SIGNIFICANCE OF FACTORS IN PRICE OF AUTOMOBILES IN THE USA SALES DATA

Davaajav Jamsran
School of Computer, Engineering, and Mathematics
Western Sydney University, Australia

Abstract - The automotive market is an ever-changing dynamic environment, especially, the market is experiencing an emerging rise of new technologies, sustainability and climate policies, and changes in consumer preferences around ownership.
A PwC study from 2022 predicts revolutionary changes in the automotive sector by 2030, particularly in developed regions like Europe and the USA. This transformation is expected to be influenced by key performance indicators shaping competition in vehicle sales. These indicators include (1) users' mobility behaviors, incorporating social personas that endorse environmentally friendly technologies with low CO2 emissions, (2) vehicle-related factors such as mileage and frequency of usage, (3) aspects of car inventory and replacement cycles, and (4) the broader implications for manufacturers, suppliers, service providers, and their corresponding business models. This research work aims to explore and determine the significant technical factors that influence the pricing of automobiles by leveraging advanced data analytics and machine learning techniques (Multiple Linear Regression and Classification Analysis through Decision Tree, Principal Component Analysis), this research seeks to uncover the underlying patterns and relationships within the data to shed light on the primary factors affecting automobile prices. By identifying the significant influences, the research also seeks whether insights from the historical data can assist manufacturers in making informed pricing decisions and enable consumers to make more educated choices when purchasing vehicles, contributing to a more transparent and competitive automotive market.

***Index Terms*—Classification, Regression, Cross Validation, Principal Component Analysis**

I. INTRODUCTION

THE AUTOMOBILE market in the USA is a dynamic and multifaceted landscape that reflects the diverse preferences and behaviors of American consumers. Over the years, the U.S. has consistently been one of the largest automotive markets globally, with millions of vehicles sold annually. Automobile sales statistics in the USA provide valuable data that reflects the ever-changing dynamics of the automotive industry and the evolving needs and preferences of American consumers. The determination of automobile prices is a complex process influenced by various factors. It begins with the manufacturing cost, which includes expenses related to materials, labor, research and development, and overhead. Beside base cost with manufacturers' a margin for profit, Market dynamics play a pivotal role, as supply and demand fluctuations, along with the consumer preferences, impact pricing decisions. Consumer preferences for specific features, performance, and technology contribute to price variations among different models.

This research work aims to explore and determine the significant factors that influence the pricing of automobiles in the USA, utilizing sales data. As the automobile industry continues to be a critical component of the nation's economy and consumer lifestyle, understanding the key drivers behind price fluctuations is of paramount importance for both manufacturers and consumers.
Leveraging advanced data analytics and machine learning techniques, this research seeks to uncover the underlying patterns and relationships within the data to shed light on the primary factors affecting automobile prices.

The purpose of identifying the significant influences, the research also seeks whether valuable insights from the historical data can assist manufacturers in making informed pricing decisions and enable consumers to make more educated choices when purchasing vehicles, contributing to a more transparent and competitive automotive market.

In this study, an automobile data set is facilitated by IBM machine learning cloud, which is composed of data related to various numerical and categorical variables such as make, engine size, wheel, length, and width etc. Two models were created using two different methods of supervised learning, which were Multiple Linear Regression and Classification Tree, for an accurate prediction of price. These models were used to determine whether to accept or reject the hypothesis using the significance level of 5%.

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

The null hypothesis (H0) states that the predictor variables do not have any significant relationship with the target variable. The alternate hypothesis (H1) states that one or more predictor variables have a contributing relationship with the target variable.

A technique of unsupervised learning was applied to visualize patterns existing among the explanatory variables. Lastly, the models were compared based on their results and accuracy.

II. DATA PROCESSING AND DATA EXPLORATION

A. *Data Description*

The dataset on automobiles comprised 201 observations, each entry of data to a vehicle with associated attributes. Among the 29 predictor variables, 16 were numeric, including features like height, curb-weight, stroke, and peak-rpm. The remaining variables were categorical, involving aspects such as symbolling, make, aspiration, engine location, body-style, and number of doors, as detailed in fig. 2.1.

```
> str(data)
'data.frame':   201 obs. of  29 variables:
 $ symboling        : int  3 3 1 2 2 2 1 1 1 2 ...
 $ normalized.losses: int  122 122 122 164 164 122 158 122 158 192 ...
 $ make             : chr  "alfa-romero" "alfa-romero" "alfa-romero" "audi" ...
 $ aspiration       : chr  "std" "std" "std" "std" ...
 $ num.of.doors     : chr  "two" "two" "two" "four" ...
 $ body.style       : chr  "convertible" "convertible" "hatchback" "sedan" ...
 $ drive.wheels     : chr  "rwd" "rwd" "rwd" "fwd" ...
 $ engine.location  : chr  "front" "front" "front" "front" ...
 $ wheel.base       : num  88.6 88.6 94.5 99.8 99.4 ...
 $ length           : num  0.811 0.811 0.823 0.849 0.849 ...
 $ width            : num  0.89 0.89 0.91 0.919 0.922 ...
 $ height           : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 54.3 ...
 $ curb.weight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 2395 ...
 $ engine.type      : chr  "dohc" "dohc" "ohcv" "ohc" ...
 $ num.of.cylinders : chr  "four" "four" "six" "four" ...
 $ engine.size      : int  130 130 152 109 136 136 136 136 131 108 ...
 $ fuel.system      : chr  "mpfi" "mpfi" "mpfi" "mpfi" ...
 $ bore             : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.5 ...
 $ stroke           : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 2.8 ...
 $ compression.ratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 8.8 ...
 $ horsepower       : num  111 111 154 102 115 110 110 110 140 101 ...
 $ peak.rpm         : num  5000 5000 5000 5500 5500 5500 5500 5500 5500 5800 ...
 $ city.mpg         : int  21 21 19 24 18 19 19 19 17 23 ...
 $ highway.mpg      : int  27 27 26 30 22 25 25 25 20 29 ...
 $ price            : num  13495 16500 16500 13950 17450 ...
 $ city.L.100km     : num  11.19 11.19 12.37 9.79 13.06 ...
 $ horsepower.binned: chr  "Medium" "Medium" "Medium" "Medium" ...
 $ diesel           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ gas              : int  1 1 1 1 1 1 1 1 1 1 ...
```

