Employee Attrition Analysis

- By *Team 7*

Abdul Shaik	ashaik@umd.edu
Adhira Ranjithkumar	adhira@umd.edu
Asmita Samanta	asamanta@umd.edu
Mingxi Xu	mercyxmx@umd.edu

COURSE - INST627-IM01: Data Analytics for Information Professionals

SEMESTER - Fall 2022

Table of Contents

1. Introduction

Employee attrition is one of the main factors that can demolish even the most stable organizations in a short period of time. Employers spend considerable effort and money training new workers to fill open roles. It is in a firm's best interest to keep the attrition rate under control. So, understanding the motivation behind employee attrition is the critical first step in understanding and properly addressing the problem.

We use healthcare industry employee data to compare & reveal the typical causes of employee exit. Factors examined include work-life balance, environmental satisfaction, employee involvement, job level, gender, marital status, education, job satisfaction, age, & business travel.

1.1. Research Question

We bifurcated our research question to best evaluate the entire dataset -

a. Does poor work life balance result in higher attrition rate?
b. Does greater environment satisfaction result in lower attrition rate?
c. Is there a lower attrition rate for a higher job level?
d. Does age impact the attrition rate of employees depending upon their past travel records?
e. Do employees who are involved & satisfied in their job tend to switch companies faster?

2. Methods & data

2.1. Dataset

For our study, we use Watson IBM's employee data compiled for the year of 2021. It is a synthetic dataset from Kaggle that is updated annually, & the fields specifically are designed for the healthcare domain. It includes 1676 employee records & 37 fields (categorical & numerical) out of which two are created by us for our analysis.

2.2. Data Preprocessing

For us to feed the data to R for further processing, we converted the categorical values to numerical values. Below given table describes the conversion we facilitated in our dataset to make it usable for modelling.

Steps	Variable Name	Outcome/Predictor?	Type	Range	Conversion
1	Attrition	Outcome	Categorical	Yes/No	Yes: 1 No: 0
2	Business Travel	Predictor	Categorical	Non_Travel Travel_Rarely, Travel_Frequently	Non_Travel: 0 Travel_Rarely: 1 Travel_Frequently: 2
3	Gender	Predictor	Categorical	Male/Female	Male: 1 Female: 0
4	Marital Status	Predictor	Categorical	Single or Divorced/Married	Married: 1 Single/Divorced: 0
5	Age & Business Travel	Both Predictors	Categorical	Age (Less than or equal to 45/Greater than 45) + Business Travel (Travel Rarely/Travel Frequently	Travel_Rarely + (<=45): 1 Travel_Frequently + (<=45): 2 Travel_Rarely + (>45): 3 Travel_Frequently + (<=45): 4
6	Job Involvement & Job Satisfaction	Both Predictors	Categorical	Involvement (High/Low) + Satisfaction (High/Low)	High JI + High JS: 1 High JI + Low JS: 2 Low JI + High JS: 3 Low JI + Low JS: 4

Table 1. Conversion Criteria for different variables

2.3. Analysis Strategy

We fixed the dependent and potential independent variables –

- Dependent variable: Attrition Rate (Categorical - Yes/No)
- Independent variables:
    a. Age_Business Travel
    b. Environment Satisfaction
    c. Job Level

d. Job Involvement_Satisfaction
e. Work Life Balance

2.4. Methods

- **Chi-Square Test of Independence** (to decipher which variables affect attrition)
- **Logistic Regression** (to make predictions and find direction of affect)

3. Result

3.1. Chi-square test

We used the Chi-Square Test of independence to find out the relationship between all our independent variables and our dependent variable.

```
> chisq.test(Attrition,WorkLifeBalance)

        Pearson's Chi-squared test

data:  Attrition and WorkLifeBalance
X-squared = 25.063, df = 3, p-value = 1.498e-05

> chisq.test(Attrition,EnvironmentSatisfaction)

        Pearson's Chi-squared test

data:  Attrition and EnvironmentSatisfaction
X-squared = 23.315, df = 3, p-value = 3.471e-05

> chisq.test(Attrition,Age_BusinessTravel)

        Pearson's Chi-squared test

data:  Attrition and Age_BusinessTravel
X-squared = 33.667, df = 3, p-value = 2.329e-07

> chisq.test(Attrition,JobInvolvement_Satisfaction)

        Pearson's Chi-squared test

data:  Attrition and JobInvolvement_Satisfaction
X-squared = 44.56, df = 3, p-value = 1.147e-09

> chisq.test(Attrition,JobLevel)

        Pearson's Chi-squared test

data:  Attrition and JobLevel
X-squared = 116.5, df = 4, p-value < 2.2e-16
```

