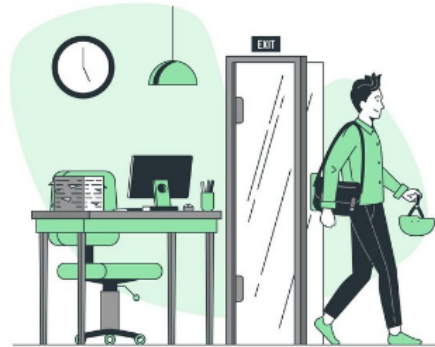


EMPLOYEE ATTRITION

How to reduce it with HR Analytics



INST 627 - Data Analytics for Information Professionals - Fall 2022

Employee Attrition

Team 6

Angela Tseng - atseng@umd.edu

Harshitha Ramachandra - harshi07@umd.edu

Sakshi Patil - spatil1@umd.edu

Srikanth Parvathala - psrikant@umd.edu

Under the guidance of Professor Christopher Antoun

CONTENTS

I. Introduction	2
A. Research question	2
B. Justification	2
C. Importance	2
II. Methods and Data	2
A. Data Source	2
B. Analysis strategy	3
III. Results	3
A. Findings of main variables	3
B. Further Analysis of other Independent variables	4
IV. Conclusion	7
A. Summary	7
B. Limitations	7
C. Practical Implications	7
D. Recommendations	7
V. References	7
VI. Appendix	8
Figure 1: Percentage of Attrition across the levels (yes and no)	8
Figure 2: Count of Attrition across levels (yes and no)	8
Figure 3 : Histogram for “MonthlyIncome”	9
Figure 4 :Exploring Data to find relation between Income and Attrition	9
Figure 5: Logistic Regression Model results for “MonthlyIncome”	10
Figure 6: Logistic Regression Model for “Monthly Income”, “JobSatisfaction” and “WorkLifebalance”	10

I. Introduction

A. Research question

Does monthly income have an impact on employee job attrition?

B. Justification

In early 2021, many of the employees voluntarily resigned from their jobs which is termed as **“The Great Resignation”**. According to Forbes, the attrition rate in the last year is nearly 23-25% which is a concerning number for organizations.

Attrition is an inevitable part of any business. Employee attrition occurs when the size of the workforce diminishes over time due to unavoidable factors such as employee resignation for personal or professional reasons.

We were curious to consider and perform statistical analysis for the most common type of attrition, voluntary attrition where employees decide to quit their jobs. Our research question will help to find whether the hike in attrition rate is because of the “Monthly Income”. For further analysis, we considered other potential factors like “Work life balance” and “Job satisfaction”. This will help organizations to focus on the areas which are causing employees to give up their jobs and take measures to improve those factors and keep their talent.

C. Importance

Having a clear view on the employee attrition rate is essential to understand where the employer stands in terms of candidate retention. If the employee attrition rate is high, for instance, it can mean that the organization is not providing enough benefits or the best work environment to keep the top-performing employees.

By measuring and analyzing the attrition rate, organizations can identify the problems to be attended to make sure to keep the employees from leaving.

II. Methods and Data


A. Data Source

Our dataset is from [Kaggle](#), **“IBM HR Analytics Employee Attrition and Performance”**. It consists of **35** different attributes from **1470** employees. This is a fictional dataset

created by IBM data scientists. There are many significant categorical variables such as Job Satisfaction, Work Life Balance, Environment Satisfaction, Job involvement, Education, Performance Rating which can be instrumental in predicting Job Attrition. There are four levels present in each of the variables, low, medium, high and very high which are encoded as 1,2,3 and 4 respectively in the final dataset. Monthly income for each employee is a numeric continuous variable. Variables like Attrition and Gender are categorical with only two levels. There are no missing values, duplicate values present in our dataset, hence no further data cleaning is required.

B. Analysis Strategy

Unit of analysis is “individuals” (employees).

The dependent variable is “Attrition” (has two levels  “Yes” and “No”). It is an ordinal scaled categorical variable.

The Independent variable is “Monthly Income”. It is numeric and discrete.

The **mean is 6503** and the **median is 4919**. Most of the employees are paid below the mean and the median of Monthly Income, which is normal in most of the organizations.

It is observed from [Figure 3](#) that “Monthly Income” is positively skewed,

Based on the variables considered, **Logistic regression** would be the best fit model for the statistical analysis.

