DDS Case Study 2 Analysis

Akbar Thobani

04/04/2020

#Problem Introduction DDS Analytics has tasked us with to uncover insights and trends specific to Job Roles within the company and how those factors contribute to turnover/attrition.

#Analysis Objectives 1) Uncover any intersting trends specific to Job Role 2) Report the top 3 factors that contribute to turnover 3) Build a model to predict attrition 4) Build a model to predict salary

#Packages

```
library(ggplot2)
library(corrplot)
library(dplyr)
library(caret)
library(MASS)
library(randomForest)
library(e1071)
library(tidyverse)
```

#Data Import

```
dfTrain <- read.csv(file="CaseStudy2-data.csv", header=TRUE, stringsAsFactors=TRUE)
dfVal <- read.csv(file="CaseStudy2validation.csv", header=TRUE, stringsAsFactors=TRUE)

dfCompAtt <- read.csv(file="CaseStudy2CompSet No Attrition.csv", header=TRUE, stringsAsFactors=TRUE)
dfCompSal <- read.csv(file="CaseStudy2CompSet No Salary.csv", header=TRUE, stringsAsFactors=TRUE)
```

#Data Check We want to make sure the dataset provided does not have any missing values or mixed data types before we begin our explaoratory and modeling exercises.

```
str(dfVal)
```

```
## 'data.frame':    300 obs. of  37 variables:
##  $ ID                      : int  1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 ...
##  $ Age                     : int  43 35 55 48 37 44 36 27 39 20 ...
##  $ Attrition               : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 2 ...
##  $ BusinessTravel          : Factor w/ 3 levels "Non-Travel","Travel_Frequently",..: 3 3 2 3 2 2 3 3
##  $ DailyRate               : int  1001 619 1091 530 1231 383 676 269 945 1362 ...
##  $ Department              : Factor w/ 3 levels "Human Resources",..: 2 3 2 3 3 3 2 2 2 2 ...
##  $ DistanceFromHome        : int  7 1 2 29 21 1 1 5 22 10 ...
##  $ Education               : int  3 3 1 1 2 5 3 1 3 1 ...
##  $ EducationField          : Factor w/ 6 levels "Human Resources",..: 2 3 2 4 4 3 5 6 4 4 ...
##  $ EmployeeCount           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ EmployeeNumber          : int  451 600 1096 473 900 1481 823 844 1043 701 ...
##  $ EnvironmentSatisfaction : int  3 2 4 1 3 1 3 3 4 4 ...
##  $ Gender                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 1 1 2 1 2 ...
##  $ HourlyRate              : int  43 85 65 91 54 79 35 42 82 32 ...
##  $ JobInvolvement          : int  3 3 3 3 3 3 3 2 3 3 ...
##  $ JobLevel                : int  3 2 3 3 1 2 2 3 3 1 ...
##  $ JobRole                 : Factor w/ 9 levels "Healthcare Representative",..: 1 8 5 4 9 8 5 6 5 7
##  $ JobSatisfaction         : int  1 3 2 3 4 3 2 4 1 3 ...
##  $ MaritalStatus           : Factor w/ 3 levels "Divorced","Married",..: 2 2 2 2 2 2 2 1 3 3 ...
##  $ MonthlyIncome           : int  9985 4717 10976 12504 2973 4768 5228 12808 10880 1009 ...
##  $ MonthlyRate             : int  9262 18659 15813 23978 21222 9282 23361 8842 5083 26999 ...
##  $ NumCompaniesWorked      : int  8 9 3 3 5 7 0 1 1 1 ...
##  $ Over18                  : Factor w/ 1 level "Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OverTime                : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ PercentSalaryHike       : int  16 11 18 21 15 12 15 16 13 11 ...
##  $ PerformanceRating       : int  3 3 3 4 3 3 3 3 3 3 ...
##  $ RelationshipSatisfaction: int  1 3 2 2 2 3 1 2 3 4 ...
##  $ StandardHours           : int  80 80 80 80 80 80 80 80 80 80 ...
##  $ StockOptionLevel        : int  1 0 1 1 1 1 1 1 0 0 ...
##  $ TotalWorkingYears       : int  10 15 23 15 10 11 10 9 21 1 ...
##  $ TrainingTimesLastYear   : int  1 2 4 3 3 4 2 3 2 5 ...
##  $ WorkLifeBalance         : int  2 3 3 1 3 2 3 3 3 3 ...
##  $ YearsAtCompany          : int  1 11 3 0 5 1 9 9 21 1 ...
##  $ YearsInCurrentRole      : int  0 9 2 0 4 0 7 8 6 0 ...
##  $ YearsSinceLastPromotion : int  0 6 1 0 0 0 0 0 2 1 ...
##  $ YearsWithCurrManager    : int  0 9 2 0 0 0 5 8 8 1 ...
##  $ Rand                    : num  -0.0245 -0.3341 0.0462 1.831 1.2296 ...
```

```
#No missing values
#sum(is.na(dfVal))
#colSums(is.na(dfTrain))
#colSums(is.na(dfVal))

#View(summary(df))
```

#Data Preparation There are a few variables that seem useless for the purposes of this analysis. ID, Standard Hours, Employee Number and Employee Count will be removed from the table

```
#Recode Attrition Column to numeric if necessary
#df$Attrition2 <- ifelse(df$Attrition == "Yes", 1, 0)

#Drop ID, StandardHours, EmployeeCount, Over18 columns
#Most values do not change so SD is 0

df_stage <- dfTrain[,!(names(dfTrain) %in% c("ID", "StandardHours", "EmployeeNumber", "EmployeeCount",
df_val_stage <- dfVal[,!(names(dfVal) %in% c("ID", "StandardHours", "EmployeeNumber", "EmployeeCount",
```

An additional dataframe was created with only numeric values to be read by a correlation heatmap later in the analysis.

```
#Return numeric values only
df_numeric <- df_stage[, sapply(df_stage, is.numeric)]
df_val_numeric <- df_val_stage[, sapply(df_val_stage, is.numeric)]
```

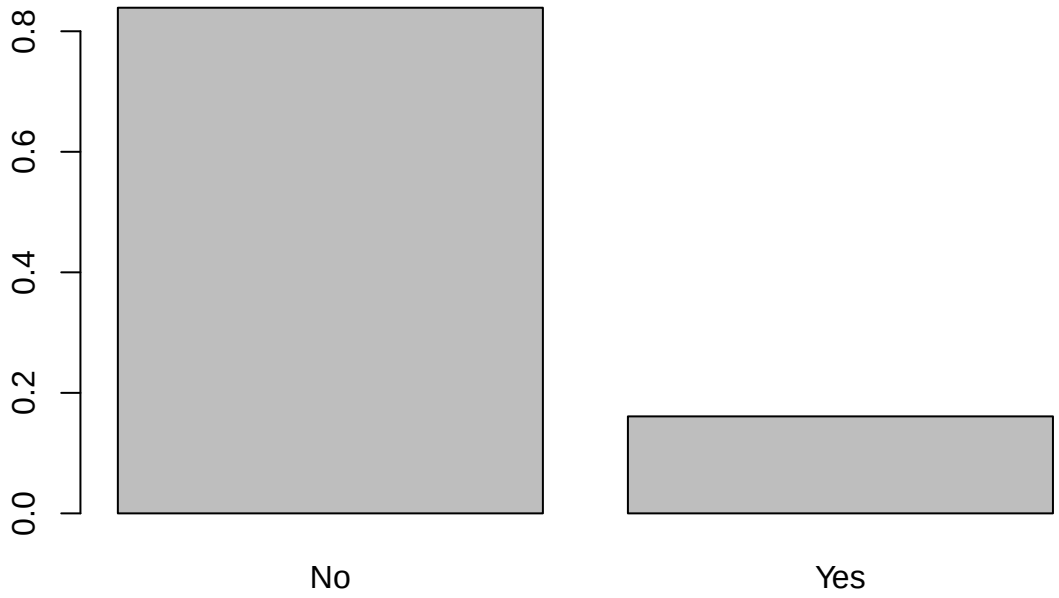
```r
#Correlation Plot
df_corr <- round(cor(df_numeric),2)
```

#Data Exploration More than 80% of the training dataset consist of employees who are still retained

```r
table(dfTrain$Attrition)
```

```
##
##  No Yes
## 730 140
```

```r
barplot(prop.table(table(dfTrain$Attrition)))
```



We created a correlation matrix to view possible multicollinearity between variables that need to be addressed before the modeling phase to avoid redundancy.

The variables below seem to have high collinearity so we will remove some of them for the Custom model at a later phase:

MonthlyIncome corr JobLevel PercentSalaryHike corr PerformanceRating TotalWorkingYears corr JobLevel Age corr TotalWorkingYears YearsInCurrentRole corr TotalWorkingYears

```r
corrplot(df_corr, order="FPC", title="Variable Corr Heatmap",tl.srt=45,method = "pie")
```

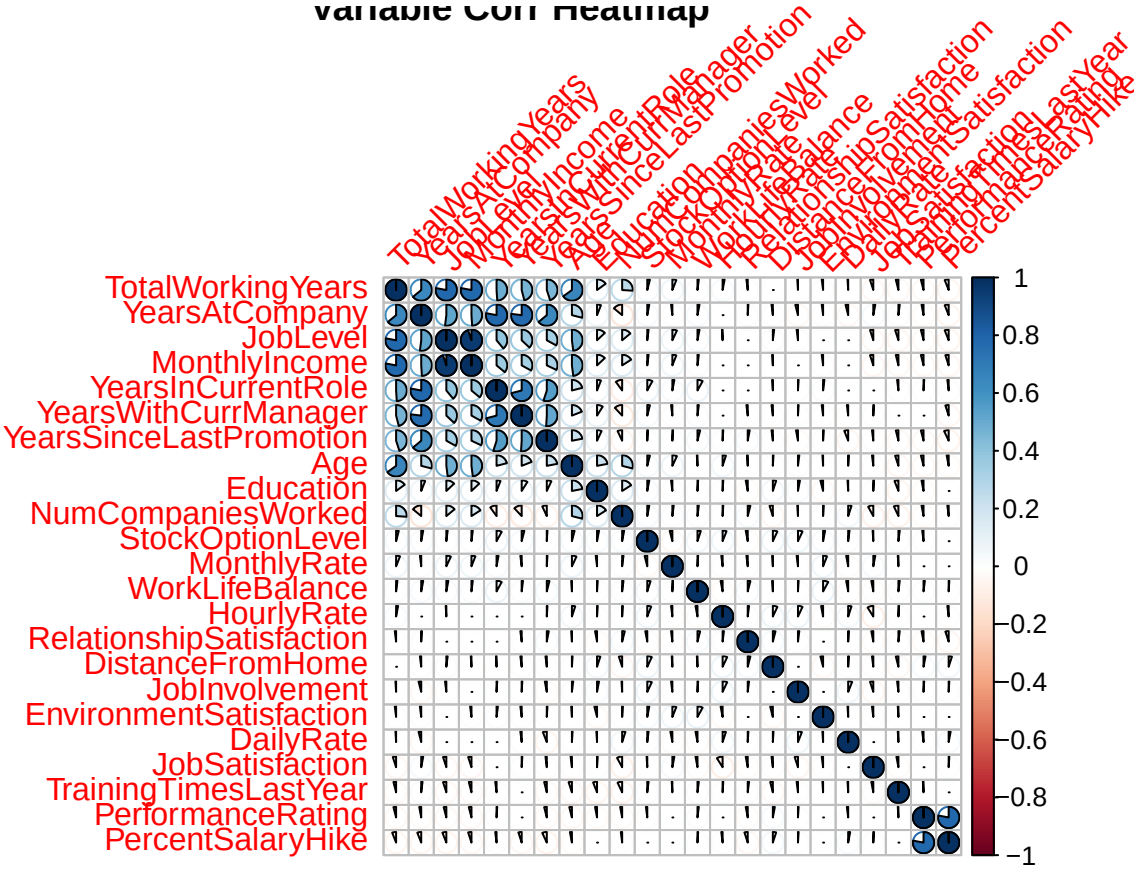
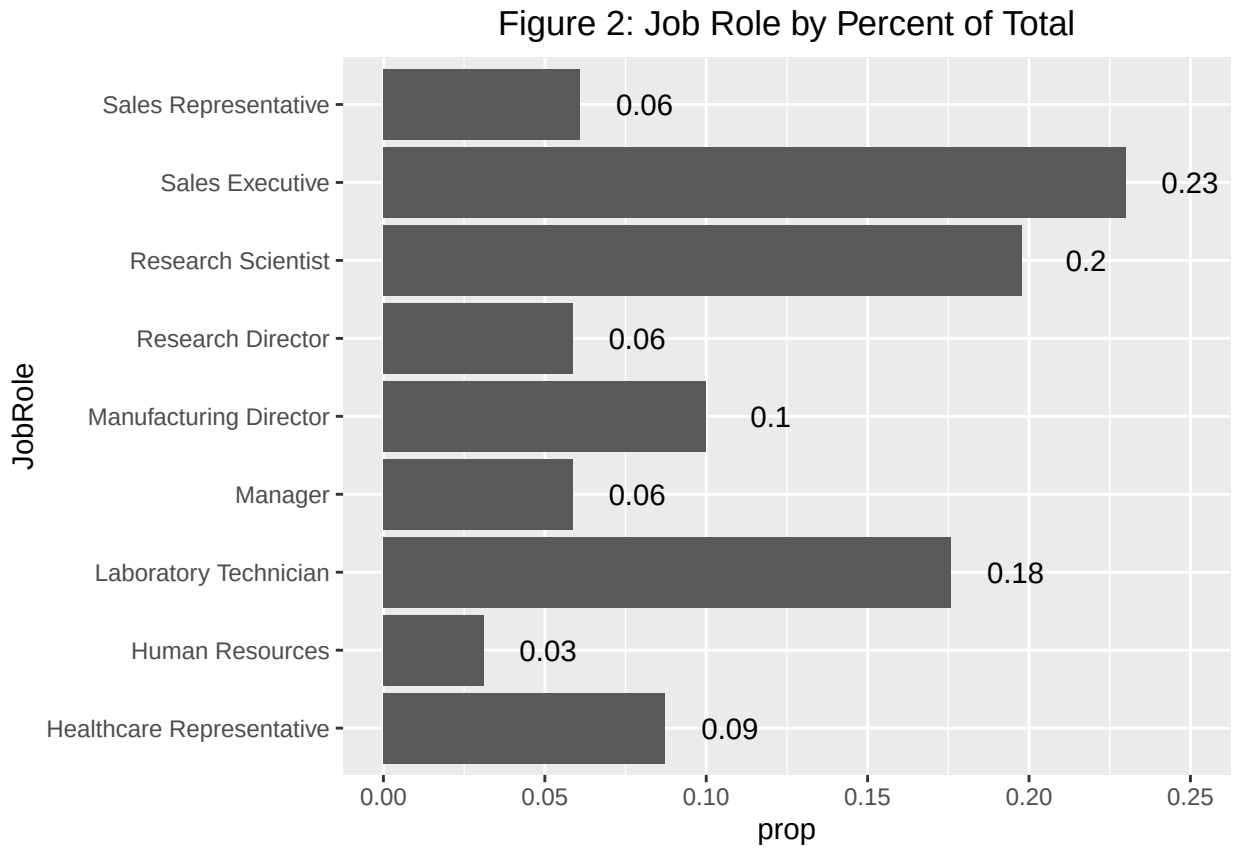


Figure 1 shows the count distribution of employees by job role. Sales Executive jobs are the most prevalent at 22% of all Job Roles followed closely by Research Scientist at 20% and Lab Technicians at 18% rounding out the top 3.

```
p1 <- ggplot(dfTrain, aes(x=JobRole), color=JobRole) + ggtitle("Figure 1: Job Role") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..), fill=JobRole),  width = 0.7) +
  labs(y="Percentage") +
  coord_flip() +
  theme_linedraw() +
  theme(plot.title = element_text(hjust = 0.7))
p1
```

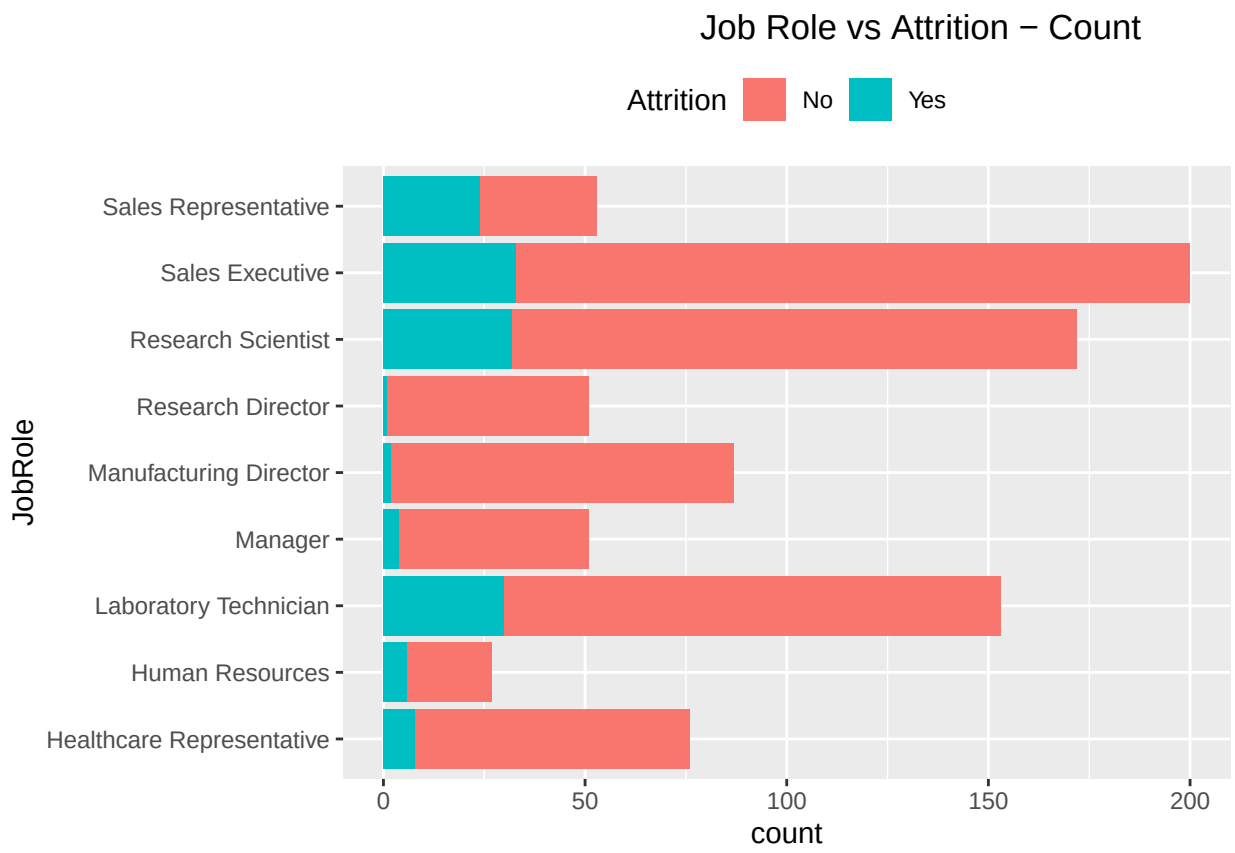
Figure 2 is another view showing the percentage of total by each Job Role

```
p2 <- ggplot(dfTrain, aes(x=JobRole, y = ..prop.., group=1)) +
  geom_bar() +
  geom_text(stat = "count",
            aes(label = round(..prop.., 2), y = ..prop.. + 0.02)) +
            coord_flip() +
  ggtitle("Figure 2: Job Role by Percent of Total") +
  theme(plot.title = element_text(hjust = 0.5))
p2
```



Although Lab Technician roles account for almost 17% of all Job Roles, the attrition total and porportion seems to abnormally high.

```
p3 <- ggplot(dfTrain,aes(x = JobRole,fill = Attrition)) +
    geom_bar(position = position_stack(reverse = FALSE)) +
    ggtitle("Job Role vs Attrition - Count") +
    coord_flip() +
    theme(legend.position = "top") +
    theme(plot.title = element_text(hjust = 0.8))

p3
```



Naive Bayes Model

```
Naive_Bayes_Model=naiveBayes( Attrition~., data=dfTrain)

#Summary of model
Naive_Bayes_Model
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##        No       Yes
## 0.8390805 0.1609195
##
## Conditional probabilities:
##      ID
## Y         [,1]     [,2]
##   No  430.3014 251.3245
##   Yes 462.6071 250.2665
##
##      Age
## Y         [,1]     [,2]
```

```
##   No  37.41233 8.673382
##   Yes 33.78571 9.614726
##
##      BusinessTravel
## Y     Non-Travel Travel_Frequently Travel_Rarely
##   No  0.11369863        0.16849315    0.71780822
##   Yes 0.07857143        0.25000000    0.67142857
##
##      DailyRate
## Y         [,1]     [,2]
##   No  821.1603 401.4137
##   Yes 784.2929 399.5637
##
##      Department
## Y     Human Resources Research & Development      Sales
##   No       0.03972603             0.66712329 0.29315068
##   Yes      0.04285714             0.53571429 0.42142857
##
##      DistanceFromHome
## Y          [,1]     [,2]
##   No   9.028767 7.982869
##   Yes 10.957143 8.748995
##
##      Education
## Y         [,1]     [,2]
##   No  2.923288 1.024865
##   Yes 2.785714 1.009207
##
##      EducationField
## Y     Human Resources Life Sciences  Marketing    Medical      Other
##   No       0.01506849    0.41780822 0.10958904 0.31917808 0.05890411
##   Yes      0.02857143    0.37857143 0.14285714 0.26428571 0.06428571
##      EducationField
## Y     Technical Degree
##   No        0.07945205
##   Yes       0.12142857
##
##      EmployeeCount
## Y     [,1] [,2]
##   No     1    0
##   Yes    1    0
##
##      EmployeeNumber
## Y          [,1]     [,2]
##   No  1035.8658 606.5168
##   Yes  998.3714 596.8576
##
##      EnvironmentSatisfaction
## Y         [,1]     [,2]
##   No  2.738356 1.077915
##   Yes 2.507143 1.190468
##
##      Gender
## Y        Female      Male
```

```
##   No  0.4123288 0.5876712
##   Yes 0.3785714 0.6214286
##
##      HourlyRate
## Y         [,1]     [,2]
##   No  65.29178 20.20311
##   Yes 67.29286 19.71214
##
##      JobInvolvement
## Y         [,1]      [,2]
##   No  2.780822 0.6655665
##   Yes 2.421429 0.8141541
##
##      JobLevel
## Y         [,1]      [,2]
##   No  2.116438 1.0943819
##   Yes 1.635714 0.9760493
##
##      JobRole
## Y     Healthcare Representative Human Resources Laboratory Technician
##   No                0.093150685     0.028767123           0.168493151
##   Yes               0.057142857     0.042857143           0.214285714
##      JobRole
## Y         Manager Manufacturing Director Research Director Research Scientist
##   No  0.064383562            0.116438356       0.068493151        0.191780822
##   Yes 0.028571429            0.014285714       0.007142857        0.228571429
##      JobRole
## Y     Sales Executive Sales Representative
##   No      0.228767123          0.039726027
##   Yes     0.235714286          0.171428571
##
##      JobSatisfaction
## Y         [,1]     [,2]
##   No  2.761644 1.111436
##   Yes 2.435714 1.094201
##
##      MaritalStatus
## Y       Divorced    Married     Single
##   No  0.24520548 0.48219178 0.27260274
##   Yes 0.08571429 0.41428571 0.50000000
##
##      MonthlyIncome
## Y         [,1]     [,2]
##   No  6702.000 4675.472
##   Yes 4764.786 3786.389
##
##      MonthlyRate
## Y         [,1]     [,2]
##   No  14460.12 7126.983
##   Yes 13624.29 6993.816
##
##      NumCompaniesWorked
## Y         [,1]     [,2]
##   No  2.660274 2.465606
```

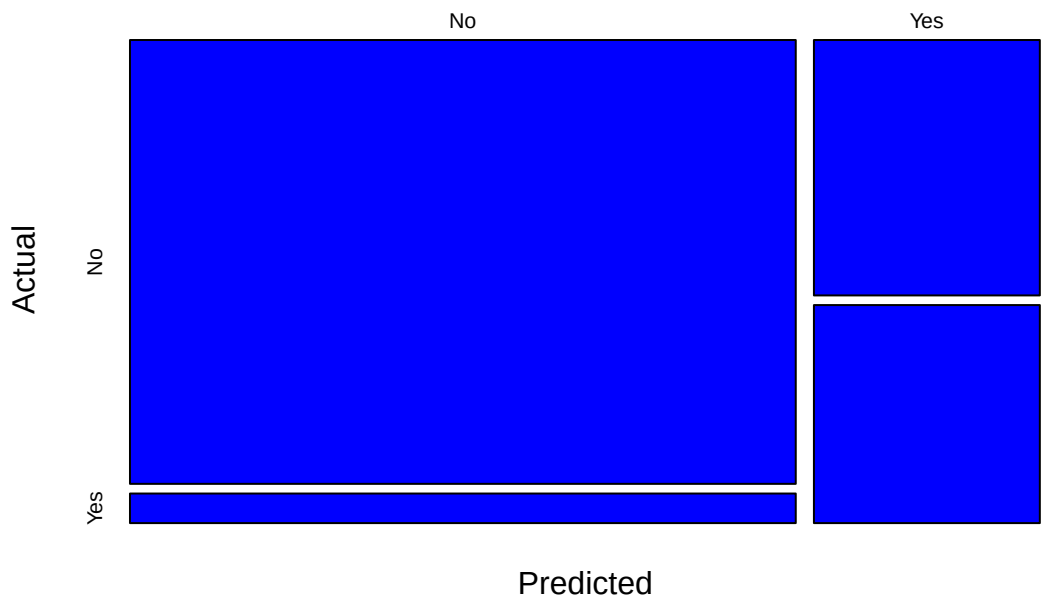
```
##   Yes 3.078571 2.772080
##
##      Over18
## Y     Y
##   No  1
##   Yes 1
##
##      OverTime
## Y           No       Yes
##   No  0.7643836 0.2356164
##   Yes 0.4285714 0.5714286
##
##      PercentSalaryHike
## Y         [,1]     [,2]
##   No  15.17534 3.627277
##   Yes 15.32857 3.928210
##
##      PerformanceRating
## Y         [,1]      [,2]
##   No  3.149315 0.3566431
##   Yes 3.164286 0.3718651
##
##      RelationshipSatisfaction
## Y         [,1]     [,2]
##   No  2.726027 1.090680
##   Yes 2.607143 1.161099
##
##      StandardHours
## Y     [,1] [,2]
##   No    80    0
##   Yes   80    0
##
##      StockOptionLevel
## Y          [,1]      [,2]
##   No  0.8397260 0.8382554
##   Yes 0.4928571 0.9016087
##
##      TotalWorkingYears
## Y          [,1]     [,2]
##   No  11.602740 7.458968
##   Yes  8.185714 7.161634
##
##      TrainingTimesLastYear
## Y         [,1]     [,2]
##   No  2.867123 1.277703
##   Yes 2.650000 1.234545
##
##      WorkLifeBalance
## Y         [,1]      [,2]
##   No  2.809589 0.6874665
##   Yes 2.635714 0.8154155
##
##      YearsAtCompany
## Y         [,1]     [,2]
```

```
##   No  7.301370 5.936068
##   Yes 5.192857 6.171292
##
##      YearsInCurrentRole
## Y         [,1]     [,2]
##   No  4.453425 3.644888
##   Yes 2.907143 3.332630
##
##      YearsSinceLastPromotion
## Y         [,1]     [,2]
##   No  2.175342 3.146526
##   Yes 2.135714 3.395322
##
##      YearsWithCurrManager
## Y         [,1]     [,2]
##   No  4.369863 3.590900
##   Yes 2.942857 3.244855
```

```
#Prediction on validation dataset
dfPreds0=predict(Naive_Bayes_Model,dfVal)
confusionMatrix(table(dfPreds0,dfVal$Attrition))
```

```
## Confusion Matrix and Statistics
##
##
## dfPreds0  No Yes
##      No  210  14
##      Yes  41  35
##
##                Accuracy : 0.8167
##                  95% CI : (0.7682, 0.8588)
##     No Information Rate : 0.8367
##     P-Value [Acc > NIR] : 0.8449784
##
##                   Kappa : 0.451
##
##  Mcnemar's Test P-Value : 0.0004552
##
##             Sensitivity : 0.8367
##             Specificity : 0.7143
##          Pos Pred Value : 0.9375
##          Neg Pred Value : 0.4605
##              Prevalence : 0.8367
##          Detection Rate : 0.7000
##    Detection Prevalence : 0.7467
##       Balanced Accuracy : 0.7755
##
##        'Positive' Class : No
##
```

```
cMatrix<-table(dfPreds0, dfVal$Attrition)
plot(cMatrix, col="blue", ylab="Actual", xlab="Predicted", main='Naive Bayes Confusion Matrix')
```




```
#Test against competition dataset
dfPreds_Comp_Att=predict(Naive_Bayes_Model,dfCompAtt)
```

KNN nearest neighbor classification

```
indx <- sapply(dfTrain, is.factor)
dfTrain[indx] <- lapply(dfTrain[indx], function(x) as.numeric(as.factor(x)))
dfTrain <- dfTrain[, sapply(dfTrain, is.numeric)]

indx <- sapply(dfVal, is.factor)
dfVal[indx] <- lapply(dfVal[indx], function(x) as.numeric(as.factor(x)))
dfVal <- dfVal[, sapply(dfVal, is.numeric)]

indx <- sapply(dfCompAtt, is.factor)
dfCompAtt[indx] <- lapply(dfCompAtt[indx], function(x) as.numeric(as.factor(x)))
dfCompAtt <- dfCompAtt[, sapply(dfCompAtt, is.numeric)]

# k = 10
#classifications = knn(dfTrain[,c(4:35)],dfVal[,c(4:35)],dfTrain$Attrition,k = 9, l=0, prob = FALSE, us
#table(dfVal$Attrition,classifications)
#confusionMatrix(table(dfVal$Attrition,classifications))
```

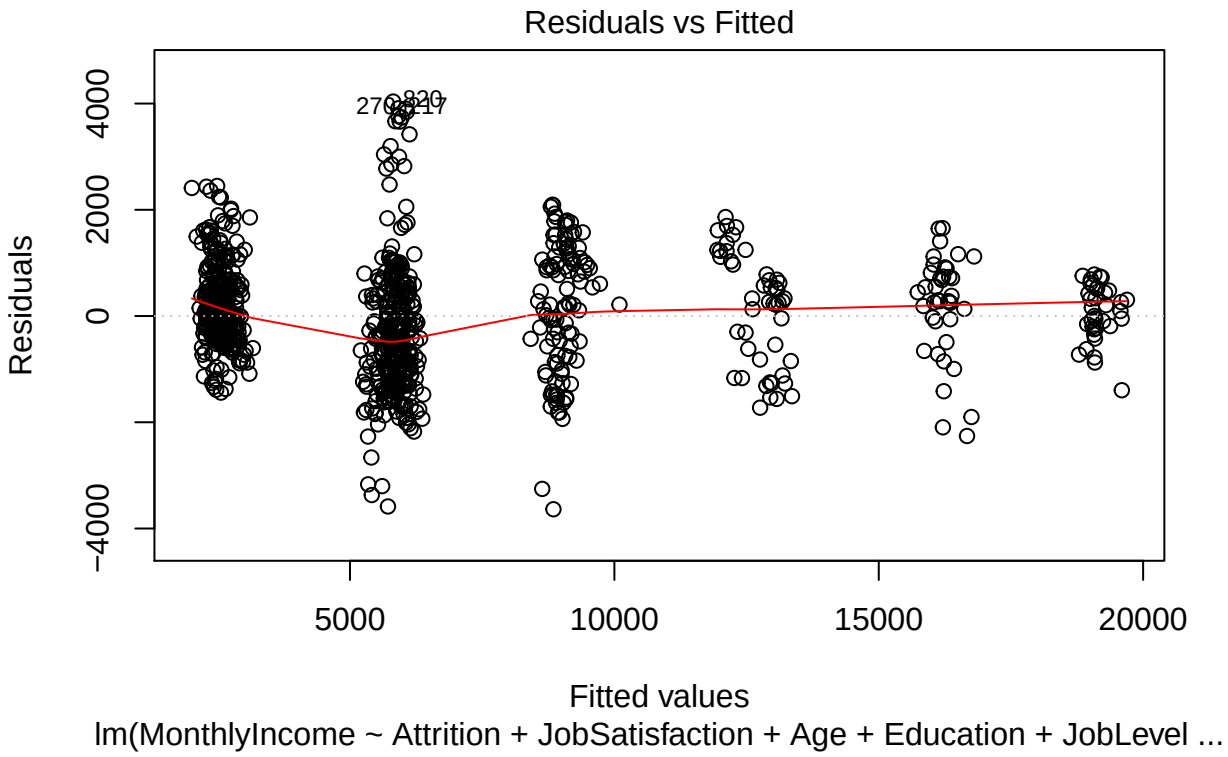
#Output Dataset

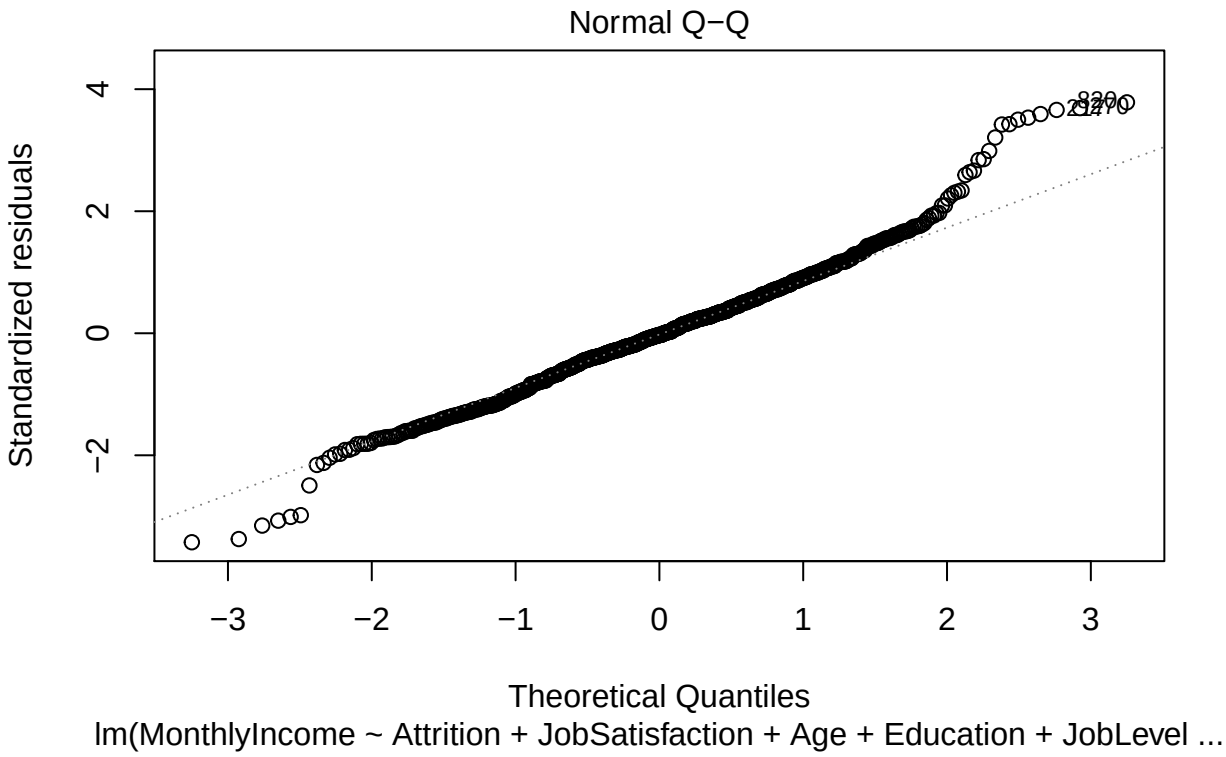
Linear Regression Model to Predict Salary

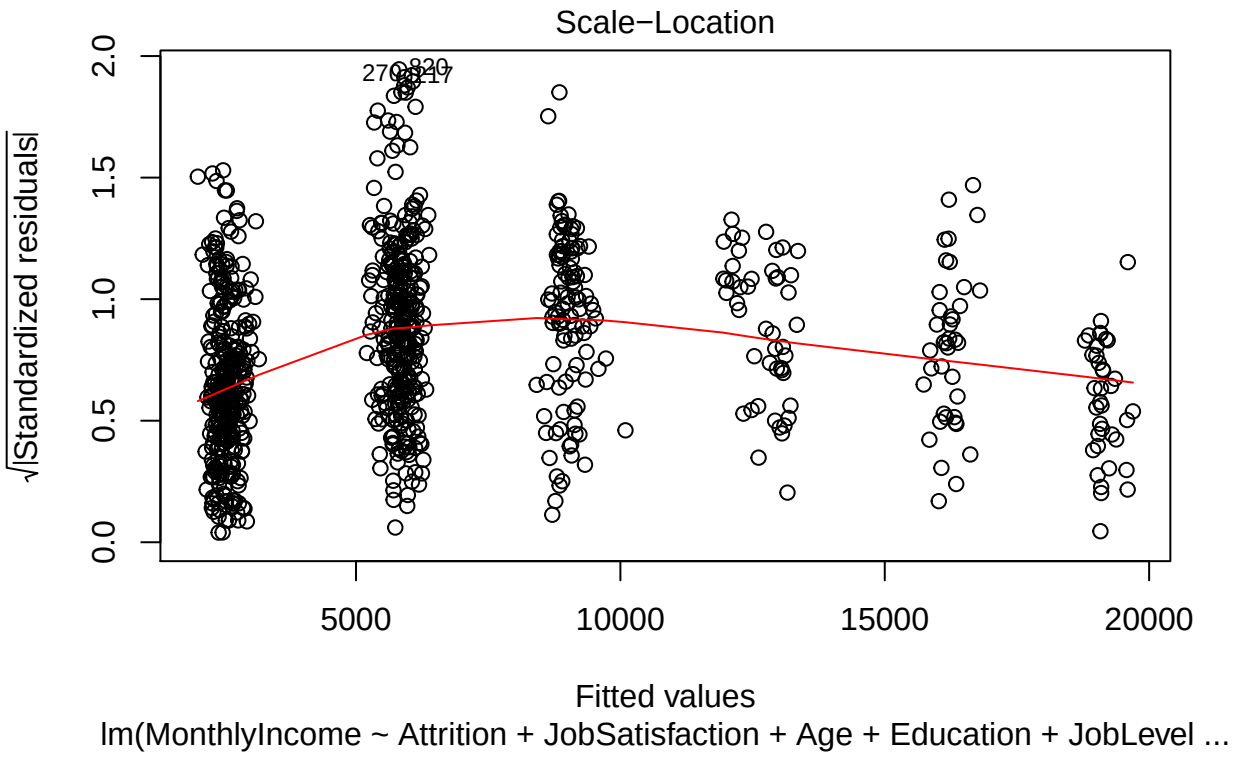
```r
dfTrain <- read.csv(file="CaseStudy2-data.csv", header=TRUE, stringsAsFactors=TRUE)

lr_mod <- lm(MonthlyIncome ~ Attrition + JobSatisfaction + Age + Education + JobLevel +
  PerformanceRating + YearsSinceLastPromotion + YearsWithCurrManager + WorkLifeBalance +
  YearsAtCompany + JobRole + EducationField , data=dfTrain)

predictMonthlyIncome <- predict(lr_mod, dfCompSal, se.fit = TRUE)
summary(lr_mod)
```

