information

MDPI

Article

Identification of Insider Trading Using Extreme Gradient Boosting and Multi-Objective Optimization

Shangkun Deng [1,2,*], Chenguang Wang [1], Jie Li [1], Haoran Yu [1], Hongyu Tian [1,2], Yu Zhang [1], Yong Cui [1,*], Fangjie Ma [3] and Tianxiang Yang [4]

1 College of Economics and Management, China Three Gorges University, Yichang 443002, China; wcg9828@126.com (C.W.); ljie0821@outlook.com (J.L.); FaTianYuff@163.com (H.Y.); ctguthy@126.com (H.T.); zhangyu@ctgu.edu.cn (Y.Z.)

2 Financial Research Institute, College of Economics and Management, China Three Gorges University, Yichang 443002, China

3 College of Science and Technology of China Three Gorges University, Yichang 443002, China; sxdxmfj@163.com

4 School of Creative Science & Engineering, Waseda University, Tokyo 169-8555, Japan; you_tensyou@akane.waseda.jp

* Correspondence: dsk8672@163.com (S.D.); cuiyong826@126.com (Y.C.)

Received: 28 October 2019; Accepted: 22 November 2019; Published: 25 November 2019

Abstract: Illegal insider trading identification presents a challenging task that attracts great interest from researchers due to the serious harm of insider trading activities to the investors' confidence and the sustainable development of security markets. In this study, we proposed an identification approach which integrates XGboost (eXtreme Gradient Boosting) and NSGA-II (Non-dominated Sorting Genetic Algorithm II) for insider trading regulation. First, the insider trading cases that occurred in the Chinese security market were automatically derived, and their relevant indicators were calculated and obtained. Then, the proposed method trained the XGboost model and it employed the NSGA-II for optimizing the parameters of XGboost by using multiple objective functions. Finally, the testing samples were identified using the XGboost with optimized parameters. Its performances were empirically measured by both identification accuracy and efficiency over multiple time window lengths. Results of experiments showed that the proposed approach successfully achieved the best accuracy under the time window length of 90-days, demonstrating that relevant features calculated within the 90-days time window length could be extremely beneficial for insider trading regulation. Additionally, the proposed approach outperformed all benchmark methods in terms of both identification accuracy and efficiency, indicating that it could be used as an alternative approach for insider trading regulation in the Chinese security market. The proposed approach and results in this research is of great significance for market regulators to improve their supervision efficiency and accuracy on illegal insider trading identification.

Keywords: sustainable development; identification; insider trading; security market; XGboost; multi-objective optimization

1. Introduction

Security exchange is a financial market where securities such as equities and bonds are issued and traded. Through this intermediary, companies have alternative access to raise capital from the public, and it also provides investors access to make investments or purchase the ownership of the listed companies. In the last few decades, with the rapid development of Chinas economy, the Chinese security exchange has become one of the most essential security exchanges in the world. Similar to the

Securities and Exchange Commission (SEC) for the U.S., in China, the China Securities Regulatory Commission (CSRC) is the regulatory body that oversees the Chinese security market [1]. Nonetheless, there were many illegal insider trading activities that occurred in the Chinese security market every year [2]. Illegal insider trading is the security transaction performed based on non-public information (e.g., financial condition, or merger and acquisition) before the information is made public [3], which has adverse influences toward the investment confidence on the market participants, thus it would be extremely harmful to the sustainable development of security markets.

In the past few decades, insider trading activities have severely damaged the fairness of stock market trading, which attracted the attention of China security regulators, and they have continuously strengthened supervision and regulations over the insider trading activities. According to statistics published by CSRC, in the first half of 2019 the CSRC has issued a total of thirty-six penalties for insider trading cases [4], which accounts for more than 50% of all penalties in the Chinese security market in that period. Therefore, although market regulators have adopted measures to the regulation of insider trading activities, they might not have been efficiently controlled. Since the current situation of insider trading becomes more serious, and it is quite harmful to the sustainable development of the Chinese security market, it would be of vast significance to develop an intelligent system to identify and control it in the early stage. It is considered to be extremely beneficial for investors to trade in a fairer trading circumstance.

However, illegal insider trading identification is generally regarded as very challenging work because of the complex, nonlinear, and non-stationary characteristics of stock markets [5]. In the last few decades, numerous researchers have started to employ machine learning (ML) methods to address classification problems in many application fields. Among the ML methods, the artificial neural network (ANN) [6], support vector machine (SVM) [7], and adaptive boosting (Adaboost) [8] are well-known and efficient algorithms that have been frequently applied in many identification problems. For instance, Farooq et al. proposed an ANN-based method for damage detection and identification in smart structures [9], and Li and Yang developed an ANN-based method for beam damage identification [10]. Nonetheless, generally it is necessary to employ a large number of training samples for ANN to perform well in many applications, thus a small number of insider trading cases may result in an over-fitting of the trained model [11]. Compared with ANN, SVM is found to produce superior results for classification problem [12–14], and it has been widely applied by researchers for identification work. For instance, Rodriguez et al. developed an SVM-based system for posture identification [15]. Jiang et al. proposed an SVM-based approach for liver cancer identification [16]. Nonetheless, SVM also has several drawbacks in applications, such as the classification performance of SVM sometimes is very sensitive to the kernel selection, therefore selecting an appropriate kernel for SVM becomes a problem [17]. Additionally, Adaboost is a well-known ML method for classification problems, which has been successfully applied to many application fields [18–20]. Nonetheless, it also has the central disadvantage that it is very sensitive to noise, thus it tends to produced over-fitting results during the model training period if the training samples consist of many noises [21].

For the purpose of identifying the illegal insider trading activities, a vast number of related variables of quoted companies could be employed. In the present study, among the available information, a total of sixteen indicators are employed as the variables for illegal insider trading identification. However, for the well-known machine learning approaches such as SVM or ANN, they often encounter a problem that if there are redundant variables involved in the training samples, performance for classification may become worse. To overcome this problem, an alternative method called random forest (RF) [22] has been increasingly adopted by many researchers. RF is a tree-based method for classification or regression problems, which is operated by constructing a group of decision trees from the training samples, and the output is the class that was voted by the individual trees. Compared with other ML methods, RF corrects the over-fitting habit of the decision tree on training samples, and it is recognized as one of the most outstanding classification methods and has been widely applied in the current research. For example, Deng et al. have developed an RF-based method

for insider trading identification [23]. Murugan et al. have employed an RF-based method for skin cancer detection [24]. In recent years, other than the RF, a state-of-the-art tree-based ensemble learning method called XGboost (eXtreme Gradient boosting) has been proposed to address lots of identification problems [25,26]. Although its computation speed is relatively lower than linear models, XGboost is a relatively new, robust, and accurate method in the ML field, which is similar to the tree-based method RF. Indeed, in a lot of literature, XGboost was found to have the advantage of not easily overfitting and it outperforms RF in terms of both classification accuracy and computation speed [27].

In the present study, we factored in the possibility that certain features of some relevant indicators may have correlations with insider trading activities. Therefore, a vast number of related indicators were employed as the input variables of XGboost for identifying insider trading in the experiments. In addition, there are several parameters of XGboost, and the classification accuracy experiment performances were found to be extremely sensitive to the selection of parameters [28], that is, they may markedly affect the performance of XGboost. Hence, it would be very crucial to optimize the initial parameters of XGboost.

