Math-331 Project 2

Malaika N

About the Dataset

- My project is based on an automobile company that wants to understand the factors on which the pricing of cars depends
- Specifically determine the factors affecting the pricing of cars in the American market using a large dataset of different types of cars across the American market
- I imported the Car Price Prediction dataset from Kaggle: https://www.kaggle.com/code/goyalshalini93/car-price-prediction-linear-regression-rfe/data

Objectives

1. Which variables are significant in predicting the price of a car?
2. How well those variables describe the price of a car?

Libraries Imported

- library(ggplot2)
- library(tidyverse)
- library(dplyr)
- library(stringr)
- library(class)
- library(lubridate)
- library(splines) -> Regression spline functions and classes.
- library(mgcv) -> Generalized additive models.
- library(randomForest) -> Model the training and test sets
- library(rpart) -> Used for building classification and regression trees
- library(rpart.plot)

Cleaning the dataset

- I stored the dataset in the dataframe **carPrice**
- After viewing the dataset it was evident that **price** was the response variable and as it is a numerical variable I would be applying **regression models** and not classification
- I removed the column **car_ID** as it does not contribute to the analysis
- I further split the **CarName** column into **Car Company** and **Car Model** to make it easier to make predictions

```
```{r}
carPrice<- (carPrice %>%
  separate(CarName, c('CarCompany', 'CarModel'),' ')%>%
    drop_na())
```
```
```
```

Further Cleaning

- Cleaning up unique values
  - I found some spelling errors in the **Car Company** column that need to be corrected
- Changing the data types of some of the variables to factors to be able to analyze and manipulate them correctly

```
```{r}
carPrice$symboling <- as.factor(carPrice$symboling)
carPrice$CarCompany<-as.factor(carPrice$CarCompany)
carPrice$fueltype<-as.factor(carPrice$fueltype)
carPrice$aspiration<-as.factor(carPrice$aspiration)
carPrice$doornumber<-as.factor(carPrice$doornumber)
carPrice$carbody<-as.factor(carPrice$carbody)
carPrice$drivewheel<-as.factor(carPrice$drivewheel)
carPrice$enginelocation<-as.factor(carPrice$enginelocation)
carPrice$enginetype<-as.factor(carPrice$enginetype)
carPrice$cylindernumber<-as.factor(carPrice$cylindernumber)
carPrice$fuelsystem<-as.factor(carPrice$fuelsystem)
```
```

Further Cleaning

I wasn't sure what the symboling variable meant so I made a plot to see how the symboling and price were related:

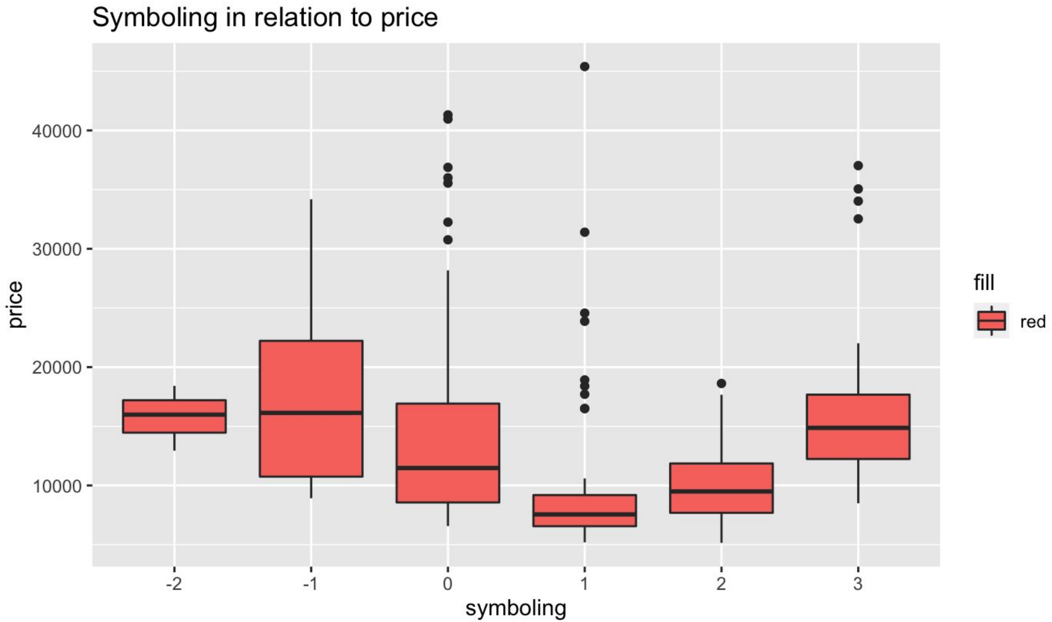
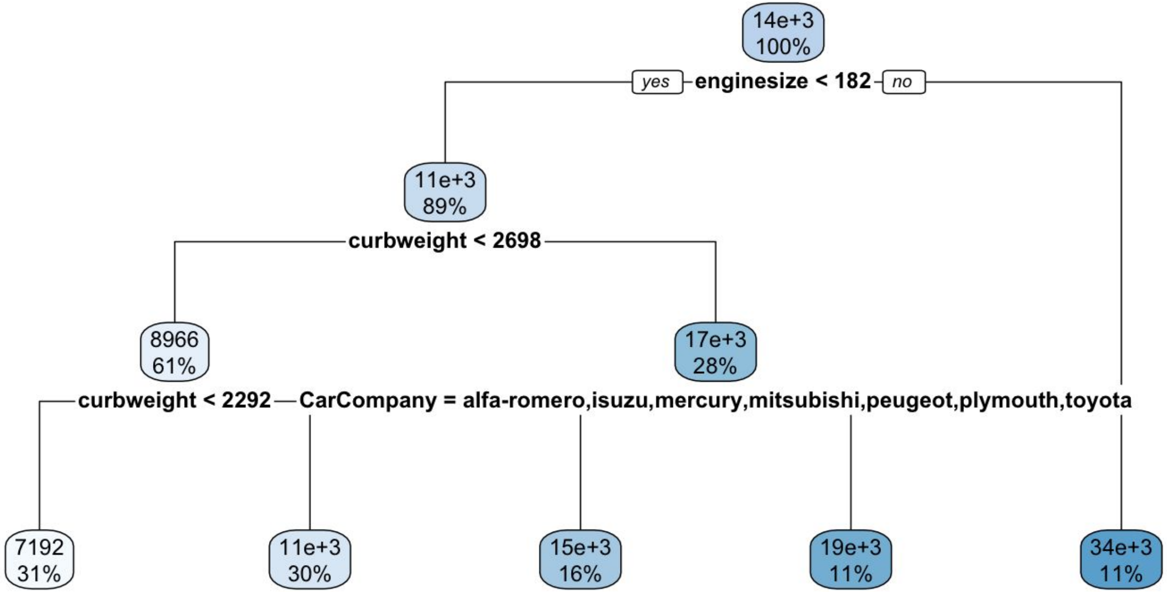
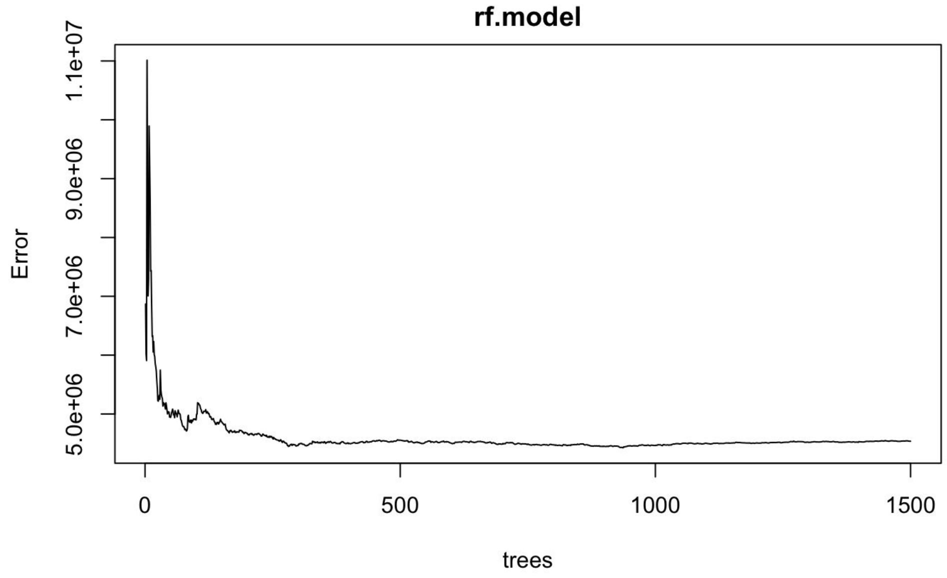


Figure 1

- **-1** symboling have the highest price
- **3** has the price range similar to **-2**.
- There is a dip in price at value **1**

Training the model

- Since we're creating a predictive model we need to **split the data into a train and test set**

