Financial Engineering and Risk Management (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/ferm.2023.061101
ISSN 2523-2576 Vol. 6 Num. 11

Big Data Automobile Price Prediction Based on Elastic Network Regression Model

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

[1]*School of Software, Taiyuan University of Technology, Taiyuan, 030600, China*
[2]*School of Digital Economy Industry, Guangzhou College of Commerce, Guangzhou, 511363, China*
[*]*Corresponding author: A1424576738@163.com*

Keywords: Elastic Network Regression; Used Car; Prediction

Abstract: At present, with the continuous improvement of people's living standards, cars have become an essential travel tool for every family, and may even become the third biggest life scene. At the same time, the number of cars flowing into the used car market is growing, and the used car trading market is also growing rapidly. However, the price of used cars is affected by many different factors, and there is no uniform pricing standard. In view of this, in the used car trading market, it is very important to accurately predict the price of used cars for both sellers and buyers. In this paper, the elastic network regression model is used to establish the used car price prediction model. The RMSE value of the test data is 0.089497. Among the model coefficients, the characteristics of model and year have the greatest impact on the used car price, which are 0.87491361and -0.74483197, respectively.

1. Introduction

The development of the transition from linear regression models to regularization can be traced back to statistical and machine learning research since the 1970s.

As we all know, linear regression models are described.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \quad (1)$$

Where y is the target variable (or dependent variable),$x_1, x_2, \ldots, x_p$ is the characteristic variable (or independent variable),$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ is the coefficient of the model and ε is the error term. The goal of linear regression is to find the optimal coefficient value so that the difference between the predicted value and the actual observed value is minimal.

The earliest regularization method is Ridge Regression, developed by statisticians Arthur E.Hoerl and Robert W.Kennard (1970).Ridge regression introduces L2 regularization term, which controls the complexity of the model by punishing the sum of squares of model parameters, solves the multicollinearity problem, and improves the stability of the model[1].

$$Loss = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha ||\beta||^2 \quad (2)$$

After ridge regression,L_1 Lasso Regression was proposed (Tibshirani,1996), which introduced

L_1 regularization terms and achieved sparsity by punishing the absolute values of model parameters[2], even if the coefficients of some features were zero, so as to achieve the effect of feature selection.

$$Loss = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha||\beta|| \quad (3)$$

The elastic network model was proposed by statisticians Hastie, Tibshirani and Friedman in 2005.In their classic The Elements of Statistical Learning, elastic network was introduced as a combination of L1 and L2 regularization of the linear regression method[3][2].The proposed method provides an effective solution for solving highly correlated and multicollinearity problems, and has been widely used in the field of machine learning and statistics. Elastic network regression analysis is a linear regression method that combines L1 and L2 regularization. By introducing both L1 and L2 regularization terms into the loss function, it makes comprehensive use of the advantages of both to control the complexity of the model and achieve feature selection and parameter contraction. The loss function of elastic network regression can be expressed as:

$$Loss = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha(\lambda_1||\beta|| + \lambda_2||\beta||^2) \quad (4)$$

L_1 regularization terms tend to produce sparse solutions, even if the coefficients of some features are zero, thus achieving the effect of feature selection. This is because L_1 regularization penalizes the absolute value of the parameter, encouraging it to shrink to zero.

The L_2 regularization term reduces the difference between parameters by penalizing the sum of squares of the parameters, making the parameters smoother. L_2 regularization helps mitigate collinearity problems between features and provides more stable parameter estimates.

Elastic network regression can find a balance point between L_1 regularization and L_2 regularization by adjusting α , λ_1 and λ_2 hyperparameters. When α=0, elastic network regression is equivalent to ordinary linear regression. When α=1, elastic network regression is equivalent to Lasso regression. When λ_1=0, elastic network regression is equivalent to ridge regression.

By introducing $L_1$and L_2 regularization, elastic network regression has advantages in dealing with high-dimensional data and in cases where collinearity problems exist. It can effectively select related features, improve the generalization ability of the model, and for cases with highly correlated features, it can be inclined to select their common influence rather than randomly selecting one or the other. Elastic network regression is a flexible and powerful regression analysis method, which is widely used in various data modeling and prediction tasks[4].

2. Data Description

Table 1: The categorical variables

The categorical variables	Explain	Variable value
Model	Model of used car	A1~A8 Q2,Q3,Q5,Q7,Q8 RS3~RS7 R8 S3~S5,S8 SQ5,SQ7,TT
Transmission	Used car transmission type	Automatic Semi-Automatic Manual
FuelType	Used car fuel type	Diesel Hybrid Petrol

The data set is sourced from the Kaggle website and can be accessed at the followinglinks: https://www.kaggle.com/code/smailaar/auidi-vehiclespredict-regression/input.

The dataset is a predictive regression task for used cars in Audi. The data set contains used car information from Audi cars and is used to predict the price of used cars. The data set covers several characteristics, including model, year, price, transmission type, mileage, fuel type, taxes, miles per gallon and engine size. Each feature in the data set has its own unique meaning and data type. Table 2-1 lists the categorical variables and their values and meanings

Table 1 lists the categorical variables and their values and meanings.

Table 2 lists the numerical variables.

Table 2: The numerical variables.

The numerical variables	Explain	Remark
Year	Year of production of the used car	1997~2020
Price	The price of the used car	The monetary unit is expressed in US dollars
Mileage	The total number of miles driven on the used car	In miles
Tax	Tax on used cars	The monetary unit is expressed in US dollars
Mpg	Fuel efficiency of used cars	The average number of miles per gallon of gasoline or diesel
Enginesize	The displacement of the used car engine	It is usually expressed in liters (L)

3. Data Exploration

As shown in Figure 1, a comparison of car sales of three transmission types in each year shows that manual transmission was very popular before 2016, but has been declining since then.

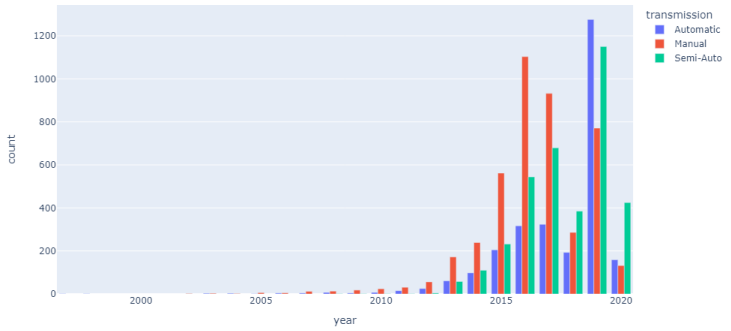

Figure 1: Sales of different types of transmission cars

As shown in the box diagram in Figure 2, the price of R8 type cars is high, and its price dispersion degree is large, and the data is skewed to the left.

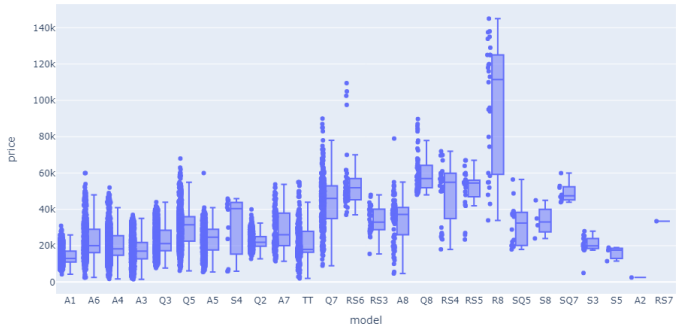

Figure 2: Box Plot

As shown in the correlation coefficient chart in Figure 3, the price of used cars is positively correlated with the year of production of used cars and engine displacement, and the correlation coefficient is as high as 0.59;

while the year of production of used cars is negatively correlated with the number of miles driven, and the coefficient is as high as 0.79; the tax rate and fuel efficiency are also negatively correlated, second only to the correlation coefficient between the year of production and the number of miles driven.

The year of production of used cars is negatively correlated with the number of miles driven, the coefficient is as high as 0.79, and the tax rate and fuel efficiency are also negatively correlated, second only to the correlation coefficient between the year of production and the number of miles driven.

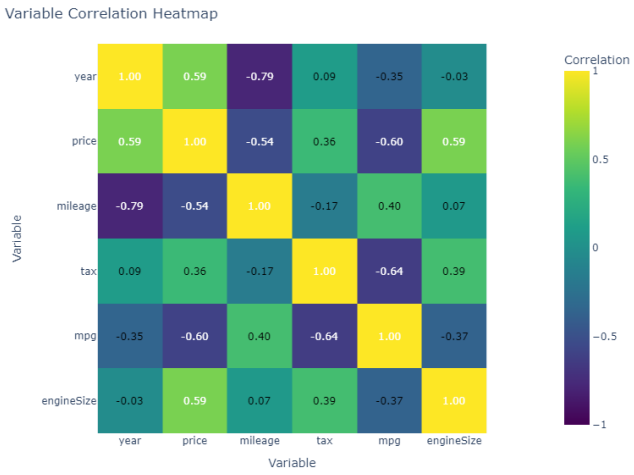

Figure 3: Correlation coefficient diagram

4. Methods

4.1 Data Standardization

Data Standardization, also known as data normalization or feature scaling, is a common data preprocessing technique used to transform features of different scales, ranges, or units into a form with uniform standards. Standardization ensures comparability between different features and eliminates the impact of scale differences on model training and performance evaluation.

The advantages of data standardization include:

(1) Eliminating dimensional differences between features and ensuring that different features are

comparable.

(2) Improve the convergence speed and stability of the model, and avoid the influence of too large or too small feature weights on model training.

(3) Improve model performance, improve prediction accuracy and generalization ability.

In this paper, we perform Min-max normalization on the ['year', 'price', 'mileage', 'tax', 'mpg', 'engineSize'] columns in the data set. Min-max normalization maps data linearly to a specified range of minimum and maximum values, typically [0,1] or [-1,1].

For each feature, we can normalize Min-max by following these steps:

(1) Find the minimum (min) and maximum (max) values for each feature.

(2) For each data point in each feature, Min-max normalization is performed using the following formula:

$$\text{standardize\_value} = \frac{\text{value} - \min}{\max - \min} \tag{5}$$

Where value is the value of the original data point.

In this way, for each feature, all data points will be mapped to the range [0,1].

By Min-max normalization of the data for the 'year','price','mileage','tax','mpg','engineSize' columns, we can eliminate the dimensional differences between these features and ensure that they are comparable, as shown in the figure below. This allows the model to better handle and interpret these features, and improves the model's performance and generalization ability.

4.2 One-hot encoding

One-Hot coding is a method to convert discrete features into binary vectors. It represents each discrete value as a binary vector where only one element is 1 and the rest are 0. The purpose of One-Hot coding is to convert discrete features into numerical representations that machine learning algorithms can handle.

For each classification feature, we can do the following One-Hot encoding:

1) Identify all the different values (categories) in each classification feature.

2) Create a new binary feature (dummy variable) for each different category.

3) For each sample, if the classification feature value of the sample matches a certain category, it is marked as 1 in the corresponding dummy variable; otherwise, it is marked as 0.

In this article, we will One-Hot code the 'model', 'transmission', and 'fuelType' columns. For each feature, we convert all of its different values (categories) into the corresponding binary feature. For example, for the 'model' feature, if there are N different models, N binary features will be created to represent the presence or absence of each model.

With One-Hot coding, we can transform classified data into the numerical form required by machine learning models for training and prediction. In this way, the model can make better use of these features and achieve better results when processing classified data.

4.3 Elastic network regression model

We use elastic network regression as a modeling method. Elastic network regression is a linear regression method that combines L_1 and L_2 regularization terms, which can simultaneously achieve variable selection and reduce the effect of overfitting. By adjusting the regularization parameters, we can balance the complexity and goodness of fit of the model to obtain better prediction results.[5-6]

Merge standardized numeric variables and One-Hot encoded categorical variables into a new data set.

Ensure that the feature order in the merged data set is consistent with the feature order before

standardization and coding to ensure correct data correspondence.

Once the data is assembled and complete, we can divide it into a training set and a test set for training and evaluation of the model. Common methods for dividing training set and test set include random partitioning and cross-validation.

Before training with an elastic network model, we need to input the training set into the model and train the model. The elastic network model combines $L_1 \, and \, L_2$ regularization, and the effect of feature selection and parameter contraction can be considered simultaneously in the model.

The steps of model training are as follows:

(1) Define the elastic network model and set the range of regularization parameters.

(2) Use the features and target variables of the training set for model training. This adjusts the weights and biases of the model to minimize the loss function on the training set.

(3) Use the trained model to predict the test set, and evaluate the performance of the model on the test set. The evaluation index for the mean square error (Mean Squared Error, MSE) and the correlation factor (Coefficient of Determination, R^2). Table 3 lists the evaluation results.

Table 3: Result

	Training data	Test data
RMSE	0.0894	0.0895
R^2	0.9123	0.9133

Through such a training and evaluation process, we can get a predictive model trained by an elastic network model and use that model to make predictions and inferences about new data.

5. Results

5.1 Model Evaluation

RMSE (root mean square error) is a measure of the model's prediction error, and the smaller the value, the more accurate the model's prediction of the target variable. The RMSE values on the training data and the test data are 0.089432 and 0.089497 respectively, indicating that the average error of the model in predicting the used car price is small.

R^2(coefficient of determination) is a measure of how well the model explains the variability of the target variable, with a value ranging from 0 to 1, with the closer to 1 indicating the better the model's ability to explain the target variable. The R^2 values on the training and test data are 0.912 and 0.913, respectively, indicating that the model can explain about 91% of the price variability of the training and test data.

5.2 Regression coefficient

According to the model coefficients provided, we can draw the following conclusion:

The features of 'model' and 'year' contribute more to the prediction results, and their coefficients are 0.87491361 and -0.74483197, respectively. A positive coefficient indicates that the price will increase as 'model' and 'year' increase, while a negative coefficient indicates that the price will decrease as 'year' increases. This shows that the model and the age of the car have a significant impact on the used car price.

The characteristics of 'mileage', 'mpg', 'engineSize' and 'tax' also have some influence on the prediction results, but relatively small. Their coefficients are -0.02852279, -0.00735012, -0.00472021 and 0.01303087, respectively. Negative coefficients indicate that the price may decrease as these eigenvalues increase, while positive coefficients indicate that the price may increase.

The near-zero coefficients for 'transmission', 'fuelType' and other features suggest that they have a

small impact on the predicted results and may not be a key factor.

6. Conclusion

6.1 Recommendations for automobile manufacturers

According to the conclusions of RMSE and model coefficient provided previously, the following recommendations can be made for automobile manufacturers:

1) The model and the age of the car have a significant impact on the used car price. Therefore, car manufacturers can consider designing and launching new models that appeal to consumers, and focus on vehicle quality and technological innovation to increase the added value of vehicles. In addition, providing a certain vehicle protection rate and after-sales service can improve the second-hand transaction price of vehicles.

2) Features such as mileage, fuel efficiency and engine size also have a certain impact on used car prices. Therefore, automobile manufacturers can focus on improving the fuel efficiency and engine technology of vehicles to reduce the cost of vehicle use and increase the market competitiveness of used cars.

3) Understand consumers' sensitivity to taxes and fees. According to the model coefficient, taxes and fees have a certain impact on the price of used cars. As a result, automakers can understand market acceptance of taxes and fees and factor them into their design and pricing strategies.

4) In addition, the near-zero model coefficient has a small impact on used car prices, including transmission type and fuel type. Therefore, these characteristics can be flexibly adjusted according to market demand and trends, but they do not have to be regarded as key competitive factors.

6.2 Recommendations for car dealers

According to the conclusions of RMSE and model coefficient provided previously, the following suggestions can be made for automobile sellers:

1) Pay attention to the impact of model and vehicle age on used car prices. According to the model coefficient, the model and the age of the car have a great influence on the used car price. Therefore, car sellers can reasonably choose a combination of models according to market demand and trends, and provide used cars of different ages for consumers to choose from. In addition, promotional campaigns can be launched to attract consumers to buy newer used cars.

2) Understand market preferences for features such as mileage, fuel efficiency and engine size. According to the model coefficient, these characteristics have a certain impact on the price of used cars. Car sellers can understand consumer preferences for lower mileage, higher fuel efficiency and moderate engine size and offer vehicle choices accordingly.

3) Provide vehicle maintenance and warranty records. Given the feature weights in the model coefficients, consumers pay a lot of attention to the vehicle's maintenance and repair history. Car sellers can provide detailed maintenance and warranty records to increase consumers' trust and sense of value in used cars.

4) Pay attention to the impact of tax factors. According to the model coefficient, taxes and fees also have a certain impact on the price of used cars. Sellers can understand and accurately convey tax information in advance, avoid price uncertainty caused by tax changes, and provide transparent pricing strategies.

5) For the characteristics of the model coefficient close to zero, such as transmission type, fuel type, etc., the seller can flexibly adjust according to market demand and trend, but it does not have to be a key competitive factor. Emphasis is placed on features that are more important to consumers, such as model, vehicle age, mileage, etc.

In general, based on the analysis of RMSE and model coefficient, automobile manufacturers can improve the used car price and market competitiveness by focusing on the improvement of model design, vehicle quality, vehicle added value, fuel efficiency, tax policy and other aspects. It is also important to understand consumer needs and trends and to be flexible in adjusting product strategies. Car sellers can increase the market competitiveness and sales success rate of used cars by understanding market needs and trends, providing models and vehicle features that meet consumer preferences, providing detailed maintenance and warranty records, and transparently communicating tax information.

References

[1] Hoerl A E, Kennard R W. Taylor & Francis Online: Ridge Regression: Applications to Nonorthogonal Problems - Technometrics - Volume 12, Issue 1[J]. Technometrics [2023-08-15].

[2] Tibshirani R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society, Series B, 1996, 58(1). DOI:10. 1111/j. 2517-6161. 1996. tb02080. x.

[3] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2001[J]. Journal of the Royal Statistical Society, 2004, 167(1):192-192. DOI:10. 1111/j. 1467-985X. 2004. 298_11. x.

[4] Jiang Shiqi, Dai Jiajia. An improved elastic net estimation of Logistic regression Model [J]. Mathematical Theory and Application [2023-08-15].

[5] Jeon S , Hong B , Chang V .Pattern graph tracking-based stock price prediction using big data[J].Future Generation Computer Systems, 2017, 80(MAR.):171-187.DOI:10.1016/j.future.2017.02.010.

[6] Han Qing, Wang Ziqi, Geng Wenjing. Research on Stock price based on Elastic net-autoregressive model [J]. Guangxi Quality Supervision Guide, 2020(10):2.