

Employee Churn Analysis

```
# Load packages

library('ggplot2') # visualization
## Warning: package 'ggplot2' was built under R version 3.4.4

library('ggthemes') # visualization
## Warning: package 'ggthemes' was built under R version 3.4.4

library('scales') # visualization
## Warning: package 'scales' was built under R version 3.4.4

library('dplyr') # data manipulation
## Warning: package 'dplyr' was built under R version 3.4.3
##
## 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('mice') # imputation
## Warning: package 'mice' was built under R version 3.4.4
## Loading required package: lattice
##
## Attaching package: 'mice'
##
## The following objects are masked from 'package:base':
##
##   cbind, rbind

library('randomForest') # classification algorithm
## Warning: package 'randomForest' was built under R version 3.4.4
## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'

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

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

dataNW<-read.csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")
head(dataNW)
```

	i..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		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
```

```
names(dataNW)[names(dataNW) == 'i..Age'] <- 'Age'
```

```
dim(dataNW)
```

```
## [1] 1470 35
```

```
names(dataNW)
```

```
## [1] "Age" "Attrition"
## [3] "BusinessTravel" "DailyRate"
## [5] "Department" "DistanceFromHome"
## [7] "Education" "EducationField"
## [9] "EmployeeCount" "EmployeeNumber"
## [11] "EnvironmentSatisfaction" "Gender"
## [13] "HourlyRate" "JobInvolvement"
## [15] "JobLevel" "JobRole"
## [17] "JobSatisfaction" "MaritalStatus"
## [19] "MonthlyIncome" "MonthlyRate"
## [21] "NumCompaniesWorked" "Over18"
## [23] "OverTime" "PercentSalaryHike"
## [25] "PerformanceRating" "RelationshipSatisfaction"
## [27] "StandardHours" "StockOptionLevel"
## [29] "TotalWorkingYears" "TrainingTimesLastYear"
## [31] "WorkLifeBalance" "YearsAtCompany"
```

```

## [33] "YearsInCurrentRole"      "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"

str(data)

## function (..., list = character(), package = NULL, lib.loc = NULL,
##      verbose = getOption("verbose"), envir = .GlobalEnv)

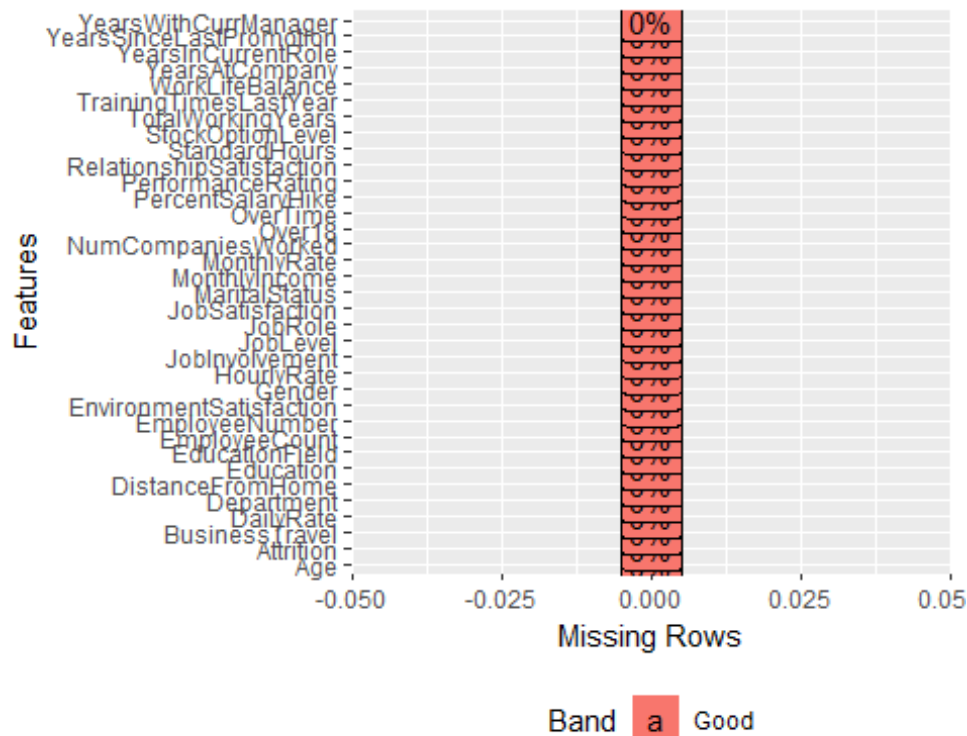
sapply(dataNW, function(x) sum(is.na(x)))

##           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

library(DataExplorer)
plot_str(dataNW)

plot_missing(dataNW)

```



```
Terminated<-as.factor(dataNW$Attrition)
summary(Terminated)

##    No    Yes
## 1233   237

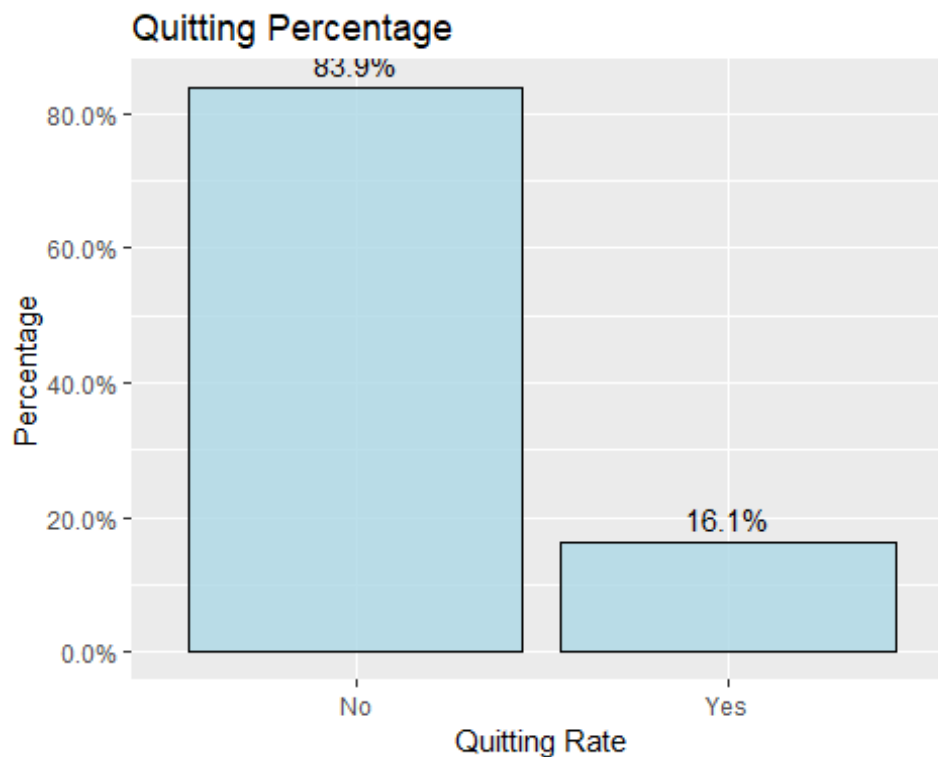
perc_attrition_rate<-sum(dataNW$Attrition)/length(dataNW$Attrition)*100

## Warning in Ops.factor(dataNW$Attrition, length(dataNW$Attrition)): '/' not
## meaningful for factors

prop.table(table(dataNW$Attrition))

##
##           No           Yes
## 0.8387755 0.1612245

Terminated<- ggplot(dataNW, aes(x=Attrition)) +
  geom_bar(aes(y=(..count..)/sum(..count..)), alpha=0.8, fill="lightblue",
color = "black") +
  scale_y_continuous(labels = scales::percent) +
  geom_text(aes(label = scales::percent((..count..)/sum(..count..)),
y= (..count..)/sum(..count..) ), stat= "count", vjust = -0.5)
+
  ylab("Percentage") + xlab("Quitting Rate")+ ggtitle("Quitting Percentage")
Terminated
```



```
library(dplyr)
library(tidyr)

## Warning: package 'tidyr' was built under R version 3.4.3
##
## Attaching package: 'tidyr'
## The following object is masked from 'package:mice':
##
##     complete

#Transforming Termination Column to Factor with True and False values
dataNW$Attrition<-factor(dataNW$Attrition,labels=c('False',"True"))
SHdf<-dataNW %>% group_by(EducationField,Attrition) %>%
  summarise(count=n())

#replacing NA value with 0
SHdf[is.na(SHdf)]<-0

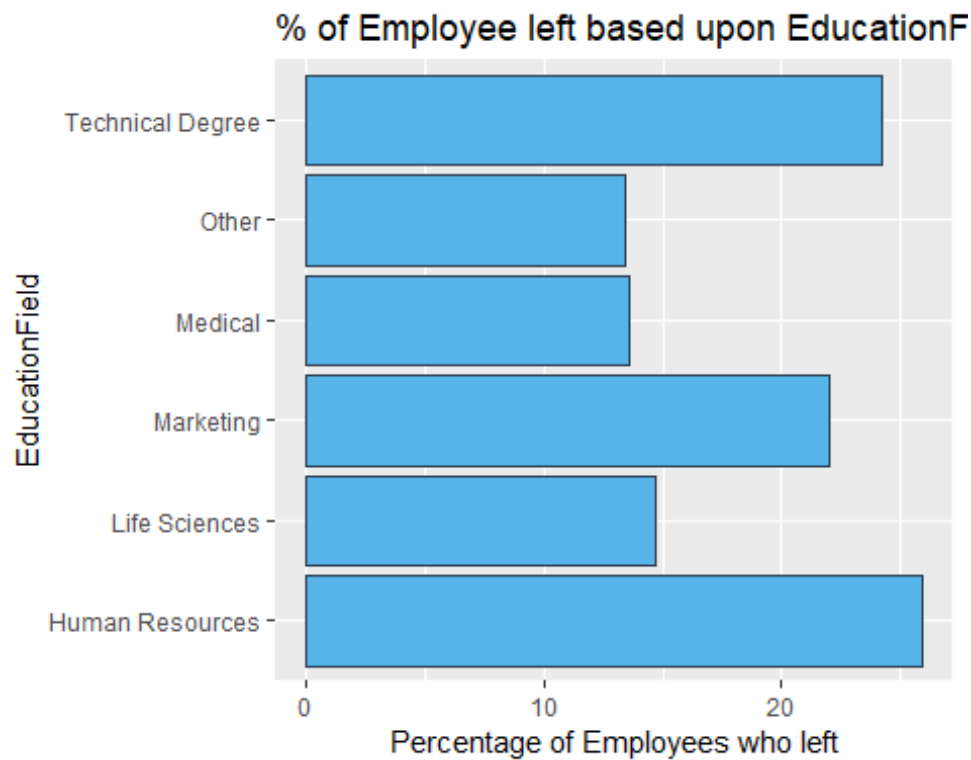
#making a data frame of Departments and the count of workers who Left or not
SHdf<-spread(SHdf,Attrition,count)

SHdf<-transform(SHdf,Perleft=(True/(True+False))*100 ,
PerWork=(False/(True+False))*100)
SHdf
```

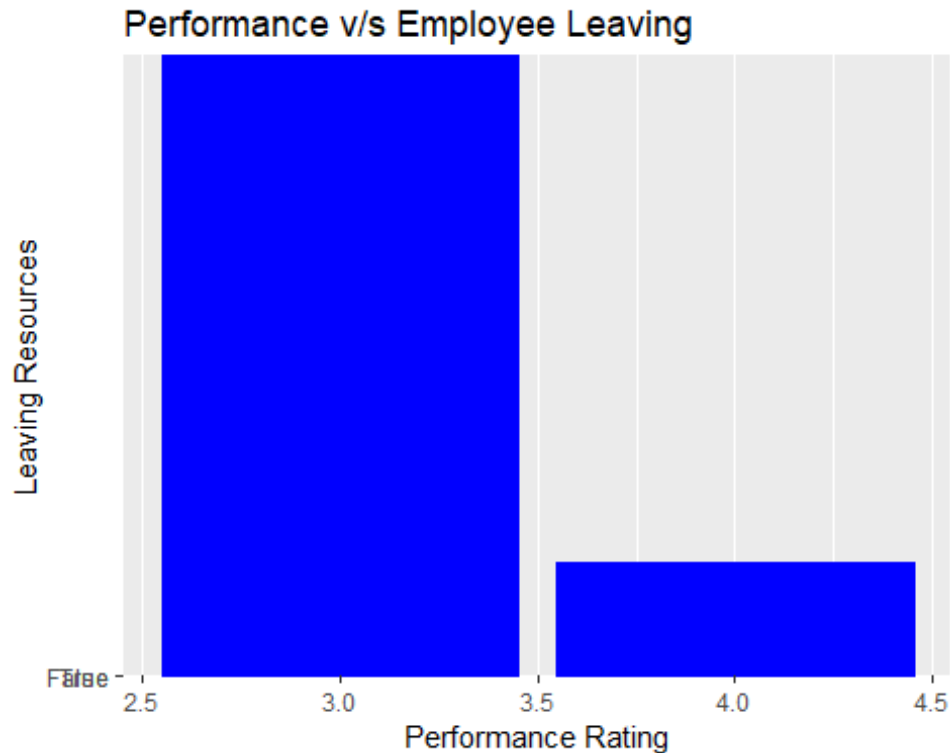
	EducationField	False	True	Perleft	PerWork
## 1	Human Resources	20	7	25.92593	74.07407
## 2	Life Sciences	517	89	14.68647	85.31353
## 3	Marketing	124	35	22.01258	77.98742
## 4	Medical	401	63	13.57759	86.42241
## 5	Other	71	11	13.41463	86.58537
## 6	Technical Degree	100	32	24.24242	75.75758

Percentage of employee who left and Employee who are working based upon Source.Of.Hire

```
#Plot of Department vs Percentage of Employees who Left
ggplot(aes(x=EducationField, y = Perleft),data = SHdf) +
  geom_col(fill='#56B4E9',color='#2f3f52') +
  coord_flip()+
  xlab("EducationField") +
  ylab("Percentage of Employees who left") +
  labs(title="% of Employee left based upon EducationField")
```



