USED CAR PRICE PREDICTION USING LINEAR REGRESSION

Business Intelligence

Professor : Giovanna Maria

Shubhrajit Pal

Engineering Management
Department of Information engineering and mathematics
University of Siena
[2023/2024]

Chapter 1 : Abstract

It is considerably easier to determine the price that a manufacturer will set for a new car due to the fixed expenses involved in its production, government taxes, the market segmentation that the company is targeting with a specific car model and other already specified factors. This is not the case with used or older vehicles, as these aspects are not always considered. It may depend on various variables and with new automobile prices constantly rising, many consumers have been driven to acquire used or older vehicles as alternatives, a trend that is on the rise. A consumer attempting to purchase an old or used car faces the issue of not understanding how the price of such a car can be calculated or predicted. This difficulty would not exist if a technology could accurately forecast the prices of such cars based on known automotive specifications. Our study attempts to create a linear regression model to estimate the prices of used cars.

Chapter 2 : Introduction

This study utilizes data from Kaggle on second-hand vehicles :
https://www.kaggle.com/code/celestioushawk/cardekho-used-cars/data?select=CAR+DETAILS+FROM+CAR+DEKHO.csv

It has 4340 rows and 8 columns. It was investigated for appropriateness.

This work has the following sections:

- *Dataset Information*

This will have the details about the information within the dataset. That includes the columns with which we will be working and the type of data point we will found for the used cars.

- *Data Cleaning*

The next step after checking the rows and columns of the dataset is to clean up the data. That means we shall check if there are null records or duplicate rows. Also if splitting or sorting of data point in specific columns is needed or not. Basically, the dataset will be explored to see if the data can be made more organised so that the machine learning model can use it without getting confused

- *Data Distributions*

Now that the dataset is arranged and error free, next we proceed to check how the data point are distributed. We would take specific columns and analyse the frequency with which those are varying, if they are correlated or not and will get an idea how we can proceed further on our analysis.

- *Data Transformation*

Once the data distribution is analysed, the data transformation of especially the categorical columns are done(as we are using regression model). As the machine learning model cannot specifically understand the string values in the categorical columns, those are encoded to make the algorithm run efficiently.

- *Machine Learning Models*

With the dataset cleaned and transformed, we sample the dataset into training set and testing set and apply the machine learning models. The price prediction is done with two ML Algorithms namely : Simple Linear Regression and Regression Tree (with and without Bagging technique).

- *Results and Conclusion*

The evaluation of above mentioned machine learning models with respect to the performance metrics such as Root Mean Square Error and R Square are done and thus concluded.

Chapter 3 : Dataset Information

The data set is about used cars and was obtained from the online Data Science and Machine Learning interactive site Kaggle.

It contains a total of 4340 records with eight attributes namely;

- *name* : The car's brand including its model.
- *selling_price* : This is the current selling price of the car.
- *year* : The year that the car was initially bought.
- *km_driven* : The number of kilometers that the car has accumulated while being driven.
- *fuel* : The type of fuel that the car consumes.
- *seller_type* : The person selling the car whether it is the owner or a dealer.
- *transmission* : This is how the car's power moves from the engine to the wheels-Automatic or Manual transmission.
- *Owner* : The number of people that have owned the car since its manufacture.

Chapter 4 : Data Cleaning

First we split the Car Make attribute (column : name) which combined both the cars' brand and model records to obtain two different columns : name and model.
Second, we checked if there is incompleteness in the dataset (NULL values).
Third, we checked the duplicate data points

