HR Data Analysis Report and Predicting of Employee Attrition

July 2023 // Prepared by Muhannad Mansour Felemban

CONTEXT:

Employee attrition can have a significant impact on an organization's bottom line, including loss of productivity, knowledge, and experience. In today's competitive job market, organizations need to take proactive measures to retain the best employees and reduce the risk of employee turnover. The purpose of this report is to analyze HR data and develop a predictive model to identify factors that contribute to employee attrition and to predict which employees are likely to leave.

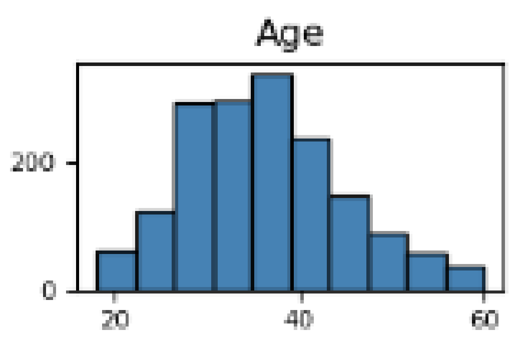
The report will begin by providing a statistical analysis to identify patterns and trends in the data. An exploration of the factors contributing to staff attrition and an assessment of the current retention strategies in place will be presented. In addition, the report will describe the development of the predictive model, including the modeling approach used, the data preparation steps taken, and the results of model evaluation.

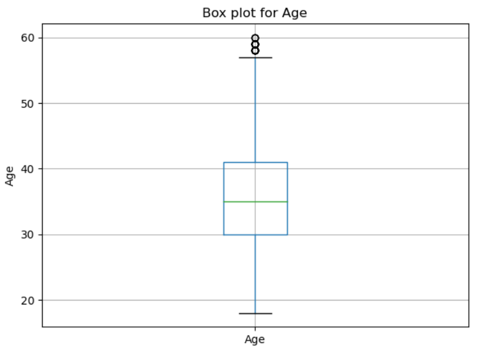
the report will conclude with a summary of the main findings and key recommendations for the organization. These recommendations will include suggestions for improving current retention strategies and implementing the predictive model to proactively identify employees who are at high risk of leaving.

Exploratory Data Analysis :

- **Age**

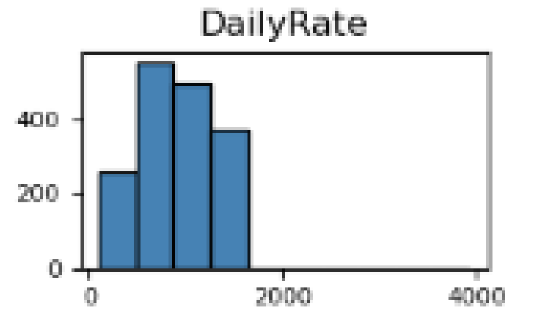
The histogram and intensity for the 'age' variable show that the distribution is roughly bell-shaped, with the distribution peaking around 30–35 years. The distribution is skewed somewhat to the right, indicating that there are more older employees in the data set than younger employees. The box plot of the 'age' variable shows that there are no outliers in the data. The box plot shows that the average data is about 35 years old, and the data range spans from 18 to 60 years.

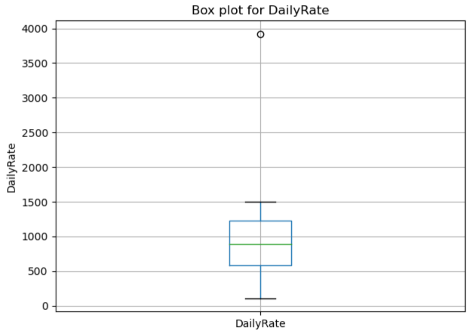




- **Daily Rate:**

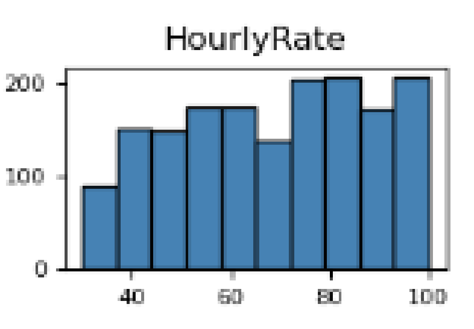
The histogram and density of the “Daily Rate” variable show that the distribution is roughly bell-shaped, with peak distribution around $500-1000 per day. The distribution skews somewhat to the right less than the mean, indicating that there are more high daily rates in the data set than low daily rates. The box plot of the 'Daily Rate' variable shows that there are many outliers in the data, with values above the top bar. The upper whisker extends to around $3,900/day, indicating that any data points with daily averages above this value could be considered outliers.

