Prediciting Used Toyota Corolla Prices in Europe

Term Paper - QNT 450

Ke Yang

Dario Cavada

Karan Prajapati

Abstract

The automotive industry is a dynamic and ever-changing landscape, with car manufacturers always seeking to improve driving quality and meet consumer demands through innovative methods. In this fiercely competitive market, price is a significant factor that plays a crucial role in determining which car a consumer purchases, and it is often based on numerous vehicle features.

One such car that has been marketed in several nations over the years and has been advancing in its characteristics as it has passed from generation to generation is the Toyota Corolla. To examine the correlation between the price of the Toyota Corolla and its specifications, we will use a dataset that details its various attributes, including its price.

Introduction

By analyzing this data, we can identify which attributes have the highest correlation with the car's price and how these elements can influence the car's total value. This information can be beneficial to consumers, dealers, and automakers alike, as it can provide insights into the variables that affect car costs and guide purchasing options.

This article examines the implications of the R code provided for the Toyota Corolla dataset regarding model selection and regression analysis. The Toyota Corolla dataset includes details on 1436 cars, such as their age, mileage, power, weight, fuel type, and price. The goal is to build a regression model with an accurate characteristic-based vehicle price prediction. This paper will cover the steps taken to develop and evaluate the regression model and use regression model selection techniques to choose the best model.

Data Exploration:

The first step in creating a regression model is exploring the data and understanding its characteristics. The Toyota Corolla dataset includes information on 1436 vehicles and nine predictor variables. We first load the dataset into R to begin the data exploration process, then use the 'create_report' function of the DataExplorer package to generate descriptive statistics and visualize the data.

There are 9 predictor variables and 1 response variable in the dataset.
The response variable is **Price** (price of car in dollars).
The predictor variables are:

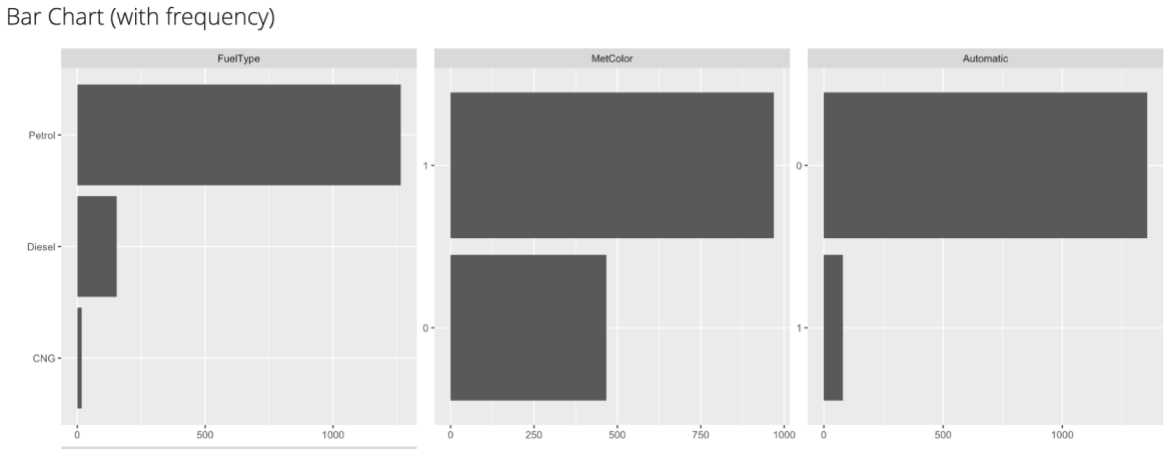
- **Age**: (age of the car)
- **KM:** (Kilometers driven)
- **FuelType:** (Diesel or Petrol or CNG)
- **HP:** (Horsepower)
- **MetCol:** (Metal Color)
- **Automatic:** (Automatic or Manual)
- **CC:** (engine size)
- **Doors:** (number of doors)
- **Weight:** (weight of the car)

Dummy Variable: Since the variable *FuelType* is a qualitative categorical data with three categories, we had to create a dummy variable for it. We named the new variable *petrol*. We changed it into 0 or 1 variables. 1 means the car uses petrol as the fuel type and 0 means they DO NOT use petrol as the fuel (they use CNG or Diesel).

```
# Creating the dummy variable
toyota$petrol<-ifelse(toyota$FuelType=='Petrol',1,0)
toyota$petrol<-as.numeric(toyota$petrol)
```