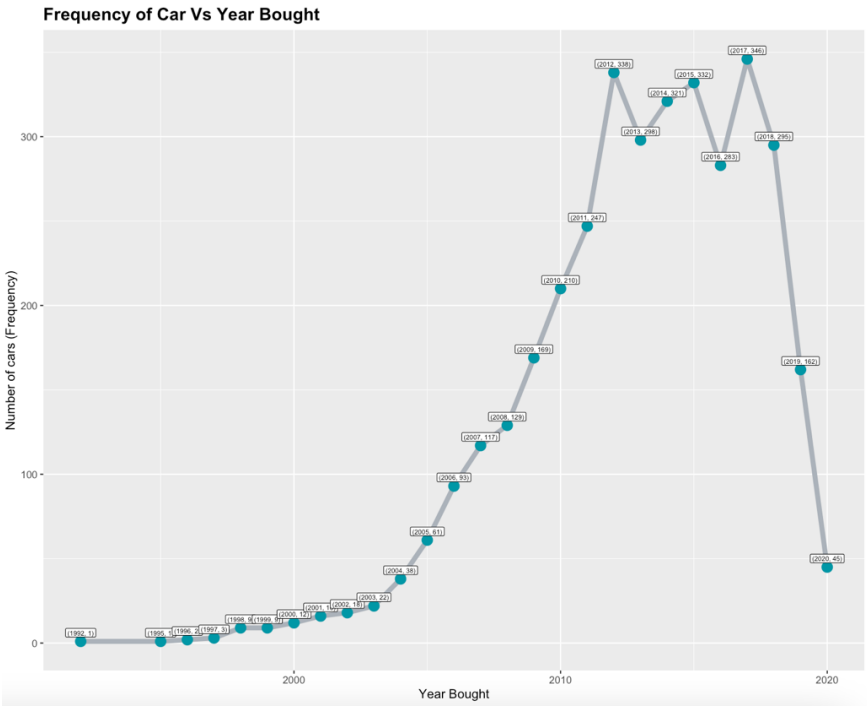
Result :
We saw there was no NULL values in any column. But we have to remove 743 duplicate rows (which we removed). **So we after data cleaning have a dataset of 3577 rows and 9 columns.**

Chapter 5 : Data Distribution

The various statistical distributions of the dataset's columns to see how they were related to each other, and, to Selling Price (column : selling_price) which is the target variable.

Plotting 'Car Count' vs 'Year'

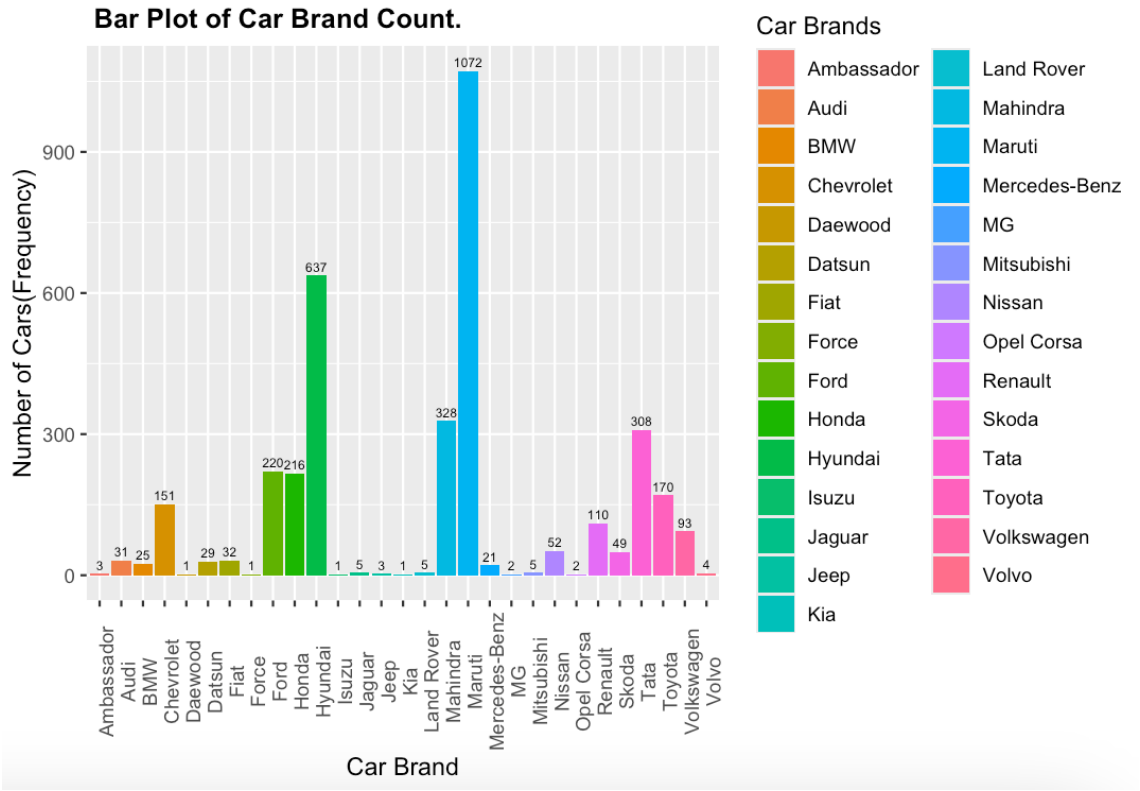
The aim was to establish the counts of cars according to the years that they were initially bought by plotting visually.



