Second-hand Car Price Prediction Based on a Mixed-Weighted Regression Model

Shengqiang Han
Department of Information Management, Shandong Normal University, Changqing, Jinan,
Shandong, China
e-mail: 2590329954@qq.com

Jianhua Qu
Department of Information Management, Shandong Normal University, Changqing, Jinan,
Shandong, China
Corresponding author: e-mail: qjh@sdnu.edu.cn

Jinyi Song
Department of Information Management, Shandong Normal University, Changqing, Jinan,
Shandong, China
e-mail: 2980559861@qq.com

Zijing Liu
Department of Information Management, Shandong Normal University, Changqing, Jinan,
Shandong, China
e-mail: 2500784423@qq.com

2022 7th International Conference on Big Data Analytics (ICBDA) | 978-1-6654-7938-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICBDA55095.2022.9760371

***Abstract*—With the development of motor vehicles, the circulation demand of motor vehicles in the form of "second-hand cars" in circulation links is increasing. As a special "e-commerce commodity", second-hand cars are more complicated than ordinary e-commerce commodities. As a result, it is difficult to estimate the price of second-hand cars, which is not only influenced by the basic configuration of the car, but also by the car conditions. At present, the state has not issued a standard to judge the value of second-hand car. To solve this problem, in this paper, first making feature engineering, which includes data preprocessing and feature screening. Data preprocessing includes data cleaning and data transformation, data cleaning includes removing outliers and filling missing values, and data transformation is used to unify data format to improve data quality. The feature screening includes correlation analysis and feature extraction based on LightMBG, and the screened features provide the basis for model building, training and prediction. Then, five regression models are constructed by using the feature attributes obtained by the feature engineering for training, and evaluated. Then, Random Forest and XGBoost are weighted and mixed to got a novel regression model, and the effect of the model is better than that of the five regression models. Finally, the novel regression model is used to predict the price of second-hand cars.**

Keywords—second-hand car price prediction, weighted regression model, LightGBM, XGBoost, random forest

I. INTRODUCTION

With the continuous growth of the number of motor vehicles in China, the per capita ownership also increases. The circulation demand of motor vehicles in the form of "second-hand cars" in circulation links, including second-hand car collection, second-hand car auction, second-hand car retail and second-hand car replacement, is increasing. As a special "e-commerce commodity", second-hand cars are much more complicated than ordinary e-commerce commodities because of their "one car, one condition" characteristics. The reason is that it is difficult to accurately estimate and set the price of second-hand cars, which is not only influenced by the basic configuration of the car itself, such as brand, car system, power, etc., but also by the car conditions such as mileage, car body damage and maintenance, etc. Even the change of new car prices will have an effect on the price of second-hand cars. At present, the state has not issued a standard to judge the value of second-hand car. Some second-hand car trading platforms and second-hand car third-party valuation platforms have established a series of valuation methods from their own perspectives to evaluate the value of second-hand car.

In a typical second-hand car retail scene, second-hand cars generally get user information through online channels such as the Internet, then offline physical stores could display and sell, commonly known as O2O store model. Stores buy second-hand cars from individuals or other channels through "buyers", and then they are priced and sold by store pricing experts. Like other commodities, second-hand cars will be discounted and promoted if the pricing is too high and unsalable, or even packaged and wholesale directly at a lower price until the commodities are finally sold.

II. RESEARCH BACKGROUND

With the rapid development of artificial intelligence, algorithms based on machine learning and deep learning have been used in the research of second-hand cars.Yang Bo[1] built a BP neural network for example analysis based on the evaluation and pricing problems in used car transactions. The results show that the built prediction model is more accurate and stable than the existing prediction models.Mao Pan etc.[2] based on the basis of the BP neural network, the correlation coefficient of model predicted price and actual price reached 0.96. Based on the ARIMA model, CHEN Daoping[3] established the Chinese automobile demand prediction model and evaluated the prediction performance of the model. The results show that the model has a very good prediction effect. Xie Yang etc. [4] use clustering, multiple regression and other algorithms to take factors such as registration time, characterization mileage, region and other factors as independent variables, and the new formation rate as the dependent variable to establish the model. Through actual evaluation, the model has good evaluation effect.

Li Shuangnan [5] used Bootstrap method to establish the second-hand car price regression model, but its sample is relatively small, and only one brand of data is selected, which is highly local. He Huanli [6] developed a commercial second-hand car appraisal and evaluation platform based on replacement cost method with the help of labview virtual instrument platform. Lin Tengfei [7] evaluates second-hand cars based on Taguchi function and analytic hierarchy process .

