Lab Report of IBM Employee Retention Data Set

Yiqiao Yin

June 21, 2019

Table of Contents

Introduction

Companies hire many employees every year. To create a positive working and learning environment, firms invest time and money in trianing the new members and also to get existing employees involved as well. The goal of these programs aim to increase the effectiveness of the employees and in doing so the firm as a whole can have better output in long run.

Human resource analytics (or HR analytics) is an important area in this field and it refers to applying statistical and analytic modelling to understand or extract the insight in hoping of improving employee performance and better return on investment. This is a field of analytics work that is not just gathering data but also to provide insight in each step of the process targeting on making sound decisions about how to improve overall satisfactory for new and existing employees.

Goal: Employee Attrition

The single most important feature we are interested in is attrition. Attrition in human resources refer to the gradual loss of employees over time. In general relatively high attrition is problematic for companies. Human Resource professionals often asume a leadership role in designing company compensation programs, work culture and motivation systems that help the organization retain top employees.

This is a significant problem because high employee attrition is a huge cost to a firm. Procedures such as job postings, hiring processes, paperwork, and new recruitment training are some of the most expensive costs of losing or replacing employees. On top of these costs, regular employee turnover prevents a firm from increasing its collective knowledge base and experience over time which in most industries can be crucial to the success of a company. Moreover, this also in some way affects customers and revenue streams because some customers prefer to interact with familiar faces. Errors and issues are more likely if you constantly have new workers.

Data: IBM HR Employee Dataset

To investigate this topic, I use IBM HR Analytics Employee Attrition and Performance Dataset. There are total of 1470 samples and 35 features. Among the target, Attrition, there are 237 candidates committed to Yes (i.e. left the company) and the rest 1233 candidates committed to No (i.e. stayed at the company).

```
# Set Working Directory
path <- "C:/Users/eagle/OneDrive/HR_Employee_Retention_IBM"
setwd(paste0(path, "/data"))

# Compile Data
all <- read.csv("data.csv")

# Define Response
all$Attrition <- as.numeric(all$Attrition) - 1L

# Define Features (i.e. these are the variables)
all <- cbind(all$Attrition, all[, -2])
colnames(all)[1] <- "Attrition"; colnames(all)[2] <- "Age"
raw_data <- all

# Convert all features into numerical value
for (j in 2:ncol(all)) {all[, j] <- as.numeric(all[, j])}
colnames(all)[1] <- "Attrition"; colnames(all)[2] <- "Age"
all <- all[, -c(9, 22, 27)] # These variables (EmployeeCount, Over18, Standar
dHours) are constants so delete them.

# Preview
print(paste0("The first three rows of the data looks like: ")); all[1:3, ]
```

```
## [1] "The first three rows of the data looks like: "

##   Attrition Age BusinessTravel DailyRate Department DistanceFromHome
## 1         1  41              3      1102          3                1
## 2         0  49              2       279          2                8
## 3         1  37              3      1373          2                2
##   Education EducationField EmployeeNumber EnvironmentSatisfaction Gender
## 1         2              2              1                        2      1
## 2         1              2              2                        3      2
## 3         2              5              4                        4      2
##   HourlyRate JobInvolvement JobLevel JobRole JobSatisfaction MaritalStatus
## 1         94              3        2       8               4             3
## 2         61              2        2       7               2             2
## 3         92              2        1       3               3             3
##   MonthlyIncome MonthlyRate NumCompaniesWorked OverTime PercentSalaryHike
## 1          5993       19479                  8        2                11
## 2          5130       24907                  1        1                23
## 3          2090        2396                  6        2                15
##   PerformanceRating RelationshipSatisfaction StockOptionLevel
## 1                 3                        1                0
## 2                 4                        4                1
## 3                 3                        2                0
##   TotalWorkingYears TrainingTimesLastYear WorkLifeBalance YearsAtCompany
## 1                 8                     0               1              6
## 2                10                     3               3             10
## 3                 7                     3               3              0
##   YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager
## 1                  4                       0                    5
## 2                  7                       1                    7
## 3                  0                       0                    0
```

```
print(paste0("The dimension (row by cols) of the data set is: ")); dim(all)
```

```
## [1] "The dimension (row by cols) of the data set is: "

## [1] 1470   32
```

