Multivariate analysis of factors affecting the price of a car

Student ID: 1619893

Department of Computer Science, University of Warwick

Published 12.12.2019

Abstract

The aim of this research project is to perform a detailed multivariate analysis on the different components and factors of a car, giving an insight to how they affect the price of a car and analysing if there are any trends between the variables, using various statistical methods on an automobile dataset. Understanding this dataset is not only important for future research application, but very advantageous for an individual when deciding on a most appropriate car to buy for themselves.

1. Introduction

In a world where 77% of households in the UK own a car [1], 75% of all adults aged over 17 own a driving license, and approximately 500000 young drivers between the age of 17-24 pass their driving test per year [2], buying a car has almost become a necessity. It is very common for people without decent background knowledge of cars to purchase a car based on its aesthetics, brand reputation and mileage and overlook the other important factors such as engine size and curb weight which may potentially affect the price of a car. As a result, money may be unnecessarily spent on components which were not required for the consumer and they could have saved money by buying a more suitable car for themselves.

As a result, the goal of this research is to address these gaps in knowledge and conduct a statistical analysis of what factors may affect the price of a car. This will be performed using various statistical software and packages, which will be explained in a later section.

2. Background

Ever since the first gasoline-powered car was produced in 1885 by Karl Benz [3], the number of cars produced per year has drastically increased and by 1927, more than 15 million cars have been produced by Ford alone [4].

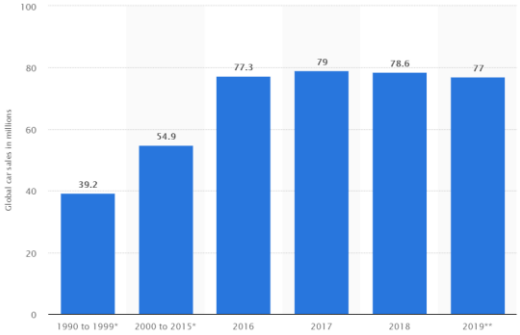


Figure 1: Global car sales (in millions) per year [5]

According to a study conducted by Scotiabank [5], the number of automobiles being purchased globally per year has doubled since 1990 from 39.2 million to 77 million.

With so many years of history, automobiles have become very customisable. 62 brands of cars are controlled by 14 major corporations and each car has its own unique combination of components. Some brands such as Mercedes, BMW and Porsche have become reputable and known to sell more luxurious expensive vehicles whilst brands such as Honda, Nissan and Volkswagen are well known to produce more affordable cars for everyday use.

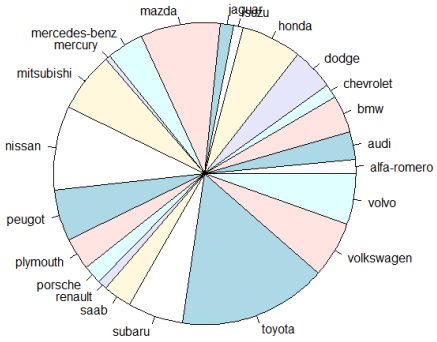


Figure 2: Piechart showing proportion of car brands in market [6]

It is also common for drivers to modify their car parts in order to enhance the performance, safety and style of the car. In the UK, one of the most common modifications is brake modifications. By increasing the size of the brake disks, it allows the car to stop more smoothly and quickly.

Another common modification applied to many cars is engine upgrades, which is performed in several ways. For example, engine reboring will generate more power and increase the RPM of your car. Grinding and polishing the engine will provide more mileage to the car, since more air can travel into the engines which increases the efficiency of the burning of fuel.

Wheels and tyres can also be changed to improve the acceleration and stopping distance of a car whilst also potentially improving the aesthetics of the car. [7]

Fuel type is also another important factor when choosing cars. A gallon of diesel fuel has up to 30% more energy than a gallon of gas, which means in theory, diesel fuel should have better mileage, although a diesel-powered car can cost as much as $700 more than a fuel gasoline-powered car of the same model [8].

