Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070915
ISSN 2371-8838 Vol. 7 Num. 9

Analysis of the Price Influence Factors of Used Audi Cars Based on Ridge Regression Model

Yi Xu[1,*], Shengyu Yan[2]

[1]School of Digital Economy Industry, Guangzhou College of Commerce, Guangzhou, 511363, China
[2]School of Software, Taiyuan University of Technology, Taiyuan, 030600, China
[*]Corresponding author: m13059584993@163.com

Keywords: Used Car; Mileage; Year; Ridge Regression Model

Abstract: This paper uses the ridge regression model to explore the factors affecting the price of second-hand Audi cars. A large number of used Audi car feature data were collected, including the *Model, Year, Mileage* and other characteristics, as well as their corresponding price. In general, since the development of these factors is homogeneous, so most of their data have multicollinearity problems. If OLS is used to estimate the parameters of the model, the parameters obtained may be difficult to objectively and accurately reflect the actual situation [6]. Using ridge regression model for modeling and prediction to solve the multicollinearity problem by introducing a regularization term. When building the model, this text considered the correlation between features and choose appropriate regularization parameters. The experimental results show that through the ridge regression model, this text analyzed the importance of the characteristics of the regression model, and found that the regression coefficient of *Mileage Year* and *Tax* is 5.17296619, -0.60579774 and 1.46868943 respectively, indicating that mileage, age and tax are important factors affecting the price of second-hand Audi cars [3]. This study provides a reliable method for predicting the price factors of the used Audi car market, which has an important reference value for both buyers and sellers.

1. Introduction

Linear regression is one of the earliest and simplest regression models that establishes a linear relationship between the independent and dependent variables. The optimal linear fit was obtained by minimizing the sum of squares.

The general form of the linear regression model can be expressed as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p * X_p + \varepsilon[1] \tag{1}$$

In the linear regression model, we estimate the model parameters by minimizing the sum of residual squares

$$\text{Minimize} \sum \left(y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p * X_{ip}) \right)^2 \tag{2}$$

However, when there is multicollinearity between the independent variables, the estimation results of the above least squares method may be unstable and have a large variance. To solve this problem, we introduce a regularization term that limits the size of the model parameters, thus improving the stability and generalization ability of the model.

Ridge Regression The L2 norm is introduced as a regularization term, and its objective function is:

$$Minimize \sum \left(y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p * X_{ip})\right)^2 + \alpha * \sum(\beta j)^2 [4] \quad (3)$$

Among α is the regularization parameter that is used to control the degree of regularization. By adjusting the value of α, we can balance the trade-off between the sum of the fitted residual squares and the regularization terms.

2. Data introduction

2.1 Data collection

The data set is from the Kaggle website and can be obtained through the following links: https: / / www.kaggle.com/code/smailaar/auidi-vehiclespredict-regression/input

The dataset contains information on used cars from Audi cars used to predict the price of used cars. The data set covered multiple features, including *Model, Year, Price, Transmission, Mileage, FuelType, Fax, Mpg, and EngineSize*. Each feature in the dataset has its own unique meaning and data type. Among (*Model, Transmission, FuelType*) are categorical variables (*Year, Price, Mileage, Tax, Mpg, EngineSize*) are numerical variables.

The categorical variables are shown in Table 1.

Table 1: For categorical variables

Specimen	Classified Variable	Variable Value
1	Model	A1~A8 Q2,Q3,Q5,Q7,Q8 RS3~RS7 R8 S3~S5,S8 SQ5,SQ7,TT
2	Transmission	Automatic Semi-Automatic Manua
3	FuelType	Diesel Hybrid Petrol

Numerical variables are shown in Table 2.

Table 2: Numerical variables

Specimen	Numerical Variable
1	Year
2	Price
3	Mileage
4	Tax
5	Mpg
6	EngineSize

2.2 Data visualization

The number of each used Audi car was obtained through data processing and analysis, and the results are shown in Figure 1.

