2020 16th IEEE International Conference on Automation Science and Engineering (CASE)
August 20-21, 2020, Online Zoom Meeting

Multi-model fusion based on Stacking: A predictive model for the price trend of natural rubber

Hongbo Fan, Xiaosong Luo, Zhao-Hui Sun, *Member, IEEE*, Xin Yuan and Siqi Qiu

***Abstract*—Natural rubber plays the most important role in manufacturing. It has an important position in the automobile tire industry, sealing industry and daily necessities industry. Therefore, its price trend has a major impact on industrial output and economy. For the manufacturing industry, a reasonable forecast of rubber prices will give enterprises a huge advantage in planning production, reducing costs and improving production efficiency. In this paper, in response to the demand for rubber price forecasts in the manufacturing industry, the relevant factors affecting rubber prices are analyzed, and the natural rubber price data sets of more than 500 sets of variables spanning 10 years are collected. The work related to data preprocessing and feature selection was completed. XGBoost was used as a super model to integrate the Random Forest, Adaboost and LSTM base models to obtain the final Stacking model. And provides an important reference for the development of natural rubber procurement strategy.**

***Index Terms*—Data Fusion; Data Driven and Data Mining; Optimization and Optimal Control.**

I. Introduction

As an important raw material for industrial production, natural rubber plays an important role in the automobile tire industry, sealing industry and daily usage industry. Therefore, a reasonable forecast of rubber prices will bring huge advantages to enterprises in planning production, reducing costs, and improving production efficiency. However, the trend of natural rubber price is affected by many factors, and the fluctuation is very large. As shown in Figure 1, it is difficult to formulate a reasonable procurement strategy, so efficient and accurate prediction methods play an important role in the cost control of the production process.

The existing methods for prediction of natural rubber price and development of procurement strategies are largely dependent on human experience, and buyers make decision by observing historical data trends and market information. This strategy is too subjective and it is difficult to gather enough information for decision making. Therefore, we collect as many data, which are related to natural rubber price, as possible through various types of databases and network information, and explore the relationship between massive data and natural rubber prices using the machine learning method.

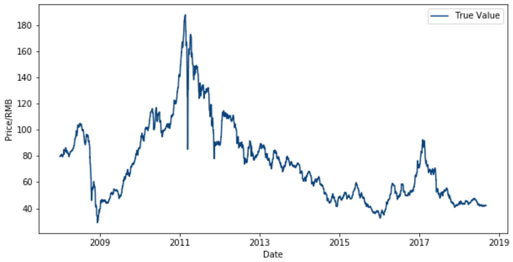

Fig. 1. Price Trend of Natural Rubber 2008-2018.

This paper first analyzes the price influencing factors of natural rubber and collects corresponding data. According to the research, we divide the influencing factors into four categories: market factors, macro factors, industrial upstream, and climate factors. To solve the problem of data dimension explosion and data sparseness during processing, principal component analysis (PCA) [1] and recursive feature selection (RFE) [2] methods are used to screen out the irrelevant data dimension, which reduces the computational cost of the model and improves the overall performance of the model. In the test process, different algorithms are compared and the performance of the single algorithm is not satisfactory. Therefore, the Stacking [3] method is used to fuse three algorithms including Random Forest [4], Adaboost [5] and LSTM(Long Short-Term Memory) [6], after which the final prediction results are obtained. By comparing with the previous corporate procurement strategies, using the model proposed in this paper can save the company 4%-6% of the procurement cost in the natural rubber procurement process, which brings obvious benefits.

The main contributions of this paper are mainly reflected in the following points:

(1) Through data analysis, we found that the factors which

This work was supported by the the National Natural Science Foundation of China (Grant No. 71632008).

H. Fan, X. Luo, Z.H. Sun, X. Yuan and S. Qiu are with the School of Mechanical Engineering and Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai, 200240 China (E-mails: 645521626@qq.com,1148212680@qq.com, zh.sun@sjtu.edu.cn, yuanxin@cqu.edu.cn, siqiqiu1988@163.com).