In this paper, the weighted and mixed regression method is proposed(WMR). Random Forest and XGBoost regression

models are weighted and mixed, and two regressions models with different properties are mixed, so the robustness is improved. At the same time, the evaluation index also proves that it has a good effect in second-hand car price prediction, so our weighted and mixed model is very suitable for second-hand car valuation.

III. Feature Engineering

A. Data Sources

The original datasets are all from Mathorcup Big Data Challenge (https://www.saikr.com/c/nd/8418), including Annex 1 and Annex 2. Annex 1 is the data set for training model. The main data includes vehicle basic information, transaction time information, price information, etc., including 36 columns of variable information, in which 15 columns are anonymous variables. See Table I for the field description of Annex 1.

According to Table I, we can find that there are a large number of features in the data set, and the final price is numerical data, which is suitable for predicting by regression. Therefore, we build a regression model for the second-hand car price prediction model.

Annex 2 is the test data set, which is the application of the built model. Compared with Annex 1, it lacks price characteristics.

TABLE I. Appendix 1 Data and Field Names and Meaning

serial number	Features	Description
1	carid	Vehicle id
2	trade time	Exhibition time
3	brand	Brand id
4	serial	Car id
5	model	Model id
6	mileage	mileage
7	color	Vehicle color
8	city id	City id of vehicle
9	car code	GB code [national standard code for Chinese characters]
10	transfer count	Number of transfers
11	seatings	Number of passengers
12	register date	date of registration
13	license date	Date of licensing
14	country	country
15	make type	Manufacturer type
16	model year	Year's money
17	displacement	displacement
18	gearbox	gearbox
19	oil type	Type of fuel
20	new price	New car price
21	anonymous feature	15 anonymous features
22	price	second-hand car transaction price (forecast target)

B. Data Preprocess

1) Data cleaning

Step 1: Remove outliers.

In order to ensure the accuracy of the subsequent prediction, the training data must conform to the actual situation. This paper assumes that there is no sudden change of features caused by external accidental factors in the prediction. Therefore, for each data, if the value of a feature obviously deviates from the normal level of this feature value, the data value can be considered as outlier data. This paper first removes the outlier data. Statistics show that all data are basically within the normal range, and only a few features have outlier values. See Fig. 1 for details.

From Fig. 1, we can see that most values of feature of V6 are distributed in an correct interval, but a few of them are not in this interval, so we can think the values that are not in this interval are outlier. At the same time, we find that all data such as V14 and V21 have the same feature value, which have little significance to the regression model, so similar attributes can be removed when feature screening.

Step 2: Fill in the missing values

According to the statistics of the original data, the data are discontinuous data sets, and there are missing data in many features, as shown in Table II for details.

TABLE II. Statistical Information of Each Feature in Annex 1

Characteristic	quantity	type
V1	30000	non-null int64
V2	30000	non-null object
V3	30000	non-null int64
V4	30000	non-null int64
V5	30000	non-null int64
V6	30000	non-null float64
V7	30000	non-null int64
V8	30000	non-null int64
V9	29991	non-null float64
V10	30000	non-null int64
V11	30000	non-null int64
V12	30000	non-null object
V13	30000	non-null object
V14	26243	non-null float64
V15	26359	non-null float64
V16	29688	non-null float64
V17	30000	non-null float64
V18	29999	non-null float64
V19	30000	non-null int64
V20	30000	non-null float64
V21	28418	non-null float64
V22	30000	non-null int64
V23	30000	non-null int64
V24	17892	non-null float64
V25	30000	non-null int64
V26	30000	non-null int64

V27	11956 non-null object
V28	26225 non-null float64
V29	26256 non-null float64
V30	23759 non-null float64
V31	29539 non-null object
V32	30000 non-null object
V33	28381 non-null float64
V34	30000 non-null int64
V35	2420 non-null object
V36	30000 non-null float64

From Table Ⅱ, we can see that many features have missing values, even some features have a huge number of missing values. Therefore, we use -1 to fill in the features with fewer missing values. For features with a large number of missing values, such as V35, we will remove them during feature screening, and we will not participate in model construction and training, so as not to affect the effect of the model.

2) Data transformation

Because there are non-numerical data in the original data set, such as date type and classification type, and the given dates are all xx/x/xx, which is not conducive to processing, it is necessary to integrate and correct errors, so the extracted time information is changed into numerical type (such as V33), which is convenient for subsequent algorithm identification and correlation analysis.