Other independent variables considered are “JobSatisfaction” and “WorkLifeBalance” which are ordinal-scaled categorical variables.

Initially pair wise **Chi-Square test** is implemented to analyse and for better prediction **Logistic Regression** is used.

III. Results

A. Findings of main variables

Null Hypothesis: Our null hypothesis states that all coefficients in the model are equal to zero. In other words, the predictor variable doesn't have a statistically significant relationship with the response variable.

Ho: $b_1=0$. There is no significant relationship between Monthly Income and Attrition.

Alternative Hypothesis: The alternative hypothesis states that not every coefficient is simultaneously equal to zero.

Ha : $b_1 \neq 0$. There is a significant relationship between Monthly Income and Attrition.

Logistic regression: `fit <- glm(df$Attrition ~ df$MonthlyIncome, family=binomial())`

	Estimate	Std.Error	z value	Pr(> z)
(Intercept)	-9.291e-01	1.292e-01	-7.191	6.43e-13***
df\$MonthlyIncome	-1.271e-04	2.162e-05	-5.879	4.12e-09***

Table 1: Logistic Regression output for MonthlyIncome

The logistic regression equation is,

$$\ln\left(\frac{\text{prob}(\text{event})}{1 - \text{prob}(\text{event})}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

ln = -0.929 + 0.000127 X1, where X1 is "MonthlyIncome"

Interpretation of results

- The p-value associated with the coefficient obtained is **4.12e-09** which is less than the significance level(0.05), hence based on our p-value **we reject the null hypothesis**.
- This implies that monthly income has a significant effect on Job Attrition.
- Each one unit change in monthly income will decrease the log odds of determining attrition by 0.000127.

B. Further Analysis of other Independent variables

Chi Square Test of Independence for "JobSatisfaction" and "Attrition"


Null hypothesis (Ho): There is no relationship between Job Satisfaction and Attrition

Alternative hypothesis (Ha): There is a relationship between Job Satisfaction and Attrition

	No	Yes
1 (Low)	0.15170068	0.04489796
2 (Medium)	0.15918367	0.03129252
3 (High)	0.25102041	0.04965986
4 (VeryHigh)	0.27687075	0.03537415

Table 2: Contingency table (JobSatisfaction and Attrition)

Output: X-squared = 17.505, df = 3, **p-value = 0.0005563**

Based on the p-value obtained we reject the null hypothesis and conclude there is a significant relationship between "JobSatisfaction" and "Attrition". 

Chi Square Test of Independence for "WorkLifeBalance" and "Attrition"


Ho: There is no relationship between WorkLifeBalance and Attrition

Ha: There is a relationship between WorkLifeBalance and Attrition

	No	Yes
1 (Low)	0.03741497	0.01700680
2 (Medium)	0.19455782	0.03945578
3 (High)	0.52108844	0.08639456
4 (VeryHigh)	0.08571429	0.01836735

Table 3: Contingency table (WorkLifeBalance and Attrition)

Output: X-squared = 16.325, df = 3, **p-value = 0.0009726**

Based on the p-value obtained we reject the null hypothesis and conclude that there is a significant relationship between "WorkLifebalance" and "Attrition" 

Logistic Regression for “MonthlyIncome”, “WorkLifeBalance” and “JobSatisfaction”

```
fit1 <- glm(formula = df$Attrition ~ df$MonthlyIncome + df$WorkLifeBalance +  
df$JobSatisfaction, family = binomial())
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.769e-01	3.026e-01	1.245	0.21300
df\$MonthlyIncome	-1.280e-04	2.176e-05	-5.884	4.01e-09 ***
df\$WorkLifeBalance2	-7.615e-01	2.896e-01	-2.630	0.00855 **
df\$WorkLifeBalance3	-9.994e-01	2.678e-01	-3.732	0.00019 ***
df\$WorkLifeBalance4	-6.967e-01	3.307e-01	-2.107	0.03511 *
df\$JobSatisfaction2	-4.535e-01	2.191e-01	-2.070	0.03844 *
df\$JobSatisfaction3	-4.283e-01	1.946e-01	-2.201	0.02771 *
df\$JobSatisfaction4	-8.910e-01	2.080e-01	-4.284	1.83e-05 ***


Table 4: Logistic Regression output for Monthly Income, WorkLife Balance and Job Satisfaction

Interpretation of results

From the below result, it is observed that all the variables have a p-value less than .05. Hence we reject the null hypothesis that the coefficient is equal to zero. **Monthly Income, Work-life balance and Job satisfaction have a significant impact on attrition.**

IV. Conclusion

A. Summary

- Monthly income has a significant impact on Attrition.
- Higher Job Satisfaction lowers the attrition rate, similarly higher Work-life balance lowers the attrition rate. 

B. Limitations

- Dataset doesn't consist of details of the company that employees joined after leaving their current job.
- Fictional data set created by IBM data scientists.
- This data will help us further analyze outsourcing or poaching which is also one of the main factors for attrition in IT companies.

C. Practical Implications

Because of the hike in attrition rate, there is a shortage of talent for professional roles. Countries like the US, Western Europe (including the UK and Ireland) and India have a talent shortage of 45%, 43% and 39% respectively. This analysis will help organizations which are facing high attrition rates to understand the reason for their employees leaving and improve in those areas. So that they won't lose their talent and spend more money and time in preparing a new employee for that particular experienced role and avoid talent shortage.

D. Recommendations

For future analysis one can consider the bottom-up approach, with other attributes from the dataset as independent variables to determine how the probability of attrition varies.

V. References

[Spiking Attrition Impact - Forbes](#)

VI. Appendix

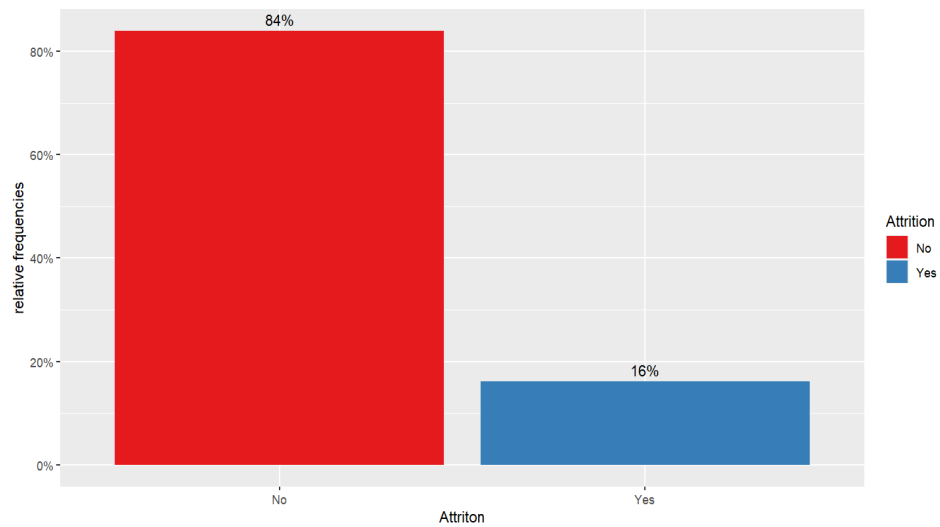


Figure 1: Percentage of Attrition across the levels (yes and no)

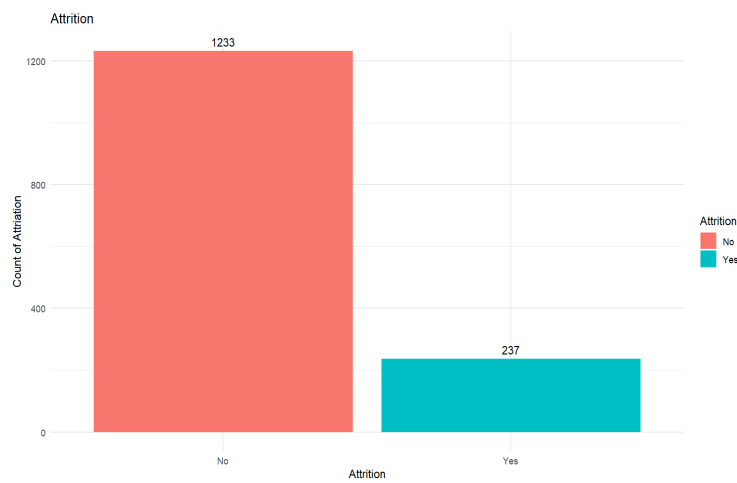


Figure 2: Count of Attrition across levels (yes and no)

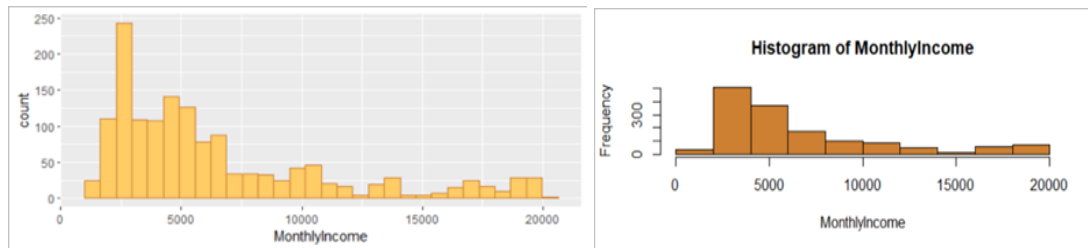


Figure 3: Histogram for “MonthlyIncome”

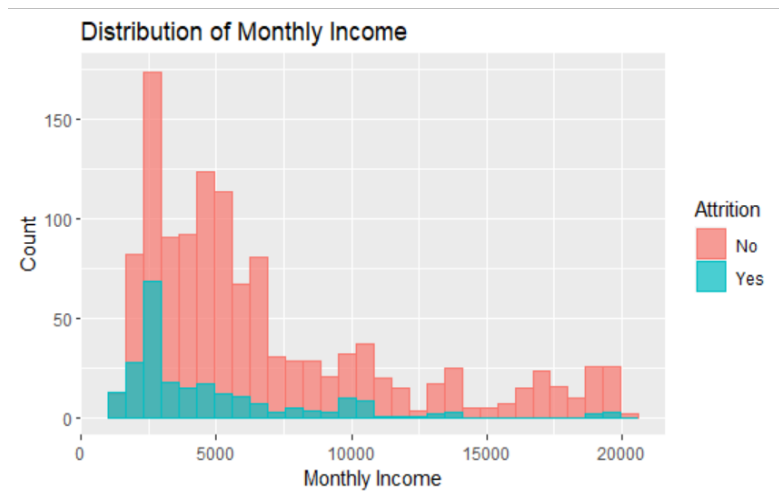


Figure 4: Exploring Data to find relation between Income and Attrition

