**LOGISTIC REGRESSION**



**By**

**Kavya Naveen Kumar**

**C00223350**

**Data Visualization CA-2**

**CW\_SRDAT\_M**

**MSc. In Data Science**

**Institute of technology Carlow**

**Supervisor: Agnes Maciocha Submission Date: 10th May 2018**

Class Assessment - 2

**Introduction**

**Logistic Regression**

Logistic regression is similar to linear regression, except that it predicts whether our chosen target variable is true or false, i.e., it is used to predict categorical variable, unlike linear regression model which predicts continuous data. Also, instead of fitting a line to the data, logistic regression fits an “S” shaped “logistic function.”Although logistic regression tells the probability that the given target variable is true or false, it commonly used for classification. For example, if the probability of positive occurrence of an event is above 50 percentage, then it is classified as true. Otherwise, it is classified as false. Just like with linear regression, we can make simple models or complicated models(more than one independent variable). It can work with continuous as well as discrete data. In logistic regression, we test to see if an independent variable’s effect on the prediction is significantly different from 0. If not, it means the variable is not helping the prediction. Various tests are carried out to figure out the significance of a variable in the model. It means we can save time and space by eliminating the variable which is not significant for the model. Logistic regression’s ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a mostly used machine learning method. One big difference between linear and logistic regression is how the line is fit to the data. With linear regression, we fit the line using “least square,”i.e., we find the line that minimizes the sum of the squares if these residuals. We also use the residuals to calculate R2 and to compare simple models to complicated models. But logistic regression uses “maximum likelihood” where we pick the probability, scaled by various independent variable, and use that to calculate the likelihood of observing all the instances and lastly multiply all of those likelihoods together. That’s the likelihood of the data given the line. Then we shift the line and calculate a new likelihood of the data repeatedly till the curve with the maximum likelihood is selected.

**About Dataset**

Attrition means losing employees as well as talent. As employees leave an organization, they take with them much-need skills and qualifications that they develop during his tenure. It is expensive from the company point of view to train the new employee. The concern of retaining one’s employees happy and satisfied is a sustained and age-old hurdle. If an employee you have invested so much time and money parts for a better opportunity, then this would mean that you would have to spend more time and money to hire somebody else. Therefore with the help of predictive modelings like logistic regression and we see if we can predict employee attrition on this synthetically created IBM dataset.

**Research question:**

**1. Can we analyse the various factors which lead an employee to turnover using regression model?**

**2. Is it work related factors or non-worked related factors with drives employees to quit the job?**

The struggle with attrition and its implications for productivity and morale is probabily the vital difficulty any manager will have to bear. Responses to counteract and to decrease attrition often require the manager to understand the causes and estimate the costs associated with staff turnover.

In his 2009 paper, Yazinski identified 7 main reasons apart from salary that contribute to employee turnover:

1)Employees undergo pressure from overtime and have a work-life imbalance.

2)Least growths and advancement opportunities.

3)Employees feel the job or workplace is not what they anticipated.

4)There is no connectivity between the position and person,

5)There are negligible coaching and feedback.

6)Employees feel depreciated and unrecognized.

This report is structured as follows:

**Exploratory Data Analysis**

In this section, we investigate the dataset by taking a look at the feature distributions, how correlated is one feature with the other and create visualizations.

Feature Engineering and Categorical Encoding: Conduct some feature engineering as well as encode all our certain categorical features into a numerical variable.

Implementing Logistic regression model: We implement a logistic regression Model and analyze the summary.

1. **Dataset overview.**

There are a total of 1,470 observations with 35 variables. Target variable is Attrition with 1232 ‘NO’ and 237 ‘Yes’ that explains the unbalance target variable.

Employee Count is equal 1 for all observation which can not produce useful information for this sample data. In this analysis, we will remove it.

Over 18 is equal to ‘Y,’ which indicates the employee is not less than 18 years old. This attribute is not necessary for this analysis, hence removing it.

EmployeeNumber is a variable for identifying the specific employee. Which is not significant for this analysis hence removing it.

Standard Hours is equal 80 for all observation. The decision for this attribute is same to Over18 and Employee Count, hence removing it. Business travel, Department, education field, Gender, job role, MaritalStatus, and OverTime are categorical data, and other variables are continues.

Some of the variables are related to the years of working which can be a useful variable for future analysis. Some of the variables are related to personal issues like WorkLifeBalance, RelationshipSatisfaction, JobSatisfaction, EnvironmentSatisfaction, etc.

