

**Predictive Model for mode of transport used by Employees**

GROUP ASSIGNMENT

Build a predictive model to ascertain if an employee will use car as mode of transport to commute to office and which variables are significant predictors behind this decision.

**CP G1 - PGP BABI Apr’19**

**Amit Sharma**

**Amitesh Bajpai**

**Sahil Sachdeva**

**Kushal Maheshwari**

**Neeraj Singh**

# Project Objective

The objective of the report is to build the best predictive model which can predict whether or not an employee will use Car as a mode of transport based on the “**Cars.csv**” in R, reflect upon the performance of the various models and find the best model and which variables are a significant predictor behind this decision. This exploration report will consist of the following:

* + Importing the dataset in R
  + Understanding the structure of dataset
  + Graphical exploration
  + Descriptive statistics
  + SMOTE
  + Logistic Regression
  + KNN
  + Naïve Bayes
  + Bagging
  + Boosting
  + Insights from the dataset

# Assumptions

The following assumptions are made for the inferential statistics:

1. Observations are independent
2. Samples are random
3. Measurements are accurate
4. For Naïve Bayes: The variables are independent and are equally important.
5. For KNN: We normalize the continuous variables

# Exploratory Data Analysis – Step by step approach

The various steps followed to analyze the case study is mentioned and explained below.

## Environment Set up and Data Import

### Install necessary Packages and Invoke Libraries

The lists of R packages used to analyze the data are listed below:

* + - * readr to Read csv data file
      * corrplot library for correlation
      * lattice for plots
      * caret to calculate confusionMatrix
      * ROCR to calculate auc, KS
      * ineq to calculate gini
      * caTools to Split data
      * naivebayes for Naive Bayes model for Numeric Predictors
      * e1071 For Naise Bayes
      * class For KNN Classifier
      * pscl to Maximum likelihood estimation
      * lmtest for diagnostic checking in linear regression models
      * purrr for Visualization
      * tidyr for Visualization
      * ggplot2 for Data Visualization
      * car for vif
      * DMwR to treat missing values by filling in NA values with the values of the nearest neighbours
      * ipred for bagging
      * gbm for boosting
      * rpart for bagging
      * vif for variable importance

### Set up working Directory

Setting up the working directory will help to maintain all the files related to the project at one place in the system.

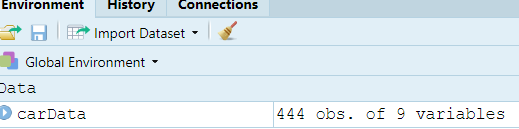
The working directory, we have setup in a local folder in the laptop.



Please refer Appendix A for Source Code.

### Import and Read the Dataset

The given datasets are in “.csv format, so to import the data in R we use the “read.csv” command. Data in file “Cars.csv” is stored in a variable called “**carData**”.



Please refer Appendix A for Source Code.

## Variable Identification

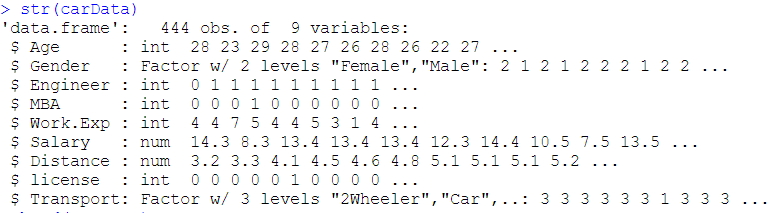
* **dim** : to check dimension (#rows/columns) of a data frame
* **str** : Display internal structure of an R object
* **head :** it will show the first n rows of a data frame or matrix in R(default is 6)
* **tail:** to verify the last 6 records of the data set
* **summary:** It gives the 5-number summary, basically the 5 statistical values, namely the minimum value, the first quartile, the median, the third quartile, and the maximum value of a data set
* **as.factor:** To convert variable to factor
* **as.data.frame:** To convert to data frame
* **histogram:** to compute histogram of the variables
* **boxplot:** to draw box plot which shows 5-number (mean, quartiles)
* **is.na:** to check if there is any missing value
* **sapply:** to apply is.na to all the objects parameter to each column of sub-data frame defined by the by input parameter
* **ifelse:** to convert variable value in 0 and 1

### Variable Identification – Inferences

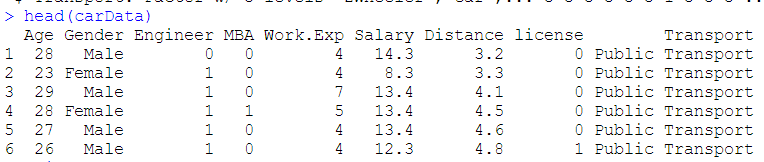
* **DIM**
  + carData data frame : There are 444 rows and 9 columns



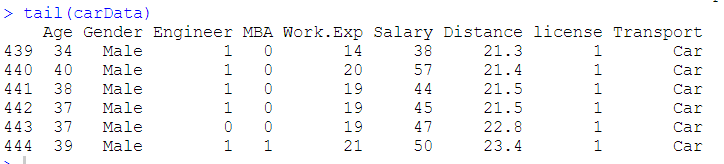
* **STR**
  + There are 9 variables in the dataset.



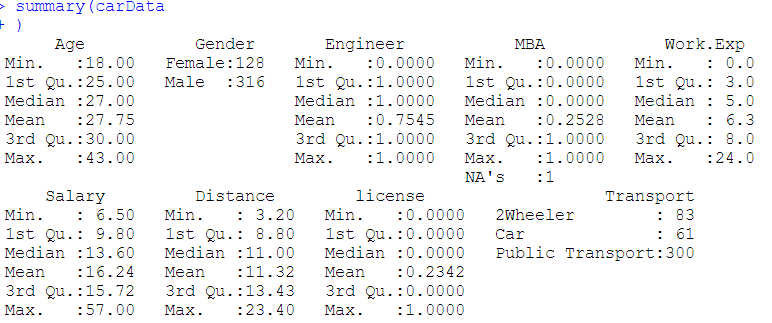
* **HEAD**
  + carData data frame: Verifying head records



* **TAIL**
  + carData data frame: Verifying tail records



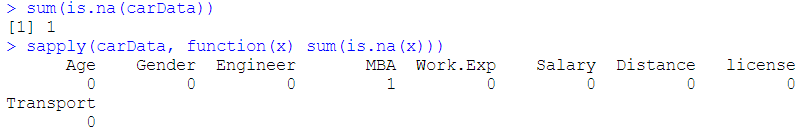
* **SUMMARY**
  + carData data frame: The variables, namely, Engineer, MBA and license should be factor. So, we will convert them to factor. The continuous variables have outliers



Please refer Appendix A for Source Code.