For using the XGboost to identify insider trading activities, we mainly encounter the following two problems: (1) There are several parameters in the XGboost model, which may highly affect the effectiveness of identification; (2) For insider trading identification in stock markets, besides the identification accuracy, market regulators generally would also take into account identification efficiency, that is, they expect a model with as few as possible insider trading cases which are incorrectly identified or classified. Nonetheless, the identification accuracy and efficiency are sometimes considered as conflicting objectives. Therefore, in the present research, it is necessary to address a multi-objective optimization problem. In this study, we decide to adopt a multi-objective optimization algorithm to optimize the parameters of XGboost. Among the multi-objective algorithms, the non-dominated sorting genetic algorithm II (NGSA-II) algorithm is an outstanding multi-objective optimization algorithm. The NSGA-II has been widely applied in many research fields. For instance, a NSGA-II-based approach was developed by Tamimi et al. for intrusion detection [29]. Lin et al. have adopted an NSGA-II-based approach to find the solutions of the optimal multi-type sensor placement for achieving outstanding performances of damage detection [30]. Guan applied the NSGA-II algorithm for parameter optimization in the main steam temperature control [31]. Successful applications of NSGA-II in the above literature suggests that it is an effective algorithm for addressing multi-objective optimization problems. Compared with multi-objective evolutionary algorithms (MOEAs) that use non-dominated sorting and sharing, the NSGA-II can find a much better spread of solutions and better convergence near the true Pareto-optimal front for most problems. In addition, compared with traditional signal objective optimization algorithms, such as GA and PSO [32], another advantage of NSGA-II is its multi-objective optimization. Existing literature on insider trading identification has hitherto set the identification accuracy as the single optimization objective, with the identification efficiency not being treated as the optimization objective. Currently, a lot of researchers have used an integrated method to improve the performance of the model. For example, Garg designed two hybrid models, which are GSA-GA (Gravational Search Algorithm-Genetic Algorithm) [33] and PSO-GA (Particle Swarm Optimization-Genetic Algorithm) [34], for constrained optimization problems. Alarifi et al. [35] proposed ANFIS-PSO (Adaptive Network-based Fuzzy Inference System-Particle Swarm Optimization) and ANFIS-GA (Adaptive Network-based Fuzzy Inference System-Genetic Algorithm) models in predicting thermo physical properties of hybrid nanofluid. Therefore, in this study, we decided to pursue the use of NSGA-II to optimize the parameters of XGboost, and the objective values for optimization are set to be two objectives: identification accuracy and efficiency.

Among the relevant indicators for insider trading identification, information of security market performances, and corporate governance and shareholding structure, as well as finance-related indicators were extracted because of the following reason: (1) Stock market performances of the company are generally intuitively recognized as the identifying indicators of insider trading activities. Numerous researchers around the world have found there were significant correlations of abnormal

price fluctuations or trading volumes with insider trading activities. For example, Chiang et al. [36] found the correlation of insider trading activities and stock return volatility. Jain and Mirman [37] discovered that several insider trading activities often had an extremely strong impact on the stock price movement. Jabbour et al. [38] performed a research of the mergers and acquisitions cases in Canada, and they found that the share prices tended to move upward to a certain extent prior to the announcement of the insider information; (2) Some indicators of corporate governance have been recognized widely as useful information for identifying insider trading activities. For instance, Dai et al. have developed an index to gauge the democracy degree of the listed companies, and they concluded that the more the company managers held its shares, the larger the probability of insider trading activities occurred [39]. Chronopoulos et al. [40] have discovered the correlation between corporation ownership structure and illegal insider trading. Lu et al. [41] have studied the correlations between illegal insider trading and corporate social responsibility; (3) Some studies also incorporated information from the financial indicators of a company, such as price-earnings (P/E) ratio. They were also employed as the input variables in our proposed method because financial performance information can sometimes reflect the profitability, growth, and investment value of quoted companies. Therefore, they have been taken into account by us as beneficial information for insider trading identification.

The main steps of the proposed approach for insider trading activity identification are as follows: (1) Firstly, the proposed method derives the indicators of insider trading cases as well as the non-insider trading samples, and the whole samples are divided into in-sample and out-of-sample datasets; (2) Then, by using the in-sample dataset, the XGboost algorithm is adopted to classify the insider trading and non-insider. In the meanwhile, the NSGA-II algorithm is employed to optimize the parameters of XGboost. This step will be stopped until the suitable initial parameters of XGboost are found; (3) Next, the trained XGboost model is generated and its identification performances (accuracy and efficiency) in the out-of-sample dataset are investigated; (4) Finally, the importance of each indicator for insider trading identification was analyzed by the importance scores of XGboost. Since the proposed method integrates the XGboost and NSGA-II, it is expected to own the advantages such as high identification accuracy, not easy to be overfittting, and capability of solving multi-objective optimization problems. Additionally, from the indicator importance results, list corporations can pay much attention to the most important indicators that are very relative to insider trading activities, then they may take measures for avoiding insider trading activities. The novelty of this research could be summarized as follows: (1) a novel approach has been developed for the identification of illegal insider trading activities. Both the identification accuracy and regulation efficiency were considered as the optimization objective in the proposed approach and both of them were empirically evaluated. (2) We performed experiments with related indicators calculated under three different time window lengths for finding a relatively superior time window length for insider trading regulations. From the experimental results, a relatively well-performed time window length has been found for insider trading identification in the Chinese stock market. (3) We have utilized information from three aspects and up to a total of sixteen relevant indicators. According to the results, we have discovered several essential indicators for illegal insider trading identification in the Chinese security market. The outstanding performance of the proposed approach in this research indicates that the it is of great significance for market regulators to improve their identification efficiency and accuracy on illegal insider trading identification.

The remainder of this paper is organized as follows. Section 2 describes the background of XGboost and NSGA-II that was used in this study. The model structure and procedures of the proposed method are described in Section 3. Section 4 explains the experimental design for insider trading identification in detail. Section 5 reports the experimental results and discussions. Lastly, Section 6 summarizes this research and provides some future research directions.

2. Background

2.1. XGboost

XGboost is a tree-based boosting machine learning algorithm that was proposed and developed by Chen and He [42]. It is an improved algorithm on the basis of the gradient boosting decision proposed by Friedman [43,44]. The XGboost can efficiently construct boosted trees, and it can address both classification and regression problems fast and accurately with parallel tree boosting. In this research, we adopted the XGboost algorithm for insider trading identification due to its superior performances in the machine learning competition Kaggle that was held in 2015.

In the training step of XGboost, gradient boosting combines weak learning models into a stronger learner in an iterative fashion [43]. At each iteration, the residual is used to the correct the previous predictor so the specified loss function could be optimized. The core algorithm of XGboost is the value optimization of the objective function. In the XGboost algorithm, an objective function consists of training loss and regularization:

$$\mathrm{O}(\theta) = \mathrm{L}(\theta) + \Omega(\theta) \tag{1}$$

where L is the training loss function and Ω is the regularization term. The XGboost employs the training loss to evaluate the model performance on training samples. In addition, the regularization term is implemented to control the complexity of the model. There are many ways to define the complexity, and in this research, the regularization term Ω for a decision tree is defined as:

$$\Omega(f) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{\mathrm{T}} {w_j}^2 \tag{2}$$

where T refers to the number of leaves in a decision tree; w is the vector of scores on leaves; γ denotes the complexity of each leaf; and λ is a parameter to scale the penalty.

Then, the objective function for calculating the structure score of XGboost can be finally derived as follows:

$$\mathrm{O} = \sum_{j=1}^{\mathrm{T}} \left[g_j w_j + \frac{1}{2}\left(h_j + \lambda\right) {w_j}^2 \right] + \gamma \mathrm{T} \tag{3}$$

where w_j are independent with each other. The form $g_j w_j + \frac{1}{2}\left(h_j + \lambda\right) w_j^2$ is quadratic and the best w_j for a given structure is $q(x)$.

2.2. NSGA-II

Non-dominated sorting genetic algorithm II (NSGA-II) [45] is an evolutionary algorithm that combines the genetic algorithm (GA) and the concept of non-dominance, which is introduced by Goldberg [46]. It has a lot of variants by using different operators for crossover, evolution, and mutation [47]. The original NSGA-II has been widely applied in many applications for addressing multi-objective optimization problems. The elitist non-dominated sorting method is the main feature of NSGA-II, and crowding distance is also incorporated to achieve the spread on the Pareto front. The main procedures of NSGA-II algorithm are as follows:

Step (1) A population $\mathrm{P_t}$ of size N is randomly initialized.

