Diploma in Data Analytics
Matthew Kavanagh
Car Price Assignment

The price of a car, as set by the manufacturer, is based on the cost of production but can also be influenced by the desirability of features and quality of performance. Depending on the market the car is sold in, the manufacturer may charge more for a car model if it has multiple features which are in demand, thus increasing profits. Therefore different features will have different influences on the price of the car.

The following analysis is conducted on a sample data set of cars for sale across the United States. The aim of the analysis is to predict the price of a car based on the given independent variables using regression analysis.

1. About the Dataset

The CarPrice_Assignment dataset comes with a data dictionary. The sample contains data for 205 cars of a variety of different specifications. There are 11 categorical features and 14 numeric features for each car.

The categorical features are as follows; symboling, car name, fuel type, aspiration, door number, car body, drive wheel, engine location, engine type, cylinder number, fuel system.

The numeric features are as follows; wheel base, car length, car width, car height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm (revolutions per minute), city mpg (miles per gallon), highway mpg, price.

Price is the dependent variable for this analysis. It is numeric and continuous, therefore a regression model is required.

Analysis on the dataset is calculated to a 0.05 significance.

2. Exploratory Data Analysis

Python 3.7.4 in Jupyter Notebook was used to conduct this analysis. Pandas, numpy and matplotlib were the most frequently used libraries but many others were used. It was found that the dataset contained no missing values using isnull().

Categorical variables

Checking the data types revealed that ‘symboling’ was stored as an integer. This differs from the data dictionary which states it is a categorical variable. This error is easily fixed with astype(str) to convert the integers to strings. Symboling, according to the data dictionary, is an insurance risk rating, -3 being the safest and +3 being risky.

- Volvo cars have the best overall rating based on this feature, making Volvo the safest brand in the sample, while Buick and Toyota are second and third.
- Alfa-Romeo, Porsche and Saab perform the worst and are the riskiest brands.

‘Car names’ include the make and model of each car. This means almost every car model in the data is unique. To make the data more usable the model was removed leaving only the make, or brand.

- The most common brand in the data set is Toyota and the least common is Mercury.

‘Fuel type’ contains 20 diesel cars and 185 gas, or petrol, cars. Gas cars are more common.

‘Aspiration’ describes how a car intakes oxygen which is then combined with the fuel for combustion. 168 cars use standard aspiration. Standard aspiration is when the car relies on atmospheric pressure to force oxygen into the combustion chambers. 37 cars use turbo aspiration. A turbocharger uses exhaust fumes, a by-product of internal combustion, to pressurise atmospheric air and force more oxygen into the combustion chamber than a standard aspiration. The common use for this is to increase horsepower and is typically found in sports cars.

‘Door number’ may also be an identifier of sports cars, which usually have two doors. There are 115 four door cars and 90 two door cars.

‘Car body’ contains five types; sedan (96), hatchback (70), wagon (25), hardtop (8), convertible (6). Sedan is the most common while convertibles are the least common.

‘Drive wheel’ has just three possibilities; front wheel drive (120), rear wheel drive (76), four wheel drive (9). These are abbreviated to fwd, rwd and 4wd respectively.

- Sedans and hatchbacks typically have front wheel drive.
- Convertibles and hardtops are predominantly rear wheel drive demonstrating their sports car nature.
- Wagons come in all three drive wheels but front wheel drive is slightly more common than the others.

‘Engine Location’ revealed that only 3 cars in the dataset had rear engined cars. The other 203 cars were front engined cars. The sample size of rear engined cars is too small to provide any insight. Therefore the variable is dropped.

‘Engine type’ has the following values; ohc (148), ohcf (15), ohcv (13), dohc (12), l (12), rotor (4), dohcv (1).

‘Cylinder number’ contains the following; four (159), six (24), five (11), eight (5), two (4), three (1), twelve (1).

‘Fuel system’ has the following; mpfi (94), 2bbl (66), idi (20), 1bbl (11), spdi (9), 4bbl (3), mfi (1), spfi (1).

‘Engine type’, ‘Cylinder number’ and ‘Fuel system’ each contain categorical values that do not appear frequently in the data. They may not be very useful in the final model but they will be left in the dataset.

Numeric variables

Using .corr() it can be seen that ‘wheel base’, ‘car length’, ‘car width’ and ‘curb weight’ have a strong positive correlation. As a car gets longer it also becomes wider and heavier. There is also a strong negative correlation between curb weight and highway mpg. This is logical as the heavier a car is, the more fuel that is required to keep it moving so it has bad efficiency. It is important to not include all of these correlated variables in the final model to avoid multicollinearity.

It is important to note that price is positively skewed. Log 10 will be used to reduce the skewed nature of price for the final model. As multiple linear regression assumes normal distribution it is important that all data is not skewed.

A new column called average mpg (miles per gallon) was created to analyse the difference, if any, of the fuel efficiency between gas and diesel cars. The average mpg column was created by taking the mean average of city and highway mpg of each car. The higher the average mpg number the more efficient the car is i.e. the car can travel further than other cars on an equal volume of fuel.

- At a 0.05 significance diesel cars get 2.01 to 8.05 more mpg than gas cars.
- The gas fueled Honda Civic is the most fuel efficient car with 51.5 average mpg.

Horsepower is a good measure of the sports performance of a car. However, power to weight ratio is a better metric to use to measure sports performance as heavier cars generally have high horsepower but are not very sporty. By dividing the horsepower of a car by its weight the power to weight ratio was found. Curb weight was converted from lbs to US ton.

- The Porsche Cayenne, with a price of $ 31,400.5, is the best sports car in the dataset with a power to weight ratio of 171.1 bhp per US ton.
- The Toyota Carina, with a price of $ 8,778.0, is the worst performing car in the dataset with a power to weight ratio of 39.9 bhp per US ton.

A one-way Anova test was used to find the difference, if any, between the horsepower of cars with different drive wheels. The different categories for drive wheels are fwd (front wheel drive), rwd (rear wheel drive) and 4wd (four wheel drive). The Anova test indicated that there was at least one pair of categories which had significantly different means. A Tukey post-hoc test was conducted to find which drive wheels have significantly different means.

- At 0.05 significance 4wd and fwd had equal means.
- Rwd had 38.4 bhp more than 4wd.
- Rwd had 47.7 bhp more than fwd.

3. Model Development

Multiple Linear Regression (MLR)

Due to the large number of variables in the dataset a MLR model was made using backwards stepwise regression using only the numeric variables. The categorical variables are then added and a further backwards stepwise regression is conducted to reduce the number of independent variables in the model and thus increase the adjusted R-squared.

It is important not to have too many independent variables in the model as there is already high multicollinearity in the data set.

Log 10 function was performed on the price column to reduce the positive skewed nature of the data.

Car width, curb weight, horsepower, engine size and city mpg were the variables in the first model against log 10 price. Backwards stepwise regression was performed until all p-values were less than the 0.05 significance level. Car width was also dropped due to the high multicollinearity with curb weight.

Plotting residuals of the remaining variables, curb weight and horsepower, shows the positively skewed nature of horsepower. The log 10 function is applied to horsepower. Using a Cook's distance plot it can be seen that three Porsches, Panamera, Cayenne and Boxter, are potential outliers but do not have a significant effect.

Plotting residuals vs fitted on the final model shows a reasonable normal distribution. Normal Q-Q plot has a good fit to the straight line demonstrating normal distribution.

Using the get_dummies() function the categorical variables were added to the model. Backwards stepwise regression was performed until all p-values of independent variables were less than 0.05.

- The final model has an adjusted R-squared of 0.924 and an AIC of -557.8.
- A one unit increase in curb weight (lbs) increases the log10 price by 0.0003
- BMWs increase the log10 price by 0.1571.
- Gas cars increase the log10 price by 0.9197.
- Ohcf engines increase the log10 price by 0.2992.
- Idi fuel systems increase the log10 price by 0.9774.
- Mpfi fuel systems increase the log10 price by 0.0406.
- One unit increase in log10 horsepower increases the log10 price by 0.2528.

- Mitsubishis reduce the log10 price by 0.0614.
- Peugeots reduce the log10 price by 0.0908.
- Subarus reduce the log10 price by 0.3451.
- Toyotas reduce the log10 price by 0.0529.
- Wagons reduce the log10 price by 0.0501.

Gas cars and idi fuel systems cause the greatest increase in log10 price.
The brand Subaru causes the largest decrease in log10 price.

Random Forest Regression

A random forest regression model was performed on the same dataset. By plotting permutation importance the most significant variables were identified. Curb weight, engine size, highway mpg, horsepower, car width, city mpg and wheel base were included in the final random forest regression. The 'n_estimators' were set to 100 for this model.

Principal Component Analysis

A final model was constructed using principal component analysis with the aim to more effectively reduce the amount of independent variables while retaining as much accuracy as possible. Using this method the number of component variables was reduced from 67 to 22 while retaining 79.8 % of the information.

This method is well suited to this dataset as there are many independent variables and is much faster than the backwards stepwise regression performed in the multiple linear regression. However, some of the information is lost through this technique and the accuracy of the model may suffer.

4. Model Accuracy Comparison

By maintaining a constant test size of 0.33 and setting the random state to 30 for each of the three predictive models their accuracy on the same train and test data can be evaluated. The R-squared score of the model when predicting the price of the cars in the test group is used to evaluate the models. Also the time for a model to make these predictions is also measured.

- Multiple Linear Regression has an R-squared of 0.91 in 0.0086 seconds.
- Random Forest Regression has an R-squared of 0.92 in 0.3179 seconds.
- Principal Component Analysis has an R-squared of 0.89 in 0.0026 seconds.

Random forest regression is the most accurate but the slowest. This may be problematic to apply to a larger data set. This dataset is relatively small at 205 when compared to the circa. 3 million cars sold in the USA per year.

Principal component analysis is the least accurate but is 100 times faster than random forest regression.

Multiple linear regression is in between the other two models in terms of accuracy and time taken to make predictions.

5. Conclusion

The price of cars in this dataset was successfully modelled using three different techniques, multiple linear regression, random forest regression and principal component analysis. Each method has benefits and limitations.

Multiple linear regression requires the most user input to evaluate the best variables to use in the model, when compared to random forest and principal component analysis which are more straightforward, but gives the best overall model in terms of accuracy and runtime. The multiple linear regression model is therefore the favoured model out of the three used. Also this method provides better insight into the data and indicates which are the most influential features on a car. Gas and idi fuel systems were found to increase the price of a car the most and if the car was a Subaru this reduced the price of the car more than any other feature.