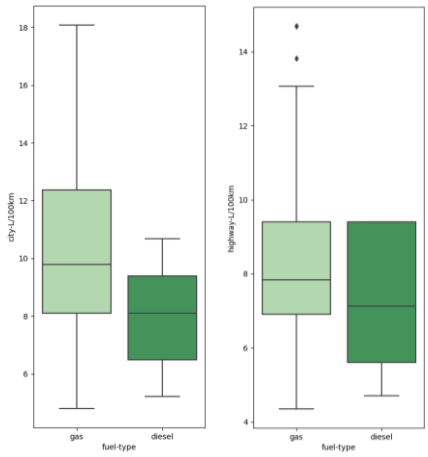


Figure 3: Boxplot of fuel type aginst litres of fuel type /100km [6]

3. Dataset

In this study, an automobile dataset will be analysed, taken from the UC Irvine Machine Learning Repository [6], which hosts 488 sets of open-source data. This particular dataset contains 26 different attributes of 205 instances of automobiles, which were already partially processed into a format that is almost ready to be implemented into statistical software. For example, the loss of value of a vehicle per year is normalised for all autos within a particular size classification (two-door small, station wagons, sports/speciality etc) and is referred to 'normalized-losses'. Although this data set from 1985, all of the components of each attribute is still commonly implemented in modern cars and the main aim of the research is to understand the trends within these attributes, therefore the data set is still sufficient.

The dataset provided consists of two files. The name file is in a .name format which describes what each column in the .data file (explained later) represents. It has a unique set of outcomes for all nominal attributes as well as the range of values for all numerical attributes. It also shows how many missing data points there are for each attribute.

The other file in a .data format which contains all the recorded data delimited by a comma.

4. Hypothesis

From this dataset, the aim is to discover which attributes affect *price* most significantly. From the background information gathered earlier in **section 2**, the hypothesis formulated is that the *fuel consumption*, *power*, *size*, *engine* and *make (brand)* are amongst the most influential factors towards the price of a car.

5. Statistical software

As the provided data is in a .name file and a .data file, it can easily be converted into a comma space delimited (CSV) file or an Attribute-Relation File Format (ARFF). The following statistical software in **section 5.1 – 5.3** are going to be used to analyse the CSV file, and the Weka will be used to analyse an ARFF file (See **section 5.4**):

5.1 R

R is a very useful language for statistical computing and graphics. Using RStudio, which provides open-source tools for R, many useful simple plots can be made, as well as some advanced statistical analysis can be performed.

5.2 Microsoft Excel

Microsoft Excel is one of the most well known and readily available data analysis tools. Excel has many simple formulas which becomes very useful when preprocessing the data set, and as the program was created in 1985, it has a

very long history, hence there are many resources and guides available online.

5.3 Python

Python is an advanced and general-purpose programming language, created in 1991. Once again, having had such a long history, there are many resources and guides available online. There is a wide range of tools within python that is useful in the field of statistical analysis. For example, matplotlib is a commonly used 2D plotting library, and seaborn is a Python data visualisation library based on matplotlib. Both of these libraries are incredibly useful as they can generate attractive and informative statistical graphics.

5.4 Weka

Released in December 22th 2017, Weka is one of the newest statistical software available on the internet. It is a GUI tool that allows the user to simply load data sets and perform data preprocessing, visualisation, regression, classification, clustering and feature selection without having to worry about the coding. All algorithms are readily available and hence, little programming knowledge is required.

6. Data Cleaning

6.1 Dealing with missing values

Given that 7 columns are containing missing data from the .name file, to easily edit the entries for each column, the .data file was opened using Microsoft Excel. There are three different methods used when dealing with missing data.

In the first case, the attribute *normalized-losses* had 41 missing data, whilst *stroke* and *bore* had 4 missing data. *Horsepower* and *peak-rpm* had 2 missing data. Since all of these attributes are numerical, all of the missing values of each attribute were replaced with their respective mean. This was simply done on excel using the formula *=AVERAGE()* for each column and replacing the missing entries with this value.

In the second case, the attribute *num-of-doors* are categorical numbers – two or four. Hence, replacing with mean would be inappropriate because a value between 2 and 4 would be returned. Therefore, the mode was used. To find the mode of a column of words, the following formula was used:

=INDEX(F1:F201,MODE(MATCH(F1:F201,F1:F201,0)))

In the third case, the row *price* had four missing entries. As stated in the hypothesis, *price* is the response variable and hence, instances without *price* data are meaningless as it cannot be used for prediction. Therefore, the corresponding four rows were simply removed in Excel.

6.2 Standardising mileage

Notice that the fuel efficiency for *highway* and *city* are given in miles per gallon. Since we are investigating fuel consumption, to make this metric easier to interpret, “miles per gallon” was transformed to “litres per 100km”. This was simply done in Excel using the following formula: L/100km = 235/mpg.

6.3 Binning

The attribute *price* was assigned into bins according to its quartiles. This was done using pandas dataframe in Python. The interquartile range of the *price* column was calculated, and any values that fall into the interquartile range ($7775-$16500) was labelled as ‘average’ and any values falling below this range was labelled as ‘cheap’. Values above the interquartile range up to the 0.9 quartile was labelled as ‘expensive’ and the remaining quartile (top 10% of data points) was labelled as ‘very expensive’ ($22470-$45400). The information was added to a new column called *price*-category. This allows Weka to perform classification and clustering later. The following line of Python code was used to perform the above

```
df['price-category'] = pd.qcut(df['price'],
q=[0, .25, .75, .9, 1], labels=['Cheap',
'Average', 'Expensive', 'Very expensive'],
precision=0)
```

6.4 Converting the DATA file to ARFF

Now that there is a processed .data file, the remaining step is to change the file to an arff format so that it can be opened in Weka. To do this, the .data file was opened with Notepad, and at the top of the file, @relation was added to give the dataset a name, followed by 26 rows of @attribute which contains by the unique set of outcomes for each attribute which can be found from the .names file, and finally, @data, above the processed entries of the dataset.

Two separate files of arff was saved, one file for where the *price* attribute was processed into bins so that group classification can be performed, and one file for where the *price* was unprocessed, such that the values were still numerical so that linear regression can be applied.

7. Exploratory Data Analysis

7.1 Initial Diagnostics and general analysis

This section provides an exploratory data analysis to the automobile data, which will be done using graphs from the listed statistical software and summary function in R.

Using Python, a heatmap was produced which shows the correlation of all attributes with each other.

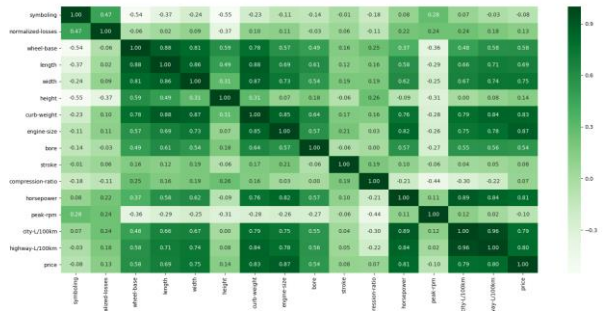


Figure 4: Heatmap of automobile data

Based on the heatmap generated by Python, the attributes *length*, *width*, *curb-weight*, *engine-size*, *horsepower*, *city-L/100km*, *highway-L/100km* have the highest correlation with *price*, which is good initial evidence for the hypothesis.

All of the attributes was fitted into a linear model against price in R and a summary of the linear model was inspected to see which attributes were the most statistically significant. From the summary, the attributes *wheel-base*, *length*, *width*, *height*, *curb-weight*, *bore*, *peak-RPM*, city-L/100km and some subattributes of *make*, *aspiration*, *style*, *engine-type* and *engine-size* had p-values less than 0.05 and were statistically significant according to the model.

7.2 Exploring individual attributes

In the following section, attributes that were related to the hypothesis and other interesting trends were analysed, which will provide a better understanding of the database.

As there are 22 makes of automobiles in the data set and the linear model indicated *make* is a significant attribute, the best method to visualise how *make* affected the price was via multiple box plots on a graph.

