**Supervised Learning:**

**Loan Default Prediction**

**Business Use Case:**

The objective is to predict whether an employee is likely to leave the company (attrition). This is a binary classification task.

**Data Understanding:**

* Data Source: This dataset has been taken from Kaggle’s HR Analytics Dataset.
* The dataset contains 1480 rows and 38 columns in total.

**Data Preparation:**

Packages: NumPy, Pandas, Matplotlib, Seaborn, Sklearn.

**Data Preprocessing:**

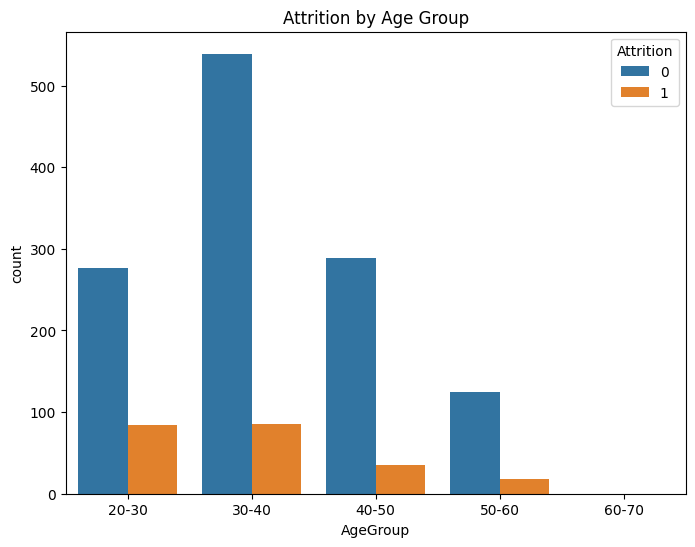
**Data Preprocessing**

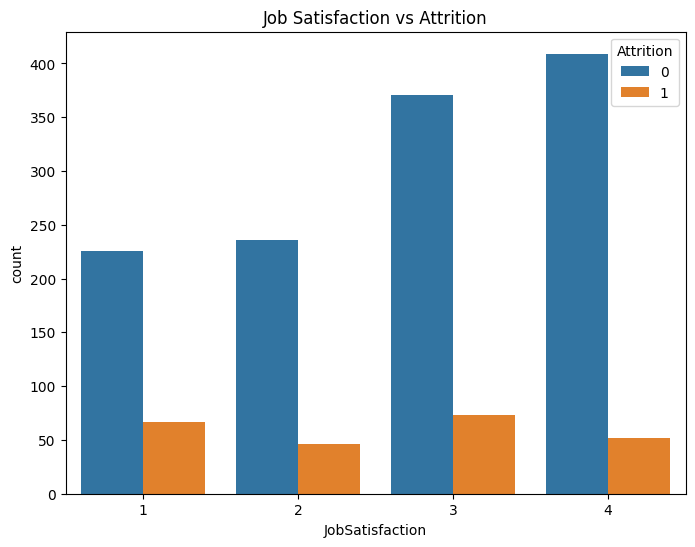
* **Initial Dataset**: The dataset contains information about employees, such as demographics (e.g., Age, Gender, Marital Status), job-related attributes (e.g., JobRole, JobSatisfaction), and performance metrics (e.g., MonthlyIncome, PerformanceRating).
* **One-Hot Encoding**: You applied one-hot encoding to multi-class categorical columns: BusinessTravel, Department, EducationField, and JobRole. This transforms these columns into multiple binary columns, one for each category, to make them suitable for machine learning models.

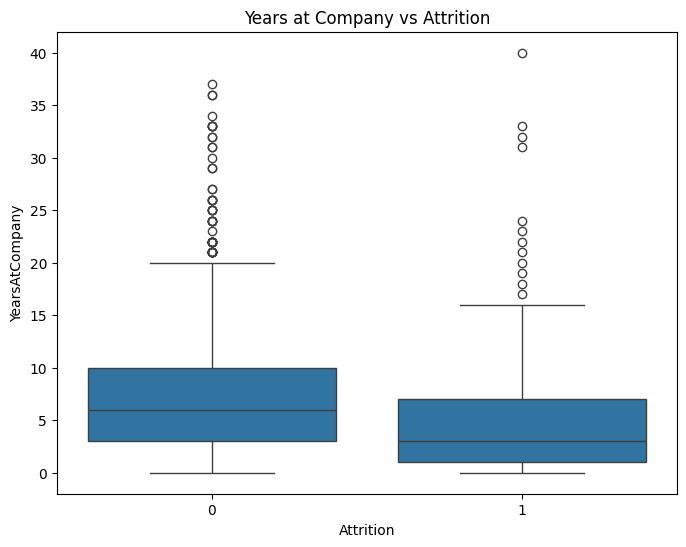
For example, the BusinessTravel column with categories Travel\_Rarely, Travel\_Frequently, and Non-Travel is encoded into separate columns, each representing one of these categories with True/False values.

* **Label Encoding**: You used label encoding for binary columns (e.g., Attrition, Gender, OverTime, etc.). Label encoding is a technique that converts categorical labels into numerical values (0 or 1), which simplifies their use in machine learning models.
* **Dropping Unnecessary Columns**: You dropped irrelevant columns such as EmpID, EmployeeNumber, AgeGroup, SalarySlab, and others that might not contribute to the model, particularly EmpID and EmployeeNumber which are unique identifiers.

**Data Visualization:**







**Building the Model**

* **Target Variable (y)**: The target variable is Attrition, which indicates whether an employee left the company or stayed (binary classification: Yes/No).
* **Features (X)**: The rest of the columns are used as features (input data) for predicting attrition.
* **Data Split**: You split the dataset into training and testing sets using an 80-20 ratio (train\_test\_split), ensuring the model can be trained and then evaluated on separate data.

**Model Training and Evaluation**

* **Model**: You used a **Random Forest Classifier** with 100 estimators to train the model on the training dataset (X\_train and y\_train).
* **Model Accuracy**: After making predictions (y\_pred) on the test data (X\_test), the accuracy of the model was found to be **86.82%**, indicating good performance in predicting employee attrition.

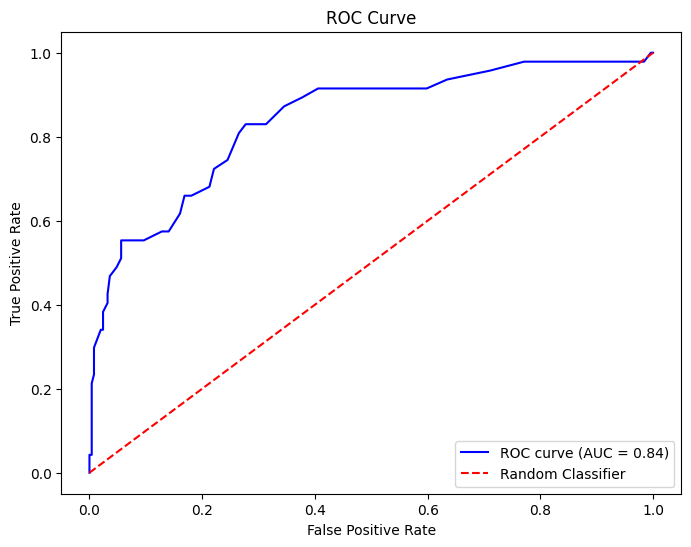
**Evaluation Metrics**

* **Confusion Matrix**: You printed the confusion matrix, which shows how well the model performs in terms of True Positives, True Negatives, False Positives, and False Negatives:

[[248 1]

[ 38 9]]

* **Classification Report**: This report gives precision, recall, F1-score, and support for both classes (Attrition = 0 and Attrition = 1):
  + **Class 0 (Attrition = 0)**: The model has a high precision and recall for this class, with 87% precision and nearly 100% recall.
  + **Class 1 (Attrition = 1)**: The model has a lower recall (19%) and precision (90%), meaning it struggles to detect attrition (employees who leave).
* **ROC Curve**: You plotted the **Receiver Operating Characteristic (ROC) curve**, which helps to evaluate the model's performance in distinguishing between the two classes (attrition vs. non-attrition). The curve compares the True Positive Rate against the False Positive Rate, with the area under the curve (AUC) providing an overall measure of model performance.



**Conclusion**

This project involved building a classification model to predict employee attrition based on various features. After preprocessing the data (one-hot encoding, label encoding, and dropping irrelevant columns), you trained a Random Forest model. While the model performed well overall, it had difficulty identifying employees who left the company (attrition), as shown by the low recall for that class. The final model achieved an accuracy of **86.82%** with a promising ROC curve.