Image 1. R Studio Screenshot of Outcome of Chi Square of Independence

In our result, we can see that p-values for all our independent variables are lesser than our reference value of 0.05, which shows the significant relationship of these with

Attrition. This step gave us the confidence that all our chosen independent variables are in fact correct to be thought of influencing attrition rate.

3.2. Logistic Regression

At this stage we finalized our research question , independent variables and ran the logistic regression test.

```
Call:
glm(formula = as.factor(Attrition) ~ as.factor(WorkLifeBalance) +
    as.factor(EnvironmentSatisfaction) + as.factor(Age_BusinessTravel) +
    as.factor(JobInvolvement_Satisfaction) + as.factor(JobLevel),
    family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6884  -0.4923  -0.3156  -0.2013   3.0485

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -0.3522     0.3301  -1.067 0.285895
as.factor(WorkLifeBalance)2           -0.8020     0.3147  -2.548 0.010829 *
as.factor(WorkLifeBalance)3           -1.2390     0.2928  -4.232 2.32e-05 ***
as.factor(WorkLifeBalance)4           -0.9605     0.3737  -2.570 0.010158 *
as.factor(EnvironmentSatisfaction)2   -0.6094     0.2437  -2.500 0.012402 *
as.factor(EnvironmentSatisfaction)3   -0.9440     0.2252  -4.191 2.77e-05 ***
as.factor(EnvironmentSatisfaction)4   -0.9012     0.2247  -4.010 6.07e-05 ***
as.factor(Age_BusinessTravel)2         0.6897     0.1993   3.460 0.000540 ***
as.factor(Age_BusinessTravel)3        -0.6959     0.3344  -2.081 0.037437 *
as.factor(Age_BusinessTravel)4         0.4532     0.5268   0.860 0.389647
as.factor(JobInvolvement_Satisfaction)2  0.7060     0.2238   3.154 0.001609 **
as.factor(JobInvolvement_Satisfaction)3  1.1760     0.2251   5.224 1.75e-07 ***
as.factor(JobInvolvement_Satisfaction)4  1.5025     0.2456   6.117 9.55e-10 ***
as.factor(JobLevel)2                  -1.6584     0.2094  -7.919 2.39e-15 ***
as.factor(JobLevel)3                  -1.2934     0.2793  -4.631 3.63e-06 ***
as.factor(JobLevel)4                  -2.2172     0.6173  -3.592 0.000328 ***
as.factor(JobLevel)5                  -2.1546     0.7469  -2.885 0.003918 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.46  on 1675  degrees of freedom
Residual deviance:  999.27  on 1659  degrees of freedom
AIC: 1033.3

Number of Fisher Scoring iterations: 6
```

Image 2. R Studio Screenshot of Outcome of Logistic Regression

We defined the result of the logistic regression model built in the form of 2 tables:

- Moving from an initial rating, its direction of affect as well as how that movement causes the log of odds to change.
- Probability Outcome of Outcome Variable v/s Predictor Variables

Please Note: To measure significance level, we have assumed the conventional level which is if p-value < 0.05, then the affect is significant.

a. Work-Life Balance

Rating of 1 to 4 where :1 = Unhealthy and 4 = Healthy

Moving from a RATING	Affect	Direction of affect	Log of odds
1 to 2	Significant	Decrease	0.802
1 to 3	Significant	Decrease	1.23
1 to 4	Significant	Decrease	0.96

Table 2. Measure of affect between Work Life Balance v/s Attrition Rate

RATING	PROBABILITY
1	0.267
2	0.14
3	0.097
4	0.121

Table 3. Probability Outcome of Work Life Balance on Attrition Rate

b. Environment Satisfaction

Rating of 1 to 4 where :1 = Unhealthy and 4 = Healthy

Moving from a RATING	Affect	Direction of affect	Log of odds
1 to 2	Significant	Decrease	0.609
1 to 3	Significant	Decrease	0.944
1 to 4	Significant	Decrease	0.901

