# **Machine Learning Group Assignment**

**Case – Predict if ‘Car’ is the preferred mode of transport for employees**

Group 7

Venkatesh Taduvayi

Deeshant Kamal

Paarth Tiwari

Note –

For easy navigation, please enable ‘Navigation’ option from ‘View’ menu in MS word

Contents

[**Machine Learning Group Assignment** 1](#_Toc45458167)

[0. Problem Statement 3](#_Toc45458168)

[1. Exploratory Data Analysis 5](#_Toc45458169)

[a) Data summary 5](#_Toc45458170)

[b) Univariate Analysis 8](#_Toc45458171)

[c) Bivariate Analysis 14](#_Toc45458172)

[d) Check for Multicollinearity - Plot the graph based on Multicollinearity. 22](#_Toc45458173)

[2. Data Preparation 25](#_Toc45458174)

[3. Building Models 26](#_Toc45458175)

[a) Applying Logistic Regression & Interpret the results - (Model Performance Measures) 26](#_Toc45458176)

[b) Applying KNN Model & Interpret results 26](#_Toc45458177)

[c) Applying Naïve Bayes 26](#_Toc45458178)

[d) Confusion matrix summary for all models 38](#_Toc45458179)

[e) Remarks on Model validation exercise <Which model performed the best> 38](#_Toc45458180)

[4. Ensemble Techniques 38](#_Toc45458181)

[a) Bagging 38](#_Toc45458182)

[b) Boosting 38](#_Toc45458183)

[5. Actionable Insights and Recommendations 38](#_Toc45458184)

[Appendix 38](#_Toc45458185)

## Problem Statement

This project requires you to understand what mode of transport employees prefers to commute to their office. The attached dataset **Cars.csv** View in a new window 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 |

**EDA**

* Perform Exploratory Data Analysis on the dataset
* Illustrate the insights based on EDA
* Multicollinearity check and summarization of problem statement for business stakeholders

**Data Preparation**

* Prepare the data for analysis

**Modeling**

Create multiple models and explore how each model perform using appropriate model performance metrics:

* KNN
* Naive Bayes (is it applicable here? comment and if it is not applicable, how can you build an NB model in this case?)
* Logistic Regression
* Apply both bagging and boosting modeling procedures to create 2 models and compare its accuracy with the best model of the above step. Actionable Insights & Recommendations

Summarize your findings from the exercise in a concise yet actionable note

Evaluation Rubric:

| **Criteria** | **Ratings** | **Pts** |
| --- | --- | --- |
| This criterion is linked to a Learning Outcome 1) EDA: a) Basic data summary, Univariate, Bivariate analysis, graphs, Check for Outliers and missing values and check the summary of the dataset |  | 5.0 pts |
| This criterion is linked to a Learning Outcome 1) EDA: b) Check for Multicollinearity - Plot the graph based on Multicollinearity |  | 2.0 pts |
| This criterion is linked to a Learning Outcome 1) EDA: c) Interpreting the business problems and sharing the observations |  | 3.0 pts |
| This criterion is linked to a Learning Outcome 2) Data Preparation |  | 6.0 pts |
| This criterion is linked to a Learning Outcome 3. a) Applying Logistic Regression & Interpret the results - (Model Performance Measures) |  | 2.0 pts |
| This criterion is linked to a Learning Outcome 3. b) Applying KNN Model & Interpret results |  | 2.0 pts |
| This criterion is linked to a Learning Outcome 3. c) Applying Naïve Bayes Model & Interpret results (is it applicable here? comment and if it is not applicable, how can you build an NB model in this case?) |  | 2.0 pts |
| This criterion is linked to a Learning Outcome 3. d) Confusion matrix interpretation |  | 2.0 pts |
| This criterion is linked to a Learning Outcome 3. e) Remarks on Model validation exercise <Which model performed the best> |  | 2.0 pts |
| This criterion is linked to a Learning Outcome 4. a) Bagging |  | 5.0 pts |
| This criterion is linked to a Learning Outcome 4. b) Boosting |  | 5.0 pts |
| This criterion is linked to a Learning Outcome 5) Actionable Insights and Recommendations |  | 4.0 pts |

## Exploratory Data Analysis

### Data summary

#### Basic Summary

**Observations** –

* Data set has 444 observations and 9 variables.
* Distribution of the variables

|  |  |
| --- | --- |
| **Category** | **Variables** |
| Continous | Age,Work.Exp,Salary,Distance |
| Categorical | Gender,Engineer,MBA,license,Transport |

* **Predictor/Independant Variables**: Age,Work.Exp,Salary,Distance,Gender,Engineer,MBA,license
* **Target/Dependant Variable**: Transport
* The 9 attributes are divided as below -
  + Age - Captured as integer and holds Age of the Employee Age Gender - Captured as character and holds Gender of Employee. (Male/Female)
  + Engineer - Captured as integer and holds whether employee is Engineer or not. For Engineer =1 , Non Engineer =0
  + MBA - Captured as integer and hold MBA status of employee. For MBA =1 , Non MBA =0
  + Work.Exp - Captured as integer and holds the work experience of employee in years
  + Salary - Captured as number and holds Salary in Lakhs per Annum
  + Distance - Captured as Number and holds Distance in Kms from Home to Office
  + license - Captured as integer and holds If Employee has Driving License -1, If not, then 0
  + Transport - Captured as character and holds Mode of Transport

#### Detailed Summary

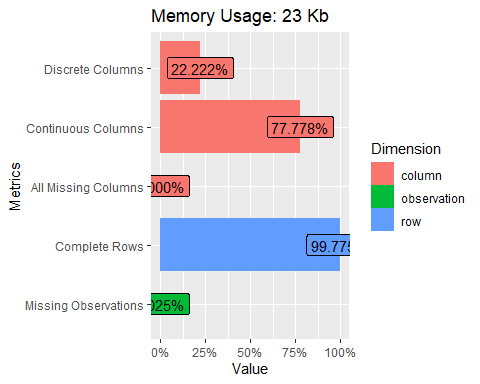
* Age - Captures Employee Age in years with minimum of 18 and max of 43 years. Overall Mean stand at 27.75 years and median at 27.
* Gender - Captures Employee Gender as either Male or Female. Male employee count is 316 and Female 128.
* Engineer - Captures employee Engineer status with 1 capturing as Engineer and 0 as Non Engineer.
* Work.Exp - Captures employee work experience. ranging from minimum of 0 to max of 24 years. Mean stand at 6.3 years and mode as 5 years.
* Salary - Captures the employee salary in Lakhs per Annum with minimum of 6.5 to max of 57 lakhs. Average salary captured in 16.24 lakhs and median as 13.6 lakhs.
* Distance - Captures the employee distance from home in kms. Minimum is at 3.4 km and max as 23.4 kms. Average distance is 11.32 kms.
* License - Captures employee license with 1 as having license and 0 as not.
* Transport - Captures the employees mode of transport.
* Among these 444 Observations, only 61 (13.7%) use car as mode of Transport. 83 (18.6%) use 2wheeler and 300 (67.5%) use Public Transport
* Count of Unique values for each column in Sample Dataset as below -

Age Gender Engineer MBA Work.Exp Salary Distance license Transport

25 2 2 3 24 122 137 2 3

#### Missing values and treatment

plot\_intro(carsdata)



**Missing Values -**

Sample dataset has 99.7% observations has all values.

**Missing value detection and treatment -**

One missing value is found in the dataset (MBA - NA’s: 1):

# Complete cases  
table(complete.cases(carsdata))

##   
## TRUE   
## 444

cars\_df[which(!complete.cases(carsdata)),]

## [1] Age Gender Engineer MBA Work.Exp Salary Distance   
## [8] license Transport  
## <0 rows> (or 0-length row.names)

We could just make one NA to Zero (No MBA) or impute using standard method.

carsdata$MBA[is.na(carsdata$MBA)] = 0

### Univariate Analysis

#### Continuous Variables

**Age,Work.Exp,Salary,Distance**

cont.vars = c("Age","Work.Exp","Salary","Distance")

1. Measures of central tendency & Spread of data

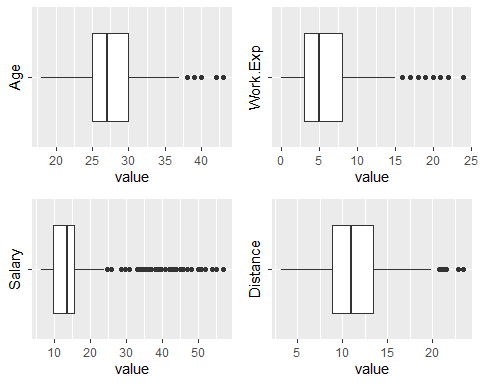
summary(cars\_df)

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

* **Age** - Captures Employee Age in years with minimum of 18 and max of 43 years. Overall Mean stand at 27.75 years and median at 27.
* **Gender** - Captures Employee Gender as either Male or Female. Male employee count is 316 and Female 128.
* **Engineer** - Captures employee Engineer status with 1 capturing as Engineer and 0 as Non Engineer.
* **Work.Exp** - Captures employee work experience. ranging from minimum of 0 to max of 24 years. Mean stand at 6.3 years and mode as 5 years.
* **Salary** - Captures the employee salary in Lakhs per Annum with minimum of 6.5 to max of 57 lakhs. Average salary captured in 16.24 lakhs and median as 13.6 lakhs.
* **Distance** - Captures the employee distance from home in kms. Minimum is at 3.4 km and max as 23.4 kms. Average distance is 11.32 kms.
* **License** - Captures employee license with 1 as having license and 0 as not.
* **Transport** - Captures the employees’ mode of transport.

Among these 444 Observations, only 61 (13.7%) use car as mode of Transport. 83 (18.6%) use 2wheeler and 300 (67.5%) use Public Transport

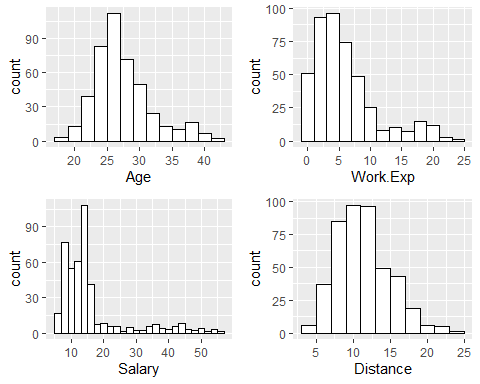
Box Plots

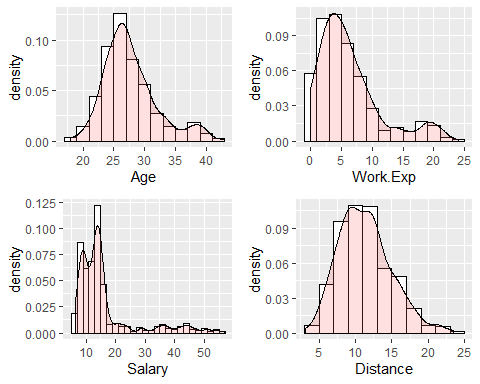


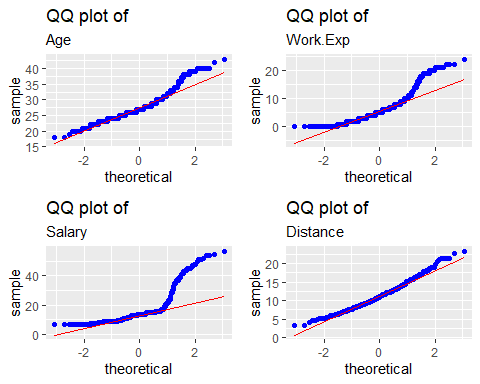
**Observations:**

There are several outliers that can be observed in the Salary distributions in the given dataset. However, since the data set is limited, we will be including all the data points for the purpose of this study.

**Histograms, density and QQ plots**







**Observations**

Following inferences can be drawn from the above graphs -

* **Age** - Age distribution is fairly uniform as can be seen from the QQ plot. The mean age is around 27 years and most the of the data points are concentrated around this point.
* **Work Experience** - Data is right-skewed with a mean work exp of 6 years and a max work experience of 24(outlier). It is likely that work experience shares a linear relationship with age and we can verify this in the bi-variate analysis.
* **Salary** - Data is heavily right-skewed with a mean salary of 16.24LPA and a max of 57LPA. Distance - Data is fairly uniform as can be seen from the QQ plot. Both mean and median values are around 11km which indicates that people prefer to reside within this range from their workplace.

#### Categorical Variables

**Gender,Engineer,MBA,license,Transport**

cat.vars = c("Gender","Engineer","MBA","license","Transport")

1. Distribution of categorical variables by Counts & proportions

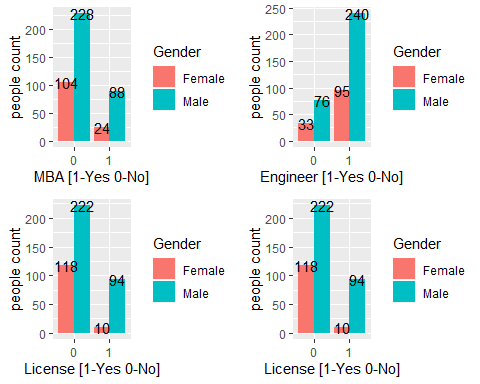
lapply(cars\_df[,cat.vars], table)

## $Gender  
##   
## Female Male   
## 128 316   
##   
## $Engineer  
##   
## 0 1   
## 109 335   
##   
## $MBA  
##   
## 0 1   
## 332 112   
##   
## $license  
##   
## 0 1   
## 340 104   
##   
## $Transport  
##   
## 2Wheeler Car Public Transport   
## 83 61 300

lapply(lapply(cars\_df[,cat.vars], table),prop.table)

## $Gender  
##   
## Female Male   
## 0.2882883 0.7117117   
##   
## $Engineer  
##   
## 0 1   
## 0.2454955 0.7545045   
##   
## $MBA  
##   
## 0 1   
## 0.7477477 0.2522523   
##   
## $license  
##   
## 0 1   
## 0.7657658 0.2342342   
##   
## $Transport  
##   
## 2Wheeler Car Public Transport   
## 0.1869369 0.1373874 0.6756757

1. Bar Plots



**Observations**

Distribution of categorical data can be summarized by the following table:

* By Gender:

|  |  |
| --- | --- |
| Male | 71.17% |
| Female | 28.82% |

* By Educational Qualification

|  |  |
| --- | --- |
| Non-Engineer | 75.45% |
| Engineer | 24.54% |
| Non-MBA | 74.71% |
| MBA | 23.42% |

* By possession of License:

|  |  |
| --- | --- |
| Have license | 23.42% |
| Don’t have license | 76.57% |

* By preferred mode of transport

|  |  |
| --- | --- |
| 2-Wheeler | 18.69% |
| car | 13.73% |
| Public Transport | 67.56% |

### Bivariate Analysis

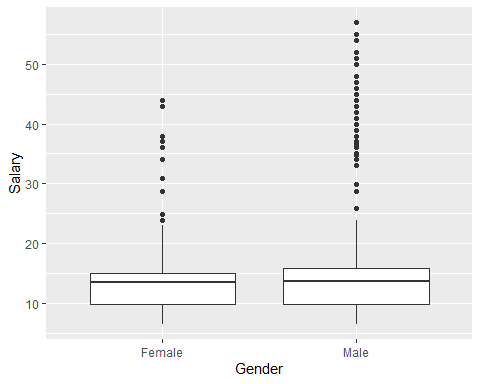
#### Categorical & Continuous

* 1. Gender and Salary

by(cars\_df$Salary, cars\_df$Gender, summary)

## cars\_df$Gender: Female  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.50 9.75 13.50 14.66 15.05 44.00   
## ------------------------------------------------------------   
## cars\_df$Gender: Male  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.50 9.80 13.60 16.88 15.90 57.00

cars\_df %>%  
 ggplot(aes(x = Gender, y= Salary)) +  
 geom\_boxplot()



**Observations**

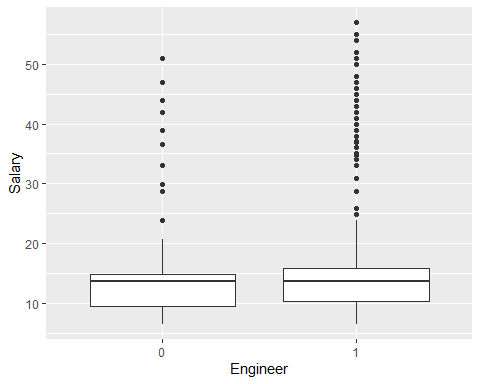
From the above boxplots, it appears that the mean income of men at 16.88LPA is slightly higher than women who earn around 14.66LPA on an average.

* 1. Engineer & Salary

by(cars\_df$Salary, cars\_df$Engineer, summary)

## cars\_df$Engineer: 0  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.50 9.50 13.60 14.65 14.80 51.00   
## ------------------------------------------------------------   
## cars\_df$Engineer: 1  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.50 10.25 13.60 16.76 15.85 57.00

cars\_df %>%  
 ggplot(aes(x = Engineer, y= Salary)) +  
 geom\_boxplot()



**Observation**

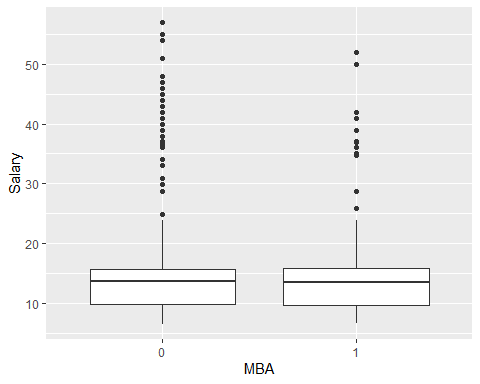
Mean salary of an engineer is 16.76LPA whereas that of a person without engineering degree is 14.65LPA. This implies that Engineers earn nearly 14% higher, on average, compared to people who do not have an engineering degree.

* 1. MBA & Salary

by(cars\_df$Salary, cars\_df$MBA, summary)

## cars\_df$MBA: 0  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.50 9.80 13.60 16.28 15.70 57.00   
## ------------------------------------------------------------   
## cars\_df$MBA: 1  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 6.60 9.65 13.55 16.11 15.82 52.00

cars\_df %>%  
 ggplot(aes(x = MBA, y= Salary)) +  
 geom\_boxplot()



**Observation**

Surprisingly, mean salaries of MBA and Non-MBA grads is nearly the same at 16.11LPA and 16.29LPA respectively. In fact, non-MBA grads earn nearly 1% higher than MBA grads on average.