Figure 2.1: Description of the data set

B. *Data Pre-Processing and Data Exploration for Continues numeric variables.*

To align with the study's emphasis on continuous numerical variables, the initial step involved filtering the dataset to exclusively include entries with numerical variables. This process was undertaken to streamline the dataset, ensuring that only numerical variables relevant to linear relationships with the target variable were considered (Fig.2.2).

Fig 2. Numeric Variable List

Name	Description	Measure
normalized.losses	relative average loss payment per insured vehicle year	$
wheel.base	distance between the front and rear axles of a vehicle	inch
length	length of a vehicle	feet
width	width of a vehicle	feet
height	height of a vehicle	mm
curb.weight	weight of the vehicle	lb
engine.size	measure of the cumulative space inside a motor's cylinders	cubic inch
bore	diameter of the circular opening in engine	diameter
stroke	depth of the hole in engine	inch
compession ratio	ratio between the volume of the cylinder and combustion chamber	cubic inch
horsepower	unit of power measure	hp
peak.rpm	revolutions per minute (how fast engine spinning)	rpm
city.mpg	city mileage	km
highway.mpg	highway mileage	km
city.L.100km	usage of fuel per 100km	Litre
price	sold price of a vehicle	$

Figure 2.2: Numerical Variables

Variables above are chosen for the new data set for the linear regression model. Few missing values were observed in the stroke variable, which is replaced by median of stroke (Fig.2.3).

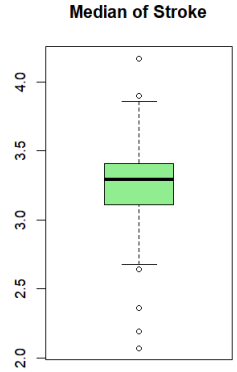


Figure 2.3: Median of stroke

Finally, a new dataset was created that contained 201 observations with 16 variables, all of which were numeric variables. (Fig.2.4)

```
 str(DataNew)
data.frame':   201 obs. of  16 variables:
$ normalized.losses: int  122 122 122 164 164 122 158 122 158 192 ...
$ wheel.base       : num  88.6 88.6 94.5 99.8 99.4 ...
$ length           : num  0.811 0.811 0.823 0.849 0.849 ...
$ width            : num  0.89 0.89 0.91 0.919 0.922 ...
$ height           : num  48.8 48.8 52.4 54.3 54.3 53.1 55.7 55.7 55.9 54.3 .
$ curb.weight      : int  2548 2548 2823 2337 2824 2507 2844 2954 3086 2395 .
$ engine.size      : int  130 130 152 109 136 136 136 136 131 108 ...
$ bore             : num  3.47 3.47 2.68 3.19 3.19 3.19 3.19 3.19 3.13 3.5 ..
$ stroke           : num  2.68 2.68 3.47 3.4 3.4 3.4 3.4 3.4 3.4 2.8 ...
$ compression.ratio: num  9 9 9 10 8 8.5 8.5 8.5 8.3 8.8 ...
$ horsepower       : num  111 111 154 102 115 110 110 110 140 101 ...
$ peak.rpm         : num  5000 5000 5000 5500 5500 5500 5500 5500 5500 5800 .
$ city.mpg         : int  21 21 19 24 18 19 19 19 17 23 ...
$ highway.mpg      : int  27 27 26 30 22 25 25 25 20 29 ...
$ city.L.100km     : num  11.19 11.19 12.37 9.79 13.06 ...
$ price            : num  13495 16500 16500 13950 17450 ...
```

Figure 2.4: Process data set

The target variable for linear regression was “Price”, therefore, the step is intended to find out the stregth of the linear relationship with the target variable. Following Fig. 2.5 illustrated the corrolation matrix (Appendix A), it can be noted that length(~0.691), width (~0.751), curb.weight(~0.834), engine.size (~0.872), horsepower (~0.81), and city.L.100km(~0.789) are a strong positive linear relationship with price while city.mpg (~-0.687) and highway.mpg(~ -0.705) are a strong negative linear relationship with the target variable.

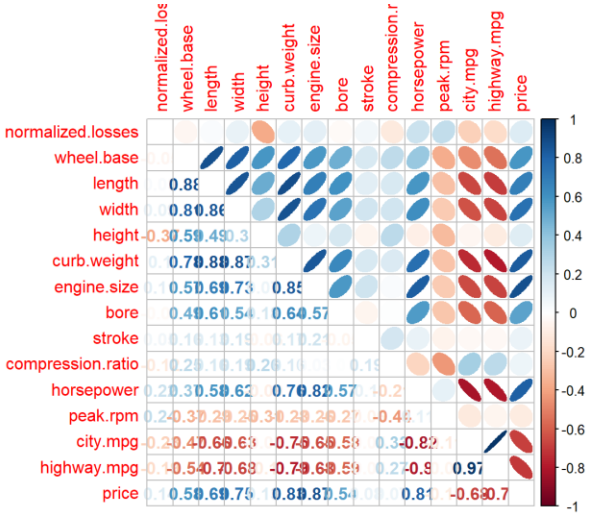


Figure 2.5: Correlation Matrix

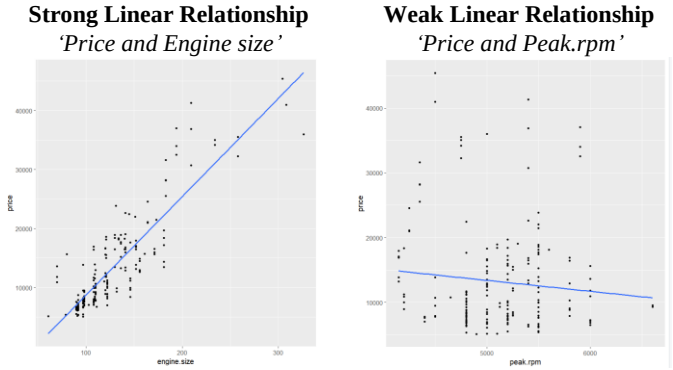


Fig 2.6 : Linear Relationship between a target variable and independent variables

For the purpose to exploring what factors have significant impact on determining price, variables above have strong linear relationships illustrated in the pair plot. (Fig.2.7) .

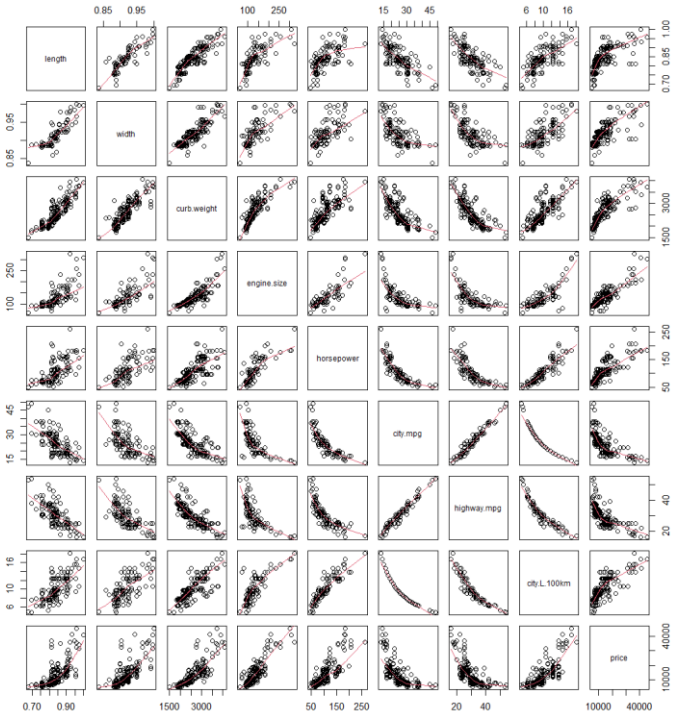


Figure 2.7: Pairwise Plot

C. Data Pre-Processing and Data Exploration for Categorical variables.

For the classification method, a factor variable, "Price" was later created that comprised two levels, "High" and "Low". These two levels resonated with the median price in terms.

Due to a given requirement that a categorical variable should not have more than 2 levels. Therefore, variables with more than 2 levels have been excluded for the classification method, namely, drive wheels, body-style, engine type, num of cylinders, make, and horsepower binned.

III. OBJECTIVES

The objective of this report is to analyze the automobile dataset to uncover the underlying patterns and relationships within the data to shed light on the primary factors affecting automobile prices using various techniques.

Multiple methods will be used to find significant results that impact price. Two models will be established in line with supervised learning. Model 1 will be built by conducting a Multiple Linear Regression Analysis using all numeric variables. Model 2 will be based on Classification Analysis using a decision tree. The models will then be compared based on the results, and recommendations will be made.

Principal Component Analysis, which is a form of unsupervised learning, will be performed on the explanatory variables to reduce the dimensionality of the dataset and to deduce a pattern in it.

IV. MODEL 1- MULTIPLE LINEAR REGRESSION ANALYSIS

The first model for supervised learning was built by using a Multiple Linear Regression whereby several explanatory variables were regressed to make a prediction of the outcome of a target variable [1].