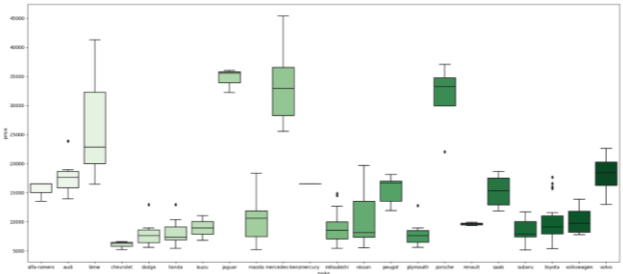


Figure 5: Boxplots of make against price

It can be inferred from the box plot that makes *BMW*, *Jaguar*, *Mercedes*, *Porsche* strongly suggests that the automobile is very expensive. *Audi* and *Volvo* makes mostly fall into the expensive range and is a fairly strong indicator. *Chevrolet* and *Plymouth* mostly fall below the interquartile range and hence are good indicators that the automobile is very cheap.

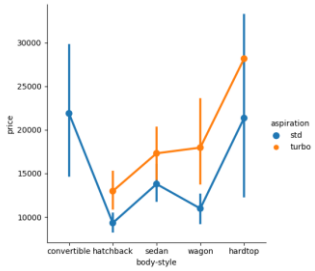


Figure 6: Catplot of body style against price for each aspiration

From the catplot generated above, there is an obvious trend, that turbo costs more than standard aspiration. Hardtop appears to be the most expensive body style and no convertible car has turbo aspiration.

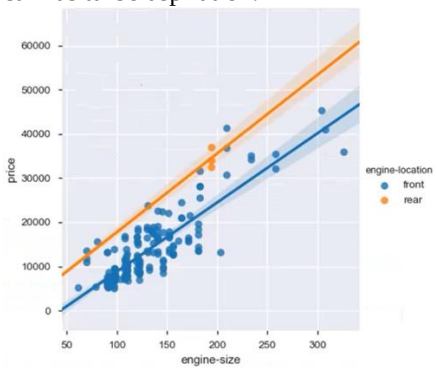
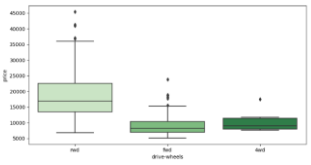


Figure 7: Scatter plot of engine size against price depending on the location of engine

It can be deduced from the plot that engines located at the rear of the automobile is more expensive than the front, and there is a clear linear trend that price increases as the engine size increases. However, there were only three instances of automobiles that had rear engines installed, therefore more data would be required to make a reliable conclusion for rear engines.



	drive-wheels	price
0	4wd	10241.000000
1	fwd	9244.779661
2	rwd	19757.613333

Figure 8: Boxplots and table of averages for drive wheels

Whilst *4wd* and *fwd* are similar, *rwd* is more targetted at luxurious cars because it corresponds to automobiles with prices mostly in the *expensive* and *very expensive* category.

8. Developing and evaluating the model

8.1 Simple Linear Regression

In this section, a simple linear regression model will be fitted and analysed for the *engine-size*, the attribute that has the highest correlation with price. A more in-depth analysis will be provided for the other models (i.e. multiple linear regression) in the later sections because the models will be more complex and should have a better fit.

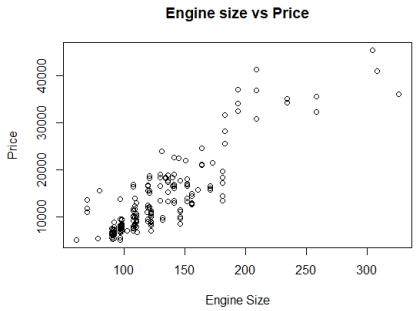


Figure 9: Scatter plot of engine size with price

The scatter plot generated by R suggests that there is a linear relationship between engine size and price. Hence, a linear regression model of engine size against price was fitted in R. The resulting regression model was *Price* = -7963.339**engine-size* + 166.860
Using the summary function, the value of R-Squared was 0.761 meaning 76.1% of the variance was explained by this model, which was decent, considering that only 1 out of 25 variables was used to predict the price.

8.2 Multiple Linear Regression (Attributes picked manually)

A more complex multiple linear regression model was fitted using all of the variables that were either in the hypothesis or had a high correlation with the price and appeared statistically significant in the summary function from **section 7.1**. This included the attributes *city-L/100km, highway-L/100km, length, width, height, curb-weight, engine-size, horsepower*. The resulting regresion model was: *Price* = 1135.181**city-L/100km* – *783.543**highway-L/100km* – *111.22**length* + *799.574width* + *320.413height* + *2.131**curb-weight* + *95.315**engine-size* + *22.011**horsepower* – *62130.144*

Using the summary function, the value of R-Squared was 0.828, meaning 82.8% of the variance was explained by this model, which was an improvement of the simple linear model fitted above.

8.3 Dealing with multicollinearity

Although the model earlier appeared to be a good fit, there lies a fundamental error when manually picking attributes – there may be cross-correlation between the predictor attributes. The following section highlights two attempts to create a model with no collinearity, under the simplification that only numerical attributes (14 out of 25 predictor variables) were considered.

8.3.1 Multiple linear regression (Attributes picked mathematically)

8.3.1.1 Initial model

Rather than manually selecting the attributes, Weka has a built-in function under linear regression which removes collinearity by comparing an attribute's standardised regression coefficient such that if it was bigger than 1.5 then it was removed. As a result, the following attributes were picked out of all numerical attributes and a regression model was produced:
Price = 1520.9113**city-L/100km* – 687.3493**highway-L/100km* – 80.5334**length* + 626.5523*width* + 252.6596**height* – 2505.3433**stroke* + 396.294**compression-ratio* + 2.0404**peak-rpm* + 122.3626**engine-size* – 61212.3744

The multiple linear regression model above gave an R-Squared value of 0.8577 meaning 85.7% of the variance was explained by this model.

8.3.1.2 Backwards Stepwise Regression of initial model

Although the model decently explained the variance, the model produced was too complex as there were 9 contributing attributes. An investigation in R was performed to determine whether any of these factors can be removed via the summary() function again, to see if any attributes had a p-value above 0.05. *Highway-L/100km* had a p-value of 0.15 which was statistically insignificant and hence was removed. This procedure was repeated, removing *length*, and then followed by *height*. The remaining attributes all had p-values below 0.05 and formed the following regression model:

Price = 980.0291**city-L/100km* + 2.0911**peak-rpm* + 398.1855**compression-ratio* – 2631.8014**stroke* + 419.1689**width* + 116.0111**engine-size* – 45053.1797

The value of R-squared is 0.8522 meaning 85.2% of the variance is explained by this simplified model. With 3 attributes removed, the variance explained only dropped by 0.5%, which shows that this simplified model is the most suitable model and the final model for multiple linear regression.

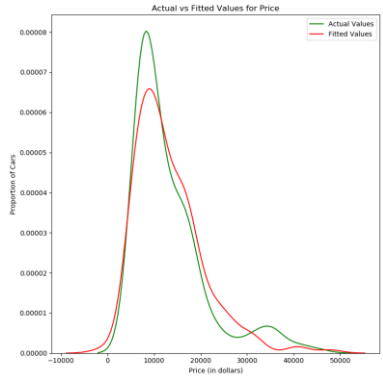


Figure 10: Distribution plot for actual price vs fitted price

The distribution of the fitted price against the actual price is very similar, although the peak of the graph is not as high, the model was simplified from 14 attributes down to 6 attributes whilst still managing to explain the majority of the variance and eliminate collinearity.

8.3.2 Principal Component Analysis

An alternative method to eliminate collinearity is via the traditional method of principal component analysis, which reduces the dimension of the dataset consisting of many variables correlated with each other, whilst retaining the variation as much as possible. As some of the variables have been standardised, the correlation matrix was used rather than the covariance matrix. The eigenvalues of the first 8 principal components are 6.809, 2.43, 1.250, 0.864 and 0.812, 0.497, 0.438 and 0.288. The respective proportion of variances is 0.486, 0.178, 0.089, 0.062, 0.058, 0.036, 0.031 and 0.021. The remaining eigenvectors explain less than 4% of the total variance and can be instantly discarded.

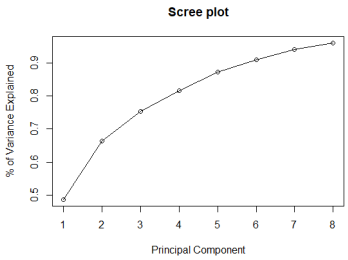


Figure 11: Scree plot of Principal Components

A scree plot was produced and the first 5 PC's explain 87.3% of the total variance and hence the remaining were discarded. The loadings of each principal component are presented in the following table.

