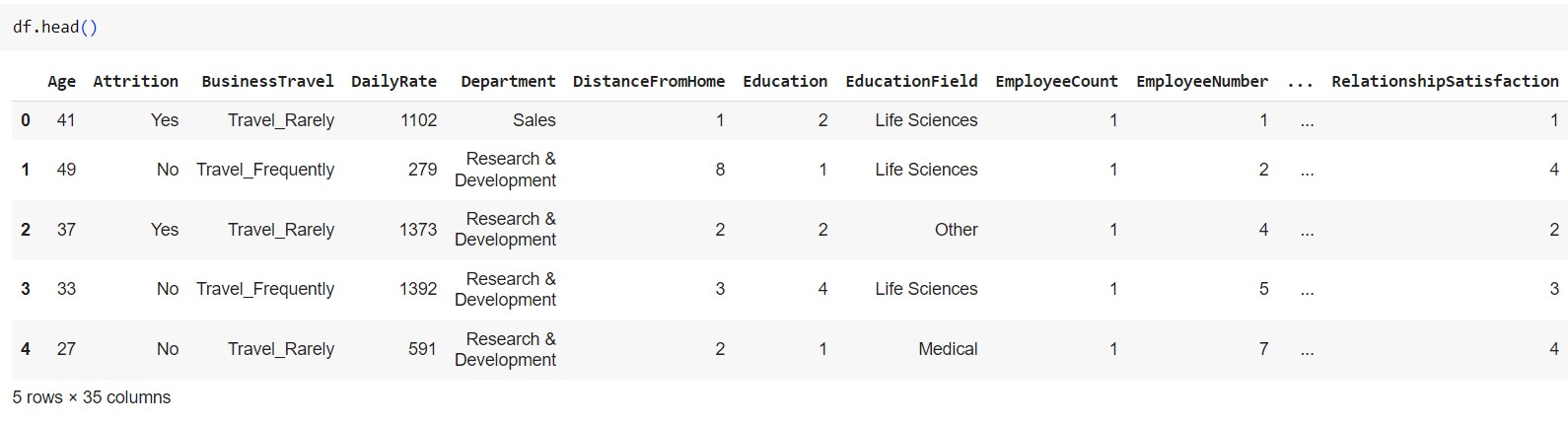
**Employee Attrition Prediction: Project Documentation**

1. **Dataset Analysis**

* Loading Dataset: The dataset "IBM HR Analytics Employee Attrition & Performance" is loaded using pandas' read\_csv function.



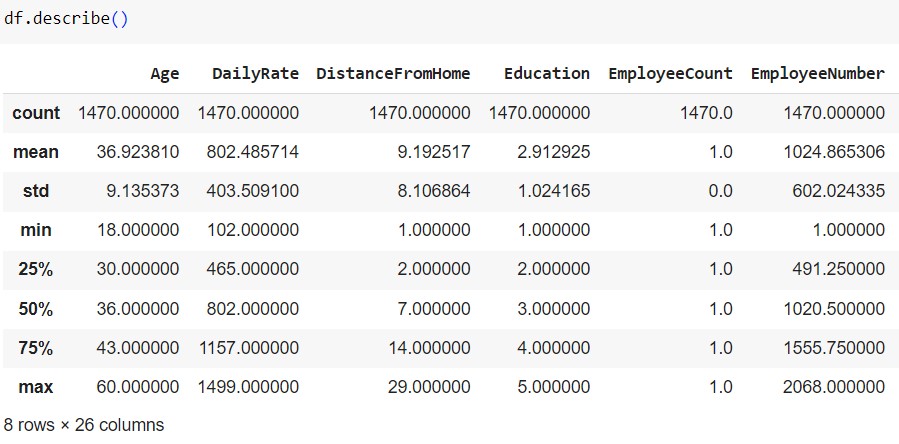
* Exploratory Data Analysis (EDA):
* head(): Displays the first few rows of the dataset.



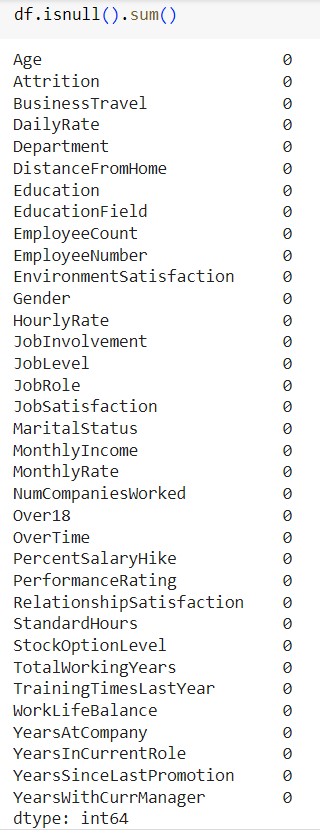
* + info(): Provides information about the dataset including data types and missing values.



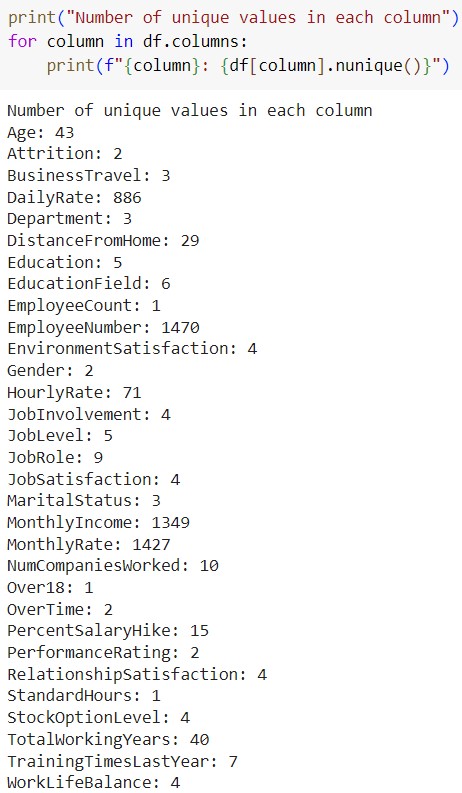
* + describe(): Generates summary statistics for numerical columns.



