

Using Predictive Modeling to Analyze Employee Churn

Frederick T. Williams

April 21, 2020

Defining the Business Problem

Within the US, employers face an average employee churn of about 10%-15% annually which can prove costly to companies especially those in their early-stages of growth (Law, 2019). Kolowich (2018) wrote that when valued employees leave abruptly, it is estimated that it costs companies 30% from its other employees' annual salary to hire junior employees. However, that percentage can be increased to 400% when replace more senior position roles (Kolowich, 2018).

For companies finding a replacement difficult because the company is trying to find someone who is as productive as their former employee. The company also must consider the loss of knowledge and business acumen about the company, and time and resources needed to teach the new hire. As a result, this process can be a serious problem for companies that are facing high rates of attrition due to the extra load management being placed on other employees (Ashe-Edmunds, 2017). However, many companies today try to resolve this issue by creating programs that provide training and career development, and improved work-life balance to boost employee retention (Regan, 2020).

The fact that employee churn has and will continue be an issue that companies face, within this data science project I will be creating an model with the IBM dataset, provided by [Kaggle] (<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>). In this analysis, I hope to use this dataset to build a model to predict when employees are going to quit by understanding the main drivers of employee churn.

Importing the data

Let's import the dataset and make of a copy of the source file for this analysis.

```
setwd("C:/Users/frede/OneDrive - Regis University/MSDS_696/MSDS696_Data_Science_Practicum_II/")

# Read Excel file
df_source <- read.csv("Data/WA_Fn-UseC_-HR-Employee-Attrition.csv")
names(df_source)

## [1] "i..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"
```

```
colnames(df_source)[1] <- "Age" # Renaming the column
```

```
# Making copy of the dataset
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.5.3
```

```
HR_data <- copy(df_source)
```

Exploratory Data Analysis

Let's look at the data and see how it is formatted before performing analysis

```
str(HR_data)
```

```
## 'data.frame':  1470 obs. of  35 variables:
## $ Age                : int  41 49 37 33 27 32 59 30 38 36 ...
## $ Attrition           : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 1 1 1 ...
## $ BusinessTravel      : Factor w/ 3 levels "Non-Travel","Travel_Frequently",...: 3 2 3 2 3 2 3 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 ...
## $ 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 ...
```

From the data, we can see that there are 1,470 observations and 35 variables with various information about the employees.

Now let us have a glimpse of the data but instead of using the `glimpse()` or `summary()` functions, let's use the `skim()` function. The reason why is because it can provide more detail about the data, such as the missing rate, complete rate, and a mini histogram of each variable (Quinn & Waring, 2019).

```
#install.packages('skimr')
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 3.5.3
```

```
skim(HR_data)
```

Table 1: Data summary

Name	HR_data
Number of rows	1470
Number of columns	35
Column type frequency:	
factor	9
numeric	26
Group variables	None

Variable type: factor

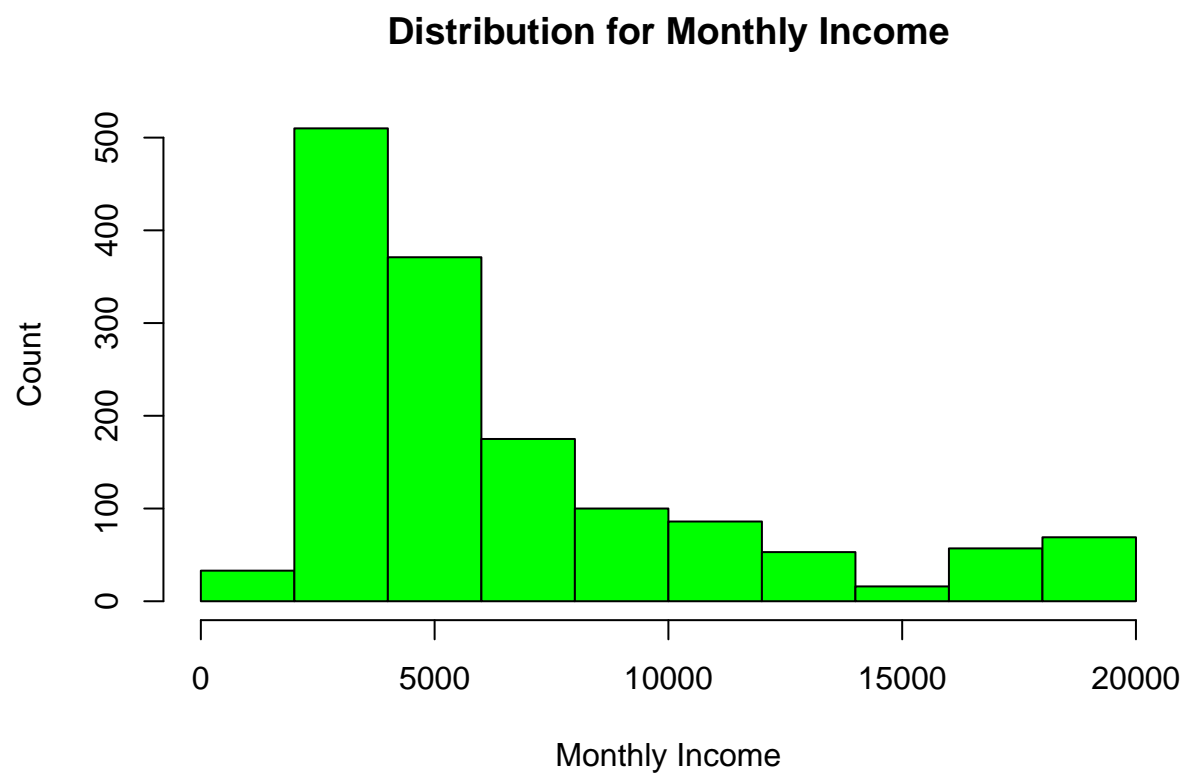
skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Attrition	0	1	FALSE	2	No: 1233, Yes: 237
BusinessTravel	0	1	FALSE	3	Tra: 1043, Tra: 277, Non: 150
Department	0	1	FALSE	3	Res: 961, Sal: 446, Hum: 63
EducationField	0	1	FALSE	6	Lif: 606, Med: 464, Mar: 159, Tec: 132
Gender	0	1	FALSE	2	Mal: 882, Fem: 588
JobRole	0	1	FALSE	9	Sal: 326, Res: 292, Lab: 259, Man: 145
MaritalStatus	0	1	FALSE	3	Mar: 673, Sin: 470, Div: 327
Over18	0	1	FALSE	1	Y: 1470
OverTime	0	1	FALSE	2	No: 1054, Yes: 416

