Project 2 Report
MATH- 331
Malaika N

I chose to analyze the "Car Price Prediction" dataset which contains various information about different car models such as the car name, engine type, door number, engine location, price, etc. The main objective of this project is to create and compare several distinct models based on the information in the dataset and analyze the relationship of a significant response variable to predictors. After briefly viewing the dataset it was evident that price would be the response variable and since price is numerical I would be using regression models and not classification ones. I broke my analysis process into steps with the first one being determining the response variable. Next I imported the libraries I would need to model the data with and perform other calculations along the way. Once I had done that I imported the dataset and stored it under the name **carPrice** and began cleaning the data.

Cleaning the data was a slightly more difficult since this was a slightly larger dataset and I had to make sure there were no errors or NA values. I removed the car_ID column as it only contained the serial count of the row and hence wasn't important in my analysis. I then separated the CarName column into two distinct columns- Car Company and Car Model, to make it easier to visualize and understand the data. To check whether there were any errors within the dataset I used **unique()** to detect any unique values and found a lot of spelling errors within Car Company and Car Model that I then fixed. I then changed the data types of variables that I determined would work better as factor variables.

I wasn't able to make much sense of the **symboling** variable specifically in how it affected the rest of the dataset and mainly the price. Hence I created a box plot to establish a relationship between **symboling and price**. Through the box plot I learned that symboling significantly affects the price as cars with a **1** in symboling are the cheapest, cars with a **3** have a price range similar to **2** and **-1** symboling have the highest price.

I then moved onto creating my train and test sets for the models. It's important to create a train set and apply the model to it first to help the model generalize and make accurate predictions on the test set (data it has never seen before) rather than have the model overfit the data. I chose to begin with a **Decision Tree** model as they are usually easy to interpret and visualize due to being set as an if else statement. The output of the decision tree was easily

understandable as it not only showed me which variables are most important in detecting the price (enginesize, curbwieght, CarCompany) but exactly which car models have the highest price and grouped the cars according to a price range.

I used a linear regression model next and found the variables that are most and least important in predicting the price of a car model such as which engine location,front/rear, etc. and this model had a high R^2 value of 0.9884 for the train set and R^2 value of 0.9863 for the test set hence the model was highly accurate.

I then applied the Random Forst Model onto my training set and then my test set. I tried the model with different ntree and mtry values and settled on ntree =1500 and mtry =5. The model prediction for price was accurate as for the training set the prediction accuracy = 4536610 and the test set = 4508604. I also checked the corelation of x which was significant at 0.9068045.

I also did a Variable Importance Plot to see the important variables and the highest ones which was very helpful as visualizing it made it a lot easier to understand. Car Company and enginesize were the variables with the highest importance scores and hence the ones that give the best prediction and contribute most to the model. This was interesting as the highest variable slightly differed with every model however they were all included in the top contributions to the response variable.

This project was a lot more interesting as I got to focus on using models and see the relationship between the response variable and other predictor variables. Since it wasn't evident from the table which were the most significant, I used all the variables (other than price) as predictors variables. I did run into some errors mainly with my linear model detecting NA values although there were none and had to look up how to fix it. After comparing the models I can determine that they all had a high accuracy with Random Forest having a lower error rate however the linear regression model had an extremely high R^2 value. For further exploration I would try going into more depth with each variable to see how they individually affect each other before seeing how they affect price and which need to be removed from the dataset completely. I would also like to see other models that can be used with this data and try applying classification models by changing the response variable just to see the results from the existing data.