* + isnull().sum(): Calculates the number of missing values in each column.



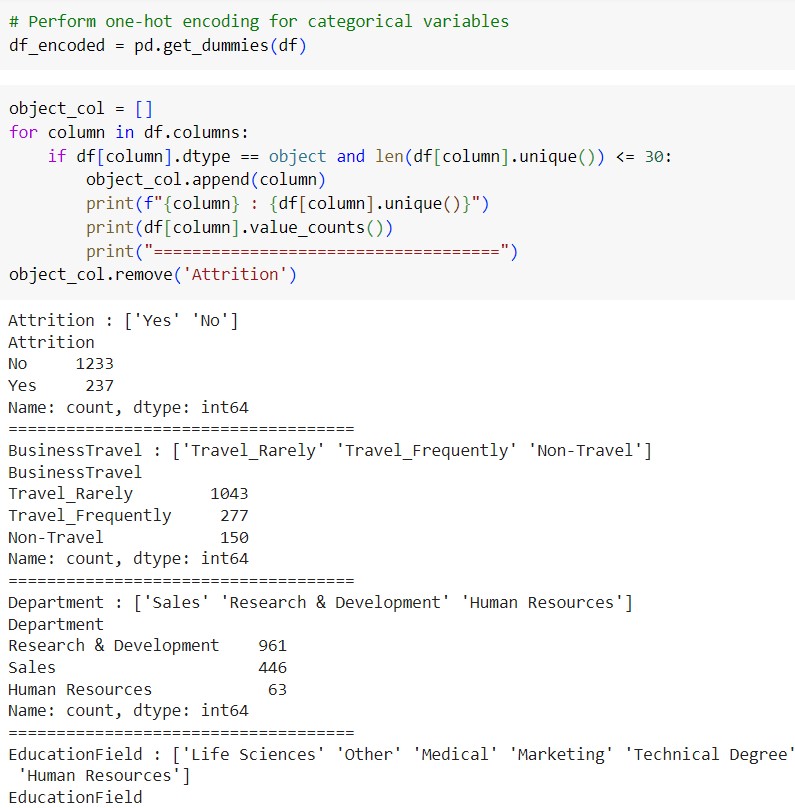
* + Number of unique values in each column is printed.

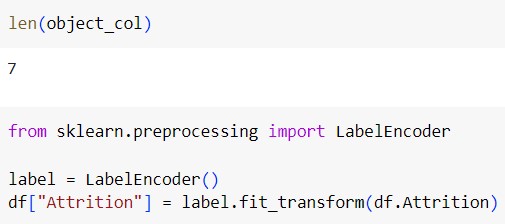


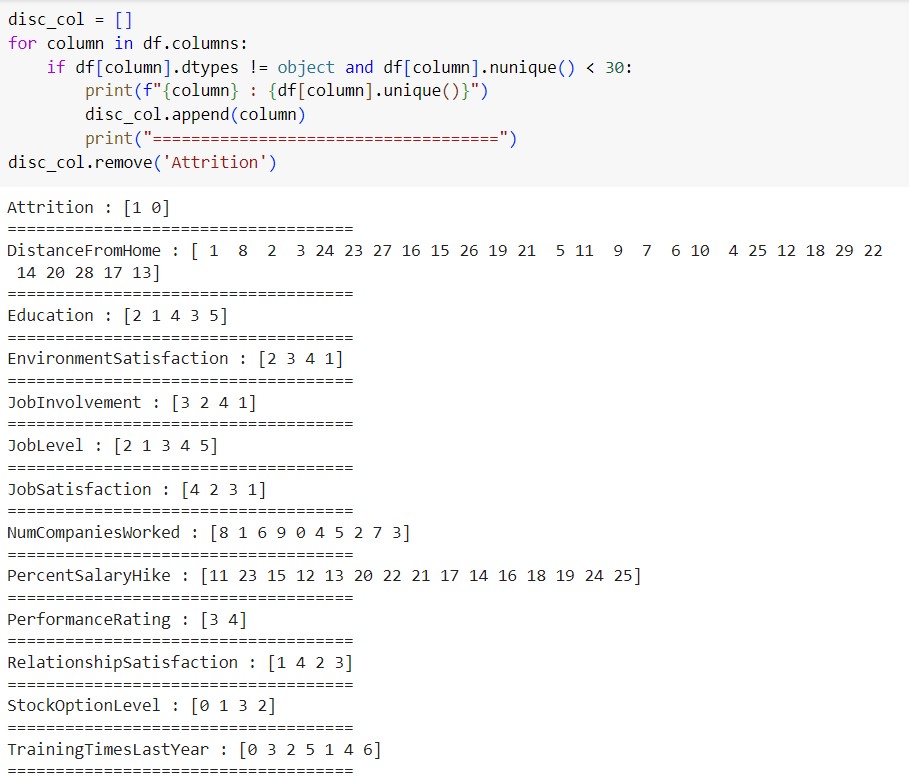
* Data Preprocessing:
  + We notice that 'EmployeeCount', 'Over18', 'StandardHours' have only one unique values and 'EmployeeNumber' has 1470 unique values. This features aren't useful for us, So we are going to drop those columns.
  + Non-essential columns ('EmployeeCount', 'EmployeeNumber', 'Over18', 'StandardHours') are dropped.

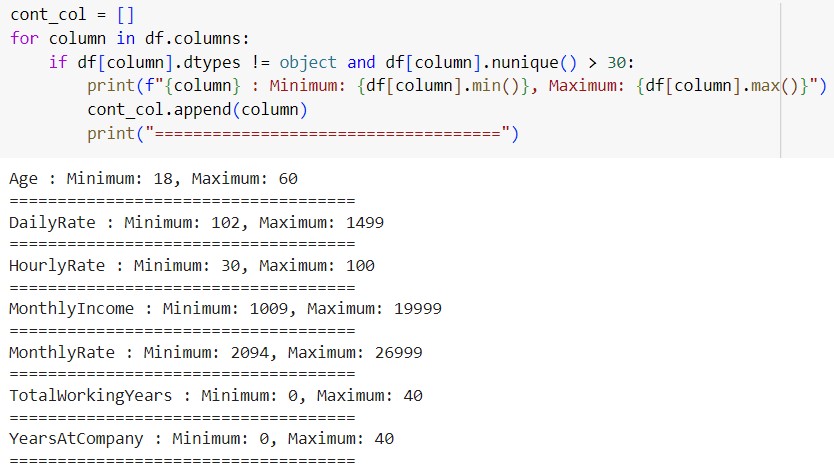


* One-hot encoding is performed for categorical variables.
* Label encoding is applied to the 'Attrition' column.
* Numeric and categorical columns are segregated.

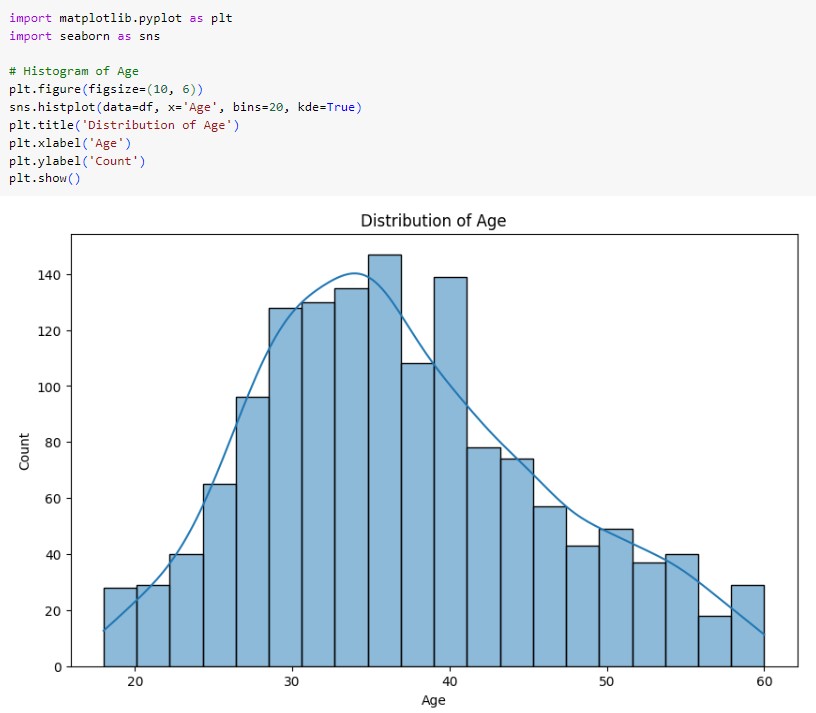


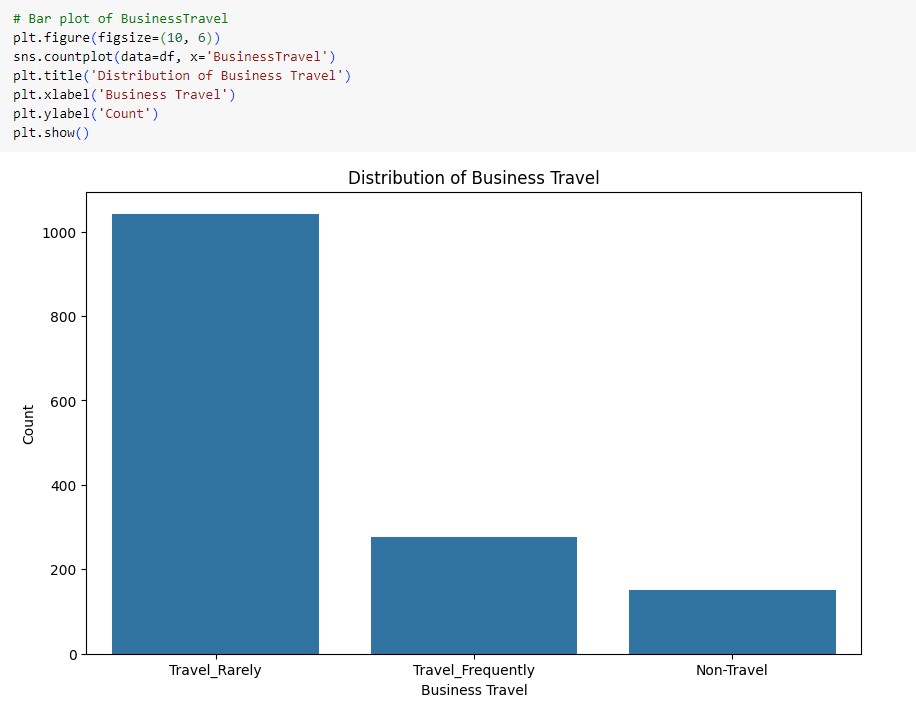


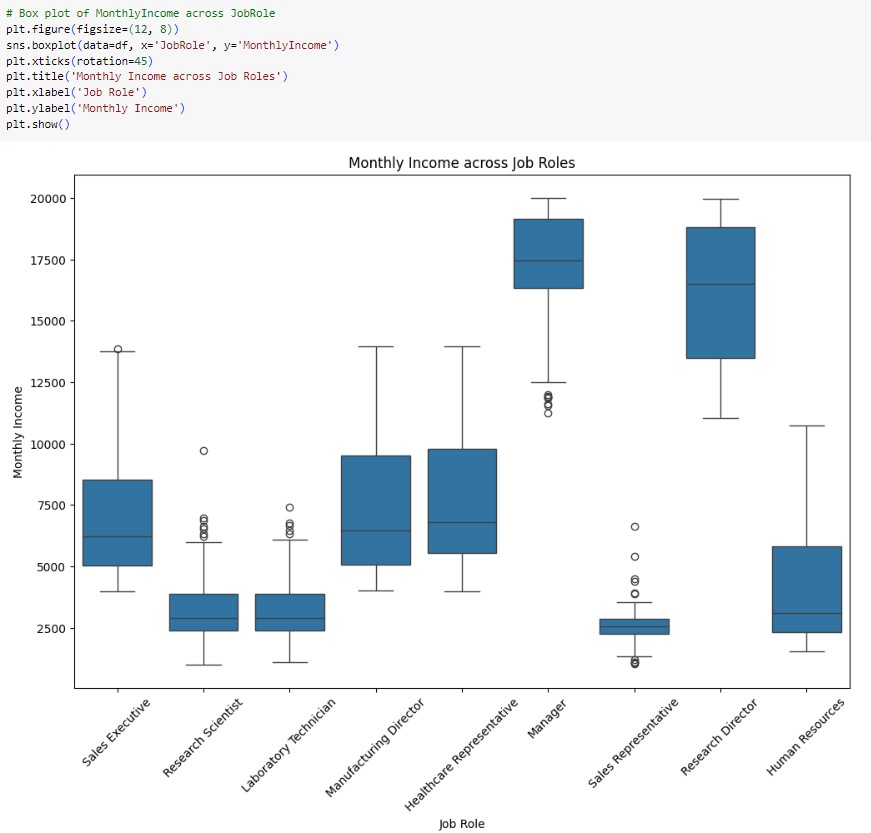


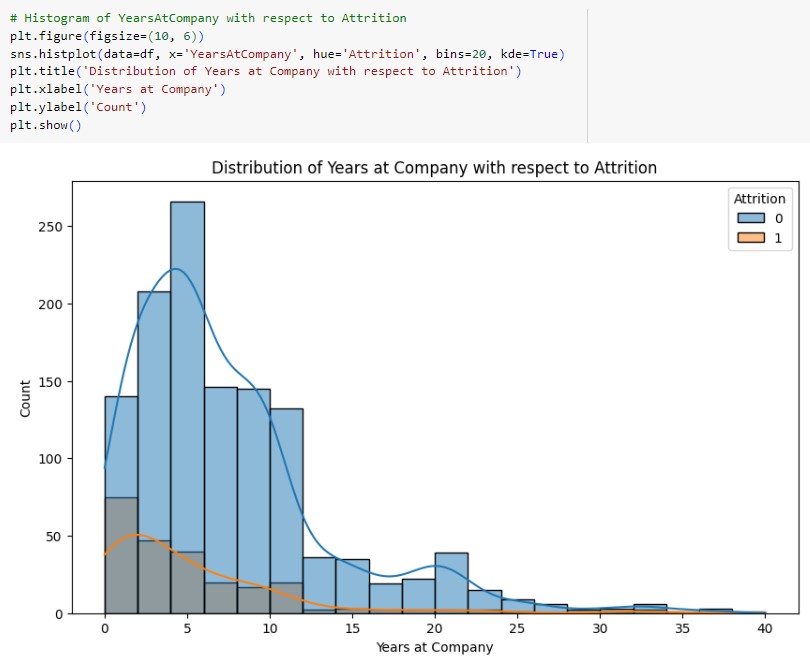


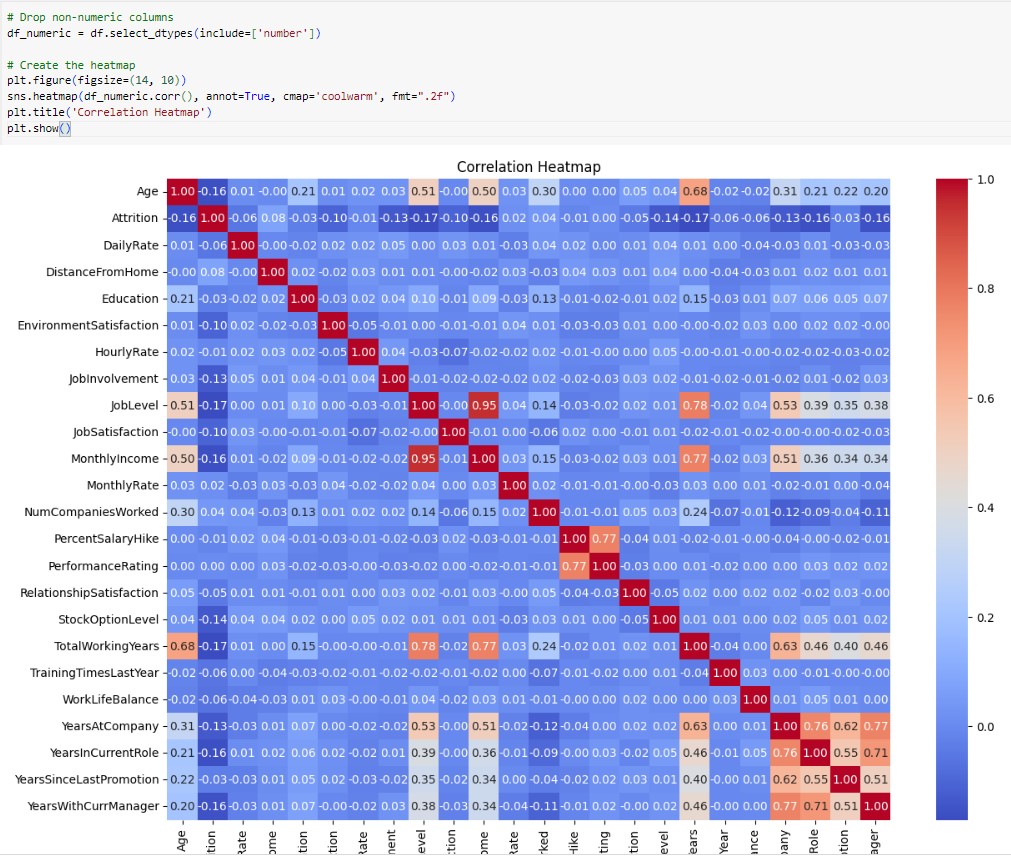
* Data Visualization:
* Histogram of 'Age', bar plot of 'BusinessTravel', and box plot of 'MonthlyIncome' across 'JobRole' are plotted.
* Histogram of 'YearsAtCompany' with respect to 'Attrition' is visualized.
* Correlation heatmap of numerical features is generated.









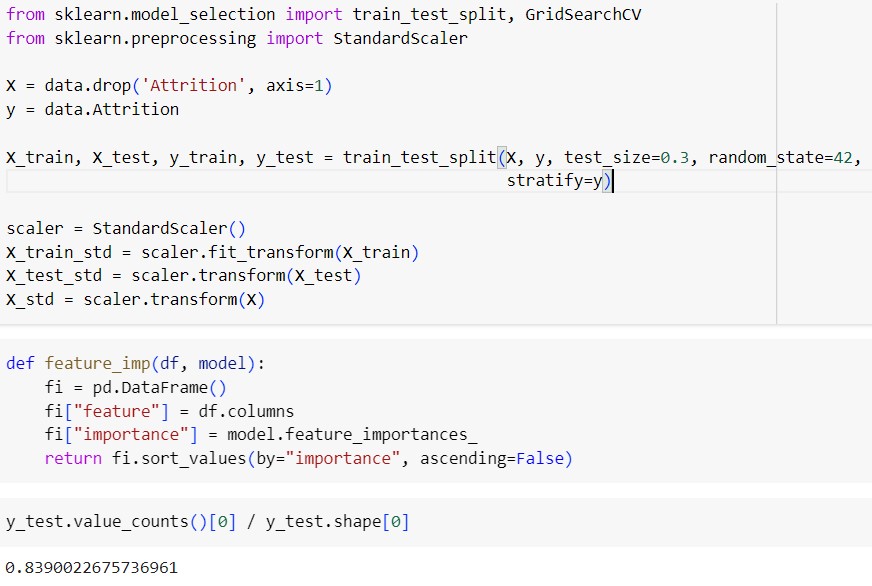


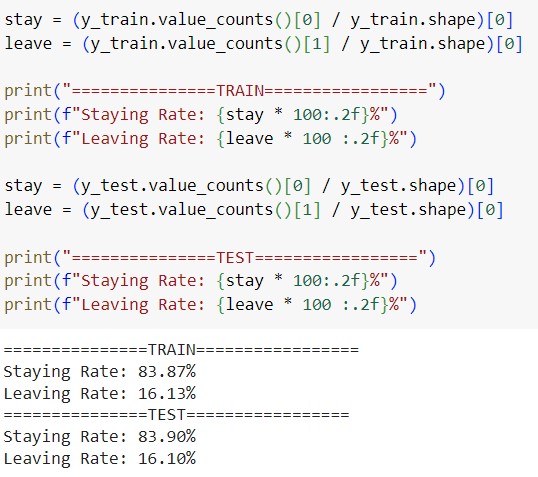
1. **Model Development**

* Feature Selection:
* Features with low correlation to the target variable 'Attrition' are removed.



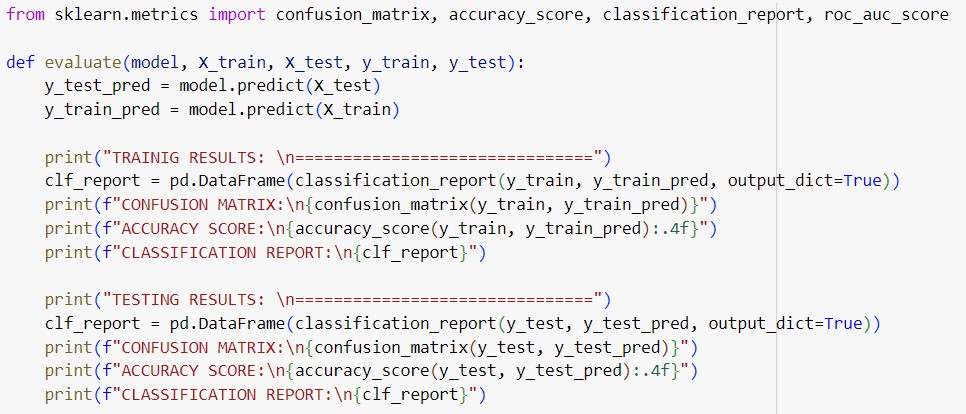
* Model Training:
* Data is split into training and testing sets.
* Standard scaling is applied to the features.
* Logistic Regression and AdaBoost classifiers are trained on the dataset.

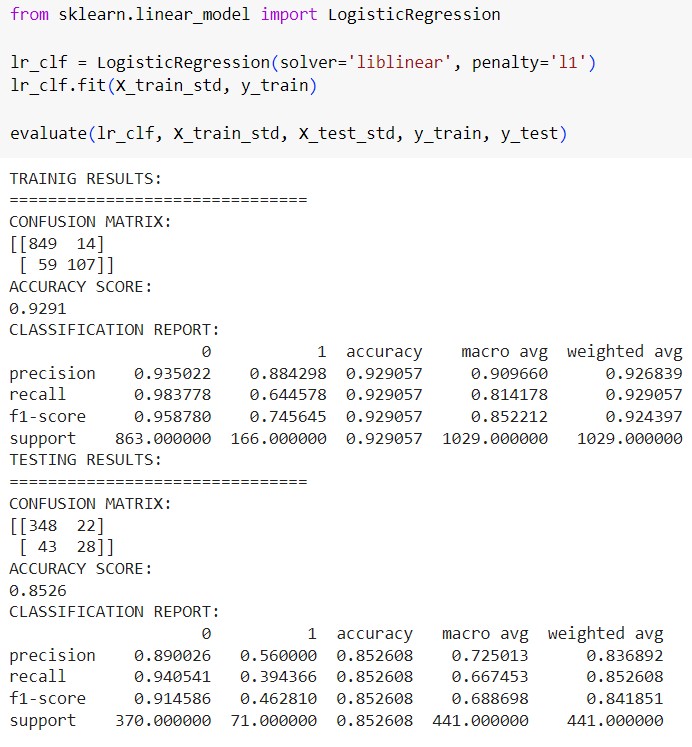


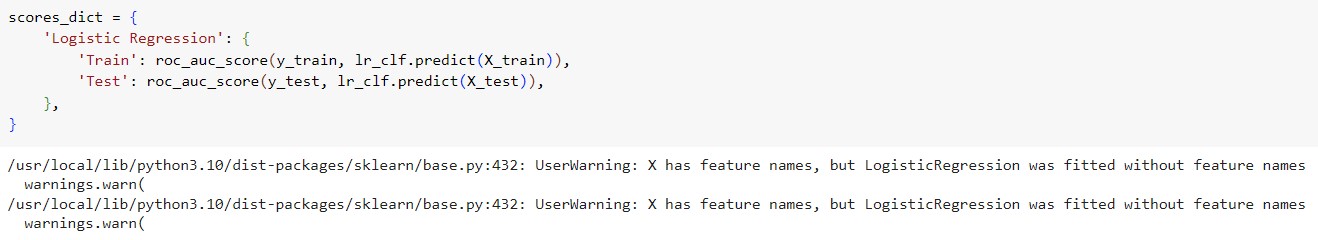


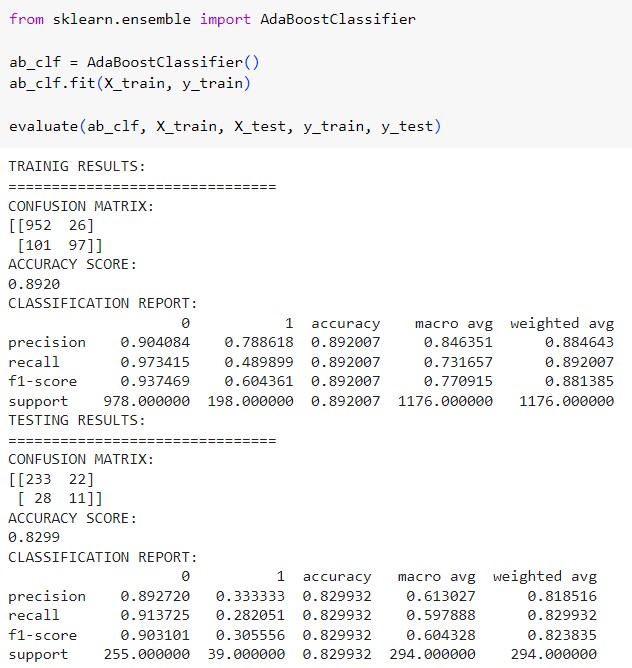
1. **Model Evaluation**

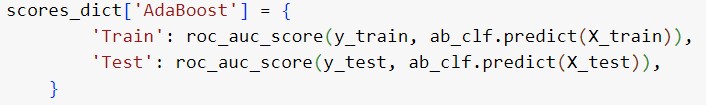
* Evaluation Metrics:
* Confusion matrix, accuracy score, precision, recall, and F1-score are computed for both training and testing sets.
* ROC AUC Scores:
* ROC AUC scores are calculated for Logistic Regression and AdaBoost classifiers.







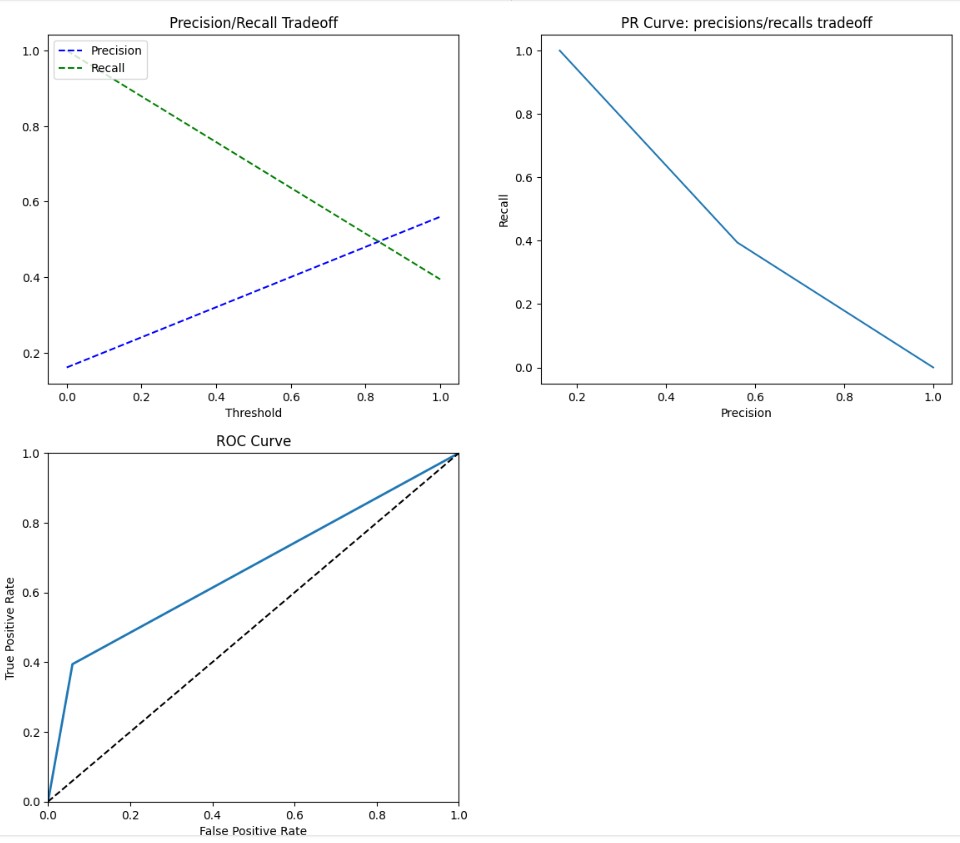


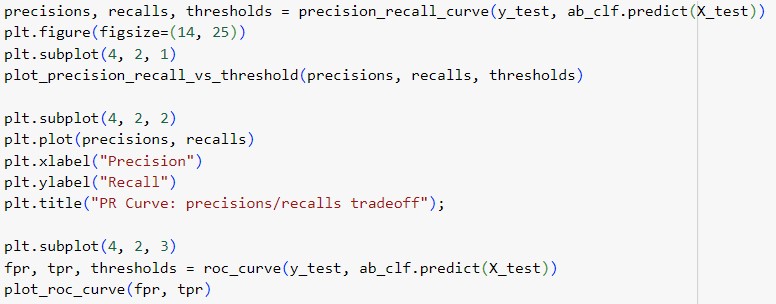


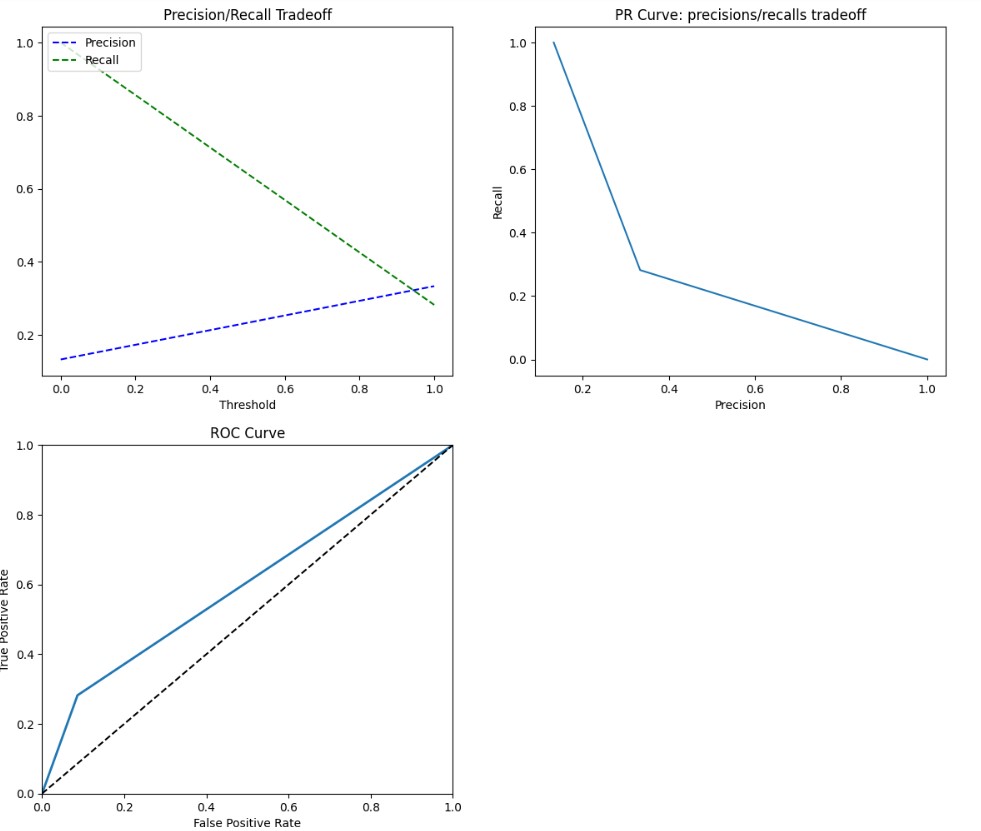
1. **Optimization Techniques**

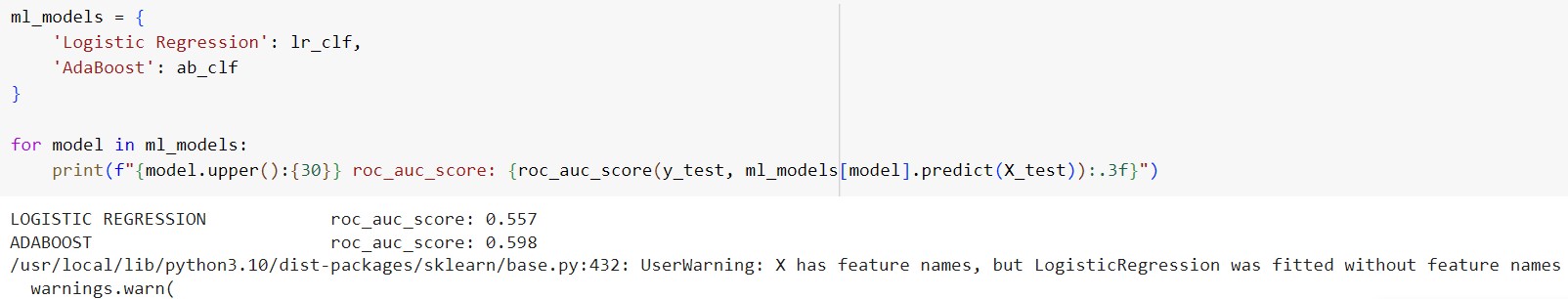
* Further Evaluation:
* Precision-recall curves and ROC curves are plotted to visualize model performance.
* Model Comparison:
* Performance metrics and ROC AUC scores are compared between the Logistic Regression and AdaBoost models.
* Model Scores Visualization:
* Model scores are visualized using a horizontal bar plot.

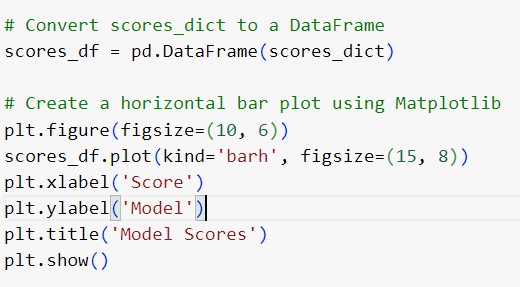


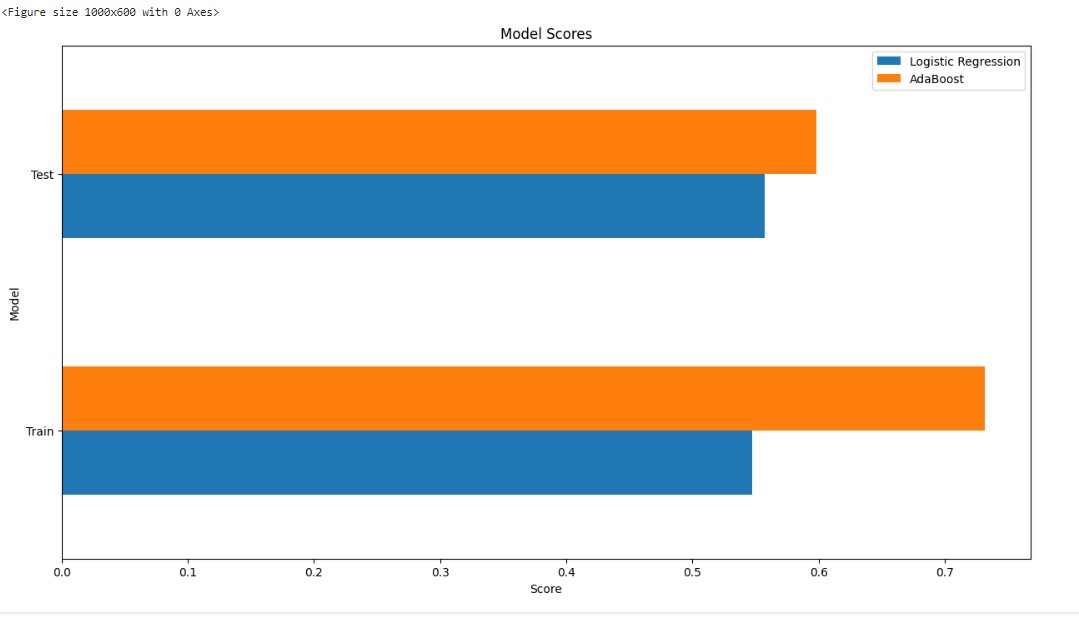












1. **Summary**

* Findings:
* The dataset is imbalanced with approximately 84% of employees staying and 16% leaving.
* Both Logistic Regression and AdaBoost models achieved reasonable accuracy, but their ROC AUC scores suggest room for improvement.
* Challenges:
* Dealing with imbalanced data and interpreting complex model results were major challenges encountered.
* Recommendations:
* Addressing imbalanced data, exploring feature importance, and further evaluation of models are recommended for improving performance and insights.