Project -5 Predicting mode of Transport (ML-1)

Submitted by,

Ajuna John

Part 1 – EDA

Description

This project helps to understand what mode of transport employees prefer to commute to their office. The dataset used in the project includes employee information about their mode of transport as well as their personal and professional details like age, salary, work exp. Here, we predict whether or not an employee will use Car as a mode of transport. Also, which variables are a significant predictor behind this decision.

Importing libraries

```
library(dplyr)

library(tidyr)

library(purrr)

library(ggplot2)

library(readr)

library(corrplot)

library (caret)

## Loading required package: lattice## Loading required package: ggplot2

library (car)

## Loading required package: carData

library (DMwR)

library(readr)

library(DMwR)

library(rattle)
```

DataSet

The dataset has 444 observations and 9 variables

Variable Names

[1] "Age" "Gender" "Engineer" "MBA" "Work.Exp" "Salary"

[7] "Distance" "license" "Transport"

Structure of dataset

'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 ...

Analysis of Dataset

Check and treatment for missing values

Here we have null/ missing values. The missing values in MBA variable is replaced with the median value

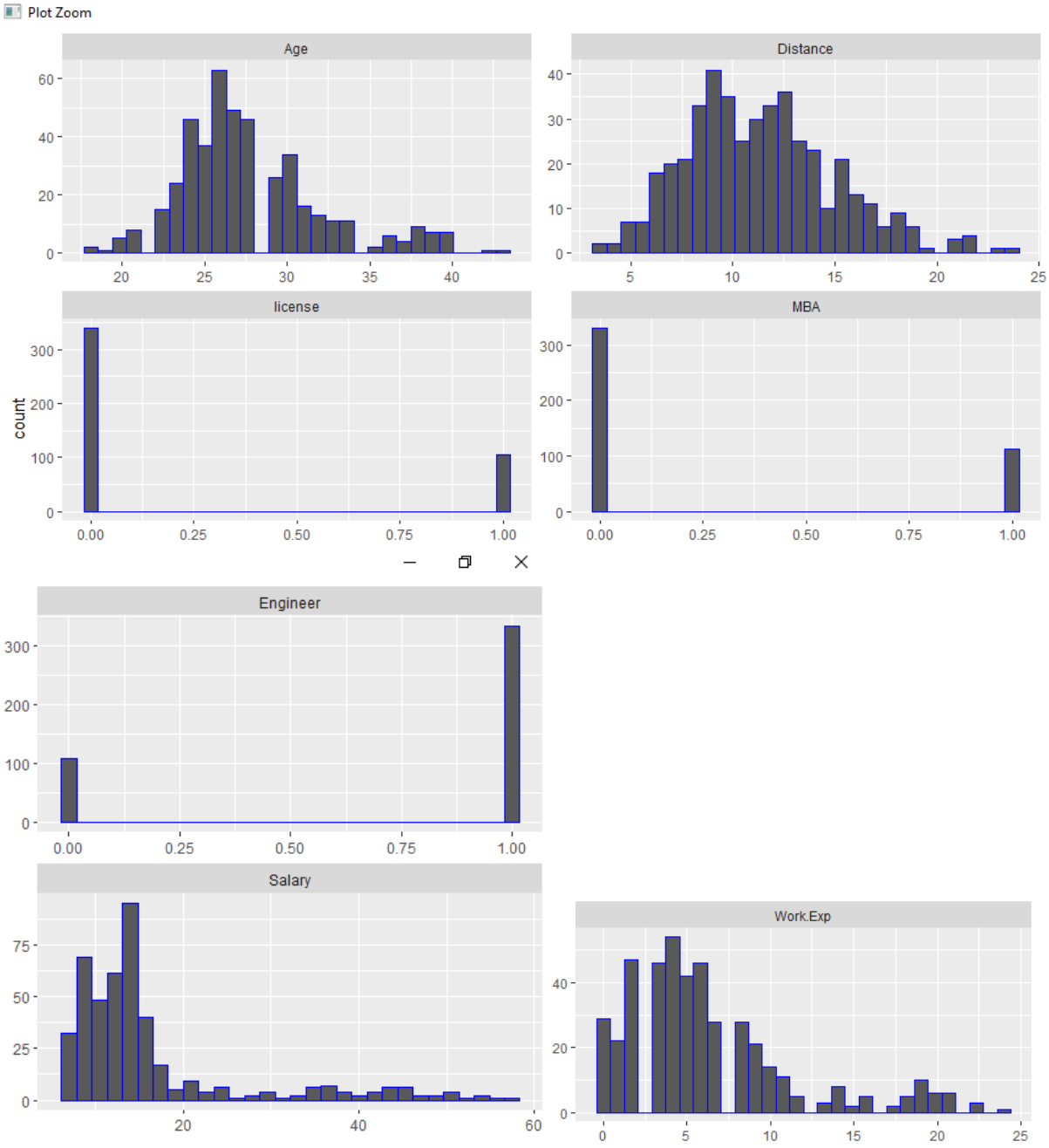
Summary of the dataset

```
> summary(data)
      Age           Gender       Engineer           MBA            Work.Exp
 Min.   :18.00   Female:128   Min.   :0.0000   Min.   :0.0000   Min.   : 0.0
 1st Qu.:25.00   Male  :316   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.: 3.0
 Median :27.00                Median :1.0000   Median :0.0000   Median : 5.0
 Mean   :27.75                Mean   :0.7545   Mean   :0.2528   Mean   : 6.3
 3rd Qu.:30.00                3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 8.0
 Max.   :43.00                Max.   :1.0000   Max.   :1.0000   Max.   :24.0
                                               NA's   :1
     Salary         Distance        license                  Transport
 Min.   : 6.50   Min.   : 3.20   Min.   :0.0000   2Wheeler        : 83
 1st Qu.: 9.80   1st Qu.: 8.80   1st Qu.:0.0000   Car             : 61
 Median :13.60   Median :11.00   Median :0.0000   Public Transport:300
 Mean   :16.24   Mean   :11.32   Mean   :0.2342
 3rd Qu.:15.72   3rd Qu.:13.43   3rd Qu.:0.0000
 Max.   :57.00   Max.   :23.40   Max.   :1.0000
```

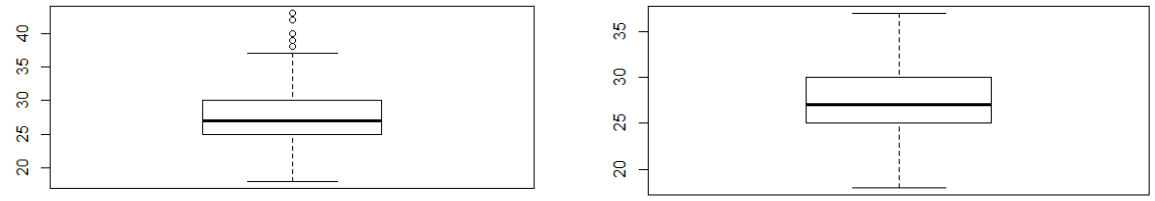
Basic Conclusions:

- The average & median age group is approx. 27
- Nearly 75% of candidates are males
- One data point of MBA is NA
- Average work experience in 6.3 years
- Average Salary is Rs. 16 Lakhs
- Public transport is the most common means of transport