## Missing Value Identification

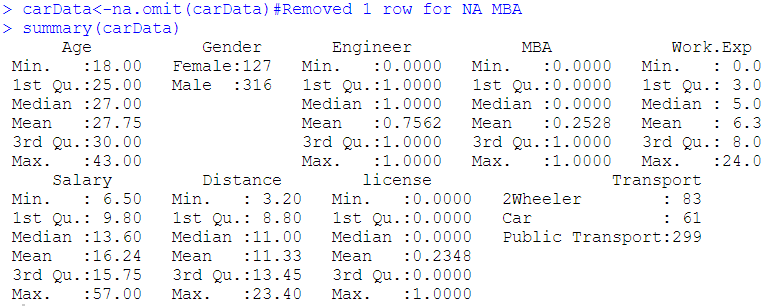
We use ‘is na’ function to check if there are any missing values. There are 1 missing value in MBA variable.



Please refer Appendix A for Source Code.

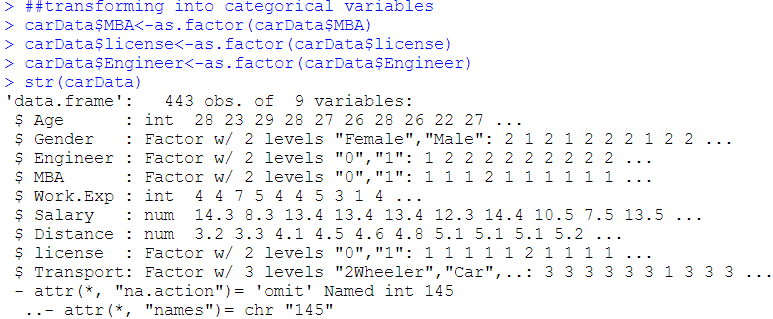
### Missing Value Treatment

As there is only one record for missing value, we can remove this.



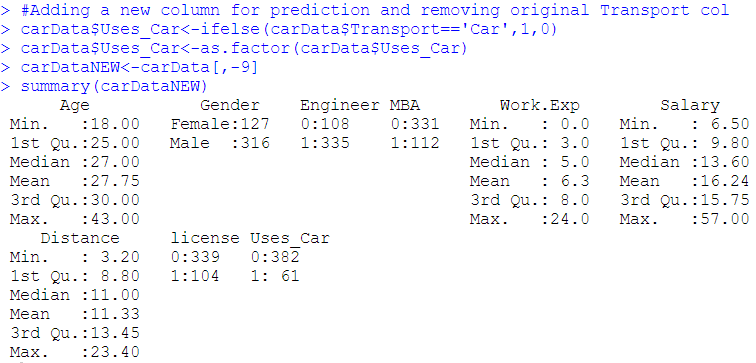
## Variable Transformation / Feature Creation

In Summary of data we have seen that the "Engineer", "MBA" and "license" variables should be factor. So we use ‘ as.factor ‘ to convert them.



Please refer Appendix A for Source Code.

Creating a new variable **Uses\_Car** as we have to predict if employee will use car or not as mode of transport. We drop the original Transport variable and create a new data set “carDataNEW”.

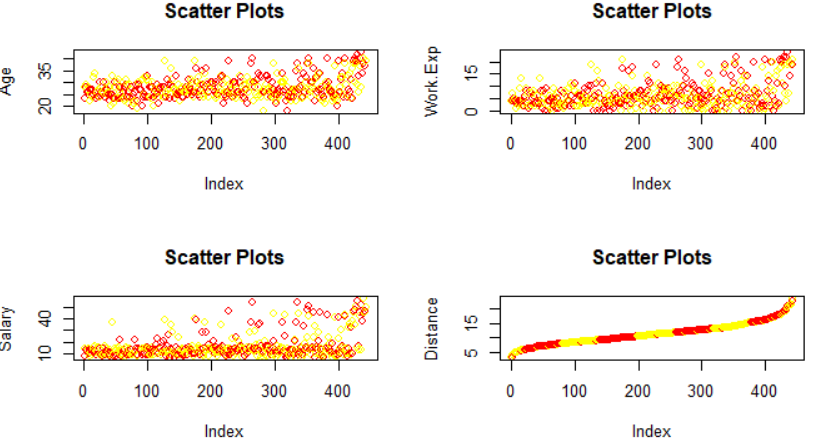


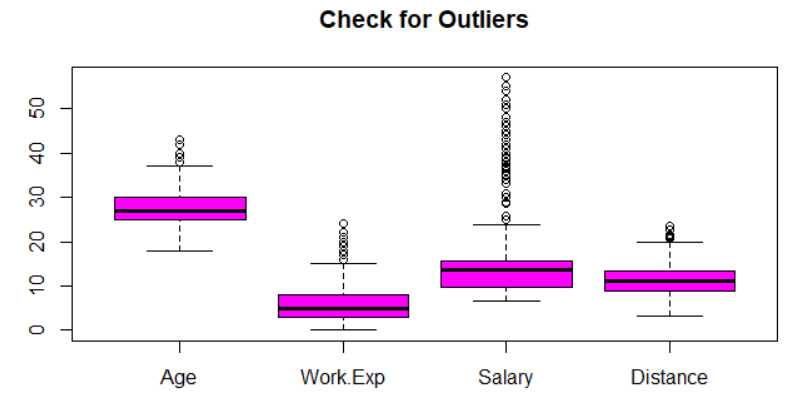
## Univariate Analysis

We are analyzing all the 8 independent variables from data set ‘carDataNEW’. The Uses\_Car variable is the dependent variable. For easy in plotting we convert the dataset to data frame **‘carDataEDA’** and remove the factor variables**.** Then we perform Univariate on the numeric variables and Bivariate analysis.

* **Age and Work.Exp has** mean and median almost the same.
* **Age and Distance** has nearly normal distribution
* **Salary and Work.Exp** are right skewed. Most of the employees have less than 5 units of experience and less than 11 units of salary (unit of variables not provided).
* The box-plot shows there are **outliers** in all the continuous variables.
* The scatter plot shows that there is **random distribution in all variables except Distance.**
* From Density plot we see that **Engineer, MBA and license are discrete variables**





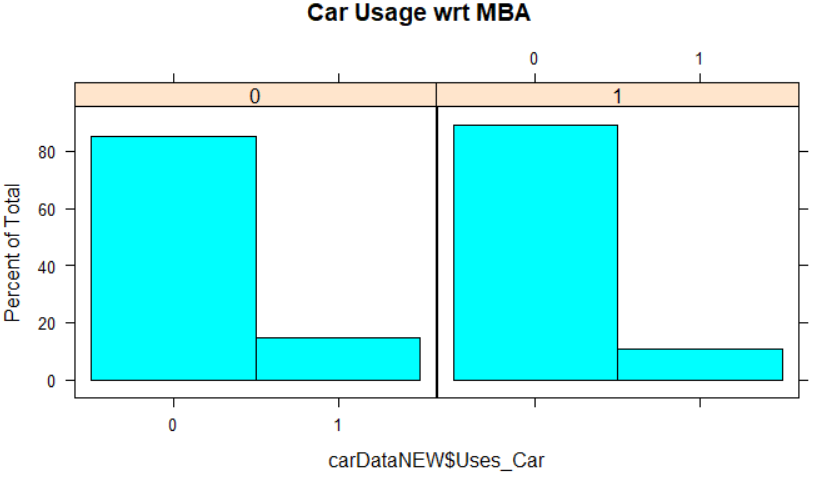


Please refer Appendix A for Source Code.

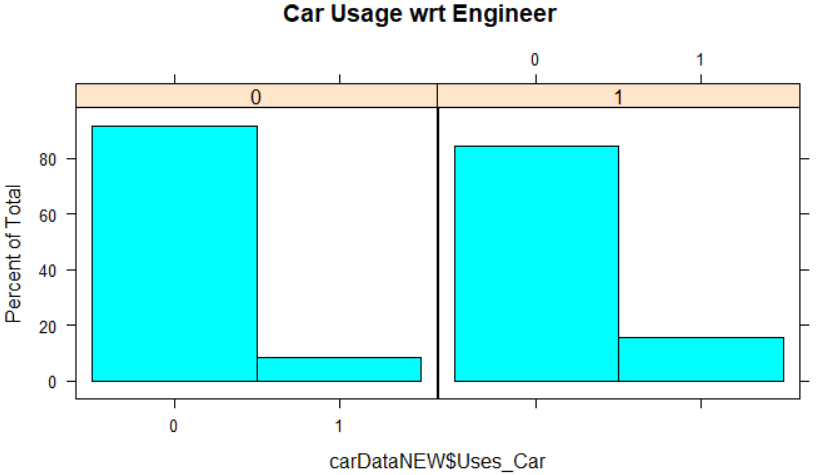
## Bi-Variate Analysis

We will analyze Uses\_Car with the other variables from data set ‘CarDataNEW’.

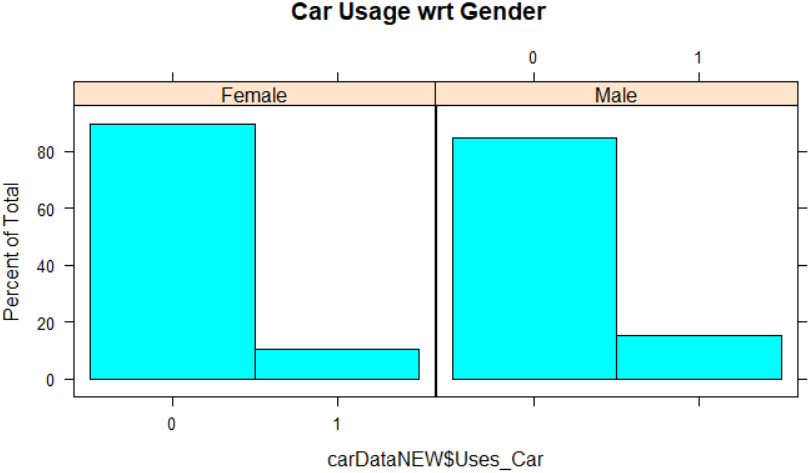
From the plot we can say that MBA is not a significant parameter to predict if employee uses a car as mode of transport.



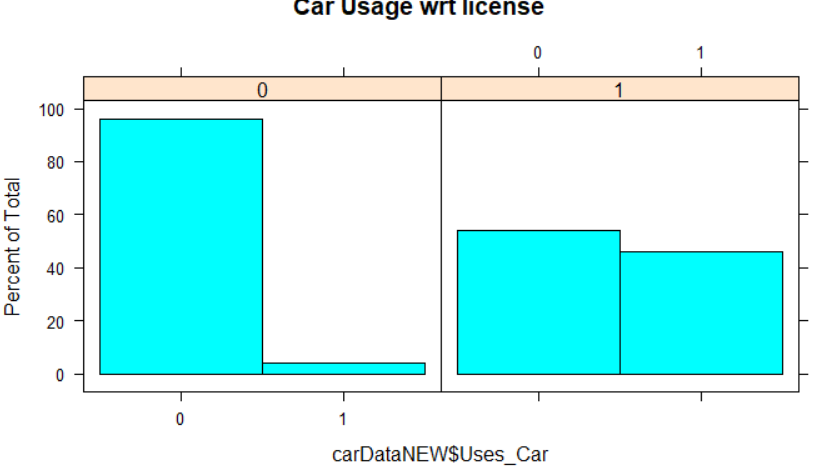
From the plot we can say that employee being an engineer is a significant parameter to predict if employee uses a car as mode of transport.



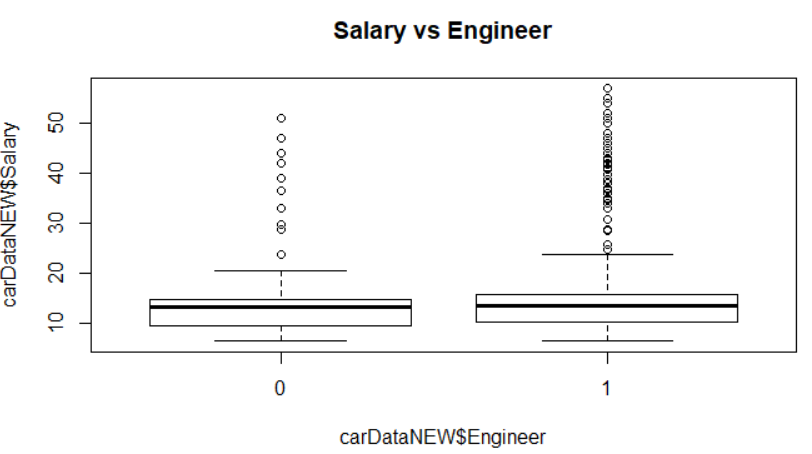
From the plot we can say that gender is a significant parameter to predict if employee uses a car as mode of transport. Car usage is higher among Male compared to female employees.



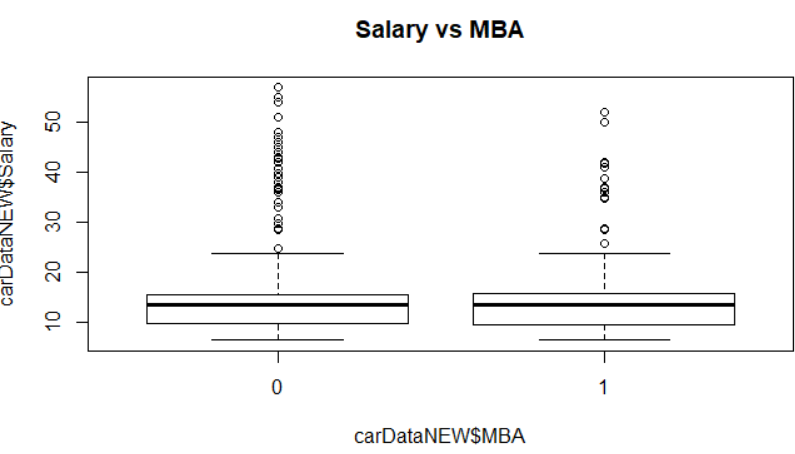
From the plot we can say that having driving license has significant influence on employee using a car as mode of transport.



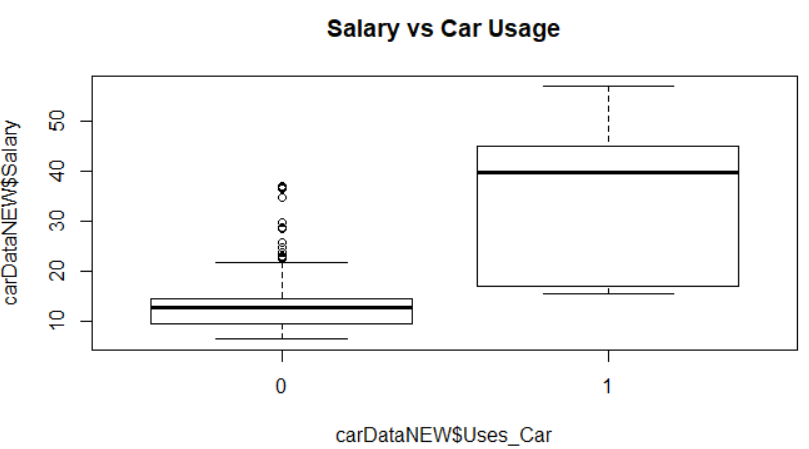
As per the plot below, median for engineer and non-engineer are nearly same but some engineers have higher salary



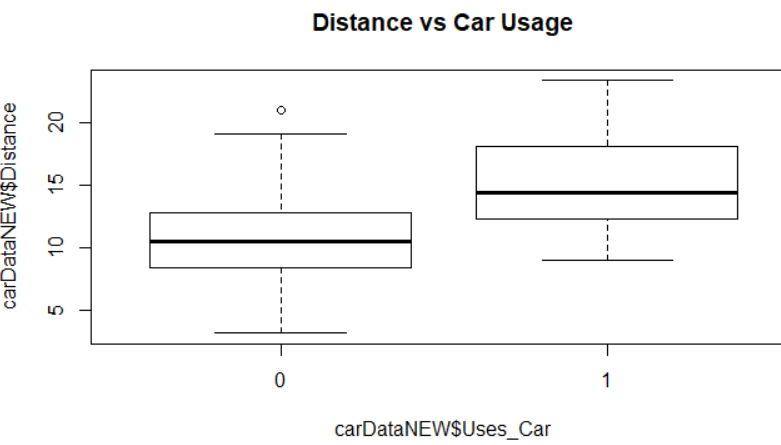
As per the plot below, median for MBA and non-MBA are nearly same but some MBAs have higher salary



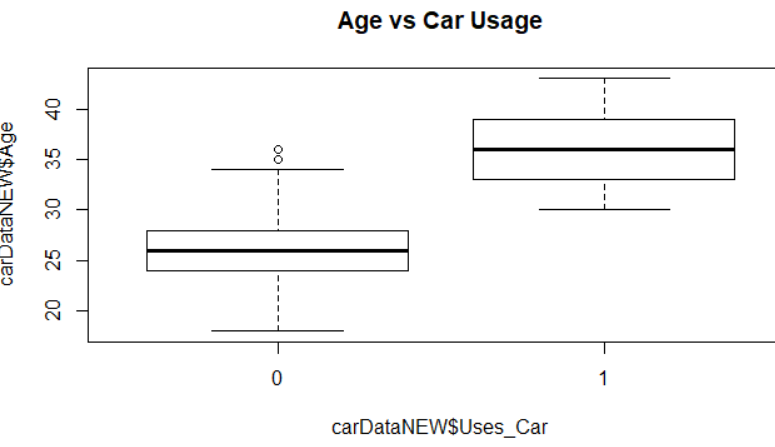
As per the plot below, employees with higher salary are more likely to use car. Employees who salary is less than 30 unit are least likely to use car as mode of transport



As per the plot below, employees who travel more distance are more likely to use car. Employees who travel distance less than 15 unit are least likely to use car as mode of transport.



As per the plot below, older employees are more likely to use car. All Employees below 30 units do not use Car as mode of transport whereas all employees above 36 units uses Car



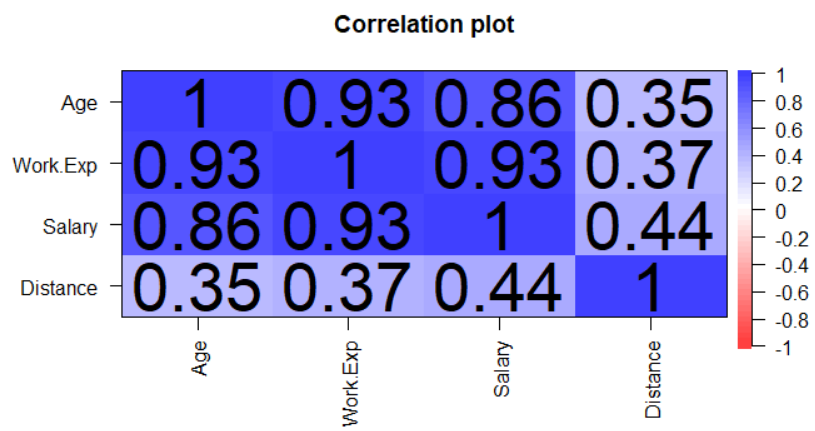
Please refer Appendix A for Source Code

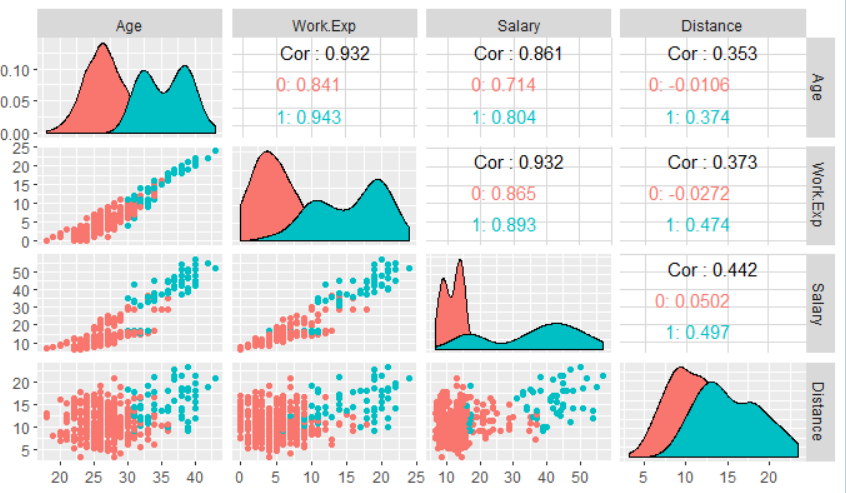
## Outlier Identification

There are outliers in all numeric variables. It is evident from the box plot as well. But all the values are possible values. The outliers do not seem to be due to some error.

Please refer Appendix A for Source Code.

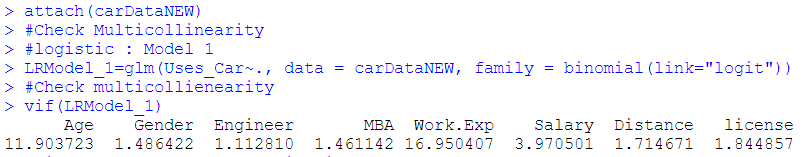
## Correlation/Multicollinearity



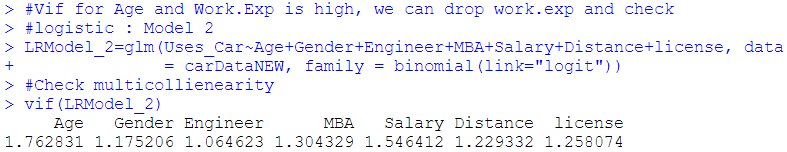


Based on the above plot we can say Age and Work.Exp are highly correlated as well as Work.Exp and Salary are highly correlated

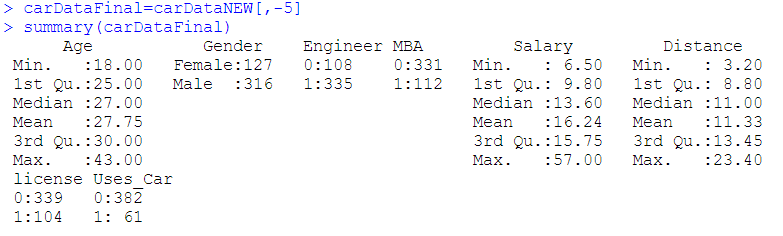
And Age and Salary are also correlated. Hence, **we will check the multicollinearity**.



We will drop Work.Exp and check multicollinearity again.



Hence we drop Work.Exp to treat multicollinearity and create a new dataset “**carDataFinal**”



Please refer Appendix A for Source Code.

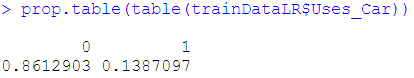
# SMOTE

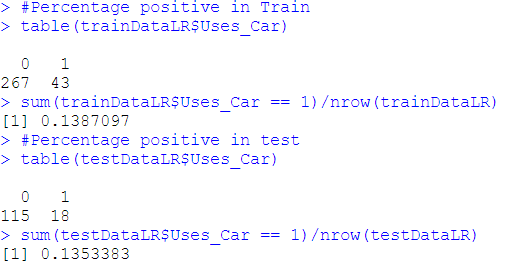
The SMOTE function oversamples your rare event by using bootstrapping and k-nearest neighbor to synthetically create additional observations of that event.

We smote the train data so that the model we build is accurate and robust

## Data Imbalance

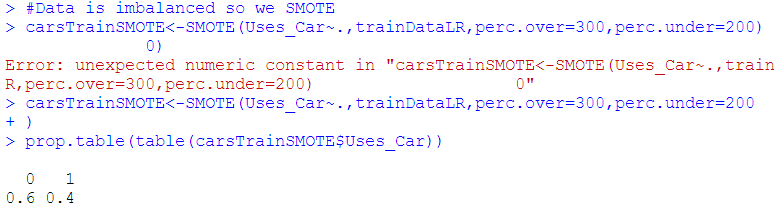
We see that the data is highly imbalanced





## Smote

We increase the ratio of 0:1 to nearly 60:40. We will build all models, except bagging and boosting, on Smote data.



# Logistic regression

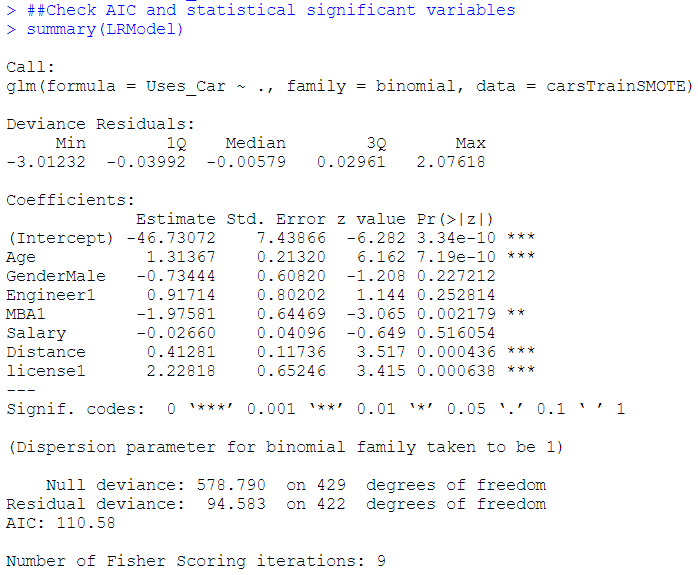
Logistic regression is part of the supervised learning. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

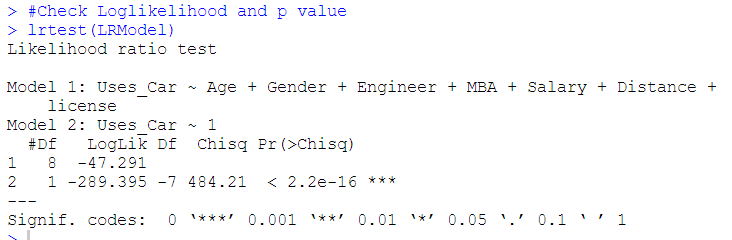
The independent variable, Transport is multilevel. Since we are concerned if employee use a car or not, we have created a new variable Uses\_Car with two values 0- Other than Car and 1 –Car.

We can scale the data to reduce the impact of outliers. While model building, we have checked with scaled data as well but there was no impact on the model due to scaling. Hence, we are not scaling the data.

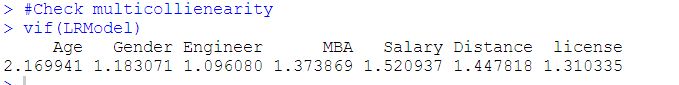
## Model Building

We will build is with all the variables on the smote data and we will check the multicollinearity





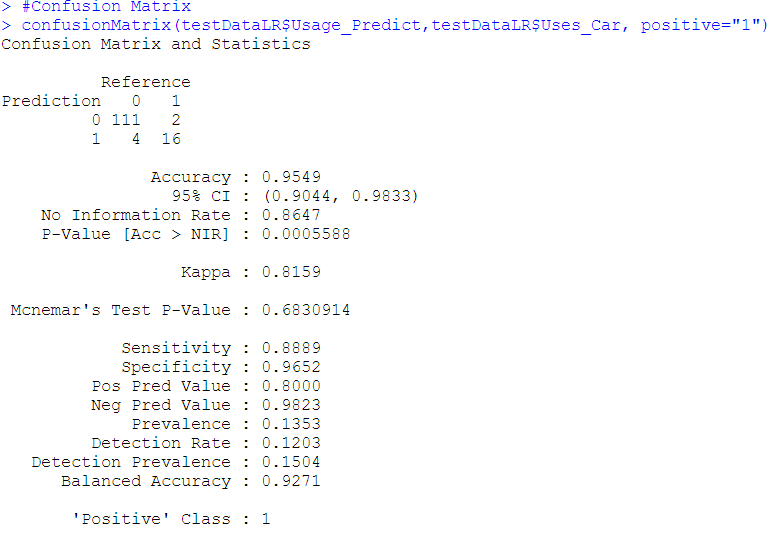
There is significant change in log likelihood from the base model. Also based on the p-value we can reject the null hypothesis. Thus, the model is valid.

