Capstone Project: Used Car Price Predication and Classification

by Casey Liu

Project Goal

I have been hired by an used car selling company to build a model that can predict the price of used cars in British market based on 9 features provided.
For Mercedes C Class & Ford Focus cars which are two of the most popular cars our company sells, I also want to make a classifier what model of the used car is from.

Data Gathering & Cleaning

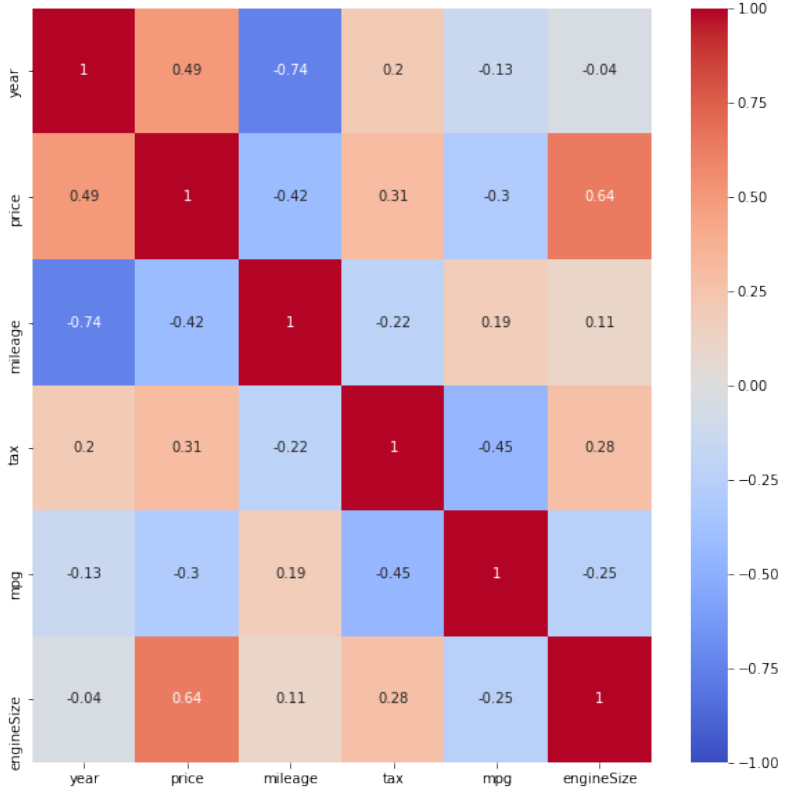
Data was collected from the 100,000 UK Used Car Data set (https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes).
There are 9 different data sets corresponding to each car manufacturer.

The 9 data sets are clean. They don't have duplicate or NaN values. I dropped the cars which have very old age and the wrong age. I combined the 9 data sets in one file and added a brand column to reflect the manufacturer information.

