27-04-2020

MACHINE LEARNING PROJECT

PREDICTING MODE OF TRANSPORT (ML)

PRESENTED BY: SHILPA GIRIDHAR

CONTENTS

DESCRIPTION

Problem Statement

This project requires you to understand what mode of transport employees prefers to commute to their office. The attached data 'Cars.csv' includes employee information about their mode of transport as well as their personal and professional details like age, salary, work exp. We need to predict whether or not an employee will use Car as a mode of transport. Also, which variables are a significant predictor behind this decision?

Data Dictionary

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 Licence -1, If not, then 0
Transport	Mode of Transport

Requirements

Perform the following :

1. **<u>EDA (15 Marks)</u>**
    - Perform an EDA on the data - Basic data summary, Univariate, Bivariate analysis, graphs, Check for Outliers and missing values and check the summary of the dataset (7 marks)
    - Illustrate the insights based on EDA (5 marks)
    - Check for Multicollinearity - Plot the graph based on Multicollinearity & treat it. (3 marks)

2. **<u>Data Preparation (10 marks)</u>**
    - Prepare the data for analysis (SMOTE)

3. **<u>Modeling (30 Marks)</u>**
    - Create multiple models and explore how each model perform using appropriate model performance metrics (15 marks)
        - Applying KNN Model & Interpret results
        - Applying Naive Bayes (is it applicable here? comment and if it is not applicable, how can you build an NB model in this case?)
        - Applying Logistic Regression & Interpret results
        - Confusion matrix interpretation
    - Apply both bagging and boosting modeling procedures to create 2 models and compare its accuracy with the best model of the above step. (15 marks)

4. **<u>Actionable Insights & Recommendations (5 Marks)</u>**
    - Summarize your findings from the exercise in a concise yet actionable note

BUSINESS OBJECTIVE

Which variables are significant in deciding whether the employees prefer Car as mode of transport?

To guide the analysis, we are going to try and answer the following questions about my customer segments:

- Does the Age of the Employee crucial in deciding preference of car as mode of transport?
- Do individuals higher Salary and Work Experience more like to use car as mode of transport?
- Does the Gender play a crucial role in transport preferences?
- Does the Distance play a crucial role in transport preferences?

DATA EXPLORATION

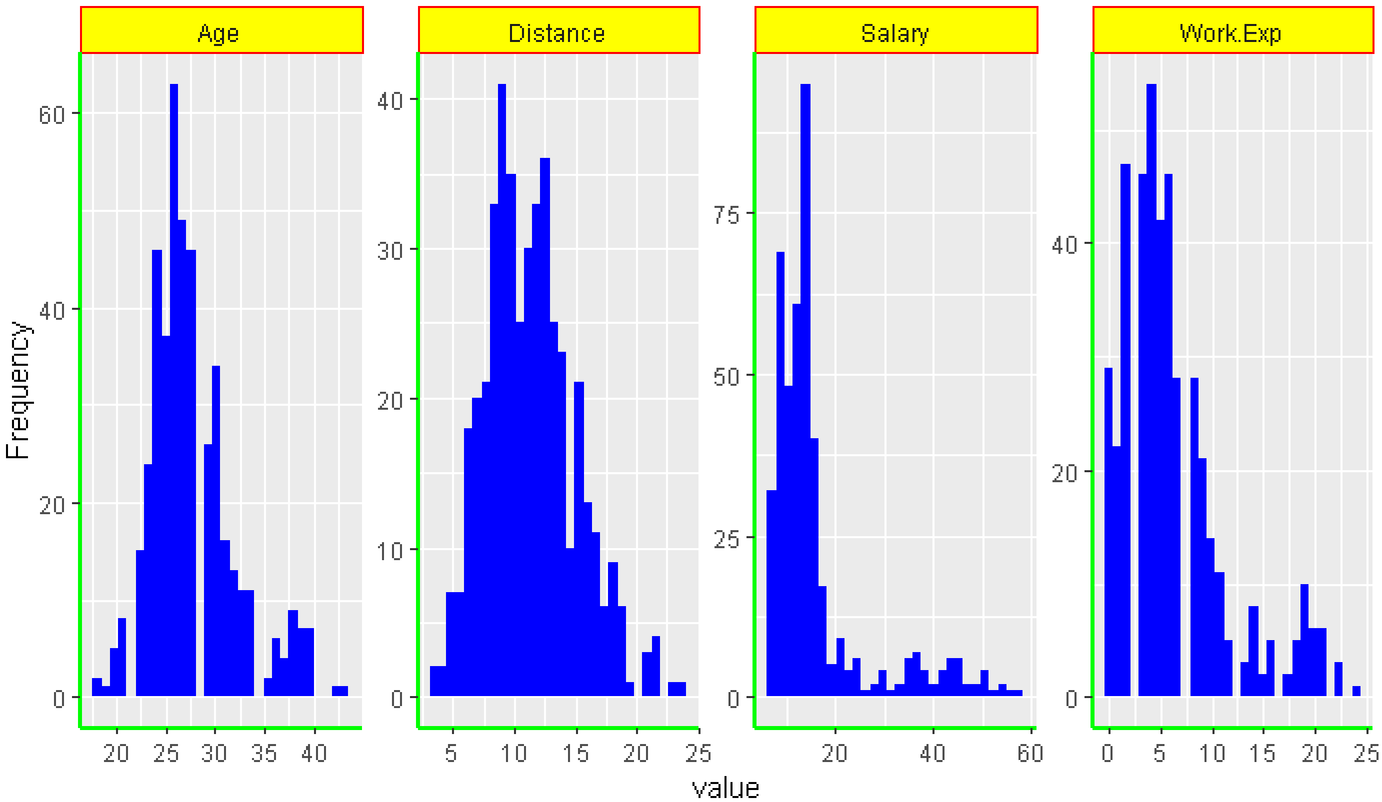
- The data shows that there are 444 observations and 9 variables
- Performed the str and summary function – Gender and Transport Mode are Factor variables, rest are of type integer variables. We can observe that Engineer, MBA, and License can also be converted to factor type. We can also observe that MBA has one 'NA' – so we can treat it before converting to factor type. Transport column is the target variable.

```
'data.frame':   444 obs. of  9 variables:
 $ Age      : int  28 23 29 28 27 26 28 26 22 27 ...
 $ Gender   : Factor w/ 2 levels "Female","Male": 2 1 2 1 2 2 2 1 2 2 ...
 $ Engineer : int  0 1 1 1 1 1 1 1 1 1 ...
 $ MBA      : int  0 0 0 1 0 0 0 0 0 0 ...
 $ Work.Exp : int  4 4 7 5 4 4 5 3 1 4 ...
 $ Salary   : num  14.3 8.3 13.4 13.4 13.4 12.3 14.4 10.5 7.5 13.5 ...
 $ Distance : num  3.2 3.3 4.1 4.5 4.6 4.8 5.1 5.1 5.1 5.2 ...
 $ license  : int  0 0 0 0 0 1 0 0 0 0 ...
 $ Transport: Factor w/ 3 levels "2Wheeler","Car",..: 3 3 3 3 3 3 1 3 3 3 ...
      Age            Gender       Engineer            MBA             Work.Exp        Salary     
 Min.   :18.00   Female:128   Min.   :0.0000   Min.   :0.0000   Min.   : 0.0   Min.   : 6.50  
 1st Qu.:25.00   Male  :316   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.: 3.0   1st Qu.: 9.80  
 Median :27.00                Median :1.0000   Median :0.0000   Median : 5.0   Median :13.60  
 Mean   :27.75                Mean   :0.7545   Mean   :0.2528   Mean   : 6.3   Mean   :16.24  
 3rd Qu.:30.00                3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 8.0   3rd Qu.:15.72  
 Max.   :43.00                Max.   :1.0000   Max.   :1.0000   Max.   :24.0   Max.   :57.00  
                                               NA's   :1                                     
    Distance        license                  Transport  
 Min.   : 3.20   Min.   :0.0000   2Wheeler        : 83  
 1st Qu.: 8.80   1st Qu.:0.0000   Car             : 61  
 Median :11.00   Median :0.0000   Public Transport:300  
 Mean   :11.32   Mean   :0.2342                         
 3rd Qu.:13.43   3rd Qu.:0.0000                         
 Max.   :23.40   Max.   :1.0000                         
```

- Treat the missing values in MBA column – Row no 145 has missing value in MBA column. We replace his with '0' value.
- Then convert Engineer, MBA, and License to factor type variables and perform Summary function of the dataset.

