Report: Predicting Prices of Used Cars

## Introduction

This report analyzes a dataset of used Toyota Corolla cars sold in the Netherlands during the late summer of 2004. The goal is to predict the price of a used Toyota Corolla based on several car specifications. To achieve this, we apply multiple regression i.e. Multiple Linear Regression. Additionally, We used Backward Elimination, Forward Elimination and Stepwise elimination to see if we can reduce the number of variable and complexity of the model.

The dataset consists of 1436 records and 38 attributes. We focus on key predictor variables like age, kilometers driven, horsepower, fuel type, and additional features such as air conditioning and CD players. The goal is to evaluate and compare different models to identify the most significant predictors of car price.1.Split the data into training (60%), and validation (40%) datasets. Run a multiple linear regression with the outcome variable Price and predictor variables Age\_08\_04, KM, Fuel\_Type, HP, Automatic, Doors, Quarterly\_ Tax, Mfr\_Guarantee, Guarantee\_Period, Airco, Automatic\_airco, CD\_Player, Powered\_Windows, Sport\_Model, and Tow\_Bar.

## Data Splitting

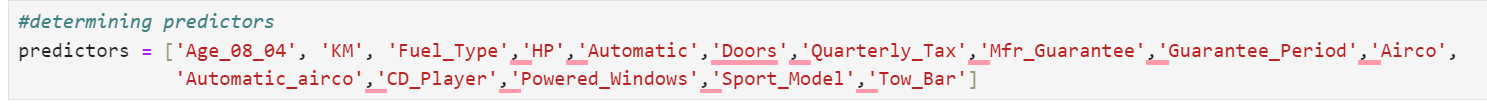
We split the dataset into two subsets:

Training Set: 60% of the data

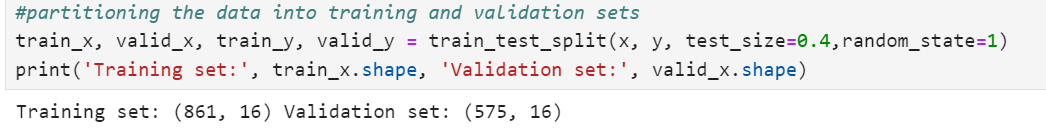
Validation Set: 40% of the data

The predictors selected for the analysis include Age\_08\_04, KM, HP, Fuel\_Type, Automatic, Doors, Quarterly\_Tax, Mfr\_Guarantee, Guarantee\_Period, Airco, Automatic\_airco, CD\_Player, Powered\_Windows, Sport\_Model, and Tow\_Bar.

Step 1: Determining the Predictors



Step 2: Splitting the data



Step 3: Normalization to scale the data

A screenshot of a computer

Description automatically generated

## Multiple Linear Regression Analysis

We ran a multiple linear regression analysis with the outcome variable being Price and the predictor variables being Age\_08\_04, KM, Fuel\_Type, HP, Automatic, Doors, Quarterly\_Tax, Mfr\_Guarantee, Guarantee\_Period, Airco, Automatic\_airco, CD\_Player, Powered\_Windows, Sport\_Model, and Tow\_Bar.

A screenshot of a computer

Description automatically generated

***Results and Key Metrics***

The regression model provides coefficients for each predictor variable, indicating their impact on the price.

Significant Variables: Based on the coefficients, the three or four most important specifications for predicting the car's price appear to be Age\_08\_04, KM, Automatic\_airco, Fuel\_Type\_Diesel, Fuel\_Type\_Petrol.

The regression model was built using the training dataset. The model used the Price as the dependent variable and the above-mentioned predictor variables.

The table below summarizes the regression coefficients for each featuXre:

•mean absolute error (MAE): 929.57

On average, the model is off by $929.57 in its price predictions

•R-squared (R²): 0.88

This indicates that 88% of the variance in car prices is explained by the model, making it a reasonably good fit for the data.

A screenshot of a computer

Description automatically generated

A screenshot of a chat

Description automatically generated

We need to determine the best subset of predictor variables by below mentioned elimination methods.

## Backward Elimination

Backward elimination was used to iteratively remove the least significant predictor. It starts with all predictors and eliminates them one at a time until only significant predictors remain.

A close-up of a computer screen

Description automatically generated

*Analysis:*

Backward elimination started with all predictors and eliminated the least significant one (Airco) based on AIC score.

Finally, we retained 15 significant predictors, including Age, Kilometers, Horsepower, Automatic, Doors, and Fuel Type (Diesel and Petrol) by eliminating Airco.

## Forward Elimination

Forward elimination was used to iteratively add significant predictors, starting with none and adding one at a time until no significant predictors were left to add. This method helps in finding out the most relevant predictors for predicting car price.