Some variables are associated with the income like MonthlyIncome, PercentSalaryHike, etc. We have to investigate that, how the company objective factors influence attrition employees, and what kind of working environment most will cause employees attrition.

getwd()

## [1] "/Users/kavya/Downloads"

setwd("/Users/kavya/Downloads")

library(ggplot2) # Data Visualization.  
library(corrplot) #visualization of correlation matrix

library(gridExtra) # Grid Graphics  
library(car) #vif function

attritiondata<-read.csv("Employee-Attrition.csv")

str(attritiondata)

**Removing Variable**

Some variables are not significant for the analysis since its standard deviation is zero. Those variable are Employee count, Over18, StandardHours. Employee Number. So removing those variables from the data frame.

attritdata <- attritiondata[,-c(9,10,22,27)]  
str(attritdata)

# Checking the missing value  
sapply(attritdata, function(x) sum(is.na(x)))

## Age Attrition BusinessTravel   
## 0 0 0   
## DailyRate Department DistanceFromHome   
## 0 0 0   
## Education EducationField EnvironmentSatisfaction   
## 0 0 0   
## Gender HourlyRate JobInvolvement   
## 0 0 0   
## JobLevel JobRole JobSatisfaction   
## 0 0 0   
## MaritalStatus MonthlyIncome MonthlyRate   
## 0 0 0   
## NumCompaniesWorked OverTime PercentSalaryHike   
## 0 0 0   
## PerformanceRating RelationshipSatisfaction StockOptionLevel   
## 0 0 0   
## TotalWorkingYears TrainingTimesLastYear WorkLifeBalance   
## 0 0 0   
## YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion   
## 0 0 0   
## YearsWithCurrManager   
## 0

**Logistic regression Model:**

# Model<-glm(outcome~predictor(s), data=dataFrame, family = name of a distribution, na.action = an action)

***family***- is the name of a distribution. Attrition is a binomial variable with ‘YES’ or ‘NO’ categories.

***na.action*** - optional command for missing values.

**Model 1:**

Logistic regression model including all the 31 variables.

Attritionmodel1<- glm(Attrition ~. , data = attritdata, family='binomial')

summary(Attritionmodel1)

##   
## Call:  
## glm(formula = Attrition ~ ., family = "binomial", data = attritdata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6450 -0.4836 -0.2532 -0.0897 3.3695   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.063e+01 3.838e+02 -0.028 0.97790  
## Age -3.088e-02 1.354e-02 -2.281 0.02256  
## BusinessTravelTravel\_Frequently 1.912e+00 4.124e-01 4.636 3.55e-06  
## BusinessTravelTravel\_Rarely 1.032e+00 3.801e-01 2.716 0.00661  
## DailyRate -2.809e-04 2.202e-04 -1.276 0.20212  
## DepartmentResearch & Development 1.274e+01 3.838e+02 0.033 0.97351  
## DepartmentSales 1.258e+01 3.838e+02 0.033 0.97385  
## DistanceFromHome 4.529e-02 1.077e-02 4.205 2.61e-05  
## Education 3.870e-03 8.761e-02 0.044 0.96477  
## EducationFieldLife Sciences -7.338e-01 8.044e-01 -0.912 0.36167  
## EducationFieldMarketing -3.353e-01 8.532e-01 -0.393 0.69432  
## EducationFieldMedical -8.256e-01 8.040e-01 -1.027 0.30450  
## EducationFieldOther -8.040e-01 8.639e-01 -0.931 0.35203  
## EducationFieldTechnical Degree 1.930e-01 8.226e-01 0.235 0.81445  
## EnvironmentSatisfaction -4.352e-01 8.289e-02 -5.250 1.52e-07  
## GenderMale 3.974e-01 1.846e-01 2.153 0.03131  
## HourlyRate 1.055e-03 4.410e-03 0.239 0.81093  
## JobInvolvement -5.253e-01 1.225e-01 -4.288 1.80e-05  
## JobLevel -8.424e-02 3.161e-01 -0.267 0.78984  
## JobRoleHuman Resources 1.400e+01 3.838e+02 0.036 0.97090  
## JobRoleLaboratory Technician 1.486e+00 4.837e-01 3.071 0.00213  
## JobRoleManager 3.550e-01 8.899e-01 0.399 0.68992  
## JobRoleManufacturing Director 2.251e-01 5.327e-01 0.422 0.67267  
## JobRoleResearch Director -1.091e+00 1.008e+00 -1.082 0.27931  
## JobRoleResearch Scientist 5.381e-01 4.951e-01 1.087 0.27706  
## JobRoleSales Executive 1.194e+00 1.125e+00 1.061 0.28850  
## JobRoleSales Representative 2.147e+00 1.179e+00 1.822 0.06852  
## JobSatisfaction -4.112e-01 8.151e-02 -5.046 4.52e-07  
## MaritalStatusMarried 3.242e-01 2.669e-01 1.215 0.22443  
## MaritalStatusSingle 1.151e+00 3.456e-01 3.329 0.00087  
## MonthlyIncome 1.034e-05 8.148e-05 0.127 0.89904  
## MonthlyRate 4.884e-06 1.250e-05 0.391 0.69598  
## NumCompaniesWorked 1.953e-01 3.875e-02 5.040 4.66e-07  
## OverTimeYes 1.982e+00 1.938e-01 10.228 < 2e-16  
## PercentSalaryHike -2.189e-02 3.928e-02 -0.557 0.57740  
## PerformanceRating 1.266e-01 3.989e-01 0.317 0.75103  
## RelationshipSatisfaction -2.567e-01 8.248e-02 -3.113 0.00185  
## StockOptionLevel -2.064e-01 1.579e-01 -1.307 0.19124  
## TotalWorkingYears -6.030e-02 2.932e-02 -2.057 0.03972  
## TrainingTimesLastYear -1.897e-01 7.304e-02 -2.597 0.00941  
## WorkLifeBalance -3.702e-01 1.239e-01 -2.987 0.00282  
## YearsAtCompany 9.376e-02 3.896e-02 2.407 0.01610  
## YearsInCurrentRole -1.484e-01 4.550e-02 -3.261 0.00111  
## YearsSinceLastPromotion 1.735e-01 4.243e-02 4.088 4.34e-05  
## YearsWithCurrManager -1.347e-01 4.699e-02 -2.867 0.00415  
##   
## (Intercept)   
## Age \*   
## BusinessTravelTravel\_Frequently \*\*\*  
## BusinessTravelTravel\_Rarely \*\*   
## DailyRate   
## DepartmentResearch & Development   
## DepartmentSales   
## DistanceFromHome \*\*\*  
## Education   
## EducationFieldLife Sciences   
## EducationFieldMarketing   
## EducationFieldMedical   
## EducationFieldOther   
## EducationFieldTechnical Degree   
## EnvironmentSatisfaction \*\*\*  
## GenderMale \*   
## HourlyRate   
## JobInvolvement \*\*\*  
## JobLevel   
## JobRoleHuman Resources   
## JobRoleLaboratory Technician \*\*   
## JobRoleManager   
## JobRoleManufacturing Director   
## JobRoleResearch Director   
## JobRoleResearch Scientist   
## JobRoleSales Executive   
## JobRoleSales Representative .   
## JobSatisfaction \*\*\*  
## MaritalStatusMarried   
## MaritalStatusSingle \*\*\*  
  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1298.58 on 1469 degrees of freedom  
## Residual deviance: 859.01 on 1425 degrees of freedom  
## AIC: 949.01  
##   
## Number of Fisher Scoring iterations: 14

Deviance statistics (-2LL) is very useful to analyze the fit of the model. Smaller the value of the deviance the better the model and vice-versa.

The summary of the logistic regression model provides two deviance statistics:

->**Null deviance\*** - the deviance of the model that contains no predictors other than the constant (-2LL (baseline))

->**The residual deviance** - deviance for the model

From above summary, we can see that there is a considerable difference between Null deviance n and residual deviance.

Null deviance: 1298.58 on 1469 degrees of freedom

Residual deviance: 859.01 on 1425 degrees of freedom

This can be better accessed using chi-square statistics:

It measures the difference between the model with predictors and without predictors. We can assess the significance of the change in a model by taking the log-likelihood of the new model and subtracting the log-likelihood of the baseline model from it.

The value of the model chi-square statistics works on this principle and = -2LL(with Intervention) - (-2LL (with constant)) = (1298.58 - 859.01 = 439.57)

This value has a chi-square distribution, and so you can calculate its statistical significance

modelChi1<- Attritionmodel1$null.deviance-Attritionmodel1$deviance  
modelChi1

## [1] 439.5731

The degrees of freedom for the model is stored in the variable df.residual, and for the null model it will be stored in df.null

chidf1<-Attritionmodel1$df.null- Attritionmodel1$df.residual  
chidf1

## [1] 44

The change in df=44 reflects that we have 44 variables in the model.

So we can say that the model is predicting whether an employees attempts attrition or not with the help of all the independent variables.

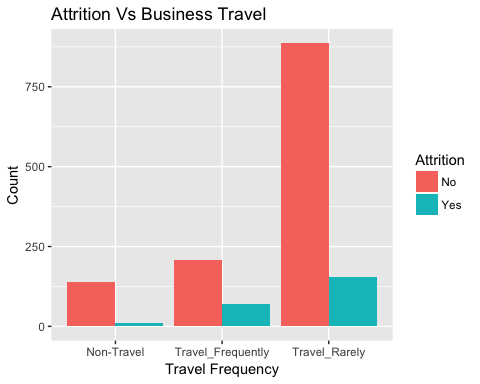
From the above summary, we can observe that some of the variables are not as significant as others by analyzing the p-value. A detailed exploratory data analysis and few tests are performed below to estimate the most significant variables and to ignore those variables which don’t contribute much to the analysis.

**Exploratory data analysis:**

**i)Attrition VS BusinessTravel**

**How frequently an employee goes to business travel.**

ggplot(attritdata,aes(BusinessTravel,fill=Attrition))+geom\_bar(position=position\_dodge())+labs(x="Travel Frequency",y="Count",title="Attrition Vs Business Travel")



table\_travel<-table(attritdata$BusinessTravel, attritdata$Attrition)  
chisq.test(table\_travel)

##   
## Pearson's Chi-squared test  
##   
## data: table\_travel  
## X-squared = 24.182, df = 2, p-value = 5.609e-06

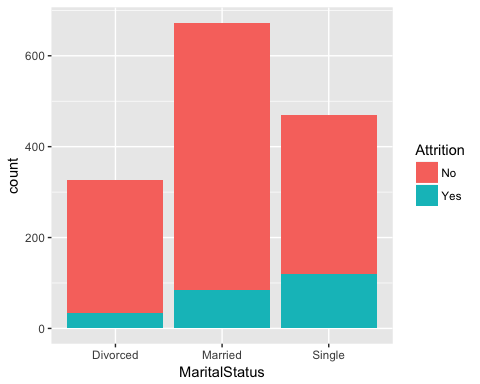
The barplot shows that employees who travel rarely do not frequently quit the job. Thus, attrition is dependent on business travel, and the chi square test proves this as the p-value is less than alpha(0.05).

**ii)Attiriton vs MaritalStatus**

Marital <- ggplot(attritdata, aes(MaritalStatus,fill = Attrition))  
Marital1 <- Marital + geom\_histogram(stat="count")

## Warning: Ignoring unknown parameters: binwidth, bins, pad

print(Marital1)



table\_marital<-table(attritdata$MaritalStatus, attritdata$Attrition)  
chisq.test(table\_marital)

##   
## Pearson's Chi-squared test  
##   
## data: table\_marital  
## X-squared = 46.164, df = 2, p-value = 9.456e-11

tapply(as.numeric(attritdata$Attrition) - 1 ,attritdata$MaritalStatus,mean)

## Divorced Married Single   
## 0.1009174 0.1248143 0.2553191

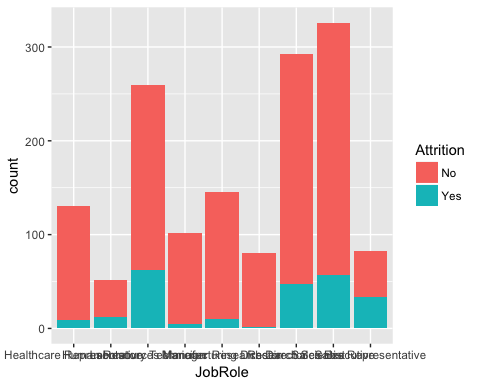
From the chi-square test we can see that p-value is 9.456e-11 which is less than 0.05 and marital status is significant fro this analysis. Single personnels have more tendency to be subject to attrition as seen in the tapply test.

1. **Attrition vs JobRole**

Jobrole <- ggplot(attritdata, aes(JobRole,fill = Attrition))  
Jobrole1 <- Jobrole + geom\_histogram(stat="count")

## Warning: Ignoring unknown parameters: binwidth, bins, pad

print(Jobrole1)



table\_jobrole<-table(attritdata$JobRole, attritdata$Attrition)  
chisq.test(table\_jobrole)

##   
## Pearson's Chi-squared test  
##   
## data: table\_jobrole  
## X-squared = 86.19, df = 8, p-value = 2.752e-15

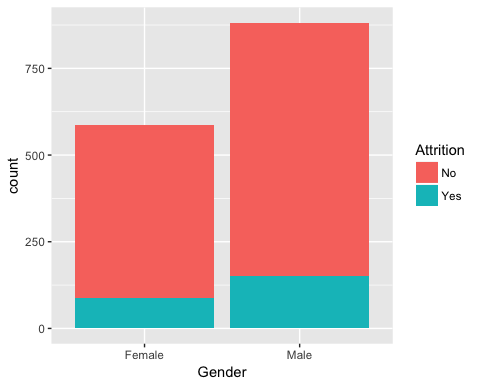
From the chi-square test we can see that p-value is 2.752e-15 which is less than 0.05 and marital status is significant fro this analysis.We see that labaratory technican, human resources and sales representative roles have more attrition.

1. **Attrition vs Gender**

Gender <- ggplot(attritdata, aes(Gender,fill = Attrition))  
Gender1 <- Gender + geom\_histogram(stat="count")

## Warning: Ignoring unknown parameters: binwidth, bins, pad

print(Gender1)



table\_gender<-table(attritdata$Gender, attritdata$Attrition)  
chisq.test(table\_gender)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: table\_gender  
## X-squared = 1.117, df = 1, p-value = 0.2906

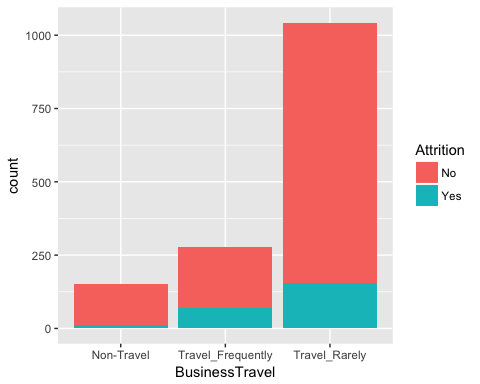
From the chi-square test we can see that p-value is 0.2906 which is more than 0.05 and Gender is significant fro this analysis. Gender variable doesn’t have so much significance for this analysis.

**v)Attrition vs BusinessTravel**

BusinessTravel <- ggplot(attritdata, aes(BusinessTravel,fill = Attrition))  
BusinessTravel1 <- BusinessTravel + geom\_histogram(stat="count")

## Warning: Ignoring unknown parameters: binwidth, bins, pad

print(BusinessTravel1)



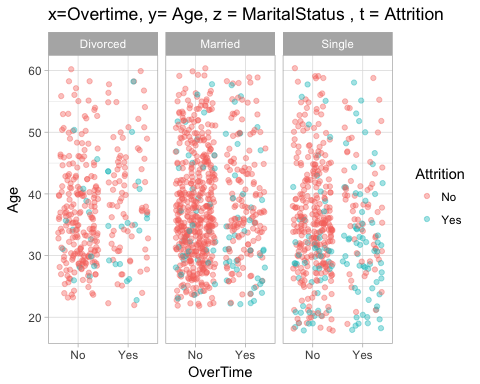
table\_BusinessTravel<-table(attritdata$BusinessTravel, attritdata$Attrition)  
chisq.test(table\_BusinessTravel)

##   
## Pearson's Chi-squared test  
##   
## data: table\_BusinessTravel  
## X-squared = 24.182, df = 2, p-value = 5.609e-06

From the chi-square test we can see that p-value is 5.609e-06 which is less than 0.05 and BusinessTravel is significant fro this analysis. A non-business traveler has little probability of having attrition. A business traveler who travels frequently has more likelihood to have attrition.