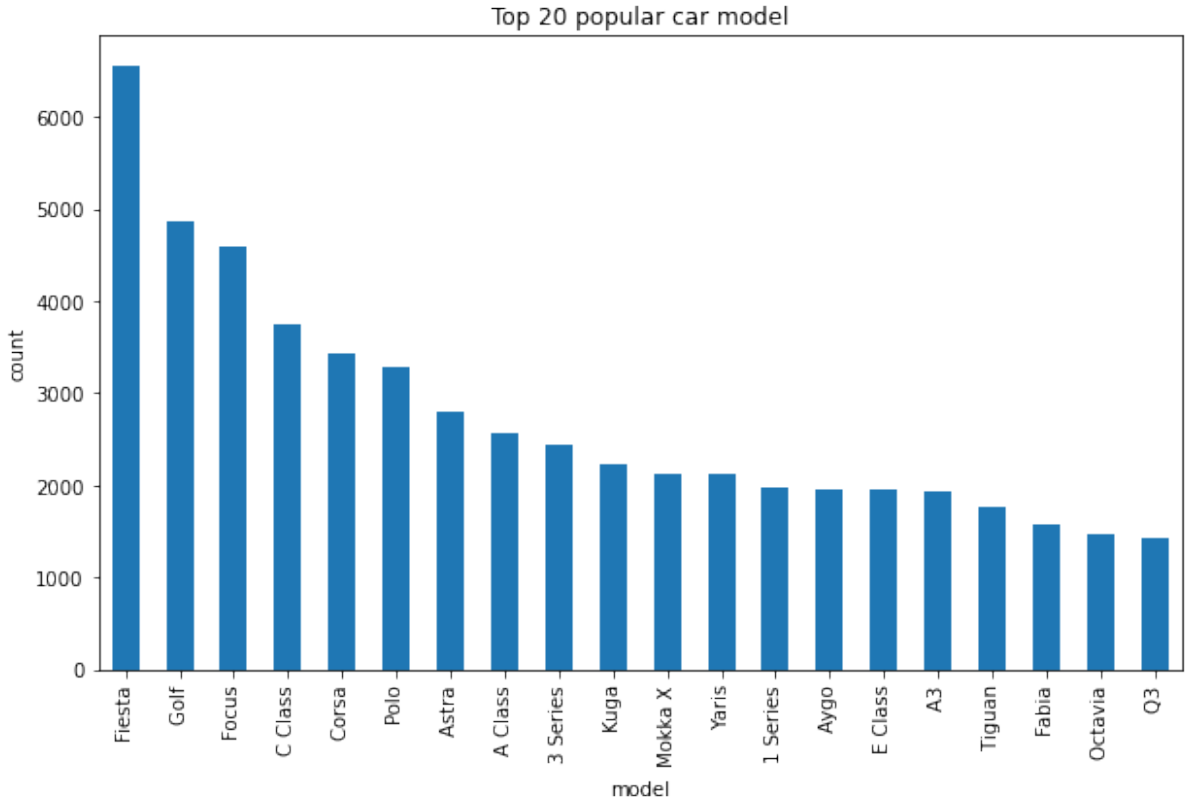
- Model : Model of the car
- Year : Registration year
- Price : Price of the car
- transmission : Type of gearbox used
- fuelType : Type of fuel used
- tax : Tax applied
- mileage : Distance the car has travelled
- mpg : Miles per gallon
- engineSize : Size of the car engine

Exploratory Data Analysis

I created a heatmap that displays the correlation between car price and all other features.