978-1-7281-6904-0/20/$31.00 ©2020 IEEE

have an important impact on the price of natural rubber can be divided into four categories and nine subcategories. The problem of data collection, storage, and analysis were solved.

(2) A fusion algorithm model based on Stacking is proposed. Compared with the previous prediction methods, the existing prediction methods are improved as a whole, which greatly reduces the procurement cost of natural rubber.

(3) Analyze the sensitivity of data features, and propose a sensitivity matrix acquisition method based on PCA and regression analysis to analyze the influence degree of different characteristics on the price trend of natural rubber.

The composition of the remaining chapters of this article is shown below. In Section II, the data preprocessing work is introduced, including data collection, data cleaning and feature processing. Section III mainly introduces our prediction model for the price trend of natural rubber. Section IV reviews and introduces the previous work on time series forecasting. The Section V summarizes and forecasts the work of this paper.

II. Data Preprocessing

A. Dataset

While natural rubber can be taken as agricultural product, it also has multiple attributes of industrial products and financial products. So, The factors that affect natural rubber price fluctuations are subdivided into four major categories and nine subcategories based on domain experts comments, as shown in Figure 2. The relevant data was collected through the national statistical bureau, financial database, network information and other sources, and more than 300 sets of daily frequency data with a time span of 2008-2018 were collected. The specific data format is shown in Table 1.

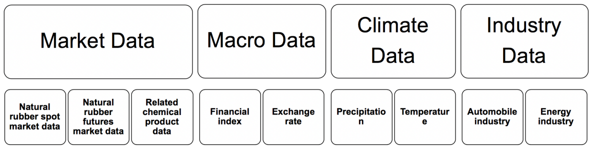

Fig. 2. Category of Raw Data Collection.

B. Data Processing

The processing of data can be divided into two steps: data cleaning and feature processing. The specific processing flow is shown in Figure 3.

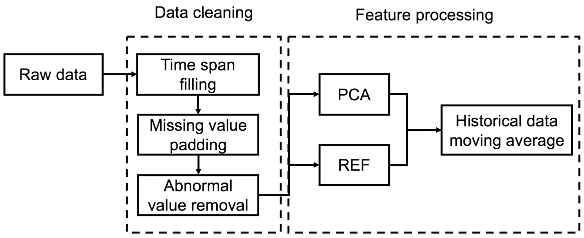

Fig. 3. Specific Process of Data Preprocessing.

Due to the difference of data source and data type, the collected data has different spans and different dates of valid data; in addition, real-life data will have more or less missing values. Therefore, the first two steps of data cleaning are to fill the time span and missing values. The padding method is using the most recent data that is in front of the time of the missing data. At the same time, because different types of data have different units, the magnitude may vary greatly. Therefore, each factor is dimensionless through a standardized method. The method is to subtract the data minmun from the original data and divide it by the scalar of data. The minmun of final data is 0 and maximun is 1.

In the process of feature processing we have adopted two methods.

Method 1 uses PCA to process the data, the purpose of which is to reduce the data dimension while retaining as much information as possible. The principle of PCA is to perform eigenvalue decomposition on the covariance matrix of the original data with n factors, and obtain n eigenvectors and their corresponding n eigenvalues. The eigenvectors are arranged according to the corresponding eigenvalues from large to small. Adding the variances corresponding to the first m eigenvectors until the cumulative variance exceeds 95%, and using the m vectors as the m-dimensional principal components of the data set, the m-dimensional principal component at this time can explain the variance of original data set by more than 95%, which means it contains more than 95% of the information in the original data set.

First, PCA is performed on more than 300 sets of features in the original data set. The advantage of such dimensionality reduction is that the calculation amount of the model can be reduced and the operation speed can be improved. Although in the implementation process, we can get the linear mapping relationship between each principal component and the original factor, such features can be poorly interpreted, and we cannot intuitively explain which principal components are used in the model. So we first classify the factors according to their attributes, and then perform principal component analysis separately in each category. The number of original factors for each category, the number of factors after dimension reduction, and the interpretable variance are as follows:

We classify the 352 original factors into six categories, and obtained a total of 48 categories of principal components as a new feature transfer model.

Method 2 uses RFE(Recursive Feature Elimination) to perform feature processing. The purpose is to screen out the predictive variables with the best prediction effect and the strongest sensitivity to achieve the best prediction effect by using the most compact model. The principle of recursive feature selection is to eliminate a factor in each cycle. The condition for eliminating the factor is that the prediction effect of the model is optimally improved after the factor is removed. The termination condition of the cycle is that

TABLE I
INTRODUCTION TO RAW DATA FORMAT

Data name	Data frequency	Time span	Data sources
Silver ETF(Exchange Traded Funds)	Day	2008-2018	English fortune
DNR.N oil index	Day	2008-2018	Wind
Car production	Month	2008-2018	Wind
Huang Chunfa barreled latex	Day	2008-2018	Choice
Ex-factory price, butadiene rubber, Jinzhou Petrochemical	Day	2008-2018	Wind
Rainfall	Month	1991-2015	World Bank Climate Change Portal
600688.SH(Shanghai) Shanghai Petrochemical	Day	2008-2018	Wind
...	...	...	...

TABLE II
NUMBER OF FACTOR CATEGORIES BEFORE AND AFTER PCA

	Rubber	Financial	Metal	Energy	Automobile	Food
Number of primitive factors	258	12	11	26	26	19
Number of factors after PCA	10	6	4	8	10	10
Interpretable variance	0.95	0.95	0.95	0.95	0.95	0.95

eliminating any factor of the residual factor the prediction effect of the model will decrease. And the feature screening method needs to be selected in combination with the model.

Since other factors and rubber prices occur at the same time, it is not possible to predict the rubber price of the day with the factor of the day when forecasting. At least the factor of the previous day should be used to predict the rubber price of the day. In addition, since the rubber price is time series data, historical data will also have a great impact on the rubber price. Therefore, it is not possible to simply use a certain day data to predict the rubber price. The moving average data of historical data should be added to improve the forecasting effect. For the above two scenarios, we added the previous day data, the exponential average of the previous 3 days of data, the exponential average of the previous 7 days of data, the exponential average of the previous 14 days of data, and the exponential average of the previous 28 days of data, corresponding to the actual cycle of company procurement: daily, weekly and monthly.

C. Data Set Partitioning and Evaluation Indicators

This paper uses the data between 2008 and 2016 data as training sets and take the data from 2017 to 2018 as test sets. The indicator used in the training is Mean Squared Error (MES).

$$MES = \frac{1}{M} \sum_{m=1}^{M} (y_m - \widehat{y}_m)^2 \qquad (1)$$

Generally, the lower the MES value, the better the effect. Where M is the number of samples, y_m is the actual value of the m^{th} sample, and $\widehat{y}_m$ is the predicted value of the m^{th} sample.

In the process of training the model, we first use the three-fold cross-validation method. The MES score on the set is repeated three times, so that each piece of data that has been separated becomes a cross-validation set, and the average of the three MES scores is taken as the cross-validation score. For the hyperparameters of each model, we use the grid search method (Grid Search) to search for the optimal parameter combination. The specific method is: put the candidate values of the hyperparameters to be adjusted into the list, each time of training selects one of the combination of the alternative values for training until all the alternative values are trained once by the three-fold cross-validation method, and the parameter combination with the smallest cross-validation score is taken as the preliminary optimal parameter. The next step is to narrow the range and step size near the value of the initial optimal parameter combination and perform the above steps again until the accuracy of each parameter reaches the expected value. At this time, the prediction effect and generalization of the model trained by the optimal parameter combination is the strongest. Thereafter, a comprehensive selection model was obtained based on the scores of the different models in the cross-validation set and the test set. If only based on the performance of the test set, the trained model will have a serious over-fitting problem, which makes it perform very well in the training set, but the prediction ability of new data is very poor. The reason why using three-fold verification is to make full use of the training set data, and to create three kinds of slightly different data distribution, so that the comprehensively trained model will have lower degree of overfitting problem than the one that relies on the test set.