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Age	0	1	36.92	9.14	18	30.00	36.0	42.0	59.0
DailyRate	0	1	802.49	403.51	102	465.00	802.0	1000.0	1500.0

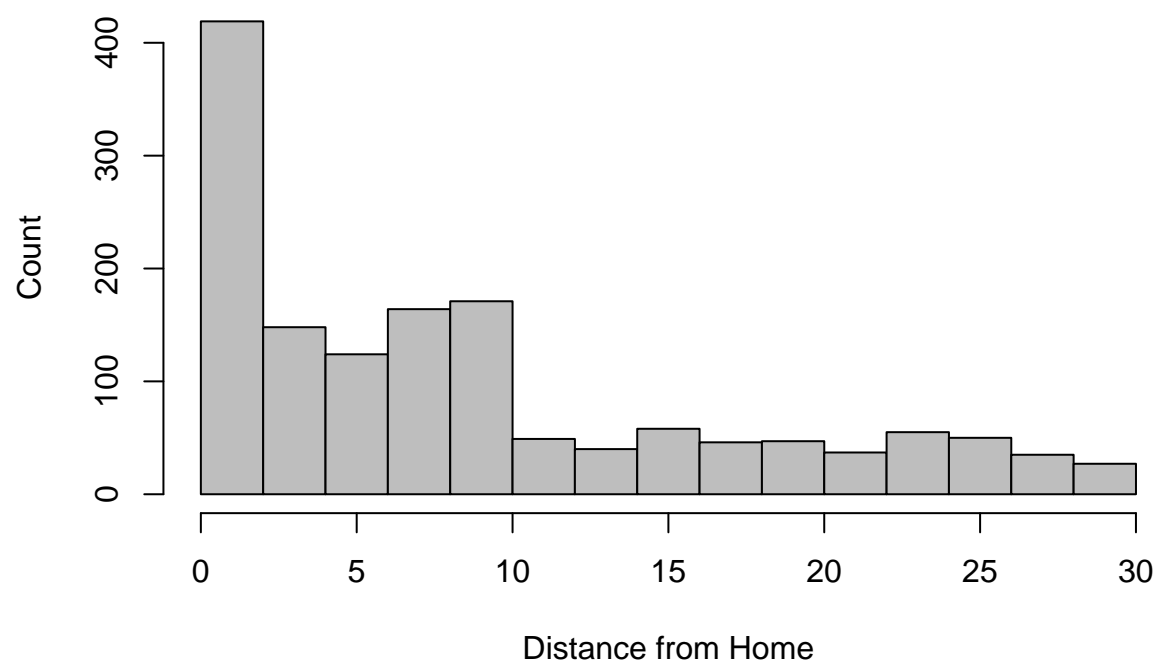
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	
DistanceFromHome	0	1	9.19	8.11	1	2.00	7.0	
Education	0	1	2.91	1.02	1	2.00	3.0	
EmployeeCount	0	1	1.00	0.00	1	1.00	1.0	
EmployeeNumber	0	1	1024.87	602.02	1	491.25	1020.5	15
EnvironmentSatisfaction	0	1	2.72	1.09	1	2.00	3.0	
HourlyRate	0	1	65.89	20.33	30	48.00	66.0	
JobInvolvement	0	1	2.73	0.71	1	2.00	3.0	
JobLevel	0	1	2.06	1.11	1	1.00	2.0	
JobSatisfaction	0	1	2.73	1.10	1	2.00	3.0	
MonthlyIncome	0	1	6502.93	4707.96	1009	2911.00	4919.0	83
MonthlyRate	0	1	14313.10	7117.79	2094	8047.00	14235.5	204
NumCompaniesWorked	0	1	2.69	2.50	0	1.00	2.0	
PercentSalaryHike	0	1	15.21	3.66	11	12.00	14.0	
PerformanceRating	0	1	3.15	0.36	3	3.00	3.0	
RelationshipSatisfaction	0	1	2.71	1.08	1	2.00	3.0	
StandardHours	0	1	80.00	0.00	80	80.00	80.0	
StockOptionLevel	0	1	0.79	0.85	0	0.00	1.0	
TotalWorkingYears	0	1	11.28	7.78	0	6.00	10.0	
TrainingTimesLastYear	0	1	2.80	1.29	0	2.00	3.0	
WorkLifeBalance	0	1	2.76	0.71	1	2.00	3.0	
YearsAtCompany	0	1	7.01	6.13	0	3.00	5.0	
YearsInCurrentRole	0	1	4.23	3.62	0	2.00	3.0	
YearsSinceLastPromotion	0	1	2.19	3.22	0	0.00	1.0	
YearsWithCurrManager	0	1	4.12	3.57	0	2.00	3.0	
From a glance of the mini histograms, it seems that several variables are tail-heavy.	stograms, it	seems that seve	ral variabl	es are tai	l-heavy	.		
So let's use the hist() function to have a better look at some of these variables.	unction to h	ave a better loo	k at some o	f these va	riables	.		

```
hist(HR_data$MonthlyIncome,main="Distribution for Monthly Income",xlab="Monthly Income",ylab="Count",col="red",border="black",breaks=10)
```