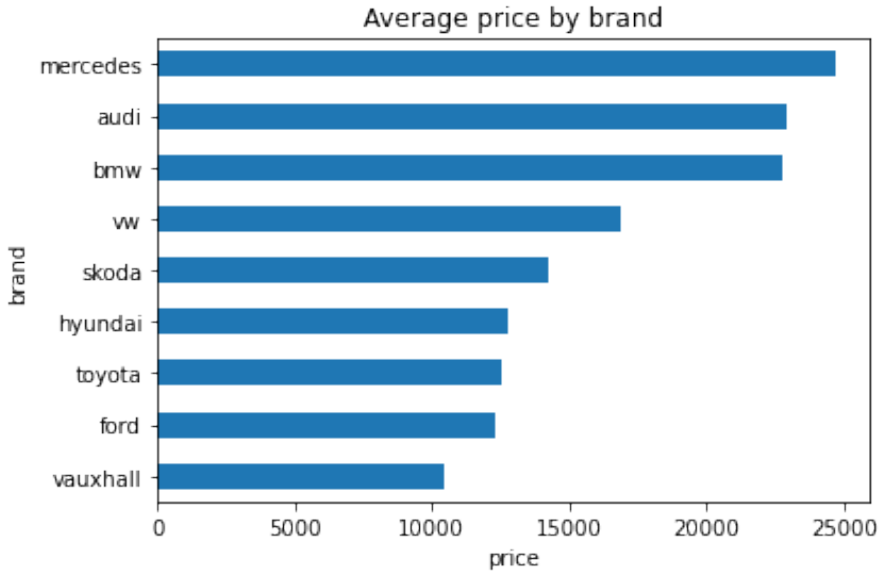


Top 20 popular car model



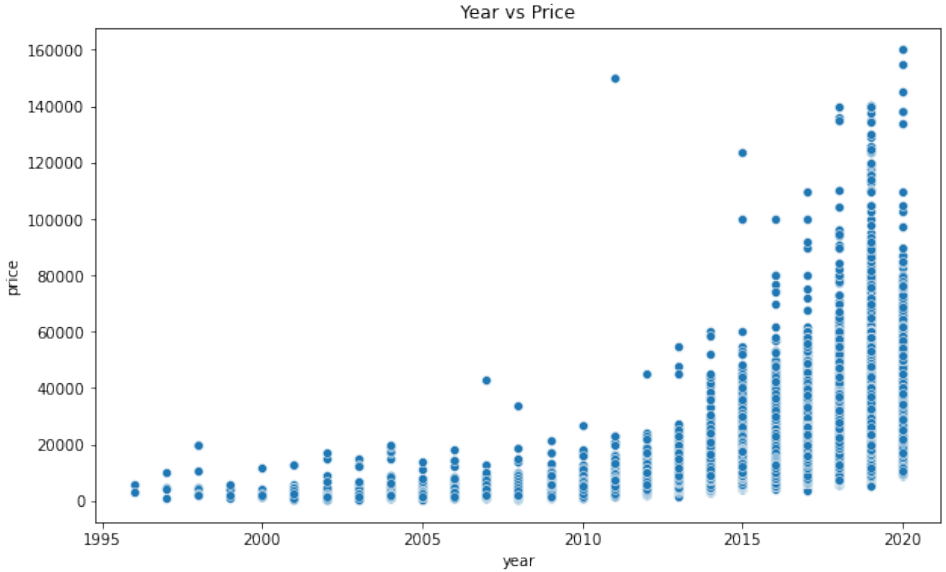
- Fiesta 6556
- Golf 4863
- Focus 4588
- C Class 3747
- Corsa 3441
- Polo 3287
- Astra 2805
- A Class 2561
- 3 Series 2443
- Kuga 2225
- Mokka X 2127
- Yaris 2122
- 1 Series 1969
- Aygo 1961
- E Class 1953
- A3 1929
- Tiguan 1765
- Fabia 1571
- Octavia 1477
- Q3 1417

Average price by brand



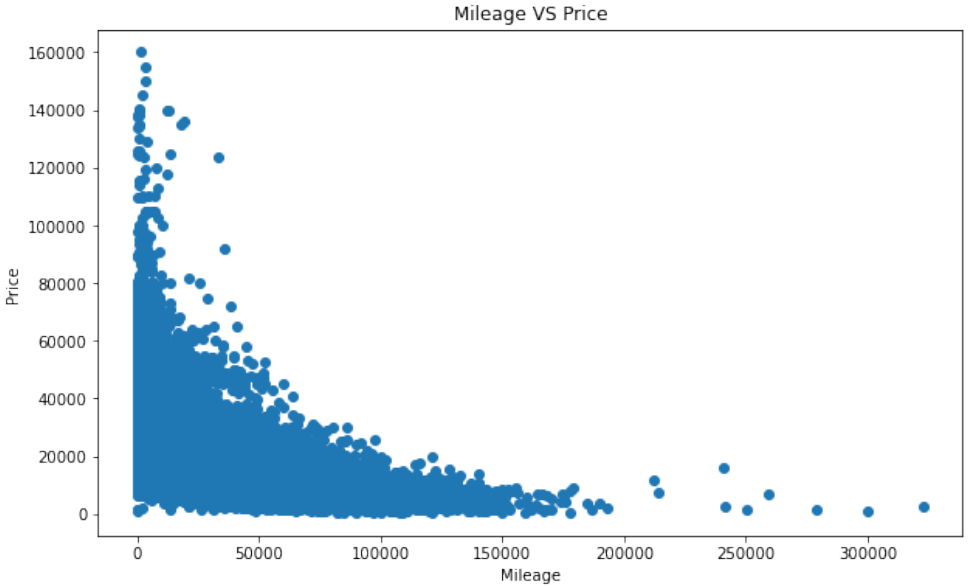
Mercedes, Audi and BMW are the most expensive brand while Toyota, Ford and Vauxhall are cheapest.

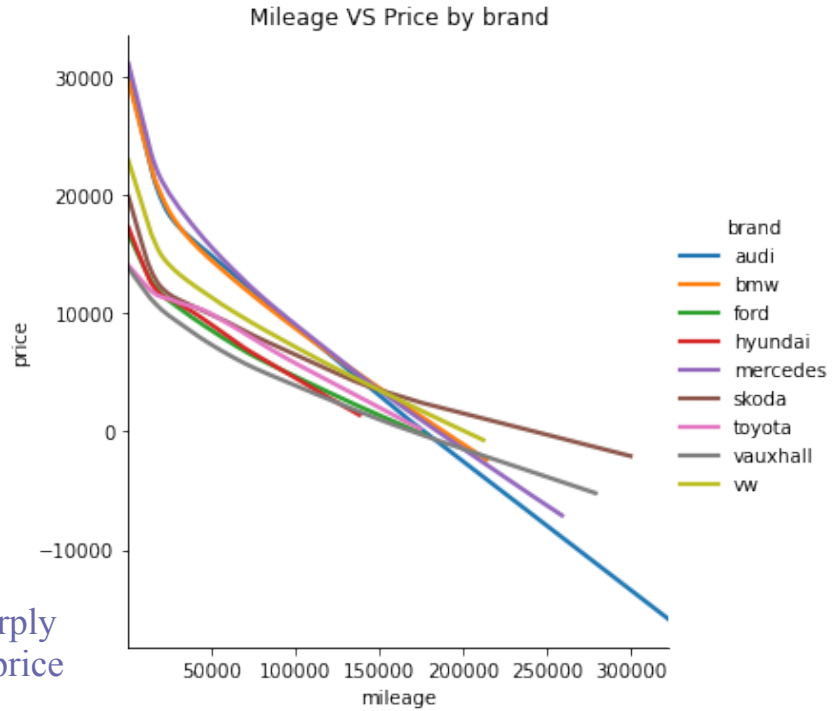
Year vs Price



The older the car, the lower the price.

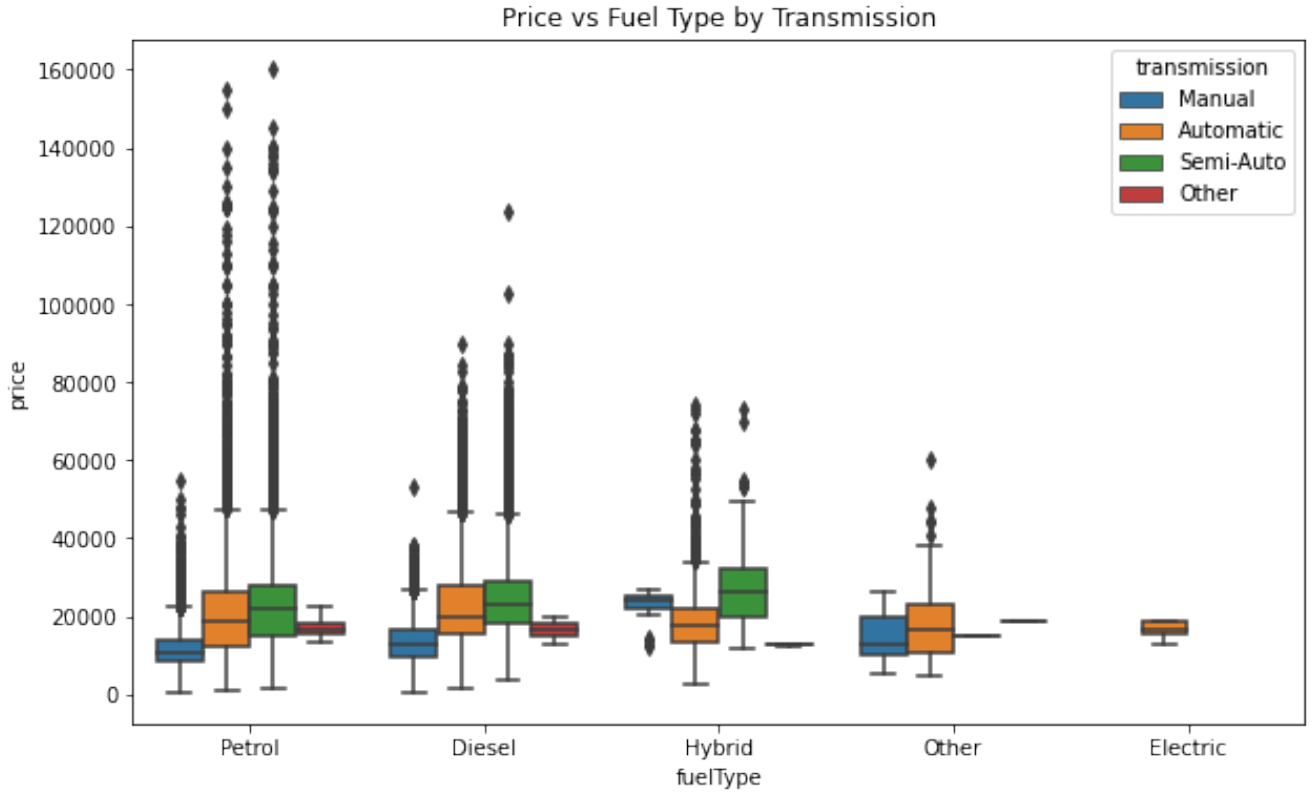
Mileage VS Price





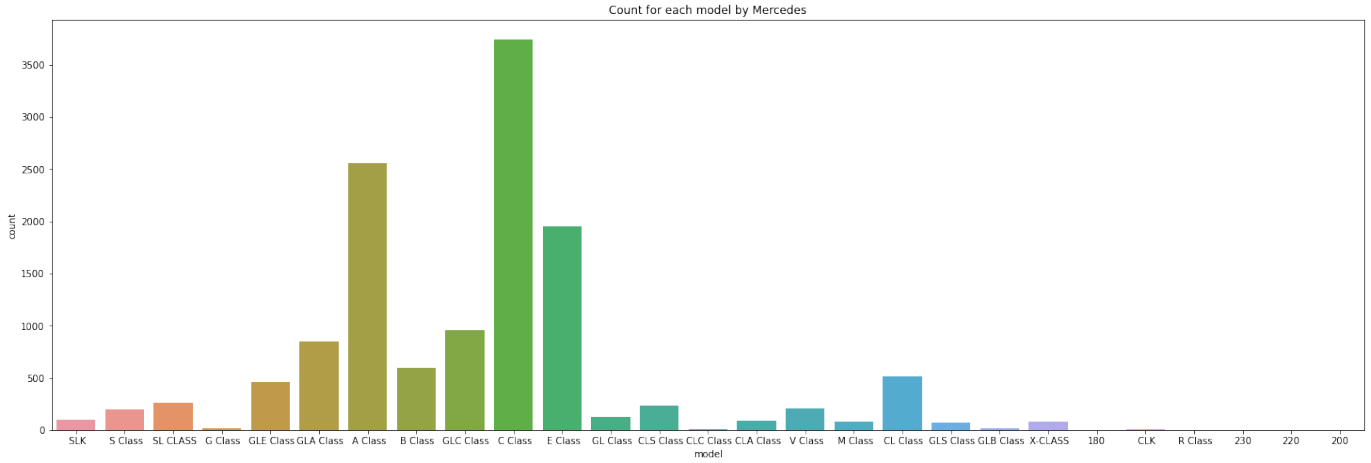
As the mileage increases, the price decreases. The Audi sharply decreases the price. At the same time, Skoda decreases the price slower compared to other brands.

Price vs Fuel Type by Transmission

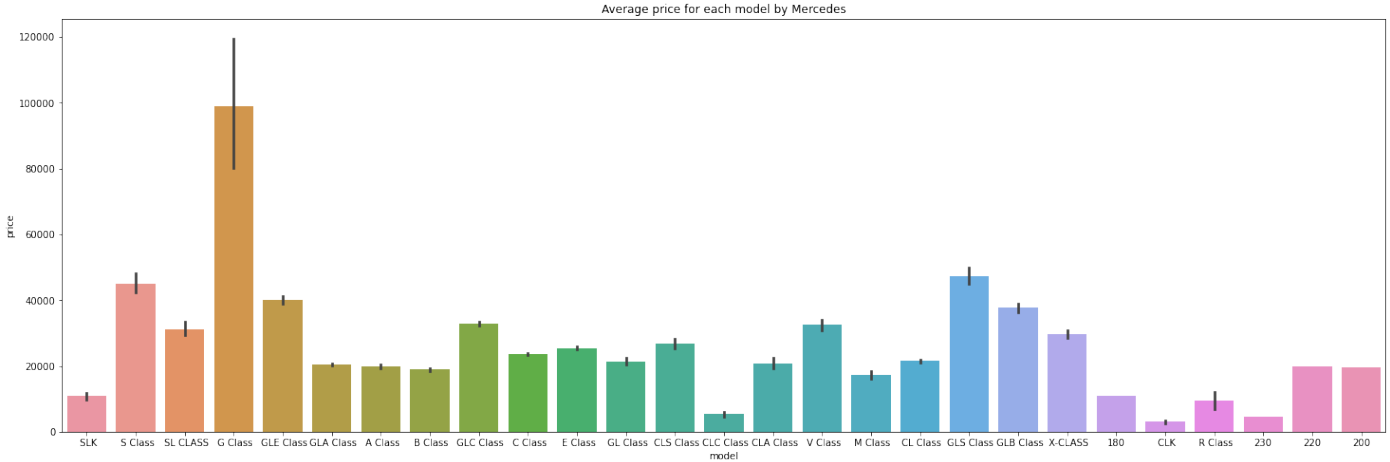


Mercedes

Count for models

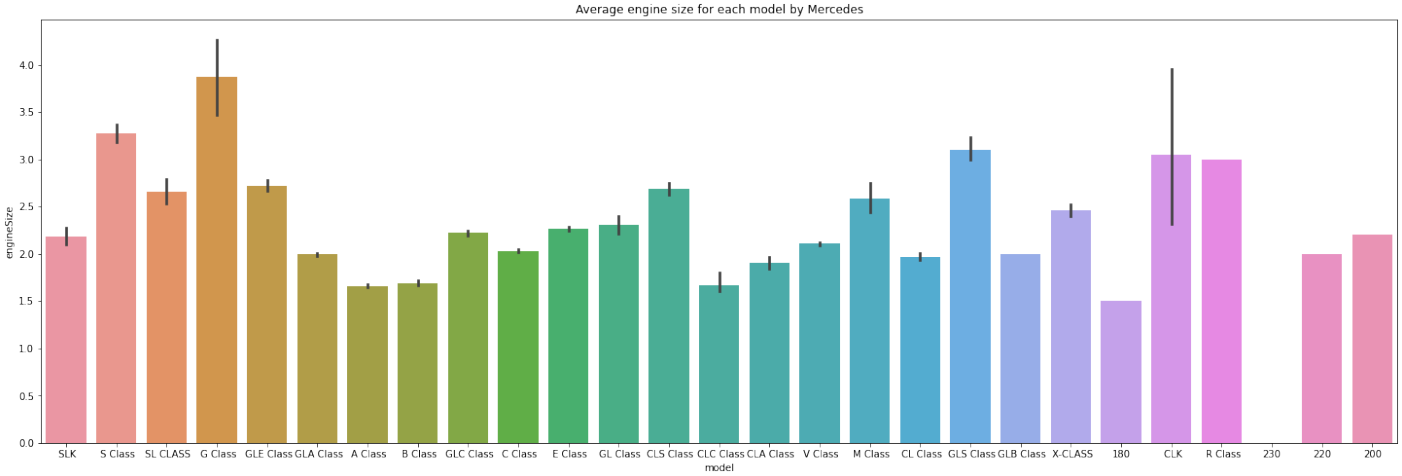


Average price

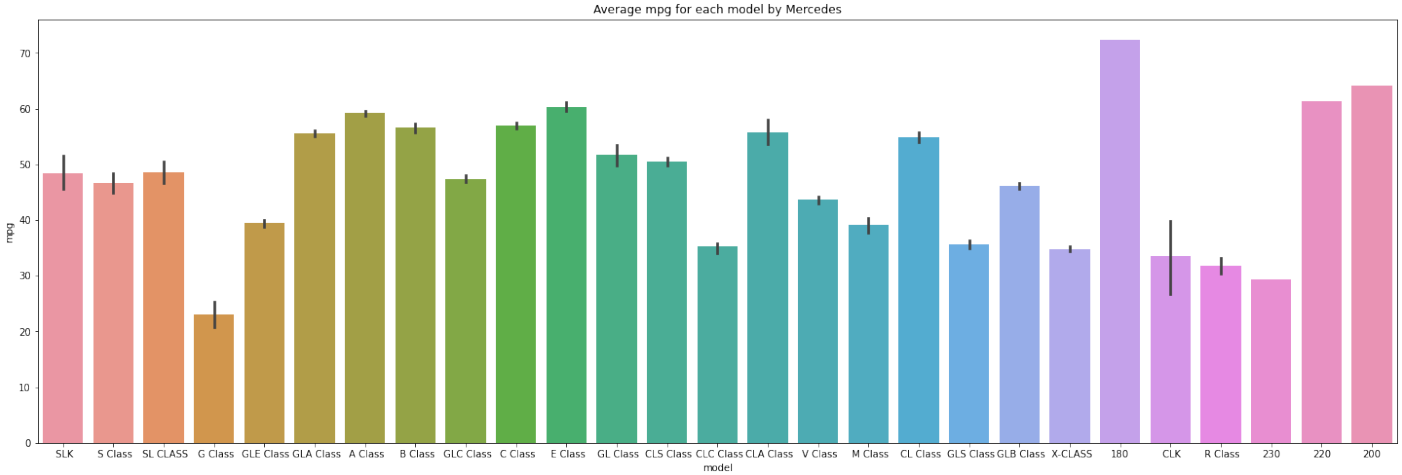


Mercedes

Average engine size



Average mpg



Regression Models

- Use different regressors in order to predict the used car price based on all of the features

- Linear Regression
- Ridge
- Lasso
- Decision Tree
- Random Forest
- AdaBoost

Regression Models

Model	R2 Score
Linear Regression (no dummies)	0.7036
Linear Regression	0.8631
Ridge	0.8628
Lasso	0.8600
Decision Tree	0.9410
Random Forest	0.9609
AdaBoost	0.2005

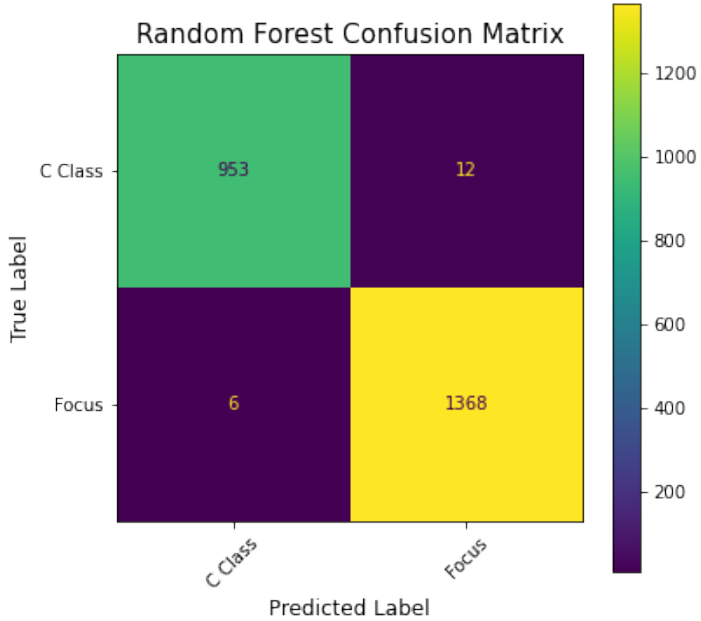
Classification Models

- Do classifier models to predict the Mercedes C Class & Ford Focus cars.

- Logistic Regression
- KNN
- Decision Tree
- Random Forest
- AdaBoost
- Support Vector Classifier

Classification Models

Model	GridSearch Best Score
Logistic Regression	0.9137
KNN	0.9693
Decision Tree	0.9882
Random Forest	0.9906
AdaBoost	0.9843
Support Vector	0.9078



Conclusion & Recommendation

The Linear Regression Model got 86% accuracy. The Ridge and Lasso didn't help improve the Linear Regression Model. The Random Forest Model preformed best with accuracy of 96%. All the classification models performed well. The Random Forest Model worked best with the score of 0.9906.

If we want to expand the data set and perform the models better, the recommendation is to add more information, such as the damage information about the used car, car service records, car add-ons and color of the car.

Thank you