Step (2) Using the genetic operators of tournament selection, crossover, and mutation to generate an offspring population $\mathrm{Q_t}$.

Step (3) Merging the parent population $\mathrm{P_t}$ and offspring population $\mathrm{Q_t}$ to create a population $\mathrm{R_t}$ of size $2N$.

Step (4) Sorting the combined population $\mathrm{R_t}$ with crowding distance to obtain different non-dominated fronts $\mathrm{F_i}$; Selecting individuals by order within the population size.

Step (5) Generating new population P_{t+1} of size N, and the non-dominated fronts F_i are included until the new population P_{t+1} is filled.

Step (6) Repeating Step 2 to Step 5 until the maximum iteration.

In the NSGA-II learning period, chromosomes represent the possible solutions for solving the problem, and crowding distances are computed and utilized to maintain population diversity in each non-dominated front. Then, all individuals in the populations are ranked, and elitism is guaranteed by combining parent population with offspring population and is selected by the crowding distance-based non-dominated sorting. By using crowding distance, individuals in the NSGA-II population can be sorted with the crowed operator [45] of non-dominated rank and crowding distance. The order of individual m is before individual n if the m_{rank} is before the n_{rank}, or $m_{distance}$ is higher than $n_{distance}$ when m_{rank} is identical with the n_{rank}, where the rank is the non-dominant rank and distance represents the crowding distance.

3. Proposed Approach

In this section, the whole structure of the proposed approach and working procedures will be first explained. Next, a list of up to sixteen indicators for insider trading is described.

3.1. Structure of the Proposed Approach

Structure of the proposed approach is shown in Figure 1. The main procedures of the proposed approach for insider trading identification and non-insider trading classification are as follows:

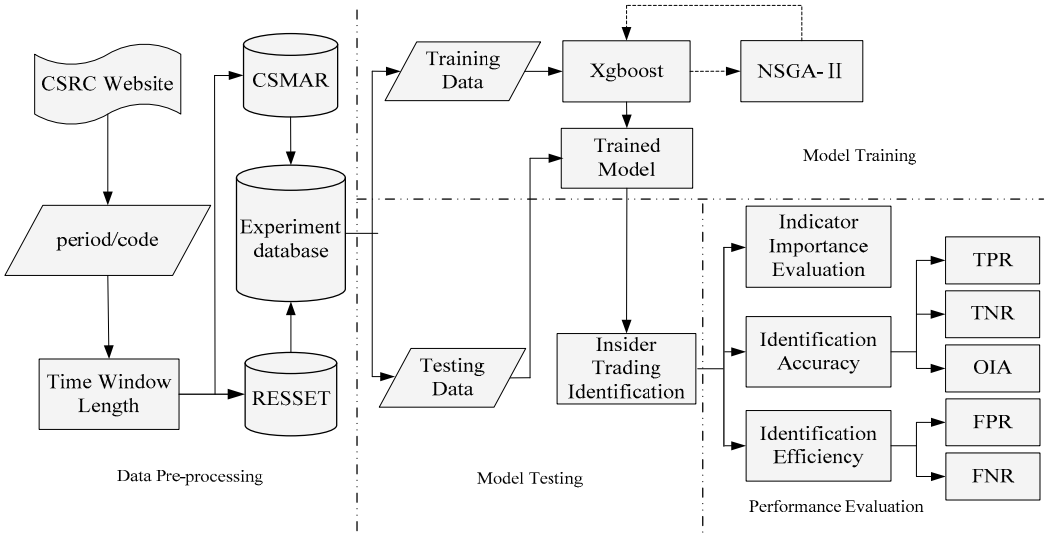

Figure 1. Structure of the proposed approach for insider trading identification and non-insider trading classification.

Step 1. Research data pre-processing

In the first step, an automatic crawling program was developed by us to derive the following information about insider trading cases from the official website of CSRC [4]: (1) Corporation name and security code that of the insider trading samples occurred from 2007 to 2018; (2) The period of the insider trading cases; (3) The disclosure time of the insider information. In addition, we also selected the non-insider trading cases in accordance with the insider trading samples, and the number ratio of insider and non-insider trading is 1:1. Next, the selected indicators of each data sample were calculated and derived from two relevant databases, which are CSMAR [48] and RESSET [49]. The relevant indicator values were calculated by using three different time window lengths, which were 30-days, 60-days, and 90-days.

Step 2. Model training

In the second step, XGboost was adopted to train the insider trading identification model by using the training samples. In the meanwhile, the NSGA-II was applied by the proposed method to optimize the parameters of XGboost. The fitness functions of the NSGA-II were designed from the aspects of identification accuracy and efficiency.

Step 3. Insider trading identification

Through the optimization of XGboost model by the NSGA-II algorithm, insider trading cases and related non-insider trading cases in the testing samples were used to be classified by the trained XGboost model.

Step 4. Identification performance evaluation

Experimental results were evaluated by using identification accuracy (TNR and TPR) and efficiency (FPR and FNR). Additionally, the importance of each indicator was also analyzed and compared for discovering the most essential indicators for insider trading identification.

3.2. Identification Indicators

After the insider trading cases were collected automatically from the website of CSRC, the proposed approach employed a total of sixteen relevant indicators for insider trading identification. These indicators could be roughly divided into the following classes:

1. Stock market performance, including the excess return compared with same market (ERCSM), beta coefficient, sigma coefficient, etc.
2. Financial performance: such as the current ratio and debt ratio.
3. Share ownership structure and corporate governance, including the H5 index, Z index, etc.

A list of the sixteen indicators is displayed in Table 1 to show the selected variables for insider trading identification. Descriptions in detail for these indicators are in the Appendix A.

Table 1. A list of relevant indicators for insider trading identification.

Insider Trading Identification	Indicators
Market performance of the stock	excess return compared with same market (ERCSM); beta coefficient, sigma coefficient; floating stock turnover rate (FSTR); volatility
Financial performance	return on assets (ROA); debt ratio (DR); total asset growth rate (TAGR); Price-earning ratio (P/E ratio); revenue growth rate (RGR); quick ratio (QR); current ratio (CR)
Company ownership structure and governance	H5 index; CR5 index; Z index; attendance ratio of the shareholders at the annual general meeting (ARAGM)

4. Experiment Design

4.1. Data

For the experiments, a total of 160 insider trading cases that occurred from the year 2007 to 2018 were derived from the punishment announcements that were located on the website of CSRC [4]. In addition, the same number of non-insider trading samples were chosen based on the following three criteria: (1) They also had significant information to disclose during or near to the insider trading period as the illegal insider trading cases; (2) These list companies belong to the same industry as the insider trading cases; (3) There was no insider trading activity in the history of the companies for non-insider trading cases. We set those criteria for non-insider trading samples selection due to the following two reasons: (1) To avoid selecting the samples that have insider trading activities but were not found by market regulators. (2) To distinguish insider and non-insider trading samples under

the conditions being as similar as possible. Note that in the whole dataset, the ratio of illegal insider trading samples number and non-insider trading samples number was 1:1. To create the training and testing datasets, the number ratio of training samples and testing samples was set to be about 4:1. We use sampling without replacement, thus each sample in the original dataset will be distributed into the training dataset with a possibility of 0.8, and the possibility for belonging to the testing dataset is 0.2.

4.2. XGboost Parameters for Optimization

Since the selection of initial parameters has a significant impact on the identification performance of XGboost [28], we adopted the NSGA-II to optimize them for obtaining a superior model for insider trading identification. Another reason for why we decided to adopt NSGA-II instead of GA is that other than the identification accuracy, we regarded the identification efficiency as an extremely essential criteria for market regulators, whereas identification efficiency and accuracy are possible to be conflicting objectives. GA is generally applied for single-objective optimization problems, thus the NSGA-II, which is a well-known and a popular multi-objective optimization method, was adopted in our proposed method. There are, in total, seven initial parameters to be optimized by NSGA-II: eta, max delta step, gamma, min child weight, colsample by tree, colsample by level, and colsample by node. A brief description of the XGboost parameters is provided in Table 2.