III. THE PROPOSED FRAMEWORK

In the experiment process, different single models are used to predict the results, but different algorithm models have different performance characteristics. After comparing the experimental results of every single model, we selected Random Forest, Adaboost and LSTM, and merged the results to improve the overall prediction effect. Model fusion refers to the integration of different model results obtained by different models for prediction, different samples for training, or different features for feature processing on the same dataset. The Stacking fusion method adopted in this paper, also called the multi-layer fusion model, is to input the output of multiple different sub-models in the first layer as the feature of the second-layer model, and then retrain the second layer in comparison with the real value. The second layer model is called the super model. Its function is to learn the performance of each sub-model of the first layer model, and give the sub-model different weights according to the error of the sub-model. Therefore, if a single model performs poorly, it will not necessarily have a detrimental effect on the results. Since different models have different understandings of the total eigenvalues on the same dataset, that is, different models assign different weights to the same features, the fusion model of the second layer has more sufficient properties for each original variable, the two-layer model tends to be better than the individual sub-models in the first layer, which also shows that the multi-model fusion has better predictive power than the single model. The algorithm flow framework is shown in Figure 4.

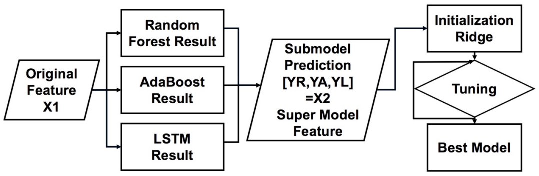

Fig. 4. Algorithm Flow Framework.

A. Base Model Selection

1) Single Model Comparison: The introduction of the Random Forest, Adaboost, and LSTM methods is not described in this article. Here, we mainly introduce the comparison of the logic performance characteristics of each algorithm.

- **Random Forest**
  Advantages: 1) the learning process is very fast; 2) the classifier is highly accurate; 3) can handle a large number of input variables. *Disadvantages:* 1) Overfitting on some of the more noisy problems; 2) Attributes with more values divided will yield lower confidence in attribute weights.
- **Adaboost**
  Advantages: 1) combined with weaker models; 2) slow over-fitting; 3) simultaneous reduction of deviations and fluctuations. *Disadvantages:* 1) More parameters need to be adjusted; 2) Easy to be affected by outliers.
- **LSTM**
  Advantages: 1) higher learning rate; 2) automatic search for new feature values. *Disadvantages:* 1) The amount of data required for training is large; 2) The computing power consumed during work is large.

2) Stacking Fusion Algorithm Strategy: Stacking can transfer the integrated knowledge to a simple classifier, and effectively prevent over-fitting, without require too many calculation. At the same time, because the application of XGBoost(eXtreme Gradient Boosting) [8] model in Stacking is more effective for model improvement, we use XGBoost model as the second layer model. The specific process of the Stacking method is shown in Figure 5.

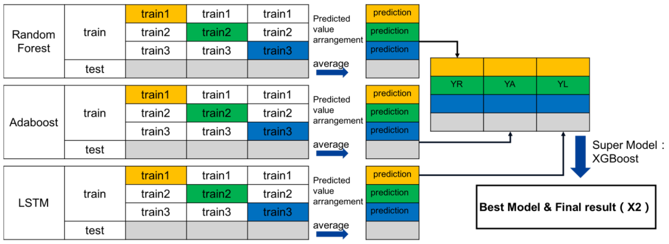

Fig. 5. Stacking Method Flow.