A screenshot of a computer

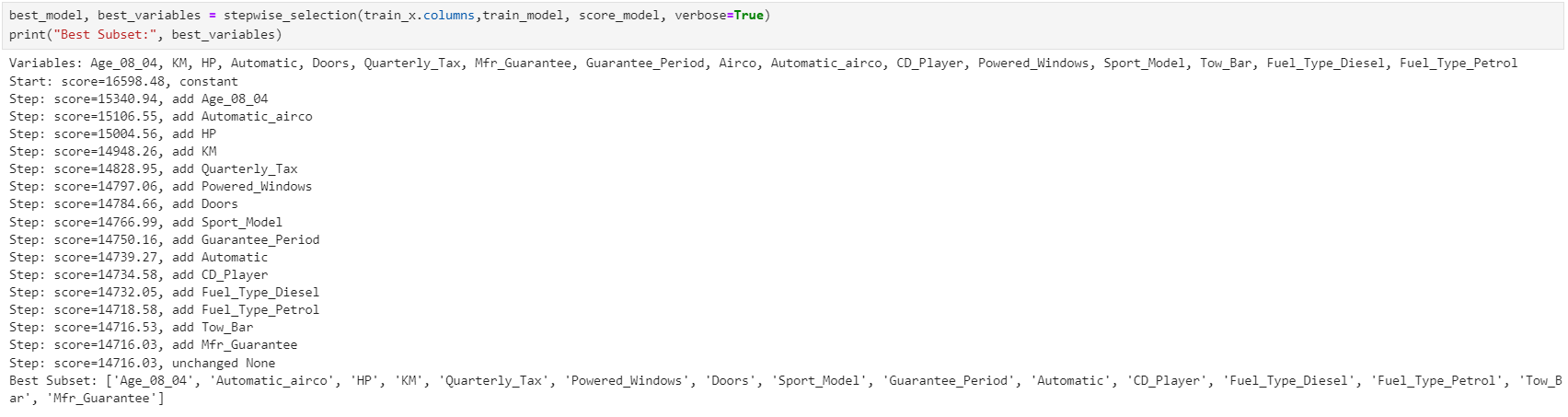
Description automatically generated ***Analysis:***

Forward elimination began with no predictors and iteratively added significant predictors based on AIC scores.

It includes 15 significant predictors such as Age, Horsepower, Automatic Air Conditioning, Kilometers, Fuel Type, and other features and eliminated Airco.

## Stepwise Regression

Stepwise regression was used to iteratively add and remove predictors based on their statistical significance. This method combines forward selection and backward elimination to identify the best set of predictors.



***Analysis:***

It started by adding the most significant variable and then moved iteratively, adding or removing variables based on their contribution to the model

Finally, it selected 15 significant predictors such as Age, Kilometers, Horsepower, Automatic air conditioning, Fuel Type, and other key car features and eliminated Airco.

All the above three models eliminated Airco variable. Henced the best subset of variables is :

['Age\_08\_04', 'Automatic\_airco', 'HP', 'KM', 'Quarterly\_Tax', 'Powered\_Windows', 'Doors', 'Sport\_Model', 'Guarantee\_Period', 'Automatic', 'CD\_Player', 'Fuel\_Type\_Diesel', 'Fuel\_Type\_Petrol', 'Tow\_Bar', 'Mfr\_Guarantee']

## mULTIPLE LINEAR rEGRESSION MODEL WITH best subset variables:

Now we will fit the linear regression model with best subset variables obtained by elimination models.

A screenshot of a computer

Description automatically generated

Even for best subset of variables we found most important specifications for predicting the car's price appear to be Age\_08\_04, KM, Automatic\_airco, Fuel\_Type\_Diesel, Fuel\_Type\_Petrol.

A screenshot of a computer

Description automatically generated

•mean absolute error (MAE): 947.24

On average, the model is off by $947.24in its price predictions

•R-squared (R²): 0.87

This indicates that 87% of the variance in car prices is explained by the model, making it a reasonably good fit for the data.

## Conclusion

In this report, we applied multiple linear regression to predict the price of used Toyota Corollas based on a dataset of car specifications.

Multiple Linear Regression provided a baseline model with all available predictors, giving us an mean absolute error of $929.57 and R-squared of 0.88. This model included 16 predictors, some of which showed minimal influence on the price.

We used Backward Elimination, Forward Elimination and Stepwise elimination to see if we can reduce the number of variable and complexity of the model.

After selecting best subset of the variables, we again ran the multiple regression model with 15 variables and found out the MEA is 947.24 & R-squared of 0.87.

After elimination process only one variable was dropped and mean absolute error also increased and R squared value decreased hence, we will choose the Model with 16 Variables where Mean absolute value is lower for the prices of the Cars.

**Key Insights:**

The following predictors consistently emerged as significant factors influencing the price of a used Toyota Corolla:

Age\_08\_04, KM, Automatic\_airco, Fuel\_Type\_Diesel, Fuel\_Type\_Petrol.