Table 2. A list of XGboost parameters that optimized by the NSGA-II algorithm.

No.	Parameters	Description
1	eta	After each boosting step, eta shrinks the feature weights to make the boosting process more conservative
2	max delta step	If the value is set to 0, it means there is no constraint. If it is set to a positive value, it can help to make the update step more conservative
3	gamma	The minimum loss reduction is required to make a further partition on a leaf node of the tree. The larger the gamma is, the more conservative of the XGboost algorithm
4	min child weight	It is the minimum sum of instance weight needed in a child. If the tree partition step results in a leaf node with the sum of instance weight less than this parameter value, then there will no further partition in the building process
5	colsample by tree	It is the subsample ratio of columns when constructing each tree. Subsampling occurs once for every tree constructed
6	colsample by level	It is the subsample ratio of columns for each level. Subsampling occurs once for every new depth level reached in a tree. Columns are subsampled from the set of columns chosen for the current tree
7	colsample by node	It is the subsample ratio of columns for each node (split). Subsampling occurs once every time a new split is evaluated. Columns are subsampled from the set of columns chosen for the current level

4.3. Evaluation Measures and Multi-Objective Functions

In this research, we employed a total of five indicators as evaluation measures, in which three were employed to gauge the performance of identification accuracy while the other two were calculated to gauge the performance of identification efficiency. Values of the true negative rate (TNR), true positive

rate (TPR), false positive rate (FPR), false negative rate (FNR), and overall identification accuracy (OIA) are calculated and used to measure the performances of insider trading identification system.

1. TNR is the ratio that the samples belong to insider trading cases are correctly identified. The calculation formula is:
$$TNR = \frac{TN}{TN + FP} \tag{4}$$

2. TPR is employed to measure the ratio that samples of non-insider trading cases are rightly classified. The calculation is:
$$TPR = \frac{TP}{TP + FN} \tag{5}$$

3. OIA is used to measure the ratio that the non-insider trading or insider trading samples are properly identified. It is calculated as:
$$OIA = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

4. FPR is a ratio that samples do not have insider trading activities that are incorrectly identified as insider trading samples. It is calculated by:
$$FPR = \frac{FP}{TN + FP} \tag{7}$$

5. FNR is employed to evaluate the ratio that insider trading samples are wrongly classified as non-insider trading samples. The calculation formula is:
$$FNR = \frac{FN}{TP + FN} \tag{8}$$

In the above Equations (4)–(8), true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*) refer to the number of right identifications for non-insider trading samples, number of correct identifications for insider trading samples, number of incorrect identifications for non-insider trading samples, and number of wrong identifications for insider trading samples, respectively. The following Table 3 displays a list of the five evaluation measures for gauging the identification performances.

Table 3. A list of the five evaluation measures for gauging the identification performances.

No	Evaluation Criteria	Calculation Formula
1	true negative rate (TNR)	$TNR = \frac{TN}{TN+FP}$
2	true positive rate (TPR)	$TPR = \frac{TP}{TP+FN}$
3	false positive rate (FPR)	$FPR = \frac{FP}{TN+FP}$
4	false negative rate (FNR)	$FNR = \frac{FN}{TP+FN}$
5	overall identification accuracy (OIA)	$OIA = \frac{TP+TN}{TP+FP+TN+FN}$

4.4. Benchmark Methods

Table 4 displays a list of the proposed approach and benchmark methods with their brief descriptions. Among the benchmark methods, many well-known and classic machine learning approaches such as ANN, SVM, Adaboost, and XGboost are included. Additionally, the performance of the RF-based method is also used to compare with our proposed approach. In addition, method 7 (XGboost-NSGA-II) is our proposed method for identification of insider trading, which adopts the use of XGboost for classification model and the NSGA-II is employed to optimize the parameters of

XGboost. Method 6 (XGboost-GA) adopts the XGboost for identification of insider trading without parameter optimization. Similar to the proposed method, GA is adopted for parameter optimization in XGboost-GA. However, since GA is a single objective optimization method, we employed the maximization of accuracy evaluation measure TNR as the fitness function. Performance evaluation of the methods 5 and 6 is utilized to determine whether the parameter optimization of XGboost can enhance the identification accuracy or efficiency. While the comparison of methods 6 and 7 is utilized to observe whether the results of multi-objective optimization (both identification accuracy and efficiency) could outperform the single objective optimization (only identification accuracy). Method 4 adopts an RF-based algorithm for identification of insider trading because it is also a well-known and an effective tree-based method. Moreover, methods 1–3 employ the ANN, SVM, and Adaboost based methods, respectively. They belong to the very classic and well-known machine learning algorithms, and they are utilized to compare their identification performances with XGboost and RF that belong to the tree-based method. Note that we calculate the error of each parameter combination by grid search algorithm and 10-fold cross-validation for the benchmark methods 1–3, the optimal parameter combination with the smallest error was chosen.

Table 4. A list of benchmark methods and their descriptions.

No	Method Name	Description
1	ANN	Identification model based on an ANN-based method
2	SVM	Identification model based on an SVM-based method
3	Adaboost	An Adaboost based approach for classification of illegal insider trading
4	RF	A random forest based approach for classification of illegal insider trading
5	XGboost	An XGboost based approach for identification of illegal insider trading
6	XGboost-GA	An XGboost based approach for identification of illegal insider trading. GA is adopted for initial parameter optimization of XGboost, and the fitness function is set to be the maximization of TPR
7	XGboost-NSGA-II	XGboost based classification approach for identification of insider trading. The NSGA-II is adopted for initial parameter optimization of XGboost. The fitness functions are designed to be the maximization of TPR and minimization of FNR

5. Experimental Results

5.1. Identification Accuracy Results

To measure the insider trading identification accuracy of the proposed method XGboost-NSGA-II, the methods ANN, SVM, Adaboost, RF, XGboost, and XGboost-GA were chosen as the benchmark methods, and the experiments were conducted under the time window lengths of 30-, 60-, and 90-days. The following Tables 5 and 6 report the TNR and TPR result, respectively to show insider trading identification accuracy.

Table 5. True negative rate (TNR) results of insider trading identification under the time window length of 30-, 60- and 90-days.

Window Length	ANN (%)	SVM (%)	Adaboost (%)	RF (%)	XGboost (%)	XGboost-GA (%)	XGboost-NSGA-II (%)
30-days	51.11	72.97	75	75.68	81.25	80.65	86.49
60-days	80	74.07	75	76.92	82.61	82.14	81.82
90-days	66.67	69.70	76.74	70.97	77.42	76.92	82.76
Average	65.93	72.25	75.58	74.52	80.43	79.90	83.69

Table 6. True positive rate (TPR) results of insider trading identification under the time window length of 30-, 60- and 90-days.

Window Length	ANN (%)	SVM (%)	Adaboost (%)	RF (%)	XGboost (%)	XGboost-GA (%)	XGboost-NSGA-II (%)
30-days	95.65	74.19	70.59	78.12	75	81.82	81.25
60-days	55.81	79.41	76.47	78.12	72.41	80	86.67
90-days	90.91	82.14	73.08	83.33	86.67	88.89	91.67
Average	80.79	78.58	73.38	79.86	78.03	83.57	86.53

Firstly, a higher TNR result, in general, represents superior identification accuracy for insider trading samples. From the TNR results shown in Table 5, we can observe that the average TNR result over the three time window lengths that were produced by the ANN, SVM, Adaboost, RF, XGboost, and XGboost-GA were 65.93%, 72.25%, 75.58%, 74.52%, 80.43%, and 79.90%, respectively, while our proposed method XGboost-NSGA-II yielded the largest average TNR of 83.69%. In addition, the XGboost-based method performed better than the other benchmark methods, which indicates that XGboost was superior to other traditional machine learning methods for identifying insider trading activities. This result is consistent with the findings in the work of Nishio et al. [26] that Xgboost was generally superior to the outstanding machine learning algorithms such as SVM. It is also observed by us that the average TNR results of XGboost (80.43%) and XGboost-GA (79.90%) were extremely close to each other. Whereas, the proposed method performed the best under the 30-days and 90-days time window lengths, and the average TNR result of the proposed approach (83.69%) was better than that of XGboost and XGboost-GA, which demonstrates that the adoption of the multi-objective optimization was beneficial for enhancing the identification accuracy on insider trading cases.