Histogram of the variables

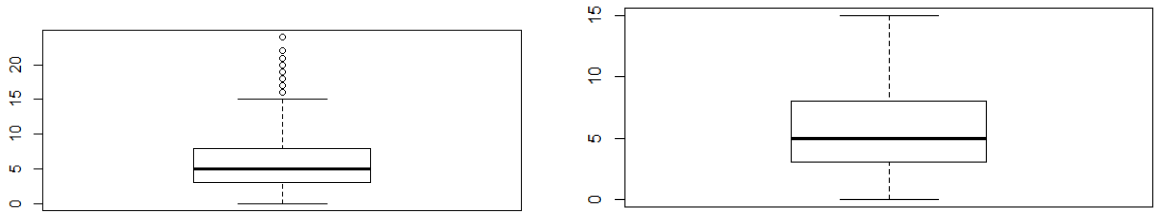

Plot Zoom
Age
Distance
license
MBA
count
Engineer
Salary
Work.Exp

Boxplot of Age



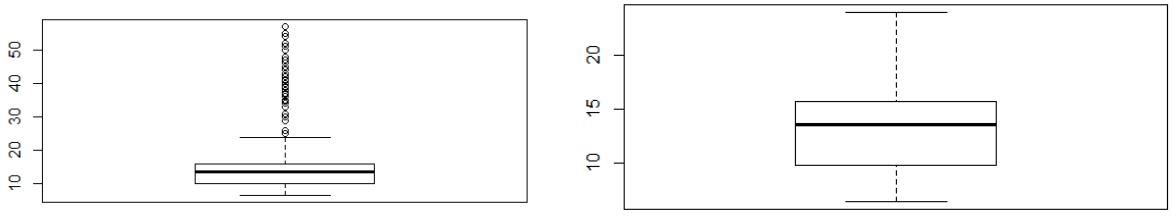
Outliers were found in Age variable. Hence treated the dataset to remove outliers

Boxplot of Work Experience



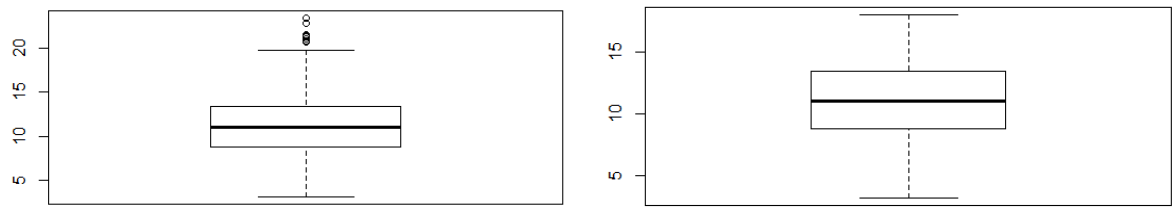
Outliers were found in Work Experience variable. Hence treated the dataset to remove outliers

Boxplot of Salary



Outliers were found in Salary variable. Hence treated the dataset to remove outliers

Boxplot of Distance



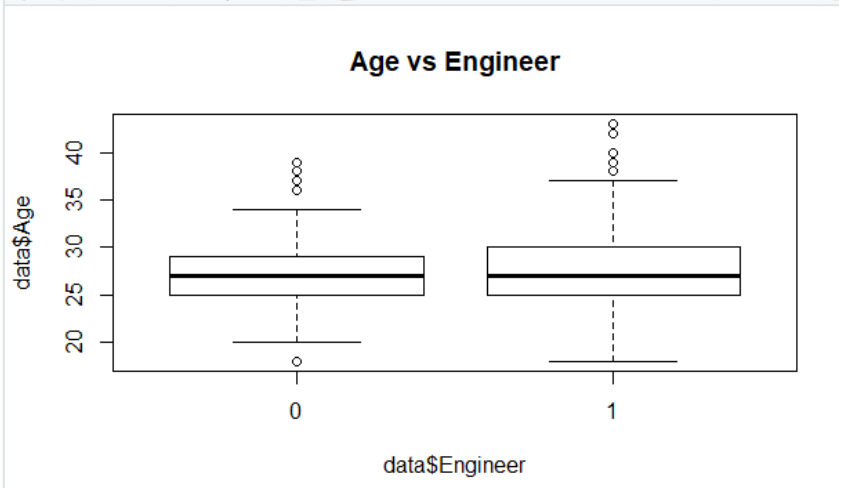
Outliers were found in Distance variable. Hence treated the dataset to remove outliers

Correlation among all variables

Work Experience is highly correlated to Age and Salary. Salary is highly correlated to Age

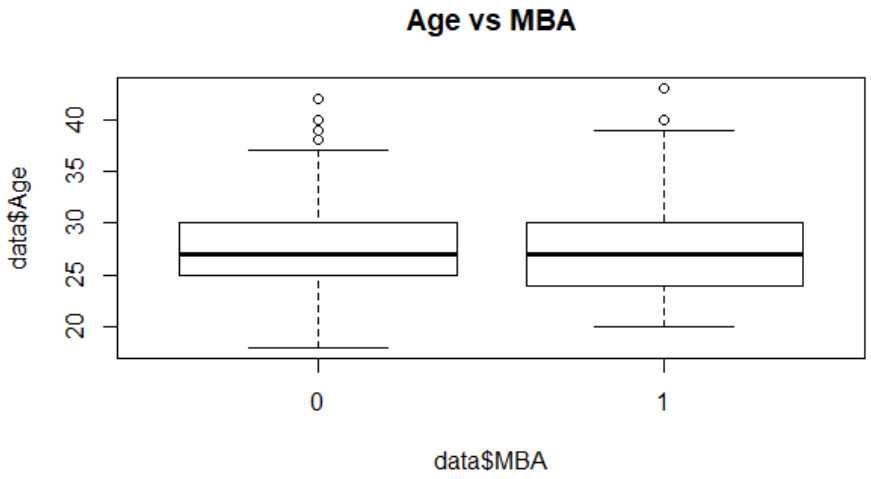
Bivariate Analysis

Boxplot of "Age vs Engineer"



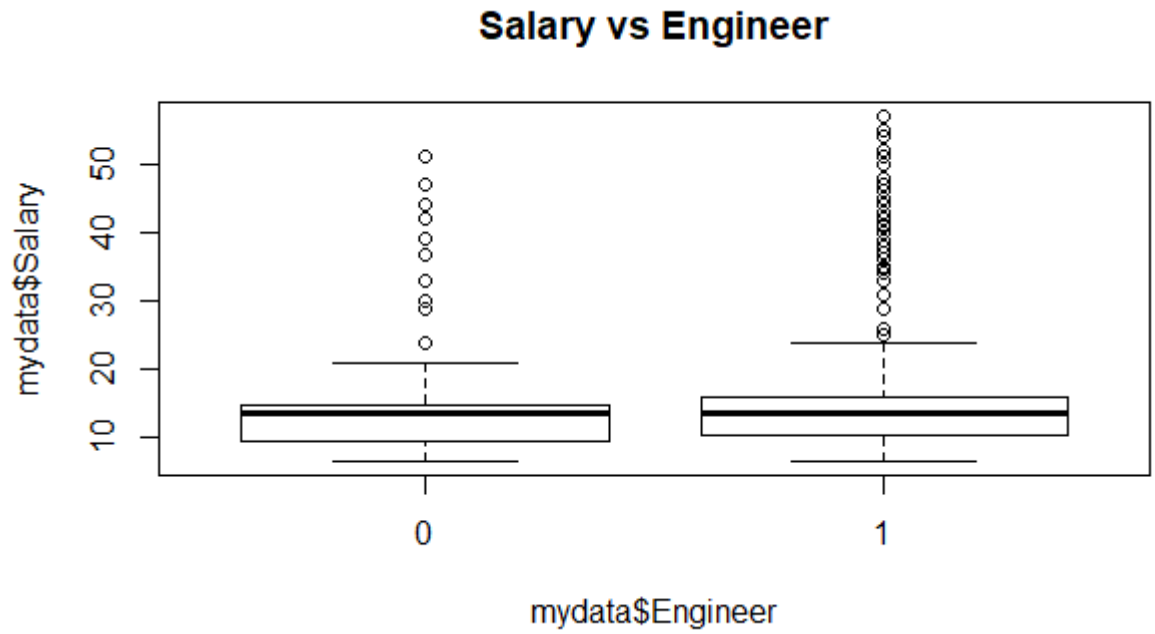
In general, the age group of Engineers vs Non-Engineers are same