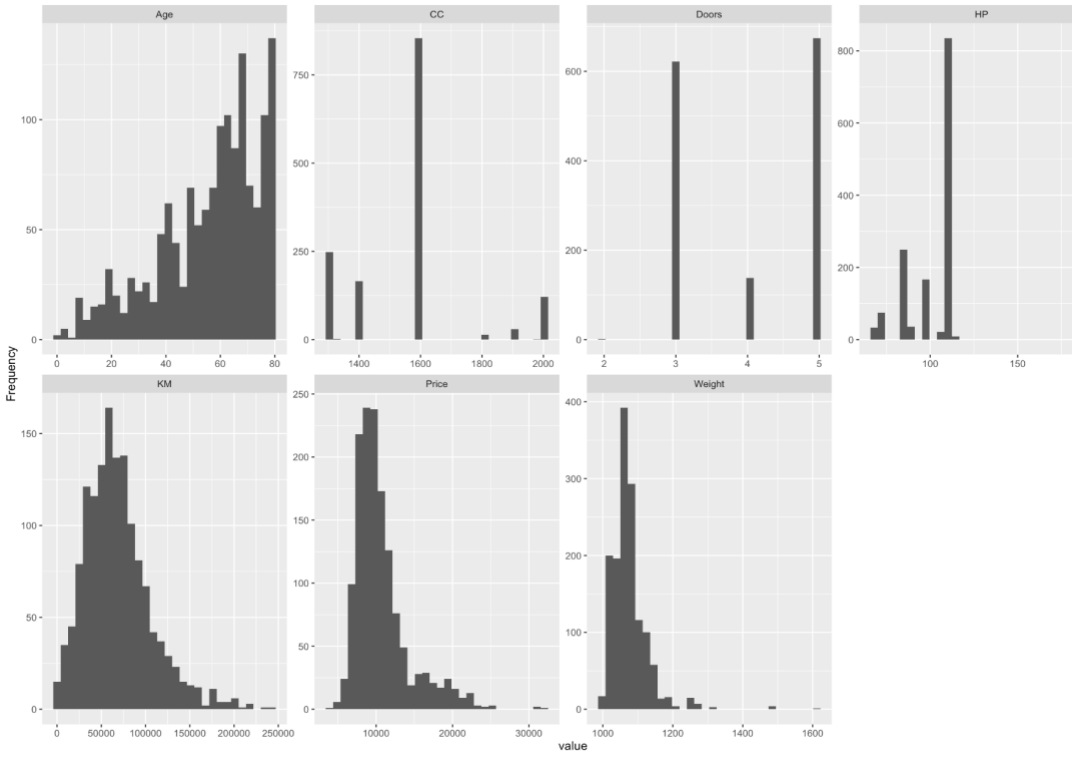
The summary statistics show that the average cost of a vehicle is 10,743, while the average cost is 9,500. From 4,000 to 24,500, there is a considerable price range. The dataset contains no missing values.