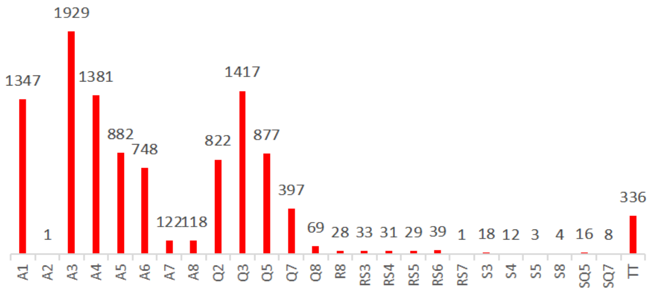

Figure 1: Number of models

Through data analysis, the four most popular models are A1, Q3, A4 and A3, and the proportion of their sales is calculated. The results are shown in Figure 2.

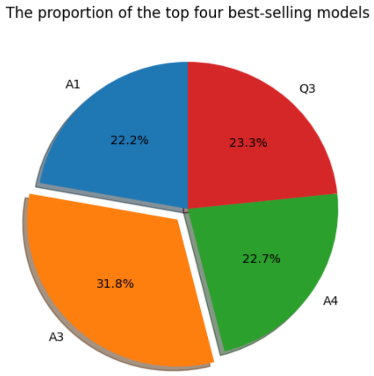

Figure 2: The four most popular models

After data analysis, in the three years from 2015 to 2018, the maximum supply of second-hand Audi cars with different gearboxes in the market every year, before 2018, the second-hand Audi cars were the most in the market and automatic transmission was the least; but in 2018, the supply was the largest and the manual transmission was the least. The results are shown in Figure 3.

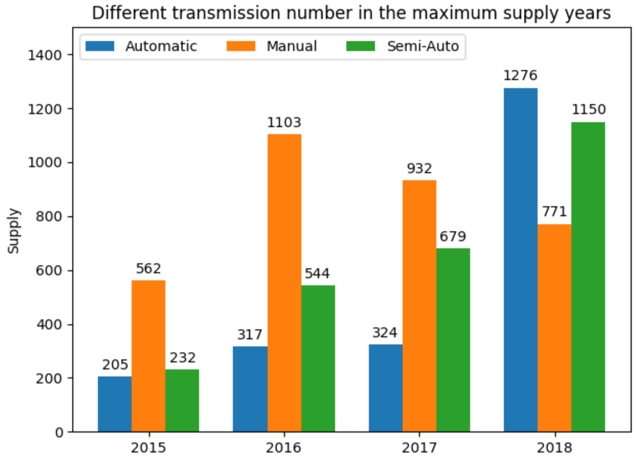

Figure 3: Maximum supply of used Audi cars with different transmissions in the market each year

By analyzing the number of different transmissions of second-hand Audi cars in the market, it is found that the most manual transmission and the least automatic transmission are the largest in the second-hand Audi car market. The results are shown in Figure 4.

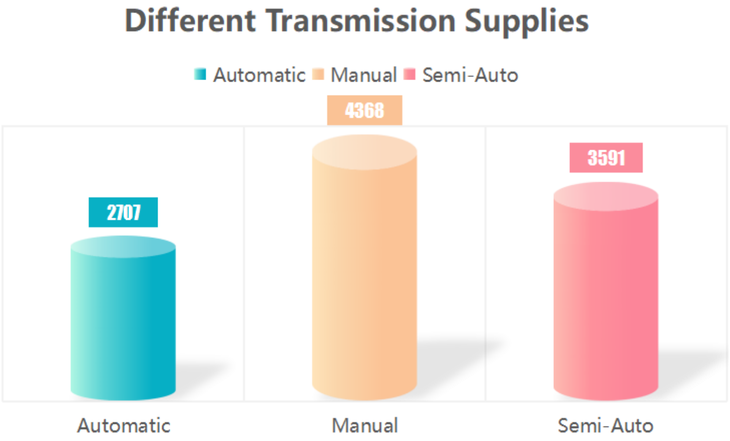

Figure 4: The number of used Audi cars with different gearboxes

By analyzing the price distribution of different second-hand Audi models, it is found that the price distribution of R8 models is the largest, that of A2 and RS7 models is the smallest, and the price is basically at the same level. The results are shown in Figure 5.

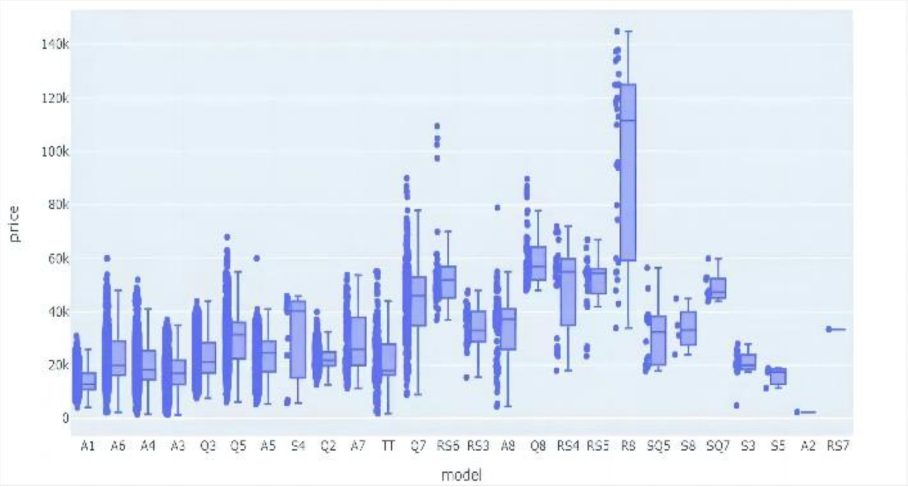

Figure 5: Price distribution of different used Audi models

3. Data research methods

3.1 Processing of the data

Firstly, the collected data were processed and the independent variables (*Model*, *Year, Transmission*, *Mileage*, *FuelType*, *Tax*, *Mpg*), fit predictive variables (*price*), and a strong collinearity between the unit price of the used Audi is too high; the common least-squares OLS regression analysis cannot be used, and a ridge regression model is needed.

3.2 Standardized processing of the data

Data Standardization, Also known as data normalization or feature scaling, is a commonly used data preprocessing technology that transforms and unifies different data according to certain rules so that they have similar scale, range or distribution.

In this paper, we need to transform six different variables (*Year, Price, Mileage, Tax, Mpg, EngineSize*) into unified standard scales. We mathematically run the raw data by Min-max normalization so that the data numerical variables are mapped in the range of 0 to 1.For each feature, we can perform the Min-max normalization in the following steps:

For each data point in each feature, Min-max was standardized using the following formula:

$$x' = (x - \min(x)) / (\max(x) - \min(x)) \tag{4}$$

Where x is the raw data, x' is the standardized data, min (x) is the minimum value of the original data, and max (x) is the maximum value of the original data.

3.3 One-hot code

One-hot code, also known as one-bit effective coding, mainly uses N-bit state register to encode N states [5]. It maps the value of each categorical variable to a new feature vector consisting of only 0 and 1, and is used to represent the different categories of the variables. The principle of single-heat encoding is to convert each category into a unique binary code.

In this paper, One-hot coding method is used to classify three features: Model, Transmission, and FuelType, establish a unique coding representation for all different values of the three classification features, and then use 0 and 1 to indicate whether this feature has this feature. For example, for the 'model' feature, if there are N different models, then N binary features are created to represent the presence or absence of each model. Single-thermal encoding can retain information about categorical features and partly avoid the influence of size relations between different values on the model.

3.4 The principle of ridge regression

$$w = (X^T + \lambda I)^{-1} X^T y[2] \tag{5}$$

λ is the ridge coefficient, I is the unit matrix (all are 1 on the diagonal, other elements are 0). The identity matrix is the full rank matrix, and multiplied by λ is still the full rank matrix.

Cost function of the ridge regression

$$\boldsymbol{J(\theta)} = \frac{\mathbf{1}}{\mathbf{2}} \sum_{i=1}^{m} \left((\boldsymbol{h_\theta(x_i)} - \boldsymbol{y_i})^{\mathbf{2}} + \boldsymbol{\lambda} \sum_{i=1}^{n} \boldsymbol{\theta_i^2} \right) \tag{6}$$

Cost function of regularization: L2 regularization (square of weights)

$$\boldsymbol{J(\theta)} = \frac{\mathbf{1}}{\mathbf{2m}} \left[\sum_{i=1}^{m} \left(\boldsymbol{h_\theta}\left(\boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)}\right) \right)^{\mathbf{2}} + \boldsymbol{\lambda} \sum_{i=1}^{n} \boldsymbol{\theta_j^2} \right] \tag{7}$$

$\lambda \sum_{i=1}^{n} \theta_j^2$Called the L2 regularization term

This penalty coefficient is a key parameter for regulating the quality of the model, and we illustrate how it regulates the model complexity through two extreme cases[4].

The λ value is 0: the loss function will be the same as the original loss function (the least squares estimation form), indicating that there is no penalty for the parameter weight θ.

λ for infinity: in the case of penalty coefficient λ infinite, in order to ensure the whole structure risk function minimized, only by minimizing the ownership weight coefficient θ, namely through the λ penalty reduces the weight of the parameter, and reduce the parameter weight while we achieve the effect of reducing the complexity of the model.

(1) By *h* _ θ (x) = θ _0 + θ _1x_1 + θ _2x_2 +... + θ _nx_n, when θ _i is small, the feature x_i is negligible.

(2) Overfitting is caused by the excessive complexity of the model.

Ridge regression was first used to handle cases where features are more than samples and is now also used to incorporate bias into the estimates to obtain better estimates. It can also solve the problem of multicollinearity, and ridge regression is a biased estimation.

Here we limit the sum of all w by introducing λ, and by introducing the penalty term, we can reduce the unimportant parameters, a technique also called reduction in statistics.

3.5 Model training

Ridge Regression Is an extended model of linear regression for processing data with collinearity. We control the complexity of the model by introducing a regularization term and reduce the variance of the parameter estimates.

The ridge regression model training steps are described as follows:

(1) The standardized numerical typed variables and categorical variables coded by One-Hot were combined into a new dataset.

(2) After the data set and completion, the dataset is divided into training set and test set for training and evaluation of the model. We divided the data set according to a certain proportion (70% is training set, 30% is test set).

(3) Ridge regression models were trained using the training set data. During training, the model optimizes the model parameters by minimizing the loss function. The ridge regression model adds an additional L2 regularization term to control the sum of squares of the model parameters.

(4) During training, methods such as cross-validation are used to select the optimal regularization parameter values. By trying different parameter values and evaluating the model performance, we select the parameters that make the model perform best on the training and test sets.

(5) The test set is predicted using the trained model, and the performance of the model on the test set is evaluated. The evaluation indicators used were the MSE and R^2. The results of the evaluation are shown in Table 3 below.

Table 3: Evaluation results

	Training data	Test data
MRSE	0.061193	0.061109
R^2	0.958	0.959

4. Results

4.1 Model evaluation

MRSE is an indicator to measure the prediction error of the model, and the smaller the value indicates the more accurate the model predicts about the target variable. The MRSE values on the training and test data were 0.061193 and 0.061109, respectively, indicating that the average error of the model was small when predicting the used car price.

R ^{2}is a measure of the model of the variability of the target variable. The values range between 0 and 1. The closer to 1, the better the model's ability to interpret the target variable. The R ^{2}values on the training and test data were 0.958 and 0.959, respectively, indicating that the model is able to explain about 96% of the price variability of the training and test data.

4.2 Regression coefficient

The regression coefficients for the different variables were obtained by model fitting, and the three features contributing the most to the prediction results are shown in Table 4.

Table 4: Regression coefficient result

	Year	Mileage	Tax
Regression coefficient	5.17296619	-0.60579774	1.46868943

5. Conclusion

5.1 Comparison of the regression coefficient

Based on the regression coefficient provided by the model, the following conclusions can be drawn:

(1)The production year, mileage and tax characteristics of used Audi cars contributed more to the forecast results, with the regression coefficients of 5.17296619, -0.60579774 and 1.46868943, respectively. This suggests that mileage, age, and taxes have significant effects on used car prices. The younger the car is, the higher the price, the less the mileage, the higher the price.

(2)The regression coefficient of other features is close to zero, indicating that other features also have a certain influence on the prediction results, but they are relatively small.

