Predicting car prices using regression and regularization methods

Melvin Biju (T00706241)

DASC 5420: Theoretical Machine Learning
Faculty of Science
Thompson Rivers University
15 April 2023

Abstract

Cars are almost a necessity in the current world. With cutthroat completion, it is imperative that a car be well priced for it to be successful in the market. This study analyzes the characteristics of cars with price in context and identifies major factors that constitute the price of a car. It also uses these identified factors to predict car prices. It makes use of the characteristic's engine size, horsepower, drive train which can be measured easily before the vehicle goes on sale. It does an exhaustive comparison of regression techniques namely – Stepwise regression, Ridge regression, LASSO Regression and Elastic Net regression to predict car prices. The performance metric RMSE and R-squared vales are used to determine which model performs better. This study can be used has a guide for car manufactures to price their cars based on the car's characteristics. This can in turn help them to position themselves better in the market and edge out competition.

Keywords: Machine Learning, Automobile, Price, Prediction, regression, regularization

Introduction

Car prices are a crucial factor in determining the success of a car. Although the buying decision relies on a host of factors, price is the deciding factor most of the time. With the introduction of data science in business sectors, companies look towards techniques to improve their revenue using the data that is readily available to them. This is where this study comes into play. All the variables in the dataset are information that the company has before the car goes on sale. While controlling the other factors, manufactures can play with the price of the car and see how much margin they can afford for each unit sale.

The response variable "price" is predicted using various car characteristics like – aspiration, fuel-type, curb weight, engine type etc. Each regression technique performs a selection of variables that it deems fit for predicting the price of the car. We choose the best model based on the performance evaluation metric R-squared and RMSE. Typically, a high R-squared value and a low RMSE score is considered as the better model. This study will also help car manufactures to determine which characteristic of the car influences the car prices and can take managerial steps to maximise profit

Data

The dataset use for this study is sourced from *UCI Machine Learning repository.* The original data is sourced from "*1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.*" [1]. The dataset is divided into three types of entities.

1. Characteristics of an automobile

2. Insurance risk rating referenced from Insurance Institute of highway safety, Washington [1]
3. Normalized values of losses in usage in comparison to other cars. It is defined as the average loss of per car per year.

The entire dataset has 26 columns and 205 instances. Here we have considered *Price* as the response variable and the rest of the attributes as predictor variables. The other attributes are given as:

1. ***symboling***: It is a categorical variable and ranges from -3 to 3. Higher value indicates higher risk
2. ***normalized-losses***: It is a continuous variable and ranges from 65 to 256.
3. ***make***: This variable is categorical and gives us the name of the car manufacturer. It has values like “alfa-romero”, “dodge” etc.
4. ***fuel-type***: It is a categorical variable. It has values - diesel, gas
5. ***aspiration***: It is a categorical variable. It denotes if the car has a turbo or is naturally aspirated. It has two values – turbo and std
6. ***num-of-doors***: It is a categorical variable and gives us the number of doors the car has. Its values are - four, two.
7. ***body-style***: The structure of the car, It has values - hardtop, wagon, sedan, hatchback, convertible.
8. ***drive-wheels***: It specifies the drive train of the car. It has values - 4wd, fwd, rwd.
9. ***engine-location***: It denotes the placement of the engine. It has values - front, rear.
10. ***wheel-base***: It is the distance from the edge of the left wheel to the edge of the right wheel. It is a continuous variable with values ranging from 86.6 to 120.9.
11. ***length***: It denotes the length of the car. It is a continuous variable with values from 141.1 to 208.1.
12. ***width***: It denotes the width of the car. It is a continuous variable with values from continuous from 60.3 to 72.3.
13. ***height***: It denotes the height of the car. It is a continuous variable with values from continuous from 47.8 to 59.8.
14. ***curb-weight***: It denotes the weight of the car. It is a continuous variable with values from 1488 to 4066.
15. ***engine-type***: It is a categorical variable. It denotes the engine type of the car. It has values - dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
16. ***num-of-cylinders***: It is a categorical variable. It denotes the number of cylinders the engine constitutes. It has values - eight, five, four, six, three, twelve, two.
17. ***engine-size***: It denotes the size of the engine. It is continuous and has values ranging from 61 to 326.
18. ***fuel-system***: It specifies the fuel technology used to power the car. It is categorical and has values - 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
19. ***bore***: It specifes the bore size of the cylinders and headers. It is a continuous variable with values from 2.54 to 3.94.
20. ***stroke***: It specifes the stroke displaced by the pistons. It is a continuous variables with values from 2.07 to 4.17.
21. ***compression-ratio***: It is a continuous variable ranging from 7 to 23.
22. ***horsepower***: It is a continuous variable ranging continuous from 48 to 288.