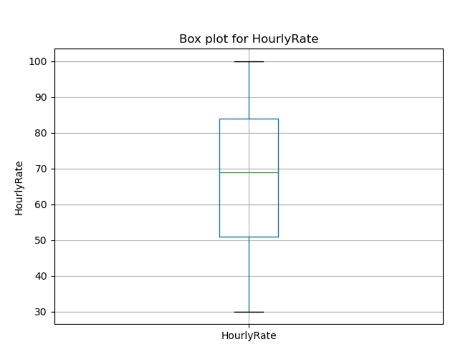




- **Hourly Rate:**

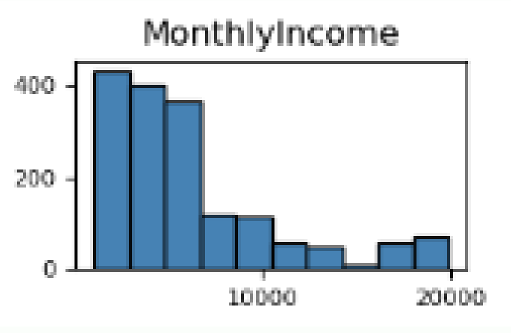
The histogram and density plot of the "Hourly Rate" variable show that the distribution is somewhat bell-shaped, with the peak of the distribution around 65-70 dollars per hour. The distribution is somewhat skewed to the left, indicating that there are more low hourly rates in the dataset than high hourly rates. The box plot of the "Hourly Rate" variable shows that there are no significant outliers in the data. The box plot shows that the median of the data is around 69 dollars per hour, and the range of the data extends from 30 to 100 dollars per hour.

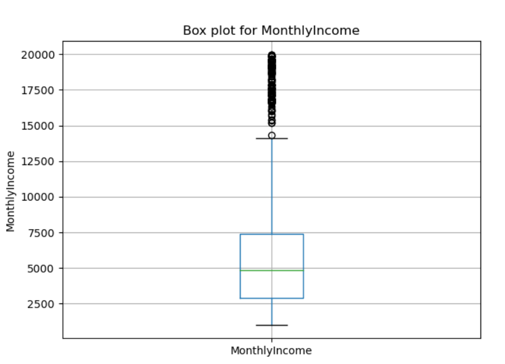




- **Monthly Income:**

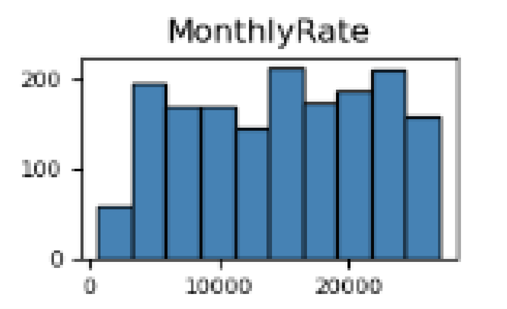
The histogram and density plot of the "Monthly Income" variable show that the distribution is heavily skewed to the right, with a long tail of values to the right of the mean. This suggests that there are a few employees who earn a very high monthly income, while most employees earn a lower monthly income. The box plot of the "Monthly Income" variable shows that there are several outliers in the data, with values above the upper whisker. The upper whisker extends to around 13000 dollars per month, indicating that any data points with monthly income above this value can be considered outliers.

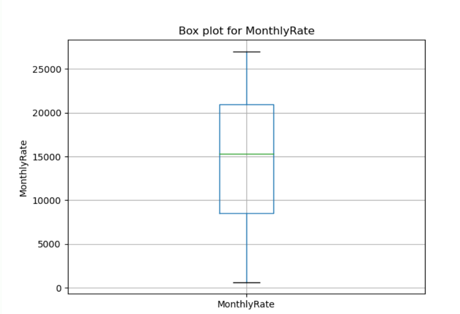




- **Monthly Rate:**

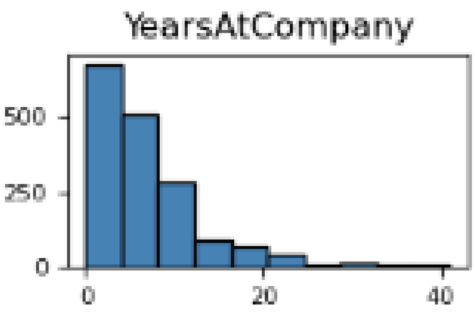
The histogram and density of the "MonthlyRate" variable show that the distribution is not bell shaped, with the peak distribution at around $15,000 per month which is roughly the median value. The distribution is slightly skewed to the left, indicating that there are lower monthly rates in the data set than high monthly rates. The box plot of the 'MonthlyRate' variable shows that there are no significant outliers in the data. The fund chart shows that the average data is about $15,000 per month, and the data range is from 600 to $27,000 per month.

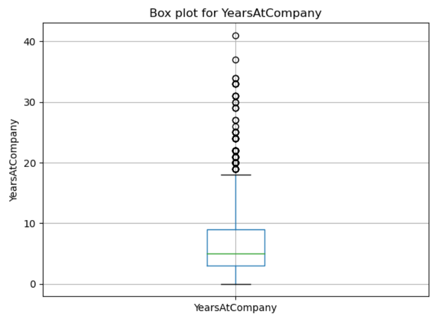




- **Years at Company:**

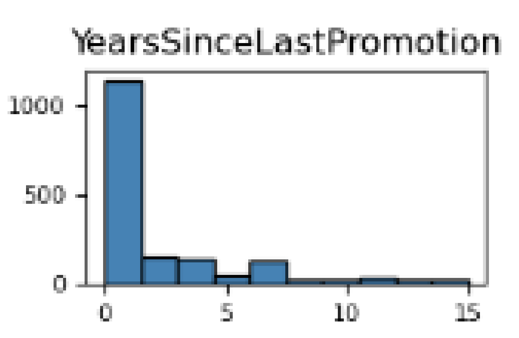
The histogram and density for the "YearsAtCompany" variable show that the distribution is shifted significantly to the left. This indicates that most employees have been with the company for a shorter period of time. The box plot of the 'YearsAtCompany' variable shows that there are many outliers in the data, with values above the upper whisker. The upper longitudinal line spans about 20 years, indicating that any data points with years in the firm above this value can be considered outliers. The scatter plot of the "YearsAtCompany" variable shows that there are many data points with long years in the company that are far from the rest of the data points.

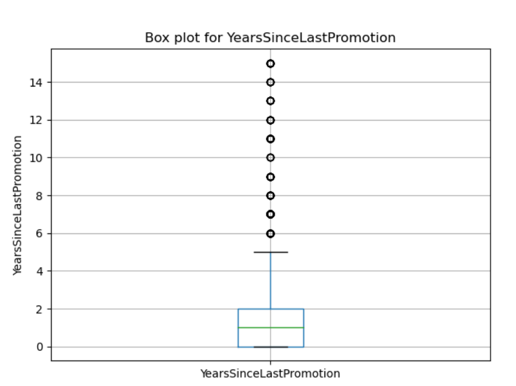




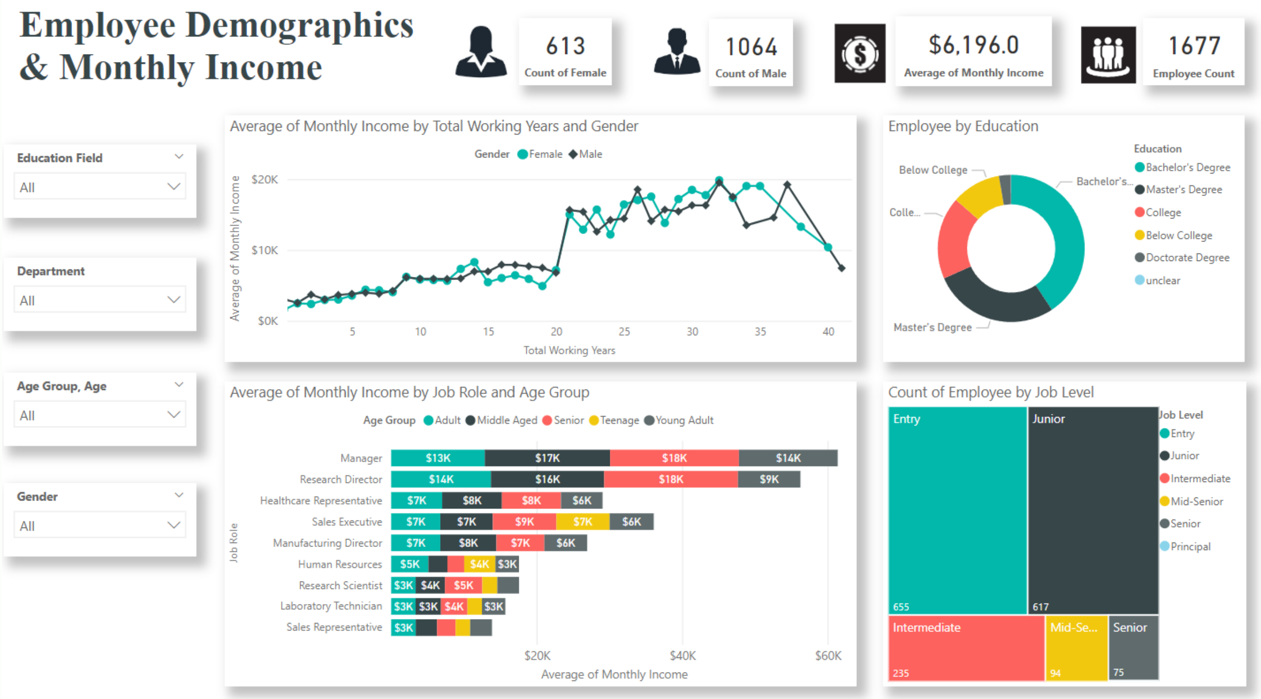
- **Years Since Last Promotion:**

The histogram and density of the "YearsSinceLastPromotion" variable show that the distribution is strongly skewed to the left, with a peak around 0-1 year and a long tail of values to the right of the mean. This indicates that there are a few employees who have been promoted recently, and most of the employees have not been promoted for a while. The box plot of the "YearsSinceLastPromotion" variable shows that there are several outliers in the data, with values above the upper whisker. The upper whisker extends to around 6 years, indicating that any data points with years since last promotion above this value can be considered outliers.





Data visualization (Power BI Report):