There are three categorical variables: Fuel Type, MetColor, Automatic

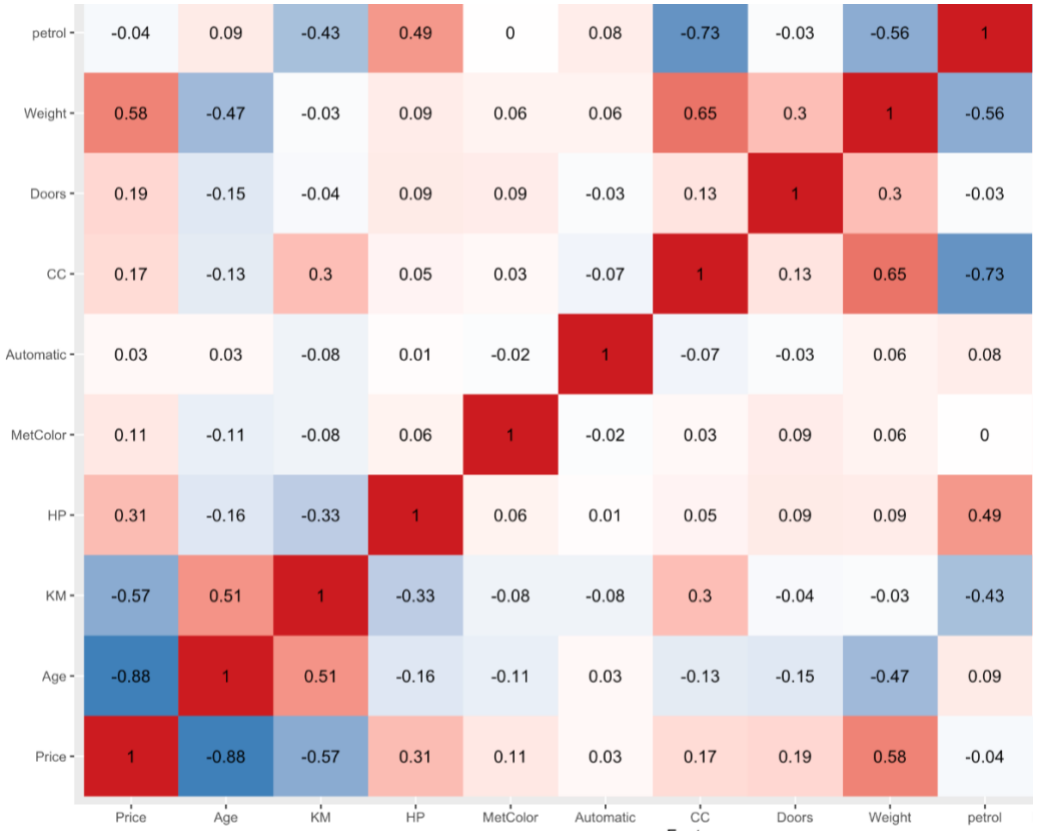


There are seven numerical variables: Price, Age, KM, HP, CC, Doors, Weight

Histogram



Most biased distributions of predictor variables are shown by histograms, which can cause problems in regression analysis. The price is also skewed to the right, as is evident. A histogram is also created to understand the price variable distribution. The histogram shows that the distribution is skewed to the right, suggesting there may be outliers in the data.



In the correlation matrix, we can see that *Price* has the strongest relation with *Age* (-0.88), *weight*(0.58) and *KM*(-0.57). This tells us that the age of the car, the weight of the car and the kilometers driven are major factors affecting the price of the car.

Data Splitting

We are separating the Toyota Corolla dataset into training and test datasets. This is a typical machine learning step where we use some of the data to train and some of the data to test our model. To ensure that the random division of the data is repeatable, we first set the seed to 7.

Then, we determine the number of rows in the dataset using the dim() function and store it in the variable n.

Next, we create a logical vector of length n with values randomly sampled from a uniform distribution between 0 and 1 using the runif() function. The threshold of 0.50 is used to divide

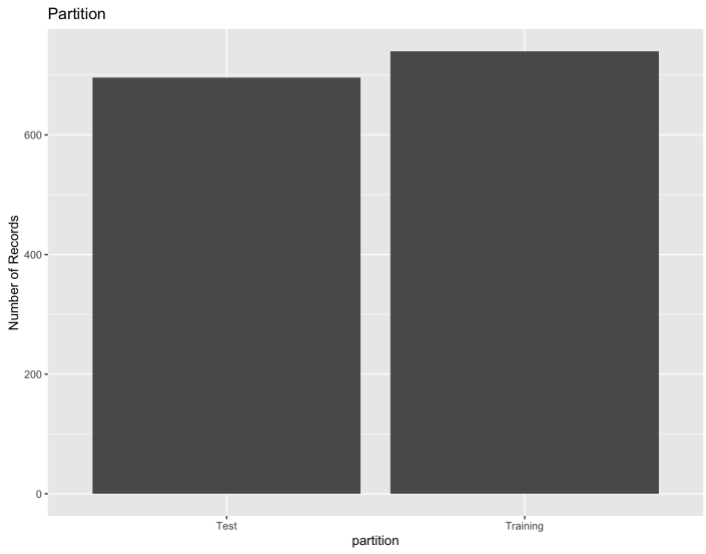
the data into two roughly equal parts. The indices where toyota_train is TRUE are used to subset the toyota dataset into train, while the indices where toyota_train is FALSE are used to subset the dataset into test.