Boxplot "Age vs MBA"

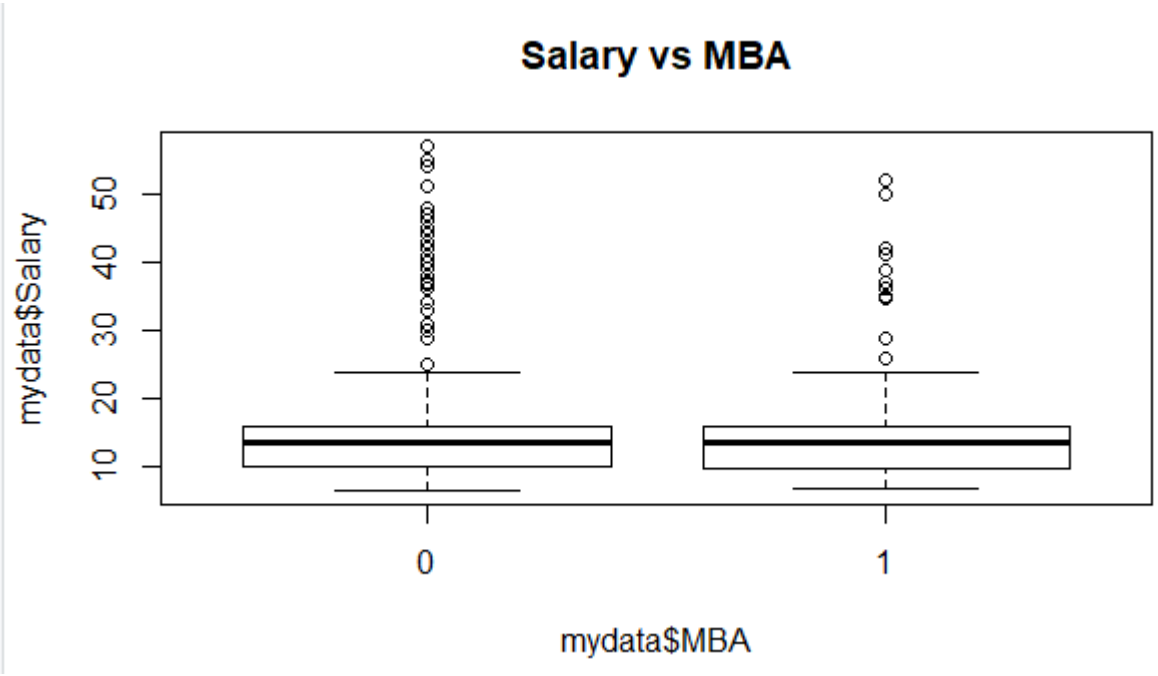


In general, the age group of MBA's vs Non-MBA's are same

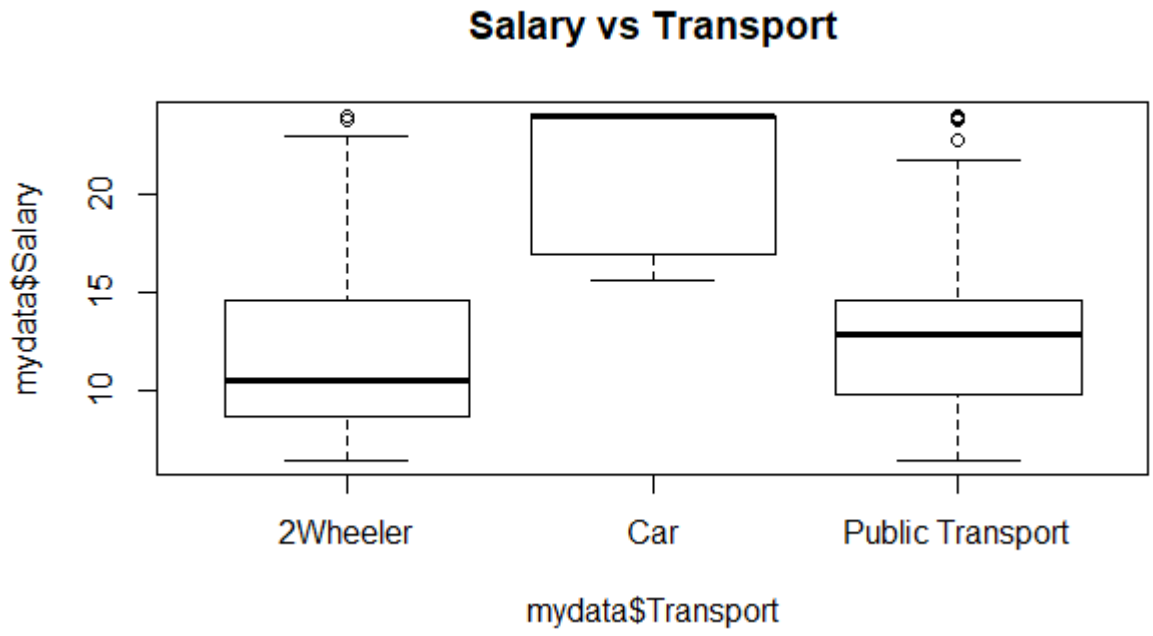
Boxplot of "Salary vs Engineer"



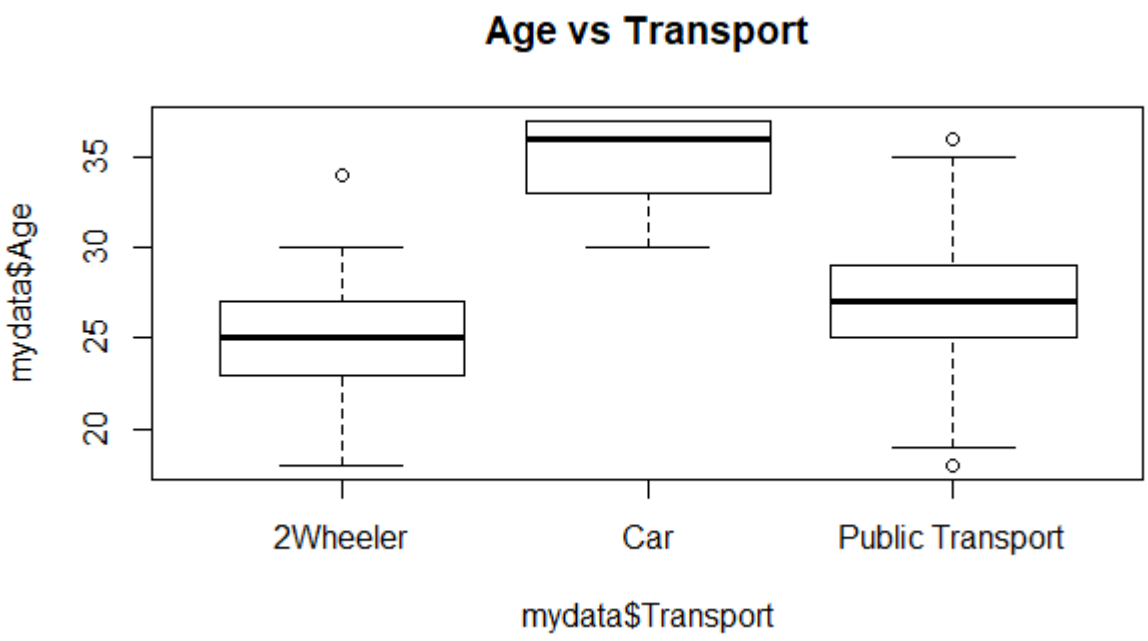
Boxplot of "Salary vs MBA"