```

> df$Attrition = as.factor(df$Attrition)
> #Logistic Regression for one variable
> fit <- glm(df$Attrition ~ df$MonthlyIncome, family=binomial())
> summary(fit)

Call:
glm(formula = df$Attrition ~ df$MonthlyIncome, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7704  -0.6646  -0.5811  -0.3430   2.6399

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.291e-01  1.292e-01  -7.191 6.43e-13 ***
df$MonthlyIncome -1.271e-04  2.162e-05  -5.879 4.12e-09 ***
---
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: 1253.1  on 1468  degrees of freedom
AIC: 1257.1

Number of Fisher Scoring iterations: 5

```

Figure 5: Logistic Regression Model results for “MonthlyIncome”

```

> #regression model for 3 variables
> df$JobSatisfaction = as.factor(df$JobSatisfaction)
> df$WorkLifeBalance = as.factor(df$WorkLifeBalance)
> fit1 <- glm(df$Attrition ~ df$MonthlyIncome + df$WorkLifeBalance + df$JobSatisfaction, family=binomial())
> summary(fit1)

Call:
glm(formula = df$Attrition ~ df$MonthlyIncome + df$WorkLifeBalance +
    df$JobSatisfaction, family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1771  -0.6537  -0.5268  -0.3302   2.8536

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.769e-01  3.026e-01   1.245  0.21300
df$MonthlyIncome -1.280e-04  2.176e-05  -5.884 4.01e-09 ***
df$WorkLifeBalance2 -7.615e-01  2.896e-01  -2.630  0.00855 **
df$WorkLifeBalance3 -9.994e-01  2.678e-01  -3.732  0.00019 ***
df$WorkLifeBalance4 -6.967e-01  3.307e-01  -2.107  0.03511 *
df$JobSatisfaction2 -4.535e-01  2.191e-01  -2.070  0.03844 *
df$JobSatisfaction3 -4.283e-01  1.946e-01  -2.201  0.02771 *
df$JobSatisfaction4 -8.910e-01  2.080e-01  -4.284 1.83e-05 ***
---
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: 1221.3  on 1462  degrees of freedom
AIC: 1237.3

Number of Fisher Scoring iterations: 5

```

Figure 6: Logistic Regression Model for “Monthly Income”, “JobSatisfaction” and “WorkLifeBalance”