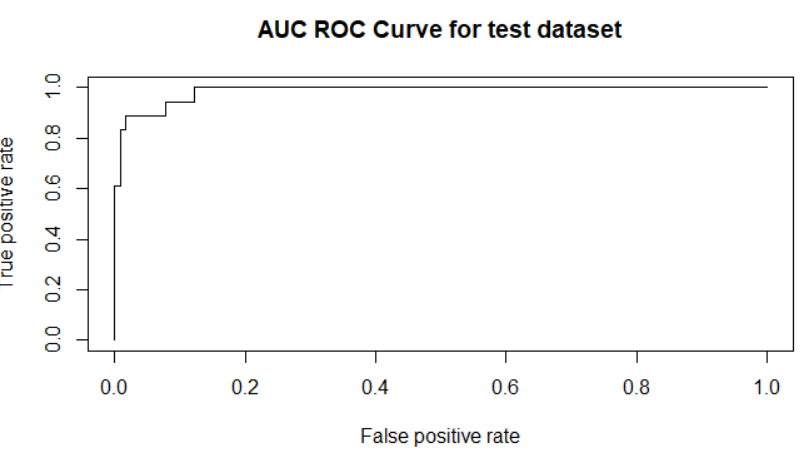


VIF is also below 5 for all variables. Please refer Appendix A for Source Code.

## Performance Metrics

**Confusion Matrix: For testing Dataset**





|  |  |
| --- | --- |
| **Metrics** | **Value for Testing Dataset** |
| Accuracy | 0.9549 |
| Sensitivity | 0.8889 |
| Specificity | 0.9652 |
| AUC | 0.9859 |
| K-S | 0.8782 |
| Gini | 0.8374 |

Please refer Appendix A for Source Code.

## Interpretation

Based on the performance metrics of the model on testing data, we can say the model is good. Based on the test metrics we can interpret that:

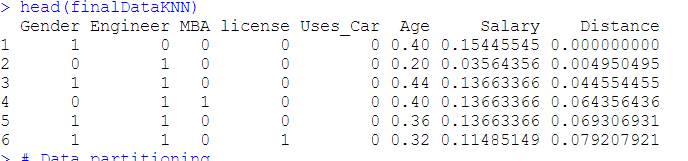
* + 1. The model will catch 88% of the employees who will use car as mode of transport.
    2. The model will catch 96% of the employees who will not use car as mode of transport
    3. Overall all accuracy is 95%
    4. Out of the employees it predicted as will use car, 80% of them will actually use car as mode of transport
    5. Out of the employees it predicted as will Not use car, 98% of them will actually not use car as mode of transport
    6. AUC is about 98%, so it is a very good classifier.
    7. K-S is 87%, the model will fairly perform to separate the employees who will use car and who will not use car as mode of transport.

# K-Nearest Neighbour

KNN which stand for K Nearest Neighbor is a Supervised Machine Learning algorithm that classifies a new data point into the target class, depending on the features of its neighboring data points.

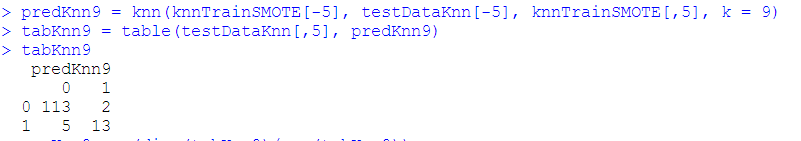
We tried the model for scaled and normalized data both. Checking the output, we have built the final model on normalized data. And also convert Gender in 0 and 1.(Male-1, Female-0)

Normalized data:



## Model Building

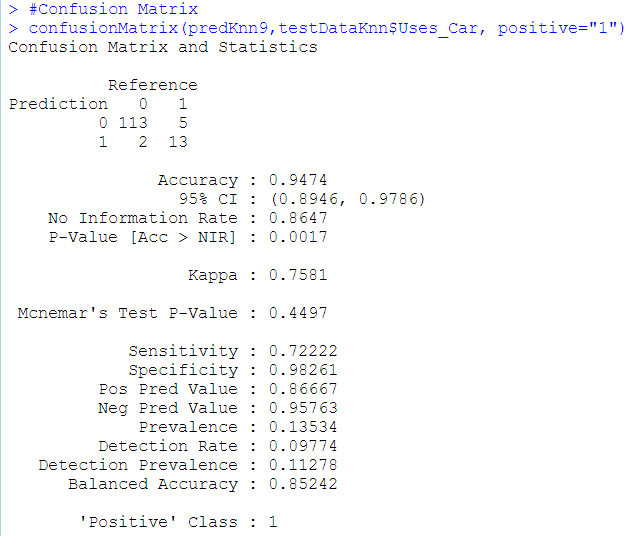
We have built the model for various values of K and found K=9 as the optimal value.



Please refer Appendix A for Source Code.

## Performance Metrics

**Confusion Matrix**



|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Accuracy | 0.9474 |
| Sensitivity | 0.7222 |
| Specificity | 0.9826 |
| AUC | NA |
| K-S | NA |
| Gini | NA |

## Interpretation

The accuracy, sensitivity and specificity of the model are high. Based on the test metrics we can interpret that

* + 1. The model will catch 94% of the employees who will use car as mode of transport.
    2. The model will catch 94% of the employees who will not use car as mode of transport
    3. Overall all accuracy is 94%
    4. Out of the employees it predicted as will use car, 86% of them will actually use car as mode of transport
    5. Out of the employees it predicted as will Not use car, 95% of them will actually not use car as mode of transport

# Naive Bayes

Naïve Bayes is a Supervised Machine Learning algorithm that classifies a new data point into the target class using Bayes’ theorem and assuming all the predictors are independent to each other.

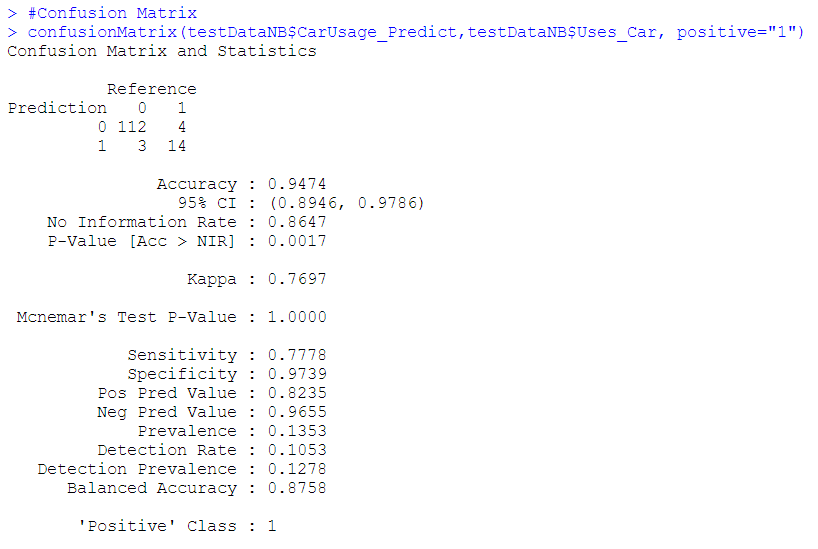
## Model Building

## 

Please refer Appendix A for Source Code.

## Performance Metrics

**Confusion Matrix for Testing Dataset**



|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Accuracy | 0.9474 |
| Sensitivity | 0.7778 |
| Specificity | 0.9739 |

* 1. **Interpretation**

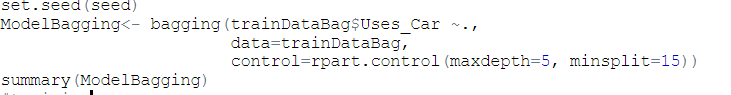
Based on the test metrics we can interpret that.

* + 1. The model will catch 77% of the employees who will use car as mode of transport.
    2. The model will catch 97% of the employees who will not use car as mode of transport
    3. Overall all accuracy is 95%
    4. Out of the employees it predicted as will use car, 82% of them will actually use car as mode of transport
    5. Out of the employees it predicted as will Not use car, 96% of them will actually not use car as mode of transport

1. **Bagging**

Bagging is also called as Bootstrap Aggregating. It is an ensemble machine learning algorithm designed to improve accuracy and stability of algorithm used in statistical classification and regression by reducing variance and avoiding overfitting.

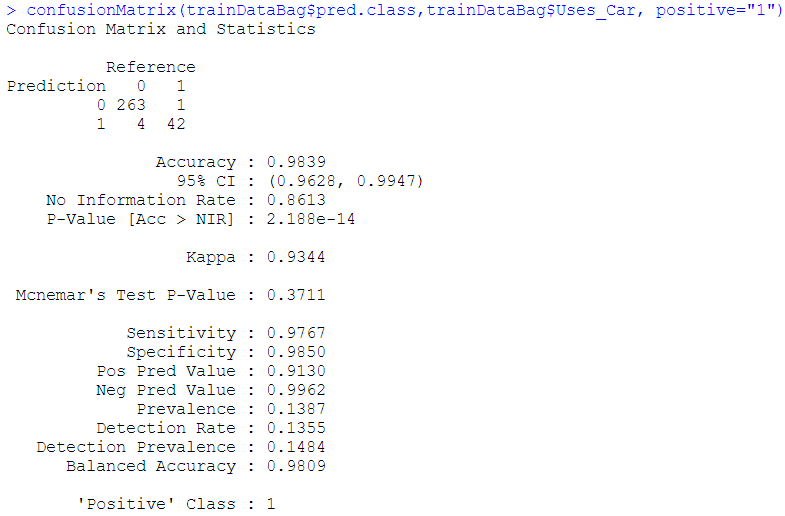
* 1. **Model Building**



Please refer Appendix A for Source Code.

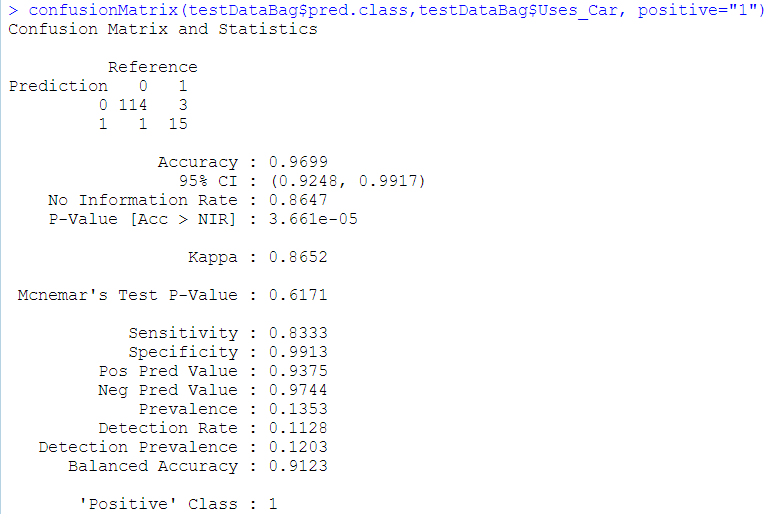
* 1. **Performance Metrics**

**Confusion Matrix for Train Dataset**



|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Accuracy | 0.9839 |
| Sensitivity | 0.9767 |
| Specificity | 0.9850 |

**Confusion Matrix For Testing Dataset**



|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Accuracy | 0.9699 |
| Sensitivity | 0.8333 |
| Specificity | 0.9913 |

* 1. **Interpretation**

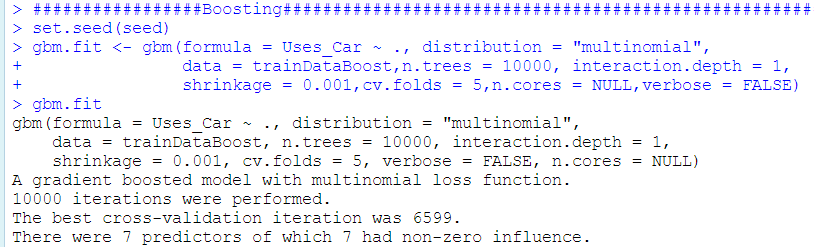
The model is stable as evident from the output of confusion matrix for training and testing dataset. Based on the test metrics we can interpret that

* + 1. The model will catch 83% of the employees who will use car as mode of transport.
    2. The model will catch 99% of the employees who will not use car as mode of transport
    3. Overall all accuracy is 97%
    4. Out of the employees it predicted as will use car, 93% of them will actually use car as mode of transport
    5. Out of the employees it predicted as will Not use car, 97% of them will actually not use car as mode of transport

1. **Boosting**

Boosting is another ensemble algorithm which is used to reduce bias and also variance, in supervised learning. In ensemble algorithm, set of weak learners are combined to form strong learner.

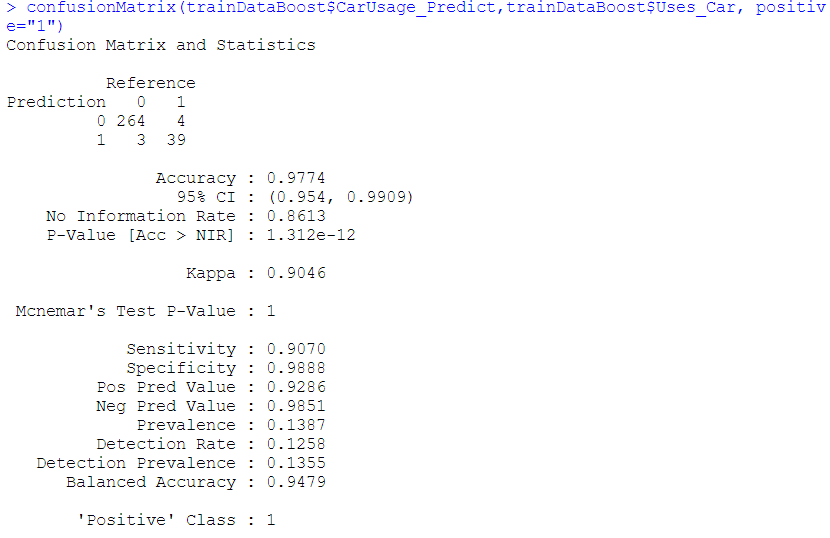
* 1. **Model Building**



Please refer Appendix A for Source Code.

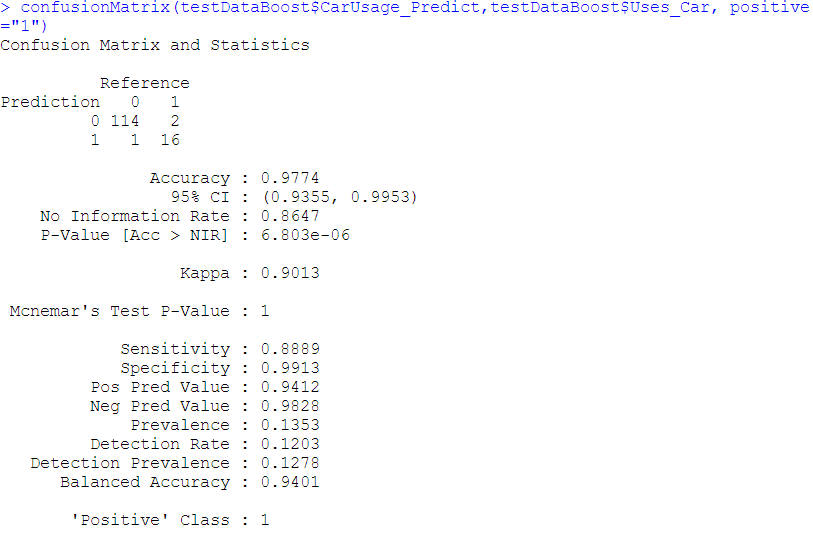
* 1. **Performance Metrics**

**Confusion Matrix for train dataset**



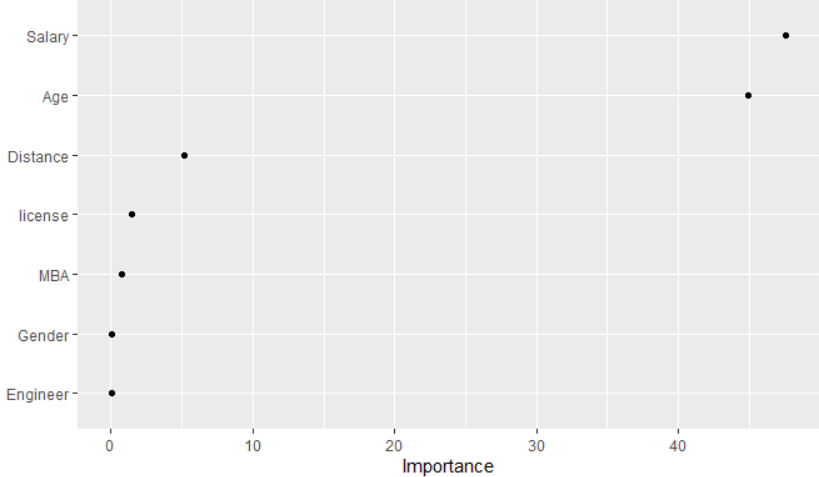
|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Accuracy | 0.9774 |
| Sensitivity | 0.9070 |
| Specificity | 0.9888 |

**Confusion Matrix for Testing Dataset**



|  |  |
| --- | --- |
| **Metrics** | **Value** |
| Accuracy | 0.9774 |
| Sensitivity | 0.8889 |
| Specificity | 0.9913 |

* 1. **Variable Importance**



We see that Salary has highest significance in the model followed by Age, Distance.

* 1. **Interpretation**

The model is stable as evident from the output of confusion matrix for training and testing dataset. Based on the test metrics we can interpret that

* + 1. The model will catch 88% of the employees who will use car as mode of transport.
    2. The model will catch 99% of the employees who will not use car as mode of transport
    3. Overall all accuracy is 98%
    4. Out of the employees it predicted as will use car, 94% of them will actually use car as mode of transport
    5. Out of the employees it predicted as will Not use car, 98% of them will actually not use car as mode of transport

1. **Model Comparison**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Performance Measure** | **Logistic**  **Regression** | **KNN** | **Naïve Bayes** | **Bagging** | **Boosting** |
|  | **Test Dataset** | **Test Dataset** | **Test Dataset** | **Test Dataset** | **Test Dataset** |
| **Confusion Matrix :**  **Accuracy** | 0.95 | 0.95 | 0.95 | 0.97 | 0.98 |
| **Confusion Matrix :**  **Sensitivity** | 0.89 | 0.72 | 0.78 | 0.83 | 0.89 |
| **Confusion Matrix :**  **Specificity** | 0.97 | 0.98 | 0.97 | 0.99 | 0.99 |

* Bagging and Boosting have the highest accuracy and have highest specificity.
* For Naïve Bayes, the base assumption is that the predictor variables are independent and equally important. For our data, we have seen that the predictors are correlated. Hence, we can say that Naïve Bayes, is not giving correct prediction.
* KNN does not give the confidence level (probabilities). It gives the class value directly. Though sensitivity is good but for other models’ sensitivity is quite similar or more.
* Out of Logistic regression, KNN and Naïve Bayes, ***Logistic Regression Model*** has the highest Accuracy and specificity. Hence, we conclude that Logistic regression model is the best among the three.
* Comparing Logistic Regression with Bagging and Boosting, we can conclude that the model developed using ***Bagging and Boosting are the best***. The performance of bagging and boosting are almost same. The accuracy and specificity are the highest.

1. **Conclusion**

We have built various models to understand the factors which influence the use of cars as a mode of transport and also explain the employee’s decision to use cars as the means of transport.

The model built using ensemble technique (Bagging and Boosting) is a good model as accuracy is about 98% and there is balance between sensitivity and specificity. Also, the model is quite stable. We are able to predict nearly 100% of the employees who will use car as mode of transport.

Key insights are –

1. Employees with higher salary tend to use car as mode of transport. It is most significant variable which influence the mode of transport used by the employees. It is evident from the bivariate analysis as well as from the variable importance for Boosting.
2. Employees who travel more distance use car as the mode of transport.
3. Older employees use car to commute to their office.
4. Gender and education do not have much influence on employees to use car as mode of transport to commute to their office.

If the report is to be used by administrator of public transport then we should include time taken to travel as a variable so that we have better understanding if distance and time are directly proportional and work on improving efficiency and reliability of public transport. They should also make travelling by public transport more convenient for elderly people.

If the report is to be used by internal administration department of the company to introduce company provided transport then further analysis is to be done and data will be required, like amount spent by employee to travel to office, time taken to travel etc.

1. **Appendix A – Source Code**

****