Employee Attrition Predictive Analysis and Data Visualization

Fraol Abebe, Data Analytics Engineering ‘23

George Mason University

Author Note

[Include any grant/funding information and a complete correspondence address.]

Contents

[1. INTRODUCTION 4](#_Toc103378649)

[2. DATASET 4](#_Toc103378651)

[3. PROBLEM STATEMENT 5](#_Toc103378654)

[4. ANALYSIS FINDINGS 5](#_Toc103378655)

[5. ANALYTICS AND ALGORITHM 8](#_Toc103378656)

[5.1 Multiple Regression Model 10](#_Toc103378657)

[5.2 Logistic Regression 11](#_Toc103378658)

[6. CONCLUSION 12](#_Toc103378659)

[7. Appendix 13](#_Toc103378660)

[7.1 References 13](#_Toc103378661)

[7.2 Code 14](#_Toc103378662)

# INTRODUCTION

## When we talk about human resources and workforce planning, attrition, and employee turnover are the topic of the conversation more often not. One of the most challenging aspects of human resource management employers still face is employment retention and understanding the underlying factors that cause work force to resign, without plans to fill or replace that vacant job position. When we definite attrition, it denotes to departure of employees from a company for any reason including voluntary resignation, termination, or retirement. This paper aims to identify answers as to why employees leave their place of employment and predicts which factors play the main role is leading employees’ attrition.

# DATASET

## For this project, we are using a dataset that was created by IBM for the purpose of human resource analytics. The dataset was recovered from Kaggle and was imported into R studio in a form of a CSV file. The dataset is a structured data with 1470 total rows and 35 columns.

Below are variables found within the dataset and the description of each variable. Some variables are not listed below due to the irrelevancy to the study.

* Age: Employee age starting at the age of 18
* Attrition: Yes or No
* Gender: Female or Male
* Business Travel: Frequency of employee travel for the job
* Department: Department employees work at
* Education: The field of education employee have
* EnviromentalSatisfaction: Environment of the company rated 1 to 5, 1 being bad 5 being good
* JobInvolvement: Level of job involvement rated 1 to 5
* HourlyRate: how much employees earn per hour
* Joblevel: level of employees job rated 1 to 5, 1 being the entry level, 5 being senior level
* JobRole: Employees Job Title
* JobSatisfaction: Employees job Satisfaction rated 1 to 4, 1 being not satisfied, 5 being very satisfied
* MaritalStatus: Employees Marital Status
* MonthlyIncome: How much employees per month including Overtime
* MonthlyRate: Fixed rate of how much employee get paid on monthly basis
* NumCompaniesWorked: Number of companies employees worked at prior to this
* OverTime: Yes or No, whether employees work overtime or not
* PercentSalaryHike: How many percent have the employee’s salary grown
* PerformanceRating: Employees performance Rate 1 to 5, 1 being not performing well, 5 being performing great
* RelationshipSatisfaction: Employee relationship satisfaction rated from 1 to 5
* TrainingTimesLastYear: How many times employees have been trained last year
* WorkLifeBalance: Work life balance rated 1 to 5
* YearsAtCompny: How many years employees stayed with current company
* YearsInCurrentRole: How many years employees stayed at current role
* YerasSinceLastPromotion: How many years since employees last promoted
* YearsWithCurrentManager: How many Years employees have stayed with their current manager

## The dataset was clean and has no missing values. In this step of data preparation, I examined the dataset to make sure that everything else is clear and that the observation values match the feature names appropriately*.* In the process of preparing the dataset for analysis, visualization and Machine learning, some of the of the columns were renamed action taken was renaming some of the variables were misspelled.

# PROBLEM STATEMENT

According to the Bureau of Labor Statistics report, in the 2021, employee the total turnover rate in united states is 57.3 %, whereas 25% employees leave their company voluntary. (*19 Employee Retention Statistics That Will Surprise You (2022 )*, 2022). The goal of this analysis is using the dataset retrieved from IBM, I plan to uncover reasons employee leave their company. Taking a closer look at the subsets of attrition data, will visualize and highlight interesting findings in the analysis. To clean and analyzing the data, Rstudio was used. The data processing

# ANALYSIS FINDINGS

Very interesting findings of the study is presented using visualizations.For the analysis, the data was explored using both numerical and non-numerical variables. To represent findings using the categorical features, bar charts and boxplots were utilized. Taking an initial look at the dataset, the dataset the analysis was broken down by gender demographic, department in the company and marital status of the employees. Within our dataset out of the total employees in the company, about 16 percent of employees have left the company voluntarily.

