Research on Big Data Power Load Based on Neural Network Prediction Model

Mingyi Ji [1, *], Jiazheng Kang [2]

[1] Nanjing University of Posts and Telecommunications, Nantong, China

[2] Shandong University of Finance and Economics, Jinan, China

* Corresponding Author Email: jimy_email@163.com

Abstract. The price of sailboats will change with time and market changes. Reasonable prediction and evaluation of the price of second-hand sailboats can have a certain impact on the price of second-hand sailboats. The better ones in the area are for sale.The paper established the RF_XGBoost second-hand sailboats price prediction model, determined the feature variables from the type of sailboat, regional economy, the random forest algorithm is used to select the feature variables, and the optimization is based on the grid search algorithm after the data normalization process obtained the minimum value.After analyzing the characteristic variables it is found that the region has little influence on the listed price, and thus has little influence on the features. Meanwhile, the Pearson correlation analysis between the variables and the price shows that there is a weak negative correlation consistency between the region and the type of sailing boats.

Keywords: RF_XGBoost prediction, grid search, Random Forest Feature Vector Selection, Coefficient of variation.

1. Introduction

Sailing boats [1-2] are very useful in many fields, and the selling price of sailing boats is closely related to other factors such as region and time. For buyers of sailing boats, it is very important to understand the sales information of sailing boats.

In Europe, the Caribbean and the United States, the sales of sailboats are variable, and the sale of monohulls and catamarans [3] has attracted much attention. There are many factors that affect the sale of second-hand monohulls and catamarans, such as the price of sailboats. Factors such as make, variant, length, geographic area, etc. all have an impact on the price a used sailboat sells for.

The forecast of the sales price of sailboats and the regional analysis of sailboats has certain references for sellers and buyers of sailboats on how to reasonably price and purchase sailboats. In addition, whether the influence of regions on sailboats is consistent remains to be explored. This article uses RF_Xgboost [4-7] related research.

2. Feature-optimized RF_XGBOOST sailing price prediction and evaluation model

2.1. Feature variable selection

Predicting the second-hand price of sailboats needs to consider many factors. The selling price of sailboats is analyzed from the regional economic factors, the type of sailboats, and the factors of the sailboat itself.

2.2. Data Merging and Normalization

The quality of the same model should be the same. Therefore, this paper merges the data according to the Make, Variant, and Country/Region/State characteristic variables. If the three columns are the same, they are added and averaged to represent a certain the general price of a second-hand sailboat in a certain area.

Different types of data are represented by numbers, and string data is processed by using label encoding.First calculate the difference between the year of manufacture and the year of sale.The year gap, manufacturer, type of sailboat, and country and region sold are digitized separately.

The minimum value normalization process [8] is performed on the selected feature variables, and the feature variables are normalized to(0,1)between.

2.3. Random forest feature selection

In this paper, based on the integrated learning theory of bagging, *RF_XGBoost* multiple decision trees are used to form a collection of "a forest", and the importance of feature variables is calculated through random forests, and feature selection is performed, and the combination of better feature variables is used as feature set input *XGBoost*. Through decision tree majority voting, random sampling with replacement is adopted to predict the price of second-hand sailboats.

The random forest is used to screen the original features, and *OOB* the feature selection principle is adopted, and the feature set is input into the XGBoost prediction model to improve the prediction accuracy of second-hand sailboats. The feature screening process of the random forest is as follows:

Measures of Feature Importance:

(1) For each decision tree, randomly select the corresponding training data and calculate the prediction error $errOOB_1$.

(2) Randomly x add feature disturbance to all sample features of the training data and calculate the prediction error again $errOOB_2$.

(3) N A random forest of trees with feature X importances of:

$$import = \frac{1}{N}\sum_{i=1}^{N} | errOOB_1 - errOOB_2 | \tag{1}$$

Among them: the larger the value, it means that the prediction error after noise is added is larger than that without noise, and the characteristic variable is very important to the sample prediction result.

Feature screening:

(1) Calculate the importance of each feature, sorted in descending order;

(2) Determine the elimination ratio, and remove the corresponding proportion of features according to the feature importance to obtain a new feature set

(3) Repeat the above process until the remaining m feature sets

(4) Calculate the prediction error of the feature set, and select the feature set with the smallest prediction error.