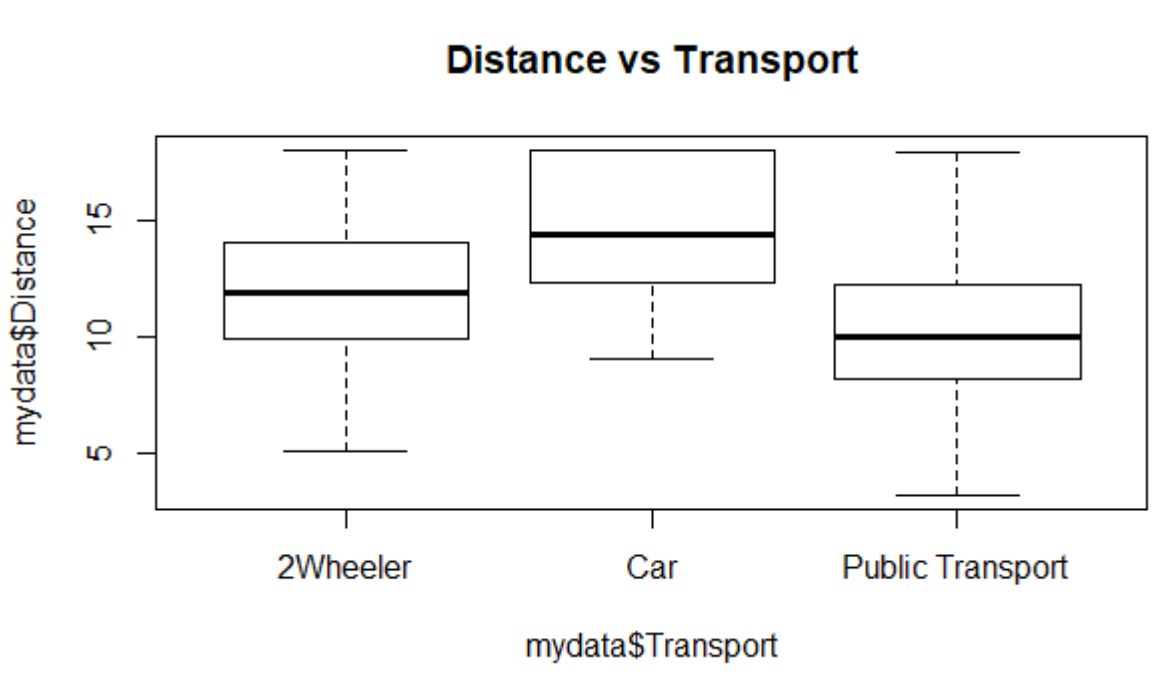
Boxplot of "Salary vs Transport"



Boxplot of "Age vs Transport"



Boxplot of "Distance vs Transport"



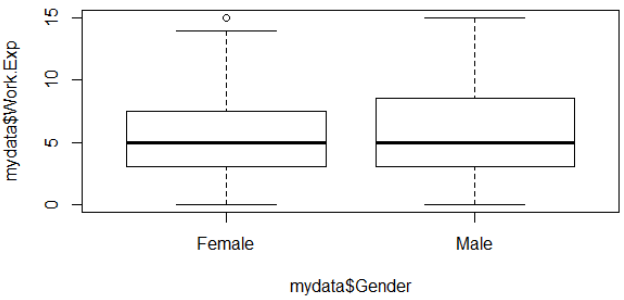
```
cor(mydata$Age,mydata$Work.Exp)
[1] 0.9165547
```

```
table(mydata$Gender, mydata$Transport)
```

	2Wheeler	Car	Public Transport
Female	38	13	77
Male	45	48	223


```
table (mydata$license,mydata$Transport)
```

	2Wheeler	Car	Public Transport
0	60	13	267
1	23	48	33



Hypothesis Testing

#Preparation of the data

Converted Engineer, MBA and Liscence to factors

carsbasedata<-knnImputation(carsbasedata)

carsbasedata$CarUsage<-ifelse(carsbasedata$Transport =='Car',1,0)

table(carsbasedata$CarUsage)

0 1

383 61

#Model Building and Data Split

Train data split is :

prop.table(table(carsdatatrain$CarUsage))

0 1

0.8621795 0.1378205

Test data split is :

prop.table(table(carsdatatest$CarUsage))

0 1

0.8636364 0.1363636

The train and test data have almost same percentage of cars usage as the base data

Apply SMOTE on Training data set

summary(carsglm$finalModel)

Call:

NULL

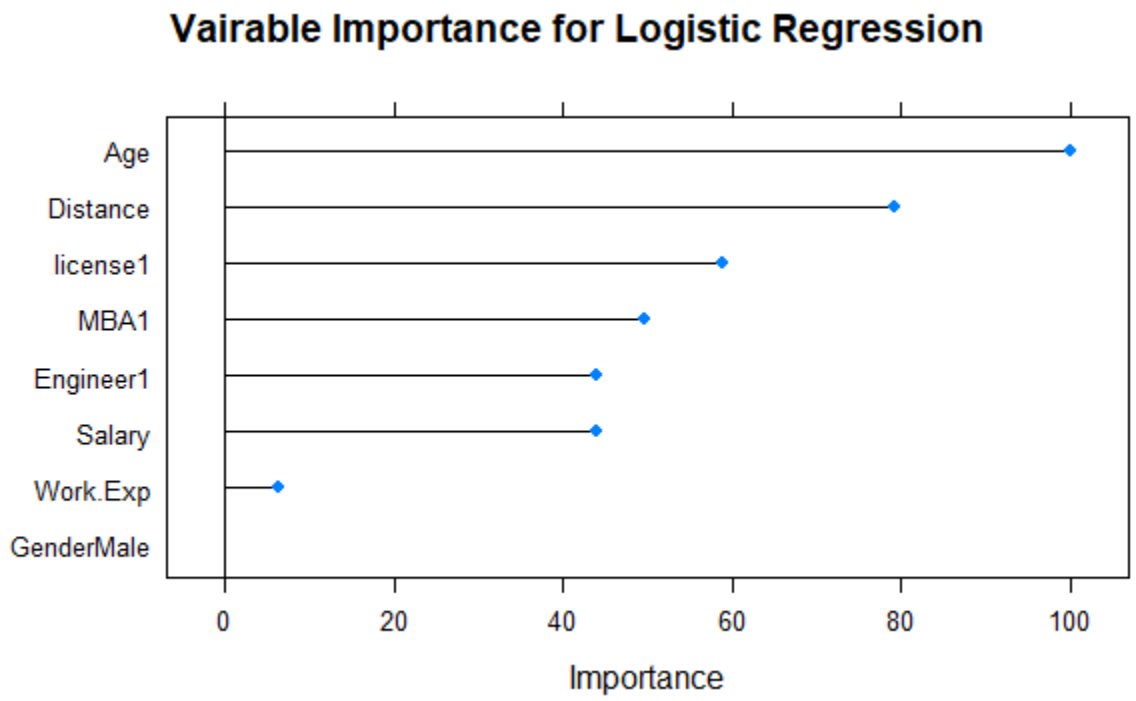
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.16989  -0.03015   0.00180   0.05402   2.35396

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -44.8238    11.2391  -3.988 6.66e-05 ***
Age           1.3016     0.3757   3.464 0.000532 ***
Work.Exp      0.3783     0.3922   0.964 0.334845
Salary       -0.4354     0.2208  -1.972 0.048664 *
Distance      0.4783     0.1645   2.908 0.003643 **
license1      2.3928     1.0099   2.369 0.017815 *
Engineer1     1.6358     0.8296   1.972 0.048619 *
MBA1         -1.7187     0.8104  -2.121 0.033926 *
GenderMale    0.6936     0.8696   0.798 0.425118
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 357.664  on 257  degrees of freedom
Residual deviance:  56.937  on 249  degrees of freedom
AIC: 74.937

Number of Fisher Scoring iterations: 9

glm variable importance

            Overall
Age         100.000
Distance     79.134
license1     58.951
MBA1         49.632
```

Engineer1 44.043

Salary 44.028

Work.Exp 6.257

GenderMale 0.000



Model Interpretation

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 109 3

1 5 15

Accuracy : 0.9394

95% CI : (0.8841, 0.9735)

No Information Rate : 0.8636

P-Value [Acc > NIR] : 0.004496

Kappa : 0.7542

Mcnemar's Test P-Value : 0.723674

Sensitivity : 0.8333
Specificity : 0.9561
Pos Pred Value : 0.7500
Neg Pred Value : 0.9732
Prevalence : 0.1364
Detection Rate : 0.1136
Detection Prevalence : 0.1515
Balanced Accuracy : 0.8947

'Positive' Class : 1

Improving the model

glmnet

258 samples
5 predictor
2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 233, 232, 232, 233, 232, 232, ...
Resampling results across tuning parameters:

alpha lambda Accuracy Kappa

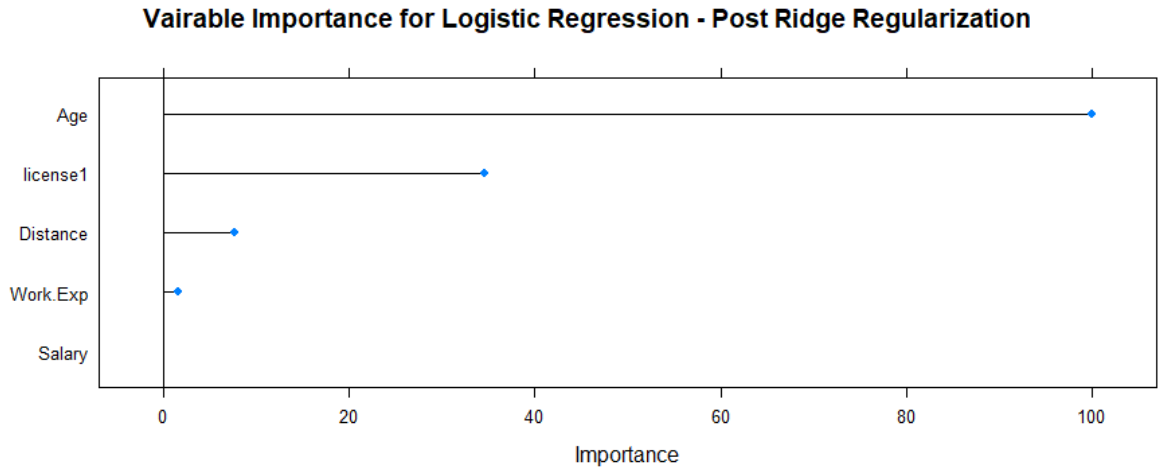
0.10	0.0008423163	0.9495385	0.8990641
0.10	0.0084231631	0.9418462	0.8836794
0.10	0.0842316313	0.9264615	0.8529102
0.55	0.0008423163	0.9456923	0.8913718
0.55	0.0084231631	0.9458462	0.8916665
0.55	0.0842316313	0.9458462	0.8916665
1.00	0.0008423163	0.9418462	0.8836794
1.00	0.0084231631	0.9496923	0.8993588
1.00	0.0842316313	0.9575385	0.9150897

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were alpha = 1 and lambda = 0.08423163.

glmnet variable importance

	Overall
Age	100.000
license1	34.544
Distance	7.693
Work.Exp	1.650
Salary	0.000



Prediction using the regularized model

Confusion Matrix and Statistics

```
          Reference
Prediction   0   1
         0 110   2
         1   4  16

               Accuracy : 0.9545
                 95% CI : (0.9037, 0.9831)
    No Information Rate : 0.8636
    P-Value [Acc > NIR] : 0.0005559

                  Kappa : 0.8156

 Mcnemar's Test P-Value : 0.6830914

            Sensitivity : 0.8889
            Specificity : 0.9649
         Pos Pred Value : 0.8000
         Neg Pred Value : 0.9821
             Prevalence : 0.1364
         Detection Rate : 0.1212
   Detection Prevalence : 0.1515
      Balanced Accuracy : 0.9269

       'Positive' Class : 1
```

Inference & Prediction Using Linear Discriminant Analysis

```
prop.table(table(cartrainlda.car$Transport))

        Car Public Transport
  0.1699605       0.8300395

prop.table(table(cartrainlda.twlr$Transport))

     2Wheeler Public Transport
   0.2193309       0.7806691

table(carldatwlrsm$Transport)

     2Wheeler Public Transport
          118             118

table(carldacarsm$Transport)

        Car Public Transport
         86              86

table(carsdatatrainldasm$Transport)

     2Wheeler Public Transport             Car
          118             118              86

carslda$finalModel
Call:
lda(x, grouping = y)

Prior probabilities of groups:
     2Wheeler Public Transport             Car
```

0.3664596 0.3664596 0.2670807

Group means:

	Age	Work.Exp	Salary	Distance	license1	GenderMale	Engineer1
2Wheeler	25.25901	4.134194	12.60951	12.14888	0.2711864	0.5338983	0.7542373
Public Transport	27.26271	5.449153	13.55932	10.54576	0.1355932	0.7118644	0.7627119
Car	35.22263	15.071805	33.26627	14.95268	0.6744186	0.7441860	0.8488372

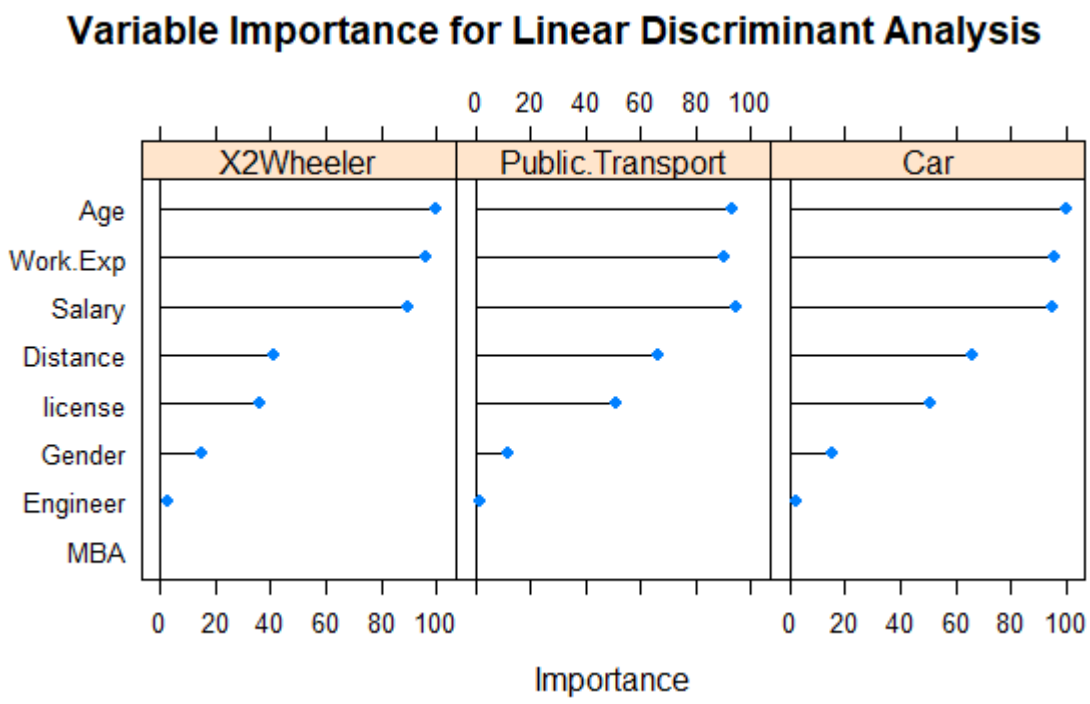
	MBA1
2Wheeler	0.2796610
Public Transport	0.2796610
Car	0.2093023

Coefficients of linear discriminants:

	LD1	LD2
Age	0.27748223	-0.38658144
Work.Exp	0.01447360	0.17068708
Salary	0.01886957	0.03499292
Distance	0.05788050	0.17945541
license1	0.08065257	1.24984845
GenderMale	-0.13649878	-0.82615178
Engineer1	0.09880066	0.09530187
MBA1	-0.45658224	-0.11376118

Proportion of trace:

LD1	LD2
0.902	0.098



Confusion Matrix and Statistics

	Reference		
Prediction	2Wheeler	Car	Public Transport
2Wheeler	17	1	28
Car	1	15	2
Public Transport	6	2	60

Overall Statistics

Accuracy : 0.697

95% CI : (0.611, 0.7739)

No Information Rate : 0.6818

P-Value [Acc > NIR] : 0.393723

Kappa : 0.4654

Mcnemar's Test P-Value : 0.002602

Statistics by Class:

	Class: 2Wheeler	Class: Car	Class: Public Transport
Sensitivity	0.7083	0.8333	0.6667
Specificity	0.7315	0.9737	0.8095
Pos Pred Value	0.3696	0.8333	0.8824
Neg Pred Value	0.9186	0.9737	0.5312
Prevalence	0.1818	0.1364	0.6818
Detection Rate	0.1288	0.1136	0.4545
Detection Prevalence	0.3485	0.1364	0.5152
Balanced Accuracy	0.7199	0.9035	0.7381

Improve LDA Model by Regularization

Call:

```
mda::fda(formula = as.formula(".outcome ~ ."), data = dat, method = mda::gen.ridge,
    lambda = param$lambda)
```

Dimension: 2

Percent Between-Group Variance Explained:

```
   v1    v2
 90.2 100.0
```

Degrees of Freedom (per dimension): 7.992552

Training Misclassification Error: 0.24224 (N = 322)

Penalized Discriminant Analysis

322 samples

8 predictor

3 classes: '2Wheeler', 'Public Transport', 'Car'

No pre-processing

Resampling: Cross-Validated (10 fold)

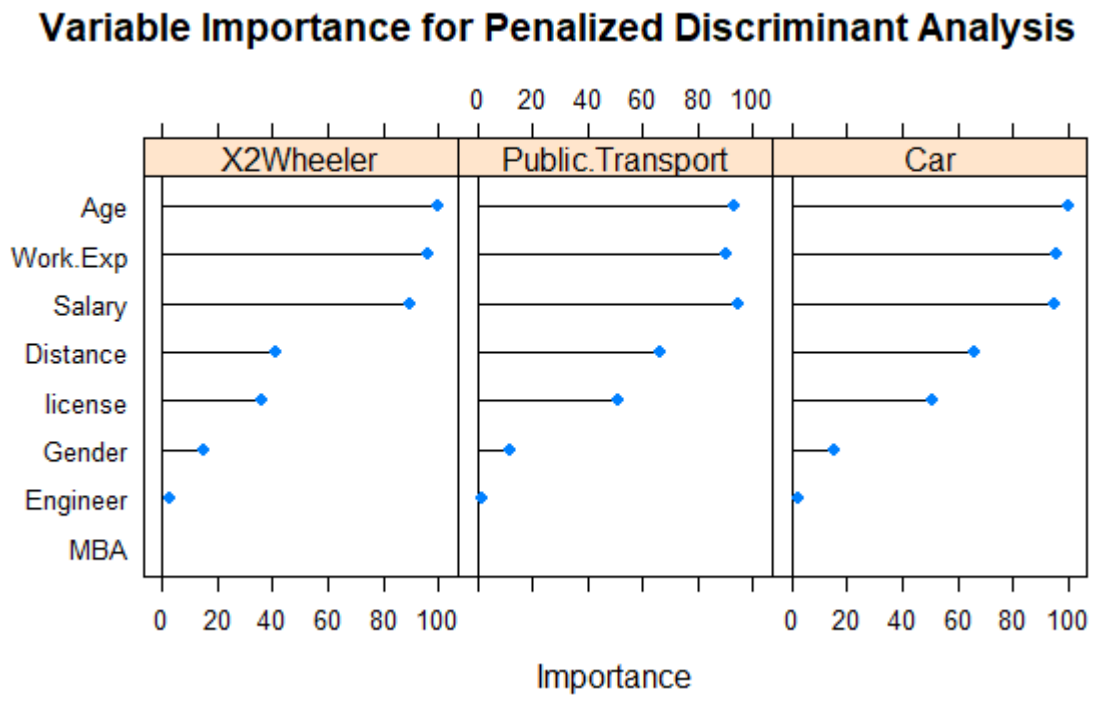
Summary of sample sizes: 289, 290, 290, 290, 290, 289, ...

Resampling results across tuning parameters:

lambda	Accuracy	Kappa
0e+00	0.7266618	0.5831623
1e-04	0.7266618	0.5831623
1e-01	0.7297868	0.5879527

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was lambda = 0.1.



Confusion Matrix and Statistics

Reference

Prediction	2Wheeler	Car	Public Transport
2Wheeler	17	1	28
Car	1	15	2
Public Transport	6	2	60

Overall Statistics

Accuracy : 0.697

95% CI : (0.611, 0.7739)

No Information Rate : 0.6818

P-Value [Acc > NIR] : 0.393723

Kappa : 0.4654

Mcnemar's Test P-Value : 0.002602

Statistics by Class:

	Class: 2Wheeler	Class: Car	Class: Public Transport
Sensitivity	0.7083	0.8333	0.6667
Specificity	0.7315	0.9737	0.8095
Pos Pred Value	0.3696	0.8333	0.8824
Neg Pred Value	0.9186	0.9737	0.5312
Prevalence	0.1818	0.1364	0.6818
Detection Rate	0.1288	0.1136	0.4545
Detection Prevalence	0.3485	0.1364	0.5152
Balanced Accuracy	0.7199	0.9035	0.7381

#Prediction using CART

carscart$finalModel

n= 322

node), split, n, loss, yval, (yprob)

* denotes terminal node

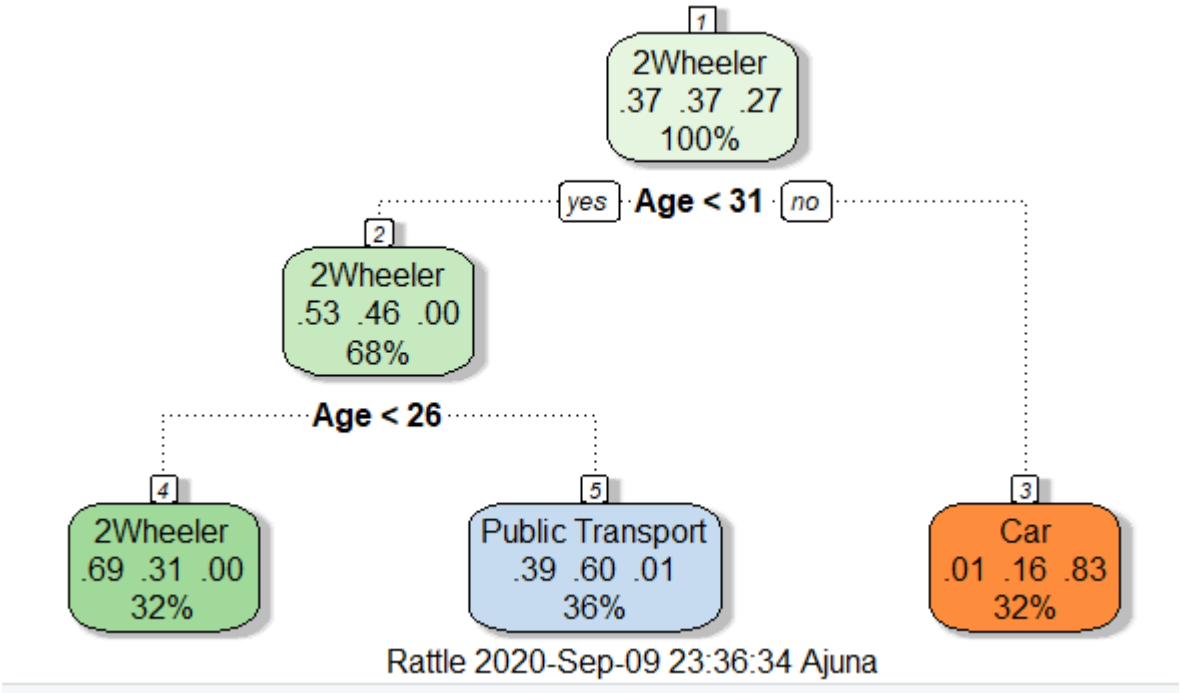
1) root 322 204 2Wheeler (0.366459627 0.366459627 0.267080745)

2) Age< 30.5005 220 103 2Wheeler (0.531818182 0.463636364 0.004545455)

4) Age< 25.95421 103 32 2Wheeler (0.689320388 0.310679612 0.000000000) *

5) Age>=25.95421 117 47 Public Transport (0.393162393 0.598290598 0.008547009) *

3) Age>=30.5005 102 17 Car (0.009803922 0.156862745 0.833333333) *



Confusion Matrix and Statistics

Reference

Prediction 2Wheeler Car Public Transport

2Wheeler	14	0	36
Car	1	17	4
Public Transport	9	1	50

Overall Statistics

Accuracy : 0.6136

95% CI : (0.525, 0.6971)

No Information Rate : 0.6818

P-Value [Acc > NIR] : 0.9603274

Kappa : 0.3544

Mcnemar's Test P-Value : 0.0002734

Statistics by Class:

	Class: 2Wheeler	Class: Car	Class: Public Transport
Sensitivity	0.5833	0.9444	0.5556
Specificity	0.6667	0.9561	0.7619
Pos Pred Value	0.2800	0.7727	0.8333
Neg Pred Value	0.8780	0.9909	0.4444
Prevalence	0.1818	0.1364	0.6818
Detection Rate	0.1061	0.1288	0.3788
Detection Prevalence	0.3788	0.1667	0.4545
Balanced Accuracy	0.6250	0.9503	0.6587

#Prediction using Boosting

xgb.Booster

raw: 40.9 Kb

call:

```
  xgboost::xgb.train(params = list(eta = param$eta, max_depth = param$max_depth,
    gamma = param$gamma, colsample_bytree = param$colsample_bytree,
    min_child_weight = param$min_child_weight, subsample = param$subsample),
    data = x, nrounds = param$nrounds, num_class = length(lev),
    objective = "multi:softprob")
```

params (as set within xgb.train):

```
  eta = "0.3", max_depth = "1", gamma = "0", colsample_bytree = "0.6", min_child_weight = "1",
subsample = "1", num_class = "3", objective = "multi:softprob", validate_parameters = "TRUE"
```

xgb.attributes:

```
  niter
```

callbacks:

```
  cb.print.evaluation(period = print_every_n)
```

# of features: 8

niter: 50

nfeatures : 8

xNames : Age GenderMale Engineer1 MBA1 Work.Exp Salary Distance license1

problemType : Classification

tuneValue :

	nrounds	max_depth	eta	gamma	colsample_bytree	min_child_weight	subsample
1	50	1	0.3	0	0.6	1	1

obsLevels : 2Wheeler Public Transport Car

param :

```
          list()
```

Predict using Test Dataset

Confusion Matrix and Statistics

Prediction	Reference 2Wheeler	Car	Public Transport
2Wheeler	18	1	30
Car	1	17	2
Public Transport	5	0	58

Overall Statistics

Accuracy : 0.7045

95% CI : (0.6189, 0.7807)

No Information Rate : 0.6818

P-Value [Acc > NIR] : 0.3234721

Kappa : 0.4962

Mcnemar's Test P-Value : 0.0001817

Statistics by Class:

	Class: 2Wheeler	Class: Car	Class: Public Transport
Sensitivity	0.7500	0.9444	0.6444
Specificity	0.7130	0.9737	0.8810
Pos Pred Value	0.3673	0.8500	0.9206
Neg Pred Value	0.9277	0.9911	0.5362
Prevalence	0.1818	0.1364	0.6818
Detection Rate	0.1364	0.1288	0.4394
Detection Prevalence	0.3712	0.1515	0.4773
Balanced Accuracy	0.7315	0.9591	0.7627

Prediction Using Multinomial Logistic Regression

Call:

nnet::multinom(formula = .outcome ~ ., data = dat, decay = param$decay)

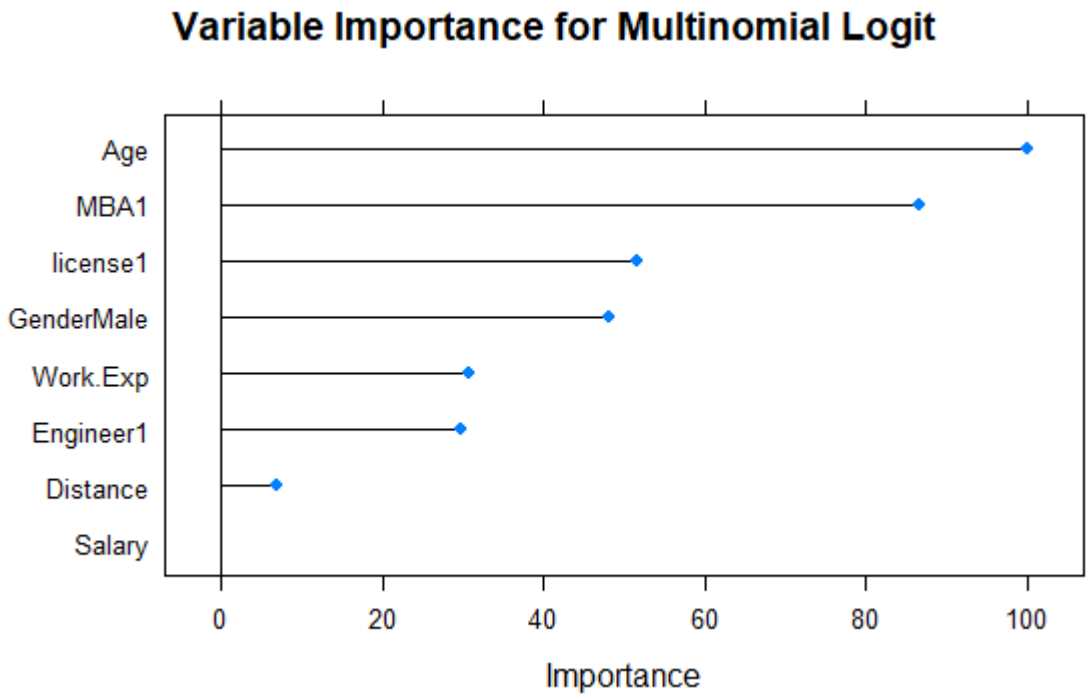
Coefficients:

	(Intercept)	Age	GenderMale	Engineer1	MBA1	Work.Exp
Public Transport	-9.455486	0.4838898	0.9389725	-0.08550113	-0.06189137	-0.07722469
Car	-73.345822	2.5199401	-0.6174967	0.95642590	-2.56344517	-0.99172522

	Salary	Distance	license1
Public Transport	-0.09642089	-0.1601295	-1.4161838
Car	0.11339876	0.2443606	-0.2361116

Residual Deviance: 320.6077

AIC: 356.6077



Predict using the test data

Confusion Matrix and Statistics

Reference

Prediction	2Wheeler	Car	Public Transport
2Wheeler	19	1	23
Car	1	17	2
Public Transport	4	0	65

Overall Statistics

Accuracy : 0.7652

95% CI : (0.6835, 0.8345)

No Information Rate : 0.6818

P-Value [Acc > NIR] : 0.022700

Kappa : 0.5834

Mcnemar's Test P-Value : 0.001526

Statistics by Class:

	Class: 2Wheeler	Class: Car	Class: Public Transport
Sensitivity	0.7917	0.9444	0.7222
Specificity	0.7778	0.9737	0.9048
Pos Pred Value	0.4419	0.8500	0.9420
Neg Pred Value	0.9438	0.9911	0.6032
Prevalence	0.1818	0.1364	0.6818
Detection Rate	0.1439	0.1288	0.4924
Detection Prevalence	0.3258	0.1515	0.5227
Balanced Accuracy	0.7847	0.9591	0.8135

Prediction using Random Forest

Call:

randomForest(x = x, y = y, mtry = param$mtry)

Type of random forest: classification

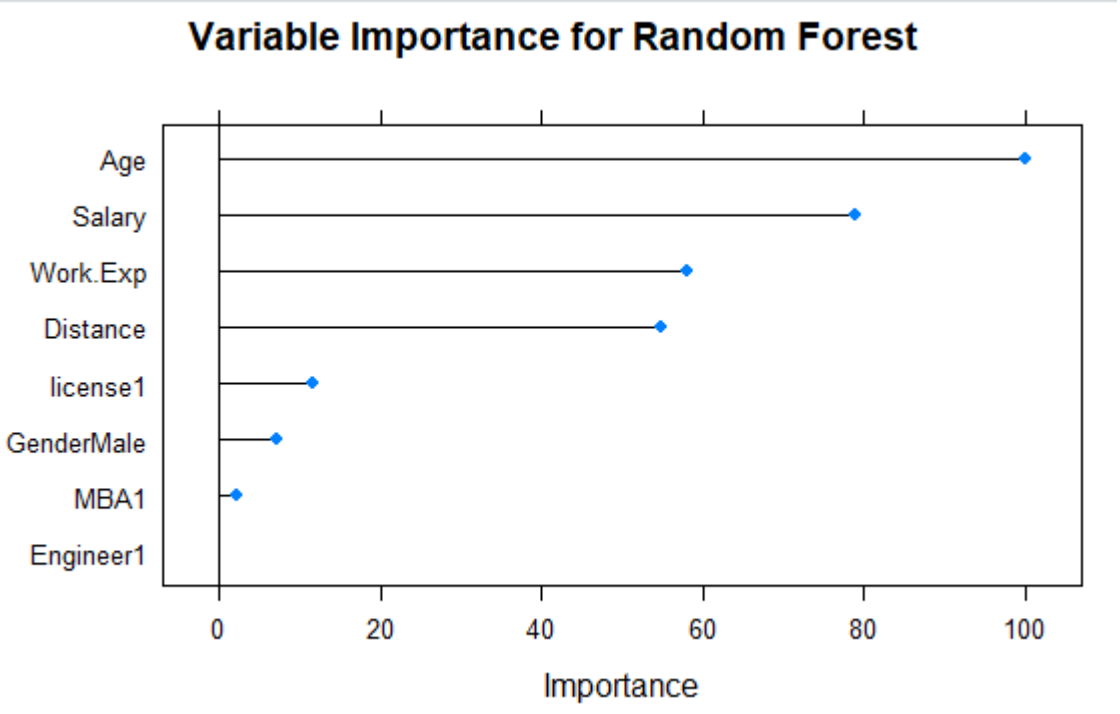
Number of trees: 500

No. of variables tried at each split: 3

OOB estimate of error rate: 17.7%

Confusion matrix:

	2Wheeler	Public Transport	Car	class.error
2Wheeler	92	25	1	0.22033898
Public Transport	24	89	5	0.24576271
Car	0	2	84	0.02325581



 # Predict for test data

Confusion Matrix and Statistics

Prediction	Reference 2Wheeler	Car	Public Transport

2Wheeler	17	1	29
Car	1	17	1
Public Transport	6	0	60

Overall Statistics

Accuracy : 0.7121

95% CI : (0.6269, 0.7876)

No Information Rate : 0.6818

P-Value [Acc > NIR] : 0.258725

Kappa : 0.4991

Mcnemar's Test P-Value : 0.001074

Statistics by Class:

	Class: 2Wheeler	Class: Car	Class: Public Transport
Sensitivity	0.7083	0.9444	0.6667
Specificity	0.7222	0.9825	0.8571
Pos Pred Value	0.3617	0.8947	0.9091
Neg Pred Value	0.9176	0.9912	0.5455
Prevalence	0.1818	0.1364	0.6818
Detection Rate	0.1288	0.1288	0.4545
Detection Prevalence	0.3561	0.1439	0.5000
Balanced Accuracy	0.7153	0.9635	0.7619