ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0007

The Pricing Prediction Method of Automobile Products based on the Multi-regression Algorithm Model

Hong Gao[1,a], Shenglin Wang[1,b], Keyu Yan[1,c] and Xing Liu[1,*]

[1]Nanjing Vocational College of Information Technology, Nanjing, China.

[a]1466116293@qq.com, [b]1643844478@qq.com, [c]308595564@qq.com, *414806943@qq.com

Abstract

The data analyzed are the price of a car production company in Country B, including 205 examples, 16 feature variables (no missing value), 10 metavariates, and price as the target variable. This case uses the Blue Whale Big Data Mining Platform to regression analysis of automotive historical price data, including data pre-processing, feature correlation analysis, and the establishment of automotive price prediction model and model evaluation. Using scatterplot, distribution chart, box chart and so on to analyze various variables affecting the price of automobile products, find out the variables that affect the price of automobile products the most, and finally compare the results by linear regression, neural network, constant, KNN and random forest seed algorithm, the analysis has a certain degree of accuracy, through the above algorithm comparison results, evaluate the model effect of the algorithm.

Keywords

Car Price Forecast; Blue Whale Mining Platform; Linear Regression Analysis; Neural Network; Random Forest.

1. Introduction

The theme of this paper is the prediction method of automobile product pricing based on multiple regression algorithm model. The data and pictures listed in this paper are mainly completed on a big data mining platform software of blue whale in China. The data background of the case is the price of the car sold by an automobile production company in country a in country B. We use this software to do regression analysis on the historical automobile price data, including data preprocessing, feature correlation analysis, establishment of automobile price prediction model and model evaluation.

Using scatter diagram, distribution diagram, box line diagram and so on, this paper analyzes various variables that affect the price of automobile products, and finds out the variables that most affect the price of automobile products. Finally, among the numerous data models presenting the results, we choose the most expressive models, including linear regression, neural network, constant, KNN and random forest model, by comparing the mean square error, root mean square error, mean absolute error and determination coefficient.

1.1 Pricing forecast method of automobile products

Background: Company A is a Chinese automobile manufacturing company. In recent years, with the development and growth of the company, the management hopes to expand the overseas market and produce and sell cars in country B. As the car pricing principles in different countries are different, company a needs to be familiar with the main factors influencing the car pricing in country B.

1.2 Data acquisition and processing

In the ‘data’ component of the ‘blue whale’ module, select the data table ‘automobile product data set’. We set the ‘insurance risk rating’, ‘automobile name’ and other data types as ‘category type’,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0007

the role as 'feature variable', 'price' role as 'target variable', and 'automobile number' role as 'meta variable', The data will be renamed 'automotive product data sheet'. According to the link between 'automobile product data table' and the data table in the module, there are 205 instances, 24 feature variables and 1 meta variable.

Then we classify the data. We divide the wheelbase [vehicle length, vehicle width, vehicle height, servicing mass, engine size, caliber ratio, stroke, compression ratio, maximum revolutions per minute, city mileage, highway mileage] into continuous variables, and divide [insurance risk rating, vehicle name, fuel type, supercharger type, number of doors, body type, wheel drive, Engine location, engine type, fuel system classification] are classified variables. The standard of classification is to see whether the value is continuous within a certain range and whether it is selected from certain categories. Next, we carry out the operation and processing of 'remove useless features'.

2. Data exploratory analysis

Explore the correlation of some variables: connect the 'scatter diagram' and 'feature setting' in the visualization module, select 'engine size' and 'price', find that the positive correlation is relatively strong, select 'car height' and 'price', find that the positive correlation is weak.Correlation: a positive value indicates a positive correlation between two variables, that is, one increases with the increase of the other, and decreases with the decrease, with the same change trend; A negative value indicates a negative correlation between two variables, that is, one decreases with the increase of the other, and the change trend is opposite.

Explore the correlation of numerical variables: connect 'feature selection' with 'correlation analysis' in the data module, select 'Pearson correlation coefficient' and 'all combinations', we can clearly see that: The correlation coefficient of 'city mileage' and 'Gaogao road mileage' was $0.971 > 0.8$, which has a strong positive correlation.The correlation coefficient between 'city mileage' and 'horsepower' was $-0.801 < 0$, and there was a negative correlation.

Constructing classification features: in order to analyze the correlation of variables more intuitively, we need to construct classification features. Feature engineering refers to the process of transforming the original data into the training data of the model.

Its purpose is to obtain better features of the training data and make the machine learning model approach the upper limit. Feature engineering can improve the performance of the model, sometimes even in a simple model can achieve good results. Feature engineering plays a very important role in machine learning, which generally includes three parts: feature construction, feature extraction and feature selection. Feature construction is troublesome and needs some experience. Feature extraction and feature selection are to find the most effective features from the original features. The difference between them is that feature extraction emphasizes to get a group of features with obvious physical or statistical significance through feature transformation;

Feature selection is to select a set of feature subsets with obvious physical or statistical significance from the feature set. Both of them can help reduce the dimension and data redundancy of features. Feature extraction can sometimes find more meaningful feature attributes. The process of feature selection can often show the importance of each feature for model construction.

3. Construction of pricing forecast model

Automobile brand and fuel type data distribution: we need to add automobile brand variables to the data model, and then connect the automobile brand name conversion with the 'distribution map' in the data module. After opening the distribution map, we can see the specific values.

