ISSN: 2634 - 8853

Journal of Engineering and Applied Sciences Technology

SCIENTIFIC Research and Community

Review Article

Open Access

Analyzing IBM HR Data: Employee Attrition and Performance Insights

Fatbardha Maloku* and Besnik Maloku

Master of Science in Business Analytics, Ageno School of Business, Golden Gate University, San Francisco, California 94105, USA

ABSTRACT

Employee turnover is often perceived as detrimental to organizational efficiency, but is this always the case? This research explores the multifaceted factors influencing employee attrition and examines whether lower turnover invariably leads to greater efficiency. Utilizing the "IBM HR Analytics Employee Attrition and Performance" dataset, which includes variables such as employee age, department, education level, job satisfaction, gender, job role, marital status, and overtime hours, we conduct a comprehensive descriptive analysis to predict employee retention. By understanding these factors, HR and management can make informed decisions to mitigate attrition. This study aims to identify novel strategies to reduce employee turnover, providing actionable insights and predictions to help IBM retain talent and maintain productivity and success.

***Corresponding author**

Fatbardha Maloku, Master of Science in Business Analytics, Ageno School of Business, Golden Gate University, San Francisco, California 94105, USA.

Received: August 09, 2024; **Accepted:** August 12, 2024; **Published:** August 25, 2024

Introduction

In general, it is considered that lower organizational turnover leads to greater efficiency, but is this always the case? Are there any other elements or reasons that have an impact on employee attrition, either directly or indirectly? In this study paper, the answers to these open questions will be revealed. The purpose of this research is to look into the factors that influence employee turnover. We'll then use the descriptive analysis' data to determine whether or not an employee will stay with the company. Employee attrition is the biggest difficulty for any company, yet HR or management may make prompt decisions to keep personnel. The same issue is happening within the IBM company also. The main attributes of the entire process are included within the "IBM HR Analytics Employee Attrition and Performance" dataset. This set contains variables such as the age of employees, the department they belong to, the education level, job satisfaction level, gender, job role, marital status, and the overtime worked hours. In this study, we'll investigate new approaches to decrease the attrition rate as well as give suggestion and predictions on how IBM can continue being productive and successful by retaining their talent.

Problem Statement

High employee turnover is often considered detrimental to organizational efficiency, but there is ambiguity regarding whether reducing turnover universally enhances organizational effectiveness. This research aims to investigate the complex factors influencing employee attrition and assess whether lower turnover consistently correlates with improved efficiency.

Business Problem Background

Employee turnover is a pervasive concern in organizational management, impacting productivity and stability. Traditional views suggest that minimizing turnover enhances operational efficiency by stabilizing workforce continuity and reducing recruitment costs. However, recent studies indicate nuances in this relationship, questioning whether low turnover always equates to optimal organizational performance. This study utilizes the "IBM HR Analytics Employee Attrition and Performance" dataset, encompassing variables such as employee demographics, job characteristics, and work-related factors. Through comprehensive descriptive analysis and predictive modeling, the research seeks to uncover underlying patterns and predictive indicators of employee retention. By gaining insights into these factors, HR professionals and management can implement targeted strategies to mitigate attrition, thereby fostering a stable and productive workforce environment at IBM.

Project Aim

The aim of this project is to analyze the factors influencing employee attrition using the "IBM HR Analytics Employee Attrition and Performance" dataset. By conducting a thorough examination of variables such as employee demographics, job characteristics, and work-related factors, the study seeks to identify

key determinants of turnover. Ultimately, the project aims to provide actionable insights and predictive models that can assist IBM in implementing effective strategies to reduce employee turnover, enhance retention rates, and sustain organizational productivity and success.

Data Collection

The dataset used in this analysis was sourced from Kaggle.com and comprises a total of 35 columns. For this study, our focus is on key variables believed to influence employee attrition rates. The following essential variables are considered:

- **Age:** Numeric variable indicating the age of employees.
- **Attrition:** Categorical variable with options "Yes" and "No".
- **Education:** Categorical variable with options including "Below College", "College", "Bachelor", "Master", and "Doctor".
- **Employee Satisfaction:** Categorical variable with options "Low", "Medium", "High", and "Very High".
- **Job Involvement:** Categorical variable describing the level of employee engagement, with options "Low", "Medium", "High", and "Very High".
- **Work Life Balance:** Categorical variable assessing work-life balance, categorized as "Bad", "Good", "Better", and "Best".
- **Performance Rating:** Categorical variable indicating performance level, with options "Low", "Good", "Excellent", and "Outstanding".

In addition to these variables, the dataset includes several others that will be explored to uncover factors strongly associated with employee attrition. This research aims to identify significant predictors and further investigate their relationships to inform strategies for reducing attrition rates and enhancing organizational retention practices.

Model Selection

During my analysis of each variable in the dataset, I focused on addressing the fundamental question: Is there a strategy to reduce employee attrition rates? In essence, I sought to uncover trends that may contribute to both employee attrition and performance. Any patterns identified are likely indicative of broader trends affecting organizations where employees choose to leave, rather than isolated factors specific to individual cases. It is crucial to acknowledge that historical data, while informative, does not guarantee future outcomes. These patterns may have evolved over time, but their persistence in the future is uncertain. Based on these considerations, the research proceeded with the following methodologies, detailed below. Initially, a descriptive analysis model was applied to Kaggle.com's historical dataset. This stage involved a thorough examination of each variable to identify key factors influencing employee attrition rates, providing insights for organizations aiming to retain their top talent over extended periods. Subsequently, a predictive analytics model was implemented to assess how findings from the descriptive study could forecast future employee attrition and performance rates within a company.

Solution Process

The initial phase of our analysis will involve conducting a descriptive analysis of the dataset. This includes checking for missing data, assessing the current data types of each variable, and generating summary statistics to gain a comprehensive understanding of the dataset. Visualizations will be utilized to better understand the distribution of each variable. Subsequently, we will implement predictive modeling to identify highly correlated factors that predict employee attrition rates within the company. Additionally, ANOVA tests will be conducted across multiple models to determine which model yields optimal results aligned with our research objectives.

Descriptive Analysis

The distribution of employee's age is shown below.

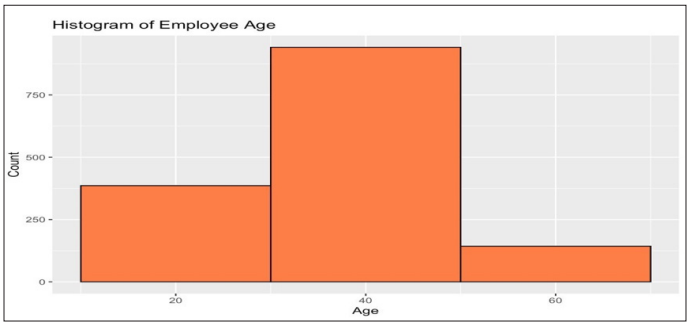

Figure 1: Distribution of Employees Age

The average age of IBM employees in this graph goes from thirty to fifty years old. We have a few individuals in their twenties as well as a few other individuals over the age of sixty.

The distribution of the number of employees per business travel sector is shown below. The bulk of employees, as seen on the graph below belong to "Travel Rarely" group. We have some individuals that belong to the "Travel Frequently" group. Lastly, we also have a group of individuals who do not travel as much as the previous two groups, and they belong the "Non-Traveler" group.

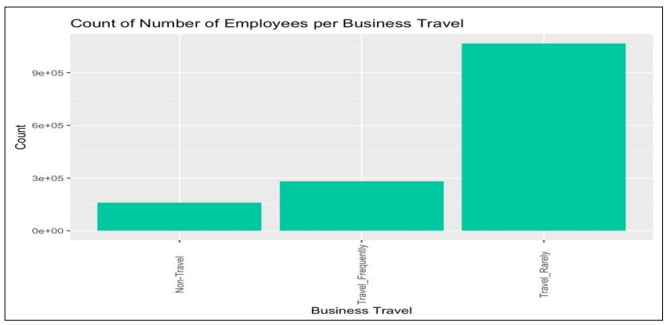

Figure 2: Distribution of the Number of Employees per Business Travel sector

The distribution of employees per department is shown below.