We can take a look at the list of variables. Variables such as Attrition, BusinessTravel, Department, and so on are discrete. The rest of variables such as DailyRate or EmployeeNumber are considered as continuous.

```
# Check Levels
levels <- c()
for (j in 1:ncol(all)) {levels <- c(levels, nrow(plyr::count(all[,j])))}
levels <- data.frame(cbind(colnames(all), levels))
colnames(levels) <- c("Variable_Names","Number_of_Levels"); levels
```

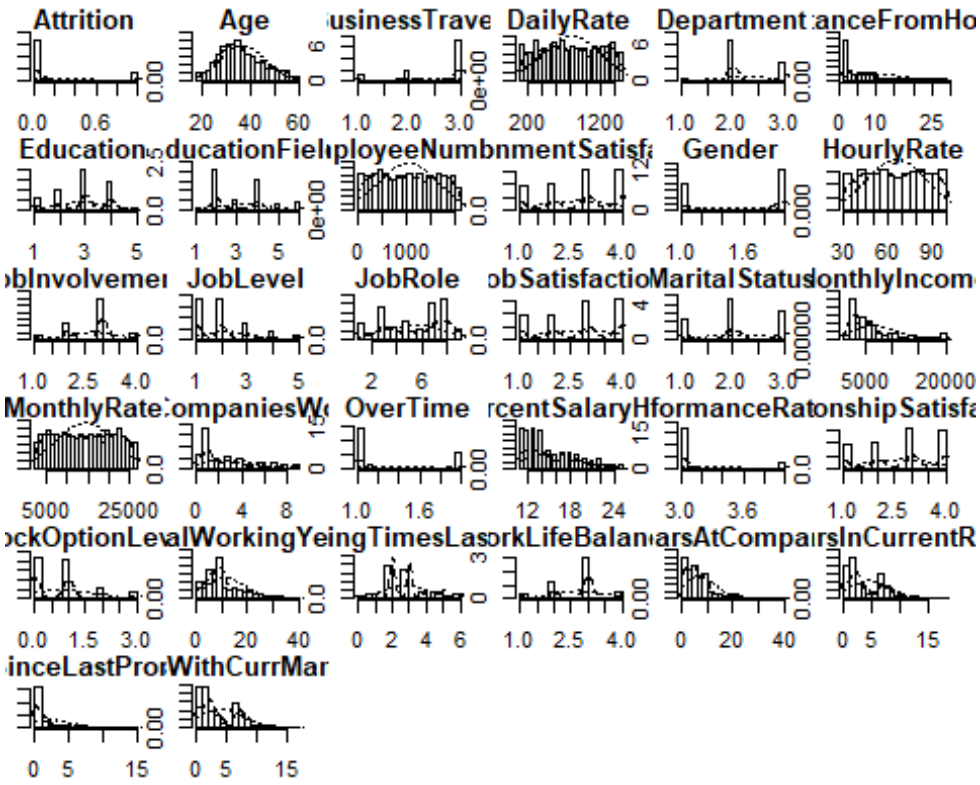
```
##              Variable_Names Number_of_Levels
## 1                 Attrition                2
## 2                       Age               43
## 3            BusinessTravel                3
```

```
## 4                 DailyRate               886
## 5                Department                 3
## 6          DistanceFromHome                29
## 7                 Education                 5
## 8             EducationField                 6
## 9            EmployeeNumber              1470
## 10  EnvironmentSatisfaction                 4
## 11                    Gender                 2
## 12              HourlyRate                71
## 13          JobInvolvement                 4
## 14                JobLevel                 5
## 15                 JobRole                 9
## 16          JobSatisfaction                 4
## 17           MaritalStatus                 3
## 18           MonthlyIncome              1349
## 19             MonthlyRate              1427
## 20       NumCompaniesWorked                10
## 21                 OverTime                 2
## 22        PercentSalaryHike                15
## 23        PerformanceRating                 2
## 24 RelationshipSatisfaction                 4
## 25         StockOptionLevel                 4
## 26        TotalWorkingYears                40
## 27    TrainingTimesLastYear                 7
## 28          WorkLifeBalance                 4
## 29           YearsAtCompany                37
## 30       YearsInCurrentRole                19
## 31  YearsSinceLastPromotion                16
## 32     YearsWithCurrManager                18
```