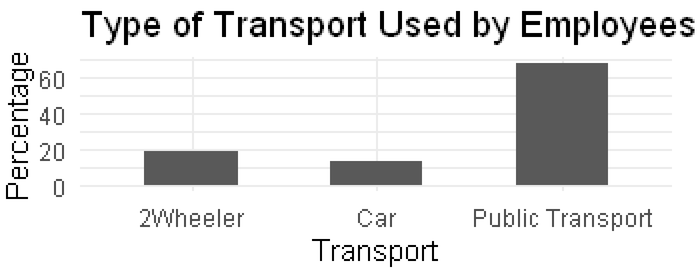
```
      Age           Gender     Engineer MBA        Work.Exp        Salary          Distance    
 Min.   :18.00   Female:128   0:109    0:332   Min.   : 0.0   Min.   : 6.50   Min.   : 3.20  
 1st Qu.:25.00   Male  :316   1:335    1:112   1st Qu.: 3.0   1st Qu.: 9.80   1st Qu.: 8.80  
 Median :27.00                                 Median : 5.0   Median :13.60   Median :11.00  
 Mean   :27.75                                 Mean   : 6.3   Mean   :16.24   Mean   :11.32  
 3rd Qu.:30.00                                 3rd Qu.: 8.0   3rd Qu.:15.72   3rd Qu.:13.43  
 Max.   :43.00                                 Max.   :24.0   Max.   :57.00   Max.   :23.40  
 license            Transport  
 0:340   2Wheeler        : 83  
 1:104   Car             : 61  
         Public Transport:300  
```

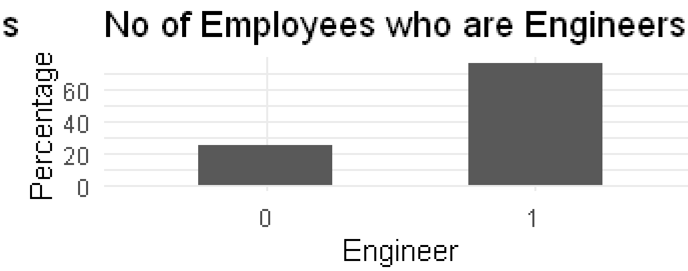
- Use Histogram plot to understand continous variables

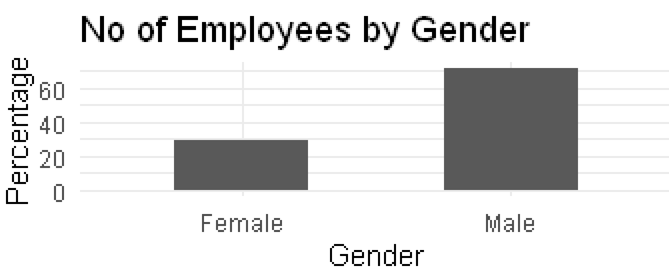


- Use Bar plot to understand categorical variables

  Majority of the employees are Engineers, Male, and majority prefer to use Public Transport. Only about 12% of employees use car for mode of transport, while 70% of employees use Public transport.





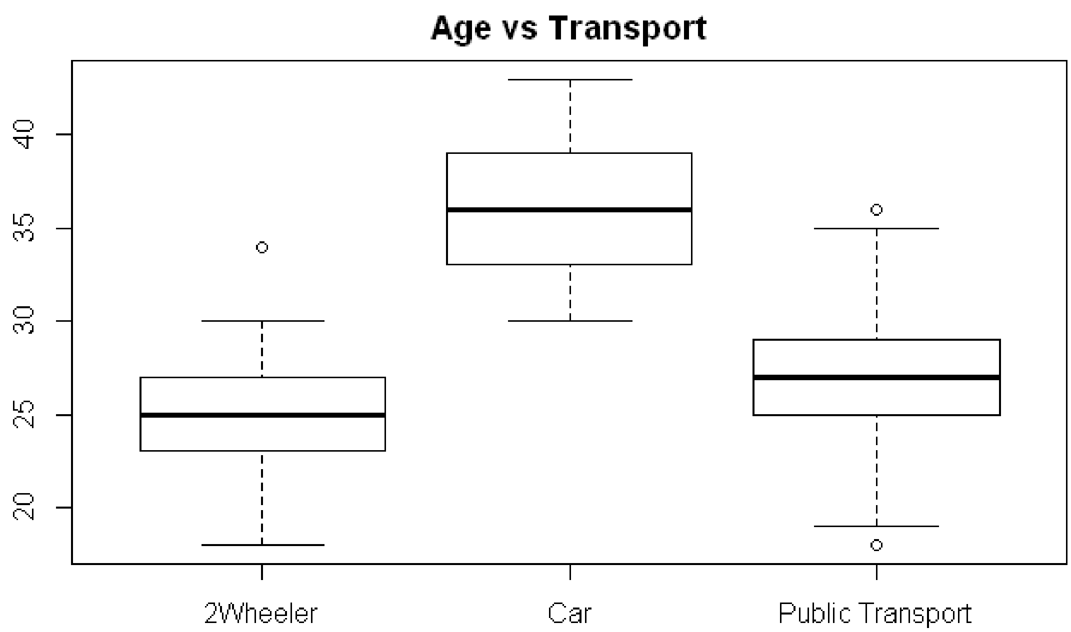


- Proportion Table shows that the number of records for people travelling by car is only about 13.7%. The given dataset is imbalanced.

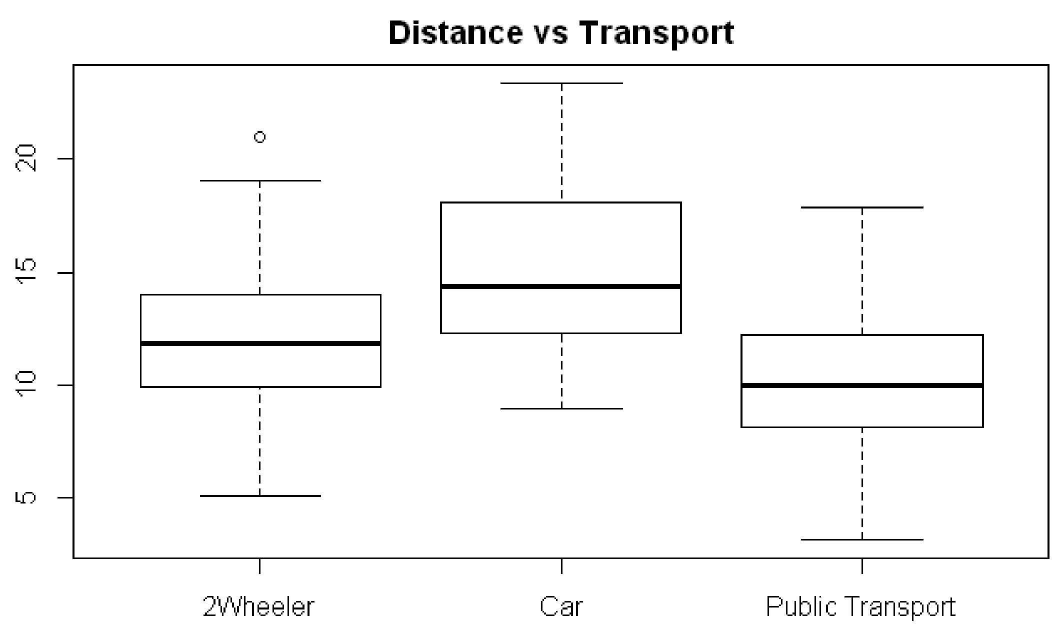
```
2Wheeler             Car          Public Transport
0.1869369      0.1373874          0.6756757
```

- Bivariate analysis using boxplot grouped by “Transport” column data

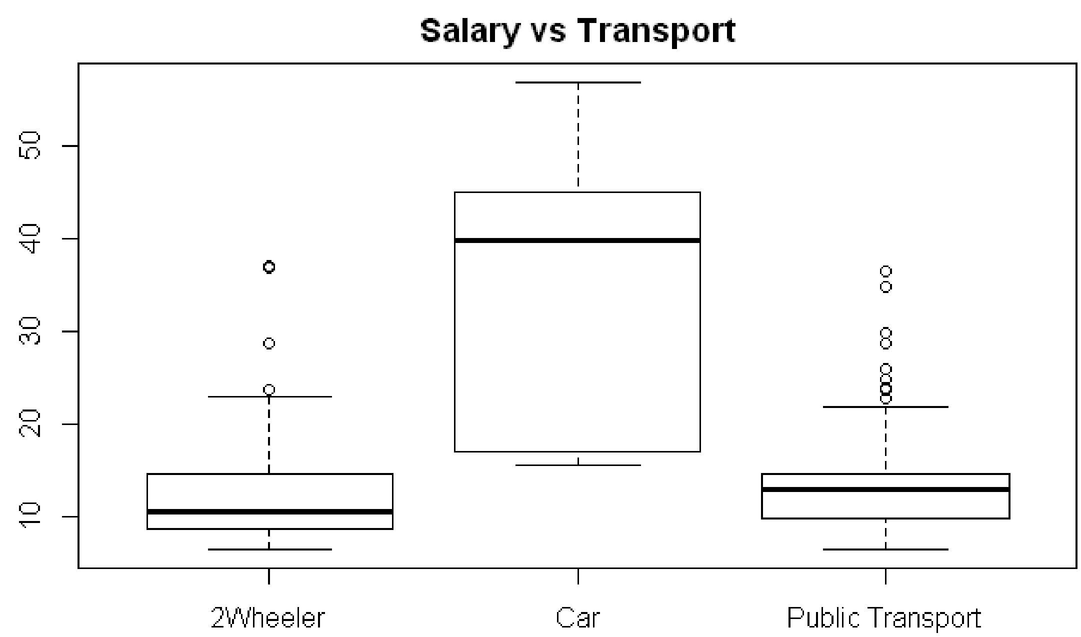
  We can perform the bivariate analysis to understand the significant factors that affects the choice of transport used by the employees using boxplot. Age seems to be a significant factor in employees who are using Car as mode of transport, as range of age of employees using car lies between 33 to 40 years. While majority of those using 2-Wheeler and Public Transport lies approximately between 23 to 28 years. So higher age seems to be a driving factor for transport mode selection.