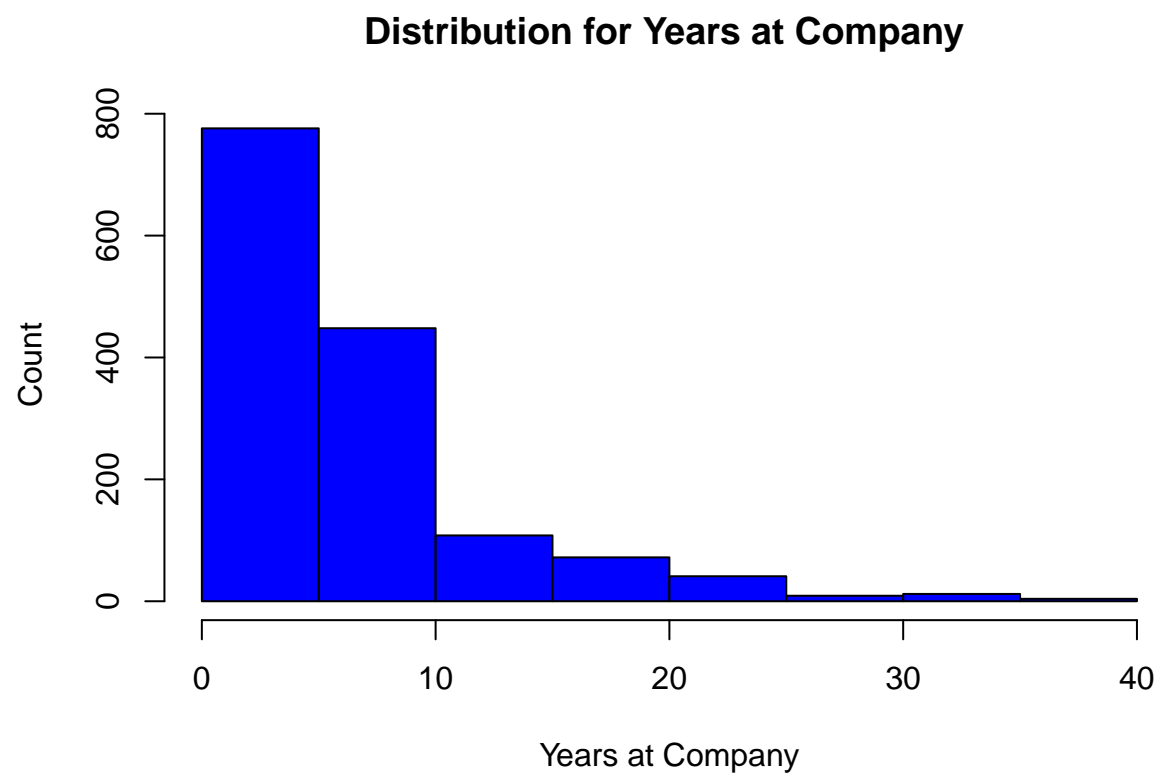


```
hist(HR_data$DistanceFromHome,main="Distribution for Distance from Home ",xlab="Distance from Home",ylab="Count")
```

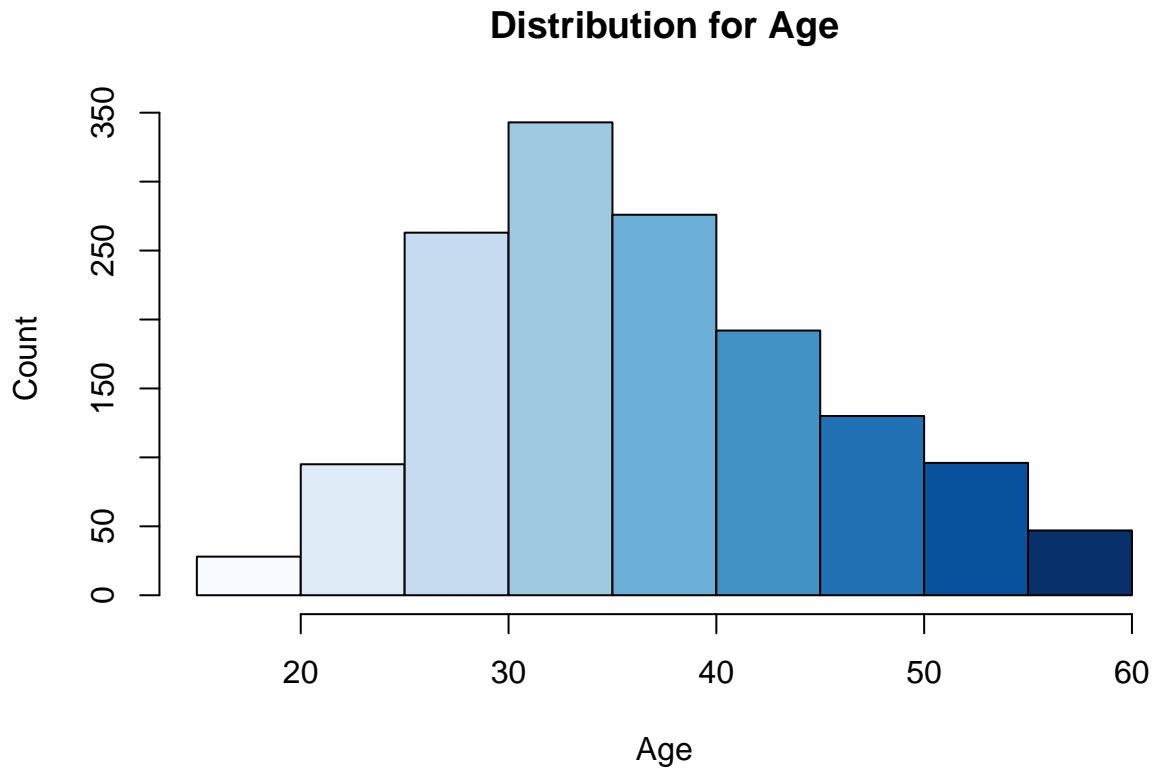
Distribution for Distance from Home



```
hist(HR_data$YearsAtCompany,main="Distribution for Years at Company ",xlab="Years at Company",ylab="Count")
```



```
hist(HR_data$Age,main="Distribution for Age",xlab="Age",ylab="Count",col=blues9)
```



After looking at the MonthlyIncome, DistanceFromHome, and YearsAtCompany they do show a right-skewed in their distributions. The distribution for Age also has normal distribution that looks slightly right-skewed with majority being in the age range of 30 to 40.

```
prop.table(table(HR_data$Gender)) #Percentage of Gender
```

```
##
## Female    Male
##    0.4    0.6
```

The table above show that 60% of the dataset gender is male.