For category data, our processing is as follows: reconstruct the data matrix, and the car Id corresponding to the data index ,the value of each feature is the columns, and the content is filled with 0 and 1, which means whether the feature value is in this vehicle,include{0,1},and finally merge it with the original data to get new data.

C. Feature Screening

In the data transformation, we added many variables, but not all of them are useful for the final construction of the regression model. If all variables are putted into the model, on the one hand, it will take more time, on the other hand, it will not achieve good results. All the features will be screened here. The specific steps are: firstly, through correlation analysis, the variables with weak correlation with the price are deleted (because many of the variables constructed are very weak correlation with the transaction price, so a large number of variables can be excluded by this step 1, which can reduce the time for feature selection according to the model later), and then the variables with low importance are removed according to the importance of variables in lightGBM model.

1) correlation test

We use Pearson correlation coefficient to judge the correlation between each feature and price. Pearson correlation coefficient is calculated as follows:

$$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-EX)(Y-EY)]}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sqrt{E(X^2)-E^2(X)}\sqrt{E(Y^2)-E^2(Y)}}$$

The calculated correlation coefficient between features and prices is shown in Table III (because there are many variables introduced, only part of them are shown in this paper).

TABLE III. CORRELATION COEFFICIENT BETWEEN FEATURES AND PRICE (PART)

characteristic	correlation coefficient
V3_1	-0.001661674
V3_2	-0.000890231
V3_5	-0.002698169
V3_6	1
V3_7	-0.002859633
V3_10	-0.001719157
V3_8	-0.004752317
V3_13	-0.003005901
V3_12	-0.002656005
V3_15	-0.001209971
V3_16	0.036174757
V3_21	-0.000164762
V3_22	-0.001535389
V3_23	-0.002957482
...	...
V_325	-0.002296238

2) Feature extraction based on LightGBM

Under the framework of GBDT, LightGBM optimizes the segmentation point searching process and tree growth mode of the base learner, that is, using Histogram algorithm to speed up the segmentation point searching process and reduce memory consumption, and using Leaf-wise/Best-first leaf growth strategy with depth limitation to improve the accuracy of the base learner and generate the decision tree more efficiently. On this basis, the foundation of LightGBM-LGB_baseline is established. Then two algorithms are proposed from the angles of reducing samples and features respectively: Gradient-based One-Side Sampling (GOSS) to focus on samples with larger gradient; And Exclusive Feature Bundling (EFB) to reduce the number of mutually exclusive features in sparse data. These two algorithms will greatly speed up the training process and save memory consumption without losing accuracy.

The importance of features based on LightGBM is shown in Table 4(due to the large number of features, only part of them are shown in this article).

According to Table IV, we select the features whose importance accumulates to 0.99 as the features of model construction and training.

