

**Group Assignment**

**Machine Learning- using R**

**Post-Graduate programme in Business Analytics & Business Intelligence**

**Batch-October 2019.**

**By: Group 4**

**Ashish Nath Km**

**Athul Aravind PM**

**Indu Mohan**

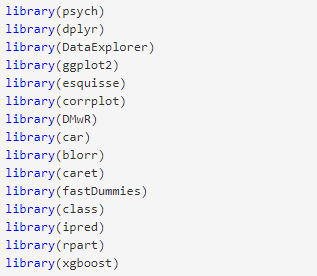
**Karthik R**

**Nishitha Ramesh**

**CAR USAGE PREDICTION CASE STUDY**

This case study attempts to understand the mode of transport that an employee prefers to commute to office. The main objective of the case study is to predict whether an employee prefers car as the mode of transport or not.

For the analysis of the problem of Cars dataset, we have utilised the below R-packages from the library.



The data “Cars\_edited.csv” is a comma separated file that includes employee information about their mode of transport as well as their personal and professional details like age, salary, work exp etc.,

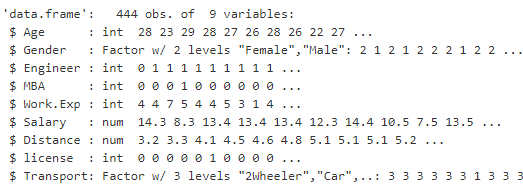
Let us take a look into the variables that are part of the Cars dataset.

|  |  |
| --- | --- |
| Age | Age of the Employee in Years |
| Gender | Gender of the Employee |
| Engineer | For Engineer =1 , Non Engineer =0 |
| MBA | For MBA =1 , Non MBA =0 |
| Work Exp | Experience in years |
| Salary | Salary in Lakhs per Annum |
| Distance | Distance in Kms from Home to Office |
| license | If Employee has Driving License -1, If not, then 0 |
| Transport | Mode of Transport |

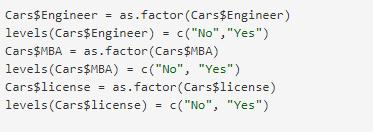
1. **Exploratory Data Analysis**

**Structure of the data**

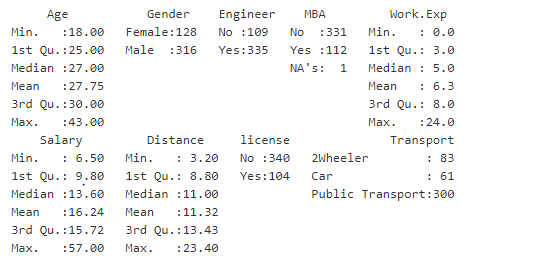
The ‘Cars\_edited.csv’ consists of 444 observations and 9 variables.  
Gender and Transport are factors. Age, Engineer, MBA, Work.Exp, license are integers. Salary and Distance are numerical



Engineer, MBA and license is fed as integer type, hence converting them to factors and adding levels.



**Summary of the data**



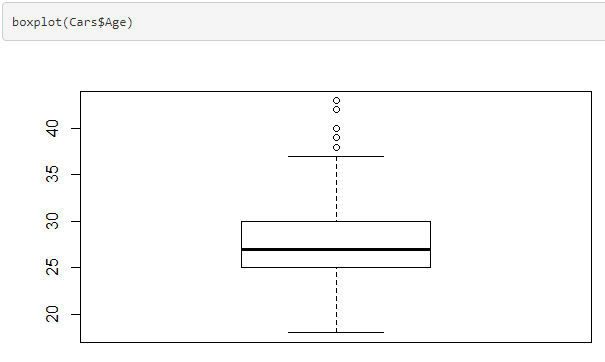
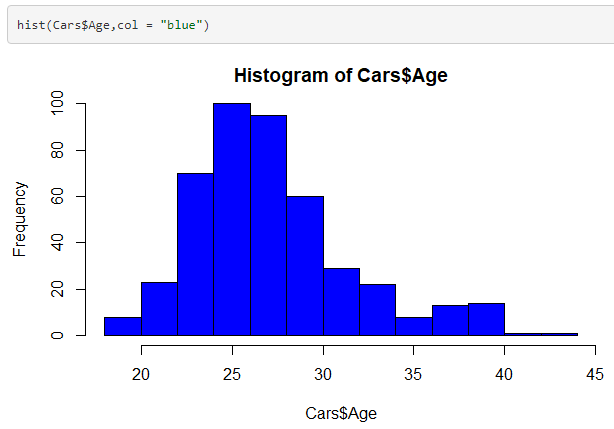
**Observation:**

* Public Transport is the most preferred ways of commute by the employees, 2-wheeler being second and followed by Cars.
* Majority of the employees are Males i.e., around 71%.
* Engineers have outnumbered the MBA graduates
* There is a missing value in MBA

**Univariate and Bivariate analysis**

Univariate and Bivariate analysis on the variables to understand the distribution and empirical relationship between the variables respectively

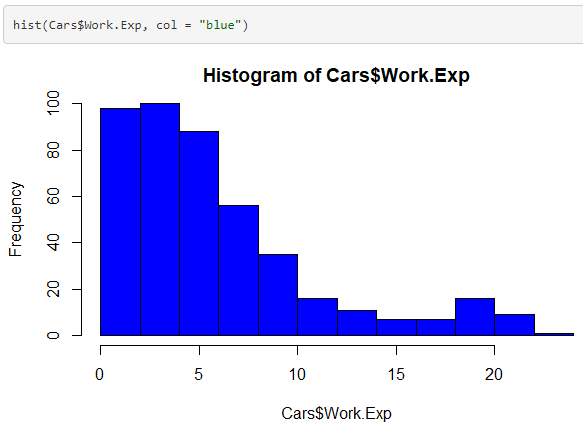
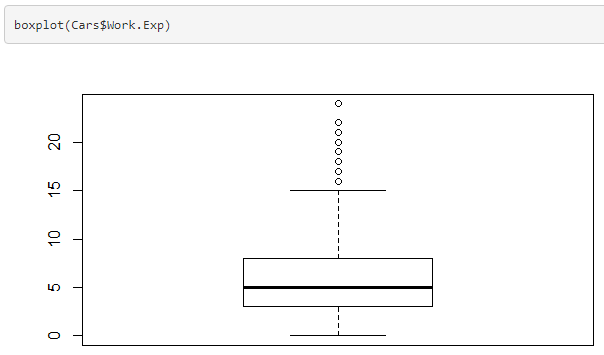
Univariate Analysis on Age