Employee Demographics & Monthly Income

613 Count of Female

1064 Count of Male

$6,196.0 Average of Monthly Income

1677 Employee Count

Education Field: All

Department: All

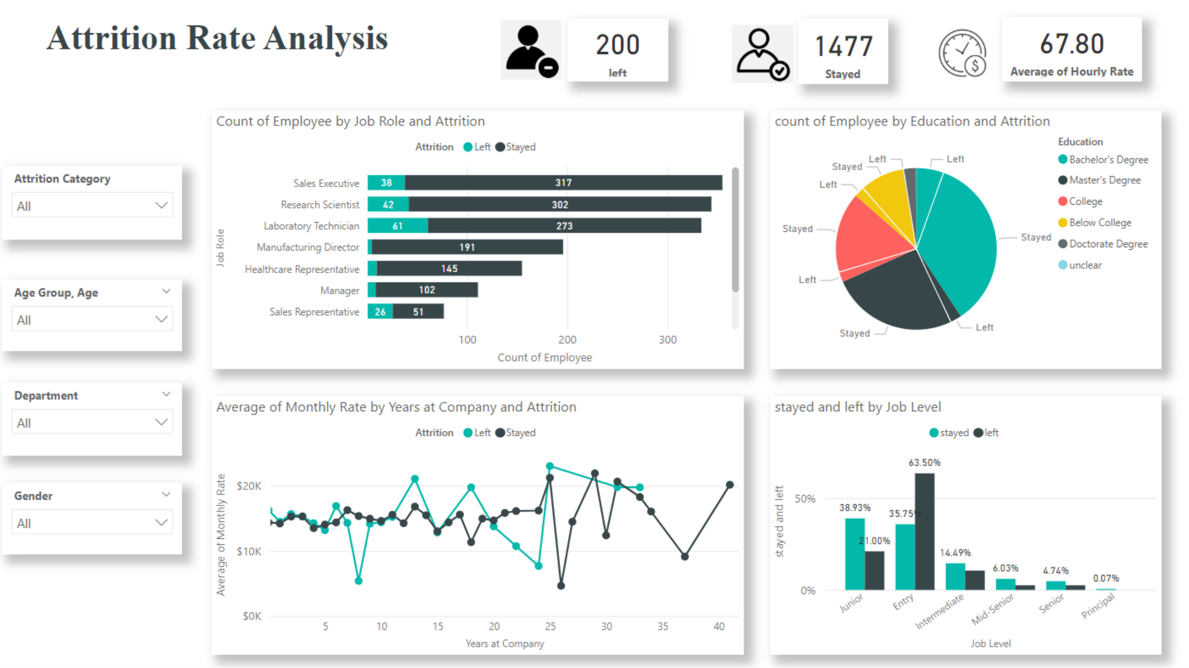
Age Group, Age: All

Gender: All

Average of Monthly Income by Total Working Years and Gender
Gender ● Female ◆ Male
$20K
$10K
$0K
Average of Monthly Income
5
10
15
20
25
30
35
40
Total Working Years
Employee by Education
Below College
Bachelor's...
Colle...
Master's Degree
Education
● Bachelor's Degree
● Master's Degree
● College
● Below College
● Doctorate Degree
● unclear
Average of Monthly Income by Job Role and Age Group
Age Group ● Adult ● Middle Aged ● Senior ● Teenage ● Young Adult
Job Role
Manager $13K $17K $18K $14K
Research Director $14K $16K $18K $9K
Healthcare Representative $7K $8K $8K $6K
Sales Executive $7K $7K $9K $7K $6K
Manufacturing Director $7K $8K $7K $6K
Human Resources $5K $4K $3K
Research Scientist $3K $4K $5K
Laboratory Technician $3K $3K $4K $3K
Sales Representative $3K
$20K
$40K
$60K
Average of Monthly Income
Count of Employee by Job Level
Entry 655
Junior 617
Intermediate 235
Mid-Se... 94
Senior 75
Job Level
● Entry
● Junior
● Intermediate
● Mid-Senior
● Senior
● Principal

- **Employee Demographics & Monthly Income Dashboard:**

The dashboard shows that the number of male employees is higher than that of females, and holders of a bachelor's degree constitute about 41% of the employees, followed by a master's degree by about 28%. The average monthly income increases compared to the total years of work from 20 to 21 years, by 85.71% for males. and for females 72.72%, and the Entry level job the largest percentage about 39%, and Junior level job abut 37%.
The two highest job role in terms of average monthly income are manager with average 16.500$ and Research director abut average 15.200$.



Attrition Rate Analysis

200 left

1477 Stayed

67.80 Average of Hourly Rate

Attrition Category: All

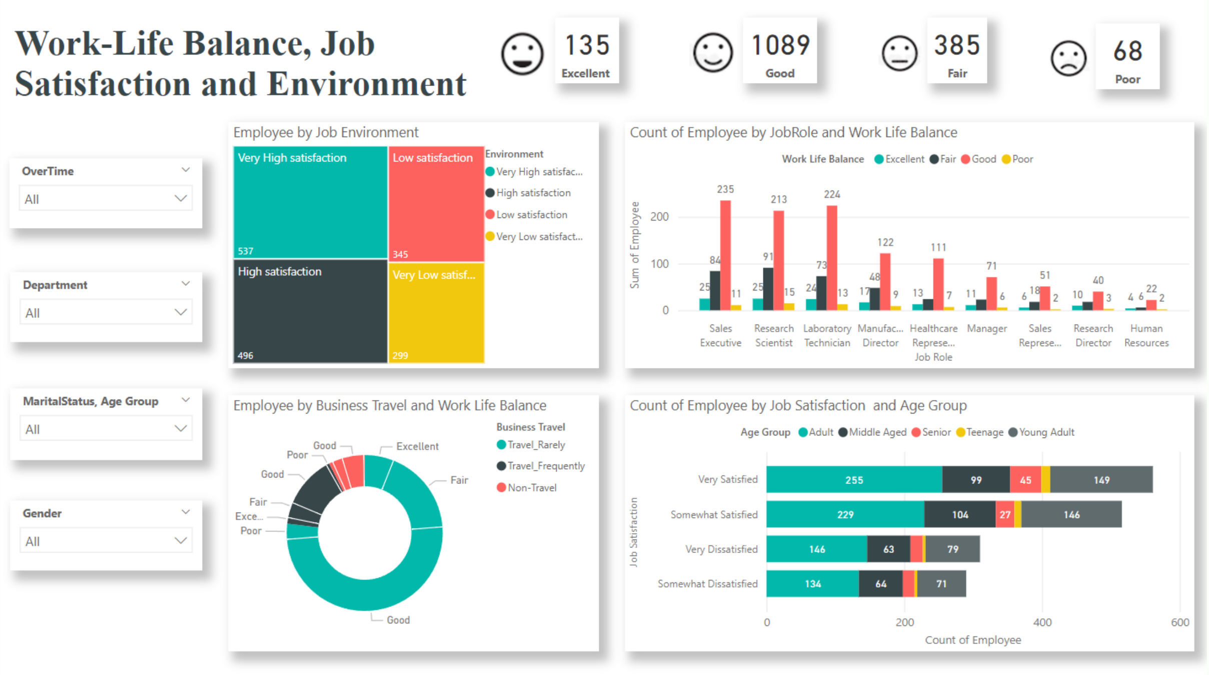
Age Group, Age: All

Department: All

Gender: All

- **Attrition Rate Analysis:**

The dashboard shows that the number employees stayed greater than left, As we can see, there are 18% of the technical laboratory employees who left and 84% stayed of the total number of employees in this job role (334), and there are 34% of the Sales Representative employees who left and 66% stayed of the total number of employees in this job role (77).
And where he was also the most left employees from the Entry level constituting about 63% from total left count and follows junior level, Most of them was are young and Adult age groups.



Work-Life Balance, Job Satisfaction and Environment

135 Excellent

1089 Good

385 Fair

68 Poor

OverTime: All

Department: All

MaritalStatus, Age Group: All

Gender: All

Work-Life Balance:

The dashboard shows that most employees have a good level of work life balance, So that it represented the highest percentage of a 'good' vote in terms of the job role, which is the Human Resources, with about 65% according to the number of employees for this job role (34), and also sales representative employees, where the votes were 66% of the total number of employees (77).

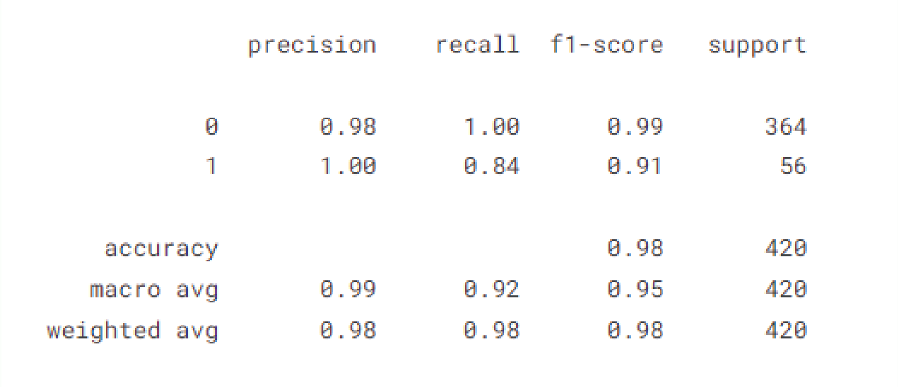
Employee Attrition Prediction Results Explained:

Overview:
Our model has been trained to predict whether employee attrition will occur or not. The key metrics to focus on are precision, recall, and the F1-score. We'll also discuss the confusion matrix, which displays the number of correct and incorrect predictions, and feature importance, which highlights the top factors contributing to the prediction.

Key Metrics :
1- Precision: This represents the accuracy of the positive predictions made by the model. In our case, for employees not leaving the company (class 0), the precision is 0.98. For those leaving (class 1), it's 1.00. The higher the precision, the better the model is at correctly identifying employees who will leave or stay.

2- Recall: This measures the ability of the model to find all the relevant instances within the data. For class 0, the recall is 1.00, and for class 1, it's 0.84. The higher the recall, the better the model is at identifying the true positive cases.

3- F1-score: This is the harmonic mean of precision and recall, giving equal importance to both. For class 0, the F1-score is 0.99, and for class 1, it's 0.91. The closer the F1-score is to 1, the better the model's overall performance.



```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       364
           1       1.00      0.84      0.91        56

    accuracy                           0.98       420
   macro avg       0.99      0.92      0.95       420
weighted avg       0.98      0.98      0.98       420
```

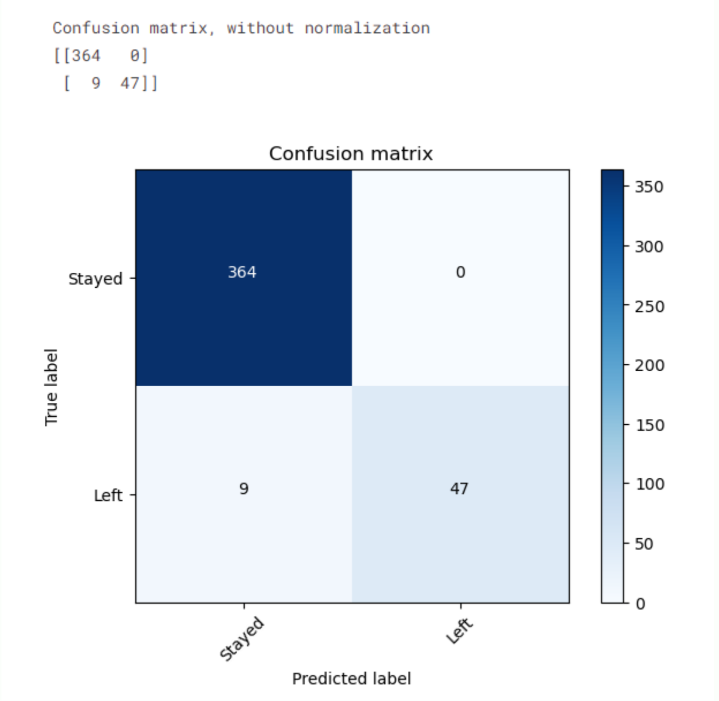
Confusion Matrix :

The confusion matrix helps visualize the performance of our model. Here's how to read it:

[[True Negatives (TN) | False Positives (FP)]
[False Negatives (FN) | True Positives (TP)]]

This means that:

364 employees were correctly predicted as not leaving the company (TN). 0 employees were incorrectly predicted as leaving the company (FP). 9 employees were incorrectly predicted as not leaving the company (FN). 47 employees were correctly predicted as leaving the company (TP).



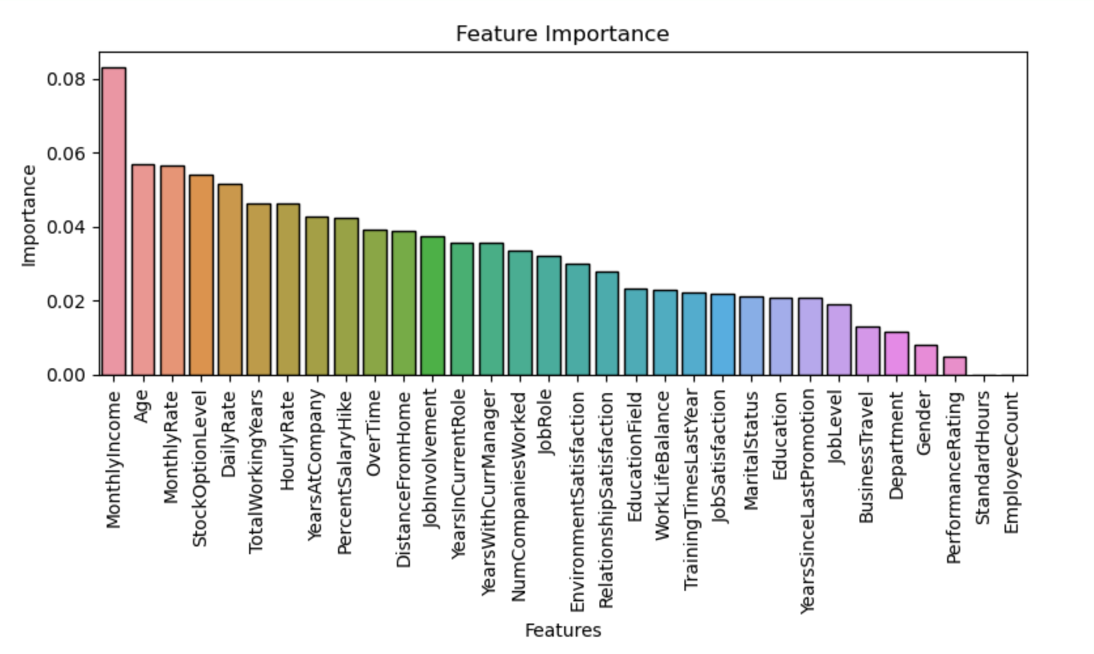
```
Confusion matrix, without normalization
[[364   0]
 [  9  47]]
```

Feature Importance :

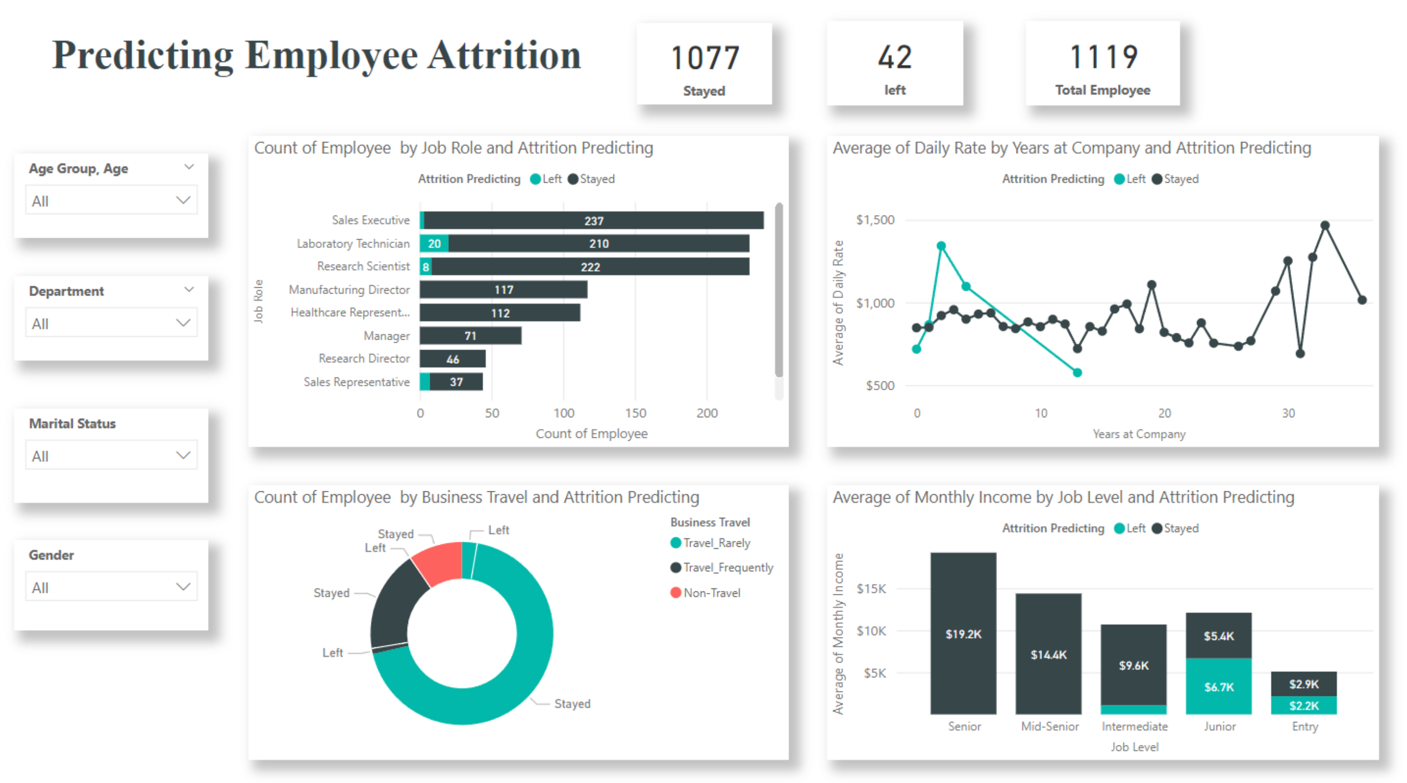
This section highlights the top factors that contribute to predicting employee attrition. The higher the importance value, the more significant the feature is for the prediction.

MonthlyIncome: 0.083045
Age: 0.056775
MonthlyRate: 0.056519

These three factors have the most significant impact on the model's predictions. It is essential to consider these factors when making decisions to reduce employee attrition.



Prediction Dashboard using Test data set:



Predicting Employee Attrition

1077 Stayed

42 left

1119 Total Employee

Age Group, Age: All

Department: All

Marital Status: All

Gender: All