```
ggplot(dataNW, aes(x = PerformanceRating, y = Attrition)) + geom_bar(stat =
"identity", fill = 'blue', colour = 'blue') + ggtitle("Performance v/s
Employee Leaving") + labs(y = "Leaving Resources", x =
"Performance Rating")
```



```
dataNW$Attrition <- as.factor(dataNW$Attrition)

dataNW %>%
  select(Department,Attrition) %>%
  group_by(Department, Attrition) %>%
  summarise(count=n()) %>%
  mutate(dep_pct = count/sum(count)) %>%
  ggplot(aes(x=Department, y=dep_pct, fill = Attrition)) +
  geom_bar(stat="identity", alpha = 0.7) +
  geom_text(aes(label = paste0(round(dep_pct*100,0),"%"),
    y=dep_pct+0.02)) +
  scale_fill_brewer(palette="Paired")+
  ylab("Percentage of Employees") + xlab("Department") +
  ggtitle("Resignation per
Department")+theme(axis.text.x=element_text(angle=45,hjust=1))

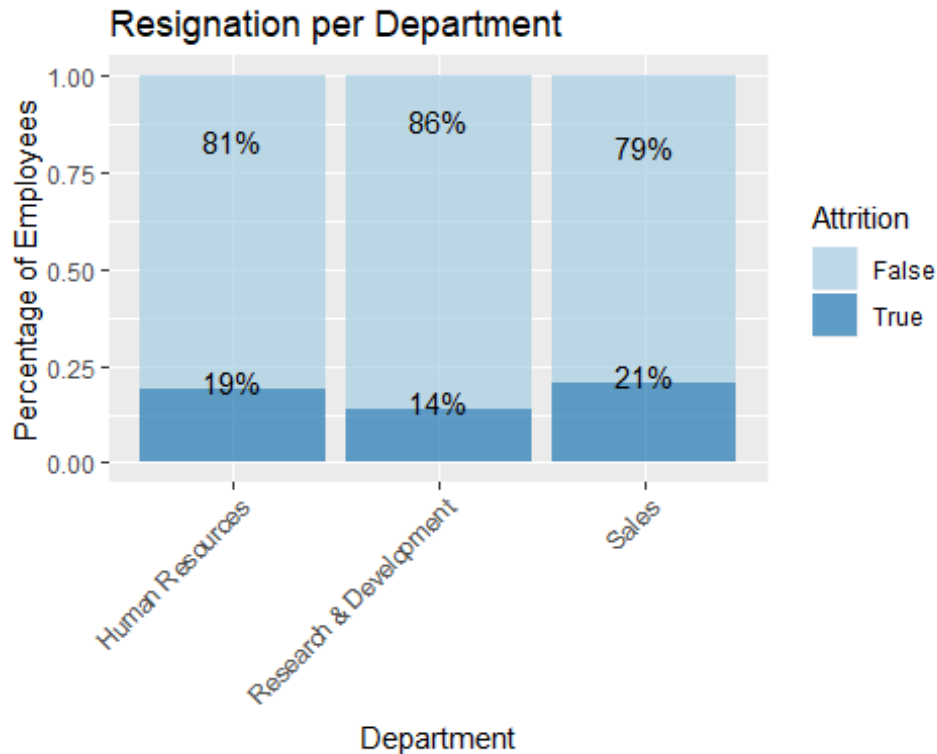
## Warning: package 'bindrcpp' was built under R version 3.4.3

## Warning: `as_dictionary()` is soft-deprecated as of rlang 0.3.0.
## Please use `as_data_pronoun()` instead
## This warning is displayed once per session.

## Warning: `new_overscope()` is soft-deprecated as of rlang 0.2.0.
## Please use `new_data_mask()` instead
## This warning is displayed once per session.
```

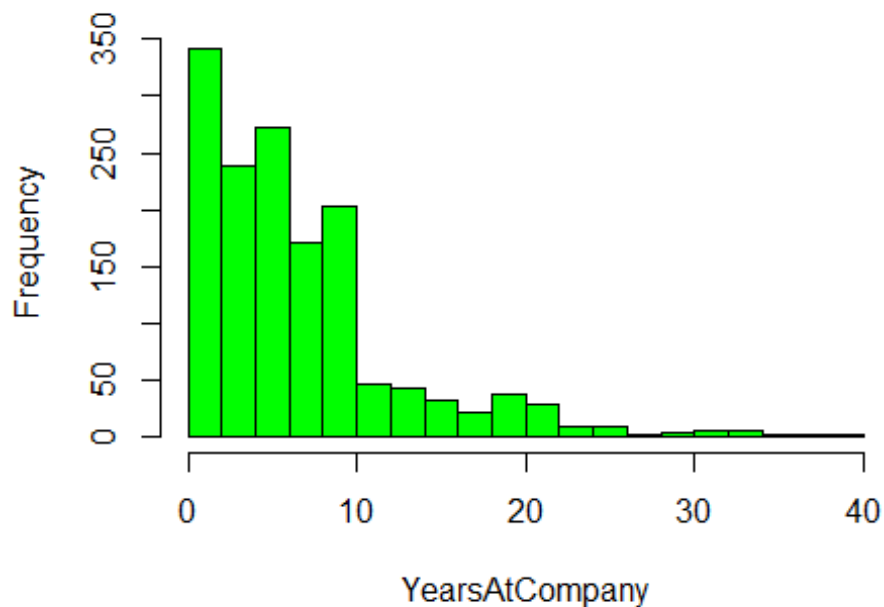


```
## Warning: The `parent` argument of `new_data_mask()` is deprecated.  
## The parent of the data mask is determined from either:  
##  
## * The `env` argument of `eval_tidy()`  
## * Quosure environments when applicable  
## This warning is displayed once per session.  
  
## Warning: `overscope_clean()` is soft-deprecated as of rlang 0.2.0.  
## This warning is displayed once per session.
```



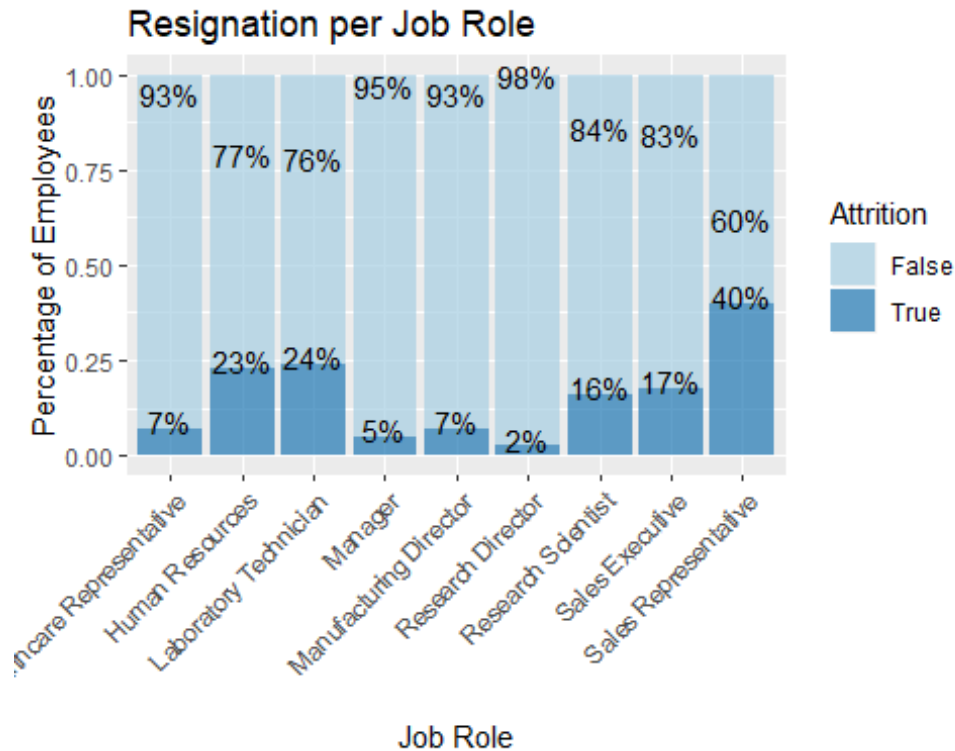
```
hist(dataNW$YearsAtCompany, breaks = 15, col = 'green', main = "Analysis of  
Years At Company Variable", xlab = "YearsAtCompany")
```

Analysis of Years At Company Variable



```
dataNW$Attrition <- as.factor(dataNW$Attrition)

dataNW %>%
  select(JobRole,Attrition) %>%
  group_by(JobRole, Attrition) %>%
  summarise(count=n()) %>%
  mutate(dep_pct = count/sum(count)) %>%
  ggplot(aes(x=JobRole, y=dep_pct, fill = Attrition)) +
  geom_bar(stat="identity", alpha = 0.7) +
  geom_text(aes(label = paste0(round(dep_pct*100,0),"%"),
    y=dep_pct+0.02)) +
  scale_fill_brewer(palette="Paired")+
  ylab("Percentage of Employees") + xlab("Job Role") +
  ggtitle("Resignation per Job
Role")+theme(axis.text.x=element_text(angle=45,hjust=1))
```



```
str(dataNW)

## 'data.frame':    1470 obs. of  35 variables:
##  $ Age                : int  41 49 37 33 27 32 59 30 38 36 ...
##  $ Attrition           : Factor w/ 2 levels "False","True": 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            : int  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 ...
##  $ EmployeeCount        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ EmployeeNumber       : int  1 2 4 5 7 8 10 11 12 13 ...
##  $ EnvironmentSatisfaction : int  2 3 4 4 1 4 3 4 4 3 ...
##  $ Gender               : Factor w/ 2 levels "Female","Male": 1 2 2 1 2
2 1 2 2 2 ...
##  $ HourlyRate           : int  94 61 92 56 40 79 81 67 44 94 ...
##  $ JobInvolvement        : int  3 2 2 3 3 3 4 3 2 3 ...
##  $ JobLevel             : int  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       : int  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 ...
## $ Over18             : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1
...
## $ 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 ...
## $ PerformanceRating   : int    3 4 3 3 3 3 4 4 4 3 ...
## $ RelationshipSatisfaction: int   1 4 2 3 4 3 1 2 2 2 ...
## $ StandardHours       : int   80 80 80 80 80 80 80 80 80 80 ...
## $ StockOptionLevel    : int    0 1 0 0 1 0 3 1 0 2 ...
## $ TotalWorkingYears   : int    8 10 7 8 6 8 12 1 10 17 ...
## $ TrainingTimesLastYear : int    0 3 3 3 3 2 3 2 2 3 ...
## $ WorkLifeBalance     : int    1 3 3 3 3 2 2 3 3 2 ...
## $ 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 ...

```

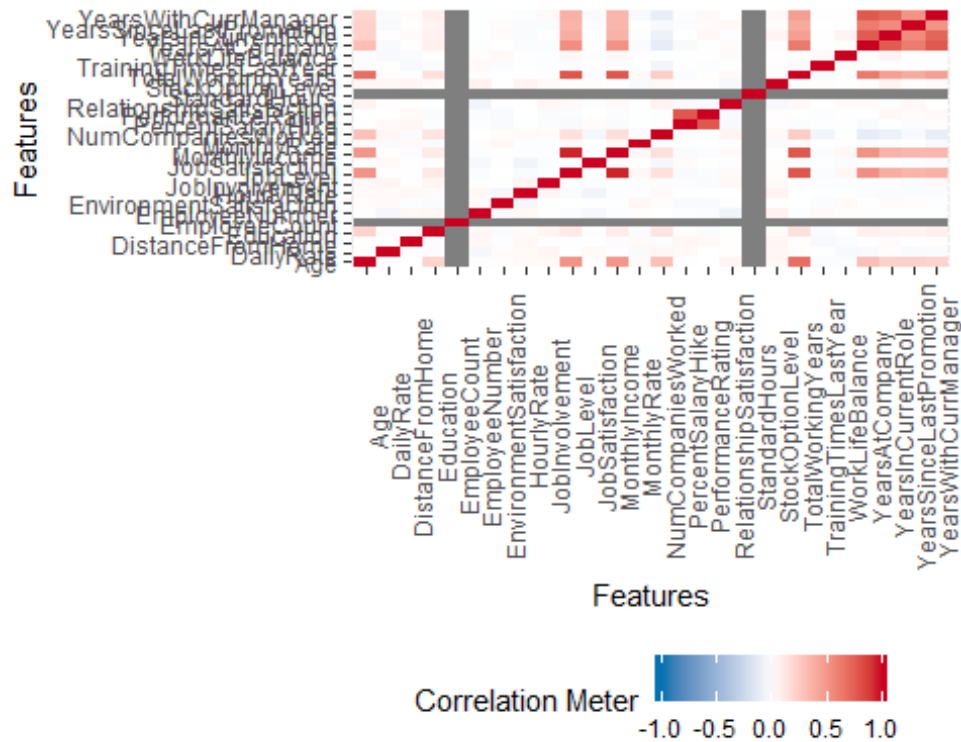
```
df1<-dataNW
```

```
plot_correlation(df1, type = 'continuous')
```

```

## Warning in cor(x = structure(list(Age = c(41L, 49L, 37L, 33L, 27L, 32L, :
## the standard deviation is zero

```



```
names(df1)
```

```
## [1] "Age" "Attrition"
## [3] "BusinessTravel" "DailyRate"
## [5] "Department" "DistanceFromHome"
## [7] "Education" "EducationField"
## [9] "EmployeeCount" "EmployeeNumber"
## [11] "EnvironmentSatisfaction" "Gender"
## [13] "HourlyRate" "JobInvolvement"
## [15] "JobLevel" "JobRole"
## [17] "JobSatisfaction" "MaritalStatus"
## [19] "MonthlyIncome" "MonthlyRate"
## [21] "NumCompaniesWorked" "Over18"
## [23] "OverTime" "PercentSalaryHike"
## [25] "PerformanceRating" "RelationshipSatisfaction"
## [27] "StandardHours" "StockOptionLevel"
## [29] "TotalWorkingYears" "TrainingTimesLastYear"
## [31] "WorkLifeBalance" "YearsAtCompany"
## [33] "YearsInCurrentRole" "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 3.4.3
```

```
#Splitting the data
set.seed(123)
```

```
indices = sample.split(df1$Attrition, SplitRatio = 0.7)
train = df1[indices,]
validation = df1[!(indices),]
```

MODEL1

```
colnames(train)
```

```
## [1] "Age" "Attrition"
## [3] "BusinessTravel" "DailyRate"
## [5] "Department" "DistanceFromHome"
## [7] "Education" "EducationField"
## [9] "EmployeeCount" "EmployeeNumber"
## [11] "EnvironmentSatisfaction" "Gender"
## [13] "HourlyRate" "JobInvolvement"
## [15] "JobLevel" "JobRole"
## [17] "JobSatisfaction" "MaritalStatus"
## [19] "MonthlyIncome" "MonthlyRate"
## [21] "NumCompaniesWorked" "Over18"
## [23] "OverTime" "PercentSalaryHike"
## [25] "PerformanceRating" "RelationshipSatisfaction"
## [27] "StandardHours" "StockOptionLevel"
## [29] "TotalWorkingYears" "TrainingTimesLastYear"
## [31] "WorkLifeBalance" "YearsAtCompany"
## [33] "YearsInCurrentRole" "YearsSinceLastPromotion"
## [35] "YearsWithCurrManager"
```

#Build the first model using all variables

```
model_1 = glm(Attrition ~ Age
+BusinessTravel+DailyRate+Department+DistanceFromHome+Education+EducationFiel
d+EnvironmentSatisfaction+Gender+HourlyRate+JobInvolvement+JobLevel+JobRole+J
obSatisfaction+MaritalStatus+MonthlyIncome+MonthlyRate+NumCompaniesWorked+Ove
rTime+PercentSalaryHike+PerformanceRating+RelationshipSatisfaction+StandardHo
urs+StockOptionLevel+TotalWorkingYears+TrainingTimesLastYear+WorkLifeBalance+
YearsAtCompany+YearsInCurrentRole+YearsSinceLastPromotion+YearsWithCurrManage
r, data = train, family = "binomial")
summary(model_1)
```

```
##
## Call:
## glm(formula = Attrition ~ Age + BusinessTravel + DailyRate +
##     Department + DistanceFromHome + Education + EducationField +
##     EnvironmentSatisfaction + Gender + HourlyRate + JobInvolvement +
##     JobLevel + JobRole + JobSatisfaction + MaritalStatus + MonthlyIncome +
##     MonthlyRate + NumCompaniesWorked + OverTime + PercentSalaryHike +
##     PerformanceRating + RelationshipSatisfaction + StandardHours +
##     StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##     WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
YearsSinceLastPromotion +
##     YearsWithCurrManager, family = "binomial", data = train)
##
```

```

## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6681  -0.4819  -0.2502  -0.0908   3.4600
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.059e+01  7.008e+02  -0.015  0.987940
## Age           -1.804e-02  1.582e-02  -1.140  0.254165
## BusinessTravelTravel_Frequently  1.989e+00  5.396e-01   3.687  0.000227
## BusinessTravelTravel_Rarely     1.037e+00  5.100e-01   2.034  0.041932
## DailyRate     -1.839e-04  2.630e-04  -0.699  0.484450
## DepartmentResearch & Development  1.348e+01  7.008e+02   0.019  0.984649
## DepartmentSales  1.423e+01  7.008e+02   0.020  0.983796
## DistanceFromHome  4.144e-02  1.302e-02   3.184  0.001454
## Education      2.685e-02  1.075e-01   0.250  0.802788
## EducationFieldLife Sciences  -9.295e-01  1.073e+00  -0.866  0.386263
## EducationFieldMarketing  -7.117e-01  1.123e+00  -0.634  0.526292
## EducationFieldMedical  -1.368e+00  1.069e+00  -1.279  0.200766
## EducationFieldOther  -1.014e+00  1.131e+00  -0.896  0.370034
## EducationFieldTechnical Degree  -2.719e-01  1.098e+00  -0.248  0.804510
## EnvironmentSatisfaction  -4.078e-01  9.939e-02  -4.103  4.09e-05
## GenderMale      3.583e-01  2.204e-01   1.625  0.104058
## HourlyRate      9.030e-04  5.231e-03   0.173  0.862952
## JobInvolvement  -4.798e-01  1.496e-01  -3.207  0.001343
## JobLevel        6.848e-02  3.702e-01   0.185  0.853224
## JobRoleHuman Resources  1.503e+01  7.008e+02   0.021  0.982884
## JobRoleLaboratory Technician  2.157e+00  6.548e-01   3.294  0.000987
## JobRoleManager   7.738e-01  1.121e+00   0.690  0.490189
## JobRoleManufacturing Director  8.420e-01  7.104e-01   1.185  0.235910
## JobRoleResearch Director  -8.839e-01  1.338e+00  -0.661  0.508748
## JobRoleResearch Scientist  8.761e-01  6.704e-01   1.307  0.191231
## JobRoleSales Executive  8.892e-01  1.314e+00   0.676  0.498732
## JobRoleSales Representative  1.985e+00  1.373e+00   1.446  0.148235
## JobSatisfaction  -4.100e-01  9.862e-02  -4.157  3.22e-05
## MaritalStatusMarried  3.086e-01  3.209e-01   0.962  0.336192
## MaritalStatusSingle  9.227e-01  4.069e-01   2.268  0.023357
## MonthlyIncome    -2.048e-05  9.401e-05  -0.218  0.827550
## MonthlyRate      1.907e-05  1.519e-05   1.255  0.209312
## NumCompaniesWorked  1.672e-01  4.741e-02   3.528  0.000419
## OverTimeYes      1.887e+00  2.354e-01   8.014  1.11e-15
## PercentSalaryHike  -2.948e-02  4.652e-02  -0.634  0.526186
## PerformanceRating  -3.232e-01  5.014e-01  -0.645  0.519218
## RelationshipSatisfaction  -2.236e-01  9.700e-02  -2.305  0.021171
## StandardHours      NA             NA             NA             NA
## StockOptionLevel  -2.891e-01  1.925e-01  -1.502  0.133094
## TotalWorkingYears  -6.822e-02  3.474e-02  -1.964  0.049547
## TrainingTimesLastYear  -1.188e-01  8.587e-02  -1.384  0.166388
## WorkLifeBalance   -5.179e-01  1.491e-01  -3.473  0.000515
## YearsAtCompany     6.794e-02  4.760e-02   1.427  0.153494
## YearsInCurrentRole  -1.445e-01  5.601e-02  -2.581  0.009861

```

```

## YearsSinceLastPromotion      1.391e-01  5.060e-02   2.750  0.005968
## YearsWithCurrManager        -8.400e-02  5.929e-02  -1.417  0.156520
##
## (Intercept)
## Age
## BusinessTravelTravel_Frequently ***
## BusinessTravelTravel_Rarely    *
## DailyRate
## DepartmentResearch & Development
## DepartmentSales
## DistanceFromHome             **
## Education
## EducationFieldLife Sciences
## EducationFieldMarketing
## EducationFieldMedical
## EducationFieldOther
## EducationFieldTechnical Degree
## EnvironmentSatisfaction       ***
## GenderMale
## HourlyRate
## JobInvolvement                **
## JobLevel
## JobRoleHuman Resources
## JobRoleLaboratory Technician  ***
## JobRoleManager
## JobRoleManufacturing Director
## JobRoleResearch Director
## JobRoleResearch Scientist
## JobRoleSales Executive
## JobRoleSales Representative
## JobSatisfaction               ***
## MaritalStatusMarried
## MaritalStatusSingle           *
## MonthlyIncome
## MonthlyRate
## NumCompaniesWorked            ***
## OverTimeYes                   ***
## PercentSalaryHike
## PerformanceRating
## RelationshipSatisfaction       *
## StandardHours
## StockOptionLevel
## TotalWorkingYears             *
## TrainingTimesLastYear
## WorkLifeBalance               ***
## 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: 909.34  on 1028  degrees of freedom
## Residual deviance: 602.56  on  984  degrees of freedom
## AIC: 692.56
##
## Number of Fisher Scoring iterations: 15
```

Using stepAIC for variable selection, which is a iterative process of adding or removing variables, in order to get a subset of variables that gives the best performing model.

```
library(MASS)

##
## Attaching package: 'MASS'

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

model_2<- stepAIC(model_1, direction="both")

## Start:  AIC=692.56
## Attrition ~ Age + BusinessTravel + DailyRate + Department +
DistanceFromHome +
##      Education + EducationField + EnvironmentSatisfaction + Gender +
##      HourlyRate + JobInvolvement + JobLevel + JobRole + JobSatisfaction +
##      MaritalStatus + MonthlyIncome + MonthlyRate + NumCompaniesWorked +
##      OverTime + PercentSalaryHike + PerformanceRating +
RelationshipSatisfaction +
##      StandardHours + StockOptionLevel + TotalWorkingYears +
TrainingTimesLastYear +
##      WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
YearsSinceLastPromotion +
##      YearsWithCurrManager
##
## Step:  AIC=692.56
## Attrition ~ Age + BusinessTravel + DailyRate + Department +
DistanceFromHome +
##      Education + EducationField + EnvironmentSatisfaction + Gender +
##      HourlyRate + JobInvolvement + JobLevel + JobRole + JobSatisfaction +
##      MaritalStatus + MonthlyIncome + MonthlyRate + NumCompaniesWorked +
##      OverTime + PercentSalaryHike + PerformanceRating +
RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
YearsSinceLastPromotion +
##      YearsWithCurrManager
```

```

##
##
## Df Deviance AIC
## - Department 2 604.10 690.10
## - HourlyRate 1 602.59 690.59
## - JobLevel 1 602.59 690.59
## - MonthlyIncome 1 602.60 690.60
## - Education 1 602.62 690.62
## - PercentSalaryHike 1 602.96 690.96
## - PerformanceRating 1 602.97 690.97
## - DailyRate 1 603.05 691.05
## - Age 1 603.88 691.88
## - EducationField 5 612.11 692.11
## - MonthlyRate 1 604.14 692.14
## - TrainingTimesLastYear 1 604.50 692.50
## - YearsWithCurrManager 1 604.54 692.54
## <none> 602.56 692.56
## - YearsAtCompany 1 604.58 692.58
## - StockOptionLevel 1 604.89 692.89
## - Gender 1 605.24 693.24
## - MaritalStatus 2 608.52 694.52
## - TotalWorkingYears 1 606.58 694.58
## - RelationshipSatisfaction 1 607.91 695.91
## - YearsInCurrentRole 1 609.19 697.19
## - YearsSinceLastPromotion 1 610.31 698.31
## - DistanceFromHome 1 612.63 700.63
## - JobInvolvement 1 612.95 700.95
## - NumCompaniesWorked 1 614.75 702.75
## - WorkLifeBalance 1 614.77 702.77
## - JobRole 8 632.01 706.01
## - EnvironmentSatisfaction 1 619.91 707.91
## - JobSatisfaction 1 620.43 708.43
## - BusinessTravel 2 624.80 710.80
## - OverTime 1 673.26 761.26
##
## Step: AIC=690.1
## Attrition ~ Age + BusinessTravel + DailyRate + DistanceFromHome +
## Education + EducationField + EnvironmentSatisfaction + Gender +
## HourlyRate + JobInvolvement + JobLevel + JobRole + JobSatisfaction +
## MaritalStatus + MonthlyIncome + MonthlyRate + NumCompaniesWorked +
## OverTime + PercentSalaryHike + PerformanceRating +
RelationshipSatisfaction +
## StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
## WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
YearsSinceLastPromotion +
## YearsWithCurrManager
##
## Df Deviance AIC
## - JobLevel 1 604.12 688.12
## - HourlyRate 1 604.14 688.14
## - MonthlyIncome 1 604.15 688.15

```

```

## - Education          1    604.15 688.15
## - PercentSalaryHike  1    604.45 688.45
## - PerformanceRating  1    604.53 688.53
## - DailyRate          1    604.56 688.56
## - Age                1    605.49 689.49
## - MonthlyRate        1    605.73 689.73
## - TrainingTimesLastYear 1    605.91 689.91
## - YearsWithCurrManager 1    605.96 689.96
## - EducationField      5    614.03 690.03
## <none>                604.10 690.10
## - YearsAtCompany     1    606.12 690.12
## - StockOptionLevel   1    606.61 690.61
## - Gender             1    606.72 690.72
## - MaritalStatus      2    609.96 691.96
## - TotalWorkingYears  1    608.00 692.00
## + Department         2    602.56 692.56
## - RelationshipSatisfaction 1    609.53 693.53
## - YearsInCurrentRole  1    610.79 694.79
## - YearsSinceLastPromotion 1    611.87 695.87
## - DistanceFromHome   1    613.88 697.88
## - JobInvolvement     1    615.14 699.14
## - WorkLifeBalance    1    616.16 700.16
## - NumCompaniesWorked  1    616.50 700.50
## - EnvironmentSatisfaction 1    621.45 705.45
## - JobSatisfaction     1    622.01 706.01
## - BusinessTravel      2    626.34 708.34
## - JobRole            8    642.99 712.99
## - OverTime           1    676.05 760.05
##
## Step:  AIC=688.12
## Attrition ~ Age + BusinessTravel + DailyRate + DistanceFromHome +
##      Education + EducationField + EnvironmentSatisfaction + Gender +
##      HourlyRate + JobInvolvement + JobRole + JobSatisfaction +
##      MaritalStatus + MonthlyIncome + MonthlyRate + NumCompaniesWorked +
##      OverTime + PercentSalaryHike + PerformanceRating +
##      RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
##      YearsSinceLastPromotion +
##      YearsWithCurrManager
##
##              Df Deviance    AIC
## - MonthlyIncome      1    604.15 686.15
## - HourlyRate          1    604.16 686.16
## - Education           1    604.17 686.17
## - PercentSalaryHike   1    604.48 686.48
## - PerformanceRating   1    604.55 686.55
## - DailyRate           1    604.59 686.59
## - Age                 1    605.51 687.51
## - MonthlyRate         1    605.78 687.78

```

```

## - TrainingTimesLastYear      1    605.93 687.93
## - YearsWithCurrManager       1    605.98 687.98
## <none>                        1    604.12 688.12
## - EducationField             5    614.16 688.16
## - YearsAtCompany             1    606.27 688.27
## - StockOptionLevel           1    606.63 688.63
## - Gender                     1    606.75 688.75
## - MaritalStatus              2    610.06 690.06
## + JobLevel                   1    604.10 690.10
## - TotalWorkingYears          1    608.11 690.11
## + Department                 2    602.59 690.59
## - RelationshipSatisfaction    1    609.58 691.58
## - YearsInCurrentRole         1    611.00 693.00
## - YearsSinceLastPromotion     1    611.87 693.87
## - DistanceFromHome           1    614.01 696.01
## - JobInvolvement             1    615.14 697.14
## - WorkLifeBalance            1    616.17 698.17
## - NumCompaniesWorked         1    616.54 698.54
## - EnvironmentSatisfaction     1    621.49 703.49
## - JobSatisfaction             1    622.03 704.03
## - BusinessTravel             2    626.40 706.40
## - JobRole                    8    643.05 711.05
## - OverTime                   1    676.05 758.05
##
## Step:  AIC=686.15
## Attrition ~ Age + BusinessTravel + DailyRate + DistanceFromHome +
##      Education + EducationField + EnvironmentSatisfaction + Gender +
##      HourlyRate + JobInvolvement + JobRole + JobSatisfaction +
##      MaritalStatus + MonthlyRate + NumCompaniesWorked + OverTime +
##      PercentSalaryHike + PerformanceRating + RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
##      YearsSinceLastPromotion +
##      YearsWithCurrManager
##
##              Df Deviance    AIC
## - HourlyRate      1    604.18 684.18
## - Education        1    604.20 684.20
## - PercentSalaryHike 1    604.51 684.51
## - PerformanceRating 1    604.57 684.57
## - DailyRate        1    604.62 684.62
## - Age              1    605.54 685.54
## - MonthlyRate      1    605.80 685.80
## - TrainingTimesLastYear 1    605.96 685.96
## - YearsWithCurrManager 1    605.98 685.98
## <none>              1    604.15 686.15
## - EducationField    5    614.17 686.17
## - YearsAtCompany    1    606.27 686.27
## - StockOptionLevel  1    606.66 686.66
## - Gender            1    606.76 686.76

```

```

## - MaritalStatus          2    610.06 688.06
## + MonthlyIncome          1    604.12 688.12
## + JobLevel               1    604.15 688.15
## + Department             2    602.61 688.61
## - TotalWorkingYears      1    609.14 689.14
## - RelationshipSatisfaction 1    609.58 689.58
## - YearsInCurrentRole     1    611.01 691.01
## - YearsSinceLastPromotion 1    611.88 691.88
## - DistanceFromHome       1    614.05 694.05
## - JobInvolvement         1    615.15 695.15
## - WorkLifeBalance        1    616.25 696.25
## - NumCompaniesWorked     1    616.55 696.55
## - EnvironmentSatisfaction 1    621.59 701.59
## - JobSatisfaction         1    622.14 702.14
## - BusinessTravel         2    626.40 704.40
## - JobRole                8    648.20 714.20
## - OverTime               1    676.05 756.05
##
## Step:  AIC=684.18
## Attrition ~ Age + BusinessTravel + DailyRate + DistanceFromHome +
##      Education + EducationField + EnvironmentSatisfaction + Gender +
##      JobInvolvement + JobRole + JobSatisfaction + MaritalStatus +
##      MonthlyRate + NumCompaniesWorked + OverTime + PercentSalaryHike +
##      PerformanceRating + RelationshipSatisfaction + StockOptionLevel +
##      TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
##      YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
##      YearsWithCurrManager
##
##              Df Deviance    AIC
## - Education          1    604.23 682.23
## - PercentSalaryHike   1    604.54 682.54
## - PerformanceRating   1    604.61 682.61
## - DailyRate          1    604.65 682.65
## - Age                1    605.56 683.56
## - MonthlyRate        1    605.85 683.85
## - TrainingTimesLastYear 1    606.02 684.02
## - YearsWithCurrManager 1    606.03 684.03
## <none>                604.18 684.18
## - EducationField     5    614.29 684.29
## - YearsAtCompany     1    606.35 684.35
## - StockOptionLevel    1    606.69 684.69
## - Gender              1    606.77 684.77
## - MaritalStatus      2    610.10 686.10
## + HourlyRate         1    604.15 686.15
## + MonthlyIncome      1    604.16 686.16
## + JobLevel           1    604.18 686.18
## + Department         2    602.64 686.64
## - TotalWorkingYears  1    609.19 687.19
## - RelationshipSatisfaction 1    609.63 687.63
## - YearsInCurrentRole 1    611.07 689.07

```

```

## - YearsSinceLastPromotion 1 611.89 689.89
## - DistanceFromHome 1 614.12 692.12
## - JobInvolvement 1 615.21 693.21
## - WorkLifeBalance 1 616.30 694.30
## - NumCompaniesWorked 1 616.56 694.56
## - EnvironmentSatisfaction 1 621.62 699.62
## - JobSatisfaction 1 622.31 700.31
## - BusinessTravel 2 626.41 702.41
## - JobRole 8 648.22 712.22
## - OverTime 1 676.10 754.10
##
## Step: AIC=682.23
## Attrition ~ Age + BusinessTravel + DailyRate + DistanceFromHome +
## EducationField + EnvironmentSatisfaction + Gender + JobInvolvement +
## JobRole + JobSatisfaction + MaritalStatus + MonthlyRate +
## NumCompaniesWorked + OverTime + PercentSalaryHike + PerformanceRating
## +
## RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears +
## TrainingTimesLastYear + WorkLifeBalance + YearsAtCompany +
## YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager
##
## Df Deviance AIC
## - PercentSalaryHike 1 604.58 680.58
## - PerformanceRating 1 604.67 680.67
## - DailyRate 1 604.71 680.71
## - Age 1 605.56 681.56
## - MonthlyRate 1 605.86 681.86
## - TrainingTimesLastYear 1 606.09 682.09
## - YearsWithCurrManager 1 606.10 682.10
## <none> 604.23 682.23
## - EducationField 5 614.40 682.40
## - YearsAtCompany 1 606.41 682.41
## - StockOptionLevel 1 606.70 682.70
## - Gender 1 606.79 682.79
## + Education 1 604.18 684.18
## + HourlyRate 1 604.20 684.20
## + MonthlyIncome 1 604.21 684.21
## - MaritalStatus 2 610.21 684.21
## + JobLevel 1 604.23 684.23
## + Department 2 602.69 684.69
## - TotalWorkingYears 1 609.22 685.22
## - RelationshipSatisfaction 1 609.64 685.64
## - YearsInCurrentRole 1 611.08 687.08
## - YearsSinceLastPromotion 1 611.96 687.96
## - DistanceFromHome 1 614.31 690.31
## - JobInvolvement 1 615.22 691.22
## - WorkLifeBalance 1 616.32 692.32
## - NumCompaniesWorked 1 616.69 692.69
## - EnvironmentSatisfaction 1 621.66 697.66
## - JobSatisfaction 1 622.33 698.33

```

```

## - BusinessTravel          2   626.49 700.49
## - JobRole                 8   648.32 710.32
## - OverTime                1   676.19 752.19
##
## Step:  AIC=680.58
## Attrition ~ Age + BusinessTravel + DailyRate + DistanceFromHome +
##      EducationField + EnvironmentSatisfaction + Gender + JobInvolvement +
##      JobRole + JobSatisfaction + MaritalStatus + MonthlyRate +
##      NumCompaniesWorked + OverTime + PerformanceRating +
##      RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
##      YearsSinceLastPromotion +
##      YearsWithCurrManager
##
##
##      Df Deviance   AIC
## - DailyRate          1   605.07 679.07
## - Age                1   605.96 679.96
## - MonthlyRate        1   606.06 680.06
## - YearsWithCurrManager 1   606.44 680.44
## - TrainingTimesLastYear 1   606.46 680.46
## <none>                1   604.58 680.58
## - YearsAtCompany     1   606.73 680.73
## - EducationField      5   614.92 680.92
## - Gender              1   607.12 681.12
## - StockOptionLevel    1   607.13 681.13
## - PerformanceRating   1   607.29 681.29
## + PercentSalaryHike   1   604.23 682.23
## + Education           1   604.54 682.54
## + HourlyRate          1   604.55 682.55
## + MonthlyIncome       1   604.55 682.55
## - MaritalStatus       2   610.56 682.56
## + JobLevel            1   604.58 682.58
## + Department          2   603.09 683.09
## - TotalWorkingYears   1   609.51 683.51
## - RelationshipSatisfaction 1   610.08 684.08
## - YearsInCurrentRole  1   611.35 685.35
## - YearsSinceLastPromotion 1   612.42 686.42
## - DistanceFromHome    1   614.53 688.53
## - JobInvolvement      1   615.55 689.55
## - WorkLifeBalance     1   616.44 690.44
## - NumCompaniesWorked  1   617.00 691.00
## - EnvironmentSatisfaction 1   621.77 695.77
## - JobSatisfaction     1   622.63 696.63
## - BusinessTravel      2   626.92 698.92
## - JobRole             8   648.61 708.61
## - OverTime            1   677.31 751.31
##
## Step:  AIC=679.07
## Attrition ~ Age + BusinessTravel + DistanceFromHome + EducationField +

```

```

##      EnvironmentSatisfaction + Gender + JobInvolvement + JobRole +
##      JobSatisfaction + MaritalStatus + MonthlyRate + NumCompaniesWorked +
##      OverTime + PerformanceRating + RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
YearsSinceLastPromotion +
##      YearsWithCurrManager
##
##
##      Df Deviance    AIC
## - Age                1    606.56 678.56
## - MonthlyRate        1    606.59 678.59
## - YearsWithCurrManager 1    606.85 678.85
## - TrainingTimesLastYear 1    606.94 678.94
## <none>                605.07 679.07
## - YearsAtCompany      1    607.24 679.24
## - Gender              1    607.63 679.63
## - StockOptionLevel    1    607.64 679.64
## - EducationField      5    615.71 679.71
## - PerformanceRating   1    607.88 679.88
## + DailyRate           1    604.58 680.58
## + PercentSalaryHike   1    604.71 680.71
## + Education           1    605.01 681.01
## + MonthlyIncome       1    605.04 681.04
## + HourlyRate          1    605.05 681.05
## + JobLevel            1    605.07 681.07
## - MaritalStatus       2    611.19 681.19
## + Department          2    603.62 681.62
## - TotalWorkingYears   1    610.13 682.13
## - RelationshipSatisfaction 1    610.70 682.70
## - YearsInCurrentRole  1    612.13 684.13
## - YearsSinceLastPromotion 1    613.29 685.29
## - DistanceFromHome    1    615.28 687.28
## - JobInvolvement      1    616.21 688.21
## - WorkLifeBalance     1    616.84 688.84
## - NumCompaniesWorked  1    617.62 689.62
## - EnvironmentSatisfaction 1    622.20 694.20
## - JobSatisfaction     1    623.24 695.24
## - BusinessTravel      2    627.76 697.76
## - JobRole             8    649.00 707.00
## - OverTime            1    677.82 749.82
##
## Step:  AIC=678.56
## Attrition ~ BusinessTravel + DistanceFromHome + EducationField +
##      EnvironmentSatisfaction + Gender + JobInvolvement + JobRole +
##      JobSatisfaction + MaritalStatus + MonthlyRate + NumCompaniesWorked +
##      OverTime + PerformanceRating + RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
YearsSinceLastPromotion +
##      YearsWithCurrManager

```



```

##
##
## - YearsWithCurrManager      1    608.11 678.11
## - MonthlyRate                1    608.17 678.17
## <none>                      606.56 678.56
## - TrainingTimesLastYear     1    608.62 678.62
## - YearsAtCompany            1    608.88 678.88
## - Gender                    1    608.91 678.91
## - EducationField            5    617.02 679.02
## - StockOptionLevel          1    609.06 679.06
## + Age                       1    605.07 679.07
## - PerformanceRating         1    609.39 679.39
## + DailyRate                 1    605.96 679.96
## + PercentSalaryHike         1    606.14 680.14
## + MonthlyIncome             1    606.53 680.53
## + Education                  1    606.55 680.55
## + HourlyRate                 1    606.55 680.55
## + JobLevel                   1    606.55 680.55
## + Department                 2    605.03 681.03
## - MaritalStatus             2    613.09 681.09
## - RelationshipSatisfaction   1    612.51 682.51
## - YearsInCurrentRole        1    613.61 683.61
## - YearsSinceLastPromotion    1    614.49 684.49
## - DistanceFromHome          1    616.56 686.56
## - TotalWorkingYears         1    617.49 687.49
## - WorkLifeBalance           1    617.92 687.92
## - JobInvolvement            1    617.93 687.93
## - NumCompaniesWorked        1    618.15 688.15
## - EnvironmentSatisfaction    1    623.84 693.84
## - JobSatisfaction            1    625.31 695.31
## - BusinessTravel            2    629.84 697.84
## - JobRole                    8    650.35 706.35
## - OverTime                   1    678.86 748.86
##
## Step:  AIC=678.11
## Attrition ~ BusinessTravel + DistanceFromHome + EducationField +
##      EnvironmentSatisfaction + Gender + JobInvolvement + JobRole +
##      JobSatisfaction + MaritalStatus + MonthlyRate + NumCompaniesWorked +
##      OverTime + PerformanceRating + RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsAtCompany + YearsInCurrentRole +
##      YearsSinceLastPromotion
##
##
##
## - YearsAtCompany            1    609.28 677.28
## - MonthlyRate                1    609.87 677.87
## - TrainingTimesLastYear     1    610.09 678.09
## <none>                      608.11 678.11
## + YearsWithCurrManager      1    606.56 678.56
## - Gender                    1    610.60 678.60

```

```

## - EducationField          5    618.78 678.78
## - StockOptionLevel        1    610.84 678.84
## + Age                      1    606.85 678.85
## - PerformanceRating        1    611.05 679.05
## + DailyRate                1    607.60 679.60
## + PercentSalaryHike        1    607.70 679.70
## + Education                1    608.08 680.08
## + JobLevel                 1    608.10 680.10
## + HourlyRate               1    608.10 680.10
## + MonthlyIncome            1    608.10 680.10
## - MaritalStatus            2    614.37 680.37
## + Department                2    606.70 680.70
## - RelationshipSatisfaction  1    614.08 682.08
## - YearsSinceLastPromotion  1    615.36 683.36
## - YearsInCurrentRole       1    617.34 685.34
## - DistanceFromHome         1    618.00 686.00
## - WorkLifeBalance          1    619.03 687.03
## - TotalWorkingYears        1    619.43 687.43
## - NumCompaniesWorked       1    620.12 688.12
## - JobInvolvement           1    620.20 688.20
## - EnvironmentSatisfaction  1    625.21 693.21
## - JobSatisfaction           1    626.50 694.50
## - BusinessTravel            2    631.09 697.09
## - JobRole                   8    652.03 706.03
## - OverTime                  1    680.54 748.54
##
## Step:  AIC=677.28
## Attrition ~ BusinessTravel + DistanceFromHome + EducationField +
##      EnvironmentSatisfaction + Gender + JobInvolvement + JobRole +
##      JobSatisfaction + MaritalStatus + MonthlyRate + NumCompaniesWorked +
##      OverTime + PerformanceRating + RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsInCurrentRole + YearsSinceLastPromotion
##
##              Df Deviance    AIC
## - MonthlyRate      1    610.89 676.89
## - TrainingTimesLastYear  1    611.24 677.24
## <none>              609.28 677.28
## - EducationField    5    619.36 677.36
## - Gender             1    611.75 677.75
## + Age               1    607.82 677.82
## - StockOptionLevel  1    611.97 677.97
## + YearsAtCompany    1    608.11 678.11
## - PerformanceRating  1    612.24 678.24
## + DailyRate         1    608.71 678.71
## + YearsWithCurrManager 1    608.88 678.88
## + PercentSalaryHike  1    608.89 678.89
## + JobLevel          1    609.18 679.18
## + HourlyRate        1    609.25 679.25
## + Education          1    609.26 679.26

```

```

## + MonthlyIncome          1  609.27 679.27
## - MaritalStatus          2  615.65 679.65
## + Department             2  607.78 679.78
## - RelationshipSatisfaction 1  615.21 681.21
## - YearsInCurrentRole     1  618.06 684.06
## - DistanceFromHome       1  618.92 684.92
## - TotalWorkingYears      1  619.59 685.59
## - YearsSinceLastPromotion 1  620.11 686.11
## - WorkLifeBalance        1  620.15 686.15
## - NumCompaniesWorked     1  620.19 686.19
## - JobInvolvement         1  621.64 687.64
## - EnvironmentSatisfaction 1  626.61 692.61
## - JobSatisfaction        1  627.56 693.56
## - BusinessTravel         2  632.66 696.66
## - JobRole                8  652.61 704.61
## - OverTime               1  681.53 747.53
##
## Step: AIC=676.89
## Attrition ~ BusinessTravel + DistanceFromHome + EducationField +
##      EnvironmentSatisfaction + Gender + JobInvolvement + JobRole +
##      JobSatisfaction + MaritalStatus + NumCompaniesWorked + OverTime +
##      PerformanceRating + RelationshipSatisfaction + StockOptionLevel +
##      TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
##      YearsInCurrentRole + YearsSinceLastPromotion
##
##              Df Deviance   AIC
## - EducationField      5  620.88 676.88
## <none>                 610.89 676.89
## - TrainingTimesLastYear 1  612.98 676.98
## - Gender               1  613.20 677.20
## + MonthlyRate          1  609.28 677.28
## + Age                  1  609.37 677.37
## + YearsAtCompany       1  609.87 677.87
## - StockOptionLevel     1  613.88 677.88
## - PerformanceRating    1  614.18 678.18
## + DailyRate            1  610.31 678.31
## + YearsWithCurrManager 1  610.38 678.38
## + PercentSalaryHike    1  610.65 678.65
## + JobLevel             1  610.75 678.75
## + HourlyRate           1  610.85 678.85
## + MonthlyIncome        1  610.88 678.88
## + Education            1  610.89 678.89
## - MaritalStatus        2  617.20 679.20
## + Department           2  609.39 679.39
## - RelationshipSatisfaction 1  616.90 680.90
## - YearsInCurrentRole   1  620.03 684.03
## - TotalWorkingYears    1  621.06 685.06
## - DistanceFromHome     1  621.17 685.17
## - YearsSinceLastPromotion 1  621.51 685.51
## - NumCompaniesWorked   1  621.77 685.77

```

```

## - WorkLifeBalance      1  621.78 685.78
## - JobInvolvement       1  623.11 687.11
## - EnvironmentSatisfaction 1  627.65 691.65
## - JobSatisfaction       1  629.19 693.19
## - BusinessTravel       2  634.18 696.18
## - JobRole              8  653.70 703.70
## - OverTime             1  683.35 747.35
##
## Step: AIC=676.88
## Attrition ~ BusinessTravel + DistanceFromHome + EnvironmentSatisfaction +
##      Gender + JobInvolvement + JobRole + JobSatisfaction + MaritalStatus +
##      NumCompaniesWorked + OverTime + PerformanceRating +
RelationshipSatisfaction +
##      StockOptionLevel + TotalWorkingYears + TrainingTimesLastYear +
##      WorkLifeBalance + YearsInCurrentRole + YearsSinceLastPromotion
##
##              Df Deviance    AIC
## <none>              620.88 676.88
## + EducationField    5  610.89 676.89
## + MonthlyRate       1  619.36 677.36
## - Gender            1  623.45 677.45
## - TrainingTimesLastYear 1  623.53 677.53
## + Age              1  619.59 677.59
## + YearsWithCurrManager 1  620.00 678.00
## + DailyRate        1  620.03 678.03
## - StockOptionLevel  1  624.17 678.17
## - PerformanceRating 1  624.23 678.23
## + YearsAtCompany    1  620.38 678.38
## + PercentSalaryHike  1  620.46 678.46
## + JobLevel          1  620.48 678.48
## - MaritalStatus     2  626.66 678.66
## + HourlyRate        1  620.77 678.77
## + MonthlyIncome     1  620.78 678.78
## + Education         1  620.85 678.85
## + Department        2  619.24 679.24
## - RelationshipSatisfaction 1  626.95 680.95
## - YearsInCurrentRole 1  629.24 683.24
## - WorkLifeBalance   1  630.94 684.94
## - YearsSinceLastPromotion 1  630.98 684.98
## - DistanceFromHome  1  631.00 685.00
## - TotalWorkingYears  1  631.27 685.27
## - NumCompaniesWorked 1  631.67 685.67
## - JobInvolvement    1  633.48 687.48
## - EnvironmentSatisfaction 1  636.69 690.69
## - JobSatisfaction   1  639.04 693.04
## - BusinessTravel    2  644.84 696.84
## - JobRole           8  671.13 711.13
## - OverTime          1  692.20 746.20

```

```
summary(model_2)
```

```
##
## Call:
## glm(formula = Attrition ~ BusinessTravel + DistanceFromHome +
##     EnvironmentSatisfaction + Gender + JobInvolvement + JobRole +
##     JobSatisfaction + MaritalStatus + NumCompaniesWorked + OverTime +
##     PerformanceRating + RelationshipSatisfaction + StockOptionLevel +
##     TotalWorkingYears + TrainingTimesLastYear + WorkLifeBalance +
##     YearsInCurrentRole + YearsSinceLastPromotion, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.7010  -0.4917  -0.2623  -0.0993   3.8093
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   2.00424    1.48474   1.350 0.177051
## BusinessTravelTravel_Frequently 2.04122    0.53275   3.831 0.000127 ***
## BusinessTravelTravel_Rarely    1.09637    0.50460   2.173 0.029800 *
## DistanceFromHome               0.04036    0.01265   3.190 0.001422 **
## EnvironmentSatisfaction        -0.37848    0.09630  -3.930 8.49e-05 ***
## GenderMale                     0.34154    0.21435   1.593 0.111076
## JobInvolvement                 -0.51003    0.14446  -3.531 0.000414 ***
## JobRoleHuman Resources         1.94824    0.75083   2.595 0.009465 **
## JobRoleLaboratory Technician  2.10220    0.59416   3.538 0.000403 ***
## JobRoleManager                 1.15109    0.81172   1.418 0.156165
## JobRoleManufacturing Director  0.76195    0.70332   1.083 0.278646
## JobRoleResearch Director      -0.65599    1.21877  -0.538 0.590413
## JobRoleResearch Scientist      0.86667    0.60489   1.433 0.151919
## JobRoleSales Executive         1.77094    0.58694   3.017 0.002551 **
## JobRoleSales Representative    2.82936    0.65114   4.345 1.39e-05 ***
## JobSatisfaction               -0.40153    0.09558  -4.201 2.66e-05 ***
## MaritalStatusMarried          0.27415    0.31461   0.871 0.383545
## MaritalStatusSingle           0.86616    0.39500   2.193 0.028323 *
## NumCompaniesWorked            0.14774    0.04445   3.324 0.000887 ***
## OverTimeYes                   1.86137    0.23053   8.074 6.79e-16 ***
## PerformanceRating             -0.58703    0.33093  -1.774 0.076083 .
## RelationshipSatisfaction       -0.23271    0.09473  -2.456 0.014031 *
## StockOptionLevel              -0.32569    0.18334  -1.776 0.075664 .
## TotalWorkingYears             -0.07367    0.02384  -3.090 0.001999 **
## TrainingTimesLastYear         -0.13615    0.08428  -1.615 0.106218
## WorkLifeBalance               -0.45746    0.14461  -3.163 0.001559 **
## YearsInCurrentRole            -0.12571    0.04450  -2.825 0.004730 **
## YearsSinceLastPromotion        0.14330    0.04499   3.185 0.001446 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 909.34  on 1028  degrees of freedom
```

```
## Residual deviance: 620.88  on 1001  degrees of freedom
## AIC: 676.88
##
## Number of Fisher Scoring iterations: 6
```

VIF:

We can use variance inflation factor (vif) to get rid of redundant predictors or the variables that have high multicollinearity between them. Multicollinearity exists when two or more predictor variables are highly related to each other and then it becomes difficult to understand the impact of an independent variable on the dependent variable. The Variance Inflation Factor(VIF) is used to measure the multicollinearity between predictor variables in a model. A predictor having a VIF of 5 or less is generally considered safe and it can be assumed that it is not correlated with other predictor variables. Higher the VIF, greater is the correlation of the predictor variable w.r.t other predictor variables. However, Predictors with high VIF may have high p-value(or highly significant), hence, we need to see the significance of the Predictor variable before removing it from our model.

```
library(car)
```

```
##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##      recode
```

```
vif(model_2)
```

	GVIF	Df	GVIF^(1/(2*Df))
## BusinessTravel	1.159661	2	1.037726
## DistanceFromHome	1.109299	1	1.053232
## EnvironmentSatisfaction	1.058276	1	1.028726
## Gender	1.038340	1	1.018990
## JobInvolvement	1.028626	1	1.014212
## JobRole	2.056495	8	1.046093
## JobSatisfaction	1.066460	1	1.032696
## MaritalStatus	2.032717	2	1.194041
## NumCompaniesWorked	1.260061	1	1.122524
## OverTime	1.240102	1	1.113598
## PerformanceRating	1.037761	1	1.018706
## RelationshipSatisfaction	1.072754	1	1.035738
## StockOptionLevel	1.928597	1	1.388739
## TotalWorkingYears	2.257093	1	1.502363
## TrainingTimesLastYear	1.062476	1	1.030765
## WorkLifeBalance	1.074822	1	1.036736
## YearsInCurrentRole	1.789946	1	1.337889
## YearsSinceLastPromotion	1.816638	1	1.347827

```
final_model <- model_2
```

Accuracy

```
prob_pred=predict(final_model,type='response', newdata = validation[-2])
y_pred = ifelse(prob_pred>0.5,"Yes","No")

accuracy <- table(y_pred, validation[, "Attrition"])
accuracy

##
## y_pred False True
##    No    364   42
##    Yes     6   29

sum(diag(accuracy))/sum(accuracy)

## [1] 0.8911565

res=predict(final_model,train, type="response")

library(ROCR)

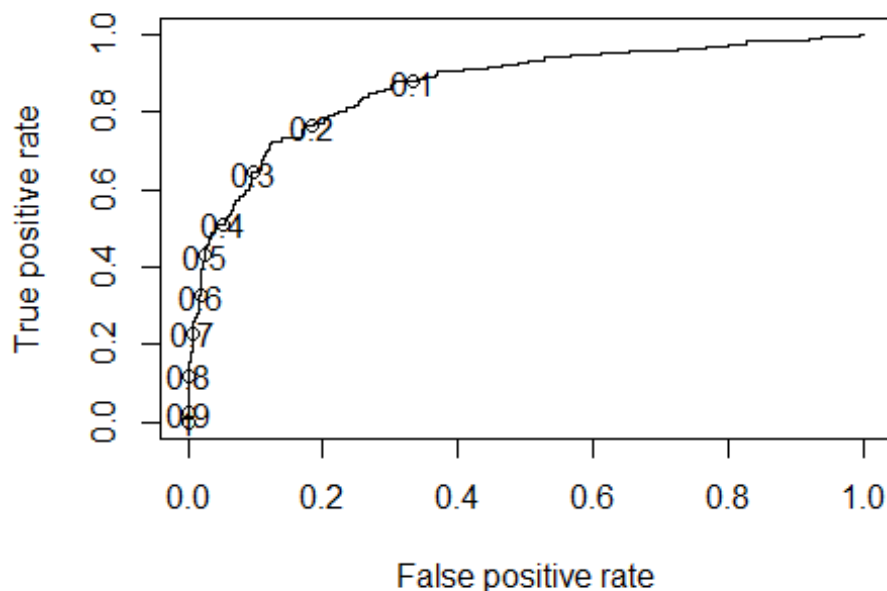
## Warning: package 'ROCR' was built under R version 3.4.3

## Loading required package: gplots

##
## Attaching package: 'gplots'

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

ROCRPred = prediction(res,train$Attrition)
ROCRPref<- performance(ROCRPred,"tpr","fpr")
plot(ROCRPref, colorsize=TRUE,print.cutoffs.at=seq(0.1, by=0.1))
```



```
prob_pred1=predict(final_model,type='response', newdata = validation[-2])
y_pred1 = ifelse(prob_pred1>0.2,"Yes","No")

accuracy1 <- table(y_pred1, validation[, "Attrition"])
accuracy1

##
## y_pred1 False True
##      No    306   18
##      Yes    64   53

sum(diag(accuracy1))/sum(accuracy1)

## [1] 0.814059
```

MODEL BUILDING 2 Decision Tree- Splits the data into multiple sets and each set is further split into subsets to arrive at a tree like structure and make a decision. Homogeneity is the basic concept that helps to determine the attribute on which a split should be made. A split that results into the most homogenous subset is often considered better and step by step each attribute is chosen that maximizes the homogeneity of each subset. Further, this homogeneity is measured using different ways such as Gini Index, Entropy and Information Gain. Hide

```
set.seed(123)
df1$Attrition <- as.factor(df1$Attrition)
indices = sample.split(df1$Attrition, SplitRatio = 0.7)
train = df1[indices,]
```



```
validation = df1[!(indices),]
head(validation)
```

```
##      Age Attrition      BusinessTravel DailyRate      Department
## 2    49      False Travel_Frequently      279 Research & Development
## 3    37       True  Travel_Rarely      1373 Research & Development
## 5    27      False  Travel_Rarely      591 Research & Development
## 11   35      False  Travel_Rarely      809 Research & Development
## 13   31      False  Travel_Rarely      670 Research & Development
## 14   34      False  Travel_Rarely      1346 Research & Development
##      DistanceFromHome Education EducationField EmployeeCount EmployeeNumber
## 2              8          1 Life Sciences          1             2
## 3              2          2      Other          1             4
## 5              2          1      Medical          1             7
## 11             16          3      Medical          1            14
## 13             26          1 Life Sciences          1            16
## 14             19          2      Medical          1            18
##      EnvironmentSatisfaction Gender HourlyRate JobInvolvement JobLevel
## 2              3      Male      61          2          2
## 3              4      Male      92          2          1
## 5              1      Male      40          3          1
## 11             1      Male      84          4          1
## 13             1      Male      31          3          1
## 14             2      Male      93          3          1
##      JobRole JobSatisfaction MaritalStatus MonthlyIncome
## 2    Research Scientist          2      Married      5130
## 3    Laboratory Technician          3      Single      2090
## 5    Laboratory Technician          2      Married      3468
## 11   Laboratory Technician          2      Married      2426
## 13    Research Scientist          3    Divorced      2911
## 14   Laboratory Technician          4    Divorced      2661
##      MonthlyRate NumCompaniesWorked Over18 OverTime PercentSalaryHike
## 2      24907              1      Y      No          23
## 3      2396              6      Y      Yes          15
## 5     16632              9      Y      No          12
## 11     16479              0      Y      No          13
## 13     15170              1      Y      No          17
## 14      8758              0      Y      No          11
##      PerformanceRating RelationshipSatisfaction StandardHours
## 2              4              4          80
## 3              3              2          80
## 5              3              4          80
## 11             3              3          80
## 13             3              4          80
## 14             3              3          80
##      StockOptionLevel TotalWorkingYears TrainingTimesLastYear
## 2              1              10          3
## 3              0              7          3
## 5              1              6          3
## 11             1              6          5
```

```

## 13          1          5          1
## 14          1          3          2
##      WorkLifeBalance YearsAtCompany YearsInCurrentRole
## 2          3          10          7
## 3          3          0          0
## 5          3          2          2
## 11         3          5          4
## 13         2          5          2
## 14         3          2          2
##      YearsSinceLastPromotion YearsWithCurrManager
## 2          1          7
## 3          0          0
## 5          2          2
## 11         0          3
## 13         4          3
## 14         1          2

options(repr.plot.width = 10, repr.plot.height = 8)
library(rpart)
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.4.4

#Training
Dtree = rpart(Attrition ~., data = train, method = "class")
summary(Dtree)

## Call:
## rpart(formula = Attrition ~ ., data = train, method = "class")
##      n= 1029
##
##      CP nsplit rel error   xerror   xstd
## 1 0.03614458      0 1.0000000 1.000000 0.07107939
## 2 0.03012048      3 0.8915663 1.012048 0.07142339
## 3 0.02409639      4 0.8614458 1.006024 0.07125184
## 4 0.01807229      7 0.7891566 1.066265 0.07292739
## 5 0.01000000     13 0.6807229 1.090361 0.07357350
##
## Variable importance
##      JobRole      MonthlyIncome      TotalWorkingYears
##      12      12      10
##      OverTime      JobLevel      MaritalStatus
##      9      8      6
##      Age      DailyRate      Department
##      6      6      5
##      StockOptionLevel      TrainingTimesLastYear      BusinessTravel
##      4      4      3
##      DistanceFromHome      EmployeeNumber      EnvironmentSatisfaction
##      2      2      2
##      YearsAtCompany      PercentSalaryHike      JobInvolvement
##      2      2      1

```

```

##           HourlyRate           MonthlyRate           JobSatisfaction
##                   1                   1                   1
##           EducationField           PerformanceRating           NumCompaniesWorked
##                   1                   1                   1
##
## Node number 1: 1029 observations,      complexity param=0.03614458
##   predicted class=False expected loss=0.1613217 P(node) =1
##   class counts:  863  166
##   probabilities: 0.839 0.161
##   left son=2 (747 obs) right son=3 (282 obs)
##   Primary splits:
##       OverTime           splits as LR,           improve=14.48644, (0
missing)
##       JobRole           splits as LLRLLLLLR,      improve=13.92655, (0
missing)
##       MonthlyIncome      < 2802   to the right, improve=13.85877, (0
missing)
##       TotalWorkingYears < 1.5     to the right, improve=13.53070, (0
missing)
##       JobLevel          < 1.5     to the right, improve=12.38312, (0
missing)
##   Surrogate splits:
##       EmployeeNumber < 22.5      to the right, agree=0.729, adj=0.011, (0
split)
##       DailyRate        < 104.5   to the right, agree=0.727, adj=0.004, (0
split)
##       MonthlyRate      < 26923.5 to the left,  agree=0.727, adj=0.004, (0
split)
##       YearsAtCompany   < 26.5    to the left,  agree=0.727, adj=0.004, (0
split)
##
## Node number 2: 747 observations,      complexity param=0.01807229
##   predicted class=False expected loss=0.1097724 P(node) =0.7259475
##   class counts:  665  82
##   probabilities: 0.890 0.110
##   left son=4 (675 obs) right son=5 (72 obs)
##   Primary splits:
##       TotalWorkingYears < 2.5     to the right, improve=7.964730, (0
missing)
##       YearsAtCompany    < 2.5     to the right, improve=7.802820, (0
missing)
##       YearsWithCurrManager < 0.5   to the right, improve=7.364425, (0
missing)
##       MonthlyIncome     < 2059.5  to the right, improve=6.199431, (0
missing)
##       JobRole           splits as LLRLLLLLR,      improve=5.979504, (0
missing)
##   Surrogate splits:
##       MonthlyIncome < 2009.5  to the right, agree=0.933, adj=0.306, (0
split)

```

```

##      Age          < 21.5    to the right, agree=0.926, adj=0.236, (0
split)
##
## Node number 3: 282 observations,    complexity param=0.03614458
## predicted class=False expected loss=0.2978723 P(node) =0.2740525
##   class counts:   198    84
##   probabilities: 0.702 0.298
##   left son=6 (189 obs) right son=7 (93 obs)
##   Primary splits:
##       MonthlyIncome    < 3751.5  to the right, improve=15.953690, (0
missing)
##       JobLevel         < 1.5     to the right, improve=14.774430, (0
missing)
##       JobRole          splits as LRRLLLLLR,   improve=12.439970, (0
missing)
##       TotalWorkingYears < 1.5     to the right, improve= 7.851383, (0
missing)
##       Age              < 26.5    to the right, improve= 6.737057, (0
missing)
##   Surrogate splits:
##       JobLevel         < 1.5     to the right, agree=0.933, adj=0.796, (0
split)
##       JobRole          splits as LRRLLLRRLR,   agree=0.865, adj=0.591, (0
split)
##       TotalWorkingYears < 3.5     to the right, agree=0.780, adj=0.333, (0
split)
##       YearsAtCompany   < 2.5     to the right, agree=0.723, adj=0.161, (0
split)
##       Age              < 23.5    to the right, agree=0.716, adj=0.140, (0
split)
##
## Node number 4: 675 observations
## predicted class=False expected loss=0.08592593 P(node) =0.6559767
##   class counts:   617    58
##   probabilities: 0.914 0.086
##
## Node number 5: 72 observations,    complexity param=0.01807229
## predicted class=False expected loss=0.3333333 P(node) =0.06997085
##   class counts:    48    24
##   probabilities: 0.667 0.333
##   left son=10 (62 obs) right son=11 (10 obs)
##   Primary splits:
##       BusinessTravel splits as LRL,           improve=5.058065, (0 missing)
##       JobRole        splits as -RR---L-R,      improve=4.500000, (0 missing)
##       HourlyRate     < 58.5    to the right, improve=4.266667, (0 missing)
##       DailyRate      < 258.5   to the right, improve=4.254945, (0 missing)
##       MaritalStatus  splits as LLR,           improve=2.427245, (0 missing)
##   Surrogate splits:
##       EmployeeNumber < 1901.5  to the left,  agree=0.889, adj=0.2, (0
split)

```

```

##      MonthlyRate      < 26306.5 to the left,  agree=0.875, adj=0.1, (0
split)
##
## Node number 6: 189 observations,      complexity param=0.02409639
## predicted class=False expected loss=0.1798942 P(node) =0.1836735
## class counts:  155   34
## probabilities: 0.820 0.180
## left son=12 (114 obs) right son=13 (75 obs)
## Primary splits:
##      JobRole           splits as  LLRLLLLRL,   improve=4.881582, (0
missing)
##      MaritalStatus     splits as  LLR,         improve=4.654420, (0
missing)
##      StockOptionLevel  < 0.5      to the right, improve=4.562328, (0
missing)
##      Department        splits as  LLR,         improve=3.841677, (0
missing)
##      EducationField    splits as  LLRLLL,      improve=3.565686, (0
missing)
## Surrogate splits:
##      Department        splits as  LLR,         agree=0.884, adj=0.707, (0
split)
##      EducationField    splits as  LLRLLL,      agree=0.698, adj=0.240, (0
split)
##      Age               < 28.5    to the right, agree=0.651, adj=0.120, (0
split)
##      MonthlyIncome     < 5841.5  to the right, agree=0.646, adj=0.107, (0
split)
##      TotalWorkingYears < 8.5     to the right, agree=0.640, adj=0.093, (0
split)
##
## Node number 7: 93 observations,      complexity param=0.03614458
## predicted class=True  expected loss=0.4623656 P(node) =0.09037901
## class counts:  43   50
## probabilities: 0.462 0.538
## left son=14 (47 obs) right son=15 (46 obs)
## Primary splits:
##      JobRole           splits as  -LR---L-R,   improve=4.545532, (0
missing)
##      EnvironmentSatisfaction < 1.5    to the right, improve=4.398897, (0
missing)
##      Age               < 33.5    to the right, improve=3.185902, (0
missing)
##      MonthlyIncome     < 2124    to the right, improve=2.876944, (0
missing)
##      NumCompaniesWorked < 0.5     to the left,  improve=2.626303, (0
missing)
## Surrogate splits:
##      Department        splits as  LLR,         agree=0.699, adj=0.391, (0
split)

```

```

##      EmployeeNumber    < 674      to the right, agree=0.624, adj=0.239, (0
split)
##      TotalWorkingYears < 1.5      to the right, agree=0.624, adj=0.239, (0
split)
##      Age                < 25      to the right, agree=0.591, adj=0.174, (0
split)
##      DailyRate         < 1285     to the left,  agree=0.591, adj=0.174, (0
split)
##
## Node number 10: 62 observations,    complexity param=0.01807229
##   predicted class=False expected loss=0.2580645 P(node) =0.06025267
##   class counts:    46    16
##   probabilities: 0.742 0.258
##   left son=20 (51 obs) right son=21 (11 obs)
##   Primary splits:
##       DailyRate      < 343.5     to the right, improve=3.827497, (0
missing)
##       HourlyRate     < 58.5      to the right, improve=3.021843, (0
missing)
##       JobRole        splits as  -RR---L-R,    improve=2.391058, (0
missing)
##       StockOptionLevel < 0.5      to the right, improve=2.391058, (0
missing)
##       MaritalStatus  splits as  LLR,          improve=2.341935, (0
missing)
##   Surrogate splits:
##       MonthlyIncome < 1162.5     to the right, agree=0.839, adj=0.091, (0
split)
##
## Node number 11: 10 observations
##   predicted class=True  expected loss=0.2 P(node) =0.009718173
##   class counts:        2        8
##   probabilities: 0.200 0.800
##
## Node number 12: 114 observations
##   predicted class=False expected loss=0.0877193 P(node) =0.1107872
##   class counts:    104    10
##   probabilities: 0.912 0.088
##
## Node number 13: 75 observations,    complexity param=0.02409639
##   predicted class=False expected loss=0.32 P(node) =0.0728863
##   class counts:     51    24
##   probabilities: 0.680 0.320
##   left son=26 (49 obs) right son=27 (26 obs)
##   Primary splits:
##       MaritalStatus  splits as  LLR,          improve=8.870769, (0
missing)
##       StockOptionLevel < 0.5      to the right, improve=6.518571, (0
missing)
##       MonthlyIncome  < 5234.5     to the left,  improve=3.681667, (0

```

```

missing)
## DistanceFromHome < 23.5 to the left, improve=3.332308, (0
missing)
## Gender splits as LR, improve=2.182735, (0
missing)
## Surrogate splits:
## StockOptionLevel < 0.5 to the right, agree=0.880,
adj=0.654, (0 split)
## HourlyRate < 82.5 to the left, agree=0.720,
adj=0.192, (0 split)
## EmployeeNumber < 68 to the right, agree=0.693,
adj=0.115, (0 split)
## NumCompaniesWorked < 7.5 to the left, agree=0.693,
adj=0.115, (0 split)
## TrainingTimesLastYear < 1.5 to the right, agree=0.693,
adj=0.115, (0 split)
##
## Node number 14: 47 observations, complexity param=0.03012048
## predicted class=False expected loss=0.3829787 P(node) =0.04567541
## class counts: 29 18
## probabilities: 0.617 0.383
## left son=28 (40 obs) right son=29 (7 obs)
## Primary splits:
## DistanceFromHome < 16 to the left, improve=3.698480, (0
missing)
## EnvironmentSatisfaction < 1.5 to the right, improve=3.470076, (0
missing)
## BusinessTravel splits as -RL, improve=2.693285, (0
missing)
## MonthlyRate < 5384 to the right, improve=2.597381, (0
missing)
## MonthlyIncome < 2469.5 to the right, improve=2.183675, (0
missing)
## Surrogate splits:
## DailyRate < 159.5 to the right, agree=0.894, adj=0.286, (0 split)
##
## Node number 15: 46 observations, complexity param=0.01807229
## predicted class=True expected loss=0.3043478 P(node) =0.0447036
## class counts: 14 32
## probabilities: 0.304 0.696
## left son=30 (20 obs) right son=31 (26 obs)
## Primary splits:
## Age < 33.5 to the right, improve=4.270569, (0
missing)
## TotalWorkingYears < 11 to the right, improve=3.846682, (0
missing)
## EducationField splits as -RRLRR, improve=2.871118, (0
missing)
## YearsSinceLastPromotion < 2.5 to the right, improve=2.774964, (0
missing)

```

```

##      DailyRate          < 1104    to the right, improve=2.263975, (0
missing)
##  Surrogate splits:
##      TotalWorkingYears < 9        to the right, agree=0.783, adj=0.50, (0
split)
##      DailyRate          < 1315    to the right, agree=0.674, adj=0.25, (0
split)
##      PercentSalaryHike < 19.5     to the right, agree=0.674, adj=0.25, (0
split)
##      PerformanceRating < 3.5      to the right, agree=0.674, adj=0.25, (0
split)
##      StockOptionLevel  < 0.5      to the right, agree=0.674, adj=0.25, (0
split)
##
## Node number 20: 51 observations
##  predicted class=False  expected loss=0.1764706  P(node) =0.04956268
##  class counts:    42    9
##  probabilities: 0.824 0.176
##
## Node number 21: 11 observations
##  predicted class=True   expected loss=0.3636364  P(node) =0.01068999
##  class counts:     4    7
##  probabilities: 0.364 0.636
##
## Node number 26: 49 observations
##  predicted class=False  expected loss=0.1428571  P(node) =0.04761905
##  class counts:    42    7
##  probabilities: 0.857 0.143
##
## Node number 27: 26 observations,    complexity param=0.02409639
##  predicted class=True   expected loss=0.3461538  P(node) =0.02526725
##  class counts:     9   17
##  probabilities: 0.346 0.654
##  left son=54 (8 obs) right son=55 (18 obs)
##  Primary splits:
##      TrainingTimesLastYear < 2.5    to the right, improve=3.769231, (0
missing)
##      MonthlyIncome          < 5791   to the left,  improve=3.211655, (0
missing)
##      JobLevel                < 2.5    to the left,  improve=2.769231, (0
missing)
##      YearsSinceLastPromotion < 2.5    to the left,  improve=2.295547, (0
missing)
##      DistanceFromHome        < 6.5    to the left,  improve=1.054945, (0
missing)
##  Surrogate splits:
##      Department             splits as -LR,          agree=0.808, adj=0.375, (0
split)
##      JobRole                 splits as --L----R-,    agree=0.808, adj=0.375, (0
split)

```



```

##      JobInvolvement < 1.5      to the left,  agree=0.769, adj=0.250, (0
split)
##      JobSatisfaction < 1.5      to the left,  agree=0.769, adj=0.250, (0
split)
##      MonthlyIncome < 5052      to the left,  agree=0.769, adj=0.250, (0
split)
##
## Node number 28: 40 observations,      complexity param=0.01807229
## predicted class=False expected loss=0.3 P(node) =0.03887269
## class counts:      28      12
## probabilities: 0.700 0.300
## left son=56 (33 obs) right son=57 (7 obs)
## Primary splits:
##      EnvironmentSatisfaction < 1.5      to the right, improve=2.912554, (0
missing)
##      MonthlyRate < 5384      to the right, improve=2.912554, (0
missing)
##      DailyRate < 527.5      to the right, improve=2.190313, (0
missing)
##      BusinessTravel splits as -RL,      improve=1.828213, (0
missing)
##      Gender splits as LR,      improve=1.609524, (0
missing)
## Surrogate splits:
##      MonthlyRate < 3251.5      to the right, agree=0.875, adj=0.286,
(0 split)
##      DistanceFromHome < 12      to the left,  agree=0.850, adj=0.143,
(0 split)
##      JobSatisfaction < 1.5      to the right, agree=0.850, adj=0.143,
(0 split)
##      StockOptionLevel < 1.5      to the left,  agree=0.850, adj=0.143,
(0 split)
##      YearsWithCurrManager < 4      to the left,  agree=0.850, adj=0.143,
(0 split)
##
## Node number 29: 7 observations
## predicted class=True expected loss=0.1428571 P(node) =0.006802721
## class counts:      1      6
## probabilities: 0.143 0.857
##
## Node number 30: 20 observations,      complexity param=0.01807229
## predicted class=False expected loss=0.45 P(node) =0.01943635
## class counts:      11      9
## probabilities: 0.550 0.450
## left son=60 (10 obs) right son=61 (10 obs)
## Primary splits:
##      DailyRate < 1121      to the right, improve=2.500000, (0
missing)
##      YearsSinceLastPromotion < 0.5      to the right, improve=2.500000, (0
missing)

```

```

##      PercentSalaryHike      < 14.5      to the right, improve=2.400000, (0
missing)
##      EducationField        splits as  -RRLRL,      improve=2.031868, (0
missing)
##      MaritalStatus         splits as   RLR,        improve=1.697980, (0
missing)
##      Surrogate splits:
##      MaritalStatus         splits as   RLL,        agree=0.85, adj=0.7,
(0 split)
##      PercentSalaryHike      < 17          to the right, agree=0.80, adj=0.6,
(0 split)
##      TrainingTimesLastYear < 2.5          to the right, agree=0.80, adj=0.6,
(0 split)
##      Department            splits as  -RL,        agree=0.70, adj=0.4,
(0 split)
##      JobInvolvement         < 2.5          to the right, agree=0.70, adj=0.4,
(0 split)
##
## Node number 31: 26 observations
##   predicted class=True   expected loss=0.1153846  P(node) =0.02526725
##   class counts:      3    23
##   probabilities: 0.115 0.885
##
## Node number 54: 8 observations
##   predicted class=False  expected loss=0.25   P(node) =0.007774538
##   class counts:      6    2
##   probabilities: 0.750 0.250
##
## Node number 55: 18 observations
##   predicted class=True   expected loss=0.1666667  P(node) =0.01749271
##   class counts:      3    15
##   probabilities: 0.167 0.833
##
## Node number 56: 33 observations
##   predicted class=False  expected loss=0.2121212  P(node) =0.03206997
##   class counts:     26    7
##   probabilities: 0.788 0.212
##
## Node number 57: 7 observations
##   predicted class=True   expected loss=0.2857143  P(node) =0.006802721
##   class counts:      2    5
##   probabilities: 0.286 0.714
##
## Node number 60: 10 observations
##   predicted class=False  expected loss=0.2   P(node) =0.009718173
##   class counts:      8    2
##   probabilities: 0.800 0.200
##
## Node number 61: 10 observations
##   predicted class=True   expected loss=0.3   P(node) =0.009718173

```

```
##      class counts:      3      7
##      probabilities: 0.300 0.700

#Predicting
DTPred <- predict(Dtree,type = "class", newdata = validation[,-2])

library(caret)

## Warning: package 'caret' was built under R version 3.4.4

confusionMatrix(validation$Attrition, DTPred)

## Confusion Matrix and Statistics
##
##              Reference
## Prediction False True
##      False   352   18
##      True     48   23
##
##              Accuracy : 0.8503
##              95% CI : (0.8136, 0.8823)
##      No Information Rate : 0.907
##      P-Value [Acc > NIR] : 0.9999501
##
##              Kappa : 0.332
##      McNemar's Test P-Value : 0.0003575
##
##              Sensitivity : 0.8800
##              Specificity : 0.5610
##      Pos Pred Value : 0.9514
##      Neg Pred Value : 0.3239
##      Prevalence : 0.9070
##      Detection Rate : 0.7982
##      Detection Prevalence : 0.8390
##      Balanced Accuracy : 0.7205
##
##      'Positive' Class : False
##
```

MODEL BUILDING 3: RANDOM FOREST- Often known as an ensemble of a large number of Decision Trees, that uses bootstrapped aggregation technique to choose random samples from a dataset to train each tree in the forest. The final prediction in a RandomForest is an aggregation of prediction of individual trees. One of the advantages of RandomForest is that, it gives out-of-bag(OOB) error estimates, which is the mean prediction error on a training sample, using the trees that do not have that training sample in their bootstrap sample. It may act as a cross validation error and eliminate the need of using test/validation data, thereby increasing the training the data. However, I am still going to use train and validation concept here as well, like I did in the above two Models. Hide

```
library(randomForest)
set.seed(123)
```

```

df1$Attrition <- as.factor(df1$Attrition)
indices = sample.split(df1$Attrition, SplitRatio = 0.7)
train = df1[indices,]
validation = df1[!(indices),]

#Training the RandomForest Model
model.rf <- randomForest(Attrition ~ ., data=train,
proximity=FALSE,importance = FALSE,
                        ntree=500,mtry=4, do.trace=FALSE)
model.rf

##
## Call:
## randomForest(formula = Attrition ~ ., data = train, proximity = FALSE,
importance = FALSE, ntree = 500, mtry = 4, do.trace = FALSE)
##              Type of random forest: classification
##              Number of trees: 500
## No. of variables tried at each split: 4
##
##              OOB estimate of  error rate: 15.26%
## Confusion matrix:
##      False True class.error
## False   858    5 0.005793743
## True    152   14 0.915662651

#Predicting on the validation set and checking the Confusion Matrix.
testPred <- predict(model.rf, newdata=validation[,-2])
table(testPred, validation$Attrition)

##
## testPred False True
##      False   368    2
##      True     62    9

confusionMatrix(validation$Attrition, testPred)

## Confusion Matrix and Statistics
##
##              Reference
## Prediction False True
##      False   368    2
##      True    62    9
##
##              Accuracy : 0.8549
##              95% CI : (0.8185, 0.8864)
##      No Information Rate : 0.9751
##      P-Value [Acc > NIR] : 1
##
##              Kappa : 0.1843
##      Mcnemar's Test P-Value : 1.643e-13
##

```

```
##          Sensitivity : 0.8558
##          Specificity : 0.8182
##          Pos Pred Value : 0.9946
##          Neg Pred Value : 0.1268
##          Prevalence : 0.9751
##          Detection Rate : 0.8345
##          Detection Prevalence : 0.8390
##          Balanced Accuracy : 0.8370
##
##          'Positive' Class : False
##
```

Variable Importance Plot: Below is the variable importance plot, that shows the most significant attribute in decreasing order by mean decrease in Gini. The Mean decrease Gini measures how pure the nodes are at the end of the tree. Higher the Gini Index, better is the homogeneity. Hide

```
#Checking the variable Importance Plot
varImpPlot(model.rf)
```



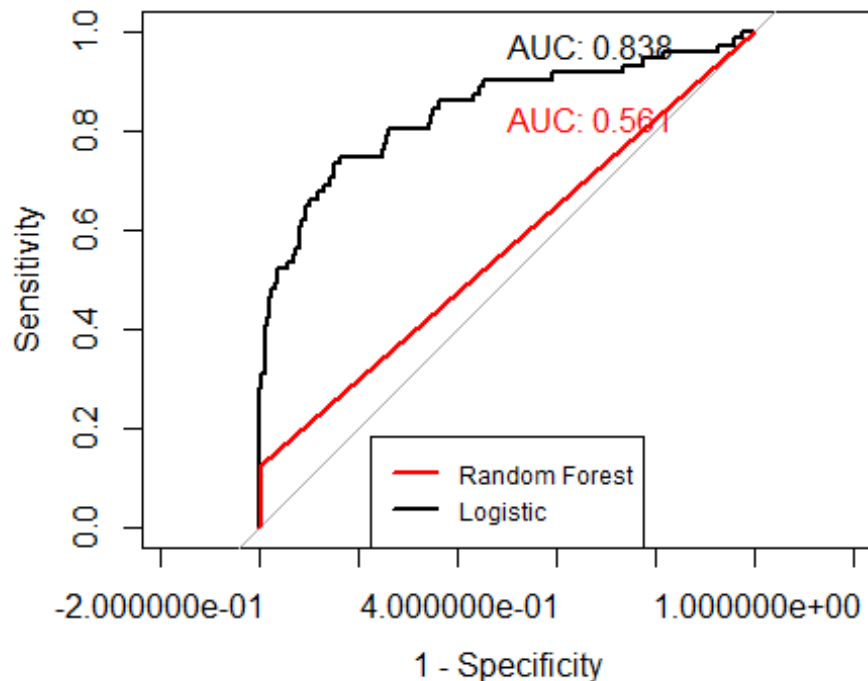
```
library(pROC)

## Warning: package 'pROC' was built under R version 3.4.4
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##      cov, smooth, var

options(repr.plot.width = 10, repr.plot.height = 8)
glm.roc <- roc(response = validation$Attrition, predictor =
as.numeric(prob_pred1))
rf.roc <- roc(response = validation$Attrition, predictor =
as.numeric(testPred))
plot(glm.roc,      legacy.axes = TRUE, print.auc.y = 1.0, print.auc = TRUE)

plot(rf.roc, col = "red" , add = TRUE, print.auc.y = 0.85, print.auc = TRUE)
legend("bottom", c("Random Forest", "Logistic"),
      lty = c(1,1), lwd = c(2, 2), col = c("red", "black"), cex = 0.75)
```



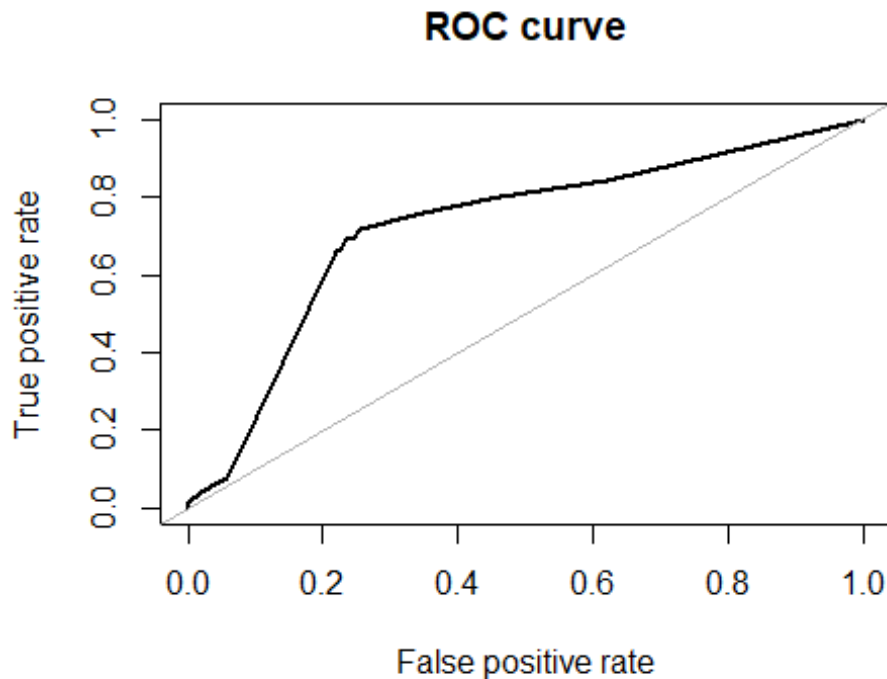
```
library(ROSE)

## Warning: package 'ROSE' was built under R version 3.4.4
## Loaded ROSE 0.0-3

data.rose<-ROSE(Attrition~., data=train,seed=1)$data
table(data.rose$Attrition)

##
## False  True
##   533   496
```

```
library(rpart)
tree.rose <- rpart(Attrition ~ ., data = data.rose)
pred.tree.rose <- predict(tree.rose, newdata = validation)
roc.curve(validation$Attrition, pred.tree.rose[,2])
```



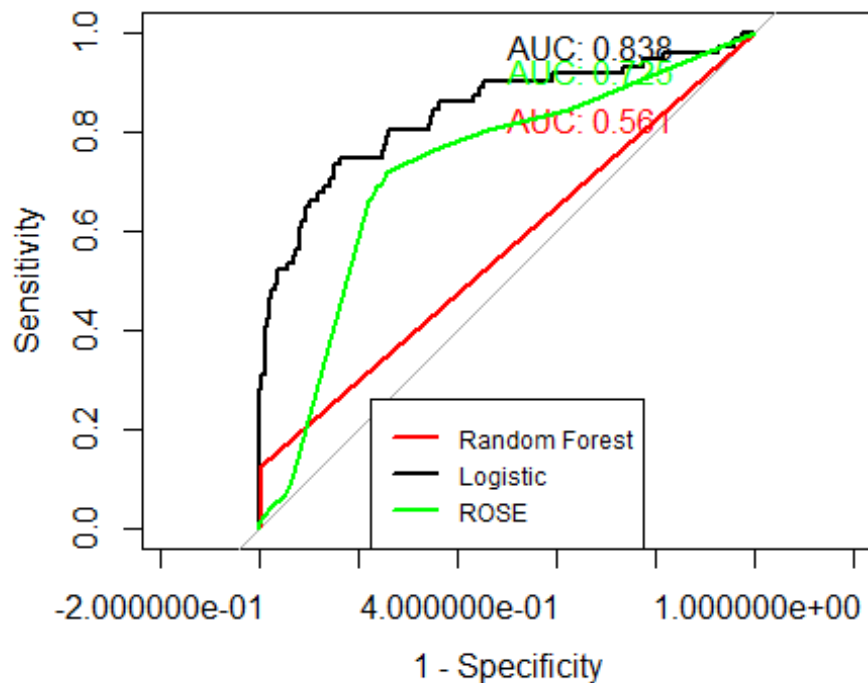
```
## Area under the curve (AUC): 0.725
```

```
library(pROC)
options(repr.plot.width = 10, repr.plot.height = 8)
glm.roc <- roc(response = validation$Attrition, predictor =
as.numeric(prob_pred1))
rf.roc <- roc(response = validation$Attrition, predictor =
as.numeric(testPred))
rose.roc <- roc(response = validation$Attrition, predictor =
as.numeric(pred.tree.rose[,2]))
plot(glm.roc, legacy.axes = TRUE, print.auc.y = 1.0, print.auc = TRUE)

plot(rf.roc, col = "red", add = TRUE, print.auc.y = 0.85, print.auc = TRUE)
plot(rose.roc, col = "green", add = TRUE, print.auc.y = 0.95, print.auc =
TRUE)

legend("bottom", c("Random Forest", "Logistic", "ROSE"),
```

```
lty = c(1,1), lwd = c(2, 2), col = c("red", "black","green"), cex = 0.75)
```



So we can see here ROSE(oversampling) on decision tree increases the performance when compared to normal decision tree but still Logistic regression wins the race by best AUC value.

```
library(survival)

##
## Attaching package: 'survival'

## The following object is masked from 'package:caret':
##
##   cluster

dataNw$YearsAtCompany=as.numeric(dataNw$YearsAtCompany)
dataNw$Attrition=as.numeric(dataNw$Attrition)
dataNw$Age=as.numeric(dataNw$Age)

time = dataNw$YearsAtCompany
event= dataNw$Attrition

mySurv<-Surv(time,event)
class(mySurv)

## [1] "Surv"
```



```

head(mySurv,20) # plus sign means censored data- there is no informaton

## [1] 6 10+ 0 8+ 2+ 7+ 1+ 1+ 9+ 7+ 5+ 9+ 5+ 2+ 4 10+ 6+
## [18] 1+ 25+ 3+

myfit<-survfit(mySurv~dataNW$OverTime)
myfit

## Call: survfit(formula = mySurv ~ dataNW$OverTime)
##
##              n events median 0.95LCL 0.95UCL
## dataNW$OverTime=No 1054    110     40     32     NA
## dataNW$OverTime=Yes 416    127     24     16     NA

survdiff(mySurv~dataNW$OverTime)

## Call:
## survdiff(formula = mySurv ~ dataNW$OverTime)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## dataNW$OverTime=No 1054    110    171.4    22.0    81.7
## dataNW$OverTime=Yes 416    127     65.6    57.5    81.7
##
## Chisq= 81.7 on 1 degrees of freedom, p= 0

summary(myfit)

## Call: survfit(formula = mySurv ~ dataNW$OverTime)
##
##              dataNW$OverTime=No
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    0  1054     6   0.994 0.00232   0.990   0.999
##    1  1024    31   0.964 0.00578   0.953   0.976
##    2   908    12   0.951 0.00677   0.938   0.965
##    3   816     6   0.944 0.00730   0.930   0.959
##    4   728    11   0.930 0.00836   0.914   0.947
##    5   648     7   0.920 0.00909   0.902   0.938
##    6   509     4   0.913 0.00971   0.894   0.932
##    7   457     3   0.907 0.01025   0.887   0.927
##    8   388     3   0.900 0.01094   0.879   0.922
##    9   332     6   0.884 0.01260   0.859   0.909
##   10   269     8   0.857 0.01527   0.828   0.888
##   11   181     1   0.853 0.01590   0.822   0.884
##   13   146     2   0.841 0.01770   0.807   0.876
##   14   126     2   0.828 0.01978   0.790   0.867
##   17    89     1   0.818 0.02163   0.777   0.862
##   18    82     1   0.808 0.02356   0.763   0.856
##   20    65     1   0.796 0.02627   0.746   0.849
##   22    35     1   0.773 0.03397   0.709   0.843
##   23    24     1   0.741 0.04532   0.657   0.835
##   32     8     1   0.648 0.09528   0.486   0.865

```

```
##      33      6      1      0.540 0.12663      0.341      0.855
##      40      1      1      0.000      NaN      NA      NA
##
##                               dataNW$OverTime=Yes
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    0    416     10    0.976 0.00751    0.961    0.991
##    1    402     28    0.908 0.01423    0.881    0.936
##    2    347     15    0.869 0.01684    0.836    0.902
##    3    312     14    0.830 0.01903    0.793    0.868
##    4    272      8    0.805 0.02034    0.766    0.846
##    5    242     14    0.759 0.02265    0.716    0.804
##    6    185      5    0.738 0.02383    0.693    0.786
##    7    161      8    0.702 0.02593    0.653    0.754
##    8    140      6    0.672 0.02757    0.620    0.728
##    9    116      2    0.660 0.02829    0.607    0.718
##   10     97     10    0.592 0.03254    0.531    0.659
##   11     65      1    0.583 0.03329    0.521    0.652
##   15     50      1    0.571 0.03460    0.507    0.643
##   16     43      1    0.558 0.03626    0.491    0.634
##   19     32      1    0.540 0.03909    0.469    0.623
##   21     22      1    0.516 0.04437    0.436    0.611
##   24     13      1    0.476 0.05595    0.378    0.599
##   31      7      1    0.408 0.07916    0.279    0.597
```

```
library(ggplot2)
require("survival")
library(survival)
library(survminer)
```

```
## Warning: package 'survminer' was built under R version 3.4.4
```

```
## Loading required package: ggpubr
```

```
## Warning: package 'ggpubr' was built under R version 3.4.4
```

```
## Loading required package: magrittr
```

```
##
```

```
## Attaching package: 'magrittr'
```

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

```
##
```

```
##      extract
```

```
fit1 <- survfit(mySurv ~ dataNW$OverTime)
summary(fit1)
```

```
## Call: survfit(formula = mySurv ~ dataNW$OverTime)
```

```
##
```

```
##                               dataNW$OverTime=No
```

```
## time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    0   1054      6    0.994 0.00232    0.990    0.999
```

```
##      1  1024      31    0.964 0.00578      0.953      0.976
##      2   908      12    0.951 0.00677      0.938      0.965
##      3   816       6    0.944 0.00730      0.930      0.959
##      4   728      11    0.930 0.00836      0.914      0.947
##      5   648       7    0.920 0.00909      0.902      0.938
##      6   509       4    0.913 0.00971      0.894      0.932
##      7   457       3    0.907 0.01025      0.887      0.927
##      8   388       3    0.900 0.01094      0.879      0.922
##      9   332       6    0.884 0.01260      0.859      0.909
##     10   269       8    0.857 0.01527      0.828      0.888
##     11   181       1    0.853 0.01590      0.822      0.884
##     13   146       2    0.841 0.01770      0.807      0.876
##     14   126       2    0.828 0.01978      0.790      0.867
##     17    89       1    0.818 0.02163      0.777      0.862
##     18    82       1    0.808 0.02356      0.763      0.856
##     20    65       1    0.796 0.02627      0.746      0.849
##     22    35       1    0.773 0.03397      0.709      0.843
##     23    24       1    0.741 0.04532      0.657      0.835
##     32     8       1    0.648 0.09528      0.486      0.865
##     33     6       1    0.540 0.12663      0.341      0.855
##     40     1       1    0.000      NaN      NA      NA
```

```
##
##               dataNW$OverTime=Yes
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    0   416    10    0.976 0.00751    0.961    0.991
##    1   402    28    0.908 0.01423    0.881    0.936
##    2   347    15    0.869 0.01684    0.836    0.902
##    3   312    14    0.830 0.01903    0.793    0.868
##    4   272     8    0.805 0.02034    0.766    0.846
##    5   242    14    0.759 0.02265    0.716    0.804
##    6   185     5    0.738 0.02383    0.693    0.786
##    7   161     8    0.702 0.02593    0.653    0.754
##    8   140     6    0.672 0.02757    0.620    0.728
##    9   116     2    0.660 0.02829    0.607    0.718
##   10    97    10    0.592 0.03254    0.531    0.659
##   11    65     1    0.583 0.03329    0.521    0.652
##   15    50     1    0.571 0.03460    0.507    0.643
##   16    43     1    0.558 0.03626    0.491    0.634
##   19    32     1    0.540 0.03909    0.469    0.623
##   21    22     1    0.516 0.04437    0.436    0.611
##   24    13     1    0.476 0.05595    0.378    0.599
##   31     7     1    0.408 0.07916    0.279    0.597
```

```
ggsurvplot(fit1, data = dataNW, pval = TRUE)
```