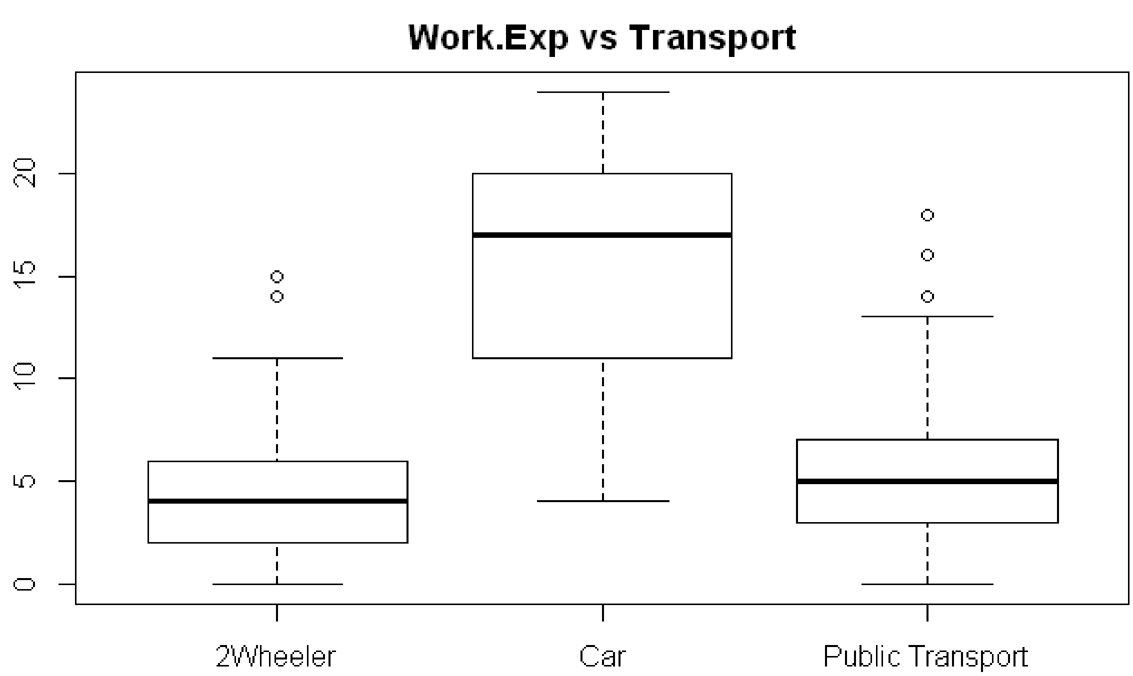


Further boxplot on Distance reveals majority of employees using Car as mode of transport travel longer distances compared to others using 2-Wheeler and Public Transport.



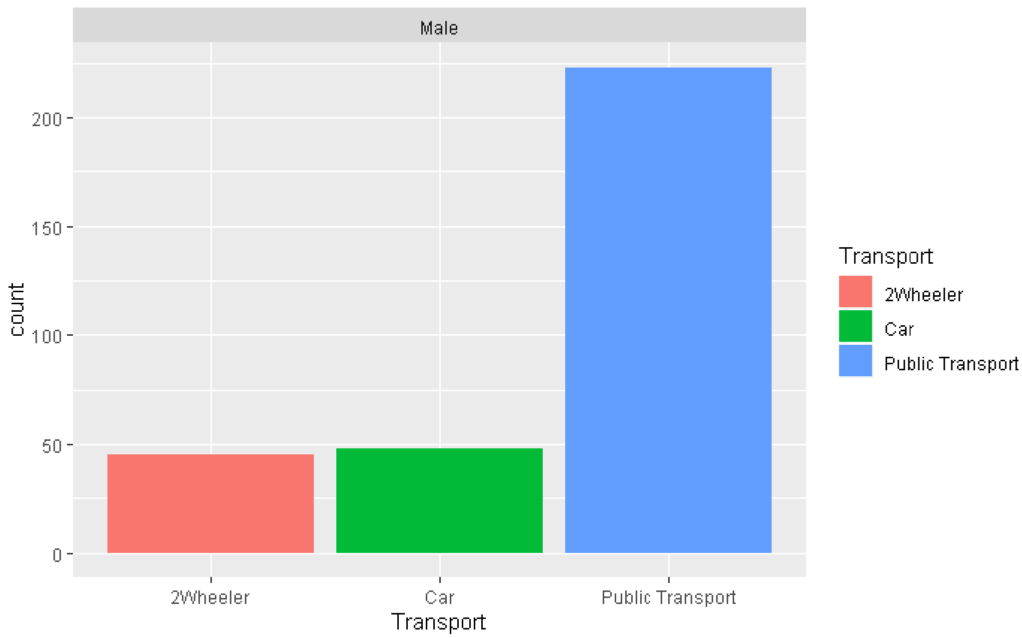
Salary and Work Experience has outliers for category of employees who have used Public Transport and 2-Wheeler as transport mode. Most of these employees are having salary ranging between 10 – 15 Lakh per annum, and work exp ranging between 2 to 7.5 years. While majority of using employees using Car as mode of transport are having salary ranging between 16 – 45 Lakh per annum, and work experience ranging between 10 to 20 years. So higher work experience, and higher salary seems to be a driving factor for transport mode selection. This is also clear that all three factors – Age, Work experience and Salary are highly correlated to each other.

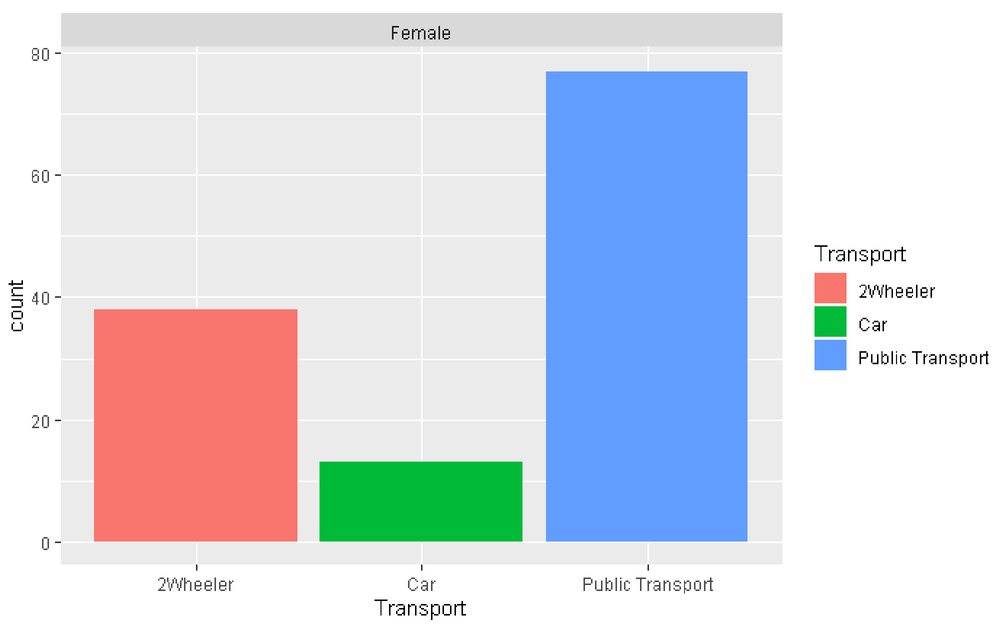




Further boxplot and proportion table on Gender vs Transport reveals that female employees prefer 2-Wheeler more in comparison with male employees, while proportion of male employees using Public transport is higher. Both proportion of male employees preferring Car as mode of transport is slightly higher than the female employees.

```
                 Female Male
2Wheeler           0.30 0.14
Car                0.10 0.15
Public Transport   0.60 0.71
Sum                1.00 1.00
```





- From boxplots, we see that variables have outliers and need to be treated before proceeding to building models

```
 detect_outliers(Age)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   18.00   25.00   27.00   27.75   30.00   43.00 
  [1] 39 39 39 38 40 38 38 38 38 40 40 39 40 38 39 38 40 39 38 42 40 43 40 38 39
```

```
detect_outliers(`Work.Exp`)
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0.0     3.0     5.0     6.3     8.0    24.0 
  [1] 19 16 21 17 16 18 19 18 21 16 19 19 18 19 20 22 16 20 18 21 20 20 16 17 21 18
20 21 19 22
[31] 22 19 24 20 19 19 19 21
```

```
detect_outliers(Salary)
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     6.50    9.80   13.60   16.24   15.72   57.00 
  [1] 36.6 38.9 25.9 34.8 28.8 39.9 39.0 28.7 36.9 28.7 34.9 47.0 28.8 36.9 54.0 29.9
34.9 36.0
[19] 44.0 37.0 24.9 43.0 37.0 54.0 44.0 34.0 48.0 42.0 51.0 45.0 34.0 28.8 45.0 42.9
41.0 40.9
[37] 30.9 41.9 43.0 33.0 36.0 33.0 38.0 46.0 45.0 48.0 35.0 51.0 51.0 55.0 45.0 42.0
52.0 38.0
[55] 57.0 44.0 45.0 47.0 50.0
```

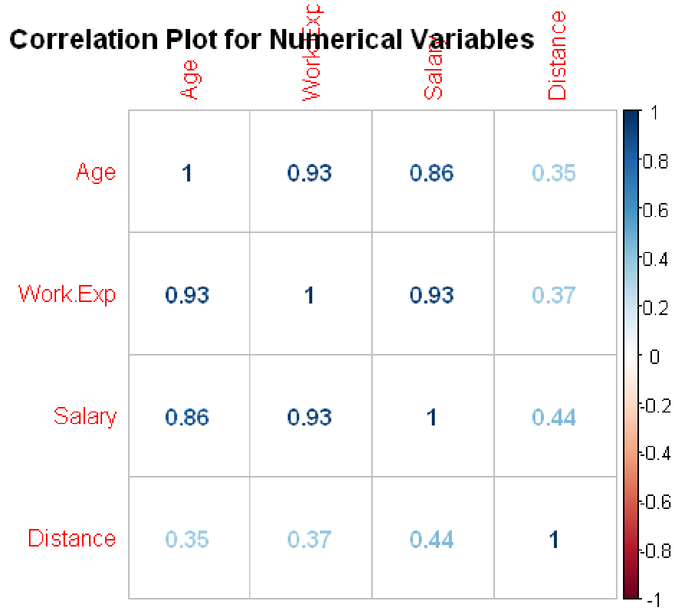
```
detect_outliers(Distance)
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     3.20    8.80   11.00   11.32   13.43   23.40 
[1] 20.7 20.8 21.0 21.3 21.4 21.5 21.5 22.8 23.4
```

After treatment of outliers

```
 detect_outliers(Age)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   22.00   25.00   27.00   27.75   30.00   38.00 
  [1] 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38
detect_outliers(`Work.Exp`)
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.000   3.000   5.000   6.227   8.000  19.000 
  [1] 19 16 19 17 16 18 19 18 19 16 19 19 18 19 19 19 16 19 18 19 19 19 16 17 19 18
19 19 19 19
[31] 19 19 19 19 19 19 19 19
detect_outliers(Salary)
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     7.60    9.80   13.60   16.00   15.72   43.00 
  [1] 36.6 38.9 25.9 34.8 28.8 39.9 39.0 28.7 36.9 28.7 34.9 43.0 28.8 36.9 43.0 29.9
34.9 36.0
[19] 43.0 37.0 24.9 43.0 37.0 43.0 43.0 34.0 43.0 42.0 43.0 43.0 34.0 28.8 43.0 42.9
41.0 40.9
[37] 30.9 41.9 43.0 33.0 36.0 33.0 38.0 43.0 43.0 43.0 35.0 43.0 43.0 43.0 43.0 42.0
43.0 38.0
[55] 43.0 43.0 43.0 43.0 43.0
detect_outliers(Distance)
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     6.10    8.80   11.00   11.27   13.43   17.80 
numeric(0)
```

- Used the library(corrplot) to plot correlation and check for high correlation between numerical variables

  We observe that Work experience, Age, and Salary are highly correlated, which does not signify much for the problem solving.



