Employee\_Atrittion\_final\_.R

Shivani Malandkar & Sojwal Shetye

Tue Dec 11 04:09:06 2018

#Final Project  
#Shivani Malandkar & Sojwal Shetye  
  
#Cleaning   
#No NAs in dataset  
EmployeeAttrition=read.csv("C:\\Users\\Shiva\\Desktop\\employee\_attrition\_prediction-master\\HR-Employee-Attrition.csv")  
summary(EmployeeAttrition)

## ï..Age Attrition BusinessTravel DailyRate   
## Min. :18.00 No :1233 Non-Travel : 150 Min. : 102.0   
## 1st Qu.:30.00 Yes: 237 Travel\_Frequently: 277 1st Qu.: 465.0   
## Median :36.00 Travel\_Rarely :1043 Median : 802.0   
## Mean :36.92 Mean : 802.5   
## 3rd Qu.:43.00 3rd Qu.:1157.0   
## Max. :60.00 Max. :1499.0   
##   
## Department DistanceFromHome Education   
## Human Resources : 63 Min. : 1.000 Min. :1.000   
## Research & Development:961 1st Qu.: 2.000 1st Qu.:2.000   
## Sales :446 Median : 7.000 Median :3.000   
## Mean : 9.193 Mean :2.913   
## 3rd Qu.:14.000 3rd Qu.:4.000   
## Max. :29.000 Max. :5.000   
##   
## EducationField EmployeeCount EmployeeNumber   
## Human Resources : 27 Min. :1 Min. : 1.0   
## Life Sciences :606 1st Qu.:1 1st Qu.: 491.2   
## Marketing :159 Median :1 Median :1020.5   
## Medical :464 Mean :1 Mean :1024.9   
## Other : 82 3rd Qu.:1 3rd Qu.:1555.8   
## Technical Degree:132 Max. :1 Max. :2068.0   
##   
## EnvironmentSatisfaction Gender HourlyRate JobInvolvement  
## Min. :1.000 Female:588 Min. : 30.00 Min. :1.00   
## 1st Qu.:2.000 Male :882 1st Qu.: 48.00 1st Qu.:2.00   
## Median :3.000 Median : 66.00 Median :3.00   
## Mean :2.722 Mean : 65.89 Mean :2.73   
## 3rd Qu.:4.000 3rd Qu.: 83.75 3rd Qu.:3.00   
## Max. :4.000 Max. :100.00 Max. :4.00   
##   
## JobLevel JobRole JobSatisfaction  
## Min. :1.000 Sales Executive :326 Min. :1.000   
## 1st Qu.:1.000 Research Scientist :292 1st Qu.:2.000   
## Median :2.000 Laboratory Technician :259 Median :3.000   
## Mean :2.064 Manufacturing Director :145 Mean :2.729   
## 3rd Qu.:3.000 Healthcare Representative:131 3rd Qu.:4.000   
## Max. :5.000 Manager :102 Max. :4.000   
## (Other) :215   
## MaritalStatus MonthlyIncome MonthlyRate NumCompaniesWorked  
## Divorced:327 Min. : 1009 Min. : 2094 Min. :0.000   
## Married :673 1st Qu.: 2911 1st Qu.: 8047 1st Qu.:1.000   
## Single :470 Median : 4919 Median :14236 Median :2.000   
## Mean : 6503 Mean :14313 Mean :2.693   
## 3rd Qu.: 8379 3rd Qu.:20462 3rd Qu.:4.000   
## Max. :19999 Max. :26999 Max. :9.000   
##   
## Over18 OverTime PercentSalaryHike PerformanceRating  
## Y:1470 No :1054 Min. :11.00 Min. :3.000   
## Yes: 416 1st Qu.:12.00 1st Qu.:3.000   
## Median :14.00 Median :3.000   
## Mean :15.21 Mean :3.154   
## 3rd Qu.:18.00 3rd Qu.:3.000   
## Max. :25.00 Max. :4.000   
##   
## RelationshipSatisfaction StandardHours StockOptionLevel TotalWorkingYears  
## Min. :1.000 Min. :80 Min. :0.0000 Min. : 0.00   
## 1st Qu.:2.000 1st Qu.:80 1st Qu.:0.0000 1st Qu.: 6.00   
## Median :3.000 Median :80 Median :1.0000 Median :10.00   
## Mean :2.712 Mean :80 Mean :0.7939 Mean :11.28   
## 3rd Qu.:4.000 3rd Qu.:80 3rd Qu.:1.0000 3rd Qu.:15.00   
## Max. :4.000 Max. :80 Max. :3.0000 Max. :40.00   
##   
## TrainingTimesLastYear WorkLifeBalance YearsAtCompany YearsInCurrentRole  
## Min. :0.000 Min. :1.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.:2.000 1st Qu.:2.000 1st Qu.: 3.000 1st Qu.: 2.000   
## Median :3.000 Median :3.000 Median : 5.000 Median : 3.000   
## Mean :2.799 Mean :2.761 Mean : 7.008 Mean : 4.229   
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.: 9.000 3rd Qu.: 7.000   
## Max. :6.000 Max. :4.000 Max. :40.000 Max. :18.000   
##   
## YearsSinceLastPromotion YearsWithCurrManager  
## Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: 2.000   
## Median : 1.000 Median : 3.000   
## Mean : 2.188 Mean : 4.123   
## 3rd Qu.: 3.000 3rd Qu.: 7.000   
## Max. :15.000 Max. :17.000   
##

#After looking into Summary ,We can see that there are no NA'S and Some variables   
#such as Education,JobInvolvement..etc which are factors are   
##stored as integers ,so Converting these continuos variables to Categorical data##  
  
names <- c('WorkLifeBalance' ,'StockOptionLevel','PerformanceRating','JobSatisfaction',  
 'RelationshipSatisfaction','JobLevel','JobInvolvement','EnvironmentSatisfaction','Education')  
EmployeeAttrition[,names] <- lapply(EmployeeAttrition[,names],factor)  
head(EmployeeAttrition)

## ï..Age Attrition BusinessTravel DailyRate Department  
## 1 41 Yes Travel\_Rarely 1102 Sales  
## 2 49 No Travel\_Frequently 279 Research & Development  
## 3 37 Yes Travel\_Rarely 1373 Research & Development  
## 4 33 No Travel\_Frequently 1392 Research & Development  
## 5 27 No Travel\_Rarely 591 Research & Development  
## 6 32 No Travel\_Frequently 1005 Research & Development  
## DistanceFromHome Education EducationField EmployeeCount EmployeeNumber  
## 1 1 2 Life Sciences 1 1  
## 2 8 1 Life Sciences 1 2  
## 3 2 2 Other 1 4  
## 4 3 4 Life Sciences 1 5  
## 5 2 1 Medical 1 7  
## 6 2 2 Life Sciences 1 8  
## EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel  
## 1 2 Female 94 3 2  
## 2 3 Male 61 2 2  
## 3 4 Male 92 2 1  
## 4 4 Female 56 3 1  
## 5 1 Male 40 3 1  
## 6 4 Male 79 3 1  
## JobRole JobSatisfaction MaritalStatus MonthlyIncome  
## 1 Sales Executive 4 Single 5993  
## 2 Research Scientist 2 Married 5130  
## 3 Laboratory Technician 3 Single 2090  
## 4 Research Scientist 3 Married 2909  
## 5 Laboratory Technician 2 Married 3468  
## 6 Laboratory Technician 4 Single 3068  
## MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike  
## 1 19479 8 Y Yes 11  
## 2 24907 1 Y No 23  
## 3 2396 6 Y Yes 15  
## 4 23159 1 Y Yes 11  
## 5 16632 9 Y No 12  
## 6 11864 0 Y No 13  
## PerformanceRating RelationshipSatisfaction StandardHours  
## 1 3 1 80  
## 2 4 4 80  
## 3 3 2 80  
## 4 3 3 80  
## 5 3 4 80  
## 6 3 3 80  
## StockOptionLevel TotalWorkingYears TrainingTimesLastYear WorkLifeBalance  
## 1 0 8 0 1  
## 2 1 10 3 3  
## 3 0 7 3 3  
## 4 0 8 3 3  
## 5 1 6 3 3  
## 6 0 8 2 2  
## YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion  
## 1 6 4 0  
## 2 10 7 1  
## 3 0 0 0  
## 4 8 7 3  
## 5 2 2 2  
## 6 7 7 3  
## YearsWithCurrManager  
## 1 5  
## 2 7  
## 3 0  
## 4 0  
## 5 2  
## 6 6

