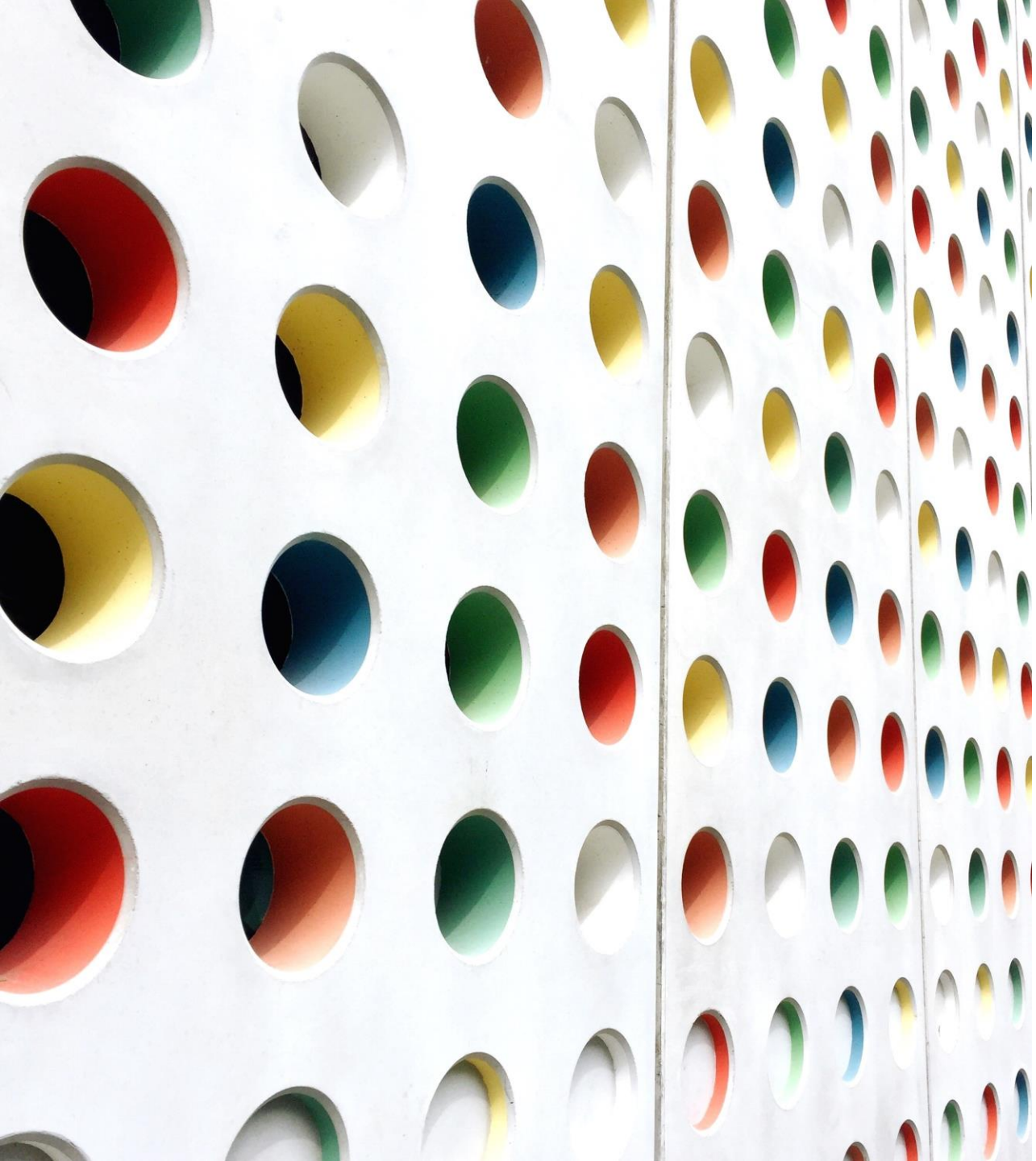


Build a predictive model for employee attrition on Python

PRESENTED BY: AJIT KASHID

BATCH : DS-7

Introduction

Overview: Introduction to the importance of predicting employee attrition for tech firms.