EDA: Exploratory Data Analysis

Let us take a look at the distribution matrix of all the variables. We can Age follows a distribution that is similar to normal distribution. However, MonthlyIncome may look more like a Poisson process in the sense that most of the sample falls on the lower end of the distribution.

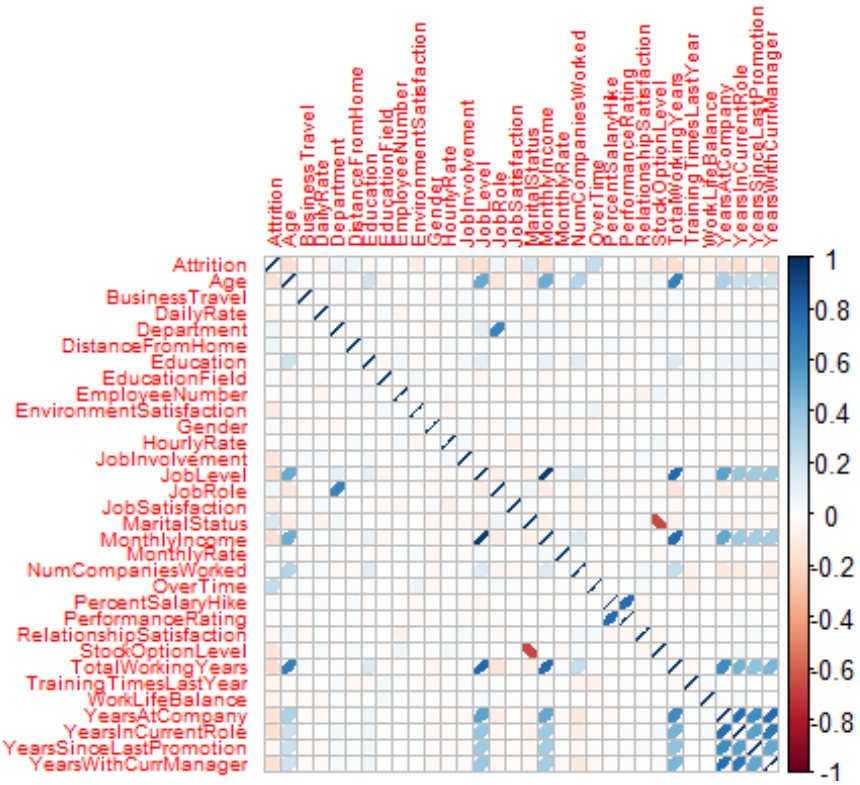
```
psych::multi.hist(all)
```



With each distribution in mind, let us also take a look at the correlation plot. For visualization purpose, we align the variables in alphabetical order on both axis. The diagonal is 100% because that is the correlation for each variable and itself. The matrix is symmetric with legend labeled on the right. The legend is color coded from -1 (red) to 1 (blue). Blue means positively correlated while red means negatively correlated. We can see that Education is positively correlated with Age. We can also see that JobLevel is highly correlated Age as well. It is the same with MOnthlyIncome and TotalWorkingYears. Based on the correlation table, I make the following observations:

- Based on correlation, Attrition is associated negatively with Age, JobInvolvement, JobLevel, Jobsatisfaction, MonthlyIncome, StockOptionLevel, TotalWorkingYears, YearsAtCompany, YearsInCurrentCompany, YearsInCurrentRole, and YearsWithCurrManager. However, Attrition is positively associated with MaritalStatus and OverTime.

```
M <- cor(all); corrplot::corrplot(M, method = "ellipse", tl.cex = 0.6)
```



- Among those who contribute to Attrition, the highest percentage falls in Life Sciences and the second falls in Medical. In other words, at IBM, the employees that left the firm mostly come from Life Sciences and then perhaps Medical department.

```
sub_data <- cbind(raw_data$Attrition, raw_data$EducationField); levels(raw_da
ta$EducationField)

## [1] "Human Resources"  "Life Sciences"    "Marketing"
## [4] "Medical"          "Other"            "Technical Degree"

Pi <- apply(sub_data,1,paste0,collapse="_"); plyr::count(Pi)

##       x freq
## 1  0_1   20
## 2  0_2  517
## 3  0_3  124
## 4  0_4  401
## 5  0_5   71
## 6  0_6  100
## 7  1_1    7
## 8  1_2   89
## 9  1_3   35
## 10 1_4   63
## 11 1_5   11
## 12 1_6   32
```

```
data.frame(EducationField = levels(raw_data$EducationField),
           Percentage = round(plyr::count(Pi)[7:12, 2]/sum(plyr::count(Pi)[7:
12, 2]),3))
```

```
##     EducationField Percentage
## 1  Human Resources      0.030
## 2    Life Sciences      0.376
## 3        Marketing      0.148
## 4          Medical      0.266
## 5            Other      0.046
## 6 Technical Degree      0.135
```

- Based on Age and WorkLifeBalance, we discovered that for those who committed to Attrition age 29 and 31 with WorkLifeBalance to be 3 happened the most frequently, both at 18.6%.

```
sub_data <- cbind(raw_data$Attrition, raw_data$Age, raw_data$WorkLifeBalance)
Pi <- apply(sub_data,1,paste0,collapse="_"); # plyr::count(Pi)
tmp <- plyr::count(Pi)[152:nrow(plyr::count(Pi)), ]
tmp_top_6 <- head(tmp[order(tmp$freq, decreasing = TRUE), ])
tmp_top_6$percent <- round(c(tmp_top_6$freq / sum(tmp_top_6$freq)), 3)
data.frame(tmp_top_6)
```

```
##          x freq percent
## 180 1_29_3    8   0.186
## 187 1_31_3    8   0.186
## 172 1_26_3    7   0.163
## 175 1_28_2    7   0.163
## 195 1_33_3    7   0.163
## 191 1_32_3    6   0.140
```

- For the people who left the firm (committed to Yes to Attrition), the most common JobRole is Laboratory Technician and Sales Representative. From our analysis below, we see that the Laboratory Technician who spent a year at the firm and than left sat on a high of 30.9% among those who committed Yes to Attrition. The second is Sales Representative that stayed at the firm for a year, at 9.1%. The third group of people who stayed at the firm for a year and left are Research Scientist, at a shy of 17.6%. These are the top three demographics that contribute to the Attrition the highest.

```
sub_data <- cbind(raw_data$Attrition, raw_data$JobRole, raw_data$YearsAtCompa
ny)
levels(raw_data$JobRole)
```

```
## [1] "Healthcare Representative" "Human Resources"
## [3] "Laboratory Technician"     "Manager"
## [5] "Manufacturing Director"    "Research Director"
## [7] "Research Scientist"        "Sales Executive"
## [9] "Sales Representative"
```

```
Pi <- apply(sub_data,1,paste0,collapse="_")
tmp <- plyr::count(Pi)[193:nrow(plyr::count(Pi)), ]
tmp_top_6 <- head(tmp[order(tmp$freq, decreasing = TRUE), ])
```

```
tmp_top_6$percent <- round(c(tmp_top_6$freq / sum(tmp_top_6$freq)), 3)
data.frame(tmp_top_6)

##         x freq percent
## 206 1_3_1   21   0.309
## 261 1_9_1   13   0.191
## 230 1_7_1   12   0.176
## 250 1_8_2    8   0.118
## 205 1_3_0    7   0.103
## 209 1_3_2    7   0.103
```