2017 had the largest count of used cars at 346 while 1995 and 1992 each had one car each respectively.

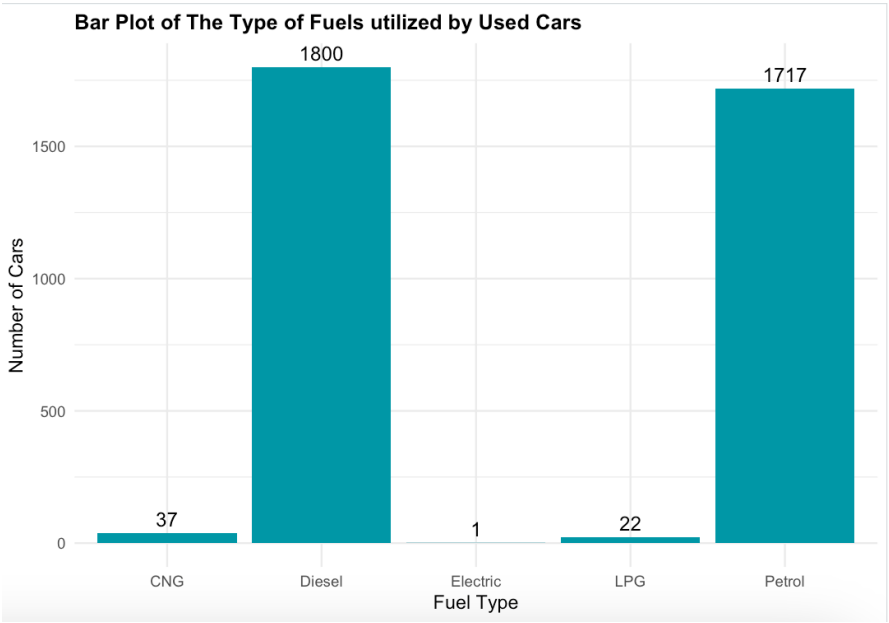
Distribution of Car Brand

To establish the car brands that had the highest and lowest counts we visualized their results in a plot.