**Observation:**

* 50% of the employees are in between the Age group of 25-30.
* Very few employees above 40 years of age.
* The distribution of Age looks to be a normal distribution

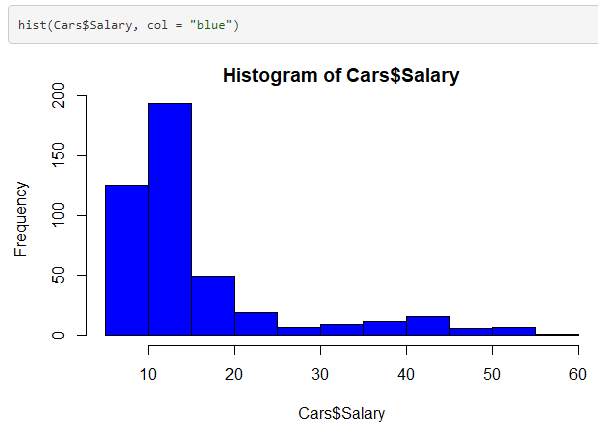
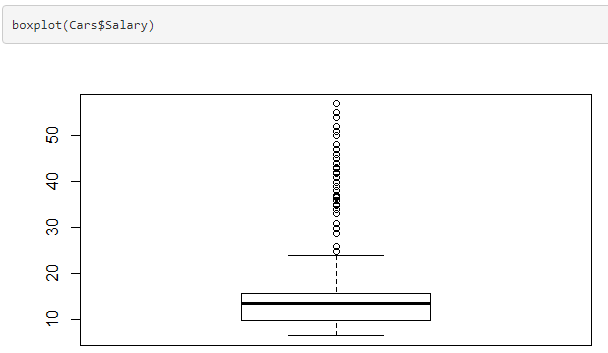
Univariate Analysis on Work Experience



**Observation:**

* Employees with Work Experience of 3-8 years constitutes to 50%.
* The distribution of Work Experience of the employees are skewed towards the right

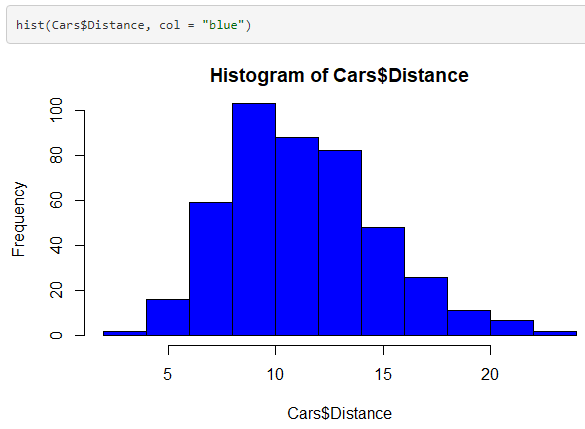
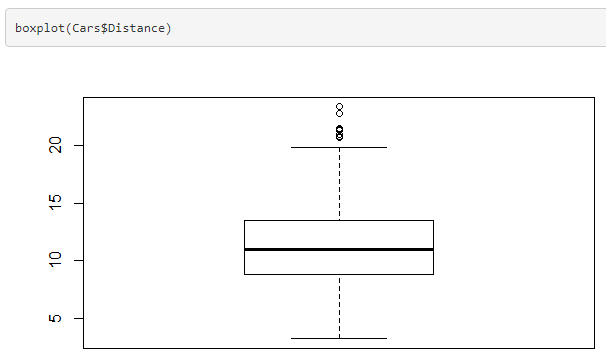
Univariate Analysis on Salary



**Observation:**

* 50% of the Salaries of the Employees lies between 10-15 lakhs.
* There is a large number of outliers on Salary which could be because of the few senior level employees who has higher salary compared to junior employees.
* The distribution of Salary of the employees are skewed towards the right.

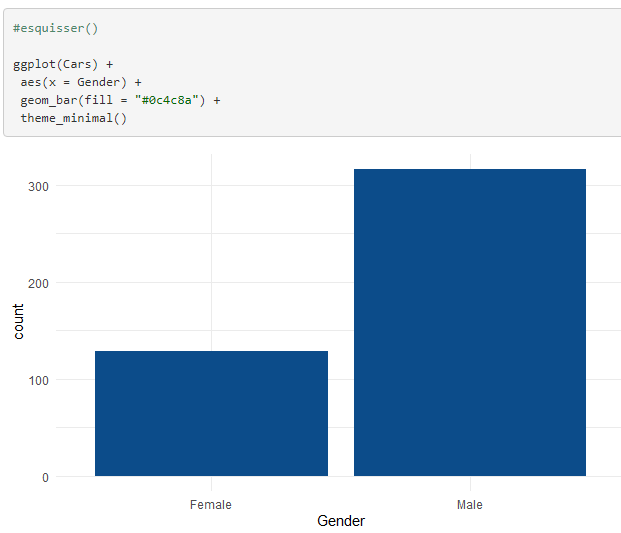
Univariate Analysis on Distance



**Observation:**

* Employees with a distance between 9-14 kms constitutes to about 50%.
* The distribution of Distance between home and office for the employees looks almost like a normal distribution.