Chi-square test for categorical variables – or Fichers Exact test - this gives p-value less than 0.05

PREPARE THE DATA FOR ANALYSIS (SMOTE)

APPLYING SMOTE

- To make sure that we get desired results in our model, we'll apply SMOTE to handle the imbalanced dataset. We have 'y' variable that contains 3 levels – "Cars", "2-Wheeler" and "Public transport". To apply SMOTE we reduce this Y variable into binary levels by clubbing non-cars transport mode as '0' and cars mode of transport as '1'. We add a new variable "CarUse" that contains this information.

```
  Var1 Freq
1    0  383
2    1   61
```

Proportion table with the new variable -

```
        0         1 
0.8626126 0.1373874 
```

- We split the main data set into 70% of the data to be our training set, and 30% to be our test set. We'll train the model on the training set, and then test out its performance on the test set. To create the split we'll use the Caret package. Train and Test data set gives the following results. We can observe that proportion of train dataset is same as the complete data set.

```
> prop.table(table(carsDStrain$CarUse))

        0         1 
0.8621795 0.1378205 
```

- We use the library "DMwR" and apply SMOTE to get a blanaced data set from the training data set. Perc.over means that 1 minority class will be added for every value of perc.over. We have increased the minority class by adding 4 for every minority class sample - perc.over. We are reducing from the majority class by subtracting 30 for every 100 - perc.under.

```
balanced.data <- SMOTE(CarUse ~., carsDStrain, perc.over = 400, k = 5, perc.under = 300)
> as.data.frame(table(balanced.data$CarUse))
  Var1 Freq
1    0  516
2    1  215
> prop.table(table(balanced.data$CarUse))
        0         1 
0.7058824 0.2941176 
 summary(balanced.data)
 Age          Gender   Engineer MBA        Work.Exp          Salary          Distance    
 Min.   :18.00   1:209   0:177    0:541   Min.   : 0.000   Min.   : 6.50   Min.   : 3.30  
 1st Qu.:26.00   2:522   1:554    1:190   1st Qu.: 3.000   1st Qu.:10.70   1st Qu.: 9.30  
 Median :28.00                            Median : 6.000   Median :14.60   Median :12.20  
 Mean   :29.27                            Mean   : 8.019   Mean   :19.47   Mean   :12.24  
 3rd Qu.:33.00                            3rd Qu.:12.000   3rd Qu.:25.03   3rd Qu.:14.66  
 Max.   :42.00                            Max.   :22.000   Max.   :57.00   Max.   :23.40  
 license CarUse 
 0:494   0:516  
 1:237   1:215  
```

MODEL EVALUATION

1. Logistic Regression
2. K- Nearest Neighbors
3. Naïve Bayes

LOGISTIC REGRESSION

We will build the Logistic regression model on the SMOTE data to understand the factors influencing car usage. Since we have only limited variable, we will use them all in model building.

Model 1 – with Balanced SMOTE dataset

param1<- CarUse ~ Age + Gender + Engineer + MBA + Work.Exp + Salary + Distance + license

```
Call:
glm(formula = param1, family = binomial, data = balanced.data)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.28566  -0.06473  -0.01142   0.01228   1.96640  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -59.008492   8.244439  -7.157 8.22e-13 ***
Age           1.852131   0.291336   6.357 2.05e-10 ***
Gender2      -1.168354   0.488492  -2.392 0.016768 *  
Engineer1     0.008641   0.457513   0.019 0.984932    
MBA1         -0.794233   0.466727  -1.702 0.088810 .  
Work.Exp     -0.900721   0.228901  -3.935 8.32e-05 ***
Salary        0.142423   0.044911   3.171 0.001518 ** 
Distance      0.472017   0.082469   5.724 1.04e-08 ***
license1      1.648653   0.447373   3.685 0.000229 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 885.68  on 730  degrees of freedom
Residual deviance: 169.68  on 722  degrees of freedom
AIC: 187.68

Number of Fisher Scoring iterations: 8
```

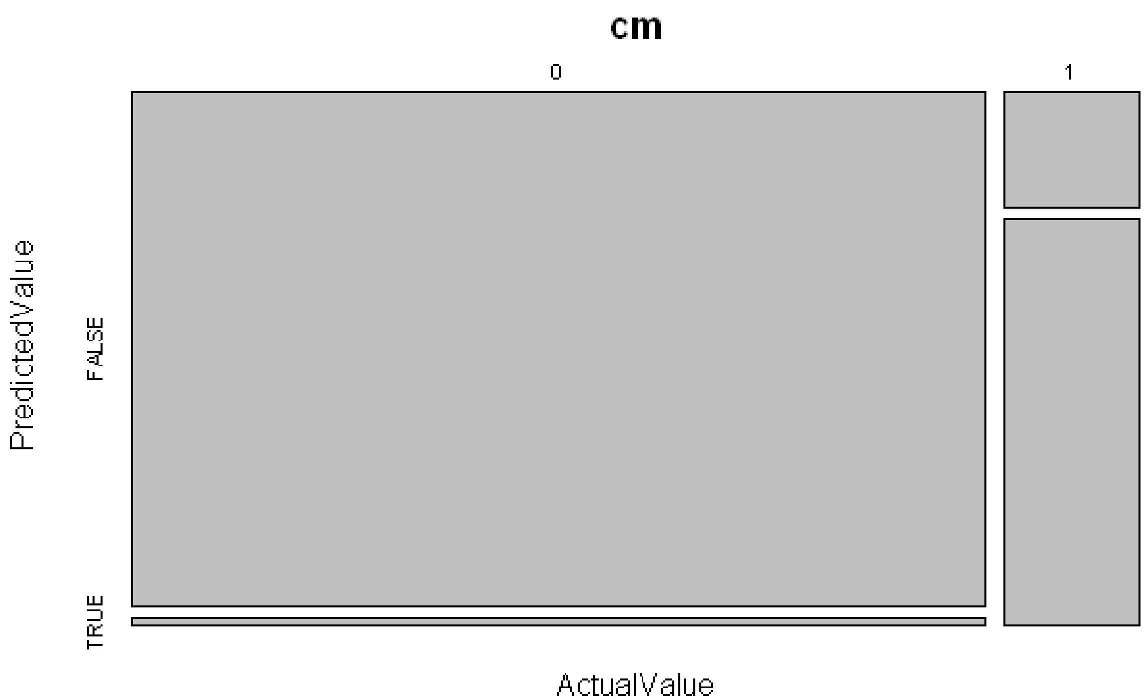
VIF – Age, Work.exp and Salary have high VIF

```
> vif(logit_model)
      Age    Gender  Engineer       MBA  Work.Exp    Salary  Distance   license 
10.016108  1.174902  1.071174  1.299387 16.237764  4.249760  1.505069  1.298327 
```

```
> table(carsDStest$CarUse, predict > 0.5)
           PredictedValue
ActualValue FALSE TRUE
          0   112    2
          1     4   14
```

Accuracy

```
"Confusion Matrix for Logistic Regression"
[1] "Accuracy :- 95.4545454545455"
[1] "FNR :- 22.2222222222222"
[1] "FPR :- 1.75438596491228"
[1] "precision :-  87.5"
[1] "recall//TPR :-  87.5"
[1] "Sensitivity :-  77.7777777777778"
[1] "Specificity :-  98.2456140350877"
```



<u>Model 2 – with Balanced SMOTE dataset</u>

We remove the variables with high VIF, and rebuild model.

param2<- CarUse ~ Gender + Engineer + MBA + Distance + license

```
Call:
glm(formula = param2, family = binomial, data = balanced.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7330  -0.4931  -0.2441   0.3888   2.7682

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.52342    0.67020 -11.226  < 2e-16 ***
Gender2     -0.27126    0.25607  -1.059  0.28947
Engineer1    0.96524    0.31148   3.099  0.00194 **
MBA1        -0.33531    0.26104  -1.284  0.19897
Distance     0.38733    0.03983   9.725  < 2e-16 ***
license1     2.46993    0.23272  10.613  < 2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 885.68  on 730  degrees of freedom
Residual deviance: 509.22  on 725  degrees of freedom
AIC: 521.22

Number of Fisher Scoring iterations: 6
```