Secondly, similar to the TNR results, a larger TPR result indicates a better classification accuracy of the non-insider trading samples. From Table 6, we can find that the average TPR result of the ANN, SVM, Adaboost, RF, XGboost, and XGboost-GA was 80.79%, 78.58%, 73.38%, 79.86%, 78.03%, and 83.57%, respectively, while the best TPR result was still produced by the proposed method XGboost-NSGA-II with an average TPR of 86.53%. Those results demonstrate that the proposed method was substantially superior to benchmark methods in terms of the TNR and TPR. Although the ANN-based method produced the best TPR result under the 30-days time window length and a TPR result larger than 90% under the 90-days time window length, it obtained an extremely bad result under 60-days time window length, which indicates that it was not sufficiently powerful for insider trading identification. From Tables 5 and 6, we can observe that SVM outperformed ANN in average TNR results whereas ANN generated better average TPR results than SVM. This result is consistent with the findings of [50] that in some cases SVM performs better classification and in some cases ANN performs better than SVM. In addition, the XGboost-GA outperformed the performance of XGboost, indicating that the parameter optimization conducted by GA helped enhancing the classification accuracy of non-insider trading samples. Although the XGboost-based method performed slightly better than the proposed method under 30-days time window length, the average TPR result of the proposed method was superior to XGboost-GA-II, which demonstrates that the adoption of NSGA-II was better than GA adoption for producing superior classification accuracy on the non-insider trading samples.

Finally, we compared the overall identification accuracy of both the insider and non-insider trading samples for the proposed method and benchmark methods that are reported in Table 7. It is observed that under all three time window lengths, the proposed method XGboost-NSGA-II always produced the best OIA results, and subsequently it outperformed other benchmark methods in terms of average OIA results. In addition, we found that XGboost-GA obtained better results than XGboost under all three time window lengths, indicating that the optimization of initial parameters enhanced the overall identification accuracy. While the proposed method XGboost-NSGA-II outperformed, XGboost-GA demonstrates that multi-objective optimization successfully enhanced the overall identification performances. This result is consistent with the findings in other literatures [51,52] that the NSGA-II performed better than simple GA when multi-objectives were considered.

Table 7. Overall identification accuracy results of insider trading identification under the time window length of 30-, 60- and 90-days.

Window Length	ANN (%)	SVM (%)	Adaboost (%)	RF (%)	XGboost (%)	XGboost-GA (%)	XGboost-NSGA-II (%)
30-days	66.18	73.52	73.17	76.81	78.13	81.25	84.06
60-days	66.67	77.05	75.71	77.59	76.92	81.03	84.13
90-days	75.86	75.41	75.36	77.05	81.97	83.02	86.79
Average	69.57	75.33	74.75	77.15	79.01	81.77	84.99

To examine whether the overall identification accuracy improvement of the proposed XGboost-NSGA-II model is statistically significant, the Friedman test [53] is conducted on the overall identification results. In Friedman test, the calculation of statistic F is as follows:

$$F = \frac{12b}{a(a+1)}\left[\sum_{i=1}^{a} R_i^2 - \frac{a(a+1)^2}{4}\right] \quad (9)$$

where a represents the number of compared models; b represents the total number of identification results; R_i represents the average rank sum received from each identification accuracy for each model. The null hypothesis for Friedman's test on insider trading identification results is the equality of overall identification accuracy results among all methods, and the alternative hypothesis is defined as the negation of the null hypothesis. The Friedman test results at the 5% significance level are shown in Table 8. It could be found that the proposed method XGboost-NSGA-II is significantly better than other compared methods in terms of the overall identification accuracy.

Table 8. Friedman test on overall identification accuracy results for XGboost-NSGA-II against benchmark methods.

Compared Models	Significant Level α= 0.05
Overall Identification Accuracy XGboost-NSGA-II vs. ANN XGboost-NSGA-II vs. SVM XGboost-NSGA-II vs. Adaboost XGboost-NSGA-II vs. RF XGboost-NSGA-II vs. XGboost XGboost-NSGA-II vs. XGboost-GA	H_0: $n_1 = n_2 = n_3 = n_4 = n_5 = n_6 = n_7$ F = 16.143 p = 0.013 (reject H_0)

5.2. Identification Efficiency Results

Similar to evaluating the accuracy of insider trading identification for the proposed approach XGboost-NSGA-II and benchmark methods, the identification efficiency under the time window lengths of 30-, 60-, and 90-days were utilized for performance comparison. The following Tables 9 and 10

display the FPR and FNR result, respectively. We then compared the insider trading identification efficiency of the proposed method and benchmark methods.

Table 9. False positive rate (FPR) results of insider trading identification under the time window length of 30-, 60- and 90-days.

Window Length	ANN (%)	SVM (%)	Adaboost (%)	RF (%)	XGboost (%)	XGboost-GA (%)	XGboost-NSGA-II (%)
30-days	48.89	27.03	25	24.32	18.75	19.35	13.51
60-days	20	25.93	25	23.08	17.39	17.86	18.18
90-days	33.33	30.30	23.26	29.03	22.58	23.08	17.24
Average	34.07	27.75	24.42	25.48	19.57	20.10	16.31

Table 10. False negative rate (FNR) results of insider trading identification under the time window length of 30-, 60- and 90-days.

Window Length	ANN (%)	SVM (%)	Adaboost (%)	RF (%)	XGboost (%)	XGboost-GA (%)	XGboost-NSGA-II (%)
30-days	4.35	25.81	29.41	21.88	25	18.18	18.75
60-days	44.19	20.59	23.53	21.88	27.59	20	13.33
90-days	9.09	17.86	26.92	16.67	13.33	11.11	8.33
Average	19.21	21.42	26.62	20.14	21.97	16.43	13.47

In general, a smaller FPR result indicates a better identification efficiency of non-insider trading samples. For FPR results displayed in Table 9, we can observe that the average FPR under three time window lengths that produced by the ANN, SVM, Adaboost, RF, XGboost, and XGboost-GA were 34.07%, 27.75%, 24.42%, 25.48%, 19.57%, and 20.10%, respectively, while our proposed method XGboost-NSGA-II obtained the smallest average TNR of 16.31%. Additionally, for the FNR results which represent the identification error rate of insider trading samples, a smaller FNR result indicates a better identification efficiency of insider trading samples. From Table 10 we find that the average FNR result of the ANN, SVM, Adaboost, RF, XGboost, and XGboost-GA were 19.21%, 21.42%, 26.62%, 20.14%, 21.97%, and 16.43%, respectively. The best FNR result was still produced by the proposed method XGboost-NSGA-II and it had an average FNR of 13.47%, which indicates that it could be used as an effective method for insider trading. In summary, the FPR and FNR results demonstrate that the proposed method was substantially superior to benchmark methods at accuracy and efficiency for insider and non-insider trading identifications.

5.3. Performance of Different Time Window Length

In the previous section, we found that the proposed method produced the best average identification results for insider trading and non-insider trading samples in terms of both accuracy and efficiency. Next, we focus on the OIA, TNR, TPR, FPR, and FNR results of the proposed method under different time window lengths. Table 11 and Figure 2 report the identification accuracy and efficiency results under the time window lengths of 30-, 60-, and 90 days. We can observe that the proposed approach produced the best OIA result under the 90-days time window length, TNR result under the 30-days time window length, and the best TPR result was produced under the 90-days time window length. It produced the best FPR result under the 30-days and the best FNR result under the 90-days time window length. Those experimental results could be extremely beneficial for market regulators since they can select an appropriate time window length based on their regulation needs. For instance, if they consider the identification accuracy of insider trading is more important, they can set the time window length to be around 30-days, or they could select the 90-days time window length if a better balance between the identification accuracy and efficiency is preferred.