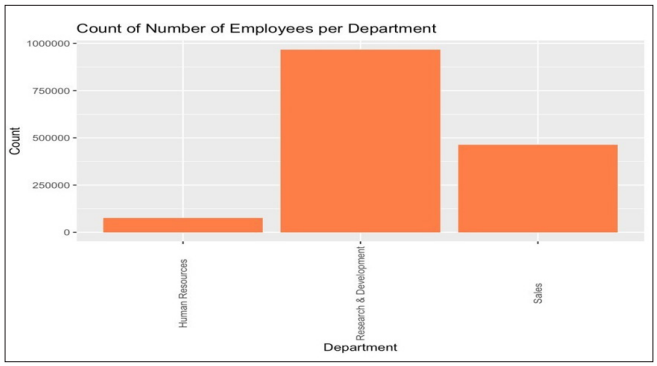

Figure 3: Distribution of the Number of Employees per Department

We examined the impact of the Gender variable on attrition. The graph below illustrates the distribution of attrition by gender and

employee count. From the graphic, it is evident that females with lower scores are more likely to leave their jobs. In contrast, males tend to remain in their jobs with slightly higher scores.

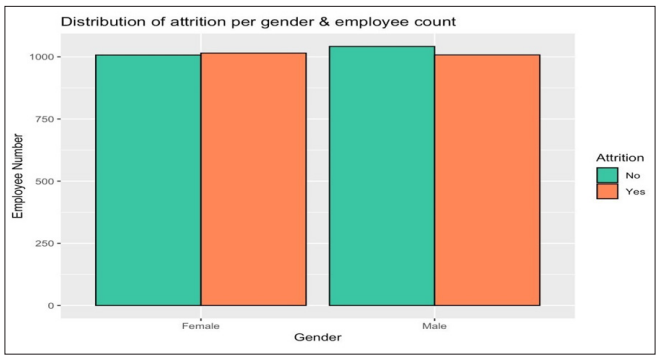

Figure 4: Distribution of Attrition per Gender and Employee Count

The graph below depicts the distribution of attrition rates across different employee job roles and the total count of employees in the organization. It illustrates that the "Yes" values for attrition are lower among research directors compared to the human resources role. This visualization indicates that research directors tend to exhibit lower turnover rates compared to other job titles within the organization.

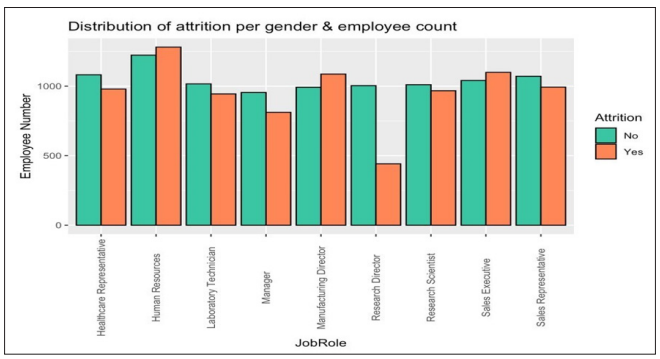

Figure 5: Distribution of the Attrition variable per Gender and Employee Count

Next, we'll look at the distribution of older vs. younger generations within organizations and compare the two groups' ages to see how that distribution appears.

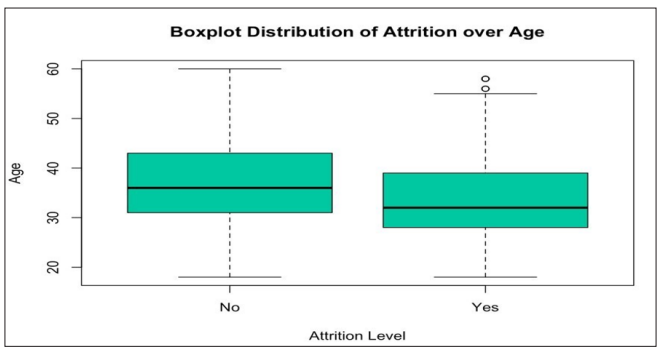

Figure 6: Box Plot Distribution of the Attrition Rate Over Age

People of younger ages are more willing to leave a job than those of older ages, as shown in the graph above. After that, we'll make a single-variable plot of the Age variable. The graph looks like this after graphing the values:

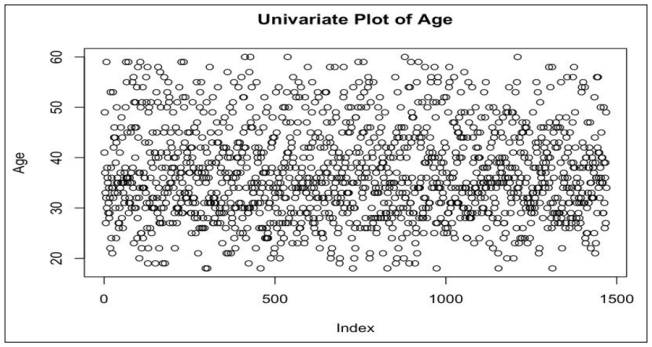

Figure 7: Univariate Plot of Age

Next, we have the Bivariate Plot of the performance rating variables and the Years in the Current Years of the employees. From the visualization we can see that the distribution of the variables is proportional.

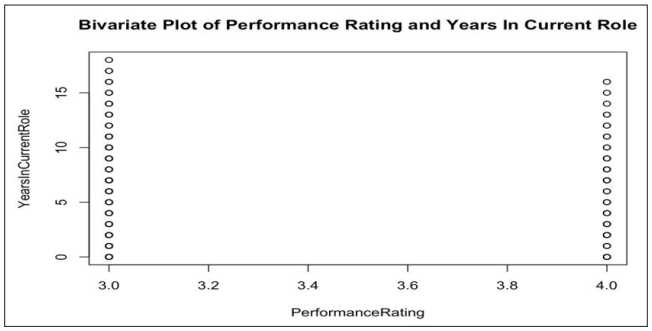

Figure 8: Bivariate Plot of Performance Rating Score and the Number of Years in the Current Role

We have generated a multivariate plot using the attrition and performance rating dataset. The graph below illustrates an example of such a plot featuring numerical variables. This includes Age, Employee Number, Environment Satisfaction, Monthly Income, Performance Rating, Work Life Balance, Years at Company, Years in Current Role, Years Since Last Promotion, Years with Current Manager, and Stock Option Level variables.

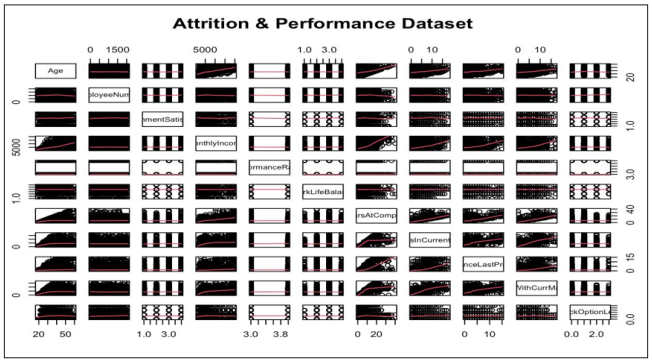

Figure 9: Multivariate Plot of the Attrition and Performance Score

Predictive Analysis

The predictive analysis will be carried out utilizing the two criteria listed below:

- Identify the correlation co-efficient values between variables.
- Create a regression model to predict the attrition rate based on the provided variables.

The correlation and regression model has been used to test the relationship between variables such as Age, Employee Number,

Environment Satisfaction, Monthly Income, Performance Rating, Work Life Balance, Years at Company, Years in Current Role, Years Since Last Promotion, Years with Current Manager, and Stock Option Level etc. By exploring these potential explanatory variables, we will know what variables help the attrition rate in an organization.

Heatmap

Below we have created a heatmap about the correlation between the attrition values and the other explanatory variables such as (Age, Total Number of Employees, Environment Satisfaction, Monthly Income, Work Life Balance, Years at the Company, Years Since Last Promotion, Stock Level options)

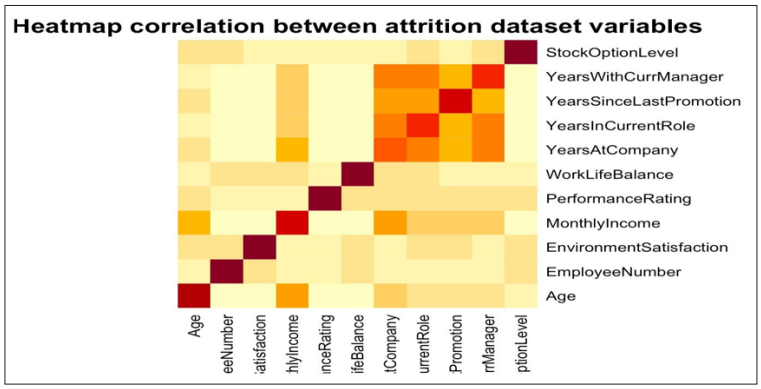

Figure 10: Heatmap Between the Explanatory Variables

As demonstrated in the heatmap above, the attrition rate is strongly connected with the variables like years with current manager, years since last promotion, years in current role and a total number of years at company variables. We also see that the attrition rate is less so connected with the age variable and the employee satisfaction variable. We may also observe that the variables with lighter colors in the heatmap are less highly associated with one another than those with deeper.

Attrition Variable Correlation

We produced a correlation chart with numbers to determine the real values of the correlation between the explanatory components, as shown below. The correlation coefficient between different variables varies a lot, as shown in the graph below. The Years With Current Manager variable has a correlation coefficient of 0.77, and the Years In Current Role variable has a correlation coefficient of 0.76. Following those two factors is Years Since Last Promotion, which has a 0.62 association coefficient. Stock Option Level and Work Life Balance variables are not substantially connected, as shown in the graph below.

Figure 11: Correlation Numeric Values Between Explanatory Variables

Regression

The Regression Model was chosen for the following reasons:

- The procedure will assist us in identifying the most strongly connected possible variables that will have an impact on IBM employee attrition rates.
- This method will help us focus our efforts on areas that will boost the likelihood of retaining talented employees within a company.
- It will have an effective and may save a company from losing productivity if employee retention is maintained.

Model 1: Attrition Rate ~ Age

```
predicted_attrition = attrition$Attrition
age = attrition$Age
model1 = lm(predicted_attrition~age)
```

Getting the summary for the regression model of model1:

```
summary(model1)
```

```
##
## Call:
## lm(formula = predicted_attrition ~ age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28254 -0.19279 -0.14791 -0.07739  0.97389
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.397938   0.039466  10.083  < 2e-16 ***
## age         -0.006411   0.001038  -6.179 8.36e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3633 on 1468 degrees of freedom
## Multiple R-squared:  0.02535,    Adjusted R-squared:  0.02468
## F-statistic: 38.18 on 1 and 1468 DF,  p-value: 8.356e-10
```

Figure 12: Model 1: Attrition Rate ~ Age

Model 2: Attrition ~ Years At Company

```
predicted_attrition = attrition$Attrition
years_at_company = attrition$YearsAtCompany
model2 = lm(predicted_attrition~years_at_company)
```

Getting the summary for the regression model of model2:

```
summary(model2)
```

```
##
## Call:
## lm(formula = predicted_attrition ~ years_at_company)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21778 -0.18550 -0.16129 -0.09673  1.10500
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.217777   0.014453  15.068  < 2e-16 ***
## years_at_company -0.008069   0.001553  -5.196 2.32e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3646 on 1468 degrees of freedom
## Multiple R-squared:  0.01806,    Adjusted R-squared:  0.01739
## F-statistic:    27 on 1 and 1468 DF,  p-value: 2.319e-07
```

Figure 13: Model 2: Attrition ~ YearsAtCompany

Model 3: Attrition ~ Job Satisfaction

```
predicted_attrition = attrition$Attrition
job_satisfaction = attrition$JobSatisfaction
model3 = lm(predicted_attrition~job_satisfaction)
```

Getting the summary for the regression model of model3:

```
summary(model3)
```

```
##
## Call:
## lm(formula = predicted_attrition ~ job_satisfaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2284 -0.1652 -0.1643 -0.1133  0.8867
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                0.1651584  0.0174108   9.486   <2e-16 ***
## job_satisfactionLow        0.0632153  0.0276903   2.283   0.0226 *
## job_satisfactionMedium    -0.0008727  0.0279581  -0.031   0.9751
## job_satisfactionVery High -0.0518686  0.0243935  -2.126   0.0336 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.366 on 1466 degrees of freedom
## Multiple R-squared:  0.01191,    Adjusted R-squared:  0.009886
## F-statistic: 5.889 on 3 and 1466 DF,  p-value: 0.0005399
```