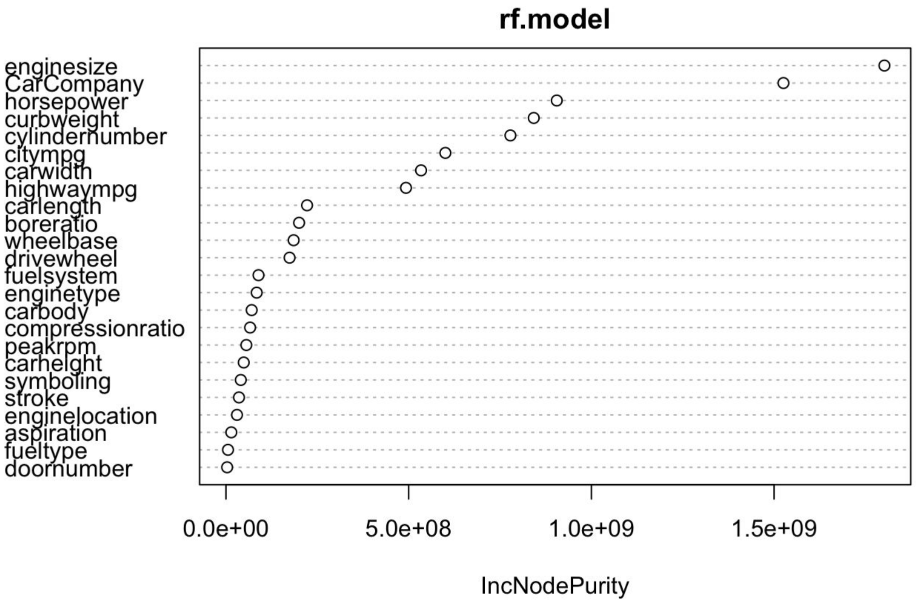
```r
```{r}
n <- round( 0.6 * nrow(carPrice)) #no of rows for the sample set=60%
  in_train <- sample(1:nrow(carPrice),n) #sample not the rows directly but draw a sample of the row numbers
#select the rows and columns
  train <- carPrice[in_train,]#rows
  test <- carPrice[-in_train,]#not the rows
  rm(carPrice,in_train)#remove extra variables added and remove carPrice since we've split our dataset
  #train2<-train %>% select(-name,-category)

#write sets to files
```
```

Decision Trees



- I wasn't able to determine which variables were specifically affecting the price the most by just looking at the table
- Hence I included all the variables to visualize the data and classify

Linear Regression Model

- I summarized the train set, and use that information to determine which factor variables are important in predicting the response

```
symboling      CarCompany   CarModel          fueltype   aspiration doornumber        carbody   drivewheel enginelocation
-2: 2     toyota    :20   Length:122        diesel: 17   std  :99   four:68    convertible: 4   4wd: 5     front:119
-1:12     honda     : 9   Class :character  gas   :105   turbo:23   two :54    hardtop    : 6   fwd:64     rear :  3
0 :44     mitsubishi: 9   Mode  :character                                     hatchback  :37   rwd:53
1 :29     nissan    : 9                                                        sedan      :58
2 :20     peugeot   : 9                                                        wagon      :17
3 :15     volkswagen: 8
          (Other)   :58
```

- Technically since **enginelocation** has only 2 levels it's not important however when included in the linear regression model it has a high significance
- When creating the linear model I used **price as the response variable** and all others as predictors
- I then summarized the linear model, and use that information to determine which variables are most important in predicting the response

Linear Regression Model

Train Adjusted R^2 = 0.9884	Test Adjusted R^2 = 0.9863

- The **P-Value is small** hence this establishes that there exists a relationship
- The **R^2 value is extremely high**
- Variables most important in predicting the response:
    - CarCompanybmw
    - CarCompanypeugeot
    - Carbodyhatchback
    - Carbodysedan
    - Carbodywagon
    - Enginelocationrear
    - Enginetypel
    - Enginetyperotor
    - Enginesize
    - boreratio

Random Forest Model

- Training the dataset using the random forest model.
- This model builds decision trees on different samples and takes their average in case of regression and provides the highest accuracy.




```
```{r}
rf.model <- randomForest(price ~ .-CarModel, data=train,
                         ntree=1500,
                         mtry=5)

predictions <- predict(rf.model)
```
```

- Prediction accuracy on the **train set = 4536610**
- Assess on the **test set = 4508604**

Variable Importance Plot

- **Visualizing the rf.model** (random forest model) using a Variable Importance Plot
- The feature importance describes which **features are relevant**



- Using **cor** to **compute the correlation of x = 0.9068045**
- **Significant variables** after Visual analysis:

- Engine Type
- Fuel type
- Car Body
- Aspiration
- Cylinder Number
- Drivewheel
- Fuel Economy
- Curbweight
- Car Length
- Car width
- Engine Size
- Boreratio
- Horse Power
- Wheel base

Further Exploration

- Conduct a bivariate analysis and apply the model to certain variables
- Do a cluster model and table on some variables to see how they affect the dataset
- Further explore how the *ntree* and *mtry* values affect the random forest model accuracy