Table 4. Measure of affect between Environment Satisfaction v/s Attrition Rate

RATING	PROBABILITY
1	0.193
2	0.116
3	0.093

4	0.096

Table 5. Probability Outcome of Environment Satisfaction on Attrition Rate

c. Age + Business Travel

Rating of 1 to 4 where:
1 = <45 travelling rarely or not at all
2 = <45 travelling frequently
3 = 45+ travelling rarely or not at all

Moving from a RATING	Affect	Direction of affect	Log of odds
1 to 2	Significant	Increase	0.689
1 to 3	Significant	Decrease	0.695

Table 6. Measure of affect between Age & Business Travel v/s Attrition Rate

RATING	PROBABILITY
1	0.123
2	0.194
3	0.039

Table 7. Probability Outcome of Age + Business Travel on Attrition Rate

d. Job Involvement + Job Satisfaction

Rating of 1 to 4 where:
1 = High job involvement & high job satisfaction
2 = High job involvement & low job satisfaction
3 = Low job involvement & high job satisfaction
4 = Low job involvement & low job satisfaction

Moving from a RATING	Affect	Direction of affect	Log of odds
1 to 2	Significant	Increase	0.706
1 to 3	Significant	Increase	1.716
1 to 4	Significant	Increase	1.502

Table 8. Measure of affect between Job Involvement & Satisfaction v/s Attrition Rate

RATING	PROBABILITY
1	0.070
2	0.112
3	0.168
4	0.222

Table 9. Probability Outcome of Job Involvement & Satisfaction on Attrition Rate

e. Job Level

Rating of 1 to 4 where :1 = Lower and 5 = Higher

Moving from a RATING	Affect	Direction of affect	Log of odds
1 to 2	Significant	Decrease	1.65
1 to 3	Significant	Decrease	1.29
1 to 4	Significant	Decrease	2.21
1 to 5	Significant	Decrease	2.15

Table 10. Measure of affect between Job Level v/s Attrition Rate

RATING	PROBABILITY
1	0.228
2	0.056
3	0.073
4	0.024
5	0.024

Table 9. Probability Outcome of Job Level on Attrition Rate

4. Discussion/Conclusion

4.1. Summary

a. With an increase in work-life balance the employee attrition will decrease.
b. With an increase in Environment Satisfaction the probability of attrition reduces.
c. After checking for travel pattern for work across different age groups, we arrived on the conclusion that with frequent travel there is a high resultant attrition.
d. High levels of Job involvement and Job satisfaction when put together can exponentially reduce the attrition of employees.
e. Entry level jobs have the highest attrition rate. It then has an overall decreasing trend as the job roles become critical.

4.2. Practical Implications

a. With work from home, though efficiency has increased the work-life balance has suffered. But amidst that people do realize the importance of good well-being and quit .
b. Toxic environments can encourage people to leave because one cannot be productive if they feel comfortable in the environment.
c. Travelling always takes a toll on body, be it young or old. Hence any job role having frequent travel makes people leave the roles very quickly.
d. With an increase on awareness amongst today's generation related to career growth, higher level of job involvement and satisfaction is very important to retain employees irrespective of salary figures.
e. Since critical jobs ensure quicker and higher career and financial growth, people tend to leave very less at that stage compared to someone who has just joined.

4.3. Recommendations

a. Companies should have strict work hours to prevent overworking as well as employee activities once a while for relaxation.
b. Companies should have frequent in-person sessions with each employee or surveys to understand if they have an issue related to their environment, job roles etc.
c. Companies should more and more shift to hybrid culture and ensure proper transportation for employees.

d. Companies should focus on entry level people, train them, understand their goals and help them grow in the company.

4.4. Limitations

a. Since we worked on synthetic data, the results on real data might vary.
b. There may be other factors than can affect an employee attrition which we didn't come across and maybe talking to real employees would have helped us better
c. Every country has a different set of economic and social factors that influence workforce in the country