- Moreover, we can look at how JobRole and YearsSinceLastPromotion affect Attrition. From results below, we see that for those who committed to Attrition, the highest group of people are Lab Technician that did not have any promotion. This makese sense as this is correlated with the previous finding. If a person stayed at a firm for a year or less, chances are this person did not receive any promotion before the job ended. The next is Sales Representative who did not receive promotions. The top two groups described about takes up 29.8% and 18.2% of those who committed to Yes to Attrition.

```
sub_data <- cbind(raw_data$Attrition, raw_data$JobRole, raw_data$YearsSinceLa
stPromotion)
levels(raw_data$JobRole)

## [1] "Healthcare Representative" "Human Resources"
## [3] "Laboratory Technician"     "Manager"
## [5] "Manufacturing Director"    "Research Director"
## [7] "Research Scientist"        "Sales Executive"
## [9] "Sales Representative"

Pi <- apply(sub_data,1,paste0,collapse="_")
tmp <- plyr::count(Pi)[111:nrow(plyr::count(Pi)), ]
tmp_top_6 <- head(tmp[order(tmp$freq, decreasing = TRUE), ])
tmp_top_6$percent <- round(c(tmp_top_6$freq / sum(tmp_top_6$freq)), 3)
data.frame(tmp_top_6)

##         x freq percent
## 121 1_3_0   36   0.298
## 159 1_9_0   22   0.182
## 139 1_7_0   20   0.165
## 122 1_3_1   17   0.140
## 147 1_8_0   17   0.140
## 148 1_8_1    9   0.074
```

Lab Procedure

The following section we search for an algorithm to further explore our target, Attrition, which is measured by 0 if the employee stays and 1 if the employee leaves.

Bagging

Consider a regression problem. Suppose we fit a model to our training data $Z = \{x_1, y_1), \ldots, (x_N, y_N)\}$, obtaiing the prediction $\hat{f}(x)$ at input x. Bootstrap aggregation or bagging averages this prediction over a collection of bootstrap samples, thereby reducing its variance. For each bootstrap sample Z^{*b} with $b = 1,2, \ldots, B$,we fit out model, giving prediction $\hat{f}^{*b}(x)$. The baggting estimate is defined by

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x).$$

Gradient Boosting Machine

To tackle a data set with Gradient Boosting Machine, we first need to define loss function using $f(x)$ to predict y on the training data, which is

$$L(f) = \sum_{i=1}^{N} L\left(y_i, f(x_i)\right).$$

Here the goal is to mimnimize $L(f)$ with respect to f while f can be trained with a sum of trees. Hence, the algorithm has the following objective function

$$\hat{f} = \underset{f}{\text{argmin}} L(f)$$

Naive Bayes

The naive Bayes model assumes given a class $G = j$, the features X_k are independent:

$$f_j(X) = \prod_{k=1}^{p} f_{jk}\left(X_k\right)$$

Then we can derive the following

$$\begin{aligned}
\log\frac{\mathrm{P}(G = l|X)}{\mathrm{P}(G = J|X)} &= \log\frac{\pi_l f_l(X)}{\pi_J f_J(X)} \\
&= \log\frac{\pi_l}{\pi_J} + \sum_{k=1}^{p} \log\frac{f_{lk}(X_k)}{f_{JK}(X_k)} \\
&= \alpha_l + \sum_{k=1}^{p} g_{lk}\left(X_k\right)
\end{aligned}$$

This has the form of a generalized additive model which can train machine to learn it.

Linear Model or Least Squares

Linear model is the most common and famous algorithm and remained mainstream of statistics for decades. Given features $X^T = (X_1, \ldots, X_p)$, we predict the output Attrition using the following model

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j.$$

In this case, we can simply write this as inner product of two matrices, i.e. $\hat{Y} = X^T \hat{\beta}$. We fit this model on training set and find the weights of features by picking coefficients β to minimize the following residual sum of squares (RSS)

$$\text{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2,$$

which is a quadratic function of the parameters and there is always a minimum (the most optimal point).

