Employee Attrition: Understanding and Addressing the Challenge

Employee attrition is a significant challenge for organizations of all sizes. It can lead to decreased productivity, higher costs, and a loss of valuable knowledge and experience. Understanding the causes, measuring the impact, and implementing strategies to mitigate attrition is crucial for long-term success.

by Aditi Kolhe & Prachi Shewale

Data Collection

Our data was sourced from the Kaggle website.
Link : https://www.kaggle.com/datasets/patelprashant/employee-attrition/data

Dataset Size:

The dataset contains 1470 rows and 35 columns .

Key Features

1) Employee Information :

- Age, Gender, Marital Status, Job Role, Department

2) Job Performance & Satisfaction:

- Job Satisfaction, Performance Rating, Job Involvement, Job Level.

3) Compensation:

- Monthly Income, Hourly Rate, Stock Option Level, Percent Salary Hike.

4) Work-Life Balance:

- Work-Life Balance, OverTime, Years at Company, Training Times Last Year.

5) Attrition:

- The target variable indicating whether the employee left the company (Yes/No).

6) Categorical Variables and Numerical Variables:

- Attributes such as Age, MonthlyIncome, YearsAtCompany, and TotalWorkingYears are numerical. Attributes like BusinessTravel, EducationField, JobRole, and Gender are categorical

Data Preprocessing: Handling Challenges in Employee Attrition Data

Handling Missing Values

No missing values were found in this dataset,so no imputation was necessary.

Encoding Categorical Variables

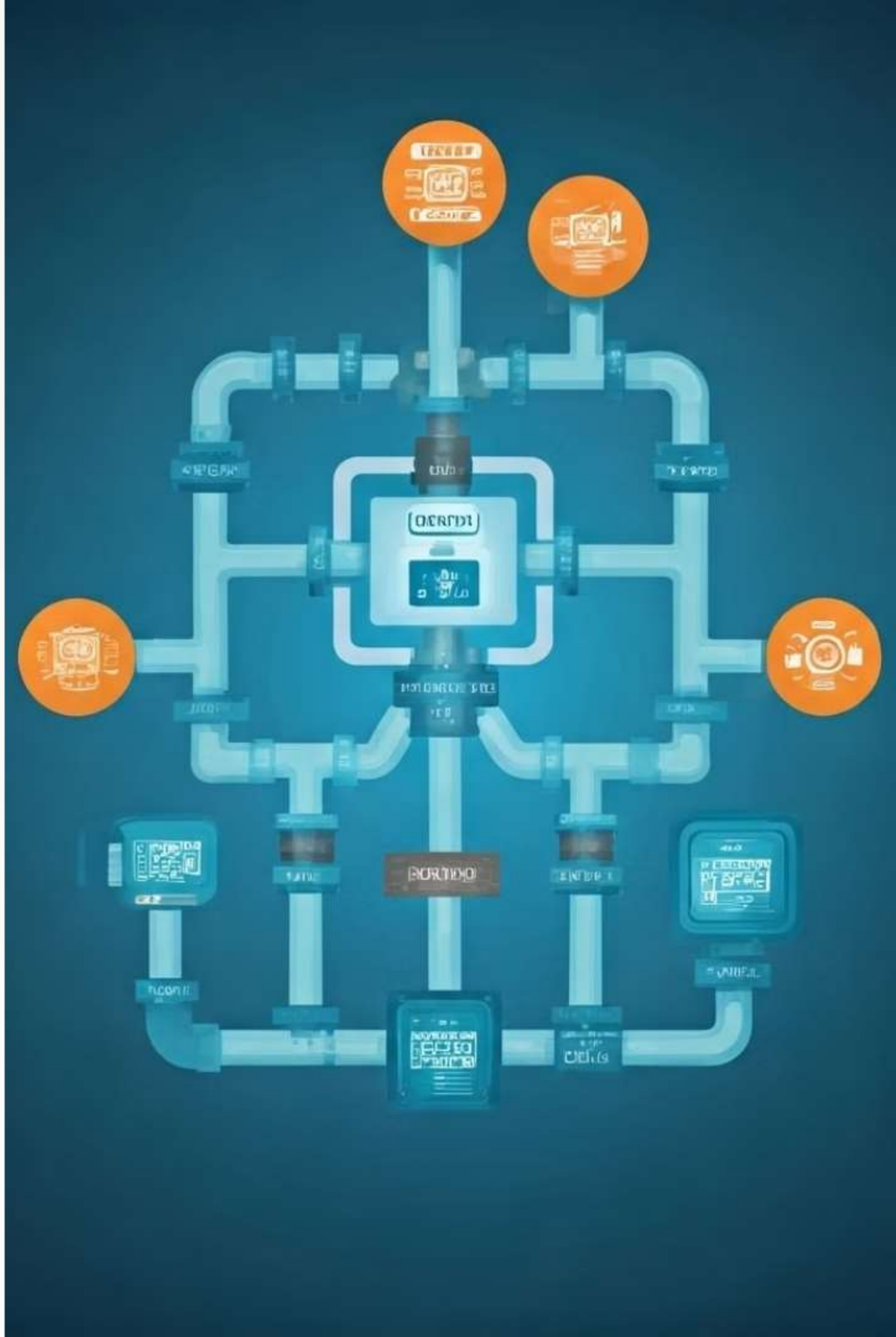
Method Used: One-hot encoding (for non-binary categories like JobRole, EducationField) or Label encoding (for binary categories like Gender).

Feature Scaling

Numerical features like 'MonthlyIncome', 'YearsAtCompany', and 'Age' were scaled using Min-Max scaling to normalize the range of values."

Outlier Detection

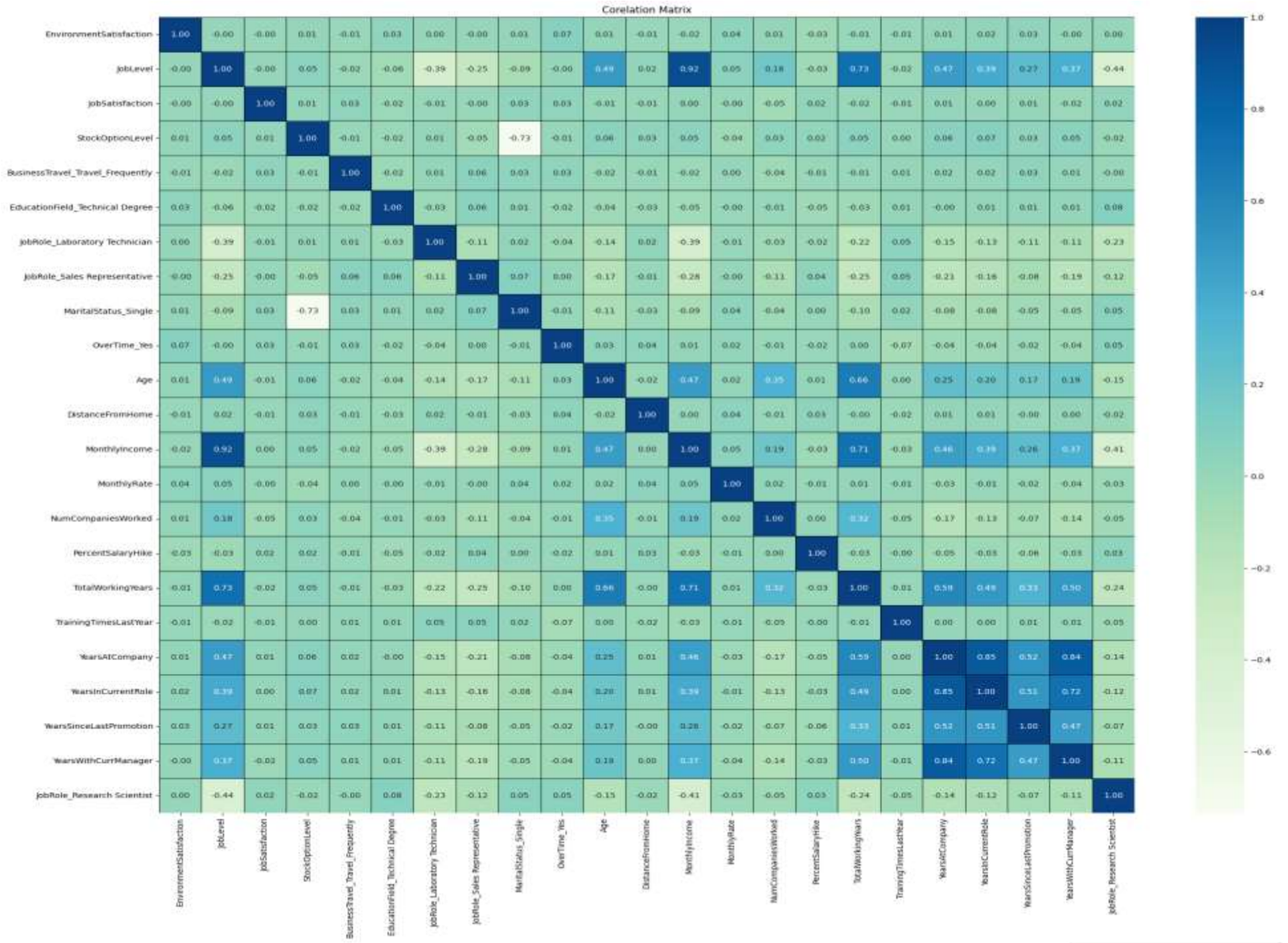
We found outlier in one of the attribute,”MonthlyRate”.



