The Big Apple Group

Staff Attrition

Analysis on IBM attrition data.

Manthan Shah, Rajan Singh & Ravina Gaikwad
11-20-2017

Table of Contents

Introduction 2
Data 2
Data Source: 2
Description: 2
Exploratory Analysis 3
Predictive Analysis 7
Conclusion 11

Staff Attrition

Introduction

Documenting employee activities not only help the productivity of the company but also help in understanding employee behavior. This forms a base on which the retention strategies are planned. Better retention strategies create better work environment and strengthen employee commitment to the organization.

But we can't just increase the income and expect employees to work over time. The reasons for increase in attrition must be understood before focusing on retention strategies. We can find articles online and follow them word to word in order to make changes but it won't make an impact as every organization runs differently. In such situations employee data can be analyzed extensively to control the attrition rate. Also help us predict who are likely to leave next.

What are the reasons for employee attrition? Are the common assumptions legit? Are the reasons plain monetary or are they impacting personal lives of the employees? These are some of the questions that we'd want to tackle in our analysis.

Data

Data Source:

https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition

Description:

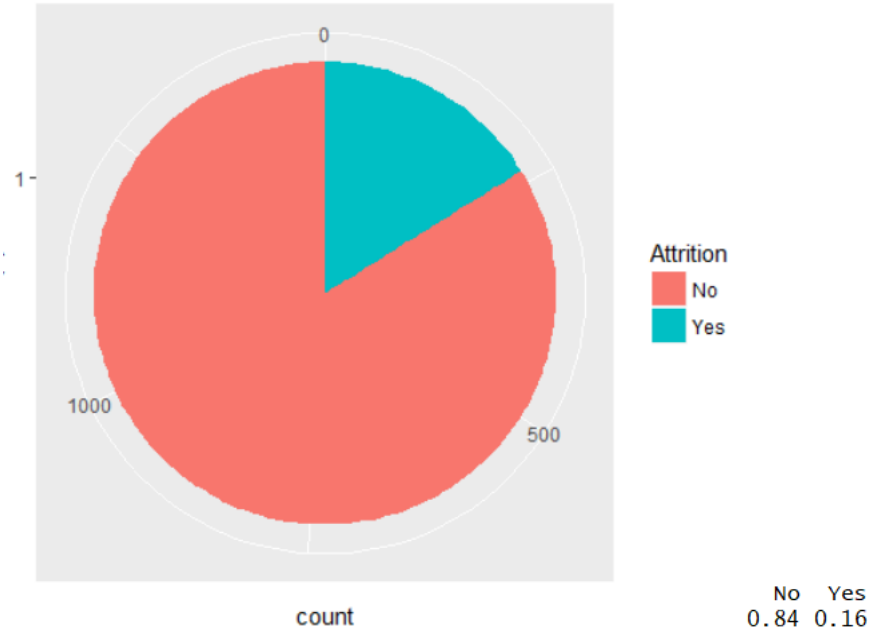
Our categorical values include variables such as Attrition, Business travel, Departments and Education fields. And numerical values include the age of the employee, percentage salary hike, wage paid per our which in our data set is the hourly rate and the variables that specify the years, for example years spent at the company, years spent at the current position, years since last promotion.

The interesting portion of the data set are the variables that need to be converted such as work life balance which consist of a number range between 1-4 (1 being the lowest value) which needs to a categorically value as it shows different levels of the said field. Similarly, we have job satisfaction, job involvement, performance rating, work environment that need to be converted. The reason these variables cannot be numerical is because finding out the mean of the entire column or the median would not make sense.

Factors that affect personal lives or work environment include work environment, work life balance, over time, work environment and job satisfaction. These are a few variables we focus on while performing our analysis. We want to find if employees in general are happy and not stressed. Another gripping question that we need answered is if the stress or displeasure is despite being paid evenly. This would help us prove the point that only paying for over time hours doesn't mean the organizations employees are working with a positive attitude which is an important factor to be considered.

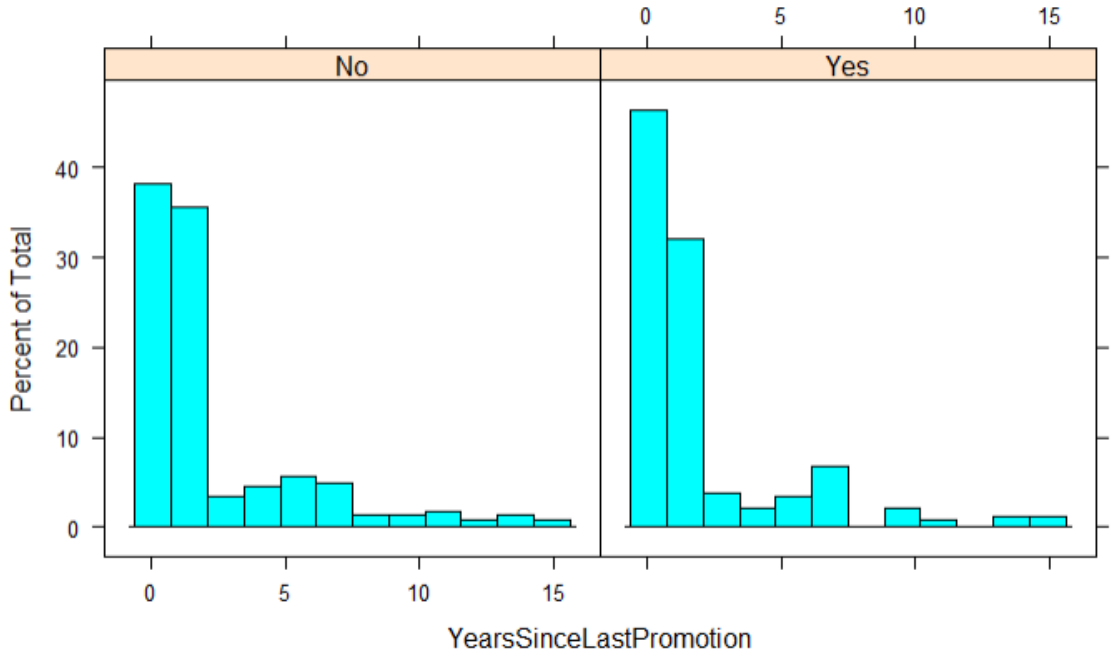
Exploratory Analysis

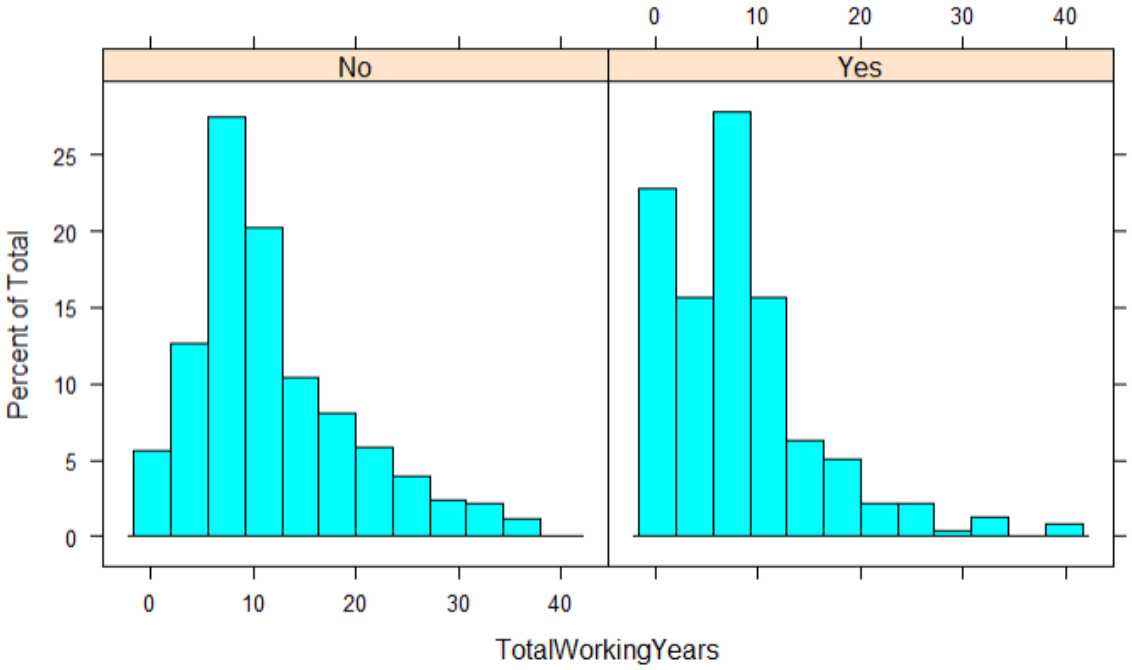
We started exploring major variables in our data sets that could possibly be the reason for staff attrition. In our analysis, we interpreted the data and found interesting results.