Taking a closer look at the income distribution of employees, the visual below will give us insight in the relationship between employee’s annual income and how long they stay at the company. This income density of employee’s annual income graphic compares employees that remained at the company with the ones that left. There is difference between employees that left in comparison to the ones that stayed, highlighting the distribution of employees that left the company is more concentrated on the lower end of annual income. The Scatter plot represents the relationship between income and how long they stay in the company. Based on scatter the plot, there is positive trend between years at a company and how much an employees’ earn. While the scatter plot highlights the expected relationship between the two features, the density plot on the right depicts how that may paly a factor into employees’ departure from a company.

Chart, histogram

Description automatically generated Chart, scatter chart

Description automatically generated

* 1. Gender Demographic Analysis

In the following visualization, the analysis dives into gender income disparity and if income variance in different sex is a contributing factor to their job satisfaction. We can observe from the plot there is a significant difference between men and women income growth. The observation also highlights, majority of men in the company are significantly more satisfied with heir job in comparison to woman at the company. Surprisingly, more men tend to leave the company looking at the plot and job satisfaction for woman is is not strongly related to the leave the company.

Chart, bar chart

Description automatically generated

* 1. Attrition By Overtime and Department

From a healthy work perspective, not a lot of individuals enjoy working overtime. The bottom graph explores monthly income percentile by department and answer the question of which department the company mostly needs to observe and make necessary changes in terms of individuals working a lot of hours. From the diagram, we can see that more employees at the research and development department put overtime in comparison to the rest. However, looking at their attrition distribution, we see an interesting finding that the sales department are more likely to leave the company.

Chart, box and whisker chart

Description automatically generated

which department has the highest turnover rate. The plot identifies the sales department to have high attrition. This answers the question of which department the company mostly needs to observe and make necessary changes. When we break down the Income percentile by department and overtime, the plot does support our expectation of employees with more overtime hours tend to leave the company.

# ANALYTICS AND ALGORITHM

The goals of this project were to see if we could predict the factors that lead to employee attrition based on some features withing our data. To achieve the study goals and predict these factors, Multiple Regression and Logistic Regression models used for analysis of the dataset.

Spearman Correlation is an analysis that measures the strength of association between two variables and the direction of the relationship.  The application of this correlation matrix is utilized to see features that were significantly correlated and drop some variables that deemed to be insignificant for the analysis. By looking at the correlation matrix, some variables were significantly correlated while some features had no relationship. For instance, MonthlyIncome is positively correlated with the JobLevel at .92, YearsWithCurrentManager is strongly correlated with YearsAtCompany at .82 as well and PerformanceRating is correlated with PercentSalaryHike at .62. Some features that have completely no relationships and was expecting to see some relationship were Age with PerformanceRating and WorkLifeBalance with any of the features.

Chart

Description automatically generated.

To further highlight the relationship between some of the features a correlation chart below has been plotted. The frequency of promotion can influence people to stay and see a positive trend in there. Number of companies worked at didn’t really have a correlation, but it shows more people tend to switch companies less frequently. Moreover, Age and years at company have a slightly positive trend showing as more people age, the tend to stay at a company longer.

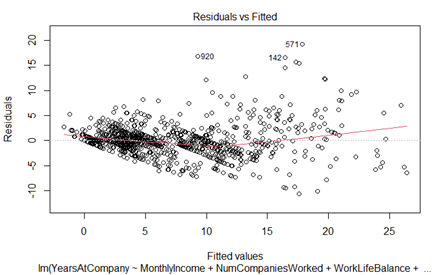
Chart, scatter chart

Description automatically generated

### Multiple Regression Model

To decide which features are the most important, the scatterplot analysis was conducted on some of the features. By randomly dividing the data into training set and the test set, we run a regression model to predict features that lead to employee attrition.

Graphical user interface, text

Description automatically generated 

Using some of the features that had relationship with years at company, were able to predict if the selected features are driving factors to employee turnover. The R-square shows 74% variance in our response variable which means 74% of the variance in our dependent variable could be explained using this model. Looking at the stastical outputs, which is the t-test value for each of the selected variables we can see Age, YearsSinceLastPromotion as well as YearsWithCurrManager are statistically significant at confidence level of 0.05. Our overall model could be undermined due to the risk of multicollinearity which is when some of the selected features are correlated (Weedmark). The scatter plot of residuals vs fitted plot of our multiple regression model is to how far the predicted values deviate from the fitted line. The closer the datapoints are to the fitted line, the more accurate our prediction is.

### Logistic Regression

Logistic Regression usually deals with the problem of how likely features are categorized into a group. This model is commonly used to predict the likelihood of an event occurring, and in our model, we use it to predict what features are more likely to drive employee out of the company. Unlike our multiple regression, the output of logistic regression is transformed with a logit function, which makes our response variable either 0 or 1. For this purpose we converted our attrition value into a binary value, changing a “Yes” to 1 and “No” to 0.

From the output below, we can see, people who answered yes to overtime are more likely to leave the company. Number of companies worked at signifies that employee that tend to switch more frequently are more likely to leave this current company. Our analysis also implies department of at company and role of an employee is a contributing factor to employee attrition

Text

Description automatically generated

# CONCLUSION

In conclusion, based on the machine learning model results, as well as our findings in our visualizations we can determine factors that lead to employee attrition, which some of these features align with the nature of human and the current job market. First, in the visualization analysis, we have identified Overtime and Income to be the major factors leading to employment departure. Although we have learned women tend to stay at the company regardless of their salary growth, we can conclude that there is a significant difference between the salary growth of men and women, and which leads men to be happier with their job.

We can also conclude, Annual income, frequency of promotion as well as number of companies worked indicate whether an employee will leave a company. Individuals who worked in 2 - 4 companies are less likely to leave, people who answered they work overtime have high attrition rate. This project could be improved using different machine learning models to increase the accuracy score.

Overall, this paper demonstrates applications of several analysis method, machine learning models, and the use of different visualization. Although, the analysis could narrow down the focus of the study, the paper answers some interesting questions and provided multiple answers with different perspective. Ggplot is heavily used for the visualization purpose. Several statistical packages such as dplyr, tidyr, and care were used when conducting analysis. The paper provides have a compelling output to readers supported with visualization and statistical output to understand factors that predict employee turnovers.

# Appendix

# References

*19 Employee Retention Statistics That Will Surprise you (2022 )*. (2022, January 6). Apollo Technical LLC. <https://www.apollotechnical.com/employee-retention-statistics/>

*CRAN Packages By Name*. (n.d.). Retrieved May 13, 2022, from <https://cran.r-project.org/web/packages/available_packages_by_name.html>

Vulpen, E. (2016, August 15). What Drives Employee Turnover? Part 2. *AIHR*. <https://www.aihr.com/blog/what-drives-employee-turnover/>

ZACH. (2021, June 7). How to Use facet\_wrap in R (With Examples). *Statology*. <https://www.statology.org/facet_wrap/>

Weedmark, D. (2018, March 13). *The Advantages & Disadvantages of a Multiple Regression Model*. Sciencing. Retrieved November 18, 2021, from <https://sciencing.com/advantages-disadvantages-multiple-regression-model-12070171.html>

*19 Employee Retention Statistics That Will Surprise you (2022 )*. (2022, January 6). Apollo Technical LLC. <https://www.apollotechnical.com/employee-retention-statistics/>

*What Is the Meaning of Attrition Used in HR?* (n.d.). Small Business - Chron.Com. Retrieved May 13, 2022, from <https://smallbusiness.chron.com/meaning-attrition-used-hr-61183.html>

# Code

The code for this project can be found using the following link. Due to the length of the, I am providing link to my code where I pushed all of my work.

* https://github.com/Fraolabebe/HR-Analytics-\_R/tree/main/HR-Analytics%20Project%20\_R