1. **Plot to understand the effects of Overtime, Age, MaritalStatus of the employees on Attrition**

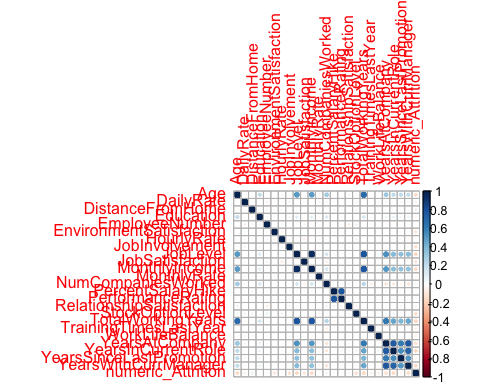
ggplot(attritdata, aes(OverTime, Age)) +   
 facet\_grid(.~MaritalStatus) +  
 geom\_jitter(aes(color = Attrition),alpha = 0.4) +   
 ggtitle("x=Overtime, y= Age, z = MaritalStatus , t = Attrition") +   
 theme\_light()



From the above can see that attrition is accumulating in the region where OverTime =Yes & Marial\_Status=Single & Age< 35.

**Correlation model:**

numeric\_mydata <- attritiondata[,c(1,4,6,7,10,11,13,14,15,17,19,20,21,24,25,26,28:35)]  
numeric\_Attrition = as.numeric(attritdata$Attrition)- 1  
numeric\_mydata = cbind(numeric\_mydata, numeric\_Attrition)  
M <- cor(numeric\_mydata)  
corrplot(M, method="circle")



# Finding how many correlations are bigger than 0.70

k = 0  
for(i in 1:25){  
for(r in 1:25){  
 if(M[i,r]> 0.70 & i != r){  
 k= k + 1  
 }  
} }  
print(k/2)

## [1] 7

**We can deduce that :**

* The most outstanding result is between JobLevel and Monthly income, whose correlation is 0.95.
* The more performance rating, the more Performance salary hike, whose correlation is 0.773 More the total working hours.
* The more Job Level, whose correlation is 0.782 More the total working hours.
* The more monthlyIncome, whose correlation is 0.772 the more yearswithcurrmanager.
* The more yearsatcompany , whose correlation is 0.769.
* The more yearsatcompany, the more yearsInCurrentRole , whose correlation is 0.758.
* The more yearswithcurrmanager, the more yearsincurrentrole, whose correlation is 0.71.

**Model2**

**Attritionmodel with work-related variable(Professional life)**

Attritionmodel2 <- glm(Attrition ~ Age+BusinessTravel+DistanceFromHome+JobInvolvement+JobRole+JobSatisfaction+NumCompaniesWorked+OverTime+TotalWorkingYears+TrainingTimesLastYear+YearsAtCompany+YearsInCurrentRole+YearsSinceLastPromotion+YearsWithCurrManager+PercentSalaryHike+PerformanceRating, data = attritdata, family='binomial')

summary(Attritionmodel2)

##   
## Call:  
## glm(formula = Attrition ~ Age + BusinessTravel + DistanceFromHome +   
## JobInvolvement + JobRole + JobSatisfaction + NumCompaniesWorked +   
## OverTime + TotalWorkingYears + TrainingTimesLastYear + YearsAtCompany +   
## YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager +   
## PercentSalaryHike + PerformanceRating, family = "binomial",   
## data = attritdata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7559 -0.5548 -0.3098 -0.1353 3.3068   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.462037 1.093771 -0.422 0.672714   
## Age -0.038090 0.012875 -2.958 0.003093 \*\*   
## BusinessTravelTravel\_Frequently 1.662379 0.375346 4.429 9.47e-06 \*\*\*  
## BusinessTravelTravel\_Rarely 0.864461 0.348406 2.481 0.013094 \*   
## DistanceFromHome 0.036286 0.009914 3.660 0.000252 \*\*\*  
## JobInvolvement -0.564528 0.113791 -4.961 7.01e-07 \*\*\*  
## JobRoleHuman Resources 1.373792 0.543878 2.526 0.011540 \*   
## JobRoleLaboratory Technician 1.512666 0.424925 3.560 0.000371 \*\*\*  
## JobRoleManager 0.097419 0.640400 0.152 0.879091   
## JobRoleManufacturing Director 0.110180 0.514316 0.214 0.830370   
## JobRoleResearch Director -0.994723 0.882622 -1.127 0.259739   
## JobRoleResearch Scientist 0.736531 0.428743 1.718 0.085818 .   
## JobRoleSales Executive 1.228149 0.413330 2.971 0.002965 \*\*   
## JobRoleSales Representative 2.150369 0.478467 4.494 6.98e-06 \*\*\*  
## JobSatisfaction -0.366697 0.075028 -4.887 1.02e-06 \*\*\*  
## NumCompaniesWorked 0.160675 0.035629 4.510 6.49e-06 \*\*\*  
## OverTimeYes 1.662175 0.174203 9.542 < 2e-16 \*\*\*  
## TotalWorkingYears -0.040655 0.024222 -1.678 0.093253 .   
## TrainingTimesLastYear -0.176890 0.068952 -2.565 0.010306 \*   
## YearsAtCompany 0.069061 0.035305 1.956 0.050451 .   
## YearsInCurrentRole -0.144985 0.042331 -3.425 0.000615 \*\*\*  
## YearsSinceLastPromotion 0.144242 0.038828 3.715 0.000203 \*\*\*  
## YearsWithCurrManager -0.104643 0.042813 -2.444 0.014518 \*   
## PercentSalaryHike -0.021557 0.036481 -0.591 0.554572   
## PerformanceRating 0.178078 0.368730 0.483 0.629132   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1298.58 on 1469 degrees of freedom  
## Residual deviance: 964.79 on 1445 degrees of freedom  
## AIC: 1014.8  
##   
## Number of Fisher Scoring iterations: 6

**Model 3:**

Attritionmodel3 with non-work related variables(Personal life)

Attritionmodel3 <- glm(Attrition ~ EnvironmentSatisfaction+Gender+MaritalStatus+RelationshipSatisfaction+WorkLifeBalance, data = attritdata, family='binomial')

summary(Attritionmodel3)

##   
## Call:  
## glm(formula = Attrition ~ EnvironmentSatisfaction + Gender +   
## MaritalStatus + RelationshipSatisfaction + WorkLifeBalance,   
## family = "binomial", data = attritdata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.1254 -0.6197 -0.4915 -0.3884 2.4462   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.65471 0.41000 -1.597 0.1103   
## EnvironmentSatisfaction -0.25699 0.06547 -3.925 8.66e-05 \*\*\*  
## GenderMale 0.21940 0.15088 1.454 0.1459   
## MaritalStatusMarried 0.23716 0.21926 1.082 0.2794   
## MaritalStatusSingle 1.16180 0.21446 5.417 6.05e-08 \*\*\*  
## RelationshipSatisfaction -0.13069 0.06660 -1.962 0.0497 \*   
## WorkLifeBalance -0.24304 0.10059 -2.416 0.0157 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1298.6 on 1469 degrees of freedom  
## Residual deviance: 1226.5 on 1463 degrees of freedom  
## AIC: 1240.5  
##   
## Number of Fisher Scoring iterations: 4

From above summary, we can see that there is a small difference between Null deviance and residual deviance. Null deviance: 1298.58 on 1469 degrees of freedom Residual deviance: 1226.5 on 1425 degrees of freedom.

This can be better accessed using chi-square statistics.

It measures the difference between the model with predictors and without predictors. We can assess the significance of the change in a model by taking the log-likelihood of the new model and subtracting the log-likelihood of the baseline model from it.

The value of the model chi-square statistics works on this principle and = -2LL(with Intervention) - (-2LL (with constant)) = (1298.58 - 1226.5 = 72.08)

This value has a chi-square distribution, and so you can calculate its statistical significance.

**Lets calculate anova test between Model2 and Model3:**

anova(Attritionmodel3, Attritionmodel2)

## Analysis of Deviance Table  
##   
## Model 1: Attrition ~ EnvironmentSatisfaction + Gender + MaritalStatus +   
## RelationshipSatisfaction + WorkLifeBalance  
## Model 2: Attrition ~ Age + BusinessTravel + DistanceFromHome + JobInvolvement +   
## JobRole + JobSatisfaction + NumCompaniesWorked + OverTime +   
## TotalWorkingYears + TrainingTimesLastYear + YearsAtCompany +   
## YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager +   
## PercentSalaryHike + PerformanceRating  
## Resid. Df Resid. Dev Df Deviance  
## 1 1463 1226.52   
## 2 1445 964.79 18 261.73

Anova compares two models and confirms that the difference between the models is 261.73 which means that model2 is better than model1. i.e., work-related variable affects the employees more than non-work related variables. Employees tend to leave the company mainly because of work-related issues.

**Validation:**

The data is split into testing and trai ning sets. 70 % of the dataset divided into the training set, and rest is used to test the models.

data1 = sort(sample(nrow(attritdata), nrow(attritdata)\*.7))  
train<-attritdata[-data1,]  
test<-attritdata[-data1,]

attm1=glm(Attrition~.,data=train,family = binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

predictionGlm=predict(attm1,type="response",newdata=test)  
#Confusion matrix is then computed to check the accuracy of the model.  
  
  
table(test$Attrition,predictionGlm>.5)

##   
## FALSE TRUE  
## No 371 8  
## Yes 27 35

(371+35)/(371+35+27+8)

## [1] 0.9274376

It is observed that the logistic regression model gives an accuracy of 0.9274376 which is quite decent. It means GLM behaves quite well at predicting which employees might leave the organization.

anova(Attritionmodel2,test = 'Chisq')

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Attrition  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 1469 1298.58   
## Age 1 39.508 1468 1259.07 3.268e-10 \*\*\*  
## BusinessTravel 2 23.753 1466 1235.32 6.952e-06 \*\*\*  
## DistanceFromHome 1 9.711 1465 1225.61 0.0018319 \*\*   
## JobInvolvement 1 23.992 1464 1201.62 9.675e-07 \*\*\*  
## JobRole 8 56.670 1456 1144.95 2.089e-09 \*\*\*  
## JobSatisfaction 1 18.460 1455 1126.49 1.735e-05 \*\*\*  
## NumCompaniesWorked 1 18.625 1454 1107.86 1.591e-05 \*\*\*  
## OverTime 1 99.535 1453 1008.33 < 2.2e-16 \*\*\*  
## TotalWorkingYears 1 4.783 1452 1003.55 0.0287340 \*   
## TrainingTimesLastYear 1 6.161 1451 997.38 0.0130601 \*   
## YearsAtCompany 1 0.005 1450 997.38 0.9451606   
## YearsInCurrentRole 1 13.217 1449 984.16 0.0002774 \*\*\*  
## YearsSinceLastPromotion 1 13.151 1448 971.01 0.0002873 \*\*\*  
## YearsWithCurrManager 1 5.871 1447 965.14 0.0153897 \*   
## PercentSalaryHike 1 0.119 1446 965.02 0.7305426   
## PerformanceRating 1 0.233 1445 964.79 0.6289951   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From the above chi-square anova test, we can tell which work-related variable affects an employee to leave the job most. The top 3 factors are:

1. OverTime

2. JobRole

3. BusinessTravel

anova(Attritionmodel3,test = 'Chisq')

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Attrition  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 1469 1298.6   
## EnvironmentSatisfaction 1 15.502 1468 1283.1 8.241e-05 \*\*\*  
## Gender 1 1.295 1467 1281.8 0.25521   
## MaritalStatus 2 45.486 1465 1236.3 1.327e-10 \*\*\*  
## RelationshipSatisfaction 1 4.008 1464 1232.3 0.04528 \*   
## WorkLifeBalance 1 5.772 1463 1226.5 0.01629 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

From the above chi-square ANOVA test, we can tell which non-work related variable affects an employee to leave the job most. The top 3 factors are:

1. MaritalStatus

2. EnvironmentSatisfaction

3. WorkLifeBalance

**Conclusion:**

1)After comparing both the models with work-related variables and non-work related variables we can conclude that work-related variable affects the employees more.

2)From the analysis one of the main factors which drives the employees towards attrition is overtime. It is crucial for any individual to have a balanced work life. Working late hours in the office affects his/her personal life, mental and physical health too.

3) From the exploratory analysis, it is evident that the employees who are over 35 years and single and works late at the office are more prone to quit the job.

**References**:

Statistics How To. (2018). *ANOVA Test: Definition, Types, Examples*. [online] Available at: http://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/anova/ [Accessed 10 May 2018].

Kaggle.com. (2018). *IBM Attrition | Kaggle*. [online] Available at: https://www.kaggle.com/sidbhargava/ibm-attrition/data [Accessed 10 May 2018].

Qualtrics. (2018). *Employee Engagement | Qualtrics*. [online] Available at: https://www.qualtrics.com/ie/lp/employee-engagement/?utm\_medium=cpc&utm\_source=google&utm\_campaign=ee+bmm+irl&utm\_term=employee+attrition+data&gclid=Cj0KCQjw28\_XBRDhARIsAEk21FjcaPTdlf2YN5Q17i8pJHoNs9csC\_co8FJLQ643kRooEyxXSMW1IK0aAn-pEALw\_wcB [Accessed 10 May 2018].

<https://www.researchgate.net/publication/259761946_Factors_Affecting_Employee_Retention_A_Comparative_Analysis_of_two_Organizations_from_Heavy_Engineering_Industry>