The above pie chart represents 84% of the total number of staff that didn't leave the firm. They are marked in red. The remaining 16% marked in blue represent staff that did leave the firm.

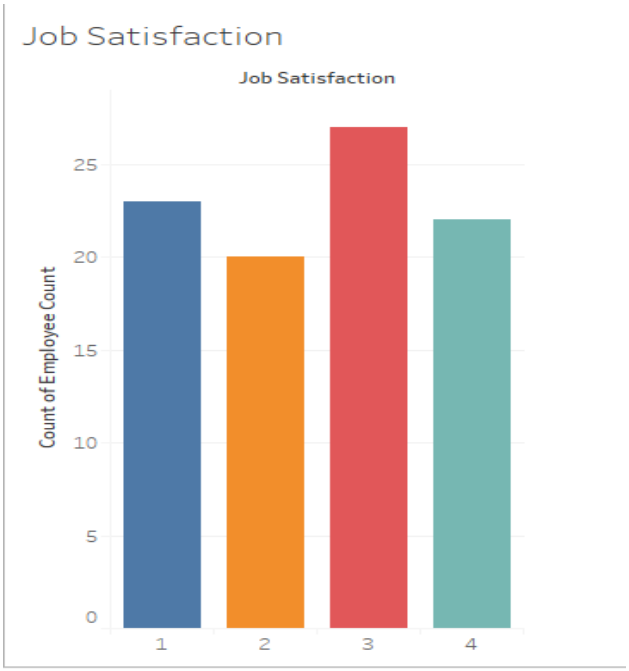
We started evaluating all the variables in our data set. Utilities used to visualize the relationship between independent and dependent variables were histogram, tableau, and boxplots. We have used these utilities extensively over exploratory analysis to find the interesting variables. Starting with Years since last promotion and total working years.


No
Yes
Percent of Total
YearsSinceLastPromotion


No
Yes
Percent of Total
TotalWorkingYears

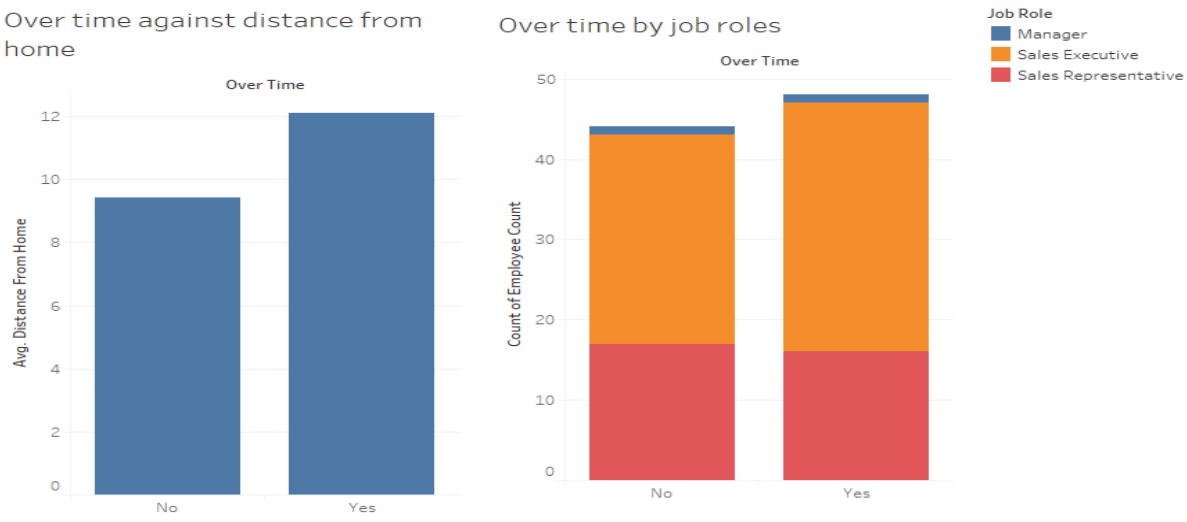
There were some results which are not obvious in nature and were interesting facts. Like the Job Satisfaction chart below. The average job satisfaction depicted by the blue line is 2.7 and we can see that maximum employees lie after the average meaning that maximum employees seemed to be happy with there jobs.

Next, we focused on Sales department first and observed how satisfied these employees were. These results are for employees who have left the firm. The difference isn't much but again, maximum people have given 3 as their job satisfaction rating.



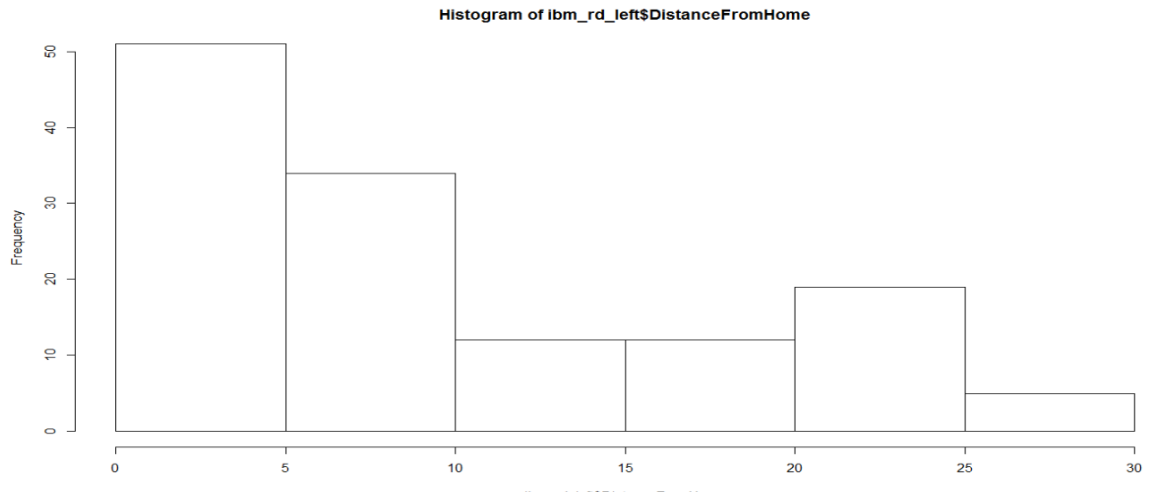
The above graph represents staff that left the firm. We can clearly observe that job satisfaction could be an essential component as percentage of staff who has job satisfaction level 1 were more than staff whose job satisfaction level 4.

Next, we observed if they were asked to work over time. Considering that this may have been one of the reasons to cause discomfort especially if they live far from the office.



Notice that people whose average distance was far from workplace worked more. Another point of observation is the sales executive form major part of that bar.

This hints that people staying closer to their offices should not have a problem with overtime considering that they will be paid hourly, but our histogram from R shows some interesting points. We considered employee from “Research & Development” department and observed that employees who were living far from office left the organization. However, majority of the employees from that department were staying closer to office location and also left the organization. So having a office close to the home has not proved to be a positive point and would require further analysis in direction of work environment at the organization or other variables.



Above histogram represent employees from Reseach & Development department who left the organization and their distance from home. Further, in addition, we explored the above histogram using overtime variable which provides information about majority of the employee who were close to office also left the organization may be due to overtime which is again an interesting fact. Over time is a contributing variable but not the average distance from home.