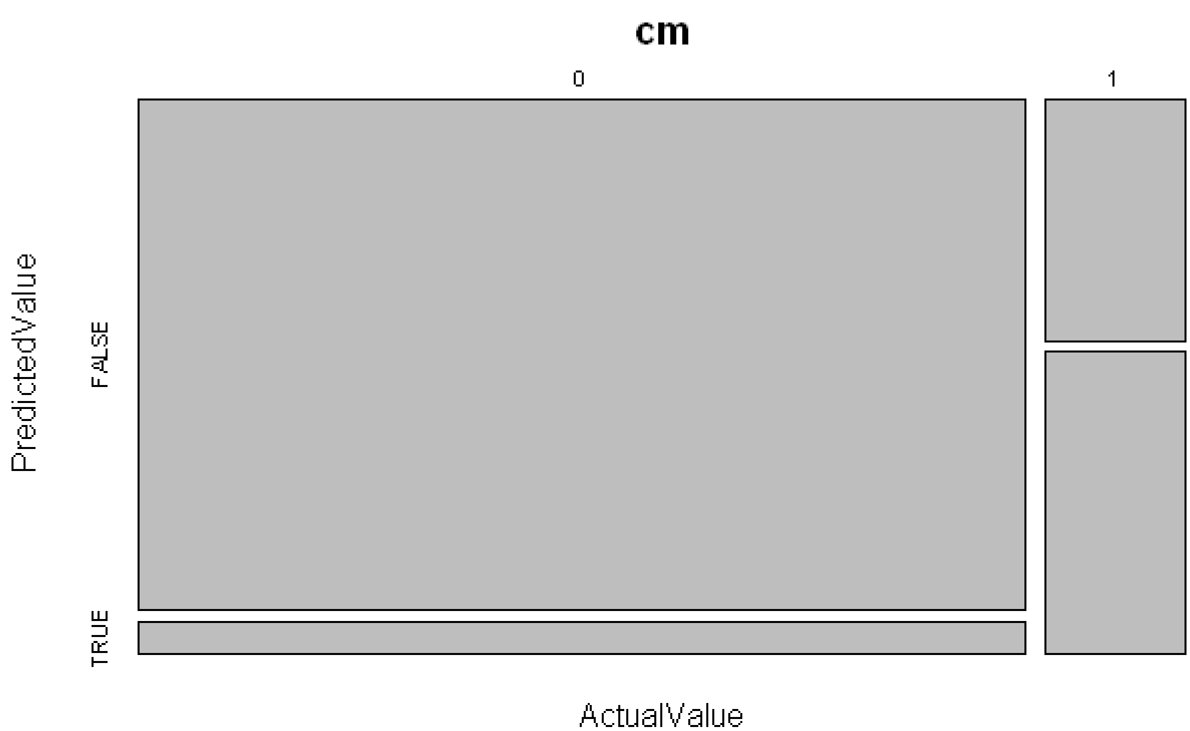
```
> vif(logit_model2)
  Gender Engineer      MBA Distance  license
1.021890 1.043260 1.031803 1.025964 1.073193
```

Confusion Matrix

```
           PredictedValue
ActualValue FALSE TRUE
          0   107    7
          1     8   10
```

Accuracy

```
"Confusion Matrix for Logistic Regression"
[1] "Accuracy :- 88.6363636363636"
[1] "FNR :- 44.4444444444444"
[1] "FPR :- 6.14035087719298"
[1] "precision :-  58.8235294117647"
[1] "recall//TPR :-  58.8235294117647"
[1] "Sensitivity :-  55.5555555555556"
[1] "Specificity :-  93.859649122807"
```



Model 3

param3 <- CarUse ~ Age + Gender + Engineer + MBA + Distance + license

```
Call:
glm(formula = param3, family = binomial, data = balanced.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.33756  -0.09295  -0.02303   0.01805   1.77596

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -34.99572    3.62086  -9.665  < 2e-16 ***
Age           0.91539    0.09899   9.247  < 2e-16 ***
Gender2      -0.95394    0.46501  -2.051   0.0402 *
Engineer1     0.04050    0.44395   0.091   0.9273
MBA1         -1.02101    0.43964  -2.322   0.0202 *
Distance      0.43582    0.07342   5.936 2.93e-09 ***
license1      1.65748    0.40301   4.113 3.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 885.68  on 730  degrees of freedom
Residual deviance: 187.05  on 724  degrees of freedom
AIC: 201.05

Number of Fisher Scoring iterations: 8
```

```
> vif(logit_model3)
     Age   Gender Engineer      MBA Distance  license
1.288645 1.088905 1.050391 1.277473 1.369810 1.137260
```

Confusion Matrix

```
           PredictedValue
ActualValue FALSE TRUE
          0   113    1
          1     4   14
```

Accuarcy

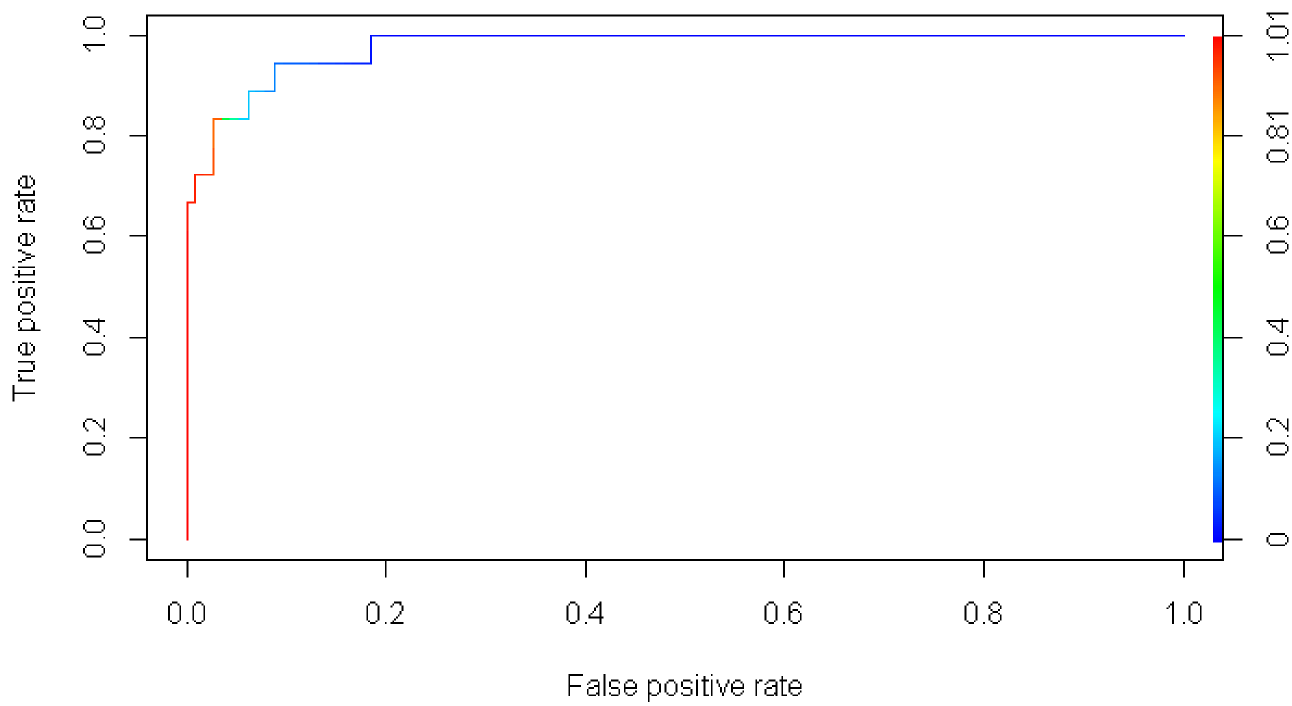
```
"Confusion Matrix for Logistic Regression"
[1] "Accuracy :- 96.2121212121212"
[1] "FNR :- 22.2222222222222"
[1] "FPR :- 0.87719298245614"
[1] "precision :-  93.3333333333333"
[1] "recall//TPR :-  93.3333333333333"
[1] "Sensitivity :-  77.7777777777778"
[1] "Specificity :-  99.1228070175439"
```

Comparing all three models, we can infer hat Model 3 performs better than the other two models, providing accuracy of 96%, Precision and Recall of 93%. Age, Distance, and License are the significant factors affecting the use of car as transport mode. This model also has lower AIC value than the other in the sense that it is less complex but still a good fit for the data.

Lets plot ROC curve for all 3 models –

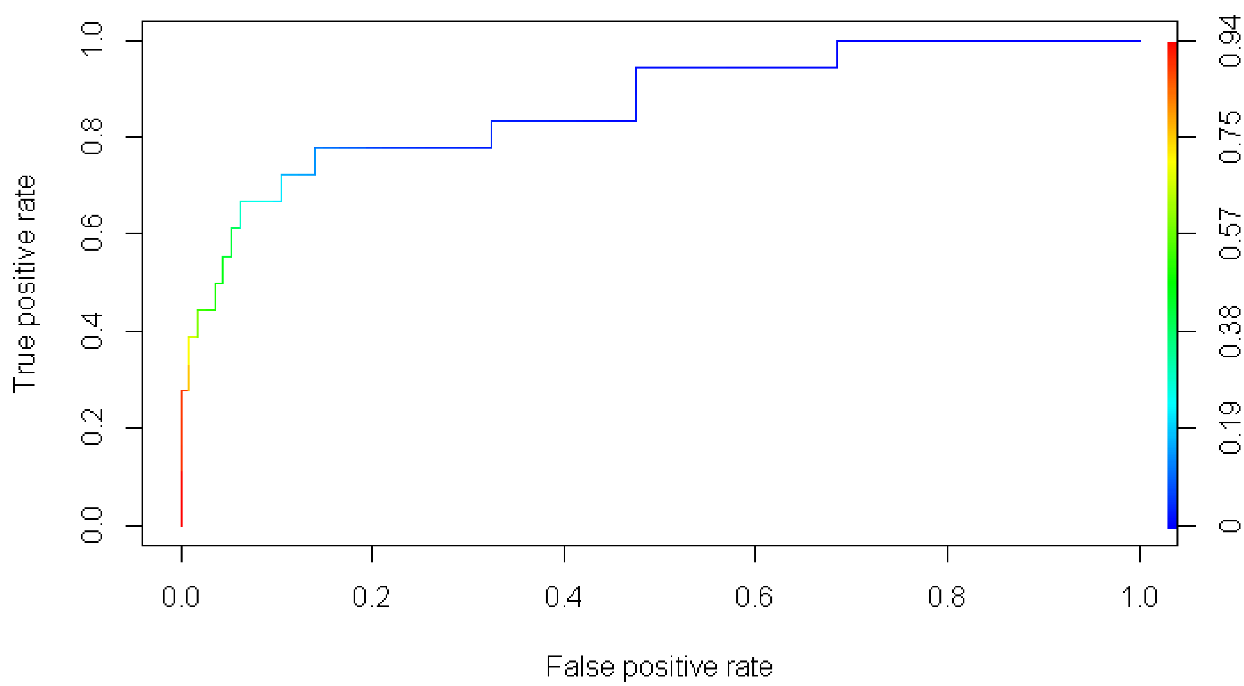
Model 1 ROC curve

```
[1] "Area Under the Curve for test Dataset: 0.974171539961014"
[1] "K-S Value for test Dataset 0.85672514619883"
```