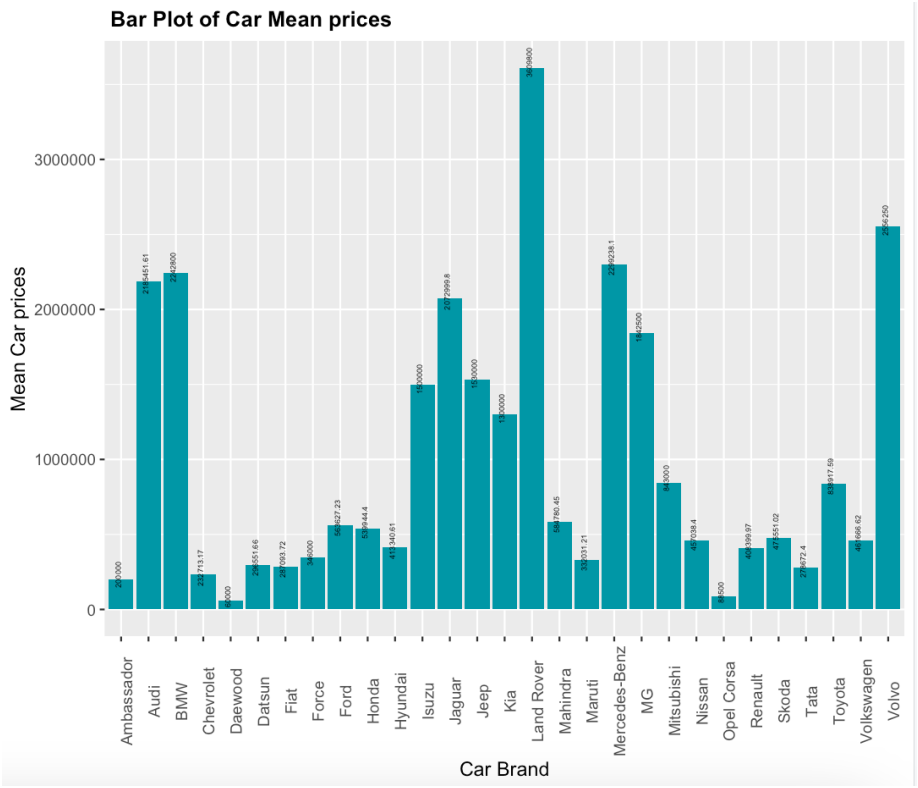
Maruti has the highest number of car (1072), second is Hyundai (637), while Kia, Isuzu, Force, and Daewood all had one record

Plotting 'Fuel Type' vs 'Frequency of Car'



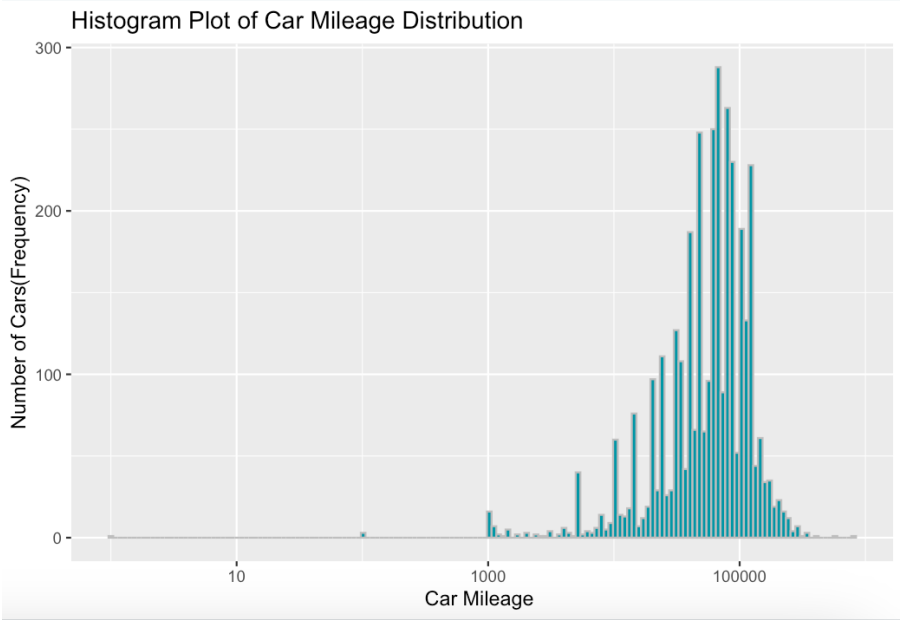
The used cars are almost divided into two categories-diesel and petrol powered. There was only one car that utilized electric power.

Distribution of the mean Car Prices according to the Car Brand



Land Rover had the highest mean price of all the car brands followed by Volvo whereas Daewood had the lowest mean prices.

