Used Car Price Prediction Analysis Based on Machine Learning

Jingwen Huang[1](✉), Zhiwen Yu[2], Zhaopeng Ning[3], and Dinghuo Hu[4]

[1] College of Transportation Engineering, Dalian Maritime University, Dalian, Liaoning, China
2013448479@qq.com

[2] College of Electronic Information and Electrical Engineering, Tianshui Normal University, Tianshui, Gansu, China

[3] College of Information Science and Technology, Dalian Maritime University, Dalian, Liaoning, China

[4] College of Naval Architecture and Ocean Engineering, Dalian Maritime University, Dalian, Liaoning, China

Abstract. In order to solve the problem of price evaluation in the Chinese used car retail scene, this paper constructs the feature engineering of machine learning (XGBoost, CatBoost, LightGBM) and Artificial Neural Network Models based on the collected data of 30,000 Chinese used car transactions. The forward feature selection algorithm is used to solve the optimal feature combination; the grid search algorithm is used to optimize the hyperparameters of each model; the evaluation indicators of the four single models are compared and analyzed by leave-one-out cross-validation; analyze and compare the fusion results of the three models and the four models; finally, the model with the highest goodness of fit, that is, the fusion model of XGBoost, CatBoost, LightGBM, and ANN with $R^2 = 0.9845$, is selected as the best model for used car price prediction. Through the prediction model established in this paper, it can guide used car dealers and the financial and insurance industry to establish a used car price evaluation system, and promote the improvement of a reasonable and standardized used car trading market system.

Keywords: Valuation Models for Used Cars · Xgboost · Catboost · Lightgbm · Model Fusion

1 Introduction

In recent years, China's auto market has developed vigorously, and the per capita ownership of motor vehicles has continued to rise. With the change of individual consumption concept, the demand for the circulation of motor vehicles in the form of "used cars" is increasing [4]. In the used car retail scene, if the price is too high, the product will be unsalable and sold at a discount. If the price is too low, the profit cannot be guaranteed. Therefore, it is particularly important to be able to accurately and reasonably evaluate the price of used cars. Previously, some scholars have explored, Yang Sirui used the characteristics of Artificial Neural Network self-learning ability and strong nonlinear

N. Radojević et al. (Eds.): ICAID 2022, AHIS 7, pp. 356–364, 2023.
https://doi.org/10.2991/978-94-6463-010-7_37

mapping ability to establish a used car valuation model, and the effect was good [3]. Jia Pengxiang proposed the LightGBM used car valuation model, and proved its strength, the model effect is better [1]. Wang Jingna used the random forest algorithm to build a used car valuation model, and gave the importance measure of the variables in the model, but it is inefficient when dealing with large-scale data and needs to be improved [2]. Therefore, this paper proposes a valuation model for used cars based on machine learning, and validates it with examples.

2 Explanation to Model Methods

2.1 Artificial Neural Network Model

Multilayer Perceptron, that is, Artificial Neural Network, in addition to the input layer and output layer, there can be multiple hidden layers in the middle. The simplest MLP only contains one hidden layer, that is, a three-layer structure, as shown in Fig. 1.

From Fig. 1, the layers of the multilayer perceptron are fully connected. L1 is the input layer, L2 is the hidden layer, and L3 is the output layer. The neurons of the hidden layer are fully connected to the input layer. Assuming that the input layer is represented by a vector x, the output of the hidden layer is $f(w_1x + b_1)$, w_1 is the weight, and b_1 is the bias. This paper selects the Tanh function as the function f for processing.

$$Tanx = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1}$$

$$(Tanhx)' = \mathrm{sech}^2 x = 1 - \tanh^2 x \tag{2}$$

2.2 Integrated Learning

2.2.1 XGBoost

XGBoost efficiently implements the GBDT algorithm and improves it, so it is widely used in many machine learning competitions. It is one of the fastest and best open source

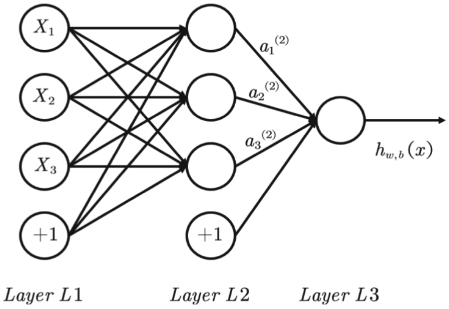

Fig. 1. Multilayer Perceptron Schematic

Boosted Tree toolkits out there. It can be applied to both classification problems and regression problems.

