DOI: 10.23977/jeis.2022.070119
ISSN 2371-9524 Vol. 7 Num. 1

Research on Second-hand car problem based on neural network model

Fengting Bai

Faculty of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin, 300222, China

Keywords: second-hand car price, correlation coefficient, neural network

Abstract: In order to promote the stable development of second-hand car market, this paper constructs a mathematical model to quickly and reasonably predict and evaluate each index of second-hand car. Firstly, the collected large-scale data are cleaned, and then the cleaned data are tested by J-B test. The results show that the P values of all variables are less than the significance level, and the dimension of the data can be reduced by calculating the correlation coefficient. Then, by calculating and selecting 15 variables with large Pearson correlation coefficient with the second-hand car transaction price, the neural network model is established and trained. By adjusting the super parameters and using various methods to optimize the performance of the neural network, it is finally used for the regression prediction of the data, and a more reasonable result is obtained.

1. Introduction

With the continuous increase of the scale of China's automobile market, the trading volume of second-hand cars has gradually increased. However, due to the complexity of second-hand car goods, it lacks a set of standards to judge the value of second-hand car assets. Therefore, it is an urgent need for buyers and sellers in the market to build a mathematical model to make a rapid and reasonable prediction and evaluation based on various indicators of second-hand cars. In this paper, a neural network model is established according to the provided training set for training and testing, and the model is used to predict the price of the given vehicles.

2. Model Establishment and Solution

2.1 Data processing

In this paper, 30000 data records are provided as training data. Each record has 36 features including transaction price, but some features contain certain missing values. In order to make full and rational use of training data, we need to process the data to meet our needs. The data processing stage is mainly composed of five steps, namely data cleaning, normalization processing, correlation coefficient calculation, data dimensionality reduction and data set division. In the data cleaning step, we should remove or complete the missing values in the data, modify the format and type of data, delete or modify the wrong data by consulting relevant professionals, analyzing the semantics and value range of each index, combined with the data distribution map, and then normalize the cleaned data by using SPSS software. Thus, it is convenient to analyze the characteristics related to the transaction price according to the data.

2.2 Data dimensionality reduction

We finally left 30 variables through screening and deletion, but the number of variables is still large, so we need to reduce the dimension of variables. The method we adopted here is to calculate the Pearson correlation coefficient between all variables and the second-hand car transaction price, and select 15 variables with large correlation coefficient for analysis.

The Pearson correlation coefficient must be used to ensure that the data obey the normal distribution, and the two calculated variables must have a linear relationship. In order to distinguish the linear correlation between each variable and the transaction price, we draw the scatter diagram of each variable and the transaction price. The image shows that each variable is linearly correlated with the transaction price to a certain extent, so the Pearson correlation coefficient can be used as the basis to measure the correlation. In addition, we also use matlab to conduct J-B test on the data of the remaining 30 variables, and set the significance level to 0.05. The results show that the significance of all variables is less than the level of P. Therefore, we reject the original assumption that the data conforms to the normal distribution, and the Pearson correlation coefficient can be calculated. The calculation formula is as follows:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))}{n} \tag{1}$$

X and Y are the two variables involved in the calculation, E(X) and E(Y) are their respective expectations, x_i and y_i are the i-th values of the two variables, and n is the total number of vehicles.

At this time, the covariance of the data is large. In order to solve the problem of dimensionality, we standardize the data to obtain a dimensionless quantity, and then obtain the Pearson correlation coefficient. The specific steps are as follows:

First, import the data processed in the previous step into SPSS, use SPSS for standardization, and store the standardized variables as new variables. Then use bivariate correlation analysis to calculate Pearson correlation coefficient. The formula is as follows:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - E(X))(y_i - E(Y))}{\sqrt{\sum_{i=1}^{n}(x_i - E(X))^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - E(Y))}^2} \tag{2}$$

By comparing the Pearson correlation coefficient between each variable and the transaction price, we select 15 variables with the greatest correlation with the transaction price of used cars for neural network modeling and analysis. The following table shows the correlation coefficients of these 15 variables and pairs:

Table 1: Top 15 variables of correlation coefficient

Variable name	correlation coefficient	Variable name	correlation coefficient
Brand ID	0.144	Anonymous variable 2	0.447
Train ID	0.166	Anonymous variable 4	0.077
Model year	0.263	Anonymous variable 5	0.239
City ID of vehicle	-0.099	Anonymous variable 8	0.210
mileage	-0.171	Anonymous variable 9	-0.147
displacement	0.523	Anonymous variable 10	-0.189
transmission case	0.144	Anonymous variable 11	0.367
New car price	0.765		

2.3 Establishment of evaluation index

There are three evaluation indicators selected in this paper, namely:

(1) Average relative error ($Mape$), let the true value of the i-th record in the test set be y_i the predicted value of neural network be $\hat{y}_i$, and the total number of records in the test set be n.

$$Mape = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i}, i = 1, 2 \ldots n \tag{3}$$