The number of 'Toyota' vehicles is 32, accounting for 15.61% of the total number of vehicles. The number of vehicles using 'gas' fuel is 185, accounting for 90.24% of the total number of vehicles.

Price distribution of different automobile brands: we can see in the box line graph of the model that the average value of 'BMW' price is 26118.8, the median value is 22835, and the floating range is

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0007

large' The average price of 'dodge' is 7875.44, the median is 7609, and the floating range is small. Different brands of cars, there are great differences in price.

Table 1. Car brand name price share

Car brand	Price (mean)	Price (25%)	Price (75%)	Price (middle)
bmw	26118.8	18947.5	33820	22835
buick	33647	28212	38008	32892
jaguar	34600	32250	36000	35550
dodge	7875.44	6377	8558	7609
cheverolet	6007	5151	6575	6295

Feature ranking: we adopt the rrelieff method. The rrelieff method deals with the regression problem where the target attribute is a continuous value, and gives higher weight to the feature with high correlation with the target attribute' 741, which has the highest correlation with 'price'.

4. Expriment and Evaluation of pricing forecast model

Cross validation: select a part of samples in the training set to test the model, test the parameters generated by the training set, and objectively judge the consistency of these parameters with the data outside the training set.

Hierarchical cross validation: first classify the data, and then divide it according to the quantity proportion between the original data categories, so that each layer can better represent the whole and avoid random division. It can be divided into two types: F and m; The ratio of F: m in the original data is 1:3; The ratio of F and M is 1:3. In the blue whale software, among the more than ten models that the software comes with, we finally choose five algorithms with better model effect through evaluation. They are linear regression, neural network, KNN, constant and random forest algorithm.

In the evaluation results, we use four parameters to evaluate the model.

4.1 Mean square error (MSE)

In mathematical statistics, mean square error refers to the expected value of the square of the difference between the estimated value and the true value of the parameter, which is recorded as MSE. The statistical parameter is the mean value of the square sum of the errors of the corresponding points of the predicted data and the original data.

MSE is a convenient method to measure the 'average error'. MSE can evaluate the change degree of data. The smaller the MSE value is, the closer the MSE is to 0, which indicates that the model selection and fitting are better, the description of experimental data has better accuracy, and the more successful the data prediction is.

In linear regression, our goal is to minimize the loss function.The calculation formula is as follows:

$$\mathbf{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{w_i}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}} \quad (1)$$

4.2 Root mean square error (RMSE)

The statistical parameter is the mean value of the sum of squares of the errors of the corresponding points of the predicted data and the original data. In fact, it is the same as MSE in essence, but it will be used to better describe the data.

RMSE is a measure of accuracy, which is related to proportion. The influence of each error on RMSE is proportional to the square error. Therefore, RMSE is sensitive to outliers.

It is always non negative.The calculation formula is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{m} w_i(y_i - \hat{y}_i)^2} \quad (2)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202109_7(9).0007

4.3 Mean absolute error (MAE)

It represents the average of absolute errors between predicted and observed values. It calculates the average of residuals directly, and RMSE will punish more for high difference than MAE. The average absolute error can better reflect the actual situation of the predicted value error. The calculation formula is as follows:

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i| \tag{3}$$

4.4 Coefficient of determination (R2)

It is also called decision coefficient or goodness of fit. It reflects how much percentage of the fluctuation of y can be described by the fluctuation of X. It represents the degree to which the regression equation explains the change of the dependent variable, or the fitting degree of the equation to the observed value. The validity of goodness of fit usually requires: the number of independent variables: the number of samples > 1:10.The value of the coefficient of determination happens to be equal to the square of the correlation coefficient.

The coefficient of determination is suitable for linear regression. The greater the goodness of fit, the higher the degree of explanation of independent variables to dependent variables, and the higher the percentage of changes caused by independent variables in the total changes. The denser the observation points are near the regression line. Its value range is 0 to 1.

The above is a brief introduction to the selected parameters, while the figure below shows some machine learning algorithms based on ‘blue whale’ software, and compares each algorithm. Linear regression is one of the most well-known and easy to understand algorithms in statistics and machine learning, We found that the advantage of linear regression is that we can directly see the correlation between variables and target variables through the model, and the linear regression model is very easy to understand, and the results have good interpretability, which is conducive to decision analysis.

5. Conclusion

In this case, we build a prediction algorithm of automobile product price based on linear regression and neural network, and use cross validation to evaluate the model. By comparing various parameters of eight different prediction models, we finally select five better prediction models, which are random forest, linear regression, KNN, constant and neural network, It can be found that random forest is the smallest compared with other MSE, and the R2 of random forest is also the highest.

Acknowledgments

The authors acknowledge the support by Nanjing Vocational College of Information Technology Foundation (YK20210601) and the fifth ‘changfeng’ Cup.

References

[1] Wang Yongtian, Lin jingdun, Chen Jing, et al. Study on the performance of random tree feature matching operator [J]. Journal of Beijing University of technology, 2009 (11): 988-993.

[2] Liaw A, Wiener M . Classification and Regression by randomForest[J]. R News, 2002, 23(23).

[3] Hastie T J, Tibshirani R . Discriminant Adaptive Nearest Neighbor Classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996.

[4] Zhou Kaili. Neural network model and MATLAB simulation program design [M]. Tsinghua University Press, 2005.