A. Building the model

The variable "Price" was used as the response variable and the 8 variables which have strong linear relationships with the independent variable such as "width", "length", "curb weight", "engine size", "horsepower" "city.L.100k", "city.mpg", and "highway.mpg", were used as the independent variable (Fig. 4.1).

	length	width	curb.weight	engine.size	horsepower	city.mpg	highway.mpg	city.L.100km	price
length	1.0000000	0.8571703	0.8806648	0.6850248	0.5798215	-0.6651924	-0.6981418	0.6573726	0.6906284
width	0.8571703	1.0000000	0.8662011	0.7294356	0.6150767	-0.6335306	-0.6806352	0.6733628	0.7512653
curb.weight	0.8806648	0.8662011	1.0000000	0.8490717	0.7579756	-0.7495431	-0.7948889	0.7853533	0.8344145
engine.size	0.6850248	0.7294356	0.8490717	1.0000000	0.8226756	-0.6505460	-0.6795713	0.7450589	0.8723352
horsepower	0.5798215	0.6150767	0.7579756	0.8226756	1.0000000	-0.8222143	-0.8045748	0.8894883	0.8095746
city.mpg	-0.6651924	-0.6335306	-0.7495431	-0.6505460	-0.8222143	1.0000000	0.9720437	-0.9497129	-0.6865710
highway.mpg	-0.6981418	-0.6806352	-0.7948889	-0.6795713	-0.8045748	0.9720437	1.0000000	-0.9300279	-0.7046923
city.L.100km	0.6573726	0.6733628	0.7853533	0.7450589	0.8894883	-0.9497129	-0.9300279	1.0000000	0.7898975
price	0.6906284	0.7512653	0.8344145	0.8723352	0.8095746	-0.6865710	-0.7046923	0.7898975	1.0000000

Fig 4.1: Correlation coefficient

A backward selection approach was taken to select the significant variables for the model (Fig.4.2).

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -68531.168  13939.274  -4.916 1.87e-06 ***
width         36514.797  16235.523   2.249   0.0256 *
curb.weight       2.578      1.279   2.016   0.0452 *
engine.size      71.434     13.579   5.261 3.78e-07 ***
horsepower       26.547     16.561   1.603   0.1106
city.L.100km   1753.474    381.553   4.596 7.76e-06 ***
city.mpg        495.984    124.786   3.975 9.94e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4.2: Summary of output of the initial model

From the summary output in Fig 4.2, the initial model could be carved out as follows:

Price^ = -68531.169 + 36514.797*width + 2.578*curb. weight +71.434*engine. size + 26.547*horsepower + 1753.474*city.L.100km + 495.984*city.mpg.

All the variables made a positive contribution to predicting the price. A hypothesis test was carried out to evaluate the significance of the explanatory variables in the model.

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

According to the null hypothesis (H0), predictor variables had no implication on defining price. The alternate hypothesis (H1), on the other hand, stated otherwise.

The p-values for the coefficients corresponding to "engine. size", "city.L.100km" and "city.mpg" were significantly lower than the 5% significance levels. Therefore, the null hypothesis could be rejected for the variables at a significance level of 1%, inferring their high contribution to price.

Similarly, the explanatory variables "curb. weight" and "weight" were statistically significant at between approx. 2.5% and 4.5% confidence interval but were slightly less significant than "engine. size", "city.L.100km" and "city.mpg". However, "horsepower" was found to be statistically insignificant as the p-values corresponding to their coefficient parameters were more than the 5% significance level. The null hypothesis had to be accepted as there was insufficient evidence supporting their contribution to life expectancy.

Since "horsepower" was shown to be a non-contributing member of the model, it was removed to enhance the model's performance and to eliminate redundancy. A total of four models were built with significant variables to perform K-fold cross validation (Fig. 4.3).

```
mod1 <- glm(price~engine.size+city.L.100km+city.mpg+curb.weight+width,
            data = newdata)
mod2 <- glm(price~engine.size+city.L.100km+city.mpg+curb.weight,
            data = newdata)
mod3 <- glm(price~engine.size+city.L.100km+city.mpg+width,
            data = newdata)
mod4 <- glm(price~engine.size+city.L.100km+city.mpg, data = newdata)
```

Figure 4.3: Combination of Regression Model

B. *Cross Validation*

The K-Fold Cross Validation method was undertaken to validate each of the 4 models in Fig. 4.3 to check for their errors and the dataset was partitioned into 5 equal groups. A plot was created to map out the errors of each of the models after testing the models (Fig.4.4).

