IBM Employee Attrition Prediction Analysis Report

1. Introduction

Attrition is a major problem for all businesses regardless of location, size, and industry sector. A well-managed human resources department is critical to any organization. Companies spend a lot of time and money to nurture their most valuable asset: their employees. When employees leave, companies face huge losses. High turnover means high cost. Hence, it is important to predict turnover rates. The Human Resources Department would greatly benefit from the ability of data scientists to predict employee attrition risk and identify the underlying factors contributing to it, such as inadequate compensation, job dissatisfaction, and commuting distance. This insight will enable them to implement targeted support measures aimed at enhancing employee retention and safeguarding our valuable workforce. In this report, we summarize the findings and insights gained from analyzing the IBM HR Analytics Employee Attrition & Performance dataset.

The original dataset has 1470 rows and 35 columns. The final dataset has 1470 rows and 30 columns. I removed a total of 5 columns (target features: Attrition and 4 unneeded columns of uniform values: EmployeeCount, EmployeeNumber, StandardHours, and Over18).

After removing these columns, the dataset is left with 9 non-numeric columns and 21 numeric columns. It is clean, with no missing values or duplicates.

2. Dataset Analysis and Preprocessing

Dataset Description:

For this project, the dataset, IBM HR Analytics Employee Attrition and Performance, is a fictional dataset created by IBM data scientists to indicate if there is an attrition or not. The data, WA_Fn-UseC_-HR-Employee-Attrition.csv, is obtained from the Kaggle website. The dataset contains 1470 entries and 35 columns, which consist of 34 features and 1 target variable.

The 34 features include: Age, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Number of Companies Worked, Over 18, Overtime, Percentage of Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years at Company, Years in Current Role, Years Since Last Promotion, and Years with Current Manager.

The target variable is Attrition, which is labeled with either ‘yes’ or ‘no’. It is also important to note that there is an imbalance in the dataset with 84% of employees who stayed and 16% of employees who left. This class imbalance is taken into consideration when splitting the data into training and testing data by using stratification.

Data Exploration:

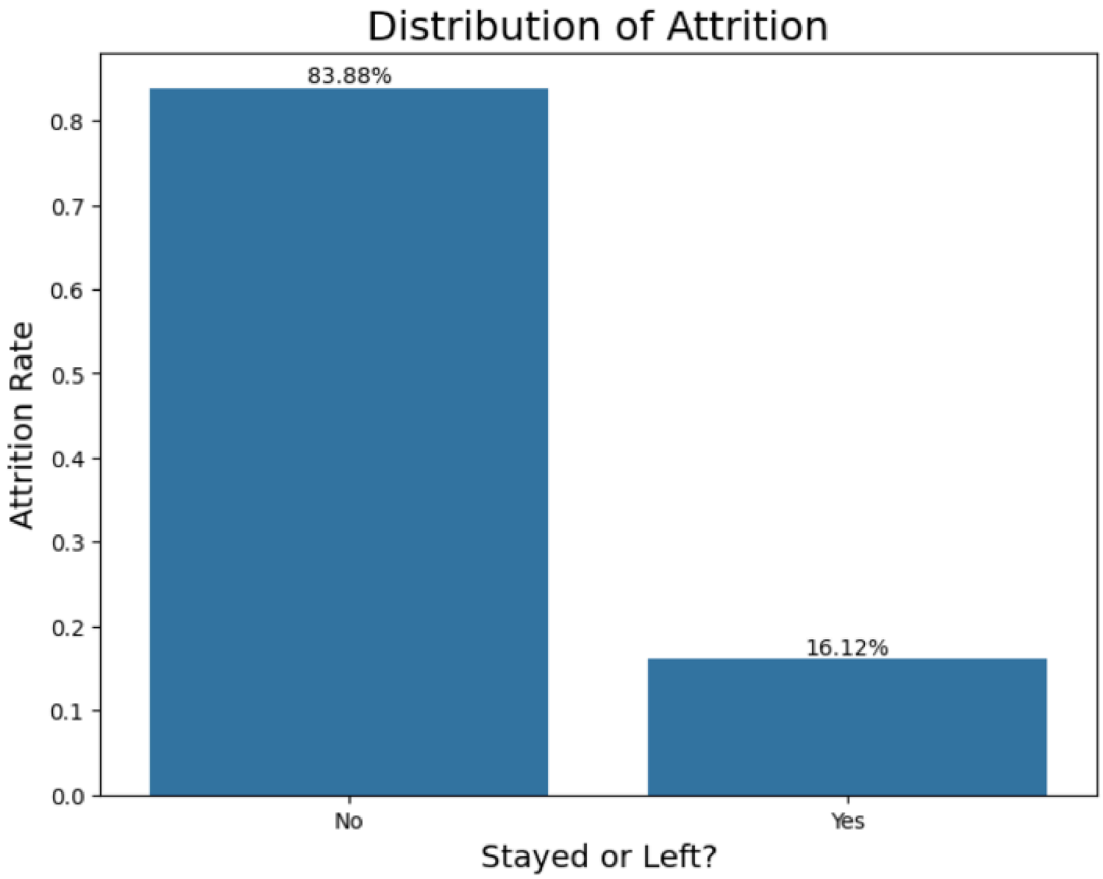
We explored the dataset to understand its structure, features, and distribution. This involved examining descriptive statistics, checking for missing values, and visualizing relationships between variables. Fortunately, our small dataset is pretty clean with no duplicates or missing values.

Preprocessing Steps:

We handled missing values, encoded categorical variables using one-hot encoding, and split the dataset into training and testing sets for model development.

2.1 Exploratory Data Analysis (EDA) and Discovery

First, we created a univariate chart of our target variable, Attrition, to look at its distribution.



As can be seen from the figure, we have an imbalanced dataset with 83.88% of employees choosing to stay (Attrition=No), while the remaining 16.12% chose to leave the company (Attrition=Yes). The imbalance should be addressed prior to model training to prevent overemphasis towards the Attrition='No' class when our study focus is on the employees who chose 'Yes'.

Then we created various bivariate and multivariate charts to see how each feature relates to the target variable, Attrition rate. We began by examining demographic data, including age, gender, marital status, academic field of study, and education level, etc.

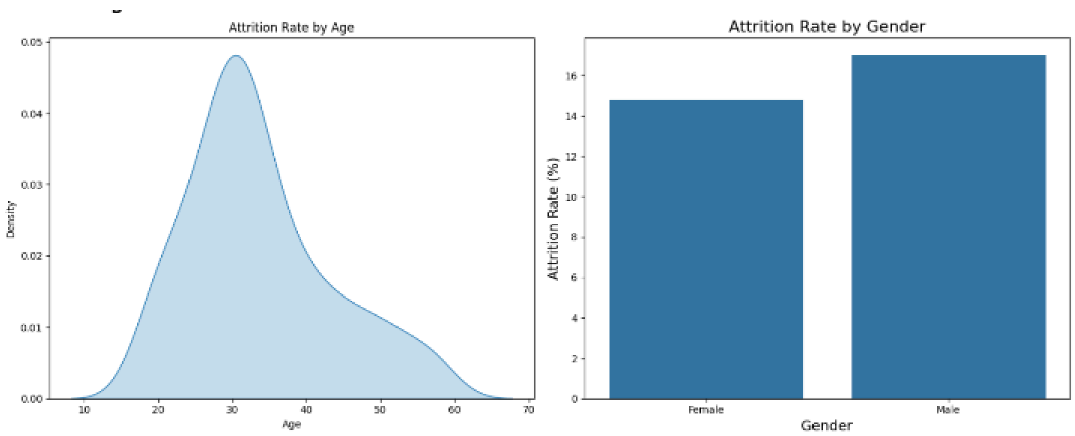
Attrition by Age:

The age range within this dataset is between 18 to 60. This gives a wide age range from young recent graduate employees who have little or no work experience to senior employees about to retire.

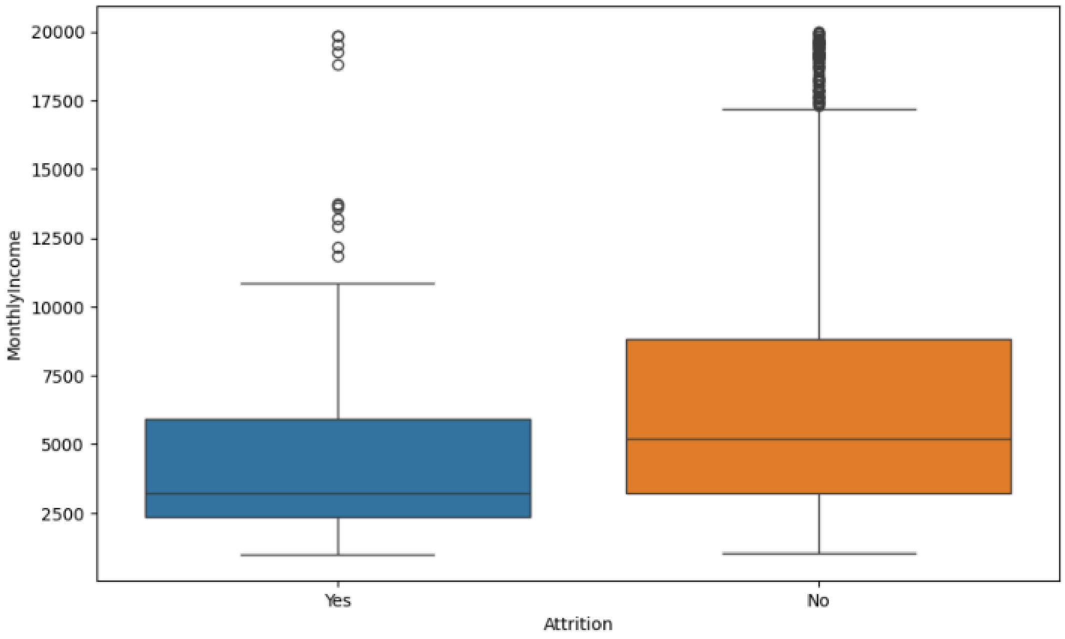
The younger employees, especially those in the 30-35 age group, appear to be more likely than other age groups to leave a company. This could be due to a number of factors, including a search for new experiences, dissatisfaction with salary or career path, or a more attractive job offer elsewhere. Older employees tend to have greater job stability. This may be due to a number of factors, such as a higher level of commitment to the company, the difficulty of finding a new job at an older age, or the existence of mandatory retirement benefits.

Attrition by Gender:

There is a significant difference in turnover rates between men and women. Male employees tend to leave more often than female employees. However, this could be due to the imbalance in the sample of both groups. Male employees are slightly more likely to attrite than female employees. This has no working research explanation and would warrant further investigation in the company. However, taking a look across employee data might suggest that males may have a higher tendency to score higher on other relevant features affecting attrition.

