**ML Group Assignment**

This project requires you to understand what mode of transport employees prefers to commute to their office. The attached dataset [Cars\_edited.csv](https://olympus.greatlearning.in/courses/4746/files/504692/download?verifier=ydNlsRK6AQDbeq6DNf9KghOKoe99Iza7VaruU8un&wrap=1) 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?

Following is expected out of the candidate in this assessment.

**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

**Modelling**

* 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 modelling 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

**Group Members**

1. Aditya Godse
2. Akshay Surve
3. Amay Saxena

**Answer**

**1] Exploratory Data Analysis**

**A. Basic data summary, Univariate, Bivariate analysis, graphs, Check for Outliers and missing values and check the summary of the dataset**

setwd("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment 7 ML")

getwd()

cars\_data <- read.csv("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment 7 ML/Cars.csv")

str(cars\_data)

library(caret)

library(DMwR)

Converting variables Engineer, MBA and license in to factor variables since they are of categorical type.

cars\_data$Engineer<-as.factor(cars\_data$Engineer)

cars\_data$MBA<-as.factor(cars\_data$MBA)

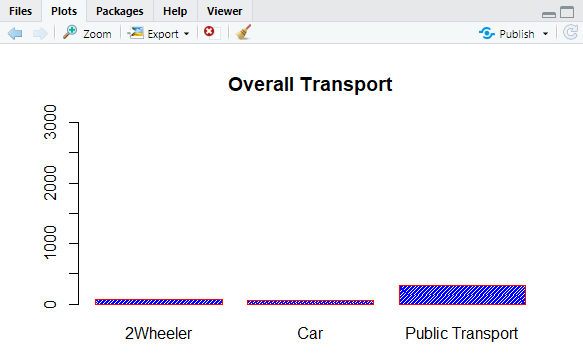
cars\_data$license<-as.factor(cars\_data$license)

cars\_data<-knnImputation(cars\_data)

**Bar plots:**

barplot(table(cars\_data$Transport),main = "Overall Transport",col="Blue",border="Red",

density=100,ylim = c(0,3000))



It can be seen that most of the employees use public transport in order to travel between home and office followed by 2 wheelers and then cars.

summary(cars\_data$Transport)

2Wheeler Car Public Transport

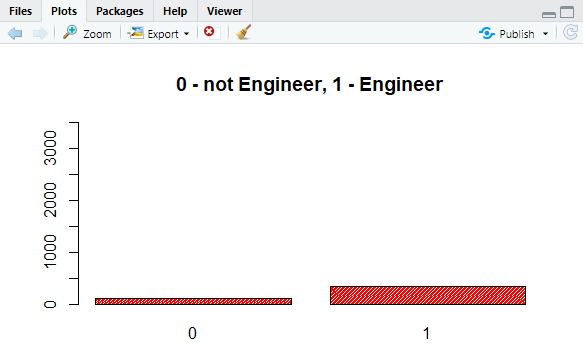
83 61 300

From summary of churn rate of customers, it can be seen that number of employees who use car as a mode of transport are 61 i.e 61/444 = 0.137387 (13.7387%).

Total employees travelling by car are 13.7387 %.

Now, considering only factor variables,

barplot(table(cars\_data$Engineer),density = 100,col="red",main = "0 - not Engineer, 1 - Engineer", ylim = c(0,3500))



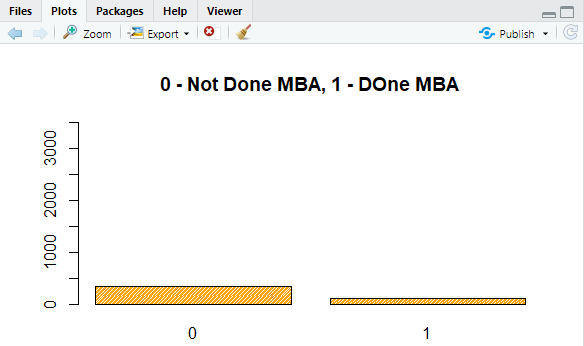
From Boxplot it can be seen that most of the employees come from engineering background.

summary(cars\_data$Engineer)

0 1

109 335

barplot(table(cars\_data$MBA),density = 100,col="orange", main = "0 - Not Done MBA, 1 - DOne MBA", ylim = c(0,3500))



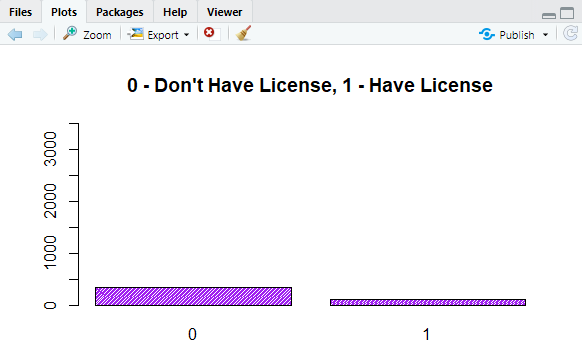
From the boxplot it seems Most of the employees have not done MBA.

summary(cars\_data$MBA)

0 1

332 112

barplot(table(cars\_data$license),density = 100,col="purple",main = "0 - Don't Have License, 1 - Have License", ylim = c(0,3500))



Boxplot indicates that the number of employees having license are less than 1/3 rd. of the total employees.

summary(cars\_data$license)

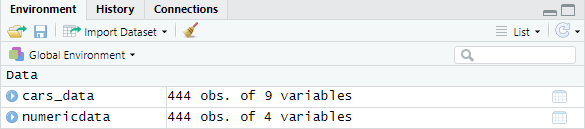
0 1

340 104

**B. Check for Multicollinearity - Plot the graph based on Multicollinearity**

Plot the graph based on Multicollinearity. For checking multi-collinearity of the dataset, considering only numeric variables. Therefore, discarding the factor variable columns.

numericdata<-cars\_data[,c(-2,-3,-4,-8,-9)]



print(cor(numericdata),digits = 3)

Age Work.Exp Salary Distance

Age 1.000 0.932 0.861 0.353

Work.Exp 0.932 1.000 0.932 0.373

Salary 0.861 0.932 1.000 0.442

Distance 0.353 0.373 0.442 1.000

library(corrplot)

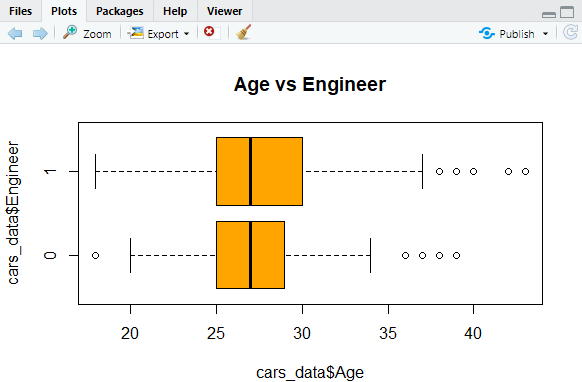
corrplot(cor(numericdata), method = c("number"), type = c("full"))



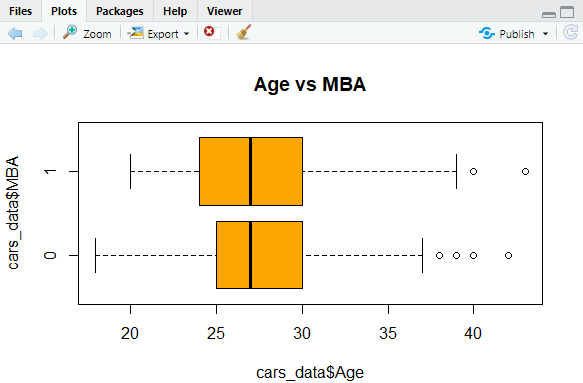
It can be seen that Age is highly collinear with work.Exp and Salary. Also, work.exp is highly collinear with salary and Age.

**C. Interpreting the business problems and sharing the observations**

boxplot(cars\_data$Age ~ cars\_data$Engineer,cars\_data$MBA, main = "Age vs Engineer", col = 'orange', horizontal = TRUE)



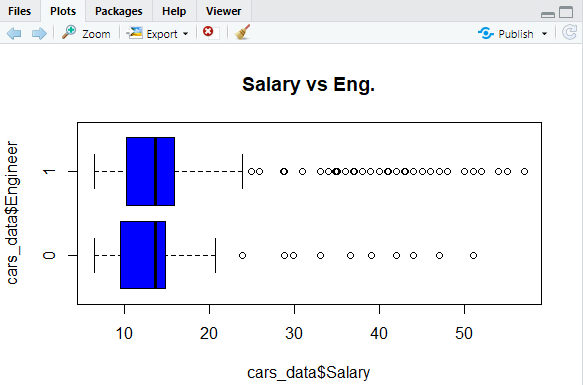
boxplot(cars\_data$Age ~ cars\_data$MBA, main = "Age vs MBA", col = 'orange', horizontal = TRUE)



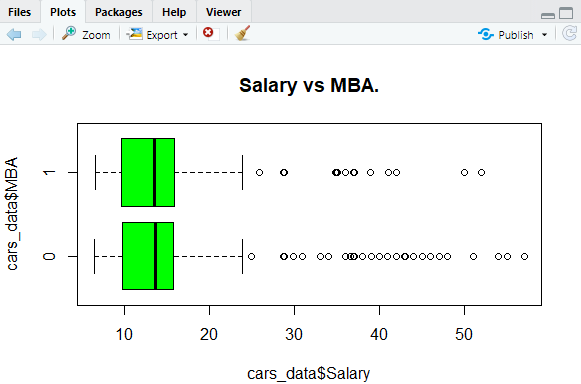
Not much difference is seen when Age is compared against Occupation i.e. Enginner and Education i.e. MBA.

Let us see the avg difference in salary for two profession,

boxplot(cars\_data$Salary ~cars\_data$Engineer, main = "Salary vs Eng.", col = "blue", horizontal = TRUE)

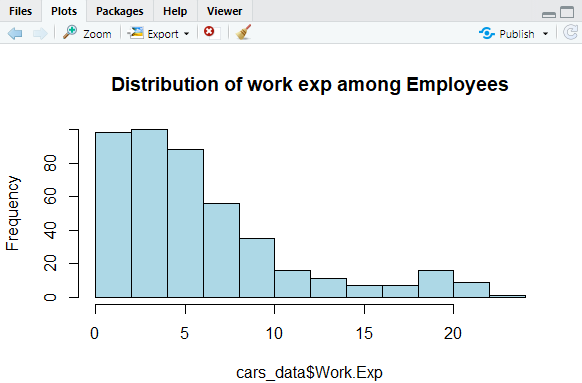


boxplot(cars\_data$Salary ~cars\_data$MBA, main = "Salary vs MBA.", col = "Green", horizontal = TRUE)



Bxplots doesn’t give any significant difference in salary of Engineers Vs Non-Engineers or MBA vs Non-MBA’s. Also, mean salary for both MBA’s and Engineer is around 16.

hist(cars\_data$Work.Exp, col = "light blue", main = "Distribution of work exp")



Histogram is skewed to the right. This indicates presence of employees with low work experience. Most of the Employees have around 5 years of work experience and the number goes down with increase in work Experience.

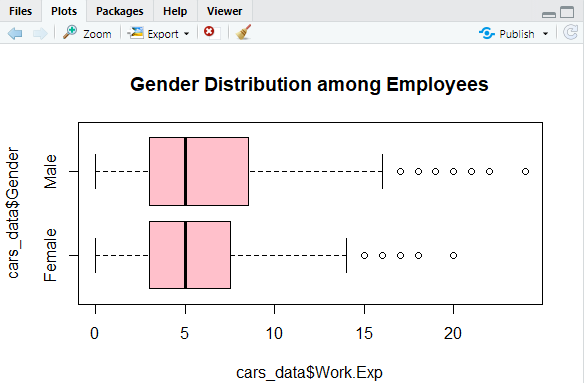
table(cars\_data$license,cars\_data$Transport)

2Wheeler Car Public Transport

0 60 13 267

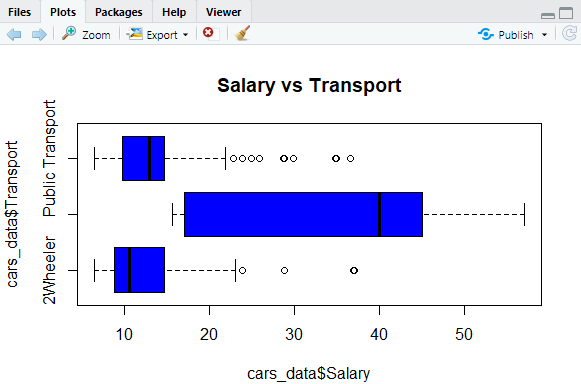
1 23 48 33

boxplot(cars\_data$Work.Exp ~ cars\_data$Gender, main = "Gender Distribution among Employees", col = 'pink', horizontal = TRUE)

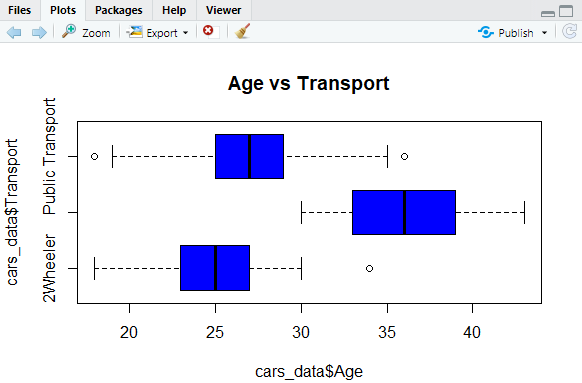


Boxplot suggests that number of female Employees are slightly lesser than male employees. But, overall mean population is equal.

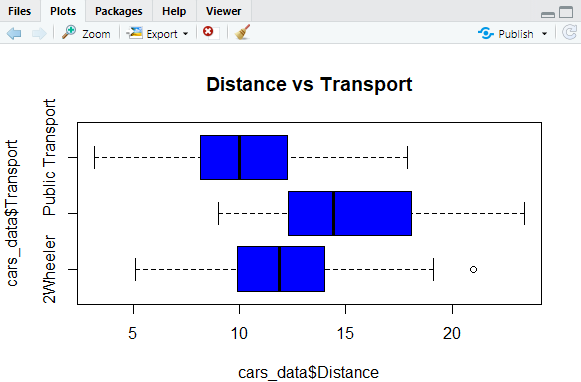
boxplot(cars\_data$Salary~cars\_data$Transport, main="Salary vs Transport", horizontal = TRUE, col = "blue")



boxplot(cars\_data$Age~cars\_data$Transport, main="Age vs Transport", horizontal = TRUE, col = "blue")



boxplot(cars\_data$Distance~cars\_data$Transport, main="Distance vs Transport", horizontal = TRUE, col = "blue")



As was the case with salary, we could see clear demarcation in usage of transport. With lower age group 2-wheeler is preferable and with higher work experience car is preferred.

table(cars\_data$Gender,cars\_data$Transport)

2Wheeler Car Public Transport

Female 38 13 77

Male 45 48 223

Our primary interest as per problem statement is to understand the factors influencing car usage. Hence we will create a new column for Car usage. It will take value 0 for Public Transport & 2 Wheeler and 1 for car usage.

cars\_data$CarUsage<-ifelse(cars\_data$Transport =='Car',1,0)

table(cars\_data$CarUsage)

0 1

383 61

sum(cars\_data$CarUsage == 1)/nrow(cars\_data)

[1] 0.1373874

cars\_data$CarUsage<-as.factor(cars\_data$CarUsage)

The number of records for people travelling by car is in minority. Hence we need to use an appropriate sampling method on the train data. Using SMOTE We will use logistic regression, boosting, KNN and NB.

**Interpreting the business problem**

To understand which variables are significant predictor behind the decision of employees to use Car as a mode of transport.

**2] Data Preparation**

Split the data into test and train datasets

set.seed(123)

cars\_split<-createDataPartition(cars\_data$CarUsage, p=0.7,list = FALSE,times = 1)

cars\_train<-cars\_data[cars\_split,]

cars\_test<-cars\_data[-cars\_split,]

dim(cars\_train)

311 9

dim(cars\_test)

133 9

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

prop.table(table(cars\_train$CarUsage))

0 1

0.8585209 0.1414791

Here, it can be seen that about 14.14% employees use car and other use different means of transport.

The new column CarUsage here indicates the total number of employees who drive cars. Also, it combines the remaining modes of transport, i.e. 2 Wheeler and Public Transport. Thus, indicating employees with 1 who have cars and other employees as 0.

Discard transport column from cars\_data since we are utilising newly created column CarUsage

cars\_train<-cars\_train[,c(1:8,10)]

cars\_test<-cars\_test[,c(1:8,10)]

**Apply SMOTE on Training data set,**

library(DMwR)

cars\_SMOTE<-SMOTE(cars\_train$CarUsage~., cars\_train, perc.over = 250,perc.under = 150)

prop.table(table(cars\_SMOTE$CarUsage))

0 1

0.5 0.5

There is equal split in the data between car users and non-car users.

**3] A. Applying Logistic Regression & Interpret results**

Create control parameter for GLM,

outcome\_variable <-'CarUsage'

regressors<-c("Age","Work.Exp","Salary","Distance","license","Engineer","MBA","Gender")

trainctrl<-trainControl(method = 'repeatedcv',number = 10,repeats = 3)

glm<-train(cars\_SMOTE[,regressors],cars\_SMOTE[,outcome\_variable],

method = "glm", family = "binomial",trControl = trainctrl)

summary(glm$finalModel)

Call:

NULL

Deviance Residuals:

Min 1Q Median 3Q Max

-2.50978 -0.02596 0.00000 0.06032 1.66304

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -58.62969 13.43393 -4.364 1.28e-05 \*\*\*

Age 1.89036 0.46410 4.073 4.64e-05 \*\*\*

Work.Exp -0.85922 0.33250 -2.584 0.00976 \*\*

Salary 0.24095 0.09164 2.629 0.00855 \*\*

Distance 0.17020 0.13063 1.303 0.19262

license1 2.27392 0.81389 2.794 0.00521 \*\*

Engineer1 0.32231 0.89788 0.359 0.71962

MBA1 -0.96837 0.72662 -1.333 0.18263

GenderMale -0.13497 0.73319 -0.184 0.85395

---

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: 65.573 on 249 degrees of freedom

AIC: 83.573

Number of Fisher Scoring iterations: 9

glmcoeff

(Intercept) Age Work.Exp Salary Distance license1

3.447055e-26 6.621739e+00 4.234924e-01 1.272457e+00 1.185537e+00 9.717415e+00

Engineer1 MBA1 GenderMale

1.380311e+00 3.797024e-01 8.737418e-01

varImp(object = glm)

glm variable importance

Overall

Age 100.000

license1 67.106

Salary 62.877

Work.Exp 61.712

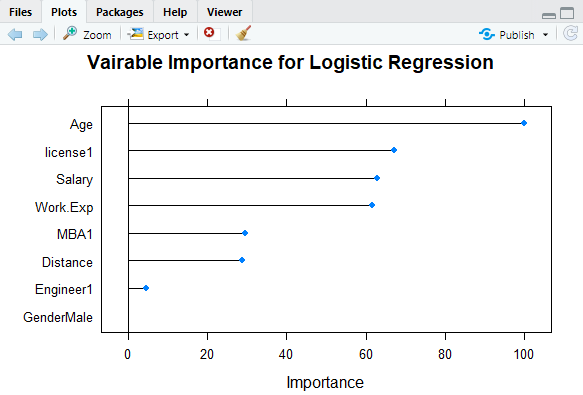
MBA1 29.534

Distance 28.767

Engineer1 4.497

GenderMale 0.000

plot(varImp(object = glm), main="Vairable Importance for Logistic Regression")



Age is the most important variable followed by license, salary and work experience, while gender happens to be least important variable.

When we look at the odds and probabilities table, we get to see that Increase in age by 1 year implies that there is a 98% probability that the employee will use a car. As expected, if the employee has a license, then it implies a 99% probability that he/she will use a car.

One lakh increase in salary increases the probability of car usage by 72%. The null deviance of this model is 357.664 and the residual deviance is 17.959. This yields a McFadden R Square of almost 0.94 yielding a very good fit. We get to see Accuracy and Kappa values are high.

We will do the prediction based on this model.

carusage\_pred<-predict.train(object = glm,cars\_test[,regressors],type = "raw")

confusionMatrix(carusage\_pred,cars\_test[,outcome\_variable], positive='1')

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 110 1

1 4 17

Accuracy : 0.9621

95% CI : (0.9138, 0.9876)

No Information Rate : 0.8636

P-Value [Acc > NIR] : 0.000156

Kappa : 0.8497

Mcnemar's Test P-Value : 0.371093

Sensitivity : 0.9444

Specificity : 0.9649

Pos Pred Value : 0.8095

Neg Pred Value : 0.9910

Prevalence : 0.1364

Detection Rate : 0.1288

Detection Prevalence : 0.1591

Balanced Accuracy : 0.9547

'Positive' Class : 1

This model gives accuracy of 96.21 % which seems it is over fitting the model. In order to improve result further apply ridge regression model on using glmnet method.

For this caret package is required.

trainctrlgn<-trainControl(method = 'cv',number = 10,returnResamp = 'none')

glmnet<-train(CarUsage~Age+Work.Exp+Salary+Distance+license, data = cars\_SMOTE,

method = 'glmnet', trControl = trainctrlgn)

glmnet

glmnet

258 samples

5 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 232, 232, 233, 232, 233, 232, ...

Resampling results across tuning parameters:

alpha lambda Accuracy Kappa

0.10 0.0008061155 0.9455385 0.8912802

0.10 0.0080611546 0.9300000 0.8603672

0.10 0.0806115458 0.9184615 0.8372903

0.55 0.0008061155 0.9455385 0.8912802

0.55 0.0080611546 0.9338462 0.8680595

0.55 0.0806115458 0.9300000 0.8603672

1.00 0.0008061155 0.9416923 0.8835879

1.00 0.0080611546 0.9455385 0.8912802

1.00 0.0806115458 0.9453846 0.8910361

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

The final values used for the model were alpha = 1 and lambda = 0.008061155.

varImp(object = glmnet)

glmnet variable importance

Overall

license1 100.000

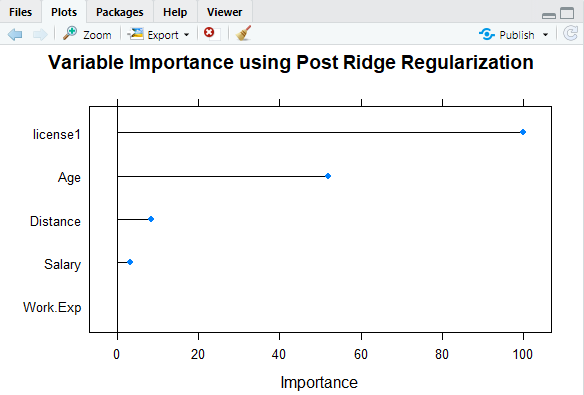
Age 52.008

Distance 8.266

Salary 3.106

Work.Exp 0.000

plot(varImp(object = glmnet), main="Vairable Importance for Logistic Regression using Post Ridge Regularization")



The license and Age are the most significant variables followed by distance. Further variables gender, engineer, MBA are neglected. Work experience appears to be of least importance.

**Regularization model for Logistic Regression:**

carusage\_pred<-predict.train(object = glmnet,cars\_test[,regressors],type = "raw")

confusionMatrix(carusage\_pred,cars\_test[,outcome\_variable], positive='1')

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 111 2

1 3 16

Accuracy : 0.9621

95% CI : (0.9138, 0.9876)

No Information Rate : 0.8636

P-Value [Acc > NIR] : 0.000156

Kappa : 0.8429

Mcnemar's Test P-Value : 1.000000

Sensitivity : 0.8889

Specificity : 0.9737

Pos Pred Value : 0.8421

Neg Pred Value : 0.9823

Prevalence : 0.1364

Detection Rate : 0.1212

Detection Prevalence : 0.1439

Balanced Accuracy : 0.9313

'Positive' Class : 1

The regularization of model retains the same accuracy as before running ridge regression model.

With **Accuracy** of **96.21%.**

**B. Applying KNN Model & Interpreting results**

Before applying KNN we need to normalize the continuous variables.

Therefore, normalizing variables Salary.

cars\_data$Salary = log(cars\_data$Salary)

cars\_test$Salary = log(cars\_test$Salary)

Splitting data into test and train dataset

set.seed(123)

cars\_split<-createDataPartition(cars\_data$CarUsage, p=0.7,list = FALSE,times = 1)

cars\_train<-cars\_data[cars\_split,]

cars\_test<-cars\_data[-cars\_split,]

loading required library

library(caret)

trControl <- trainControl(method = "cv", number = 10)

fit.knn <- train(cars\_data$Transport ~ ., method = "knn", data = cars\_data,

trControl = trControl,

metric = "Accuracy",

preProcess = c("center","scale"))

fit.knn

k-Nearest Neighbors

312 samples

8 predictor

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

Pre-processing: centered (8), scaled (8)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 281, 281, 280, 281, 280, 282, ...

Resampling results across tuning parameters:

k Accuracy Kappa

2 0.7365457 0.4543489

3 0.7855712 0.5248631

4 0.7629839 0.4800127

5 0.7828562 0.5081854

6 0.7734812 0.4905393

7 0.7634005 0.4624704

8 0.7408065 0.4118105

9 0.7534005 0.4199273

10 0.7536022 0.4116860

11 0.7598454 0.4168749

12 0.7662970 0.4266860

13 0.7662970 0.4213708

14 0.7566129 0.3930122

15 0.7661895 0.4135919

16 0.7660887 0.4090611

17 0.7566129 0.3862387

18 0.7629637 0.3926229

19 0.7661895 0.4026549

20 0.7661895 0.3942178

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

The final value used for the model was k = 3.

Number of cluster/ centroids required in this case are 3.

KNN\_predictions = predict(fit.knn,cars\_train)

table(KNN\_predictions, cars\_train$Transport)

KNN\_predictions 2Wheeler Car Public Transport

2Wheeler 37 0 8

Car 0 35 2

Public Transport 22 8 200

KNN\_predictions = predict(fit.knn,cars\_test)

table(KNN\_predictions, cars\_test$Transport)

predict(fit.knn,cars\_data)

KNN\_predictions 2Wheeler Car Public Transport

2Wheeler 9 0 11

Car 1 15 3

Public Transport 14 3 76

In above confusion matrix values which are correctly predicted i.e. for 2wheeler (9), car (15) and public transport (76) indicates True Positives.

Hence Accuracy = (9+15+76) / (9+11+1+15+3+14+3+76) = 0.75757575 i.e. 75.75 %.

**Accuracy** in this case is **75.75%** when model is run on test data.

**C. Applying Naïve Baye’s Model & Interpret results**

library(e1071)

NB\_Model=naiveBayes(cars\_train$Transport ~., data=cars\_train)

NB\_Model

Naive Bayes Classifier for Discrete Predictors

Call:

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

A-priori probabilities:

Y

2Wheeler Car Public Transport

0.1826923 0.1378205 0.6794872

Conditional probabilities:

Age

Y [,1] [,2]

2Wheeler 25.03509 2.865839

Car 35.74419 3.579658

Public Transport 26.91038 2.976446

Gender

Y Female Male

2Wheeler 0.4385965 0.5614035

Car 0.2093023 0.7906977

Public Transport 0.2924528 0.7075472

Engineer

Y 0 1

2Wheeler 0.2982456 0.7017544

Car 0.1162791 0.8837209

Public Transport 0.2594340 0.7405660

MBA

Y 0 1

2Wheeler 0.8070175 0.1929825

Car 0.7906977 0.2093023

Public Transport 0.7547170 0.2452830

Work.Exp

Y [,1] [,2]

2Wheeler 3.842105 3.320873

Car 15.627907 4.928166

Public Transport 5.113208 3.221141

Salary

Y [,1] [,2]

2Wheeler 2.415753 0.3838427

Car 3.493113 0.4290057

Public Transport 2.533174 0.3289971

Distance

Y [,1] [,2]

2Wheeler 12.53860 3.537036

Car 15.40233 3.745759

Public Transport 10.43585 2.875709

license

Y 0 1

2Wheeler 0.73684211 0.26315789

Car 0.25581395 0.74418605

Public Transport 0.91037736 0.08962264

CarUsage

Y 0 1

2Wheeler 1 0

Car 0 1

Public Transport 1 0

Prediction on the test dataset,

NB\_Predictions=predict(NB\_Model,cars\_test)

NB\_Predictions

[1] Public Transport Public Transport Public Transport Public Transport

[5] Public Transport Public Transport Public Transport Public Transport

[9] Public Transport Public Transport Public Transport Public Transport

[13] Public Transport Public Transport Public Transport Public Transport

[17] Public Transport Public Transport Public Transport Public Transport

[21] Public Transport Public Transport Public Transport Public Transport

[25] Public Transport Public Transport Public Transport Public Transport

[29] Public Transport Public Transport Public Transport Public Transport

[33] Public Transport Public Transport Public Transport Public Transport

[37] Public Transport Public Transport Public Transport Public Transport

[41] Public Transport Public Transport Public Transport Public Transport

[45] Public Transport Public Transport Public Transport Public Transport

[49] Public Transport Public Transport Public Transport Public Transport

[53] Public Transport Public Transport Public Transport Public Transport

[57] Car Public Transport Public Transport Car

[61] Car Public Transport Public Transport Public Transport

[65] Public Transport Public Transport Public Transport Public Transport

[69] 2Wheeler Public Transport Public Transport 2Wheeler

[73] Public Transport Public Transport Public Transport Public Transport

[77] Public Transport Public Transport Public Transport Public Transport

[81] Public Transport Public Transport Public Transport 2Wheeler

[85] Public Transport Public Transport Public Transport Car

[89] Car Public Transport Public Transport Car

[93] Public Transport Car Public Transport Public Transport

[97] 2Wheeler Public Transport Public Transport Public Transport

[101] Car Public Transport Public Transport 2Wheeler

[105] Public Transport Public Transport Public Transport 2Wheeler

[109] Car Car Car Public Transport

[113] Public Transport Car 2Wheeler 2Wheeler

[117] Public Transport 2Wheeler 2Wheeler Car

[121] Public Transport Public Transport Public Transport Car

[125] Public Transport 2Wheeler 2Wheeler Car

[129] Car Car Car Car

Levels: 2Wheeler Car Public Transport

summary(NB\_Predictions)

2Wheeler Car Public Transport

12 19 101

table(NB\_Predictions,cars\_test$Transport)

NB\_Predictions 2Wheeler Car Public Transport

2Wheeler 5 0 7

Car 1 17 1

Public Transport 20 1 80

Hence Accuracy = (5+17+80) / (5+7+1+17+1+20+1+80) = 0.772727 i.e. 77.27 %.

**Accuracy** in this case is **77.27 %** when model is run on test data.

**D. Confusion matrix interpretation & Remarks on Model validation exercise**

|  |  |  |
| --- | --- | --- |
| **Model** | **Confusion Matrix** | **Accuracy** |
| **Logistic Regression** | Reference  Prediction 0 1  0 111 2  1 3 16 | **96.21%** |
| **KNN** | KNN\_predictions 2Wheeler Car Public Transport  2Wheeler 9 0 11  Car 1 15 3  Public Transport 14 3 76 | **75.75%** |
| **Naïve Bayes’** | NB\_Predictions 2Wheeler Car Public Transport  2Wheeler 5 0 7  Car 1 17 1  Public Transport 20 1 80 | **77.27%** |

From the above table it seems Logistic regression model gives highest accuracy followed by Naïve Bayes’ and further KNN which has least accuracy.

In case of logistic regression both glmnet and ridge regression model surprisingly gives similar accuracies or model.

**4] Bagging and Boosting**

**Bagging**

cars\_data\_lda <- read.csv("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment 7 ML/Cars.csv")

cars\_data\_lda$Gender<-as.factor(cars\_data\_lda$Gender)

cars\_data\_lda$Engineer<-as.factor(cars\_data\_lda$Engineer)

cars\_data\_lda$MBA<-as.factor(cars\_data\_lda$MBA)

cars\_data\_lda<-knnImputation(cars\_data\_lda)

Split the original base data into test and train samples again

set.seed(123)

cars\_split\_lda<-createDataPartition(cars\_data\_lda$Transport, p=0.7,list = FALSE,times = 1)

cars\_train\_lda<-cars\_data\_lda[cars\_split\_lda,]

cars\_test\_lda<-cars\_data\_lda[-cars\_split\_lda,]

cars\_train\_lda$license<-as.factor(cars\_train\_lda$license)

cars\_test\_lda$license<-as.factor(cars\_test\_lda$license)

cartrain\_lda.car<-cars\_train\_lda[cars\_train\_lda$Transport %in% c("Car", "Public Transport"),]

cartrain\_lda.twlr<-cars\_train\_lda[cars\_train\_lda$Transport %in% c("2Wheeler", "Public Transport"),]

cartrain\_lda.car$Transport<-as.character(cartrain\_lda.car$Transport)

cartrain\_lda.car$Transport<-as.factor(cartrain\_lda.car$Transport)

cartrain\_lda.twlr$Transport<-as.character(cartrain\_lda.twlr$Transport)

cartrain\_lda.twlr$Transport<-as.factor(cartrain\_lda.twlr$Transport)

prop.table(table(cartrain\_lda.car$Transport))

Car Public Transport

0.1699605 0.8300395

prop.table(table(cartrain\_lda.twlr$Transport))

2Wheeler Public Transport

0.2193309 0.7806691

car\_lda\_twlrsm <- SMOTE(Transport~., data = cartrain\_lda.twlr, perc.over = 150, perc.under=200)

table(car\_lda\_twlrsm$Transport)

2Wheeler Public Transport

118 118

car\_lda\_carsm <- SMOTE(Transport~., data = cartrain\_lda.car, perc.over = 175, perc.under=200)

table(car\_lda\_carsm$Transport)

Car Public Transport

86 86

car\_lda<-car\_lda\_carsm[car\_lda\_carsm$Transport %in% c("Car"),]

cars\_train\_ldasm<-rbind(car\_lda\_twlrsm,car\_lda)

str(cars\_train\_ldasm)

|  |
| --- |
| 'data.frame': 322 obs. of 9 variables:  $ Age : num 27 28 25 23 26 26 25 27 26 26 ...  $ Gender : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 1 1 1 ...  $ Engineer : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  $ MBA : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 2 1 1 ...  $ Work.Exp : num 8 4 3 2 3 2 3 4 3 3 ...  $ Salary : num 21.8 14.8 9.9 8.5 10.7 9.6 10.7 13.7 10.5 10.5 ...  $ Distance : num 13.4 13 17.2 6.1 12.2 9.5 10.8 13.3 5.1 5.1 ...  $ license : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...  $ Transport: Factor w/ 3 levels "2Wheeler","Public Transport",..: 2 2 2 2 2 ... |
|  |
| |  | | --- | |  | |

**Boosting**

boostcontrol <- trainControl(number=10)

xgbGrid <- expand.grid(

eta = 0.3,

max\_depth = 1,

nrounds = 50,

gamma = 0,

colsample\_bytree = 0.6,

min\_child\_weight = 1, subsample = 1

)

cars\_boosting <- train(Transport ~ .,cars\_train\_ldasm,trControl = boostcontrol,

tuneGrid = xgbGrid,metric = "Accuracy",method = "xgbTree")

cars\_boosting

|  |
| --- |
| eXtreme Gradient Boosting  322 samples  8 predictor  3 classes: '2Wheeler', 'Public Transport', 'Car'  No pre-processing  Resampling: Bootstrapped (10 reps)  Summary of sample sizes: 322, 322, 322, 322, 322, 322, ...  Resampling results:  Accuracy Kappa  0.7591986 0.6364491  Tuning parameter 'nrounds' was held constant at a value of 50  Tuning  0.6  Tuning parameter 'min\_child\_weight' was held constant at a value of 1  Tuning parameter 'subsample' was held constant at a value of 1 |
|  |
| |  | | --- | |  | |

Predict using the test dataset,

predictions\_boosting<-predict(cars\_boosting,cars\_test\_lda)

confusionMatrix(predictions\_boosting,cars\_test\_lda$Transport)

Confusion Matrix and Statistics

Reference

Prediction 2Wheeler Car Public Transport

2Wheeler 18 2 25

Car 0 16 1

Public Transport 6 0 64

Overall Statistics

Accuracy : 0.7424

95% CI : (0.6591, 0.8146)

No Information Rate : 0.6818

P-Value [Acc > NIR] : 0.078662

Kappa : 0.5391

Mcnemar's Test P-Value : 0.002146

Statistics by Class:

Class: 2Wheeler Class: Car Class: Public Transport

Sensitivity 0.7500 0.8889 0.7111

Specificity 0.7500 0.9912 0.8571

Pos Pred Value 0.4000 0.9412 0.9143

Neg Pred Value 0.9310 0.9826 0.5806

Prevalence 0.1818 0.1364 0.6818

Detection Rate 0.1364 0.1212 0.4848

Detection Prevalence 0.3409 0.1288 0.5303

Balanced Accuracy 0.7500 0.9401 0.7841

The **overall accuracy** is at **74.24 %** with accuracy of **car usage predicted** at **91.43 %** Let us try and predict for the two unknown cases.

**5] Actionable Insights and Recommendations**

* **Most significant variables** which are influencing employees to use car as mode of transport are **Age, License, Distance and Work Experience**.
* Employees with high work Experience are likely to take Car as mode of transport.
* Employees who are traveling maximum distance between office and home are using car as mode of transport.
* 74.41 % of employees who drive car have license while remaining employees from this segment drive car without license.