Finally, we create a data frame called df with two columns: partition and num.records. The partition column specifies whether each row belongs to the training or testing set, and the num.records column shows the number of records in each set. We use the gg.plot2 package to create a bar chart that visualizes the partitioning of the data.

```
> set.seed(7)
> n<-dim(toyota)[1]
> toyota_train<-runif(n)<0.50
> train<-toyota[toyota_train,]
> test<-toyota[!toyota_train,]
> df <- data.frame(partition = c("Training","Test"),
+                  num.records = c(dim(train)[1],dim(test)[1]))
> table(df)
          num.records
partition  696 740
  Test       1   0
  Training   0   1
```

We used the table() function to show the counts of records in each partition. The output shows that the data has been split into roughly equal training and testing sets, with X records in the training set and X records in the testing set.

The plot below shows the data split:



Model Evaluation

Model 1 (Using all variables)

The summary statistics are obtained after running the regression model using all variables. The R-squared value of the model is 0.8571, which means that predictor variables can account for 85.71 percent of the variation in the price of a vehicle. We find that there are four statistically insignificant variables in this model (*MetColor, Automatic, Doors, petrol)*

```
Call:
lm(formula = Price ~ Age + KM + HP + MetColor + Automatic + CC +
    Doors + Weight + petrol, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-10618.6   -714.0     -4.1    761.7   6008.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.407e+03  1.827e+03  -4.601 4.95e-06 ***
Age         -1.205e+02  3.674e+00 -32.808  < 2e-16 ***
KM          -1.543e-02  1.807e-03  -8.538  < 2e-16 ***
HP           2.965e+01  5.538e+00   5.353 1.16e-07 ***
MetColor     6.489e+01  1.044e+02   0.622 0.534328
Automatic   -3.453e+01  2.149e+02  -0.161 0.872400
CC          -1.915e+00  5.532e-01  -3.463 0.000566 ***
Doors       -2.107e+01  5.408e+01  -0.390 0.696914
Weight       2.490e+01  1.616e+00  15.405  < 2e-16 ***
petrol       3.099e+02  4.042e+02   0.767 0.443590
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1320 on 730 degrees of freedom
Multiple R-squared:  0.8588,    Adjusted R-squared:  0.8571
F-statistic: 493.4 on 9 and 730 DF,  p-value: < 2.2e-16
```

Model 2 (Using Forward Selection Method)

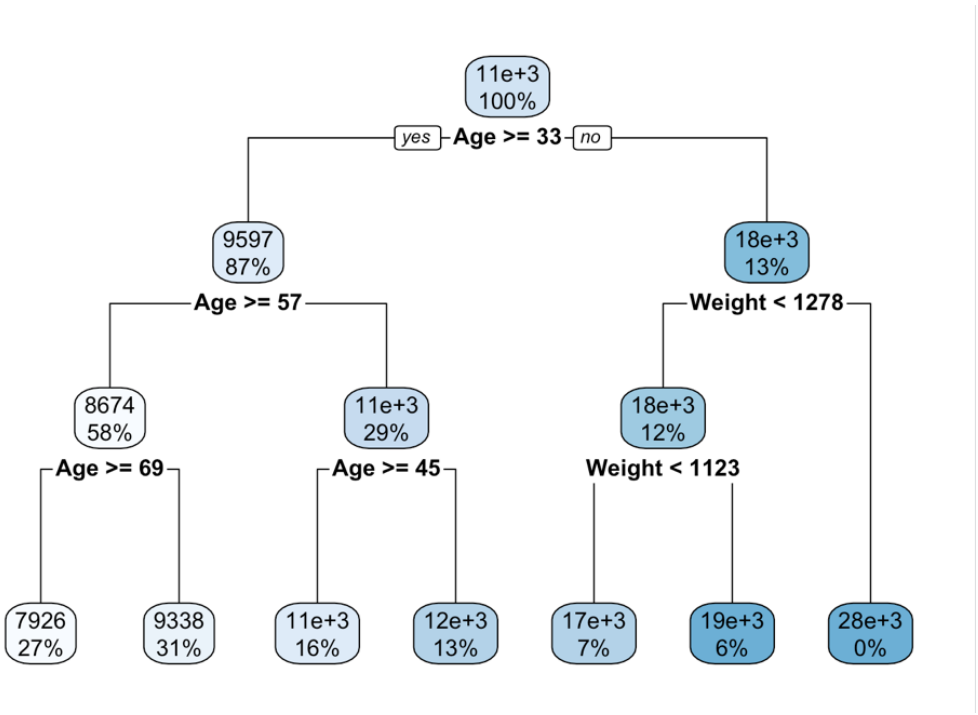
To increase the accuracy of the regression model, we can use model selection techniques to find the ideal subset of predictor variables to incorporate. To choose the best subset of predictor variables, direct selection is used in the provided code, which uses the "regsubsets" function of the "leaps" package. To use the Forward Selection Method, we set the method = “forward”.

```
model02<-regsubsets(Price~Age
                    +KM
                    +HP
                    +MetColor
                    +Automatic
                    +CC
                    +Doors
                    +Weight
                    +petrol,data=train, method='forward')
```

