North American Journal of Economics and Finance 64 (2023) 101848

Contents lists available at ScienceDirect

North American Journal of Economics and Finance

journal homepage: www.elsevier.com/locate/najef

Stock index direction forecasting using an explainable eXtreme Gradient Boosting and investor sentiments

Shangkun Deng[a,*], Xiaoru Huang[a], Yingke Zhu[a], Zhihao Su[b], Zhe Fu[c], Tatsuro Shimada[d]

[a] College of Economics and Management, China Three Gorges University, Yichang 443002, China
[b] School of Economics, Shandong University, Jinan 250100, China
[c] School of History, Beijing Normal University, Beijing 100875, China
[d] Graduate School of Science and Technology, Keio University, Yokohama 2238522, Japan

ARTICLE INFO

Keywords:
Investor sentiment
Stock index
Direction prediction
Essential features
Explainable XGBoost

ABSTRACT

In this article, an explainable eXtreme Gradient Boosting (XGBoost) method is proposed for stock index prediction and trading simulation in the Chinese security market. Sentiment features of three types of investors, including institutional, individual, and foreign investors, are utilized as explanatory variables, and a binary classification model based on XGBoost is constructed to predict the direction of the Shanghai composite index and Shenzhen composite index movements. Additionally, the Gain function of XGBoost and SHapley Additive exPlanations (SHAP) are employed to estimate the importance of sentimental factors affecting index direction forecasting. Experimental results demonstrate that the XGBoost-based approach using multiple investor sentiments achieved the best forecasting accuracy, and sentiment features of the institutional investor were relatively more essential than the individual investor and foreign investor sentiments for index direction prediction in most out-of-sample periods. It demonstrates that the method proposed in this research can provide useful references for market participants to support their investment in the Chinese security market.

1. Introduction

The Chinese stock market is one of the emerging capital markets with an apparent speculative phenomenon, and the magnitude of the stock prices is susceptible to irrational factors, notably the behavior of market investors. As of the year 2020, the number of accounts opened by individual investors was still greater than 99 % of the overall investors who participated in the Chinese stock market, accounting for the overwhelming majority of market investors.[1] However, in recent years, with the continuous development of the Chinese capital market, the capital proportion of different types of investors in the Chinese stock market has changed substantially. Among the market participants, institutional and foreign investors have gradually become to be essential parts. Although the proportions of institutional and foreign investor accounts are still fewer than that of individual investors, their market capitalization share respectively reached 28.6 % and 8 % at the end of the year 2020, and their trading behavior in the Chinese security market could

* Corresponding author.
E-mail address: dsk8672@163.com (S. Deng).

[1] The data could be derived from the website of the Chinese Securities Depository and Clearing Corporation Limited (CSDC), which is https://www.chinaclear.cn/.

https://doi.org/10.1016/j.najef.2022.101848
Received 23 December 2021; Received in revised form 29 October 2022; Accepted 21 November 2022
Available online 24 November 2022

impact market returns to a certain extent (Han & Li, 2017; Gao et al., 2020).

Extant literature recognizes investor sentiment as a significant factor influencing stock markets, and several empirical pieces of research have been conducted to verify the impacts of investor sentiment on stock returns (Fisher & Statman, 2000; Chi et al., 2012). Looking forward to expanding the research on investor sentiment, numerous scholars have provided evidence that investor sentiment is capable of forecasting market returns. For instance, a previous study (Bollen et al., 2011) argued that investor sentiment could be exploited for stock market returns prediction by constructing an individual investor sentiment proxy. Dergiades (2012) also found that the investor sentiment index constructed from the US economic indicators showed significant predictive power to stock returns. Additionally, the effectiveness of lagged investor sentiment for stock market returns forecasting has been further demonstrated (Chung et al., 2012). Moreover, many researchers have found that institutional and foreign investors could also be considered as sentimental traders, and their participation and behavior in the market could influence the movement of stock returns (Baker et al., 2012; Ryu et al., 2017). It has been widely reported by a large number of empirical studies that market investor sentiments could be the predictors of stocks, whereas, most researchers have only utilized a single kind of investor sentiment in one research, e.g., only the individual investor sentiment data were employed to forecast the movement of the stock market. Therefore, the empirical research of the security market return forecasting from the perspective of market investors is of great practical significance. To address this gap, we attempt to expand the research on the return direction forecasting of the Chinese stock index by utilizing sentiment variables of three kinds of prominent investors, which includes the institutional, individual, and foreign investors.

In recent decades, many commonly applied models have provided scholars with beneficial tools for predicting time series in financial markets, such as machine learning techniques and traditional time series analysis approaches. In general, regarding the non-linear and non-stationary characteristics of financial time series, machine learning algorithms are considered capable of producing more excellent performance than traditional analysis methods in forecasting problems (Li & Tang, 2020). The classical machine learning techniques, including SVM (Support Vector Machine), RF (Random Forest), KNN (K-Nearest Neighbors), and ANN (Artificial Neutral Network), have been evidenced to generate outstanding performances in many application fields (Tanaka et al., 2016; Deng, Wang, et al., 2021; Fang & Taylor, 2021). Additionally, XGBoost (eXtreme Gradient Boosting), which is an advanced ensemble learning approach, is known to deliver comparable or better performance than traditional machine learning methods (Chen & Guestrin, 2016). Recently, in many applications, numerous scholars have proved that XGBoost outperforms other machine learning algorithms in terms of computing speed and prediction accuracy (Fan et al., 2018; Meng et al., 2020). Despite the fact that the computation speed of XGBoost is relatively slower than traditional linear approaches, such as OLS (Ordinary Least Squares), it is more robust and accurate. Indeed, in many empirical studies, XGBoost was proved to be robust to the over-learning problem, and it outperformed a lot of traditional machine learning approaches in terms of classification accuracy (Chen & Guestrin, 2016; Deng, Huang, et al., 2021; Kwon et al., 2022). Hence, XGBoost is implemented for the stock index forecasting task in the current research. Overall, five machine learning algorithms and one traditional regression method, including XGBoost, ANN, SVM, RF, KNN, and OLS, are designed as the baseline methods to implement the index direction prediction task.

As a stock index predictor, it should be noted that even if it could yield a high accuracy for index direction prediction, it might be unable to generate a profit in the case that it fails to yield more positive returns in correct predictions than the loss incurred in the incorrect direction predictions. Consequently, besides the prediction accuracy, another proper indicator of the index predictor accurateness should be the trading return. The transaction executed by a participant in stock markets is either short-selling or long, while for simulation trading by a model, the return depends on a trading strategy or rule, which should be designed based on the signals generated from the direction predictor. For the proposed approach, a simple trading strategy is constructed as follows: suppose an approach forecasts that the stock index will rise when the predicted label is “rising”, then a “long” trading will be executed. On the contrary, if the approach predicts the stock index will decline when the predicted label is “falling”, a transaction “short-selling” will be carried out. During trading simulation, the position holding period of each transaction is identical to the forecasting horizon for the investigated methods. Additionally, if the transaction performed at the current time point T is identical to the one performed at the previous time point T-1, the position of the previous day will be maintained. Otherwise, the position of the previous day is closed, and the model will open a new position. Furthermore, for results comparison and evaluation, two classical naïve trading strategies, Short-and-Hold (SAH) and Buy-and-Hold (BAH), are adopted as the benchmarks in the experiments of trading simulation.

Although machine learning algorithms have significant advantages in classification and regression problems, most of the existing machine learning approaches belong to the kind of “black-box” model. That is, machine learning models are used to generate prediction outputs using the input features, which can effectively facilitate prediction accuracy than the traditional linear method. Nonetheless, they are generally challenging to illustrate the decision basis of model prediction well, which makes the prediction results of machine learning models lack interpretability. This drawback limits the uses of machine learning approaches. In recent years, Lundberg and Lee (2017) proposed a model explanation approach called SHAP (SHapley Additive exPlanations) to further enhance the interpretability of machine learning approaches by measuring the exact contribution of each feature input to model output. SHAP considers each feature in the model as a contributor, calculates the contribution value of each feature, and sums it up to obtain the final prediction value for model interpretation. Since the XGBoost model employed in this study cannot explain the specific impact of each sentiment feature on the stock index direction prediction, the SHAP approach is incorporated in the XGBoost-based predictor to address the problem of model explainability for its direction predictions.

Focusing on the prediction and trading simulation of the stock index in the Chinese stock market, the rest of this article is structured as follows. The background of methods and the proposed approach is illustrated in Section 2. The empirical data, evaluation criteria, and benchmarks design are illustrated explicitly in Section 3. Section 4 presents the experimental results and related discussions. Section 5 summarizes the findings and provides several research directions to expand this research.

2. Methods

2.1. XGBoost

XGBoost was developed by Chen and Guestrin (2016), and it is one of the most pratical ensemble learning algorithms. It comprises many classifiers, which can be linearly fused to a relatively more robust classifier. XGBoost has many superiorities, such as large accuracy, strong generalization, and quick computation, make it an outstanding approach in machine learning and data mining (Li & Zhang, 2019).

Suppose the prediction result $\widehat{y}_i$ is the summary of the leaf scores from K weak learners generated by a boosting approach, which is given by Eq. (1):

$$\widehat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in F \tag{1}$$

where $f_k(.)$ denotes the score for the k-th weak learner, while F means the set of all decision trees. For constructing an XGBoost model, the primary work is to seek for the optimal parameters by minimizing the objective function. In Eq. (2), the objective function *Obj* of XGBoost comprises two parts: l is the term of the loss function, and Ω is the term of the model regularization.

$$Obj = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

On the right side of Eq. (2), The first item is utilized to measure the differences between the prediction $\widehat{y}_i$ and the target y_i, and the second item is employed to punish the complexity of the tree structure. The regularization term Ω is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

where γ represents the parameter of complexity, ω is the score of the leaf nodes, T denotes the total number of the leaf nodes, and λ is the regularization coefficient.

Specifically, the model is optimized in an additive manner, by adding a new $f_t(x_i)$ to improve the current model and set up a new loss function. Subsequently, we can express the objective function as:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

By carrying out a second-order Taylor expansion in gradient boosting, the loss function could be extended. To simplify the objective function, the constant term is removed. Hence, the final *Obj* is expressed in Eq. (5). For g_i and $h_{i,}$ which represent the gradient statistics of the loss function, are defined in Eq. (6).

$$\begin{aligned} Obj^{(t)} &\cong \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)w_j^2\right] + \gamma T \end{aligned} \tag{5}$$

$$g_i = \partial_{\widehat{y(t-1)}} l\left(y_i, \widehat{y}_i^{(t-1)}\right), h_i = \partial \frac{2}{y(t-1)} l\left(y_i, \widehat{y}_i^{(t-1)}\right) \tag{6}$$

Suppose that an instance set I_j in the leaf j is$I_j = \{iq(x_i) = j\}$. When a tree structure is given, the calculation way of the optimal leaf weight $w{*}j$ and the loss function at the leaf node j is respectively shown in Eqs. (7) and (8).

$$w_j^* = -\frac{\sum_{i\in I_j} g_j}{\sum_{i\in I_j} h_j + \lambda} \tag{7}$$

$$Obj^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{8}$$

After splitting, the information gain of leaf nodes can be recorded by Eq. (9). A greedy search algorithm is adopted to estimate the split candidates. Thus, Gain function of XGBoost can be applied as a measure of the importance of the features. Subsequently, Gain function of XGBoost is utilized for feature selection from the multiple sentiment features.

$$Gain = \frac{1}{2}\left(\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right) - \gamma \tag{9}$$

Furthermore, since the hyper-parameters selection of XGBoost can have a significant impact on its forecasting accuracy, we decide to pursue the use of an optimization method grid search to find the best hyper-parameter combination in the XGBoost during the model training process.

2.2. SHAP approach

To promote the interpretability of the XGBoost model, we incorporate the SHAP approach to gauge the influences of sentiment features on the forecasting results. The SHAP approach is a model interpretation framework proposed by Lundberg and Lee (2017). It originates from cooperative game theory (Shapley, 2016), in which the shapley value was designed to gauge the contribution of each player in the alliance to the coalition. For the aim of model interpretation by the SHAP approach, each input feature becomes the participant in the prediction outcome, and the SHAP value is a measure of the feature contribution to the model outcomes.

The SHAP value of each feature should meet the following three properties: 1) Local accuracy. It ensures that the sum of all feature contributions equals the prediction results; 2) Missingness. It guarantees that the absence of input features will not affect the model outcomes; 3) Consistency. It ensures the stability of the SHAP value, which indicates that even if the model changes, the attributions of that input feature will not decrease as long as the input's contributions to the outcomes remain the same. The equation for the attribute value of sample *i* of each feature is shown in Eq. (10):

$$\Phi_i(f,x) = \sum_{x\subseteq x} \frac{z(M-z-1)}{M}\left[f_x(z) - f_x(z/i)\right] \tag{10}$$

where Φ_i represents the SHAP value, and $|z'|$ denotes the number of non-zero terms in z', $z' \in \{0,1\}^M$; M means the number of input features, and S represents the set of non-zero indexes in z'.

2.3. Proposed method

In this study, the stock forecasting and trading simulation model mainly comprises four sections: 1) Data Collection and Pre-processing (DC&P) section; 2) Model Training and Optimization (MT&O) section; 3) Model Testing and Results Evaluation (MT&RE) section; and 4) Model Interpretation (MI) section. The main flowchart of the proposed approach is depicted in Fig. 1, and the primary function of each section is illustrated as follows.

(1) DC&P section. This section is developed for the acquisition and preliminary processing of the stock index price and investor sentiment feature. Historical trading data and capital flow data for the Shanghai/Shenzhen stock market at daily frequency are collected from the database. Next, the Gain function is utilized for the feature selection of multiple investor sentiments. Additionally, a walk-forward validation approach is employed to separate the entire dataset into several sub-datasets. Each sub-dataset is further separated into an in-sample dataset (training period) and an out-of-sample dataset (testing period). Details of data separation are described in **Section 3.1**.

(2) MT&O section. For each training dataset, XGBoost is used as the basic predictor to optimize the stock index direction classification model, and the Grid Search algorithm is employed to find the optimal hyperparameter combinations for the basic predictor.

(3) MT&RE section. In this section, using each testing dataset, the optimized models for stock index forecasting are verified. The performance of the optimized models is investigated by three indicators: accuracy, F1-score, and accumulated return.

(4) MI section. In this section, using the Gain function of XGBoost, the proposed method ranks the importance of the original variables to identify the essential sentiment features. In addition, the SHAP approach is incorporated in the XGBoost model to estimate the effects of each investor sentiment on the forecasting outputs of the stock index direction forecasting.

3. Data and experiment design

3.1. Data and experiment dataset

The historical data adopted in the experiments include the daily prices (open, highest, lowest, and close values) of the Shanghai/Shenzhen composite index, the daily capital flow data of the Shanghai-HongKong/Shenzhen-HongKong Stock Connect Program, and the daily capital flow data of Small Quantity Transaction (SQT) and Super-Large Quantity Transaction (SLQT) of the Chinese security market.[2] The whole data period for the Shanghai stock index starts on November 17th, 2014, and ends on December 31st, 2021. While for the Shenzhen stock index, because Shenzhen-Hongkong connect program was launched on 5th December 2016, the data period for

[2] SQT denotes a single transaction of fewer than twenty thousand shares or forty thousand CNY, and SLQT means a single transaction of more than a half million shares or one million CNY.

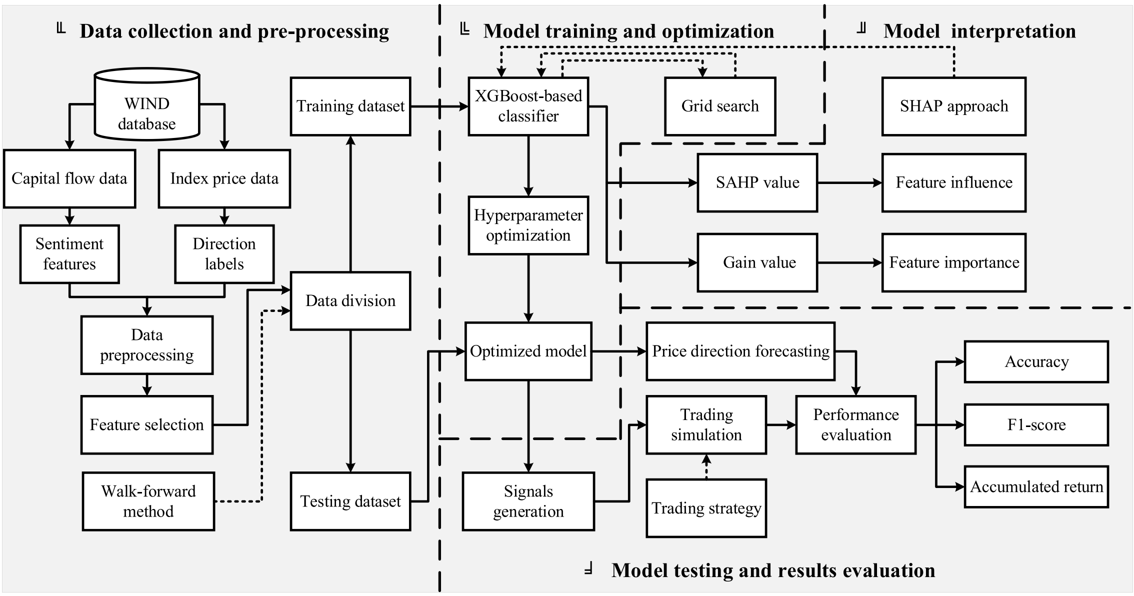


Fig. 1. The flowchart of the proposed explainable XGBoost-based method for stock index forecasting and trading simulation.

Shenzhen composite index is from January 3rd, 2017 to December 31st, 2021. The original data are derived from the WIND Database.[3] Inspired by Kumar and Lee (2006) and Frazzini and Lamont (2008), the daily capital inflow and outflow data of SQT and SLQT are used to calculate the individual and institutional sentiment index, respectively, while the daily capital inflow and outflow data of the Shanghai-Hongkong and Shenzhen-Hongkong Stock Connect Programs are utilized to construct the foreign investor sentiment index.[4] Then, the open and close prices are used to calculate the prediction return of the Shanghai/Shenzhen composite index.[5] The formulas for sentiment features and stock index return are shown in the following Eqs. (11) and (12), respectively:

$$Investor\ Sentiment\ Index = \frac{CapitalInflows - CapitalOutflows}{CapitalInflows + CapitalOutflows} \tag{11}$$

$$R_{t+5} = \frac{Close_{t+5} - Open_{t+1}}{Open_{t+1}} \tag{12}$$

Throughout the experiments, the forecasting models construct input features using the lagged values of investor sentiment, which are designed as three kinds of investor sentiment indices from the period from (*i*-5)-th day to *i*-th day. In addition, daily movements of the stock index are classified in two directions: “falling” or “rising”, representing the classification labels 0 and 1, respectively. If the stock return is smaller than zero, the method records a “falling” label. Otherwise, a “rising” label will be recorded if the stock return is larger than zero. As recommended by Thomason (1999), a time horizon of five days is chosen for index prediction. Thereby, our task is the movement direction prediction of the Shanghai/Shenzhen composite index after five business days. Additionally, we suppose that the direction changes depending on investor sentiment features in the latest six business days. An example of the eighteen sentiment features and their corresponding forecasting labels is explained in Fig. 2.

As depicted in Fig. 3, for investigating the prediction performance of investor sentiment on the stock index in different periods, a walk-forward validation approach is utilized to establish the forecasting models throughout the experiments. Subsequently, the whole experiment data is divided into several consecutive sub-datasets for model training and prediction. In each sub-dataset, the length ratio of the in-sample and out-of-sample period is 3:1. The in-sample dataset is used for model training. In contrast, the out-of-sample dataset is designed to evaluate the forecasting ability. Once the experiment of one sub-dataset is finished, the sliding window will be moved forward by one year to carry out a rolling experiment for a new round of model learning and evaluation.

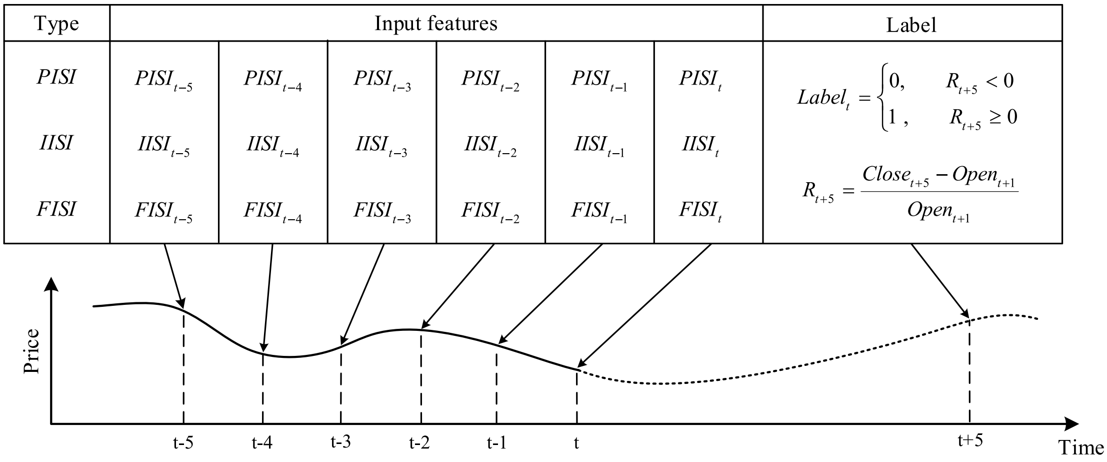
3.2. Evaluation measures

Additionally, as presented in Eqs. (13) and (14), two extensively used evaluation indicators, accuracy and F1-score, are employed

[3] The website of the WIND database is https://www.wind.com.cn/.

[4] Foreign investors here refer to investors who trade A-shares in Chinese security market through the Shanghai-HongKong (Shenzhen-HongKong) Stock Connect Program from HongKong.

[5] Note that it is difficult to determine a trading signal until the closing price is known. Thus, for daily return calculations, we suppose that market participants execute the transaction at the opening price of the trading day T+1 if the trading signal is provided on T.



Type	Input features						Label
PISI	$PISI_{t-5}$	$PISI_{t-4}$	$PISI_{t-3}$	$PISI_{t-2}$	$PISI_{t-1}$	$PISI_{t}$	$Label_t = \begin{cases} 0, & R_{t+5} < 0 \\ 1, & R_{t+5} \geq 0 \end{cases}$
IISI	$IISI_{t-5}$	$IISI_{t-4}$	$IISI_{t-3}$	$IISI_{t-2}$	$IISI_{t-1}$	$IISI_{t}$	
FISI	$FISI_{t-5}$	$FISI_{t-4}$	$FISI_{t-3}$	$FISI_{t-2}$	$FISI_{t-1}$	$FISI_{t}$	$R_{t+5} = \frac{Close_{t+5} - Open_{t+1}}{Open_{t+1}}$

Fig. 2. An example of the model input features and forecasting labels. *PISI, IISI*, and *FISI* represent the individual, institutional, and foreign investor sentiment, respectively.

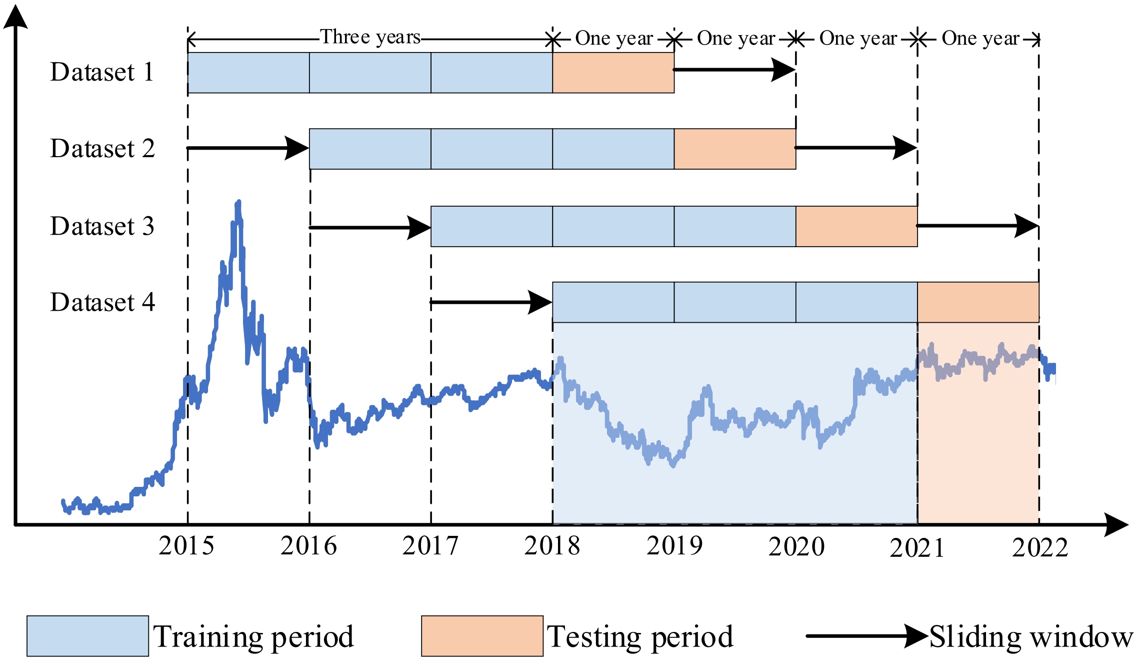


Fig. 3. Data separation of the training and testing periods using a walk-forward validation approach.

to measure the direction forecasting ability of each method on the Shanghai/Shenzhen composite index movements:

$$Accuracy = \frac{\text{Number of correct ``1'' predictions} + \text{Number of correct ``0'' predictions}}{\text{Total number of direction predictions}} \tag{13}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{14}$$

where *precision* indicates the proportion of actual “1″ samples among the predicted “1” samples, *recall* represents the proportion of actual “1” samples that are predicted correctly, and the F1-score is a weighted average of the model *precision* and *recall.*

In addition to yielding a considerable direction prediction accuracy by a predictor, almost all the market traders would pay great attention to its profitability. Thus, other than accuracy and F1-score, accumulated return (AR) is also adopted to investigate the profit-making ability of the proposed approach. In this research, a trading strategy is designed based on the predicted signal of each method: if the predicted movement signal is “1″ at time point *T*, the proposed method will open a long position with the opening price at time point $T + 1$; On the contrary, if the predicted movement signal is ”0″, a short-selling transaction would be executed; Additionally, we adopt a simple “separation strategy” for the trading position management, which means that the whole position is uniformly distributed for each holding days averagely. By using this trading position management, it can be ensured that the models can execute trading by the designed trading strategy on every trading day. For instance, if a 5-day holding period is adopted and a “rising” signal is suggested by the predictor at trading date *T*, the model will uniformly distribute 1/5 (20 %) position to the stock index at trading day $T + 1$. Therefore, the simulation trading return of throughout each testing period is calculated for the proposed method. Furthermore, a cost of 0.5 % per round trip for each transaction is considered for each transaction in the simulation trading. Subsequently, for each testing period, the accumulated return could be calculated by Eq. (15):

$$AR = \sum_{i=1}^{n}\left(p \times \frac{(close_{i+5} - open_{i+1})}{open_{i+1}} \times sign_i - \text{cost}\right) sign_i = \begin{cases} -1, hifsignal = 0 \\ 1, hhhifsignal = 1 \end{cases} \tag{15}$$

where *i* is for the *i*-th transaction in a testing period, $close_{i+5}$ represents the closing price five days after each trading point, $open_{i+1}$ represents the opening price on day $i + 1$, *p* is the trading position, and *cost* represents the cost for each transaction. In addition, $sign_i$ is the symbolic function, which equals 1 for the long signals and 0 for the short-selling signals. Thus, based on Eq. (15), the accumulated return results for all *n* transactions over each testing period can be calculated.

3.3. Hyper-parameters optimization design

Generally, the design of hyper-parameters of the model could have a significant impact on forecasting accuracy. It is recognized that the hyper-parameters selected for a model could be better optimized based on the historical data, instead of referring to experts' experience. Thus, the Grid Search algorithm is adopted as an optimization approach to determine the optimal hyper-parameters of XGBoost in each testing period. As for XGBoost, the hyper-parameters ‘nrounds’, ‘max_depth’, ‘eta’, and ‘subsample’ are selected and optimized. A brief description of those four hyper-parameters, along with their search ranges, is explained in Table 1.

3.4. Benchmarks

Other than the XGBoost-based method, nine benchmarks are designed for results comparison, and a summary of all methods investigated in the experiments is listed in Table 2. Specifically, Method 1 (XGBoost-S) represents the proposed method, while Methods 2–10 belong to benchmarks. Method 2 adopts the traditional OLS (Ordinary Least Squares) regression method, while Methods 3–6 belong to traditional machine learning techniques. Similarly, we also employ the Grid Search method to find the best hyper-parameter combinations for ANN, KNN, SVM, and RF. The hidden units ‘size’ and the weight decay ‘decay’ are two main hyper-parameters that need to be optimized for ANN. As for SVM, the regularization term ‘c’ and the hyper-parameter of kernel function ‘sigma’ are optimized. As for RF, the optimized hyper-parameter is the minimum number of distinct row references to split a node and the number of candidate variables for a split. For KNN, the optimized hyper-parameter is the ‘k’, which represents the number of neighbors that will be checked to determine the classification of a specific query point. Method 7 employs the indicators of investor sentiment and return as features for the proposed method. By contrast, only the indicators of historical return are employed as predictive features for Method 8. Methods 7 and 8 are designed and adopted as baselines to verify whether the indicators of multiple investor sentiment are more satisfactory than historical returns as features for the proposed method, and it should be noted that in Methods 7 and 8, the hyper-parameter value ranges of the XGBoost are consistent with the proposed method. Additionally, to keep consistency, the predicted time horizon and input features used by the benchmarks are identical to the proposed method. Furthermore, BAH and SAH, which are two classical trading strategies, are introduced and designed as Methods 9 and 10 to compare with the proposed method for determining whether the proposed method could perform better than those renowned trading strategies. Note that BAH and SAH are employed only for simulation trading.

Table 1
The hyper-parameters of the XGBoost model that selected for model training.

Hyper-parameter	Description	Function	Search range
nrounds	The max number of boosting iterations.	It controls the maximum number of iterations.	[150, 200]
max_depth	The maximum depth of a tree.	It controls the depth of the tree to prevent overfitting.	[4, 10]
eta	It controls the learning rate, and it scales the contribution of each tree by a factor of eta.	It controls the learning rate, i.e., the rate at which the model learns patterns in data.	[0.01, 0.10]
subsample	The subsample ratio of the training instance.	It controls the number of samples supplied to a tree to prevent overfitting.	[0.5, 1.0]

Table 2
A list of the XGBoost based approach and benchmarks. DF denotes index direction forecasting, and ST means simulation trading.

No	Method	Function	Description
1	XGBoost-S (proposed method)	DF and ST	An index direction classifier based on an XGBoost method; Trading simulation is executed based on the predicted direction.
2	OLS	DF and ST	An index direction predictor based on an OLS regression method; Trading simulation is performed based on the predicted direction.
3	ANN	DF and ST	An index direction classifier based on an ANN method; Trading simulation is carried out based on the predicted direction.
4	KNN	DF and ST	An index direction classifier based on a KNN method; Trading simulation is performed based on the predicted direction.
5	SVM	DF and ST	An index direction classifier based on an SVM method; Trading simulation is executed based on the predicted direction.
6	RF	DF and ST	An index direction classifier based on an RF method; Trading simulation is carried out based on the predicted direction.
7	XGBoost-RS	DF and ST	The indicators of investor sentiment and return are employed as features for the XGBoost model.
8	XGBoost-R	DF and ST	Only the indicators of return are employed as features for the XGBoost model.
9	BAH	ST	A “long” transaction is carried out at the start of each testing period, and the position is closed at the end.
10	SAH	ST	A “short-selling” transaction is carried out at the start of each testing period, and the position is closed at the end.

4. Results

4.1. Forecasting results

The experimental results of accuracy and F1-score for the Shanghai/Shenzhen composite index forecasting in the out-of-sample periods are presented in Table 3, from which we observe that among all methods, the proposed approach XGBoost-S obtained the best average accuracy and F1-score results for movement directions prediction of the Shanghai composite index and the Shenzhen composite index. It demonstrates that the forecasting ability of XGBoost-S was superior to OLS regression and other traditional machine learning techniques, which is consistent with the findings of Meng et al. (2020). Additionally, although SVM generated a larger F1-score result than XGBoost-S in the out-of-sample period of 2021 for the Shanghai composite index, it failed to produce outstanding results in the other three model training periods. Likewise, although the benchmark XGBoost-RS produced a larger F1-score than the proposed method XGBoost-S in the out-of-sample period of 2021 for the Shenzhen composite index, its performance in 2020 was extremely worse than XGBoost-S. Overall, the proposed method XGBoost-S showed an extremely better forecasting ability to predict the movement directions of the Shanghai composite index and the Shenzhen composite index.

Additionally, we focus on the direction forecasting performances of the XGBoost-based approaches, which include XGBoost-S, XGBoost-R, and XGBoost-RS. From Table 3, it is observed that XGBoost-S and XGBoost-RS produced comparable average accuracy and F1-score results, while they were both apparently better than that of the benchmark XGBoost-R. XGBoost-R adopted only the historical return as the predictive features, while investor sentiment and historical return are utilized by XGBoost-RS. It demonstrates that the beneficial features for XGBoost to yield superior direction accuracy were the investor sentiments, not the historical returns. Overall, the accuracy and F1-score results of the Shanghai composite index and the Shenzhen composite index demonstrate that those benchmarks were not sufficiently robust for stock index prediction, and the proposed method outperformed all benchmark methods in terms of direction prediction accuracy.

Furthermore, the Friedman test (Derrac et al., 2011) is conducted to investigate whether the proposed method is significantly superior to the benchmarks. The null hypothesis H_0 of the Friedman test on Accuracy/F1-score states the equality of movement direction forecasting among the compared models, and the alternative hypothesis H_1 is defined as the negation of the null hypothesis. The Friedman test results are reported in Table 4, from which it is found that the proposed method was significantly better than the benchmarks in terms of accuracy and F1-score at the 0.01 and 0.05 confidence level, respectively.

4.2. Trading simulation results

Table 5 reports the accumulated returns of the proposed and benchmark approaches for trading simulation on the Shanghai composite index and the Shenzhen composite index. The values of the yearly overall return displayed in Table 5 are calculated relative to the initial investment at the start of each out-of-sample period. Note that the simulation trading results of the benchmark methods SAH and BAH are provided for results comparison with the proposed approach.

Generally, a larger accumulated return represents the superior profitability of a method. From the simulation trading results for the Shanghai composite index shown in Panel A of Table 5, it is observed that the annual accumulated return of four years’ testing that yielded by RF, SVM, ANN, KNN, OLS, BAH, and SAH were 0.09 %, −8.54 %, −1.72 %, 2.23 %, −3.10 %, 3.37, and − 4.37 %, respectively, while XGBoost-S gained the best average return of 8.71 % per year. Although SAH, BAH, and BAH respectively yielded the best overall return in 2018, 2019, and 2020, they suffered vast losses in some years. For instance, SAH produced a negative return of more than −10 % in 2019 and 2020, which ultimately led to a vast loss at the end of the whole out-of-sample period. Similarly, BAH incurred a huge loss of −25.25 % in 2018. It demonstrates that these approaches were not sufficiently robust for yielding stable positive returns in the Chinese stock market. Additionally, except for XGBoost-RS, only the proposed method XGBoost-S consistently produced positive returns in four testing years, and it outperformed all the benchmarks in terms of average return.

Furthermore, it is found that the overall return per year gained by the proposed method XGBoost-S (8.71 %) was superior to that of

Table 3
Forecasting performance of the proposed method and benchmark methods in the out-of-sample periods.

Period	ANN		SVM		RF		KNN		OLS		XGBoost-RS		XGBoost-R		XGBoost-S	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Panel A The forecasting results for the Shanghai composite index																
2018	48.56 %	53.53 %	49.79 %	59.60 %	54.32 %	61.05 %	55.14 %	60.36 %	51.03 %	54.05 %	57.61 %	57.96 %	54.32 %	56.81 %	60.08 %	62.84 %
2019	54.51 %	60.50 %	50.82 %	55.88 %	56.97 %	66.67 %	54.10 %	62.16 %	50.82 %	50.00 %	64.34 %	75.49 %	52.87 %	52.67 %	63.93 %	76.60 %
2020	50.66 %	59.21 %	47.60 %	55.22 %	49.34 %	56.72 %	57.64 %	66.89 %	50.22 %	44.66 %	57.64 %	63.12 %	55.02 %	56.90 %	59.39 %	71.38 %
2021	51.97 %	53.78 %	51.97 %	68.39 %	53.71 %	62.41 %	55.46 %	63.83 %	56.33 %	65.28 %	57.56 %	61.60 %	53.78 %	62.59 %	58.08 %	66.20 %
Average	51.42 %	56.75 %	50.04 %	59.77 %	53.59 %	61.71 %	55.59 %	63.31 %	52.10 %	53.50 %	59.29 %	64.54 %	54.00 %	57.24 %	60.37 %	69.25 %
Panel B The forecasting results for the Shenzhen composite index																
2020	51.44 %	60.93 %	54.32 %	60.50 %	50.42 %	59.59 %	55.56 %	62.76 %	51.85 %	50.21 %	58.02 %	63.04 %	59.26 %	73.60 %	60.91 %	71.12 %
2021	53.36 %	62.63 %	52.67 %	61.79 %	57.61 %	61.99 %	54.62 %	65.38 %	54.62 %	61.97 %	56.72 %	68.31 %	57.56 %	71.87 %	57.98 %	67.32 %
Average	52.40 %	61.78 %	53.50 %	61.15 %	54.02 %	60.79 %	55.09 %	64.07 %	53.24 %	56.09 %	57.37 %	65.68 %	58.41 %	72.73 %	59.44 %	69.22 %

Table 4
Friedman Test of the forecasting performance for the proposed method and benchmark methods.

Compared models	Accuracy testing results	F1-score testing results
	Significant level α = 0.01	Significant level α = 0.05
XGBoost-S versus ANN	H_0: n1 = n2 = n3 = n4 = n5 = n6 = n7 = n8	H_0: n1 = n2 = n3 = n4 = n5 = n6 = n7 = n8
XGBoost-S versus SVM		
XGBoost-S versus RF		
XGBoost-S versus KNN	F = 21.08	F = 13.92
XGBoost-S versus OLS		
XGBoost-S versus XGBoost-RS		
XGBoost-S versus XGBoost-R	p < 0.01 (reject H_0)	p = 0.05 (reject H_0)

Table 5
Simulation trading results of the proposed method and benchmark methods in the out-of-sample periods.

Period	ANN	SVM	RF	KNN	OLS	SAH	BAH	XGBoost-RS	XGBoost-R	XGBoost-S
Panel A. The simulation trading results of Shanghai composite index										
2018	−13.22 %	−16.40 %	−6.70 %	−0.09 %	−6.72 %	24.25 %	−25.25 %	8.87 %	8.75 %	10.43 %
2019	3.75 %	3.53 %	3.96 %	0.22 %	−1.54 %	−22.61 %	21.61 %	8.15 %	3.27 %	8.17 %
2020	0.45 %	−4.59 %	2.24 %	2.91 %	−4.05 %	−13.76 %	12.76 %	3.77 %	4.05 %	6.88 %
2021	2.12 %	−16.68 %	0.86 %	5.86 %	−0.07 %	−5.35 %	4.35 %	8.81 %	−4.03 %	9.38 %
Average	−1.72 %	−8.54 %	0.09 %	2.23 %	−3.10 %	−4.37 %	3.37 %	7.40 %	3.01 %	8.71 %
Panel B. The simulation trading results of Shenzhen composite index										
2020	−7.20 %	0.09 %	−2.59 %	4.50 %	−6.45 %	−7.21 %	6.21 %	6.68 %	8.93 %	7.53 %
2021	−0.33 %	−4.09 %	5.45 %	−1.72 %	4.86 %	−34.79 %	33.79 %	4.19 %	−6.47 %	8.63 %
Average	−3.77 %	−2.00 %	1.43 %	1.39 %	−0.79 %	−21.00 %	20.00 %	5.44 %	1.23 %	8.08 %

XGBoost-RS (7.40 %) and XGBoost-R (3.01 %). The performance of XGBoost-S was slightly better than the benchmark XGBoost-RS, whereas it was markedly superior to that of the benchmark XGBoost-R. Furthermore, XGBoost-S and XGBoost-RS yielded positive returns in all four testing periods, while XGBoost-R suffered enormous losses in 2021. It reveals that the beneficial indicators for producing outstanding results were the multiple investor sentiments, not historical returns.

Next, we focus on the profitability of each approach traded on the Shenzhen composite index. From the experimental results presented in Panel B of Table 5, it is found that the average accumulated return of the proposed method XGBoost-S was superior to that of the baseline methods, including ANN, SVM, KNN, OLS, and SAH. Although BAH produced a superior average return than the proposed method XGBoost-S because it generated a considerable profit in 2021, it incurred huge losses in some years. For instance, BAH obtained a negative loss of –33.89 % in 2018. Similarly, as for the Shanghai composite index, among the three XGBoost-based approaches, the best return was also yielded by the proposed method XGBoost-S, which yielded an average accumulated return of 8.08 %. It demonstrates that the beneficial features are the sentiment variables, and the proposed method outperforms the benchmarks in terms of profitability.

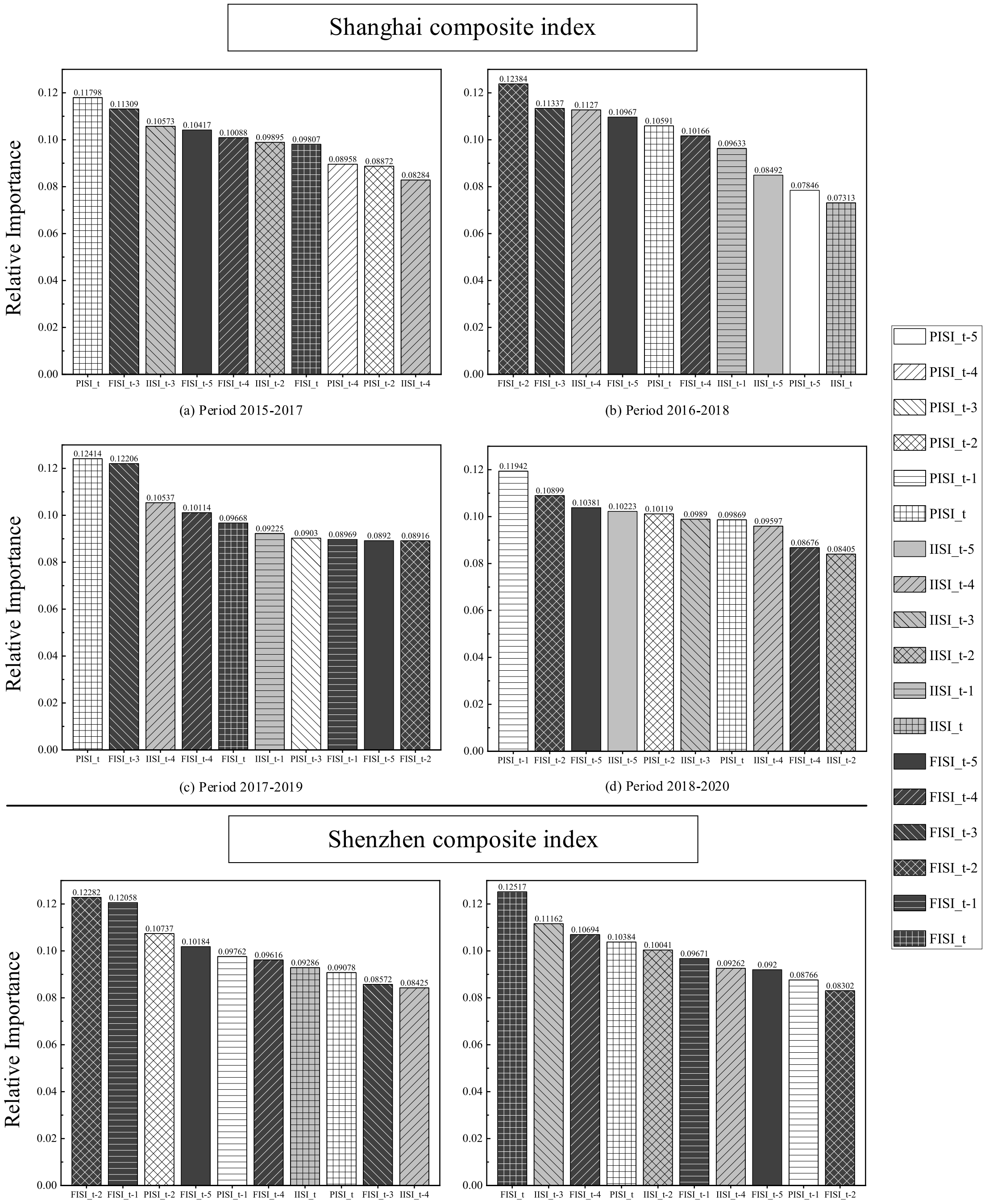
Similarly, Friedman test is conducted to further test whether the proposed method is significantly better than the benchmarks in terms of profit-making ability. The null hypothesis of the Friedman test on profitability results states the equality of trading returns among the compared models, while the alternative hypothesis is defined as the negation of the null hypothesis. The Friedman test results are provided in Table 6, from which it is found that the null hypothesis was rejected with a confidence level of 0.05 in the one-tail test. Therefore, we had a high level of confidence that the proposed method is significantly superior to the benchmarks in terms of profit-making ability.

Table 6
Friedman test of the trading simulation performance for the proposed method and benchmark methods.

Compared models	Significant level α = 0.05
XGBoost-S versus ANN	H_0: n1 = n2 = n3 = n4 = n5 = n6 = n7 = n8 = n9 = n10
XGBoost-S versus SVM	
XGBoost-S versus RF	
XGBoost-S versus KNN	
XGBoost-S versus OLS	F = 18.00
XGBoost-S versus SAH	
XGBoost-S versus BAH	
XGBoost-S versus XGBoost-RS	
XGBoost-S versus XGBoost-R	p = 0.03 (reject H_0)

4.3. Feature importance results

Furthermore, since the relative importance score of features is available by calculating the *Gain* of XGBoost, we then concentrated on investigating which features are the essential indicators to forecast the direction movements of the Shanghai/Shenzhen composite index, especially the features with the five largest importance scores in the training periods. The relative importance scores of the ten sentiment indicators after feature selection and the relative importance scores for specific sentiment indicators with their time lags in the training periods are respectively plotted in Figs. 4 and 5. Overall, The Gain values obtained using XGBoost show that there was a possible correlation between the stock index direction change and the sentiments of three kinds of investors in the Chinese security



(a) Period 2015-2017

(b) Period 2016-2018

(c) Period 2017-2019

(d) Period 2018-2020

(e) Period 2017-2019

(f) Period 2018-2020

Fig. 4. Relative importance scores of the selected sentiment indicators in the in-sample periods.

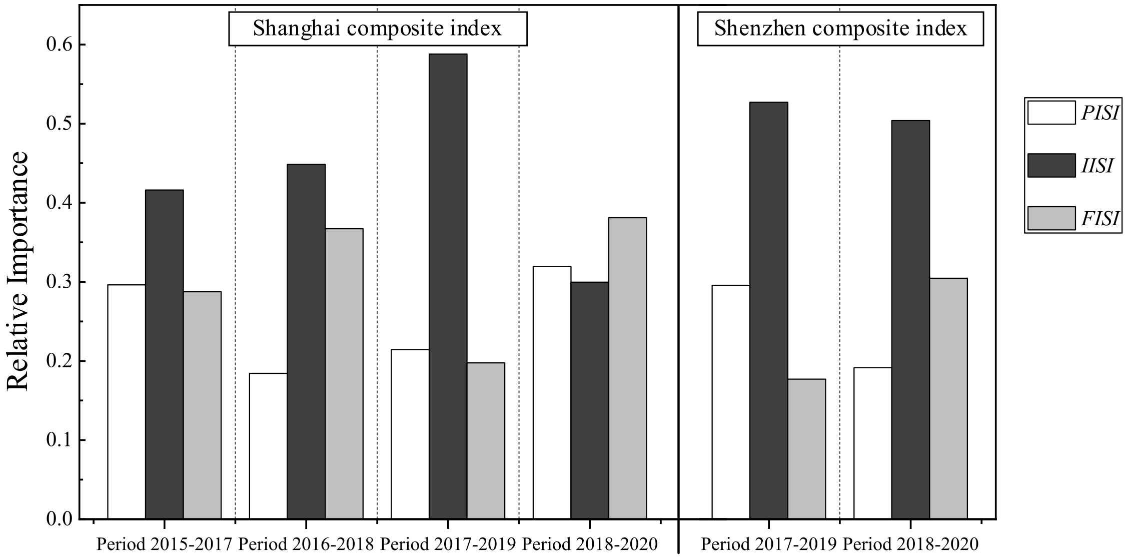
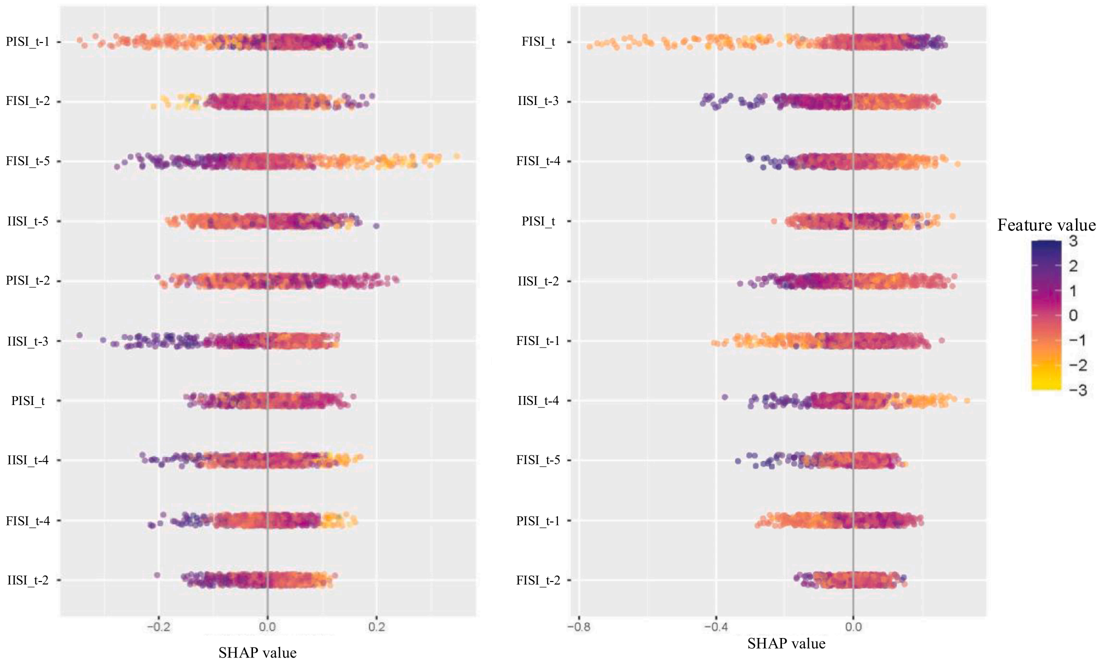


Fig. 5. The relative importance scores of three types of investor sentiment features obtained in the in-sample periods (The score of *PISI*, *IISI*, and *FISI* represents the sum of the relative importance score of the individual, institutional, and foreign investor sentiments, respectively).

market. In addition, as shown in Fig. 4, sentiment features of the individual, institutional, and foreign investors are lagged from the period of the (*i*-5)-th day to the *i*-th day. Among them, it can be observed that individual investor sentiment with a zero-day lag consistently ranked in the top eight relative importance score over the training periods, indicating that individual investor sentiment of the current day could be one of the most influential and essential factors for forecasting the Shanghai/Shenzhen composite index five days later. Besides, it is found that for the Shanghai composite index, the proportion of the three investor sentiments in the Top-five relative importance score was very comparative, demonstrating the stable importance of each investor sentiment on the Shanghai



(a) Shanghai composite index

(b) Shenzhen composite index

Fig. 6. The SHAP summary plot of XGBoost-based classifier for direction prediction of the Shanghai composite index and Shenzhen composite Index.

composite index in recent years. Furthermore, it can be seen from Fig. 5 that for the Shanghai composite index, the relative importance scores of institutional investor sentiment were larger than that of individual and foreign investor sentiments in the training periods 2015 to 2017, 2016 to 2018, and 2017 to 2019, while foreign investor sentiment was distinctly more critical than the other two kinds of sentiment in the training period 2018 to 2020. For the Shenzhen composite index, it is found that the relative influence of foreign investor sentiment also increased from the period 2017 to 2020, suggesting an increasing importance of foreign investor sentiment on those two stock indices.

4.4. SHAP results

The importance ranking results of the features for index direction forecasting of the Shanghai composite index and Shenzhen composite index are analyzed in Section 4.3 using the Gain values of XGBoost. However, the Gain values of XGBoost could be used to illustrate the significance of the features on the prediction results, it is difficult to understand the exact influence of each sentiment feature on stock index movement, and how the influence is affected by the feature values. Therefore, the SHAP approach is then employed to precisely assess and visualize the contributions of the sentiment indicators of three types of investors to the index movements. The SHAP summary plots for the Shanghai composite index and Shenzhen composite index under the latest training period (2018–2020) are presented as an example in Fig. 6. The horizontal axis value of the two subplots in Fig. 6 represents the SHAP value, and the features are ranked by the average absolute value of their SHAP values on the vertical axis (the symbols for features are listed in Fig. 1), with each point representing one sample from the dataset. Feature values are represented using the colors from yellow to purple, whereas the intensity of color presents the weight of feature influence on the prediction result. A lighter/deeper color denotes a smaller/larger feature value.

As can be seen from Fig. 6, the ranking results in the SHAP summary plots of the training dataset 4 for the Shanghai composite index and Shenzhen composite index are consistent with the results of importance ranking in Fig. 4, which also validates the accuracy of the importance ranking for the model. From Fig. 6(a), it is evident that the most influential feature for direction prediction of the Shanghai composite index was PISI_t-1. The SHAP value gradually increased as the value of PISI_t-1 increased, indicating the increasing positive

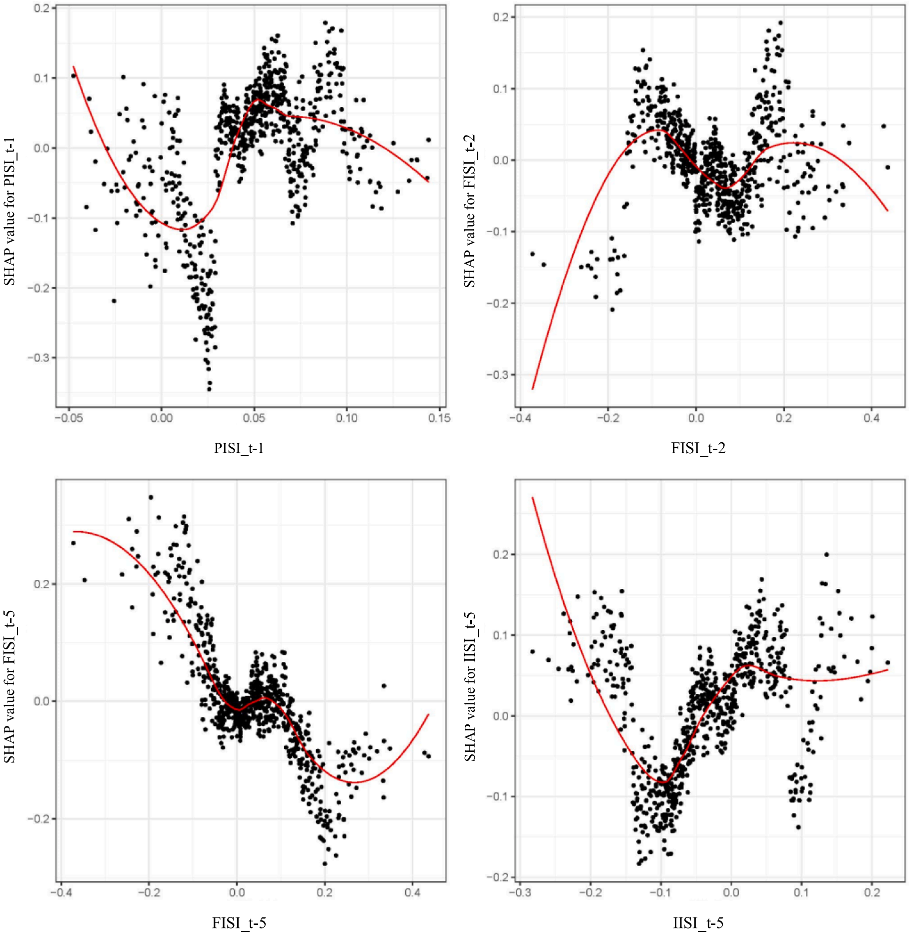


Fig. 7. The SHAP dependence plot of the PISI_t-1, FISI_t-2, FISI_t-5, and IISI_t-5 for the direction forecasting of the Shanghai composite index.

contributions to the prediction signal “rising”, and it reveals that the probability of “rising” predicted by the proposed method becomes larger. In addition, the SHAP value decreased as the value of feature FISI_t-5 increased, which indicates that the negative contributions to the prediction results gradually increased. Similarly, the positive contributions of features PISI_t-1, IISI_t-5, and PISI_t-2 increased as their feature increased, while for the features FISI_t-5, IISI_t-3, and IISI_t-2, the negative contributions to the predictions increased as their feature value become larger. As shown in Fig. 6(b), for the Shenzhen composite index, the positive contributions of the features FISI_t, FISI_t-1, and PISI_t-1 tended to become larger as their feature values increased. In contrast, the features IISI_t-3, FISI_t-4, IISI_t-2, and IISI_t-4 all had more negative contributions to the model predictions as their values increased. In addition, for the Shanghai composite index and Shenzhen composite index, eight features (PISI_t, PISI_t-1, FISI_t-2, FISI_t-4, FISI_t-5, IISI_t-2, IISI_t-3, and IISI_t-4) consistently remained after feature selection. Moreover, it is clear from the two subplots in Fig. 6 that the distribution of their SAHP values is similar, demonstrating that those sentiment features had similar changes in their contributions to the prediction results of the Shanghai composite index and Shenzhen composite index using the XGBoost model.

Furthermore, to analyze the complex relationship between feature values and SHAP values, the dependence plots of the four most influential features (PISI_t-1, FISI_t-2, FISI_t-5, and IISI_t-5) for the Shanghai composite index prediction, and the four most influential features (FISI_t, IISI_t-3, FISI_t-4, and PISI_t) for Shenzhen composite index prediction, are respectively displayed in Fig. 7 and Fig. 8. The horizontal axis of each subplot in Figs. 7 and 8 represents the feature value, and the vertical axis indicates their SHAP value. Each point in the figures refers to one prediction sample, and a smooth red curve is drawn for each feature to represent the relationship between feature values and SHAP values. As shown in Fig. 7 (a), for the Shanghai composite index, when the most critical sentiment feature PISI_t-1 was increased from -0.05 to 0.012, its SHAP value decreased, indicating that its negative contribution to the prediction results gradually increased. Thus, the probability of “falling” for the index price was raised as PISI_t-1 value was increased. As the feature PISI_t-1 continued to grow from approximately 0.012 to 0.05, its SHAP value was increased, indicating that the probability of “rising” continued to rise as the feature value was increased within this range. Thereafter, as the feature PISI_t-1 continued to increase, the probability of “falling” predicted by the XGBoost-based classifier was increased. It is also found that when FISI_t-2 gradually increased in the range of (-0.4, -0.083) or (0.067, 0.234), the probability of “falling” for the Shanghai composite index grew, and when FISI_t-2 was gradually increased in the range of (-0.083, 0.067) or (0.234, 0.476), the probability of “falling” was increased. For FISI_t-5, the probability of the “rising” prediction of the Shanghai composite index tended to increase when the FISI_t-5 value increased in the range of (0, 0.064) or (0.268, 0.476), and the probability of “falling” prediction was increased as its value grew in the range of (-0.4, 0) or (0.064, 0.268). For IISI_t-5, the probability of “rising” prediction about the Shanghai composite index was increased when the feature value IISI_t-5 raised in the range of (-0.1, 0.025) or (0.125, 0.248), and the probability of “falling” prediction was increased as its value raised in the range of (-0.3, -0.1) or (0.025, 0.125).

For the most influential feature FISI_t, in direction forecasting of the Shenzhen composite index, the probability of the “rising” prediction was increased when the feature value FISI_t grew in the range of (-0.169, 0.013) or (0.044, 0.25), whereas the “falling” probability was increased as it continues to grow in the ranges of (-0.288, -0.169), (0.013, 0.044), or (0.25, 0.4). For the feature IISI_t-3, the probability of “rising” was increased when it grew in the range of (-0.324, -0.999), and the “falling” probability increased when it grew in the range of (-0.999, 0.3). As the feature FISI_t-4 increased in the range of (-0.288, -0.106), the probability of the “rising” prediction for the Shenzhen composite index was increased, while the probability of the “falling” prediction was increased when its value grew in the range of (-0.106, 0.394). For the feature PISI_t, the probability of “rising” prediction was increased in its feature value ranges of (-0.068, -0.011), (0.035, 0.066), or (0.094, 0.170), and the probability for making a “falling” was increased in its feature value ranges of (-0.011, 0.035) or (0.066, 0.094). Therefore, the complex and non-linear effects of the sentiment varied in the individual, institutional, and foreign investors on the index movements of the Shanghai composite index and Shenzhen composite index could be revealed and understood based on the SHAP dependence plots. Based on the explainable XGBoost prediction model proposed in this study, the influences of investor sentiments on stock index prices could be visualized. The SHAP values of different kinds of investor sentiment and the different time lags, as well as their SHAP values produced over different training periods, should be a practical reference for identifying the index direction changes during actual trading.

5. Conclusion

In this research, we reasoned that sentiment features of more kinds of investors than those exploited by numerous researchers should be incorporated into the model functions to facilitate forecasting and trading performance. Thus, sentiment features of the individual, institutional, and foreign investors were utilized to forecast the Shanghai composite index and Shenzhen composite index. On this basis, an explainable XGBoost based approach was adopted for direction forecasting and simulation trading. Experimental results show that the XGBoost model outperformed the traditional methods of OLS, KNN, ANN, SVM, and RF based models throughout the testing periods. Experimental results demonstrate that the proposed method can be applied as a reliable approach for direction forecasting and simulation trading of the Shanghai composite index and Shenzhen composite index. Moreover, the relative importance scores and SHAP results provide the most influential sentiment factors for market participants to forecast the direction changes of the Shanghai composite index and Shenzhen composite index.

In the future, other scholars could apply the explainable XGBoost approach for stock index prediction by using market investor sentiment features with other lag lengths to expand the current research. In addition, the validity of the XGBoost approach using investor sentiments on individual stocks or industry indices could be further investigated. Furthermore, due to the lack of capital flow information to construct the three types of investor sentiment, this research mainly investigated the effectiveness of the proposed method in the Chinese security market. To improve the robustness of the findings in this research, other researchers could further provide evidence for index forecasting and trading simulation from developed countries if they could obtain those investor sentiment

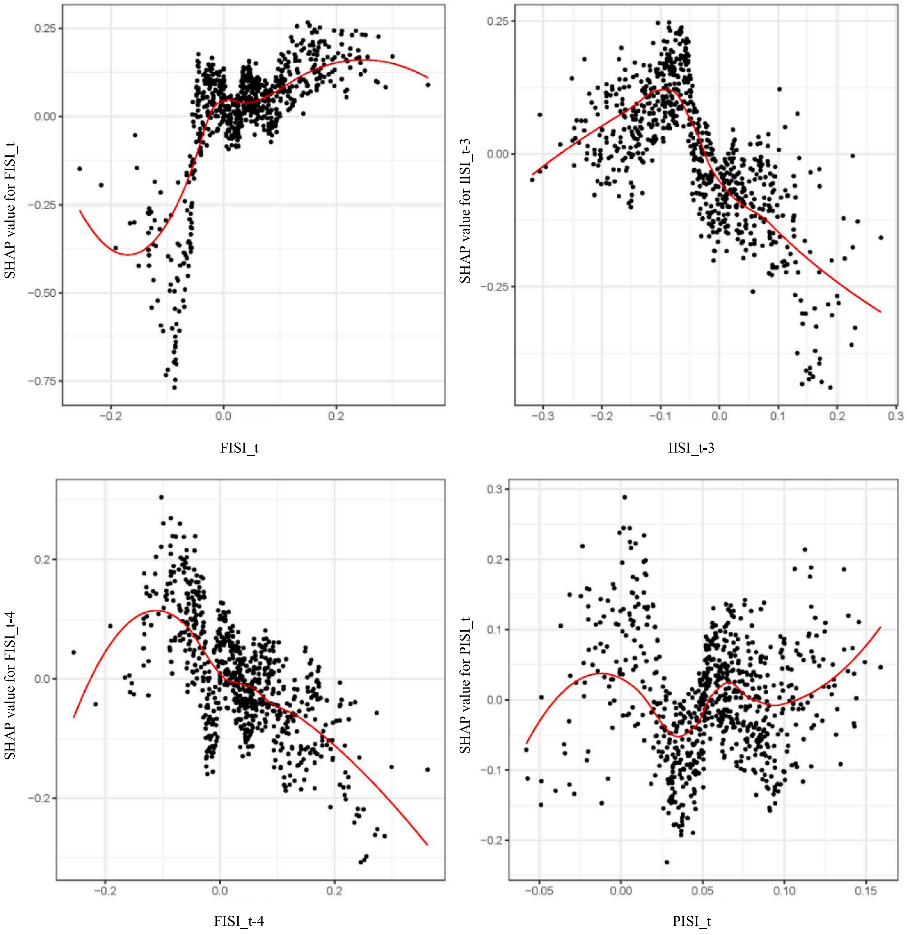


Fig. 8. The SHAP dependence plot of the FISI_t, IISI_t-3, FISI_t-4, and PISI_t for the direction forecasting of the Shenzhen composite index.

indicators.

CRediT authorship contribution statement

Shangkun Deng: Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition. **Xiaoru Huang:** Data curation, Writing – original draft, Software, Visualization. **Yingke Zhu:** Writing – review & editing, Formal analysis. **Zhihao Su:** Validation, Writing – review & editing. **Zhe Fu:** Investigation, Data curation, Software. **Tatsuro Shimada:** Writing – review & editing, Funding acquisition, Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is supported by Hubei Provincial Natural Science Foundation of China (grant number 2021CFB175), Philosophy and Social Science Research Project of the Department of Education of Hubei Province (grant number 21Q035), and Natural Science Foundation of Yichang City (grant number A22-3-011). In addition, the authors would like to thank the anonymous reviewers for their comments and discussions.

References

Baker, M., Wurgler, J., & Yuan, Y. (2012). Global, local, and contagious investor sentiment. *Journal of Financial Economics, 104*(2), 272–287.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science SCI-NETH., 2*(1), 1–8.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In the 22nd ACM SIGKDD International Conference. 785-794.

Chi, L., Zhuang, X., & Song, D. (2012). Investor sentiment in the Chinese stock market: An empirical analysis. *Applied Economics Letters, 19*(4), 345–348.

Chung, S. L., Hung, C. H., & Yeh, C. Y. (2012). When does investor sentiment predict stock returns? *Journal of Empirical Finance, 19*(2), 217–240.

Deng, S., Huang, X., Qin, Z., Fu, Z., & Yang, T. (2021). A novel hybrid method for direction forecasting and trading of Apple Futures. *Applied Soft Computing, 110*, Article 107734.

Deng, S., Wang, C., Fu, Z., & Wang, M. (2021). An intelligent system for insider trading identification in Chinese security market. *Computational Economics, 57*(2), 593–616.

Dergiades, T. (2012). Do investors' sentiment dynamics affect stock returns? Evidence from the US economy. *Economic Letters, 116*(3), 404–407.

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation, 1*(1), 3–18.

Fan, J., Wang, X., Wu, L., Zhou, H., et al. (2018). Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Conversion and Management, 164*, 102–111.

Fang, M., & Taylor, S. (2021). A machine learning based asset pricing factor model comparison on anomaly portfolios. *Economic Letters*, 109919.

Fisher, K. L., & Statman, M. (2000). Investor sentiment and stock returns. *Financial Analysts Journal, 56*(2), 16–23.

Frazzini, A., & Lamont, O. A. (2008). Dumb money: mutual fund flows and the cross-section of stock returns. *Journal of Financial Economics, 88*(2), 299–322.

Gao, X., Gu, C., & Koedijk, K. (2020). Institutional investor sentiment and aggregate stock returns. *European Financial Management, 27*(5), 899–924.

Han, X., & Li, Y. (2017). Can investor sentiment be a momentum time-series predictor? Evidence from China. *Journal of Empirical Finance, 42*, 212–239.

Kumar, A., & Lee, C. M. C. (2006). Retail investor sentiment and return comovements. *The Journal of Finance, 61*(5), 2451–2486.

Kwon, J., Zhao, S., & Li, Z. (2022). Predicting crowd funding success with visuals and speech in video ads and text ads. *European Journal of Marketing, 56*(6), 1610–1649.

Li, X., & Tang, P. (2020). Stock index prediction based on wavelet transform and FCD-MLGRU. *J. Forecasting, 39*(8), 1229–1237.

Li, S., & Zhang, X. (2019). Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. *Neural Computing and Applications, 32*, 1971–1979.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems.*, 4765–4774.

Meng, M., Zhong, R., & Wei, Z. (2020). Prediction of methane adsorption in shale: Classical models and machine learning based models. *Fuel, 278*, Article 118358.

Ryu, D., Kim, H., & Yang, H. (2017). Investor sentiment, trading behavior and stock returns. *Applied Economics Letters, 24*(12), 826–830.

Shapley, L.S., 2016. 17. A value for n-person games. In H. Kuhn & A. Tucker (Ed.), Contributions to the Theory of Games (AM-28), Volume II (pp. 307-318).

Tanaka, K., Kinkyo, T., & Hamori, S. (2016). Random forests-based early warning system for bank failures. *Economics Letters, 148*, 118–121.

Thomason, M. (1999). The practitioner methods and tool. *Journal of Computational Finance, 7*(3), 36–45.