ProjectIST707\_Discussion2

Sangam\_Matt\_Subbu

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   
  
IBMHRData <- read.csv("HREmployeeAttrition.csv")   
names(IBMHRData)

## [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(IBMHRData)[1] <- "Age"  
   
library(caret) # models

## Loading required package: lattice

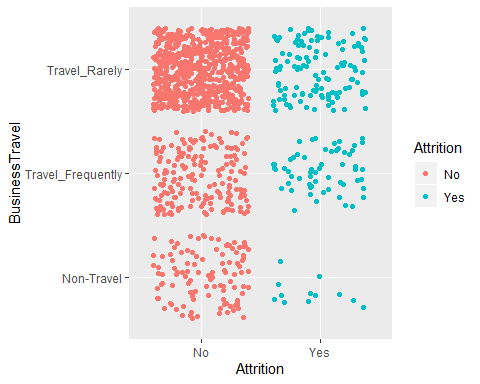
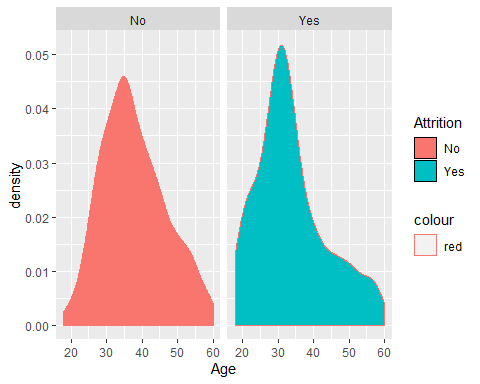
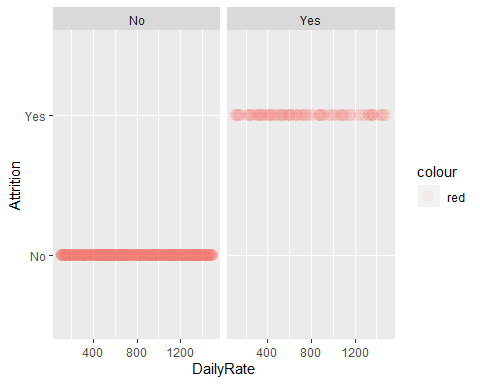
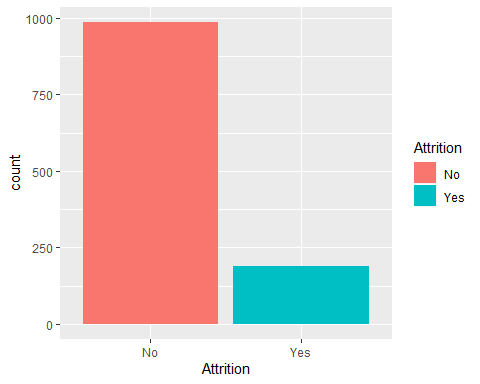
## Loading required package: ggplot2

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(IBMHRData$Attrition,p=0.80,list = FALSE)  
Training <- IBMHRData[inTrain,]   
Testing <- IBMHRData[-inTrain,]  
str(Training)

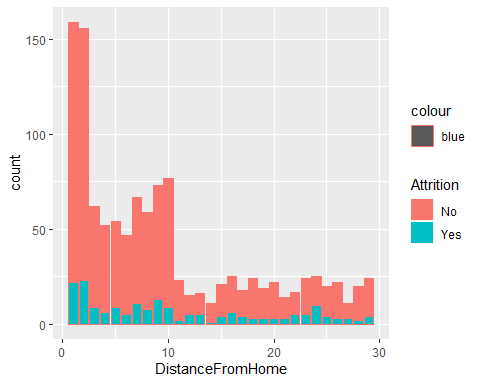
## 'data.frame': 1177 obs. of 35 variables:  
## $ Age : int 49 37 27 32 59 30 38 36 35 29 ...  
## $ Attrition : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 1 ...  
## $ BusinessTravel : Factor w/ 3 levels "Non-Travel","Travel\_Frequently",..: 2 3 3 2 3 3 2 3 3 3 ...  
## $ DailyRate : int 279 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

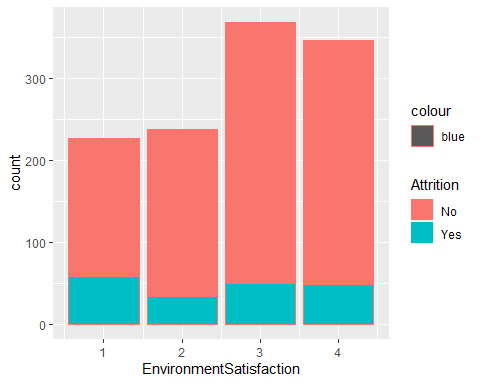
Looking at the attrition percentage from the training data set, we could see the attrition percentage is around 16% (i.e., 16% - employees leaving organization.) In our Data exploration process, we are trying to understand how each variable in the dataset influencing the attrition of the organization. We ’re using ggplot2 to plot and measure this and also using lattice, grid, gridExtra for simple and better visualization.

