EPABA Batch 2 Group17 HR_Analytics

Vikas Goel, Srikanth Shetty, Nilima Ghosh

Tue Feb 15 18:00:00 2019

```
#Set the working Directory
setwd("C:\\Data\\Learning\\DataScience\\IIMA\\Project\\EPABA2_Grp17_FinalProject_HR_Analytics")

#load the common functions file. This file abstrats all the common functions.
# This file contains  algorithms functions of all the models.
source("Code\\CommonFunctionsVikas.R")

#Load the data
dataOriginal = read.csv("Data\\HR_Analytics.csv", stringsAsFactors = TRUE
)

#make of copy of original Data
data<-dataOriginal

#Understand data
DataFamilarization(dataOriginal)
```

```
## [1] "Size of data : Columns =  10 Rows =  14999"
## 'data.frame':    14999 obs. of  10 variables:
##  $ satisfaction_level    : num  0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...
##  $ last_evaluation       : num  0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...
##  $ number_project        : int  2 5 7 5 2 2 6 5 5 2 ...
##  $ average_montly_hours  : int  157 262 272 223 159 153 247 259 224 142 ...
##  $ time_spend_company    : int  3 6 4 5 3 3 4 5 5 3 ...
##  $ Work_accident         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ left                  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ promotion_last_5years: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Department            : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ salary                : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...
```

```
##
## 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
```

```
## Observations: 14,999
## Variables: 10
## $ satisfaction_level    <dbl> 0.38, 0.80, 0.11, 0.72, 0.37, 0.41, 0.10...
## $ last_evaluation       <dbl> 0.53, 0.86, 0.88, 0.87, 0.52, 0.50, 0.77...
## $ number_project        <int> 2, 5, 7, 5, 2, 2, 6, 5, 5, 2, 2, 6, 4, 2...
## $ average_montly_hours  <int> 157, 262, 272, 223, 159, 153, 247, 259, ...
## $ time_spend_company    <int> 3, 6, 4, 5, 3, 3, 4, 5, 5, 3, 3, 4, 5, 3...
## $ Work_accident         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ left                  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ promotion_last_5years <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Department            <fct> sales, sales, sales, sales, sales, sales...
## $ salary                <fct> low, medium, medium, low, low, low, low,...
```

```
## NULL
```

```
# Compute the Attrition and its frequency
dataAttritionFreq <- dataOriginal
attrition <- as.factor(dataAttritionFreq$left) ; summary(attrition)
```

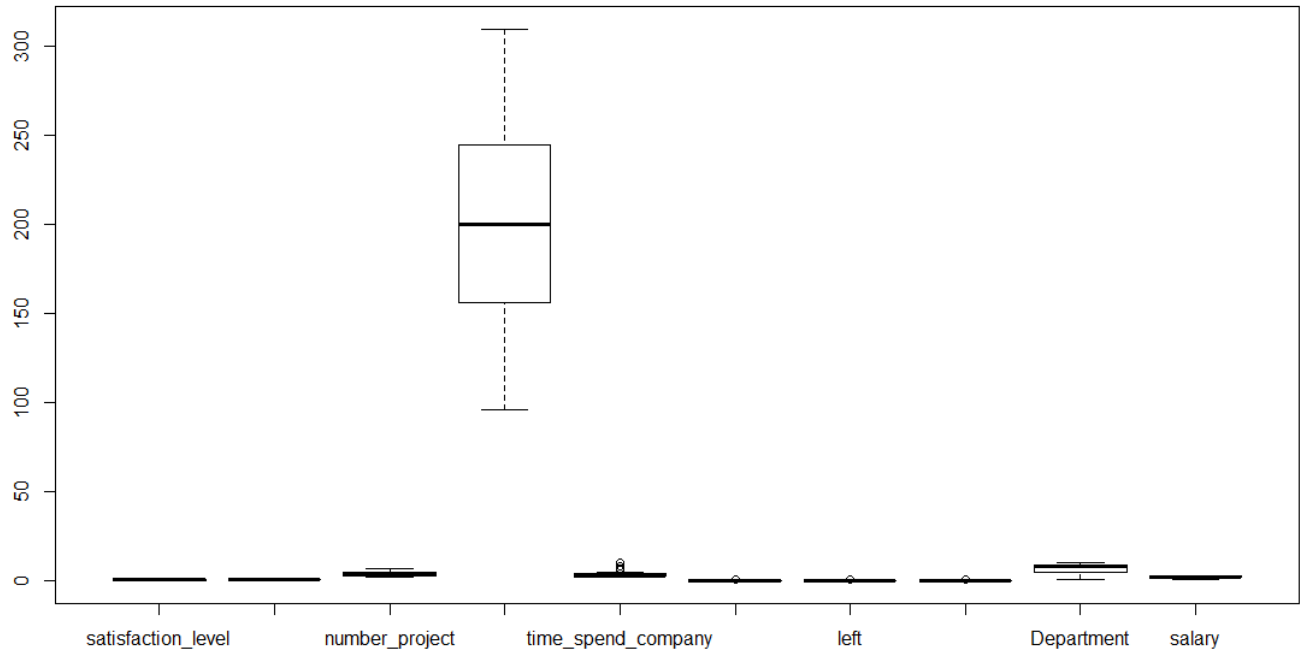
```
##     0     1
## 11428  3571
```

```
AttritionRate <- sum(dataAttritionFreq$left / length(dataAttritionFreq$left)) * 100

print(paste("Attrition Rate = ",round(AttritionRate,2),"%"))
```

```
## [1] "Attrition Rate =  23.81 %"
```

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```



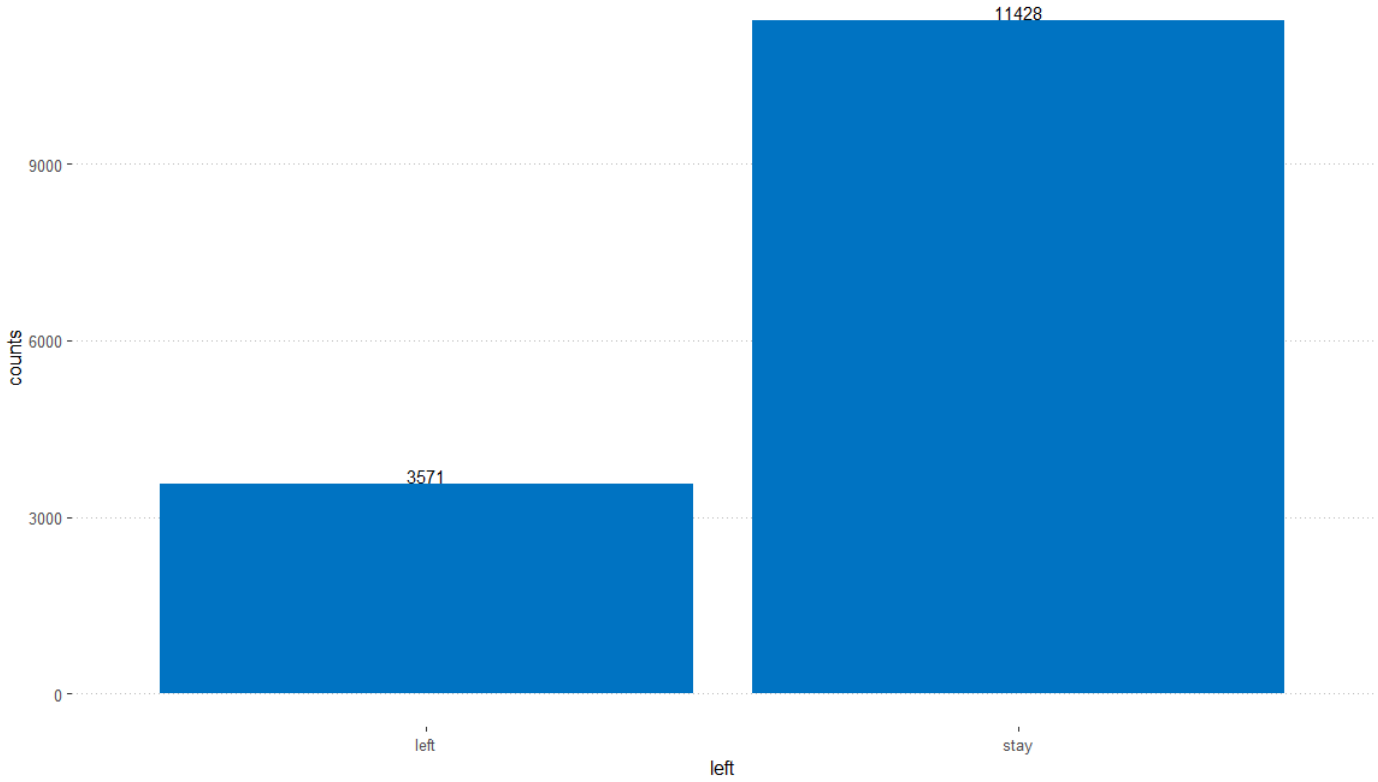

```
library(ggpubr)
```

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

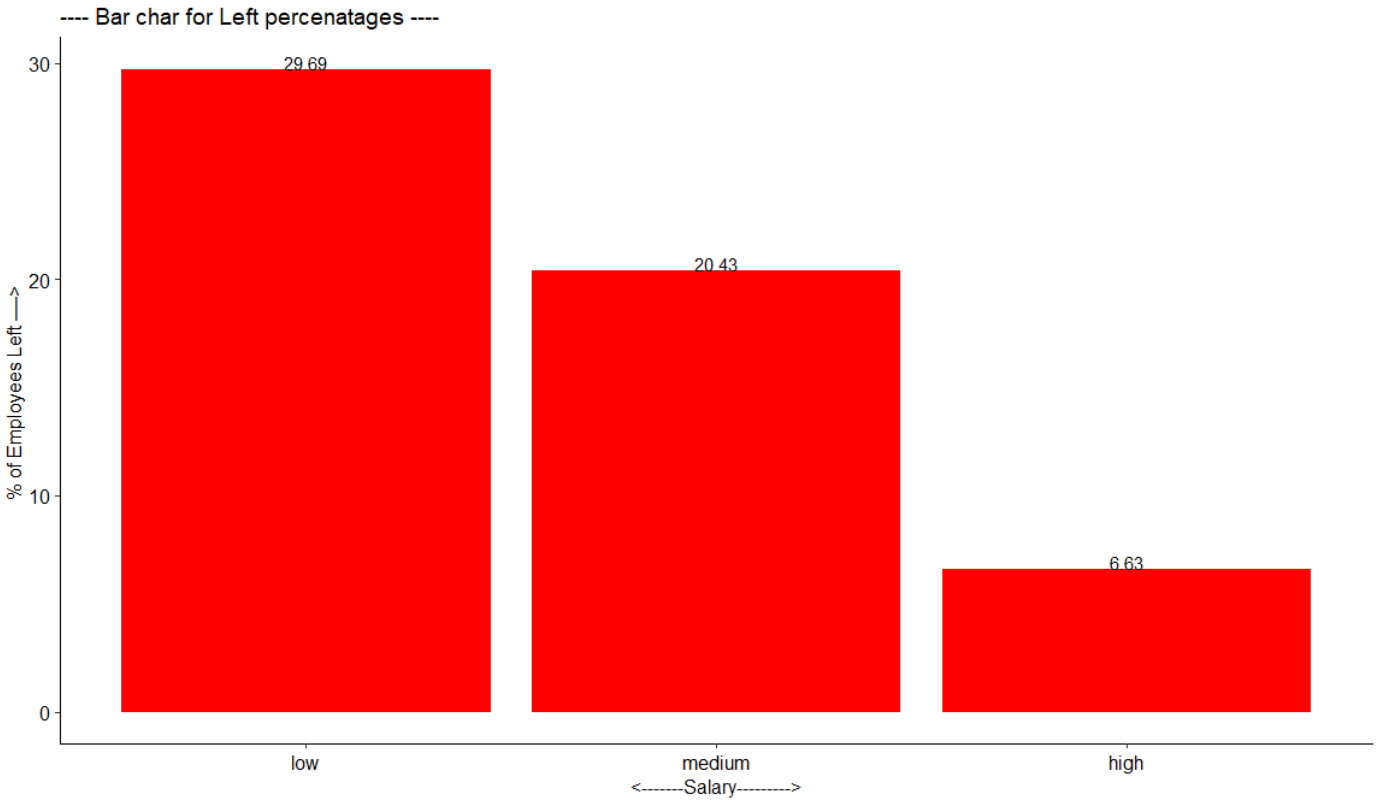
```
library(dplyr)
theme_set(theme_pubr())
dataAttritionFreq$left <- ifelse(dataAttritionFreq$left == '1',"left","stay"
)
df <- dataAttritionFreq %>%
  group_by(left) %>%
%
  summarise(counts = n()
)
df
```

```
## # A tibble: 2 x 2
##   left  counts
##   <chr>  <int>
## 1 left    3571
## 2 stay   11428
```

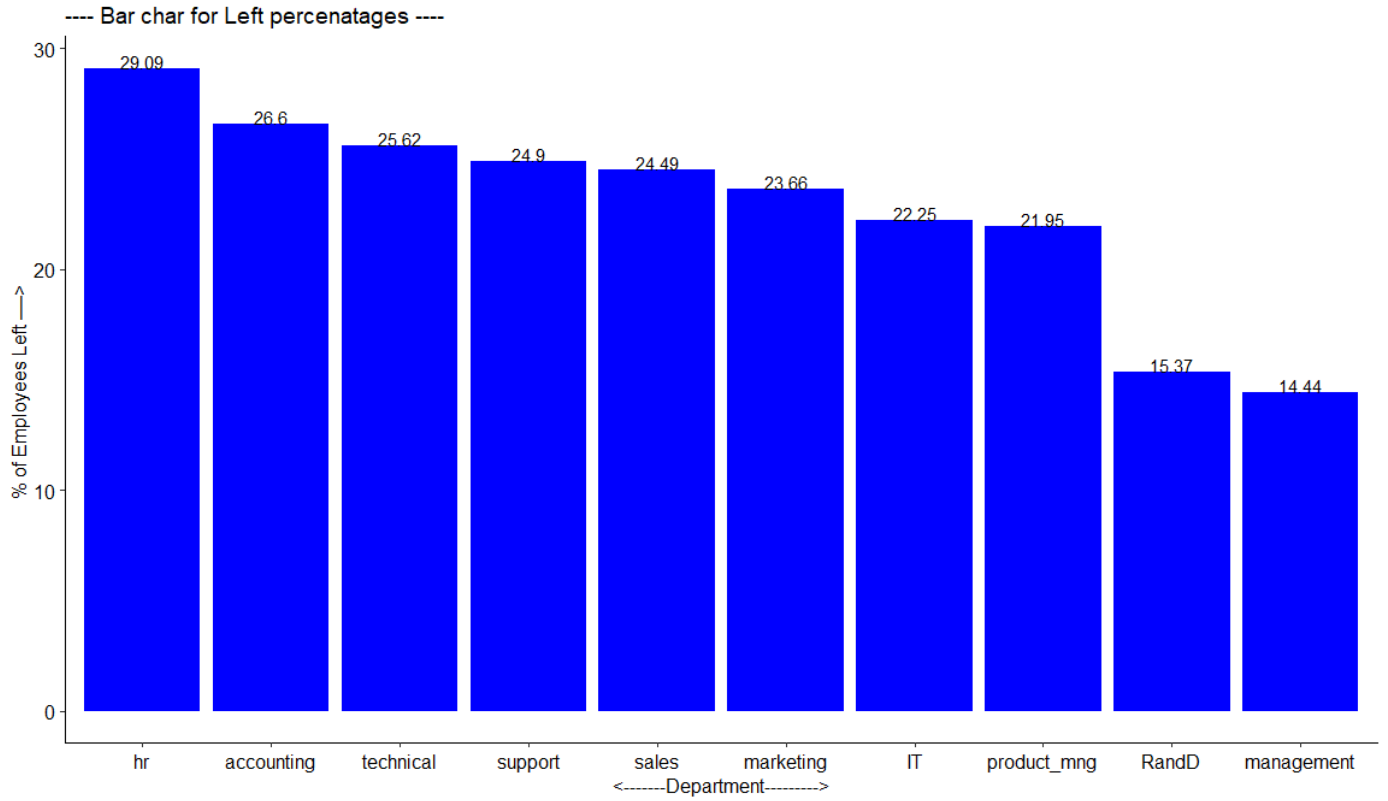
```
ggplot(df, aes(x = left, y = counts))
+
  geom_bar(fill = "#0073C2FF", stat = "identity")
+
  geom_text(aes(label = counts), vjust = -0.1) +

  theme_pubclean()
```



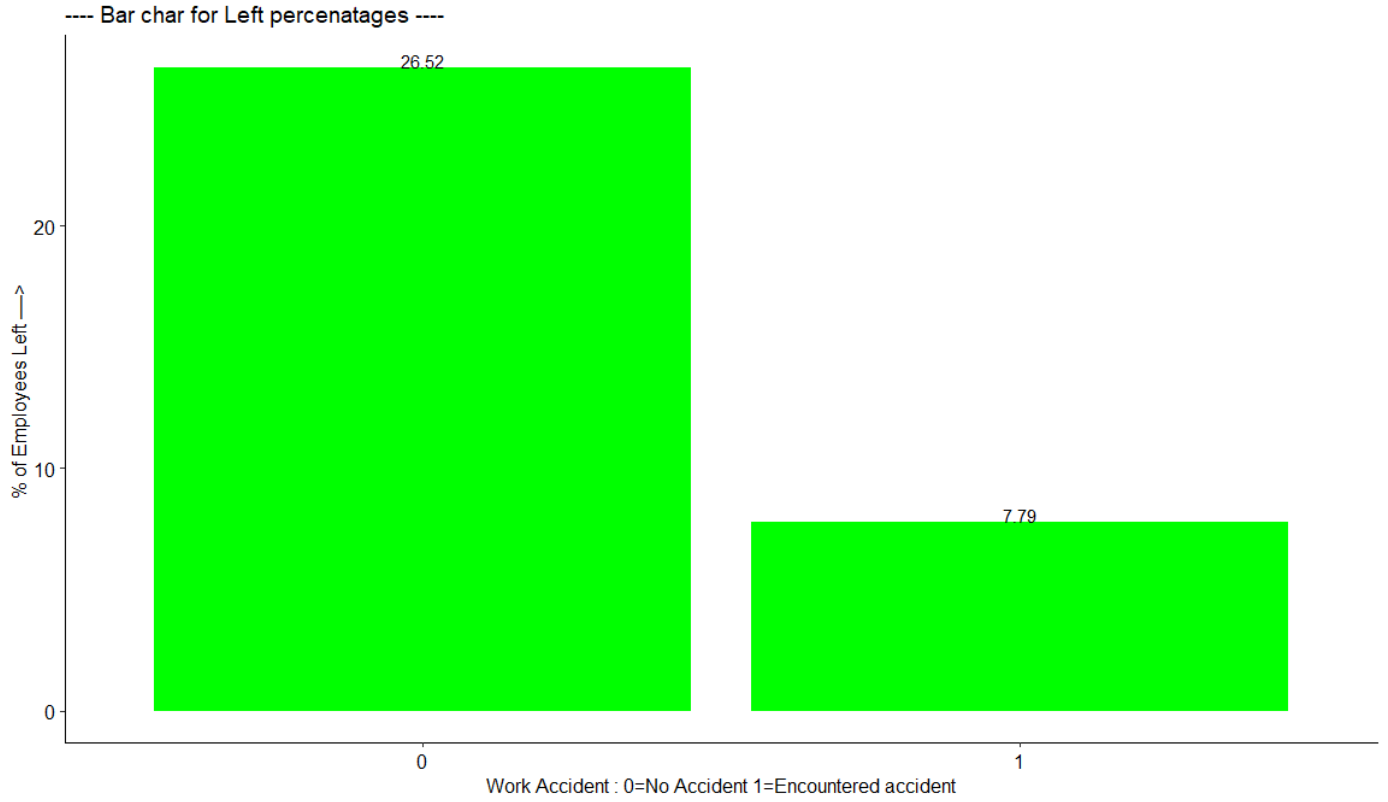
```
#############[   Basic EDA    ]################################################
PercentLeft <- function(X,Y,colnames,color)
{
  library(ggplot2)
  DataWithLeft<-as.data.frame(prop.table(table(X,Y, dnn=c("Var1","Left")), 1)
)
  DataWithLeft <- subset(DataWithLeft, Left==1
)
  names(DataWithLeft) <- c("Var1","Left","PercentLeft"
)
  DataWithLeft$PercentLeft <- round(DataWithLeft$PercentLeft * 100.00,2
)
  ggplot(DataWithLeft, aes(x=reorder(Var1, -PercentLeft),y=PercentLeft,fill=Left))
+
    geom_bar(stat='identity', fill=color)
+
    geom_text(aes(label=PercentLeft), vjust=0) +

    ggtitle("---- Bar char for Left percenatages ----") + xlab(colnames) +

    ylab("% of Employees Left ----> "
)
}
PercentLeft(data$salary,data$left, "<-------Salary--------->", "Red"
)
```



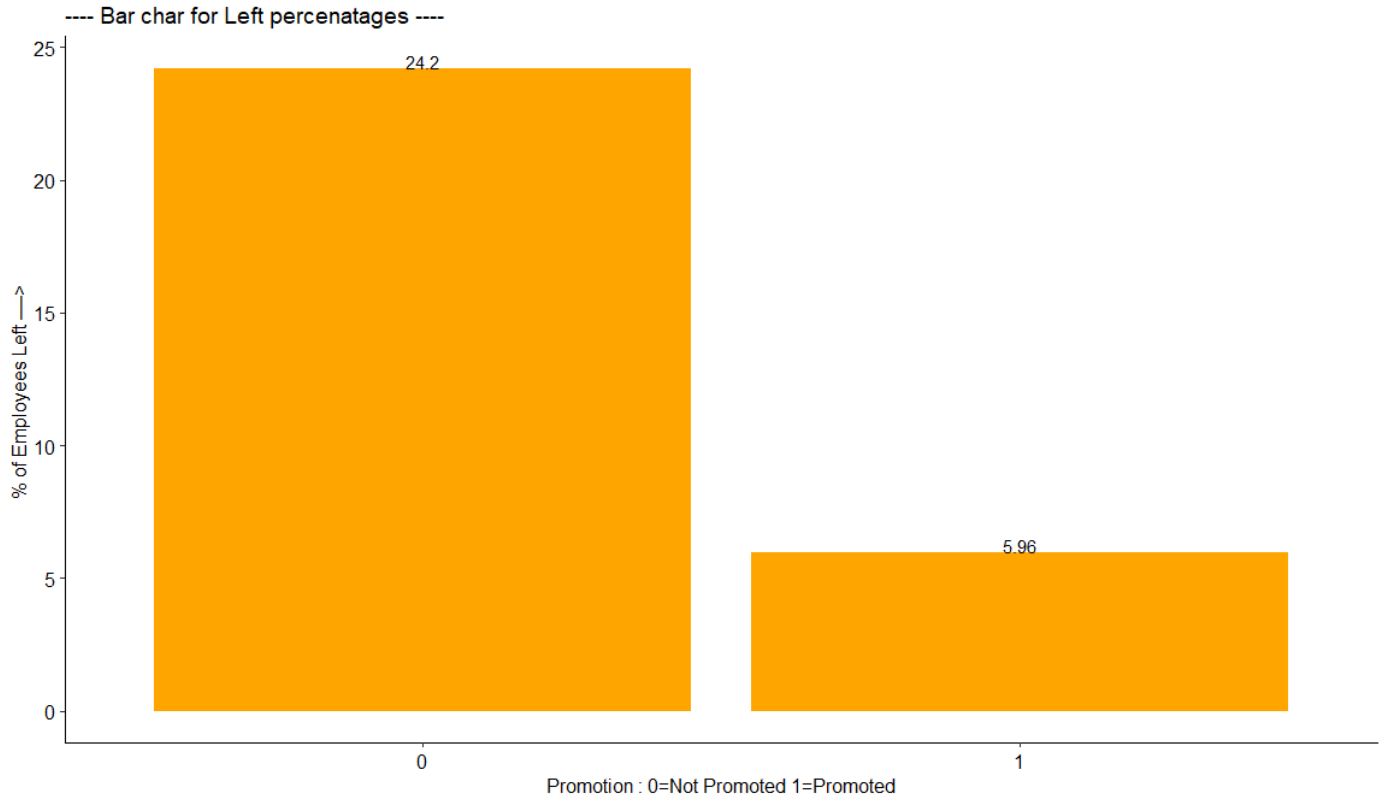

```
PercentLeft(data$Department,data$left, "<-------Department--------->", "Blue"
)
```



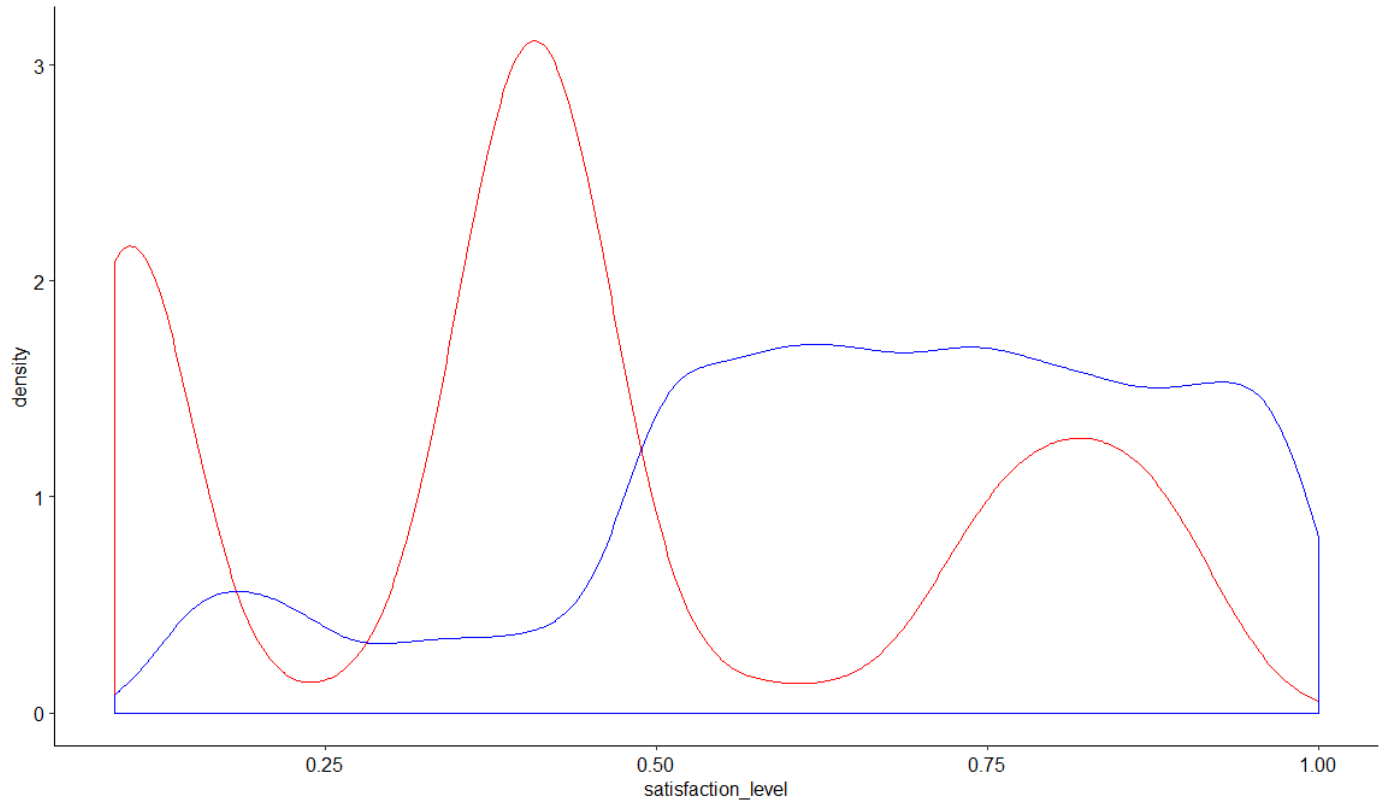

```
PercentLeft(data$Work_accident, data$left,
            "Work Accident : 0=No Accident 1=Encountered accident", "Green"
)
```



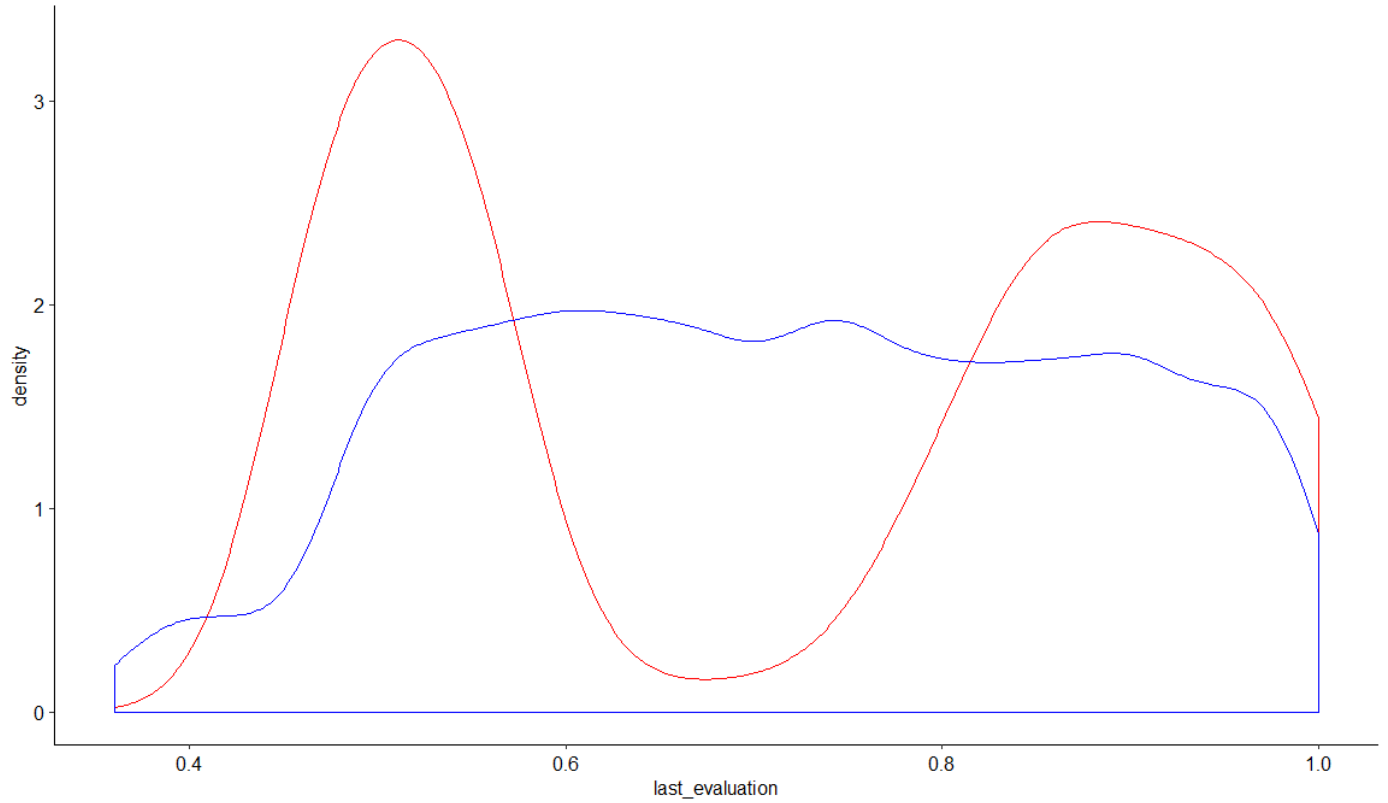

```
PercentLeft(data$promotion_last_5years,  data$left,
            "Promotion : 0=Not Promoted 1=Promoted", "Orange"
)
```



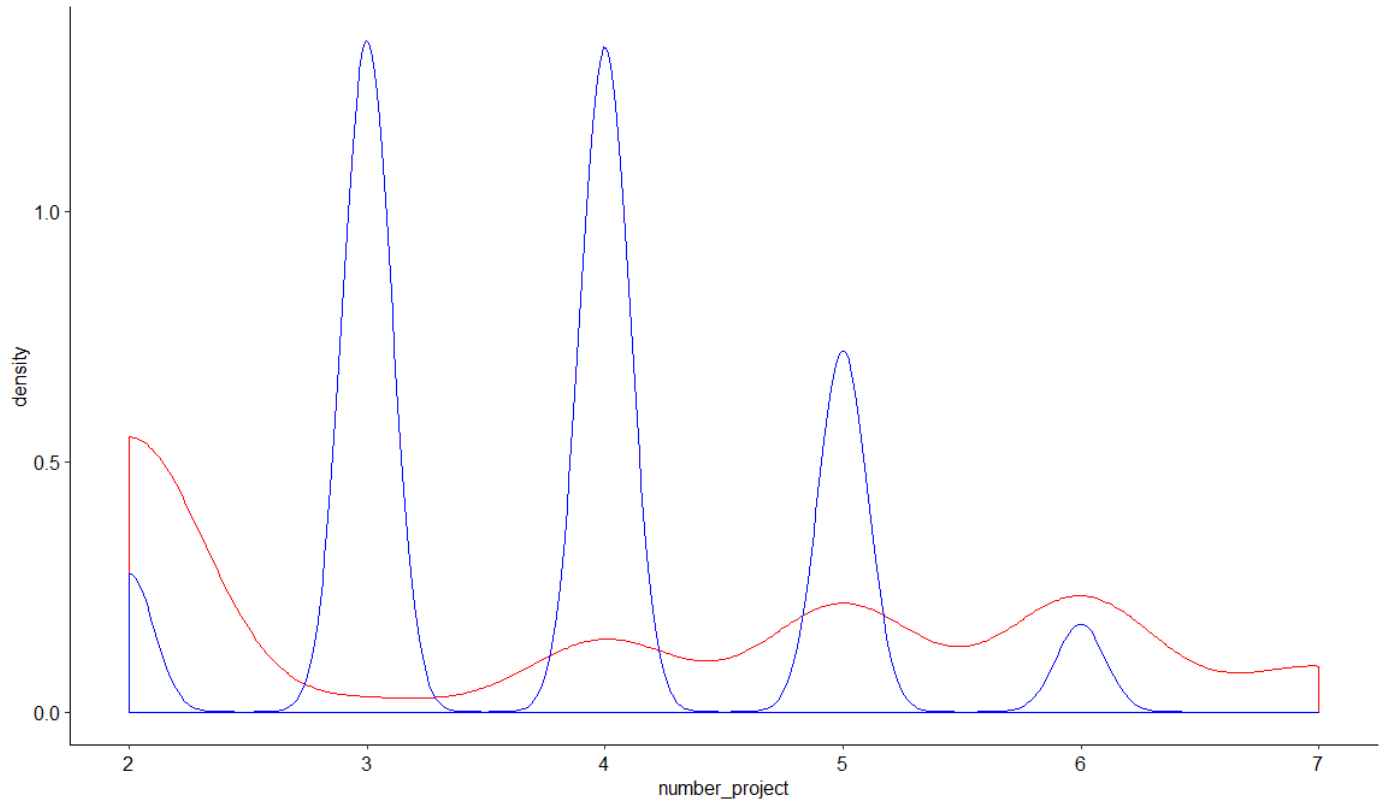

```
#Prepare Data of employess Left and Stayed
LeftData <- subset(data, left == 1
)
StayData <- subset(data, left == 0
)

ggplot() + geom_density(aes(x = satisfaction_level), colour = "red"  , data = LeftData) +

  geom_density(aes(x = satisfaction_level), colour = "blue" , data = StayData
)
```



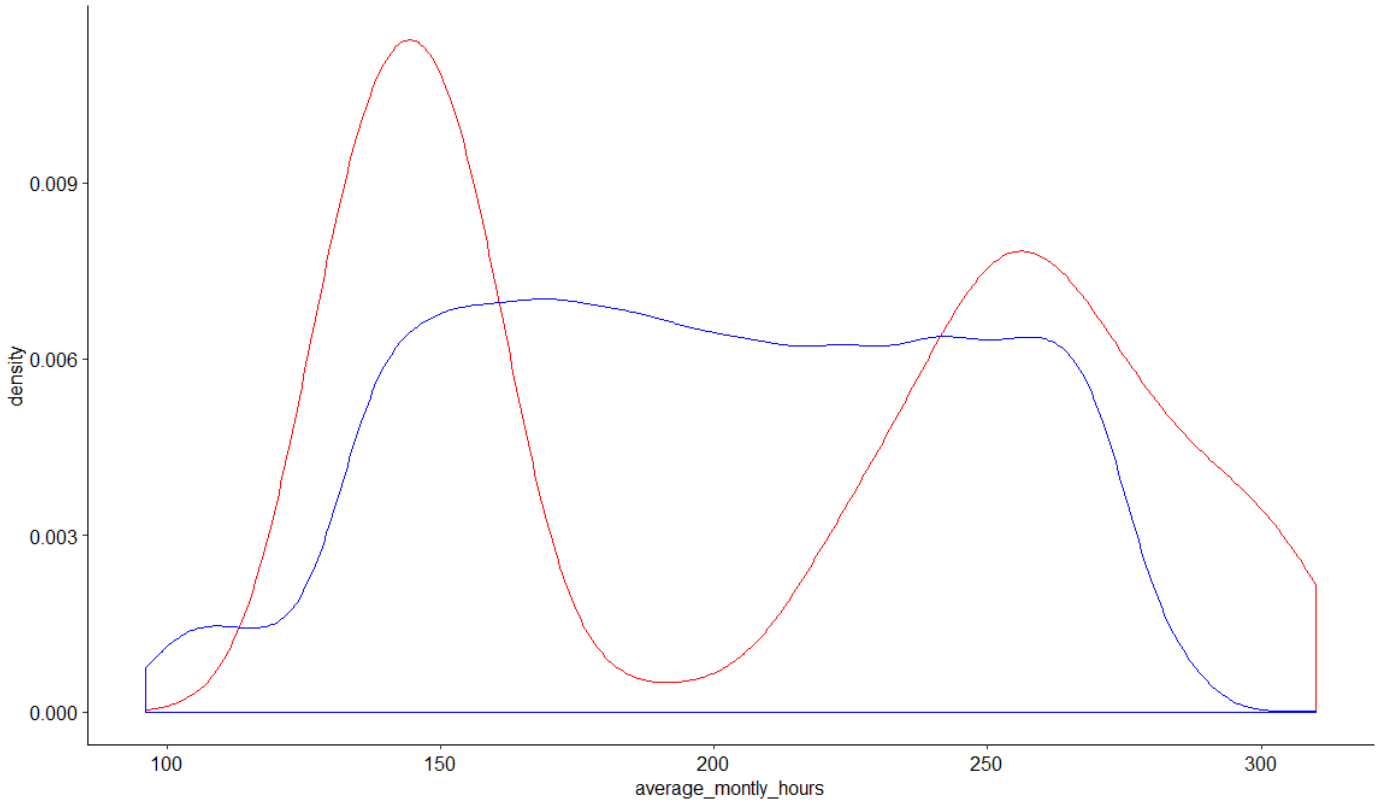
```
ggplot() + geom_density(aes(x = last_evaluation), colour = "red"  , data = LeftData) +

  geom_density(aes(x = last_evaluation), colour = "blue" , data = StayData)
```



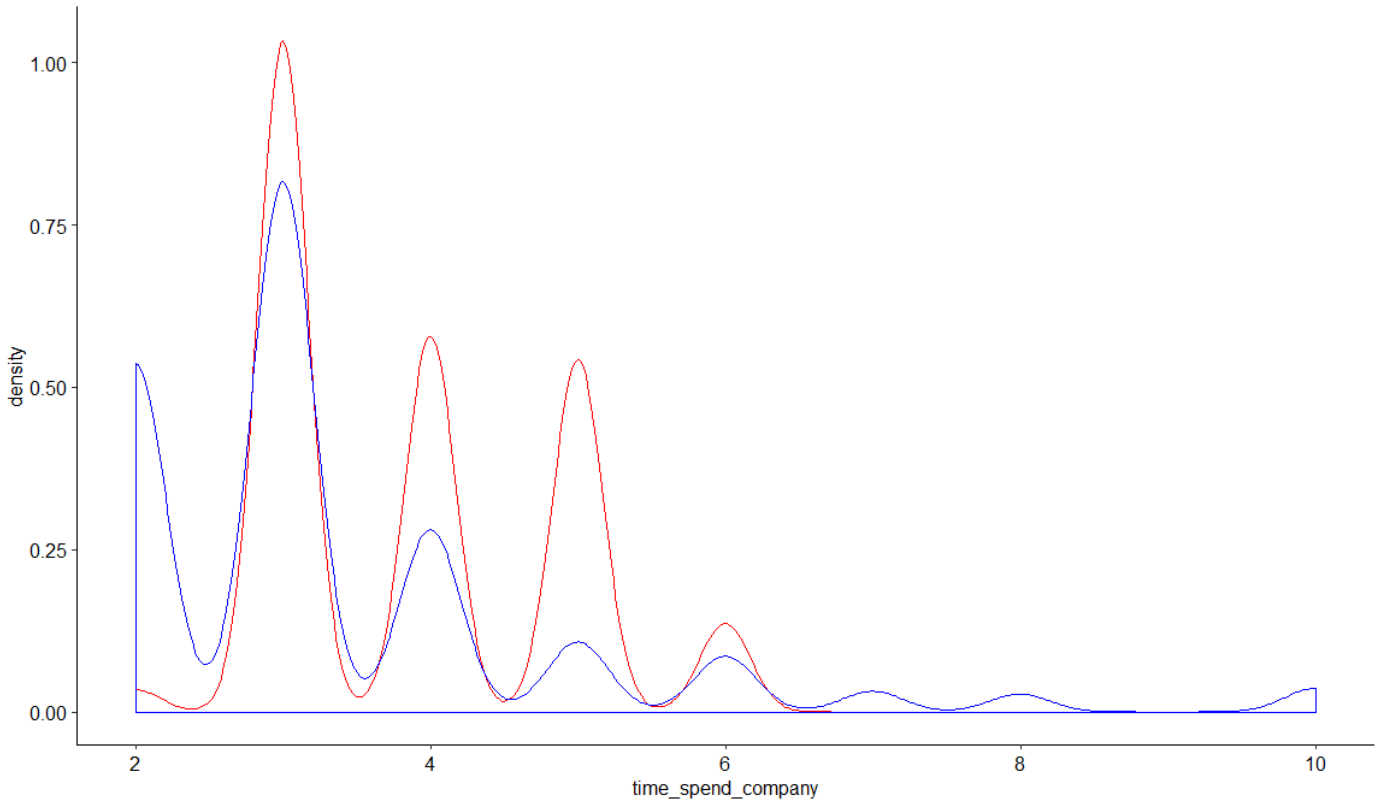

```
ggplot() + geom_density(aes(x = number_project), colour = "red"  , data = LeftData) +

  geom_density(aes(x = number_project), colour = "blue" , data = StayData
)
```



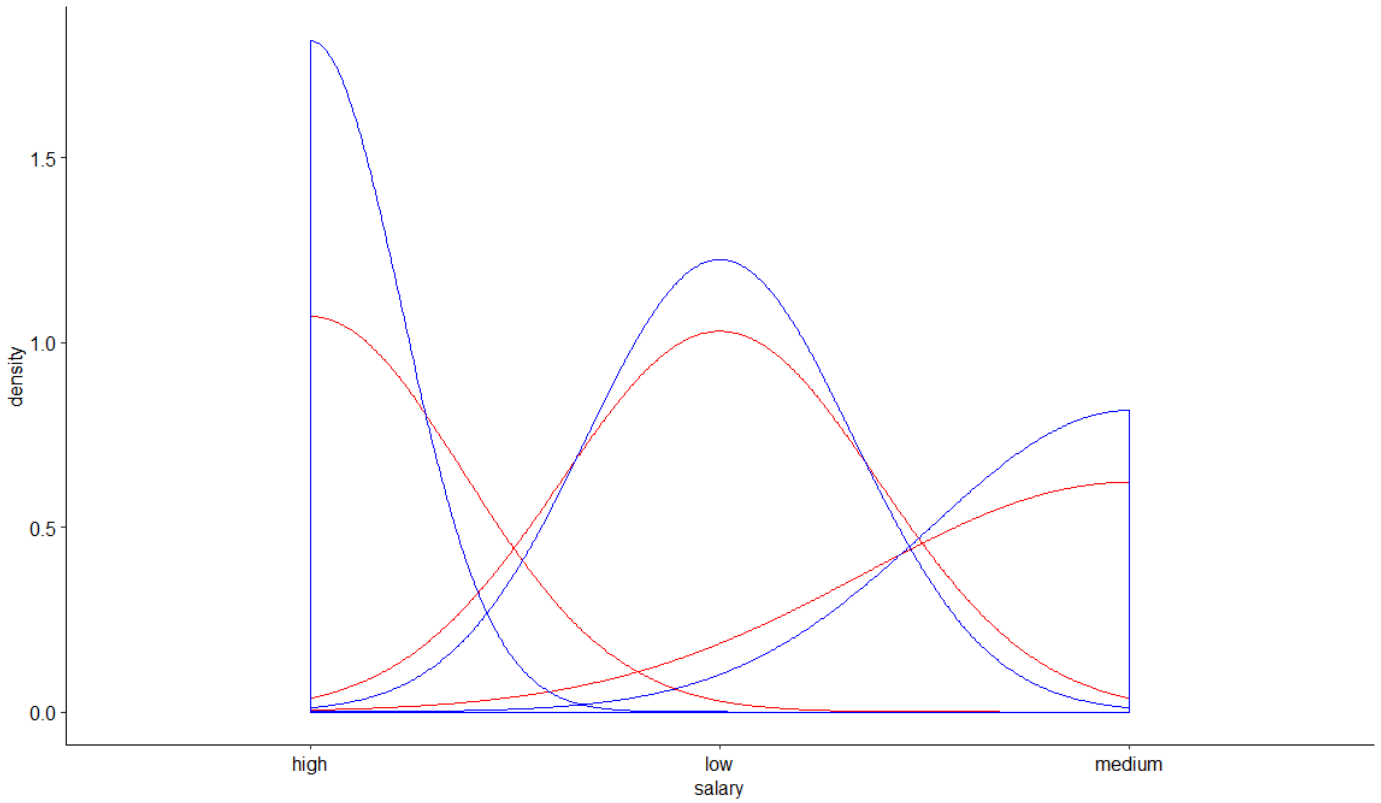
```
ggplot() + geom_density(aes(x = average_montly_hours), colour = "red"  , data = LeftData) +

  geom_density(aes(x = average_montly_hours), colour = "blue" , data = StayData)
```



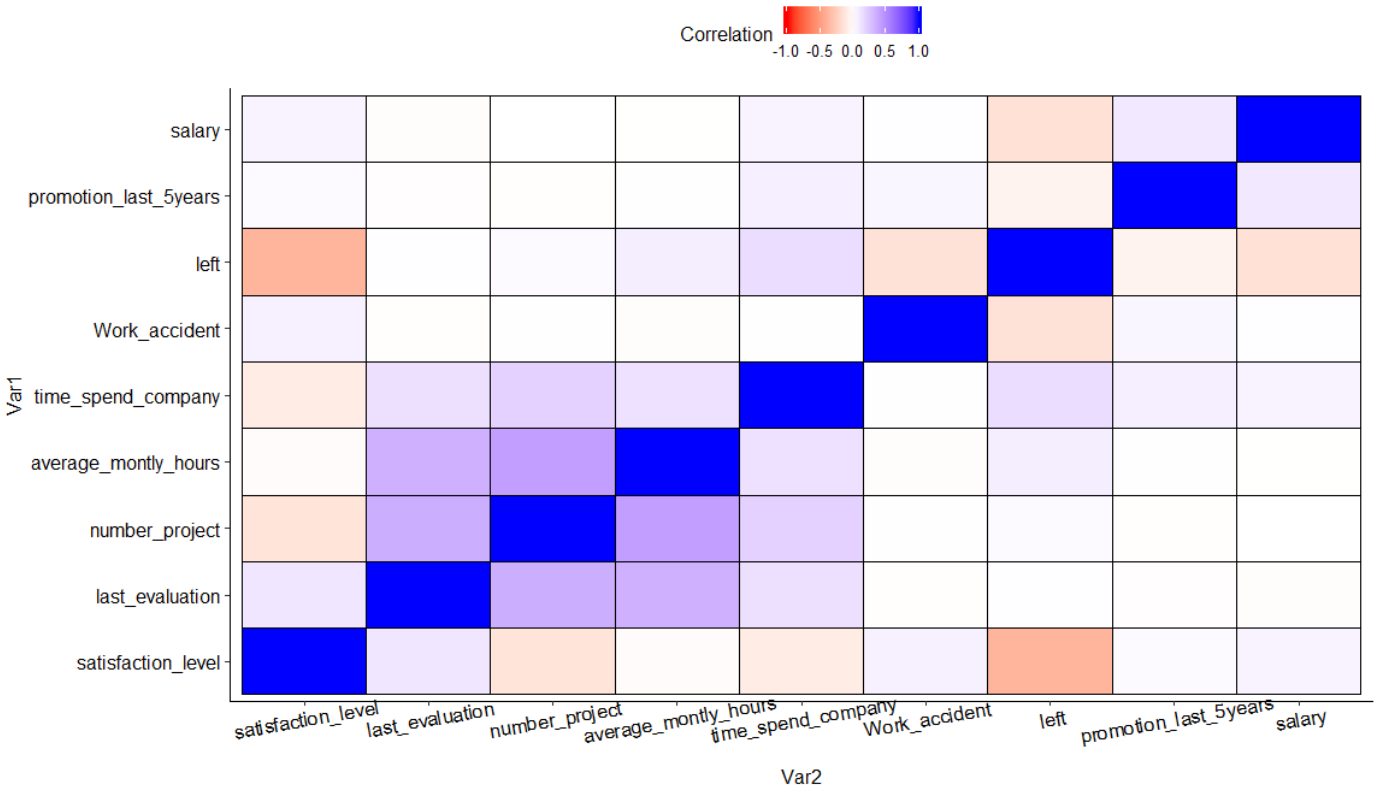

```
ggplot() + geom_density(aes(x = time_spend_company), colour = "red"  , data = LeftData) +

  geom_density(aes(x = time_spend_company), colour = "blue" , data = StayData
)
```



```
ggplot() + geom_density(aes(x = salary), colour = "red"  , data = LeftData) +

  geom_density(aes(x = salary), colour = "blue" , data = StayData
)
```




```
#################[    Predictive Modeling    ]#########################################################

#convert salary into numeric low=1, medium=2, High=3
data$salary <- ifelse(data$salary == 'low', 1
,
                      ifelse(data$salary == 'medium', 2
,
                             ifelse(data$salary == 'high', 3, 0)
))

#Print correlation heat-map;
PrintCorrelationHeatmap(data[,-9]) #remove department as it is still category
```




```
#Create one-hot-encoding for categorical data (Department)
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

```
dataWithDummies <- dummy.data.frame(data, sep = "."
)

#Normalize the data by diving max(x)
data_norm <- as.data.frame(lapply(dataWithDummies, normalize_DivideByMax))



#Final Data
dataFinal<- data_norm
dataFinal$left <- factor(dataFinal$left) #required for random forest and Navie Bayes

#Split data into Train and Test
train<-TrainTestSplit(dataFinal, splitFactor = 0.7, train = TRUE
)
test<-TrainTestSplit(dataFinal, splitFactor = 0.7, train = FALSE
)

write.csv(train, file = "Data\\train.csv")
write.csv(test, file = "Data\\test.csv"
)

#Prepare Data for Models
trainLabel<- "train$left"
numericIndeDependentVariables <-paste("satisfaction_level+last_evaluation+number_project+average_montl
)

departments <-  paste("Department.accounting+Department.hr+Department.IT+Department.management+Departm
)

inputVariables <- paste(numericIndeDependentVariables, "+", departments)

###########[    Classification Models      ]################################################
CreateLogisticRegressionModel ( trainLabel, test$left, inputVariables, train, test    )
```

```
## [1] "Creating Model for  Logistic Regression"
##
## Call:
## glm(formula = formula, family = binomial(), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1896  -0.6569  -0.4019  -0.1282   3.1024
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              1.30624    0.16617   7.861 3.81e-15 ***
## satisfaction_level      -4.14566    0.11743 -35.305  < 2e-16 ***
## last_evaluation          0.65586    0.17789   3.687 0.000227 ***
## number_project          -2.22439    0.17951 -12.391  < 2e-16 ***
## average_montly_hours     1.47944    0.19241   7.689 1.48e-14 ***
## time_spend_company       2.60540    0.18617  13.995  < 2e-16 ***
## Work_accident           -1.63682    0.11263 -14.532  < 2e-16 ***
## promotion_last_5years   -1.50753    0.30168  -4.997 5.82e-07 ***
## salary                  -2.04524    0.13666 -14.966  < 2e-16 ***
## Department.accounting   -0.06023    0.12529  -0.481 0.630697
## Department.hr            0.05269    0.12696   0.415 0.678134
## Department.IT           -0.26533    0.11141  -2.382 0.017238 *
## Department.management   -0.50431    0.16337  -3.087 0.002022 **
## Department.marketing    -0.06760    0.12676  -0.533 0.593855
## Department.product_mng  -0.35443    0.12643  -2.803 0.005059 **
## Department.RandD        -0.55714    0.14448  -3.856 0.000115 ***
## Department.sales        -0.13530    0.07878  -1.717 0.085896 .
## Department.support      -0.03872    0.09032  -0.429 0.668154
## Department.technical          NA         NA      NA       NA
```

```
## Department.Technical          NA          NA       NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11488.5  on 10498  degrees of freedom
## Residual deviance:  8975.5  on 10481  degrees of freedom
## AIC: 9011.5
##
## Number of Fisher Scoring iterations: 5
##
## [1] "Logistic Regression"
## [1] "Test data Confusion Matrix for  Logistic Regression  Model:"
## Confusion Matrix and Statistics
##
##           Actual
## Prediction    0    1
##          0 3289  896
##          1  124  191
##
##                Accuracy : 0.7733
##                  95% CI : (0.7608, 0.7855)
##     No Information Rate : 0.7584
##     P-Value [Acc > NIR] : 0.009914
##
##                   Kappa : 0.1839
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9637
##             Specificity : 0.1757
##          Pos Pred Value : 0.7859
##          Neg Pred Value : 0.6063
##              Prevalence : 0.7584
##          Detection Rate : 0.7309
##    Detection Prevalence : 0.9300
##       Balanced Accuracy : 0.5697
##
##        'Positive' Class : 0
##
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

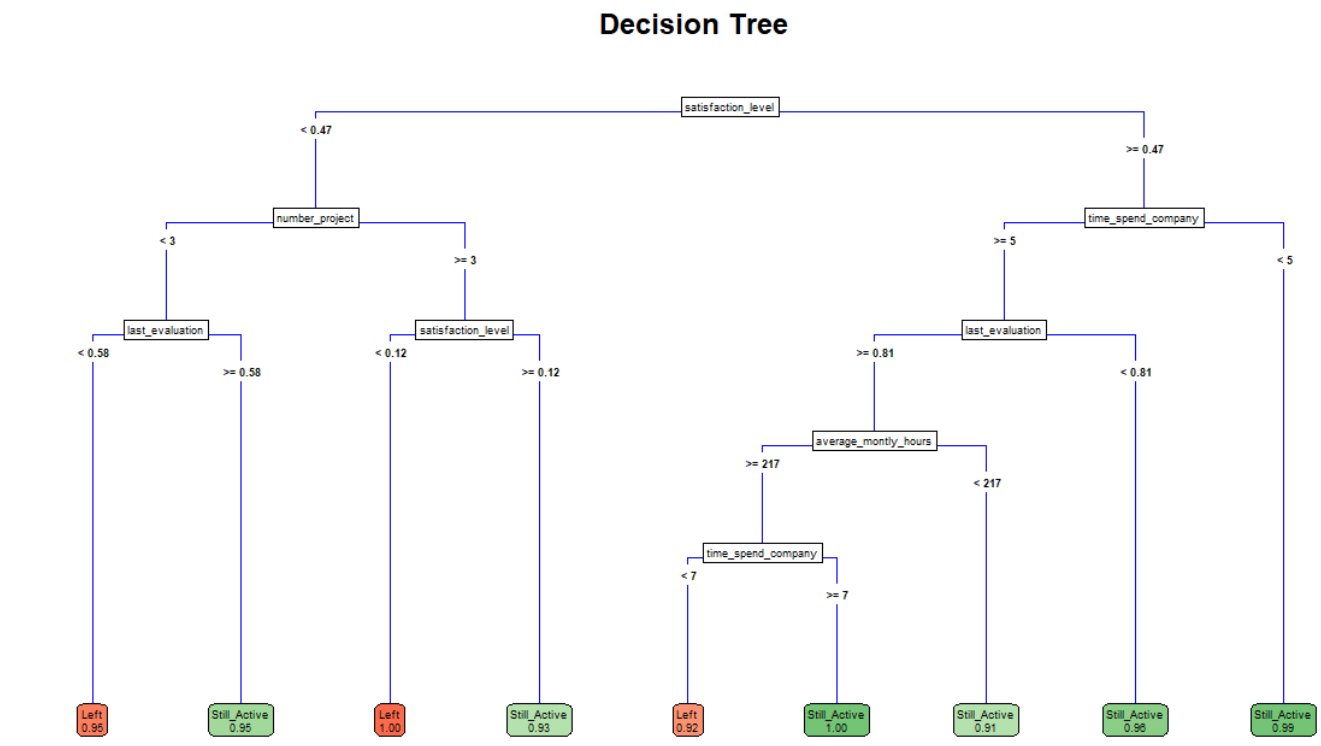
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
## [1] "Model Performance:"
##                 MODEL Accuracy  AUC Specificity Precision
## 1 Logistic Regression  77.33 % 0.57      0.1757    0.7859
##   Sensitivity_Recall CostToCompany
## 1             0.9637     $ 18540 K
## [1] "False Negative =  896    False Positive =  124  Cost To Company =  $ 18540 K"
```

```
##################################################
#To get maximum benefit of decision Tree, we need original data and convert numeric left to
# "Still_Active" and "left". Also we wont use department for decision tree as from
# logistic regression, we came to know that department is not that important variable
dataDecisionTree <- data[,-9]
dataDecisionTree$left <- ifelse(dataDecisionTree$left == '0',"Still_Active"
,
                                ifelse(dataDecisionTree$left == "1",'Left',"NA"
))
#Split data into Train and Test
trainDecisionTree <- TrainTestSplit(dataDecisionTree, splitFactor = 0.7, train = TRUE
)
testDecisionTree <- TrainTestSplit(dataDecisionTree, splitFactor = 0.7, train = FALSE
)

CreateDecisionTreeModel ( "trainDecisionTree$left",

                          testDecisionTree$left, numericIndeDependentVariables,
                          trainDecisionTree, testDecisionTree)
```

```
## [1] "Creating Model for  Decision Tree"
## box.palette "RdGn" (diverging pal.thresh 0.5): #FB6A4A (near tomato) to #74C476 (near palegreen3)
```

```
## cex 0.58   xlim c(0, 1)   ylim c(0, 1)
```



```
## [1] "Decision Tree"
## [1] "Test data Confusion Matrix for  Decision Tree  Model:"
## Confusion Matrix and Statistics
##
##              Actual
## Prediction     Left Still_Active
##   Left         1001           55
##   Still_Active   86         3358
##
##                Accuracy : 0.9687
##                  95% CI : (0.9632, 0.9736)
##     No Information Rate : 0.7584
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9136
##  Mcnemar's Test P-Value : 0.01152
##
##             Sensitivity : 0.9209
##             Specificity : 0.9839
##          Pos Pred Value : 0.9479
##          Neg Pred Value : 0.9750
##              Prevalence : 0.2416
##          Detection Rate : 0.2224
##    Detection Prevalence : 0.2347
##       Balanced Accuracy : 0.9524
##
##        'Positive' Class : Left
##
## [1] "Model Performance:"
##           MODEL Accuracy  AUC Specificity Precision Sensitivity_Recall
## 1 Decision Tree  96.87 % 0.95      0.9839    0.9479             0.9209
##   CostToCompany
## 1      $ 1530 K
## [1] "False Negative =  55    False Positive =  86  Cost To Company =  $ 1530 K"
```

```
################################################
CreateRandomForestModel( trainLabel, test$left, inputVariables, train, test, 50
)
```

```
## [1] "Creating Model for  Random Forest"
```

```
## randomForest 4.6-14
```

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

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

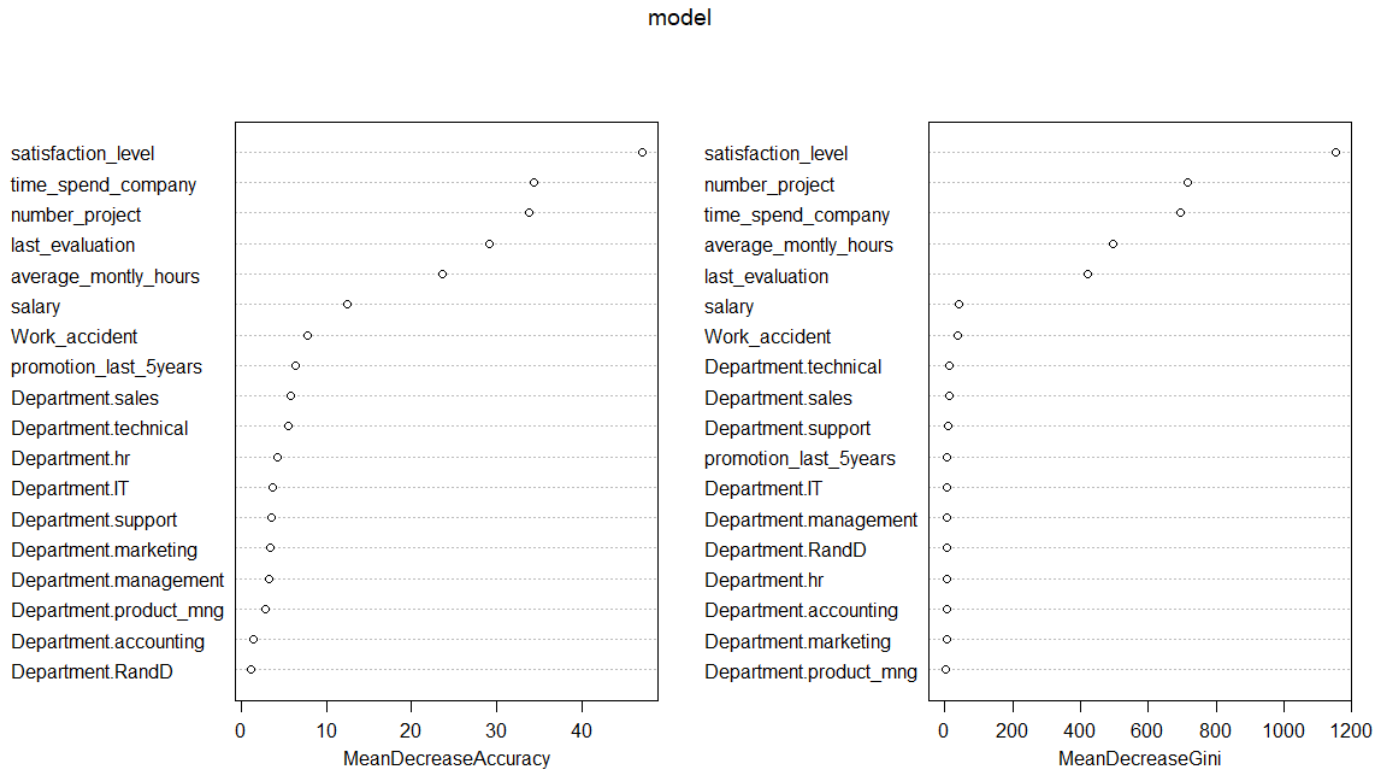
```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

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

```
##
## Call:
##  randomForest(formula = formula, data = train, importance = TRUE,      ntree = numTree)
##                Type of random forest: classification
```

```
##                     Number of trees: 50
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 1.49%
## Confusion matrix:
##      0    1 class.error
## 0 7997   18 0.002245789
## 1  138 2346 0.055555556
##                               0        1 MeanDecreaseAccuracy
## satisfaction_level     21.39253 47.15920             47.07956
## last_evaluation         7.97914 29.25799             29.12640
## number_project         14.03402 33.30982             33.71134
## average_montly_hours   14.55875 21.64710             23.64089
## time_spend_company     19.66713 32.90225             34.35793
## Work_accident           2.58234  8.19856              7.78696
## promotion_last_5years   1.71960  5.77242              6.34747
## salary                  8.04755 11.43809             12.45844
## Department.accounting   0.42877  1.89176              1.40559
## Department.hr          -1.67058  8.12664              4.30421
## Department.IT           0.77785  4.72663              3.74608
## Department.management   0.48139  4.74849              3.29094
## Department.marketing    1.73787  3.54013              3.35388
## Department.product_mng  1.83870  2.43538              2.77763
## Department.RandD       -1.30132  2.93253              1.19474
## Department.sales        0.99773  8.40349              5.83756
## Department.support      0.56785  5.89373              3.62263
## Department.technical   -0.94608  9.84358              5.59713
##                       MeanDecreaseGini
## satisfaction_level          1154.22282
## last_evaluation              422.57371
## number_project               715.46101
## average_montly_hours         497.40994
## time_spend_company           696.22476
## Work_accident                 36.50093
## promotion_last_5years          7.74975
## salary                        42.92405
## Department.accounting          4.64416
## Department.hr                  5.44878
## Department.IT                  6.10819
## Department.management          5.97108
## Department.marketing           4.35694
## Department.product_mng         4.15566
## Department.RandD               5.58715
## Department.sales              11.76015
## Department.support            10.34282
## Department.technical          12.00047
## [1] "Random Forest"
## [1] "Test data Confusion Matrix for  Random Forest  Model:"
## Confusion Matrix and Statistics
##
##          Actual
## Prediction    0    1
##          0 3404   36
##          1    9 1051
##
##                Accuracy : 0.99
##                  95% CI : (0.9866, 0.9927)
##     No Information Rate : 0.7584
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9725
##  Mcnemar's Test P-Value : 0.0001063
##
##             Sensitivity : 0.9974
##             Specificity : 0.9669
```

```
##          Pos Pred Value : 0.9895
##          Neg Pred Value : 0.9915
##              Prevalence : 0.7584
##          Detection Rate : 0.7564
##    Detection Prevalence : 0.7644
##       Balanced Accuracy : 0.9821
##
##        'Positive' Class : 0
##
## [1] "Model Performance:"
##           MODEL Accuracy  AUC Specificity Precision Sensitivity_Recall
## 1 Random Forest     99 % 0.98      0.9669    0.9895             0.9974
##   CostToCompany
## 1       $ 765 K
## [1] "False Negative =  36    False Positive =  9  Cost To Company =  $ 765 K"
```




```
################################################
CreateKernalSvmModel ( trainLabel, test$left, inputVariables, train, test    )
```

```
## [1] "Creating Model for  Kernal SVM"
## [1] "Kernal SVM"
## [1] "Test data Confusion Matrix for  Kernal SVM  Model:"
## Confusion Matrix and Statistics
##
##           Actual
## Prediction    0    1
##          0 3290  108
##          1  123  979
##
##                Accuracy : 0.9487
##                  95% CI : (0.9418, 0.9549)
##     No Information Rate : 0.7584
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8606
##  Mcnemar's Test P-Value : 0.357
##
##             Sensitivity : 0.9640
##             Specificity : 0.9006
##          Pos Pred Value : 0.9682
##          Neg Pred Value : 0.8884
##              Prevalence : 0.7584
##          Detection Rate : 0.7311
##    Detection Prevalence : 0.7551
##       Balanced Accuracy : 0.9323
##
##        'Positive' Class : 0
##
## [1] "Model Performance:"
##        MODEL Accuracy  AUC Specificity Precision Sensitivity_Recall
## 1 Kernal SVM  94.87 % 0.93      0.9006    0.9682              0.964
##   CostToCompany
## 1      $ 2775 K
## [1] "False Negative =  108    False Positive =  123  Cost To Company =  $ 2775 K"
```

```
################################################
CreateNaiveBayesModel( trainLabel, test$left, inputVariables, train, test    )
```

```
## [1] "Creating Model for  Naive Bayes"
## [1] "Naive Bayes"
## [1] "Test data Confusion Matrix for  Naive Bayes  Model:"
## Confusion Matrix and Statistics
##
##           Actual
## Prediction    0    1
##          0 2465  251
##          1  948  836
##
##                Accuracy : 0.7336
##                  95% CI : (0.7204, 0.7464)
##     No Information Rate : 0.7584
##     P-Value [Acc > NIR] : 0.9999
##
##                   Kappa : 0.4032
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7222
##             Specificity : 0.7691
##          Pos Pred Value : 0.9076
##          Neg Pred Value : 0.4686
##              Prevalence : 0.7584
##          Detection Rate : 0.5478
##    Detection Prevalence : 0.6036
##       Balanced Accuracy : 0.7457
##
##        'Positive' Class : 0
##
## [1] "Model Performance:"
##         MODEL Accuracy  AUC Specificity Precision Sensitivity_Recall
## 1 Naive Bayes  73.36 % 0.75      0.7691    0.9076             0.7222
##   CostToCompany
## 1      $ 9760 K
## [1] "False Negative =  251    False Positive =  948  Cost To Company =  $ 9760 K"
```

```
##################################################
CreateDeepnetNNModel(train,test,targetColumnNumber=7,hiddenLayers=c(50, 30, 10),numepochs =700
)
```

```
## ####loss on step 10000 is : 0.174080
```

```
## ####loss on step 20000 is : 0.065343
```

```
## ####loss on step 30000 is : 0.024764
```

```
## ####loss on step 40000 is : 0.027589
```

```
## ####loss on step 50000 is : 0.000757
```

```
## ####loss on step 60000 is : 0.034999
```

```
## ####loss on step 70000 is : 0.000882
```

```
## [1] "Train Accuracy of  Deepnet Neural Network  is  97.49 %"
## [1] "Test Accuracy of  Deepnet Neural Network  is  96.64 %"
## [1] "Deepnet Neural Network"
## [1] "Test data Confusion Matrix for  Deepnet Neural Network  Model:"
## Confusion Matrix and Statistics
##
##           Actual
## Prediction    0    1
##          0 3356   94
##          1   57  993
##
##                Accuracy : 0.9664
##                  95% CI : (0.9608, 0.9715)
##     No Information Rate : 0.7584
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9073
##  Mcnemar's Test P-Value : 0.003394
##
##             Sensitivity : 0.9833
##             Specificity : 0.9135
##          Pos Pred Value : 0.9728
##          Neg Pred Value : 0.9457
##              Prevalence : 0.7584
##          Detection Rate : 0.7458
##    Detection Prevalence : 0.7667
##       Balanced Accuracy : 0.9484
##
##        'Positive' Class : 0
##
## [1] "Model Performance:"
##                    MODEL Accuracy  AUC Specificity Precision
## 1 Deepnet Neural Network  96.64 % 0.95      0.9135    0.9728
##   Sensitivity_Recall CostToCompany
## 1             0.9833      $ 2165 K
## [1] "False Negative =  94    False Positive =  57  Cost To Company =  $ 2165 K"
```

```
################################################
CreateKerasNNModel(train,test,targetColumnNumber=7,batchSize=128,numepochs=25,validationSplit=0.2,
                   lossFunction= "categorical_crossentropy",errorMetrics= "accuracy"
)
```

```
## [1] "Keras Neural Network"
## [1] "Test data Confusion Matrix for  Keras Neural Network  Model:"
## Confusion Matrix and Statistics
##
##           Actual
## Prediction    0    1
##          0 3267   81
##          1  146 1006
##
##                Accuracy : 0.9496
##                  95% CI : (0.9428, 0.9558)
##     No Information Rate : 0.7584
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8651
##  Mcnemar's Test P-Value : 2.159e-05
##
##             Sensitivity : 0.9572
##             Specificity : 0.9255
##          Pos Pred Value : 0.9758
##          Neg Pred Value : 0.8733
##              Prevalence : 0.7584
##          Detection Rate : 0.7260
##    Detection Prevalence : 0.7440
##       Balanced Accuracy : 0.9414
##
##        'Positive' Class : 0
##
## [1] "Model Performance:"
##                  MODEL Accuracy  AUC Specificity Precision
## 1 Keras Neural Network  94.96 % 0.94      0.9255    0.9758
##   Sensitivity_Recall CostToCompany
## 1             0.9572      $ 2350 K
## [1] "False Negative =  81    False Positive =  146  Cost To Company =  $ 2350 K"
```

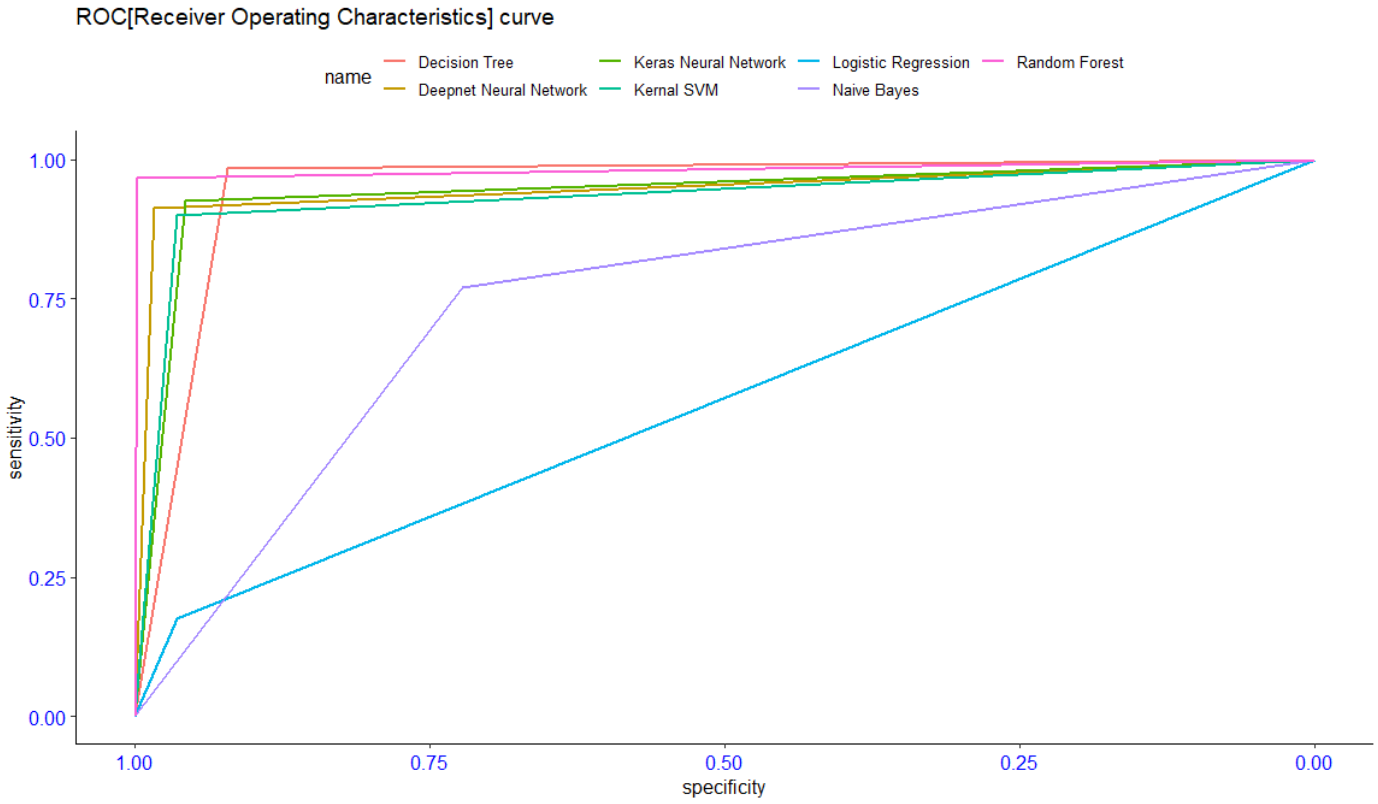
```
################################################
CompareModelsAndPlotCombinedRocCurve()
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:randomForest':
##
##     combine
```

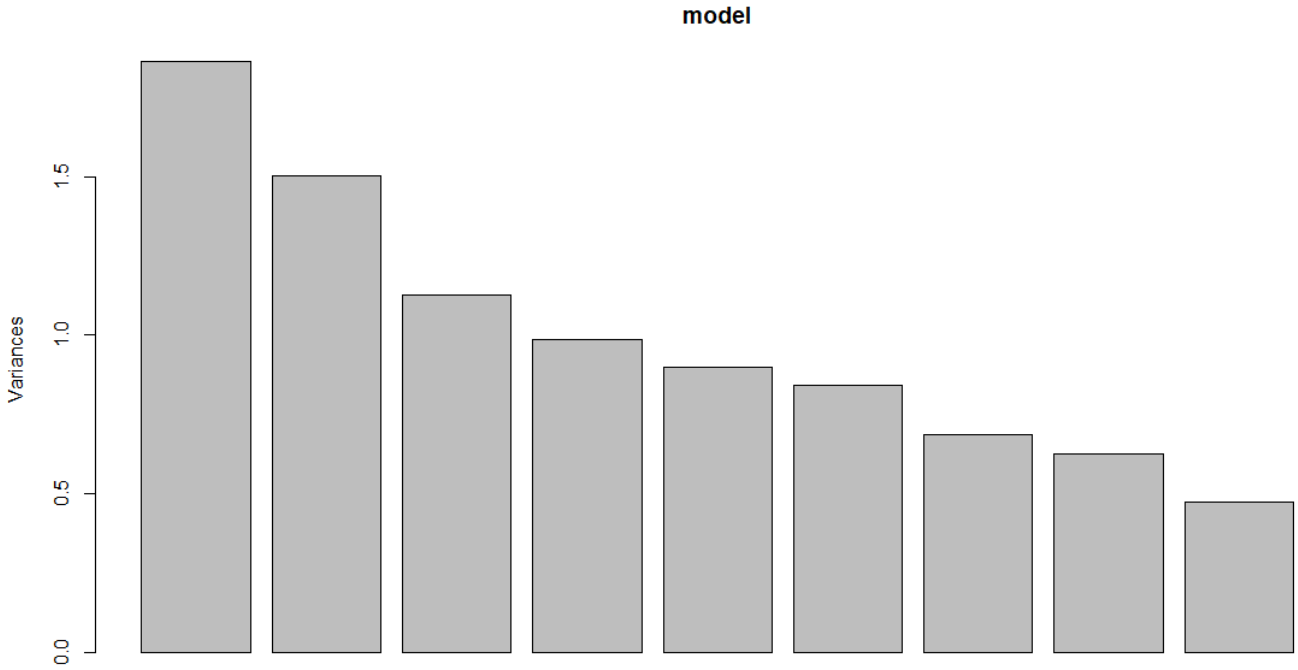
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

	MODEL	Accuracy	AUC	Specificity	Precision	Sensitivity_Recall	CostToCompany
1	Keras Neural Network	94.96 %	0.94	0.9255	0.9758	0.9572	$ 2350 K
2	Deepnet Neural Network	96.64 %	0.95	0.9135	0.9728	0.9833	$ 2165 K
3	Naive Bayes	73.36 %	0.75	0.7691	0.9076	0.7222	$ 9760 K
4	Kernal SVM	94.87 %	0.93	0.9006	0.9682	0.964	$ 2775 K
5	Random Forest	99 %	0.98	0.9669	0.9895	0.9974	$ 765 K
6	Decision Tree	96.87 %	0.95	0.9839	0.9479	0.9209	$ 1530 K
7	Logistic Regression	77.33 %	0.57	0.1757	0.7859	0.9637	$ 18540 K

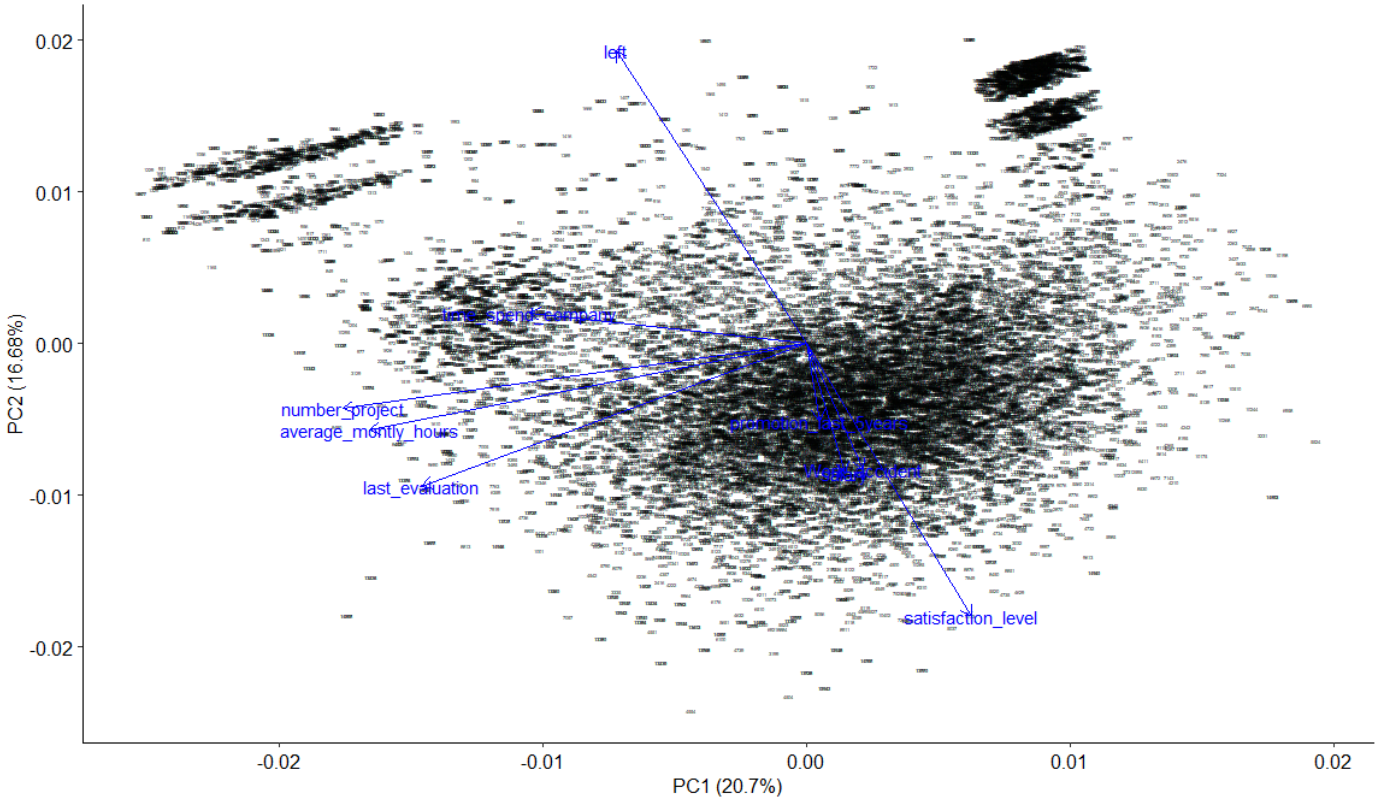



```
#########[    Principle Components Analysis    ]########################################
#for PCA, remove the department
CreatePCA(data[,-9], numComponents=3)
```

```
## [1] "Creating Principle Components Analysis(PCA)"
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.365 1.2251 1.0612 0.9922 0.94816 0.91804 0.82727
## Proportion of Variance 0.207 0.1668 0.1251 0.1094 0.09989 0.09364 0.07604
## Cumulative Proportion  0.207 0.3737 0.4989 0.6082 0.70812 0.80177 0.87781
##                           PC8     PC9
## Standard deviation     0.79149 0.68792
## Proportion of Variance 0.06961 0.05258
## Cumulative Proportion  0.94742 1.00000
```



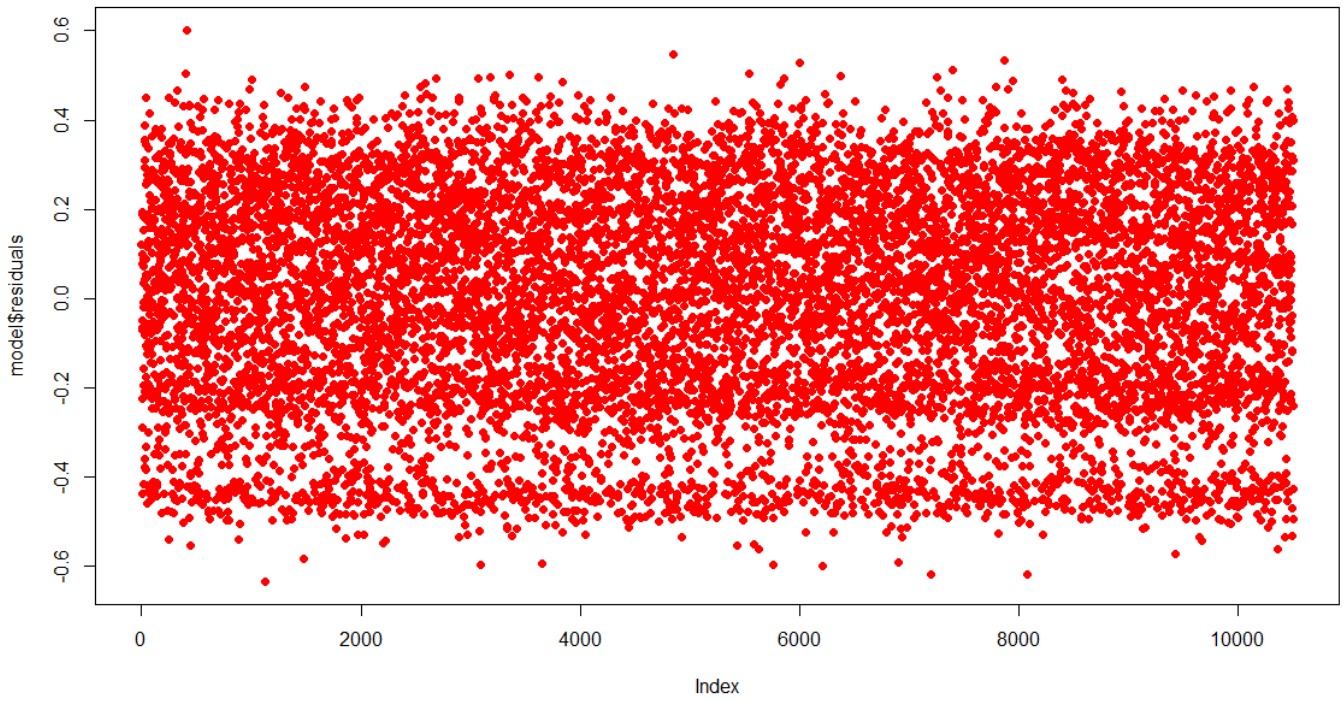

```
##                                PC1        PC2         PC3
## satisfaction_level     0.19723126 -0.5667338  0.23539398
## last_evaluation       -0.46006448 -0.2959920  0.19856579
## number_project        -0.55431462 -0.1345452  0.01627272
## average_montly_hours  -0.52278654 -0.1795027  0.11612682
## time_spend_company    -0.33143549  0.0645894 -0.43833959
## Work_accident          0.06767655 -0.2622693 -0.09761638
## left                  -0.22815071  0.6079660 -0.04655295
## promotion_last_5years  0.01532004 -0.1605496 -0.63070885
## salary                 0.04640023 -0.2692992 -0.53831188
```




```
#########################[   Models for Satisfaction_level   ]##############################
#make of copy of original Data
data<-dataOriginal

#convert salary into numeric low=1, medium=2, High=3
data$salary <-ifelse(data$salary == 'low',1, ifelse(data$salary == 'medium', 2
,
                                                    ifelse(data$salary == 'high', 3, 0
)))

#eliminate left and department
dataFinal <- data[,-c(7,9)]

#Split data into Train and Test
train<-TrainTestSplit(dataFinal, splitFactor = 0.7, train = TRUE
)
test<-TrainTestSplit(dataFinal, splitFactor = 0.7, train = FALSE
)
#Prepare Data for Models
trainLabel  <- "train$satisfaction_level"
test_Y      <- test$satisfaction_level
inputVariables <-paste("last_evaluation+number_project+average_montly_hours+time_spend_company+Work_ac
)

#without polynomial Regression
CreateStepwiseLinearRegressionModel(dataFinal,targetColumnNumber=1,isPoly=FALSE)
```

```
##
## Call:
## lm(formula = satisfaction_level ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6361 -0.1862  0.0190  0.1968  0.6024
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.770e-01  1.404e-02  41.096  < 2e-16 ***
```

```
## last_evaluation        2.595e-01  1.502e-02  17.280  < 2e-16 ***
## number_project        -3.895e-02  2.191e-03 -17.778  < 2e-16 ***
## average_montly_hours   8.871e-05  5.311e-05   1.671   0.0948 .
## time_spend_company    -1.634e-02  1.652e-03  -9.887  < 2e-16 ***
## Work_accident          4.232e-02  6.714e-03   6.303 3.03e-10 ***
## promotion_last_5years  3.305e-02  1.596e-02   2.071   0.0384 *
## salary                 1.894e-02  3.681e-03   5.146 2.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2401 on 10491 degrees of freedom
## Multiple R-squared:  0.06236,    Adjusted R-squared:  0.06174
## F-statistic: 99.68 on 7 and 10491 DF,  p-value: < 2.2e-16
##
## Start:  AIC=-29949.92
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     salary
##
##                        Df Sum of Sq    RSS    AIC
## <none>                              604.78 -29950
## - average_montly_hours   1    0.1609 604.94 -29949
## - promotion_last_5years  1    0.2473 605.03 -29948
## - salary                 1    1.5266 606.31 -29926
## - Work_accident          1    2.2904 607.07 -29912
## - time_spend_company     1    5.6352 610.42 -29855
## - last_evaluation        1   17.2134 621.99 -29657
## - number_project         1   18.2209 623.00 -29640
##
## Call:
## lm(formula = satisfaction_level ~ last_evaluation + number_project +
##     average_montly_hours + time_spend_company + Work_accident +
##     promotion_last_5years + salary, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6361 -0.1862  0.0190  0.1968  0.6024
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.770e-01  1.404e-02  41.096  < 2e-16 ***
## last_evaluation        2.595e-01  1.502e-02  17.280  < 2e-16 ***
## number_project        -3.895e-02  2.191e-03 -17.778  < 2e-16 ***
## average_montly_hours   8.871e-05  5.311e-05   1.671   0.0948 .
## time_spend_company    -1.634e-02  1.652e-03  -9.887  < 2e-16 ***
## Work_accident          4.232e-02  6.714e-03   6.303 3.03e-10 ***
## promotion_last_5years  3.305e-02  1.596e-02   2.071   0.0384 *
## salary                 1.894e-02  3.681e-03   5.146 2.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2401 on 10491 degrees of freedom
## Multiple R-squared:  0.06236,    Adjusted R-squared:  0.06174
## F-statistic: 99.68 on 7 and 10491 DF,  p-value: < 2.2e-16
```




```
##        RMSE     R2.fit     R2.lwr     R2.upr
## 1 0.4544963 0.06231374 0.06230637 0.06232022
```

```
#With Polynomial regression
CreateStepwiseLinearRegressionModel(dataFinal,targetColumnNumber=1,isPoly=TRUE)
```

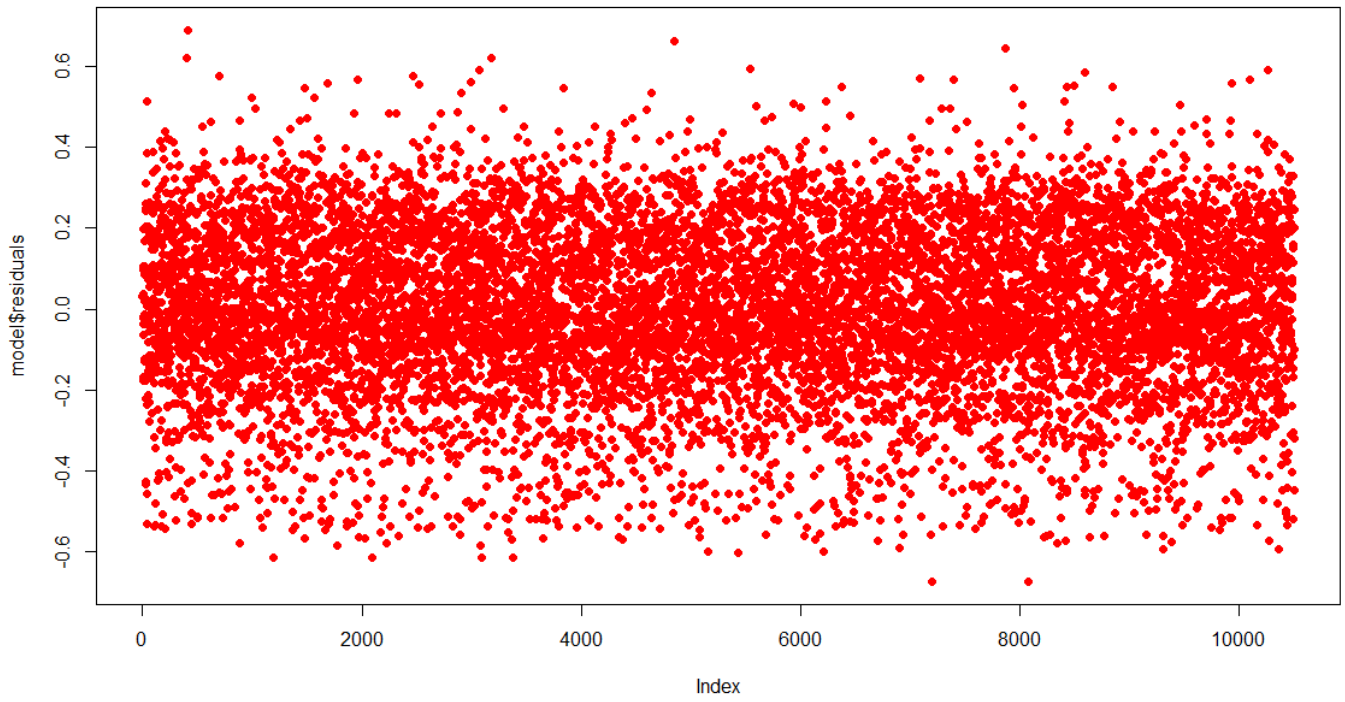
```
##
## Call:
## lm(formula = satisfaction_level ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67467 -0.12006 -0.00556  0.15177  0.69072
##
## Coefficients: (5 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2.324e-01  1.917e-01  -1.212    0.226
## last_evaluation               3.939e+00  6.670e-01   5.905 3.63e-09 ***
## number_project                4.694e-01  4.083e-02  11.498  < 2e-16 ***
## average_montly_hours         -1.407e-02  1.826e-03  -7.706 1.41e-14 ***
## time_spend_company           -1.307e-01  1.874e-02  -6.974 3.27e-12 ***
## Work_accident                 6.984e-03  5.663e-03   1.233    0.218
## promotion_last_5years         1.104e-02  1.341e-02   0.824    0.410
## salary                        5.791e-03  1.722e-02   0.336    0.737
## last_evaluation_Square       -5.328e+00  9.673e-01  -5.508 3.72e-08 ***
## number_project_Square        -6.532e-02  1.020e-02  -6.402 1.60e-10 ***
## average_montly_hours_Square   8.764e-05  9.264e-06   9.461  < 2e-16 ***
## time_spend_company_Square     2.084e-02  3.861e-03   5.399 6.85e-08 ***
## Work_accident_Square                 NA         NA      NA       NA
## promotion_last_5years_Square         NA         NA      NA       NA
## salary_Square                -1.276e-03  4.733e-03  -0.270    0.788
## last_evaluation_Cube          2.408e+00  4.547e-01   5.295 1.22e-07 ***
## number_project_Cube           5.110e-04  8.068e-04   0.633    0.527
## average_montly_hours_Cube    -1.673e-07  1.526e-08 -10.965  < 2e-16 ***
## time_spend_company_Cube      -1.008e-03  2.342e-04  -4.303 1.70e-05 ***
## Work_accident_Cube                   NA         NA      NA       NA
## promotion_last_5years_Cube           NA         NA      NA       NA
## salary_Cube                          NA         NA      NA       NA
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2014 on 10482 degrees of freedom
## Multiple R-squared:  0.3407, Adjusted R-squared:  0.3397
## F-statistic: 338.6 on 16 and 10482 DF,  p-value: < 2.2e-16
##
## Start:  AIC=-33629.97
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     salary + last_evaluation_Square + number_project_Square +
##     average_montly_hours_Square + time_spend_company_Square +
##     Work_accident_Square + promotion_last_5years_Square + salary_Square +
##     last_evaluation_Cube + number_project_Cube + average_montly_hours_Cube +
##     time_spend_company_Cube + Work_accident_Cube + promotion_last_5years_Cube +
##     salary_Cube
##
##
## Step:  AIC=-33629.97
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     salary + last_evaluation_Square + number_project_Square +
##     average_montly_hours_Square + time_spend_company_Square +
##     Work_accident_Square + promotion_last_5years_Square + salary_Square +
##     last_evaluation_Cube + number_project_Cube + average_montly_hours_Cube +
##     time_spend_company_Cube + Work_accident_Cube + promotion_last_5years_Cube
##
##
## Step:  AIC=-33629.97
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     salary + last_evaluation_Square + number_project_Square +
##     average_montly_hours_Square + time_spend_company_Square +
##     Work_accident_Square + promotion_last_5years_Square + salary_Square +
##     last_evaluation_Cube + number_project_Cube + average_montly_hours_Cube +
##     time_spend_company_Cube + Work_accident_Cube
##
##
## Step:  AIC=-33629.97
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     salary + last_evaluation_Square + number_project_Square +
##     average_montly_hours_Square + time_spend_company_Square +
##     Work_accident_Square + promotion_last_5years_Square + salary_Square +
##     last_evaluation_Cube + number_project_Cube + average_montly_hours_Cube +
##     time_spend_company_Cube
##
##
## Step:  AIC=-33629.97
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     salary + last_evaluation_Square + number_project_Square +
##     average_montly_hours_Square + time_spend_company_Square +
##     Work_accident_Square + salary_Square + last_evaluation_Cube +
##     number_project_Cube + average_montly_hours_Cube + time_spend_company_Cube
##
##
## Step:  AIC=-33629.97
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     salary + last_evaluation_Square + number_project_Square +
##     average_montly_hours_Square + time_spend_company_Square +
##     salary_Square + last_evaluation_Cube + number_project_Cube +
##     average_montly_hours_Cube + time_spend_company_Cube
##
```

```
##                               Df Sum of Sq    RSS    AIC
## - salary_Square                1    0.0029 425.24 -33632
## - salary                       1    0.0046 425.24 -33632
## - number_project_Cube          1    0.0163 425.25 -33632
## - promotion_last_5years        1    0.0275 425.26 -33631
## - Work_accident                1    0.0617 425.30 -33630
## <none>                                     425.23 -33630
## - time_spend_company_Cube      1    0.7510 425.98 -33613
## - last_evaluation_Cube         1    1.1373 426.37 -33604
## - time_spend_company_Square    1    1.1825 426.42 -33603
## - last_evaluation_Square       1    1.2308 426.46 -33602
## - last_evaluation              1    1.4147 426.65 -33597
## - number_project_Square        1    1.6629 426.90 -33591
## - time_spend_company           1    1.9730 427.21 -33583
## - average_montly_hours         1    2.4093 427.64 -33573
## - average_montly_hours_Square  1    3.6309 428.86 -33543
## - average_montly_hours_Cube    1    4.8776 430.11 -33512
## - number_project               1    5.3634 430.60 -33500
##
## Step:  AIC=-33631.89
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     salary + last_evaluation_Square + number_project_Square +
##     average_montly_hours_Square + time_spend_company_Square +
##     last_evaluation_Cube + number_project_Cube + average_montly_hours_Cube +
##     time_spend_company_Cube
##
##                               Df Sum of Sq    RSS    AIC
## - salary                       1    0.0063 425.24 -33634
## - number_project_Cube          1    0.0166 425.25 -33633
## - promotion_last_5years        1    0.0274 425.26 -33633
## - Work_accident                1    0.0615 425.30 -33632
## <none>                                     425.24 -33632
## + salary_Square                1    0.0029 425.23 -33630
## + salary_Cube                  1    0.0029 425.23 -33630
## - time_spend_company_Cube      1    0.7541 425.99 -33615
## - last_evaluation_Cube         1    1.1377 426.37 -33606
## - time_spend_company_Square    1    1.1854 426.42 -33605
## - last_evaluation_Square       1    1.2309 426.47 -33604
## - last_evaluation              1    1.4145 426.65 -33599
## - number_project_Square        1    1.6670 426.90 -33593
## - time_spend_company           1    1.9759 427.21 -33585
## - average_montly_hours         1    2.4082 427.64 -33575
## - average_montly_hours_Square  1    3.6296 428.87 -33545
## - average_montly_hours_Cube    1    4.8761 430.11 -33514
## - number_project               1    5.3717 430.61 -33502
##
## Step:  AIC=-33633.74
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     last_evaluation_Square + number_project_Square + average_montly_hours_Square +
##     time_spend_company_Square + last_evaluation_Cube + number_project_Cube +
##     average_montly_hours_Cube + time_spend_company_Cube
##
##                               Df Sum of Sq    RSS    AIC
## - number_project_Cube          1    0.0172 425.26 -33635
## - promotion_last_5years        1    0.0303 425.27 -33635
## - Work_accident                1    0.0613 425.30 -33634
## <none>                                     425.24 -33634
## + salary                       1    0.0063 425.24 -33632
## + salary_Square                1    0.0047 425.24 -33632
## + salary_Cube                  1    0.0030 425.24 -33632
## - time_spend_company_Cube      1    0.7545 426.00 -33617
## - last_evaluation_Cube         1    1.1351 426.38 -33608
## - time_spend_company_Square    1    1.1877 426.43 -33606
## - last_evaluation_Square       1    1.2286 426.47 -33605
```

```
## - last_evaluation_Square       1    1.2286 426.47 -33605
## - last_evaluation              1    1.4124 426.66 -33601
## - number_project_Square        1    1.6756 426.92 -33594
## - time_spend_company           1    1.9806 427.22 -33587
## - average_montly_hours         1    2.4185 427.66 -33576
## - average_montly_hours_Square  1    3.6436 428.89 -33546
## - average_montly_hours_Cube    1    4.8937 430.14 -33516
## - number_project               1    5.3933 430.64 -33503
##
## Step:  AIC=-33635.31
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + promotion_last_5years +
##     last_evaluation_Square + number_project_Square + average_montly_hours_Square +
##     time_spend_company_Square + last_evaluation_Cube + average_montly_hours_Cube +
##     time_spend_company_Cube
##
##                              Df Sum of Sq    RSS    AIC
## - promotion_last_5years       1     0.031 425.29 -33637
## - Work_accident               1     0.062 425.32 -33636
## <none>                                    425.26 -33635
## + number_project_Cube         1     0.017 425.24 -33634
## + salary                      1     0.007 425.25 -33633
## + salary_Square               1     0.005 425.25 -33633
## + salary_Cube                 1     0.003 425.26 -33633
## - time_spend_company_Cube     1     0.753 426.01 -33619
## - last_evaluation_Cube        1     1.124 426.38 -33610
## - time_spend_company_Square   1     1.189 426.45 -33608
## - last_evaluation_Square      1     1.218 426.48 -33607
## - last_evaluation             1     1.404 426.66 -33603
## - time_spend_company          1     1.989 427.25 -33588
## - average_montly_hours        1     2.401 427.66 -33578
## - average_montly_hours_Square 1     3.629 428.89 -33548
## - average_montly_hours_Cube   1     4.884 430.14 -33517
## - number_project              1    77.154 502.41 -31887
## - number_project_Square       1    89.440 514.70 -31633
##
## Step:  AIC=-33636.55
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
##     time_spend_company + Work_accident + last_evaluation_Square +
##     number_project_Square + average_montly_hours_Square + time_spend_company_Square +
##     last_evaluation_Cube + average_montly_hours_Cube + time_spend_company_Cube
##
##                              Df Sum of Sq    RSS    AIC
## - Work_accident               1     0.066 425.36 -33637
## <none>                                    425.29 -33637
## + promotion_last_5years       1     0.031 425.26 -33635
## + promotion_last_5years_Square 1    0.031 425.26 -33635
## + promotion_last_5years_Cube  1     0.031 425.26 -33635
## + number_project_Cube         1     0.018 425.27 -33635
## + salary                      1     0.010 425.28 -33635
## + salary_Square               1     0.008 425.28 -33635
## + salary_Cube                 1     0.005 425.29 -33635
## - time_spend_company_Cube     1     0.764 426.05 -33620
## - last_evaluation_Cube        1     1.118 426.41 -33611
## - time_spend_company_Square   1     1.205 426.50 -33609
## - last_evaluation_Square      1     1.212 426.50 -33609
## - last_evaluation             1     1.397 426.69 -33604
## - time_spend_company          1     2.009 427.30 -33589
## - average_montly_hours        1     2.399 427.69 -33580
## - average_montly_hours_Square 1     3.625 428.92 -33549
## - average_montly_hours_Cube   1     4.880 430.17 -33519
## - number_project              1    77.222 502.51 -31887
## - number_project_Square       1    89.529 514.82 -31633
##
## Step:  AIC=-33636.92
## satisfaction_level ~ last_evaluation + number_project + average_montly_hours +
```

```
##     time_spend_company + last_evaluation_Square + number_project_Square +
##     average_montly_hours_Square + time_spend_company_Square +
##     last_evaluation_Cube + average_montly_hours_Cube + time_spend_company_Cube
##
##                               Df Sum of Sq    RSS    AIC
## <none>                                     425.36 -33637
## + Work_accident                1     0.066 425.29 -33637
## + Work_accident_Square         1     0.066 425.29 -33637
## + Work_accident_Cube           1     0.066 425.29 -33637
## + promotion_last_5years        1     0.035 425.32 -33636
## + promotion_last_5years_Square 1     0.035 425.32 -33636
## + promotion_last_5years_Cube   1     0.035 425.32 -33636
## + number_project_Cube          1     0.019 425.34 -33635
## + salary                       1     0.010 425.35 -33635
## + salary_Square                1     0.008 425.35 -33635
## + salary_Cube                  1     0.006 425.35 -33635
## - time_spend_company_Cube      1     0.772 426.13 -33620
## - last_evaluation_Cube         1     1.111 426.47 -33612
## - last_evaluation_Square       1     1.206 426.56 -33609
## - time_spend_company_Square    1     1.220 426.58 -33609
## - last_evaluation              1     1.391 426.75 -33605
## - time_spend_company           1     2.033 427.39 -33589
## - average_montly_hours         1     2.400 427.76 -33580
## - average_montly_hours_Square  1     3.629 428.99 -33550
## - average_montly_hours_Cube    1     4.887 430.24 -33519
## - number_project               1    77.760 503.12 -31876
## - number_project_Square        1    90.136 515.49 -31621
##
## Call:
## lm(formula = satisfaction_level ~ last_evaluation + number_project +
##     average_montly_hours + time_spend_company + last_evaluation_Square +
##     number_project_Square + average_montly_hours_Square + time_spend_company_Square +
##     last_evaluation_Cube + average_montly_hours_Cube + time_spend_company_Cube,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67443 -0.12052 -0.00641  0.15218  0.68945
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.947e-01  1.868e-01  -1.042    0.297
## last_evaluation             3.903e+00  6.664e-01   5.856 4.88e-09 ***
## number_project              4.455e-01  1.017e-02  43.785  < 2e-16 ***
## average_montly_hours       -1.397e-02  1.817e-03  -7.692 1.58e-14 ***
## time_spend_company         -1.324e-01  1.871e-02  -7.079 1.55e-12 ***
## last_evaluation_Square     -5.267e+00  9.660e-01  -5.453 5.08e-08 ***
## number_project_Square      -5.905e-02  1.253e-03 -47.141  < 2e-16 ***
## average_montly_hours_Square  8.718e-05  9.216e-06   9.459  < 2e-16 ***
## time_spend_company_Square    2.114e-02  3.855e-03   5.484 4.25e-08 ***
## last_evaluation_Cube         2.376e+00  4.540e-01   5.234 1.69e-07 ***
## average_montly_hours_Cube   -1.666e-07  1.518e-08 -10.976  < 2e-16 ***
## time_spend_company_Cube     -1.021e-03  2.339e-04  -4.364 1.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2014 on 10487 degrees of freedom
## Multiple R-squared:  0.3405, Adjusted R-squared:  0.3398
## F-statistic: 492.3 on 11 and 10487 DF,  p-value: < 2.2e-16
```




```
##        RMSE    R2.fit    R2.lwr    R2.upr
## 1 0.3815705 0.338463 0.3383658 0.3385583
```

```
#using XGBoost
CreateXGBoostModel(train,test,test$satisfaction_level,number=10,classification=FALSE)
```

```
## Loading required package: lattice
```

```
##        RMSE       R2
## 1 0.1854616 0.451467
```

```
#=====================================================================================================
```