(2) 5% error accuracy ($Accuracy_5$)

$$Accuracy_5 = \frac{m}{n} \tag{4}$$

Where m is the number of records with Ape less than or equal to 5% in n test set records. The Ape value Ape_i of i-th record is calculated as follows:

$$Ape_i = \frac{|\hat{y}_i - y_i|}{y_i} \tag{5}$$

(3) $Criterion$ combines the first two indicators, converts the very small indicator $Mape$ into a very large indicator, and sums it with the very large indicator $Accuracy_5$ with the weight of 0.2 and 0.8 respectively. The calculation formula is as follows:

$$Criterion = 0.2*(1 - Mape) + 0.8*Accuracy_5 \tag{6}$$

2.4 Establishment of neural network model

After removing unnecessary features, we randomly divide the remaining 30000 records into training set and test set according to the ratio of 7:3. The training set is used to input the neural network, optimize the weight of each layer of the neural network through the back propagation method, and the test set is used to detect the performance of the neural network according to the evaluation index to obtain a model with excellent performance.

The neural network constructed by us consists of four layers. The first layer is the input layer, which contains 15 neurons, corresponding to 15 relevant features of the input data. The second and third layers are two hidden layers containing 18 neurons. The activation function selected by each layer is the relu function. Finally, the output layer containing one neuron corresponds to the final transaction price. The structure diagram of the whole neural network is as follows:

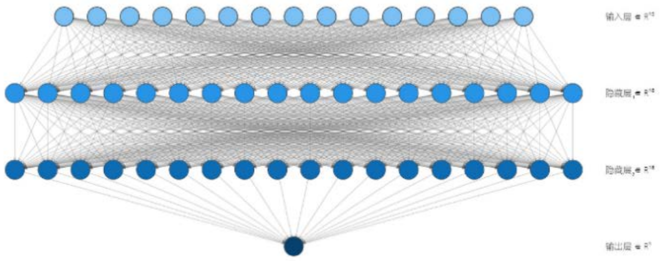

Figure 1: Structure of neural network

In order to accelerate the convergence speed of neural network training process, avoid various common problems and improve the performance of neural network model, we adopt batch processing, batch normalization, weight attenuation, Kaiming initialization and other methods (the two batch processing layers are not represented in the structure diagram). At first, we set the epoch as 20, the batch size as 256, and the learning rate (LR) as 0.01. Input the training set into the neural network for training, and then use the trained neural network to predict the test set. After that, the index values of the data in the test set are as follows:

$$\begin{cases} Mape = 0.26482418 \\ Accuracy_5 = 0.17155555555555554 \\ Criterion = 0.2842796080854204 \end{cases}$$

After that, we tried various common batch sizes, dynamically adjusted the corresponding learning rate according to the batch size, and tried to increase epoch to improve the accuracy of the model. Finally, after continuous attempts combined with experience, we determined the final super parameters as follows:

Table 2: Super parameter setting

Super parameter	**Value**
Epoch	100
BatchSize	128
LearningRate	0.001
WeightDecay	0.001

Under the condition of this super parameter, the corresponding index values of the optimal model we trained for the prediction results recorded in the test set are as follows:

$$\begin{cases} Mape = 0.19093445 \\ Accuracy_5 = 0.2208888888888888 \\ Criterion = 0.3385242212242511 \end{cases}$$

It can be found that the index value has been greatly improved compared with before. At this time, the neural network model has excellent performance. We save the model and use it for the subsequent regression prediction of the data, so as to obtain the final prediction result.

3. Conclusion

This paper adopts a reasonable way to clean and reduce the dimension of the data, then the neural network model is established and the evaluation index is established. After the training, it is tested with the test set, and the parameters are continuously adjusted according to the test results to improve the accuracy and generalization performance of the model. In addition, we also use a large number of methods to optimize the performance of neural network. Finally, we make regression prediction through the neural network model and get more reasonable prediction results.

The model can be extended to the second-hand goods trading platform. It can reasonably predict the price of second-hand goods that are difficult to be accurately valued based on the previous data. It plays a strong guiding and normative role in second-hand goods trading. At the same time, the methods in the model can be improved according to the actual situation of model application, such as using principal component analysis to reduce dimension, so as to improve the performance of the model under different service conditions.

References

[1] Liu Sen Research on used car price evaluation method based on neural network [J] Automotive industry research, 2019 (1): 21-24

[2] Zhang Cuijuan, Feng Xuejun, Sheng min Development steps of factor analysis and implementation of R language program code [J] Journal of Anqing Normal University (NATURAL SCIENCE EDITION), 2013 (2): 28-31

[3] He Xiaoqun Modern statistical analysis methods and applications 3rd Edition [M] Renmin University of China Press, 2012

[4] Edited by Si Shoukui, sun Xijing, Zhang Decun, Zhou Gang and Han Qinglong Mathematical modeling algorithm and application problem solving [M] Beijing: National Defense Industry Press, 2015.08