Within our exploratory analysis, the Attrition column will be used as our target variable. Before continuing out analysis of the data, we should find out the distribution and percentage of the Attrition variable.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##    between, first, last
```



```
## The following objects are masked from 'package:stats':  
##  
##   filter, lag
```

```
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

```
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 3.5.2
```

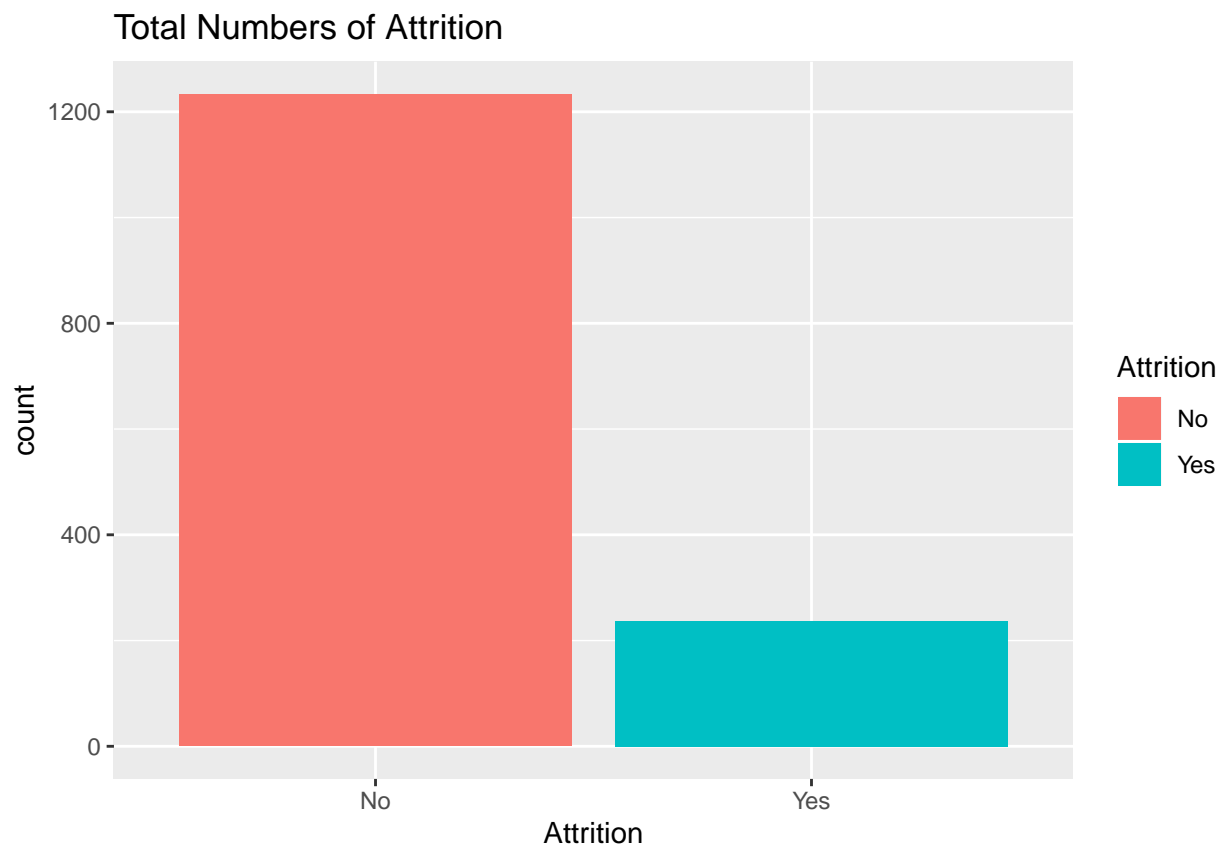
```
HR_data %>% group_by(Attrition) %>% summarise(Total = n()) %>% print()
```

```
## # A tibble: 2 x 2  
##   Attrition Total  
##   <fct>      <int>  
## 1 No         1233  
## 2 Yes         237
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
ggplot(HR_data,aes(Attrition,fill=Attrition))+geom_bar() + ggtitle("Total Numbers of Attrition")
```



```
prop.table(table(HR_data$Attrition)) #Percentage of Attrition
```

```
##  
##           No           Yes  
## 0.8387755 0.1612245
```

From the table above, we see approximately 16% of IBM employees are leaving

Now that we have set the Attrition as our target variable, we can see how it affects the other variables in the dataset. In order to reduce time producing single graphs for these variables, we are going to use the `grid()` and `gridExtra()` functions to help arrange multiple grid-based plots on a page (Phiri, 2013).

```
library(grid)  
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3
```

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

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

```
age_graph <- ggplot(HR_data,aes(Age,fill=Attrition))+geom_density()+facet_grid(~Attrition)  
gender_graph <- ggplot(HR_data,aes(Gender,fill=Attrition))+geom_bar()  
marital_graph <- ggplot(HR_data,aes(MaritalStatus,fill=Attrition))+geom_bar()  
business_graph <- ggplot(HR_data,aes(BusinessTravel,fill=Attrition))+geom_bar()  
grid.arrange(age_graph,gender_graph,marital_graph,business_graph,ncol=2, bottom = "Figure 1")
```

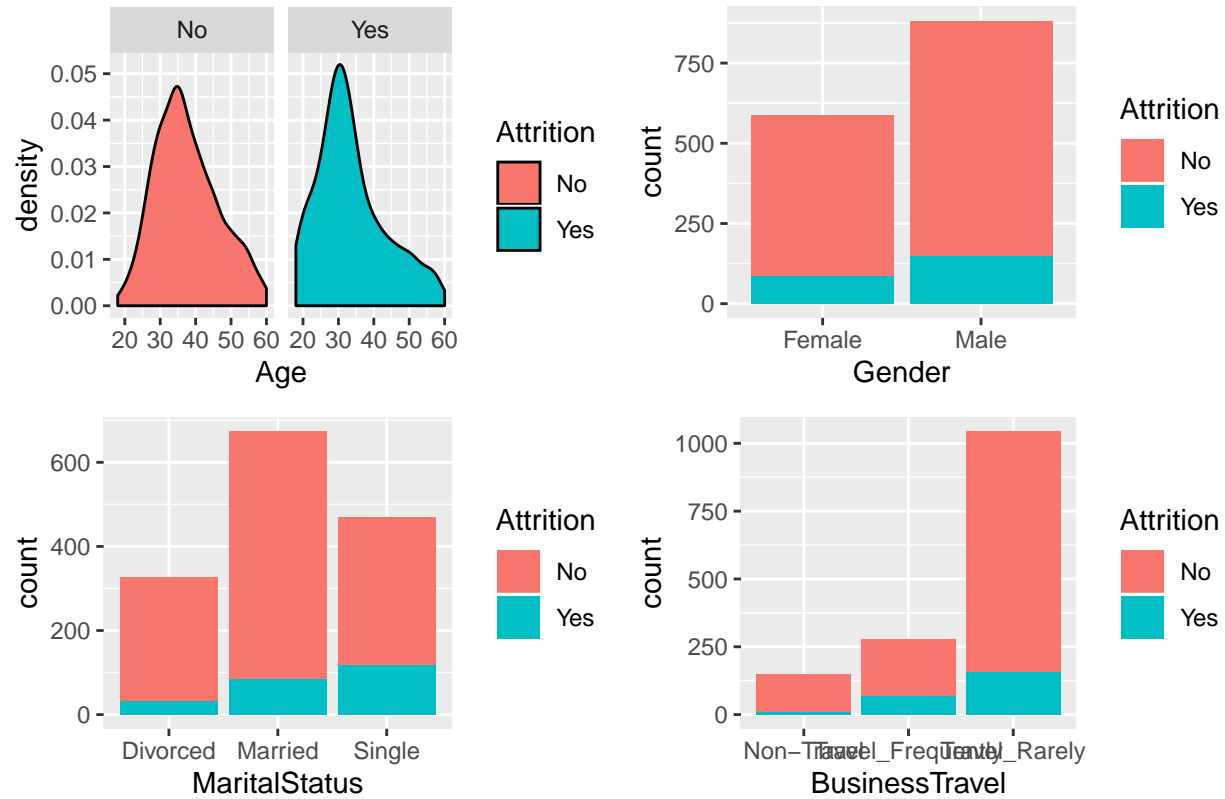


Figure 1

In Figure 1, we see the following: 1. Age: Most employees that leave IBM are around 30 years old. 2. Gender: We see that majority of separated employees are Male and that is due to our dataset being comprised of 60% Male. 3. Marital Status: Employees that are Single show the highest signs of Attrition, while Divorced employees are the lowest. 4. Business Travel: Among employee who leave IBM, most travel.

```
YAC_graph <- ggplot(HR_data,aes(YearsAtCompany,fill = Attrition))+geom_bar()
YSP_graph <- ggplot(HR_data,aes(YearsSinceLastPromotion,fill = Attrition))+geom_bar()
YCM_graph <- ggplot(HR_data,aes(YearsWithCurrManager,fill = Attrition))+geom_bar()
MTHincome_graph <- ggplot(HR_data,aes(MonthlyIncome,fill=Attrition))+geom_density()
OT_graph<- ggplot(HR_data,aes(OverTime,fill=Attrition))+geom_bar()
grid.arrange(YAC_graph,YSP_graph,YCM_graph,MTHincome_graph,OT_graph,ncol=2, bottom = "Figure 2")
```



Figure 2

In Figure 2, we see the following: 5. Years at Company: Employees who have been with IBM for <3 years make up a larger proportion of those quitting the company. 6. Years Since Last Promotion: Employees that have been recently promoted are making up a larger proportion of those who quit IBM. 7. Years With Current Manager: Newly hired managers are also a reason for employees to quit. 8. Monthly Income: We see higher levels of attrition among the lower segment of monthly income. 9. Over Time: Employees who work overtime also have a larger proportion that are quitting.

Preprocessing the Data

Before we start modelling the data, we should check if there are any missing values in the data which interfere with the predictive model.

```
sum(is.na(HR_data))
```

```
## [1] 0
```

We see that there are no missing values in the data after checking it, but we also would like to perform some data transformation. That is convert the type of some columns into a proper format.

```
HR_data$Education <- as.factor(HR_data$Education)
HR_data$EnvironmentSatisfaction <- as.factor(HR_data$EnvironmentSatisfaction)
HR_data$JobInvolvement <- as.factor(HR_data$JobInvolvement)
HR_data$JobLevel <- as.factor(HR_data$JobLevel)
HR_data$JobSatisfaction <- as.factor(HR_data$JobSatisfaction)
HR_data$StockOptionLevel <- as.factor(HR_data$StockOptionLevel)
```

```
HR_data$PerformanceRating <- as.factor(HR_data$PerformanceRating)
HR_data$RelationshipSatisfaction <- as.factor(HR_data$RelationshipSatisfaction)
HR_data$WorkLifeBalance <- as.factor(HR_data$WorkLifeBalance)
```

Due to some columns only having a single value in their columns, we will remove them.

```
HR_data <- HR_data %>% select(-EmployeeCount, -StandardHours, -Over18)
```

Feature Engineering

Feature engineering can be defined as the science of extracting more information from existing data. This newly extracted information can be used as input to our prediction model (Bock, 2017). Thereby, creating the outcome to have more impact than the model. Now based on my assumptions, we can create two features with existing variables.

1. Tenure per job: People who worked at several companies but only for a short period time usually leave the company early maybe for a change of pace or building up enough experience through these companies to help them land a job at a major company.
2. Years without Change: People who went through role or job level changes probably enjoy the thought of taking on more responsible task as they gain seniority within a company. This variable will see the years a employee went without kind of change using the Role, Job Change and Promotion, as the metrics to determine change.

```
HR_data_feng <- HR_data
```

```
HR_data_feng$TenurePerJob <- ifelse(HR_data_feng$NumCompaniesWorked!=0, HR_data_feng$TotalWorkingYears/HR_data_feng$NumCompaniesWorked, 0)
HR_data_feng$YearWithoutChange <- HR_data_feng$YearsInCurrentRole - HR_data_feng$YearsSinceLastPromotion
HR_data_feng$YearsWithoutChange2 <- HR_data_feng$TotalWorkingYears - HR_data_feng$YearsSinceLastPromotion
```

```
tenurePlot <- ggplot(HR_data_feng, aes(TenurePerJob)) + geom_density() + facet_grid(~Attrition)
changePlot <- ggplot(HR_data_feng, aes(YearWithoutChange)) + geom_density() + facet_grid(~Attrition)
change2Plot <- ggplot(HR_data_feng, aes(YearsWithoutChange2)) + geom_density() + facet_grid(~Attrition)
grid.arrange(tenurePlot, changePlot, change2Plot, ncol=2, bottom = "Figure 3")
```

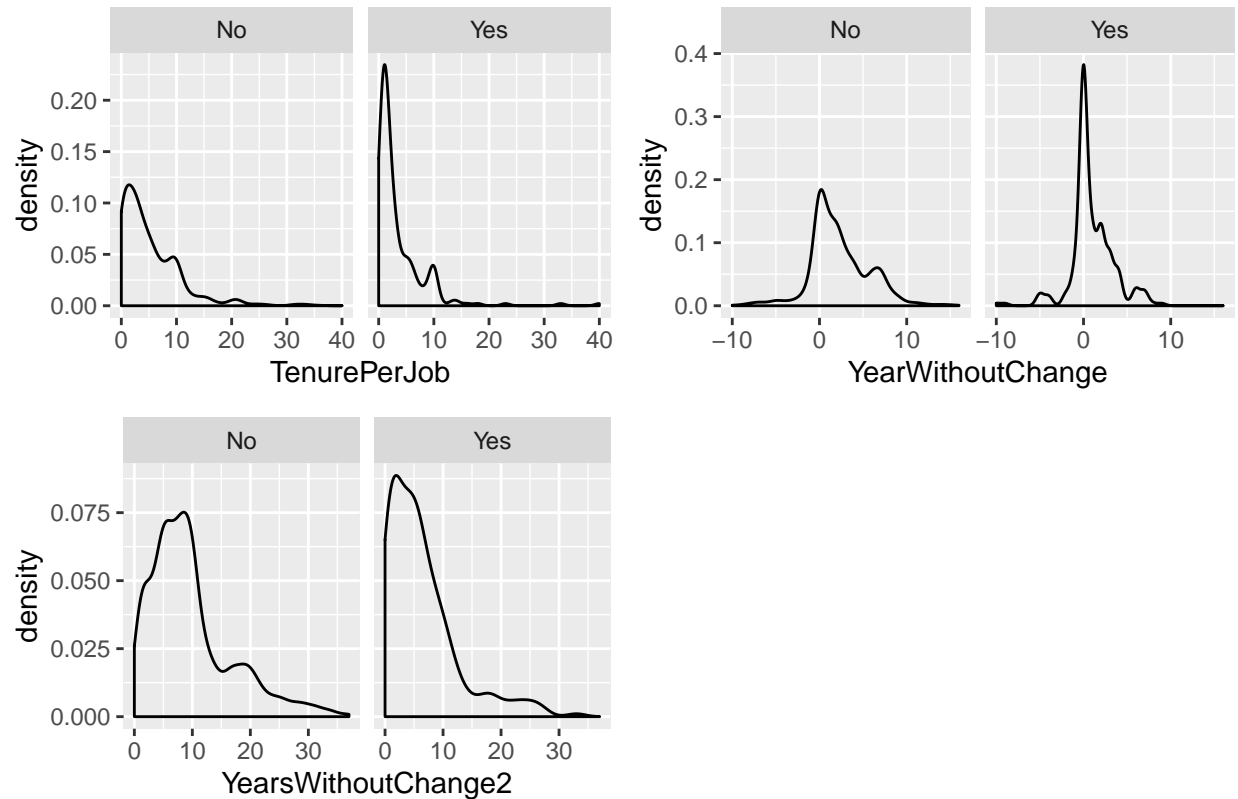


Figure 3

In figure 3, we see that the Attrition variable does have an affect on these new features.

Logistic Regression

Now we split the data into a 20% testing set and 80% training set with the `sample()` function which takes a vector as input; then you tell it how many samples to draw from that list ('R Function', 2019). We also use the `set.seed()` function which produces the same sample again and again. The purpose of creating two sets of data is so the training set is the one on which we train and fit our model basically to fit the parameters whereas test data is used only to assess performance of model (Shah, 2017).

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.3
```

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

```
## Warning: package 'lattice' was built under R version 3.5.3
```

```
# Splitting the data
set.seed(18)
attr_training <- sample(nrow(HR_data), nrow(HR_data)*.8)
train_attr <- HR_data %>% slice(attr_training)
test_attr <- HR_data %>% slice(-attr_training)
```

```
# Check the portion and percentage of Attrition in train data
table(train_attr$Attrition)
```

```
##
## No Yes
## 996 180
```

```
prop.table(table(train_attr$Attrition))
```

```
##
## No Yes
## 0.8469388 0.1530612
```

From the information displayed, we see that the data is imbalanced with Yes cases at 15%. However, we could try to fix this imbalance sample by using up-sampling or down-sampling techniques. Keep in mind, there are pros and cons when using those techniques (Altini, 2015). With that in mind, we will try to make it balanced by using an upsampling technique with `ovun.sample()` function from ROSE package (Analytics, 2019).

```
#install.packages("ROSE")
library(ROSE)
```

```
## Warning: package 'ROSE' was built under R version 3.5.3
```

```
## Loaded ROSE 0.0-3
```

```
library(brglm2)
```

```
## Warning: package 'brglm2' was built under R version 3.5.3
```

```
balanced_attr <- ovun.sample(Attrition ~ ., data = train_attr, method = "over",
                             N = 996*2, seed = 1)$data
table(balanced_attr$Attrition)
```

```
##
## No Yes
## 996 996
```

After balancing the data, we will perform our first Logistic Regression model by using all the predictors in the formula.

```
log_regress <- glm(Attrition ~ ., family = "binomial", data = balanced_attr)
summary(log_regress)
```

```
##
## Call:
## glm(formula = Attrition ~ ., family = "binomial", data = balanced_attr)
##
## Deviance Residuals:
```

```

##      Min      1Q   Median      3Q      Max
## -2.8672  -0.5173  0.0246   0.5847   3.4560
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -9.801e+00  3.256e+02  -0.030  0.975983
## Age            -3.858e-02  1.031e-02  -3.741  0.000183
## BusinessTravelTravel_Frequently  2.260e+00  3.203e-01   7.054  1.74e-12
## BusinessTravelTravel_Rarely    1.224e+00  2.896e-01   4.226  2.38e-05
## DailyRate      -4.255e-04  1.751e-04  -2.429  0.015120
## DepartmentResearch & Development  1.521e+01  3.256e+02   0.047  0.962735
## DepartmentSales    1.505e+01  3.256e+02   0.046  0.963137
## DistanceFromHome   4.566e-02  8.898e-03   5.132  2.87e-07
## Education2        -1.167e-02  2.495e-01  -0.047  0.962693
## Education3        -9.487e-02  2.207e-01  -0.430  0.667304
## Education4         2.541e-01  2.422e-01   1.049  0.294043
## Education5         2.890e-01  4.320e-01   0.669  0.503559
## EducationFieldLife Sciences  -1.471e+00  6.599e-01  -2.229  0.025842
## EducationFieldMarketing  -6.816e-01  6.877e-01  -0.991  0.321555
## EducationFieldMedical  -1.346e+00  6.509e-01  -2.068  0.038676
## EducationFieldOther    -8.496e-01  6.986e-01  -1.216  0.223902
## EducationFieldTechnical Degree  1.691e-01  6.741e-01   0.251  0.801992
## EmployeeNumber    -2.245e-04  1.224e-04  -1.834  0.066652
## EnvironmentSatisfaction2  -1.244e+00  2.258e-01  -5.509  3.60e-08
## EnvironmentSatisfaction3  -1.090e+00  2.059e-01  -5.296  1.18e-07
## EnvironmentSatisfaction4  -1.111e+00  2.029e-01  -5.473  4.41e-08
## GenderMale         4.716e-01  1.448e-01   3.258  0.001124
## HourlyRate       -1.655e-03  3.419e-03  -0.484  0.628366
## JobInvolvement2    -1.240e+00  3.224e-01  -3.846  0.000120
## JobInvolvement3    -1.580e+00  3.078e-01  -5.133  2.85e-07
## JobInvolvement4    -1.678e+00  3.808e-01  -4.408  1.05e-05
## JobLevel2         -1.025e+00  2.986e-01  -3.433  0.000596
## JobLevel3          1.023e+00  5.256e-01   1.946  0.051666
## JobLevel4        -2.091e+00  9.785e-01  -2.137  0.032572
## JobLevel5          3.386e+00  1.189e+00   2.848  0.004400
## JobRoleHuman Resources   1.642e+01  3.256e+02   0.050  0.959764
## JobRoleLaboratory Technician  2.041e+00  4.764e-01   4.284  1.83e-05
## JobRoleManager     -3.228e-01  8.494e-01  -0.380  0.703927
## JobRoleManufacturing Director  1.068e+00  4.715e-01   2.264  0.023555
## JobRoleResearch Director  -2.529e+00  9.117e-01  -2.774  0.005531
## JobRoleResearch Scientist   1.107e+00  4.789e-01   2.312  0.020776
## JobRoleSales Executive    2.459e+00  1.007e+00   2.442  0.014620
## JobRoleSales Representative  3.035e+00  1.061e+00   2.862  0.004213
## JobSatisfaction2    -1.168e+00  2.168e-01  -5.387  7.17e-08
## JobSatisfaction3    -8.921e-01  1.891e-01  -4.718  2.38e-06
## JobSatisfaction4    -1.750e+00  2.043e-01  -8.564  < 2e-16
## MaritalStatusMarried    5.933e-01  2.022e-01   2.935  0.003338
## MaritalStatusSingle    6.732e-01  2.930e-01   2.297  0.021591
## MonthlyIncome      -1.466e-04  7.005e-05  -2.093  0.036321
## MonthlyRate         9.625e-06  9.522e-06   1.011  0.312118
## NumCompaniesWorked    2.380e-01  3.096e-02   7.687  1.50e-14
## OverTimeYes         2.020e+00  1.534e-01  13.166  < 2e-16
## PercentSalaryHike    -6.368e-02  3.011e-02  -2.115  0.034450
## PerformanceRating4     2.976e-01  3.191e-01   0.933  0.351078

```



```

## RelationshipSatisfaction2      -1.149e+00  2.183e-01  -5.264  1.41e-07
## RelationshipSatisfaction3      -1.045e+00  1.955e-01  -5.347  8.93e-08
## RelationshipSatisfaction4      -1.094e+00  1.970e-01  -5.552  2.82e-08
## StockOptionLevel1             -1.274e+00  2.405e-01  -5.299  1.16e-07
## StockOptionLevel2             -1.131e+00  3.153e-01  -3.586  0.000336
## StockOptionLevel3             -3.910e-01  3.662e-01  -1.068  0.285618
## TotalWorkingYears             -3.083e-02  2.078e-02  -1.483  0.138032
## TrainingTimesLastYear         -1.209e-01  5.217e-02  -2.318  0.020468
## WorkLifeBalance2              -9.685e-01  2.946e-01  -3.288  0.001008
## WorkLifeBalance3              -1.549e+00  2.815e-01  -5.503  3.73e-08
## WorkLifeBalance4              -6.003e-01  3.325e-01  -1.805  0.071037
## YearsAtCompany                 2.011e-01  3.016e-02   6.670  2.56e-11
## YearsInCurrentRole            -2.519e-01  4.021e-02  -6.266  3.70e-10
## YearsSinceLastPromotion        1.024e-01  3.173e-02   3.225  0.001258
## YearsWithCurrManager          -1.281e-01  3.721e-02  -3.444  0.000573
##
## (Intercept)
## Age                            ***
## BusinessTravelTravel_Frequently ***
## BusinessTravelTravel_Rarely    ***
## DailyRate                      *
## DepartmentResearch & Development
## DepartmentSales
## DistanceFromHome              ***
## Education2
## Education3
## Education4
## Education5
## EducationFieldLife Sciences    *
## EducationFieldMarketing
## EducationFieldMedical          *
## EducationFieldOther
## EducationFieldTechnical Degree
## EmployeeNumber                 .
## EnvironmentSatisfaction2        ***
## EnvironmentSatisfaction3        ***
## EnvironmentSatisfaction4        ***
## GenderMale                     **
## HourlyRate
## JobInvolvement2                 ***
## JobInvolvement3                 ***
## JobInvolvement4                 ***
## JobLevel2                      ***
## JobLevel3                      .
## JobLevel4                      *
## JobLevel5                      **
## JobRoleHuman Resources
## JobRoleLaboratory Technician   ***
## JobRoleManager
## JobRoleManufacturing Director  *
## JobRoleResearch Director       **
## JobRoleResearch Scientist      *
## JobRoleSales Executive          *
## JobRoleSales Representative     **

```

```

## JobSatisfaction2          ***
## JobSatisfaction3          ***
## JobSatisfaction4          ***
## MaritalStatusMarried     **
## MaritalStatusSingle      *
## MonthlyIncome            *
## MonthlyRate
## NumCompaniesWorked       ***
## OverTimeYes              ***
## PercentSalaryHike        *
## PerformanceRating4
## RelationshipSatisfaction2  ***
## RelationshipSatisfaction3  ***
## RelationshipSatisfaction4  ***
## StockOptionLevel1        ***
## StockOptionLevel2        ***
## StockOptionLevel3
## TotalWorkingYears
## TrainingTimesLastYear    *
## WorkLifeBalance2         **
## WorkLifeBalance3         ***
## WorkLifeBalance4         .
## 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: 2761.5  on 1991  degrees of freedom
## Residual deviance: 1511.3  on 1928  degrees of freedom
## AIC: 1639.3
##
## Number of Fisher Scoring iterations: 14

```

```

glm(formula = Attrition ~ ., family = "binomial", data = train_attr,
     method = "detect_separation", linear_program = "dual")

```

```

## Warning: 'detect_separation' will be removed from 'brglm2' at version
## 0.8. A new version of 'detect_separation' is now maintained in the
## 'detectseparation' package.

```

```

## Separation: FALSE
## Existence of maximum likelihood estimates
##              (Intercept)                Age
##              -Inf                0
## BusinessTravelTravel_Frequently BusinessTravelTravel_Rarely
##              0                0
##              DailyRate DepartmentResearch & Development
##              0                Inf
##              DepartmentSales DistanceFromHome

```

##	Inf	0
##	Education2	Education3
##	0	0
##	Education4	Education5
##	0	0
##	EducationFieldLife Sciences	EducationFieldMarketing
##	0	0
##	EducationFieldMedical	EducationFieldOther
##	0	0
##	EducationFieldTechnical Degree	EmployeeNumber
##	0	0
##	EnvironmentSatisfaction2	EnvironmentSatisfaction3
##	0	0
##	EnvironmentSatisfaction4	GenderMale
##	0	0
##	HourlyRate	JobInvolvement2
##	0	0
##	JobInvolvement3	JobInvolvement4
##	0	0
##	JobLevel2	JobLevel3
##	0	0
##	JobLevel4	JobLevel5
##	0	0
##	JobRoleHuman Resources	JobRoleLaboratory Technician
##	Inf	0
##	JobRoleManager	JobRoleManufacturing Director
##	0	0
##	JobRoleResearch Director	JobRoleResearch Scientist
##	0	0
##	JobRoleSales Executive	JobRoleSales Representative
##	0	0
##	JobSatisfaction2	JobSatisfaction3
##	0	0
##	JobSatisfaction4	MaritalStatusMarried
##	0	0
##	MaritalStatusSingle	MonthlyIncome
##	0	0
##	MonthlyRate	NumCompaniesWorked
##	0	0
##	OverTimeYes	PercentSalaryHike
##	0	0
##	PerformanceRating4	RelationshipSatisfaction2
##	0	0
##	RelationshipSatisfaction3	RelationshipSatisfaction4
##	0	0
##	StockOptionLevel1	StockOptionLevel2
##	0	0
##	StockOptionLevel3	TotalWorkingYears
##	0	0
##	TrainingTimesLastYear	WorkLifeBalance2
##	0	0
##	WorkLifeBalance3	WorkLifeBalance4
##	0	0
##	YearsAtCompany	YearsInCurrentRole

```

##                                0                                0
##      YearsSinceLastPromotion      YearsWithCurrManager
##                                0                                0
## 0: finite value, Inf: infinity, -Inf: -infinity

```

The separation in the our model returned False so there extist no perfect separation