First, divide the training set into 3 parts: train1, train2, train3; then select the base model, we select the Random Forest, Adaboost and LSTM which have been tested well, and select train1, train2, train3 as the verification set, and the rest. Two training sets were used as the training set, and the model was trained by 3-fold cross-validation. The predicted results of the base model YR, YA and YL were used to train with the XGBoost model to establish the XGBoost model. Finally, the optimal model parameters and prediction results were selected through the tuning parameters.

B. Sensitivity Matrix

The vector obtained by PCA has orthogonal characteristics, that is, the problem of correlation between various influencing factors on the original data set is eliminated, and then the dimension reduction is adopted in the process of analyzing the influencing factors of natural rubber price trend. The latter data is used as an independent variable to avoid the problem of mutual interference between the analyzed independent variables. In the work of quantifying the influencing factors, the main focus is on the impact of macro factors on the price trend of natural rubber. Through the use of PCA dimensionality reduction data, the linear regression of natural rubber price trends is obtained, and the main component of each dimension reduction is obtained. The linear weights are multiplied by the matrix of the influencing factors of the individual eigenvectors obtained during the PCA dimension reduction process, and then quantified to obtain the weight of each original influencing factor in linearly fitting the natural rubber price trend.

The mathematical expression of the above method is as follows. It is assumed that the influencing factors of m group length n need to be analyzed, so that they constitute the original influencing factor matrix, which is recorded as B_{m*n}; the matrix is transformed into k group unrelated variables by PCA, which is recorded as A_{k*n}; Since the PCA dimensionality reduction is a linear change process, its linear change matrix is denoted as P_{k*m}.

$$A_{k*n} = P_{k*m} B_{m*n} \tag{2}$$

The target values are linearly fitted using the uncorrelated variables obtained by dimensionality reduction, and the weight matrix w_{1*k} and the deviation b_{1*k} are obtained.

$$y_{1*n} = w_{1*k} A_{k*n} + b_{1*k} \tag{3}$$

The linear fitting weight matrix is multiplied by the PCA linear transformation matrix to obtain the weight matrix of the influencing factors, which is denoted as W_{1*m}.

$$W_{1*m} = w_{1*k} P_{k*m} \tag{4}$$

The weights of the influencing factors obtained through the above calculations are shown in Table III.

From Table III, we can see that the industrial environment has a strong positive impact on the price trend of natural rubber. The exchange rate of RMB against the US dollar and the economic environment represented by shanghai composite is negatively correlated with the price of natural rubber. When the economic environment weakens, the price of natural rubber will increase. At the same time, as the natural rubber of agricultural products is dispersed due to the origin, we directly use the global rubber production as its characteristic value, and from the perspective of its impact weight, it will not have a greater impact on price fluctuations. However, as a demand, the Chinese rubber consumption of the end data is negatively correlated with the price trend. This is very intuitive. When demand increases, the price will naturally fall.

C. Analysis of prediction results

1) Evaluation of prediction results: We take the actual natural rubber price trend in Malaysia in 2008-2018 as the target for model testing, as shown in Figure 6. We selected the data from 2008/01-2016/10 as the training set and the data for 2016/10-2018/05 as the test set.

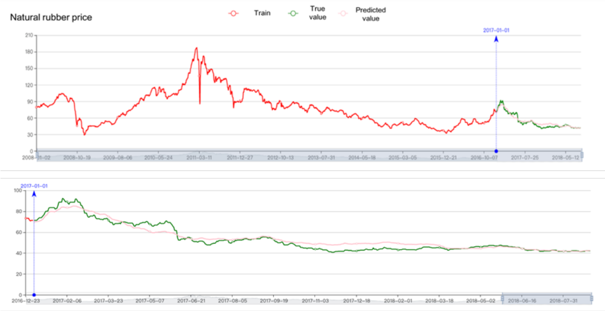

Fig. 6. Effect of the Fusion Model based on Stacking on the Prediction of Natural Rubber Price.