 ## Lets continue with second set of variables and see it’s influence on the attrition

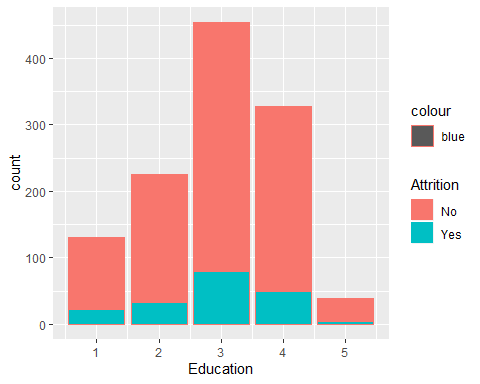
#Distance from Home : Unlike the norms, most of employees who have left the organization were closer to the Office.   
  
distPlot <- ggplot(Training,aes(DistanceFromHome,fill=Attrition, color = "blue"))+geom\_bar()   
distPlot



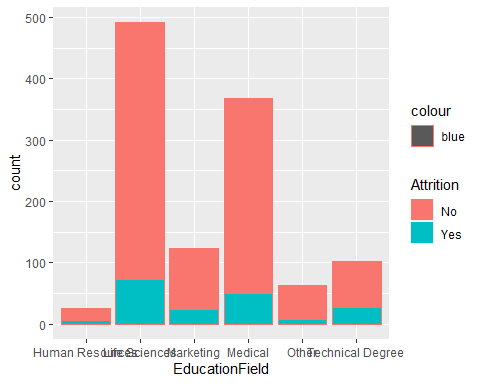
#Environment Satisfaction: Ratings stand for - 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’ . We don’t see any distinguishable feature  
  
envPlot <- ggplot(Training,aes(EnvironmentSatisfaction,fill=Attrition,color = "blue"))+geom\_bar()  
envPlot



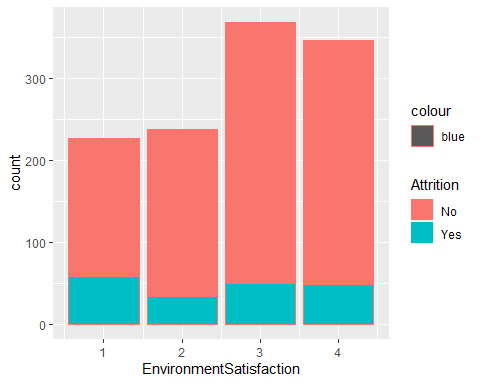
#Education : Docts in the lowest numbers - probably due to very low percentage of doctors working.  
eduPlot <- ggplot(Training,aes(Education,fill=Attrition,color = "blue"))+geom\_bar()  
eduPlot



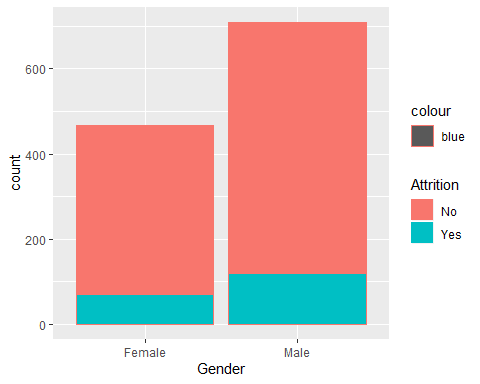
#Education Field: On lines of the trend in Departments, a minority of HR educated employees leave and it is majorly because of low number of people  
  
edufieldPlot <- ggplot(Training,aes(EducationField,fill=Attrition,color = "blue"))+geom\_bar()   
edufieldPlot



#Environment Satisfaction: Ratings stand for - 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’ . We don’t see any distinguishable feature  
  
envPlot <- ggplot(Training,aes(EnvironmentSatisfaction,fill=Attrition,color = "blue"))+geom\_bar()   
envPlot



#Gender: As you can see majority of them left are Male and the reason might be because around 61% of employees in our dataset are Male  
  
genPlot <- ggplot(Training,aes(Gender,fill=Attrition,color = "blue"))+geom\_bar()   
genPlot



#grid.arrange(distPlot,eduPlot,edufieldPlot,envPlot,genPlot,ncol=2,top = " Attrition by Variable - Contd.. ")

## Remaining workflow

* Sampling - (if required ONLY) Not sure if we need to use techniques like Oversampling or undersampling, because there are advantages and disadvantages. And for this analysis we will not get improved results using these techniques so we can continue using the sample we have for model building.

#Correlation (Ideas) - Years at Company, Years in Curr Role, Years with Curr Manager & Years Since Last Promotion  
- Job Level & Monthly Income  
- Percent Salary Hike & Performance Ratiing

Model Building (Decision Tree Models),

Select right model

& final result