Project Final Report

Team 29

Haris Saeed (hs8tuf)

Hamsini Muralikrishnan (hm7qgr)

Richard He (yh9vhg)

Bella Binder (imb6bwd)

1. Executive Summary

Questions of Interest

Regression Question:

Which predictors influence the selling price of a used car the most and should be accounted for when a used car is being sold?

Motivations

This question is worth exploring because car sellers can understand what characteristics increase or decrease the selling price of a car. Sellers can use this information when bringing in cars to the business to ensure a wide variety of cars to appeal to various customers. Furthermore, this data can be used by sellers to market the used cars a certain way by emphasizing the influential characteristics when speaking to a customer so that they are persuaded to buy the car.

Classification Question:

Which characteristics of a car influence whether a used car is sold to the customer or not?

Motivations

This question is worth exploring because sellers can see what characteristics of their car can enhance it being bought. Sellers can use the data for when they should sell and for what price at that time. They can also see if the characteristics of their car are similar to a car that did not sell and put cars in the storage or discard them if they are not worth putting on the market at that time.

Answering the Questions of Interest

This study found that the most important aspects when assigning a selling price to a car is the maximum power output and the model year of a car, with higher prices being attached to higher maximum power outputs and more recent model years. This study also found that the selling price of a car, how many kilometers the car has been driven, and the maximum power output of the car are important aspects for whether a car is sold or not, aiding sellers in determining which cars would be good on the market. In addition to these findings, the study also determined that other factors on whether a car is bought or not are related to human behavior, such as impulse buying, and socioeconomic factors, such as their financial status.

2. Data Description

Data Info

The data is about the Indian Used-Car market based on the information a consulting firm has gathered from various market surveys alongside the usage of a site known as Car Dheko, an online car dealing website. This dataset contains 18 different variables concerning used cars such as the mileage, make, engine type, car's selling price in rupees, and whether a car was sold or not to a customer.

Data Source

The dataset was found on Kaggle and the link is included below. The dataset is labeled Used-Car Data. We did not use data from any R package, nor the cereal data set from kaggle, nor the Western Collaborative Group Study (wcgs) dataset. None of the group members have previously worked with this dataset in previous classes. The dataset does not include any data related to time. Furthermore, the data included in this dataset is not simulated. Dataset Link: https://www.kaggle.com/datasets/shubham1kumar/usedcar-data?select=UserCarData.csv

Description of Variables Used

Response Variables

- The response variable for the regression question is the selling price column indicating the selling price of the used car measured in rupees. A new variable is created to convert rupees to USD to aid in the understanding of the data, and will be used in place of the original response variable.
- The response variable for the categorical question is the sold column indicating whether the used car was sold to the customer or not. The variable is binary and the two classes are "Y", referring to the car being sold, or "N", referring to the car not being sold.

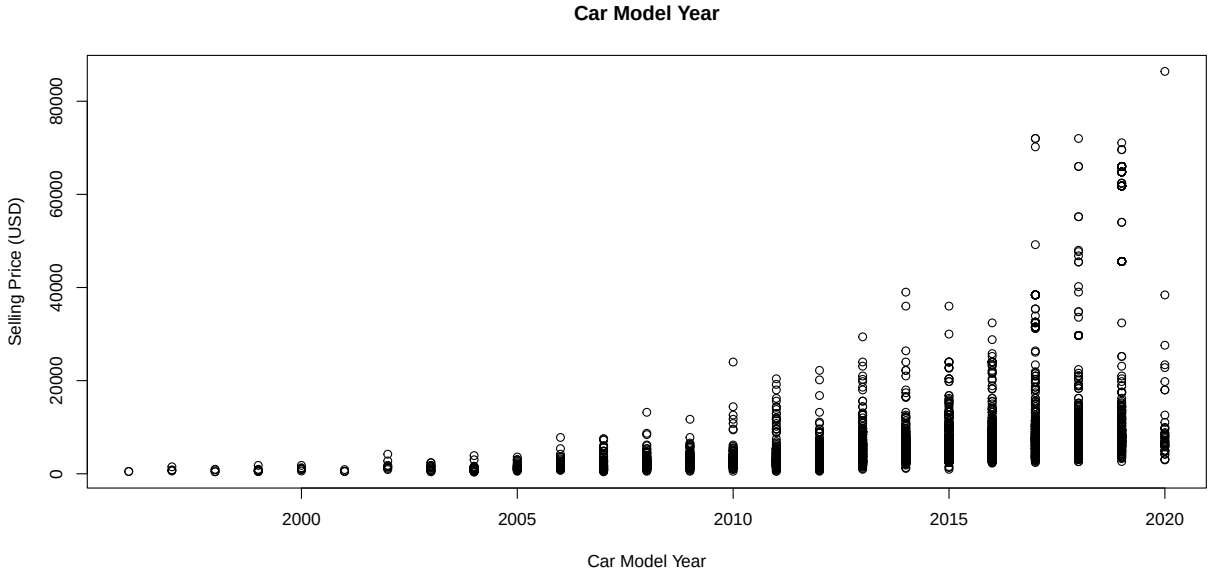
Other Important Variables

- Mileage: A continuous variable that refers to the number of miles a vehicle can travel with one gallon of fuel.
- Seats: The number of seats in the car. The values of this predictor are discrete as there can only be whole number/integer values.
- Engine: The displacement of the engine measured in cubic centimeters and its values are discrete.
- Region: The area in India the car used in and is categorical.
- Max Power: The power the engine can output and its values are continuous and measured in horsepower.
- Km Driven: Number of kilometers the vehicle has been driven before being put on sale and its values are discrete.
- Transmission: A categorical variable that refers to whether the car's transmission is automatic or manual.
- Fuel Type: The type of fuel the car uses, such as diesel or petrol and this is a categorical variable.
- Year: The year the manufacturer puts on the car, also known as the model year.
- Seller Type: The type of dealer selling the car, can be an individual, a dealer, or a trust mark dealer.
- Owner: Which owner of the car is selling the car, can be from the dealership classified as test drive car, the first owner, second, third, or fourth and above owner.

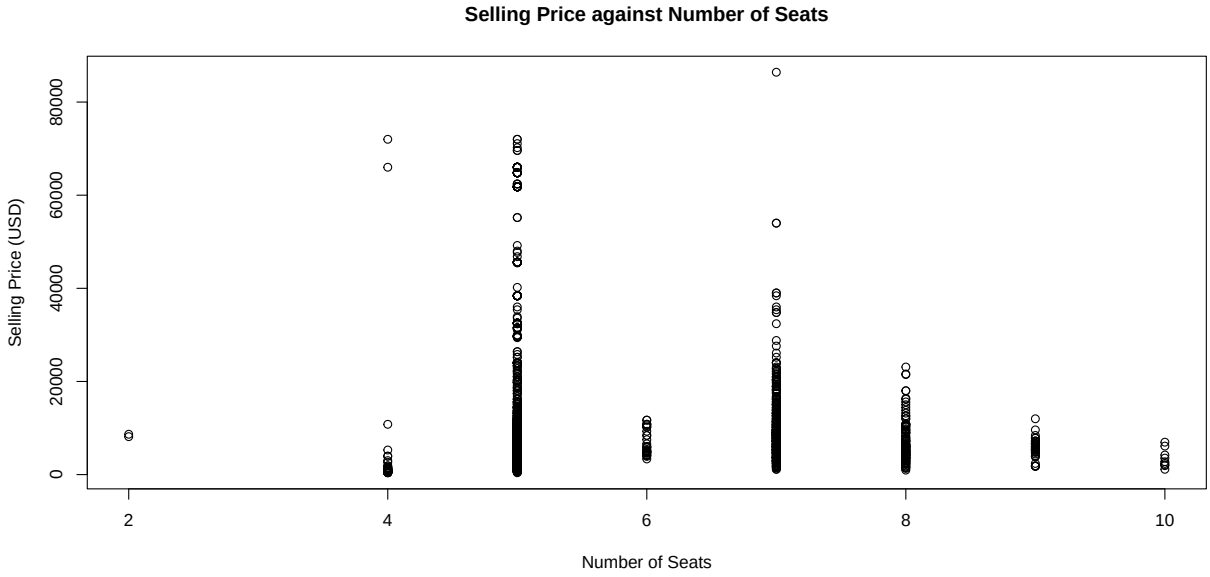
3. Regression Question

Data Cleaning

EDA

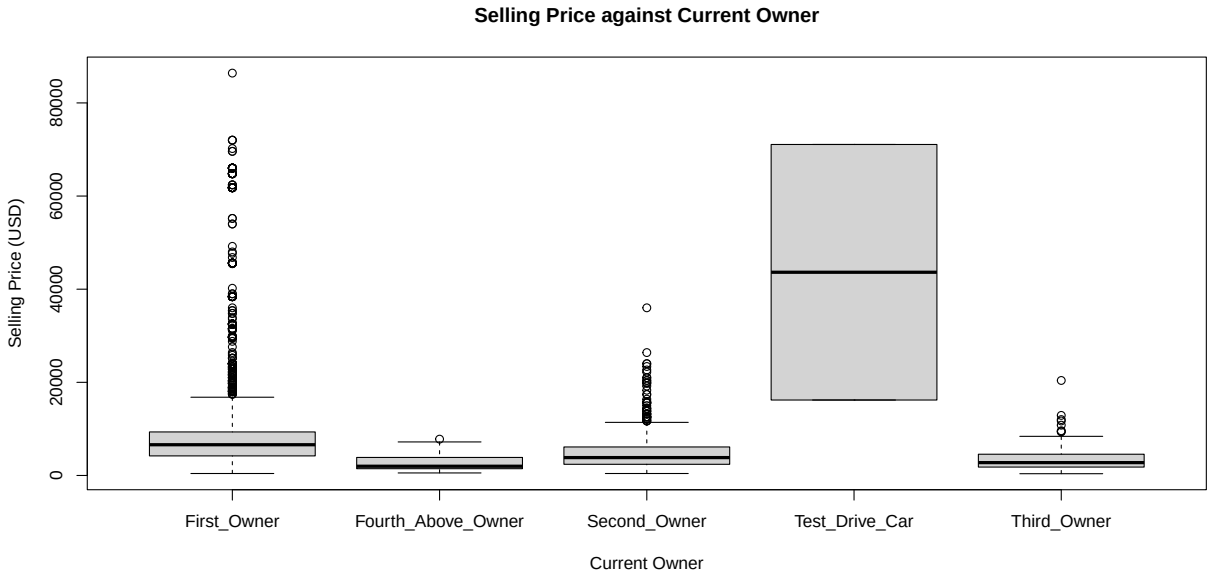


This scatter plot shows the selling price compared to the model year of the car was purchased by the current seller. It shows that there may be a strong positive correlation between a more recent model year and a higher selling price. This is not surprising since newer cars sell for more in the market. Additionally, cars that are newer are bound to be less used by the seller, making buyers see the car as if they are buying a brand new car, thus making sellers keep relatively the same price as when they bought it or even higher for a profit.

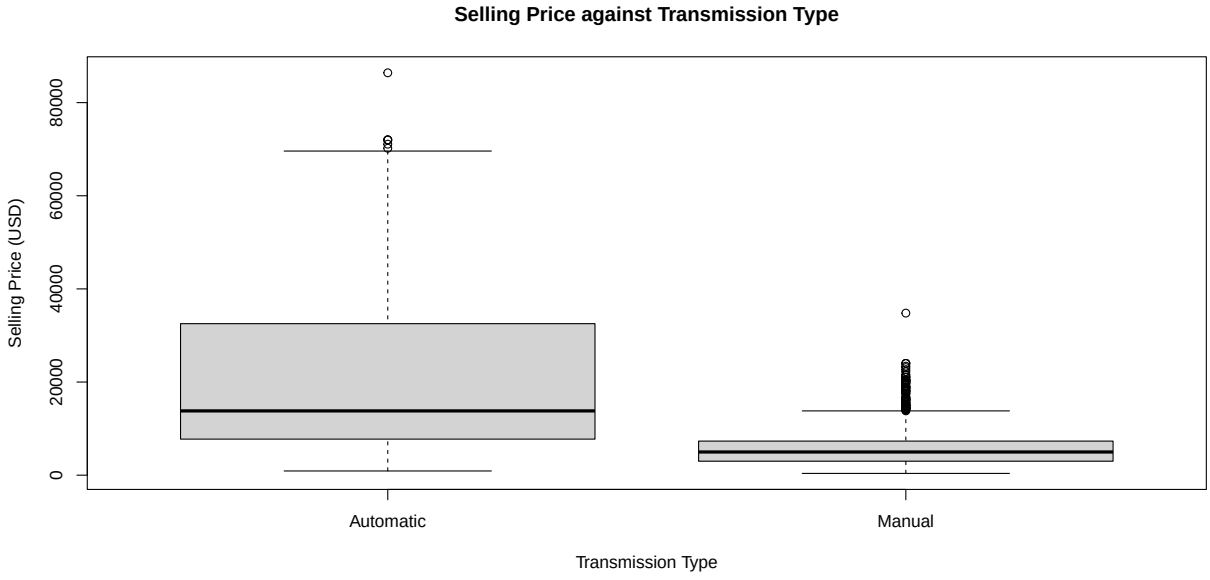


This scatter plot shows the selling price compared to the number of seats. It shows that there is an increase

in the selling price of 5-seat cars and a decrease in the selling price of 6+ seat cars. This is surprising since you would expect a bigger car to increase the selling price. Likewise, you wouldn't expect many families to buy cars with almost 10 seats.



This boxplot shows the selling price compared to different current owners that are putting the car on sale. There is a surprising difference between the car being sold from the dealer, labeled above as a "Test Drive Car", and the car already being used by any number of owners. We expected there to be a difference between each of the owners and it gradually decreasing with the car being passed down to more owners. Instead we see a huge drop from the dealership to the next owner and then relatively the same selling price from the first owner and rest



This boxplot shows the selling price compared to the transmission type. It shows that there is a large difference between the selling prices of cars with automatic transmission. Likewise, it shows the difference

between the selling prices of manual transmission is about half of automatic transmission.

Predictor Correlations with Selling Price

```
            [,1]
year       0.414
km_driven -0.247
mileage   -0.121
engine     0.453
max_power  0.757
seats      0.035
```

This numerical summary shows different predictors and their correlation with the selling price. There seems to be a positive correlation with year, engine, and max power predictors. There seems to be a negative correlation with the km driven and mileage (MPG). Seats has a positive correlation as well but it is fairly small compared to other predictors.

Predictor Correlations

```
##                   year   km_driven    mileage     engine   max_power
## year       1.000000000 -0.42854848  0.3285438  0.0182631  0.22659780
## km_driven -0.428548483  1.00000000 -0.1729803  0.2060307 -0.03815852
## mileage    0.328543848 -0.17298035  1.0000000 -0.5764079 -0.37462089
## engine     0.018263100  0.20603073 -0.5764079  1.0000000  0.70397453
## max_power  0.226597796 -0.03815852 -0.3746209  0.7039745  1.00000000
## seats     -0.007923033  0.22725939 -0.4517005  0.6111034  0.19199918
## price_usd  0.412301558 -0.22215848 -0.1262799  0.4556818  0.74967378
##                  seats   price_usd
## year      -0.007923033  0.41230156
## km_driven  0.227259388 -0.22215848
## mileage   -0.451700469 -0.12627995
## engine     0.611103386  0.45568180
## max_power  0.191999183  0.74967378
## seats      1.000000000  0.04161669
## price_usd  0.041616694  1.00000000
```

This numerical summary shows the correlation between predictors including the response variable of price in USD to see if there is any multicollinearity that needs to be dealt with. There seems to be a high correlation between the engine and max power and engine and seats. This makes sense since with more max power and increased weight of a car from seats, a better engine is needed. The selling price in USD seems to have a high correlation with max_power and moderate correlations with engine and year, indicating those predictors might be the significant predictors for selling price. Multicollinearity will be checked for prior to the model building to determine if engine and other predictors need to be removed to fix multicollinearity in the model.

Shrinkage Methods

Data Cleaning

Comparing MSE

```
##      ridge_MSE lasso_MSE  OLS_MSE
## [1,]  29067003  29284698 29364154
```

Regression Trees

Data Cleaning

Reason for Proposed Model

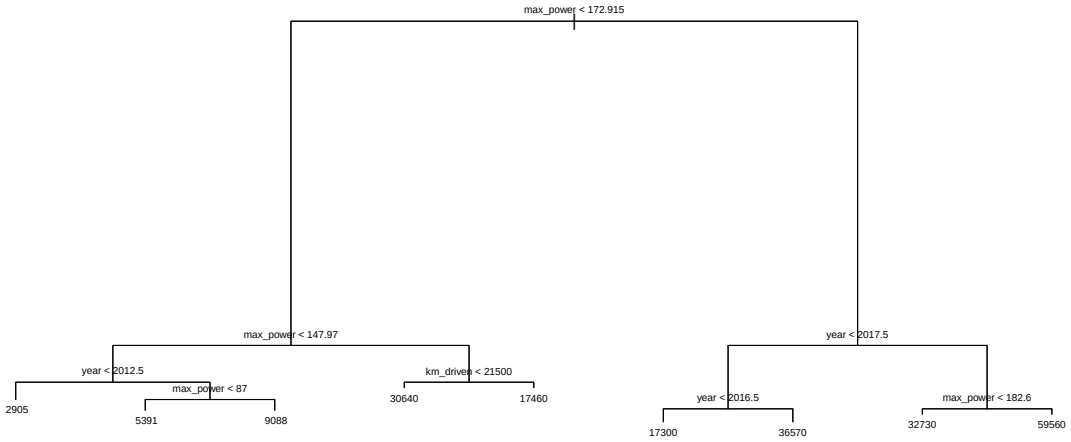
We chose to present the tree built with recursive binary splitting because after we did a 10-fold CV on our training tree model, we found the tree size to be 9, which is the same tree as before. We couldn't prune the tree since both recursive binary splitting and the CV have 9 terminal nodes.

Model Summary

```
##
## Regression tree:
## tree(formula = price_usd ~ ., data = train.tree)
## Variables actually used in tree construction:
## [1] "max_power" "year"      "km_driven"
## Number of terminal nodes:  9
## Residual mean deviance:  10970000 = 4.328e+10 / 3944
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -19360.0  -1588.0   -231.5      0.0   1415.0  35430.0
```

There are a total of 9 terminal nodes in our tree with 3 predictors estimating car sale price: the max_power, year, and km_driven.

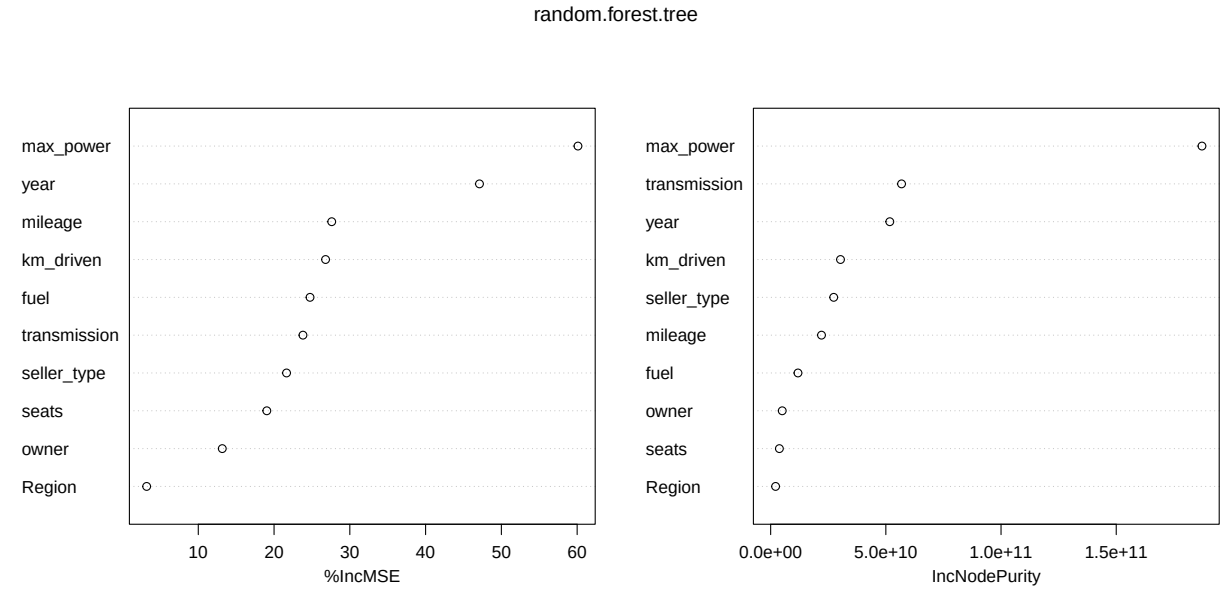
Graphical Output



Random Forest

```
##                %IncMSE IncNodePurity
## year         47.110560   51711230771
## km_driven    26.797927   30318176301
## Region        3.186816    2139404948
## fuel         24.747815   11848783429
```

```
## seller_type  21.649299   27370581606
## transmission 23.813822   56817053271
## owner        13.165892    5026756715
## mileage      27.610731   22072690114
## max_power    60.106845  187210980440
## seats        19.044222    3786127816
```



Comparing MSE

```
##   binary.splitting random.forest
## 1         15599804       5405648
```

Summary of Findings

Comparing MSE

```
##   linear.regression ridge.regression lasso.regression
## 1          29364154         29067003         29284698
##   recursive.binary.splitting random.forests
## 1                   15599804        5405648
```

Linear regression had the largest test MSE. Linear regression, ridge regression, and lasso regression are similar test MSEs around 29 million. Random forests had the smallest test MSE with 5.4 million. Recursive binary splitting did better than the linear, ridge, and lasso regressions with a MSE of 15.6 million, but worse than random forests. MSE squares the original unit, so the error value is reasonable considering the fact that the selling price in USD are values that are several thousands of USD.

How Proposed Methods Answers Our Question of Interest

The most important predictors are the max power of a car and the model year of the car when it comes to influencing the selling price. The recursive binary splitting tree has km_driven as an additional predictor, but the MSE is higher with the predictor included with the max power and year. These predictors show that sellers base their selling price based on how powerful the car is (the max power) and the model year of the car with newer cars that are more powerful with the highest prices. This is surprising as we expected

other predictors such as the km_driven to be more significant, but with the random forest regression tree and km_driven being less significant, the test MSE was halved.

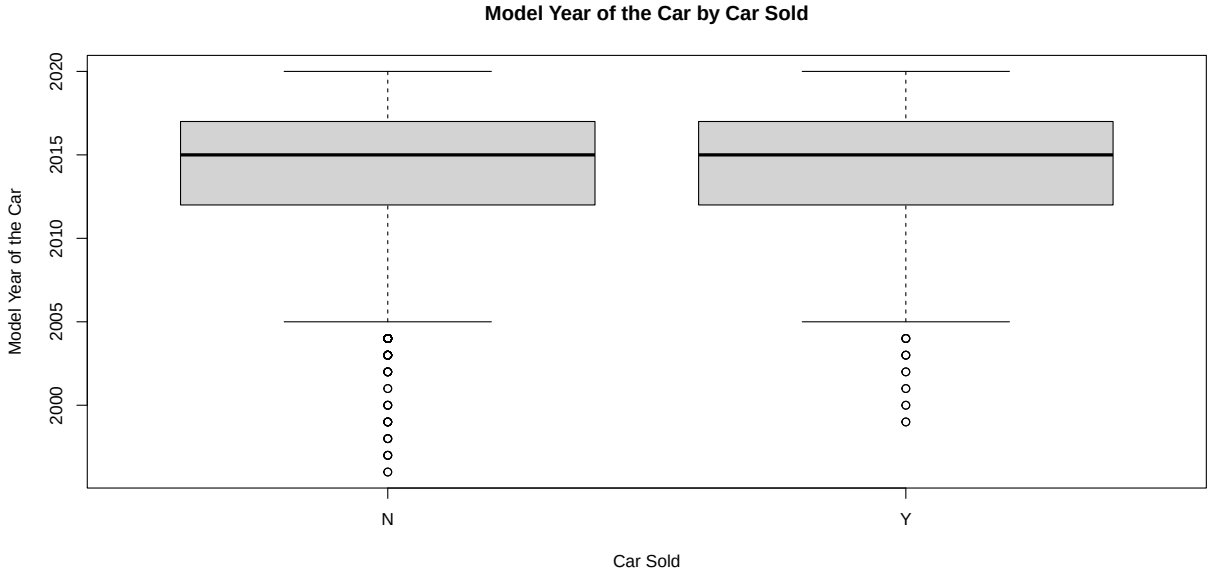
Proposed Best Method

Random forests is the best method that shows the max power and year of a car are the best predictors when it comes to influencing the selling price. Recursive binary splitting tree was also a good method, but with its involvement km_driven, the MSE was higher than random forests. Through the shrinkage methods, we see that ridge regression performs better compared to linear and lasso regression, but the MSE values are similar. Since ridge regression performed the best out of those methods, this shows that the model does better with less variance through the reduction of residual sum of squares. This brings in bias and this method does the best when there is some multicollinearity which means some multicollinearity is still present even after dealing with it before model building.

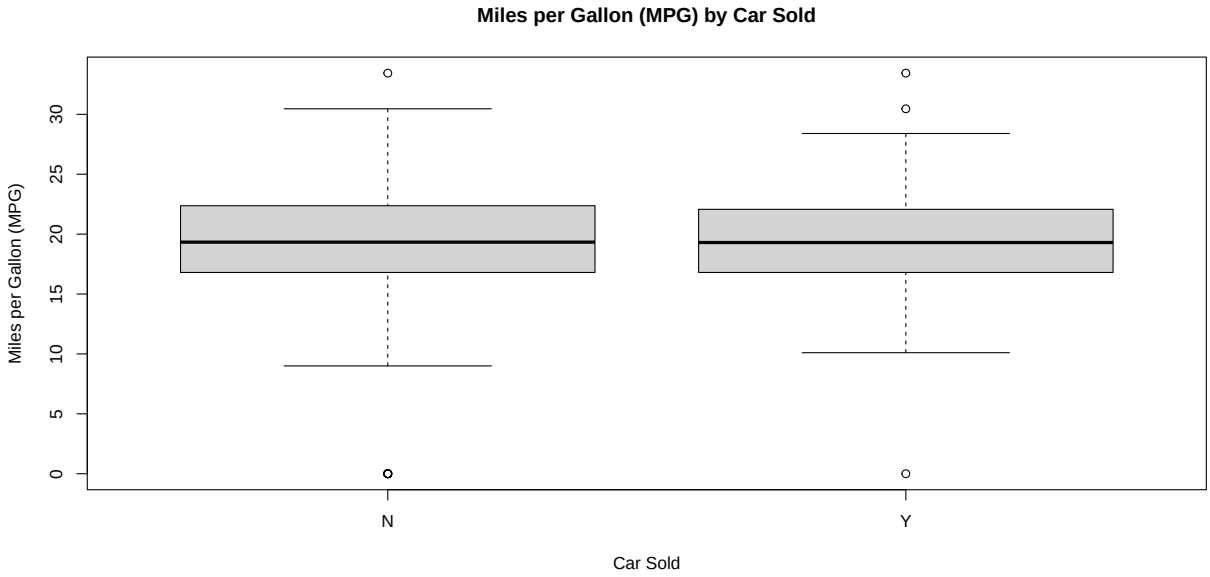
4. Classification Question

Data Cleaning

EDA



The box plot above displays the relationship between whether a vehicle was sold and the model year of the car (on training data). The distribution in mileage between its sold status seems similar, with a few sold and unsold vehicles as outliers when cars were purchased less recently. The outliers are explainable, since customers are more inclined to purchase the newer models, especially for a second hand car. However, the main distribution such as median and interquartile range are more similar, which means that year might not be as important for customers when considering whether to purchase a used car or not.



The boxplot above displays the relationship between whether a vehicle was sold or not based on its mileage (on training data). The distribution in mileage between a vehicle's sold status seems similar, with a few sold vehicles as higher outliers in mileage. Generally, for used vehicles, the more mileage the vehicle has, the cheaper it is, which could prompt more customers to purchase the car.

Region's Proportion with Sold Status

```
                N         Y
Central 0.3141892 0.2618328
East    0.2168919 0.3262840
South   0.2250000 0.1550856
West    0.2439189 0.2567976
```

This proportion table shows the car's region. Most cars are sold in the Eastern United States. Likewise, most unsold cars are in the Central United States. It is surprising that the most unsold cars are in the Central United States because it can be rural.

Fuel Type's Proportion with Sold Status

```
                 N           Y
CNG    0.005743243 0.004028197
Diesel 0.550675676 0.533736153
LPG    0.004054054 0.003021148
Petrol 0.439527027 0.459214502
```

This numerical summary shows the relationship between fuel type and whether a vehicle was sold or not. Not many vehicles in general, that were sold or not, ran on CNG or LPG as the fuel. Most cars in the dataset ran with diesel or petrol as their fuel. But, as displayed, there was a generally even split on whether the car was sold or not between diesel or petrol as their fuel type.

Predictor Correlations

```
##                    year   km_driven    mileage     engine   max_power
## year       1.000000000 -0.42854848  0.3285438  0.0182631  0.22659780
## km_driven -0.428548483  1.00000000 -0.1729803  0.2060307 -0.03815852
## mileage    0.328543848 -0.17298035  1.0000000 -0.5764079 -0.37462089
## engine     0.018263100  0.20603073 -0.5764079  1.0000000  0.70397453
## max_power  0.226597796 -0.03815852 -0.3746209  0.7039745  1.00000000
## seats     -0.007923033  0.22725939 -0.4517005  0.6111034  0.19199918
## price_usd  0.412301558 -0.22215848 -0.1262799  0.4556818  0.74967378
##                   seats   price_usd
## year      -0.007923033  0.41230156
## km_driven  0.227259388 -0.22215848
## mileage   -0.451700469 -0.12627995
## engine     0.611103386  0.45568180
## max_power  0.191999183  0.74967378
## seats      1.000000000  0.04161669
## price_usd  0.041616694  1.00000000
```

This numerical summary shows the correlation between predictors to see if there is any multicollinearity that needs to be dealt with. There seems to be a high correlation between the engine and max power and engine and seats. This makes sense since with more max power and increased weight of a car from seats, a better engine is needed. Selling price and the max power of a car also seem to have a high correlation. Multicollinearity will be checked for prior to the model building to determine if engine and other predictors need to be removed to fix multicollinearity in the model.

Logistic Regression

Model Summary

```
##
## Call:
## glm(formula = sold ~ year + Region + transmission + mileage +
##     max_power, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0399  -0.7773  -0.6882   1.3327   1.9255
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -39.420573  22.600561  -1.744   0.0811 .
## year                 0.019337   0.011277   1.715   0.0864 .
## RegionEast           0.592571   0.097958   6.049 1.46e-09 ***
## RegionSouth         -0.190658   0.113775  -1.676   0.0938 .
## RegionWest           0.236501   0.101294   2.335   0.0196 *
## transmissionManual  -0.245939   0.127850  -1.924   0.0544 .
## mileage             -0.018435   0.011090  -1.662   0.0964 .
## max_power           -0.002526   0.001370  -1.844   0.0652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4456.2  on 3952  degrees of freedom
## Residual deviance: 4387.3  on 3945  degrees of freedom
## AIC: 4403.3
##
## Number of Fisher Scoring iterations: 4
```

Confusion Matrix

The error rate is 0.2547. The FPR is 0 and the FNR is 1.

```
##
##     FALSE
##   N  2946
##   Y  1007
```

```
## [1] "The error is 0.254743232987604"
```

```
## [1] "The FPR is 0"
```

```
## [1] "The FNR is 1"
```

The threshold should be changed. This is because it would be better for car sellers to identified cars that may not actually sell to be marked as sold rather than cars that sell to be marked as not sold. This is because sellers want to earn a profit and thus the threshold should be changed. The new suggested threshold is 0.21, and that achieves an error rate of 0.5851, an FPR of 0.6901, and FNR of 0.2781.

```
##
##     FALSE TRUE
##   N   913 2033
##   Y   280  727
```

```
## [1] "The error is 0.585125221350873"
## [1] "The FPR is 0.690088255261371"
## [1] "The FNR is 0.278053624627607"
```

Classification Trees

Data Cleaning

Reason for Proposed Model

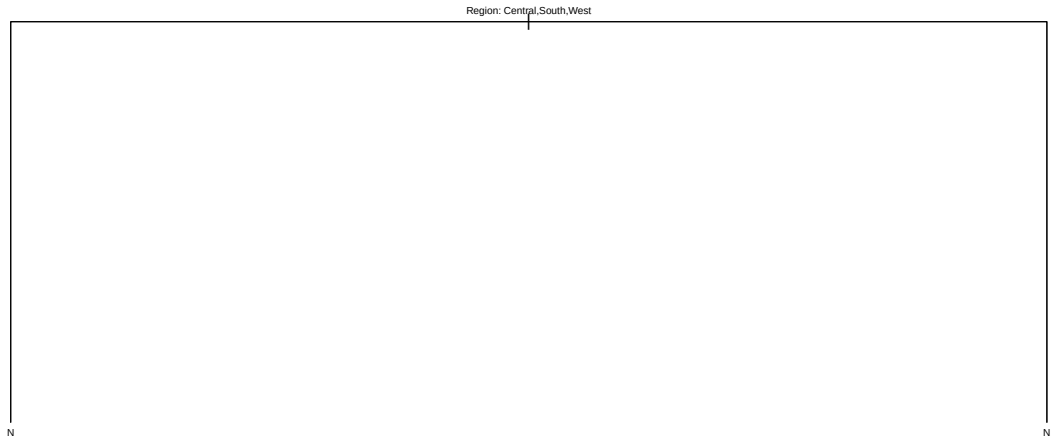
We chose to present the tree built with recursive binary splitting because we have tried multiple combinations of predictors, and the best model we got was with 2 terminal nodes, so there is no point of pruning our tree.

Model Summary

The tree has 2 terminal nodes in it with Region as the only variable used in the tree construction.

```
##
## Classification tree:
## tree(formula = sold ~ ., data = train.tree.class)
## Variables actually used in tree construction:
## [1] "Region"
## Number of terminal nodes:  2
## Residual mean deviance:  1.116 = 4410 / 3951
## Misclassification error rate: 0.2512 = 993 / 3953
```

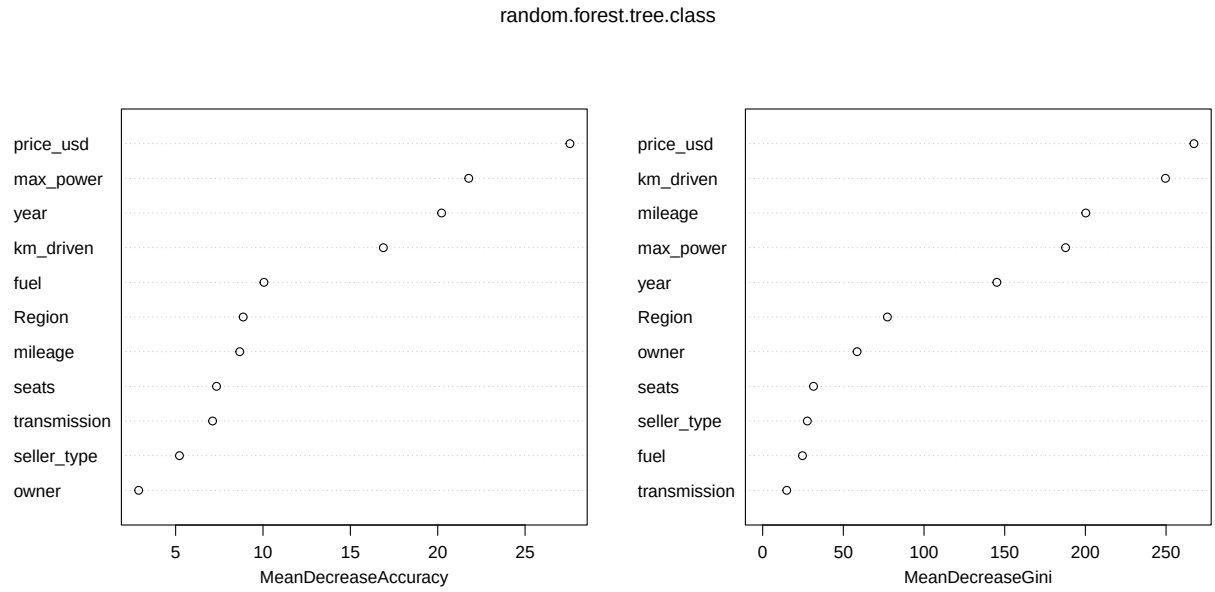
Graphical Output



Random Forest

```
##                     N           Y MeanDecreaseAccuracy MeanDecreaseGini
## year         27.700149 -23.0882516            20.223451        145.14341
## km_driven    23.406826 -14.2495985            16.888092        249.63535
## Region        4.138183  10.8517353             8.866267         77.29581
```

```
## fuel         13.967174 -11.2125055           10.055431         24.67223
## seller_type   6.403536  -2.3844291            5.219286         27.68372
## transmission 13.034392 -14.7505014            7.114429         14.93366
## owner         3.238460  -0.2867159            2.883127         58.52237
## mileage      14.087747 -11.5000563            8.668892        200.24374
## max_power    26.837940 -17.4436967           21.776737        187.73081
## seats         9.598041  -5.7208689            7.340311         31.52209
## price_usd    34.602288 -25.0812661           27.568411        267.28488
```



Confusion Matrix & Summary Stats

Recursive Binary Splitting The error rate is 0.2547, the FPR is 0, and the FNR is 1. [1] "FNR is 1"

```
##    pred.tree.class
##        N    Y
##   N 2946    0
##   Y 1007    0
```

```
## [1] "Error rate is 0.254743232987604"
```

```
## [1] "FPR is 0"
```

```
## [1] "FNR is 1"
```

The threshold does not need to be changed from 0.5. As shown below, adjusting it lower and higher does not make a difference. The results also make sense due to the fact that the sold variable is heavily dominated by N's indicating the vehicle not being sold, while there are a small amount of Y's, indicating the vehicle being sold.

```
## [1] "Threshold 0.01"
```

```
##
##     TRUE
##   N 2946
##   Y 1007
```

```
## [1] "Threshold 0.1"
```

```
##
##     TRUE
##   N 2946
##   Y 1007
```

```
## [1] "Threshold 0.9"
```

```
##
##     FALSE
##   N  2946
##   Y  1007
```

Random Forest The error rate is 0.2750, the FPR is 0.0479, and the FNR is 0.9394.

```
##    pred.random.forest.class
##        N    Y
##   N 2804  142
##   Y  947   60
```

```
## [1] "Error rate is 0.275486971920061"
```

```
## [1] "FPR is 0.0482009504412763"
```

```
## [1] "FNR is 0.940417080436941"
```

We believe that the threshold needs to be adjusted such that the FNR decreases, because a lot of cars that are sold are being identified as not sold. This is done by lowering the threshold. If the model is inaccurately identifying sold cars as unsold, this will reflect negatively on the company. With a threshold of 0.18, we now have an error rate of 0.5492, FPR of 0.6246, and FNR of 0.3287, as shown below.

```
##
##     FALSE TRUE
##   N  1106 1840
##   Y   331  676
```

```
## [1] "Error is  0.549203136858083"
```

```
## [1] "FPR is 0.624575695858792"
```

```
## [1] "FNR is 0.328699106256207"
```

Summary of Findings

Comparing Error Rates & FPR & FNR

The error rate for all 3 methods along with their FPR and FNR rates are similar. Binary splitting and logistic regression perform the same under a threshold of 0.5, while random forests has a higher error rate with an FNR of about 0.95 and FPR of about 0.05. Discussion of why the threshold should be changed and what it was changed to is in the next section.

```
##       binary.splitting random.forest logistic.reg
## Error        0.2547432    0.27548697    0.2547432
## FPR          0.0000000    0.04820095    0.0000000
## FNR          1.0000000    0.94041708    1.0000000
```

Threshold Discussion

We believe that the threshold needs to be adjusted for all of them. We want cars to be identified as sold even if they would not sell to make sellers put those cars out and make a profit rather than identifying cars that would sell as not being sold so they aren't sold and make less than they should. For binary splitting, as explained in an earlier section, regardless of the threshold, the results do not change. For random forests, a

threshold of 0.18 gives us an error of 0.5492, a FPR of 0.6246, and FNR of 0.3287. For logistic regression, a threshold 0.21 gives us an error of 0.5851, FPR of 0.6901, and a FNR of 0.2781. These thresholds were chosen to have a higher FPR value when compared to the FNR value, while also keeping error rates as low as they could be. Although the error is higher than before, this is more beneficial for a seller as described above. Random forests has a lower error, lower FPR, and higher FNR than the logistic regression model. Based on the new error rate, FPR, and FNR values, the better model is the random forest model.

```
##       binary.splitting random.forest logistic.reg
## Error        0.2547432     0.5492031    0.5851252
## FPR          0.0000000     0.6245757    0.6900883
## FNR          1.0000000     0.3286991    0.2780536
```

How Proposed Methods Answers Our Question of Interest

The most important predictors for whether a car is sold or not is the selling price, km driven, and max power of a car. This is based on the random forest classification tree since the binary splitting tree showed that no predictors had any significance. The logistic regression method found the year, region, transmission, mileage, and max power as significant predictors on a significance level of 0.1 rather than the desired 0.05 level which only shows region as significant, similar to the binary splitting tree.

Proposed Best Method

The best model in answering this question is the random forest classification tree. The random forest classification tree has the lowest error with a good FPR and FNR rate for the classification question that helps sellers determine what aspects of a car get the car sold in order to make a profit. These predictors are the selling price, km driven, and max power. Cars with a good selling price, less km driven, and are powerful are bound to be bought over other cars.
The logistic regression model is good at including multiple predictors (year, region, transmission, mileage, and max power), but on a 0.1 significance level rather than a 0.05. On a 0.05 significance level, only the region is involved, making it a poorer model than the 0.1 significance level. With this, the error is still higher than the random forest classification tree method. The binary splitting classification tree was poor in all aspects as it only saw region as an important factor, but even then, regardless of the region, the output of the tree was not sold.
Aside from this, it seems that there are other external factors that have not been considered that may be better predictors. Other factors that were not accounted for could be the buyer themselves and their financial status or even aspects of human behavior like buying on impulse that could ignore all predictors.

5. Challenges and Further Work

Some troubles we had as a group were loading packages into our file. We forgot to load some packages and couldn't use functions. It took a lot of time to finally figure out the issue and finally we were able to add the package for our function. Another challenge was to address the initial model being too uninformative for the classification question since it only had one significant predictor. We had to try various ways to uncover predictors that can potentially be significant. When we first started making the classification trees, we struggled to determine why we were getting a tree with only two nodes and how to analyze it. We tried using all the variables in the data set and compared it to our selected variables, and still there was no difference. After going to office hours, we determined that the predictors were not useful when it came to answering the question through classification trees. We continued from there and then struggled to determine how to adjust thresholds. After a long period of time, we realized that classification trees using different methods had different ways of changing the threshold.
If we had more time to work on this project, we would have tried to use a different dataset with different variables or even try to balance out the current dataset between cars that were sold and cars that weren't sold. A different dataset with different variables could inform us about other influential categories or even just indicate that the dataset we initially used is poor for predicting. Balancing the current dataset would allow for better analysis and test/train splits. Another aspect that we could dive deeper into is using datasets

from other countries and compare those countries and what predictors are the most influential for their cars and their selling prices and whether they are sold or not. This could add a global aspect to the project and see if the car market differs significantly based on country.