In Figure 6, we selected the Stacking fusion model with XGBoost as the super model and the prediction of the 28-day offset with the time span as the final result. The upper part of Figure 6 contains the training set and test set and prediction results. At the same time, based on the different demands of natural rubber procurement decision makers on the predicted frequencies, we provide model prediction results with time offsets of 1 day, 14 days and 28 days respectively. The specific prediction effect will be explained in the next section, along with the performance comparison of different models.

2) Comparison with other models: During the experiment, we separately predicted the performance of the single model, the stacking-based fusion model proposed in this paper, and compared to other papers, such as EMD-LSSVR-ADD (Zhu et al. 2017) [9] and Lasso Selection (Miao et al. 2017) [10]. Moreover, the MES method was used to compare the time offsets of 1, 14, and 28 respectively, and the comparison results are shown in Figure 7.

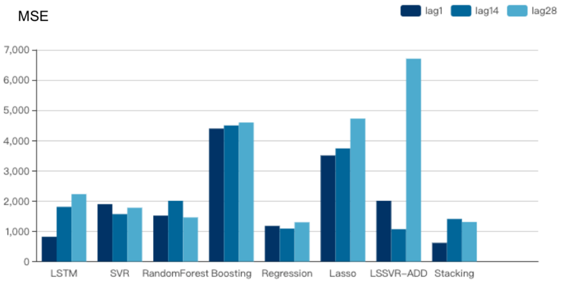

Fig. 7. MES Comparison Results of Different Methods and Time Offsets.

It can be clearly seen that the Stacking fusion model has improved performance at different time offsets compared to some typical single algorithm, especially in the model with offset 1 being the most obvious.

In the comparison of complex models, we selected EMD-LSSVR-ADD based on carbon price prediction and Lasso Selection based on oil price prediction. Both models performed well in their respective prediction scenarios, but were not ideal for natural rubber. The main reason is that the forecasting problem has strong knowledge requirements in the professional field, and different forecasting models have strong pertinence. So how to realize the transfer of knowledge is also the focus of the work.

D. Model application effect analysis

In actual procurement activities, the procurement cycle is basically on a monthly basis. To this end, we compare the historical procurement strategy of the actual enterprise with the procurement strategy at the lowest point of the forecasting result, and randomly select the results of three months, as shown in Figure 8.

Through comparison, we can find that our forecast results can bring about a 4%-6% reduction in procurement costs, which has very important practical significance. The original purchasing strategy of the enterprise is formulated by the purchasing staff, and the data source is the enterprise's enterprise resource planning system.

TABLE III
INFLUENCING FACTORS WEIGHT TABLE

Type of influence factor	Influencing factor	Influence weight
Industrial situation	PPI	0.421703
Industrial situation	Second industry added value (100 million yuan)	0.072097
Supply side	World rubber production	0.027035
Upstream and downstream industries	Automobile production (10,000 units)	0.025995
Economic development	GDP (100 million yuan)	0.024567
Major energy industry	World oil production	-0.06174
Demand side	China's rubber consumption	-0.08788
Economic environment	RMB against the U.S. dollar	-0.14952
Economic environment	Shanghai Composite Index closed	-0.38411

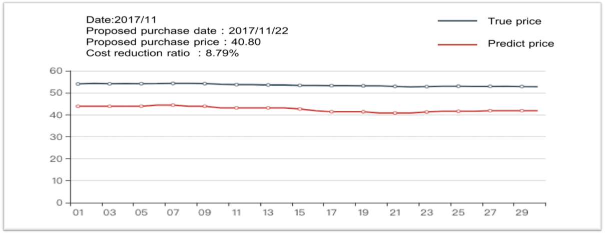

Fig. 8. Natural Rubber Procurement Strategy Formulation and Effect.

IV. RELATED WORK

Tan [11] used an econometric model of natural rubber and synthetic rubber to predict natural rubber prices. The results show that among all explanatory variables, the price of natural rubber stocks in major consuming countries, the demand and price of natural rubber are the most important explanatory variables in natural rubber price prediction model.

Krichene [12] pointed out that natural palm oil is related to the nominal effective exchange rate of the US dollar and the US international exchange rate. The study used a simultaneous equation model for the natural palm oil and natural gas market and found that interest rates and exchange rates are inelastic in short-term prices, resulting in natural palm oil and natural gas.

Shamsudin, M. N & Fatimah [13] provided short-term advance forecasts for Malaysian natural palm oil prices. The results show that the prediction of the MARMA model is more efficient than the prediction of the econometric model.

Burger and Smit [14] studied short-long-term analysis of the natural rubber market. It covers economic and price fluctuations of key countries on both supply and demand of natural rubber. When additional trees are put into production, prices will fall until they fall to the 2000 standard level.

V. SUMMARY AND OUTLOOK

This paper analyzes the main influencing factors of natural rubber price, and completes the basic work of data collection, sorting and storage. A fusion algorithm model based on Stacking is proposed, and the model is tested on the data set. Compared with other similar algorithm models, the excellent performance of the model proposed in the trend forecast of natural rubber price is proved. The comparison of the actual procurement strategies confirms that the model plays an important role in reducing the procurement cost of enterprises in practical applications. At the same time, in the process of analyzing the sensitivity of data features, a sensitivity matrix acquisition method based on PCA and regression analysis is proposed as a general sensitivity test method.

There is more work for further exploration in our research work. In the future research direction, the remaining points can be continuously optimized: 1) Optimize data collection process and standardize data processing. 2) Develop model dynamics, support dynamic update and real-time prediction, 3) Develop new data feature, 4) Explore knowledge transfer and promote the association of industrial mechanism modeling methods with other industrial knowledge.

REFERENCES

[1] Brian Sidney Everitt. Principal Components Analysis[J]. 2005.

[2] Zhang X , Lu X , Shi Q , et al. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data[J]. BMC Bioinformatics, 2006, 7(1):197-0.

[3]]Liu J , Shang W , Lin W . Improved Stacking Model Fusion Based on Weak Classifier and Word2vec[C]// 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS). IEEE Computer Society, 2018.

[4] Yuan song, 2018. Stock Trend Prediction: Based on Machine Learning Methods. UCLA.

[5] Zhu J , Arbor A , Hastie T . Multi-class AdaBoost[J]. Statistics & Its Interface, 2006, 2(3):349-360.

[6] Huang C J, Kuo P H. A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities[J]. Sensors, 2018, 18(7):2220-.

[7] Zhu J , Arbor A , Hastie T . Multi-class AdaBoost[J]. Statistics & Its Interface, 2006, 2(3):349-360. [8] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. 2016.

[8] Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. 2016.

[9] Zhu B , Han D , Wang P , et al. Forecasting carbon price using empirical mode decomposition and evolutionary least squares support vector regression[J]. Applied Energy, 2017, 191:521-530.

[10] Miao H , Ramchander S , Wang T , et al. Influential Factors in Crude Oil Price Forecasting[J]. Energy Economics, 2017:S0140988317303134.

[11] C.S. Tan, World rubber market structure and stabilization:An econometric study, World Bank Staff Commodity Papers, No. 10, 1984.

[12] N. Krichene, A simultaneous equations model for world crude oil and natural gas markets, IMF Working Paper WP/05/32, 2005.

[13] Shamsudin, M. N & Fatimah, 2000. Short Term Forecasting of Malaysian Crude Palm Oil Prices[J].

[14] Burger, K., and H. P. Smit. 2000. Long-Term and Short-Term Analysis of the Natural Rubber Market. Department of Econometrics, Economic and Social Institute, Faculty of Economics and Business Administration, Vrije University, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.