A used car repricing method based on K-Means++ clustering and multiple linear regression

Jiajia Meng[1,a,*]

[1]School of Management, Qufu Normal University, Rizhao, Shandong, 276826, China
[a]m2220707543@163.com
[*]Corresponding author

Abstract: *A predictive pricing method based on k-means++ clustering and multiple linear regression is proposed to solve the unreasonable initial pricing of used cars and evaluate the model's low accuracy. To begin, all vehicles are divided into three categories based on the transaction period of the data sample. The K-Means++ clustering algorithm is then used to cluster the best-selling cars, and a multiple linear regression equation is fitted for each category as the regression equation for prediction and evaluation. Finally, the revaluation of unsold and unsalable vehicles is examined. In addition, two used car valuation methods, XGBoost and AdaBoost, are compared. Because the model performs well, the experiment shows that using the multiple linear regression method based on clustering to estimate car price is reasonable.*

Keywords: *Used cars; Transaction period; K-Means++ clustering; Multiple linear regression*

1. Introduction

Currently, there is no unified national standard for appraising and pricing second-hand cars, nor are there any national policies or regulations for their appraisal and pricing methods. Commonly used appraisal methods include the replacement cost method, present value of income method, current market method, and liquidation price method. Of these, the current market method is based on data obtained from the second-hand car market, which provides an objective reflection of the current situation of the vehicle market. With the advancement of computer technology and artificial intelligence science, machine learning and deep learning-based current market evaluation methods for used cars have become the mainstream decision-making tools for practitioners.

Literature[1-3] proposed a price estimation model of BP neural network based on big data, but the model only considered few factors affecting price, and its accuracy was limited. In literature [4], PCA, random forest, and GDBT are used for feature selection, and SVM support vector machine prediction model is used as the evaluation model. However, the prediction effect of SVM model for cars with different price levels is not stable. Literature [5, 6] only considers a few variables affecting the price and adopts a multiple linear regression model for valuation. Literature [7, 8] proposed a random forest-based valuation model. In 1995, Yoav Freund and Robert Schapire proposed the idea of the Boosting algorithm[9], which is a classification method that combines weak classifiers to obtain a strong classifier with greatly improved classification performance. Literature [10-13] compares and analyzes three Boosting models, XGBoost, LightGBM, and AdaBoost, and other traditional machine learning methods. The Boosting method is superior to other traditional machine learning methods in terms of accuracy and time.

However, all the above evaluation methods do not consider the irrationality of the initial pricing and the length of the trading period, which may result in evaluation results that lack scientific basis and deviate from the real market situation. Therefore, this paper proposes a used car valuation method based on K-means++ clustering and multiple linear regression[14, 15], which takes the transaction period into account in the model. This method can reasonably re-evaluate unsalable vehicles and unsold vehicles, shorten the transaction period, speed up sales, and improve the liquidity of the used car market on data sets with many variables.

2. Based on clustering and multivariate linear regression, a method for valuing used cars

2.1. Data Preparation

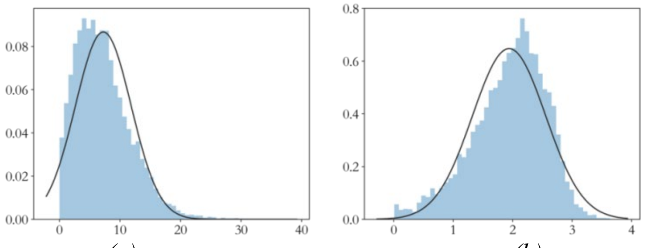
The data samples are obtained from the used car transaction records of 58.com, which includes 24 variables as shown in Table 1. Among these variables, mileage, emissions, exhibition and sales time, registration time, and licensing time require separate analysis and processing, while the other characteristics may only need to be filled with missing values. As there is no trading period information in the data, the last price is used as the starting point to calculate the trading period by subtracting the number of days from the last price. Two additional features, the number of price cuts and the range of price cuts, are constructed for each sample. The trading cycle is classified into three types. This paper will design the valuation strategy based on the first category and analyze and discuss the valuation effect on the second and third categories.