Before building the XGBoost prediction model, three types of parameters need to be set for the XGBoost model, namely regular parameters, booster parameters and task parameters. Among them, the regular parameters are related to the booster used for boosting, which is usually a tree model or a linear model, the booster parameters depend on the selected booster, and the parameters of the learning task determine the learning scenario.

2.2.2 CatBoost

CatBoost is composed of Categorical and Boosting. It is a GBDT framework with less parameters, support for categorical variables and high accuracy based on symmetric decision tree as the base learner. The main difficulty is to efficiently and reasonably process categorical features. In addition, CatBoost also solves the problem of gradient bias and prediction bias, thereby reducing the occurrence of overfitting and improving the accuracy and generalization ability of the algorithm.

2.2.3 LightGBM

The optimization part of LightGBM includes: Histogram-based decision tree algorithm, Leaf-wise leaf growth strategy with depth limit, histogram differential acceleration, direct support for category features, Cache hit rate optimization, histogram-based sparse feature optimization and multi-threading optimization.

LightGBM uses a histogram-based algorithm, for example, it buckets continuous feature values into discrete bins, making the training process faster. At the same time, it avoids the splitting method of the entire layer of nodes, but adopts the method of in-depth decomposition of the node with the largest gain, as shown in Fig. 2.

2.3 Hyperparameter Optimization Based on Grid Search and Cross-Validation

Grid search and cross-validation are a method of parameter adjustment. Within the range of all candidate parameters, the parameters are adjusted by step size in turn, and the adjusted parameters are used to train the learner. Through loop traversal, find the one with the highest accuracy from all the parameters.

The optimization results of the parameters of the Artificial Neural Network model are as follows.

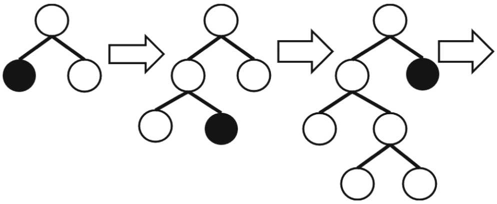

Fig. 2. Schematic diagram of the growth of a decision tree

Table 1. Parameter optimization results of the three models

	XGBoost	LightGBM	CatBoost
n_estimators _	4 00	—	—
learning_rate	0.1	—	0.05
max_depth	8	5	—
min_child_weight	7	—	—
num_leaves	—	100	—
iterations	—	—	400
depth	—	—	10

hidden_layer_sizes = (2000, 1000, 500), activation = 'relu', solver = 'adam', alpha = 0.0001, batch_size = 'auto', learning_rate = 'constant'.

The optimization results of the main parameters of the three algorithms of machine learning are shown in Table 1.

2.4 Model Fusion

Usually, the single model's own characteristics determine that the learning ability of the single model will be limited, so the prediction effect of the model fusion performance is often better than the prediction effect of the single model.

A single model is usually learned on a training set by a basic learning algorithm, which is also called a basic learner. Model fusion is a method of improving the overall performance of a model by combining a set of base classifiers in a certain way. Common model fusions mainly include averaging, voting and learning. In this paper, the weighted average method will be adopted to carry out the fusion analysis of the model.

3 Case Analysis

3.1 Data Description

This paper analyzes 30,000 used car transaction sample data collected from http://mathorcup.org//. The data includes vehicle basic information, transaction time information, price information, etc., including 36 columns of feature information, of which 15 are anonymous features, as shown in Table 2.

Table 2. Character description

Features	Description
carid	Vehicle id
tradeTime	Exhibition time
brand	Brand id
serial	car id
model	Model id
mileage	Mileage
color	Vehicle color
cityId	Vehicle city id
carCode	GB code
transferCount	Number of transfers
seatings	Number of passengers
registerDate	Registration date
licenseDate	Listing date
country	Country
maketype	Manufacturer type
modelyear	Modelyear
displacement	Displacement
gearbox	Gearbox
oiltype	Fuel type
newprice	New car price
anonymousFeature	15 anonymous features(P1, P2, …, P15)
price	Used car transaction price (forecast target)

3.2 Data Processing

3.2.1 Data Preprocessing

The data in this article are divided into text type and numerical type, among which the text type data are ‘tradeTime’, ‘registerDate’,‘licenseDate’,‘P7’, ‘P11’, ‘P12’, ‘P15’, and the rest are of numerical type. The processing is as follows.