Eigenvectors

V1	V2	V3	V4	V5	V6	V7	V8	
-0.0397	-0.3276	-0.3959	0.057	0.7428	0.3321	0.1002	-0.2307	normalized-losses
-0.3058	0.2715	0.0251	-0.2484	0.1703	0.1537	0.0097	0.3113	wheel-base
-0.3462	0.1544	0.0473	-0.1553	0.1369	0.0626	0.0773	0.1348	length
-0.3435	0.0923	-0.0958	-0.0994	0.1159	-0.0798	-0.0603	0.4869	width
-0.1137	0.4355	0.338	-0.4254	0.1261	0.115	0.0382	-0.6154	height
-0.3708	0.0457	-0.0591	0.025	0.0332	-0.0826	-0.0675	-0.0759	curb-weight
-0.3351	-0.0683	-0.1344	0.2013	-0.1844	-0.0549	-0.2644	-0.2845	engine-size
-0.2699	0.0211	0.1984	0.4063	-0.1044	0.1111	0.8179	-0.0065	bore
-0.0576	0.0588	-0.6916	-0.3742	-0.4854	0.249	0.2308	-0.0824	stroke
-0.017	0.4404	-0.3918	0.1954	0.2304	-0.6789	0.0856	-0.1074	compression-ratio
-0.3106	-0.2899	-0.0127	0.089	-0.1611	-0.2262	-0.029	-0.3252	horsepower
0.0886	-0.4301	0.0939	-0.5674	0.0898	-0.4933	0.3562	0.0353	peak-rpm
-0.3264	-0.2833	0.0892	-0.0517	-0.0379	-0.0463	-0.1007	-0.0354	city-L/100km
-0.342	-0.2076	0.0854	-0.0561	-0.0324	-0.0281	-0.1923	0.0512	highway-L/100km

Table 1: Loadings of each component

Here, it can be deduced that the first principal component is a weighted average of the car size, *curb-weight*, *engine-size*, *bore*, *horsepower* and fuel consumption. This can simply be interpreted as how powerful a car is overall.

The second principal component is a contrast between *normalized-losses*, *horsepower*, *peak-RPM* and *fuel consumption* against *car size*, *wheel-base* and *compression-ratio*. This can simply be interpreted as the efficiency of a car.

The third principal component is simply a weighted average of *normalized-losses*, *height*, *stroke* and *compression ratio*.

The fourth principal component is a contrast between *height*, *wheel-base*, *stroke*, *peak-RPM* against *engine-size*, *bore* and *compression-ratio.*

Finally, the fifth principal component is simply a contrast of mainly *normalized-losses* (with slight weighting of *compression ratio*) against *stroke* and *engine size.* This can be simply interpreted as how good the car keeps its value over time.

8.4 Classification with price categories

In the earlier subsections, linear regression was mainly used to build a model to predict a price given its numerical attributes were known. Although the model produced was easy to interpret, 9 non-numerical attributes were ignored which may potentially have an impact on its reliability. In this section, rather than using the *price* attribute as the response variable, the *price-category* attribute (See **Section 6.3**) was used, which will allow other classification methods that include all of the attributes to be used.

8.4.1 Naïve Bayes

Using the Naïve Bayes classifier, an accuracy of 81.6% was achieved, correctly classifying instances into the classes *cheap*, *average*, *expensive*, *very-expensive*, with a precision of 0.804, 0.893, 0.614, 1.0 respectively, which are all very high with the exception of the *expensive* class.

8.4.2 Decision Tree

Decision trees are very easy to interpret because no computation power is required to classify new data. A person can simply trace down the tree from the root node to predict its class label simply following the values of the attributes of the instance. The predicted class label is found at the leaf node. A J48 Decision Tree was produced for the automobile data and obtained an accuracy of 87.6%, correctly classifying instances into the categories *cheap*, *average*, *expensive* and *very-expensive* with a precision of 0.904, 0.933, 0.667 and 1 respectively, which is an improvement over the Naïve Bayes model. The decision tree is not overfitted, because it is relatively short and easy to interpret.

From the tree, it can be deduced that for an automobile with *horsepower* greater than 82, any *engine-size* above 181 will automatically indicate that the automobile is very expensive. Any automobile with a *width* greater than 66.6, or a *wheel-base* smaller than 66.6 and turbo *aspiration* will be classed as expensive. The same goes for *4wd* drive wheels. The average and cheap classes are more randomly distributed but can be simply traced down the tree and deduced.

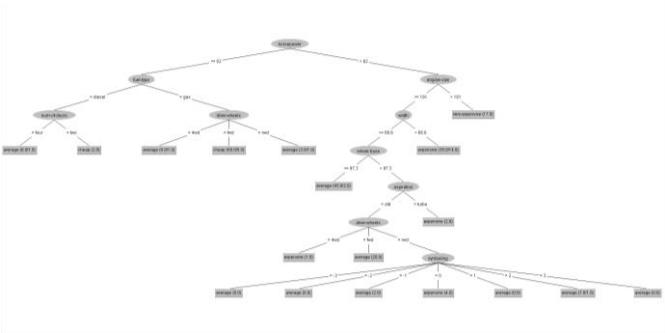


Figure 12: J48 Decision Tree

8.4.3 Support Vector Machine Classifier

Using the Support Vector Machine (SVM) Classifier with poly kernel, an accuracy of 90.05% was achieved correctly classifying instances into the categories *cheap, average, expensive* and *very-expensive* with a precision of 0.922, 0.902, 0.800, 1.000 respectively. This achieved the best accuracy and precision out of all three classifiers and is the only classifier that achieved a decent precision for the *expensive* class.

9. Conclusion

The goal of this research was to find whether fuel consumption, power, size, engine and make were amongst the most influential factors affecting the price of a car. In other words, this refers to the attributes *highway-L/100km, city-L/100km, horsepower, height, width, length, engine-size, engine type* and *make.*

Based on the analysis results, *height, length* and *engine-type* did not affect the price as much as expected because they were rarely included in any of the models created. Consumers who are looking to buy a new car should mainly consider the fuel consumption (*highway-L/100km, city-L/100km), horsepower, width, engine-size* and *make* of a car as they agreed with the hypothesis, as many of these attributes consistently appeared across all of the models, often carrying a large weight. They also have a high correlation with price.

10. Extensions

Future work could include linear regression and principal component analysis of all attributes rather than just numerical attributes. This can be done by an indicator variable for each category of an attribute, although this would be time-consuming, even higher accuracies can be reached. Factor analysis can be performed on the dataset as an alternative method for dimension reduction and to detect latent variables.

Another improvement that could be implemented to the analysis is to find a more recent data set of the same format because the value of some car parts may have changed and new car makes may have emerged into the market since this data was collected.

Whilst the factors influencing the car price at the day of the purchase were determined, the depreciation of the car was not taken into account.

According to AA, assuming the driver has driven 10000 miles per year, the average new car only retains 40% of its original purchase price after three years, meaning a vehicle loses 20% of its value per year on average [9]. A very extreme example is the *Fiat Doblo XL Combi 1.6 Multijet 120 SX*, which had an initial price of £26183 in 2016, but rapidly dropped to a price of £6825 in 2019 [10].

This can easily be analysed in the automobiles dataset because the standardised annual loss of value of the automobiles is stored as the attribute *normalized-losses*. Hence, a similar analysis can be conducted, using *normalized-losses* as the response variable instead of *price*.

References

[1] David Leibling, *Car Ownership in Great Britain*, 10 2018
[2] Department for Transport, *Facts on Young Drivers*, 04 2014
[3] "DRP patent No. 37435" (PDF). Archived from the original (PDF) on 4 February 2012.
[4] "Model T Facts" (Press release). US: Forid. Archived from the original on September 28, 2013. Retrieved April 23, 2013.
[5] Scotiabank, *Global Autoreport*, 06 2019
[6] Jeffrey C. Schlimmer, *Automobile Data Set*, May 19, 1987
[7] Insurance Revolution, *Most Popular Car Modifications*, Jan 10 2017
[8] Erik Bjornstad, *Diesel vs. Gasoline: Which Engine is a Better Fit for You?*, 25 June 2014
[9] AA, *Find out how quickly new cars lose money*, March 23, 2012
[10] What Car? Team, *The 10 fastest depreciating cars*, April 5, 2019