Figure 14: Model 3: Attrition ~ Job Satisfaction

Model 4: Attrition ~ YearsWithCurrManager

```
predicted_attrition = attrition$Attrition
years_with_curr_manager = attrition$YearsWithCurrManager
model4 = lm(predicted_attrition~years_with_curr_manager)
```

Getting the summary for the regression model of model4:

```
summary(model4)
```

```
##
## Call:
## lm(formula = predicted_attrition ~ years_with_curr_manager)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22762 -0.19541 -0.16321 -0.08269  0.99783
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.227622   0.014490  15.709  < 2e-16 ***
## years_with_curr_manager -0.016104   0.002658  -6.059 1.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3635 on 1468 degrees of freedom
## Multiple R-squared:  0.0244,    Adjusted R-squared:  0.02373
## F-statistic: 36.71 on 1 and 1468 DF,  p-value: 1.737e-09
```

Figure 14: Model 4: Attrition ~ YearsWithCurrManager

Model 5: Attrition ~ YearsInCurrentRole + YearsWithCurrManager + YearsSinceLastPromotion + YearsAtCompany

```
predicted_attrition = attrition$Attrition
model5 = lm(predicted_attrition~YearsInCurrentRole+YearsWithCurrManager+YearsSinceLastPromotion+YearsAtCompany)
```

Getting the summary for the regression model of model5:

```
summary(model5)
```

```
##
## Call:
## lm(formula = predicted_attrition ~ YearsInCurrentRole + YearsWithCurrManager +
##     YearsSinceLastPromotion + YearsAtCompany)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25885 -0.20378 -0.15469 -0.05654  1.02337
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.242727   0.015286  15.879  < 2e-16 ***
## YearsInCurrentRole       -0.012910   0.004254  -3.035  0.00245 **
## YearsWithCurrManager     -0.010037   0.004361  -2.302  0.02150 *
## YearsSinceLastPromotion   0.011734   0.003773   3.110  0.00191 **
## YearsAtCompany           -0.001597   0.002878  -0.555  0.57916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3617 on 1465 degrees of freedom
## Multiple R-squared:  0.0358,    Adjusted R-squared:  0.03317
## F-statistic:  13.6 on 4 and 1465 DF,  p-value: 6.874e-11
```

Figure 14: Model 5: Attrition ~ YearsInCurrentRole + YearsWithCurrManager + YearsSinceLastPromotion + YearsAtCompany