(1) Processing of Time Features

Set data1 to ‘tradeTime’-‘registerDate’ and data2 to ‘tradeTime’-‘licenseDate’, the purpose is to convert the time field into a continuous numerical feature for subsequent processing. At the same time, data1 and data2 have a strong correlation, so a new feature ‘use_car_time’ is constructed, that is, the average use time of the vehicle is (data1 + data2)/2, and the features ‘tradeTime’, ‘registerDate’ and ‘licenseDate’are deleted.

(2) Processing of Anonymous Features

Observing the anonymous feature ‘P7’, it can be seen that it is a text type. Therefore, it needs to be converted into a numerical type, that is, the year data is extracted, and then ‘2008–2022’ is encoded as 1–15, and the vacancy value is encoded as 0. Observing the anonymous feature ‘P11’, it can be seen that it is a string type and belongs to a text type. Therefore, it is converted into a numerical type and encoded, and the ‘1’, ‘1 + 2’, ‘1 + 2,4 + 2’, ‘1,3 + 2’, ‘3 + 2’, ‘5’of anonymous features are coded as 1–6, and their vacancies are coded as 0.

3.2.2 Data Cleaning

For the missing value, it will be divided into two types of features for processing. When the feature field is continuous, use random forest to predict the vacant value, and then fill it. When the feature field is discrete, it is filled with the mode of the feature. For outlier testing, use boxplots to handle exceptions and replace outliers.

3.3 Feature Engineering and Feature Selection

3.3.1 Feature Engineering

For the four models constructed, the way the data is processed is different. ① The Artificial Neural Network model needs to perform Onehot encoding processing on the discrete features of its input, and then normalize all the features. ② For the three models of machine learning, the 32 features obtained by data preprocessing are filtered.

3.3.2 Feature Selection Based on Recursive Feature Elimination

The model is filtered through the recursive feature elimination method, the weights of the 32 features are sorted from high to low, and then the features are eliminated from low to high to train the model, and finally the feature with the highest R^2 value can be selected, as shown in the Fig. 3.

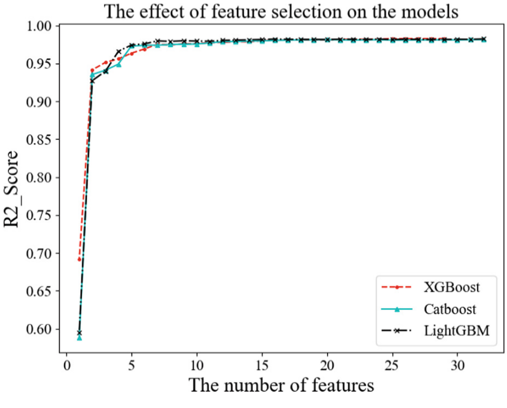

Fig. 3. The effect of feature selection

(1) Feature Screening of XGBoost

The maximum value of R^2 is 0.9827, and the number of features is 29, that is, the three features 'P1', 'country' and 'oiltype' should be eliminated according to the weight order.

(2) Feature Screening of CatBoost

The maximum value of R^2 is 0.9814, and the number of features is 31, that is, the feature 'P1' should be eliminated according to the weight sorting.

(3) Feature Screening of LightGBM

The maximum value of R^2 is 0.9828, and the number of features is 31, that is, the features 'P1', 'P10', and 'transferCount' should be eliminated according to the weight sorting.

3.4 Analysis of Model Results

3.4.1 Introduction of Model Evaluation Indicators

(1) Relative error

$$Ape = \frac{\left|\hat{y} - y\right|}{y} \tag{3}$$

(2) Average relative error

$$Mape = \frac{1}{m}\sum_{i=1}^{m} Ape_i \tag{4}$$

(3) 5% error Accuracy

$$Accuracy_5 = \frac{count(Ape <= 0.05)}{count(total)} \tag{5}$$

Among them, the true value: $y = (y_1, y_2, \ldots, y_m)$ the model predicted value $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$ count (Ape <= 0.05) is the number of samples whose Ape is within 5%, and count (total) is the total number of samples.