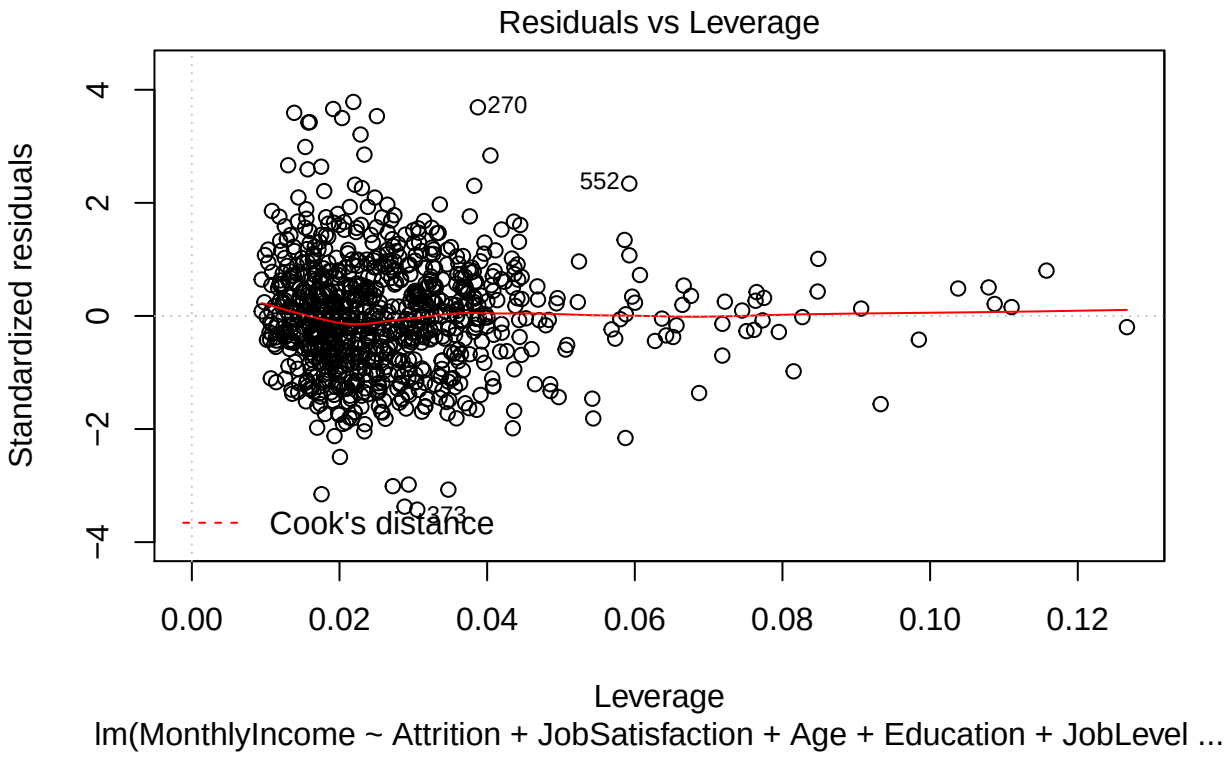
```
##
## Call:
## lm(formula = MonthlyIncome ~ Attrition + JobSatisfaction + Age +
##     Education + JobLevel + PerformanceRating + YearsSinceLastPromotion +
##     YearsWithCurrManager + WorkLifeBalance + YearsAtCompany +
##     JobRole + EducationField, data = dfTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3637.4  -652.2   -30.5   607.6  4037.0
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      87.534    575.849   0.152  0.87922
## AttritionYes                     34.819    105.806   0.329  0.74217
## JobSatisfaction                  10.721     33.336   0.322  0.74783
## Age                              13.033      4.856   2.684  0.00742 **
## Education                       -31.608     37.453  -0.844  0.39894
## JobLevel                       2967.108     75.245  39.432  < 2e-16 ***
## PerformanceRating              -136.653    103.014  -1.327  0.18501
## YearsSinceLastPromotion          33.050     15.307   2.159  0.03112 *
## YearsWithCurrManager            -26.524     16.320  -1.625  0.10449
## WorkLifeBalance                 -38.898     52.159  -0.746  0.45601
## YearsAtCompany                   14.485     11.676   1.241  0.21512
## JobRoleHuman Resources         -323.467    292.498  -1.106  0.26909
## JobRoleLaboratory Technician   -539.705    172.418  -3.130  0.00181 **
## JobRoleManager                 4041.972    235.662  17.152  < 2e-16 ***
## JobRoleManufacturing Director   140.174    170.853   0.820  0.41220
## JobRoleResearch Director       4077.131    221.581  18.400  < 2e-16 ***
## JobRoleResearch Scientist      -275.023    172.156  -1.598  0.11052
## JobRoleSales Executive          -86.039    156.183  -0.551  0.58186
## JobRoleSales Representative    -457.327    223.568  -2.046  0.04111 *
## EducationFieldLife Sciences      69.152    342.410   0.202  0.84000
## EducationFieldMarketing          -2.743    361.589  -0.008  0.99395
## EducationFieldMedical           -28.128    343.513  -0.082  0.93476
## EducationFieldOther              39.658    369.350   0.107  0.91452
## EducationFieldTechnical Degree   11.751    359.433   0.033  0.97393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1079 on 846 degrees of freedom
## Multiple R-squared:  0.9464, Adjusted R-squared:  0.945
## F-statistic: 649.7 on 23 and 846 DF,  p-value: < 2.2e-16
```

```r
plot(lr_mod)
```


Residuals vs Fitted
Residuals
4000
2000
0
−4000
5000
10000
15000
20000
Fitted values
lm(MonthlyIncome ~ Attrition + JobSatisfaction + Age + Education + JobLevel ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(MonthlyIncome ~ Attrition + JobSatisfaction + Age + Education + JobLevel ...
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5000
10000
15000
20000
Fitted values
lm(MonthlyIncome ~ Attrition + JobSatisfaction + Age + Education + JobLevel ...



#Output Monthly Income Dataset