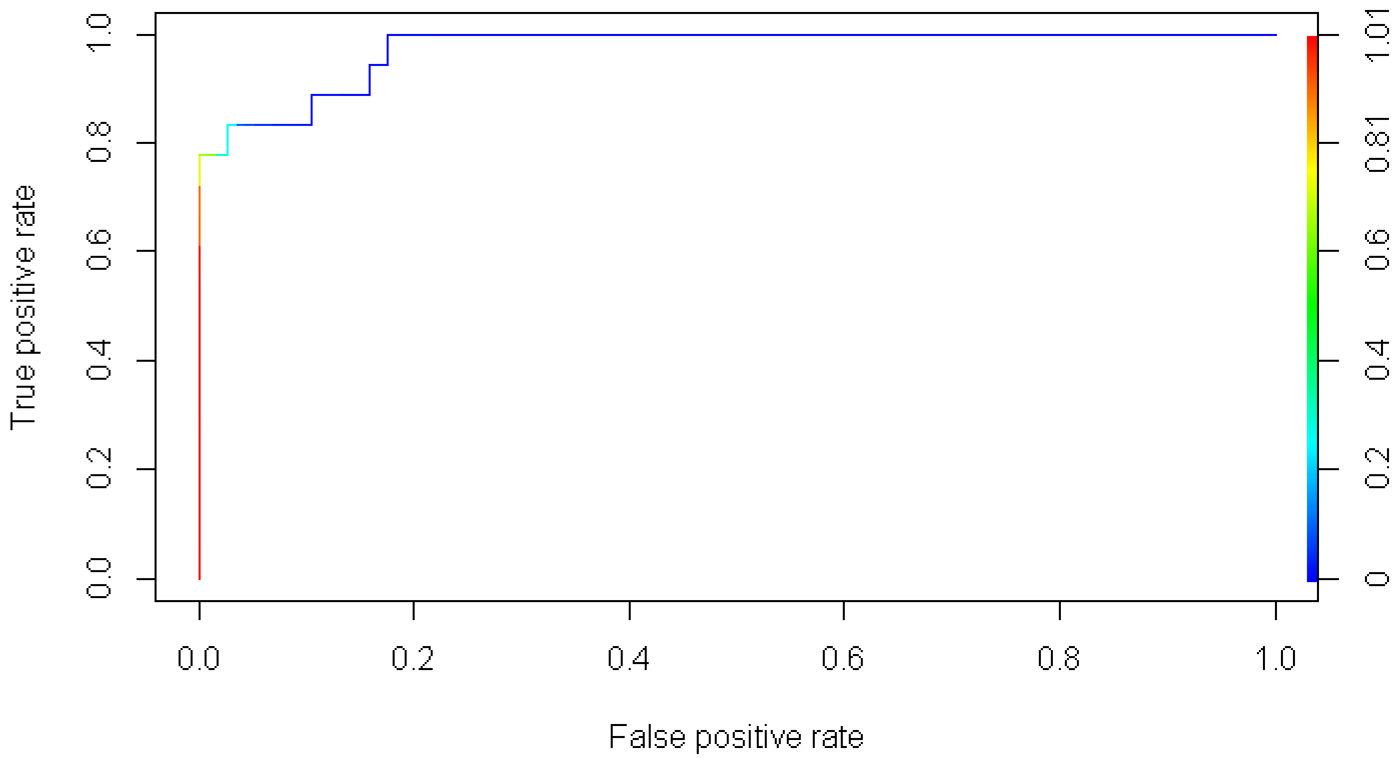
Model 2 ROC curve

```
[1] "Area Under the Curve for test Dataset: 0.865009746588694"
[1] "K-S Value for test Dataset 0.637426900584795"
```



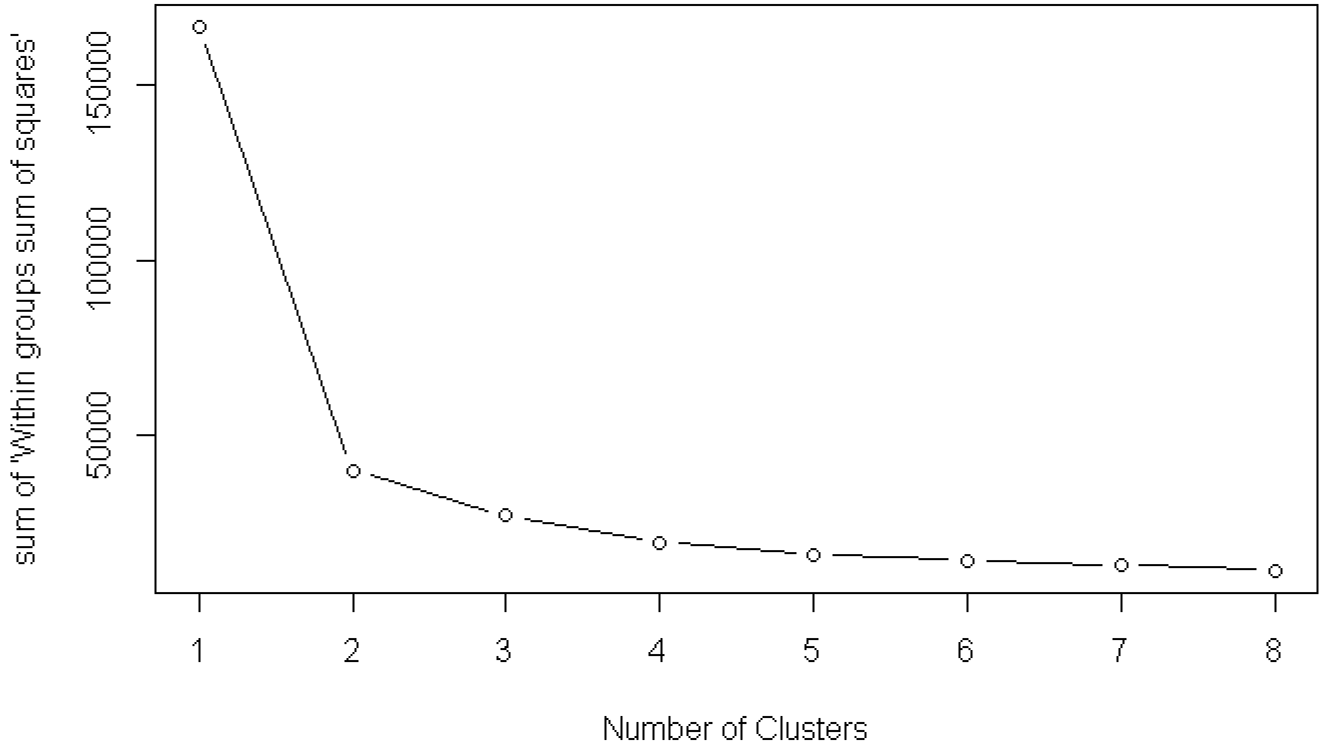
Model 3 ROC curve

```
[1] "Area Under the Curve for test Dataset: 0.974171539961014"
[1] "K-S Value for test Dataset 0.824561403508772"
```



KNN CLASSIFICATION

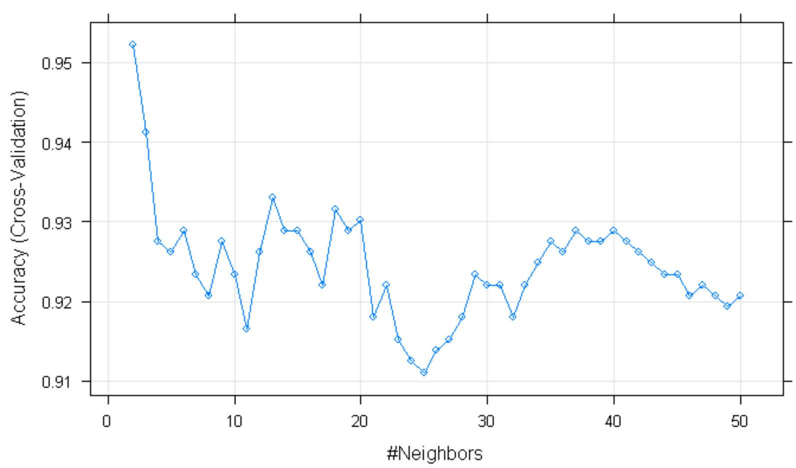
KNN classification iterations shows the best vale of k=5 gives optimum results. Accordingly, the confusion matrix is built to calculate the accuracy, precision, and recall values. Accuracy is 89% , precision and recall values are at 78% much higher than the logistic regression model.




```
> print(kn)
k-Nearest Neighbors

731 samples
  8 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 584, 585, 585, 585, 585
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   2  0.9521293  0.8825294
   3  0.9411891  0.8532634
   4  0.9274904  0.8210146
   5  0.9261485  0.8158808
   6  0.9288510  0.8247062
   7  0.9233529  0.8142532
   8  0.9206225  0.8056751
   9  0.9274811  0.8221611
  10  0.9233809  0.8103242
  11  0.9165409  0.7928262
  12  0.9261206  0.8176847
  13  0.9329699  0.8348310
  14  0.9288603  0.8252839
  15  0.9288696  0.8235565
  16  0.9261299  0.8176476
```

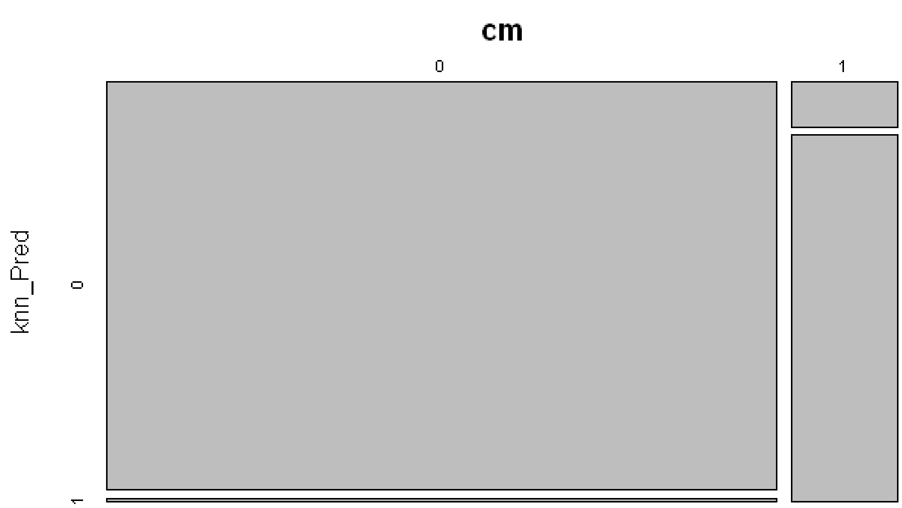
```
  17  0.9220296  0.8072978
  18  0.9316000  0.8303156
  19  0.9288603  0.8234481
  20  0.9302395  0.8250751
  21  0.9179294  0.7969138
  22  0.9220203  0.8067526
  23  0.9151990  0.7901900
  24  0.9124313  0.7830843
  25  0.9110800  0.7793493
  26  0.9138105  0.7858644
  27  0.9151896  0.7882358
  28  0.9179387  0.7956759
  29  0.9234181  0.8086534
  30  0.9220390  0.8049968
  31  0.9220390  0.8049968
  32  0.9179387  0.7952354
  33  0.9220296  0.8044071
  34  0.9247694  0.8108590
  35  0.9275091  0.8178488
  36  0.9261392  0.8139193
  37  0.9288789  0.8213542
  38  0.9275091  0.8181280
  39  0.9275184  0.8181952
  40  0.9288789  0.8220222
  41  0.9275091  0.8181280
  42  0.9261485  0.8149908
  43  0.9247787  0.8118323
  44  0.9234088  0.8083288
  45  0.9234088  0.8083288
  46  0.9206691  0.8013582
  47  0.9220390  0.8048616
  48  0.9206691  0.8013582
  49  0.9192992  0.7975607
  50  0.9206691  0.8013582

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 2.

Plotting the model result
```

It is showing Accuracy and Kappa metrics result for different k value.



We will use K value as 3 to predict classes for our test set. We can use predict() method. We are passing two arguments. The first parameter is our trained model and second parameter test dataset that holds our testing data frame. The predict() method returns a list, we are saving it in a variable.

```
Confusion Matrix and Statistics

   knn_Pred
      0   1
  0 113   1
  1   2  16

               Accuracy : 0.9773
                 95% CI : (0.935, 0.9953)
    No Information Rate : 0.8712
    P-Value [Acc > NIR] : 1.772e-05

                  Kappa : 0.9012

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9826
            Specificity : 0.9412
         Pos Pred Value : 0.9912
         Neg Pred Value : 0.8889
             Prevalence : 0.8712
         Detection Rate : 0.8561
   Detection Prevalence : 0.8636
      Balanced Accuracy : 0.9619

       'Positive' Class : 0

[1] "Accuracy :- 97.7272727272727"
[1] "FNR :- 11.1111111111111"
[1] "FPR :- 0.87719298245614"
[1] "precision :-  94.1176470588235"
[1] "recall//TPR :-  94.1176470588235"
[1] "Sensitivity :-  88.8888888888889"
[1] "Specificity :-  99.1228070175439"
```



NAIVE BAYES

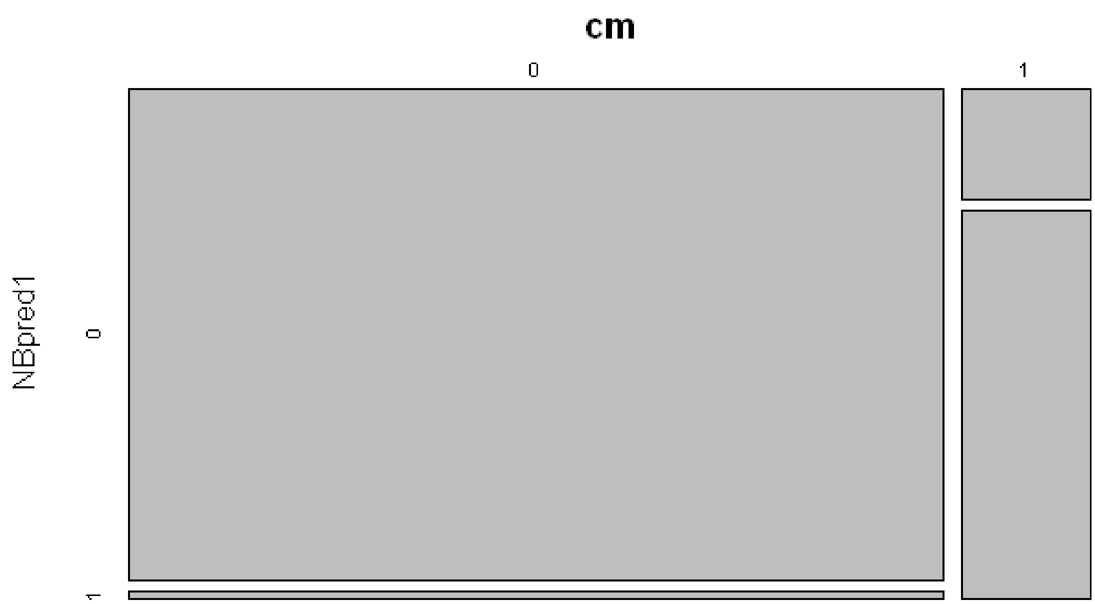
Naive Bayes Classifier for Discrete Predictors

A-priori probabilities indicates the distribution of our data. The Y values are the means and standard deviations of the predictors within each class.

Model 1 `Naive Bayes Classifier for Discrete Predictors`

```
Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.8621795 0.1378205 

Conditional probabilities:
   Age
Y       [,1]     [,2]
  0 26.56506 3.072575
  1 35.67442 3.212483

   Gender
Y           1         2
  0 0.3011152 0.6988848
  1 0.1627907 0.8372093

   Engineer
Y           0         1
  0 0.2750929 0.7249071
  1 0.1395349 0.8604651

   MBA
Y           0         1
  0 0.7211896 0.2788104
  1 0.8139535 0.1860465

   Work.Exp
Y        [,1]     [,2]
  0  4.907063 3.339428
  1 15.790698 4.548983

   Salary
Y       [,1]      [,2]
  0 13.20037  5.306076
  1 36.54651 13.103496

   Distance
Y       [,1]     [,2]
  0 10.83792 3.205623
  1 15.47907 3.671872

   license
Y           0         1
  0 0.8550186 0.1449814
  1 0.1627907 0.8372093
```

```
Contingency Table for Training Data
   NBpred1
      0   1
  0 112   2
  1   4  14

Accuracy NB Model 1
[1] "Accuracy :- 95.4545454545455"
[1] "FNR :- 22.2222222222222"
[1] "FPR :- 1.75438596491228"
[1] "precision :-  87.5"
[1] "recall//TPR :-  87.5"
[1] "Sensitivity :-  77.7777777777778"
[1] "Specificity :-  98.2456140350877"
```