5.2. Put forward some suggestions for second-hand Audi car sellers for this study

(1) The year and mileage of used cars should be reasonably considered when pricing, and the market conditions and the price level of similar models should be consulted to ensure the competitive pricing.

(2) For second-hand cars with more mileage or long use time, necessary repair and maintenance can be considered, such as replacement of worn parts, cleaning or replacement of interior decoration, appearance beautification, etc. By repairing and improving the car conditions, improve the overall quality of used cars, and increase the interest of buyers and their willingness to buy them.

(3) Pay attention to the influence of tax factors. According to the model coefficient, taxes and fees also have a certain impact on the price of used cars. Sellers should understand and accurately convey the tax information in advance, to avoid the price uncertainty caused by the change of taxes, so that consumers have doubts about the price of second-hand cars.

5.3 Opinions on Audi car manufacturers

(1) According to the model coefficient, the age and mileage have a great impact on the price of Audi used cars, indicating that consumers are very concerned about the wear and tear of second-hand car engines. Manufacturers should pay attention to improving the engine technology of cars to reduce the loss of engines.

(2) Manufacturers need to provide perfect after-sales service, such as car repair, car maintenance, to increase the retention rate of used Audi cars.

6. Summarize the subject

6.1 Innovation points of the model

(1) The traditional linear regression model is prone to overfitting phenomenon in the presence of collinearity (highly correlation between independent variables), while the ridge regression model can effectively deal with the collinearity problem and improve the generalization ability of the model by introducing regularization terms.

(2) The prediction of second-hand car price is affected by multiple characteristics, including models, mileage, number of years, car condition, etc. When using ridge regression models, these

features can be appropriately selected, transformed and standardized to improve the prediction accuracy of the model.

(3) The lambda or alpha in the ridge regression model needs to be adjusted for the best model performance. Through cross-validation, appropriate regularization parameters can be selected to improve the predictive power of the model.

6.2 Study significance

(1) Market pricing: Understanding the main factors affecting the price of used cars can help market participants (such as sellers, buyers, and dealers) to price more accurately.

(2) Policy support: For consumers, understanding the factors influencing the price of second-hand cars can provide decision support. They can better assess whether used car purchases are value for money and consider important influences when choosing models and configurations.

(3) Market forecast: With the help of the ridge regression model, we can identify the factors that have a significant impact on the price of second-hand cars, such as car age, mileage, brand, etc. Based on these factors, we can establish a model to predict the changing trend of second-hand car prices and provide the basis for market prediction for market participants.

(4) Product improvement: By analyzing the factors that affect the price of second-hand cars, we can understand the preference degree of buyers for different factors. For example, consumers may be more willing to pay higher prices to get low-mileage, recent models. These insights can provide automakers with evidence for product improvement and positioning strategies.

References

[1] Liu Hongyong, Hu Jian, Wang Peng, Guo Ji. Study on the influencing factors of real estate price in Sichuan Province based on ridge regression method [J]. The Practice and Understanding of Mathematics, 2014, 44 (12): 72-78.

[2] Wen-Zhu K , Lin L .Analysis and Prediction of the Factors That Influence the Price of the Second-hand House in Nanning[J].Journal of Guangxi Academy of Sciences, 2012.

[3] Zhang Kun, Zhou Yunlong. Main factors influencing the price of second-hand car and purchase precautions [J]. Automobile maintenance and repair, 2021, (02): 68-69.

[4] Li Yanan, Chen Jianguo. Application of Lasso and ridge regression in rice genome-wide prediction [J]. Journal of Hubei University (Natural Science Edition), 2020, 42 (04): 384-389.

[5] Liang Jie, Chen Jiahao, Zhang Xueqin, Zhou Yue, Lin Jiajun. Anomaly detection based on single-thermal encoding and convolutional neural networks [J]. Journal of Tsinghua University (Natural Science Edition), 2019, 59 (07): 523-529.

[6] Wang Yan. Analysis of the influencing factors of the retail sales of social consumer goods in Jiangsu Province based on ridge regression [J]. Journal of Xuzhou Institute of Technology (Social Science Edition), 2019, 34 (02): 52-60.