Exploratory Data Analysis: Uncovering Patterns and Insights

1) Correlation Between Features:

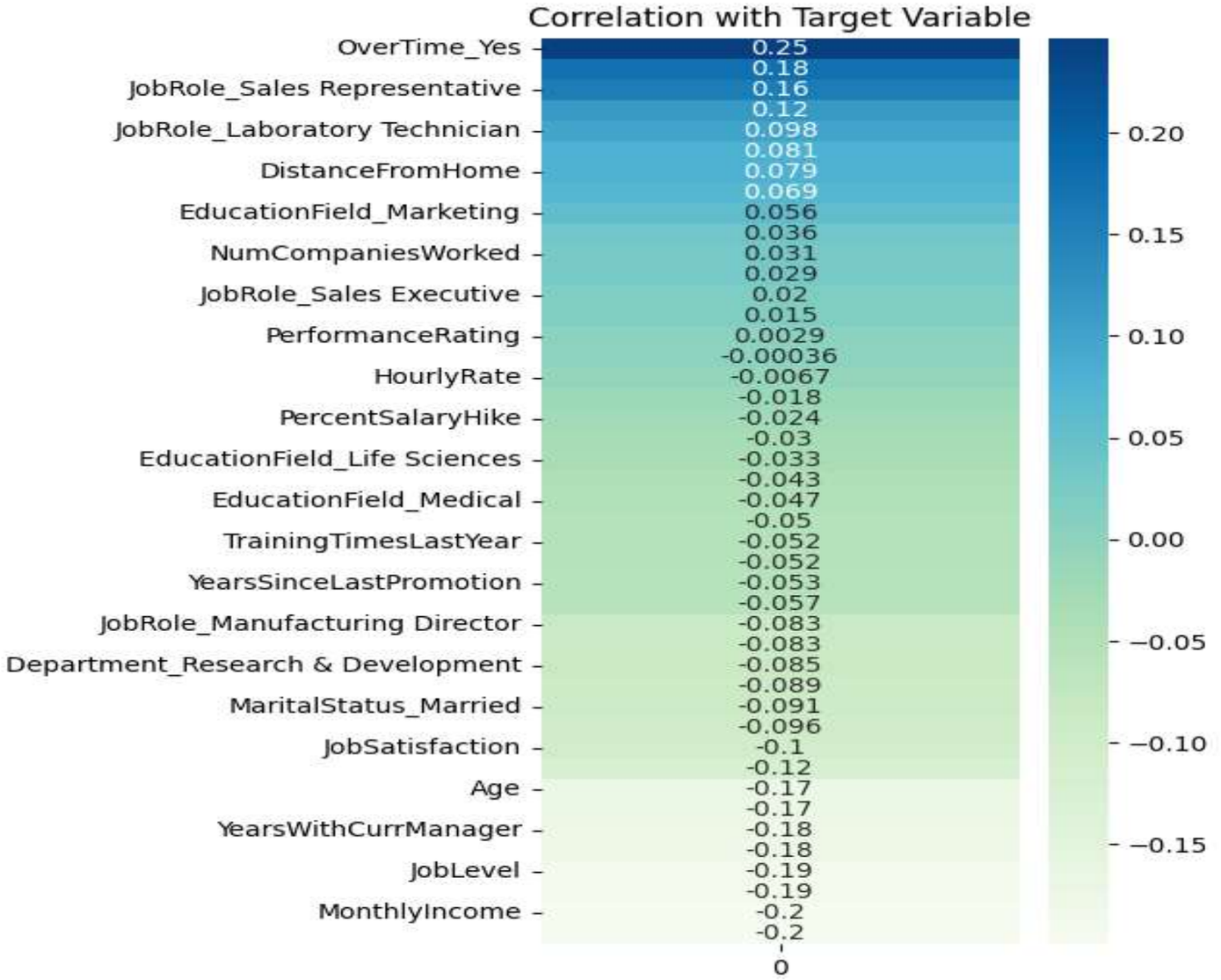
The heatmap shows strong positive correlations between **MonthlyIncome and JobLevel** (0.92) and **TotalWorkingYears and YearsAtCompany** (0.66), indicating clear patterns in salary and experience. Negative correlation is notable between **StockOptionLevel and YearsSinceLastPromotion** (-0.73). Most variables display weak or no correlation, suggesting their independence. This highlights key factors influencing employee metrics.



Exploratory Data Analysis: Uncovering Patterns and Insights

2) Correlation With Target Variable:

The graph highlights that features like “OverTime_Yes”and specific job roles (e.g., “Sales Representative”) have the strongest positive correlations with the target variable. In contrast, factors like “MonthlyIncome”, “JobLevel”, and “YearsWithCurrManager” show the most significant negative correlations, indicating they may decrease as the target variable increases.



Model Selection:

- Model Building without doing any feature selection and balancing the dataset
- Feature Selection
- XGBoost Hyperparameter tunning
- Using PCA

Model Building without doing any feature selection and balancing the dataset

Model Building Process

In this approach, we built multiple machine learning models without performing feature selection or dataset balancing. The models used include:

- **Decision Tree, Random Forest, AdaBoost, Gradient Boosting, XGBoost, and LightGBM Classifiers**.
- These models are initialized with a fixed random state to ensure consistency.
- We use evaluation metrics like accuracy, F1-score, and confusion matrix to compare their performance.

This strategy enables a diverse comparison of model effectiveness using the raw dataset, without altering features or addressing data imbalance.

train_accuracy	test_accuracy	train_f1	test_f1	
DecisionTreeClassifier	1.000000	0.768707	1.000000	0.333333
RandomForestClassifier	1.000000	0.833333	1.000000	0.140351
AdaBoostClassifier	0.908163	0.850340	0.647059	0.333333
GradientBoostingClassifier	0.956633	0.846939	0.844985	0.262295