Data Source:

For this Project, an HR dataset named '**IBM HR Analytics Employee Attrition & Performance**', has been picked, which is available on website
IBM dataset with 1,470 employees.

Objective: **Understanding HR Analytics - Scope, Applications and Impact**

- Build a predictive model to identify employees at risk of leaving.
- It has information about employee's current employment status,
- The total number of companies worked for in the past,
- Total number of years at he current company and the current roles, etc.

Tools and skills

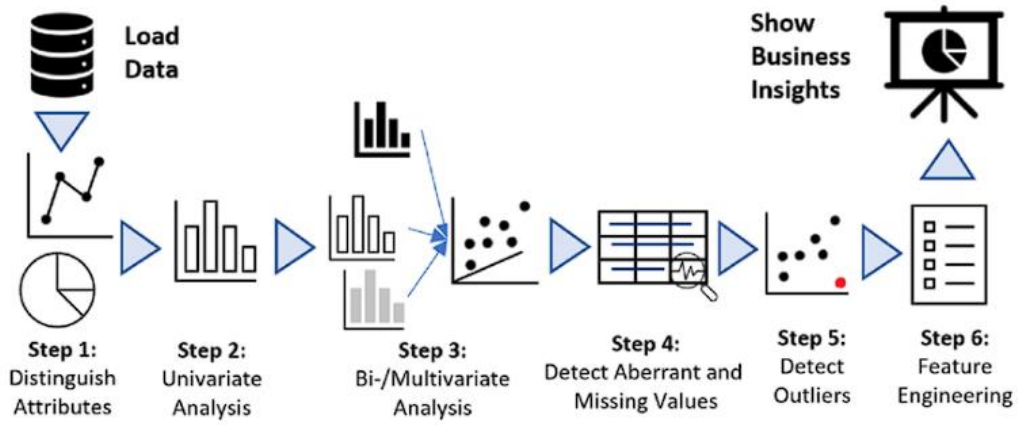
- **Created a research report for HR analysis with below points**

  Origin & history of HR analytics
  Scope of HR analytics
  Applications of HR analytics with examples of practical implementation
  Business impact of HR analytics

- **Reading your data into a Data frame in Python**

  Python includes many package such as Pandas, NumPy, Matplotlib, and seaborn

- **Running the ML Models for the Problem**
  - **Logistic Regression Model**
  - **Decision Tree Model**
  - **Random Forest Model**
  - **ROC Curves**



Data Analysis and Exploration

- **Dataset Overview:**
  - Data collection: The data gleaned was structured data, and it consisted of 1470 rows and 35 columns.
  - Key Features: Age, Monthly Income, Overtime, Gender, Marital Status, Job Role, etc

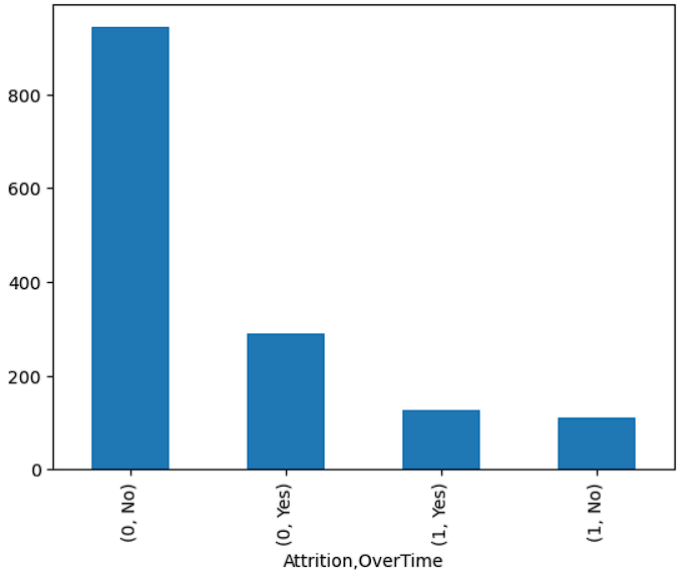
  .
- **Initial Observations:**
  No duplicated values.
  No null values.