(4) Coefficient of Determination R^2

R^2 is the ratio of the regression sum of squares to the total sum of squares, which reflects the explanatory ratio of the independent variable to the dependent variable. The closer it is to 1, the better the model effect.

3.4.2 Comparison of the Results of the Four Models

From the comparison of the results in Table 3, it can be seen that the Mape and $Accuracy_5$ indicators of ANN are less effective, while R^2 is relatively close to the other three. The three indicators of XGBoost, CatBoost and LightGBM have good effects and are very close. Therefore, this paper selects XGBoost, CatBoost and LightGBM for model fusion, and the four are used for model fusion to explore the possibility of improving the accuracy of the model.

Table 3. Comparison of indicators of the model

	Mape	Accuracy_5	R^2
XGBoost	0.0309	0.8283	0.9827
CatBoost	0.0329	0.8118	0.9814
LightGBM	0.0310	0.8228	0.9828
ANN	0.0402	0.7486	0.9713

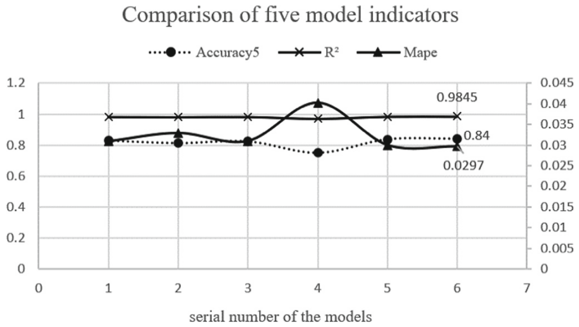

Fig. 4. Comparison of five model indicators

3.4.3 Model Fusion and Model Selection

XGBoost, CatBoost, and LightGBM are selected for model fusion and four models are fused, and the weighted average processing is performed with the R^2value, as shown in the formula (6), $w_{(i)}$ represents the single model $h_{(i)}$ weighting factor.

$$H_{(x)} = \frac{1}{T}\sum_{i=1}^{T} w_i h_{i(x)} \tag{6}$$

The model results are as follows. ① the Map value of the three-model fusion is 0.03, the Accuracy_5 is 0.837, and the R^2value is 0.9842. ② the Map value of the four-model fusion is 0.0297, the Accuracy_5 is 0.84, The R^2value is 0.9845. Contrast as shown in Fig. 4, where the serial numbers 1–6 are: XGBoost, CatBoost, LightGBM, ANN, the fusion of three models, and the fusion of four models. It can be seen that the effect of integrating the four models is the best, so this model should be selected as the used car price prediction model.

3.5 Analysis of Model Results

In this paper, 6 prediction models are established, 4 of which are single models, 2 are three-model fusion and four-model fusion respectively. The model with the best evaluation index is selected as the prediction model, and the analysis and verification are carried out with examples. The effect of the four-model fusion is optimal. The analysis shows that although the indicators of the ANN single model are not dominant, the fusion effect of the four models is better than the fusion of the three models. It can be seen that due to the differences in the construction of model feature engineering, the fusion between different types of machine learning models makes the overall better.

4 Conclusions

In this paper, by studying the price evaluation problem in the Chinese used car retail scene, a set of used car evaluation system model based on machine learning is established, that is, a model based on the weighted average fusion of XGBoost, CatBoost, LightGBM, and ANN. Its R^2 can reach 0.9845, and the prediction effect is the best. At the same time, its Mape value is only 0.0297, and the relative average error is small, which fully demonstrates the superiority of the model. Therefore, according to the model built in this paper, it can accurately predict the price of used cars, guide used car dealers and the financial and insurance industry to establish a used car price evaluation system, and promote the improvement of a reasonable and standardized used car trading market system.

References

1. Jia P (2021) Used car price prediction based on LightGBM. Shandong Normal University. https://doi.org/10.27280/d.cnki.gsdsu.2021.001186
2. Wang J (2019). Research on used car valuation model based on random forest algorithm. Beijing Jiaotong University. https://doi.org/10.26944/d.cnki.gbfju.2019.000864
3. Yang S (2020) Research on used car valuation model based on GA-MIV-BP algorithm. Chongqing University of Technology. https://doi.org/10.27753/d.cnki.gcqgx.2020.000367
4. Zhang Y (2018) A used car price evaluation model based on neural network. Tianjin University