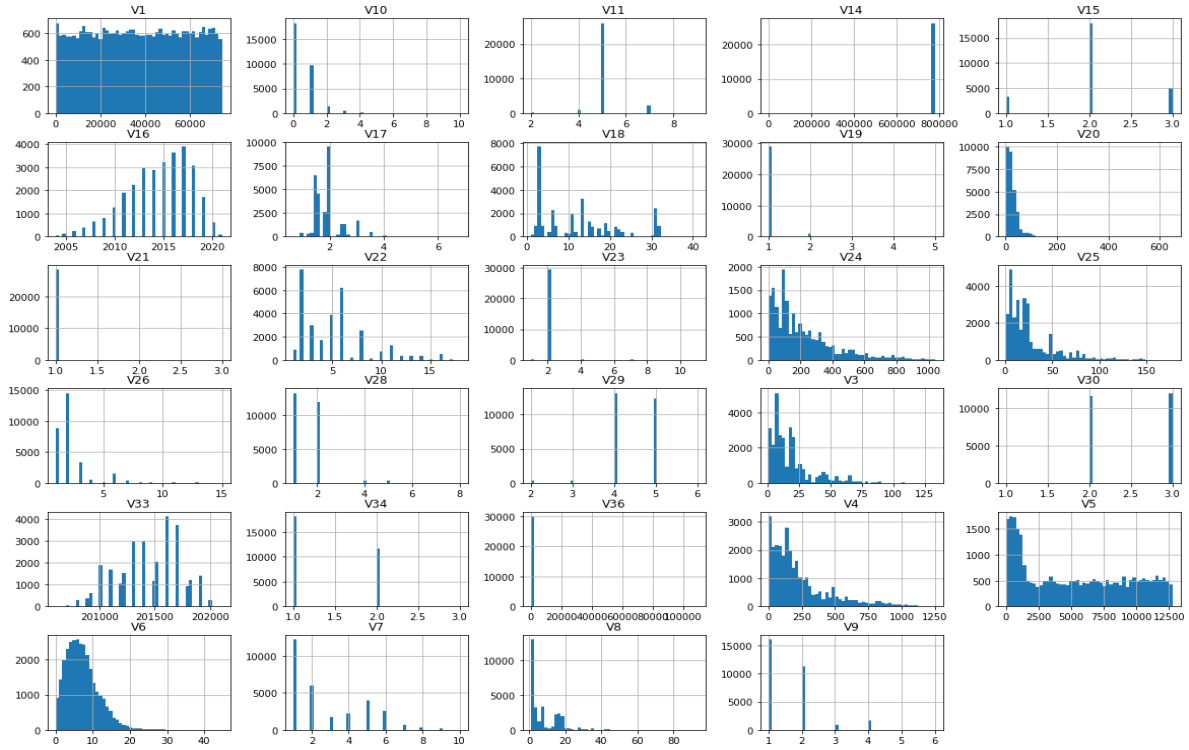

Fig. 1. Appendix 1 distribution of numerical data

TABLE IV. IMPORTANCE OF EACH FEATURE (PART)

feature	importance	normalized_importance	cumulative_importance
V36_x	311.7	0.567759563	0.567759563
V20	37.5	0.068306011	0.636065574
V6	28.9	0.052641166	0.68870674
V17	24.7	0.044990893	0.733697632
V33	23.4	0.042622951	0.776320583
V1	21.6	0.039344262	0.815664845
V24	18.5	0.033697632	0.849362477
V25	17	0.030965392	0.880327869
V27_2020/9/14	10.3	0.018761384	0.899089253
V18	9.1	0.016575592	0.915664845
V28	8.3	0.015118397	0.930783242
V22	6.8	0.012386157	0.943169399
V27_2021/1/25	3.9	0.007103825	0.950273224
V10	3.5	0.006375228	0.956648452
V27_2020/12/14	3.4	0.006193078	0.96284153
V26	3.3	0.006010929	0.968852459
V16	2.8	0.005100182	0.973952641
V34	1.9	0.003460838	0.977413479
V27_2020/9/24	1.8	0.003278689	0.980692168
V11	1.8	0.003278689	0.983970856
V30	1.7	0.003096539	0.987067395
V14779415.0	1.3	0.002367942	0.989435337
V27_2020/9/10	1.3	0.002367942	0.991803279
V27_-1	1	0.001821494	0.993624772
V19	0.9	0.001639344	0.995264117

IV. Model and Result

A. Evaluating Criterion

This paper adopts a more comprehensive evaluation standard (from mathorcup Big Data Challenge Model Evaluation Standard), and the model evaluation standard is:

$$\text{Evaluation} = 0.2 \times (1 - \text{Mape} + 0.8 \times \text{Accuracy}_5)$$

Ape (relative error):

$$\text{Ape} = \frac{|y - \hat{y}|}{y}$$

Mape (average relative error):

$$\text{Mape} = \frac{1}{m}\sum_{i=1}^{m} \text{Ape}_i$$

note: the true value y = (y1, y2, ..., ym), and the predicted value of the model is $\hat{y}$= ($\hat{y}$1,$\hat{y}$2, ..., $\hat{y}$m);

Accuracy5(5% error accuracy):

$$\text{Accuracy}_5 = \text{count}(\text{Ape} \le 0.05)/\text{count}(\text{total})$$

note: the number of samples with relative error Ape within 5% is $\text{count}(\text{Ape} \le 0.05)$, the total number of samples is $\text{count}(\text{total})$.

It is not difficult to see from the evaluation standard of the model that the closer the evaluation value is to 1, the better the model is; otherwise, the farther away from 1, the worse the model is. From the expression of the model evaluation formula, we can know that this standard pays more attention to the proportion of points with relative errors within 0.05(the weight is 0.8).

B. Weighted and Mixed Regression Model

Random Forest and XGBoost are classic models, we add them by linear weighting. The weight of random forest is α, and that of XGBoost is β:

$$\text{RP} = \alpha \times \text{RandomForest\_Predict}$$

$$\text{XP} = \beta \times \text{XGBoost\_Predict}$$

$$\text{Final\_Predict} = \text{RP} + \text{XP}$$

note: normally set α =1.09, β = -0.09, and $\alpha + \beta = 1$.