  - **Data transformation:** Categorical data was transformed using Ordinal Encoder. The data was generally clean without missing values.

  - **Data visualization:** Here, information was displayed using histograms, heatmaps, and other visual aids to facilitate clear and simple interpretation.

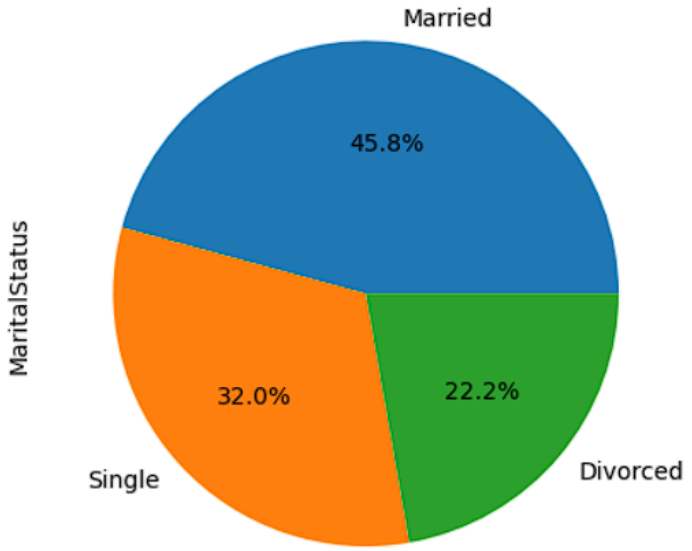
Exploratory Data Analysis (EDA)

•**Attrition vs Overtime:**
•Employees working overtime show higher attrition rates.




```
Attrition
0    37.561233
1    33.607595
Name: Age, dtype: float64
```

•**Marital Status vs Attrition**:
•Higher attrition rates among single employees.