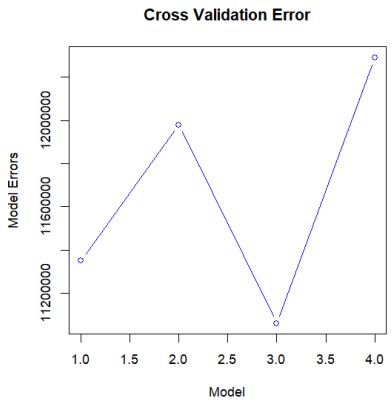


Figure 4.4: 5K-Fold CV Error Plot (Appendix B)

From Fig. 4.4, it could be seen that model 3 had the lowest cross validation error out of the 4 models with the least complexity. Therefore, model 3 was chosen as the final model. The final model included the explanatory variables that significantly impact on vehicle price with the lowest testing error (Fig. 4.5).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -76750.70   12578.95  -6.102 5.51e-09 ***
engine.size     94.75      10.00   9.472  < 2e-16 ***
city.L.100km  1954.74     350.46   5.578 8.01e-08 ***
city.mpg       431.64     122.30   3.529 0.000519 ***
width        52047.07   12362.77   4.210 3.89e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3323 on 196 degrees of freedom
Multiple R-squared:  0.8287,    Adjusted R-squared:  0.8252
F-statistic:   237 on 4 and 196 DF,  p-value: < 2.2e-16
```

Figure 4.5: Summary output of the final model

The model was as follows:

Price^ = -76750.70 + 94.75*engine. size +1954.74*city.L.100km + 431.64*city.mpg +52047.07*width.

C. *Model Evaluation and Testing*

The adjusted R-square value of the new fitted model was 0.8252, which meant that 82.52% of the variance in the data set was explained by the model. The residual error is 3323, which implies that the predicted values deviate from the true regression line by 3323 (Fig.4.5).

To cross check significances of independent variables to the target variable, ANOVA analysis was conducted to check for F-test value (Fig.4.6)

```
Analysis of Variance Table

Response: price
               Df     Sum Sq    Mean Sq F value    Pr(>F)
engine.size     1 9611926358 9611926358 870.458 < 2.2e-16 ***
city.L.100km    1  556134080  556134080  50.364 2.283e-11 ***
city.mpg        1  103091861  103091861   9.336   0.00256 **
width           1  195714738  195714738  17.724 3.889e-05 ***
Residuals     196 2164305651   11042376
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4.6: ANOVA output for new fitted model (Appendix B)

It can be seen clearly from the ANOVA table that (Fig. 4.6), F-statistics = 237 on 4 and 196 degrees of freedom which was 2.417725. This signified that the relationship between the target variable and the model was viable.

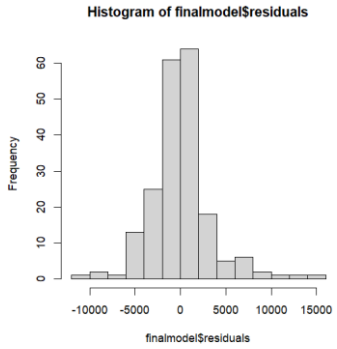


Figure 4.7 – Histogram of Residuals (Appendix B)

Price was predicted based on the newly fitted model and the histogram of residuals was computed as depicted in Fig. 4.7. The histogram exhibited that the residuals were normally distributed.

A further analysis of the error in the residuals of the fitted model was conducted to check if the model had met the key assumptions of a linear regression. The assumptions were assessed based on the residual plots in Fig.4.8.

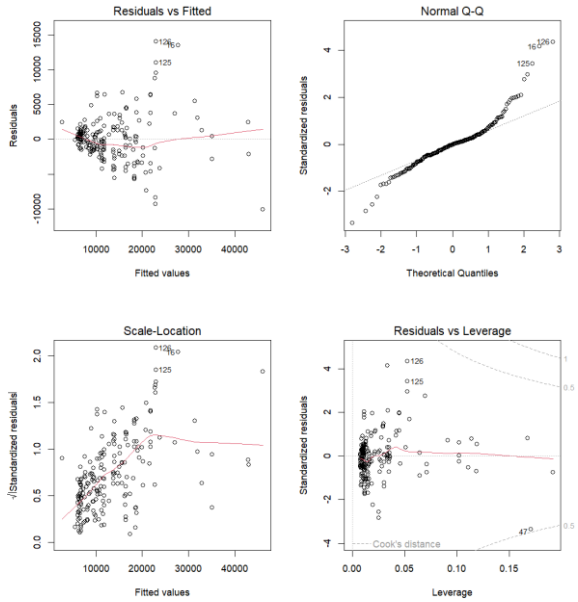


Figure 4.8: Residual plots of the fitted model (Appendix B)

Figure 4.8 comprised four residual plots, with each plot corresponding to a different assumption of linear regression. In the initial plot, "Residual vs. Fitted," it was observed that the residuals were concentrated at the beginning and lacked a discernible pattern. This suggested that the model's adherence to the linearity assumption was rather weak. The "Normal Q-Q" plot illustrates the relationship between the sample and a normal distribution. However, the residuals' points deviated from the straight line, indicating a lack of adherence to the normality assumption. The "Scale-Location" plot revealed non-constant variances in the residuals. The final plot, "Residuals vs Leverage," did not identify any notable outliers of significance.

V. Model 2- Classification Analysis

The second supervised learning model was constructed by employing a Classification Tree, which derived conclusions about the target variable by examining decision rules based on the attributes present in the data set.

A. Building model

To perform a classification analysis, a new factor variable “Price_category” was created based on the value of “price”. Price_category variable was split into levels whereby if the value of price was greater than its median, it was leveled as “High” which indicated high price. If it was below the median, it was labelled as “Low”.