* 1. Gender & Distance

wilcox.test(cars\_df$Distance~cars\_df$Gender)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: cars\_df$Distance by cars\_df$Gender  
## W = 19102, p-value = 0.3598  
## alternative hypothesis: true location shift is not equal to 0

**Inference** p-value = 0.3598 > 0.05 i.e. H0 is accepted There is no significant difference in the distance travelled amongst the two genders.

* 1. MBA & Distance

wilcox.test(cars\_df$Distance~cars\_df$MBA)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: cars\_df$Distance by cars\_df$MBA  
## W = 17847, p-value = 0.526  
## alternative hypothesis: true location shift is not equal to 0

**Inference**

p-value = 0.5392 > 0.05 i.e. H0 is accepted. There is no significant difference in the distance travelled amongst people MBA and non-MBA grads.

* 1. Engineering & Distance

wilcox.test(cars\_df$Distance~cars\_df$Engineer)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: cars\_df$Distance by cars\_df$Engineer  
## W = 16552, p-value = 0.1427  
## alternative hypothesis: true location shift is not equal to 0

**Inference**

p-value = 0.1427 > 0.05 i.e. H0 is accepted There is no significant difference in the distance travelled amongst people Engineering and non-Engineering grads.

#### Categorical & Categorical

* **By Count**

The following 3-way frequency table describes the counts of categorical variables

mytable <- xtabs(~Gender+Engineer+MBA+Transport, data=cars\_df)  
ftable(mytable)

## Transport 2Wheeler Car Public Transport  
## Gender Engineer MBA   
## Female 0 0 9 1 17  
## 1 0 0 6  
## 1 0 22 10 45  
## 1 7 2 9  
## Male 0 0 11 8 41  
## 1 3 0 13  
## 1 0 24 30 114  
## 1 7 10 55

* **By Percentage**

##### Distribution of engineers by gender

CrossTable(cars\_df$Engineer,cars\_df$Gender, prop.t = T)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 444   
##   
##   
## | cars\_df$Gender   
## cars\_df$Engineer | Female | Male | Row Total |   
## -----------------|-----------|-----------|-----------|  
## 0 | 33 | 76 | 109 |   
## | 0.079 | 0.032 | |   
## | 0.303 | 0.697 | 0.245 |   
## | 0.258 | 0.241 | |   
## | 0.074 | 0.171 | |   
## -----------------|-----------|-----------|-----------|  
## 1 | 95 | 240 | 335 |   
## | 0.026 | 0.010 | |   
## | 0.284 | 0.716 | 0.755 |   
## | 0.742 | 0.759 | |   
## | 0.214 | 0.541 | |   
## -----------------|-----------|-----------|-----------|  
## Column Total | 128 | 316 | 444 |   
## | 0.288 | 0.712 | |   
## -----------------|-----------|-----------|-----------|

##### Distribution of MBA grads by gender

CrossTable(cars\_df$MBA,cars\_df$Gender, prop.t = T)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 444   
##   
##   
## | cars\_df$Gender   
## cars\_df$MBA | Female | Male | Row Total |   
## -------------|-----------|-----------|-----------|  
## 0 | 104 | 228 | 332 |   
## | 0.718 | 0.291 | |   
## | 0.313 | 0.687 | 0.748 |   
## | 0.812 | 0.722 | |   
## | 0.234 | 0.514 | |   
## -------------|-----------|-----------|-----------|  
## 1 | 24 | 88 | 112 |   
## | 2.128 | 0.862 | |   
## | 0.214 | 0.786 | 0.252 |   
## | 0.188 | 0.278 | |   
## | 0.054 | 0.198 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 128 | 316 | 444 |   
## | 0.288 | 0.712 | |   
## -------------|-----------|-----------|-----------|  
##   
##

##### Distribution of license holders by Gender

CrossTable(cars\_df$license,cars\_df$Gender, prop.t = T)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 444   
##   
##   
## | cars\_df$Gender   
## cars\_df$license | Female | Male | Row Total |   
## ----------------|-----------|-----------|-----------|  
## 0 | 118 | 222 | 340 |   
## | 4.074 | 1.650 | |   
## | 0.347 | 0.653 | 0.766 |   
## | 0.922 | 0.703 | |   
## | 0.266 | 0.500 | |   
## ----------------|-----------|-----------|-----------|  
## 1 | 10 | 94 | 104 |   
## | 13.317 | 5.394 | |   
## | 0.096 | 0.904 | 0.234 |   
## | 0.078 | 0.297 | |   
## | 0.023 | 0.212 | |   
## ----------------|-----------|-----------|-----------|  
## Column Total | 128 | 316 | 444 |   
## | 0.288 | 0.712 | |   
## ----------------|-----------|-----------|-----------|  
##   
##

##### Distribution of transport media by gender

CrossTable(cars\_df$Transport,cars\_df$Gender, prop.t = T)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 444   
##   
##   
## | cars\_df$Gender   
## cars\_df$Transport | Female | Male | Row Total |   
## ------------------|-----------|-----------|-----------|  
## 2Wheeler | 38 | 45 | 83 |   
## | 8.276 | 3.352 | |   
## | 0.458 | 0.542 | 0.187 |   
## | 0.297 | 0.142 | |   
## | 0.086 | 0.101 | |   
## ------------------|-----------|-----------|-----------|  
## Car | 13 | 48 | 61 |   
## | 1.196 | 0.484 | |   
## | 0.213 | 0.787 | 0.137 |   
## | 0.102 | 0.152 | |   
## | 0.029 | 0.108 | |   
## ------------------|-----------|-----------|-----------|  
## Public Transport | 77 | 223 | 300 |   
## | 1.041 | 0.421 | |   
## | 0.257 | 0.743 | 0.676 |   
## | 0.602 | 0.706 | |   
## | 0.173 | 0.502 | |   
## ------------------|-----------|-----------|-----------|  
## Column Total | 128 | 316 | 444 |   
## | 0.288 | 0.712 | |   
## ------------------|-----------|-----------|-----------|  
##   
##