23. ***peak-rpm***: It is a continuous variable ranging continuous from 4150 to 6600.
24. ***city-mpg***: It is a continuous variable ranging continuous from 13 to 49.
25. ***highway-mpg***: It is a continuous variable ranging continuous from 16 to 54.
26. ***price***: This is our response variable. It is continuous and ranges from 5118 to 45400.

Exploratory Data Analysis

1. Data wrangling and pre-processing:

    a. *Handling null values*: Upon initial EDA, there was no null values. But upon closer inspection, it was found that many columns had "?" values. For the variable ***normalized-losses***, there was 41 observations of "?" value. Thus it was interpolated using the mean of the column. The variables ***num-of-doors, bore, horsepower, peak-rpm*** and ***price*** had missing values. But they were a maximum of 4 and these observations were omitted.

    b. *Outlier detection:* The dataset was checked for outliers by using a box plot for the numerical variables. The plot below (Figure 1) shows the presence of outliers in almost all numerical variables. Here the IQR (Interquartile Range) technique was used to remove the outliers. A custom function was used to achieve this.

    c. *Scaling:* The continuous variables were scaled. This was done to improve optimization. This was done using the scale () that is in-built in R

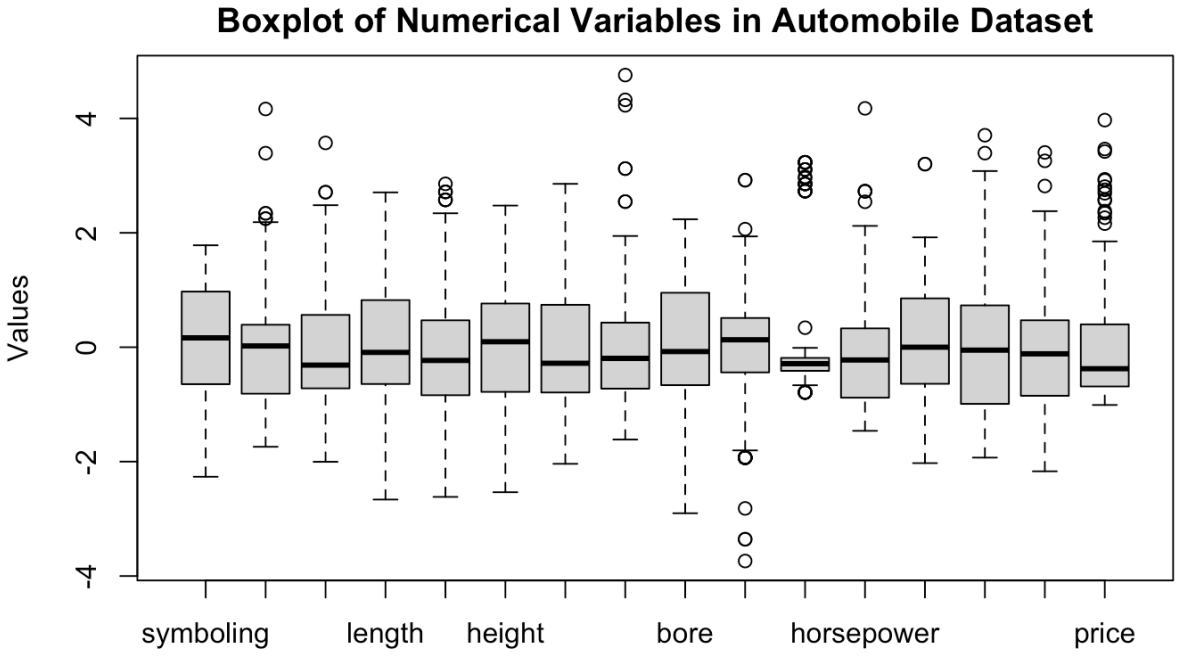

Figure 1: Boxplot of all the numerical variables

    d. *Converting categorical variables into numeric for analysis:* EDA was done on the categorical variables with the response variables to check for any imbalances.