Table 11. Insider and non-insider trading identification accuracy and efficiency results of the proposed method under three time window lengths.

Window Length	OIA (%)	TNR (%)	TPR (%)	FPR (%)	FNR (%)
30-day	84.06	86.49	81.25	13.51	18.75
60-day	84.13	81.82	86.67	18.18	13.33
90-day	86.79	82.76	91.67	17.24	8.33

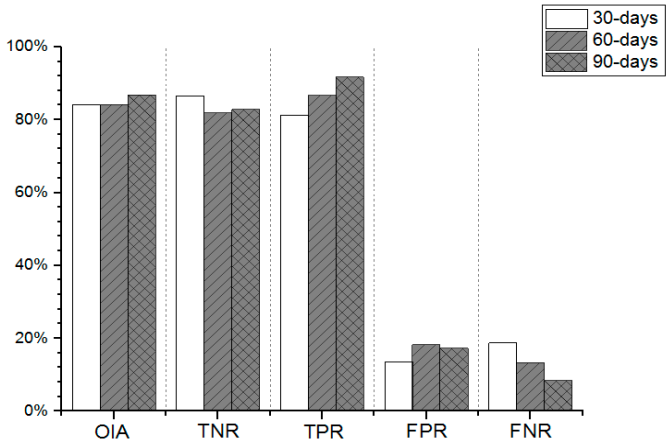

Figure 2. Accuracy and efficiency performance comparison of the proposed method under three different time window lengths (30-days, 60-days, and 90-days).

5.4. Importance of Indicators

Furthermore, we investigated which indicators that were employed were the most crucial ones for insider trading identification because it is possible to obtain the relative importance of the indicators by using XGboost. Figure 3 was plotted to display the relative importance results of relevant indicators calculated under three time window lengths. To find the most essential indicators, we focus on the indicators of the top five highest importance scores under each time window length. Under the 30-days time window length, the ERCSM, ROA, TAGR, H5 index, and DR obtained the highest five importance scores. While under the 60-days time window length, the top five indicators that obtained the highest scores were the ERSCM, H5 index, TAGR, ROA, and FSTR. When under the 90-days time window length, the five most influential indicators were ERCSM, H5 index, ROA, CR5, and the P/E ratio. We observe that ERCSM, ROA, and H5 index were always in the top five important indicators, which indicates that they could be considered as essential indicators for insider and non-insider trading classification. Additionally, ERCSM was always the most important indicator, and its relative importance proportions (17.92% under 30-days, 16.27% under 60-days, and 18.93% under 90-days time window length) were extremely larger than other indicators, demonstrating that excess rate of return over the average return of the market was the most significant indicator for market regulators to discover insider trading activities. This finding is consistent with previous literature [54,55] that abnormal returns before the insider trading information are predictive indicators for insider trading identification.

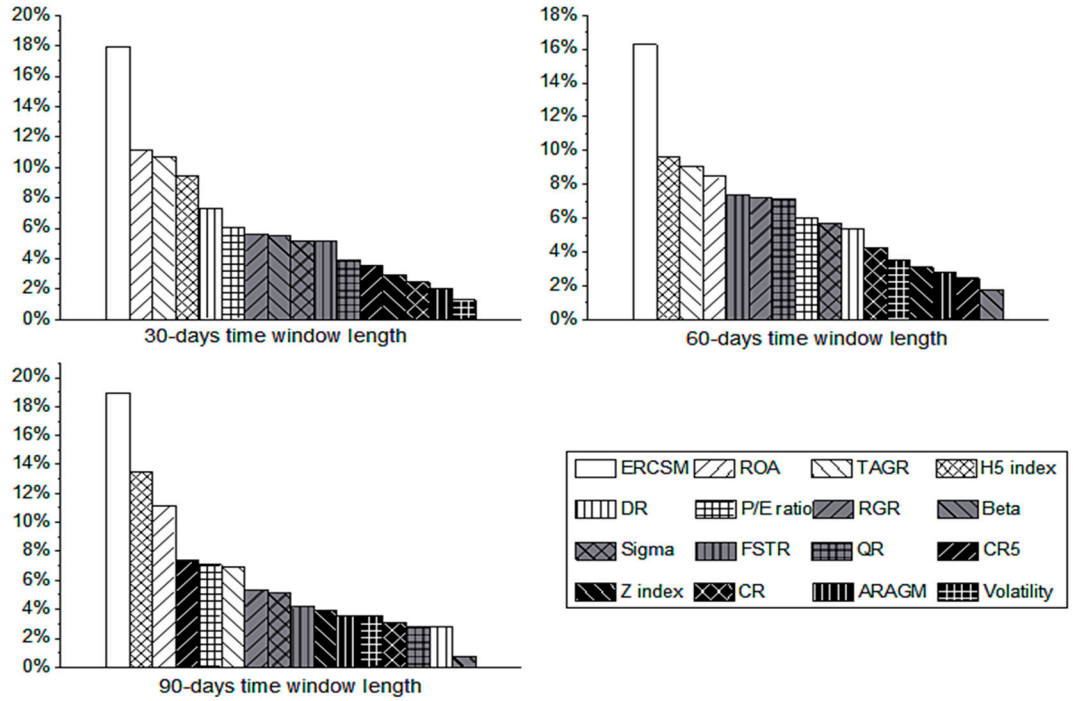

Figure 3. Relative importance scores of the selected relevant indicators for insider trading identification under 30-, 60-, and 90-days time window lengths.

6. Conclusions

In this research, a total of 160 illegal insider trading cases that were penalized by the CSRC from 2007 to 2018 were collected and identified. For identification of the insider trading cases, we constructed a combination approach of XGboost and NSGA-II, in which the XGboost was adopted for sample classification, while the initial parameters were optimized by NSGA-II algorithm. The proposed approach XGboost-NSGA-II identified the insider trading samples by employing features from the information of stock market performance, financial performance and corporate governance. Compared with the benchmark methods, our proposed method has the following advantages: (1) It can provide more robust and accurate results and it is not easy to be overfitting when compared with traditional machine learning methods such as SVM, ANN, Adaboost, and RF; (2) Compared with simple GA that is used for single objective optimization, our proposed method can solve the problem of multiple objective optimization for both identification accuracy and efficiency; (3) Our proposed method was superior to XGboost since our proposed method optimized the initial parameters of XGboost, which are considered to influence the performance of identification results considerable.

From the experimental results, some beneficial findings could be summarized: (1) The proposed XGboost-NSGA-II identification method not only performed well in terms of identification accuracy but also identified the insider trading more effectively than other benchmark methods; (2) Experimental results of the proposed approach XGboost-NSGA-II method outperformed the XGboost, which demonstrates that the parameters optimization by the NSGA-II was beneficial for enhancing the identification accuracy and efficiency; (3) The proposed approach produced the best identification accuracy results under the time window length of 30-days, demonstrating that those relevant indicators calculated with 30-days could be recognized as an alternative time window length for insider trading identification; (4) The indicator importance result of the proposed method reveals that ERCSM, ROA, and H5 index could be considered as the essential indicators for identifying the insider trading activities.

Although our research contributes to insider trading identification to some extent, this study also had some restrictions that could be investigated in the further research directions. For instance, we have investigated the performances of the proposed method under three different time window

lengths. Nonetheless, identification performances by using indicators that are calculated with other time window lengths, such as 10-day, 70-day or 100-days, could also be investigated for finding a more appropriate length to calculate related indicators. In this research, we have proposed an insider trading identification method and applied it in the Chinese security market. The effectiveness of the proposed method in security markets of other countries in developed regions (Europe, America, etc.) and developing regions (Thailand, Mexico, etc.) would be examined in the future work. Additionally, other researchers could apply the proposed method on related fields such as bankruptcy prediction [56] or design an intelligent system for money laundering identification [57].

Author Contributions: Conceptualization, S.D.; data curation, S.D., C.W., and J.L.; formal analysis, S.D. and C.W.; funding acquisition, S.D.; investigation, Y.C., T.Y., and Y.Z.; methodology, S.D., Y.C., and H.Y.; project administration, S.D. and Y.C.; resources, Y.Z. and H.T.; software, S.D. and F.M.; supervision, S.D. and Y.C.; validation, T.Y. and S.D.; writing—original draft, S.D. and C.W.; writing—review & editing, T.Y.

Funding: This work was funded by Hubei Provincial Department of Education, grant No. Q20171208; and the "Talent Excellence Program 2018" funded by Hubei Provincial Department of Education.

Acknowledgments: We are greatful to the anonymous reviewers for their comments and discussions.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

1. Excess return compared with same market (ERCSM):

   This indicator estimates the excess return over the security market return. It is calculated by:

$$ERCSM = Security\ Return - Market\ Return$$

2. Return on assets (ROA)

   The ROA is calculated to evaluate how much of the net income is yielded per unit of the total assets. It is calculated by:

$$ROA = Net\ Income / Total\ Assets$$

3. Total asset growth rate (TAGR)

   It is the ratio of the total asset growth in current year to the total assets at the start of current year, which reflects the asset growth ratio of the company in current year. It is calculated by:

$$TAGR = (Total\ Asset\ Growth / Total\ Assets) \times 100\%$$

4. H5 index

   The H5 index is the sum of squares of the largest five stockholders' share proportion. The closer of the H5 index to 1, the greater the share proportion difference between the largest five stockholders.

5. Debt ratio (DR)

   It is a ratio of company total debts and total assets. The DR is calculated as:

$$DR = (Total\ Debts / Total\ Assets) \times 100\%$$

6. Price-earning ratio (P/E ratio)

   It is a ratio of a company's stock price to the company's earnings per share. The P/E ratio is often employed in stock price valuation. The calculation formula is:

$$P/E = Company\ Stock\ Price / Earning\ Per\ Share$$

7. Revenue growth rate (RGR)

It is the rate of the company increased revenue to the total revenue in the previous year. It is calculated by:

$$RGR = (Increased\ Revenue / Total\ Revenue\ in\ the\ last\ year) \times 100\%$$

8. Beta coefficient

   A stocks beta coefficient is the ratio of the product of the covariance of the stock's returns and the benchmark's returns to the product of the variance of the benchmark's returns over a certain period.

9. Sigma coefficient

   The sigma coefficient is measured by using the standard deviation of a company's stock prices in a certain length of period.

10. Floating stock turnover rate (FSTR)

    The FSTR is generally used to evaluate the degree of the stock transfer frequency in a certain length of period. It is calculated as:

$$FSTR = (Stock\ Trading\ Volume / Floating\ Stocks\ in\ circulation) \times 100\%$$

11. Quick ratio (QR)

    It is the rate of a company's quick asset to its current liability. The calculation formula is:

$$QR = (Quick\ Asset / Current\ Liability) \times 100\%$$

12. CR5 Index

    The CR5 index is the total stock proportion of the largest five shareholders.

13. Z index

    It is the ratio of the largest shareholder's stock amount and the second-largest shareholder's stock amount.

14. Current Ratio (CR)

    The CR is the ratio of a company's current assets to its current liabilities. It is often used to evaluate whether a company has enough current assets to meet its short-term obligations.

$$CR = (Current\ Asset / Current\ Liability) \times 100\%$$

15. Attendance ratio of the shareholders at the annual general meeting (ARAGM)

    The ARAGM is a ratio of what percentage of the company's shareholders are attending at the annual general meeting.

16. Volatility

    It is the degree of stock price variation of a company's stock prices at a certain length of period that evaluated by standard deviation of logarithmic return.

References

1. Cheung, Y.L.; Jiang, P.; Limpaphayom, P.; Lu, T. Does corporate governance matter in china? *China Econ. Rev.* **2008**, *19*, 460–479. [CrossRef]
2. Howson, N.C. Enforcement without Foundation?—Insider Trading and China's Administrative Law Crisis. *Am. J. Comp. Law* **2012**, *60*, 955–1002. [CrossRef]

3. Meulbroek, L.K.; Hart, C. The Effect of Illegal Insider Trading on Takeover Premia. *Rev. Financ.* **2015**, *1*, 51–80. [CrossRef]
4. Website of CSRC. Available online: http://www.csrc.gov.cn/pub/newsite/ (accessed on 1 October 2019).
5. Islam, S.R.; Ghafoor, S.K.; Eberle, W. Mining Illegal Insider Trading of Stocks: A Proactive Approach. In Proceedings of the IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018.
6. Zhang, G.; Patuwo, B.E.; Hu, M.Y. Forecasting with artificial neural networks: The state of the art. *Int. J. Forecast.* **1998**, *14*, 35–62. [CrossRef]
7. Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
8. Collins, M.; Schapire, R.E.; Singer, Y. Logistic Regression, AdaBoost and Bregman Distances. *Mach. Learn.* **2002**, *48*, 253–285. [CrossRef]
9. Farooq, M.; Zheng, H.; Nagabhushana, A.; Roy, S.; Burkett, S.; Barkey, M.; Kotru, S.; Sazonov, E. Damage Detection and Identification in Smart Structures using SVM and ANN. In Proceedings of the Smart Sensor Phenomena, Technology, Networks, & Systems Integration, San Diego, CA, USA, 30 March 2012; Volume 8346, p. 40.
10. Li, Z.X.; Yang, X.M. Damage identification for beams using ANN based on statistical property of structural responses. *Comput. Struct.* **2008**, *86*, 64–71. [CrossRef]
11. Stoica, M.; Calangiu, G.A.; Sisak, F.; Sarkany, I. A method proposed for training an artificial neural network used for industrial robot programming by demonstration. In Proceedings of the International Conference on Optimization of Electrical & Electronic Equipment, Basov, Romania, 20–22 May 2010.
12. Das, A.B.; Bhuiyan, M.I.H.; Alam, S.M.S. A statistical method for automatic detection of seizure and epilepsy in the dual tree complex wavelet transform domain. In Proceedings of the International Conference on Informatics, Dhaka, Bangladesh, 23–24 May 2014.
13. ÇiMen, M.; KiSi, O. Comparison of two different data-driven techniques in modeling lake level fluctuations in Turkey. *J. Hydrol.* **2009**, *378*, 253–262. [CrossRef]
14. Sun, H.; Xie, L. Recognition of a Sucker Rod's Defect with ANN and SVM. In Proceedings of the International Joint Conference on Computational Sciences and Optimization, Sanya, China, 24–26 April 2009.
15. Rodriguez-Martin, D.; Samà, A.; Perez-Lopez, C.; Català, A.; Cabestany, J.; Rodriguez-Molinero, A. SVM-based posture identification with a single waist-located triaxial accelerometer. *Expert Syst. Appl.* **2013**, *40*, 7203–7211. [CrossRef]
16. Jiang, H.; Tang, F.; Zhang, X. Liver cancer identification based on PSO-SVM model. In Proceedings of the International Conference on Control Automation Robotics & Vision, Singapore, 7–10 December 2011.
17. Amiri, S.; Rosen, D.V.; Zwanzig, S. The SVM approach for Box–Jenkins Models. *Revstat-Stat. J.* **2011**, *7*, 23–36.
18. Liu, M. Fingerprint classification based on Adaboost learning from singularity features. *Pattern Recogn.* **2010**, *43*, 1062–1070. [CrossRef]
19. Kim, D.; Philen, M. Damage classification using Adaboost machine learning for structural health monitoring. In Proceedings of the Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems, San Diego, CA, USA, 14 April 2011.
20. Gutiérreztobal, G.C.; Álvarez, D.; Gómezpilar, J.; Campo, F.D.; Hornero, R. AdaBoost Classification to Detect Sleep Apnea from Airflow Recordings. In *XIII Mediterranean Conference on Medical & Biological Engineering & Computing*; Romero, L.M., Ed.; Springer: Cham, Switzerland, 2013; pp. 1829–1832.
21. Liu, X.; Dai, Y.; Zhang, Y.; Yuan, Q.; Zhao, L. A preprocessing method of AdaBoost for mislabeled data classification. In Proceedings of the 29th Chinese Control and Decision Conference (CCDC), Chongqing, China, 28–30 May 2017.
22. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
23. Deng, S.; Wang, C.; Cao, C.; Fan, Y. Identification of Insider Trading in Security Market Based on Random Forests. *J. China Three Gorges Univ. Humanit. Soc. Sci.* **2019**, *41*, 70–75. (In Chinese) [CrossRef]
24. Murugan, A.; Nair, S.A.H.; Kumar, K.P.S. Detection of Skin Cancer Using SVM, Random Forest and KNN Classifiers. *J. Med. Syst.* **2019**, *43*, 269. [CrossRef] [PubMed]
25. Choi, D.K. Data-Driven Materials Modeling with XGBoost Algorithm and Statistical Inference Analysis for Prediction of Fatigue Strength of Steels. *Int. J. Precis. Eng. Manuf.* **2019**, *20*, 129–138. [CrossRef]

26. Nishio, M.; Nishizawa, M.; Sugiyama, O.; Kojima, R.; Yakami, M.; Kuroda, T.; Togashi, K. Computer-aided diagnosis of lung nodule using gradient tree boosting and Bayesian optimization. *PLoS ONE* **2017**, *13*, e0195875. [CrossRef]
27. Mustapha, I.B.; Saeed, F. Bioactive Molecule Prediction Using Extreme Gradient Boosting. *Molecules* **2016**, *21*, 983. [CrossRef]
28. Li, Y.Z.; Wang, Z.Y.; Zhou, Y.L.; Han, X.Z. The Improvement and Application of Xgboost Method Based on the Bayesian Optimization. *J. Guangdong Univ. Technol.* **2018**, *35*, 23–28. (In Chinese)
29. Tamimi, A.; Naidu, D.S.; Kavianpour, S. An Intrusion Detection System Based on NSGA-II Algorithm. In Proceedings of the Fourth International Conference on Cyber Security, Cyber Warfare, and Digital Forensic (CyberSec), Jakarta, Indonesia, 29–31 October 2016.
30. Lin, J.F.; Xu, Y.L.; Law, S.S. Structural damage detection-oriented multi-type sensor placement with multi-objective optimization. *J. Sound Vib.* **2018**, *422*, 568–589. [CrossRef]
31. Guan, X.Z. Multi-objective PID Controller Based on NSGA-II Algorithm with Application to Main Steam Temperature Control. In Proceedings of the International Conference on Artificial Intelligence & Computational Intelligence, Shanghai, China, 7–8 November 2009.
32. Li, C.; Liu, C.; Yang, L.; He, L.; Wu, T. Particle Swarm Optimization for Positioning the Coil of Transcranial Magnetic Stimulation. *BioMed Res. Int.* **2019**, *2019*, 946101. [CrossRef]
33. Garg, H. A hybrid GSA-GA algorithm for constrained optimization problems. *Inf. Sci.* **2019**, *478*, 499–523. [CrossRef]
34. Garg, H. A hybrid PSO-GA algorithm for constrained optimization problems. *Appl. Math. Comput.* **2016**, *274*, 292–305. [CrossRef]
35. Alarifi, I.M.; Nguyen, H.M.; Bakhtiyari, A.N.; Asadi, A. Feasibility of ANFIS-PSO and ANFIS-GA Models in Predicting Thermophysical Properties of Al2O3-MWCNT/Oil Hybrid Nanofluid. *Materials* **2019**, *12*, 3628. [CrossRef] [PubMed]
36. Chiang, C.H.; Chung, S.G.; Louis, H. Insider trading, stock return volatility, and the option market's pricing of the information content of insider trading. *J. Bank. Financ.* **2017**, *76*, 65–73. [CrossRef]
37. Jain, N.; Mirman, L.J. Effects of insider trading under different market structures. *Q. Rev. Econ. Financ.* **2004**, *42*, 19–39. [CrossRef]
38. Jabbour, A.R.; Jalilvand, A.; Switzer, J.A. Pre-bid price run-ups and insider trading activity: Evidence from Canadian acquisitions. *Int. Rev. Financ. Anal.* **2004**, *9*, 21–43. [CrossRef]
39. Dai, L.; Fu, R.; Kang, J.K.; Lee, I. Corporate governance and insider trading. *SSRN Electron. J.* **2013**, *40*, 235–253. [CrossRef]
40. Chronopoulos, D.K.; McMillan, D.G.; Papadimitriou, F.I.; Tavakoli, M. Insider trading and future stock returns in firms with concentrated ownership levels. *Eur. J. Financ.* **2018**, *25*, 139–154. [CrossRef]
41. Lu, C.; Zhao, X.; Dai, J. Corporate Social Responsibility and Insider Trading: Evidence from China. *Sustainability* **2018**, *10*, 3163. [CrossRef]
42. Chen, T.; He, T. XGBoost: eXtreme Gradient Boosting, R package version 04-2, 2015. Available online: https://cran.r-project.org/src/contrib/Archive/xgboost/ (accessed on 24 November 2019).
43. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
44. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **2002**, *38*, 367–378. [CrossRef]
45. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197. [CrossRef]
46. Goldberg, D.E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, USA, 1989.
47. Nebro, A.J.; Durillo, J.J.; Machín, M.; Coello, C.A.C.; Dorronsoro, B.A.J.; Dorronsoro, B. A Study of the Combination of Variation Operators in the NSGA-II Algorithm. In *Advances in Artificial Intelligence*; Springer: Berlin, Germany, 2013; pp. 269–278.
48. CSMAR Database. Available online: http://www.gtafe.com/WebShow/ShowDataService/1 (accessed on 1 October 2019).
49. RESSET Database. Available online: http://www.resset.cn/databases (accessed on 1 October 2019).

50. Kalarani, P.; Brunda, S.S. An efficient approach for ensemble of SVM and ANN for sentiment classification. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24 October 2016.
51. Ramaswamy, P.C.; Deconinck, G. Smart grid reconfiguration using simple genetic algorithm and NSGA-II. In Proceedings of the IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe), Berlin, Germany, 14–17 October 2012.
52. Song, R.; Cui, M.; Liu, J. Single and multiple objective optimization of a natural gas liquefaction process. *Energy* **2017**, *124*, 19–28. [CrossRef]
53. Eisinga, R.; Heskes, T.; Pelzer, B.; Grotenhuis, M.T. Exact p-values for pairwise comparison of friedman rank sums, with application to comparing classifiers. *BMC Bioinform.* **2017**, *18*, 68. [CrossRef] [PubMed]
54. Meulbroek, L.K. An Empirical Analysis of Illegal Insider Trading. *J. Financ.* **1992**, *47*, 1661–1699. [CrossRef]
55. Reynolds, J. Insider trading activities around the world: A case study in East Asia. *Res. J. Financ. Account.* **2010**, *1*.
56. Klepáč, V.; Hampel, D. Prediction of Bankruptcy with SVM Classifiers Among Retail Business Companies in EU. *Acta Univ. Agric. Silvic. Mendel. Brun.* **2016**, *64*, 627–634. [CrossRef]
57. Liu, K.; Yu, T. An Improved Support-Vector Network Model for Anti-Money Laundering. In Proceedings of the Fifth International Conference on Management of E-commerce & E-government, Wuhan, China, 5–6 November 2011.