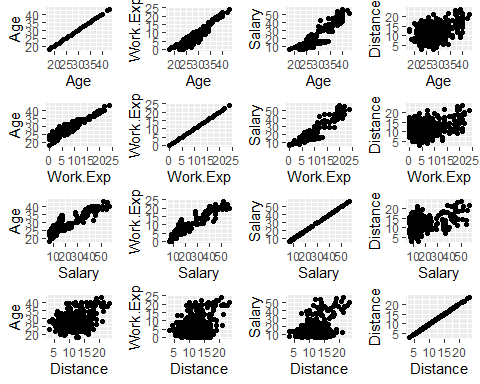
#### Continuous & Continuous

Correlation and Scatter Plots

cor(cars\_df[,cont.vars])

## Age Work.Exp Salary Distance  
## Age 1.0000000 0.9322364 0.8606732 0.3528725  
## Work.Exp 0.9322364 1.0000000 0.9319745 0.3727350  
## Salary 0.8606732 0.9319745 1.0000000 0.4423591  
## Distance 0.3528725 0.3727350 0.4423591 1.0000000





**Observation**

From the above plots a **high** correlation is observed between

* Age & Work experience
* Age & salary Work Experience and salary

This is also confirmed by the scatter plots where a **linear** positive relationship can be observed between the above-mentioned pairs.

No clear pattern is observed between Distance and other variables.

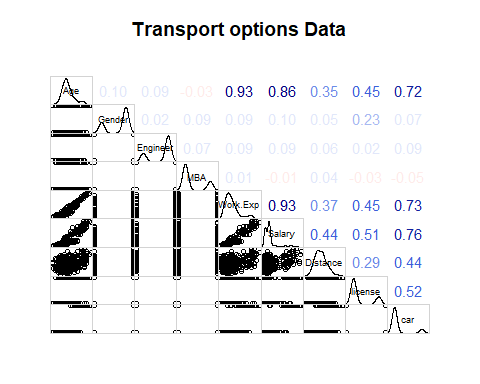
### Check for Multicollinearity - Plot the graph based on Multicollinearity.

Plot using corrgram, ggpairs and cor functions

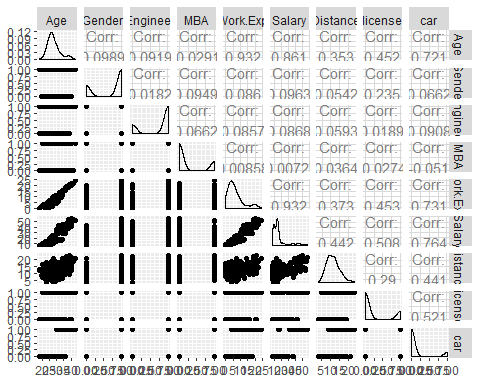
round(cor(traintest, use="complete.obs", method="kendall") ,2)

## Age Gender Engineer MBA Work.Exp Salary Distance license car  
## Age 1.00 0.06 0.05 -0.04 0.77 0.75 0.16 0.30 0.49  
## Gender 0.06 1.00 0.02 0.09 0.04 0.03 0.04 0.23 0.07  
## Engineer 0.05 0.02 1.00 0.07 0.05 0.06 0.06 0.02 0.09  
## MBA -0.04 0.09 0.07 1.00 0.01 -0.01 0.02 -0.03 -0.05  
## Work.Exp 0.77 0.04 0.05 0.01 1.00 0.82 0.15 0.27 0.47  
## Salary 0.75 0.03 0.06 -0.01 0.82 1.00 0.20 0.27 0.46  
## Distance 0.16 0.04 0.06 0.02 0.15 0.20 1.00 0.20 0.32  
## license 0.30 0.23 0.02 -0.03 0.27 0.27 0.20 1.00 0.52  
## car 0.49 0.07 0.09 -0.05 0.47 0.46 0.32 0.52 1.00

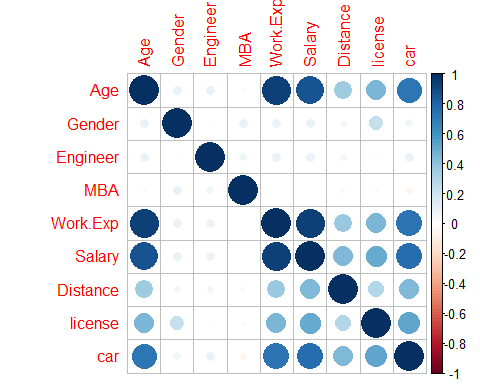
corrgram(traintest, main="Transport options Data",  
 lower.panel=panel.pts, upper.panel=panel.cor,  
 diag.panel=panel.density)



ggpairs(traintest)



corx <- cor(traintest)  
corrplot(corx)



**Observations**:

There is a high correlation observed between Independent variables. See below.

Correlation between Age and Work.Exp is 0.93 Correlation between Age and Salary is 0.86 Correlation between Age and license is 0.45 Correlation between Work.exp and Salary is 0.93 Correlation between Work.exp and Salary is 0.45 Correlation between salary and license is 0.51 Correlation between salary and distance is 0.44

## Data Preparation

#### Converting variables to factors

Converting Transport data to binary with respect to target variable (**Car**). Also converting gender variable to binary

carsdata$car = ifelse(carsdata$Transport=="Car",1,0)  
carsdata$Gender<-ifelse(carsdata$Gender=="Female",0,1)  
carsdata <- carsdata[,-9]  
  
# Converting all the factors  
  
carsdata$Engineer = as.factor(carsdata$Engineer)  
carsdata$Gender = as.factor(carsdata$Gender)  
carsdata$MBA = as.factor(carsdata$MBA)  
carsdata$license = as.factor(carsdata$license)  
carsdata$car = as.factor(carsdata$car)  
str(carsdata)

## 'data.frame': 444 obs. of 9 variables:  
## $ Age : int 28 23 29 28 27 26 28 26 22 27 ...  
## $ Gender : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 1 2 2 ...  
## $ Engineer: Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 2 2 ...  
## $ MBA : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...  
## $ 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 : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...  
## $ car : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

#### Splitting data to training and testing datasets

Splitting data into training-testing datasets in the ratio of 70-30%

sample = sample.split(carsdata, SplitRatio = .70)  
train = subset(carsdata, sample == TRUE)  
test = subset(carsdata, sample == FALSE)

Training set records and proportional distribution

nrow(train)

## [1] 297

prop.table(table(train$car))

##   
## 0 1   
## 0.8619529 0.1380471

Testing set records and proportional distribution

nrow(test)

## [1] 147

prop.table(table(test$car))

##   
## 0 1   
## 0.8639456 0.1360544

apply(carsdata, 2, function(x) length(unique(x)))

## Age Gender Engineer MBA Work.Exp Salary Distance license   
## 25 2 2 2 24 122 137 2   
## car   
## 2

**Observation:**

Class data imbalance is observed in the above datasets:

Car vs non-Car is 13.6% vs 86.39%

We may need to use Smote technique to balance data in the model building exercise.