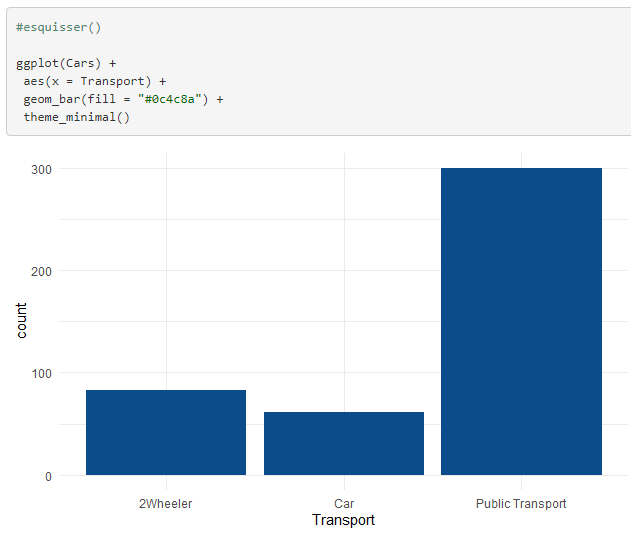
Univariate Analysis on Gender



**Observation:**

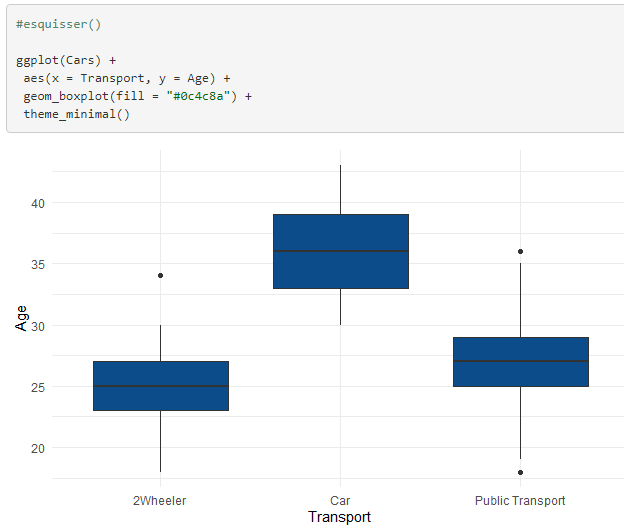
* About 70% of the employees in the company are Males

Univariate Analysis on Mode of Transport

**Observation:**

* Public transport is the most preferred modes of transport by the employees
* 2-wheeler comes second and Car is third

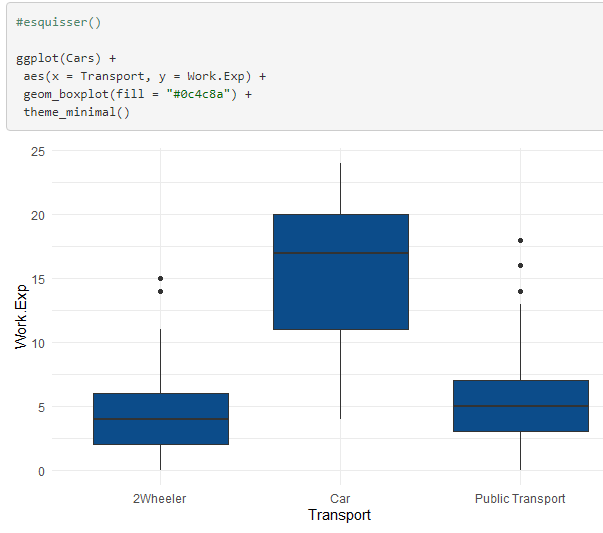
Bivariate Analysis on Mode of Transport & Age



**Observation:**

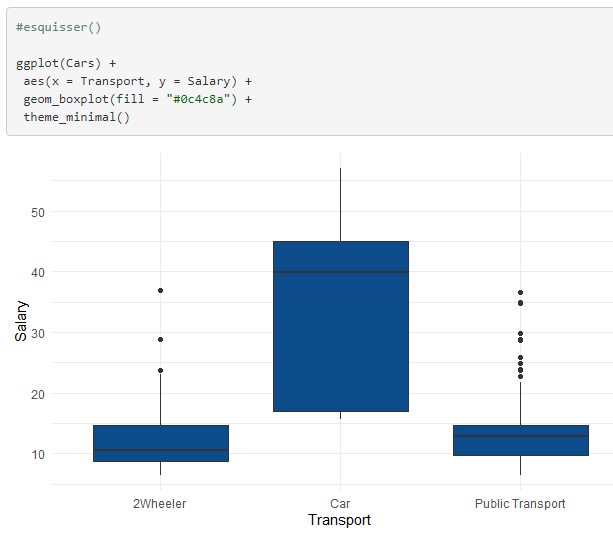
* Plot clearly shows that employees with a higher Age group prefer Car as a mode of transport.
* While 2-wheeler is mainly preferred by the employees with a lower Age group.

Bivariate Analysis on Mode of Transport & Work Experience

**Observation:**

* Similar to Age, employees with Higher Work experience use Car as a Mode of Transport.
* 2-wheeler and Public transport are mainly used by employees with lower Work Experience.

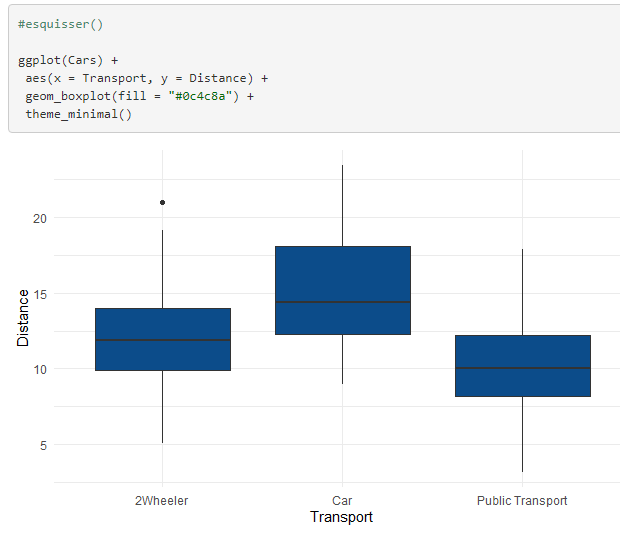
Bivariate Analysis on Mode of Transport & Salary



