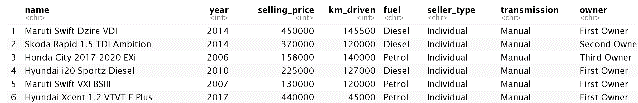
**ABSTRACT: This study analyses car data. The study makes use of R-Studio. Prior to data analysis, the data are preprocessed. Outlier analysis, data cleaning, and handling messy data are done. Also, some data manipulation techniques applied to prepare the data to prediction. Data manipulation is followed by the exploratory data analysis. Also,** **Logistic regression, Support Vector Machine****, Naïve Bayes, Decision Tree, XGBoost and Neural Network are applied to predict transmission type.**

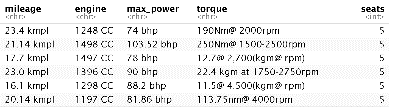
*1. INTRODUCTION*

I have worked with the car dataset. First, I showed the first six line of my dataset and structure of it. Then, I corrected and abbreviated the column names. Then, I looked the factor columns and correct the mistakes. After that, I applied exploratory data analysis. Then, I looked the number of NA values. After that, I cleaned each column and handled with messy data. Then, I detected outliers and replace them with mean of the columns where the outliers located. After, I applied PCA to reduce dimension. Then, I applied confirmatory data analysis. Finally, I split data into train and test data and conducted prediction models and performance measure.

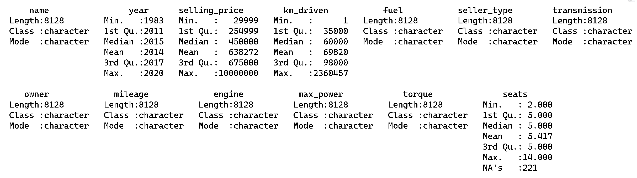
*1.1. DATA DESCRIPTION*

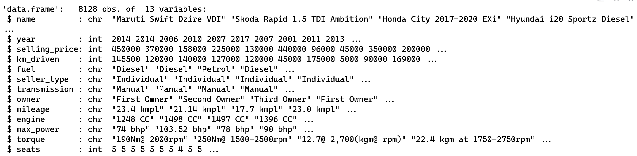
The data is about car sales. The data includes 13 columns. There is a name variable including names of the cars. The six of columns are categorical, and the six of them are numerical. Year is year in which car was sold, kmDriven is number of kilometers that the car is driven, seller\_Type tells us whether car is sold by individual or dealer, and owner tells the number of previous owners. Source the data is <https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho?select=Car+details+v3.csv>.





*Table 1 The first six lines of the data*

*Table 2 The summary of the data*

*Table 3 The structure of the data*

*1.2 RESEARCH QUESTIONS*

1) What are the distributions of the numeric variables?

2) Is the mileage distributed normally?

3) In different sale types, does transmission type have an impact on sales prices?

4) Is there a statistically significant correlation among numeric variables?

5) How is transmission type distributed according to sell type?

6) Does the engine affect max power in transmission type?

*1.3 AIM OF STUDY*

This project analyses the car dataset using statistical applications such as exploratory data analysis and confirmatory data analysis, and models which predict the transmission type of cars.

*2. METHODOLOGY/ANALYSIS*

First, data preprocessing is conducted to remove duplicate values and unused columns from the data. Then, the exploratory data analysis is conducted to answer the research questions by using visual techniques. After that, I handled with missing values by using predictive mean matching method (PMM) from mice package. Then, for data manipulation, Box-Cox transformation is used to normalize data, for numerical variables, PCA is applied to reduce dimension of data, and for categorical variables, dummy variables are created by using fastdummies package. Then, the confirmatory data analysis is conducted to find out exact answers of the research questions by using hypothesis testing such as Pearson’s Correlation Test, Spearman Correlation Test, Scheirer-Ray-Hare Test, Chi-Squared Test, and Kolmogorov-Smirnov test. Then, the dataset is splitted into two part which are train and test by validation set approach and by using createDataPartition function whose p is 0.8 from caret package. Finally, the models, such as Logistic regression, Support Vector Machine, Naïve Bayes, Decision Tree, XGBoost and Neural Network, are trained and compared to find out which model shows best performance.

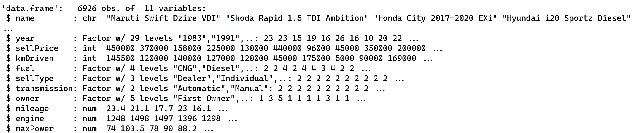
*3.RESULT AND FINDINGS*

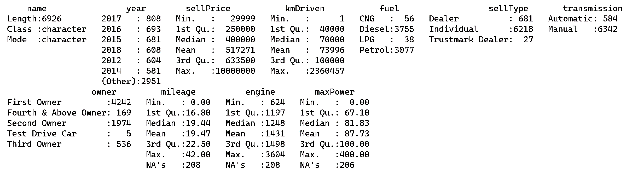
*3.1 DATA PREPROCESSING*

Firstly, duplicated data is checked. 1230 rows of data are duplicated value hence duplicated values are deleted. Then, the torque and seats are removed from data since there is no need the column. Then, column names are checked, and some column names are replaced with their different format. In mileage,engine and maxPower columns, there are their units; hence,the units are removed by using str\_remove from stringr. Also, categorical columns are checked to find a mistake by using table function, but any mistake was not found. Then, categorical columns are converted to factor. Also, numerical columns are got from data as a new data frame to compute the standard deviations of them.



After the changes:

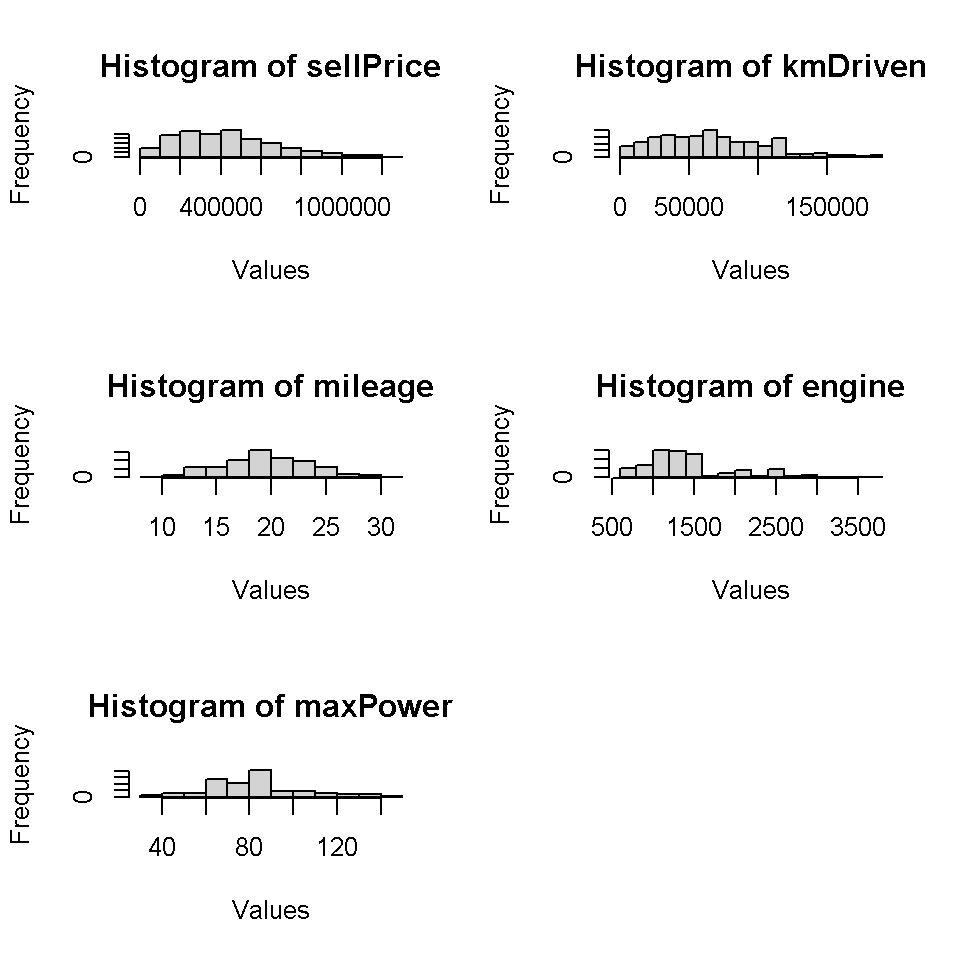
*Table 4 The structure of the data*

 *Table 5 The summary of the data*

Finally, outliers of numeric column are replaced with their mean values except engine and seats column since the percentage of their outliers is greater than ten percent of their size.

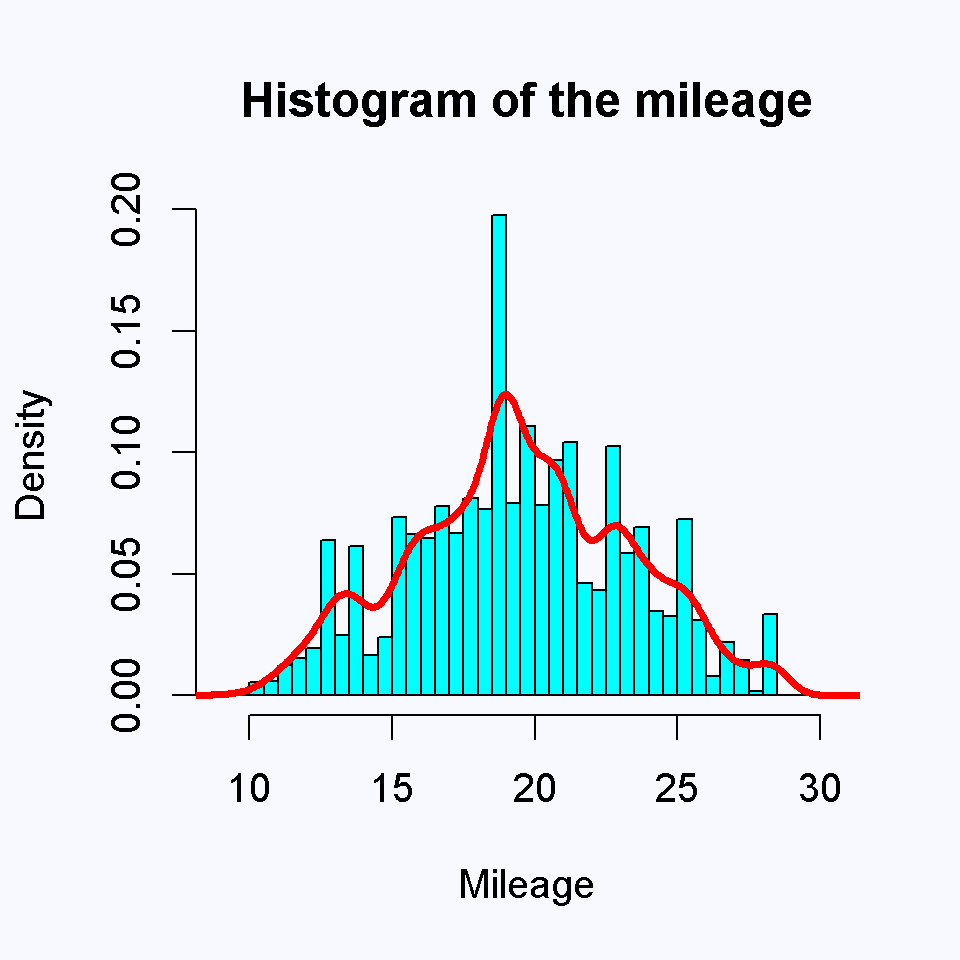
*3.2 EXPLORATORY DATA ANALYSIS*

*1) What are the distributions of the numeric variables?*

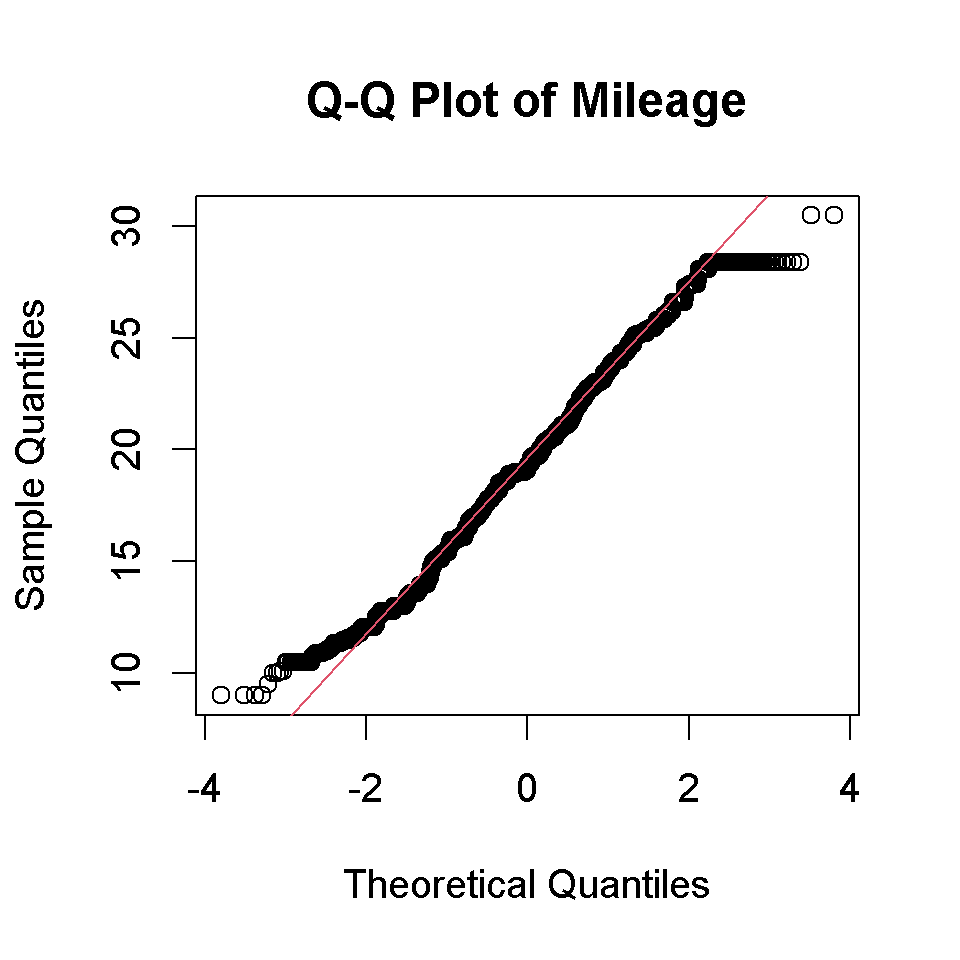
 *Figure 1 the histograms of numeric variables*

These plots show us the mileage may be distributed approximately normal. Also, price and engine columns look like right skewed.

*2) Is the mileage distributed normally?*

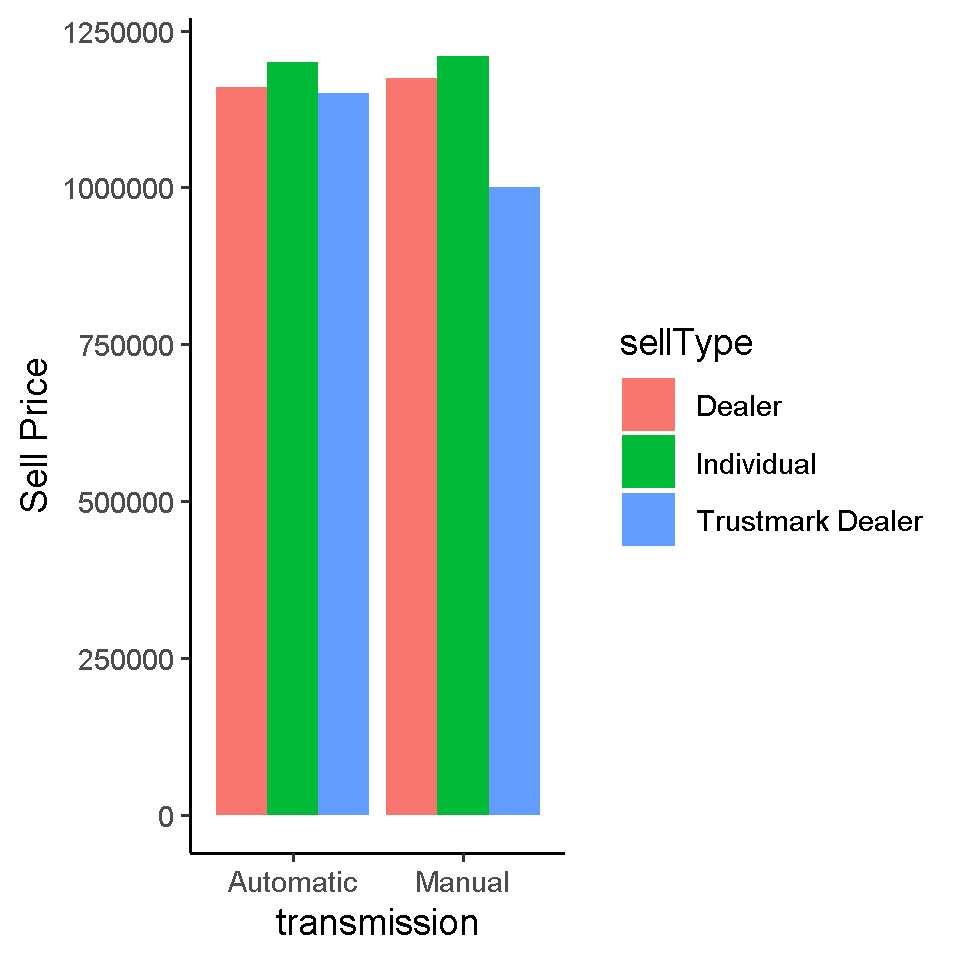


*Figure 2 Histogram and density plot of mileage*

 *Figure 3 Q-Q plot of mileage*

According to Histogram and density plot, the distribution of it is not normal. Also, when we checked the Q-Q Plot of it, there exists significant deviations from the line. Hence, according to visual evidence, the mileage does not normally distributed.

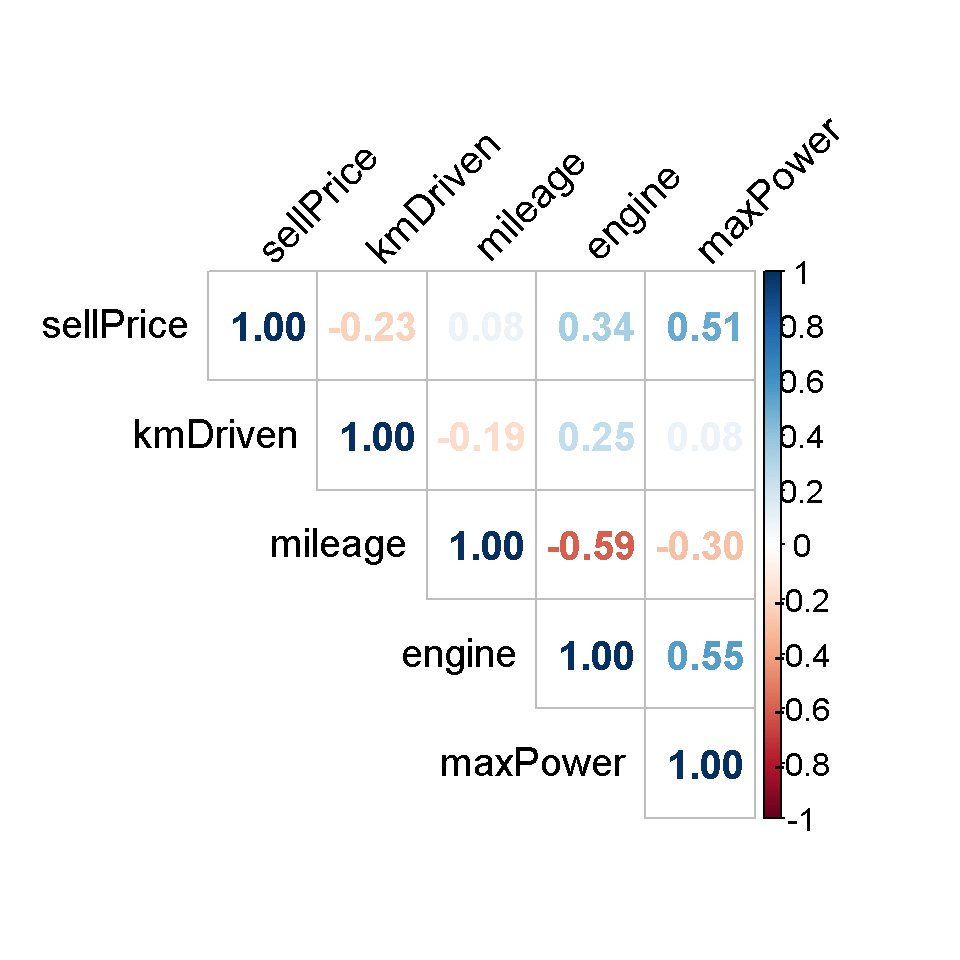
*3) In different sale types, does transmission type have an impact on sales prices?*



*Figure 4 Bar plot of sell price according to sell type and transmission type*

This plot suggests that there may be a significant difference between sell prices for each sell type depending on the transmission type. However, for each situation, the prices of cars which are sold by individual is highest and the prices of cars which are sold by Trustmark dealer are lowest. Also, the prices of cars with manual type increase for each sell type except Trustmark Dealer.

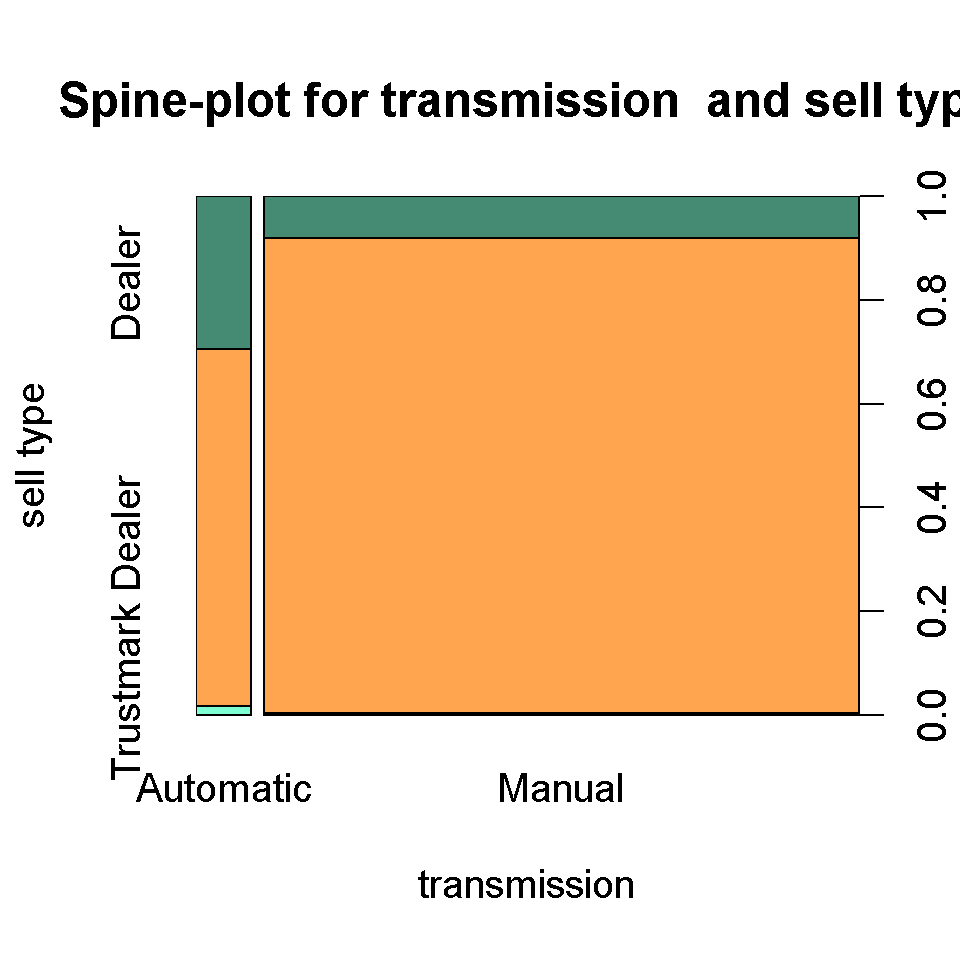
*4) Is there a statistically significant correlation among numeric variables?*



*Figure 5 Correlation plot of numeric variables*

According to the plot, there exists a high positive correlation between max power and sell prices. Also, there are high positive correlations between engine& max power and engine& seats. However, there are high negative correlations between engine& mileage and mileage& seats. Also, there is no multicollinearity problem.

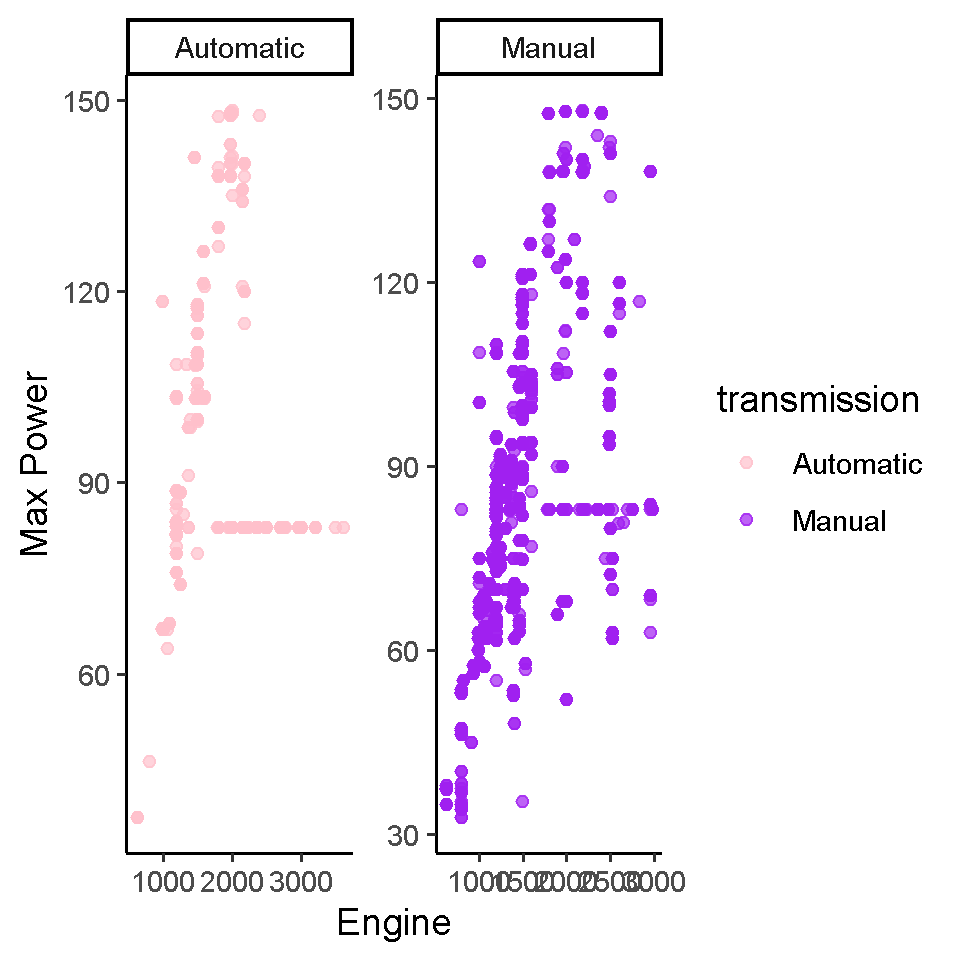
*5) How is transmission type distributed according to sell type?*



*Figure 6 Spine plot of transmission type and sell type*

According to plot, for each transmission type, the highest percentage is Trustmark dealer and the lowest is dealer. However, there is no individual seller for manual transmission type.

*6) Does the engine affect max power in transmission type?*

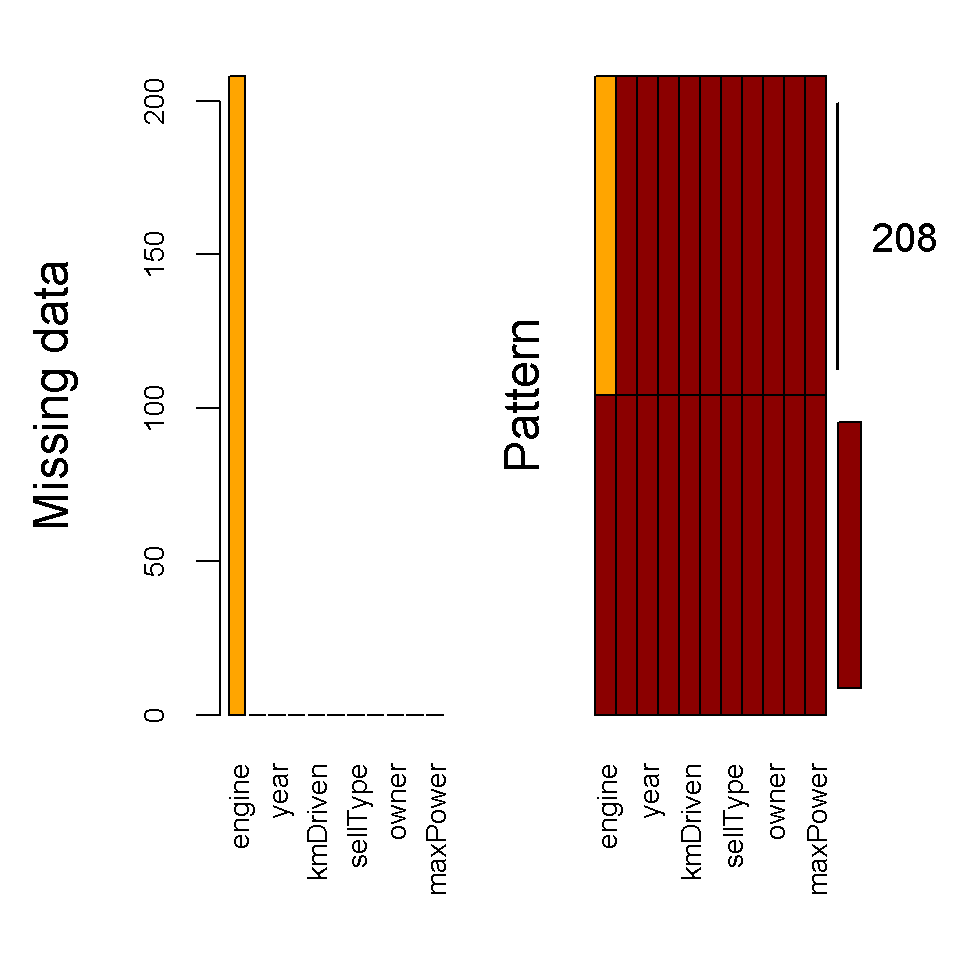


*Figure 7 Scatter plot of engine and max power according to transmission type*

For each plot, there is a positive correlation between engine and max power. Hence, we can say engine affects max power for each transmission type.

*3.3 HANDLING MISSING VALUES*

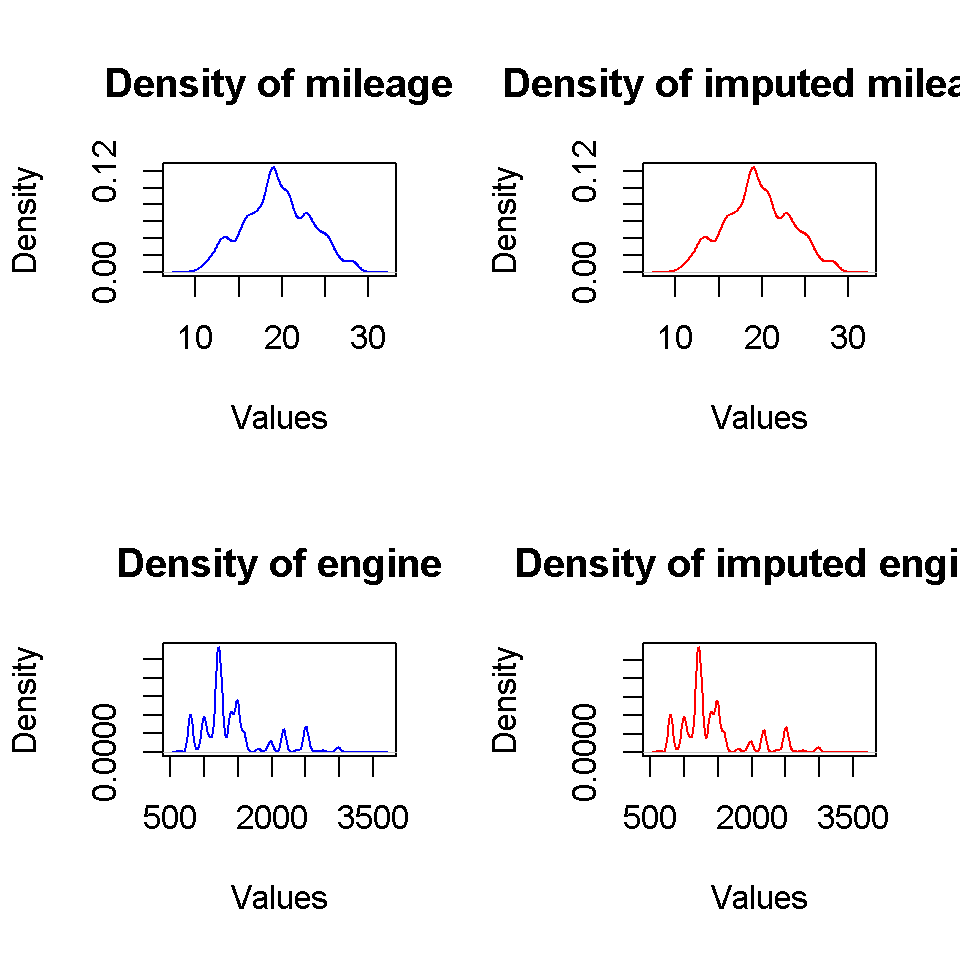
I handled with them by using mice package.



*Figure 7 Missing data plot* *Table 6 Missing data table*

According to the outputs, there exist 208 NA values in data. The values are in engine column. NA values are replaced with new values by using predictive mean matching method (PMM).

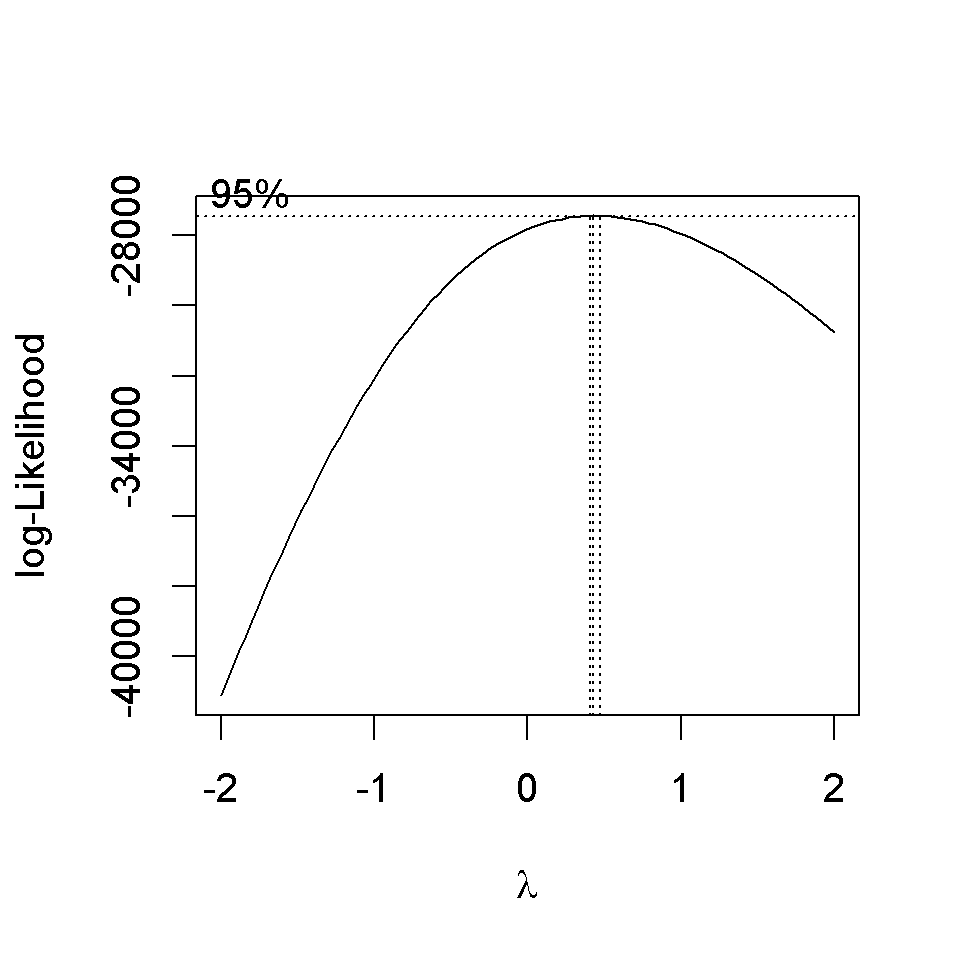
Then, the distribution of imputed engine column is checked to find out that the distribution do not change after imputation.



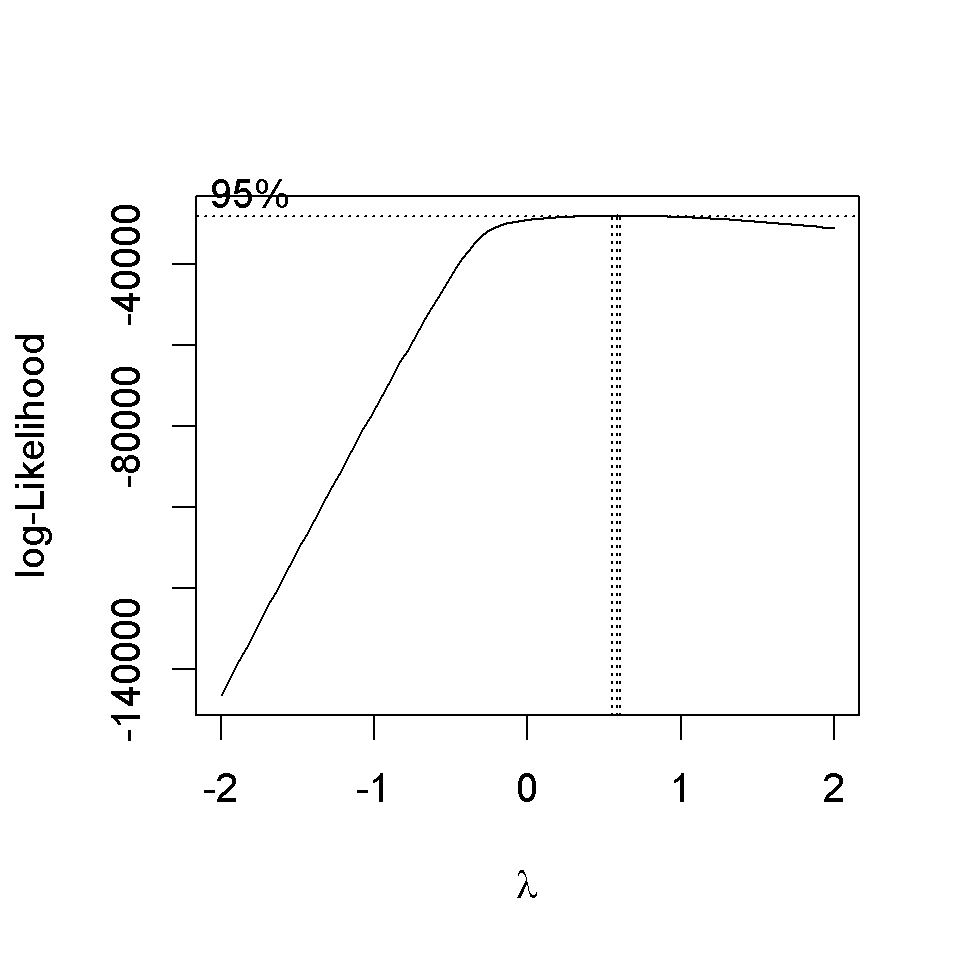
*Figure 8 Density plots of engine and imputed engine*

According to density plot, the distribution of it does not change.

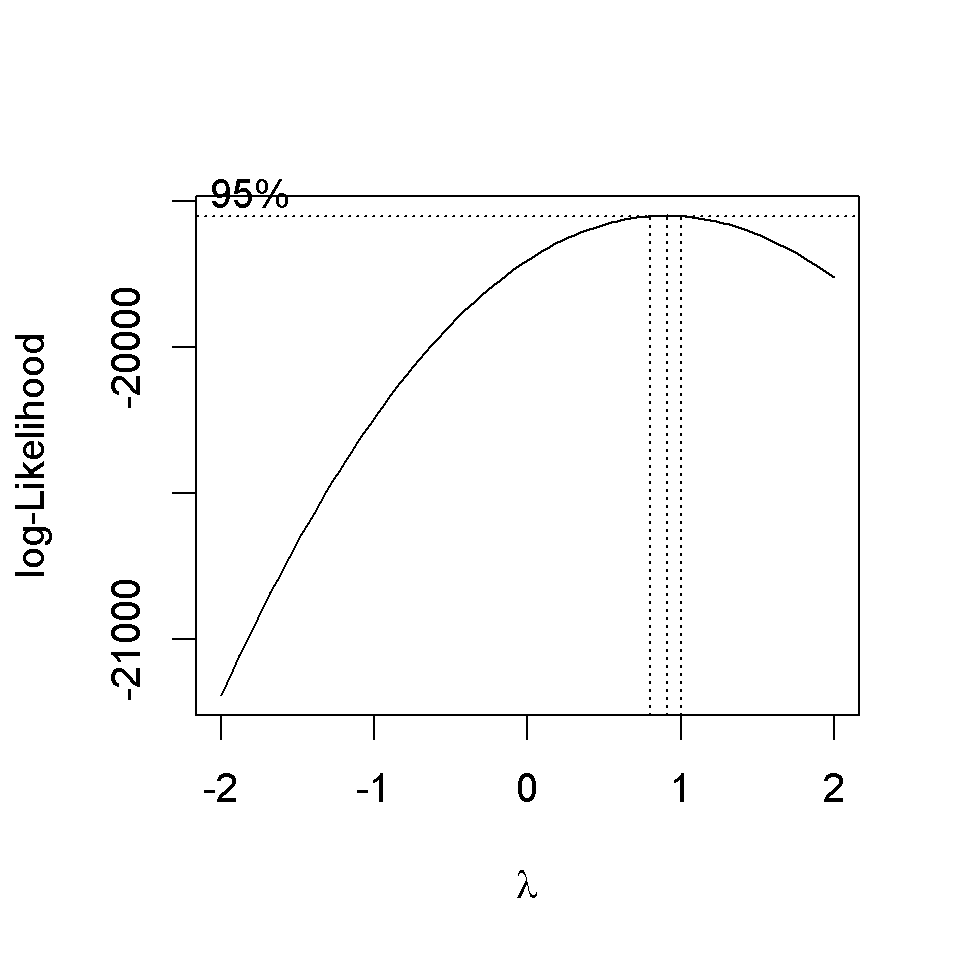
*3.4 DATA MANIPULATION*

 Firstly, I used Box-Cox transformation to normalize the numerical columns.

For sellPrice columns, since the interval does not include 1, I transformed sellPrice column.



For kmDriven column, since the interval does not include 1, I transformed kmDriven column.



For mileage column, since the interval does not include 1, I transformed mileage column.

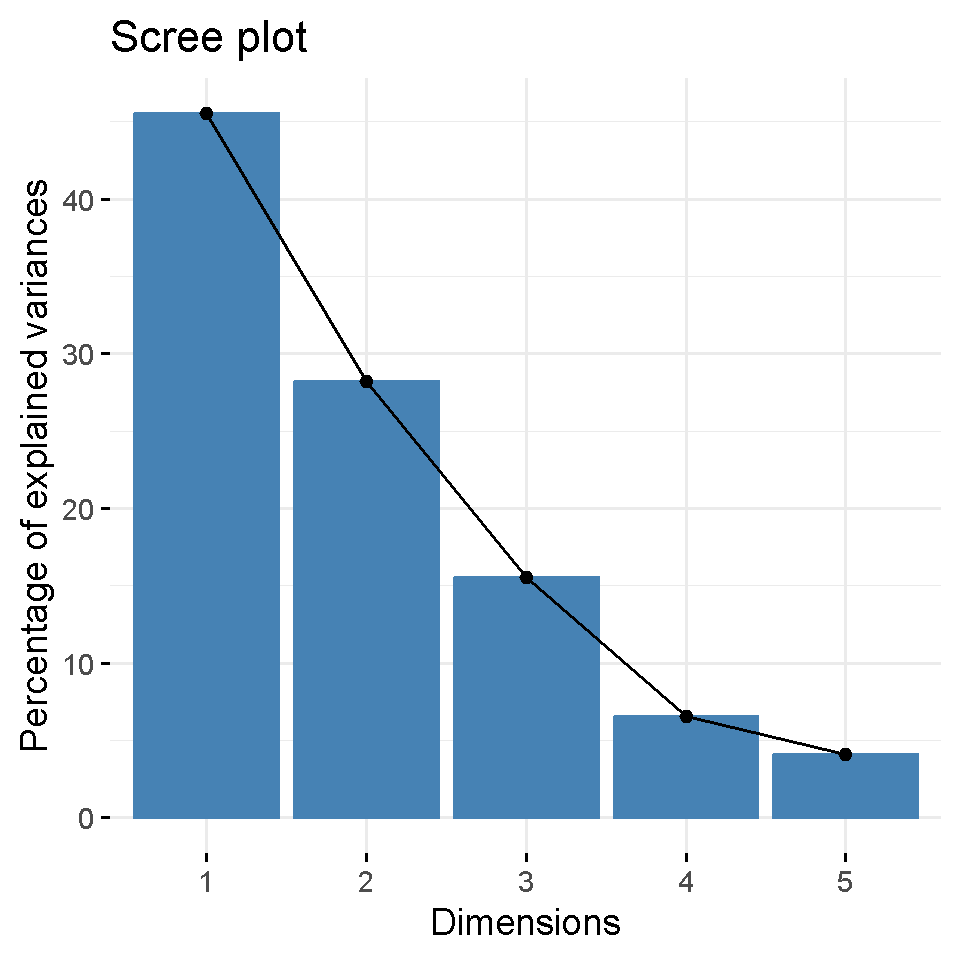


For max power column, since the interval does not include 1, I transformed max power column.



For engine column, since the interval does not include 1, I transformed engine column.

After that, in order to reduce the dimensions of the data, I used principal component analysis (PCA).



*Figure 9 PCA plot*

According to the plot, I chose first two principal components.

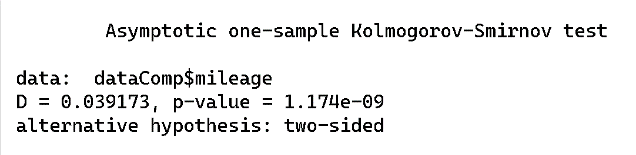
Finally, I used fastDummies package and dummy\_cols function for one hot encoding of the categorical columns.

*3.5 CONFIRMATORY DATA ANALYSIS*

*2) Is the mileage distributed normally?*

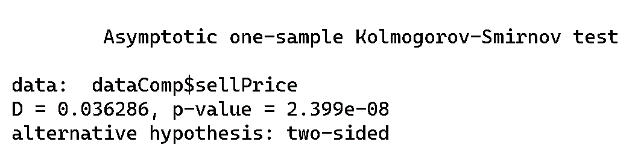
Since length of the data is greater than 5000, we cannot apply Shapiro-Wilk test; hence, we will apply Kolmogorov-Smirnov test.

Null Hypothesis (H0): Sample data comes from the normal distribution.

Alternative Hypothesis (H1): Sample data does not come from the normal distribution.

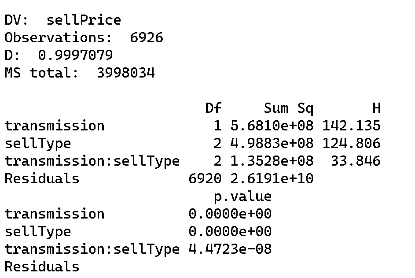
Since p value is less than 0.05, reject null hypothesis. Hence, sample data does not come from the normal distribution.

*3) In different sale types, does transmission type have an impact on sales prices?*



Since p value is less than 0.05, reject null hypothesis. Hence, sample data does not come from the normal distribution.

Since normality assumption is not provided, I used Scheirer-Ray-Hare Test which is the non-parametric version of the two way anova.



Hypothesis 1 (Sell Price does not differ depending on transmission type): 0 < .05

Hypothesis 2 (Sell price levels do not differ depending on sell type): 0 < .05

Hypothesis 3 (The combination of transmission type and sell type is not impacting the sell price): 4.4723e-08 < .05

Hypothesis 1: Reject!

Hypothesis 2: Reject!

Hypothesis 3: Reject!

Hence, in different sale types, transmission type has impact on sales prices according to hypothesis 3.

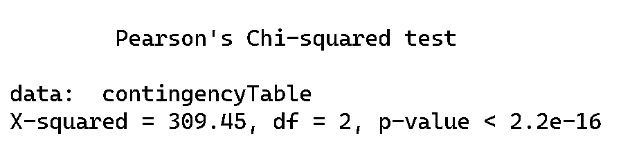
*4) Is there a statistically significant correlation among numeric variables?*

Pearson’s correlation test is conducted to test correlation between each two numeric variables. All p-values except correlation test for kmDriven and maxPower are equal to 2.2\*10^(-16) and p-value of correlation test for kmDriven and maxPower is equal to 4.028\*10^(-10).

According to the outputs, there is no correlation between each two numeric variables since all p-values of the Pearson’s correlation tests are less than 0.05.

*5) How is transmission type distributed according to sell type?*

Since there exist two categorical variables, we can apply Chi-Squared Test.



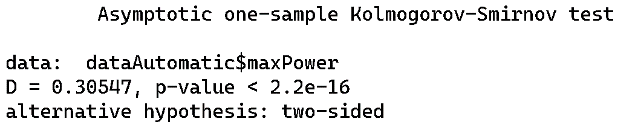
We fail to reject the null hypothesis, and we conclude that there is not a significant association between sell type and the situation that car is sold in 2020 or before.

*6) Does the engine affect max power in transmission type?*

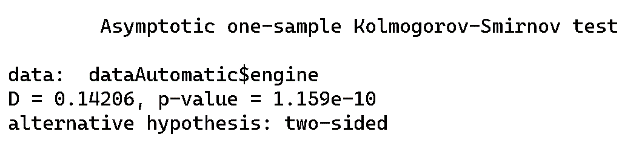
Firstly, the normalities of maxPower and engine variables are checked for different transmission types.

For automatic type, the p-values of the Kolmogorov-Smirnov test are:

For maxPower:

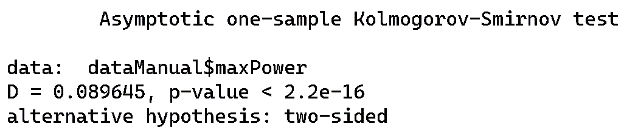


For engine:

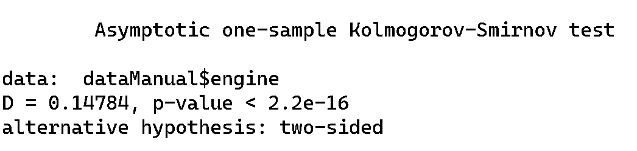


For manual type, the p-values of the Kolmogorov-Smirnov test are:

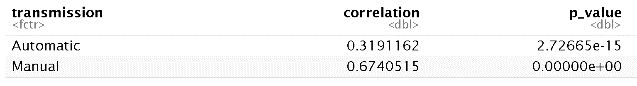
For maxPower:



For engine:



Since all variables are not normally distributed, I will use Spearman correlation test which is alternative of Pearson’s corr. test for nonnormal variables



Since p-values are smaller than 0.05, there is no correlation between variables according to transmission type.

*3.6 MODELING*

The dataset is splitted into two part which are train and test by validation set approach and by using createDataPartition function whose p is 0.8 from caret package. Logistic regression model, support vector machine (SVM), Naïve Bayes, Decision Tree, XGBoost and Neural Network are used to classify the data to predict transmission type. Tuning is applied the svm model to find the best parameters.

For Logistic Regression Model:

metin, ekran görüntüsü, yazı tipi, menü içeren bir resim

Açıklama otomatik olarak oluşturuldu

Since p values of fuel\_CNG, fuel\_LPG, sellType\_Dealer, `owner\_First Owner` , `owner\_Fourth & Above Owner`, `owner\_Test Drive Car`, `owner\_Second Owner` are greater than 0.05 and p values of fuel\_Petrol,`sellType\_Trustmark Dealer`, `sellType\_Trustmark Dealer`,`owner\_Third Owner` are NA, these columns do not affect the model hence we need to remove them.

metin, ekran görüntüsü, yazı tipi, doküman, belge içeren bir resim

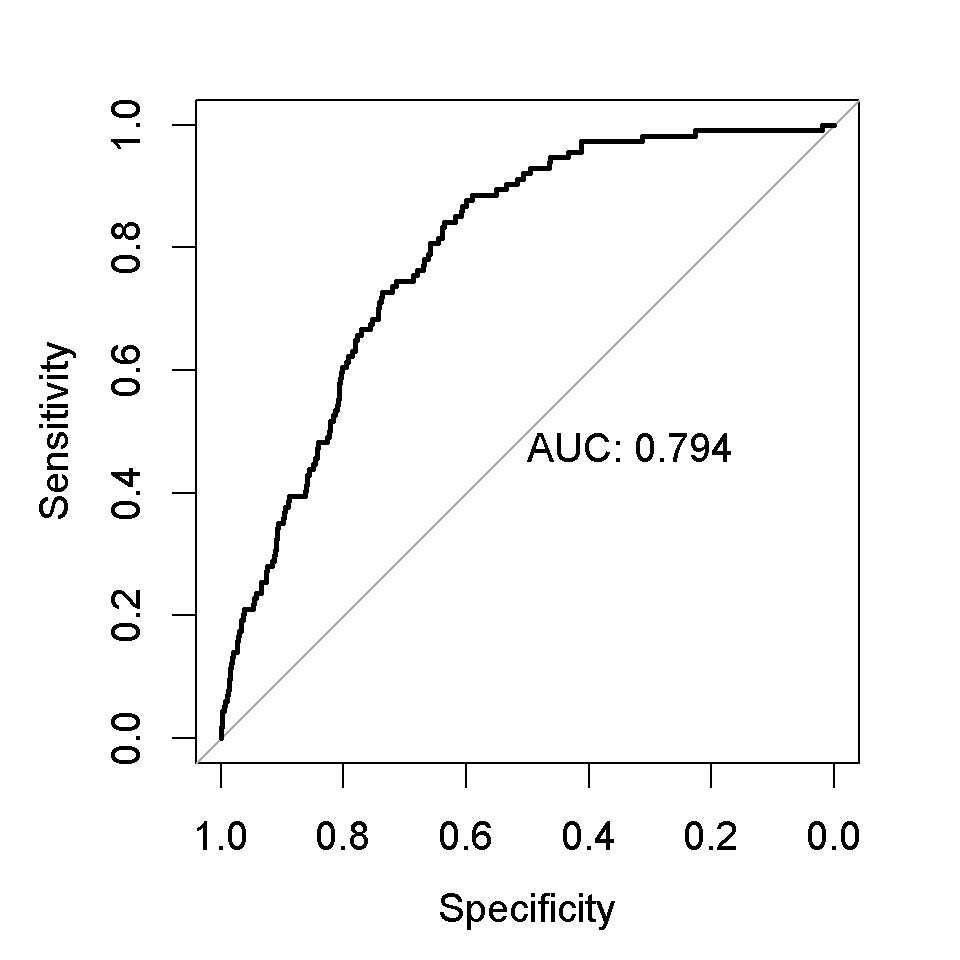
Açıklama otomatik olarak oluşturuldu

Since p values of all variables are smaller than 0.05, these columns affect the model. The performance measure tables of train and test are, respectively:

yazı tipi, metin, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu yazı tipi, metin, beyaz, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu



*Figure 10 ROC-AUC plot of Logistic model*

The ROC curve shows the model is good for the trade-off between sensitivity and specificity since the curve is close to left corner. According to AUC, degree or measure of separability is 0.794.

For SVM model:

First the model, where type is C-classification and kernel is linear, is tuned. After tuning the model, best parameters are 0.00001 for gamma, 1 for cost, and 0.01 for epsilon. The summary of the model:

metin, ekran görüntüsü, yazı tipi, cebir içeren bir resim

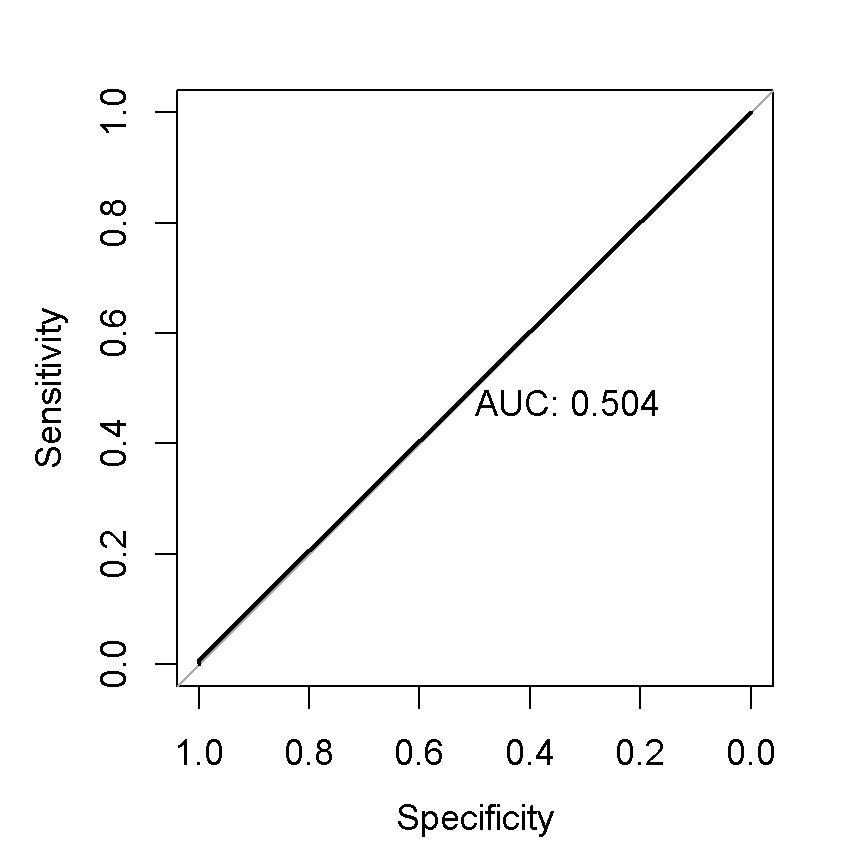
Açıklama otomatik olarak oluşturuldu

The performance measure tables of train and test are, respectively:

yazı tipi, metin, beyaz, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, yazı tipi, beyaz, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu



*Figure 11 ROC-AUC plot of SVM model*

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.504.

For Naïve Bayes model:

The summary of the model:

metin, yazı tipi, makbuz, beyaz içeren bir resim

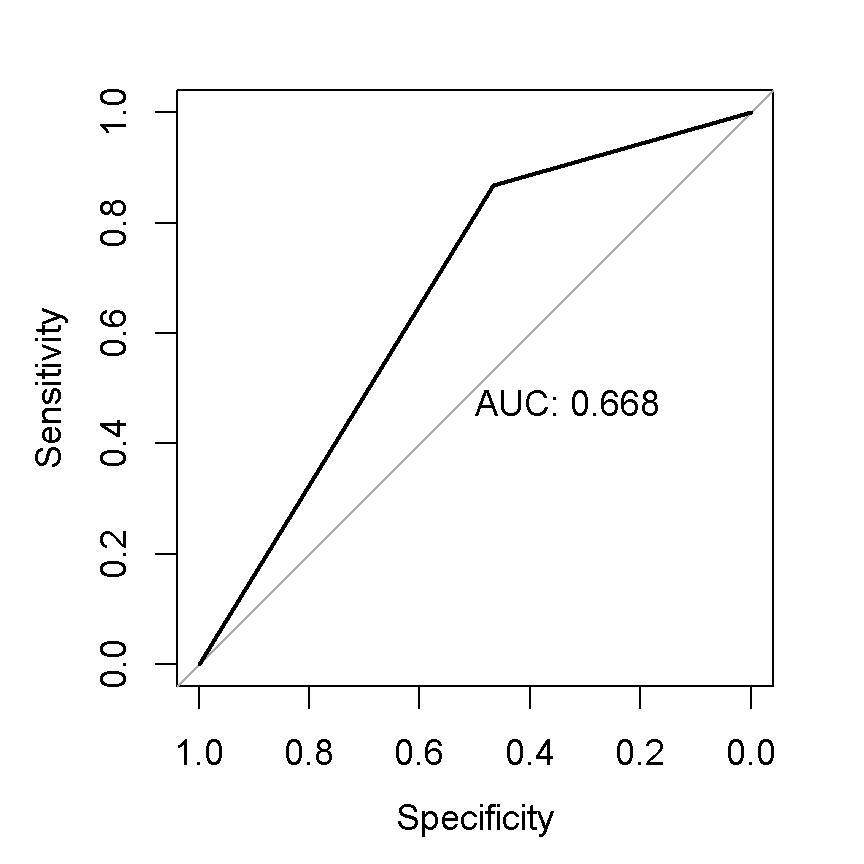
Açıklama otomatik olarak oluşturuldu

The performance measure tables of train and test are, respectively:

metin, yazı tipi, beyaz, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu yazı tipi, metin, beyaz, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu

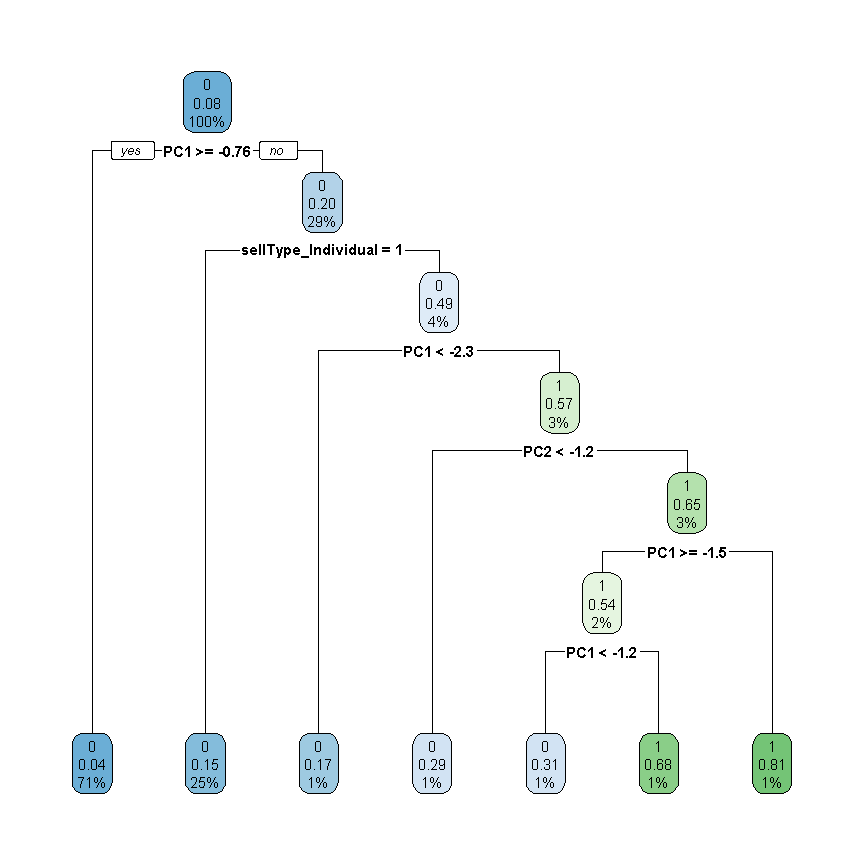


*Figure 12 ROC-AUC plot of Naïve Bayes model*

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.668.

For Decision Tree:

The plot of the decision tree:

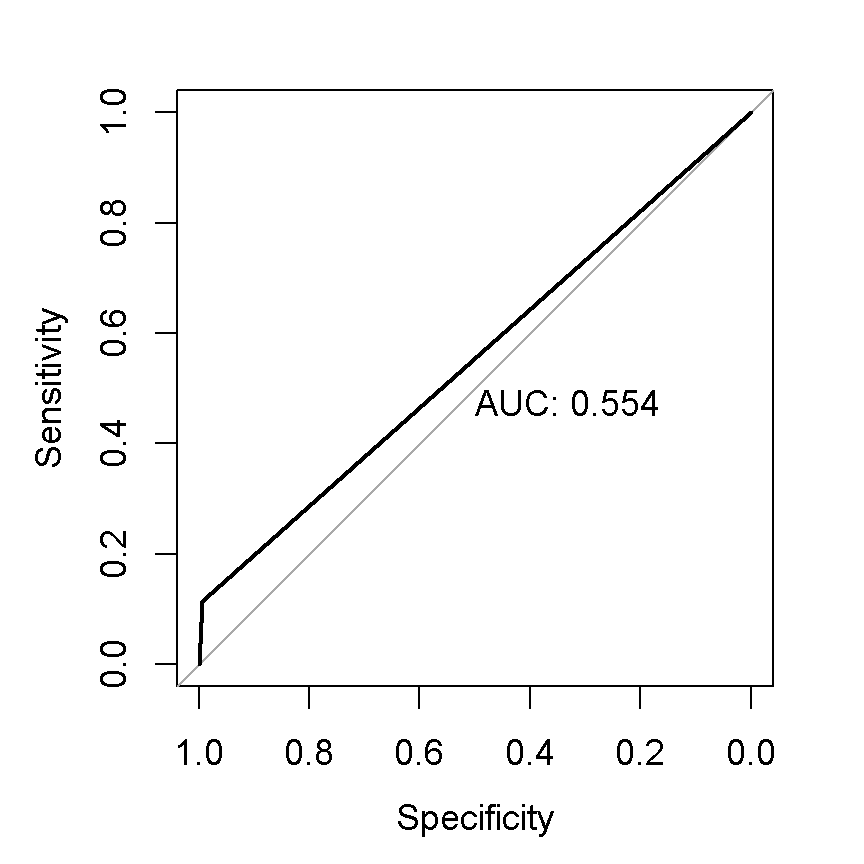


The performance measure tables of train and test are, respectively:

yazı tipi, metin, ekran görüntüsü, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, yazı tipi, beyaz, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu



*Figure 13 ROC-AUC plot of Decision Tree model*

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.554.

For XGBoost model:

The summary of the model:

metin, yazı tipi, ekran görüntüsü, makbuz içeren bir resim

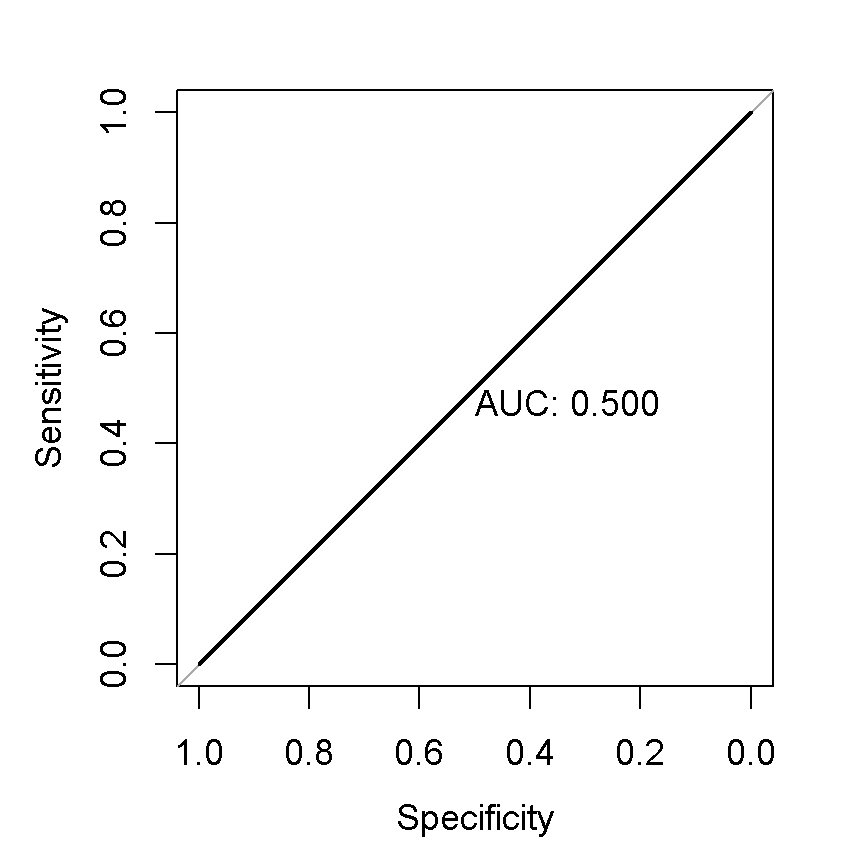
Açıklama otomatik olarak oluşturuldu

The performance measure tables of train and test are, respectively:

metin, yazı tipi, beyaz, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturulduyazı tipi, metin, beyaz, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu



*Figure 14 ROC-AUC plot of XGBoost model*

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.500.

For Neural Network model:

The summary of the model:

metin, yazı tipi, ekran görüntüsü, menü içeren bir resim

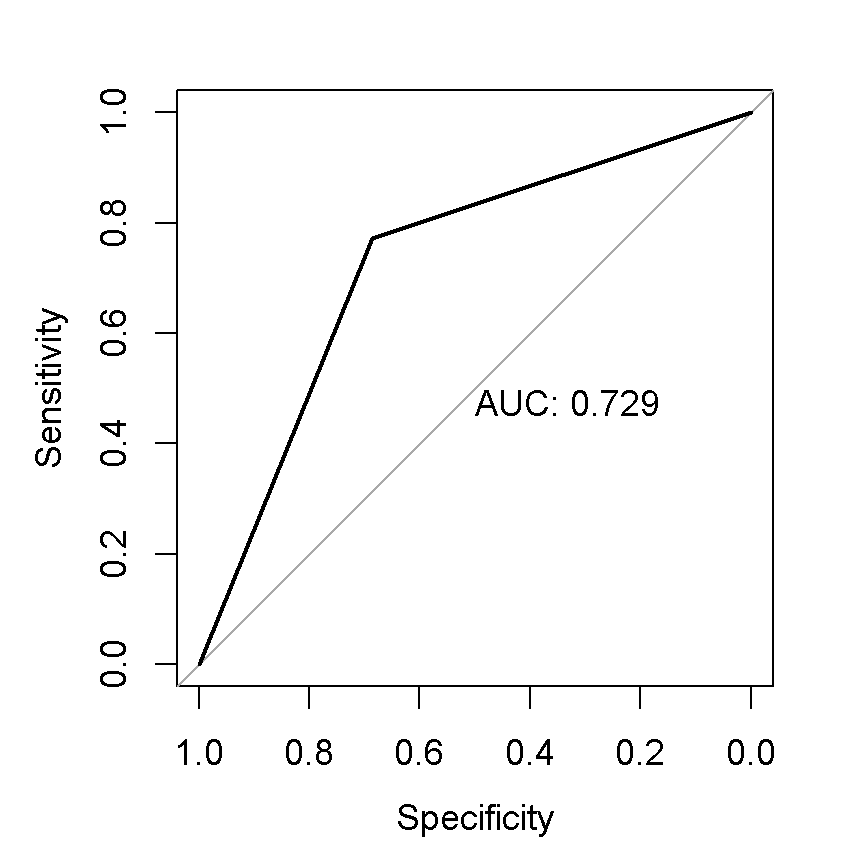
Açıklama otomatik olarak oluşturuldu

The performance measure tables of train and test are, respectively:

yazı tipi, metin, beyaz, tipografi içeren bir resim

Açıklama otomatik olarak oluşturulduyazı tipi, metin, beyaz, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu



*Figure 15 ROC-AUC plot of Neural Network model*

The ROC curve shows the model is worse than logistic model for the trade-off between sensitivity and specificity since the curve is not closer to left corner. According to AUC, degree or measure of separability is 0.729.

The performance measure metric table for the model is:

metin, makbuz, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Performance measure metric table shows us that the best model is decision tree model for both train and test data by accuracy but to predict positive values, the best model is XGBoost since the sensitivity, which is ratio of true positive value to the actual positive values, is highest for both train and test data, and SVM has best performance for train data to predict negative values since it has highest specificity of train data, which is ratio of true negatives to the actual negative values, and Decision tree has best performance for test to predict negative values since it has highest specificity of test data.

*4.CONCLUSION*

As a result, according to performance measure metric table, the best model is decision tree model for both train and test data by accuracy. However, in order to predict positive values, the best model is XGBoost since the sensitivity, which is ratio of true positive value to the actual positive values, is highest for both train and test data. Also, SVM has best performance for train data to predict negative values since it has highest specificity of train data, which is ratio of true negatives to the actual negative values, and Decision tree has best performance for test to predict negative values since it has highest specificity of test data.

*5. REFERENCES*

[1] B. KOCA, « Logistic Regression,» *Stat412 Recitation 7*, https://rpubs.com/BurcuKG/1180677

[2] B. KOCA, «Support Vector Machines,» *Stat412 Recitation 11*, <https://rpubs.com/BurcuKG/1185743>

[3] KOCA, «Naïve Bayes Classifier, Decision Tree,» *Stat412 Recitation 12*, https://rpubs.com/BurcuKG/1187989

[4] B. KOCA, «XGBoost,» *Stat412 Recitation 13*, <https://rpubs.com/BurcuKG/1190267>

[5] Z. [BOBBITT](https://www.statology.org/author/admin/), «Kolmogorov-Smirnov Test in R (With Examples),»

https://www.statology.org/kolmogorov-smirnov-test-r/

[6] G. Barabás, «13.2 The Scheirer–Ray–Hare test,» *13 Two-way ANOVA and the Scheirer–Ray–Hare test,* <https://dysordys.github.io/data-with-R/Anova_two_way.html>

[7] amiyaranjanrout, *Spearman Correlation Testing in R Programming,* <https://www.geeksforgeeks.org/spearman-correlation-testing-in-r-programming/>