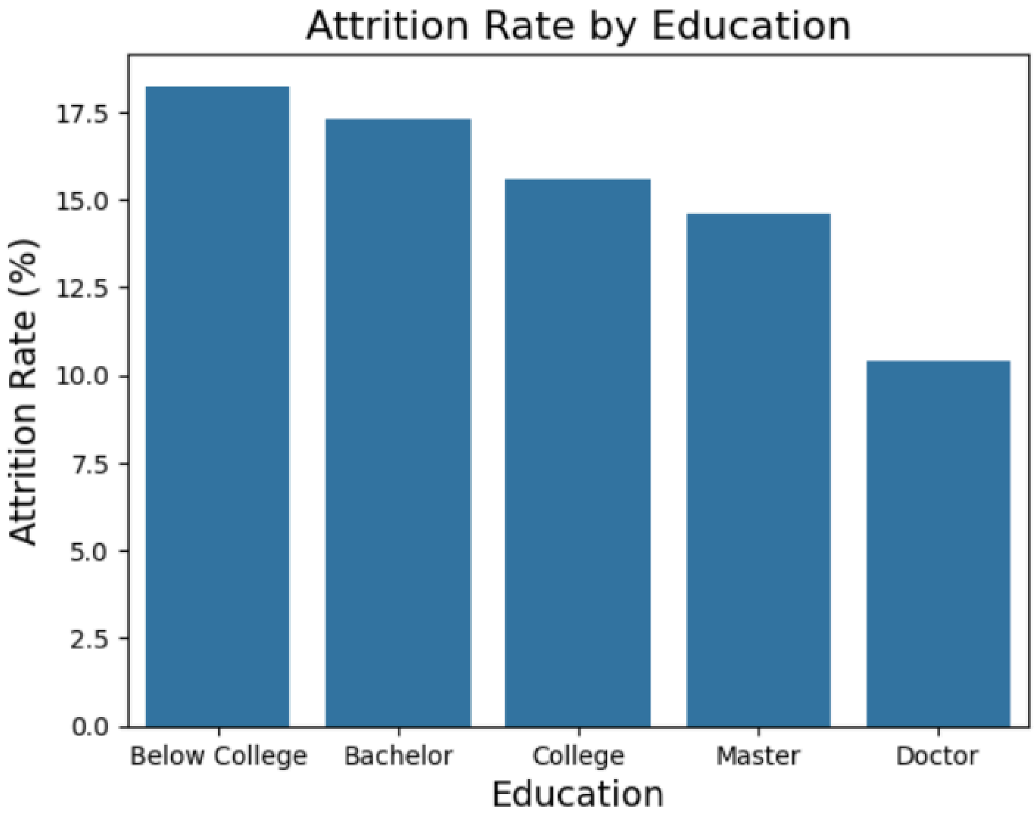


Attrition by Monthly Income: lower monthly income contributes to higher attrition rate



Attrition by Education:

Education includes five levels, from no college experience to having a master's or doctorate degree. Attrition rate breakdown by education level shows that the highest education (doctorate) has the lowest attrition rate, while below college has the highest attrition rate.

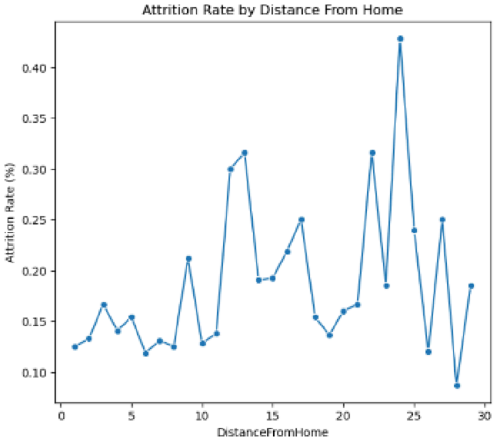


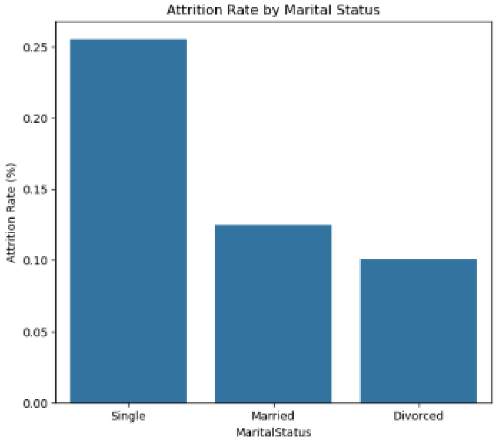
Attrition Rate by Marital Status:

The marital status includes whether an employee is single, married, or reportedly divorced. When analyzing marital status, we do see a greater proportion of single employees leaving the company versus those that are married or divorced. The lack of family commitment or other financial ties might influence single employees to look for other career opportunities elsewhere.

Attrition Rate by Distance From Home:

Commuting stress or a desire for proximity to home may play a role in influencing attrition. Although, it is interesting to discover that people who live close by have a higher attrition rate than those who live between 25-30 miles.





Attrition by Environment Satisfaction:

Concerning environment satisfaction, we do see a positive correlation between a reportedly low environmental satisfaction versus a medium, high, and very high. We see that the average attrition rate for medium, high, and very high is 14%. The attrition rate jumps to 25% with a low environmental rating. When the environment satisfaction is low, employees are likely to leave, and vice versa.

Attrition by Job Satisfaction:

Employees with low levels of job satisfaction tend to leave more often. This suggests that aspects of the job itself, such as tasks, responsibilities, and challenges, strongly influence an employee's decision to stay or leave. Job satisfaction is generally understood to be an important factor in employee attrition. After examining this variable, we see that the range between employees who report low job satisfaction and those that report very high job satisfaction, is 22.8% versus 11.3%. This is approximately double and is notably concerning, although it is not as drastically different as we might have expected. It is also important to note that while this variable is inherently indicative of an employee’s sentiment towards the job, it is also very difficult to control for. There can be an array of circumstances that affect an employee’s degree of satisfaction where, furthermore, while certain circumstances may affect one employee drastically, it may affect another employee very little.

Attrition by Job Involvement:

When the job involvement is high, the attrition is low and vice versa. In addition, the results of the analysis show a strong correlation between the level of job involvement and the level of turnover. Employees with low levels of job involvement tend to leave the organization more frequently. This suggests that a lack of job involvement, which may be caused by a lack of career development opportunities or a lack of challenge in the job, may encourage employees to seek more fulfilling work elsewhere. Looking at job involvement, we see that there is also a high correlation to attrition rate. Employees with low job involvement ratings (1 or lower) have an attrition rate of 33.5%. In stark contrast, those with very high job involvement ratings, the attrition rate hovers around 9%. If we average job involvement, combining medium, high, and very high, we see an attrition ratio of 13.6%. This is roughly two and a half times lower than employees with a low job involvement rating. Additionally, after oversampling for this variable, we found that low job involvement corresponded to an attrition rate of 70%. This is a significant finding.

Attrition by Environmental Satisfaction:

A work environment that is uncomfortable, unsupportive, or inconsistent with an employee's values may encourage them to seek employment elsewhere.

Attrition by Relationship Satisfaction:

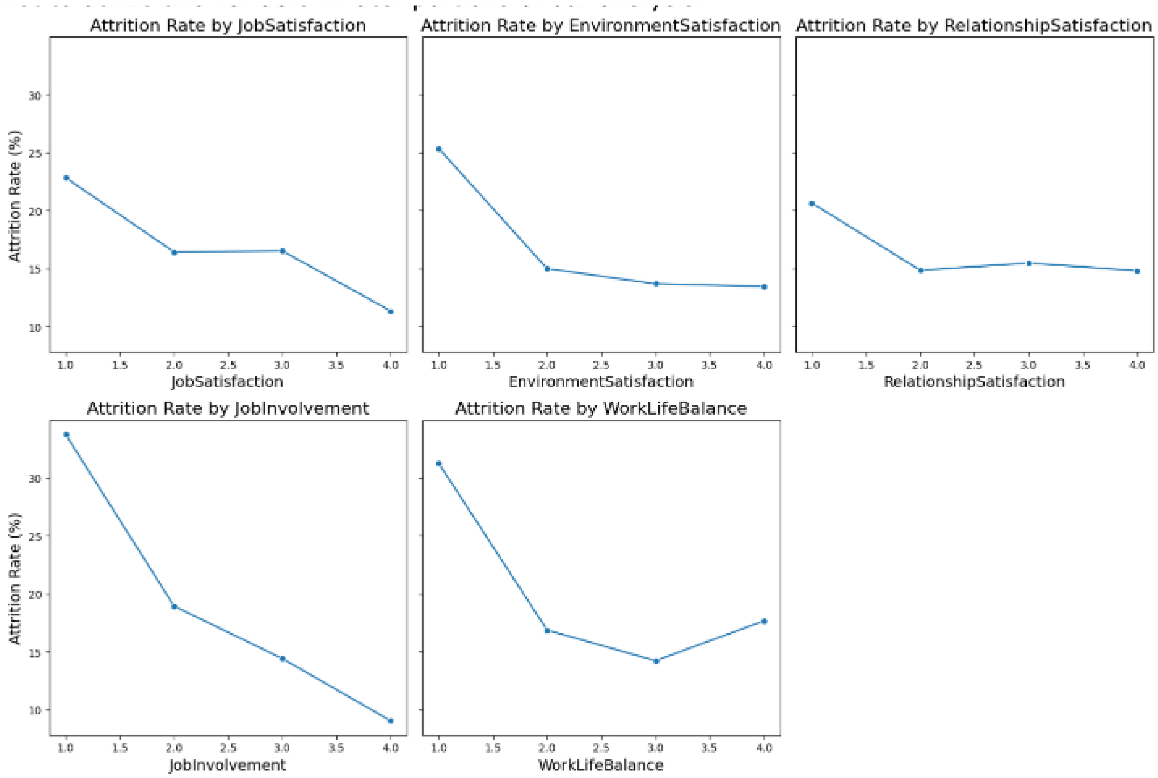
Good relationships with co-workers and supervisors can increase a sense of belonging and loyalty to the organization, thereby reducing turnover. Job Involvement: Employees who feel engaged in their work tend to be more loyal and committed to the organization.

We see that relationship satisfaction has a notable, though ambiguous connection to attrition rate. The attrition rate between when an employee who indicates a satisfaction of low versus very high is different by 9%. In comparison to job involvement, where there is a difference of 24% between the low and high end. Relationship status is concerning, yet to a lesser degree. By looking deeper at these two variables, we do find interdependent differences. For instance, when relationship satisfaction is low and job involvement is low, the attrition rate is 23.5%. When relationship satisfaction is low and job involvement is high, the attrition rate drops to 7%. When relationship satisfaction is very high and job involvement is low, the attrition rate is 43.5%. But when relationship satisfaction is high and job involvement is also high, the attrition rate is 12.5%. While we do notice this relationship, after reviewing the statistical correlation between the two variables, we see that it is only 0.03, which is very low. Additionally, relationship status is another variable that is difficult to control for, since there are many factors that go into determining an employee’s degree of satisfaction towards their manager or colleagues.