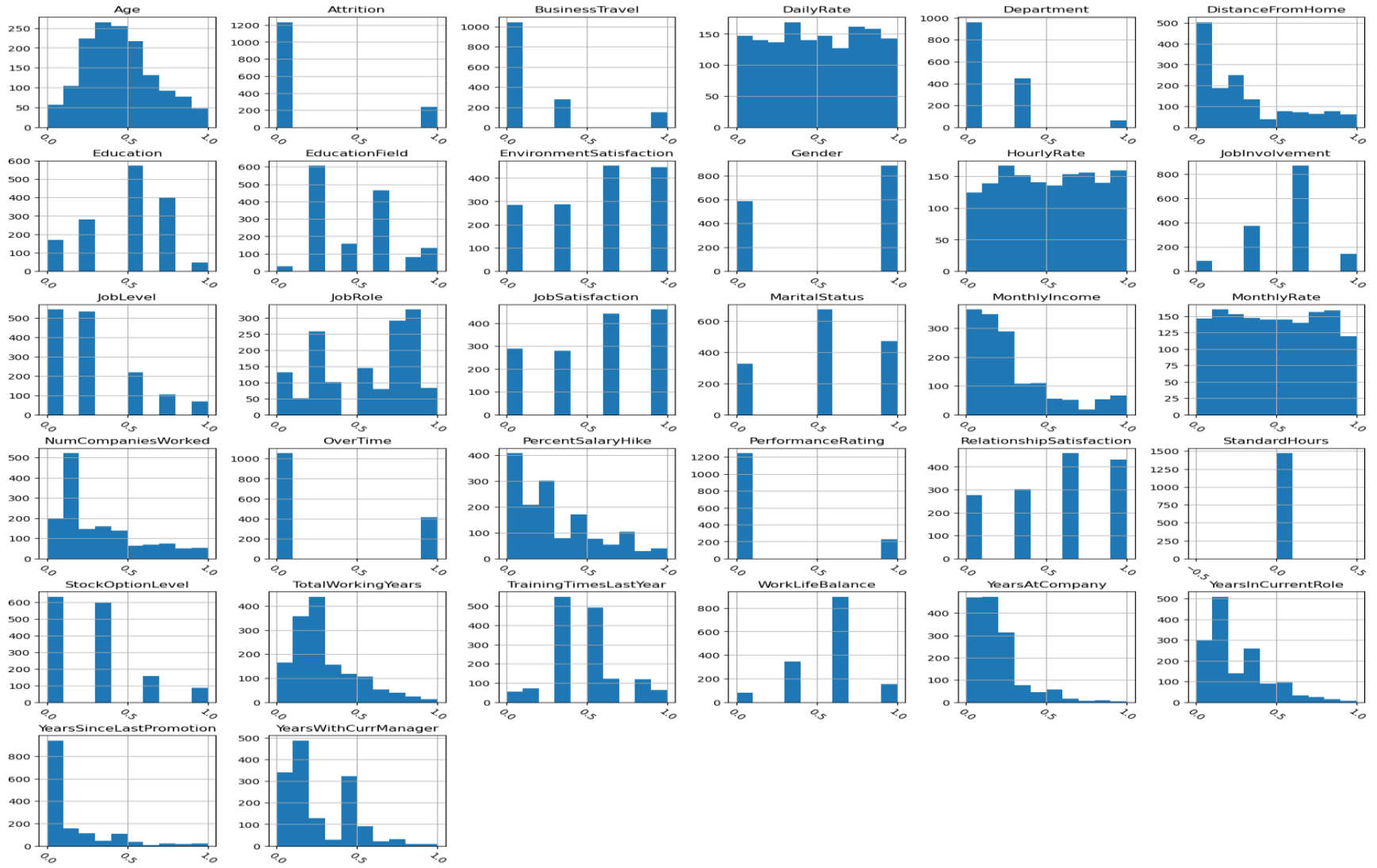


Exploratory Data Analysis (EDA)

•Monthly Income has a strong positive correlation with Job Level. This means the higher job levels usually have a higher monthly income.

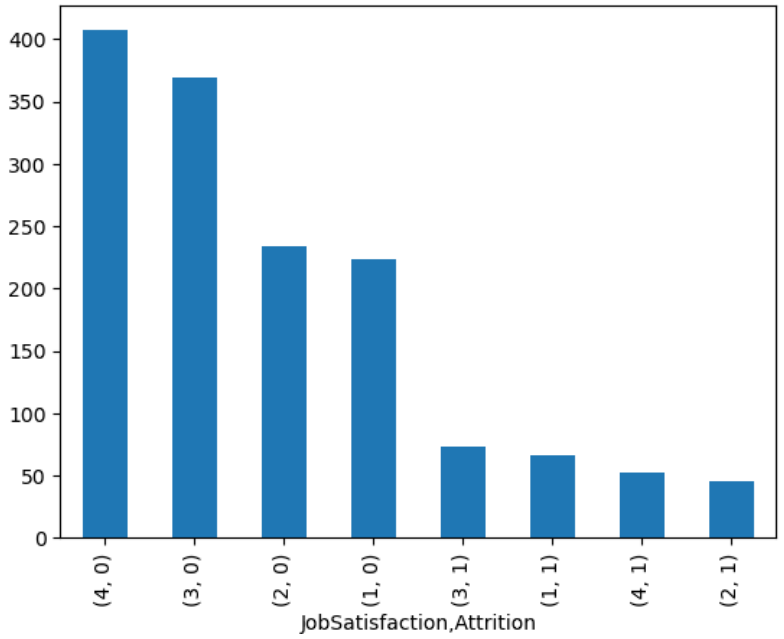
•Performance rating has a positive correlation with Percent Salary Hike. This means that an increase in salary is related with an increase in performance.

•Total Working years has a positive correlation with the employee's age, Job Level, and Monthly Income. This is understandable as you gain more experience in years and salary increases as you age.

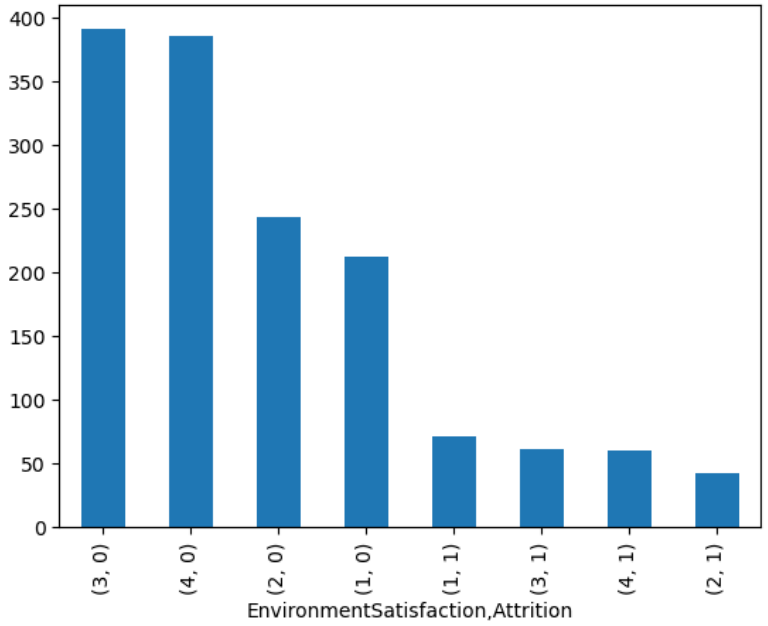


Exploratory Data Analysis (EDA)

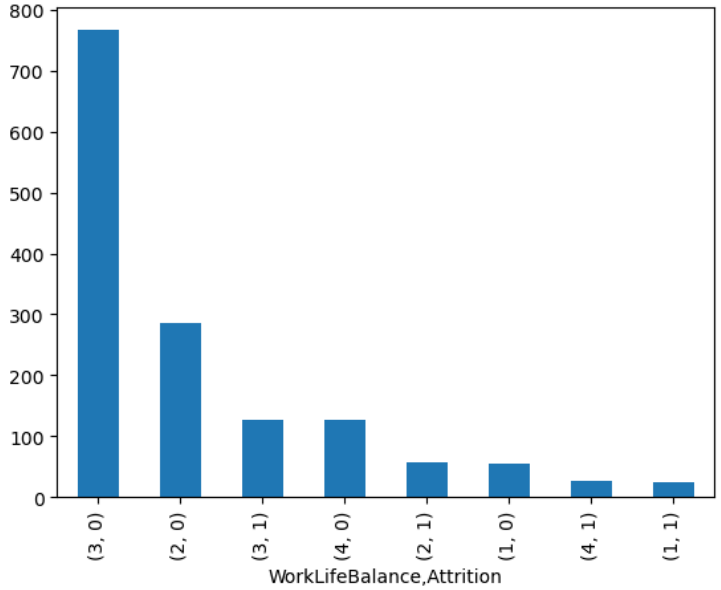
Job Satisfaction



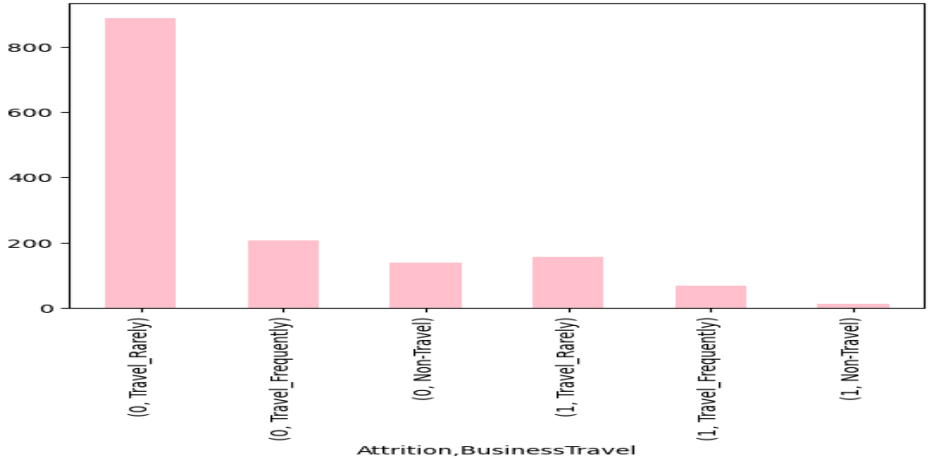
Environment Satisfaction



Work Life Balance

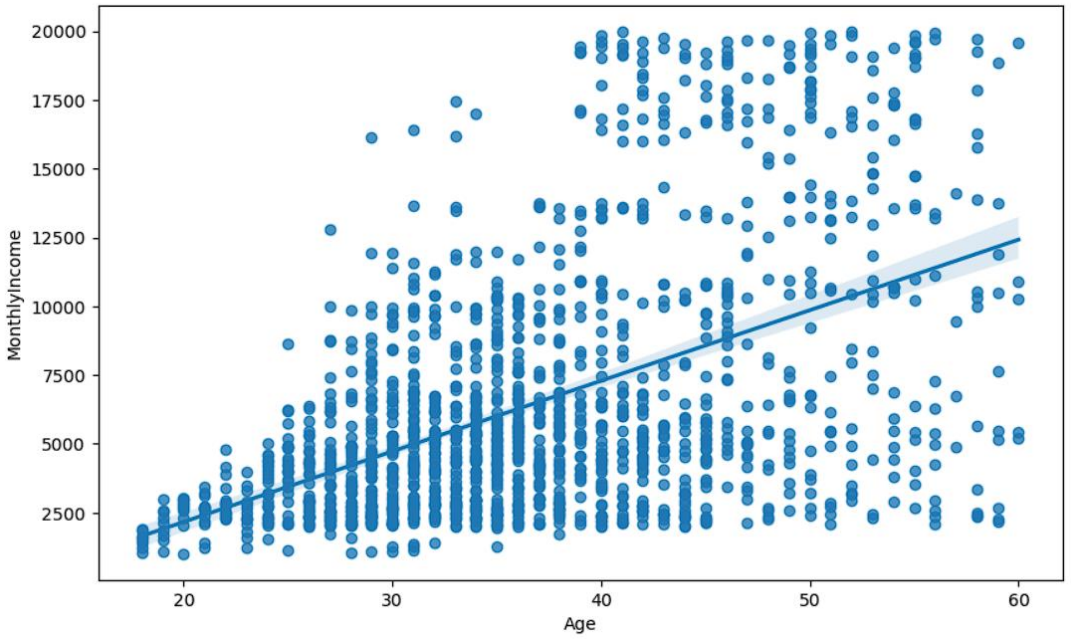


Business Travel



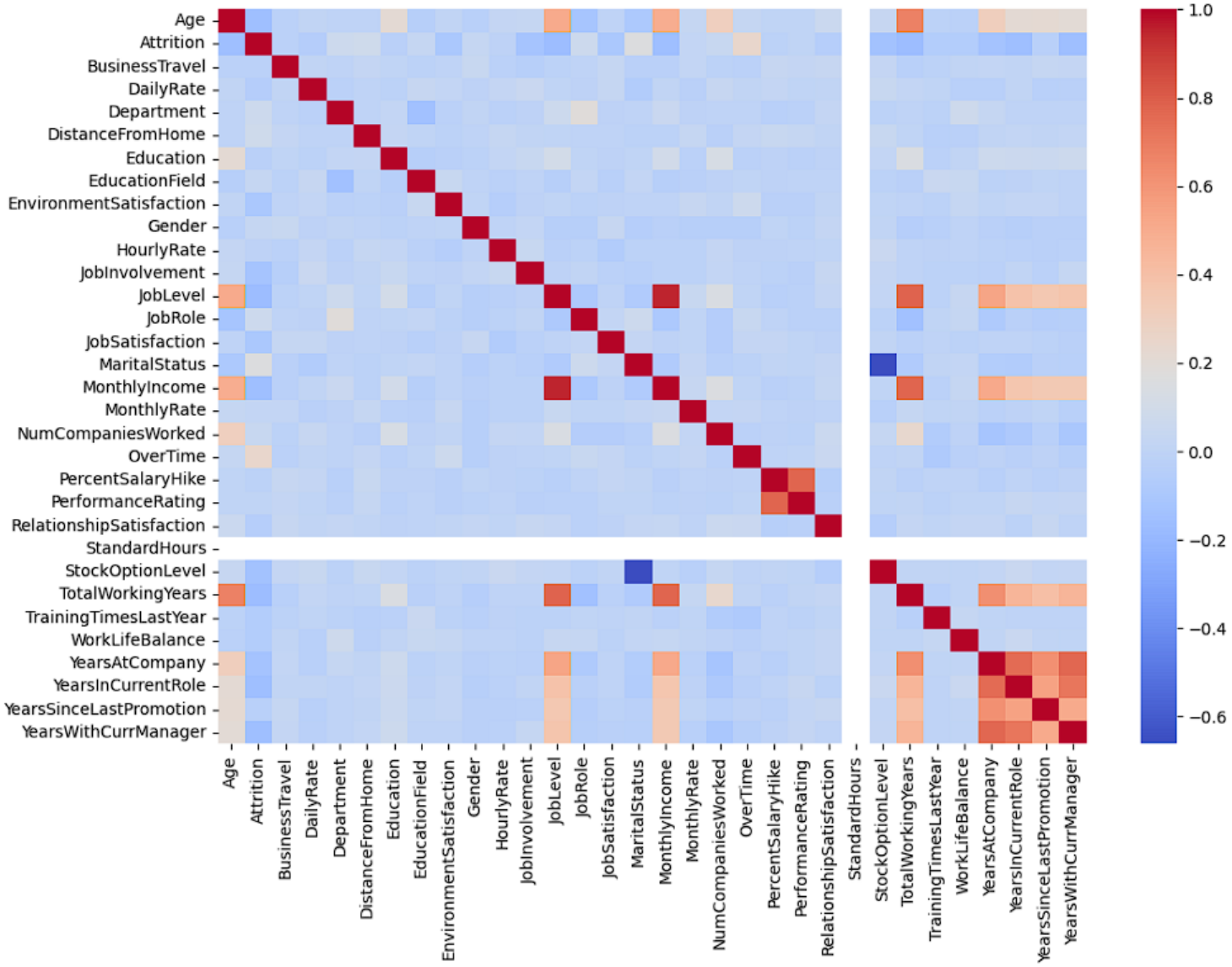
Exploratory Data Analysis (EDA)

- **Age vs Monthly Income:**
- Normal distribution observed.
- Higher attrition among lower income earners.



Exploratory Data Analysis (EDA)

This step was taken to better understand the data that had been gathered, give a more full picture of the data, uncover and comprehend patterns that would explain unexpected results.



Model Selection

- **Decision Tree:**

- Ease of understanding and visualization.
- Accuracy: 82%

- **Random Forest:**

- Robustness and accuracy through ensemble learning.
- Accuracy: 88%

- **Logistic Regression:**

- Simplicity and interpretability.
- Accuracy: 89%

```
Classifier: Decision Tree
Accuracy: 0.78
Precision: 0.20
Recall: 0.21
F1 Score: 0.20
F2 Score: 0.20
---------------------------
Classifier: Random Forest
Accuracy: 0.87
Precision: 0.67
Recall: 0.10
F1 Score: 0.18
F2 Score: 0.12
---------------------------
Classifier: Logistic Regression
Accuracy: 0.89
Precision: 0.79
Recall: 0.28
F1 Score: 0.42
F2 Score: 0.32
---------------------------
```

Model Building and Evaluation

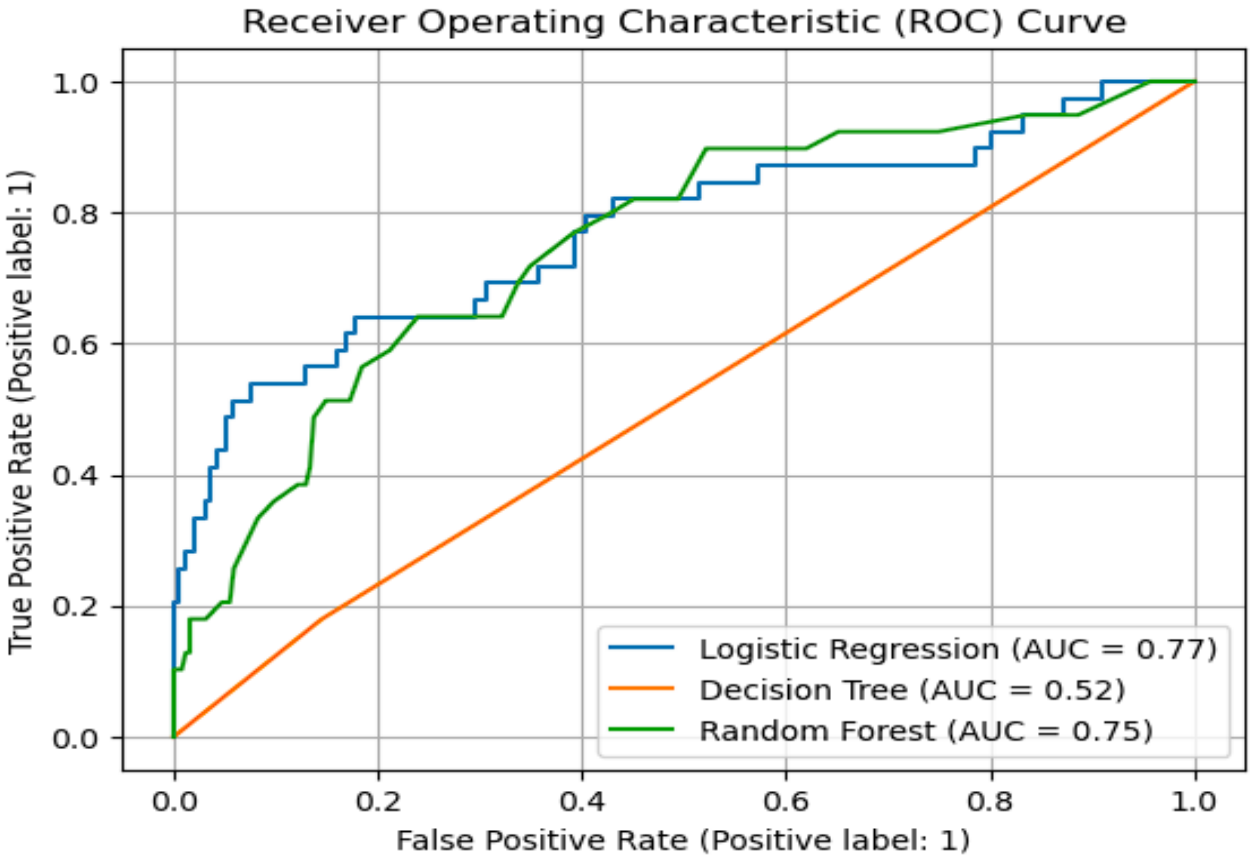
Example results

Accuracy: Evaluate the model's accuracy on the test set to gauge overall performance. Precision and Recall: Assess precision (accuracy of positive predictions) and recall (true positive rate) to understand model effectiveness in classifying 'Attrition' or other outcomes. Confusion Matrix: Analyze the confusion matrix to see how well the model predicts each class (true positives, false positives, true negatives, false negatives). Feature Importance: Random Forests can provide feature importance scores, indicating which features contributed most to the model's predictions. Generalization: Check for signs of overfitting or underfitting by comparing performance on training versus test data.

Benefits:

Robustness: Handles noisy data and missing values well. Scalability: Scales effectively to large datasets and parallel processing. Versatility: Suitable for a wide range of applications, from healthcare to finance and beyond.

ROC Curves: Visualization of model performance.



Conclusion:

Conclusion:
Based on the accuracy calculations from the confusion matrices:

1. Logistic Regression has an accuracy of approximately 89%.
2. Decision Tree has an accuracy of approximately 82%.
3. Random Forest has an accuracy of approximately 88%.

Logistic Regression has the highest accuracy among the three models based on the provided confusion matrices.