2.4. Grid search parameter optimization

Using the normalized data, it is divided into training set and test machine[9]. The training set accounts for 80%, and the test set data accounts for 20%. It is divided into 10 parts, 9 of which are used as model training sets, and the other 1 is used as a model verification set. The grid search algorithm is used for parameter optimization. Since the data volume of the data set and the training set are constant, this paper mainly focuses on N_estimators and Max_depth . search for excellence.

Algorithm 1
Input: Used Sailboat Data
Define the parameter search space D
Initialize the model
While: the convergence condition is not satisfied
slice optimization space
Find the optimization space, select representative points, calculate the order of target values, and eliminate some poor points
The first point of the heap sort divides the new optimization space

Include part of the elimination points and the representative points in the new optimization space to merge and calculate the target and sort
satisfy:
Output optimal parameters
Output square error MSE_0, fit index R_0^2, root mean square error $RMSE_0$

2.5. Prediction of used ship price based on RF_XGBoost

XGBoost is k an additive model composed of four weak classifiers. In this paper, the effects of all classifiers are averaged, and the corresponding predicted value is:

$$y_i = \sum_{t=1}^{K} f_t(x_i) \tag{2}$$

Where K is the total number of decision trees, f_t represents the first t tree, and y represents the result of the model predicting the sample x_i.

In this paper, it is defined n as the number of training samples, and y_i the general loss function is determined by the deviation between the predicted value L and the real value y_i:

$$l = \sum_{i=1}^{n} l(y_i, y_i) = (y_i - y_i)^2 \tag{3}$$

Due to the limited amount of data, in order to stabilize the variance and prevent over-fitting, L1 and L2 regularization items are added to the loss function. The regularization items of each decision tree are:

$$\Omega(f_t) = \gamma T + \frac{\lambda}{2}\sum_{j=1}^{J} w_{tj}^2 \tag{4}$$

Among them, T is the number of leaf nodes, is w_{tj} the weight of the leaf nodes of γ the first tree, t and λ is a hyperparameter. The complexity of the model is determined by the leaf nodes and the weights of the leaf nodes. Therefore, the number of leaf nodes in this paper should not be too much , and the weight of the leaf nodes should not be too extreme.

In this paper, the loss function of all decision trees is added to the regularization term, and the target loss function is:

$$Obj = \sum_{i=1}^{\mathrm{m}} l(y_i, y) + \sum_{i=1}^{t} \Omega(f_i) \tag{5}$$

During the first t iteration, the model's x_i predicted value for the sample is:

$$y_t^{(t)} = y_t^{(t-1)} + f_t(x_i) \tag{6}$$

Therefore, at t step, $y_t^{(t-1)}$ is a constant value, so the objective function is rewritten as:

$$\begin{aligned} Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_{i=1}^{t-1} \Omega(f_i) + \Omega(f_t) \\ &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + c \end{aligned} \tag{7}$$

According to the second-order expansion of Taylor's formula, it can be $l(y_i, y_t^{(t-1)} + f_t(x_i))$ rewritten as:

$$l(y_i, y_t^{(t-1)} + f_t(x_i)) \approx l(y_i, y_t^{(t-1)} + g_i f_t(x_i)) + \frac{1}{2} h_i f_t^2(x_i) \tag{8}$$

Among them, g_i is the first derivative of the loss function, h_i is the Second derivative of the loss function, according to the error loss function, deduce:

$$\begin{aligned} g_i &= \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} = -2(y_i - \hat{y}^{(t-1)}) \\ h_i &= \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} = 2 \end{aligned} \tag{9}$$

For each decision tree $f_{(t)}(x)$, there are:

$$f_t(x) = w_{q(x)}, w \in R^T, q : R^d \longrightarrow \{1,2,...T\} \tag{10}$$

Among them, w is the leaf node score value; $q(x)$ is the leaf node corresponding to T; x is the number of leaf nodes. Therefore, putting formula (11) into (8), after simplification, we can get:

$$\begin{cases} Obj^{(t)} = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T \\ G_j = \sum_{i \in I_j} g_j \\ H_j = \sum_{i \in I_j} g_i \end{cases} \tag{11}$$

Where w_j is j the score value of the th leaf node.