Result: 0.7901800575909526

Compared with the classical model, the effect of the WMR is better, and the two kinds of regressions are mixed to enhance the robustness.

C. Results

In this paper, Random Forest [8], XGboost [9], multiple linear regression [10], GBDT[11] and LightGBM [9] are used to compare with the regression model we proposed(WMR). From Table V, we find that WMR is the best, so we use this model to predict the second-hand car price based on the second-hand car data in Annex 2, and the results are shown in Table VI (due to the large number of samples, in this paper,a part predicted prices are shown)

TABLE V. Evaluation Results of Regression Model

regression model	evaluation
WMR	**0.79**
Random forest	0.73
XGBoost	0.48
multiple linear regression	-0.01
GBDT	0.10
LightGBM	-0.11

TABLE VI. Second-hand Car Price Prediction (Part)

Vehicle Id	forecast price
3	15.62199159
4	9.432364501
8	2.87983162
9	6.256158113
11	14.80823458
15	34.00462447
17	13.09818465
...	...
21	6.068566689
27	3.341252013
28	26.20324481

V. Conclusions

In this paper, first make feature engineering, obtain the datasets which could be used to train regression model. Then, through weighting and mixing Random Forest and XGBoost regression models to obtain WMR, WMR is used in training datasets. Compared with classical regression models such as Random Forest, XGboost, multiple linear regression, GBDT, LightGBM, it is found that WMR model is superior to other five regression models in second-hand car price prediction, so our mixed weighted regression model has remarkable effect in second-hand car prediction. Finally, it is applied to the second-hand car price prediction of the test datasets.

Funding Support

This study was funded by National Natural Science Foundation of China Nos. 61876101, 61802234 and the Meigu College of Shandong Normal University.

References

[1] Yang Bo. Based on the Intelligent Algorithm, the "Internet Plus"Era Used Vehicle Valuation Pricing Model [J]. Jiangsu Commercial Forum, 2017(5): 21-24.(in Chinese)

[2] Mao Pan, Cai Yun, Wan Xiong, Wang Wendi. Research on Influencing Factors of Used Car Price Evaluation Based on BP Neural Network [J]. Automobile Technology, 2020 (4), 59-63, 67.(in Chinese)

[3] Chen Daoping. Chinese Automobile Demand Prediction Based on ARIMA Model[C]// 2011 4th International Conference on Biomedical Engineering and Informatics (BMEI), Oct. 15-17, 2011, Shanghai, China. Piscataway NJ: IEEE, c2011: 2197-2201.

[4] Xie Yang, Wen Hua, Zhang Jie. Second Hand Car Price Evaluation Method Based on Machine Learning [J]. Technological Development of Enterprise, 2015, 34 (11), 116-118.(in Chinese)

[5] Li Shuangnan. Ucar price regression model based on the Bootstrap method [J]. Time Finance, 2012 (15): 202-203.(in Chinese)

[6] He Huanli, Guo Xiangtai. Development of a used car appraisal and evaluation system based on Labview [J]. Automotive Utilities Technology, 2017(19):199-201. DOI:10.16638/j.cnki.1671-7988.2017.19.069.(in Chinese)

[7] Lin Tengfei. Study on Used Car Evaluation Method Based on Tiankou Function and Hierarchical Analysis Method [J]. Automotive Utilities Technology, 2017(10):176-178.DOI:10.16638/j.cnki.1671-7988.2017.10.061.(in Chinese)

[8] Lv Hongyan, Feng Qian. Summary of Random Forest algorithm Research [J]. Journal of Hebei Academy of Sciences,2019,36(03):37-41.DOI:10.16191/j.cnki.hbkx.2019.03.005.(in Chinese)

[9] Xie Yong, Xiang Wei, Ji Mengzhong, Peng Jun, Huang Yihuai. Application analysis of predicting monthly housing rent based on Xgboost and LightGBM algorithms [J]. Computer applications and software, 2019,36 (09): 151-155 + 191.(in Chinese)

[10] Ye Feng. Application of multiple linear regression in economic and technical yield prediction [J]. Chinese and foreign energy sources, 2015,20 (02): 45-48.(in Chinese)

[11] Friedman J H . Greedy Function Approximation: A Gradient Boosting Machine[J]. The Annals of Statistics, 2001, 29(5):1189-123