The new variable was added to the dataset to be used as the target variable and “price” variable was dropped. The dataset was split into a training and a validation set in a 80:20 ratio where 80% of the dataset was used to train the model and 20% of used to test the model that was built out of it.

From Fig. 5.1, it could be assessed that “Curb. weight” was the most significant predictor in determining price. City.L.100km, Width, and Length were also second major contributors. Price was high level when curb. weight ranged higher than 2376. On the other hand, the price was low when curb. weight fell below 2376 and both City.L.100km and width fell below 9.129 years and 0.89 respectively.

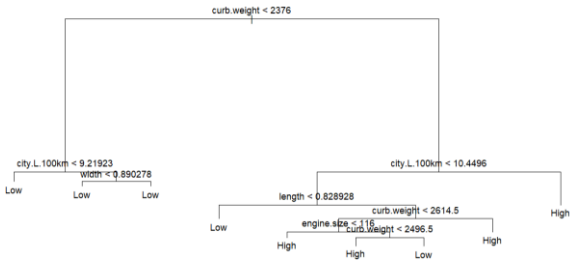


Figure 5.1 – Decision Tree Model for Classification Analysis (Appendix C)

The classification tree model had a misclassification error rate of 0.0375 which meant that 3.75% of the predictions were inaccurately predicted by the model.

B. *Cross Validation*

The training model was validated using the testing dataset, and the outcome was predicted. Upon competition of the cross-validation process, the classification rate was calculated using the predicted and actual values of 'price_category'.

	actual_class	
tree_predict	High	Low
High	22	3
Low	2	14

Figure 5.2: – Predicted vs Actual values.

Using Fig. 5.2, the misclassification error rate for the testing model was computed, which was 0.1219512. It indicated that 12.195% of the predictions were untrue, and this rate was statistically higher than that of the training model.

C. *Pruned Model*

The initial model had 9 terminal nodes in total, which did not represent a complicated tree to interpret. But is recommended to have a smaller tree with fewer splits for smoother interpretation. Therefore, to reduce redundant and non-critical splits and to achieve a less complex model, a process of pruning was conducted.

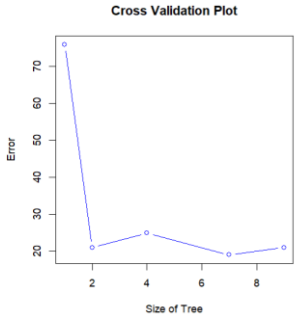


Figure 5.3: CV Error vs Tree Size (Appendix C)

Fig. 5.3 depicts the cross-validation error for each terminal node in the tree model. The model tree size with 2 and 7 terminal nodes had the lowest error out of all the other nodes, but since the tree size with 2 nodes was not adequate for the analysis, the model tree size of 7 terminal nodes was deemed best.

The tree model with a size of 7 terminal nodes was trained with the training set. The pruned model is depicted in Fig. 5.4

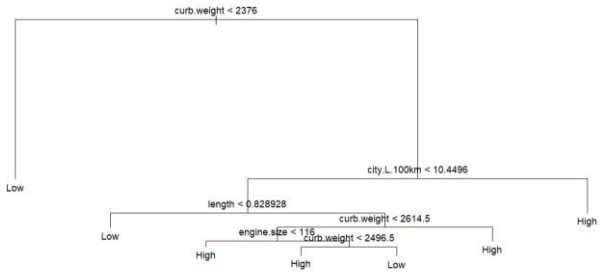


Figure 5.4: –Pruned Decision Tree Model (Appendix C)

The pruned model had 7 terminal nodes, with "Curb. weight" still being the most significant predictor. Since the difference in terminal nodes between the original tree and the pruned tree was slim, the features were not different from each other. According to the pruned model, price was high when "curb. weight" was higher than 2376. Also, the second significant variable was noted city.L.100km, and where price was predicted high if city.L.100km was higher than 10.4496. If "curb. weight" fell below 2376 and, then price was predicted to be low.

Low curb. weight and city.L.100km resulted in a low level of price. Similarly, low city.L.100km and length lower than 0.828 translated to a low level of price.

From the confusion matrix in Fig. 5.5, the misclassification error rate was still at 12.195%, which meant that the pruned model was free of any redundant terminal nodes and 87.805% of its predictions were correctly executed.

	actual_class	
tree_predict	High	Low
High	22	3
Low	2	14

Figure 5.5: Predicted vs Actual Values of Pruned Tree Model.

VI. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a form of unsupervised learning that is used to reduce the dimensionality of the life expectancy dataset to obtain a lower dimensional dataset while preserving as much of the data's variation as possible. [1].

This technique of unsupervised learning was utilized in the report to uncover any discerning patterns that might be of importance and to visualize the variation of the dataset with many predictor variables.

The original data set has 29 variables. For PCA analysis, the categorical variables have been excluded from the data set. Because the levels of categorical variables will create new variables, and as a result, this can increase the dimensionality of the data set. Secondly, the remaining 15 numerical variables also is considered to be complex. Therefore, PCA analysis is executed only on the variables ('length', 'width', 'height', 'engine. size', 'curb. weight', 'city.L.100km, city.mpg, and highway.mpg) which has strong correlations with the target variable, and the target variable has also been excluded.