Since the target loss function $Obj^{(t)}$ is a quadratic equation in one variable, its extreme value is:

$$\begin{aligned} w_j^* &= -\frac{G_j}{H_j + \lambda} \\ Obj^{(t)} &= -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \end{aligned} \tag{12}$$

When building the first t tree, introduce the split income and update the maximum income:

$$Gain = \max(gain, \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G_L^2 + G_R^2}{H_L + H_R + \lambda}\right] - \gamma) \tag{13}$$

Therefore, the algorithm flow chart of RF_XGBoost is as follows:

Algorithm 2 RF_XGBoost
1.Input the training sample set, select the original feature variable as the independent variable, and the third-hand ship price as the dependent variable output 2.Sort the importance of features after random forest processing, and select the top n feature quantities as feature variable input 3.Divide the data set and training set, and use validation for model training 4.Input XGBoost. classifier for learning and training, and constantly adjust model parameters to improve prediction accuracy

2.6. Evaluation of model implementation results

In order to verify the accuracy of the *RF_XGBoost* model prediction[10], the fitting coefficient R^2, mean square error MSE, and root mean square error $RMSE$ are selected as performance indicators for evaluation:

$$
\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)} \\
MSE &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \\
RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}
\end{aligned} \tag{14}
$$

2.7. The optimal parameter value

According to the probability experience, define the grid optimization interval *Interval* and optimization step size of monohull sailboat and catamaran sailboat *Step*, use 10-fold cross-validation for data training, and use grid search to optimize the parameters to obtain the optimal parameter value, as shown in Table.1.

Table 1: Optimal Results of Optimizing Parameters

parameter	Monomer interval	*Step 1*	The optimal value	Interval	*Step 2*	The optimal value
N_estimators	(100,200)	5	115	(50,200)	8	58
Max_depth	(9,25)	1	10	(5,30)	2	11

Then the importance of the characteristic variables of monohull sailing boats and catamaran sailing boats based on random forest screening is obtained:

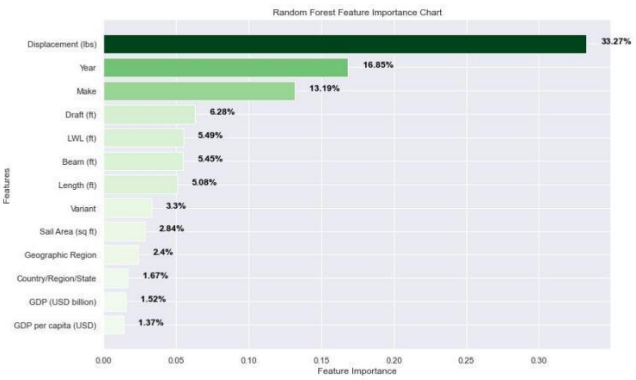

Fig 1. Monohull sailing ships

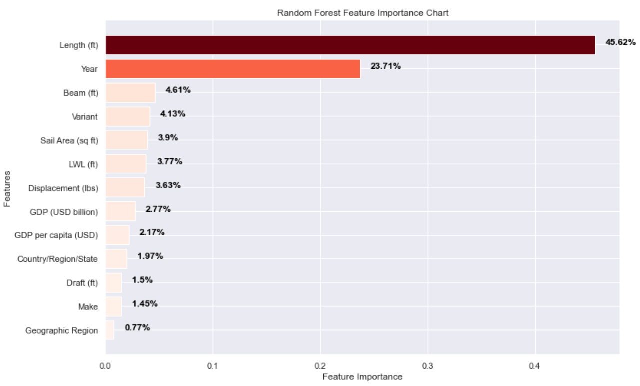

Fig 2. Catamaran sailing ships

From Fig. 1 and Fig. 2, it can be seen that for monohull sailboats, the most important characteristic variable is draft (Displacement), and for catamaran sailboats, the most important characteristic variable is the length of sailboats (Length). In practical applications, when designing monohull sailboats, , the draft of the ship often has a direct impact on the performance of this type of ship, and the length of the sailboat and the length of the waterline will have an impact on the performance of the catamaran sailboat. Therefore, the results are quite reasonable.

According to the results of the importance of feature variables, this paper selects the previous $n = 10$ feature variable as a feature set input into *XGBoost*, divides the training set and test set into a ratio of 8:2, based on the optimal parameter training, denormalizes the output data, and uses the evaluation metrics evaluate the model, as shown in Table.2:

Table 2: Evaluation Metric Results

Evaluation Index	Monohull sailing boat	Catamaran
$R_0^{\ 2}$	0.7525	0.7851
MSE_0	2535246520.9612	6207321336.9541
R^2	0.7639	0.8166
MSE	2195009295.7677	5575612469.0050
$RMSE$	46850.9263	74670.0239

The test set data prediction effect diagram is as shown in Fig.3 and Fig.4:

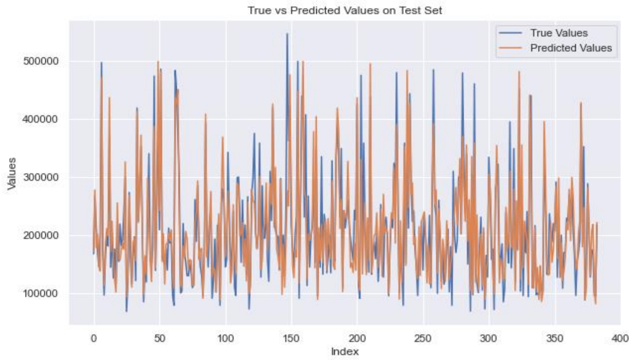

Fig 3. Monohull sailboat price forecast

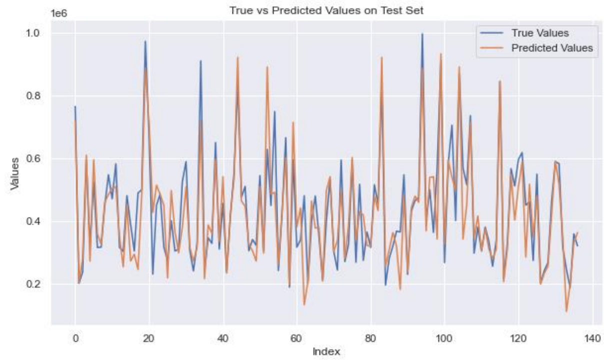

Fig 4. Catamaran sailboat price forecast

The fitting coefficients for monohull sailboats and body sailboats R_0 are both close to 0.8, and the fitting effect is good. It can be seen that compared with random forest, *RF_XGBoost* significant progress has been made in the fitting coefficient and mean square error.

The model also does a good job of predicting off-center points in used price predictions for monohulls and catamarans.

2.8. Analysis of regional effects

In order to assess the impact of the interpretation area on the price, first use the scatter plot to visualize the price distribution of second-hand monohull sailboats and catamaran sailboats in different countries. The price distribution in different regions is visualized as follows Fig 5 and Fig 6:

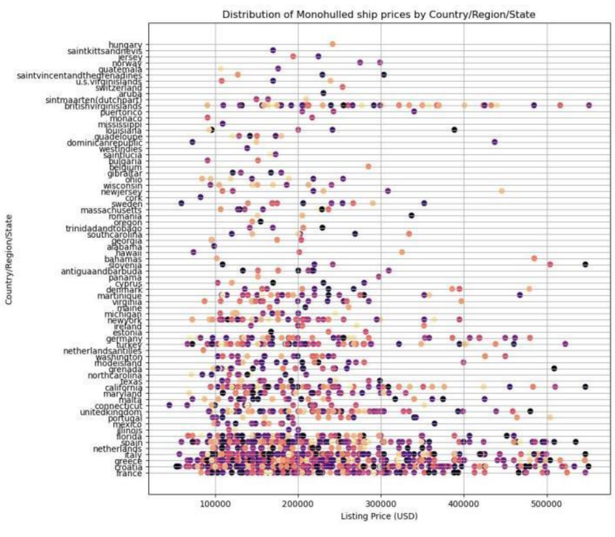

Fig 5. Scatter diagram of monohull sailing prices

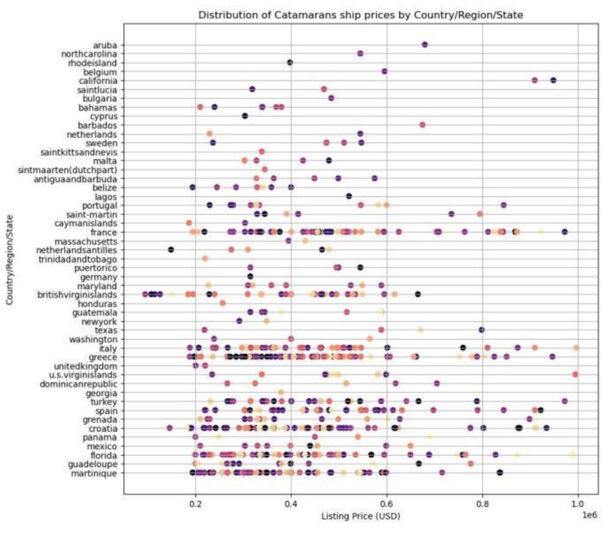

Fig 6. Scatter diagram of catamaran sales

From the preliminary scatter diagram of the regional distribution of selling prices, it is observed that the sales volume of monohull sailing boats and catamarans is higher in Germany, Italy, France and other countries, and there are more second-hand sailing boats sold at high prices. , Bahamas, a country with a small population distribution, has less sales volume, and most of the ships sold are at low prices. Therefore, this paper can preliminarily determine that regional factors have an impact on the listing price of second-hand ships.

Therefore, this article defines the regional impact index $Area$ to measure the influence of the region on the price:

$$Area = \frac{W_{Geographic\_Region+} W_{Country/Region/State}}{2} \tag{15}$$

Among them:,the importance of W_i the first i index

Furthermore, according to the random forest for feature variable selection, the importance analysis of the feature map shows that the importance of the region in the monohull sailboat is, the importance of the region 2.04% in the 2.07% catamaran sailboat is the impact is small.

2.9. Consistency analysis of regional effects

Since the selling price of second-hand sailboats is generally distributed normally, in order to further analyze the relationship between sailboat models and regional factors, the normalized data in this paper is used for correlation analysis of characteristic variables, and Pearson correlation coefficient is used to analyze the correlation coefficient of the matrix r . Plot a correlation coefficient heatmap:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{16}$$

Correlation coefficient heat map:

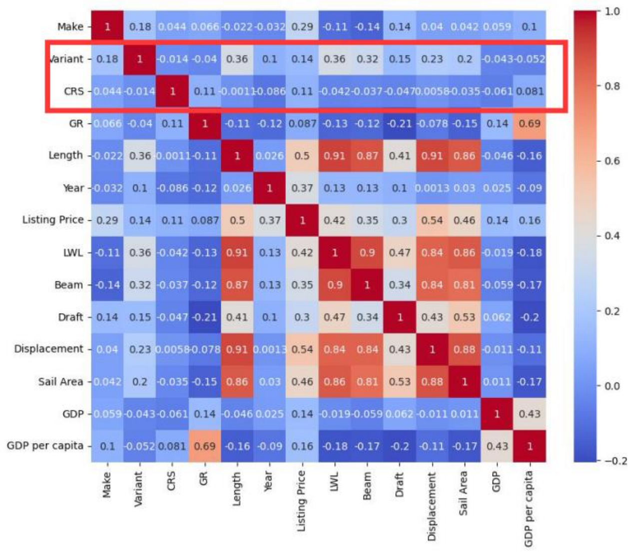

Fig 7. Monohull sailboat

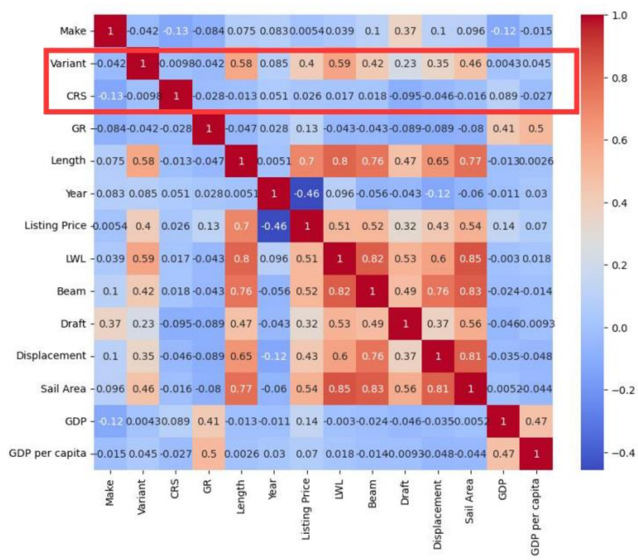

Fig 8. Catamaran sailboat

As is shown in Fig 7 and Fig.8, according to the correlation coefficient between the country and region of sales and other characteristic variables, the regional correlation coefficient between the region of the monohull sailboat and the catamaran sailboat and other factors is low, and the influence of the region on other factors has a consistent low impact sex.

3. Conclusions

In this paper, by establishing a second-hand ship price prediction model, the characteristic variables are determined from the types of sailboats, the characteristics of the sailboat itself, and the regional economy. After cleaning and normalizing the data, the random forest algorithm is used to select the characteristic variables. The algorithm finds the optimal parameters. The training set uses 10-fold cross-validation to obtain the importance ranking of the feature vectors under the optimal parameters. Select the feature vectors with the highest importance as the feature set input model. The square error and the root mean square error are used as performance indicators to predict and evaluate the listing price of second-hand ships. The evaluation shows that the model in this paper has very good performance.

Since the time complexity of the model in this paper still needs to be optimized, the fitting effect needs to be further improved. In the next work, this paper will try to improve the accuracy and generalization performance of the model as much as possible.

References

[1] A. Chongterdtoonskul and W. Tharmmaphornphilas, "Offshore Crew Boat Sailing Time Forecast using Regression Models," 2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA), Bangkok, Thailand, 2020, pp. 883-889.

[2] S. Guo, "Dynamic simulation and Analysis of Propulsion Performance of Intelligent Unmanned Boat," 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Shenyang, China, 2021, pp. 653-655.

[3] R. M. Balinda, S. Allamsetty and T. R. Choudhury, "Sizing an Electric Propulsion System for Catamaran Passenger Boat in Uganda," 2022 IEEE India Council International Subsections Conference (INDISCON), Bhubaneswar, India, 2022, pp. 1-6.

[4] Cai Qiuru. Electric Vehicle Price Prediction Based on XGBoost Integrated Learning Algorithm [J]. Journal of Jiangsu Institute of Technology, 2020, 26(06): 33-43.

[5] J. Song, L. Jin, Y. Xie and C. Wei, "Optimized XGBoost based sparrow search algorithm for short-term load forecasting," 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE), SC, USA, 2021, pp. 213-217.

[6] Ouyang Han, Liao Wenqi, Tian Qiuhong. Electric Vehicle Price Prediction Based on Machine Learning [J]. Guangxi Quality Supervision Herald, 2021, No.244(04):204-206.

[7] Li Fuqiang, Peng Haili, Yang Xi, Zhang Wenjing. Used car price prediction model and impact analysis based on deep learning [J]. Journal of Automotive Engineering, 2021,11(05):379-385.

[8] R. Oji, S. F. Razavi, S. A. Dehsorkh, A. Hariri, H. Asheri and R. Hosseini, "ParsiNorm: A Persian Toolkit for Speech Processing Normalization," 2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS), Tehran, Iran, Islamic Republic of, 2021, pp. 1-5.

[9] Zheng Jie. Prediction of Used Car Price Based on Random Forest and XGBoost Algorithm [J]. Digital Technology and Application, 2021, 39(06): 90-93+188.

[10] J. Jin, C. Zhang and Q. -J. Zhang, "An enhanced multi-circuit optimization approach for identification of microwave filter parameters," 2016 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO), Beijing, China, 2016, pp. 1-2.