**Observation:**

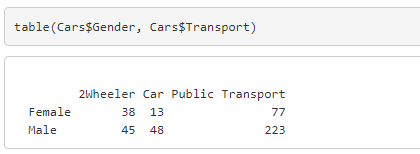
* Employees with Medium to High Salary prefer Car as the mode of Transport
* 2-wheeler and Public transport is preferred by employees with lower salary

Bivariate Analysis on Mode of Transport & Distance



**Observation:**

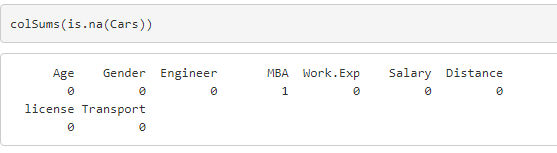
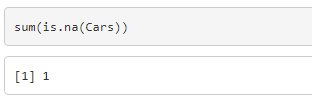
* Greater the distance, employees prefer Car as mode of transport for comfort and ease, followed by 2-wheeler and then Public Transport

Bivariate Analysis on Mode of Transport & Gender

**Observation:**

* About 60% of the Female employees use Public Transport, 30% use 2-Wheeler and 10% use Car as a Mode of Transport
* About 71% of the Female employees use Public Transport, 14% use 2-Wheeler and 15% use Car as a Mode of Transport

**Check for Missing Values**



We can observe that there is 1 missing value in the dataset, and the missing value is in the MBA column, which was also evident from the summary.

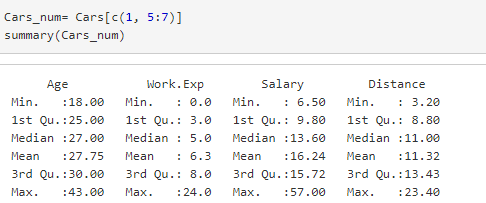
**Outliers**

From the boxplots above, we can conclude that there are Outliers in the below variables

Age  
Work Experience  
Salary  
Distance

**Check for Multicollinearity**

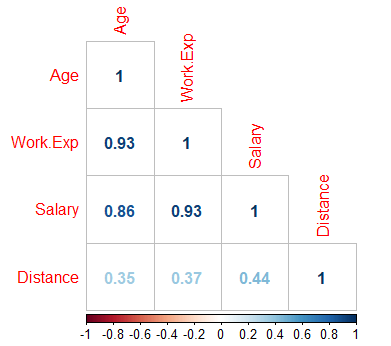
Multicollinearity generally occurs when there are high correlations between two or more [predictor variables](https://www.statisticshowto.com/independent-variable-definition/). In other words, one predictor variable can be used to predict the other.











Age, Work Experience, Salary and Distance are all predictor variables.  
We can observe that there is there is a high correlation between Age & Work.Exp, Age & Salary and Work.Exp & Salary.

Thus, showing there is a huge multicollinearity in the data set.

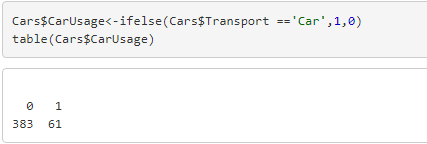
1. **Data Preparation**

Some pre-processing steps before the data is fed into the Machine Learning Model.

Imputing the missing values using knn imputation



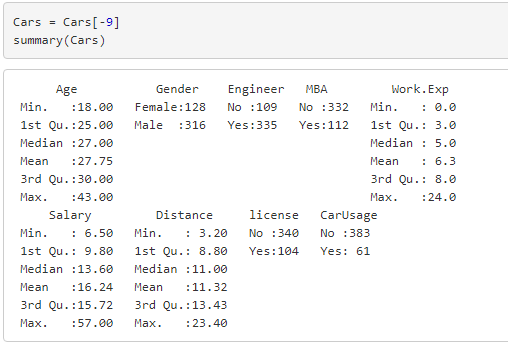
Since our main objective is to predict whether or not an employee will use Car as a mode of transport, hence we will create a new column for Car usage, which will take value 0 for Public Transport & 2-Wheeler and 1 for Car usage



Let us convert CarUsage column to factors



Now, let us remove the ‘Transport’ column from the dataset

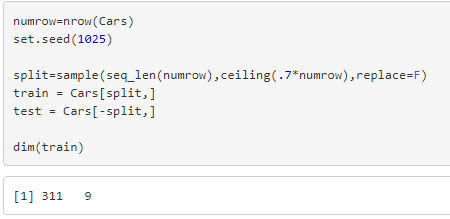


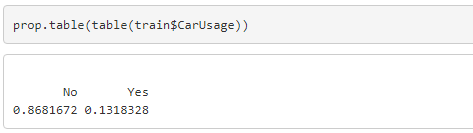
The data is ready to be fed to the models.

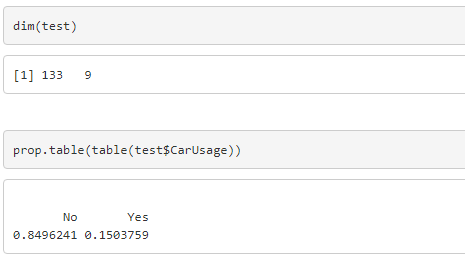
**Train-Test split**

Splitting the data into Train-Test is a most common practice for evaluating the performance of the Machine learning Algorithm.

Here, we are performing a random Train-Test split. 70% of the data is used for Training the model and rest 30% is used for Testing.







1. **Model Building**

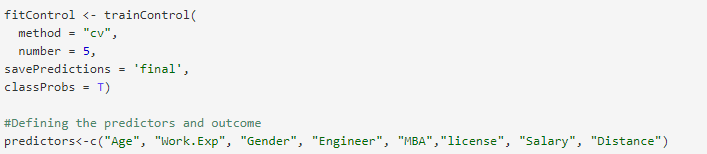
**Logistic Regression Model**

[Logistic regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-logistic-regression/) is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, the logistic regression is a predictive analysis. 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.

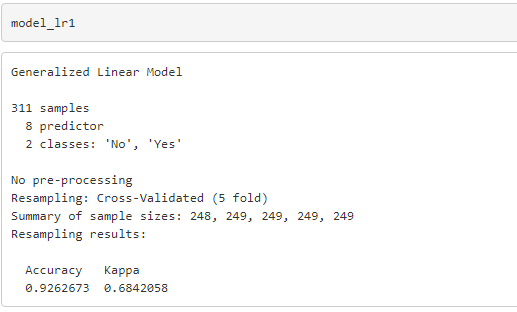
Logistic regression is a linear method, but the predictions are transformed using the logistic function.

We are using the CARET package to train the models.

Training the Logistic Regression Full Model using a 5-fold Cross validation

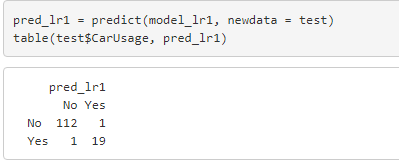




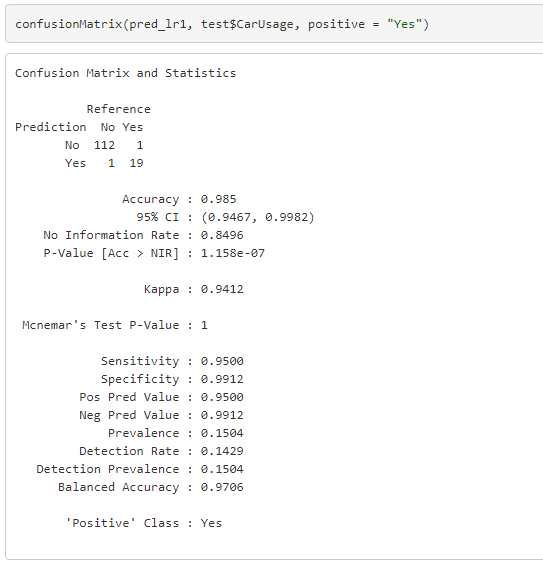


Accuracy obtained for the trained data is 92.62%

Now, let us predict the test dataset using the trained model



**Confusion Matrix**



The Accuracy, Sensitivity and Specificity obtained for the test data are as below.

Accuracy:98.5%  
Sensitivity:95%  
Specificity: 99.12%

With Kappa value with 94.12%

Kappa statistic is a measure of how closely the instances classified by the machine learning classifier matched the data labeled as ground truth

Let us check for the important variables used by the model



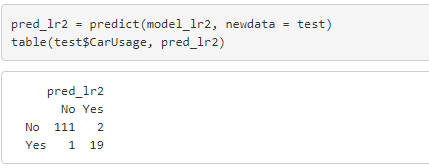
We can observe that Age, Work Experience, Distance, License are of high importance followed by MBA and Salary.

Depending on the variable importance, let us build a logistic regression model considering the significant variables.

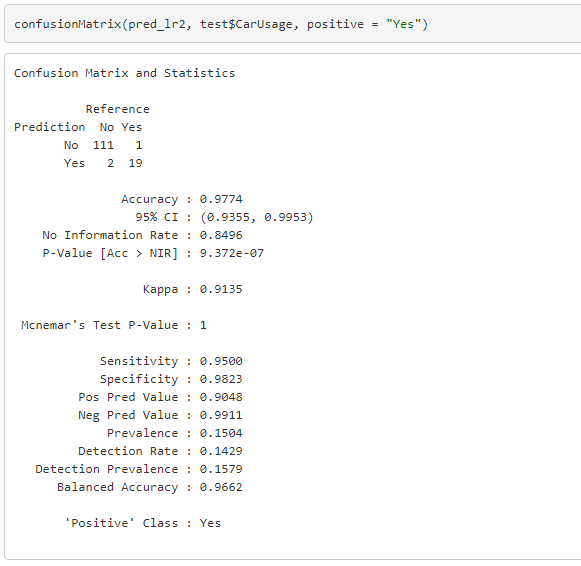


Accuracy obtained for the trained data is 94.21%, slightly higher than obtained using Full Model

Now, let us predict the test dataset using the trained model



**Confusion Matrix**



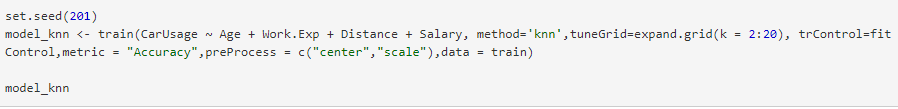
The Accuracy, Sensitivity and Specificity obtained for the test data are as below.

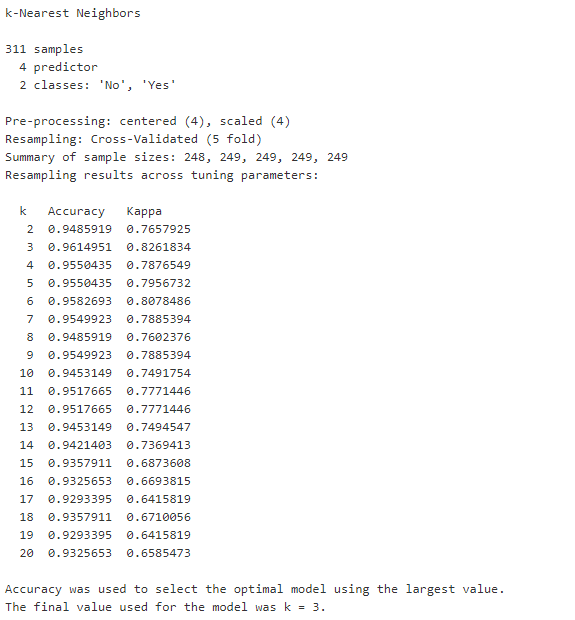
Accuracy:97.74%  
Sensitivity:95%  
Specificity: 98.23%

With Kappa value with 91.35%

**K-Nearest Neighbors (KNN) Model**KNN: K Nearest Neighbor is one of the fundamental algorithms in machine learning used for Classification. It classifies the data point on how its neighbor is classified.   
  
K in KNN represents the number of the nearest neighbors used to classify new data points. KNN uses Euclidean distance to calculate the nearest neighbor. If we have two points (x, y) and (a, b). The formula for Euclidean distance (d) will be d = sqrt((x-a)²+(y-b)²)

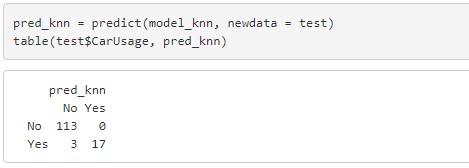
Since KNN is a distance-based model, let us consider Age, Distance, Work Experience and Salary as predictors for this model.



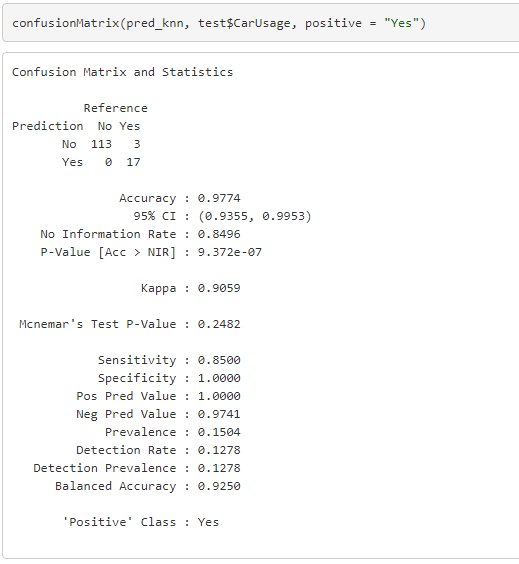


Accuracy obtained for the trained data is 96.14%,

Now, let us predict the test dataset using the trained model



**Confusion Matrix**



The Accuracy, Sensitivity and Specificity obtained for the test data are as below.

Accuracy:97.74%  
Sensitivity:85%  
Specificity: 100%

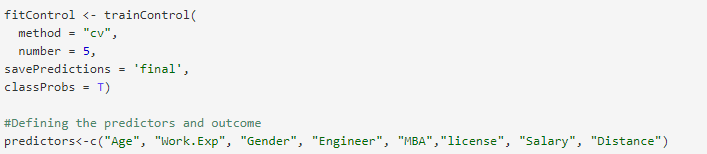
With Kappa value with 90.59%

**Naïve-Bayes Model**

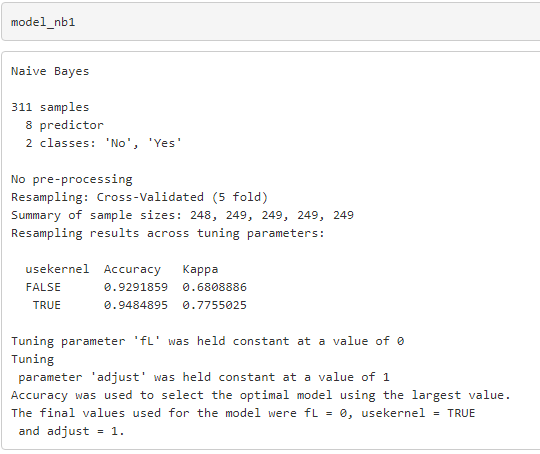
Naive Bayes is a Supervised Machine Learning algorithm based on the Bayes Theorem that is used to solve classification problems by following a probabilistic approach.  It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Using the CARET package to train the model.

Training the Naïve Bayes Full Model using a 5-fold Cross validation



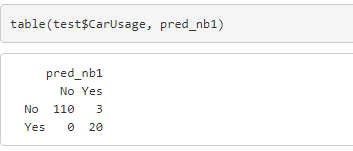




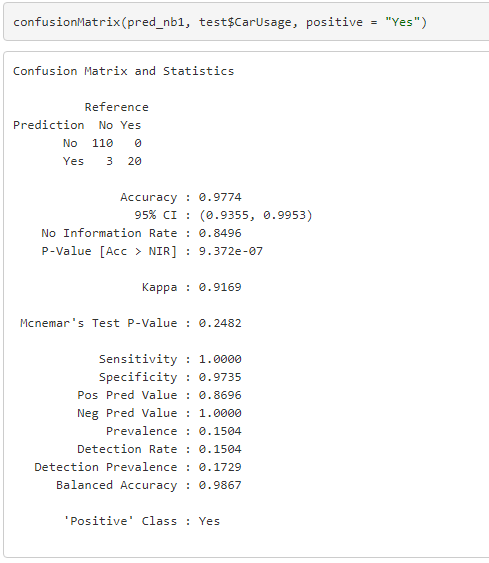
Accuracy obtained for the trained data is 94.84%

Now, let us predict the test dataset using the trained model





**Confusion Matrix**



The Accuracy, Sensitivity and Specificity obtained for the test data are as below.

Accuracy:97.74%  
Sensitivity:100%  
Specificity: 97.35%

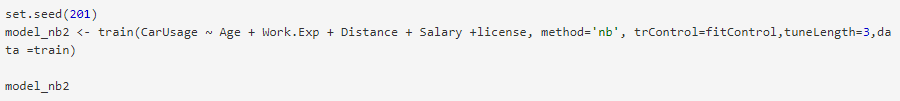
With Kappa value with 91.69%

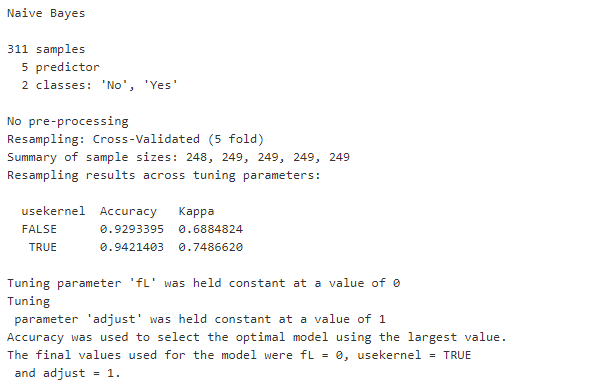
Let us check for the important variables used by the model



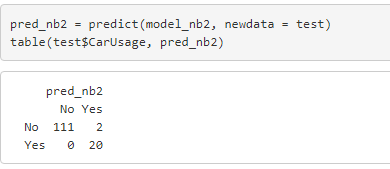
We can observe that Age, Salary, Work Experience, License, Distance are of high importance

Depending on the variable importance, let us build a Naïve Bayes model considering the significant variables.

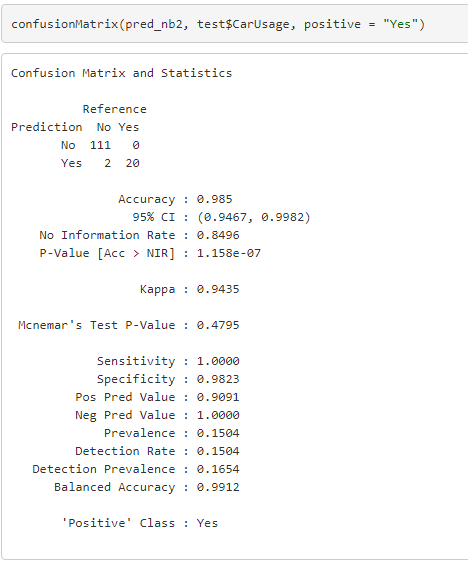




Accuracy obtained for the trained data is 94.21

Now, let us predict the test dataset using the trained model 

**Confusion Matrix**



The Accuracy, Sensitivity and Specificity obtained for the test data are as below.

Accuracy:98.5%  
Sensitivity:100%  
Specificity: 98.23%

With Kappa value with 94.35%

Comparing the Confusion Matrix, we can observe that the Naïve Bayes model with the significant variables has performed the best not just by giving good Accuracy but also good Sensitivity and Specificity.

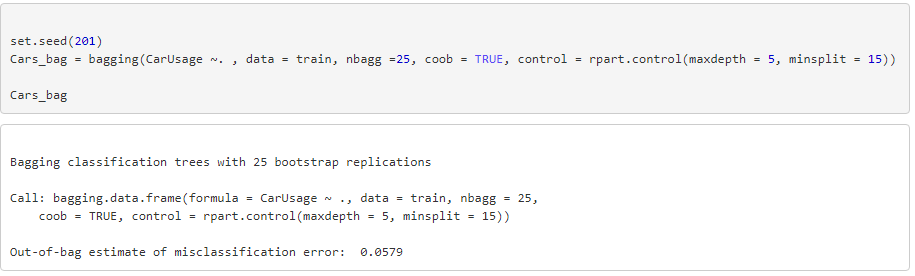
Accuracy:98.5%  
Sensitivity:100%  
Specificity: 98.23%

With Kappa value with 94.35%

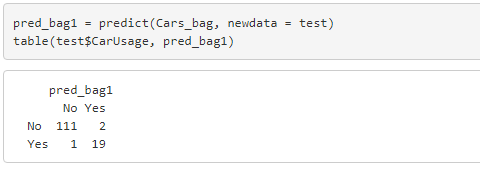
1. **Bagging and Boosting**

**Bagging**

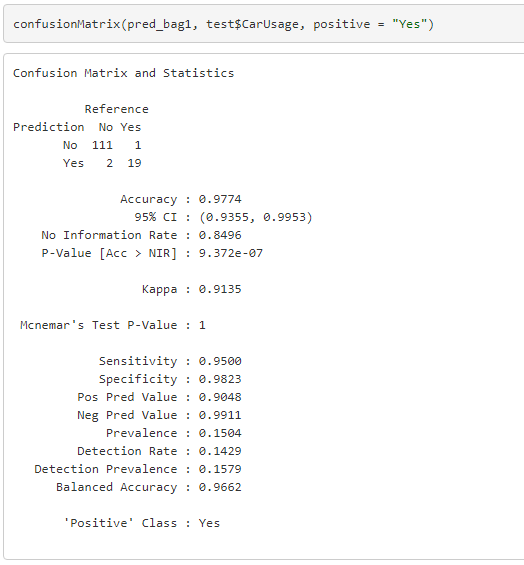
Bagging(Bootstrap Aggregation) is used when our goal is to reduce the variance of a decision tree. Here idea is to create several subsets of data from training sample chosen randomly with replacement. Now, each collection of subset data is used to train their decision trees. As a result, we end up with an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree.



Let us predict on the test data



**Confusion Matrix**



The Accuracy, Sensitivity and Specificity obtained for the test data are as below.

Accuracy:97.74%  
Sensitivity:95%  
Specificity: 98.23%

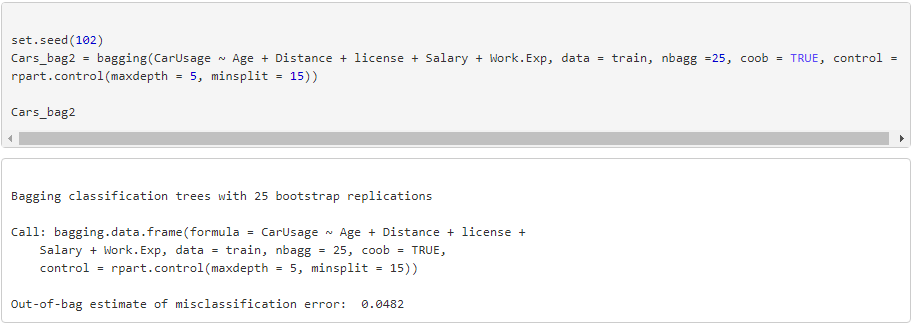
With Kappa value with 91.35%

Checking for the variable importance in Bagging model

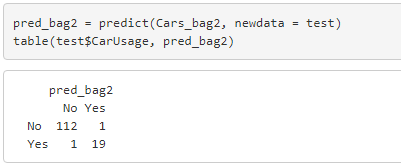


We can observe that Age, Salary, Work Experience, License, Distance are of high importance

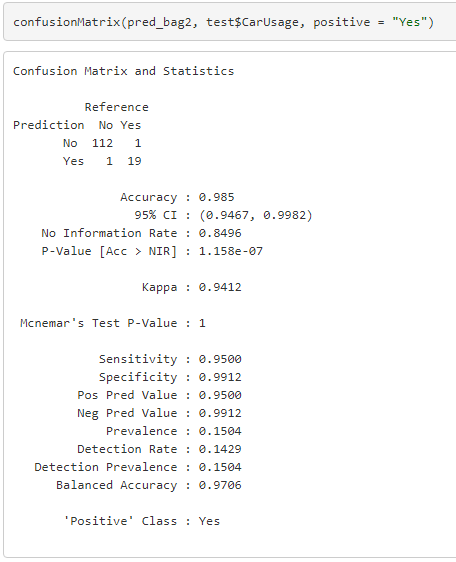
Let us consider the important/significant variables to build another bagging model to compare the results.



Let us predict on the test data



**Confusion Matrix**



The Accuracy, Sensitivity and Specificity obtained for the test data are as below.

Accuracy:98.5%  
Sensitivity:95%  
Specificity: 99.12%

With Kappa value with 94.12%

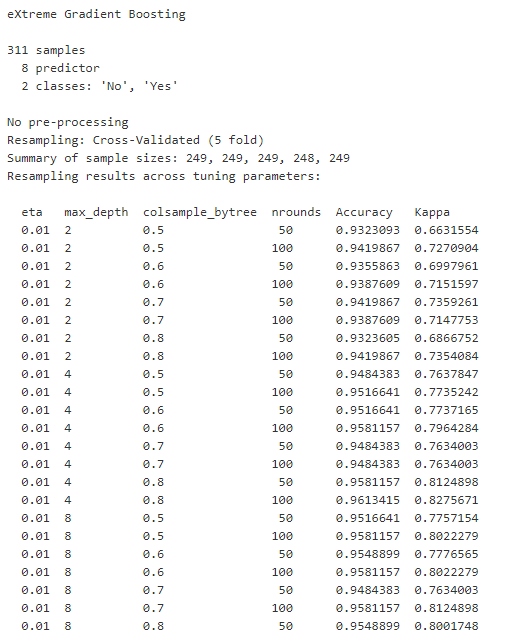
**Boosting**

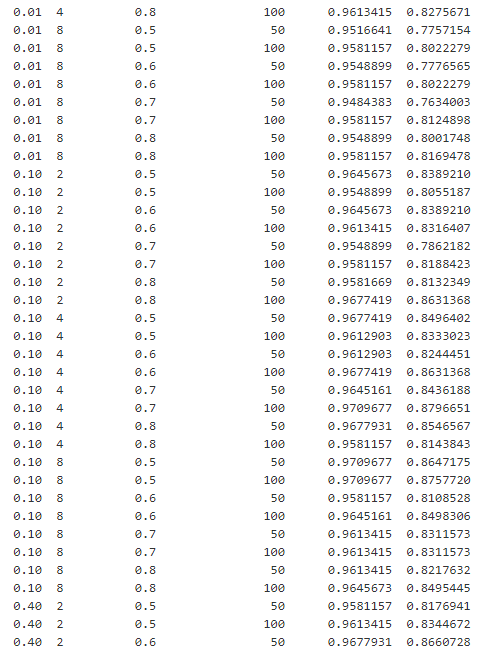
Boosting is another ensemble technique to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. In other words, we fit consecutive trees (random sample) and at every step, the goal is to solve for net error from the prior tree.

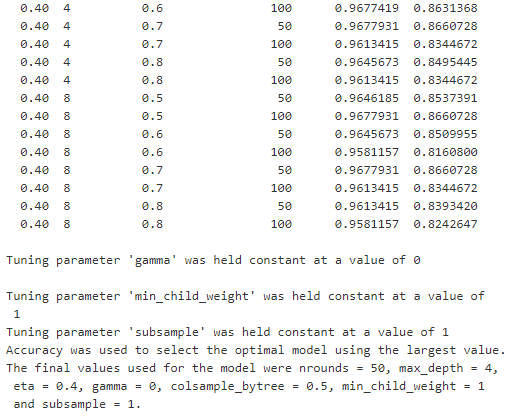
Boosting is an iterative technique which adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation. Boosting in general builds strong predictive models.

Boosting along with hyper parameters tuning

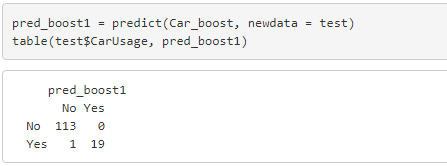




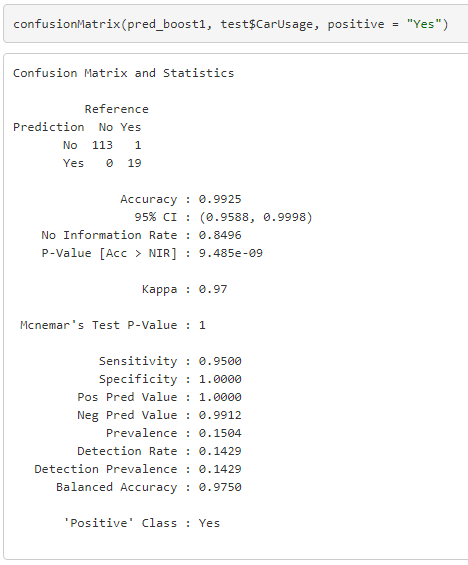




Let us predict on the test data



**Confusion Matrix**



The Accuracy, Sensitivity and Specificity obtained for the test data are as below.

Accuracy:99.25%  
Sensitivity:95%  
Specificity: 100%

With Kappa value with 97%

Let us take a look at the important variables considered for building this model.



We can observe that Age, Salary, Work Experience, Distance are of high importance

Naïve bayes model (significant variables) that had performed better among KNN and Logistic Regression had given 98.5% accuracy.

Boosting model has outperformed the Naïve Bayes giving an accuracy of 99.25%.

Important variables are Age, Salary, Work Experience, License and Distance

Employees with higher Age, Work Experience, Salary and with greater Distance use Car as a mode of Transport.