Age, mileage, power, engine displacement, weight, and fuel type are the predictor variables that the direct selection process reveals are present in the best model. The adjusted R-squared value of this model is 0.8645, slightly less than the value of the original model. The model only uses six predictor variables to minimize the possibility of overfitting and simplify the understanding of the
model.

```
         Age KM  HP  MetColor Automatic CC  Doors Weight petrol
1  ( 1 ) "*" " " " " " " "       " "       " " " "   " "    " "
2  ( 1 ) "*" " " " " " " "       " "       " " " "   "*"    " "
3  ( 1 ) "*" " " " " " " "       " "       " " " "   "*"    "*"
4  ( 1 ) "*" "*" " " " " "       " "       " " " "   "*"    "*"
5  ( 1 ) "*" "*" "*" " " "       " "       " " " "   "*"    "*"
6  ( 1 ) "*" "*" "*" " " "       " "       "*" " "   "*"    "*"
7  ( 1 ) "*" "*" "*" "*"       " "       "*" " "   "*"    "*"
8  ( 1 ) "*" "*" "*" "*"       " "       "*" "*"   "*"    "*"
> summary$adjr2
[1] 0.7477876 0.7878191 0.8372286 0.8522264 0.8554302 0.8575650 0.8574384 0.8572713
> which.max(summary$adjr2)
[1] 6
```

Decision Tree



We used the rpart and rpart.plot to create a decision tree for the model. We can see that the split in this tree. The price of the car is $19,000 when the age is less than 33 years and the weight is greater than 1123kgs, this is 6% of the data. Similarly, the price of the car is $17,000 when the age is less than 33 years and the weight is less than 1123kgs, this is 7% of the data.

Conclusion

In conclusion, this project involved the exploration, analysis, and modeling of a dataset of Toyota Corolla cars. The data exploration phase involved examining the dimensions of the dataset, creating visualizations to examine the distribution of the target variable (Price), and using the DataExplorer package to perform an automated exploratory data analysis.

After creating a dummy variable for FuelType, a multiple linear regression model was built using various predictors, including Age, KM, HP, MetColor, Automatic, CC, Doors, Weight, and petrol. The model was then split into training and testing sets, and the model was fitted using the training set. The model was then further refined using forward selection, and the final model was tested on the testing set, achieving a root mean square error (RMSE) of 1293.

Finally, a classification and regression tree (CART) model was built using the same predictors as the linear regression model. The CART model produced a decision tree that shows the most significant predictors for predicting the price of Toyota Corolla cars.

Overall, this project demonstrated the usefulness of data exploration, model building, and evaluation in understanding and predicting the price of Toyota Corolla cars. The methods and techniques used in this project can be applied to other datasets, providing insights into a wide range of research questions and business problems.

Key References

Chang, H. (2021, June 24). *What's your car worth? - now with linear regression and correlation!* Medium. Retrieved April 18, 2023, from https://towardsdatascience.com/whats-your-car-worth-part-2-ee0500d5c997

D'Allegro, J. (2022, December 19). *Just what factors into the value of your used car?* Investopedia. Retrieved April 18, 2023, from https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp