 base features
- recession period for each share and new features
- recession period and feature reduction utilizing the PCA feature extraction method.

After collecting the data and separating the boom and bust cycle (recession and prosperity period) of both companies stocks, with the assistance of the missingpy library in Python programming language and random forest technique, the missing or unassigned data values have been filled. Afterward, for the process of normalizing the data, through using the VectorAssembler function in the pyspark library, which is a designated for working with the big data, the available features were converted into a vector. Furthermore, they have given as inputs to the MimMaxScaler normalizer function; thus, a normalized vector of the features was acquired through the mentioned process.

The preprocessing step of data in the big data space relatively differs from the corresponding step in common data space. For the big data preprocessing task, after the converting of properties into a single vector property by VectorAssembler, the label property is also converted to numeric data by the StringIndexer function if it is in the string data form. Consequently, the mentioned two features are placed in a data frame by the existing Pipeline function in pyspark; moreover, they are considered as the model input.

In the field of time series prediction, by applying the machine learning algorithms after the data preprocessing step, it is necessary to convert the data into two categories: a training set and a test set. The training set is exploited for making the model, and the test set is used to validate the accuracy of the model. Of course, it is worth mentioning that the mentioned test dataset is divided into two categories of test and validation, which are used for evaluating the amount of training that has been obtained from the training data set. The evaluation outcome is solely applied to select the best training part, and after finding this part, it has been utilized for the final evaluation, which is examined with the test data. To determine the test and training set for each share, 20% of the total data have been allocated to the test set, and 80% of it has been assigned for the training set.

In financial time series, traditional cross-validation procedures such as K-fold is not utilizable due to random and unbiased selection of training data sets as well as the time dependencies in time series. As a consequence, the stages of the cross-validation method in time series are as follows:

First stage: the whole training data set has been divided into several sections. The default value of this segmentation number in the conducted implementations is three; however, after performing several experiments, the value of 10 has been selected for the segmentation number. As a result, the training samples have been separated into 10 sections, and each time, one section has been selected.

Second stage: The first two sections, which contain two-tenths of the training data, have been divided into two sections. 50% of the data (equivalent to 10% of the total training data) has been assigned for the training purpose, and the other 50% for the validation purpose and the test data.

Third stage: at this stage, one-tenth of the training data has been added at each turn, and the newly added section has been allocated to the test data, and further, the previous sections have been considered for training purposes. This process has been extended until all ten sections are validated, as displayed in Figure 1.

The classification techniques have been utilized for numerous applications in various fields of science. There are several methods for the evaluation of classification algorithms. Analysis of such criteria and their importance should be appropriately interpreted to evaluate different learning algorithms. In advance of introducing evaluation methods, it's preferable to express the fundamental and significant concept of a confusion matrix for two-class or binary classification purposes.

In the confusion matrix, four symbols are encountered: TN, FN, TP, and FP. If the sample is actually positive and is also classified as a positive sample, it is regarded that the sample has been correctly classified as positive; moreover, the "TP" symbol is assigned for it. If the corresponding sample is actually positive and is classified as negative, it is declared that the sample is conversely classified as negative, and it is displayed with the "FN" symbol. Accordingly, if the negative sample is classified as a negative sample, the sample is considered as a correctly classified negative sample; furthermore, it is represented with the "TN" symbol. Eventually, if the negative sample is classified as a positive sample, it has been considered as a misclassified positive sample and is exhibited with the "FP" symbol. As can be observed in the following, the confusion matrix is applied for the calculation of numerous standard classification criteria.

Accuracy: It is one of the most common measures for classification performance and is defined as the ratio of correctly classified samples to the total number of samples:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Recall: It represents the ratio of correctly classified positive samples to samples that have been actually labeled positive. This evaluation criterion is expressed as follows:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Precision: It indicates the ratio of correctly classified positive samples to the total samples that are correctly or incorrectly classified as positive ones. This evaluation criterion is illustrated as follows:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

AUC-ROC: A perfect and superb model has an Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) close to one, which means it has adequate separability. A deficient model has an AUC close to 0, meaning it has the worst separability. Furthermore, when the AUC is 0.5, it means that the model is not able for any class separability measures.

4. EXPERIMENTS

Afterwards, the above procedure was applied to the six stated modes, and the results of three prediction algorithms were compared. Table 1 shows, that validation and testing of Pars Oil Company in the prosperity period along with the PCA feature reduction method for all three algorithms have obtained superior results in comparison to the initial and the new data with 33 features. Additionally, among the three selected algorithms, two ensemble algorithms (random forests and gradient boosted tree) have mostly acquired better results in comparison to the singular algorithm of the decision tree. Moreover, the achieved results for the test data sets in the gradient boosted tree algorithm have been better than other algorithms that demonstrate the additional generalizability of this algorithm.

In Table 2, the obtained results of the Pars Oil Company evaluation during the recession have been examined. As illustrated in this table, the PCA method leads to better results in most cases compared to the 10 basic features and the new data, which includes 33 features. However, it is crystal clear that adding 23 technical features to the initial data had positively influenced the acquired results in all three algorithms.

TABLE 1. The evaluation results of Pars Oil Company during the prosperity period

Models	**The validation results of Pars Oil Company during the prosperity period**											
	33 features				**10 basic features**				**PCA Feature reduction algorithm**			
	Au-Roc	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***	***Auc Roc***	***Precision***	***Recall***	***Accuracy***
DT	0.864	0.802	**0.830**	0.872	0.511	**0.505**	**0.627**	0.601	0.931	0.888	0.865	0.941
RF	0.889	**0.836**	0.804	**0.873**	0.525	0.442	0.536	**0.608**	0.944	0.888	0.865	**0.955**
GBT	**0.927**	0.803	**0.830**	0.865	**0.526**	0.460	**0.627**	0.605	**0.973**	0.888	0.865	0.953
Models	**The test results of Pars Oil Company during the prosperity period**											
	33 features				**10 basic features**				**PCA Feature reduction algorithm**			
	Au-Roc	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***
DT	0.559	0.687	0.427	0.618	0.498	0.5	0349	0.502	1.0	1.0	1.0	1.0
RF	0.720	0.723	**0.533**	0.666	0.503	0.490	0.728	0.487	1.0	1.0	1.0	1.0
GBT	**0.743**	**0.776**	0.504	**0.681**	**0.526**	**0.502**	**0.980**	**0.507**	1.0	0.990	1.0	0.995

TABLE 2. The evaluation results of Pars Oil Company during the recession period

Models	**The validation results of Pars Oil Company during the recession period**											
	33 features				**10 basic features**				**PCA Feature reduction algorithm**			
	Au-Roc	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***	***Auc Roc***	***Precision***	***Recall***	***Accuracy***
DT	0.873	0.927	0.861	0.898	0.514	0.344	0.321	0.637	0.971	0.963	0.987	0.969
RF	**0.915**	**0.933**	0.788	0.899	0.532	0.421	0.231	**0.643**	1.0	1.0	1.0	0.996
GBT	0.914	0.923	**0.866**	**0.908**	**0.535**	**0.479**	**0.440**	0.611	1.0	1.0	1.0	0.996
Models	**The test results of Pars Oil Company during the recession period**											
	33 features				**10 basic features**				**PCA Feature reduction algorithm**			
	Au-Roc	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***
DT	0.383	0.529	0.246	0.609	0.468	**0.666**	**0.547**	**0.609**	0.889	0.752	1.0	0.868
RF	**0.857**	**0.931**	**0.561**	**0.807**	0.505	0.0	0.0	0.598	**1.0**	**0.858**	**1.0**	**0.934**
GBT	0.759	0.921	0.479	0.774	**0.539**	**0.666**	0.027	0.604	1.0	0.752	1.0	0.868

Furthermore, ensemble algorithms in this period, similar to the boom period, have presented a more significant performance with respect to the singular algorithm of the decision tree, and all three algorithms have presented appropriate predictability during the boom period. The most generalizability to the new data has been observed in the random forest algorithm along with the availability of 23 technical features and the application of feature reduction technique. On the other hand, for the case of 10 basic characteristics, the decision tree results in a better performance in test data sets. Generally, both periods of the prosperity and the recession, the PCA feature reduction method has demonstrated a more significant performance for the Pars Oil Company in comparison to the two other cases. Moreover, the initial data have not represented a suitable performance, and among the applied algorithms, the ensemble algorithms had achieved more excellent results in relation to the singular algorithms.

In Table 3, the obtained results of the Shazand Petrochemical Company during the boom period have been investigated. As can be observed, in almost all of the cases, the ensemble algorithms have performed better than the singular algorithms. The performance of the feature reduction algorithm on the company's data during the prosperity period is also higher than the initial data along with 10 basic features and the 23 technical features utilization; thus, the mere initial data will not be an appropriate representative for prediction of the company's situation in the future. As the obtained values in validation and testing reveal, adding 23 technical features and subsequent feature extracting from them positively affect the performance of all three selected algorithms. The class separability and generalizability of these data in the absence of technical features and solely considering the initial data are not appropriate; while in these data, the use of technical features, as well as the ensemble algorithms, are recommended in comparison to the utilization of traditional singular algorithms.

In Table 4, the evaluation results of the Shazand Petrochemical Company during the recession have been reviews. As can be noticed, in both validation and testing, the PCA feature reduction technique has demonstrated a more significant performance compared to the other two modes, and on the other hand, 10 basic features mode did not present an appropriate performance. The stated fact means that adding 23 technical features to the corresponding data is also effective. Moreover, among the chosen algorithms, in all three cases, ensemble learning algorithms have performed better in the validation process. Furthermore, these algorithms represent proper generalizability in the test process.

The two ensemble algorithms of random forest and the gradient boosted tree approximately have similar results in all three modes, and for all four input data sets. Additionally, they have exhibited higher performance in comparison to the traditional and singular algorithm of the decision tree in most cases. Generalizability and class separability in ensemble algorithms, particularly in the new data set, including 23 technical features in addition to 10 basic features, as well as the application of the PCA feature reduction method, are superior to the singular learning methods. The predictability of algorithms during the company's recession is as accurate as its prosperity period.

TABLE 3. The evaluation results of Shazand Petrochemical Company during the prosperity period

Models	**The validation results of Pars Oil Company during the prosperity period**											
	33 features				**10 basic features**				**PCA Feature reduction algorithm**			
	Au-Roc	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***	***Auc Roc***	***Precision***	***Recall***	***Accuracy***
DT	0.867	0.932	0.796	0.863	0.508	0.532	0.407	0.623	0.997	0.996	1.0	0.995
RF	**0.936**	**0.939**	0.782	**0.898**	**0.573**	0.463	0.364	0.627	**1.0**	**1.0**	1.0	**1.0**
GBT	0.920	0.894	**0.862**	0.890	0.544	**0.558**	**0.573**	**0.629**	**1.0**	0.997	1.0	0.998
Models	**The test results of Pars Oil Company during the prosperity period**											
	33 features				**10 basic features**				**PCA Feature reduction algorithm**			
	Au-Roc	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***	***Au-Roc***	***Precision***	***Recall***	***Accuracy***
DT	0.628	**0.829**	0.323	0.626	**0.5**	0.0	0.0	**0.497**	1.0	1.0	1.0	1.0
RF	**0.773**	0.706	0.733	**0.712**	0.429	0.365	0.142	0.444	1.0	1.0	1.0	1.0
GBT	0.707	0.607	**0.885**	0.655	0.490	**0.483**	**0.695**	0.473	1.0	1.0	1.0	1.0

TABLE 4. The evaluation results of Shazand Petrochemical Company during the recession period

Models	The validation results of Shazand Petrochemical Company during the recession period											
	33 features				10 basic features				PCA Feature reduction algorithm			
	Au-Roc	*Precision*	*Recall*	*Accuracy*	*Au-Roc*	*Precision*	*Recall*	*Accuracy*	*Auc Roc*	*Precision*	*Recall*	*Accuracy*
DT	0.919	**0.885**	0.867	0.9	0.499	0.496	0.543	0.625	0.993	1.0	0.993	0.995
RF	0.934	0.877	**0.912**	0.887	0.515	0.543	0.598	**0.631**	**0.997**	1.0	0.993	0.995
GBT	**0.939**	0.883	**0.912**	**0.906**	**0.546**	**0.595**	**0.818**	0.618	0.996	1.0	**1.0**	0.995
Models	**The test results of Shazand Petrochemical Company during the recession period**											
	33 features				**10 basic features**				**PCA Feature reduction algorithm**			
	Au-Roc	*Precision*	*Recall*	*Accuracy*	*Au-Roc*	*Precision*	*Recall*	*Accuracy*	*Au-Roc*	*Precision*	*Recall*	*Accuracy*
DT	0.674	0.532	**0.506**	0.638	0.5	0.0	0.0	**0.614**	0.965	0.880	1.0	0.947
RF	**0.809**	**0.944**	0.419	**0.766**	0.517	0.4	0.024	0.609	**0.990**	**0.964**	1.0	**0.985**
GBT	0.718	0.579	0.493	0.666	**0.531**	**0.431**	**0.617**	0.538	0.957	0.880	1.0	0.947

5. CONCLUSION

The results of stock evaluations of the two active companies demonstrate that the ensemble algorithms of random forest and gradient boosted tree have better predictability in comparison to the decision tree, which is a singular tree-based algorithm, during both of the boom and bust cycles. Additionally, the AUC-ROC in these algorithms represent the fact that the ensemble algorithms create further class separability; therefore, they can accurately predict the positive and negative sign of stock return. Furthermore, adding 23 technical features to 10 basic features and subsequent creation of the new data and further feature extracting utilizing the PCA method has a significant effect in improving the performance of the algorithms. In this regard, the most excellent result has been obtained from the application of the PCA method. Moreover, the decision tree algorithm possesses a higher performance speed than the ensemble algorithms since it merely exploits one tree for the prediction procedure, and among the ensemble algorithms, the decision tree is faster in case of the current study data.

6. REFERENCES

1. Khedmati. M, Seifi. F, Azizi. M.J, "Time Series Forecasting of Bitcoin Price Based on Autoregressive Integrated Moving Average and Machine Learning Approaches", ***International Journal of Engineering, Transactions A: Basics,*** Vol. 33, No. 7, (2020), 1293-1303. DOI: 10.5829/IJE.2020.33.07A.16
2. Hemati. H.R., Ghasemzadeh. M, Meinel. C, "A Hybrid Machine Learning Method for Intrusion Detection", ***International Journal of Engineering, Transactions C: Aspects,*** Vol. 29, No. 9, (2016), 1242-1246, DOI: 10.5829/idosi.ije.2016.29.09c.09
3. Liu, J, and Kemp. A, "Forecasting the sign of U.S. oil and gas industry stock index excess returns employing macroeconomic variables", ***Energy Economics***, Vol. 81, (2019), 672-686. https://doi.org/10.1016/j.eneco.2019.04.023
4. Jiang. M, Liu. J, Zhang. L, and Liu. C, "An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms", ***Physica A-Statistical Mechanics and Its Applications***, Vol. 541, (2020) 122272. https://doi.org/10.1016/j.physa.2019.122272
5. Begenau. J, Farboodi. M, and Veldkamp. L, "Big data in finance and the growth of large firms", ***Journal of Monetary Economics***, Vol. 97, (2018), 71-87. https://doi.org/10.1016/j.jmoneco.2018.05.013
6. Breiman. L, "Bagging Predictors", ***Machine Learning Archive***, Vol. 24, No. 2, (1996), 123-140.
7. Freund. Y and Schapire. R.E, "Experiments with a New Boosting Algorithm", in Proceedings of the International Conference on Machine Learning, (1996), 148-156.
8. Tsai. C.-F, Lin. Y.-C, Yen. D.C, and Chen. Y.M, "Predicting stock returns by classifier ensembles", ***Applied Soft Computing***, Vol. 11, No. 2, (2011), 2452-2459. https://doi.org/10.1016/j.asoc.2010.10.001
9. Ballings. M, Van den Poel. D, Hespeels. N, and Gryp. R, "Evaluating multiple classifiers for stock price direction prediction", ***Expert Systems With Applications***, Vol. 42, No. 20, (2015), 7046-7056. https://doi.org/10.1016/j.eswa.2015.05.013
10. Basak. S, Kar. S, Saha. S, Khaidem. L, and Dey. S.R, "Predicting the direction of stock market prices using tree-based classifiers", ***The North American Journal of Economics and Finance***, Vol. 47, (2019), 552-567. https://doi.org/10.1016/j.najef.2018.06.013
11. Weng. B, Martinez. W.G, Tsai. Y, Li. C, Lu. L, Barth. J.R., Megahed F.M, "Macroeconomic indicators alone can predict the monthly closing price of major U.S. indices: Insights from artificial intelligence, time-series analysis and hybrid models", ***Applied Soft Computing***, Vol. 71, (2018), 685-697. https://doi.org/10.1016/j.asoc.2018.07.024

Persian Abstract

چکیده

این پژوهش در رابطه با پیش‌بینی‌پذیری الگوریتم‌های یادگیری‌ماشین گروهی و منفرد مبتنی بر درخت در دوران رکود و رونق دو شرکت حاضر در بورس تهران در بستر داده‌های حجیم است. چالش موردتوجه در این حوزه، این است که مدیران اقتصادی و جامعه علمی، همچنان به دنبال مدل‌های پیش‌بینی با دقت بیشتر و در زمان کمتر می باشند و همچنین پیش‌بینی دوران رکود شرکت‌ها در بازار بورس از اهمیت به سزایی برخوردار است. الگوریتم‌های یادگیری ماشین باید بتوانند علامت بازده سهام را در روزهای رکود بازار به‌خوبی روزهای رونق پیش‌بینی کنند. رفع چالش یادشده موجب ارتقاء کیفیت خرید سهام و به سبب آن موجب ارتقای سودآوری می‌شود. راه‌حل پیشنهادی، تکیه‌بر به‌کارگیری الگوریتم‌های یادگیری ماشین مبتنی بر درخت در بستر داده‌های حجیم دارد. در راه‌حل پیشنهادی از الگوریتم درخت تصمیم که یک الگوریتم سنتی و منفرد مبتنی بر درخت است و همچنین دو الگوریتم مدرن و گروهی مبتنی بر درخت یعنی جنگل تصادفی و درخت گرادیان تقویتی برای پیش‌بینی پذیری علامت بازده سهام در روزهای رکود و رونق، استفاده‌شده است. موارد یادشده، با به‌کارگیری ابزارهای یادگیری ماشین به زبان پایتون و در کتابخانه PYSPARK که مخصوص داده‌های حجیم است، پیاده‌سازی گردیدند. داده‌های تحقیق که در این پژوهش به کار گرفته شدند، اطلاعات مربوط به سهام دو شرکت از بورس تهران می‌باشند. نتایج نشان می‌دهند که الگوریتم‌های گروهی استفاده‌شده عملکرد بهتری نسبت به الگوریتم منفرد داشته‌اند. همچنین اضافه کردن ۲۳ ویژگی فنی به دادگان اولیه و سپس استفاده از روش کاهش ویژگی PCA بهترین عملکرد را در بین حالت‌های دیگر داشته است. این مطلب موجب پیش‌بینی با دقت بالاتری می‌گردد. دراین‌بین به این نتیجه می‌رسیم که دادگان اولیه چه در دوران رونق و چه در دوران رکود از قابلیت تفکیک‌پذیری و تعمیم‌پذیری مناسبی برخوردار نیست.