Attrition by Work-Life Balance:

A good work-life balance is very important to employees. Employees who feel that their work interferes with their personal lives are more likely to leave the company. Work-life balance is presumably an important factor in modern day company culture. This variable carries a rating of between 1 and 4, where 1 is equivalent to a bad work-life balance and a 4 is equivalent to the best work-life balance. Looking at the data, we find that the attrition rate when work-life balance is bad is 31.25%; furthermore, on average, when work-life balance is good, better, or best, the attrition rate is 16.2%. This is approximately a two-fold difference.

When reviewing the statistical correlation between work-life balance and attrition, we find that there is again, a very low correlation of 0.064. For this reason, and due to the fact that the dataset does not state what factors lead to a bad or a good work-life balance, we chose not to utilize this variable in later portions of our analysis.

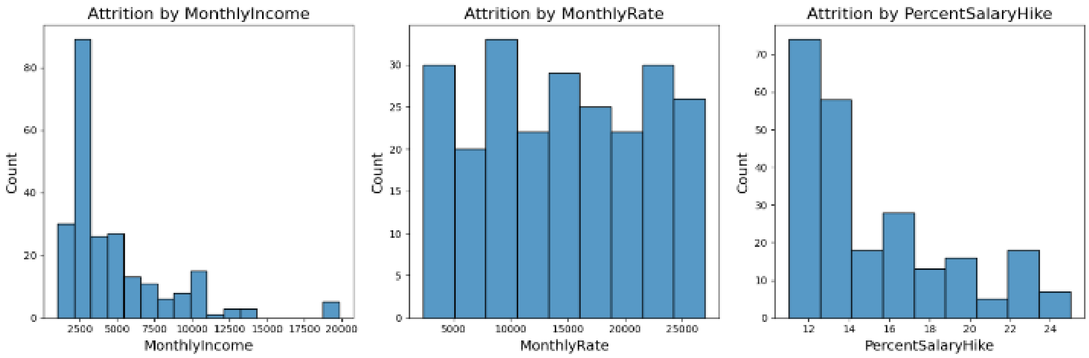


Performance rating in context to this dataset is not notably correlated to attrition. This seems mainly due to the fact that we do not have performance ratings of low and good available. The dataset only contains employees with a performance rating of excellent and outstanding. Between these two instances, we see the same attrition rate of 16%. We would note that while performance rating is typically a very good indicator of how an employee fits within a company, for our analysis, it is not practical to use within our analysis.

Lastly, we examined work-life balance, since this is presumably an important factor in modern day company culture. This variable carries a rating of between 1 and 4, where 1 is equivalent to bad work-life balance and a 4 is equivalent to the best work-life balance. Looking at the data, we find that the attrition rate when work-life balance is bad, is 31.25%; furthermore, on average, when work-life balance is good, better, or best, the attrition rate is 16.2%. This is approximately a two-fold difference.

When reviewing the statistical correlation between work-life balance and attrition, we find that there is again, a very low correlation of 0.064. For this reason, and due to the fact that the dataset does not state what factors lead to a bad or a good work-life balance, we chose not to utilize this variable in later portions of our analysis.

Looking at salary and benefit-related factors

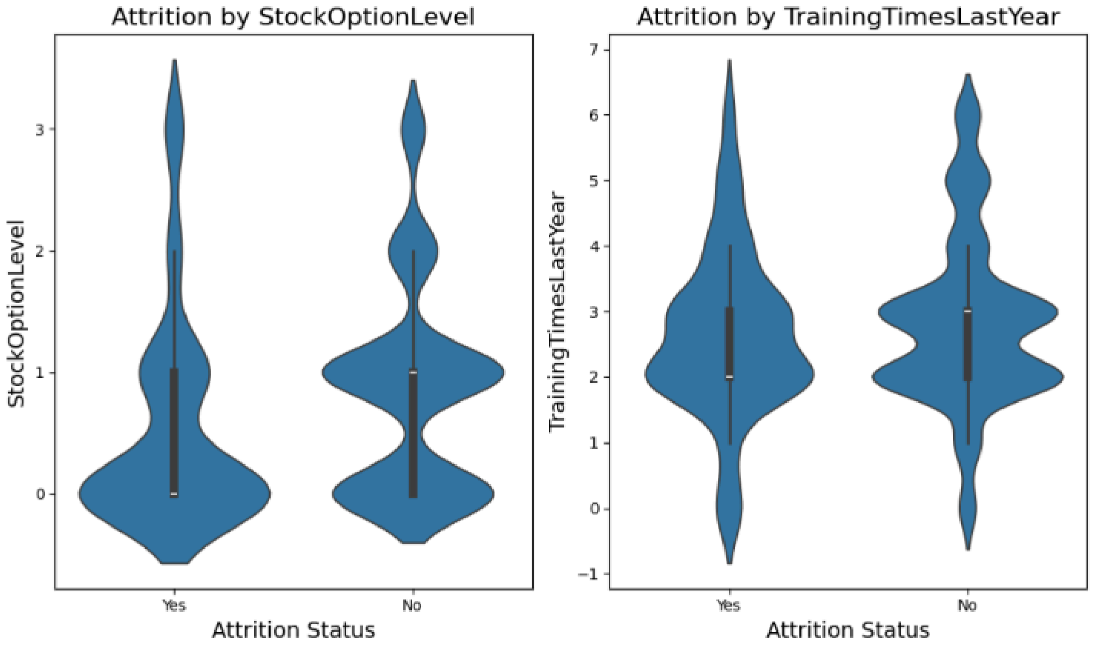


Turnover by Monthly Income:

This chart shows that most of the employees who left had a monthly income in the range of 5,000 to 7,500. There is a significant decrease in the turnover rate for employees with a monthly income above 7,500, indicating that employees with higher salaries tend to stay with the company longer. Turnover by Monthly Rate: The Turnover by Monthly Rate graph does not show a clear pattern between salary levels and turnover rates. Turnover fluctuates randomly across different salary ranges.

Turnover by Percent Salary Increase:

This chart shows that employees who receive lower salary increases (below 16%) tend to have higher turnover rates. The higher the percentage increase, the lower the turnover rate. This shows that a significant salary increase can be an effective retention factor.



Turnover by Stock Option:

Stock options have a positive impact on employee retention. Employees who own more shares tend to be more loyal and stay with the company longer.

Turnover by Training Benefit:

Training opportunities also play an important role in employee retention. Employees who have more training opportunities tend to be happier and more motivated to stay with the company.
Attrition by Performance Rating:

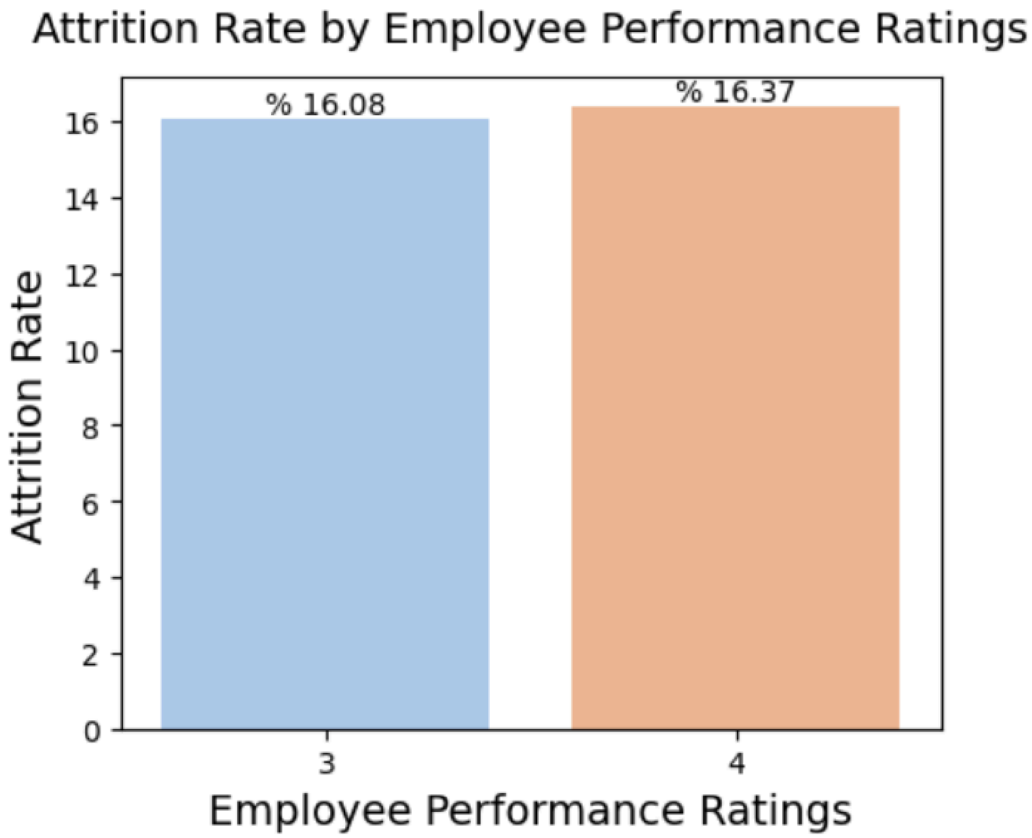
Hypothesis: Performance rating has a significant effect on attrition

1. Contrary to the normal belief that employees having higher ratings will not leave the company, it may be seen that there is no significant difference between the performance rating and Attrition Rate.

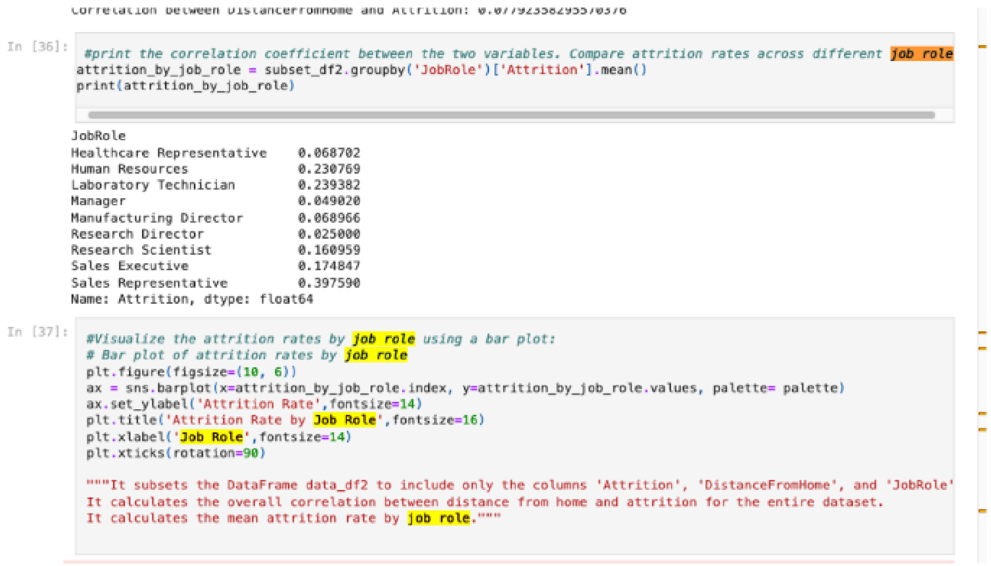
Code:

```
# Relationship between performance ratings and attrition
# Calculate the correlation coefficient between 'Attrition' and 'PerformanceRating'
correlation_coefficient = round(data_df2['Attrition'].corr(data_df2['PerformanceRating']), 3)
print(f"Correlation coefficient: {correlation_coefficient}")
```

Correlation coefficient: 0.003 Correlation coefficient: 0.003 is too close to zero, and it means a weak correlation between Employee Performance Ratings and Attrition.