Distribution of cars mileage

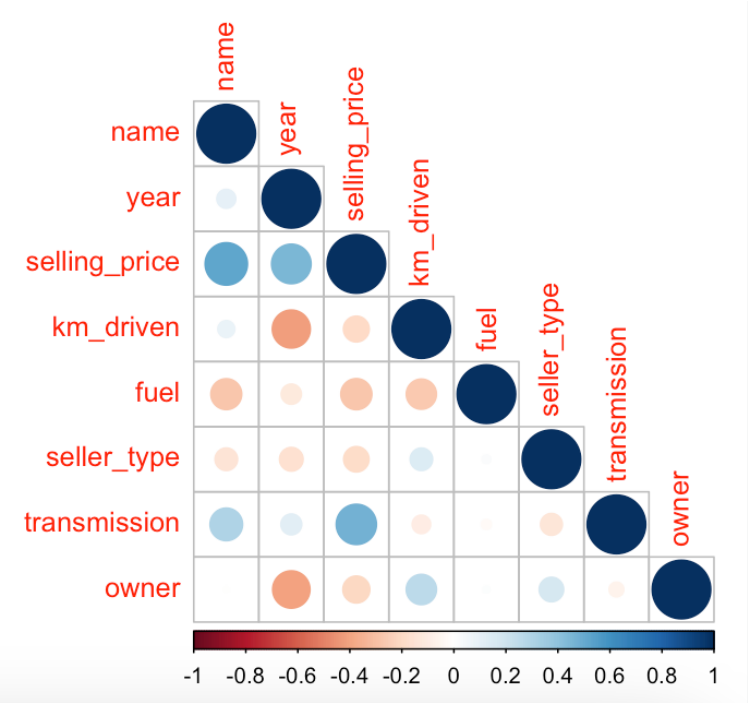


The plotted histogram pointed out that the car mileages were skewed to the right which is an indication that most of the cars in the data set had accumulated more than 100,000 km while being driven on the road.

Chapter 6 : Data Transformation

Our data set had five attributes which were categorical. These were name, seller_type, fuel, transmission, and owner and we converted them to integers through ordinal encoding since regression only utilizes numeric variables. Thus, a label encoding algorithm was implemented to help normalize these attributes.
Secondly, The model column (representing Car Model) had so many unique records in it making it hard to carry ordinal encoding on it and we had to drop it from the data as we already had the Brands in the column "name".

Then we obtained the correlation matrix from our data set to see how the predictor variables correlated with Price as well as with each other.



From the resulting correlation plot, we can see that the target variable(selling_price) is very positively correlated with three predictor variables : name, transmission, and year. Then, 'km_driven' and 'owner', 'km_driven' and 'seller_type', 'owner' and 'seller_type', 'name' and 'transmission' all had very little correlation respectively.

Chapter 7 : Machine Learning Models

The first step we do is we sample the data in 2 parts : Training set and Test set.
The Training set has 2416 rows and 8 columns and our Test set has 1137 rows and 8 columns.

Then, we run two machine learning algorithm on them :

A) Simple Linear Regression
B) Regression Tree

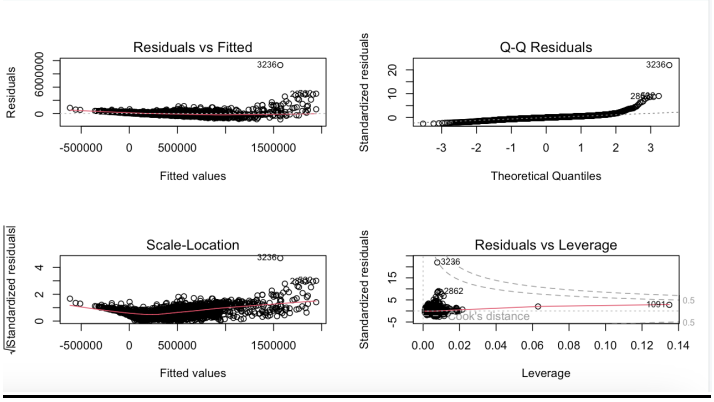
SIMPLE LINEAR REGRESSION:

After training the model on train dataset , use it to predict the price in the test dataset.
We plot predicted selling price with the original selling price of the test data :

Simple Linear Regression : Plot of Actual Selling Price vs Predicted Selling Price

