ProjectIST707\_Discussion2

Subbu Kandhaswamy

11/12/2019

## Summary

Retaining human capital and talent management is increasingly becoming a competitive game within an already complex and adaptive economic system. In a tight labor market perks like flexibility to work from home or anywhere, signing bonuses, and equity in the form of restricted stock units are just a couple of the benefits that companies are offering their employees.

The data set that the data analytics team is exploring is the IBM HR Analytics Employee Attrition & Performance. The aim is to use this data to predict the attrition of the most valuable employees.

Questions: 1. What are the key factors that lead to employee attrition?

1. How can companies improve and in what areas in order to retain their best employees and keep them from accepting a more attractive offer from a competitor?
2. How can we predict which employees are at a higher risk for leaving?
3. What can we do to proactively address these employees with a high risk for leaving to course correct their potential actions?

\*There are likely many more questions that will arise with further exploration of the data. The team believes that these preliminary questions will open the doors to other insightful analysis.

Data Set: Contains 1,470 observations and 35 variables and a total of 51,450 values.

#Step:1 Import Dataset   
  
HRdata <- read.csv("HREmployeeAttrition.csv")   
names(HRdata)

## [1] "ï..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"

# We cam see column name for age "ï..Age" need to be updated on this dataset.  
colnames(HRdata)[1] <- "Age"  
library(caret) # models

## Loading required package: lattice

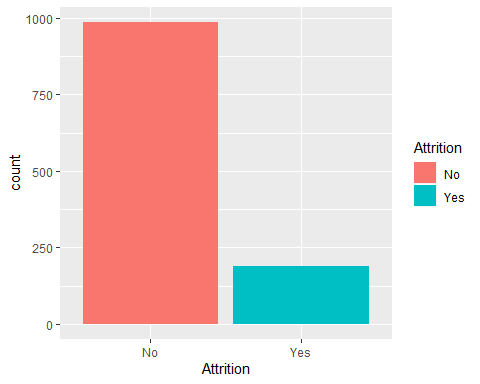
## Loading required package: ggplot2

library(lattice) # robust & elegant visualization  
library(ggplot2) # plotting graphs  
library(grid) # modify specific elements of a plot  
library(gridExtra) # to arrange multiple grid-based plots  
#Set seed number as starting point to generate random numbers  
set.seed(12345)   
# Divide our dataset into training(75%) and testing(25%)  
inTrain <- createDataPartition(HRdata$Attrition,p=0.80,list = FALSE)  
Training <- HRdata[inTrain,]   
Testing <- HRdata[-inTrain,]  
str(Training)

## 'data.frame': 1177 obs. of 35 variables:  
## $ Age : int 49 37 27 32 59 30 38 36 35 29 ...  
## $ Attrition : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...  
## $ BusinessTravel : Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 2 3 3 2 3 3 2 3 3 3 ...  
## $ DailyRate : int 279 1373 591 1005 1324 1358 216 1299 809 153 ...  
## $ Department : Factor w/ 3 levels "Human Resources",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ DistanceFromHome : int 8 2 2 2 3 24 23 27 16 15 ...  
## $ Education : int 1 2 1 2 3 1 3 3 3 2 ...  
## $ EducationField : Factor w/ 6 levels "Human Resources",..: 2 5 4 2 4 2 2 4 4 2 ...  
## $ EmployeeCount : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ EmployeeNumber : int 2 4 7 8 10 11 12 13 14 15 ...  
## $ EnvironmentSatisfaction : int 3 4 1 4 3 4 4 3 1 4 ...  
## $ Gender : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 2 2 2 2 1 ...  
## $ HourlyRate : int 61 92 40 79 81 67 44 94 84 49 ...  
## $ JobInvolvement : int 2 2 3 3 4 3 2 3 4 2 ...  
## $ JobLevel : int 2 1 1 1 1 1 3 2 1 2 ...  
## $ JobRole : Factor w/ 9 levels "Healthcare Representative",..: 7 3 3 3 3 3 5 1 3 3 ...  
## $ JobSatisfaction : int 2 3 2 4 1 3 3 3 2 3 ...  
## $ MaritalStatus : Factor w/ 3 levels "Divorced","Married",..: 2 3 2 3 2 1 3 2 2 3 ...  
## $ MonthlyIncome : int 5130 2090 3468 3068 2670 2693 9526 5237 2426 4193 ...  
## $ MonthlyRate : int 24907 2396 16632 11864 9964 13335 8787 16577 16479 12682 ...  
## $ NumCompaniesWorked : int 1 6 9 0 4 1 0 6 0 0 ...  
## $ Over18 : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...  
## $ OverTime : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 1 1 1 2 ...  
## $ PercentSalaryHike : int 23 15 12 13 20 22 21 13 13 12 ...  
## $ PerformanceRating : int 4 3 3 3 4 4 4 3 3 3 ...  
## $ RelationshipSatisfaction: int 4 2 4 3 1 2 2 2 3 4 ...  
## $ StandardHours : int 80 80 80 80 80 80 80 80 80 80 ...  
## $ StockOptionLevel : int 1 0 1 0 3 1 0 2 1 0 ...  
## $ TotalWorkingYears : int 10 7 6 8 12 1 10 17 6 10 ...  
## $ TrainingTimesLastYear : int 3 3 3 2 3 2 2 3 5 3 ...  
## $ WorkLifeBalance : int 3 3 3 2 2 3 3 2 3 3 ...  
## $ YearsAtCompany : int 10 0 2 7 1 1 9 7 5 9 ...  
## $ YearsInCurrentRole : int 7 0 2 7 0 0 7 7 4 5 ...  
## $ YearsSinceLastPromotion : int 1 0 2 3 0 0 1 7 0 0 ...  
## $ YearsWithCurrManager : int 7 0 2 6 0 0 8 7 3 8 ...

## Data Exploration

Checking the attrition percentage using our Training data set we could see 16% approximation of employees leaving organization.



##   
## No Yes   
## 0.8385726 0.1614274