Before starting the PCA processes, the mean and variance of the dataset were analyzed to check for any contrasting values (Fig. 6.1).

```
#Mean
#length      width     curb.weight   engine.size  horsepower    city.mpg     highway.mpg   city.L.100km
#0.8371023   0.9151258 2555.6666667  126.8756219  103.4055339   25.1791045   30.6865672    9.9441455

#Variance
#length      width       curb.weight    engine.size horsepower   city.mpg     highway.mpg  city.L.100km
#3.506151e-03 8.518865e-04 2.675959e+05 1.726139e+03 1.396195e+03 4.125776e+01 4.644627e+01 6.424193e+00
```

Figure 6.1: Mean and Variance of Explanatory Variables

It could be noted that the mean and variance of all the variables differed from each other.

As PCA loads only on the large variance, therefore, the dataset was scaled and normalized to avoid any dominance by the first principal component.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.5231 0.8977 0.66605 0.38231 0.32362 0.25979 0.21836 0.13597
Proportion of Variance 0.7957 0.1007 0.05545 0.01827 0.01309 0.00844 0.00596 0.00231
Cumulative Proportion  0.7957 0.8965 0.95193 0.97020 0.98329 0.99173 0.99769 1.00000
```

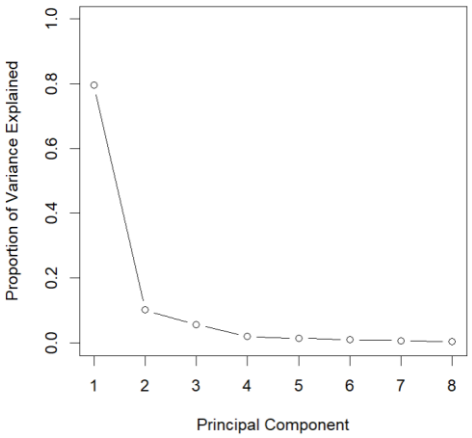
Figure 6.2: Summary of PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
length	0.3332897	0.4766261	0.32551680	0.49422027	-0.42209007	0.34804899	0.02562255	-0.10759191
width	0.3349434	0.4814955	0.12517231	-0.79044028	0.01408091	0.05130477	0.09945659	0.05281134
curb.weight	0.3708339	0.2871132	-0.05449399	0.25501568	0.18507604	-0.77598658	-0.24037931	0.13362293
engine.size	0.3417099	0.1439066	-0.67681708	0.17227879	0.44435811	0.41201561	0.07328211	0.04671855
horsepower	0.3503835	-0.3049415	-0.44657404	-0.09895905	-0.66764011	-0.20375211	0.29336777	-0.04120580
city.mpg	-0.3594670	0.3798938	-0.32629992	-0.04803211	-0.09414733	-0.14344419	-0.10921057	-0.75915651
highway.mpg	-0.3657836	0.2929573	-0.32950789	-0.05019111	-0.36469529	0.09319563	-0.46385932	0.55895458
city.L.100km	0.3696818	-0.3386946	0.05333191	-0.14703195	-0.04821558	0.18314229	-0.78303405	-0.27413589

Figure 6.3: Principal Components with each variable

This produced the principal component loading vectors. Each column contained the corresponding principal component loading vector. Eight principal components were formed and almost all variables had an equal contribution to PC1. "Curb. weight" had the highest contribution in PC1, whereas in PC2, "width" and "length" had the highest contribution.

From the plots in fig. 6.4 and fig. 6.5 (Appendix), it was evident that PC1 captured 79.6% of the variation in the dataset, while PC2 explained about 10.1% of the variation, and so forth. The first and second component together explained 89.7% of the variation in the automobile dataset. All the components slowly explained the entire dataset, but the first two principal components explained the largest portion of the variation.



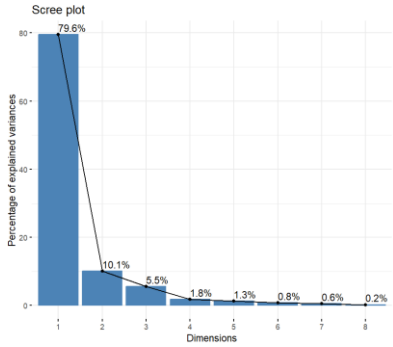


Figure 6.4: Proportion of Variance Explained by Principal Components

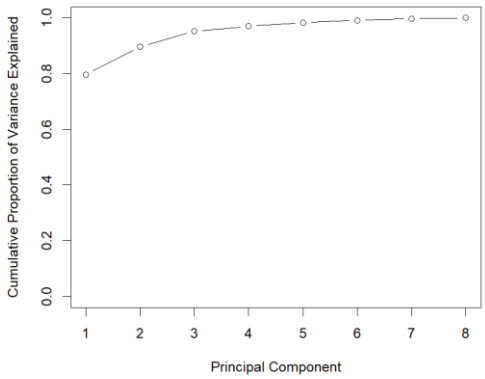


Figure 6.5 – Cumulative Proportion of Variance Explained by Principal Components (Appendix D)

A biplot was drawn to visualize the results of PCA. A PCA biplot combines both the PCA score plot and the loading plot, making it easier to capture the influence of the vectors on principal components (Fig. 6.6)

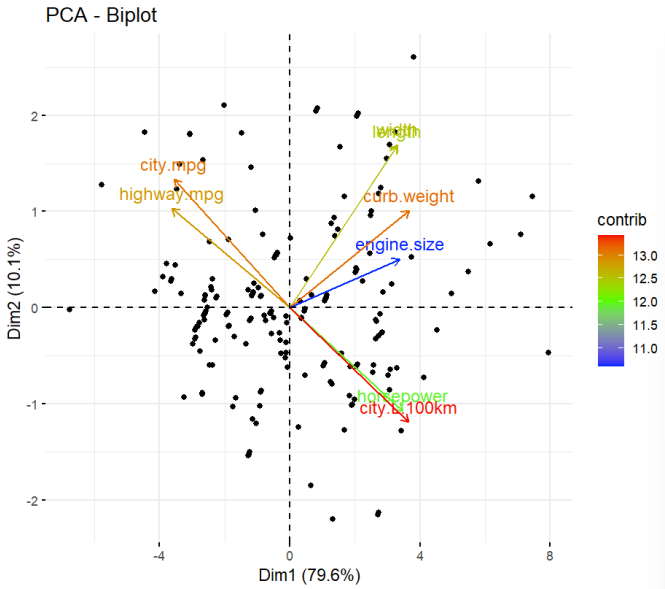


Figure 6.6: Principal Component Analysis Biplot (Appendix D)

From the PCA biplot (Fig. 25), it could be assessed that the first loading vector (PC1) placed the most weight on "curb.

weight" and "city.L.100km" while all the other variables contributed almost equally to it. "City.mpg' and 'Highway' were located very close together, which indicated that they were negatively correlated. The second loading factor, PC2, placed most of its weight on "Width" and "Length" while, 'horsepower' and 'city.L.100km' were located close.

VII. MODEL COMPARISON

Two models were built to analyze and predict the effects of vehicles' attributes on determining price.

Model 1 was built by using Multiple Linear Regression (MLR). All the predictor variables were regressed on the target variable, from which it was noted that all the variables had a somewhat impact on the target variables, however, 8 variables out of 15 numerical variables had a strong linear relationship, and on which the model was built and cross validated. As a result, engine size, city.L.100km, city.mpg, and width had the strongest influences on determining price. In addition, 82.52% of the variables were explained by the model.

Model 2 was built by using a Classification Tree for which a factor variable of price was created with two levels, "High" and "Low". After training, validating, and pruning the tree, the model predicted an outcome whereby "Curb. weight", was indicated to be of utmost significance. "horsepower", which was removed from the classification because of its insignificance, appeared to be a less significant contributor to price in the decision tree. "Curb. Weight", "city.L.100km", "length", and "engine. size" variables were used in the best model of tree construction.

Both the models concluded with similar results that the "Engine. size" and "City.L.100km" were significant contributors to determining price. In both models, horsepower has the least significance in the price.

The models suggest that if a car with a higher engine size and higher city. L.100km tends to be higher price whereas higher city mpg and higher highway mpg tends to be deductive in price. Although both models had a resemblance in identifying significant factors in price, 'Curb. weight' had been identified significantly in the tree classification model.

VIII. CONCLUSION AND RECOMMENDATION

The analysis in this paper was conducted to find out what factors had significant impact on price. Both the models, Multiple Linear Regression and Classification Tree suggested that "Engine size" and "City.L.100km" had significant impacts on price.

It indicated that variables above were positively correlated with the target variable. Therefore, the null hypothesis (Ho), which was that none of the explanatory variables had an impact on the price, was rejected.

Surprisingly, horsepower has less significant on price amongst variables indicated strong linear relationships with the target variable. In addition, there were some variables that indicate no or less significance in determining price such as stroke, bore, compression ratio and peak. rpm etc. based on the correlation matrix.

Nevertheless, it is essential to acknowledge the limitations associated with both models, rendering them less reliable tools for predicting precise pricing. One notable limitation is the exclusion of categorical variables with more than two levels. This decision restricts the models' ability to encompass the entirety of variables that should ideally be taken into account in the pricing prediction process.

Overall, the models illustrate that price can be predicted accurately based on the research methods. For instance, Model 1 was explaining 82.52% of variables and F-statistics = 237 on 4 and 196 degrees of freedom which was 2.417725. whereas 87.805% of its predictions were correctly executed in model 2.

Consequently, this indicates that training and testing the accurate model may be able to help manufacturers and seller companies to make right pricing decisions in relation to their business models as well as predicting market competitors. Yet, further study is recommended for better understanding of the determinants.

REFERENCES

1. Dalgard, P 2008, "Introductory Statistics with R", Springer, 2nd Edition
2. Gareth. J, Daniela. W, Trevor. H, and Robert. T, 2013, "An Introduction to Statistical Learning with Applications in R", 2nd Edition.
3. McKinsey, 2016 "Automotive revolution – perspective towards 2030" < https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/disruptive-trends-that-will-transform-the-auto-industry/de-DE>
4. PwC, five trends transforming the Automotive Industry < https://pwc.to/2RbDS7g>