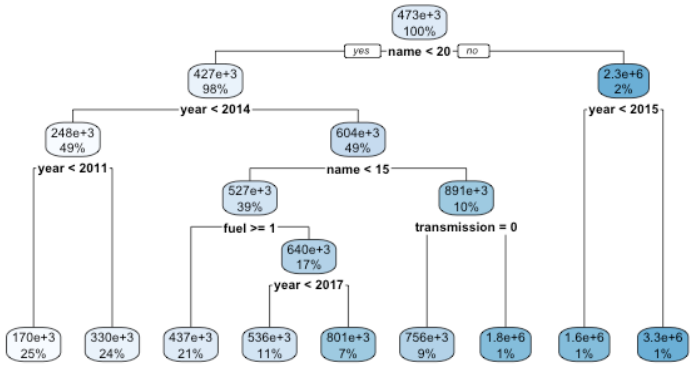
The R square value of the predicted test data using this model is : 0.535625752140327

The Root Mean Square value of the predicted test data using this model is : 280359.70764486



REGRESSION TREE :

To establish whether our regression model could generalize better, we built a Regression Tree to predict Price as a function of the seven predictor variables namely Year, Car_Make, Mileage, Fuel_Type, Seller_Type, Transmission, and Num_Owners. We fitted the Regression Tree using rpart by setting the method as 'anova'.

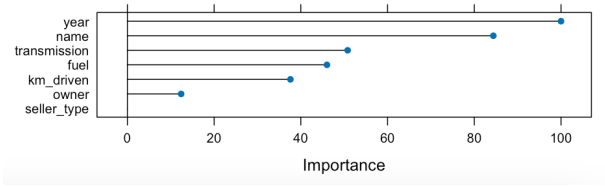


The fitted regression tree model started at 2416 observations as the root node and the first variable that entered the split was "name". The use of 'name'(representing the car brands in the dataset) in the first split suggests that there are significant differences in 'Selling Price' based on different car brands. This initial split helps to create more homogeneous subsets in terms of 'Selling Price', leading to a lower overall variance(SSE). This indicates that the make of the car is a strong predictor of its price, which is logical given that different car brands often have different market values *(as we saw earlier while plotting car mean prices w.r.t car brands).*

The R square value of the predicted test data using this model is : 0.595194613660132

The Root Mean Square value of the predicted test data using this model is : 261760.840138064

Next we improved the model by using the ***Bootstrap Aggregation or bagging technique in the Regression Tree***. This algorithm was used to improve the model i.e. to reduce chances of overfitting. Secondly, and most importantly, we can know exactly or rank the importance of the predictor variables in the model.



We see the most important features for the model are 'year', 'name'(which means : car brand), transmission. This is was we exactly saw in our correlation plot also.

After using this model to predict the selling price on the test dataset:

The R square value of the predicted test data using this mode is : 0.675882531876479

The Root Mean Square value of the predicted test data using this model is : 95239.0151303883

Conclusion

The dataset was split in 2:3 ratio (testing set : training sets).

In the Linear regression model, six of the seven predictor variables had p-values less than the statistically significant lambda value of 0.05. It is only 'seller_type' variable that yielded a higher p-value indicating that it was contributing very little in explaining the variance in the model. The same (together with owner attribute) is replicated in the improved Regression Tree built with bagging.

The below table shows the comparison of the ML models with respect to the performance metrics :

ID	Model Name	R Square Value (%)	Root Mean Square Error Value
1	Simple Linear Regression	53.5%	280359.7
2	Regression Tree	59.5%	261760.8
3	Bootstrap Aggregation Regression Tree	67.5%	95239.01

We see that using the Bootstrap Aggregation Regression Tree had a significant improvement in R square value than using Simple Linear Regression. Also the RMSE shows improvement. The bagged Regression Tree model's accuracy is superior as compared to the Linear Regression model or the basic Regression Tree.

Also the top two attribute for predicting the **selling price** for all the models was : **Year** (column : year) **and Car Brand** (column : name), and next three features was the **Mileage** (column: km_driven), **Fuel Type**(column: fuel), **Transmission** (column: transmission). The Ownership Type (column: owner) and Seller Type(column: seller_type) was not considered as significant by all the models.