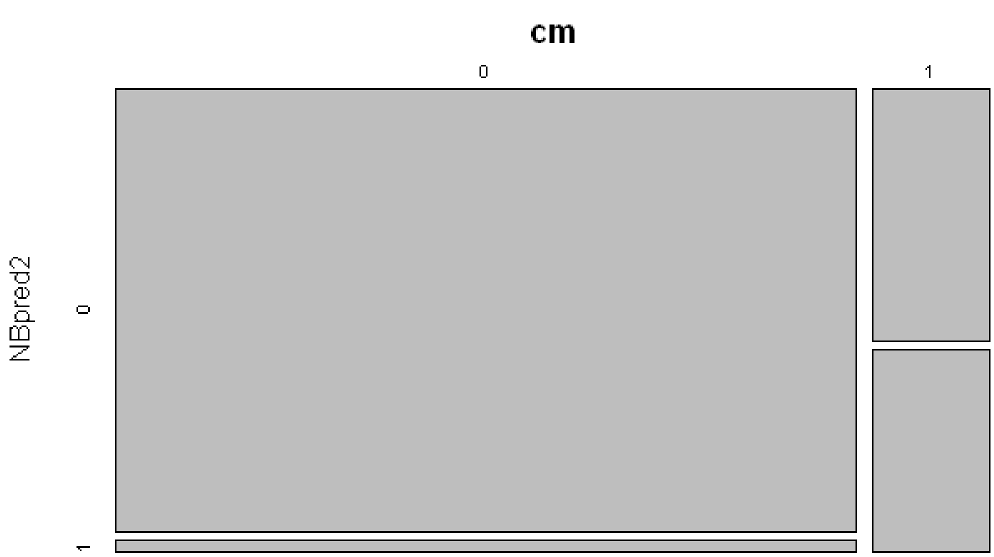


Model 2 Naive Bayes Classifier for Discrete Predictors

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.8621795 0.1378205 

Conditional probabilities:
   Gender
Y           1         2
  0 0.3011152 0.6988848
  1 0.1627907 0.8372093
```

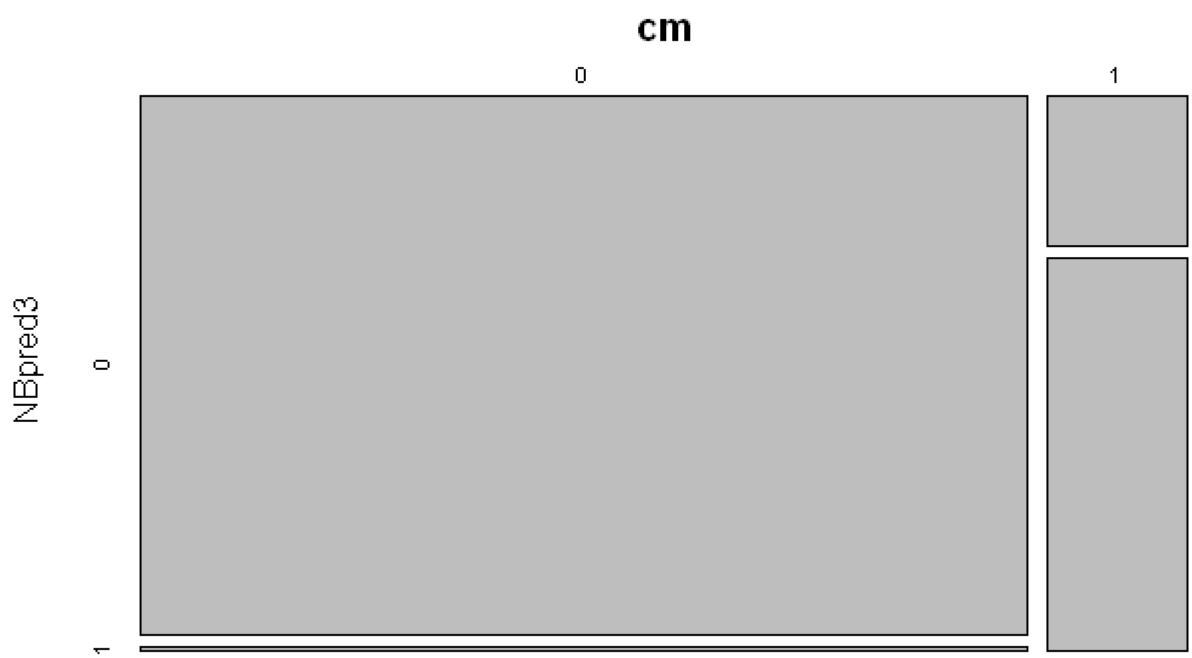
```
   Engineer
Y           0         1
  0 0.2750929 0.7249071
  1 0.1395349 0.8604651

   MBA
Y           0         1
  0 0.7211896 0.2788104
  1 0.8139535 0.1860465

   Distance
Y       [,1]     [,2]
  0 10.83792 3.205623
  1 15.47907 3.671872

   license
Y           0         1
  0 0.8550186 0.1449814
  1 0.1627907 0.8372093
```

```
 Contingency Table for Training Data
  NBpred2
      0   1
  0 111   3
  1  10   8
Accuracy NB Model 2
[1] "Accuracy :- 90.1515151515152"
[1] "FNR :- 55.5555555555556"
[1] "FPR :- 2.63157894736842"
[1] "precision :-  72.7272727272727"
[1] "recall//TPR :-  72.7272727272727"
[1] "Sensitivity :-  44.4444444444444"
[1] "Specificity :-  97.3684210526316"
```



Model 3 Naive Bayes Classifier for Discrete Predictors

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1 
0.8621795 0.1378205 

Conditional probabilities:
   Age
Y       [,1]     [,2]
  0 26.56506 3.072575
  1 35.67442 3.212483

   Gender
Y           1         2
  0 0.3011152 0.6988848
  1 0.1627907 0.8372093

   Engineer
Y           0         1
  0 0.2750929 0.7249071
  1 0.1395349 0.8604651

   MBA
Y           0         1
  0 0.7211896 0.2788104
  1 0.8139535 0.1860465

   Distance
Y       [,1]     [,2]
  0 10.83792 3.205623
  1 15.47907 3.671872

   license
Y           0         1
  0 0.8550186 0.1449814
  1 0.1627907 0.8372093

Contingency Table for Training Data
   NBpred3
      0   1
  0 113   1
  1   5  13

Accuracy NB Model 3
[1] "Accuracy :- 95.4545454545455"
[1] "FNR :- 27.7777777777778"
[1] "FPR :- 0.87719298245614"
[1] "precision :-  92.8571428571429"
[1] "recall//TPR :-  92.8571428571429"
[1] "Sensitivity :-  72.2222222222222"
[1] "Specificity :-  99.1228070175439"
```



As in Logistic Regression, we have a highest test accuracy in Model3 of about 95.5% , and precision and recall values at 93%.

BAGGNG AND BOOSTING

Bagging is a prediction model for fitting multiple versions of a prediction model and then combining (or ensembling) them into an aggregated prediction. We can also apply bagging within caret to see how well our ensemble will generalize.

```
set.seed(123)

# train bagged model
cars_bag1 <- bagging(
  formula = CarUse ~ .,
  data = carsDStrain,
  nbagg = 100,
  coob = TRUE,
  control = rpart.control(minsplit = 2, cp = 0)
)
cars_bag1


cars_bag2 <- bagging(
  formula = CarUse ~ .,
  data = carsDStrain,
  nbagg = 25,
  coob = TRUE,
  control = rpart.control(maxdepth=5, minsplit=4)
)
cars_bag2
```

```
carsDStest$pred.class <- predict(cars_bag1, carsDStest)
```

```
Bagging classification trees with 100 bootstrap replications

Call: bagging.data.frame(formula = CarUse ~ ., data = carsDStrain,
    nbagg = 100, coob = TRUE, control = rpart.control(minsplit = 2,
        cp = 0))

Out-of-bag estimate of misclassification error:  0.0385
```

```
Bagging classification trees with 25 bootstrap replications

Call: bagging.data.frame(formula = CarUse ~ ., data = carsDStrain,
    nbagg = 25, coob = TRUE, control = rpart.control(maxdepth = 5,
        minsplit = 4))

Out-of-bag estimate of misclassification error:  0.0417
```

```
##Boosting


gbm.fit <- gbm(
  formula = CarUse ~ .,
  distribution = "bernoulli",#we are using bernoulli because we are doing a logistic and
want probabilities
  data = carsDStrain,
  n.trees = 10000, #these are the number of stumps
  interaction.depth = 1,#number of splits it has to perform on a tree (starting from a
single node)
  shrinkage = 0.001,#shrinkage is used for reducing, or shrinking the impact of each
additional fitted base-learner(tree)
  cv.folds = 5,#cross validation folds
 objective = "binary:logistic",  # for regression models
 n.cores = NULL, # will use all cores by default
  verbose = FALSE#after every tree/stump it is going to show the error and how it is
changing
)


carsDStest$pred.class <- predict(gbm.fit, CarUse, type = "response")

table(carsDStest$CarUse,carsDStest$pred.class>0.5)
```

MODEL COMPARISON USING - CONFUSION MATRIX INTERPRETATION FOR ALL MODELS

CONFUSION MATRIX	LOGISTIC REGRESSION (Model 3)	KNN	NAÏVE BAYES (Model 3)
Accuracy	96.2%	97.7%	95.5%
FNR	22.2%	11.1%	27.8%
FPR	0.9%	0.9%	0.9%
PRECISION	93.3%	94.1%	92.9%
RECALL (TPR)	93.3%	94.1%	92.9%
SENSITIVITY (TNR)	77.7%	88.9%	72.2%
SPECIFICITY	99.1%	99.1%	99.1%

Accuracy seems to be good in all models. Based on the precision, recall value, KNN model is better than other models is classifying the cluster of employees who are likely to use car for transport. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant ones, while high recall means that an algorithm returned most of the relevant results. However, in this problem, we cannot explain variable importance in KNN model. As our problem statement is to build a model that explains best the employee's decision to use car as means of transport, we should go with next best model, which is Logistic regression Model3, which gives that Age, Distance and License are most important variables for predicting car usage.

ACTIONABLE INSIGHTS AND RECOMMENDATIONS

To guide the recommendation, let's go back and try to answer our queries about predictors:

- Does the Age of the Employee crucial in deciding preference of car as mode of transport?

Yes. Age is significant factor or a predictor that indicates that range of age of employees using car lies between 33 to 40 years. While majority of those using 2-Wheeler and Public Transport lies approximately between 23 to 28 years. So higher age seems to be a driving factor for transport mode selection.

- Do individuals higher Salary and Work Experience more like to use car as mode of transport?

Work experience and Salary are highly correlated to the Age variable and hence these variables may be ignored, given that Age alone can significantly predict the category of Employees who are likely to use car as mode of transport.

- Does the Gender play a crucial role in transport preferences?

Due to the disparity in Gender employment in this particular dataset, the data set contains fewer Female employees compared to Male Employees. Looking at the proportion table, there is not much difference in either Gender for choosing Car as mode of transport.

- Does the Distance or License play a crucial role in transport preferences?
  Yes. Models show that both distance and valid driving license play a crucial role in choosing car as mode of transport. Those with driving license and those travelling longer distance are more likely to choose car as means of transport.