Table 1: Results of multiple regression evaluation indicators

Serial number	Variables	Serial number	Variables	Serial number	Variables
1	Vehicle ID	9	GB code	17	Transmission gear box
2	Exhibition and sales time	10	Carrying passengers Number of people	18	Type of fuel oil
3	Brand ID	11	Date of registration	19	Time on shelves
4	Vehicle model ID	12	Date of license	20	Price on shelves
5	Model ID	13	Country	21	Time of transaction
6	Mileage	14	Manufacturer Type	22	New car price
7	City ID	15	Style	23	Transaction price
8	Number of transfers	16	Volume of displacement	24	Price adjustment time: adjusted price

Through data analysis, visualization in Figure 1 (a) shows that mileage belongs to the right-skewed distribution, and the value distribution of variables is not uniform, which will greatly affect the regression estimation.

For the data with right-skewed distribution, log transformation is generally used. The comparison results are shown in Figure 1 (b), and the distribution tends to be more normal, which is conducive to the fitting of subsequent regression models.



(a) (b)

Figure 1: Mileage Normal Distribution Fitted Plot

Emissions are numerical characteristics, and vehicles can be roughly classified by emissions. Vehicles with an displacement of less than 1L are mini cars, vehicles with an displacement of 1.0-1.6L are ordinary cars, vehicles with an displacement of 1.6-2.5L are intermediate cars, vehicles with an displacement of 2.5-4.0L are medium-high class cars, and vehicles with an displacement of more than 4L are advanced cars. According to this rule, the emissions data are coded in buckets. The corresponding former vehicle category is coded as 1-5 types. For the three features of exhibition time, registration time and license time, year, month and day are proposed as three new separate features. A total of 9 new features will be used in subsequent models, and the original variable data will be discarded. In the data set, the missing fields of national standard code, year type, country, manufacturer and transmission are completed by the method of mode.

2.2. Selection of significant variables for regression analysis

After the above preprocessing, 33 feature variables are obtained. Not all features are beneficial for regression, so it is necessary to screen the features and select the key factors with high correlation with price as the key factors for regression analysis. The adopted method is Spearman's rank correlation coefficient method [16], and 20 features with a threshold greater than 0.05 are selected, and the correlation is shown in Table 2 below.

Talbe 2: Table of correlation between characteristics and prices

Features	Brand ID	Model ID	Mileage	Color	City ID	GB code	Carrying passengers Number of people	Country	Manufacturer Type	Style
Relevance	0.0934	0.0848	0.2506	0.0832	0.1808	0.2128	0.0522	0.3632	0.4030	0.4330
Feature	Number of price cuts	Level of price	Class of displacement	Category of mileage	Year of registration	Time of registration	Month of registration	Volume of displacement	Transmission gear box	New car price
Relevance	0.0647	0.8112	0.5662	0.1711	0.4172	0.4358	0.0672	0.5305	0.2302	0.8137

Secondly, SPSS analysis software was used to analyze the collinearity of all features, and Variance Inflation Factor (VIF) was used to measure the severity of multicollinearity in multiple linear regression models[17]. It represents the ratio of the variance of the regression coefficient estimator compared to the variance when the independent variables are assumed not to be linearly dependent. Usually, 10 is used as the judgment boundary. When$VIF < 10$, there is no multicollinearity; When $10 \leqslant VIF < 100$, there is strong multicollinearity; When $VIF \geqslant 100$, there is severe multicollinearity. The variance inflation coefficient is calculated in Eq. 1.

$$VIF = \frac{1}{1 - R_i^2} \tag{1}$$

The resulting values are shown in Table 3.

Table 3: The calculation results

Independent variable	VIF	Independent variable	VIF	Independent variable	VIF
Brand ID	1.068	Carrying passengers Number of people	1.13	New car price	3.631
Model ID	1.095	Country	1.309	Number of price cuts	1.029
Mileage	2.282	Manufacturer	1.698	Year of registration	70.404
Color	1.075	Style	17.105	Time of registration	72.583
City ID	1.033	Volume of displacement	2.765	Month of registration	1.412
GB code	1.173	Transmission gear box	1.149		

Among them, the VIF of the year, the registration year and the license year are all greater than 10, so there is collinearity. So in the regression analysis, only the registration years with high correlation with price are kept. Therefore, 15 key factors affecting the price of used cars are obtained through screening, including brand, model ID, mileage, color, city ID, national standard code, passenger number, country ID, manufacturer type, displacement, transmission, new car price, reduction times, registration year, and license month.

