Angela Boakye Danquah, Farhan Zaman, Zhongxuan Wang
STAT 4630
Fall 2020
Dr. Jeffery Woo

A Statistical Analysis of the UK Used Car Market

Section 1: Executive Summary

The data set used in this analysis was obtained from Kaggle.com. It is a public data set that contains information about 100,000 used cars from the United Kingdom, and it contains 9 data sheets corresponding to different car brands and manufacturers. The car brands are Audi, BMW, Ford, Hyundai, Mercedes-Benz, Skoda, Toyota, Vauxhall, and Volkswagen. For each of the 9 brands of vehicle analyzed in this project, each data sheet contains values for 9 variables: the model of the car, year of the model, retail price, transmission type, mileage, fuel type, engine size in litres, road tax, and miles per gallon. The total number of observations analyzed throughout the data set after compiling the different car manufacturer's distinct data sheets and entries is 99,188.

Questions of Interest

1. Regression Problem: Which features of a car most greatly affect its retail price?
2. Classification Problem: Which features of a vehicle most greatly contribute in differentiating between brands, specifically BMW and Ford?

For our first problem, we seek to determine which feature of a car most greatly affects its resale price. As all the cars in our data set are used, the price that they are sold for varies more greatly than with new cars. We ask this question since most people will own a car at some point in their life. If a person buys a new car and is planning on reselling it, they may want to consider the features of the car that will give them the highest resale value. If a person buys a car used, they may care to know about those features as well since it will affect the price at which they buy the car and may potentially affect the price at which they can resell if they choose to. To answer this question, we created regression models to firstly identify if any of the features influence the price (if the features in our data are relevant to our question). We then used regression to identify the features with the largest influence on price.

Our second question seeks to ascertain whether we can differentiate between car brands given the features in our data set. To do this we will isolate BMW and Ford data and analyze them. We chose the car brands BMW and Ford since they are brands that cater to different consumer bases. We ask this question to discover whether or not there truly is a difference between the performance of normal and luxury brands since luxury car brands generally promote themselves as being superior performance-wise. To answer this question, we built different classification models, logistic regression, LDA, and classification trees to attempt to classify our observations as either Ford or BMW. We then analyzed these models to determine which feature played the largest role in differentiating between the two car brands.

Section 2: Data Processing & Cleaning

To start off our regression analysis, we had to perform some slight cleaning to our data set. We combed through our data and removed the vehicles that had a year greater than 2020 as we wanted to focus on 2020 models and prior. For our classification analysis and classification tree, in addition to removing the vehicles with a year greater than 2020, we changed the response variable of "Make" (the brand of the car) into a factor. This resulted in two classes, BMW and Ford. We then had to do the most cleaning for our regression tree as we had to change the two categorical variables amongst our 7 predictors into binary variables. Both transmission and fuel type have "other" as one of the classes, so we decided to drop "other because "other" doesn't contain much information about the car. After removing "other" from the data set, we decided to use "manual" as the reference class for transmission as after analyzing the data, we found that "manual" has the lowest average price among the 3 types of transmissions. Similarly, we chose "petrol" as the reference class for fuel type, as we found that "petrol" has the lowest average prices. We also selected Audi as our reference class for the variable 'make' for our regression models.

Section 3: Regression

3.1 Exploratory Data Analysis

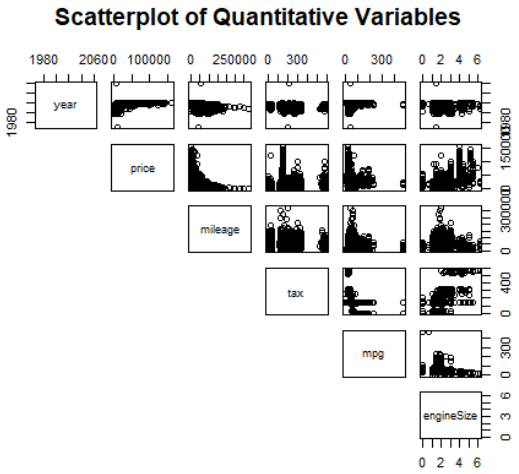
Here we present graphical summaries of some interesting variables used in our regression models.

Histograms of Price

The histogram for Price is skewed to the right. This is to be expected since the prices of cars are around the same level with the exception of luxury cars, which tend to be priced higher. The histogram also shows a large variance in price which we also expected since used cars vary greatly in price based on their condition.

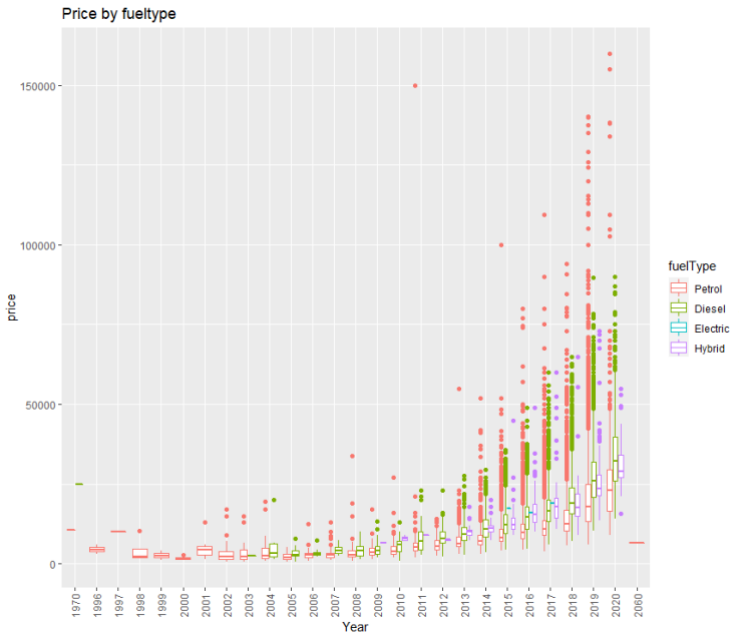
Scatterplots of Quantitative Variables

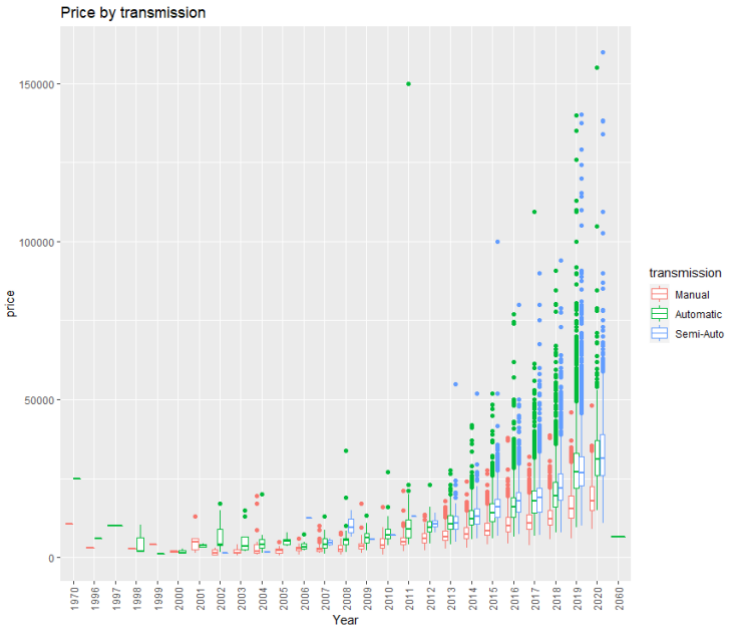
We created scatterplots for all quantitative variables because we would like to see the relationship between all the quantitative variables and the response variable “price”. No one variable appears to have a linear relationship with price, however this may be because we have very observations or because these variables truly do not have a relationship with price. We also observe other interesting non-linear relationships between our quantitative predictors.



Boxplots of fuel type and transmission

Boxplots of price by transmission and price by fuel type are created to determine which class in the two categorical variables should be used as the reference class in our regression model.



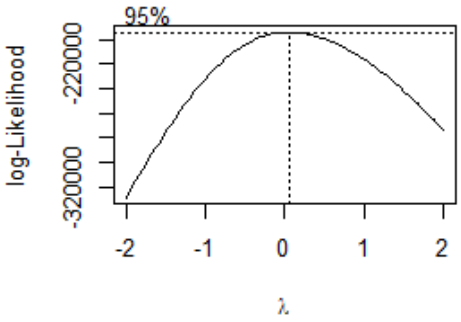


The boxplots show that petrol cars and automatic and semi-automatic cars tend to be priced higher and hybrid and manual cars tend to be priced lower.

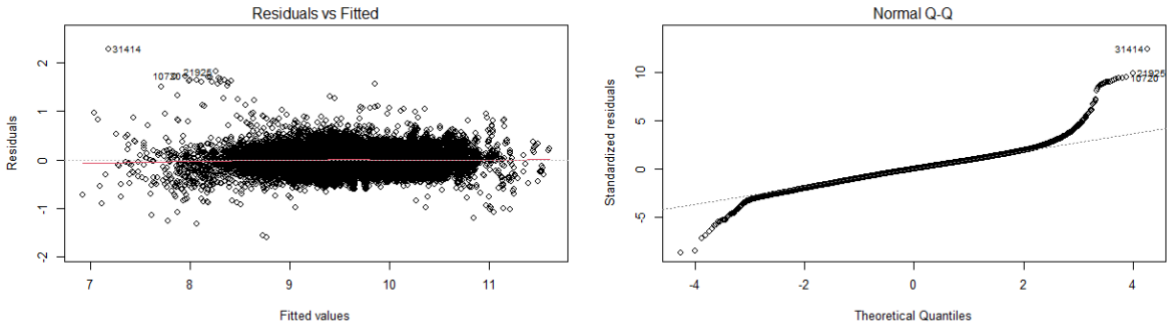
3.2 Regression Model

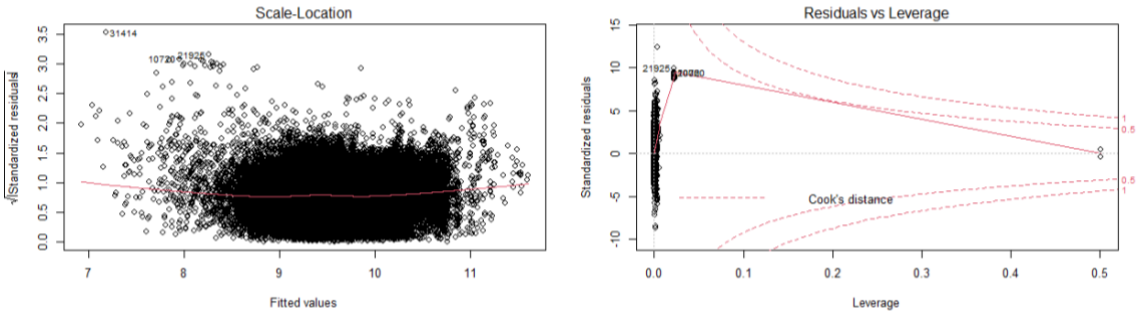
To figure out how to transform the response variable, we used the boxcox plot. According to the boxcox plot, $\lambda = 0$, so we transformed our response variable price into log(price).

\



After transforming the response variable, we also added car brands (in the data set, it's called "make") into our improved regression model. In addition, we excluded outliers. Then we regressed log(price) on make, year, transmission, mileage, fuel type, tax, mpg, and engine size. We created residual plots and the Q-Q plot to check if the assumption of our model is met. The red line in the first graph is close to 0 after our transformation so the mean zero assumption for the error term is now reasonable and the vertical spread of plot in the 3rd graph seems consistent. The assumptions of the model seem reasonable and there are no influential outliers.





From the results of our model, it is shown that all our predictors are statistically significant. According to the last diagnostic plots, the assumptions of mean zero assumption for the error term and constant variance are reasonable in the model.

```
Call:
lm(formula = log(price) ~ make + year + transmission + mileage +
    fuelType + tax + mpg + engineSize, data = training2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.59500 -0.11329  0.00301  0.11426  2.29014

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -2.231e+02  1.213e+00 -183.95   <2e-16 ***
makebmw               -1.161e-01  3.635e-03  -31.95   <2e-16 ***
makeford              -2.647e-01  3.429e-03  -77.20   <2e-16 ***
makehyundi            -3.903e-01  4.689e-03  -83.25   <2e-16 ***
makemercedes benz     -3.642e-02  3.480e-03  -10.46   <2e-16 ***
makeskoda             -2.964e-01  4.252e-03  -69.71   <2e-16 ***
maketoyota            -4.327e-01  4.558e-03  -94.94   <2e-16 ***
makevauxhall          -4.686e-01  3.608e-03 -129.89   <2e-16 ***
makevw                -1.690e-01  3.369e-03  -50.16   <2e-16 ***
year                   1.153e-01  6.007e-04  191.88   <2e-16 ***
transmissionAutomatic  1.366e-01  2.673e-03   51.12   <2e-16 ***
transmissionSemi-Auto  1.412e-01  2.565e-03   55.04   <2e-16 ***
mileage               -4.873e-06  6.144e-08  -79.31   <2e-16 ***
fuelTypeDiesel         6.682e-02  2.343e-03   28.52   <2e-16 ***
fuelTypeElectric       1.842e+00  1.343e-01   13.72   <2e-16 ***
fuelTypeHybrid         3.840e-01  6.766e-03   56.76   <2e-16 ***
tax                    2.870e-04  1.576e-05   18.21   <2e-16 ***
mpg                   -2.932e-03  7.677e-05  -38.20   <2e-16 ***
engineSize             3.394e-01  2.368e-03  143.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1847 on 49459 degrees of freedom
Multiple R-squared:  0.8805,    Adjusted R-squared:  0.8804
F-statistic: 2.024e+04 on 18 and 49459 DF,  p-value: < 2.2e-16
```

Test MSE: 378,700,539

We began our regression analysis by asking the question "which feature of a car influences its resale price the most?" The regression analysis we conducted led us to conclude that although all of the predictors were proven to be significant, the most significant predictors were make, year, and engine size. With Audi as the reference class for make, all the coefficients for make are negative as the other car brands tend to be priced lower than Audi. However, it makes sense that the make of the car largely influences the price because different brands have different reputations for reliability or luxury. We also expected the coefficient of year to be positive as cars tend to be more expensive as the year they were made becomes more recent. The coefficient for engine size is also as we expected because larger engines usually accompany faster cars which can also add to the expense of a car.

3.3 Regression Trees

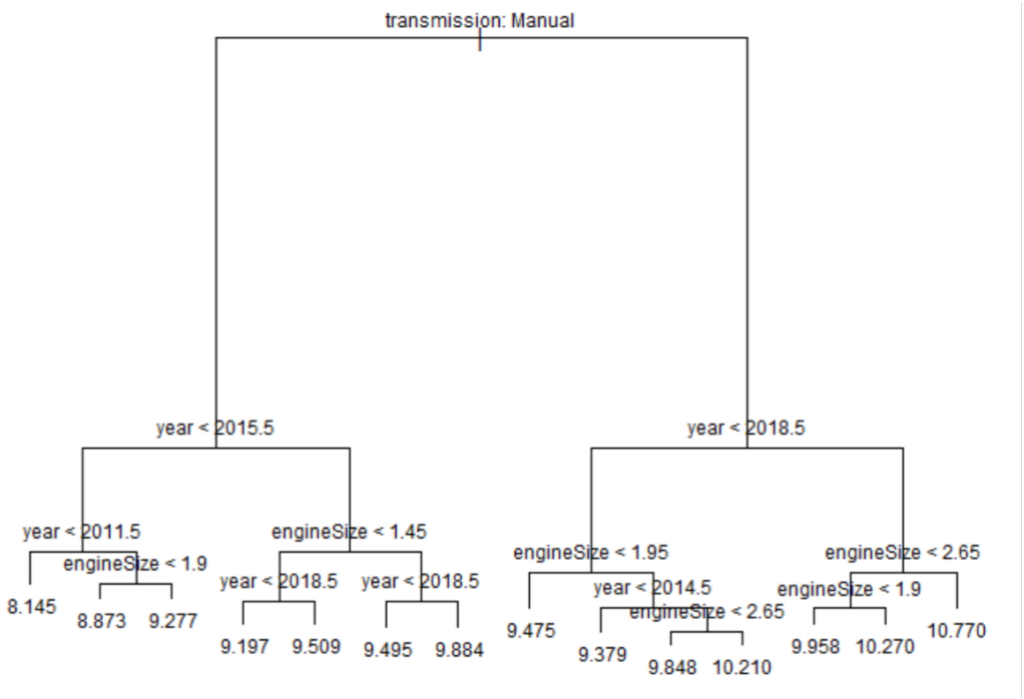
Since our regression tree using recursive binary splitting is the same as our pruned tree, here we present our regression tree using recursive binary splitting.

Regression tree using recursive binary splitting

Predictors used in the tree: transmission, year and enginSize

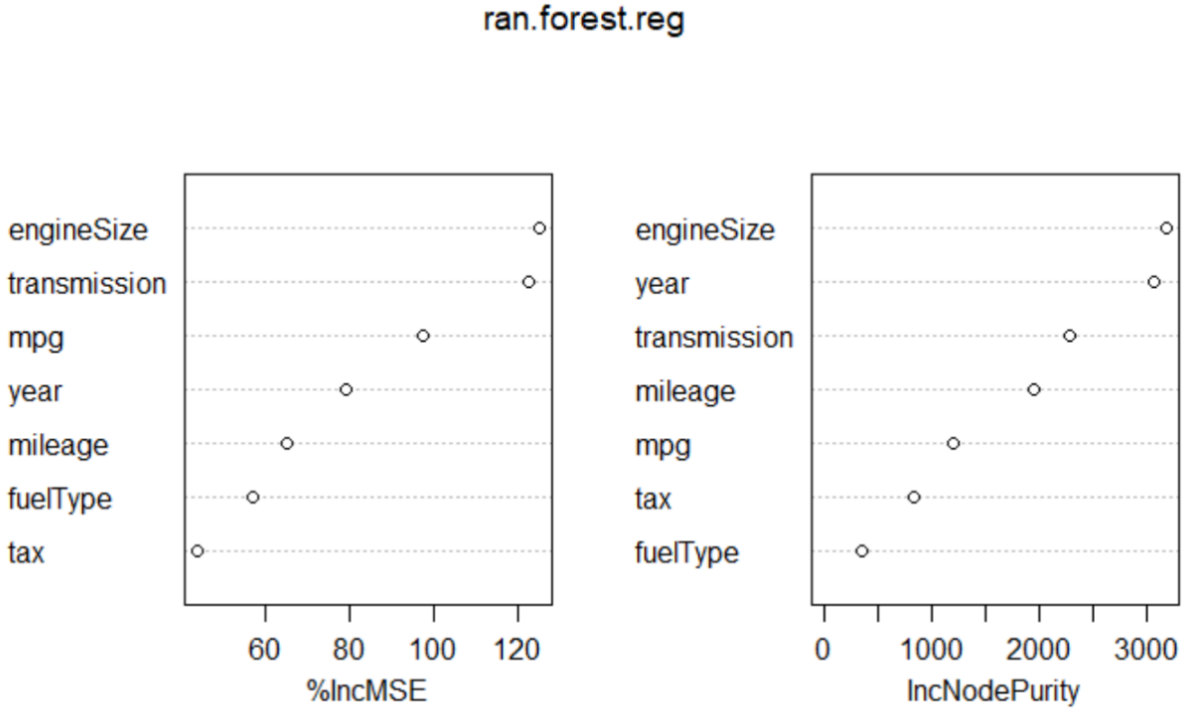
```
Regression tree:
tree(formula = log(price) ~ year + transmission + mileage + fuelType +
    tax + mpg + engineSize, data = training1)
Variables actually used in tree construction:
[1] "transmission" "year"         "engineSize"
Number of terminal nodes:  14
Residual mean deviance:  0.07437 = 3678 / 49460
Distribution of residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.385000 -0.172800  0.004255  0.000000  0.171400  2.539000
```

Our regression tree using recursive binary splitting has 14 terminal nodes. The predictors that R utilized for the tree are transmission, year, and engineSize. The following graph shows a graphical output of the regression tree. From this graph, we know that transmission type is the most important predictor in determining used car prices. Notice because we transformed the response variable prices into log(price), all the prices showing in the graph are log prices. Cars having a manual transmission and were produced before mid 2015 have a log price range from 8.145 to 9.277; cars having a manual transmission and were produced after mid 2015 have a log price range from 9.197 to 9.884; cars having automatic or semi-auto transmission and were produced before mid 2018 have a log price range from 9.475 to 10.21; cars having automatic or semi-auto transmission and were produced after mid 2018 have a log price range from 9.958 to 10.770. This answers our first question of interest: "which feature of a car influences its resale price the most?" According to the regression tree, transmission type influences its resale price the most, followed by years. Older cars having manual transmission tend to have lower prices; newer cars having automatic and semi-auto transmission tend to have higher prices.



Test MSE: 379,376,473

Improved regression tree using random forest



Using random forests, we found the most important predictors to be enginSize, followed by transmission and year. Random forests has a slightly smaller test MSE of: 379,374,906.

3.4: Summary of Findings

Our regression model produced the lowest test MSE followed by random forests while our regression tree produced the highest test MSE.

Comparisons of Test MSEs in Regression			
category	Regression model	Regression tree using recursive binary splitting	Regression tree using random forests
Test MSE	378,700,539	379,376,473	379,374,906

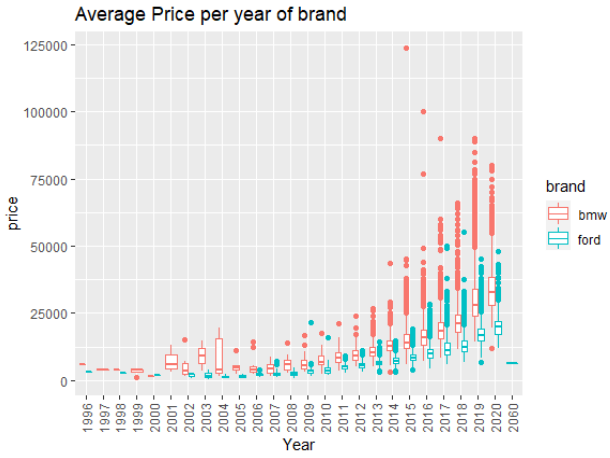
We used these methods to answer the question "which feature of a car influences its resale price the most?". Our regression model found that all of our variables were significant, so although it had the lowest test MSE, our regression trees are more helpful in answering our question of interest. Our regression tree found transmission type to be most influential in determining price whereas random forests determined engine size was most influential in determining the price.

Section 4: Classification

4.1 Exploratory Data analysis

There are 10 car brands in our data set and since we are only trying to classify two of them, we subset the large data set and extracted BMW and Ford's data. We then made sure that there were no null or missing values in our data. We created the following graphs with our data to view the distribution of our predictors. We present a few interesting variables here.

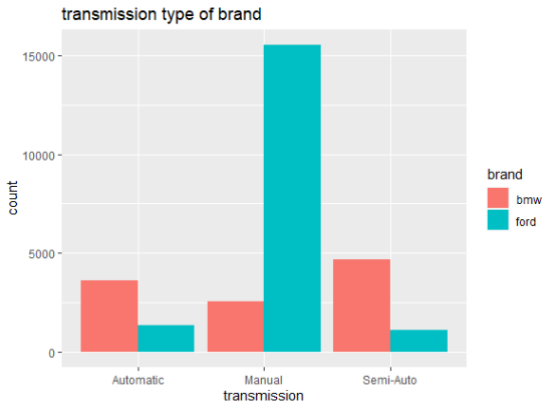
Price per brand of vehicle



The box plot provided outlines the average price per year of model for a specific brand of vehicle. The two brands analyzed are BMW and Ford, and are represented by the difference in color. The box plot indicates that BMW has a higher average and distribution of retail price across the years of 1996 to 2020 in comparison to Ford which has a cheaper average and distribution of retail price across the years of 1996 to 2020. Therefore, it can be assumed that the retail price for the BMW car manufacturing brand is, on average, higher and more expensive than the retail price for the Ford car manufacturing brand. We chose this to explore first as we knew from background knowledge that BMW being a

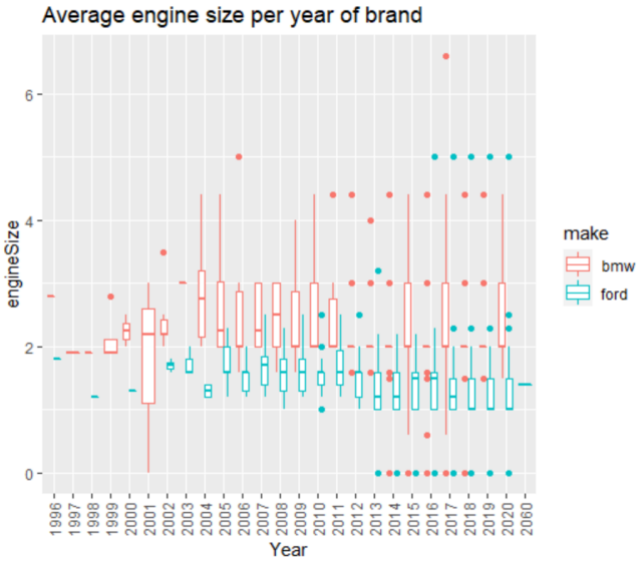
foreign luxury car brand and Ford being an American everyday car/truck brand, a feature of BMW vehicles would be their expensive price tag.

Transmission type per brand of vehicle



The bar graphs provided compare the numerical count of transmission types between vehicles produced by BMW and Ford. They are divided into Automatic, Manual, and Semi-Auto transmission types. Through the graph indicated, it can be concluded that Ford manufactured cars mostly have a Manual transmission and BMW cars are more evenly distributed amongst Automatic, Manual, and Semi-Auto transmissions compared to Ford vehicles, however, BMW has less distribution across Manual transmissions compared to Automatic and Semi-Auto. To try and find other features that could help us distinguish between BMW and Ford vehicles, we thought back to the vehicle transmissions. Automatic transmissions have been overtaking the manual transmissions in recent years and BMW being a luxury brand, we hypothesized that they would be more likely classified with automatic transmissions over manual transmissions. Exploring this leads to a fascinating insight into just how many Ford cars use the Manual transmissions in comparison.

Average Engine Size per Year of Brand



The boxplots shown to the right demonstrate the differences between the engine sizes used for BMW vehicles vs Ford vehicles. As we can see, on average throughout 1996 to 2020, Ford typically went with smaller engine sizes in their cars than BMW. On average, the median engine size per year of the BMW cars were larger than the median engine size of the Ford cars. Looking at this boxplot, an assumption can be made when trying to identify if the vehicle at question is a Ford or BMW; if they have a larger engine size, it is more likely that they are a BMW vehicle. There are, however, outliers engine sizes for both BMW and Ford but again looking at the boxplots

we see that the BMW outliers are typically larger than the Ford outliers.

Average MPG per Year of Brand

Looking at this plot we see that there has been a steady increase for both brands of average mpg throughout the years but not by much. The median mpg for Ford is slightly higher than the median mpg for BMW but BMW has many more outliers in the recent years.

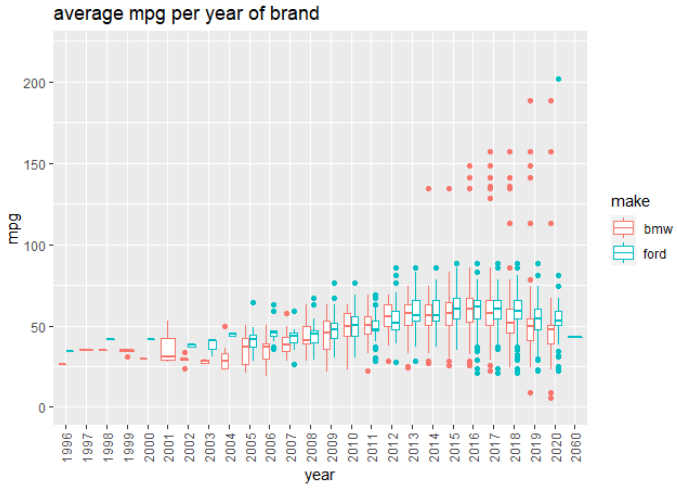
4.2 Logistic Regression

Our logistic regression model had a higher AUC value, so we present this model below.

Predictors in logistic regression

Make, either BMW or Ford, is our response variable. We chose to use 6 numerical variables as our predictors since LDA can only be performed with numerical variables. Our explanatory variables are year, price, mileage, tax, mpg, and engine size.

Logistic Regression Output:



```
Call:
glm(formula = make ~ year + price + mileage + tax + mpg + engineSize +
    transmission + fuelType, family = binomial, data = cltrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6300  -0.1964   0.1780   0.3811   6.8615

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)           -9.426e+02  4.973e+01 -18.954  < 2e-16 ***
year                   4.773e-01  2.475e-02  19.283  < 2e-16 ***
price                 -3.911e-04  1.254e-05 -31.189  < 2e-16 ***
mileage               -3.151e-05  1.998e-06 -15.769  < 2e-16 ***
tax                    5.106e-03  6.254e-04   8.165 3.21e-16 ***
mpg                   -1.320e-01  5.467e-03 -24.152  < 2e-16 ***
engineSize            -3.177e+00  1.327e-01 -23.940  < 2e-16 ***
transmissionManual     1.479e+00  8.182e-02  18.077  < 2e-16 ***
transmissionSemi-Auto -2.904e-01  9.947e-02  -2.919  0.00351 **
fuelTypeElectric       2.785e+01  2.813e+04   0.001  0.99921
fuelTypeHybrid         1.974e+00  4.738e-01   4.167 3.08e-05 ***
fuelTypeOther         -2.753e+00  1.590e+02  -0.017  0.98618
fuelTypePetrol        -2.561e+00  1.174e-01 -21.815  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19051.9  on 14371  degrees of freedom
Residual deviance:  7695.9  on 14359  degrees of freedom
AIC: 7721.9

Number of Fisher Scoring iterations: 13
```

Area under curve: 0.9589

Estimated test error rate using k fold cross validation for logistic regression

k=5	0.0742
k=10	0.0742

Confusion matrix for logistic regression		
	Classified as bmw	Classified as ford
bmw	4336	1020
ford	511	8506

FPR: 1020 / (1020 + 4336) = 0.1904 (The rate at which BMWs are misclassified as Fords over the total number of BMWs in our data)
FNR: 511 / (511 + 8506) = 0.0567 (The rate at which Fords are misclassified as BMWs over the total number of Fords in our data)

The output from our logistic model shows that the coefficients for year, tax, transmission manual, and fuel type hybrid are positive which means the larger these values are, the more likely the car will be classified as Ford. This conclusion aligns with our predictions. Since BMW is a luxury car brand, it is less likely to produce cars with manual transmissions which involve more work to operate and aren't typically seen as luxurious. In addition, since many of the cars Ford produces are work vehicles like trucks, it is more likely to produce cars with manual transmissions or hybrid engines. The variable fuel type electric also has a positive coefficient; however, this variable is not significant at the alpha=.05 level of significance so we believe it to be irrelevant in differentiating between the two brands.

The coefficients for price, mileage, mpg, engine size, fuel type petrol, and transmission semi-auto, are negative. This means the larger these values the more likely the car will be classified as BMW. This conclusion also aligns with our predictions. Since we know that BMW is a luxury brand and Ford is not, we expect that a higher price would increase the odds that the car is classified as BMW. In addition, we expect that the variables associated with higher quality makes such as high mpg and larger engines would also increase the odds of a car being classified as BMW. Fuel type other also has a negative coefficient, however, since the p-value for this variable is insignificant at the alpha=.05 level of significance, we believe this variable to be irrelevant in differentiating between the two brands.

The confusion matrix above summarizes the results of our logistic regression. From the matrix we obtain a false positive rate of 0.1904 and a false negative rate of 0.0567. The model, and therefore these rates were calculated using a threshold value of 0.5.

4.3 Classification Trees

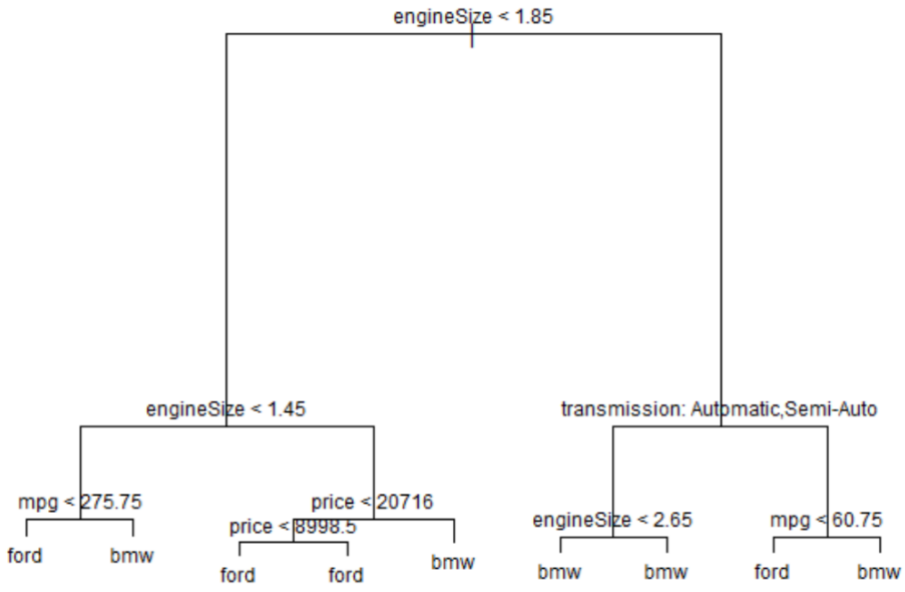
Since our classification tree using recursive binary splitting is the same as our pruned tree, we present our classification tree with recursive binary splitting.

Classification tree using recursive binary splitting

Predictors used in the tree: engineSize, mpg, price, transmission

```
Classification tree:
tree(formula = make ~ year + price + mileage + tax + mpg + engineSize +
    transmission + fuelType, data = cltrain)
Variables actually used in tree construction:
[1] "engineSize"   "mpg"          "price"        "transmission"
Number of terminal nodes:  9
Residual mean deviance:  0.5729 = 8219 / 14340
Misclassification error rate: 0.1204 = 1728 / 14354
```

Our classification tree using recursive binary splitting has 9 terminal nodes. The predictors that were used in the tree are: engineSize, mpg, price, and transmission. The following graph displays graphical output of the classification tree. According to the graph, we know that engineSize is the most important predictor in classifying car brands. This answers our second question of interest: "which features of a vehicle most greatly contribute to determining whether its car manufacturer is BMW or Ford?" If enginSize is smaller than 1.45 and mpg is less than 275.75, then the car is a Ford; if a car's engineSize is smaller than 1.45 and mpg is greater than 275.75, then the car is a BMW; if a car's engineSize is smaller than 1.85 but greater than 1.45, and its price is less than 20716, then it is a Ford; if a car's engineSize is smaller than 1.85 but greater than 1.45 and, and its price is more than 20716, then it is a BMW; if a car's engineSize is greater than 1.85 and has automatic or semi-auto transmission, then it is a BMW; if a car's engine Size is greater than 1.85 and has manual transmission and its mpg is smaller than 60.75, then it is a Ford; if a car's engineSize is greater than 1.85 and have manual transmission and its mpg is greater than 60.75, then it is a BMW.

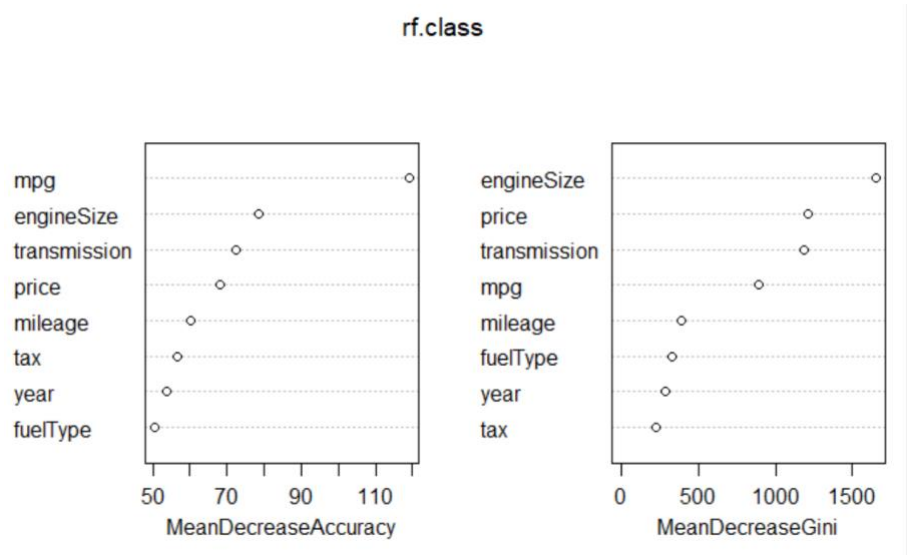


Confusion matrix for the classification tree using recursive binary splitting		
	Classified as bmw	Classified as ford
bmw	4397	951
ford	703	8304

Overall error rate: (703+951)/(4397+951+703+8304) = 0.1152
FPR: 951 / (951 + 4397) = 0.1778 (The rate at which BMWs are misclassified as Fords over the total number of BMWs in our data)
FNR: 703 / (703 + 8304) = 0.078 (The rate at which Fords are misclassified as BMWs over the total number of Fords in our data)

The threshold used in this confusion matrix is 0.5. The false negative rate is 7.8%, which is pretty low considering it is smaller than 10%. The false positive rate is about two times of the false negative rate. Thus, if we want to lower the false positive rate, we would need to increase the threshold.

Improved classification tree using random forest



The graph provides two different measures for importance, and they each produce a different result. The graph on the left determined the most important predictor in random forests is mpg and the graph on the right determined the most important predictor in random forests is enginSize.

Confusion matrix for the classification tree using random forests		
	Classified as bmw	Classified as ford
bmw	5080	268
ford	219	8788

Overall error rate: (219+268)/(5080+8788+219+268) = 0.0339
FPR: 268 / (268 + 5080) = 0.05 (The rate at which BMWs are misclassified as Fords over the total number of BMWs in our data)
FNR: 219 / (219 + 8788) = 0.024 (The rate at which Fords are misclassified as BMWs over the total number of Fords in our data)

To answer the question "which features of a vehicle most greatly contribute to determining whether its car manufacturer is BMW or Ford?", we used classification trees to categorize our data. Our classification trees found that engine size to be this feature. Using random forests, we found these determining features to be miles per gallon and engine size.

4.4 Summary of Findings

The following table summarizes results from all the confusion matrices in the classification section. All the thresholds used here are 0.5. It is obvious that the classification tree using random forests performs the best. It has the lowest overall error rate, false negative rate and false negative rate. As for the other two methods, the classification tree using recursive binary splitting has a lower false positive rate than the logistic regression, but it has a higher overall error rate and false negative rate. If we choose to use the classification tree using random forests, we don't need to adjust the threshold because it has a very low false positive rate and a low false negative rate: both small or equal to 5%. If we choose to use the other two methods, whether to adjust the threshold is contingent upon the model. A seller would be most interested in minimizing the false positive rate by increasing the threshold since selling a car that is a BMW as a Ford would cost them and reduce their profits. Increasing the threshold would increase the probability of a car being classified as BMW. In contrast, a buyer would be most interested in reducing the false negative rate, or the proportion of cars classified as BMW that are actually Fords. A buyer who believed their car to be a BMW when it is actually a Ford would be disappointed to learn that they overpaid for

their car. To reduce the FNR, buyers would prefer that the threshold be decreased, thereby increasing the probability that a car would be classified as a Ford.

When answering our second research question: "Which features of a vehicle most greatly contribute in determining whether its brand and car manufacturer is BMW or Ford?", these three methods give different answers. The output from our logistic regression shows that the coefficients for year, tax, transmission manual, and fuel type hybrid are positive which means the larger these values are, the more likely the car will be classified as Ford. According to the classification tree using recursive binary splitting, engine size is the most important predictor in classifying car brands. According to random forests, the determining features are miles per gallon and engine size. Random forests performs the best in terms of answering our question because it has the lowest error rates and directly provide the most important predictors, making it easier to interpret.

Comparisons of Test Error in Classification			
	Logistic regression	Classification tree using recursive binary splitting	Classification tree using random forests
Overall error rate	**0.1065**	**0.1152**	**0.0339**
False positive rate	**0.1904**	**0.1778**	**0.05**
False negative rate	**0.0567**	**0.078**	**0.024**

Section 5: Further Work

During our regression analysis, we considered adding interaction variables as predictors. We could not determine, however, what combination of interaction terms would help us to answer our regression problem, so we decided to build our model without them. Had we had more time, we would've explored different combinations of predictors. Our classification problem involved us differentiating between BMW and Ford. It would be interesting to create more models with different brands to determine whether the feature we concluded was most influential in differentiating between BMW and Ford is the same feature that is most influential in differentiating between any other combination of brands.

So far, our analysis focused on supervised learning methods, which explore the relationship between a response variable and a number of predictors. It would be nice if we could apply unsupervised learning methods to discover interesting features with our

variables. We could use principal component analysis to produce a lower dimensional representation of our data in our EDA. Furthermore, considering we have 9 brands of vehicles and numerous models, we can use K means clustering to discover how to divide all the brands into several subgroups. We could also use hierarchical clustering to discover subgroups among all models within each brand, or models across different brands.