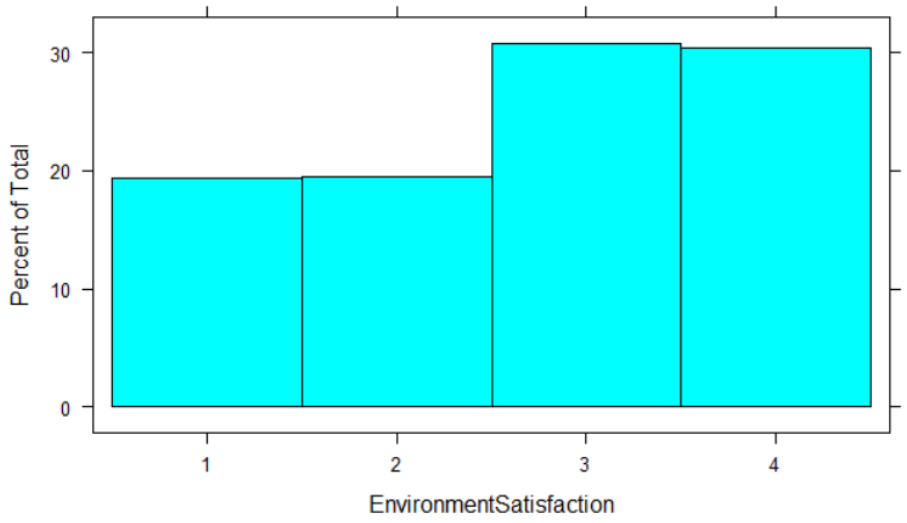
Predictive Analysis

After looking at the data, it is ironically clear that there are some unclear working conditions and personal factors which are contributing to employees leave the firm.

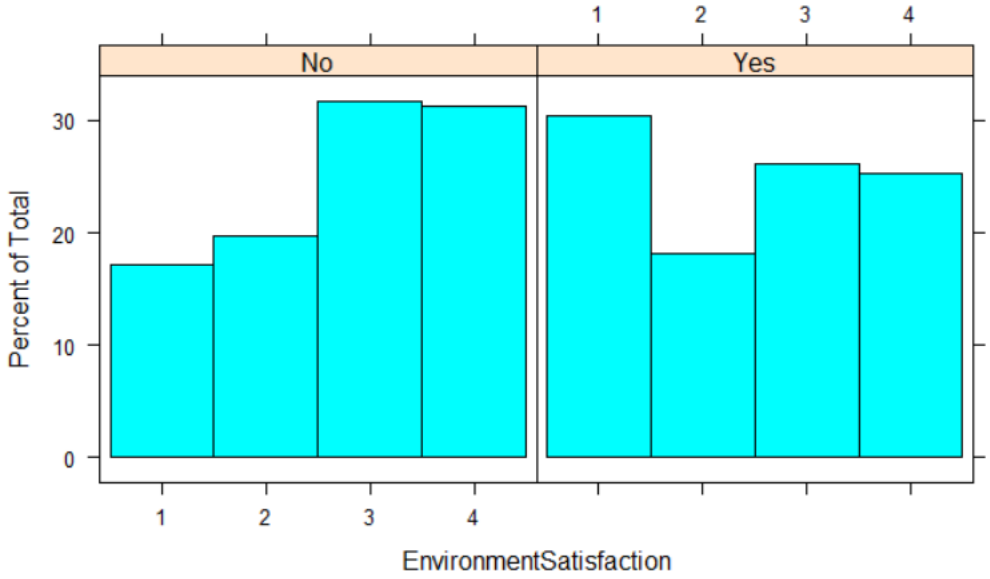
To find out what those factors are and to make a predictive model to save future employees from leaving the company, we will use Logistic Regression as our method. Since, we have a wide range of independent variables and our output dependent variable is categorical, Logistic Regression would work well to find the best fitting model.

To perform Logistic Regression, we choose only those variables in our model whose data suggest some contrasting and unusual phenomenon. Also we divided the dataset into Train data and Test data. Train data to make the model from and test data to run and compare the model output and check accuracy. In order to pick the best predictive model we choose AIC (Akaike's Information Criterion) as the criteria and also level of significance (P-value) of all the variables.

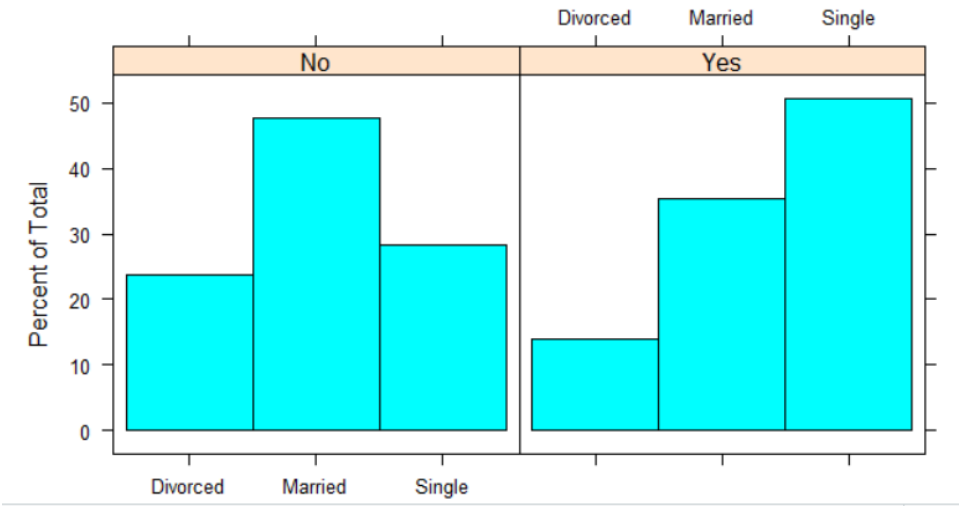
During our exploratory analysis we found some such interesting variables which we considered for creating our Attrition Prediction model. We had grouped some variables by those who left the company and those who didn't. Take Environment satisfaction as an example. If you see the first histogram very less employees seem to have very low environment satisfaction throughout the company.



But when we group them by their Attrition status, we found that the many employees, who left the company, had a very low Environment Satisfaction.

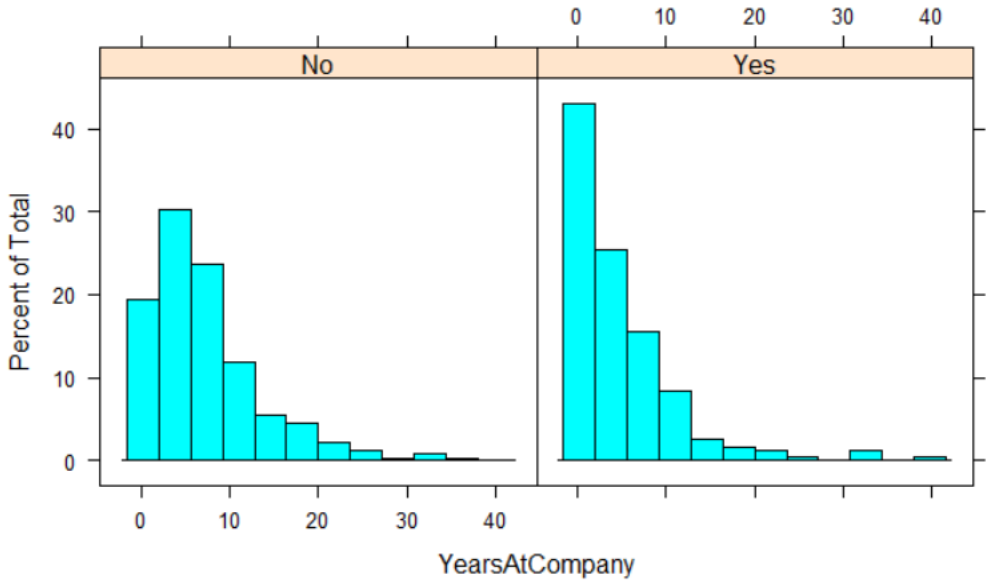


Similarly, more than 50% with their Single marital status left the company.



	Divorced	Married	Single
No	0.24	0.48	0.28
Yes	0.14	0.35	0.51

Also, a huge number of employees who hardly worked more than a year in the company left prematurely.



After shortlisting few variables, and trying to make a generalized linear model on train data, we were able to come up with a model with a lowest AIC score compared to others and significant variables. Our model looks like:

```
Attrition ~ 1 + EnvironmentSatisfaction + JobLevel + JobRole +
    JobSatisfaction + MaritalStatus + OverTime + YearsAtCompany +
    YearsInCurrentRole + YearsSinceLastPromotion + YearsWithCurrManager
<environment: 0x000000005631ac50>
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1678  -0.5518  -0.3142  -0.1260   3.1368

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.14875    0.65782  -1.746 0.080757 .
EnvironmentSatisfaction2      -1.03562    0.27617  -3.750 0.000177 ***
EnvironmentSatisfaction3      -1.30073    0.25935  -5.015 5.29e-07 ***
EnvironmentSatisfaction4      -1.36289    0.26259  -5.190 2.10e-07 ***
JobLevel2                     -1.34731    0.40367  -3.338 0.000845 ***
JobLevel3                     -0.63951    0.50033  -1.278 0.201186
JobLevel4                     -1.86105    0.78937  -2.358 0.018391 *
JobLevel5                      0.12474    1.12735   0.111 0.911894
JobRoleHuman Resources         0.79410    0.66810   1.189 0.234599
JobRoleLaboratory Technician   0.73719    0.57765   1.276 0.201891
JobRoleManager                -1.01751    0.95205  -1.069 0.285180
JobRoleManufacturing Director  0.29484    0.56740   0.520 0.603311
JobRoleResearch Director      -1.89700    1.05445  -1.799 0.072011 .
JobRoleResearch Scientist     -0.24449    0.59132  -0.413 0.679267
JobRoleSales Executive         1.35078    0.44711   3.021 0.002518 **
JobRoleSales Representative    1.30450    0.62747   2.079 0.037617 *
JobSatisfaction2              -0.65528    0.29183  -2.245 0.024744 *
JobSatisfaction3              -0.61410    0.25052  -2.451 0.014235 *
JobSatisfaction4              -1.10480    0.25862  -4.272 1.94e-05 ***
MaritalStatusMarried           0.56160    0.27749   2.024 0.042987 *
MaritalStatusSingle            1.45212    0.27747   5.233 1.66e-07 ***
OverTimeYes                    1.82697    0.19854   9.202  < 2e-16 ***
YearsAtCompany                 0.07027    0.03433   2.047 0.040681 *
YearsInCurrentRole            -0.13971    0.04794  -2.914 0.003566 **
YearsSinceLastPromotion        0.13575    0.04251   3.193 0.001407 **
YearsWithCurrManager          -0.14596    0.04803  -3.039 0.002375 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1065.40  on 1173  degrees of freedom
Residual deviance:  776.53  on 1148  degrees of freedom
AIC: 828.53
```

To check the accuracy of the model we applied the model on test data. To talk about our test data it consisted of 296 observations who had their Attrition status as follows:

Count and percentages:

```
 No Yes
257  39

  No  Yes
0.87 0.13
```

Model Prediction and confusion matrix:

```
predicted.classes  No Yes
                0 251  31
                1   6   8
```

Accuracy:

```
0.875
```

Conclusion

We started with 35 variables to predict staff attrition. As a part of our exploratory analysis, we evaluated the list of variables and eliminated those that have significantly less impact on our response variable. Secondly, we designed a predictive model using logistic regression and found the number of variables that have contributed to the attrition. They are as follows and are arranged in order of effectiveness.

- Environment satisfaction
- Job satisfaction
- Overtime
- Marital status
- Job role
- Years in current role
- Years since last promotion
- Years with current manager
- Years at company
- Job level

To test our model, we selected 80% of our data set to be training data set and remaining 20% to be our testing data sets. Our test dataset consisted 87% of “No” and 13% of “Yes” and we have achieved efficiency of 87.5 % with our predictive model.