Model Building without doing any feature selection and balancing the dataset

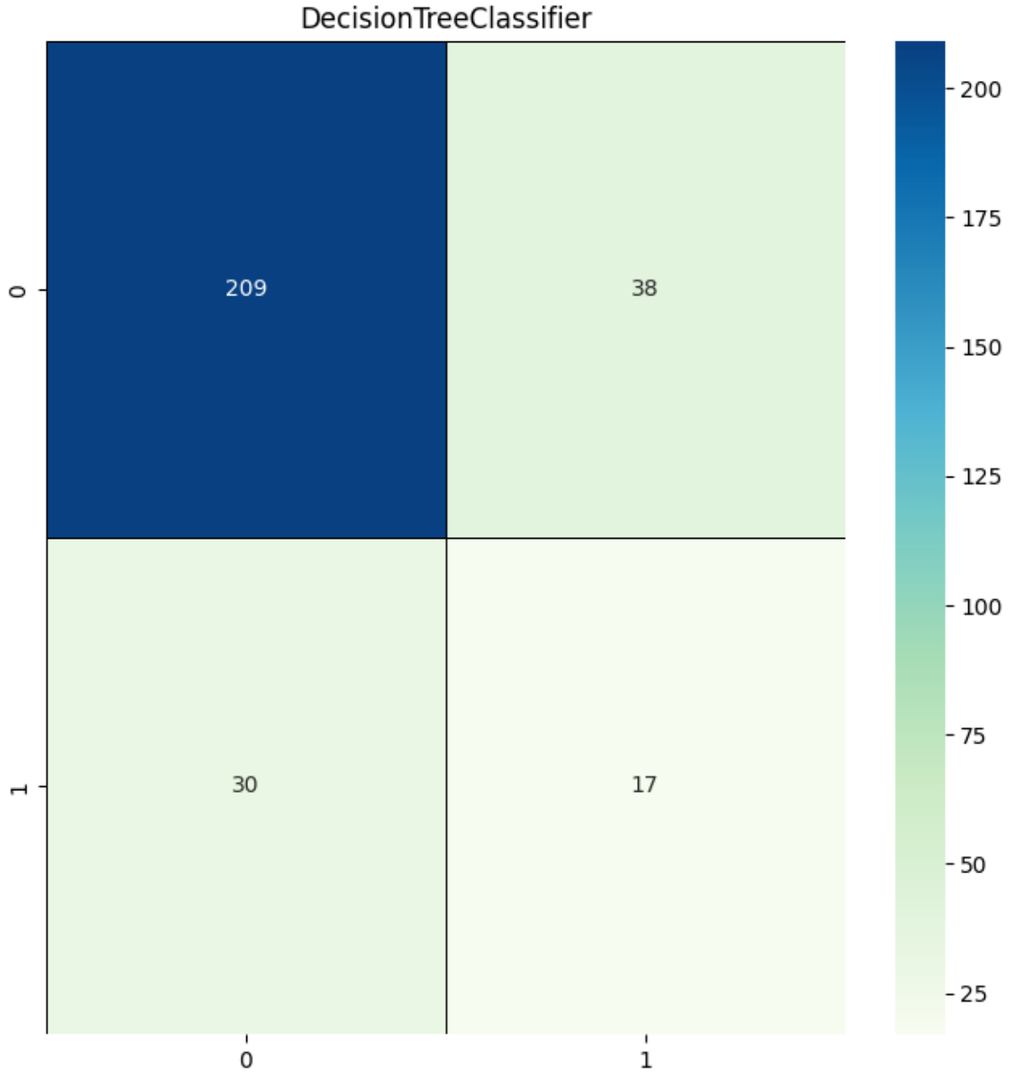
DecisionTreeClassifier classification report

	precision	recall	f1-score	support
0	0.87	0.85	0.86	247
1	0.31	0.36	0.33	47
accuracy			0.77	294
macro avg	0.59	0.60	0.60	294
weighted avg	0.78	0.77	0.78	294

DecisionTreeClassifier Report:

•The classification report for the **DecisionTreeClassifier** shows an accuracy of **0.77**.

•The precision, recall, and F1-score are significantly higher for class 0 (precision of 0.87) compared to class 1 (precision of 0.31), indicating an imbalance in the model's performance towards predicting the majority class more accurately.



This is a **confusion matrix** for a DecisionTreeClassifier, which summarizes the performance of a binary classification model.

•**Top-left (209)**: True Negatives (correctly predicted 0s)

•**Top-right (38)**: False Positives (predicted 1 but actual 0)

•**Bottom-left (30)**: False Negatives (predicted 0 but actual 1)
•**Bottom-right (17)**: True Positives (correctly predicted 1s)

Key insights:
•The model performs well in identifying class 0 but struggles with class 1.
•**False negatives (30)** are higher than **true positives (17)**, indicating poor performance in detecting class 1.
•It could benefit from tuning or rebalancing if predicting class 1 is important.

Model Building without doing any feature selection and balancing the dataset

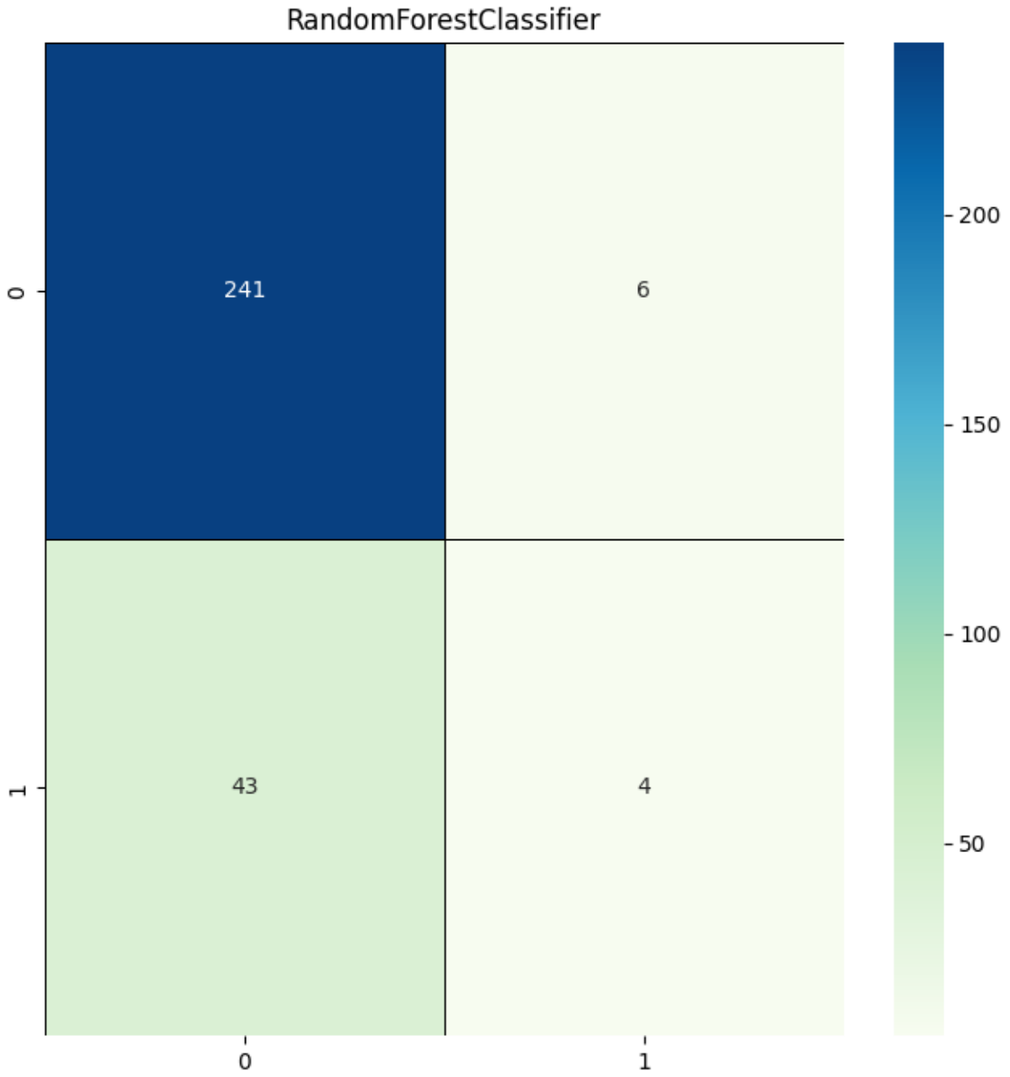
RandomForestClassifier classification report

	precision	recall	f1-score	support
0	0.85	0.98	0.91	247
1	0.40	0.09	0.14	47
accuracy			0.83	294
macro avg	0.62	0.53	0.52	294
weighted avg	0.78	0.83	0.79	294

RandomForestClassifier Interpretation:

- **Class 0**:
  - F1-score: 0.91 – Performs well.
- **Class 1**:
  - F1-score: 0.14 – Struggles with many false negatives.
- **Accuracy**: 83%

The model favors class 0, performing poorly on class 1. Consider class balancing or hyperparameter tuning for improvement.



The graph you provided is a confusion matrix for a **RandomForestClassifier**.

- **True Negatives (0,0)**: 241 instances were correctly predicted as class 0.
- **False Positives (0,1)**: 6 instances were incorrectly predicted as class 1, but they belong to class 0.
- **False Negatives (1,0)**: 43 instances were incorrectly predicted as class 0, but they belong to class 1.
- **True Positives (1,1)**: 4 instances were correctly predicted as class 1.

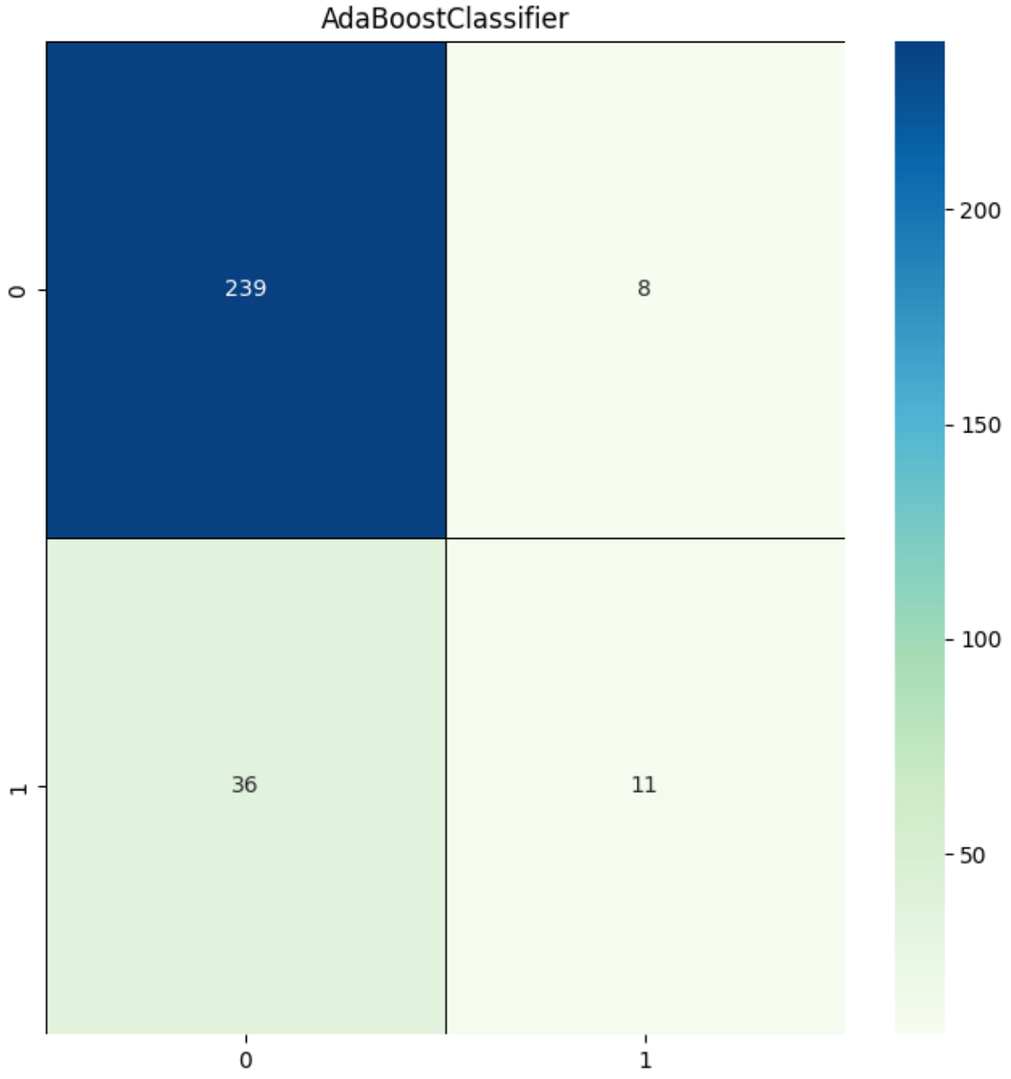
The classifier shows strong performance in predicting class 0 but struggles more with class 1, indicating some imbalance or difficulty in distinguishing class 1 correctly.

Model Building without doing any feature selection and balancing the dataset

AdaBoostClassifier classification report

	precision	recall	f1-score	support
0	0.87	0.97	0.92	247
1	0.58	0.23	0.33	47
accuracy			0.85	294
macro avg	0.72	0.60	0.62	294
weighted avg	0.82	0.85	0.82	294

The AdaBoost classifier achieves **high performance on Class 0** (Precision: 0.87, Recall: 0.97, F1-Score: 0.92), but **struggles with Class 1** (Precision: 0.58, Recall: 0.23, F1-Score: 0.33), showing poor detection of the minority class. Overall accuracy is **85%**, indicating strong performance for the majority class but significant imbalance in handling Class 1 predictions.



The confusion matrix evaluates the performance of an AdaBoostClassifier, a machine learning ensemble method.

The matrix provides a summary of prediction results on a classification problem with two classes (binary classification: 0 and 1). Each cell reflects the number of instances in the predicted vs. actual categories:True Negatives (239): The classifier correctly identified 239 instances as class 0 (actual class 0).False Positives (8): The classifier mistakenly predicted 8 instances as class 1 when they were actually class 0.

False Negatives (36): The classifier predicted 36 instances as class 0 when they were actually class 1.True Positives (11): The classifier correctly predicted 11 instances as class 1 (actual class 1).The heatmap uses color intensity, with darker blue for higher values, to visually emphasize the distribution of correctly and incorrectly classified cases. This confusion matrix can be used to calculate various metrics like accuracy, precision, recall, and F1-score for the classifier.

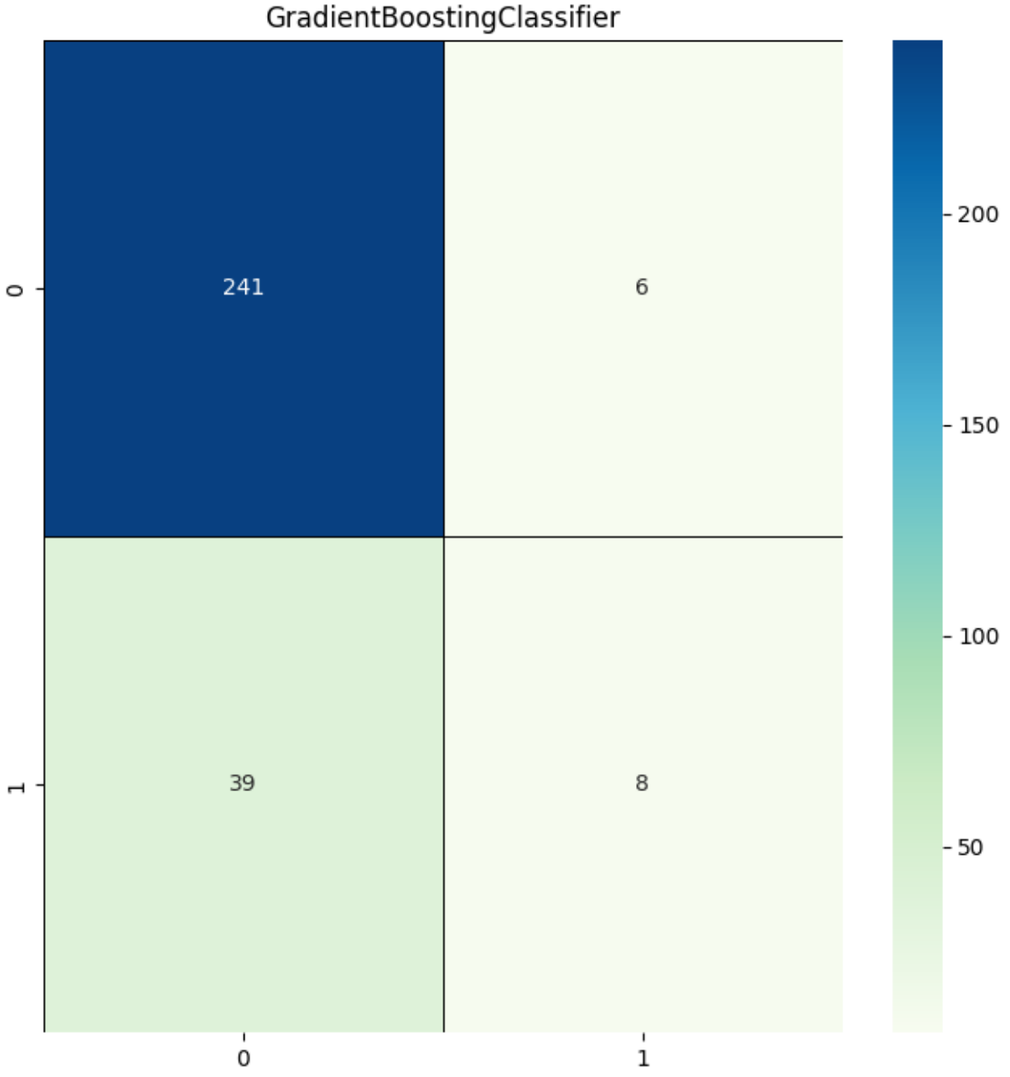
Model Building without doing any feature selection and balancing the dataset

GradientBoostingClassifier classification report

	precision	recall	f1-score	support
0	0.86	0.98	0.91	247
1	0.57	0.17	0.26	47
accuracy			0.85	294
macro avg	0.72	0.57	0.59	294
weighted avg	0.81	0.85	0.81	294

The **GradientBoostingClassifier** performs well for the majority class (`0`) with a high precision of 0.86 and recall of 0.98, leading to a strong F1-score of 0.91. However, its performance significantly drops for the minority class (`1`), with a low recall of 0.17 and F1-score of 0.26, indicating difficulty in identifying this class.

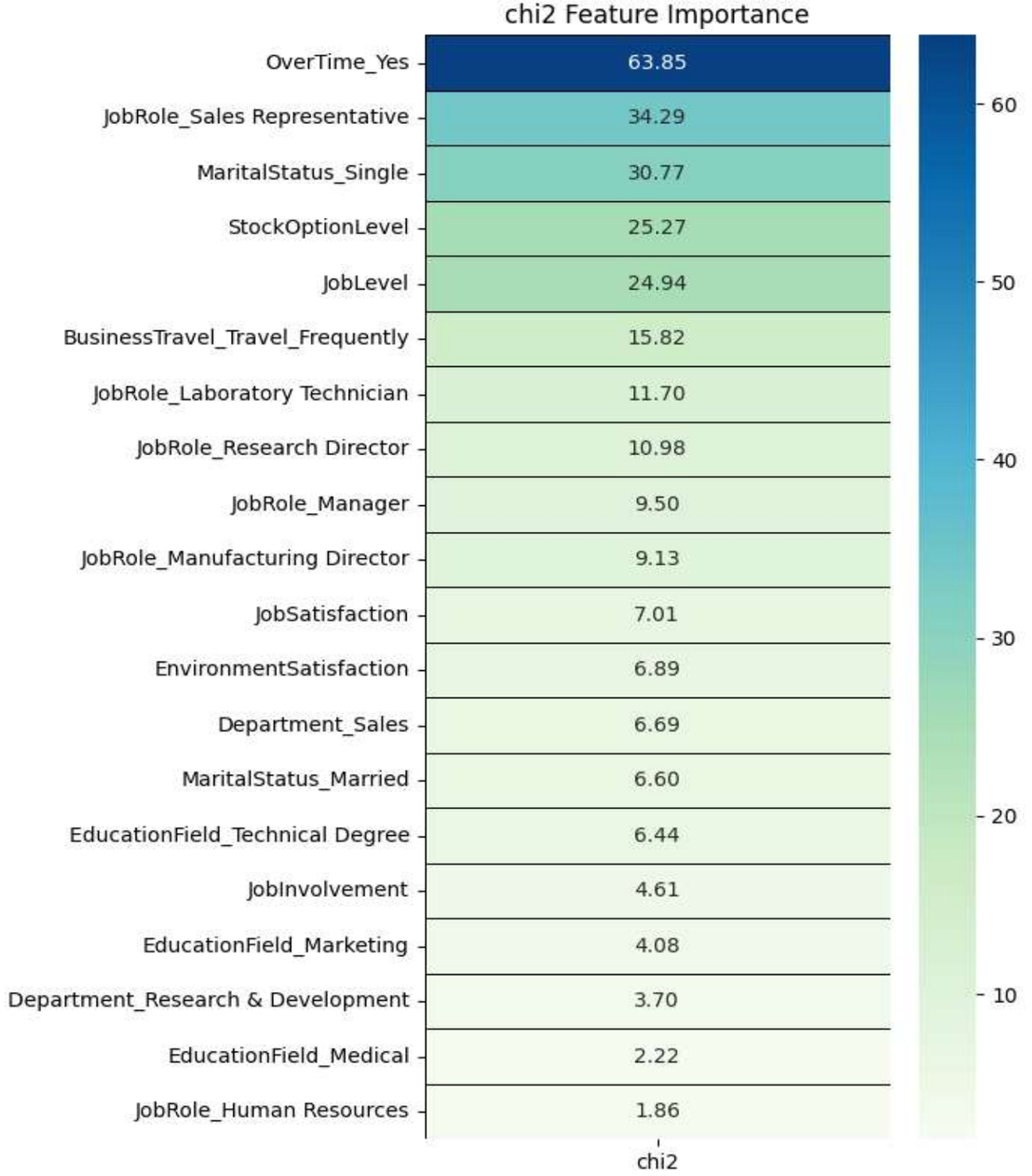
The overall accuracy is 0.85, but the macro average highlights the model's bias towards the majority class. This imbalance suggests that the model favors the dominant class, affecting its predictive ability for the minority class.

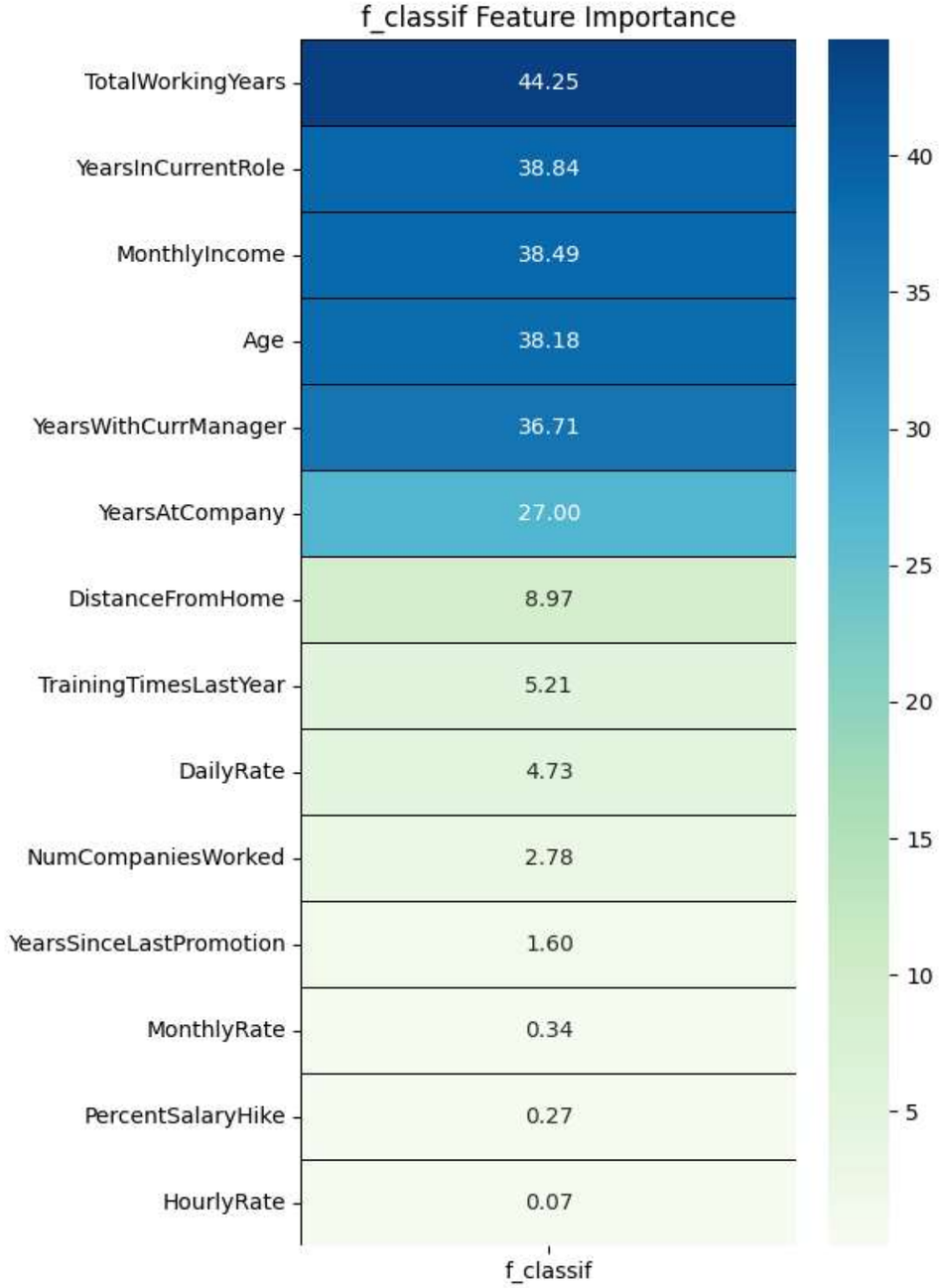


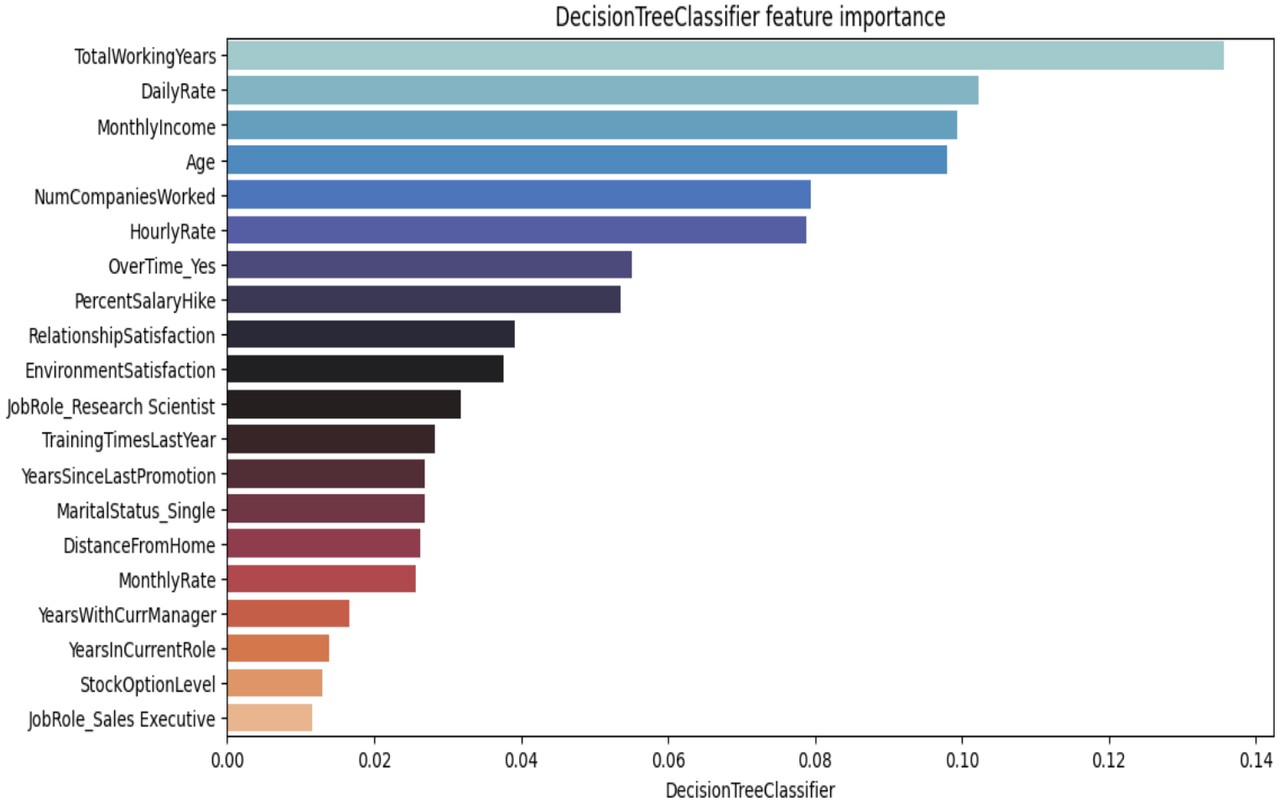
This confusion matrix shows the performance of a GradientBoostingClassifier. The model correctly identified 241 true negatives and 8 true positives, while misclassifying 6 false positives and 39 false negatives. The classifier performs well on class 0, but struggles with class 1, as indicated by a higher number of false negatives. This suggests the model is biased towards predicting class 0 over class 1.

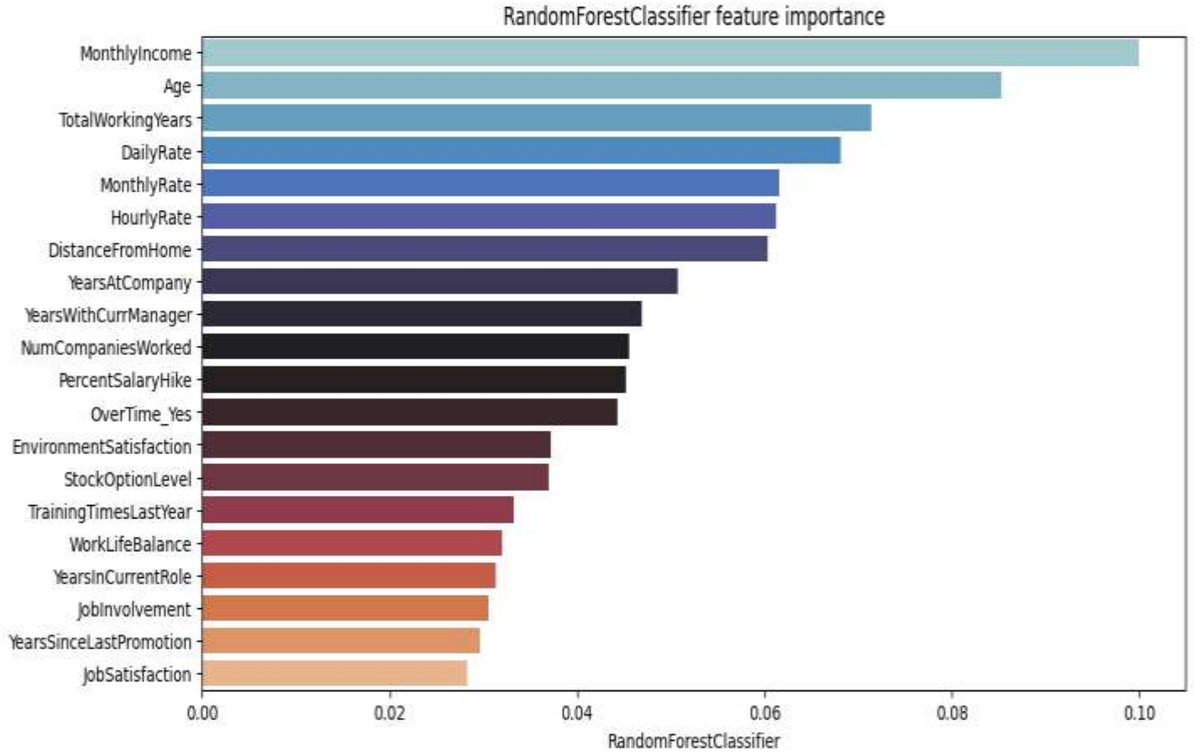
True Negatives (241): The model correctly predicted 241 instances as class 0 (actual class 0).False Positives (6): The model incorrectly predicted 6 instances as class 1, but they were actually class 0.False Negatives (39): The model predicted 39 instances as class 0, but they were actually class 1.True Positives (8): The model correctly predicted 8 instances as class 1 (actual class 1).

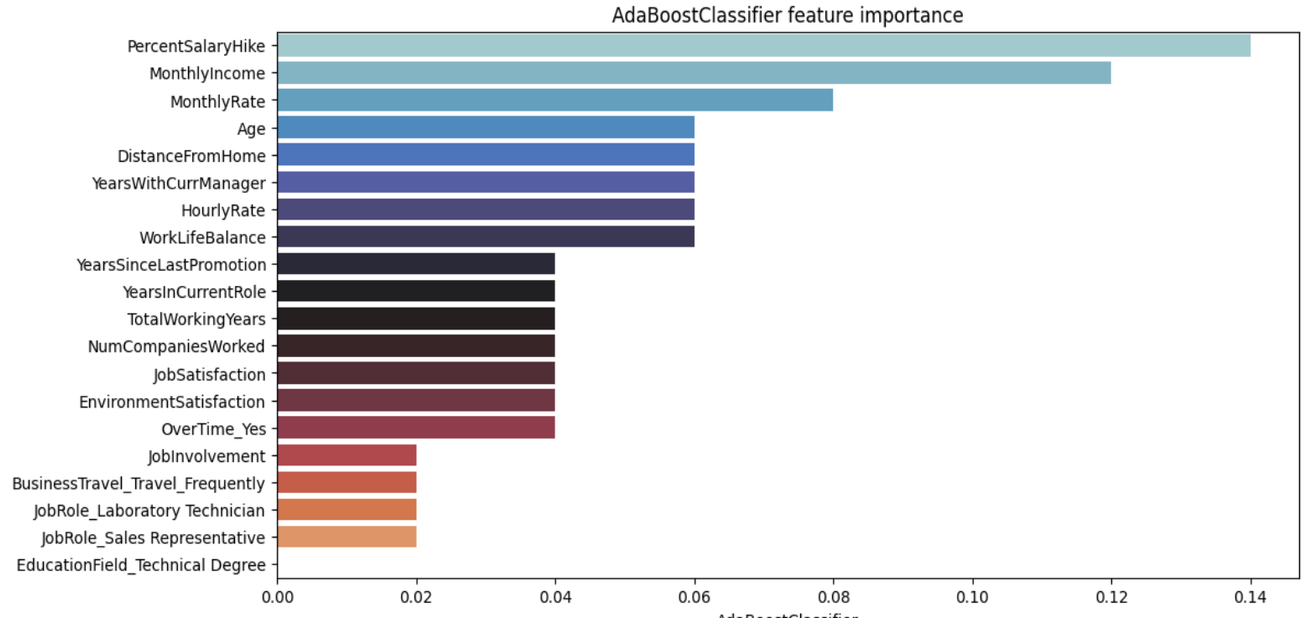
Feature Selection

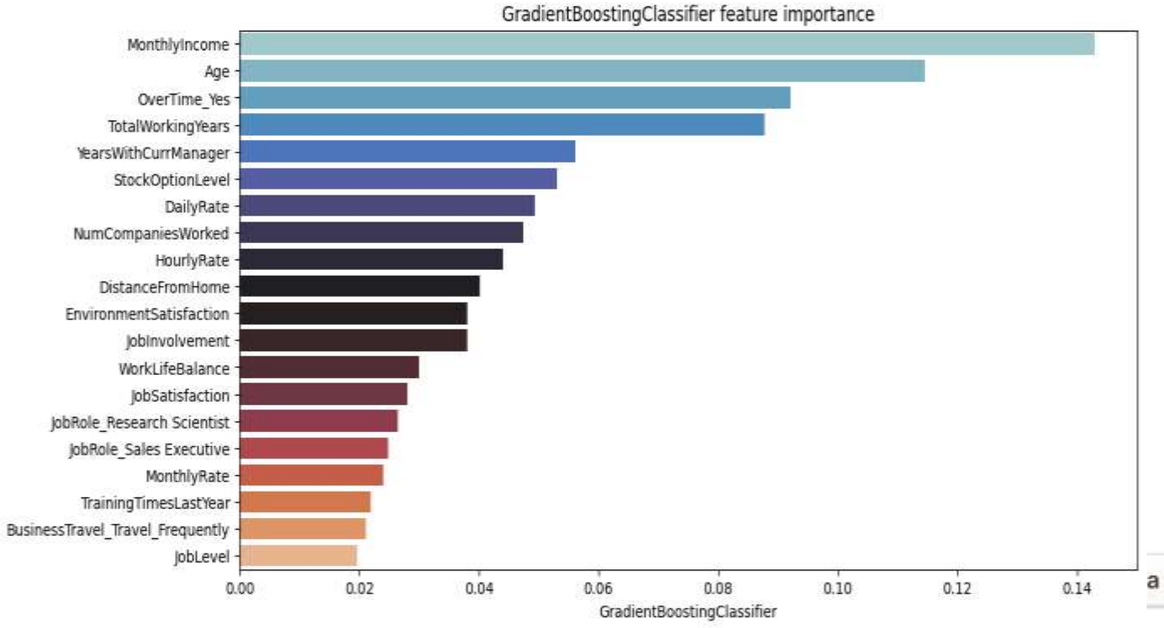




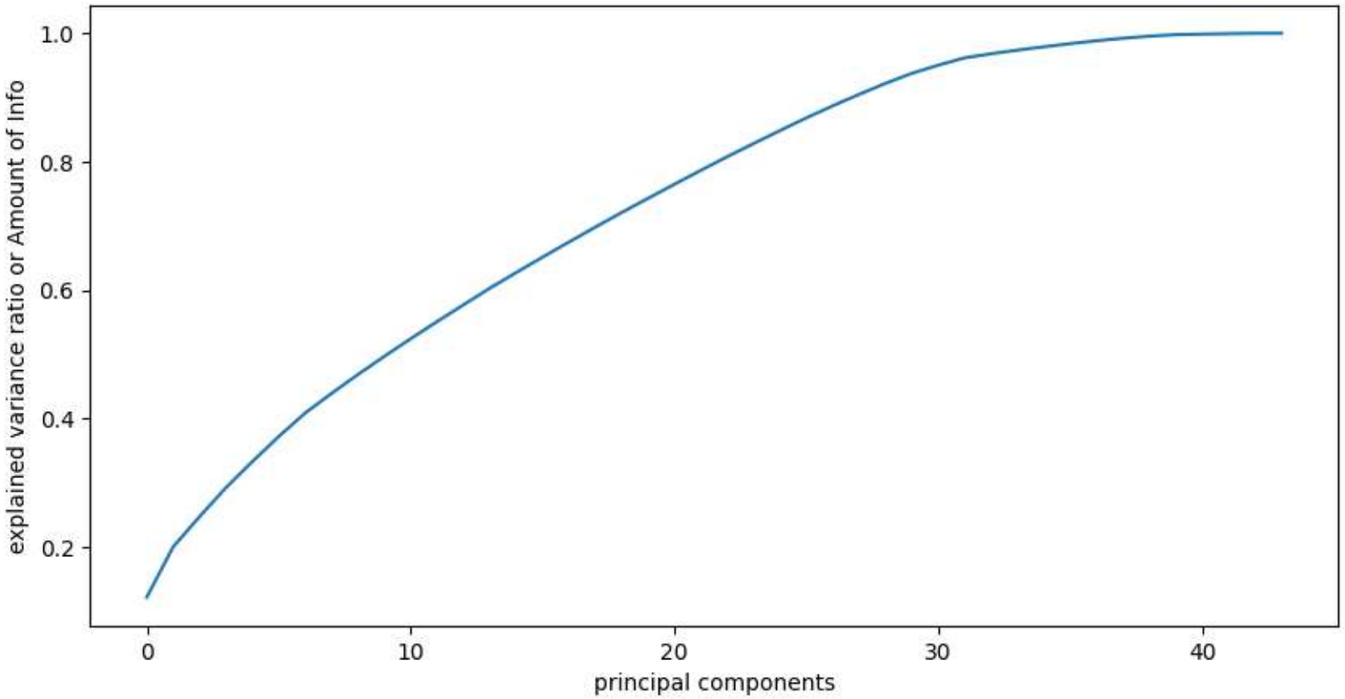

DecisionTreeClassifier feature importance
TotalWorkingYears
DailyRate
MonthlyIncome
Age
NumCompaniesWorked
HourlyRate
OverTime_Yes
PercentSalaryHike
RelationshipSatisfaction
EnvironmentSatisfaction
JobRole_Research Scientist
TrainingTimesLastYear
YearsSinceLastPromotion
MaritalStatus_Single
DistanceFromHome
MonthlyRate
YearsWithCurrManager
YearsInCurrentRole
StockOptionLevel
JobRole_Sales Executive
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
DecisionTreeClassifier


RandomForestClassifier feature importance
MonthlyIncome
Age
TotalWorkingYears
DailyRate
MonthlyRate
HourlyRate
DistanceFromHome
YearsAtCompany
YearsWithCurrManager
NumCompaniesWorked
PercentSalaryHike
OverTime_Yes
EnvironmentSatisfaction
StockOptionLevel
TrainingTimesLastYear
WorkLifeBalance
YearsInCurrentRole
JobInvolvement
YearsSinceLastPromotion
JobSatisfaction
0.00
0.02
0.04
0.06
0.08
0.10
RandomForestClassifier


AdaBoostClassifier feature importance
PercentSalaryHike
MonthlyIncome
MonthlyRate
Age
DistanceFromHome
YearsWithCurrManager
HourlyRate
WorkLifeBalance
YearsSinceLastPromotion
YearsInCurrentRole
TotalWorkingYears
NumCompaniesWorked
JobSatisfaction
EnvironmentSatisfaction
OverTime_Yes
JobInvolvement
BusinessTravel_Travel_Frequently
JobRole_Laboratory Technician
JobRole_Sales Representative
EducationField_Technical Degree
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14


GradientBoostingClassifier feature importance
MonthlyIncome
Age
OverTime_Yes
TotalWorkingYears
YearsWithCurrManager
StockOptionLevel
DailyRate
NumCompaniesWorked
HourlyRate
DistanceFromHome
EnvironmentSatisfaction
JobInvolvement
WorkLifeBalance
JobSatisfaction
JobRole_Research Scientist
JobRole_Sales Executive
MonthlyRate
TrainingTimesLastYear
BusinessTravel_Travel_Frequently
JobLevel
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
GradientBoostingClassifier

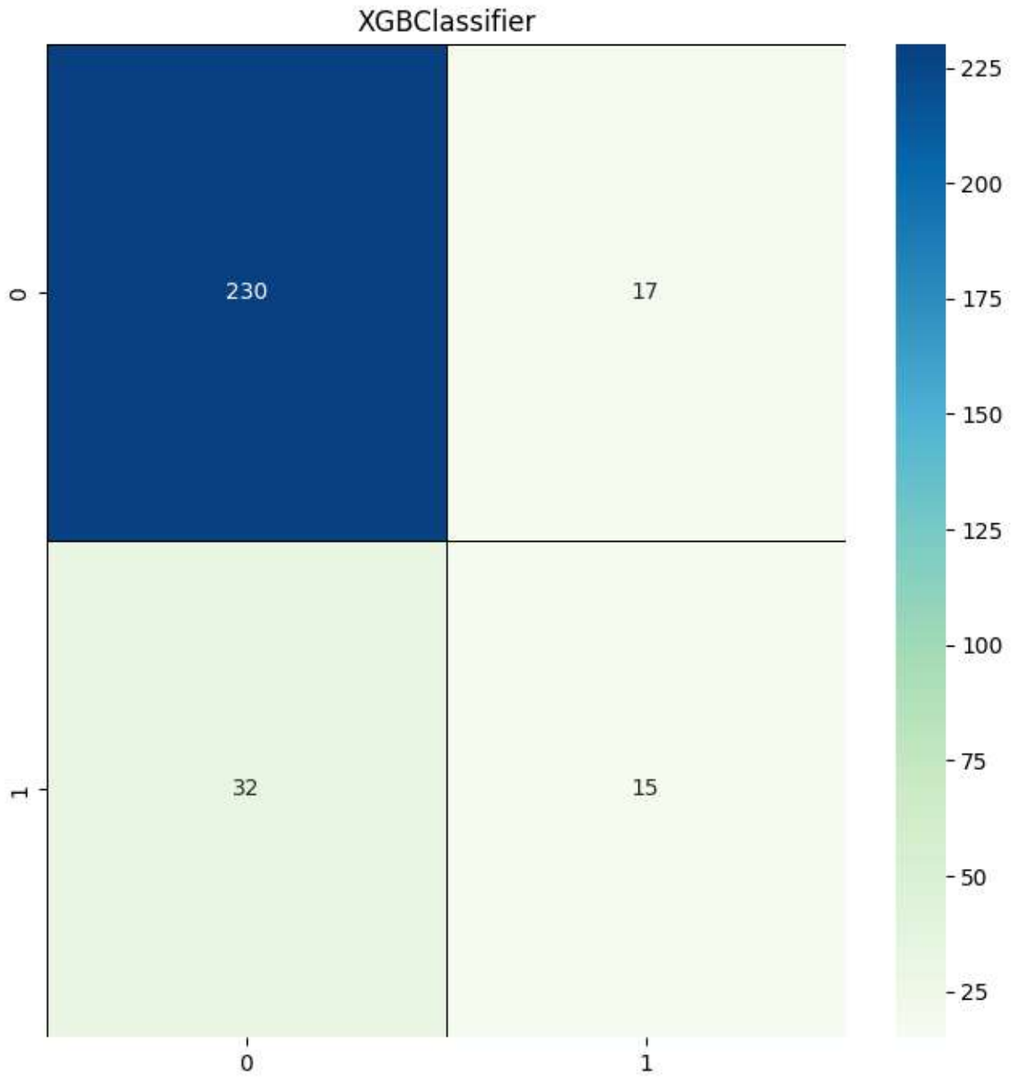
Using PCA



The x-axis represents the "principal components," which are the new variables created from the original dataset after performing PCA.

The y-axis shows the "explained variance ratio" or the amount of information captured by each principal component.The curve increases as more principal components are added, indicating that each additional component explains more variance in the data.

The curve flattens around 30 to 40 components, suggesting that adding more components beyond this point offers diminishing returns in terms of explained variance.This suggests that most of the variance in the dataset can be explained using the first 30 or so principal components, making it a good point for dimensionality reduction.



- The rows represent the *actual classes* (0 or 1).- The columns represent the *predicted classes* (0 or 1).- The values inside the matrix are the number of instances for each combination of actual and predicted class

- *230* true negatives (top-left): The model correctly predicted class 0.- *17* false positives (top-right): The model predicted class 1, but the actual class was 0.- *32* false negatives (bottom-left): The model predicted class 0, but the actual class was 1.- *15* true positives (bottom-right): The model correctly predicted class 1

- The model performs well at predicting class 0 (with 230 true negatives), but struggles more with class 1, misclassifying 32 instances as class 0.- The balance of the confusion matrix suggests that the model may be biased towards predicting class 0, possibly because of class imbalance in the data.

Conclusion and Key Takeaways

•**Job Role and Compensation**: Certain job roles, such as Sales Representatives, were highly correlated with attrition. Additionally, lower income levels and fewer opportunities for growth (as indicated by low Job Level) contributed to higher attrition rates.

•**Career Development and Job Satisfaction**: Factors like years since the last promotion and job satisfaction were influential in predicting whether an employee would leave. Providing clear career growth opportunities and addressing dissatisfaction at work are crucial in retaining employees.

Fostering a Positive Work Environment

A positive work environment is crucial for employee well-being and retention. Cultivating a culture of respect, open communication, and collaboration can significantly reduce attrition rates.