Tree-based Algorithm

Random Forest (RF), iterative Random Forest (iRF), and Bayesian Additive Regression Tree (BART) are all tree-based algorithms. To make metters simple, let us use X_1 and X_2, any two variables in a given data set, as an example. We split at $X_1 = t_1$. Then the region $X_1 \leq t_1$ is split at $X_2 = t_2$ and the region $X_1 > t_1$ is split at $X_1 = t_3$ and so on. What this does is to partition into regresions $R_1, R_2, \ldots$. The corresponding regression model predicts Y with a constant c_m in region R_m, that is,

$$\hat{f}(X) = \sum_{m=1}^{M} c_m I\{(X_1, X_2) \in R_m\}.$$

Lab Result

For lab procedure, we use k-fold cross validation to compare the results of handful of algorithms. We cut data set into k folds. Iteratively, each fold is used as test and the rest folds as training. We test all k folds using all machine learning algorithms. In the end, we average k-fold validating or test set average accuracy as final metric for comparisons.

Important Features

We select features using importance measure based on partitions.

- The first three important variables are EducationField, JobRole, and YearsAtCompany.
- The second two important variables are JobRole, and YearsAtCompany.

```
selected_variables <- data.frame(read.csv(paste0(path, "/results/selected_var
iables.csv")))[, -1]
selected_variables$Variables_Names <- NA
for (i in 1:nrow(selected_variables)) {
  selected_variables[i, 3] <- paste0(colnames(all)[c(as.numeric(unlist(strspl
it(as.character(selected_variables[i, 1]), split="_"))))], collapse = "_") }
selected_variables[1:2, ]
```

```
##   Top.Module  Measure                       Variables_Names
## 1    8_15_29 13.86004 EducationField_JobRole_YearsAtCompany
## 2      15_29 13.30205               JobRole_YearsAtCompany
```

Measurement: AUC

For measurement of Attrition, we use area under curve or AUC which computes the area under curve formed by the axis of recall and precision.

Result

We present test result in Area-Under-Curve (AUC) for selected machine learning algorithm. The best approach, at 90% accuracy, is to use the following variables and use bagging which is aggregate averages from bootstrap results.

```
setwd(paste0(path))
how.many.folds = 5; used_iscore = "used"
final_table <- read.csv(paste0(path, "/results/performance_5_fold_used_iscore
_top_7_var.csv"))[, -1]
data.frame(final_table)
```

```
##      Name Result
## 1 Bagging  0.904
## 2     GBM  0.526
## 3      NB  0.452
## 4      LM  0.745
## 5      RF  0.544
## 6     iRF  0.808
## 7    BART  0.890
```

Summary

This report investigated IBM HR Employee Attrition data set. The analysis extracts insights from data set and conclude the following:

- My analysis identified the following important variables: EducationField, JobRole, and YearsAtCompany.

-For the people who left the firm (committed to Yes to Attrition), the most common JobRole is Laboratory Technician and Sales Representative. From our analysis below, we see that the Laboratory Technician who spent a year at the firm and than left sat on a high of 30.9%

among those who committed Yes to Attrition. The second is Sales Representative that stayed at the firm for a year, at 9.1%. The third group of people who stayed at the firm for a year and left are Research Scientist, at a shy of 17.6%. These are the top three demographics that contribute to the Attrition the highest.

- Based on Age and WorkLifeBalance, we discovered that for those who committed to Attrition age 29 and 31 with WorkLifeBalance to be 3 happened the most frequently, both at 18.6%.

We also tested a variety of approaches in machine learning using linear model, naive Bayes, tree-based algorithms and so on. In the end, we propose using bagging algorithm and such algorithm deliver a solution to predict Attrition given candidates information at 90%.