Used Cars Price Prediction - Report

Roxy Zhang

3/25/2022

Contents

Introduction

Since the market of used cars is growing, there is an increasing need of the information of used cars prices. Therefore, I want to build a model predicting the price using attributes of the car as predictors, and thus provide information for the potential buyers for decision making.

Our data is from Kaggle. This dataset contains 22 (1998-2019) years of used cars information. There are 5872 rows and 11 predictors, among which 5 are numeric and 6 are categorical.

Data Cleaning

For better illustration, I transformed the original unit (100000 Indian Rupee) of response variable **price** to USD, using the exchange rate of 100000 Indian Rupee equals to 1309.75 USD on March 21, 2022. For numeric variables with unit, the units are deleted for model-building. I also delete the car model from the original `name` variable, keeping only the brand name, since the model can be greatly explained by the predictors I are using. The type `Fourth` in owner_type is actually Fourth and above. For analysis completeness, I dropped NA values. I also filtered out a row containing an extremely large value of kilometers_driven.

Below are the variable used:

name: The brand of the car.
location: The location in which the car is being sold or is available for purchase.
year: The year or edition of the model.
kilometers_driven: The total kilometres driven in the car by the previous owner(s) in KM.
fuel_type: The type of fuel used by the car.
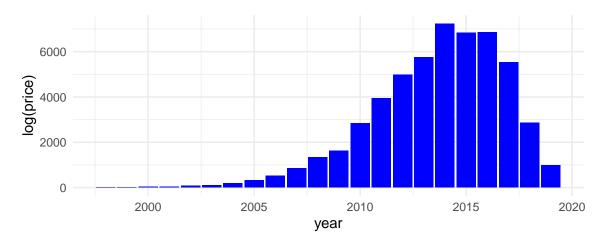**transmission**: The type of transmission used by the car.
owner_type: Whether the ownership is Firsthand, Second hand or other.
mileage: The standard mileage offered by the car company in kmpl or km/kg.
engine: The displacement volume of the engine in cc.
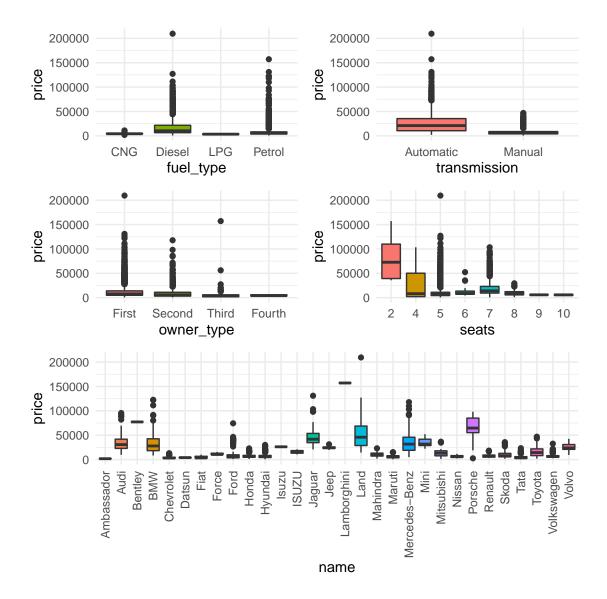**power**: The maximum power of the engine in bhp.
seats: The number of seats in the car.
price: The price of the used car in US Dollar.

Exploratory Data Analysis

First, I plot the price against year. For better scaling, the price is log-transformed. There is an increasing trend of price from 1998-2014, while the price goes down rapidly since 2015.

Then I look at the price distribution for 5 categorical predictors respectively. There are some interesting findings:

1. There are several attributes associated with higher price: fuel type diesel, automatic transmission, first-hand owner, and with 2 seats.

2. The price range for each brand varies a lot. For example, Lamborghini has a condensed price range slightly above 150k (which is not surprising), while the majority stays below 25k.

Data Partitioning

I split the data into training data and testing data by the proportion of 0.8:0.2.

Model Building

What predictor variables did you include? What technique did you use? What assumptions, if any, are being made by using this technique? If there Ire tuning parameters, how did you pick their values? Discuss the training/test performance if you have a test data set. Which variables play important roles in predicting the response? What are the limitations of the models you used (if there are any)? Are the models flexible enough to capture the underlying truth? . . .

5 models are used to fit the training data: Lasso, Elastic Net, Principal components regression (PCR), Partial least squares (PLS), and Multivariate Adaptive Regression Spline (MARS). Cross validation is used to select the best parameter or parameter combination for each model.

Lasso

The tuning parameter λ controls the L1 regularization, as λ increases, the number of predictors in the model decreases. Setting the candidate values of λ to be from 0.1353353 to 54.59815 with 100 steps, the best-tune λ is 2.069.

Elastic Net

Conclusions

What Ire your findings? Are they what you expect? What insights into the data can you make?