Attrition by Job Role: People working in the Sales department are most likely to quit the company followed by Laboratory Technicians and Human Resources their attrition rates are 40%, 24%, and 22% respectively



```
Correlation between DistanceFromHome and Attrition: 0.07792358295570376
```

In [36]:

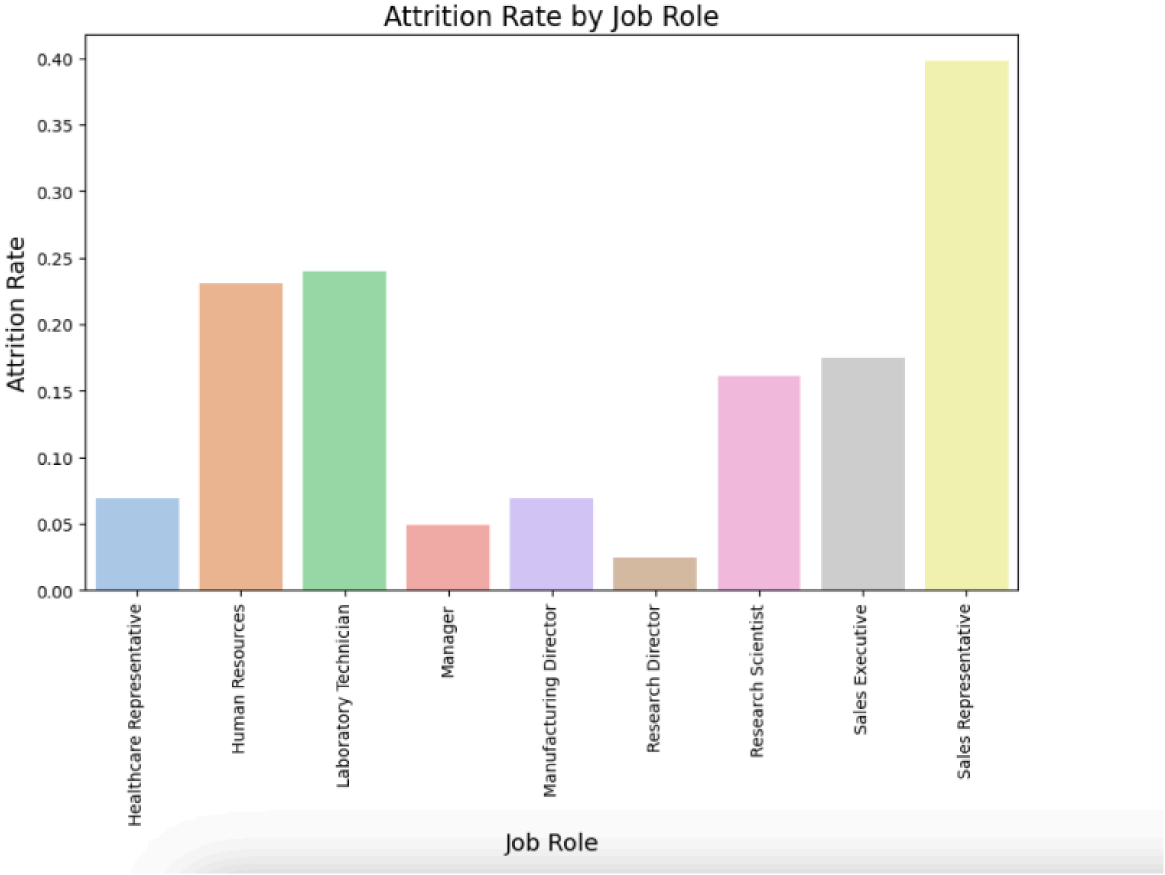
```
#print the correlation coefficient between the two variables. Compare attrition rates across different job role
attrition_by_job_role = subset_df2.groupby('JobRole')['Attrition'].mean()
print(attrition_by_job_role)
```

```
JobRole
Healthcare Representative    0.068702
Human Resources              0.230769
Laboratory Technician        0.239382
Manager                      0.049020
Manufacturing Director       0.068966
Research Director            0.025000
Research Scientist           0.160959
Sales Executive              0.174847
Sales Representative         0.397590
Name: Attrition, dtype: float64
```

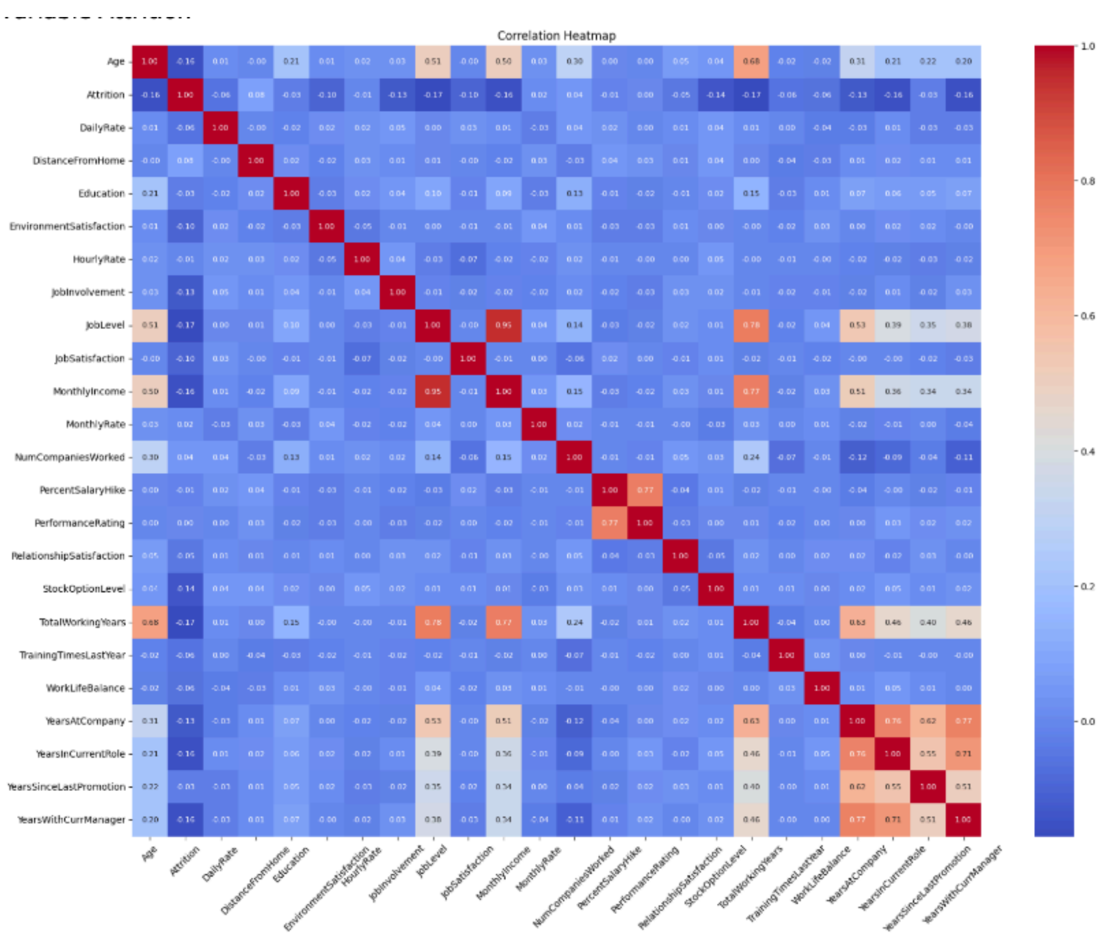
In [37]:

```
#Visualize the attrition rates by job role using a bar plot:
# Bar plot of attrition rates by job role
plt.figure(figsize=(10, 6))
ax = sns.barplot(x=attrition_by_job_role.index, y=attrition_by_job_role.values, palette= palette)
ax.set_ylabel('Attrition Rate',fontsize=14)
plt.title('Attrition Rate by Job Role',fontsize=16)
plt.xlabel('Job Role',fontsize=14)
plt.xticks(rotation=90)

"""It subsets the DataFrame data_df2 to include only the columns 'Attrition', 'DistanceFromHome', and 'JobRole'
It calculates the overall correlation between distance from home and attrition for the entire dataset.
It calculates the mean attrition rate by job role."""
```

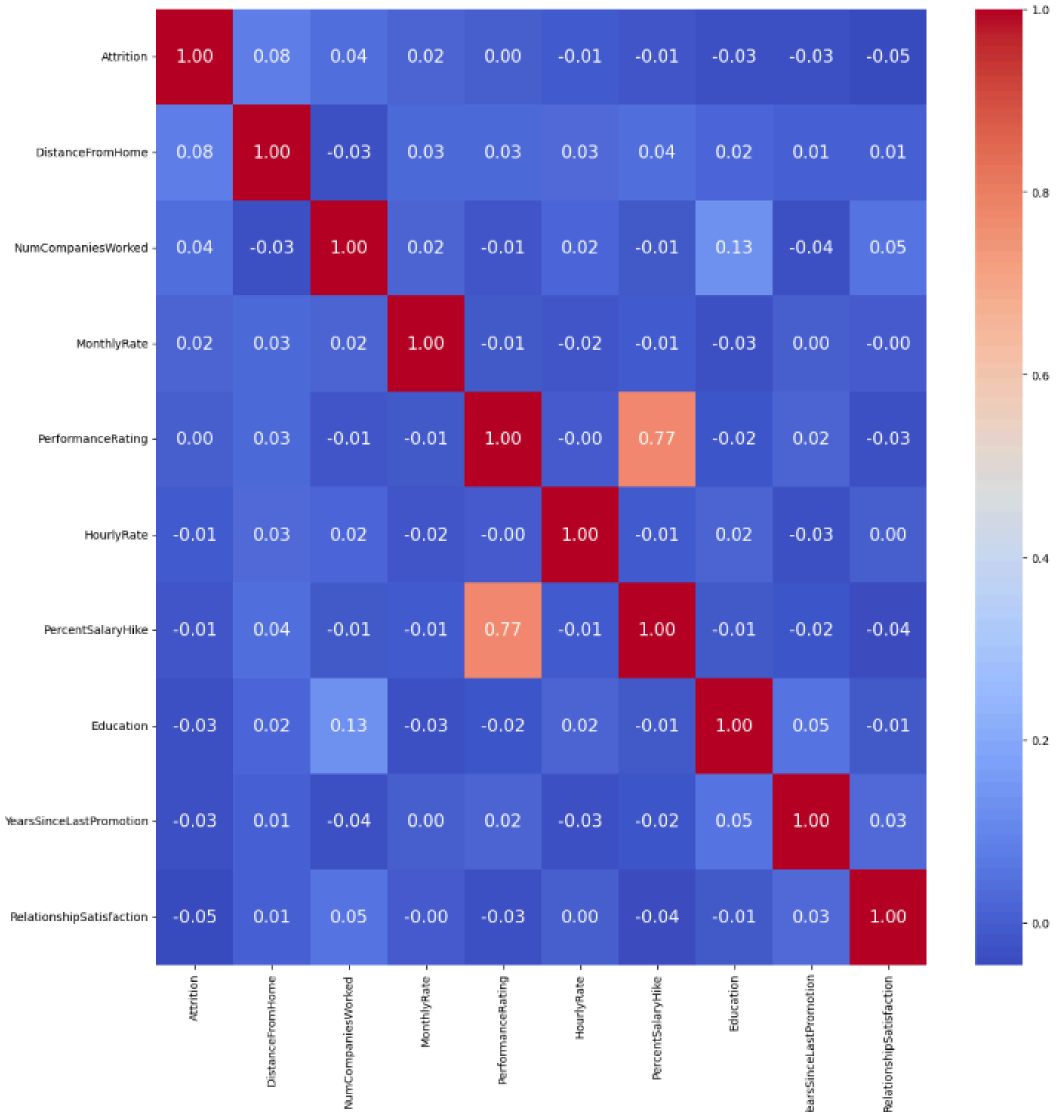


Dependent Correlation: We created a correlation matrix heatmap to see how the dependent variables relate to our target variable Attrition.