## Building Models

Steps go here…

### Applying Logistic Regression

Assasd

Asd

Asd

asd

### Applying KNN Model

Asdasdasd

### Applying Naïve Bayes

Naive Bayes Model is based out of Probabilistic computation and favors classification technique. Strictly should be used for classification and not for regression because it borrowed the logic from Bayes theorem. NB uses Bayes theorem If independent variable is Categorical and Gaussian theorem if independent variable is Continuous model favors (works better) categorical independent variables over continuous.

Naive Bayes Model assumes the independent variables do not have any relationship between them. There is a Naive-ness in the computation hence the name Naive Bayes.

Categorical - Categorical: Chi Square test (Test of Independence)

Categorical - Continuous: Correlation test

Continuous - Categorical: Two Sample test /ANOVA (test of means)

Continuous - Continuous: Correlation test

Theory says that one should not run NB if there is a high correlation between Independent variables as this will affect the end result in positive or negative way. It is recommended that we apply FA/PCA crate factors and then possibly run NB model.

We will try below approaches for this project- Take data as it is and run NB model using K fold approach NB model based on Train/test datasets NB model using Cross validation approach Apply FA/PCA and then run NB model using K fold approach NB model using Cross validation approach

#### Data prep for NB

Create two consolidated datasets and update missing MBA record value to Zero (No value defaults to Zero)

traintest = rbind(train, test)  
traintestfa = traintest  
train$MBA[is.na(train$MBA)] = 0  
traintest$MBA[is.na(traintest$MBA)] = 0  
traintestfa$MBA[is.na(traintestfa$MBA)] = 0

#### NB model using K fold approach

Create folds – on traintest dataset and create 10 fields

carsdatafolds = traintest  
set.seed(123)  
carsfolds<-createFolds(carsdatafolds$car,k=10)

##### Try NB on first fold for testing

testdatanb = carsdatafolds[carsfolds$Fold01,]  
 traindatanb = carsdatafolds[-carsfolds$Fold01,]  
 nbmodel = naive\_bayes(traindatanb$car ~ ., traindatanb)  
 Predictnb = predict(nbmodel,testdatanb[-9],type="class")  
 confusionMatrix(Predictnb, testdatanb$car, positive = "1", mode="everything")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 35 1  
## 1 3 5  
##   
## Accuracy : 0.9091   
## 95% CI : (0.7833, 0.9747)  
## No Information Rate : 0.8636   
## P-Value [Acc > NIR] : 0.2654   
##   
## Kappa : 0.6615   
##   
## Mcnemar's Test P-Value : 0.6171   
##   
## Sensitivity : 0.8333   
## Specificity : 0.9211   
## Pos Pred Value : 0.6250   
## Neg Pred Value : 0.9722   
## Precision : 0.6250   
## Recall : 0.8333   
## F1 : 0.7143   
## Prevalence : 0.1364   
## Detection Rate : 0.1136   
## Detection Prevalence : 0.1818   
## Balanced Accuracy : 0.8772   
##   
## 'Positive' Class : 1

##### Now try NB in kfold way

nbkfold = lapply(carsfolds,function(x){  
 testdatanb = carsdatafolds[x,]  
 traindatanb = carsdatafolds[-x,]  
 nbmodel = naive\_bayes(traindatanb$car ~ ., traindatanb)  
 Predictnb = predict(nbmodel,testdatanb[-9],type="class")  
# confusionMatrix(Predictnb, testdatanb$car, positive = "1", mode="everything")  
 sumnb=table(testdatanb$car,Predictnb)  
 sum(diag(sumnb))/sum(sumnb)  
})  
nbaccuracy = mean(unlist(nbkfold))  
nbaccuracy

## [1] 0.9389899

**Observation** –

Accuracy comes to 0.9389899. Which is decent and seems to be working okay on this dataset. Also model captures the importance of the variables. Clearly Age, Salary, Work.Exp, license and Distance are driving the end result and we observed that all these variables have very high correlation among them

Try NB on Train and Test dataset

train$car = as.factor(train$car)  
test$car = as.factor(test$car)  
train$Gender = as.factor(train$Gender)  
test$Gender = as.factor(test$Gender)  
train$Engineer = as.factor(train$Engineer)  
test$Engineer = as.factor(test$Engineer)  
train$MBA = as.factor(train$MBA)  
test$MBA = as.factor(test$MBA)  
train$license = as.factor(train$license)  
test$license = as.factor(test$license)  
nbmodel2 = naive\_bayes(train$car ~ ., train)

Predictnb2 = predict(nbmodel2,test[-9],type="class")  
confusionMatrix(Predictnb2, test$car, positive = "1", mode="everything")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 124 1  
## 1 3 19  
##   
## Accuracy : 0.9728   
## 95% CI : (0.9318, 0.9925)  
## No Information Rate : 0.8639   
## P-Value [Acc > NIR] : 6.359e-06   
##   
## Kappa : 0.8889   
##   
## Mcnemar's Test P-Value : 0.6171   
##   
## Sensitivity : 0.9500   
## Specificity : 0.9764   
## Pos Pred Value : 0.8636   
## Neg Pred Value : 0.9920   
## Precision : 0.8636   
## Recall : 0.9500   
## F1 : 0.9048   
## Prevalence : 0.1361   
## Detection Rate : 0.1293   
## Detection Prevalence : 0.1497   
## Balanced Accuracy : 0.9632   
##   
## 'Positive' Class : 1   
##

**Observation**-

Accuracy comes to 0.97 which is again very good for this dataset. Also model captures the importance of the variables. Clearly Age, Salary, Work.Exp, license and Distance are driving the end result and we observed that all these variables have very high correlation among them.