#Checking for missing value and removing non value attribute  
apply(is.na(EmployeeAttrition),2,sum)

## ï..Age Attrition BusinessTravel   
## 0 0 0   
## DailyRate Department DistanceFromHome   
## 0 0 0   
## Education EducationField EmployeeCount   
## 0 0 0   
## EmployeeNumber EnvironmentSatisfaction Gender   
## 0 0 0   
## HourlyRate JobInvolvement JobLevel   
## 0 0 0   
## JobRole JobSatisfaction MaritalStatus   
## 0 0 0   
## MonthlyIncome MonthlyRate NumCompaniesWorked   
## 0 0 0   
## Over18 OverTime PercentSalaryHike   
## 0 0 0   
## PerformanceRating RelationshipSatisfaction StandardHours   
## 0 0 0   
## StockOptionLevel TotalWorkingYears TrainingTimesLastYear   
## 0 0 0   
## WorkLifeBalance YearsAtCompany YearsInCurrentRole   
## 0 0 0   
## YearsSinceLastPromotion YearsWithCurrManager   
## 0 0

EmployeeAttrition$EmployeeNumber=NULL  
EmployeeAttrition$StandardHours=NULL  
EmployeeAttrition$Over18=NULL  
EmployeeAttrition$EmployeeCount=NULL  
cat("Data Set has",dim(EmployeeAttrition)[1],"Rows and",dim(EmployeeAttrition)[2],"Columns")

## Data Set has 1470 Rows and 31 Columns

sum(is.na(duplicated(EmployeeAttrition)))#No missing values and no duplicate

## [1] 0

#Removing columns which have same value for all  
cleaned\_data=EmployeeAttrition[,-c(9,10,22,27)]  
#replacing all blank cells with NA  
cleaned\_data[cleaned\_data==""]=NA  
#removing all rows with any blank cell  
cleaned\_data=cleaned\_data[complete.cases(cleaned\_data), ]  
str(cleaned\_data)

## 'data.frame': 1470 obs. of 27 variables:  
## $ ï..Age : int 41 49 37 33 27 32 59 30 38 36 ...  
## $ Attrition : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 1 ...  
## $ BusinessTravel : Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 3 2 3 2 3 2 3 3 2 3 ...  
## $ DailyRate : int 1102 279 1373 1392 591 1005 1324 1358 216 1299 ...  
## $ Department : Factor w/ 3 levels "Human Resources",..: 3 2 2 2 2 2 2 2 2 2 ...  
## $ DistanceFromHome : int 1 8 2 3 2 2 3 24 23 27 ...  
## $ Education : Factor w/ 5 levels "1","2","3","4",..: 2 1 2 4 1 2 3 1 3 3 ...  
## $ EducationField : Factor w/ 6 levels "Human Resources",..: 2 2 5 2 4 2 4 2 2 4 ...  
## $ HourlyRate : int 94 61 92 56 40 79 81 67 44 94 ...  
## $ JobInvolvement : Factor w/ 4 levels "1","2","3","4": 3 2 2 3 3 3 4 3 2 3 ...  
## $ JobLevel : Factor w/ 5 levels "1","2","3","4",..: 2 2 1 1 1 1 1 1 3 2 ...  
## $ JobRole : Factor w/ 9 levels "Healthcare Representative",..: 8 7 3 7 3 3 3 3 5 1 ...  
## $ JobSatisfaction : Factor w/ 4 levels "1","2","3","4": 4 2 3 3 2 4 1 3 3 3 ...  
## $ MaritalStatus : Factor w/ 3 levels "Divorced","Married",..: 3 2 3 2 2 3 2 1 3 2 ...  
## $ MonthlyIncome : int 5993 5130 2090 2909 3468 3068 2670 2693 9526 5237 ...  
## $ MonthlyRate : int 19479 24907 2396 23159 16632 11864 9964 13335 8787 16577 ...  
## $ NumCompaniesWorked : int 8 1 6 1 9 0 4 1 0 6 ...  
## $ OverTime : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 2 1 1 1 ...  
## $ PercentSalaryHike : int 11 23 15 11 12 13 20 22 21 13 ...  
## $ RelationshipSatisfaction: Factor w/ 4 levels "1","2","3","4": 1 4 2 3 4 3 1 2 2 2 ...  
## $ StockOptionLevel : Factor w/ 4 levels "0","1","2","3": 1 2 1 1 2 1 4 2 1 3 ...  
## $ TotalWorkingYears : int 8 10 7 8 6 8 12 1 10 17 ...  
## $ TrainingTimesLastYear : int 0 3 3 3 3 2 3 2 2 3 ...  
## $ YearsAtCompany : int 6 10 0 8 2 7 1 1 9 7 ...  
## $ YearsInCurrentRole : int 4 7 0 7 2 7 0 0 7 7 ...  
## $ YearsSinceLastPromotion : int 0 1 0 3 2 3 0 0 1 7 ...  
## $ YearsWithCurrManager : int 5 7 0 0 2 6 0 0 8 7 ...

library(ggplot2)  
  
library(tidyr)  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(corrplot)

## corrplot 0.84 loaded

library(miscset)

##   
## Attaching package: 'miscset'

## The following object is masked from 'package:dplyr':  
##   
## collapse

library(purrr)  
require(gridExtra)

## Loading required package: gridExtra

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(caTools)  
  
  
library(e1071)  
  
library(glmnet)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':  
##   
## expand

## Loading required package: foreach

##   
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':  
##   
## accumulate, when

## Loaded glmnet 2.0-16

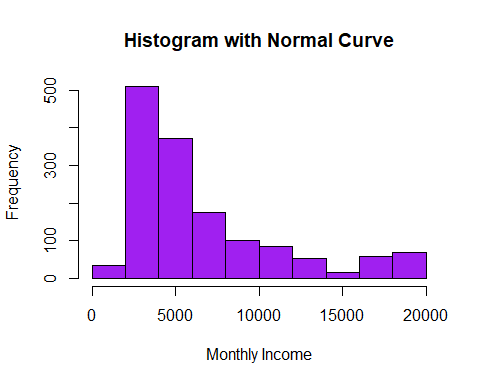
#Exploring the data   
dim(cleaned\_data)

## [1] 1470 27

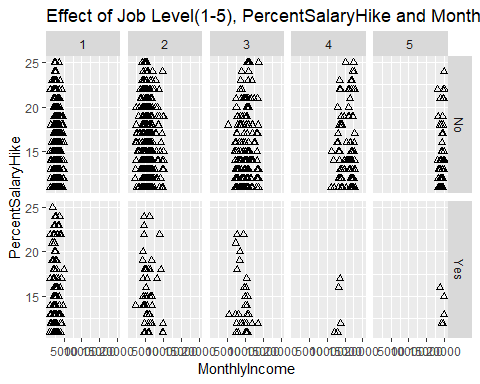
#performing cross validation, its important to maintain this turnover ratio  
attrition<-as.factor(cleaned\_data$Attrition)  
summary(attrition)#We can see that 237 employees have been retained whereas 1233 employees have been let go of.

## No Yes   
## 1233 237

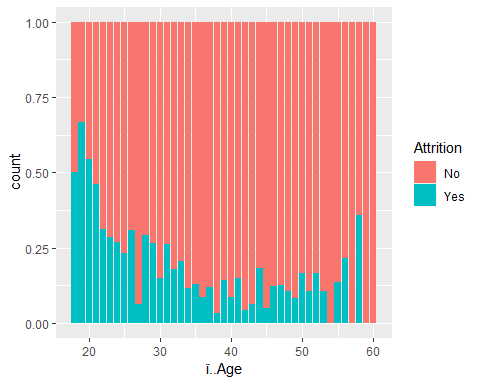
#Exploratory data plots   
# Histogram with normal curve for monthly income  
# Histogram  
histogram.curve <- hist(cleaned\_data$MonthlyIncome, breaks = 10, col = "purple", xlab = "Monthly Income", main = "Histogram with Normal Curve")  
# Adding normal curve to the histogram  
xfit <- seq(min(cleaned\_data[,19]), max(cleaned\_data[,19]), length=40)  
yfit <- dnorm(xfit, mean=mean(cleaned\_data[,19]), sd=sd((cleaned\_data[,19])))  
yfit <- yfit\*diff(histogram.curve$mids[1:2])\*length(cleaned\_data$MonthlyIncome)  
lines(xfit, yfit, col ="black", lwd=2)



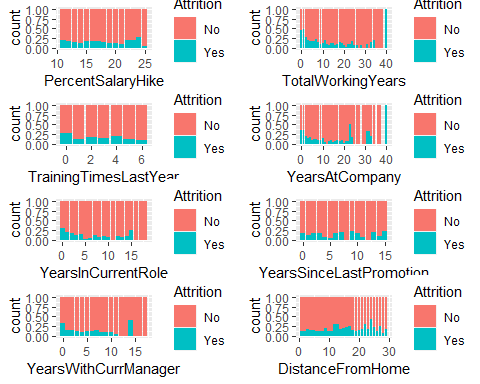
# Plot showing relationships between employees leaving the company with respect to monthly income, percent salary hike and job level  
library(ggplot2)  
pl <- ggplot(cleaned\_data, aes(x=MonthlyIncome, y=PercentSalaryHike)) + geom\_point(shape=2)+ ggtitle("Effect of Job Level(1-5), PercentSalaryHike and MonthlyIncome on Attrition(Y/N)")  
pl + facet\_grid(Attrition ~ JobLevel)



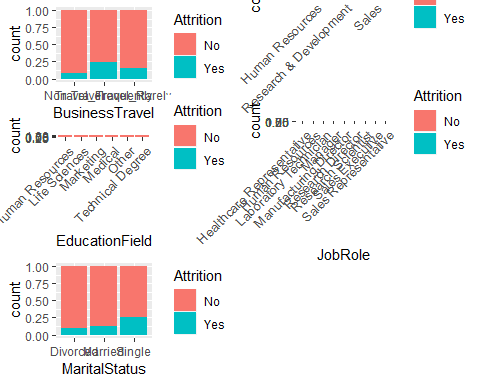
ggplot(cleaned\_data,aes(x = ï..Age ,fill = Attrition)) + geom\_bar(position = "fill")



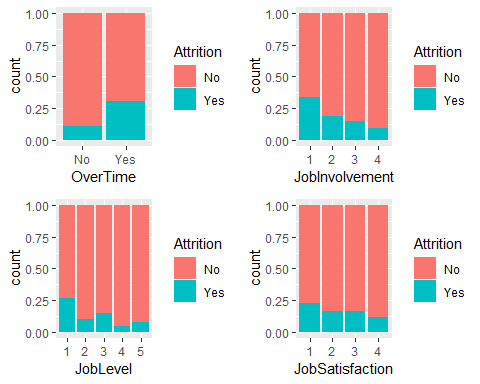
co2 <- ggplot(cleaned\_data,aes(x = PercentSalaryHike,fill = Attrition)) +   
 geom\_bar(position = "fill")   
  
co3 <- ggplot(cleaned\_data,aes(x = TotalWorkingYears,fill = Attrition)) +   
 geom\_bar(position = "fill")  
  
co4 <- ggplot(cleaned\_data,aes(x = TrainingTimesLastYear,fill = Attrition)) +   
 geom\_bar(position = "fill")  
  
co5 <- ggplot(cleaned\_data,aes(x = YearsAtCompany,fill = Attrition)) +   
 geom\_bar(position = "fill")  
  
co6 <- ggplot(cleaned\_data,aes(x = YearsInCurrentRole,fill = Attrition)) +   
 geom\_bar(position = "fill")  
  
co7 <- ggplot(cleaned\_data,aes(x = YearsSinceLastPromotion,fill = Attrition)) +   
 geom\_bar(position = "fill")  
  
co8 <- ggplot(cleaned\_data,aes(x = YearsWithCurrManager,fill = Attrition)) +   
 geom\_bar(position = "fill")  
  
co9 <- ggplot(cleaned\_data,aes(x = DistanceFromHome,fill = Attrition)) +   
 geom\_bar(position = "fill")  
  
co10 <- ggplot()  
  
grid.arrange(co2,co3,co4,co5,co6,co7,co8,co9,ncol=2)



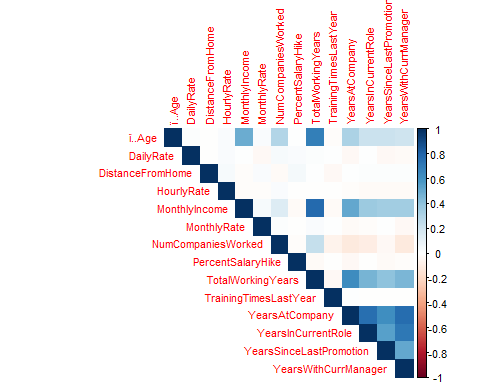
pc1 <- ggplot(cleaned\_data,aes(x = BusinessTravel,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill")   
  
  
pc2 <- ggplot(cleaned\_data,aes(x = Department,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill") +   
 theme(axis.text.x = element\_text(size = 10, angle = 45,hjust = 1,vjust = 1))  
  
pc3 <- ggplot(cleaned\_data,aes(x = EducationField,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill") +  
 theme(axis.text.x = element\_text(size = 10, angle = 45,hjust = 1,vjust = 1))  
  
  
  
pc5 <- ggplot(cleaned\_data,aes(x = JobRole,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill") +  
 theme(axis.text.x = element\_text(size = 10, angle = 45,hjust = 1,vjust = 1))  
  
pc6 <- ggplot(cleaned\_data,aes(x = MaritalStatus,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill")  
  
pc7 <- ggplot(cleaned\_data,aes(x = OverTime,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill")  
  
  
pc9 <- ggplot(cleaned\_data,aes(x = JobInvolvement,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill")  
  
pc10 <- ggplot(cleaned\_data,aes(x = JobLevel,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill")  
  
pc11 <- ggplot(cleaned\_data,aes(x = JobSatisfaction,..count..)) +   
 geom\_bar(aes(fill = Attrition),position = "fill")  
  
grid.arrange(pc1,pc2,pc3,pc5,pc6,ncol =2)



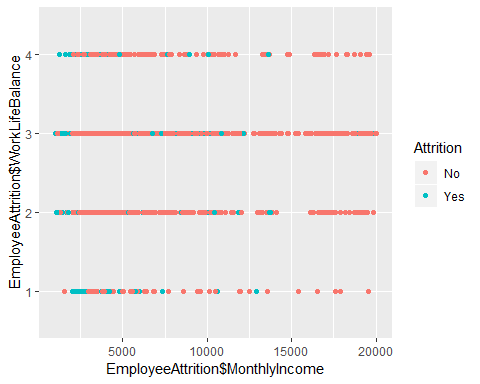
####From these graphs, we can Infer that Education and Performance Rating,Training times since  
#last year doesnot impact on Employee Attrition  
  
grid.arrange(pc7,pc9,pc10,pc11,ncol =2)



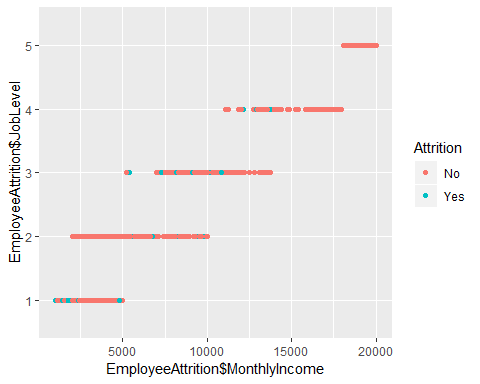
#From this ,We can infer that employees who travel frequently will leave  
#company when company when compared to Non-Travellers. More than 25% of Employees who work Overtime leave the company  
  
#So, Education Field, Gender, Department ,Trainingtimessincelastyear, Performance rating and   
##Education Field are not strong predictors and I will not be including these variables.  
  
  
#Checking if there is Multi-Co linearity - High Correlation between   
#ndependent variables  
library(corrplot)  
empn <- which(sapply(cleaned\_data,is.numeric))  
corrplot(cor(cleaned\_data[empn]),type = "upper",method='color',tl.cex = .7,cl.cex = .7,number.cex = 0.7)



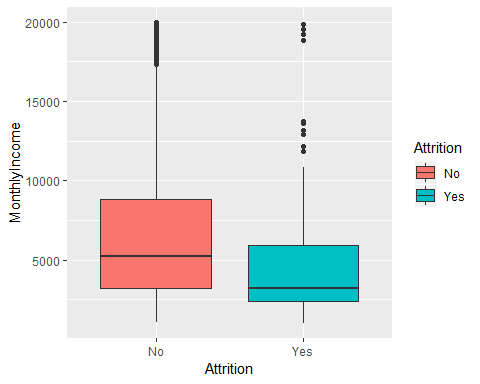
#scatter plot between monthly income, work life balance and attrition  
ggplot(EmployeeAttrition,aes(EmployeeAttrition$MonthlyIncome,EmployeeAttrition$WorkLifeBalance, color=Attrition))+geom\_point()



#scatter plot between monthly income, JobLevel and attrition  
ggplot(EmployeeAttrition,aes(EmployeeAttrition$MonthlyIncome,EmployeeAttrition$JobLevel, color=Attrition))+geom\_point()



#boxplot between monthly income and attrition  
ggplot(cleaned\_data,aes(Attrition,MonthlyIncome,fill=Attrition))+geom\_boxplot()



#Logistic Regression  
#We split the data into two chunks: training and testing set. The training set will be used to fit our model.  
  
#install.packages('caret')  
#load package  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

trainIndex = createDataPartition(cleaned\_data$Attrition,  
 p=0.7, list=FALSE,times=1)  
  
train = cleaned\_data[trainIndex,]  
test = cleaned\_data[-trainIndex,]  
  
model <- glm(Attrition ~.,family=binomial(link='logit'),data=train)  
summary(model)

##   
## Call:  
## glm(formula = Attrition ~ ., family = binomial(link = "logit"),   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9126 -0.4269 -0.1948 -0.0514 3.5904   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.261e+01 6.619e+02 -0.019 0.984795  
## ï..Age -5.729e-02 1.834e-02 -3.124 0.001786  
## BusinessTravelTravel\_Frequently 1.232e+00 4.865e-01 2.532 0.011341  
## BusinessTravelTravel\_Rarely 2.529e-01 4.478e-01 0.565 0.572191  
## DailyRate -2.902e-04 2.800e-04 -1.037 0.299926  
## DepartmentResearch & Development 1.606e+01 6.619e+02 0.024 0.980640  
## DepartmentSales 1.459e+01 6.619e+02 0.022 0.982419  
## DistanceFromHome 6.011e-02 1.371e-02 4.384 1.16e-05  
## Education2 8.145e-01 4.304e-01 1.892 0.058451  
## Education3 4.570e-01 3.794e-01 1.205 0.228380  
## Education4 4.247e-01 4.161e-01 1.021 0.307394  
## Education5 4.888e-02 8.208e-01 0.060 0.952513  
## EducationFieldLife Sciences -1.753e+00 1.087e+00 -1.613 0.106667  
## EducationFieldMarketing -1.201e+00 1.144e+00 -1.050 0.293899  
## EducationFieldMedical -1.799e+00 1.088e+00 -1.653 0.098301  
## EducationFieldOther -1.939e+00 1.163e+00 -1.667 0.095502  
## EducationFieldTechnical Degree -7.081e-01 1.112e+00 -0.637 0.524086  
## HourlyRate 1.123e-03 5.737e-03 0.196 0.844828  
## JobInvolvement2 -1.230e+00 4.256e-01 -2.890 0.003854  
## JobInvolvement3 -1.612e+00 4.029e-01 -4.001 6.31e-05  
## JobInvolvement4 -1.840e+00 5.605e-01 -3.283 0.001029  
## JobLevel2 -2.022e+00 5.882e-01 -3.437 0.000587  
## JobLevel3 -2.488e-01 8.827e-01 -0.282 0.778075  
## JobLevel4 -1.823e+00 1.491e+00 -1.223 0.221424  
## JobLevel5 8.995e-01 1.949e+00 0.462 0.644412  
## JobRoleHuman Resources 1.644e+01 6.619e+02 0.025 0.980182  
## JobRoleLaboratory Technician 7.020e-01 7.216e-01 0.973 0.330589  
## JobRoleManager 7.329e-01 1.138e+00 0.644 0.519523  
## JobRoleManufacturing Director 2.809e-01 7.002e-01 0.401 0.688278  
## JobRoleResearch Director -1.413e+00 1.583e+00 -0.892 0.372151  
## JobRoleResearch Scientist -4.255e-01 7.453e-01 -0.571 0.568113  
## JobRoleSales Executive 3.032e+00 1.446e+00 2.097 0.036018  
## JobRoleSales Representative 2.286e+00 1.548e+00 1.477 0.139785  
## JobSatisfaction2 -7.109e-01 3.532e-01 -2.013 0.044123  
## JobSatisfaction3 -4.475e-01 3.029e-01 -1.477 0.139629  
## JobSatisfaction4 -1.393e+00 3.442e-01 -4.048 5.17e-05  
## MaritalStatusMarried 3.674e-01 3.495e-01 1.051 0.293075  
## MaritalStatusSingle 6.768e-01 5.043e-01 1.342 0.179623  
## MonthlyIncome -7.064e-05 1.151e-04 -0.614 0.539414  
## MonthlyRate 1.506e-05 1.578e-05 0.954 0.339928  
## NumCompaniesWorked 2.403e-01 4.872e-02 4.932 8.15e-07  
## OverTimeYes 2.456e+00 2.584e-01 9.504 < 2e-16  
## PercentSalaryHike -2.753e-03 3.100e-02 -0.089 0.929232  
## RelationshipSatisfaction2 -9.808e-01 3.492e-01 -2.808 0.004978  
## RelationshipSatisfaction3 -1.439e+00 3.228e-01 -4.457 8.32e-06  
## RelationshipSatisfaction4 -1.439e+00 3.220e-01 -4.468 7.89e-06  
## StockOptionLevel1 -1.140e+00 3.970e-01 -2.871 0.004092  
## StockOptionLevel2 -1.042e+00 5.549e-01 -1.879 0.060285  
## StockOptionLevel3 -2.177e-01 5.991e-01 -0.363 0.716282  
## TotalWorkingYears -1.873e-02 3.579e-02 -0.523 0.600780  
## TrainingTimesLastYear -2.565e-01 9.353e-02 -2.743 0.006097  
## YearsAtCompany 1.373e-01 5.266e-02 2.608 0.009119  
## YearsInCurrentRole -1.964e-01 6.271e-02 -3.132 0.001736  
## YearsSinceLastPromotion 9.630e-02 5.455e-02 1.765 0.077503  
## YearsWithCurrManager -1.509e-01 6.117e-02 -2.466 0.013644  
##   
## (Intercept)   
## ï..Age \*\*   
## BusinessTravelTravel\_Frequently \*   
## BusinessTravelTravel\_Rarely   
## DailyRate   
## DepartmentResearch & Development   
## DepartmentSales   
## DistanceFromHome \*\*\*  
## Education2 .   
## Education3   
## Education4   
## Education5   
## EducationFieldLife Sciences   
## EducationFieldMarketing   
## EducationFieldMedical .   
## EducationFieldOther .   
## EducationFieldTechnical Degree   
## HourlyRate   
## JobInvolvement2 \*\*   
## JobInvolvement3 \*\*\*  
## JobInvolvement4 \*\*   
## JobLevel2 \*\*\*  
## JobLevel3   
## JobLevel4   
## JobLevel5   
## JobRoleHuman Resources   
## JobRoleLaboratory Technician   
## JobRoleManager   
## JobRoleManufacturing Director   
## JobRoleResearch Director   
## JobRoleResearch Scientist   
## JobRoleSales Executive \*   
## JobRoleSales Representative   
## JobSatisfaction2 \*   
## JobSatisfaction3   
## JobSatisfaction4 \*\*\*  
## MaritalStatusMarried   
## MaritalStatusSingle   
## MonthlyIncome   
## MonthlyRate   
## NumCompaniesWorked \*\*\*  
## OverTimeYes \*\*\*  
## PercentSalaryHike   
## RelationshipSatisfaction2 \*\*   
## RelationshipSatisfaction3 \*\*\*  
## RelationshipSatisfaction4 \*\*\*  
## StockOptionLevel1 \*\*   
## StockOptionLevel2 .   
## StockOptionLevel3   
## TotalWorkingYears   
## TrainingTimesLastYear \*\*   
## YearsAtCompany \*\*   
## YearsInCurrentRole \*\*   
## YearsSinceLastPromotion .   
## YearsWithCurrManager \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 909.69 on 1029 degrees of freedom  
## Residual deviance: 540.41 on 975 degrees of freedom  
## AIC: 650.41  
##   
## Number of Fisher Scoring iterations: 15

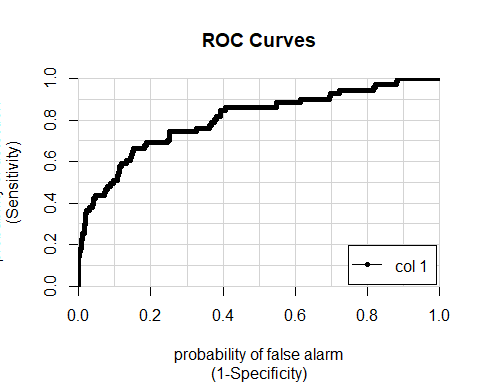
# Predicting the results using testing dataset  
LR\_model.predict <- predict(model, test, type = "response")  
length(LR\_model.predict)

## [1] 440

length(test$Attrition)

## [1] 440

library(caTools)  
colAUC(LR\_model.predict,test$Attrition, plotROC=TRUE)



## [,1]  
## No vs. Yes 0.8055269

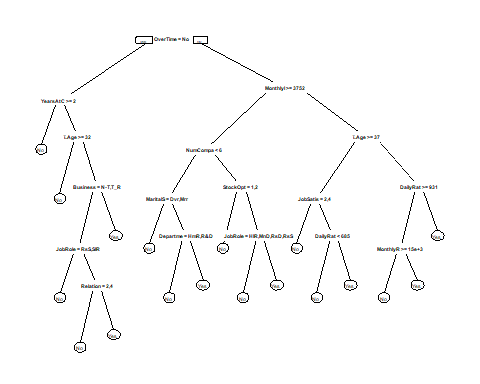
#Column under ROC  
  
#Make use of the confusion matrix   
conf\_mat = table(LR\_model.predict,test$Attrition)  
  
#To evaluate this model, we will use 10 repeats of 10-fold cross-validation and use the 100 holdout samples to evaluate the overall accuracy of the model.  
set.seed(123)   
  
library(devtools)  
  
library(caret)  
  
# Define train control for k fold cross validation  
train\_control <- trainControl(method="cv", number=10)  
# Fit Naive Bayes Model  
model2 <- train(Attrition ~., data=cleaned\_data, trControl=train\_control, method="glm",family=binomial())  
# Summarise Results  
summary(model2)

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6752 -0.4701 -0.2239 -0.0696 3.3655   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.227e+01 6.221e+02 -0.020 0.984261  
## ï..Age -3.197e-02 1.389e-02 -2.302 0.021322  
## BusinessTravelTravel\_Frequently 1.893e+00 4.237e-01 4.467 7.94e-06  
## BusinessTravelTravel\_Rarely 9.106e-01 3.928e-01 2.318 0.020437  
## DailyRate -3.578e-04 2.238e-04 -1.599 0.109834  
## `DepartmentResearch & Development` 1.440e+01 6.221e+02 0.023 0.981537  
## DepartmentSales 1.372e+01 6.221e+02 0.022 0.982409  
## DistanceFromHome 4.774e-02 1.110e-02 4.301 1.70e-05  
## Education2 3.803e-01 3.378e-01 1.126 0.260242  
## Education3 3.276e-01 2.983e-01 1.098 0.272166  
## Education4 4.048e-01 3.231e-01 1.253 0.210357  
## Education5 1.603e-01 6.274e-01 0.255 0.798380  
## `EducationFieldLife Sciences` -1.023e+00 8.737e-01 -1.171 0.241668  
## EducationFieldMarketing -5.524e-01 9.171e-01 -0.602 0.546937  
## EducationFieldMedical -1.108e+00 8.736e-01 -1.268 0.204805  
## EducationFieldOther -1.237e+00 9.437e-01 -1.311 0.189905  
## `EducationFieldTechnical Degree` -3.008e-02 8.913e-01 -0.034 0.973078  
## HourlyRate 2.728e-03 4.553e-03 0.599 0.549026  
## JobInvolvement2 -1.296e+00 3.546e-01 -3.655 0.000257  
## JobInvolvement3 -1.588e+00 3.354e-01 -4.734 2.20e-06  
## JobInvolvement4 -2.184e+00 4.722e-01 -4.625 3.74e-06  
## JobLevel2 -1.563e+00 4.570e-01 -3.421 0.000624  
## JobLevel3 -3.139e-02 7.075e-01 -0.044 0.964610  
## JobLevel4 -1.083e+00 1.211e+00 -0.895 0.370821  
## JobLevel5 1.734e+00 1.605e+00 1.080 0.280122  
## `JobRoleHuman Resources` 1.463e+01 6.221e+02 0.024 0.981239  
## `JobRoleLaboratory Technician` 6.411e-01 5.812e-01 1.103 0.270021  
## JobRoleManager -1.693e-01 1.080e+00 -0.157 0.875388  
## `JobRoleManufacturing Director` 1.457e-01 5.484e-01 0.266 0.790418  
## `JobRoleResearch Director` -2.033e+00 1.182e+00 -1.719 0.085532  
## `JobRoleResearch Scientist` -4.554e-01 5.983e-01 -0.761 0.446556  
## `JobRoleSales Executive` 1.909e+00 1.204e+00 1.585 0.112865  
## `JobRoleSales Representative` 1.484e+00 1.285e+00 1.154 0.248306  
## JobSatisfaction2 -6.826e-01 2.790e-01 -2.447 0.014399  
## JobSatisfaction3 -6.138e-01 2.461e-01 -2.494 0.012616  
## JobSatisfaction4 -1.209e+00 2.618e-01 -4.618 3.88e-06  
## MaritalStatusMarried 2.664e-01 2.833e-01 0.940 0.346969  
## MaritalStatusSingle 3.858e-01 3.990e-01 0.967 0.333540  
## MonthlyIncome -1.073e-04 9.193e-05 -1.168 0.242978  
## MonthlyRate 6.241e-06 1.268e-05 0.492 0.622727  
## NumCompaniesWorked 2.039e-01 3.944e-02 5.169 2.35e-07  
## OverTimeYes 1.946e+00 1.961e-01 9.925 < 2e-16  
## PercentSalaryHike -6.061e-03 2.525e-02 -0.240 0.810315  
## RelationshipSatisfaction2 -8.947e-01 2.908e-01 -3.077 0.002091  
## RelationshipSatisfaction3 -9.475e-01 2.570e-01 -3.688 0.000226  
## RelationshipSatisfaction4 -9.498e-01 2.574e-01 -3.689 0.000225  
## StockOptionLevel1 -1.277e+00 3.132e-01 -4.077 4.57e-05  
## StockOptionLevel2 -1.221e+00 4.394e-01 -2.780 0.005440  
## StockOptionLevel3 -5.425e-01 4.642e-01 -1.169 0.242572  
## TotalWorkingYears -3.835e-02 2.897e-02 -1.324 0.185606  
## TrainingTimesLastYear -1.994e-01 7.438e-02 -2.681 0.007336  
## YearsAtCompany 1.050e-01 4.020e-02 2.612 0.009014  
## YearsInCurrentRole -1.545e-01 4.879e-02 -3.166 0.001544  
## YearsSinceLastPromotion 1.336e-01 4.245e-02 3.147 0.001649  
## YearsWithCurrManager -1.358e-01 4.824e-02 -2.815 0.004880  
##   
## (Intercept)   
## ï..Age \*   
## BusinessTravelTravel\_Frequently \*\*\*  
## BusinessTravelTravel\_Rarely \*   
## DailyRate   
## `DepartmentResearch & Development`   
## DepartmentSales   
## DistanceFromHome \*\*\*  
## Education2   
## Education3   
## Education4   
## Education5   
## `EducationFieldLife Sciences`   
## EducationFieldMarketing   
## EducationFieldMedical   
## EducationFieldOther   
## `EducationFieldTechnical Degree`   
## HourlyRate   
## JobInvolvement2 \*\*\*  
## JobInvolvement3 \*\*\*  
## JobInvolvement4 \*\*\*  
## JobLevel2 \*\*\*  
## JobLevel3   
## JobLevel4   
## JobLevel5   
## `JobRoleHuman Resources`   
## `JobRoleLaboratory Technician`   
## JobRoleManager   
## `JobRoleManufacturing Director`   
## `JobRoleResearch Director` .   
## `JobRoleResearch Scientist`   
## `JobRoleSales Executive`   
## `JobRoleSales Representative`   
## JobSatisfaction2 \*   
## JobSatisfaction3 \*   
## JobSatisfaction4 \*\*\*  
## MaritalStatusMarried   
## MaritalStatusSingle   
## MonthlyIncome   
## MonthlyRate   
## NumCompaniesWorked \*\*\*  
## OverTimeYes \*\*\*  
## PercentSalaryHike   
## RelationshipSatisfaction2 \*\*   
## RelationshipSatisfaction3 \*\*\*  
## RelationshipSatisfaction4 \*\*\*  
## StockOptionLevel1 \*\*\*  
## StockOptionLevel2 \*\*   
## StockOptionLevel3   
## TotalWorkingYears   
## TrainingTimesLastYear \*\*   
## YearsAtCompany \*\*   
## YearsInCurrentRole \*\*   
## YearsSinceLastPromotion \*\*   
## YearsWithCurrManager \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1298.58 on 1469 degrees of freedom  
## Residual deviance: 831.16 on 1415 degrees of freedom  
## AIC: 941.16  
##   
## Number of Fisher Scoring iterations: 15

print(model2)

## Generalized Linear Model   
##   
## 1470 samples  
## 26 predictor  
## 2 classes: 'No', 'Yes'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 1323, 1323, 1323, 1322, 1323, 1323, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.8680586 0.4339151

#decision tree  
# Load CART packages  
library(rpart)  
## Warning: package 'rpart' was built under R version 3.3.3  
library(rpart.plot)  
## Warning: package 'rpart.plot' was built under R version 3.3.3  
decisiontree = rpart(Attrition ~ ., data=train, method="class")  
  
  
#Plot the model  
  
prp(decisiontree)



#Predict on the test data  
prediction <- predict(decisiontree, newdata=test, type="class")  
#Baseline Accuracy vs CART Accuracy  
  
  
table(test$Attrition)

##   
## No Yes   
## 369 71

369/nrow(test)

## [1] 0.8386364

#Confusion matrix   
table(test$Attrition, prediction)

## prediction  
## No Yes  
## No 348 21  
## Yes 51 20

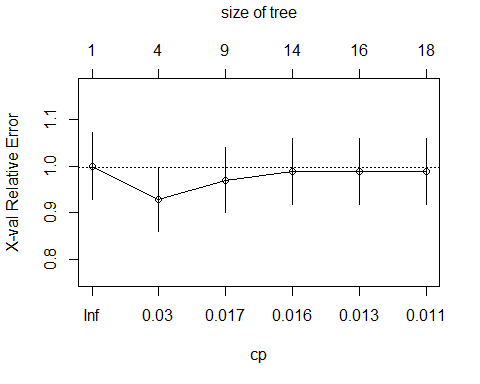
#CART model accuracy  
(336+23)/nrow(test)

## [1] 0.8159091

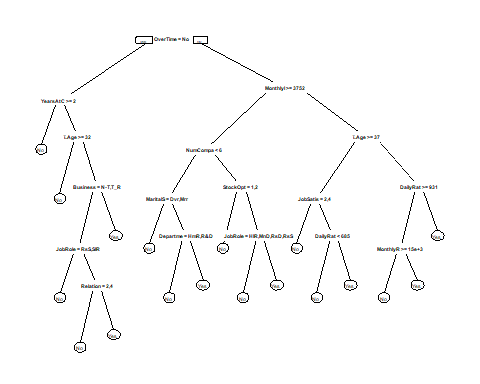
#Baseline Accuracy - If we just predict attrition as No for every observation, we will get an accuracy of 83.8%. Model Accuracy - The model gave us an accuracy of 84%, an improvement of approx. 1% over the baseline accuracy.  
  
#As a fully grown tree is prone to overfitting, lets prune the tree and see if we can improve the model.  
printcp(decisiontree)

##   
## Classification tree:  
## rpart(formula = Attrition ~ ., data = train, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] BusinessTravel DailyRate   
## [3] Department ï..Age   
## [5] JobRole JobSatisfaction   
## [7] MaritalStatus MonthlyIncome   
## [9] MonthlyRate NumCompaniesWorked   
## [11] OverTime RelationshipSatisfaction  
## [13] StockOptionLevel YearsAtCompany   
##   
## Root node error: 166/1030 = 0.16117  
##   
## n= 1030   
##   
## CP nsplit rel error xerror xstd  
## 1 0.048193 0 1.00000 1.00000 0.071086  
## 2 0.018072 3 0.85542 0.92771 0.068942  
## 3 0.016064 8 0.76506 0.96988 0.070210  
## 4 0.015060 13 0.66867 0.98795 0.070738  
## 5 0.012048 15 0.63855 0.98795 0.070738  
## 6 0.010000 17 0.61446 0.98795 0.070738

plotcp(decisiontree)



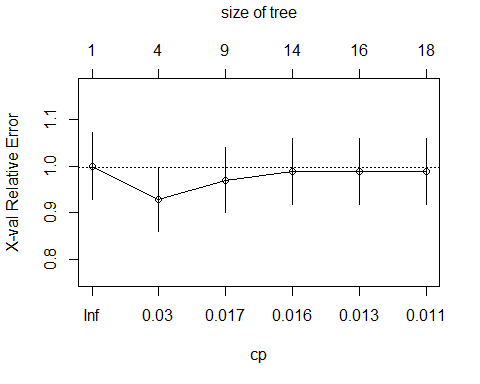
bestcp <- decisiontree$Attrition[which.min(decisiontree$Attrition[,"xerror"]),"CP"]  
prunedModel <-prune(decisiontree, cp= bestcp)  
prp(prunedModel)



printcp(prunedModel)

##   
## Classification tree:  
## rpart(formula = Attrition ~ ., data = train, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] BusinessTravel DailyRate   
## [3] Department ï..Age   
## [5] JobRole JobSatisfaction   
## [7] MaritalStatus MonthlyIncome   
## [9] MonthlyRate NumCompaniesWorked   
## [11] OverTime RelationshipSatisfaction  
## [13] StockOptionLevel YearsAtCompany   
##   
## Root node error: 166/1030 = 0.16117  
##   
## n= 1030   
##   
## CP nsplit rel error xerror xstd  
## 1 0.048193 0 1.00000 1.00000 0.071086  
## 2 0.018072 3 0.85542 0.92771 0.068942  
## 3 0.016064 8 0.76506 0.96988 0.070210  
## 4 0.015060 13 0.66867 0.98795 0.070738  
## 5 0.012048 15 0.63855 0.98795 0.070738  
## 6 0.010000 17 0.61446 0.98795 0.070738

plotcp(prunedModel)



#Predict on the test data  
prediction\_pm <- predict(prunedModel, newdata=test, type="class")  
table(test$Attrition, prediction\_pm)

## prediction\_pm  
## No Yes  
## No 348 21  
## Yes 51 20

(336+23)/nrow(test)

## [1] 0.8159091

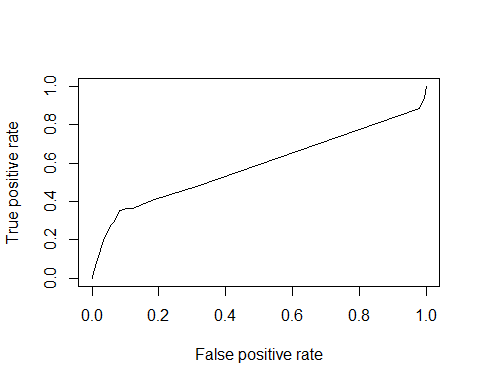
#So the pruning does not improve the model accuracy  
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

prediction\_ROC <- predict(prunedModel, newdata=test)  
pred = prediction(prediction\_ROC[,2], test$Attrition)  
perf = performance(pred, "tpr", "fpr")  
plot(perf)



#Area under the curve  
as.numeric(performance(pred, "auc")@y.values)

## [1] 0.5865873

#Random Forest   
  
   
library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':  
##   
## combine

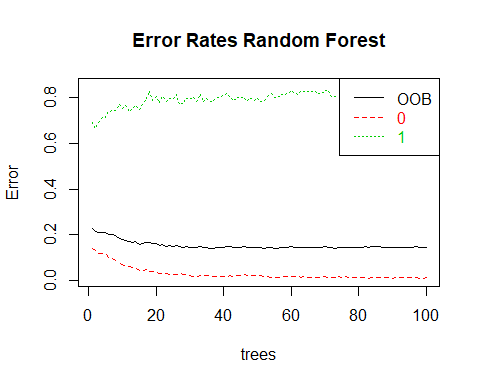
## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

randomForestModel <- randomForest(Attrition~.,data=train,ntree=100,mtry=5, importance=TRUE)  
print(randomForestModel)

##   
## Call:  
## randomForest(formula = Attrition ~ ., data = train, ntree = 100, mtry = 5, importance = TRUE)   
## Type of random forest: classification  
## Number of trees: 100  
## No. of variables tried at each split: 5  
##   
## OOB estimate of error rate: 14.56%  
## Confusion matrix:  
## No Yes class.error  
## No 854 10 0.01157407  
## Yes 140 26 0.84337349

plot(randomForestModel, main="")  
legend("topright", c("OOB", "0", "1"), text.col=1:6, lty=1:3, col=1:3)  
title(main="Error Rates Random Forest")

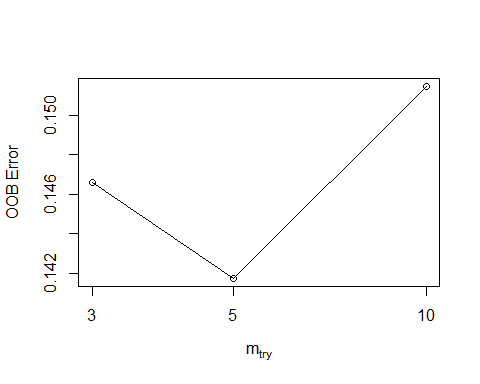


## List the importance of the variables.  
impVar <- round(randomForest::importance(randomForestModel), 2)  
impVar[order(impVar[,3], decreasing=TRUE),]

## No Yes MeanDecreaseAccuracy MeanDecreaseGini  
## OverTime 7.25 12.08 11.29 20.27  
## JobLevel 2.95 3.46 4.30 7.79  
## JobRole 3.64 2.11 4.18 14.56  
## ï..Age 2.87 3.47 3.90 18.68  
## StockOptionLevel 4.04 1.45 3.85 7.58  
## NumCompaniesWorked 2.40 3.09 3.71 10.58  
## TotalWorkingYears 2.55 2.26 3.21 12.17  
## BusinessTravel 2.75 1.52 2.89 5.07  
## YearsInCurrentRole 1.64 1.76 2.28 7.31  
## MaritalStatus 0.55 2.94 1.95 5.30  
## YearsAtCompany 1.46 1.05 1.82 11.94  
## MonthlyIncome 0.71 2.39 1.76 22.99  
## YearsWithCurrManager 1.99 0.03 1.69 7.88  
## Education 1.73 -0.19 1.35 8.52  
## Department 0.85 0.68 1.14 2.54  
## DailyRate 0.95 0.42 1.02 16.35  
## YearsSinceLastPromotion 0.65 0.66 0.92 5.69  
## EducationField 0.94 0.05 0.86 9.36  
## JobSatisfaction 0.85 0.23 0.82 8.02  
## JobInvolvement 0.56 0.56 0.75 8.26  
## DistanceFromHome 0.15 -0.14 0.08 14.29  
## PercentSalaryHike -0.20 0.28 -0.05 9.97  
## RelationshipSatisfaction -0.91 1.06 -0.16 7.59  
## HourlyRate -0.19 -1.54 -0.78 13.82  
## TrainingTimesLastYear -1.87 1.08 -1.00 8.35  
## MonthlyRate -1.25 -0.80 -1.34 13.93

# Tuning Random Forest  
tunedRf <- tuneRF(x = train[,-2],   
 y=as.factor(train$Attrition),  
 mtryStart = 5,   
 ntreeTry=60,   
 stepFactor = 2,   
 improve = 0.001,   
 trace=TRUE,   
 plot = TRUE,  
 doBest = TRUE,  
 nodesize = 5,   
 importance=TRUE  
)

## mtry = 5 OOB error = 14.17%   
## Searching left ...  
## mtry = 3 OOB error = 14.66%   
## -0.03424658 0.001   
## Searching right ...  
## mtry = 10 OOB error = 15.15%   
## -0.06849315 0.001



impvarTunedRf <- tunedRf$importance  
impvarTunedRf[order(impvarTunedRf[,3], decreasing=TRUE),]

## No Yes MeanDecreaseAccuracy  
## OverTime 1.377459e-02 0.0533664396 2.007451e-02  
## JobRole 7.077975e-03 0.0162950444 8.541189e-03  
## TotalWorkingYears 6.181258e-03 0.0042048676 5.821248e-03  
## MonthlyIncome 3.817357e-03 0.0138375883 5.430112e-03  
## JobLevel 3.823750e-03 0.0120692742 5.118708e-03  
## ï..Age 3.200071e-03 0.0145132197 4.987702e-03  
## YearsAtCompany 2.878657e-03 0.0087114865 3.832346e-03  
## StockOptionLevel 2.491679e-03 0.0078103703 3.358959e-03  
## MaritalStatus 1.539545e-03 0.0073896916 2.474629e-03  
## YearsWithCurrManager 2.104689e-03 0.0021166493 2.074422e-03  
## NumCompaniesWorked 1.839176e-03 0.0029870870 1.999315e-03  
## JobSatisfaction 1.551516e-03 0.0020640963 1.632495e-03  
## BusinessTravel 9.187148e-04 0.0042861021 1.457197e-03  
## DistanceFromHome 7.747187e-04 0.0024571642 1.040680e-03  
## YearsInCurrentRole 3.294178e-05 0.0063852817 1.038146e-03  
## JobInvolvement 4.470828e-04 0.0038400568 9.753700e-04  
## DailyRate 4.595290e-04 0.0036028798 9.330698e-04  
## Department 3.451766e-04 0.0023788559 6.619648e-04  
## YearsSinceLastPromotion 9.828311e-04 -0.0012371754 6.129680e-04  
## HourlyRate 5.999602e-04 -0.0001240287 4.610962e-04  
## RelationshipSatisfaction 1.044515e-04 0.0020617046 4.258636e-04  
## EducationField 9.650671e-04 -0.0026416780 3.580265e-04  
## Education 3.853383e-04 0.0001330482 3.447229e-04  
## TrainingTimesLastYear 1.038326e-04 -0.0003210152 4.211951e-05  
## PercentSalaryHike 3.327802e-05 -0.0014661764 -1.890161e-04  
## MonthlyRate -1.336967e-04 -0.0011267225 -2.822954e-04  
## MeanDecreaseGini  
## OverTime 19.251275  
## JobRole 14.083064  
## TotalWorkingYears 10.591924  
## MonthlyIncome 17.003264  
## JobLevel 7.062155  
## ï..Age 15.252457  
## YearsAtCompany 9.537282  
## StockOptionLevel 7.354336  
## MaritalStatus 5.115712  
## YearsWithCurrManager 6.630427  
## NumCompaniesWorked 8.765131  
## JobSatisfaction 7.350888  
## BusinessTravel 4.370969  
## DistanceFromHome 11.200321  
## YearsInCurrentRole 5.121905  
## JobInvolvement 6.177443  
## DailyRate 13.834251  
## Department 2.284567  
## YearsSinceLastPromotion 4.261893  
## HourlyRate 9.667634  
## RelationshipSatisfaction 7.273940  
## EducationField 7.657382  
## Education 6.556013  
## TrainingTimesLastYear 5.127803  
## PercentSalaryHike 7.224265  
## MonthlyRate 10.800101

predictionRf <- predict(tunedRf, test, type="class")  
  
  
  
#RandomForest Accuracy  
#Confusion matrix   
  
t2 <- table(test$Attrition, predictionRf)  
t2

## predictionRf  
## No Yes  
## No 365 4  
## Yes 55 16

#RandomForest model accuracy  
(t2[1]+t2[4])/(nrow(test))

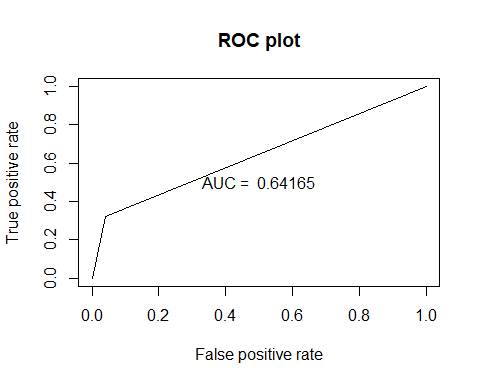
## [1] 0.8659091

#Xtreme Gradient Boosting  
  
  
library(caret)  
library(xgboost)

##   
## Attaching package: 'xgboost'

## The following object is masked from 'package:dplyr':  
##   
## slice

control <- trainControl(method="repeatedcv", number=5)  
set.seed(123)  
model\_xgb <- train(as.factor(Attrition)~., data=train, method="xgbTree", trControl=control)  
#Output Prediction  
pred\_xgb <- predict(model\_xgb, newdata=test)  
  
library(ROCR)  
ROCRpred <- prediction(as.numeric(pred\_xgb), as.numeric(test$Attrition))  
ROCRpref <- performance(ROCRpred,"auc")  
auc\_xgb <- as.numeric(ROCRpref@y.values)  
perf\_ROC <- performance(ROCRpred,"tpr","fpr") #plot the actual ROC curve  
plot(perf\_ROC, main="ROC plot")  
text(0.5,0.5,paste("AUC = ",format(auc\_xgb, digits=5, scientific=FALSE)))



#Confusion Matrix  
confusionMatrix(pred\_xgb, test$Attrition)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 354 48  
## Yes 15 23  
##   
## Accuracy : 0.8568   
## 95% CI : (0.8206, 0.8882)  
## No Information Rate : 0.8386   
## P-Value [Acc > NIR] : 0.1657   
##   
## Kappa : 0.3487   
## Mcnemar's Test P-Value : 5.539e-05   
##   
## Sensitivity : 0.9593   
## Specificity : 0.3239   
## Pos Pred Value : 0.8806   
## Neg Pred Value : 0.6053   
## Prevalence : 0.8386   
## Detection Rate : 0.8045   
## Detection Prevalence : 0.9136   
## Balanced Accuracy : 0.6416   
##   
## 'Positive' Class : No   
##