***Project-5:***

**Objective:**

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.

**EDA:**

* The numeric variables Salary and Work.Exp are Skewed to the right, some variable have been converted to factors and a new variable Tranport1 has been created by combining the levels 2-wheeler and Public transport (0=Other and 1= Car)

which will be used as a DV for further analysis, missing values have been omitted.

'data.frame': 417 obs. of 9 variables:

$ Age : int 28 24 27 25 25 21 23 23 24 28 ...

$ Gender : Factor w/ 2 levels "Female","Male": 2 2 1 2 1 2 2 2 2 2 ...

$ Engineer : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 2 1 2 2 ...

$ MBA : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...

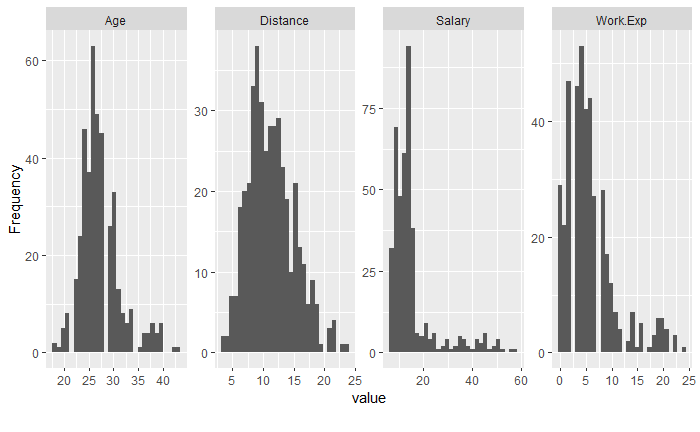
$ Work.Exp : int 5 6 9 1 3 3 3 0 4 6 ...

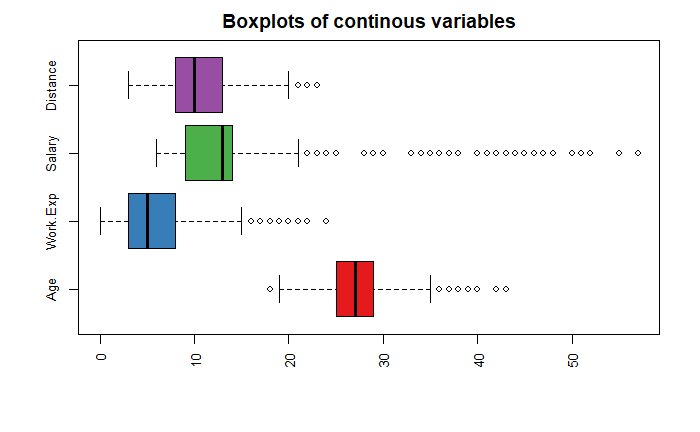
$ Salary : num 14.4 10.6 15.5 7.6 9.6 9.5 11.7 6.5 8.5 13.7 ...

$ Distance : num 5.1 6.1 6.1 6.3 6.7 7.1 7.2 7.3 7.5 7.5 ...

$ license : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...

$ Transport1: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...



***Code:***

ca=read.csv("cars.csv",header = T)

ca=na.omit(ca)

ca$Transport1=ifelse(ca$Transport==2,0,ifelse(ca$Transport=="P",0,1))

table(ca$Transport1)

library(DataExplorer)

plot\_histogram(ca)

ca$Engineer=as.factor(ca$Engineer)

ca$MBA=as.factor(ca$MBA)

ca$Gender=as.factor(ca$Gender)

ca$Transport1=as.factor(ca$Transport1)

ca$license=as.factor(ca$license)

summary(ca)

str(ca)

ca=ca[,-9]

cai=ca[,c(1,5,6,7)]

library(RColorBrewer)

boxplot(cai,

las=3,

horizontal = TRUE,

cex= 0.8,

par(cex.axis = 0.8),

col=brewer.pal(8,"Set1"),

main = "Boxplots of continous variables")

* Visualizing the factor and Numeric variables separately against the DV Transport1.

***Code:***

cef=ca[,c(2,3,4,8)]

par(mfrow=c(2,2))

for (i in names(cef)) {

print(i)

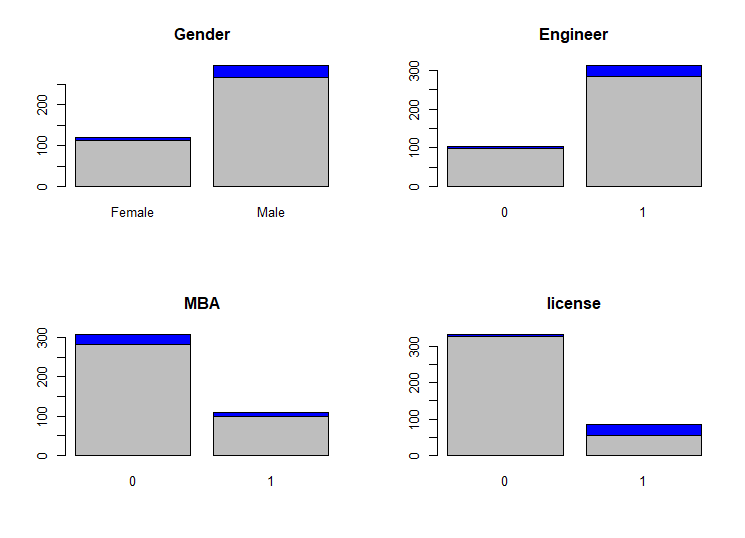
print(table(ca$Transport1, cef[[i]]))

barplot(table(ca$Transport1, cef[[i]]),

col=c("grey","blue"),

main = names(cef[i]))}

par(mfrow=c(1,1))



Observations for factor variables:

1. In terms of Graduation, Under-Graduates (Engineer) contribute 6% as compared with the Post-Graduates (MBA) who contribute only 2% to the overall Car usage which is 8%
2. In terms of Gender, Males seems to dominate cars use with a contribution of 7% to overall 8%
3. In terms of license, Employees with License contribute 7% to the overall 8%.

***Code:***

cai=ca[,c(1,5,6,7)]

ca2 <- cbind(cai, ca$Transport1)

colnames(ca2)[5] <- "Transport"

str(ca2)

**# stack the data using melt function.**

library(reshape2)

nd2.melt<- melt(ca2, id = c("Transport"))

**# box plots**

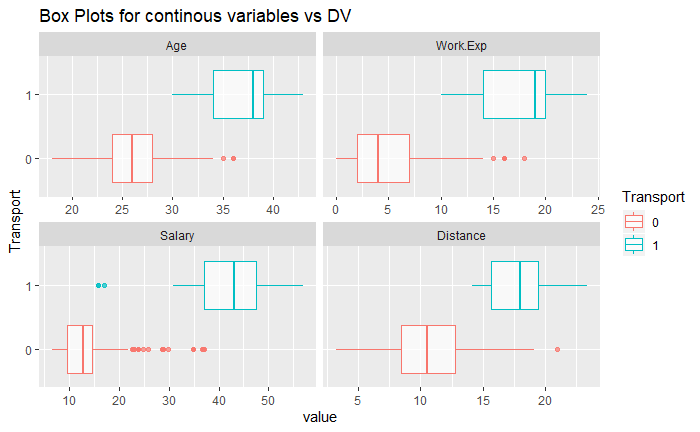
library(tidyverse)

zz <- ggplot(nd2.melt, aes(x=Transport, y=value))

zz+geom\_boxplot(aes(color = Transport), alpha=0.7 ) +

facet\_wrap(~variable,scales = "free\_x", nrow = 3)+

ggtitle("Box Plots for continous variables vs DV")+coord\_flip()



Observations for numeric variables:

1. The Interquartile range for the employees who use car to commute are higher as compared to employees who use other modes to commute.
2. Average Salary for employees who use car to commute is 41.29 as compared to the other mode of transport which is 13.05.
3. Average Age of employees who use car as a mode to commute is 36 as compare with other mode of transport which is 26.
4. Average Work Exp for employees who use car to commute is 17.5 as compare with other mode of transport which is 4.8.
5. Average Distance for employees who uses car to commute is 17 as compare with other mode of transport which is 10.

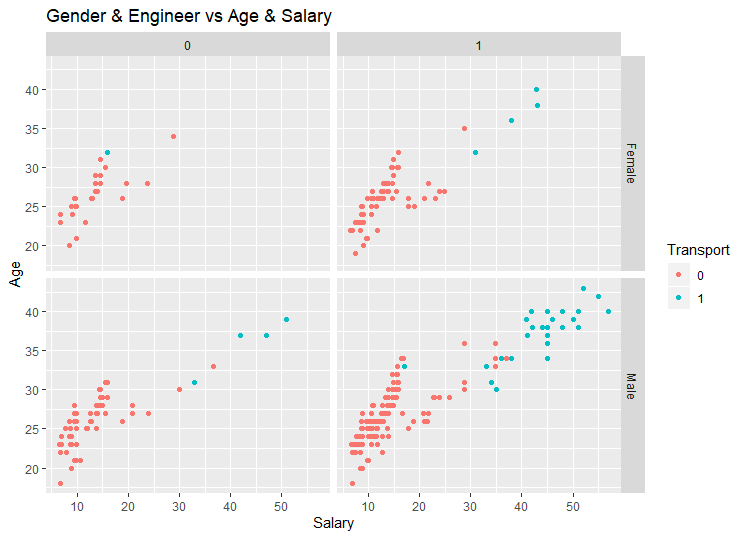
* Finding relation between variable by using Multi-variate analysis.

1. Using Transport as target variable and plotting a scatter plot against the predictor variables, Salary on the x-axis and Age on the y-axis and splitting it across factors Gender and Engineer.

*Code:*

bk=ggplot(ca,aes(x=Salary,y=Age,color=Transport1))

bk+geom\_point(aes(color=Transport1))+ggtitle("Gender & Engineer vs Age & Salary")+facet\_grid(Gender~Engineer)



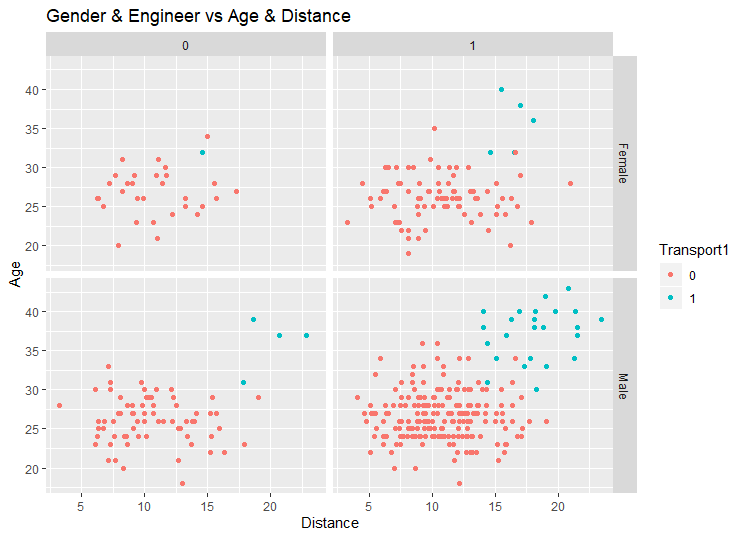
Observation: There appears to be a linear relationship between the variables.

1. Using Transport as target variable and plotting a scatter plot against the predictor variables, Distance on the x-axis and Age on the y-axis and splitting it across factors Gender and Engineer.

*Code:*

bk=ggplot(ca,aes(x=Distance,y=Age,color=Transport1))

bk+geom\_point(aes(color=Transport1))+ggtitle("Gender & Engineer vs Age & Distance")+facet\_grid(Gender~Engineer)



Observation: Male Engineer employees above 30 years in age and where distance is greater than 12.5 prefer commuting in car.

* Dropping Work.Exp from further analysis after checking for multi-collinearity and Variance Inflation using Linear regression.

*Code:*

library(corrplot)

str(ca)

ca=ca[,-9]

str(ca)

ca$Gender=ifelse(ca$Gender=="Male",1,0)

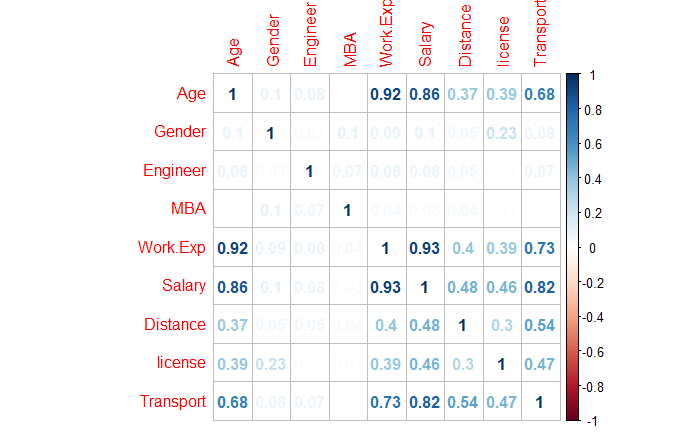
ca[,1:8]=lapply(ca[,1:8],as.integer)

str(ca)

ca$Transport1=as.numeric(ca$Transport1)

colnames(ca)[9]="Transport"

corrplot(cor(ca),method="number")



library(car)

str(ca)

**#After Dropping Work.Exp**

m=lm(Distance~Age+Gender+MBA+license+Salary+Transport+Engineer,data=ca)

summary(m)

vif(m)

ca=ca[,-5]

Age Gender MBA license Salary Transport Engineer

3.781607 1.069895 1.019641 1.380561 6.110774 3.101970 1.013003

KNN With Smote:

* Will be using the caret library and library DMwR to perform knn with smote.

*Code:*

**#Preparing the data and splitting it into 70:30 ratio.**

library(caTools)

nor <-function(x) { (x -min(x))/(max(x)-min(x))}

ca\_norm <- as.data.frame(lapply(ca[,c(1:8)], nor))

summary(ca\_norm)

ca\_norm$Transport=as.factor(ca\_norm$Transport)

set.seed(1900)

str(ca\_norm)

spl=sample.split(ca\_norm,SplitRatio = 0.7)

train=subset(ca\_norm,spl==T)

test=subset(ca\_norm,spl==F)

dim(train)

dim(test)

prop.table(table(train$Transport))

prop.table(table(test$Transport))

str(ca\_norm)

library(caret)

library(DMwR)

**#Setting the control parameter will be using 5 as number of neighbours for the best tune**

ctrl=trainControl(method = "repeatedcv",

number=10,

repeats = 10,

verboseIter = F,

sampling = "smote")

knn\_fit = train(Transport ~., data = train, method = "knn",

trControl = ctrl,

tuneLength = 5)

knn\_fit$bestTune$k

**#Confusion Matrix:**

final=data.frame(actual=test$Transport,predict(knn\_fit,newdata = test,type="prob"))

final$pred=ifelse(final$X0>0.5,"other","car")

table(test$Transport,final$pred)

pred=predict(knn\_fit,newdata=test,type="raw")

tab=table(test$Transport,pred)

tab

confusionMatrix(test$Transport,data=pred,positive="1")

varImp(knn\_fit)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 137 2

1 6 11

Accuracy : 0.9487

95% CI : (0.9015, 0.9776)

No Information Rate : 0.9167

P-Value [Acc > NIR] : 0.09006

Sensitivity : 0.84615

Specificity : 0.95804

Pos Pred Value : 0.64706

Neg Pred Value : 0.98561

Variable Importance > 50

|  |
| --- |
|  |
| |  | | --- | |  | | **Variable** | **Importance** |  |  |  | | Salary | 100.000000 |  |  |  | | Age | 99.802241 |  |  |  | | Distance | 94.462755 |  |  |  | | license | 65.260382 |  |  |  | |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Naïve Bayes with Smote:

* Will be using the caret library and library DMwR to perform Naïve Bayes with smote, have prepared the data by converting all the variable to categorical.

*Code:*

**#Preparing the data and splitting it into 70:30 ratio.**

ca=read.csv("cars.csv",header = T)

ca$Transport=ifelse(ca$Transport==2,0,ifelse(ca$Transport=="P",0,1))

summary(ca)

ca=na.omit(ca)

ca$Transport=as.factor(ca$Transport)

ca=ca[,-5]

summary(ca)

ca$Age1=as.numeric(cut(ca$Age,4))

ca$Age1=as.factor(ca$Age1)

summary(ca$Age1)

ca$Salary1=as.factor(as.numeric(cut(ca$Salary,4)))

summary(ca$Salary1)

ca$Distance1=as.factor(as.numeric(cut(ca$Distance,4)))

summary(ca)

ca$Engineer=as.factor(ca$Engineer)

ca$MBA=as.factor(ca$MBA)

ca$license=as.factor(ca$MBA)

ca$Gender=ifelse(ca$Gender=="Male",1,0)

ca$Gender=as.factor(ca$Gender)

str(ca)

can=ca[,-c(1,5,6)]

summary(can)

set.seed(100)

spl=sample.split(can,SplitRatio = 0.7)

trn=subset(can,spl==T)

tes=subset(can,spl==F)

dim(trn)

dim(tes)

prop.table(table(trn$Transport))

prop.table(table(tes$Transport))

nbb = train(Transport~., data = trn, method = "nb",

trControl = ctrl,

tuneLength = 10)

final=data.frame(actual=tes$Transport,predict(nbb,newdata = tes,type="prob"))

final$pred=ifelse(final$X1>0.5,"1","0")

table(final$pred,test$Transport)

table(test$Transport)

pre=predict(nbb,newdata=tes,type="raw")

tab=table(tes$Transport,pre)

tab

confusionMatrix(test$Transport,data=pre,positive="1")

varImp(nbb)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 132 1

1 11 12

Accuracy : 0.9231

95% CI : (0.8695, 0.9596)

No Information Rate : 0.9167

P-Value [Acc > NIR] : 0.458291

Sensitivity : 0.92308

Specificity : 0.92308

Pos Pred Value : 0.52174

Neg Pred Value : 0.99248

Variable Importance > 50

| **Variable** | **Importance** |  |  |  |
| --- | --- | --- | --- | --- |
| Age1 | 100.0000000 |  |  |  |
| Distance1 | 92.9401447 |  |  |  |
| Salary1 | 90.5064679 |  |  |  |

Logistic Regression with Smote:

* In the given data set while performing logistic regression only variables Distance and Salary seem to be significant.

*Code:*

**#Preparing the data and partitioning to 70:30 ratio**

library(caTools)

ca=read.csv("cars.csv",header = T)

str(ca)

summary(ca)

ca=na.omit(ca)

summary(ca)

ca$Transport=ifelse(ca$Transport==2,0,ifelse(ca$Transport=="P",0,1))

str(ca)

ca$Engineer=as.factor(ca$Engineer)

ca$MBA=as.factor(ca$MBA)

ca$Gender=as.factor(ca$Gender)

ca$Transport=as.factor(ca$Transport)

str(ca)

summary(ca)

ca$license=as.factor(ca$license)

summary(ca)

ca=ca[,-5]

set.seed(1979)

spl=sample.split(ca,SplitRatio = 0.7)

train=subset(ca,spl==T)

test=subset(ca,spl==F)

dim(train)

dim(test)

summary(ca)

library(DMwR)

balanced.gd <- SMOTE(Transport~.,train, perc.over = 100, k = 5, perc.under = 450)

table(balanced.gd$Transport)

logi=glm(Transport~Distance+Salary,data=balanced.gd,family=binomial(link="logit"))

summary(logi)

logi$residuals

table(test$Transport)

pred=predict(logi,newdata=test,type="response")

test$Pred=predict(logi,test,type="response")

test$Pred=ifelse(test$Pred>0.5,1,0)

test$Pred=as.factor(test$Pred)

confusionMatrix(test$Transport,test$Pred,positive = "1")

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 141 3

1 1 12

Accuracy : 0.9745

95% CI : (0.9361, 0.993)

No Information Rate : 0.9045

P-Value [Acc > NIR] : 0.0005594

Sensitivity : 0.80000

Specificity : 0.99296

Pos Pred Value : 0.92308

Neg Pred Value : 0.97917

Random Forest with Smote(Bagging Technique):

*Code:*

library(caTools)

library(caret)

set.seed(122)

spl=sample.split(ca,SplitRatio = 0.7)

train=subset(ca,spl==T)

test=subset(ca,spl==F)

dim(train)

dim(test)

library(DMwR)

balanced.gd <- SMOTE(Transport~.,train, perc.over = 150, k = 5, perc.under = 400)

table(balanced.gd$Transport)

mr=train(Transport~.,data =balanced.gd,method="rf",trcontrol=ctrl,family=binomial)

pre=predict(mr,newdata=test,type="raw")

confusionMatrix(test$Transport,data=pre,positive="1")

varImp(mr)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 143 2

1 2 10

Accuracy : 0.9745

95% CI : (0.9361, 0.993)

No Information Rate : 0.9236

P-Value [Acc > NIR] : 0.006059

Sensitivity : 0.83333

Specificity : 0.98621

Pos Pred Value : 0.83333

Neg Pred Value : 0.98621

Variable Importance > 30

|  |
| --- |
|  |
| **Variable** | **Overall** |  |  |  |
| Salary | 100.0000000 |  |  |  |
| Age | 98.5073356 |  |  |  |
| Distance | 64.2084316 |  |  |  |

Bagging using Smote:

*Code:*

library(ipred)

library(rpart)

str(ca)

str(ca)

set.seed(100)

spl=sample.split(ca,SplitRatio = 0.7)

train=subset(ca,spl==T)

test=subset(ca,spl==F)

dim(train)

dim(test)

balanced.gd <- SMOTE(Transport~.,train, perc.over = 100, k = 5, perc.under = 400)

table(balanced.gd$Transport)

bag=bagging(Transport~.,data=balanced.gd,control=rpart.control(maxdepth=5, minsplit=4))

pred=predict(bag,newdata=test,type="class")

confusionMatrix(test$Transport,data=pred,positive = "1")

varImp(bag)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 141 0

1 2 13

Accuracy : 0.9872

95% CI : (0.9545, 0.9984)

No Information Rate : 0.9167

P-Value [Acc > NIR] : 0.0001466

Sensitivity : 1.00000

Specificity : 0.98601

Pos Pred Value : 0.86667

Neg Pred Value : 1.00000

Variable Importance > 40

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Variable** | **Overall** | | --- | --- | | Age | 50.9439398 | | Distance | 45.2798953 | | Salary | 52.1529437 | |  |  | |  |  | |
|  |  |  | |  |  |
|  |  |  | |  |  |
|  |  |  | |  |  |
|  | | |

Adaptive Boosting without Smote:

*Code:*

ca=read.csv("cars.csv",header = T)

ca=na.omit(ca)

ca$Transport=ifelse(ca$Transport==2,0,ifelse(ca$Transport=="P",0,1))

str(ca)

summary(ca)

ca[,1:9]=lapply(ca[,1:9],as.integer)

nor <-function(x) { (x -min(x))/(max(x)-min(x))}

ca\_norm <- as.data.frame(lapply(ca[,c(1:9)], nor))

library(caTools)

set.seed(1547)

spl=sample.split(ca\_norm,SplitRatio = 0.7)

train=subset(ca\_norm,spl==T)

test=subset(ca\_norm,spl==F)

dim(train)

dim(test)

library(gbm)

gbm.fit <- gbm(

formula = Transport ~ .,

distribution = "bernoulli",

data = train,

n.trees = 100,

interaction.depth = 1,

shrinkage = 0.01,

cv.folds = 5,

n.cores = NULL,

verbose = FALSE

)

test$pred=predict(gbm.fit,test,type="response")

test$pred<- ifelse(test$pred<0.5,0,1)

table(test$Transport,test$pred)

confusionMatrix(data=factor(test$pred),

reference=factor(test$Transport),

positive='1')

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 127 4

1 0 8

Accuracy : 0.9712

95% CI : (0.928, 0.9921)

No Information Rate : 0.9137

P-Value [Acc > NIR] : 0.005872

Sensitivity : 0.66667

Specificity : 1.00000

Pos Pred Value : 1.00000

Neg Pred Value : 0.96947

XgBoost with Smote:

*Code:*

library(xgboost)

ca=read.csv("cars.csv",header = T)

str(ca)

summary(ca)

ca=na.omit(ca)

summary(ca)

ca$Transport=ifelse(ca$Transport==2,0,ifelse(ca$Transport=="P",0,1))

str(ca)

ca$Engineer=as.factor(ca$Engineer)

ca$MBA=as.factor(ca$MBA)

ca$Gender=as.factor(ca$Gender)

ca$Transport=as.factor(ca$Transport)

str(ca)

summary(ca)

ca$license=as.factor(ca$license)

summary(ca)

ca=ca[,-5]

ca[,1:8]=lapply(ca[,1:8],as.integer)

nor <-function(x) { (x -min(x))/(max(x)-min(x))}

ca\_norm <- as.data.frame(lapply(ca[,c(1:8)], nor))

set.seed(1478)

spl=sample.split(ca\_norm,SplitRatio = 0.7)

train=subset(ca\_norm,spl==T)

test=subset(ca\_norm,spl==F)

dim(train)

dim(test)

str(train)

train$Transport=as.factor(train$Transport)

library(DMwR)

balanced.gd <- SMOTE(Transport~.,train, perc.over = 150, k = 5, perc.under = 400)

table(balanced.gd$Transport)

str(balanced.gd)

balanced.gd[,1:8]=lapply(balanced.gd[,1:8],as.integer)

prop.table(table(balanced.gd$Transport))

str(balanced.gd)

balanced.gd$Transport=ifelse(balanced.gd$Transport==2,1,0)

table(balanced.gd$Transport)

smote\_features\_train<-as.matrix(balanced.gd[,1:7])

smote\_label\_train<-as.matrix(balanced.gd[,8])

smote\_features\_test<-as.matrix(test[,1:7])

smote.xgb.fit <- xgboost(

data = smote\_features\_train,

label = smote\_label\_train,

eta = 0.01,

max\_depth = 3,

min\_child\_weight = 3,

nrounds = 50,

nfold = 5,

objective = "binary:logistic",

verbose = 0,

early\_stopping\_rounds = 10

)

test$pred<- predict(smote.xgb.fit,smote\_features\_test,type="response")

test$pred<- ifelse(test$pred<0.5,0,1)

table(test$Transport,test$pred)

table(test$Transport)

confusionMatrix(data=factor(test$pred),

reference=factor(test$Transport),

positive='1')

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 124 2

1 21 10

Accuracy : 0.8535

95% CI : (0.7883, 0.9048)

No Information Rate : 0.9236

P-Value [Acc > NIR] : 0.9991050

Sensitivity : 0.83333

Specificity : 0.85517

Pos Pred Value : 0.32258

Neg Pred Value : 0.98413

Summary and actionable notes:

Based on the analysis done and taking sensitivity into account it can be concluded that Distance, Salary and Age are important factors that would help us determine if an employee is going to choose car as a mode to commute, In terms of model performance having compared the various model built using the smote technique bagging seems to give 100% sensitivity along with highest accuracy. At the initial data exploration, it showed that, Male engineer employees dominate car usage, also employees above 30years in age and where distance is above 12.5 prefer to use car, average Salary for an employee who prefers car as a mode of transport has a salary higher to that of employees who do not use car to commute which is intuitive.