Here we discover some highly correlated features: "YearsWithCurrManager" has a high positive correlation with 'YearsAtCompany'(0.77), and 'YearsInCurrentRole' (0.71) -- this may indicate employees tend to stay with the same manager in the same role. For the above reason, using a cutoff of 0.7 correlation coefficient, the columns MonthlyIncome, TotalWorkingYears, YearsInCurrentRole, and YearsWithCurrManager will be removed, and the columns JobLevel and YearsAtCompany will be retained. It will reduce the possibility of features multicollinearity and will improve model accuracy as well. For MonthlyIncome, the high correlation with job level and sensitivity around salary makes it both redundant and hard to work with in policy change management. For TotalWorkingYears, the data will be encapsulated in other features like YearsAtCompany. For YearsInCurrentRole and YearsWithCurrManager, the data will not pose much additional insights beyond YearsAtCompany, considering organisational structures/functional specializations (e.g. Accounting) are what keep individuals in similar Departments and roles.

We also created a correlation matrix heatmap for the top 10 continuous features:



	Attrition	DistanceFromHome	NumCompaniesWorked	MonthlyRate	PerformanceRating	HourlyRate	PercentSalaryHike	Education	earsSinceLastPromotion	RelationshipSatisfaction
Attrition	1.00	0.08	0.04	0.02	0.00	-0.01	-0.01	-0.03	-0.03	-0.05
DistanceFromHome	0.08	1.00	-0.03	0.03	0.03	0.03	0.04	0.02	0.01	0.01
NumCompaniesWorked	0.04	-0.03	1.00	0.02	-0.01	0.02	-0.01	0.13	-0.04	0.05
MonthlyRate	0.02	0.03	0.02	1.00	-0.01	-0.02	-0.01	-0.03	0.00	-0.00
PerformanceRating	0.00	0.03	-0.01	-0.01	1.00	-0.00	0.77	-0.02	0.02	-0.03
HourlyRate	-0.01	0.03	0.02	-0.02	-0.00	1.00	-0.01	0.02	-0.03	0.00
PercentSalaryHike	-0.01	0.04	-0.01	-0.01	0.77	-0.01	1.00	-0.01	-0.02	-0.04
Education	-0.03	0.02	0.13	-0.03	-0.02	0.02	-0.01	1.00	0.05	-0.01
YearsSinceLastPromotion	-0.03	0.01	-0.04	0.00	0.02	-0.03	-0.02	0.05	1.00	0.03
RelationshipSatisfaction	-0.05	0.01	0.05	-0.00	-0.03	0.00	-0.04	-0.01	0.03	1.00

These are the top 10 factors influencing the attrition rate: distanceFromHome, NumCompaniesWorked, MonthlyRate, PerformanceRating, HourlyRate, PercentSalaryHike, Education, YearsSinceLastPromotion, RelationshipSatisfaction

Data Preprocessing

Data preprocessing is necessary prior to running modeling due to constraints in the modeling process. For instance, certain models cannot process categorical data. For this, we need to convert the categorical features into a numerical representation. We use the get_dummies method to convert 7 categorical features in our dataset to int data type by creating 7 dummy vars. We use drop_first = True to reduce multicollinearity.

In [47]:

```
dummy_BusinessTravel = pd.get_dummies(data_df['BusinessTravel'],prefix ='BusinessTravel')
dummy_Department = pd.get_dummies(data_df['Department'],prefix ='Department')
dummy_EducationField = pd.get_dummies(data_df['EducationField'],prefix ='EducationField')
dummy_Gender = pd.get_dummies(data_df['Gender'],prefix ='Gender',drop_first=True)
dummy_JobRole = pd.get_dummies(data_df['JobRole'],prefix ='JobRole')
dummy_MaritalStatus = pd.get_dummies(data_df['MaritalStatus'],prefix ='MaritalStatus')
dummy_OverTime = pd.get_dummies(data_df['OverTime'],prefix ='OverTime',drop_first=True)
```

We also drop 4 irrelevant columns: Over18, EmployeeCount, EmployeeNumber, and StandardHours

---

3. Model Development

We experimented with three machine learning algorithms for discrete binary classification, including Logistic Regression, Random Forest, and Gradient Boosting Classifier. We evaluated each model's performance using metrics such as accuracy, precision, recall, F1-score as well as AUC-ROC score on the test data. Optimization techniques like hyperparameter tuning were applied to improve model performance.

Model Performance Metrics: Model performance metrics are used to evaluate the performance of a machine learning model. Because the dataset appears to be imbalanced, with attrition rates as low as 16%, selecting the right model performance measure is critical. As a result, model accuracy alone cannot determine the robustness of a machine learning model. Based on a confusion matrix created for training dataset predictions:

- **Accuracy** is defined as the number of correct predictions made by the machine learning model divided by the total number of datapoints. The best accuracy is 100 percent, which indicates that all predictions are correct. Given our dataset's conversion rate of 16%, accuracy is not a valid measure of model performance. Even if all of our predictions are incorrect, our model's accuracy would still be 84%. As a result, additional model performance measures are included.

Logistic Regression: A logistic regression model is used for predicting classes using the probability of the target variable. Unlike linear regression, which uses expected values of the response model, logistic regression uses the probability or odds of the response variable to model based on the combination of values taken by the predictors. This model uses the sigmoid function that maps predicted values to probabilities. It works well on linearly separable classes with easy implementation, making it a popular choice for classification problems.

There are two types of logistic regression models for classification: binary and multinomial. Binary logistic regression requires a dependent variable with only two possible outcomes whereas a multinomial requires three or more outcomes. In this case, the dataset is working with binary logistic regression since the target variable is binary (Yes, No). Logistic regression is applicable to this problem since we want to predict the probabilities and classify the employees into two categories (Yes, No) based on the explanatory variables. For the solver in the logistic regression model, liblinear is picked since it supports both L1 and L2 regularization. Logistic Regression model cannot handle categorical variables unlike decision trees. We used the get_dummies method to convert categorical vars to numerical. After encoding the categorical values, the entire dataset scaled using StandardScaler. This is especially useful since the data has varying scales and it prevents the algorithms like linear regression from making assumptions that the data has a Gaussian distribution. We first tried to build a logistic regression model, which is specifically designed for binary classification problems and is the most straightforward model in our case. And we also performed hyperparameter tuning via GridSearchCV to improve the result of our model.

Best Parameters for Logistic Regression: {'C': 0.5, 'penalty': 'l1'} **Accuracy Tuned for Logistic Regression:** 0.8843537414965986

Confusion Matrix Tuned for Logistic Regression:

```
[[359, 11]
[ 40, 31]]
```

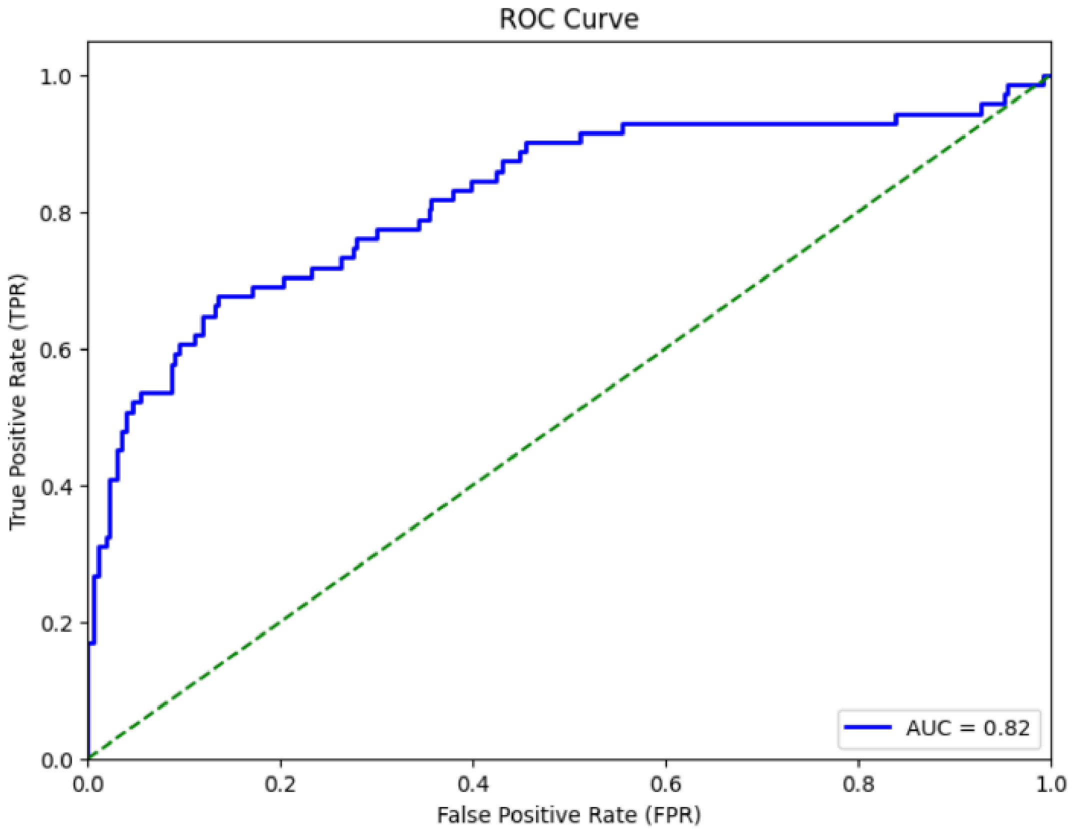
Classification Report Tuned for Logistic Regression:

	Precision	**Recall**	**F1-score**	**Support**
0	0.90	0.97	0.93	370
1	0.74	0.44	0.55	71
Accuracy			0.88	441
Macro Avg	0.82	0.70	0.74	441

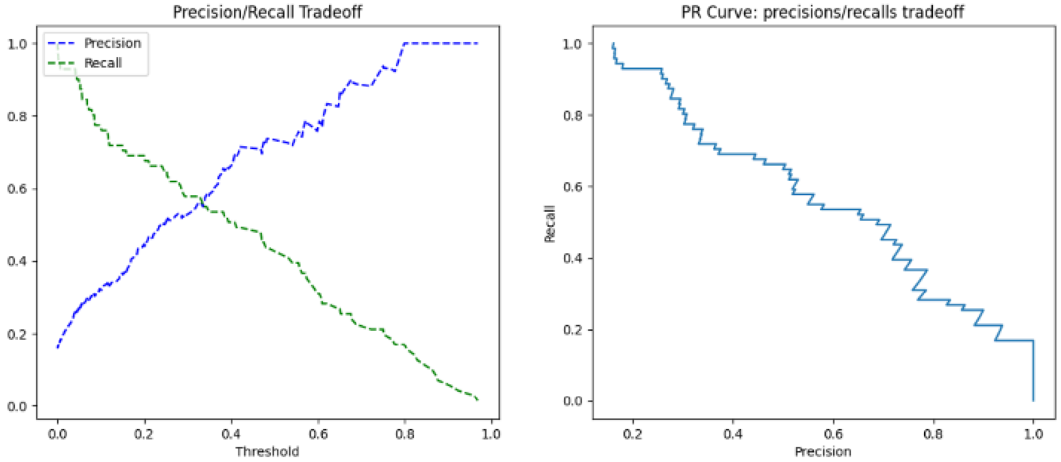
Weighted Avg	0.87	0.88	0.87	441

The tuned Logistic Regression model has an F1-score of 0.55 and AUC=0.82 for attrition="Yes". Considering only a small percentage of employees left the company, the Linear Regression model is doing a good job in predicting attrition. We have an imbalanced dataset. The F1 score is generally better than the area under the curve (AUC) for imbalanced datasets when the minority class is of interest. The F1 score is the harmonic mean of precision and recall, which balances the importance of both metrics. It's a more robust evaluation metric than accuracy because it gives a fair representation of a model's performance despite class imbalance.

AUC The AUC is a single value that summarizes a model's overall performance. It's useful for comparing the performance of multiple models. However, the AUC and ROC curve may not be well-suited for imbalanced problems because they may be biased toward the majority class. The accuracy score may also be less useful here as there is lesser emphasis on the minority class (Attrition - 'Yes') with better results in the majority class. With the relatively small sample and lack of longitudinal employee data, these scores are pretty decent in aiding the company to predict Attrition and work on tailored intervention.



AUC-ROC Score: 0.8217738865626188



Next, we examined the data through a different model - Random Forest. Considering the numerous correlations across the features, the low F1 score in the logistic regression may be attributed to more non-linear and complex relationships across features and Attrition. There may also be more noisy data in this case as our hyperparameter tuning did not pose much benefit to the linear regression. This prompted us to utilize the Random Forest model to see if we can deal better with the possible presence of complex relationships and noisy data.

RANDOM FOREST: A Random Forest Classifier model is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Trees in the forest use the best split strategy, i.e., equivalent to passing splitter="best" to the underlying DecisionTreeRegressor. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise, the whole dataset is used to build each tree. Here we went straight to tune the model by performing hyperparameter tuning via GridSearchCV to improve the result.

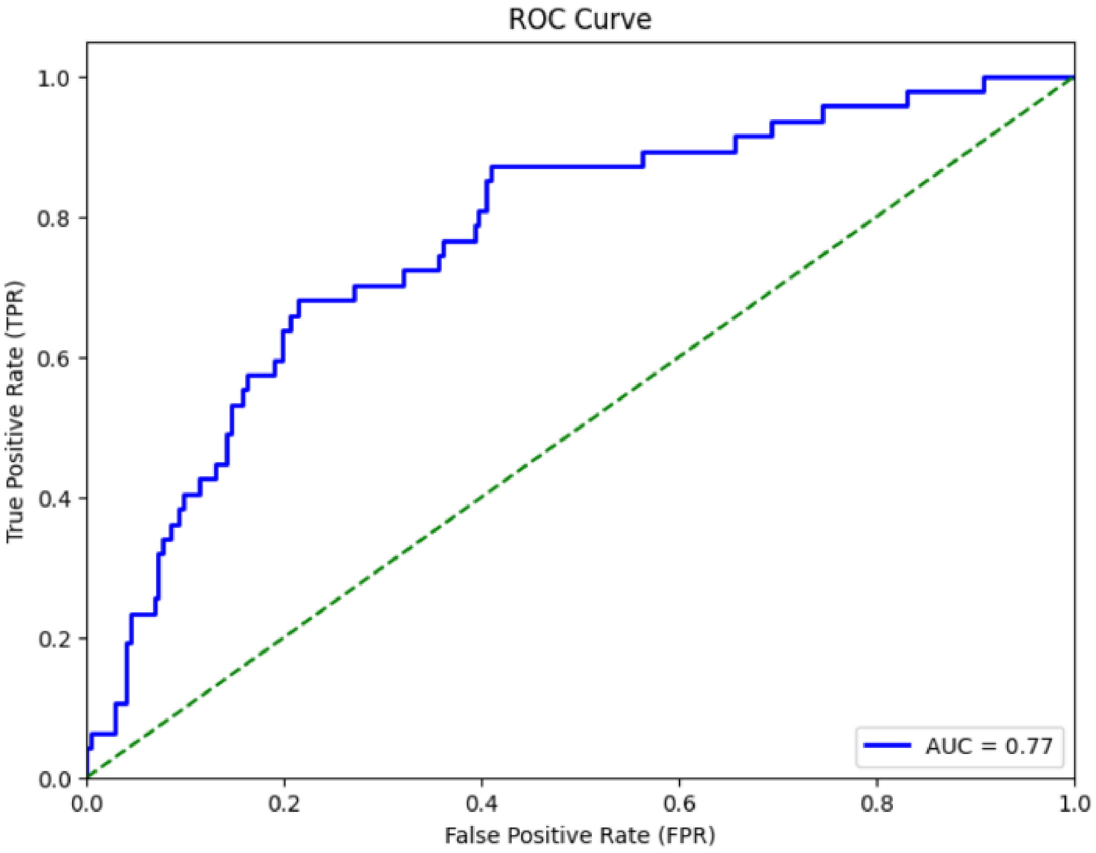
Best Parameters (Random Forest): {'max_depth': 15, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500}
Accuracy (Tuned Random Forest): 0.8299319727891157

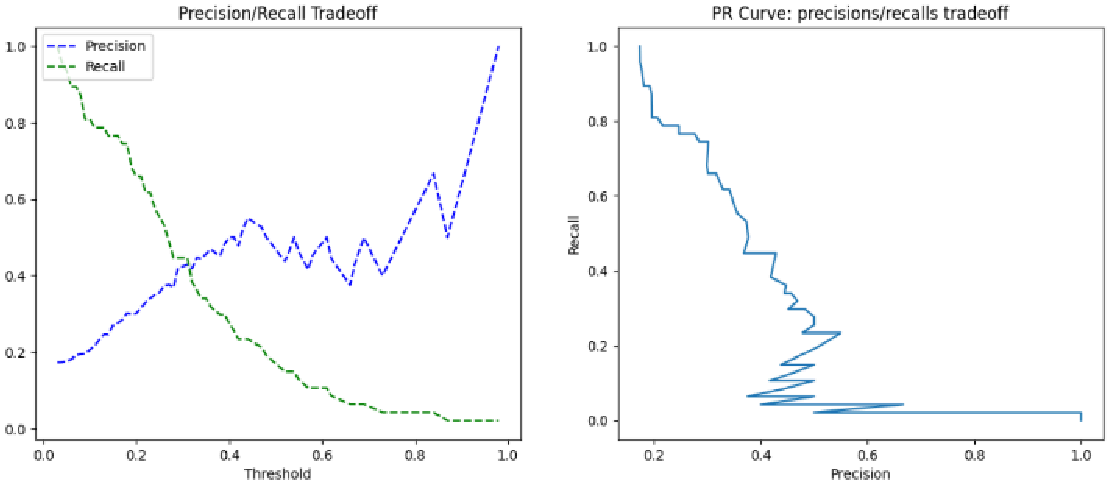
Confusion Matrix (Tuned Random Forest):
[[239, 8]
[42, 5]]

Classification Report (Tuned Random Forest):

	Precision	**Recall**	**F1-score**	**Support**
0	0.85	0.97	0.91	247
1	0.38	0.11	0.17	47
Accuracy			0.83	294
Macro Avg	0.62	0.54	0.54	294
Weighted Avg	0.78	0.83	0.79	294

The tuned Random Forest model (with StratifiedKFold) has an F1-score of 0.17 and the AUC-ROC score of 0.77. Interestingly, the hyperparameter-tuned Random Forest model did not perform as well as the logistic regression with an F1 score of 0.17 for Attrition['Yes'] and an AUC score of 0.77. The tuned Logistic Regression model has an AUC score of 0.81. While we hypothesized more complex and non-linear relationships across features and Attrition, the relatively poor performance of Random Forest in this context suggests that the underlying data may not exhibit strong non-linear dependencies.





Lastly, we will seek to improve the prediction modeling through Gradient Boosting method. As GB builds trees sequentially by weighing misclassified samples more heavily, it is possible that the model will provide higher predictive stats for Attrition.

GRADIENT BOOSTING CLASSIFIER: Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met. Gradient boosting is popular because it can handle complex relationships in data and protect against overfitting. Here we went straight to tune the model by performing hyperparameter tuning via GridSearchCV to improve model result

Best Parameters (Gradient Boosting): {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 200}
Accuracy (Tuned Gradient Boosting): 0.8401360544217688

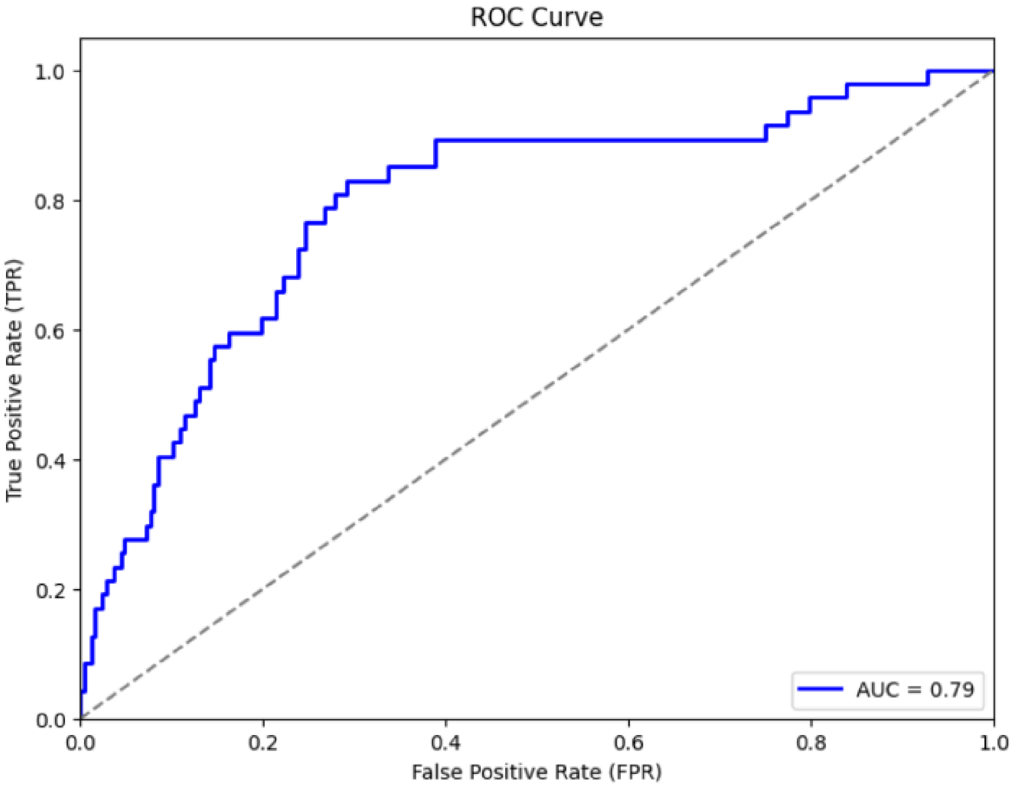
Confusion Matrix (Tuned Gradient Boosting):

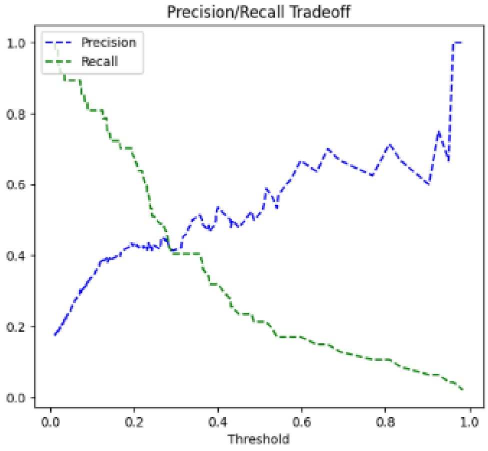
```
[[236, 11]
[ 36, 11]]
```

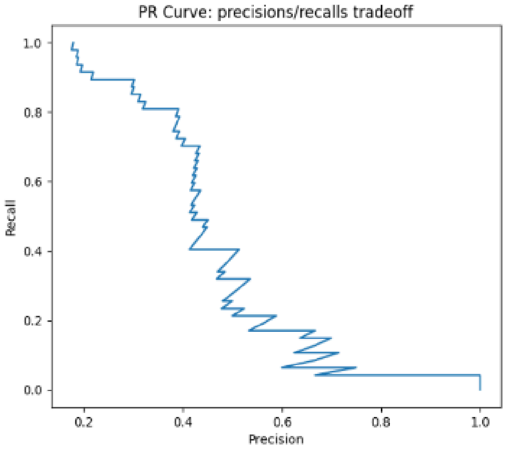
Classification Report (Tuned Gradient Boosting):

	Precision	**Recall**	**F1-score**	**Support**
0	0.87	0.96	0.91	247
1	0.50	0.23	0.32	47
Accuracy			0.84	294
Macro Avg	0.68	0.59	0.61	294
Weighted Avg	0.81	0.84	0.82	294

The hyperparameter-tuned Gradient Boosting model has an F1-score of 0.32, lower than the Logistic Regression model but higher than the Random Forest. The AUC-ROC score is 0.79. The GB model did not perform as well as the logistic regression with an F1 score of 0.50 for Attrition['Yes'] and an AUC score of 0.78. It did, however, do better than the Random Forest model, likely due to the sequential learning algorithm. Ensemble methods tend to excel with non-linear relationships and interactions between features. The relatively poor performance of Random Forest and Gradient Boosting in this context suggests that the underlying data may indeed not exhibit strong non-linear dependencies.

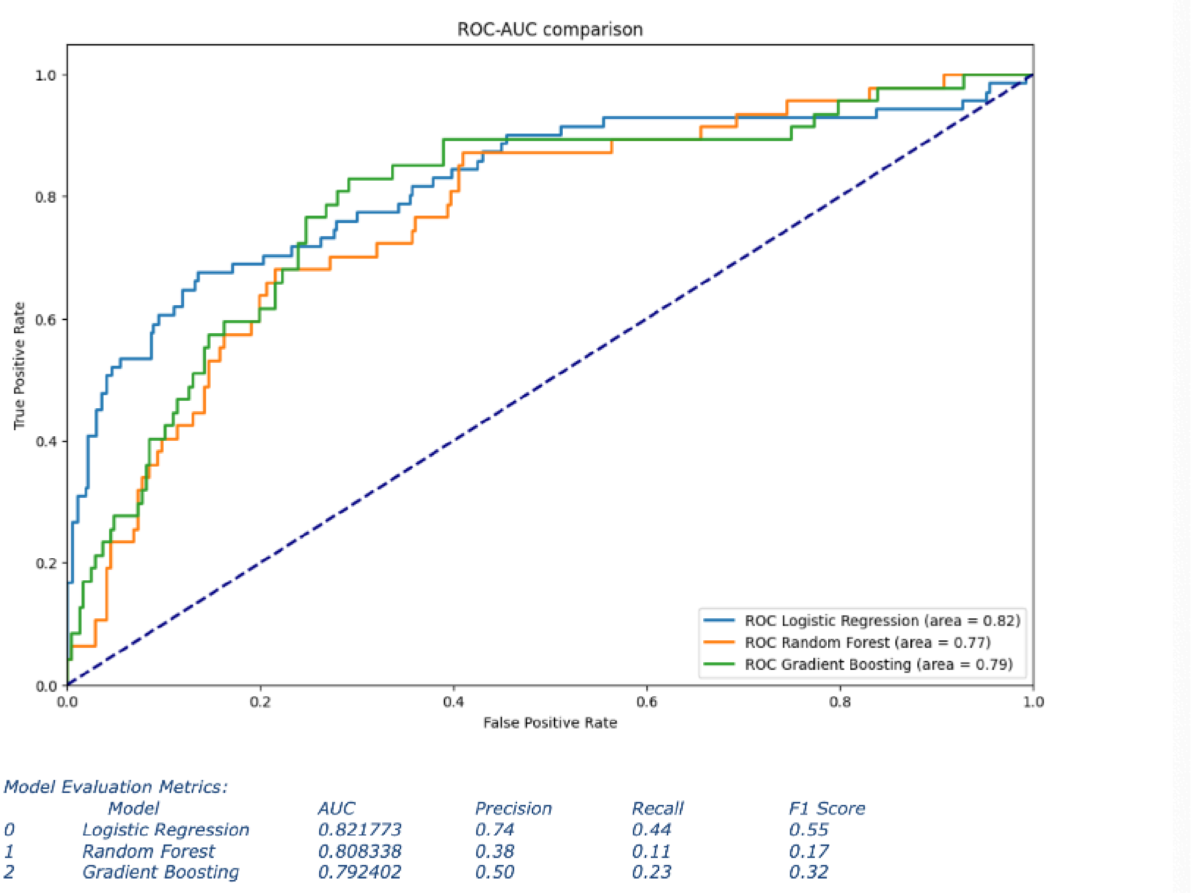






With the analysis of this binary classification problem through 3 different models and hyperparameter tuning, we will look at the results collectively in the next section.

Model Result and Comparison:



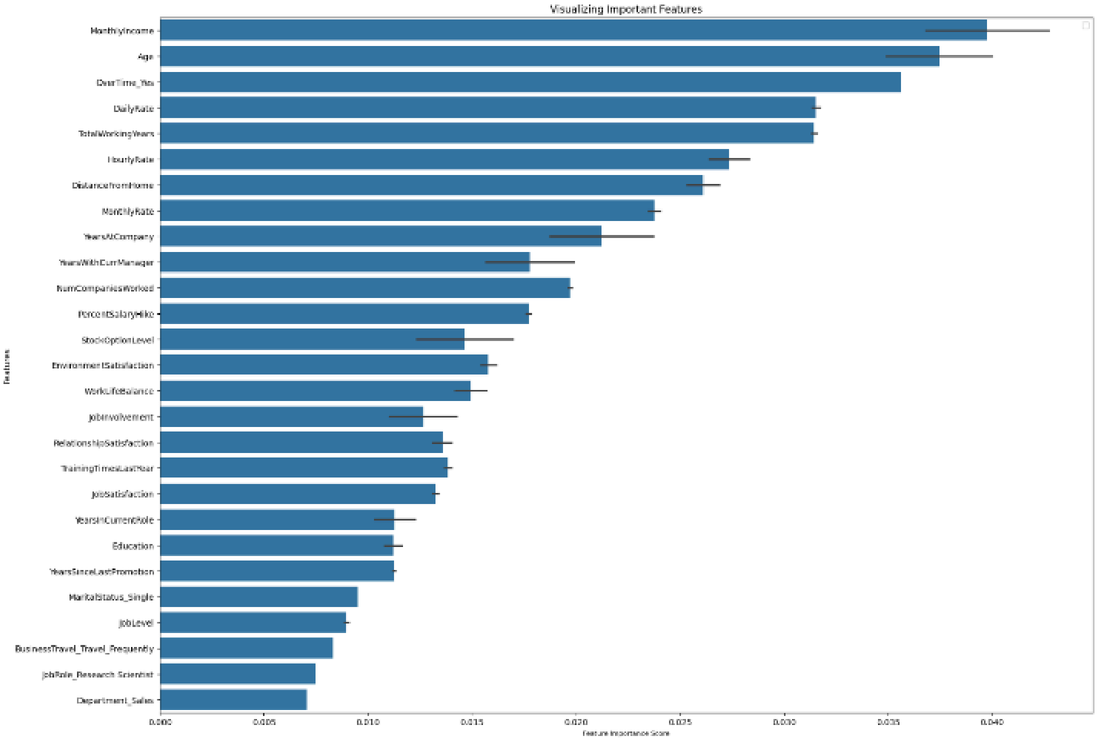
```
Model Evaluation Metrics:
             Model               AUC        Precision    Recall    F1 Score
0      Logistic Regression    0.821773     0.74         0.44      0.55
1      Random Forest          0.808338     0.38         0.11      0.17
2      Gradient Boosting      0.792402     0.50         0.23      0.32
```

Looking at the data collectively, the Logistic Regression model has the highest AUC score We can tell again that the logistic regression model emerged as the top performer, achieving an F1 score of approximately 0.55 for Attrition['Yes'] and an AUC of 0.82. This outcome suggests that there is likely the existence of linear relationships between Attrition and the features and Attrition is less likely to depend on non-linear or more complex relationships with other features. Also, when dealing with imbalanced datasets, it is important to point out again that the F1 score can be a better metric than AUC. This is because the F1 score balances precision and recall and is less affected by class imbalance. While accuracy score is more commonly used, it is less useful in this context as there is lesser emphasis on the minority class (Attrition - 'Yes') due to the overemphasis of better results in the majority class. Overall, considering the relatively small sample of Attrition['Yes'] data and lack of longitudinal employee data, the results from the logistic regression model are satisfactory and decent in aiding the company to predict Attrition and work on tailored interventions. On the other hand, Random Forest and Gradient Boosting algorithms may not have shined here because of the relatively small/moderate-sized dataset. For these 2 ensemble methods, it is likely that we will see better predictive statistics when there is a larger dataset in order to fully exploit

their capabilities. In our context, their complexity in building the model might have led to overfitting issues with limited data size.

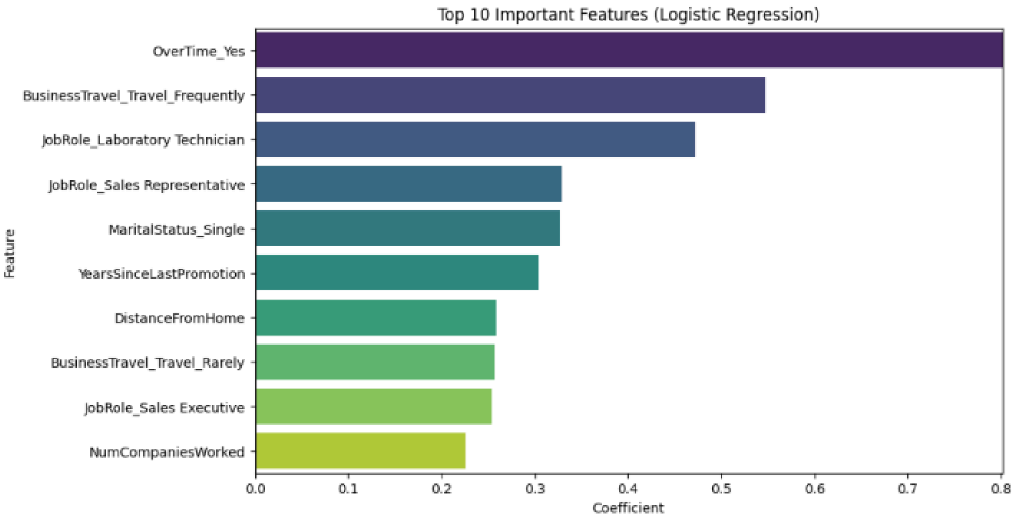
Features Importance:

Models like Random Forest can provide valuable insights into feature importance and their contributions to predicting attrition. Here are the top 10 most important features according to Random Forest



Factors such as MonthlyIncome, Age, Overtime, Distance, work-life balance, job satisfaction, etc. were identified as significant predictors of employee attrition As the logistic regression model emerged as the winner in our dataset context, we plotted the top 10 features from the model for the prediction on attrition. Below, we listed some explanation for each of the feature's possible causality in attributing attrition. We will also leave out HourlyRate and MonthlyRate due to their relatively small coefficient which reflects little to no influence vis-à-vis other features on Attrition in real-world settings. Overtime_Yes 0.801547): Overtime work may lead to burnout and decreased job satisfaction, which can contribute to attrition. PerformanceRating (0.417478): Surprisingly, better performance rating is associated with a higher likelihood of attrition. This, however, suggests that top employees may have attrited in light of better prospects for their careers elsewhere. BusinessTravel_Travel_Frequently (0.547421): Frequent travel may disrupt work-life balance, possibly contributing to attrition. NumCompaniesWorked (Coefficient:

0.322271): Employees who have worked in more companies may be more willing to seek outside opportunities in line with their working history. YearsSinceLastPromotion (0.264568): Lack of career advancement may lead to employee dissatisfaction and contribute to attrition. Gender_Male (0.071276): Male employees are slightly more likely to attrite than female employees. This has no working research explanation and would warrant further investigation in the company. However, taking a look across employee data might suggest that males may have a higher tendency to score higher on the other relevant features in affecting attrition. MaritalStatus_Single (0.048594): The lack of family commitment or other financial ties might influence single employees to look for other career opportunities elsewhere DistanceFromHome (Coefficient: 0.034816): Commuting stress or a desire for proximity to home may play a role in influencing attrition JobRole_Sales Executive (Coefficient: 0.253945):



Our analysis suggested that addressing work-life balance issues related to overtime and frequent business travel may help reduce attrition rates. Additionally, HR departments should pay attention to employees who have not received promotions for an extended period, as this group exhibits a higher likelihood of attrition. Higher performing employees are also likely to be talent-scouted by the competition. A focus on recognition of work and tangible rewards should be emphasized in performance ranking systems, while employee engagement initiatives and personalized retention strategies may be useful through predicting potential employees who are thinking of leaving. Care has to be taken, however, in how HR policies are implemented to avoid the case of over-classification or targeting of employees on a basis of higher likelihood to leave the company from past data.

Discussion:

Metrics selection: One of the more important metrics to look out for is the recall. Higher Recall is important in predicting employee churn rate. In predicting IBM employee attrition, "recall" refers to the ability of a prediction model to identify most of the employees who actually leave the company (i.e., correctly identifying "true positives"), while "precision" indicates how accurate the model is in identifying those predicted to leave, minimizing false positives (employees wrongly classified as leaving when they stay) - essentially, recall focuses on capturing most potential leavers, while precision focuses on the quality of those identified as potential leavers.

Key points about recall and precision in employee attrition prediction:

- High recall is crucial: When predicting employee attrition, prioritizing high recall means the model is less likely to miss employees who are actually going to leave, even if it might flag some false positives.

- Precision is important too: While high recall is often prioritized, having good precision ensures that the identified "at-risk" employees are more likely to be genuinely considering leaving, preventing unnecessary interventions. Example:

- High recall, low precision: A model might identify 90% of employees who leave the company but also incorrectly flag 50% of employees who stay as potential leavers, resulting in a large number of false positives.

- Low recall, high precision: A model might only identify 60% of employees who leave but accurately classify most of the flagged employees as potential leavers, potentially missing some high-risk individuals.

In the context of predicting IBM employee attrition, recall is generally considered more important than precision. This is because identifying a high percentage of employees who are actually going to leave (even if it means including some false positives) allows HR to proactively intervene and potentially prevent more departures, which is the primary goal in employee retention strategies.

Explanation:

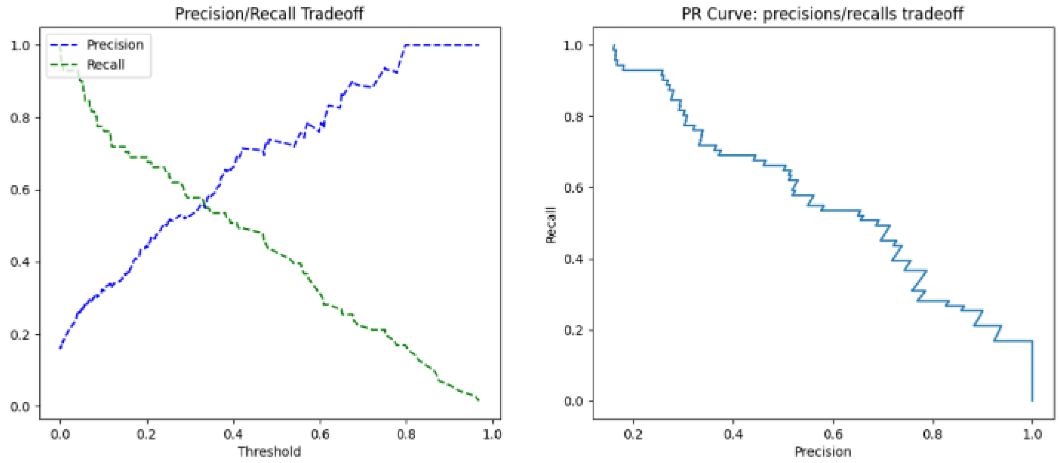
- Precision: Focuses on the accuracy of positive predictions, meaning how many of the identified "at-risk" employees actually leave. While important, a high precision might miss some potential leavers if the model is too strict in its predictions.

- Recall: Focuses on the ability to identify all actual leavers, meaning how many of the employees who are leaving are correctly identified by the model. A high recall ensures that most potential departures are flagged, allowing for targeted interventions.

Key points to consider:

- Cost of false positives: In the case of employee attrition, the cost of a false positive (identifying an employee as likely to leave when they aren't) might be relatively low compared to the cost of missing a true positive (failing to identify an employee who is actually planning to leave).

- Proactive approach: Prioritizing recall allows HR to take proactive steps to address potential issues with a wider range of employees who might be considering leaving, even if it means investigating a few false positives.



Metrics selected for model performance comparison:

Recall: recall score is calculated as: TP / (TP + FN), which tells us how many positive attritions we are able to predict using our model.

Precision: precision score is calculated as: TP / (TP + FP), which tells us among all the positive predictions of attritions, the ratio that we are getting correct.

F1: f1-score is the harmonic mean of precision and recall, it gives more weight to the low value, thus f1-score will only be higher, if both precision and recall are high and with no greater difference.

Final Model Selection:

Final model selection should not be based just on higher performance in numbers but also on fulfilling business requirement(s) and the choice can change over time to accommodate changes in data and demand.

We choose " Logistic Regression" to be the final model for this project to measure the generalization error on the test data out of the following reasons:

1. Based on experience working in multiple start-up environments, Recall is relatively more important than precision in a smaller-sized company, as each role is quite unique and the employees usually wear multiple hats and are accountable for a lot of things, therefore finding

good replacements can take a long time. As a decision-maker, I would rather have a few false alarms than miss any employee that is truly leaving.

2. However, as a company grows in size thus increasing their budget and decides to use a more aggressive retention strategy to keep the high attrition risk employees, the precision score needs to be high as well. Otherwise, resources are wasted on things that are unnecessary.

In conclusion, the final model selection is based on the assumptions of the business case. If we make absolutely no assumption about the data, then there is no reason to prefer one model over another.

ROC Curve: It is a curve between True Positive rate (Recall) and False Positive rate (1 – True Negative rate). The ROC curve is a simple plot that shows the trade-off between the true positive rate and the false positive rate of a classifier for various choices of the probability threshold. From the ROC Curve, we have a choice to make depending on the value we place on true positive and tolerance for false positive rate. If we wish to find more people who are leaving, we could increase the true positive rate by adjusting the probability cut-off for classification. However by doing so would also increase the false positive rate Hence we need to find the optimum value of cut-off for classification

Note: Recall = Sensitivity = True Positive Rate (TPR)

7. Recommendation for Reducing Employee Attrition:

• Implement proactive measures based on predictive models to identify at-risk employees and intervene early.

• Focus on improving job satisfaction, work-life balance, and career development opportunities to increase employee retention. Regularly retrain and update predictive models with new data to ensure effectiveness in capturing evolving trends in attrition

8. CONCLUSION: Predicting employee attrition is a complex but important task for organizations to manage their workforce effectively. By leveraging machine learning techniques and analyzing relevant factors, organizations can gain valuable insights into attrition patterns and take proactive steps to reduce attrition rates, improve employee satisfaction, and enhance overall organizational performance