Figure 2 shows that there are no major imbalances, and the values are distributed reasonably well. So, there is no binning required and we convert them to numeric variables.

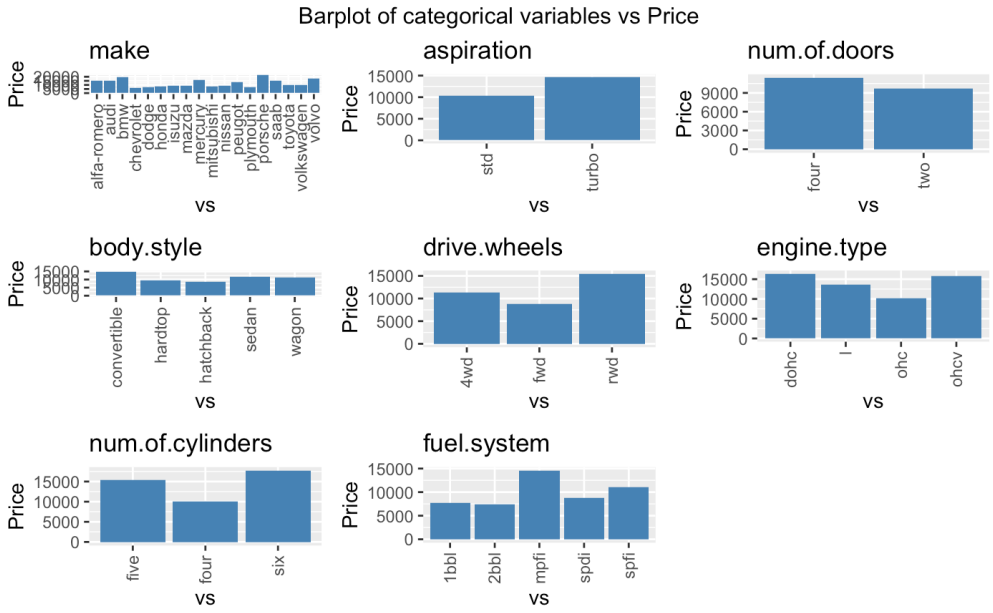

Figure 2: Barplot of all the categorical variables vs Price (response variable)

Visualizations

Distribution of response variable: From Figure 3, we establish that the distribution is slightly right skewed. This is expected Price data as there are lot of cheap cars available in the market, whereas the option for expensive cars is comparatively less.

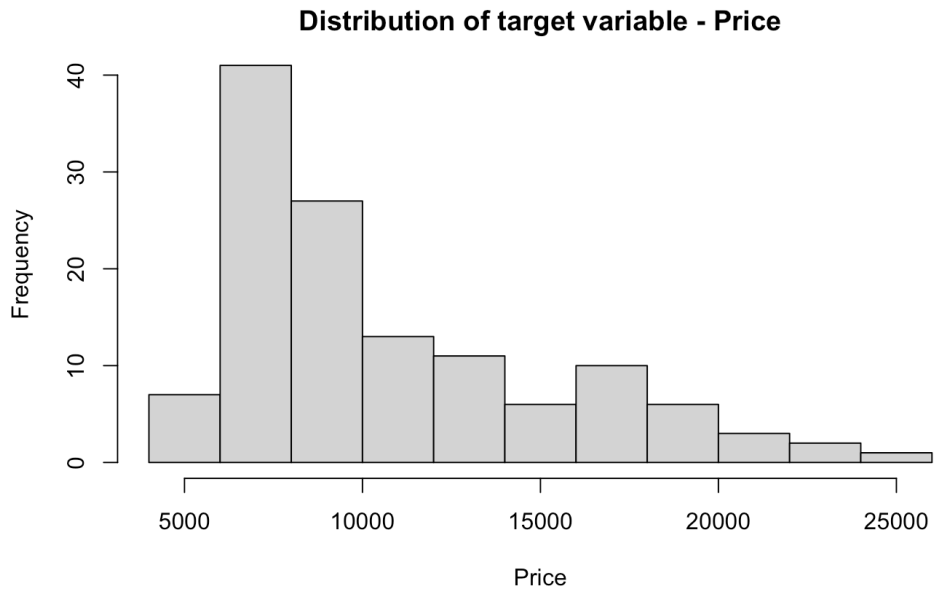

Figure 3: Histogram showing the distribution of Price variable.

Methodology

Github link: https://github.com/d250kid/DASC_5420_final_project

Initially, the cleaned data is split into training and testing data. For this study, the training data constitutes 80% and testing data constitutes of 20% of the entire dataset. We make use of the training dataset to train our model. Later we evaluate the performance of our model using test data. Here we have considered the *RMSE* and *R-squared* values as our evaluation parameters. RMSE is the root of the mean squared error between the predicted and actual values [2]. We use it here because it gives us the value in the same scale as the unit being predicted [2]

To predict the price, four regression techniques have been used. They are:

a. *Stepwise Regression:* In this method, we fit the regression model using an automatic procedure. It conducts a series of tests (i.e., t-tests, f-tests) to find predictor variables that significantly influence the response variable [3]. For this study, we have used bidirectional elimination. It is a combination of forward as well as backward elimination. It uses the AIC value to determine which variables must be included or excluded. We used the "MASS" and "caret" library to implement the model. The evaluation metrics that we use here are RMSE and R-squared values.

b. *Ridge Regression*: This is a shrinkage method. Ridge shrinks the regression coefficients of the predictor variables. This results in the coefficients of the predictor variables with minor contribution to the response variable have their coefficient close to zero [4]. For this study, we employ a sequence of λ values to find the best lambda. For this study, we have made use of the "*glmnet*" library for implementation of the model. The metrics for

evaluation of model are RMSE and R-squared values. We use ten folds cross validation to give us an understanding of how well the model will fit new data.

c. *LASSO Regression*: Least Absolute Shrinkage and Selection Operator can shrink the coefficients to zero. Thus, it technically eliminates the predicator variables that do not have a significant contribution to the response variable [4]. This model excels over ridge because it eliminates variables that are not useful. Thus, it makes the model more interpretable and easier to understand. In this study, we expect LASSO to perform better since we are dealing with a large number of predictor variables. The model is implemented using the "*glmnet*" library with ten folds cross validation.

d. *Elastic Net:* This is a combination of both Ridge and LASSO. This shrinks the value of the coefficients of predictor variables. Simultaneously it also eliminates insignificant predictor variables out of the model. This is done to increase the accuracy. Here we determine the alpha parameter to identify which penalty to choose, on qualitative grounds [5]. We also use the lambda parameter to control the weights to the penalty of the loss function. [5]. This model is also implemented using the "*glmnet*" library with ten folds cross validation.

Results

We compute the RMSE and R-squared values for the four methods. In the end, we compare the scores to estimate which model is better. Since our response variable is continuous, the R-squared value gives us the model accuracy (i.e., It can explain the % of variation explained by the predictors)

a. *Stepwise Regression:* When we perform stepwise regression, the model is run in both directions. When the AIC threshold is reached, it stops iterating and gives us the selected predictor variables and coefficients. We get a R-squared value of **0.8528** as shown in figure 5. It chose 11 significant variables. These variables have high coefficient values and are significant factors in forming the price of the car.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          6016.398   8824.493   0.682  0.49711
normalized.losses      10.421      7.249   1.438  0.15399
wheel.base            247.360     77.125   3.207  0.00185 **
bore                -1860.207   1309.961  -1.420  0.15901
stroke              -4510.993    977.583  -4.614 1.29e-05 ***
compression.ratio    1040.073    389.344   2.671  0.00895 **
horsepower             58.289     13.868   4.203 6.14e-05 ***
city.mpg             -358.561     78.559  -4.564 1.56e-05 ***
drive.wheels_num      851.273    569.465   1.495  0.13841
body_style_num       -710.581    322.145  -2.206  0.02992 *
num.of.cylinders_num -1492.987    677.478  -2.204  0.03007 *
num.of.doors_num     -1290.405    512.473  -2.518  0.01355 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1774 on 91 degrees of freedom
Multiple R-squared:  0.8528,    Adjusted R-squared:  0.835
F-statistic: 47.92 on 11 and 91 DF,  p-value: < 2.2e-16
```

Figure 4: Results of stepwise regression

b. *Ridge regression:* After specifying a list of λ, we perform ridge regression. Referring to figure 5, we get a R-squared value of **0.6835.** We also obtain an RMSE of 3047.25.

```
                              s0
(Intercept)          -1.807910e+04
symboling            -2.968794e+01
normalized.losses     8.955106e+00
wheel.base            7.193081e+01
length                2.170972e+01
width                 2.098926e+02
height                5.349178e+01
curb.weight           1.122023e+00
engine.size           1.644853e+01
bore                  7.322365e+02
stroke               -2.486624e+03
compression.ratio     1.612138e+02
horsepower            1.831491e+01
peak.rpm              6.531777e-01
city.mpg             -1.065716e+02
highway.mpg          -9.227860e+01
drive.wheels_num      7.708662e+02
engine.type_num      -2.492608e+02
body_style_num       -1.947824e+02
num.of.cylinders_num -9.986474e+01
asipration_num        7.691638e+02
num.of.doors_num     -4.971076e+02
fuel.system_num       1.462603e+02
```

RMSE <dbl>	Rsquare <dbl>
3047.255	0.6835343

Figure 5: Results of ridge regression

c. *LASSO regression:* From the figure 6, we get a R-squared value of **.7188.** We also obtain an RMSE of 2994.162. This is a significant improvement over ridge.

```
23 x 1 sparse Matrix of class "dgCMatrix"
                                    s0
(Intercept)            2.367931e+03
symboling              .
normalized.losses      .
wheel.base             1.123796e+02
length                 .
width                  9.085415e+01
height                 .
curb.weight            1.274964e+00
engine.size            .
bore                   .
stroke                -2.918661e+03
compression.ratio      .
horsepower             3.483318e+01
peak.rpm               1.935667e-02
city.mpg              -2.827571e+02
highway.mpg            .
drive.wheels_num       7.738164e+02
engine.type_num        .
body_style_num         .
num.of.cylinders_num   .
asipration_num         .
num.of.doors_num       .
fuel.system_num        .
```

RMSE <dbl>	Rsquare <dbl>
2994.162	0.7188223

Figure 6: Results of LASSO regression

d. *Elastic regression:* From the figure 6, we get a R-squared value of **0.7188.** We also obtain an **RMSE** of **2994.162**. This is a significant improvement over ridge.

Figure 7: Results of LASSO regression

Conclusion

To conclude, we compare the results of all the four techniques. We select the one with the highest R-squared value. The stepwise regression performs the best with a R-squared value of **0.85.** It gives 11 predictor variables that are considered as influential factors that make the price of a car. The model can explain about 85% of the variation in the data, thus its accuracy being 85%. This model can be used by companies to accurately price their cars and make their product successful in the market.

References

[1] Schlimmer JC. [Internet]. UCI Machine Learning Repository: Automobile Data Set. [cited 2023Apr15]. Available from: https://archive.ics.uci.edu/ml/datasets/Automobile

[2] Allwright S. What is a good RMSE value? simply explained [Internet]. Stephen Allwright. Stephen Allwright; 2022 [cited 2023Apr15]. Available from: https://stephenallwright.com/good-rmse-value/

[3] Hayes A. Stepwise regression: Definition, uses, example, and limitations [Internet]. Investopedia. Investopedia; 2022 [cited 2023Apr15]. Available from: https://www.investopedia.com/terms/s/stepwiseregression.asp#:~:text=Stepwise%20regression%20is%20the%20step,statistical%20significance%20after%20each%20iteration.

[4] Kassambara, Chandrakumaran. Penalized regression essentials: Ridge, Lasso & Elastic Net [Internet]. STHDA. 2018 [cited 2023Apr15]. Available from: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net/

[5] Page 663, Hastie T, Friedman J, Tisbshirani R. The elements of Statistical Learning: Data Mining, Inference, and prediction. New York: Springer; 2017.