Residuals vs Fitted

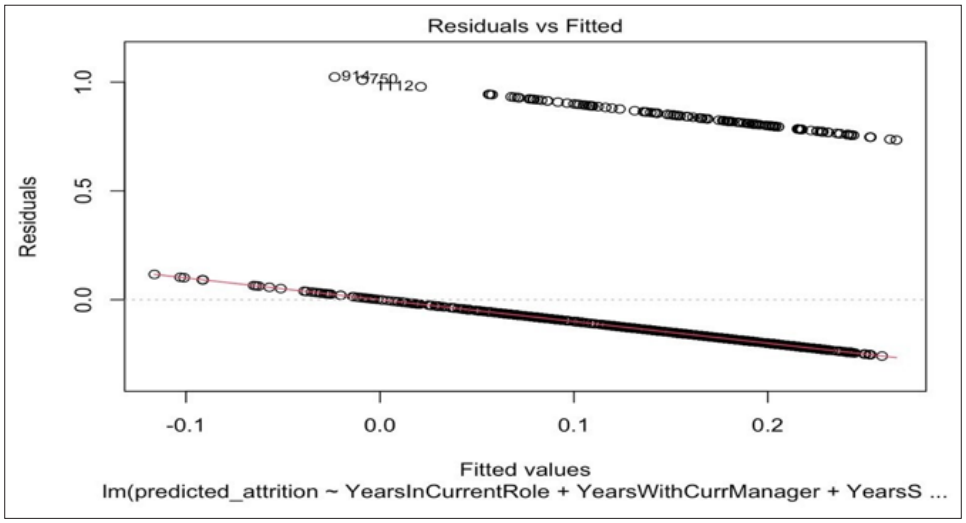

Figure 15: Regression Model of Residuals vs Fitted

Regression Model of Theoretical Quantiles

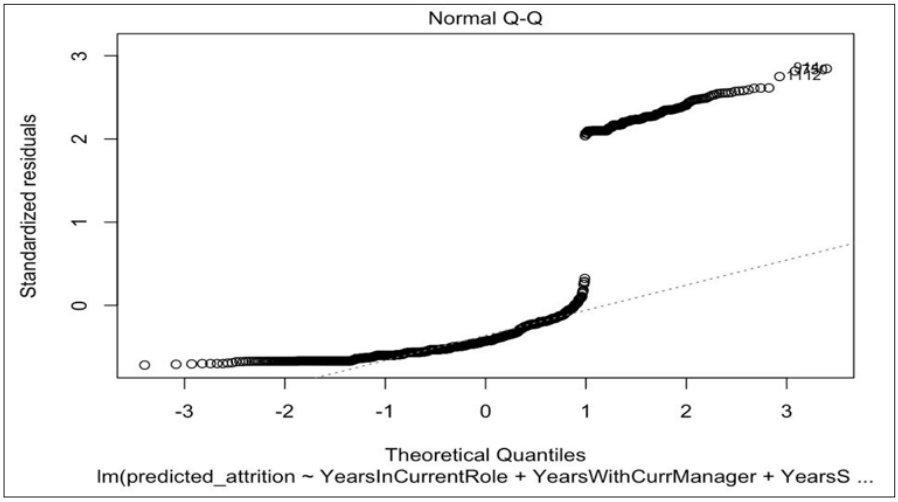

Figure 16: Regression Model of Theoretical Quantiles

Regression Model of Scale-Location

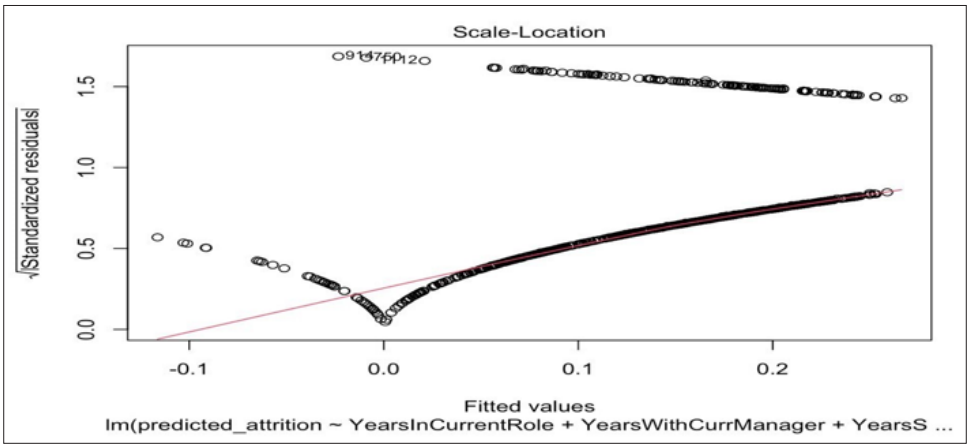

Figure 17: Regression Model of Scale-Location

Regression Model of Residuals vs Leverage

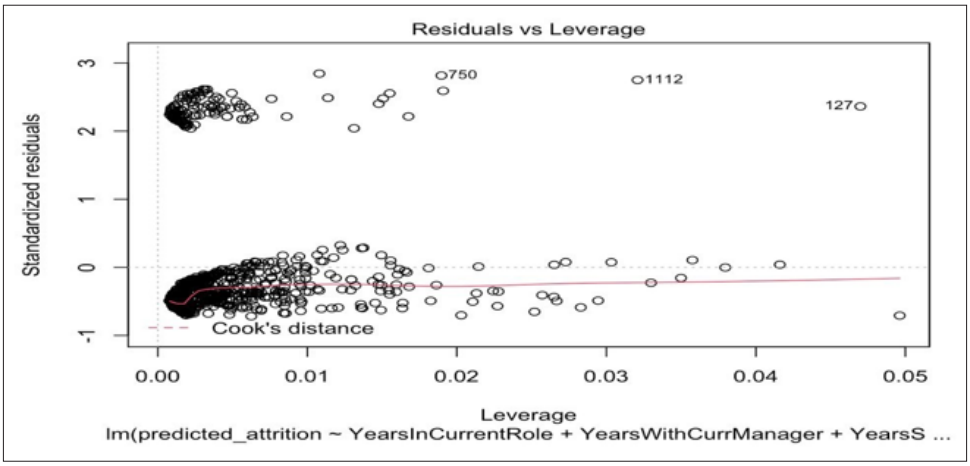

Figure 18: Regression Model of Residuals vs Leverage

ANOVA for Model 1 & Model 2

```
anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: predicted_attrition ~ age
## Model 2: predicted_attrition ~ years_at_company
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1   1468 193.75
## 2   1468 195.20  0   -1.4482
```

Figure 19: ANOVA for Model 1 & Model 2

ANOVA for Model 3 & Model 4

Run anova on (model3 and model4):

```
anova(model3,model4)
```

```
## Analysis of Variance Table
##
## Model 1: predicted_attrition ~ job_satisfaction
## Model 2: predicted_attrition ~ years_with_curr_manager
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1   1466 196.42
## 2   1468 193.94 -2    2.4829
```

Figure 20: ANOVA for Model 3 & Model 4

Model Results

The regression models that were ran on the dataset yielded the following findings. The model results in each of these five models will be summarized, and we'll observe how these discoveries affect IBM's total attrition rate.

Model 1: Attrition Rate ~ Age

After we run the regression model of user rating score against the attrition rate against age, we got an R-squared value of 0.024. From this model we understand that the model is correct for approximately 2%.

Model 2: Attrition ~ YearsAtCompany

We then run the regression model of attrition against the YearsAtCompany variable, and we got an R-squared value of 0.017. From this model we understand that the R-squared value is approximately 1% correct for our use case.

Model 3: Attrition ~ JobSatisfaction

Lastly, we run the regression model of Attrition against the JobSatisfaction variable, we got a p-value of 0.0005 and an R-squared value of 0.009. From this model we understand that the p-value is within our limit, however, is very small. The same thing happens with our R-squared value also, which indicates that our model is 0.9% right. This indicates that the model needs more work, therefore other models have been considered.

Model 4: Attrition ~ YearsWithCurrManager

Next, we run the regression model of attrition against the YearsWithCurrManager variable, where we got a very low p-value and an R-squared value of 0.023 From this model we understand that the R-squared value indicates that our model is ~2%.

Model 5: Attrition ~ YearsInCurrentRole + YearsWithCurrManager + YearsSinceLastPromotion + YearsAtCompany

The last model, we run is the regression model of Attrition against the YearsInCurrentRole, YearsWithCurrManager, YearsSinceLastPromotion and YearstAtCompany variable. Since all these values where highly correlated, we decided to investigate the regression model and see what we get for the p-value and the R-squared value. In this model, we got a very low p-value and an R-squared value of 0.033. From this model we understand that the R-squared value indicates that our model is approximately 3% correct understand that the R-squared value indicates that our model is approximately 3% correct.

Results Interpretation

In the previous phase of the analysis, we created a correlation and regression model. YearsWithCurrManager is the most highly associated value, according to the correlation model. We also discovered that the YearsWithCurrManager variable, as well as YearsSinceLastPromotion,YearsInCurrentRole,YearsAtCompay, are all closely linked. Then I ran the regression model and found that the combined variables

Conclusion

In summary, we utilized strategic tools to assess the current market landscape and potential factors influencing future employee attrition rates using the dataset sourced from Kaggle.com. Through our analysis, several key insights into the determinants of attrition have emerged. Employee turnover poses a significant challenge for organizations, but proactive decisions by HR and management can mitigate this issue. This holds true for IBM as well. Our dataset analysis allowed us to delve into explanatory variables and their impact on attrition rates. Notably, YearsWithCurrManager emerged as the most influential variable affecting attrition, based on our investigation. Additionally, YearsInCurrentRole exhibited the second highest R-squared value at 0.02, followed by YearsSinceLastPromotion at 0.01. When combining YearsInCurrentRole, YearsWithCurrManager, YearsSinceLastPromotion, and YearsAtCompany, the cumulative R-squared of 0.03 underscores the multifaceted nature of factors influencing an employee's decision to remain with or depart from the company [1].

References

1. Pavansubhash (2017) IBM HR Analytics Employee Attrition & Performance. Kaggle https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset.