

# **Predicting Employee Attrition**

**Fatima Soytemiz**

**April 1, 2021**

# Contents

|  |    |
|--|----|
| 1. Introduction .....                        | 3  |
| 2. Data Acquisition .....                    | 3  |
| 3. Data Cleaning .....                       | 3  |
| 4. Exploratory Analysis .....                | 4  |
| 4.1. Target Variable .....                   | 4  |
| 4.2. Features .....                          | 4  |
| 5. Data Pre-Processing .....                 | 16 |
| 6. Modeling Development and Evaluation ..... | 18 |
| 7. Conclusion .....                          | 21 |

# 1 Introduction

Attrition is terminology in human resources that refers to the employees leaving the company. Attrition is measured by the number of employees leaving the company either by voluntarily resigning or being laid off by the company. In general, high attrition is problematic for companies and causes huge losses. There are many reasons for which the employees leave the company, such as salary dissatisfaction, no career growth, etc. The loss is not only in terms of the money but also the company sometimes loses skilled employees who are the most valuable assets to the company (Morrison, 2014). If the company can predict employee attrition (employees who are going to leave the company) in the near future, they can also work on retention beforehand and avoid the loss of valuable employees. The prediction of attrition and retention is part of HR Analytics: how to retain talent and avoid attrition in the organizations.

In this project, I analyze the IBM Employee Attrition dataset to find the main factors why employees choose to leave and to help the company to predict whether or not a certain employee will leave the company by utilizing machine learning models.

## 2 Data Acquisition

I used IBM HR Analytics Employee Attrition & Performance dataset from Kaggle. The data is a fictional data set created by IBM data scientists. The data can be reached from this [link](#).

The data has 35 columns and 1470 observations and contains numeric and categorical data types. Each row has various attributes from each unique employee, such as age, gender, job role, educational field, marital status, monthly income, etc.

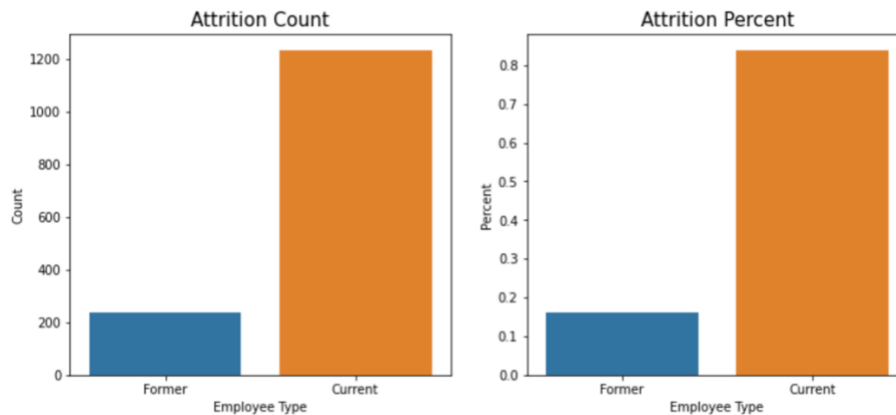
## 3 Data Cleaning

There are no missing values. Also, there are no duplicate values in the dataset. "Over18", "EmployeeCount," and "StandardHours" variables are deleted since they contain only one value, and there is no variation in these features. "EmployeeNumber" is also deleted because it has all unique numbers like ID (unique identifier), and it does not generate any value for the model. I count the number of employees who left the company vs. those who stayed in the company.

## 4 Exploratory Data Analysis

### 4.1 Target Variable (Attrition)

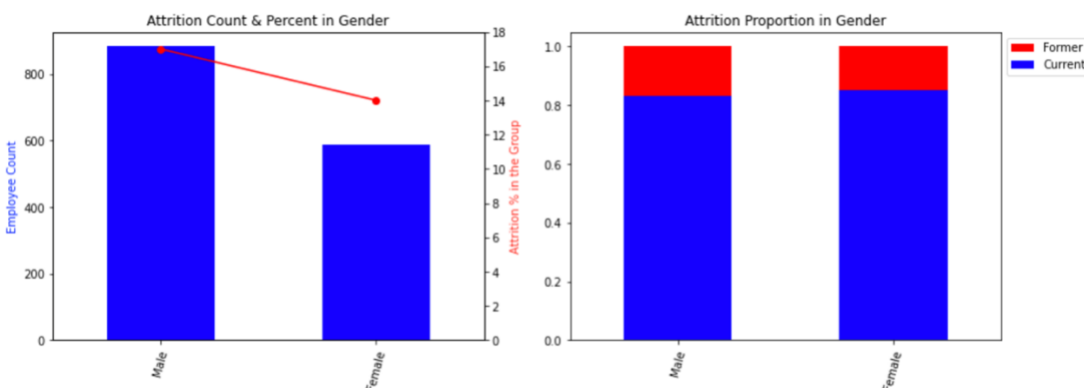
There are 1233 employees who stayed and 237 employees who left the company. As it is shown in the chart below, 16% of the employees left the company, which tells us there is quite a large skew in the target. Therefore, quite a large imbalance in the target variable is seen. I will use an oversampling method known as SMOTE to treat the imbalance in the data.



### 4.2 Features

#### 4.2.1 Gender

| Gender | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|--------|----------------|-----------------|--------------------------|----------------------------|
| Male   | 882            | 150             | 17                       | 63                         |
| Female | 588            | 87              | 14                       | 36                         |



According to the table and charts above, the number of male employees is higher than females. Also, the attrition rate in males is higher than in females.

#### 4.2.1.1 Hypothesis Testing

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that is made about the population parameter.

##### 4.2.1.1.1 Testing Employee Counts in Gender

There are 558 females and 882 males in the company. There is a difference between the employee counts. Is the difference significant? Is it too big or not really big? The answer is found by using the Chi-Square test. The Chi-Square statistics are commonly used for testing relationships between categorical variables. The goal is to see the difference in numbers between female and male employee categories.

Null Hypothesis:

$H_0$ : There is no difference in the number of males and females in the company.

Alternative Hypothesis:

$H_A$ : There is a difference in the number of males and females in the company.

| Category | Observed | Expected | Residual =<br>Observed - Expected | (Residual) <sup>2</sup> | (Residual) <sup>2</sup> /Expected | Chi Square<br>Score | Degree of<br>Freedom | P Value |
|----------|----------|----------|-----------------------------------|-------------------------|-----------------------------------|---------------------|----------------------|---------|
| Female   | 588      | 735      | -147                              | 21609                   | 29.4                              | 58.8                | 1                    | < 0.05  |
| Male     | 882      | 735      | 147                               | 21609                   | 29.4                              |                     |                      |         |

I used a 95% significance level. As in the statistical calculation table above, the p-value is less than 0.05. Therefore, the rejection of the null hypothesis fails. I am confident in accepting that there is no statistically significant difference between the number of males and females.

##### 4.2.1.1.2 Testing Attrition Rates in Gender

The attrition rate in males is 17% and 14% in females. It is obvious that male attrition is more than female attrition. Is the difference significant? Since the attrition rate and the sample size are greater than 30, it is appropriate to use a two-proportion z-test.

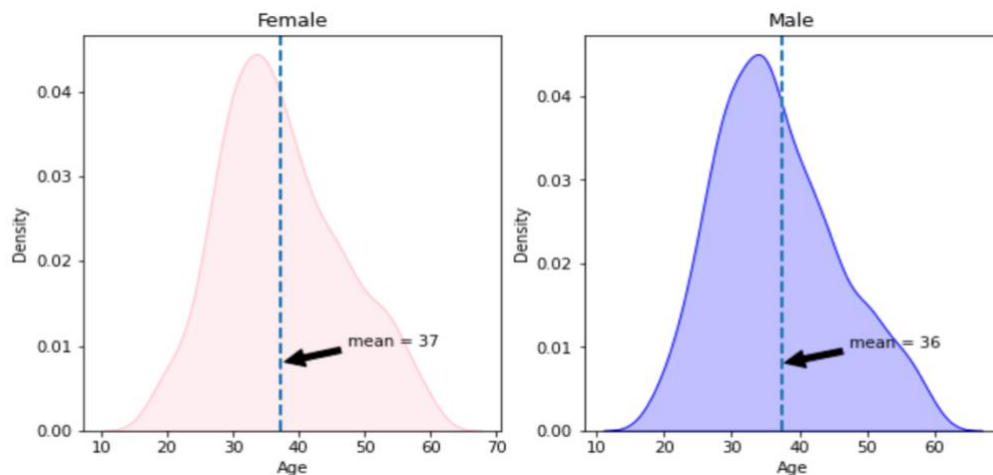
$H_0$ : There is no difference in the attrition rate for male and female employees in the company.

$H_A$ : There is a significant difference in the attrition rate for males and female employees in the company.

| Category | Observed | Attrition | Proportion  | Overall Proportion | Z Score     | P Value |
|----------|----------|-----------|-------------|--------------------|-------------|---------|
| Female   | 588      | 87        | 0.147959184 | 0.16122449         | 1.129254781 | < 0.05  |
| Male     | 882      | 150       | 0.170068027 |                    |             |         |

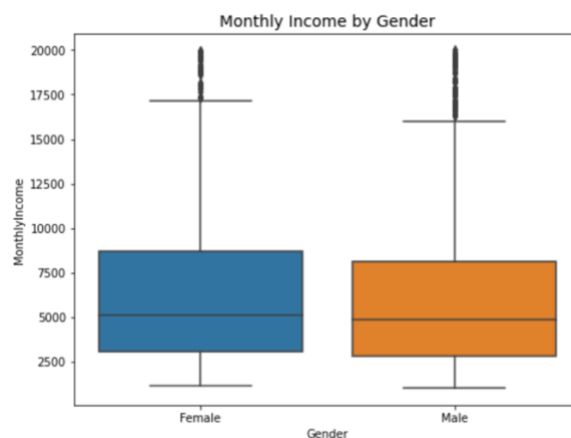
The p-value in the table above is below the significance level (0.05). Therefore, the rejection of the null hypothesis fails. I can conclude that there is no significant difference in the attrition rates between female and male employees in the company.

#### 4.2.1.2 Age Distribution by Gender



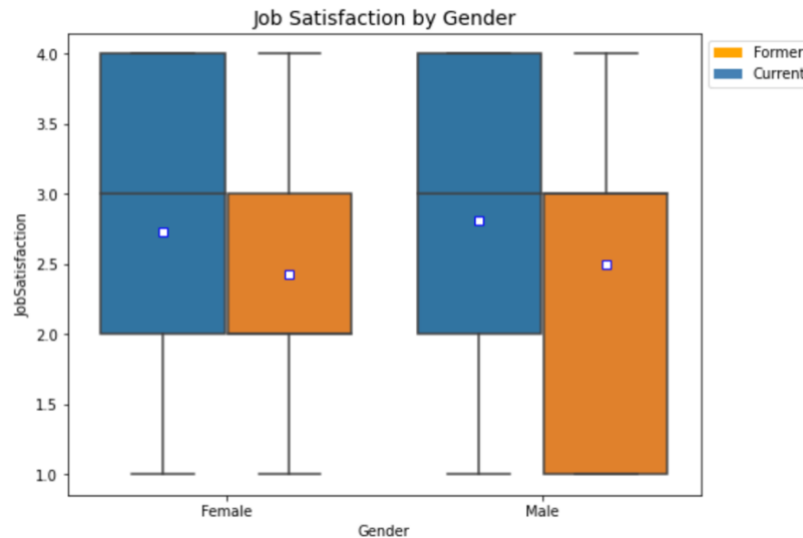
What is the age distribution between males and females? They both have the same distribution as below. The average age of females is 37 and for males is 36.

#### 4.2.1.3 Monthly Income by Gender



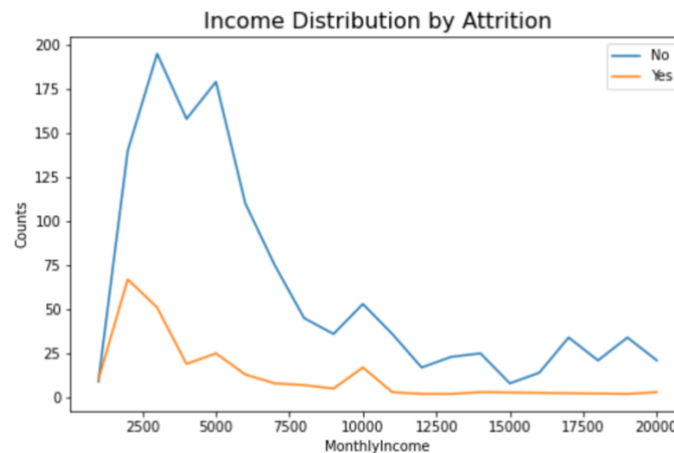
Is there any gender disparity in income? Based on the chart above, it clearly shows that males and females have equal payments on average.

#### 4.2.1.4 Job Satisfaction by Gender



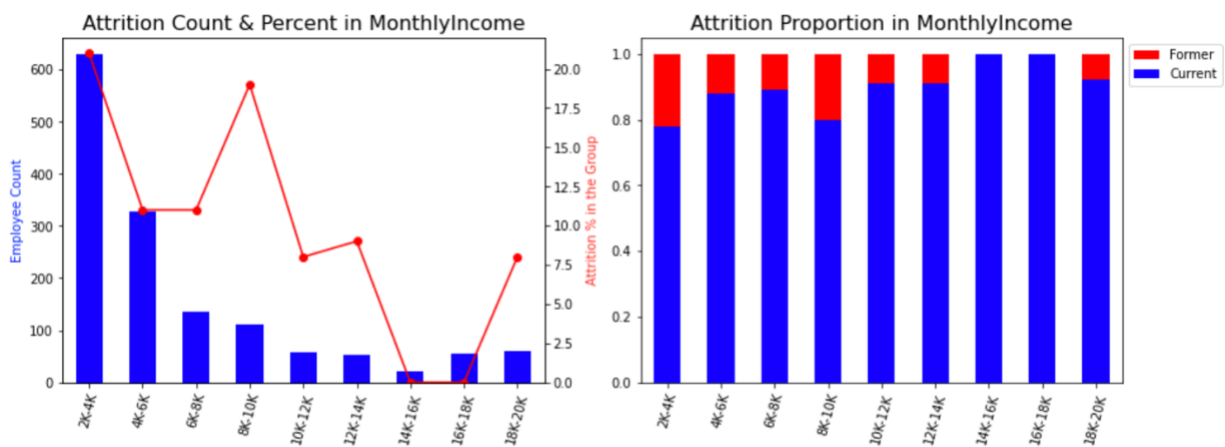
Based on the chart above, it is obvious that job satisfaction is low in males and females who left the company.

#### 4.2.2 Monthly Income

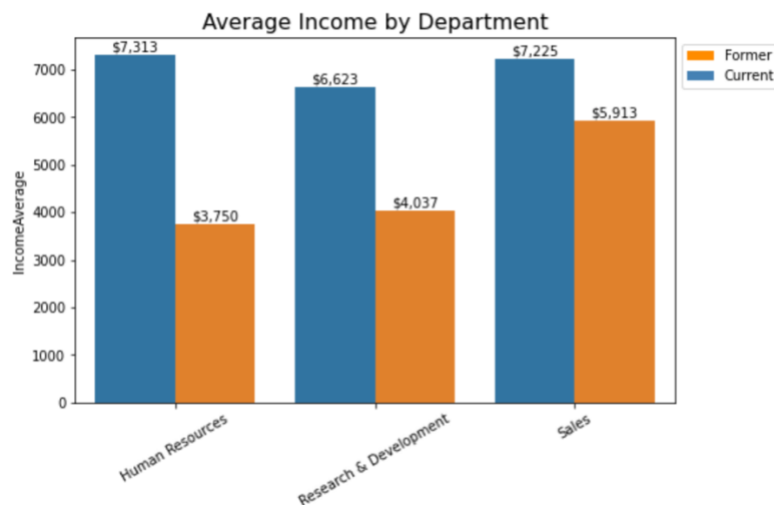


Is income the main factor toward employee attrition? As seen in the chart above, the attrition rate is obviously high at low-income levels, especially for the income levels less than \$4,000. The attrition rate decreases after \$4,000, but there is a minor spike at around \$10,000. Employees at the mid-income level tend to go to different companies to upgrade salaries or to shift toward a better standard of living. When the income is pretty decent, the chances of an employee leaving the company are low, as is seen by the flat line.

|   | MonthlyIncome | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|---|---------------|----------------|-----------------|--------------------------|----------------------------|
| 0 | 2K-4K         | 630            | 137             | 21                       | 57                         |
| 1 | 4K-6K         | 327            | 38              | 11                       | 16                         |
| 2 | 6K-8K         | 135            | 15              | 11                       | 6                          |
| 3 | 8K-10K        | 111            | 22              | 19                       | 9                          |
| 4 | 18K-20K       | 60             | 5               | 8                        | 2                          |
| 5 | 10K-12K       | 58             | 5               | 8                        | 2                          |
| 6 | 16K-18K       | 55             | 0               | 0                        | 0                          |
| 7 | 12K-14K       | 53             | 5               | 9                        | 2                          |
| 8 | 14K-16K       | 22             | 0               | 0                        | 0                          |



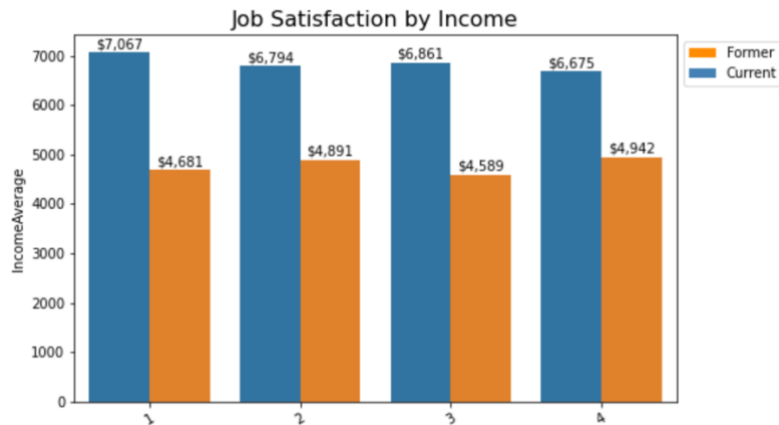
#### 4.2.2.1 Monthly Income by Department



The attrition rate is high in the Sales department despite the high average income compared to other departments.



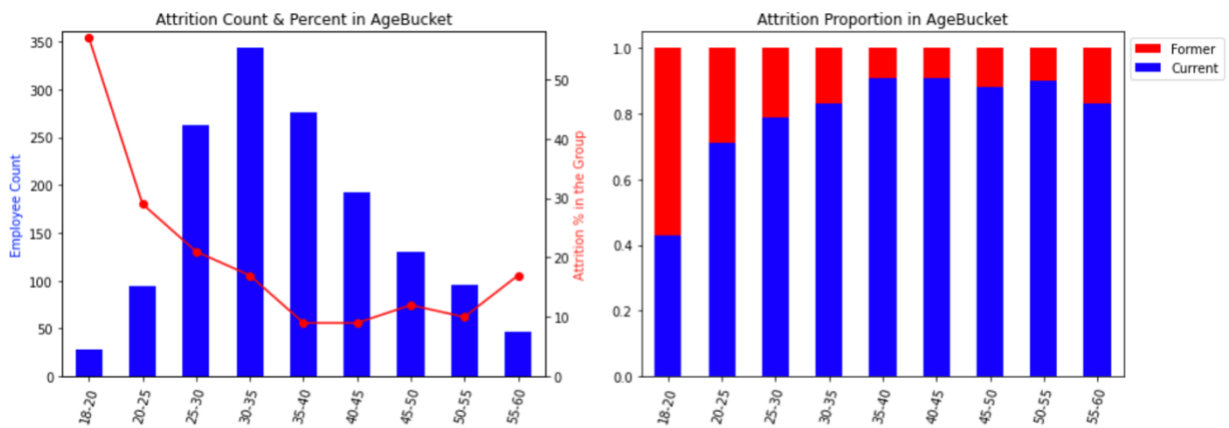
### 4.2.2.2 Job Satisfaction by Monthly Income



As it is clearly seen in the chart above, employees who left the company have low income in all satisfaction levels.

### 4.2.3 Age

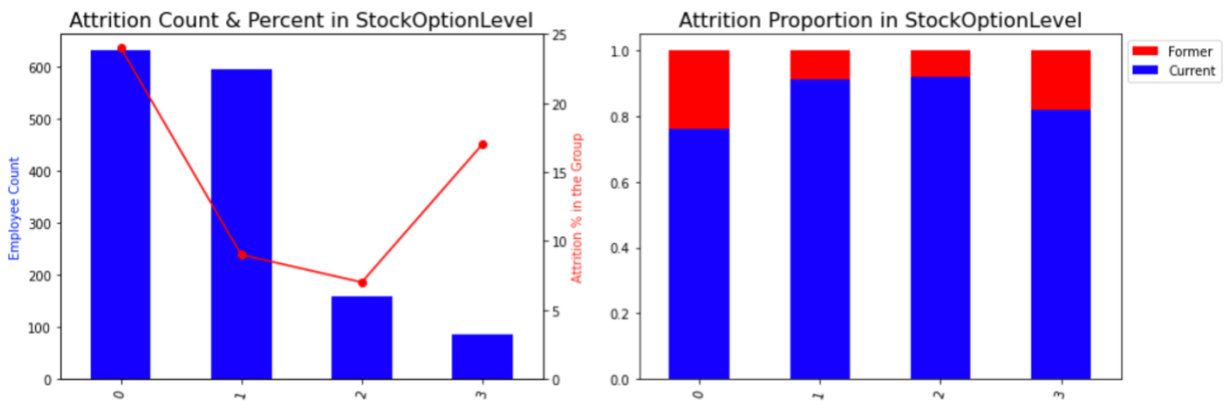
|   | AgeBucket | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|---|-----------|----------------|-----------------|--------------------------|----------------------------|
| 0 | 18-20     | 28             | 16              | 57                       | 6                          |
| 1 | 20-25     | 95             | 28              | 29                       | 11                         |
| 2 | 25-30     | 263            | 56              | 21                       | 23                         |
| 3 | 30-35     | 343            | 60              | 17                       | 25                         |
| 4 | 35-40     | 276            | 25              | 9                        | 10                         |
| 5 | 40-45     | 192            | 18              | 9                        | 7                          |
| 6 | 45-50     | 130            | 16              | 12                       | 6                          |
| 7 | 50-55     | 96             | 10              | 10                       | 4                          |
| 8 | 55-60     | 47             | 8               | 17                       | 3                          |



As it is seen in the chart above, the attrition is very high for young people who are less than 25. The attrition keeps on failing with increasing age. After around age 35, people prefer stability in their jobs.

#### 4.2.4 Stock Option

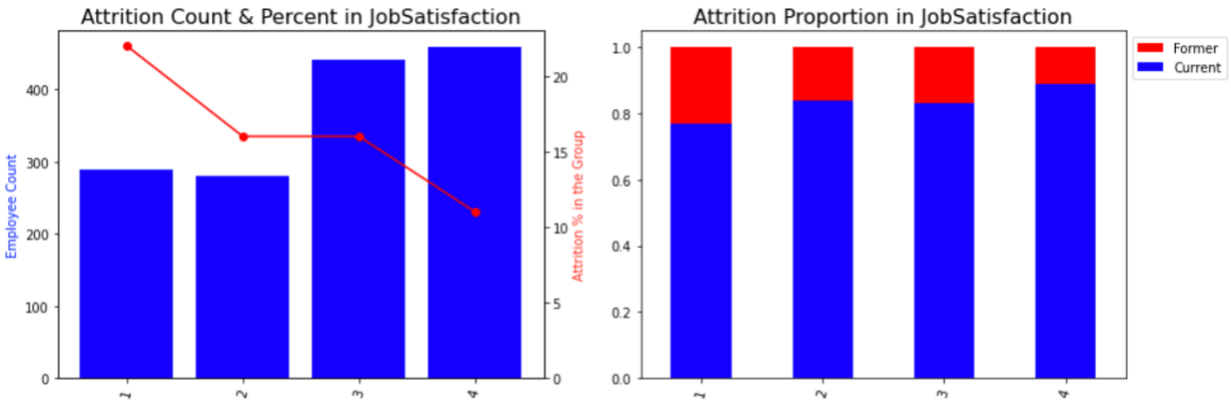
| StockOptionLevel | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|------------------|----------------|-----------------|--------------------------|----------------------------|
| 0                | 0              | 631             | 154                      | 24                         |
| 1                | 1              | 596             | 56                       | 9                          |
| 2                | 2              | 158             | 12                       | 7                          |
| 3                | 3              | 85              | 15                       | 17                         |



The tendency of employees to leave the company is high when there are low or no stock options. Since stocks contribute a huge amount of money, people do not want to lose that opportunity. People tend to leave the company if they are not happy with stock options.

#### 4.2.5 Job Satisfaction

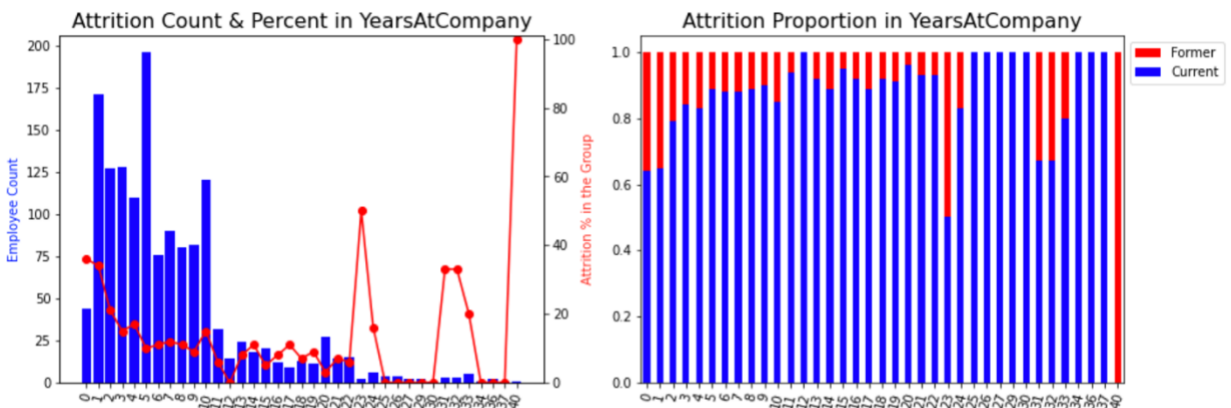
| JobSatisfaction | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|-----------------|----------------|-----------------|--------------------------|----------------------------|
| 0               | 1              | 289             | 66                       | 22                         |
| 1               | 2              | 280             | 46                       | 16                         |
| 2               | 3              | 442             | 73                       | 16                         |
| 3               | 4              | 459             | 52                       | 11                         |



There are four job satisfaction levels in the data. The value of 1 indicates low satisfaction, and 4 indicates high satisfaction. As it is seen in the chart above, employees who have low job satisfaction are more likely to leave the company than employees who have high job satisfaction. The attrition rate in low job satisfaction is 27%, and the attrition rate in high job satisfaction is 21%. With increasing job satisfaction, the attrition rates decrease.

#### 4.2.6 Years at Company

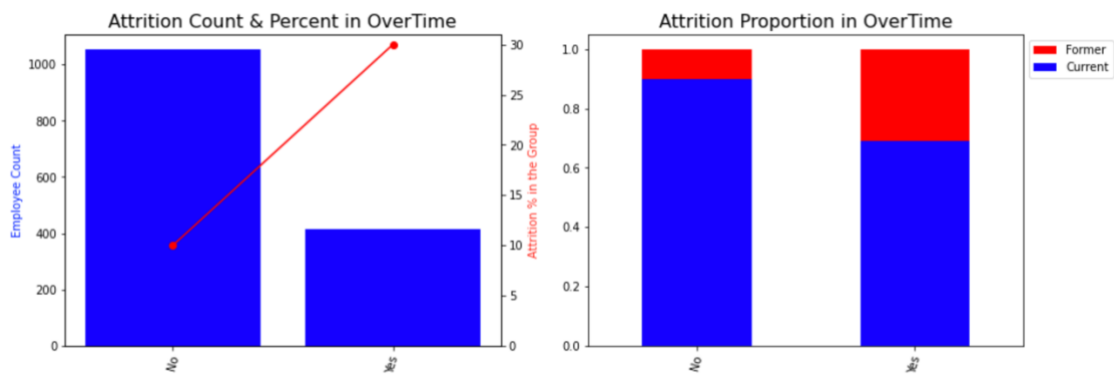
| YearsAtCompany | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|----------------|----------------|-----------------|--------------------------|----------------------------|
| 0              | 0              | 44              | 16                       | 36                         |
| 1              | 1              | 171             | 59                       | 34                         |
| 2              | 2              | 127             | 27                       | 21                         |
| 3              | 3              | 128             | 20                       | 15                         |
| 4              | 4              | 110             | 19                       | 17                         |
| 5              | 5              | 196             | 21                       | 10                         |
| 6              | 6              | 76              | 9                        | 11                         |
| 7              | 7              | 90              | 11                       | 12                         |
| 8              | 8              | 80              | 9                        | 11                         |
| 9              | 9              | 82              | 8                        | 9                          |
| 10             | 10             | 120             | 18                       | 15                         |



As it is seen in the chart and table above, employees who have two years or less working experience in the company have the highest attrition percentage. It composes 41% of the total attrition in the company. Employees who are in their initial years have a higher chance of leaving the company. People who have gained working experience tend to stay in the company.

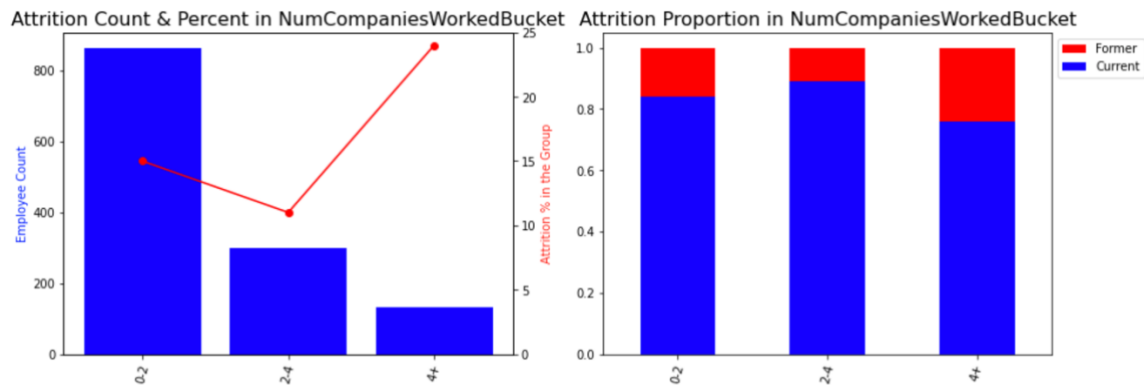
## 4.2.7 Over Time

|   | OverTime | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|---|----------|----------------|-----------------|--------------------------|----------------------------|
| 0 | No       | 1054           | 110             | 10                       | 46                         |
| 1 | Yes      | 416            | 127             | 30                       | 53                         |



## 4.2.8 Number of Companies Worked

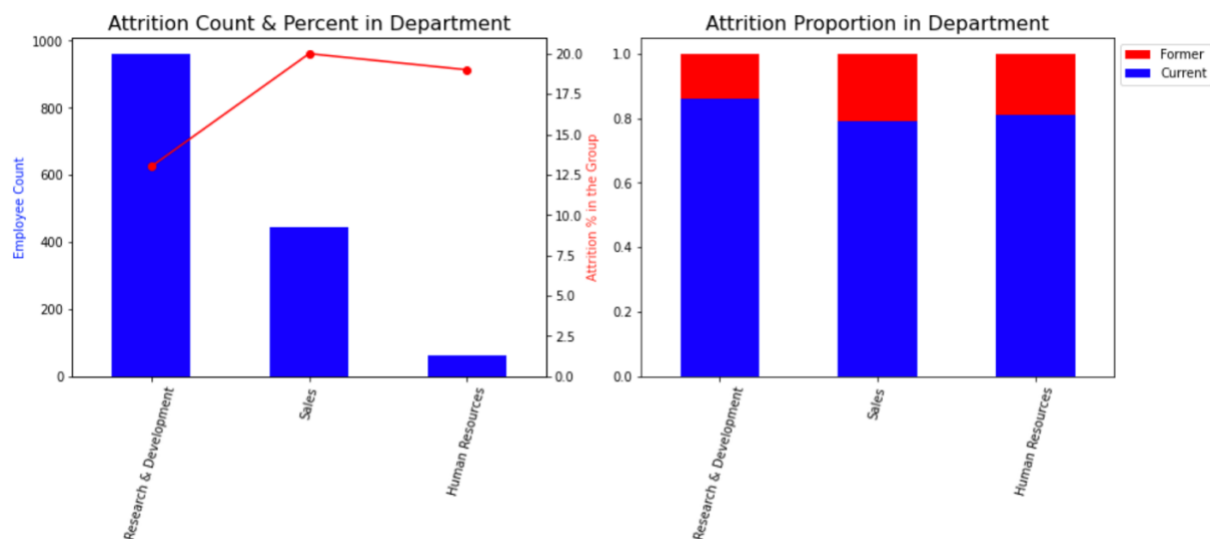
|   | NumCompaniesWorkedBucket | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|---|--------------------------|----------------|-----------------|--------------------------|----------------------------|
| 0 | 0-2                      | 864            | 137             | 15                       | 57                         |
| 1 | 2-4                      | 298            | 33              | 11                       | 13                         |
| 2 | 4+                       | 133            | 32              | 24                       | 13                         |



Employees are more likely to leave the company if the number of companies they worked for before this company increases.

## 4.2.9 Department

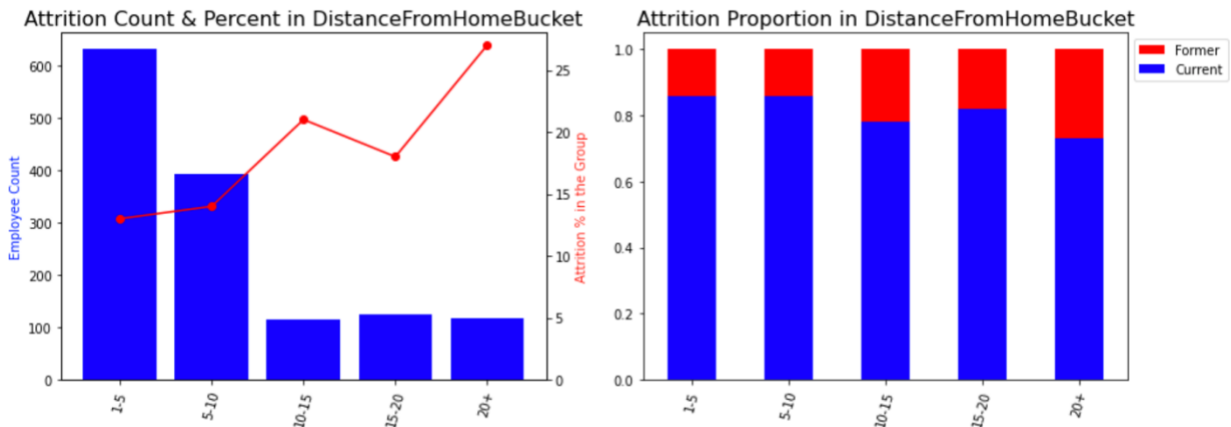
|   | Department             | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|---|------------------------|----------------|-----------------|--------------------------|----------------------------|
| 0 | Research & Development | 961            | 133             | 13                       | 56                         |
| 1 | Sales                  | 446            | 92              | 20                       | 38                         |
| 2 | Human Resources        | 63             | 12              | 19                       | 5                          |



There are three departments in the company based on the data. The Sales Department has the highest attrition rates at 20%, and it is followed by the Human Resource Department which has 19%. Research and Development have the least attrition rates, which shows employees prefer stability.

## 4.2.10 Distance from Home

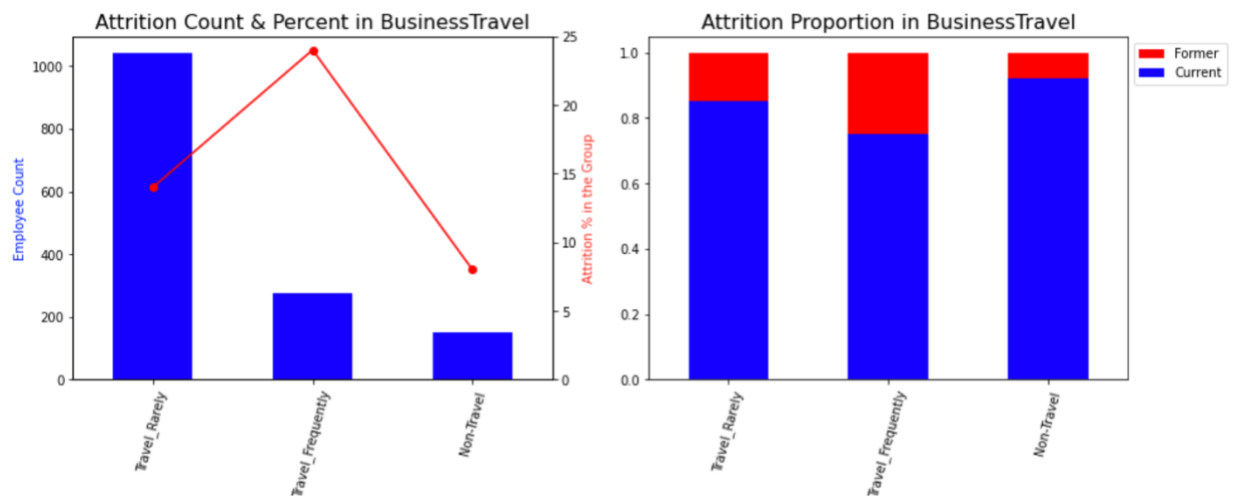
|   | DistanceFromHomeBucket | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|---|------------------------|----------------|-----------------|--------------------------|----------------------------|
| 0 | 1-5                    | 632            | 87              | 13                       | 36                         |
| 1 | 5-10                   | 394            | 57              | 14                       | 24                         |
| 2 | 10-15                  | 115            | 25              | 21                       | 10                         |
| 3 | 15-20                  | 125            | 23              | 18                       | 9                          |
| 4 | 20+                    | 117            | 32              | 27                       | 13                         |



How does the distance from home impact attrition? It is clearly shown in the chart that people who live more than 10 miles away from the company are more likely to leave the company. Employees who live more than 10 miles away from the company compose 1/3 of the whole company attrition.

#### 4.2.11 Business Travel

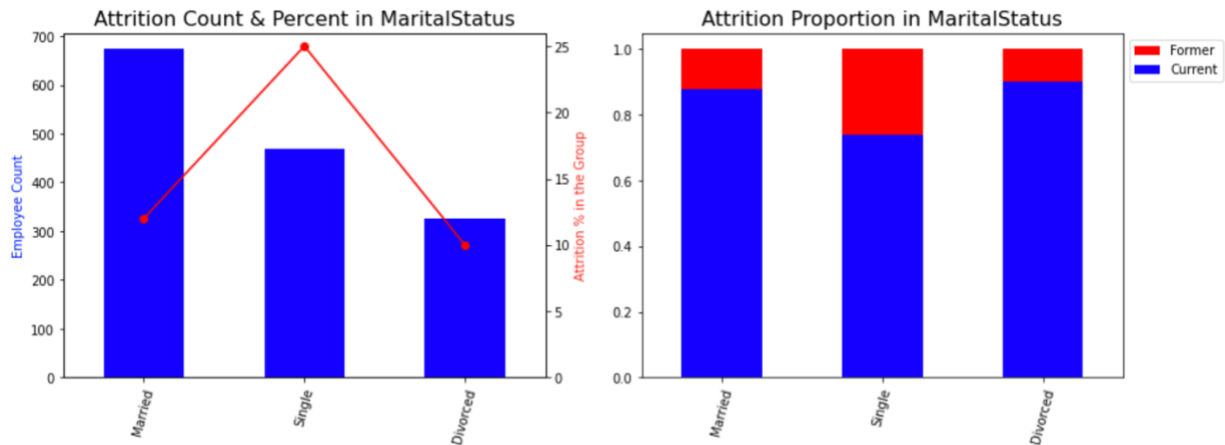
| BusinessTravel    | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|-------------------|----------------|-----------------|--------------------------|----------------------------|
| Non-Travel        | 150            | 12              | 8                        | 5                          |
| Travel_Frequently | 277            | 69              | 24                       | 29                         |
| Travel_Rarely     | 1043           | 156             | 14                       | 65                         |



What is the impact of business travel? Employees who travel frequently have the highest attrition rate, 24%. Employees who don't travel have the lowest attrition rate in the company, 8%. People do not prefer to travel in general.

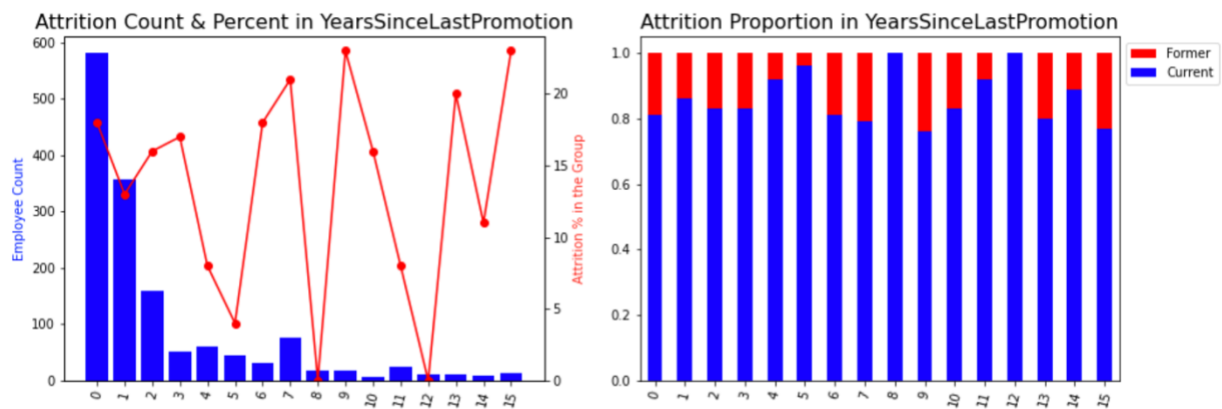
## 4.2.12 Marital Status

|   | MaritalStatus | Employee Count | Attrition Count | Attrition % in the Group | Attrition % in the Company |
|---|---------------|----------------|-----------------|--------------------------|----------------------------|
| 0 | Married       | 673            | 84              | 12                       | 35                         |
| 1 | Single        | 470            | 120             | 25                       | 50                         |
| 2 | Divorced      | 327            | 33              | 10                       | 13                         |



Single employees are more likely to leave the company. They have the highest attrition rate, which makes up 50% of all attrition in the company.

## 4.2.13 Years Since Last Promotion



There are 1097 employees who have less than 2-years since the last promotion, and 16% of those employees leave the company. Attrition rate increases if the number of years increases since the last promotion.

## 5 Data Pre-Processing

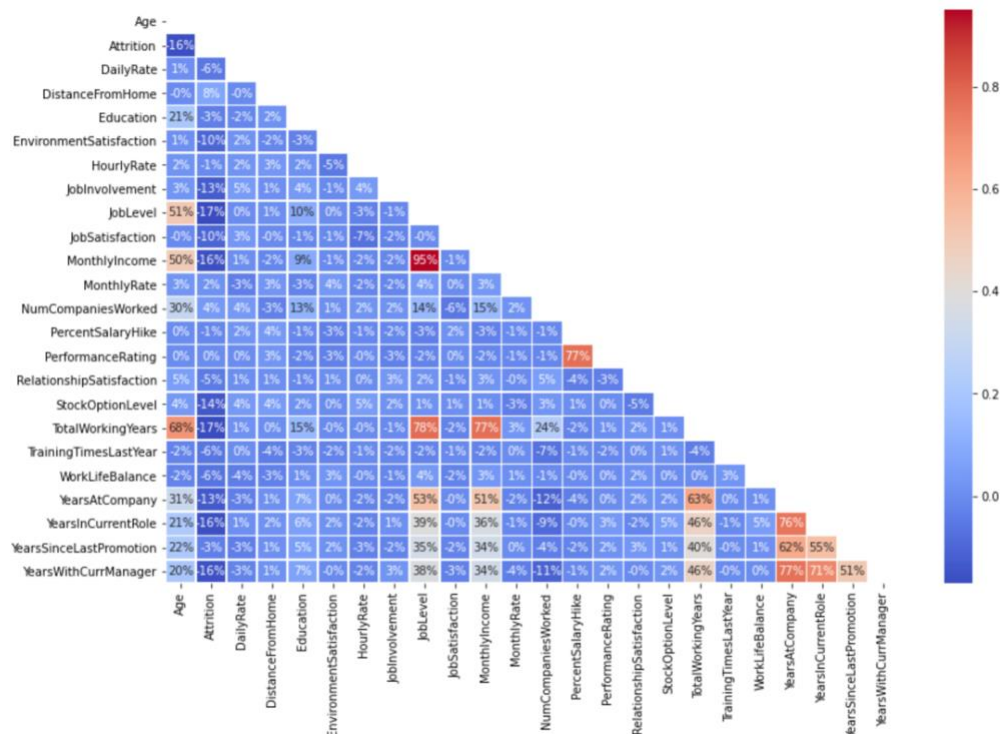
Before running the machine learning algorithms, I will do some pre-processing steps to make the dataset ready for model building.

### 5.1 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. Adding redundant variables reduces the generalization capability of the model and may also reduce the overall accuracy of a classifier. Furthermore, adding more and more variables to a model increases the overall complexity of the model.

Some features just have one data level that does not contribute anything to the model. EmployeeCount, Over18, and StandardHours, and employee numbers do not have meaning in analyzing results, so I also deleted these features.

I built a correlation matrix which is a table showing correlation coefficients between variables as in the heat map below. MonthlyIncome and JobLevel variables have a strong correlation (95%). Therefore, I can remove JobLevel and keep MonthlyIncome in the modeling part based on the results. The rest of the variables have less correlation in general; I will keep all others for now.





## 5.2 Dummy Variables

All the categorical data must be converted into numerical data for the Machine Learning model to work. I have used one hot encoding to create dummy variables. The basic strategy is to convert each category value into a new column and assign a 1 or 0 (True/False) value to the column. Dummy coding is a commonly used method for converting a categorical variable into a continuous variable.

## 5.3 Train Test Split

The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. First, I separated features and response variables as X and y. Then, I divided the dataset into the training and test sets. I have used 40% of the data for testing and 60% of the dataset for the training. Splitting the dataset is essential for an unbiased evaluation of prediction performance. I use the training set to build and train the model. Once the model is ready, I will test it on the testing set to see how well it performs.

## 5.4 Oversampling Imbalanced Data

An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. Most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. Imbalanced classifications significantly affect the model performance. This will result in poor predictive performance, specifically for the minority class. As is mentioned in the earlier sections, there is imbalanced data with 16% minority class. I have used an oversampling method known as the SMOTE (Synthetic Minority Oversampling Technique), which increases the number of observations for the minority class. The SMOTE method randomly creates synthetic instances of the minority class so that the net observations of both the classes get balanced out. I balanced only the training dataset and didn't touch the test dataset.

## 5.5 Feature Scaling

Each feature can have a different magnitude and different units. Variables that are measured at different scales do not contribute equally to the model. Some machine learning models are sensitive for scaling, and some are not. Since most of the machine learning algorithms use Euclidean distance between two data points in their computation, this is a problem. I used StandardScaler to scale the data. With this scaling method, features will have a mean of 0 and a standard deviation of 1. I fit the scaler on the training data and then used it to transform the testing data. This would avoid any data leakage during the model testing process.

## 5.6. Variance Inflation Factor (VIF)

Multicollinearity will be addressed in this section. Multicollinearity occurs when there are two or more independent variables in a multiple regression model, which have a high correlation among themselves. If there are some features that are highly correlated, there might be difficulty in distinguishing between their individual effects on the dependent variable. The Variance Inflation Factor (VIF) method is used to detect multicollinearity. Generally, a VIF above 10 indicates high multicollinearity. However, some important features in the model can be kept even if they have high VIF. It is also known from the correlation matrix that JobLevel has a 95% correlation with MonthlyIncome. VIF also shows that they are multicollinear. In this case, JobLevel is removed based on the table below.

|    | feature                  | VIF        |
|----|--------------------------|------------|
| 0  | const                    | 202.093542 |
| 1  | Age                      | 2.044890   |
| 2  | DailyRate                | 1.018198   |
| 3  | DistanceFromHome         | 1.015515   |
| 4  | Education                | 1.059352   |
| 5  | EnvironmentSatisfaction  | 1.010099   |
| 6  | HourlyRate               | 1.018318   |
| 7  | JobInvolvement           | 1.016316   |
| 8  | JobLevel                 | 11.205067  |
| 9  | JobSatisfaction          | 1.014511   |
| 10 | MonthlyIncome            | 10.800070  |
| 11 | MonthlyRate              | 1.012282   |
| 12 | NumCompaniesWorked       | 1.257737   |
| 13 | PercentSalaryHike        | 2.516385   |
| 14 | PerformanceRating        | 2.513734   |
| 15 | RelationshipSatisfaction | 1.015771   |
| 16 | StockOptionLevel         | 1.017682   |
| 17 | TotalWorkingYears        | 4.767796   |
| 18 | TrainingTimesLastYear    | 1.009917   |
| 19 | WorkLifeBalance          | 1.014884   |
| 20 | YearsAtCompany           | 4.587391   |
| 21 | YearsInCurrentRole       | 2.718604   |
| 22 | YearsSinceLastPromotion  | 1.674278   |
| 23 | YearsWithCurrManager     | 2.774587   |

## 6 Model Development and Evaluation

### 6.1 Machine Learning Models

In this part, machine learning models are built to select important features that influence employee attrition and classify the features to help us understand the main reason why people left the IBM company. The model is constructed to predict employee attrition according to the given data. Since there are binary response variables in the data, classification

models are used. The aim is to have better accuracy of predicting attrition (Attrition = 1), which is "True Positive" in the confusion matrix.

Logistic Regression, Random Forest Classification, KNN Classification, Support Vector Machine Classification, Gradient Boosting Classification, and Adaptive Boost Classification algorithms are applied to the IBM Employee Attrition data. Initially, machine learning algorithms with default parameters are applied. Then, hyperparameter tuning is used to try to improve model performance. Then, the threshold is moved to see any change in the model performance.

## 6.2 Model Evaluation Metric

Once a classification model is built, an evaluation of how good the predictions made by that model are needed. Accuracy can tell whether a model is being trained correctly and how it may perform generally. However, it does not give detailed information regarding its application to the problem. The problem with using accuracy as the main performance metric is that it does not do well with the presence of a severe class imbalance. Accuracy is used when the True Positives and True negatives are more important.

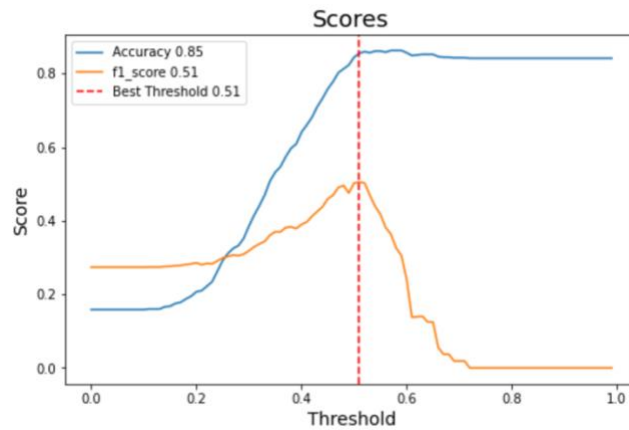
F1 is a harmonic mean of precision and recall. F1-score is used when the False Negatives and False Positives are crucial. A good F1 score means that low false positives and low false negatives are present, so correctly identifying real threats and not being disturbed by false alerts is needed. The data is highly imbalanced data. F1-score is a better metric when there are imbalanced classes. Therefore, F1-score is used to select the best model.

## 6.3 Model Comparison Table

|                            | default_model | grid_search_model | moved_threshold |
|----------------------------|---------------|-------------------|-----------------|
| <b>Logistic Regression</b> | 0.4694        | 0.5024            | 0.5052          |
| <b>Random Forest</b>       | 0.3014        | 0.4096            | 0.4889          |
| <b>KNN</b>                 | 0.2818        | 0.3305            | 0.3580          |
| <b>SVM</b>                 | 0.3380        | 0.3920            | 0.4107          |
| <b>Gradient Boosting</b>   | 0.4118        | 0.4118            | 0.4608          |
| <b>ADA Boosting</b>        | 0.3700        | 0.4615            | 0.4615          |

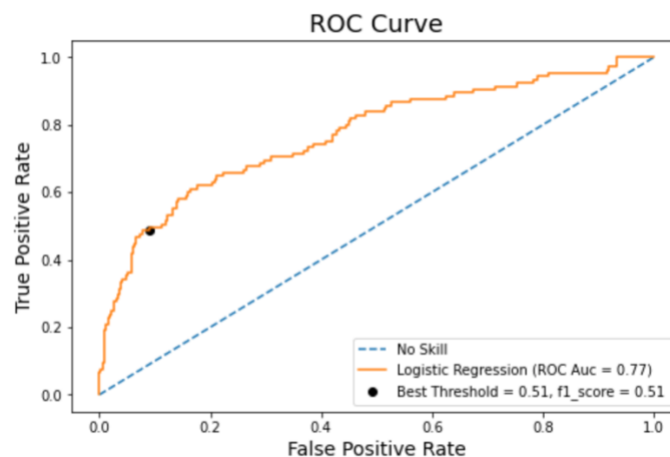
Based on the F1-score of different models, Logistic Regression gives a higher F1-score (51%) when the threshold is moved. Therefore, Logistic Regression is used to understand the factors that are impacting employee attrition.

## 6.4 Optimum Threshold



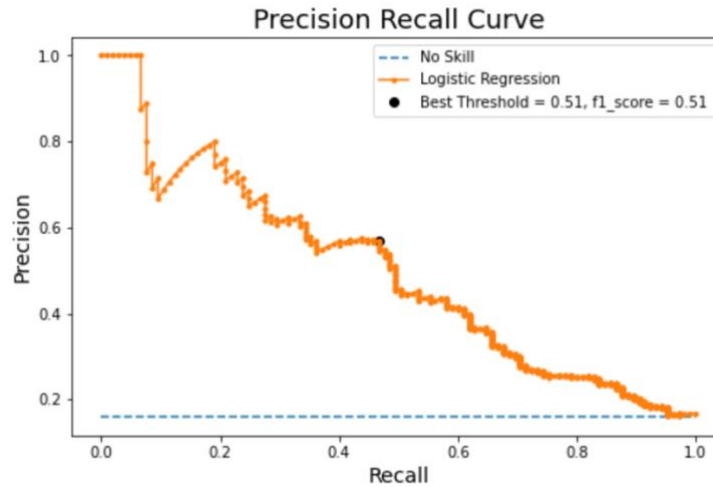
The default threshold for classification algorithms is 0.50. All predicted probabilities that are equal or greater than the threshold are mapped to 1 class, and all other values are mapped to 0 class. Based on the chart above, the F1-score is having a high value when the threshold is 0.40. F1-score is decreased if the threshold increases.

## 6.4 ROC Curve



ROC AUC is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0 and 1. It is a trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds. The AUC (Area under Curve) is 0.77. The plot shows low false positives and high true positives.

## 6.5 Precision Recall Curve



Precision-Recall is a useful measure of successful prediction when the classes are very imbalanced. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

## 7 Conclusion

### 7.1 Important Features

Coefficient values are used to explain the logistic regression model. However, coefficients are not directly related to importance. When a binary outcome variable is modeled using logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables.

$$\text{logit}(p) = \log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right) = \log\left(\frac{P(y = 1)}{P(y = 0)}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

The logit transformation squeezes the output of a linear equation between 0 and 1. The odds ratio is received by taking the exponent of both sides of the equation above.

$$\frac{P(y = 1)}{1 - P(y = 1)} = \text{odds} = e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}$$

Then, a comparison of what happens when one of the feature values is increased by 1. Next is to look at the ratio of two predictions.

$$\frac{\text{odds}_{x_j+1}}{\text{odds}} = \frac{e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_j (x_{j+1}) + \dots + \beta_n x_n)}}{e^{(\beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \dots + \beta_n x_n)}}$$

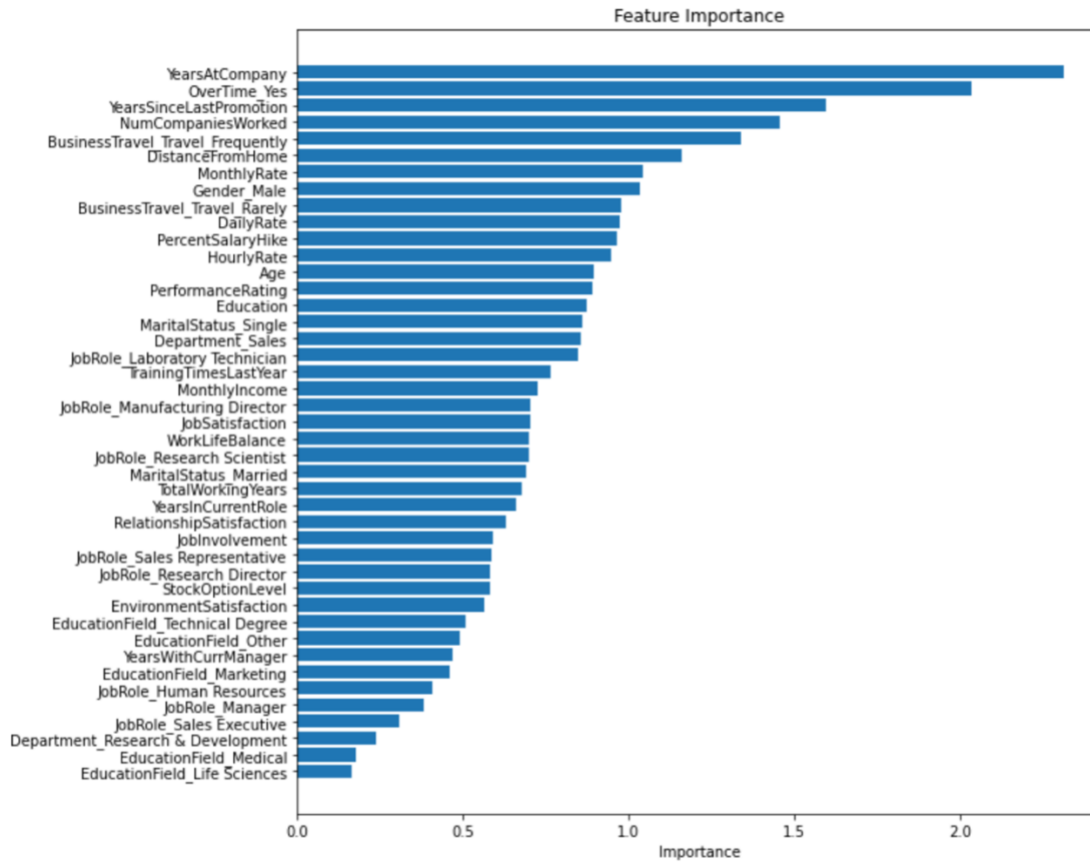
Then apply this rule:

$$\frac{e^a}{e^b} = e^{(a-b)}$$

And remove many terms:

$$\frac{\text{odds}_{x_j+1}}{\text{odds}} = e^{(\beta_j (x_{j+1}) - \beta_j x_j)} = e^{\beta_j}$$

This equation means that a change in a feature by one unit changes the odds ratio by a factor of  $e^{\beta_j}$ . This rule is applied to all coefficients to find the feature importance.



As it is seen from the feature importance plot, the strongest feature in employee attrition data is YearsAtCompany. An increase of the YearsAtCompany feature by one unit increases the odds ratio of YearsAtCompany by a factor of 2.3 when all other features remain the same.

## 7.2 Strong Factors in Employee Attrition

- 1 23% of employees have two years or less working experience in the company, and 30% of those employees leave the company. The model shows that the number of years at the company is the strongest factor in attrition. Employees who are in their first years are more likely to leave the company. Employees who have gained working experience prefer to stay in the company. Therefore, the company should understand why their new employees leave the company.
- 2 28% of employees work overtime in the company, and 30% of those employees leave the company. The model also shows that overtime work is the second strongest factor in attrition. Therefore, the company should understand the reason why they are working overtime. Is it for too high a workload, or are employees' qualifications not enough to complete the scheduled tasks on time? There might be some other reasons behind that. The recommendation will be to understand the reason(s) for overtime with detailed research and take appropriate measures to reduce the factors behind this attrition factor.
- 3 There are 1097 employees who have less than 2-years since the last promotion, and 16% of those employees leave the company. Attrition rate increases if the number of years increases since the last promotion. Because of this reason, the company should review the promotion policy. They can make promotion requirements clear to all employees how and when they can be promoted.
- 4 9% of employees worked in 4 or more companies before this company, and 24% of those employees leave the company. The company should understand why those employees are leaving the company. The company can question the applicants why they want to quit their previous job during the interview process. The company can even get in touch with the previous company.
- 5 19% of the employees travel frequently, and 70% of those employees leave the company. This is one of the strongest factors in employee attrition. The company should understand if a high number of business travels are really necessary. They can adjust the frequency. If they can't, they can give some extra incentives to motivate those employees.
- 6 24% of the employees live more than 10 miles away from the company, and 22% of those employees leave the company. The company should understand how distance is affecting their work. They can support their employees to move closer areas to the company by providing moving expenses. If they can't move closer due to their mandatory things, the company can adjust their work shift to prevent employees from wasting their time during rush hours. The company can allow some days to work from home if they can.

## **7.3 Next Steps to Improve Model**

### **7.2.1 More Data**

In the data, there are 1470 observations from the company. This is poor data to create a strong model predicting attrition before that happens. Because of this reason, the company should focus on collecting more data to improve predicting power of modeling. They can prepare surveys for the employees who are leaving and for the employees who stay in the company for longer years.

### **7.2.2 Bootstrap F1 Score**

With highly imbalanced data, the model score will have a high variance. The F1 scores are bootstrapped to mitigate this. Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples. This process allows for the calculation of standard errors, confidence intervals, and hypothesis testing. The confidence interval of the F1 score is calculated to get the stability of the results.