2.3. Model Building

In order to select a pricing strategy that accelerates sales, the data within the first week of the trading period is selected for regression analysis. Firstly, from the above data preprocessing data set, all samples with trading cycle category coding 1 are selected and K-means ++ clustering is performed. The value of k is first chosen based on the sum of the squared errors (SSE), which is calculated as shown in Eq. 2.

$$SSE = \sum_{i=1}^{k} \sum_{p_i \in C_i} |p - m_i|^2 \tag{2}$$

WhereCi is the ith cluster, p is the sample point in Ci, m_i is the centroid of Ci (the mean of all samples in Ci), and SSE is the clustering error of all samples, which represents the quality of the clustering effect. According to the elbow method, when k is less than the true cluster number, the SSE will decline greatly because the increase of k will greatly increase the aggregation degree of each cluster. When k reaches the true cluster number, the return of aggregation degree obtained by increasing k will quickly become smaller, so the decline of SSE will decrease sharply, and then become flat as k continues to increase. This point is called the elbow point, and it is chosen as the k value for this clustering method.

Because K-means ++ initializes the cluster center randomly, the random seed will also affect the effect of clustering. In this paper, the silhouette coefficient is used as the evaluation index to select the random seed. According to the intra-cluster dissimilarity a_i and inter-cluster dissimilarity b_i of sample i, the silhouette coefficient s_i is defined, and the calculation method is shown in Eq. 3.

$$s_i = \frac{b_i - a_i}{\max\{b_i, a_i\}} \tag{3}$$

The silhouette coefficient ranges between [-1,1]. The larger the value is, the more reasonable it is. When s_i is close to 1, it means that sample i is reasonably clustered. When s_i is close to -1, sample i should be classified into another cluster. If s_i is approximately 0, it means that sample i is on the boundary of two clusters. The value when the silhouette coefficient is the largest is selected as the random seed of the clustering method, denoted as r.

Through the above clustering, m clusters with cluster centers of $c_i = (i = 1, 2, \cdots, m)$ are obtained, denoted as Ci, $i = 1, 2, \cdots, m$. Then multiple linear regression analysis was performed on each cluster to establish m regression equations. Then the price evaluation equation for m clusters is as follows.

$$y_i = B_{i_0} + B_{i_1} \times x_{i_1} + \cdots + B_{i_{14}} \times x_{i_{14}} + B_{i_{15}} \times x_{i_{15}} \quad (i = 1, 2, \cdots, m) \tag{4}$$

Where y represents the selling price, B0 is a constant, and B1-B15 represent the characteristics of 15 used cars, which are brand, model ID, mileage, color, city ID, national standard code, passenger number, country ID, manufacturer type, displacement, transmission, new car price, reduction times, registration year, and license month.

3. Analysis of simulation results

3.1. Clustering and regression

The dataset has a total of 15,000 samples, and there are 3727 sample data of hot selling vehicles in the first week of the trading cycle. In the process of cluster comparison analysis, the range of *k* is selected from 2 to 8, as shown in Figure 2, the SSE decreases greatly between *k*=2 and *k*=3, and after *k* equals 3, the SSE tends to be flat. According to the definition of elbow method, the clustering effect is the best when k=3.

The random seed is taken as an integer between 2000 and 2025, and the silhouette coefficient is used as the evaluation index, as shown in Figure 3, when the random seed is 2016, the silhouette coefficient is the largest and the clustering effect is the best.

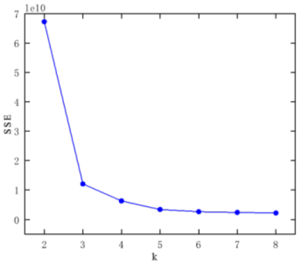


Figure 2: Diagram of elbow method with different K values

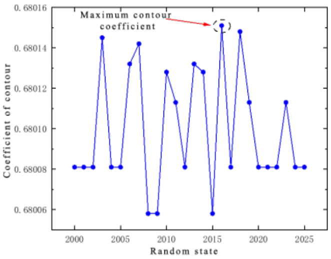


Figure 3: Contour coefficients of different random seeds

The visualization of the clustering effect is shown in Figure 4, from which it can be seen that the three categories are distinguished significantly. The first cluster has 2114 samples, the second cluster has 1603 samples, and the third cluster has 10 samples. The cluster centers of the three clusters, keeping four decimal places, are shown in Table 4 below.

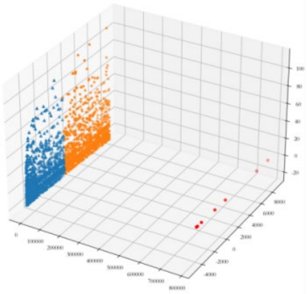


Figure 4: Clustering effect of hot vehicle transaction data

Table 4: The clustering center

Features	Brand ID	Model ID	Mileage	Color	City ID	GB code	Carrying passengers Number of people	Country	Manufacturer	Number of price cuts
Class I	17.4652	1878.5485	1.9849	2.7407	7.37916	1.7074	5.1451	779413.7098	1.9976	0.6165
Class II	20.5517	9324.2798	1.9543	2.7486	8.3969	1.7733	5.1603	779413.587	2.0068	0.6229
Class III	68	3557	1.6702	3.3	6.4	1.1	4.8	-1.16E-10	1	1
Features	Level of price	Class of displacement	Category of mileage	Year of registration	Time of registration	Price	Volume of displacement	Transmission gear box	New car price	
Class I	1.3030	2.6541	1.2878	2014.2787	6.6641	2.2531	1.9088	11.4410	3.0902	
Class II	1.3399	2.7250	1.2941	2014.1789	6.5585	2.3084	1.9768	12.2929	3.1721	
Class III	1.1	3	1	2016.5	6.7	2.3768	2.32	3.6	2.9535	

Using the clustering results for hot-selling vehicles, multiple regression equations were fitted for each cluster. SPSS software was used for multiple regression analysis, with R2 indicating the model fit (with higher values indicating better fit). The Durbin-Watson coefficient (D-W) was used to check for autocorrelation, with values close to 2 indicating less obvious autocorrelation and greater independence between the independent variables.

For the sample data of the first cluster, price was taken as the dependent variable and 15 variables were taken as independent variables. The regression equation has an R2 of 0.928 and a D-W coefficient of 1.474. Table 5 shows the standardized coefficients and significance of each independent variable, indicating good independence between variables.

Independent variables with significance less than 0.05 were selected. It includes brand ID, model ID, mileage, GB code, number of passengers, manufacturer type, displacement, transmission, new car price, reduction times, and registration year. The first multiple linear regression equation is obtained as shown in Eq. 5.

$$\begin{aligned} y = & -0.003 \times x_1 - 0.000005902 \times x_2 - 0.087 \times x_3 - 0.027 \times x_6 + 0.02 \times x_7 \\ & + 0.058 \times x_9 + 0.039 \times x_{10} + 0.001 \times x_{11} + 0.957 \times x_{12} + 0.009 \times x_{13} + 0.111 \times x_{14} \end{aligned} \quad (5)$$

Table 5: Standardized coefficients and significance of variables in multiple regression

Independent variable	Brand ID	Model ID	Mileage	Color	City ID	GB code	Carrying passengers Number of people	Country
Unstandardized coefficient B	-0.003	-5.902e-6	-0.087	- 0.004	0.000	-0.027	-.020	-0.002
Significance of significance	0.000	0.031	0.000	0.113	0.488	0.000	0.007	0.378
Independent variable	Manufacturer	Volume of displacement	Transmission gear box	New car price	Number of price cuts	Year of registration	Month of registration	Constant
Unstandardized coefficient B	0.058	0.039	0.001	0.957	0.009	0.111	-0.002	1451.322
Significance of significance	0.000	0.006	0.015	0.000	0.037	0.000	0.129	0.445

For the second cluster, the regression results in an R-square of 0.891 and a Durbin Watson value of 1.645, indicating good independence between the variables. The coefficients are shown in Table 6.

Table 6: Standardized coefficients and significance of variables in multiple regression

Independent variable	Brand ID	Model ID	Mileage	Color	City ID	GB code	Carrying passengers Number of people	Country
Unstandardized coefficient B	-0.003	1.505e-6	-0.076	0.002	-0.004	-0.005	0.050	-0.005
Significance of significance	0.000	0.622	0.000	0.521	0.000	0.481	0.000	0.167
Independent variable	Manufacturer	Volume of displacement	Transmission gear box	New car price	Number of price cuts	Year of registration	Month of registration	Constant
Unstandardized coefficient B	0.123	0.047	0.001	0.882	0.012	0.113	-0.006	3604.807
Significance of significance	0.000	0.013	0.357	0.000	0.048	0.000	0.001	0.194

Ten variables with significance less than 0.05 were selected, including brand ID, mileage, number of passengers carried, manufacturer type, displacement, new car price, number of price reductions, registration year, and license month. The second regression equation is obtained as shown in Eq. 6.

$$\begin{aligned} y = &-0.003 \times x_1 - 0.076 \times x_3 - 0.004 \times x_5 \\ &+ 0.05 \times x_7 + 0.123 \times x_9 + 0.047 \times x_{10} + 0.882 \times x_{12} \\ &+ 0.012 \times x_{13} + 0.113 \times x_{14} - 0.006 \times x_{15} \end{aligned} \tag{6}$$

Since there are only 10 samples in the third category, the third regression equation is obtained by directly solving, and the variables are model ID, color, city ID, GB code, new car price 12, number of price reduction 13, registration year 14, and month of license plate 15. As shown in Eq. 7.

$$\begin{aligned} y = &0.000003207 \times x_2 - 0.026 \times x_4 - 0.01 \times x_5 \\ &+ 0.269 \times x_6 + 0.935 \times x_{12} - 0.035 \times x_{13} \\ &+ 0.134 \times x_{14} - 0.004 \times x_{15} \end{aligned} \tag{7}$$

3.2. Revaluation analysis

Firstly, the samples of unsalable vehicles and vehicles that have not sold are classified by cluster center and divided into three clusters. Equations (5) - (7) are used to re-evaluate each cluster. Since the third cluster has fewer samples, no visualization is done. Regressing the price of the sample of unsalable vehicles sold with equations (5) and (6), partial visualizations are shown in Figures 5.

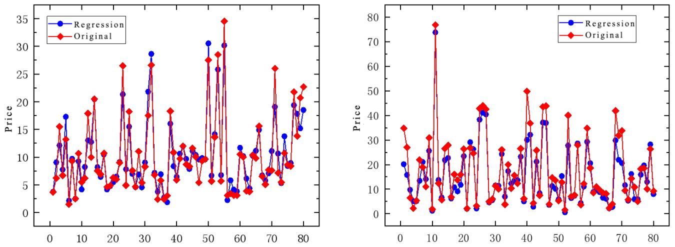


Figure 5: Prediction effect of regression equation on unsalable vehicles

Similarly, equation (5) and (6) are used to evaluate the sample set of unsold vehicles by regression, and partial visualization results are shown in Figure 6.

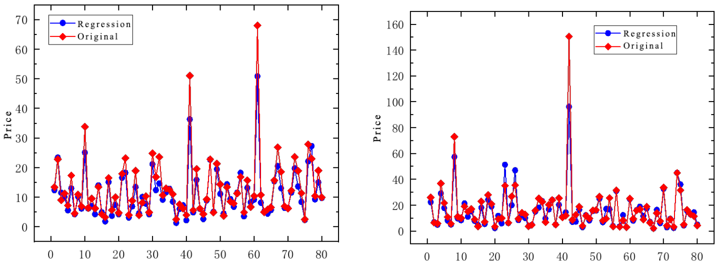


Figure 6: The predictive effect of the regression equation on unsold vehicles

Table 7: MSE of original price and return price

Regression equation	Equation I	Equation II	Equation III	Sum
Hot selling vehicle	0.0417	0.0467	0.0170	0.1050
No vehicles sold	0.0431	0.0557	0.1376	0.2364

As can be seen from the above figure, most of the original pricing is higher than the regression pricing, and the original pricing is on the high side. Table 7 shows that the mean absolute error between the original price and the regression price of unsold vehicles is 0.1050, and the mean absolute error between the original price and the regression price of vehicles that have not been sold is 0.2364. The latter is higher than the former, indicating that unreasonable pricing is the key factor affecting the transaction cycle, and the model of revaluation is statistically significant.

3.3. Comparison of multiple valuation schemes

For the prediction model, the evaluation metrics used in this paper are Mean Absolute error (MAE), Mean Absolute Percentage error (MAPE) and root mean square error (RMSE). They are given by Eq. 8-10.

$$MAE=\frac{1}{n}\sum_{k=1}^{n}\left|t_k-y_k\right| \tag{8}$$

$$MAPE=\frac{1}{n}\sum_{k=1}^{n}\left|\frac{t_k-y_k}{y_k}\right| * 100\% \tag{9}$$

$$RMAE=\sqrt{\frac{1}{n}\left(\sum_{k=1}^{n}(t_k-y_k)^2\right)} \tag{10}$$

Compared with other two used car price evaluation and prediction methods, AdaBoost[12] and XGboost[13]. The corresponding error analysis results are shown in Table 8.

Table 8: Evaluation index Results

Model of valuation	MAE	MAPE	RMSE
Our Approach	0.1004	5.5426	0.1441
XGBoost	0.1041	6.4634	0.1528
AdaBoost	0.2662	15.4742	0.3319

Table 8 shows that the total MAE of clustering-based multiple linear regression is 0.1004, which is

62.2% lower than AdaBoost and 3.5% lower than XGBoost. It is 5.5426, 27.1% lower than AdaBoost and 14.2% lower than XGBoost. It is 0.3319, which is 56.5% lower than AdaBoost and 5.6% lower than XGBoost. From the three evaluation indexes, the multiple regression model based on clustering has the best fitting effect, which shows the rationality of the multiple regression evaluation method based on clustering.

4. Conclusions

This study proposes a new pricing system for used cars using multiple linear regression and clustering, which considers the trading duration. The data set is divided into three categories based on the trading cycle: popular cars traded within the first week, unsold cars traded after the first week, and unsold cars that remain. During model training, the samples of popular cars are clustered using K-means ++, and a regression equation is developed for each cluster. When the model is tested, new samples are classified into the appropriate cluster, and the corresponding regression equation is used for prediction and evaluation. The model's re-rationality is verified by comparing it with unsold and remaining inventory of vehicles. Two commonly used car price assessment techniques, XGBoost and AdaBoost, are also compared using different evaluation indicators, and the experimental results show that the multiple linear regression approach based on clustering has a smaller error, indicating the effectiveness of the revaluation model.

References

[1] Mao pan, Cai yun, Wan xiong. Study on Influencing Factors of Second-hand Car Price Evaluation Based on BP Neural Network [J]. Automotive Practical Technology, 2020(04): 59-63+67.

[2] Zhang Yuansen. Second-hand Car Price Evaluation Model Based on Neural Network [D]. Tianjin: Tianjin University, 2018:11-15.

[3] Yang Sirui. Research on Second-hand Car Evaluation Model Based on GA-MIV-BP Algorithm [D]. Chongqing: Chongqing University of Technology, 2020:15-22.

[4] Lv Jin. Study on Used Car Price Prediction Based on Feature Optimization Combination SVM [D]. Wuhan: Zhongnan University of Economics and Law, 2019: 9-18.

[5] Hu Yu. Construction and Application of Second-hand Car Evaluation Model Based on Characteristic Price Theory [D]. Changsha: Hunan University, 2017:11-15.

[6] Xie Yang, Wen Hua, Zhang Jie. A second-hand car price evaluation method based on Machine learning [J]. Enterprise Technology Development: Mid-Day Issue, 2015, 34(4): 116-8.

[7] Cao Jie. Study on the Value Evaluation of Second-hand Cars Based on Random Forest Model [D]; Hebei University of Economics and Business, 2020.

[8] Wang Jingna. Research on Second-hand Car Evaluation Model Based on Random Forest Algorithm [D]; Beijing Jiaotong University, 2019.

[9] FREUND Y, SCHAPIRE R, ABE N. A short introduction to boosting[J]. Journal-Japanese Society For Artificial Intelligence, 1999, 14(771-780): 1612.

[10] Cui Sishuai. Analysis of Domestic Used Car Price Forecast Based on Integrated Learning [D]. Dalian: Dalian University of Technology, 202:15-21.

[11] Jia Pengxiang. Forecast of Used Car Price based on LightGBM [D]. Jinan: Shandong Normal University, 2021:14-22.

[12] Liu Cong, Cheng Ximing. A second-hand car valuation method based on AdaBoost. Journal of Beijing University of Information Science and Technology (Natural Science), 2017, 32(03): 49-53.

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

[14] ARTHUR D, VASSILVITSKII S. k-means++: The advantages of careful seeding[R]: Stanford, 2006.

[15] Wang Huiwen, Meng Jie. Multiple linear regression prediction modeling method [J]. Journal of Beijing University of Aeronautics and Astronsutics,2007(04): 500-504.

[16] KUMAR A, ABIRAMI S. Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method [J]. Information Sciences, 2018, 460: 23-41.

[17] SALMERoN R, GARCiA C, GARCiA J. Variance inflation factor and condition number in multiple linear regression [J]. Journal of Statistical Computation and Simulation, 2018, 88(12): 2365-84.