##### Try NB using Cross Validation approach

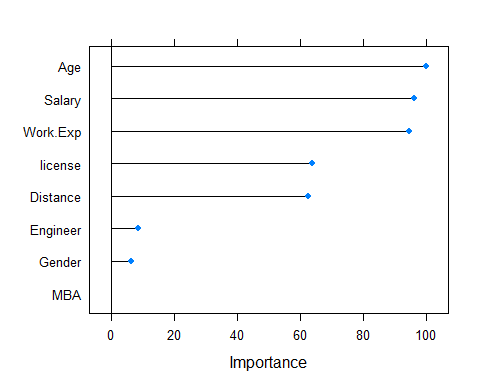
library(e1071)  
x = train[,-9]  
y = train$car  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))

Predict3 <- predict(model,newdata = test[-9] )

confusionMatrix(Predict3, test$car, positive = "1", mode="everything")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 125 0  
## 1 2 20  
##   
## Accuracy : 0.9864   
## 95% CI : (0.9517, 0.9983)  
## No Information Rate : 0.8639   
## P-Value [Acc > NIR] : 1.338e-07   
##   
## Kappa : 0.9445   
##   
## Mcnemar's Test P-Value : 0.4795   
##   
## Sensitivity : 1.0000   
## Specificity : 0.9843   
## Pos Pred Value : 0.9091   
## Neg Pred Value : 1.0000   
## Precision : 0.9091   
## Recall : 1.0000   
## F1 : 0.9524   
## Prevalence : 0.1361   
## Detection Rate : 0.1361   
## Detection Prevalence : 0.1497   
## Balanced Accuracy : 0.9921   
##   
## 'Positive' Class : 1   
##

X <- varImp(model)  
plot(X)



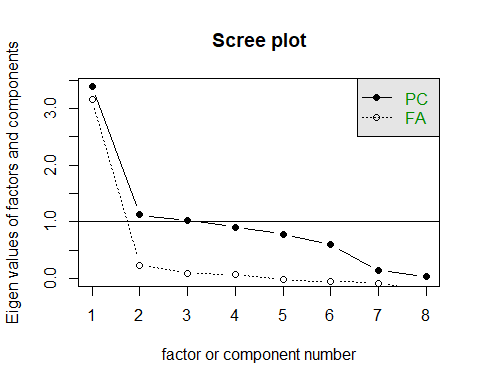
**Observations** –

Accuracy captured using Cross validation is 0.986 which is again very good for this dataset. Also model captures the importance of the variables. Clearly Age, Salary, Work.Exp, license and Distance are driving the end result and we observed that all these variables have very high correlation among them.

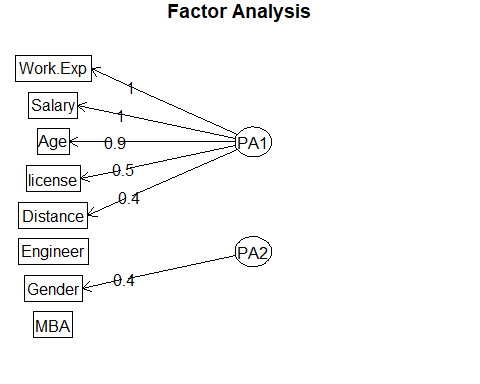
#### Try FA/PCA given there is very high relation between independent variables

Scree Plot

scree(traintestfa[-9])



fa1.out = fa(traintestfa[-9], nfactors=2,fm="pa",rotate = "none")  
fa.diagram(fa1.out)



**Observations**:

* **Bartlett** –

The null hypothesis is there is no correlation between the variables. p value is zero — rejecting the null hypothesis… default hypothesis in no correlation. So in this case reject points to there is a correlation

* **KMO** –

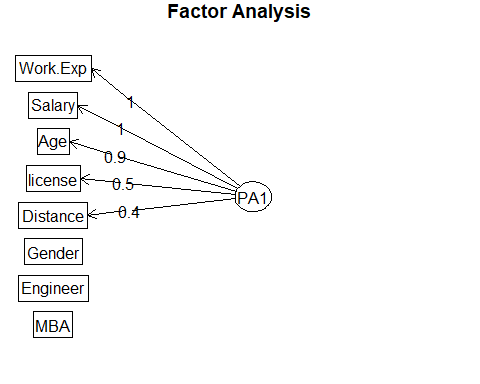
Test will give overall MSA value and MSA value for each variable. MSA value statistically is between 0 and 1 Benchmark is it should not be less than 0.7 (for practical purposes)

* **Scree plot** –

Captures one Factor using FA approach and 2 using PCA.

##### Explore the factors using FA and PCA way

fa1.out = fa(traintestfa[-9], nfactors=1,fm="pa",rotate = "none")  
fa.diagram(fa1.out)



fa1.out$communality

## Age Gender Engineer MBA Work.Exp Salary   
## 8.426981e-01 1.706600e-02 8.288204e-03 3.080955e-06 9.352321e-01 9.185186e-01   
## Distance license   
## 1.808375e-01 2.657799e-01

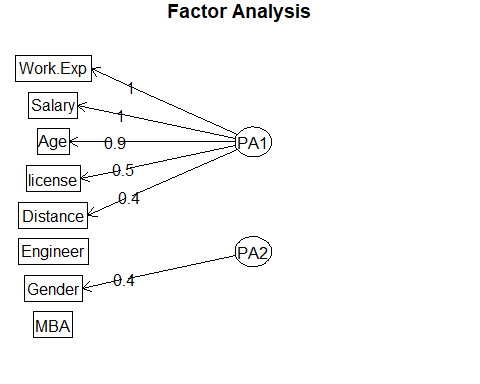
100\*fa1.out$e.values/length(fa1.out$e.values)

## [1] 42.4160602 14.0070604 12.8809509 11.2910332 9.7684917 7.4818895 1.6449008  
## [8] 0.5096133

print(fa1.out$loadings, digits=3)

##   
## Loadings:  
## PA1   
## Age 0.918  
## Gender 0.131  
## Engineer   
## MBA   
## Work.Exp 0.967  
## Salary 0.958  
## Distance 0.425  
## license 0.516  
##   
## PA1  
## SS loadings 3.168  
## Proportion Var 0.396

#fa1.out$scores  
fa2.out = fa(traintestfa[-9], nfactors=2,fm="pa",rotate = "none")  
fa.diagram(fa2.out)



fa2.out$communality

## Age Gender Engineer MBA Work.Exp Salary   
## 0.848422596 0.173338504 0.008205010 0.007823531 0.979985774 0.909750190   
## Distance license   
## 0.189523462 0.448669646

100\*fa2.out$e.values/length(fa2.out$e.values)

## [1] 42.4160602 14.0070604 12.8809509 11.2910332 9.7684917 7.4818895 1.6449008  
## [8] 0.5096133

print(fa2.out$loadings, digits=3)

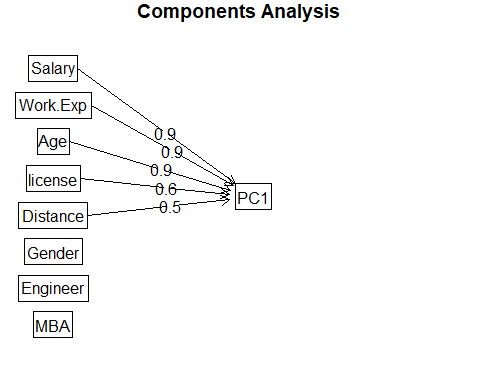
##   
## Loadings:  
## PA1 PA2   
## Age 0.914 -0.115  
## Gender 0.138 0.393  
## Engineer   
## MBA   
## Work.Exp 0.975 -0.169  
## Salary 0.953   
## Distance 0.425   
## license 0.544 0.391  
##   
## PA1 PA2  
## SS loadings 3.198 0.367  
## Proportion Var 0.400 0.046  
## Cumulative Var 0.400 0.446

**Observation** –

Two factor approach doesn’t seem to work as second factor only loaded one variable. We will go with One factor way. From the table of commonality, Work.Exp, Salary, Age are very important, while Gender, MBA, Distance, Engineer, license is the least important.

##### Now let’s Try PCA

pc1.out = principal(traintestfa[-9],nfactors = 1, rotate = 'varimax')  
fa.diagram(pc1.out)



**Observation** -

More or less both FA and PCA are giving similar results. Should look to consolidate Work.Exp, Salary, Age, License

Factor Identified is one

Factor 1 = Work.Exp, Salary, Age, license and Distance Factor 1 can be named as ‘Employee-Professional’

Now try running Naive Bayes using K fold approach

# convert into factor  
traintestpostfa=cbind(traintestfa[-c(1,5,6,7,8)], fa1.out$scores)  
traintestpostfa$Gender = as.factor(traintestpostfa$Gender)  
traintestpostfa$Engineer= as.factor(traintestpostfa$Engineer)  
traintestpostfa$MBA = as.factor(traintestpostfa$MBA)  
traintestpostfa$car = as.factor(traintestpostfa$car)

carsdatafolds = traintestpostfa  
set.seed(123)  
carsfolds<-createFolds(traintestpostfa$car,k=10)

nbkfold = lapply(carsfolds,function(x){  
 testdatanb = carsdatafolds[x,]  
 traindatanb = carsdatafolds[-x,]  
 nbmodel = naive\_bayes(traindatanb$car ~ ., traindatanb)  
 Predictnb = predict(nbmodel,testdatanb[-4],type="class")  
# confusionMatrix(Predictnb, testdatanb$car, positive = "1", mode="everything")  
 sumnb=table(testdatanb$car,Predictnb)  
 sum(diag(sumnb))/sum(sumnb)  
})

nbaccuracy = mean(unlist(nbkfold))  
nbaccuracy

## [1] 0.9412626

**Observation** –

Accuracy captured is 0.9412 Which is again seems to be okay.

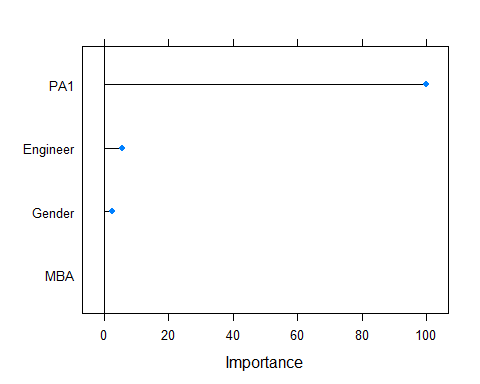
##### Now let’s Try NB using Cross Validation approach

# Divide data in training:testing 70:30  
set.seed(100)  
indices= sample(1:nrow(traintestpostfa), 0.7\*nrow(traintestpostfa))  
train\_data = traintestpostfa[indices,]  
test\_data = traintestpostfa[-indices,]

x = traintestpostfa[,-4]  
y = traintestpostfa$car  
model = train(x,y,'nb',trControl=trainControl(method='cv',number=10))  
  
confusionMatrix(Predict3, test\_data$car, positive = "1", mode="everything")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 115 4  
## 1 2 13  
##   
## Accuracy : 0.9552   
## 95% CI : (0.9051, 0.9834)  
## No Information Rate : 0.8731   
## P-Value [Acc > NIR] : 0.001226   
##   
## Kappa : 0.7872   
##   
## Mcnemar's Test P-Value : 0.683091   
##   
## Sensitivity : 0.76471   
## Specificity : 0.98291   
## Pos Pred Value : 0.86667   
## Neg Pred Value : 0.96639   
## Precision : 0.86667   
## Recall : 0.76471   
## F1 : 0.81250   
## Prevalence : 0.12687   
## Detection Rate : 0.09701   
## Detection Prevalence : 0.11194   
## Balanced Accuracy : 0.87381   
##   
## 'Positive' Class : 1   
##

X <- varImp(model)  
plot(X)



**Observation**

Captured as Accuracy: 0.9552 which is again good.

Comparison with different approaches

Take data as it is and run NB model using K fold approach : Accuracy - 0.93 NB model Based on Train/test daatsets : Accuracy - 0.97 NB model using Cross validation approach : Accuracy - 0.986 Apply FA/PCA and then run NB model using K fold approach : Accuracy - 0.9412 NB model using Cross validation approach : Accuracy -0.9552

### Confusion matrix summary for all models

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression |  |
| k-NN |  |
| Naïve Bayes |  |

### Remarks on Model validation exercise <Which model performed the best>

## Ensemble Techniques

Steps go here

### Bagging

Steps go here

### Boosting

Steps go here

## Actionable Insights and Recommendations

Steps go here -

# Appendix

* 1. Source data file:

