ISOM 835 Term Project: Predicting Employee Attrition

Course: ISOM 835  
Name: Kajol Parche  
Date: 05/06/2025

GoogleColab Link - <https://colab.research.google.com/drive/1umyE9NJbAPftulTQJZzKZNhuxFhyLrvY?usp=sharing>

# Introduction & Dataset Description

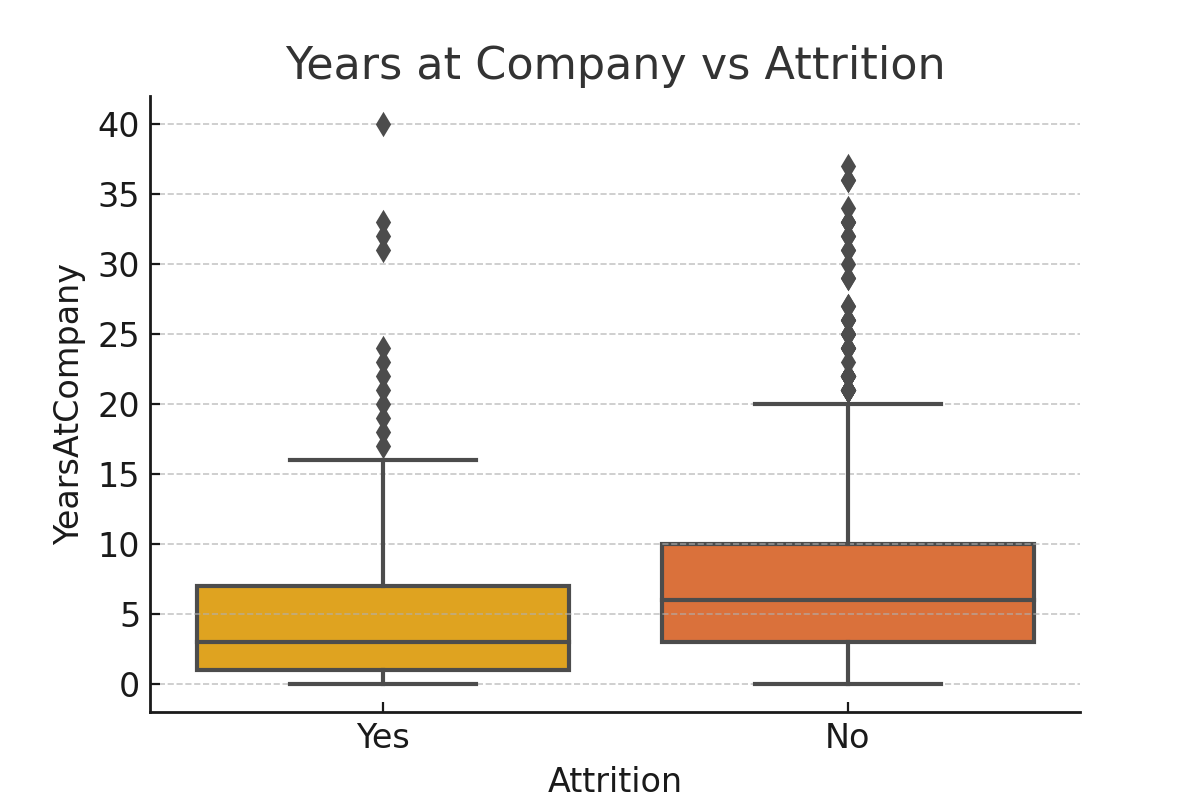
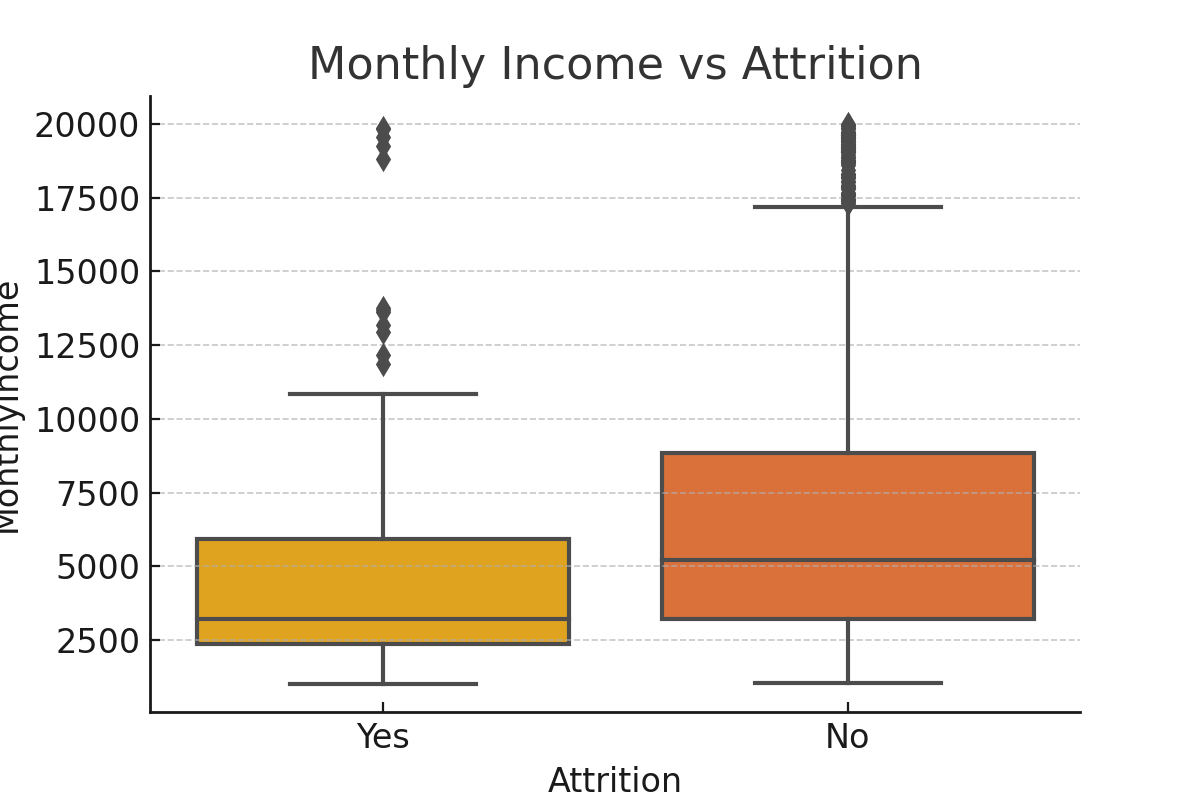
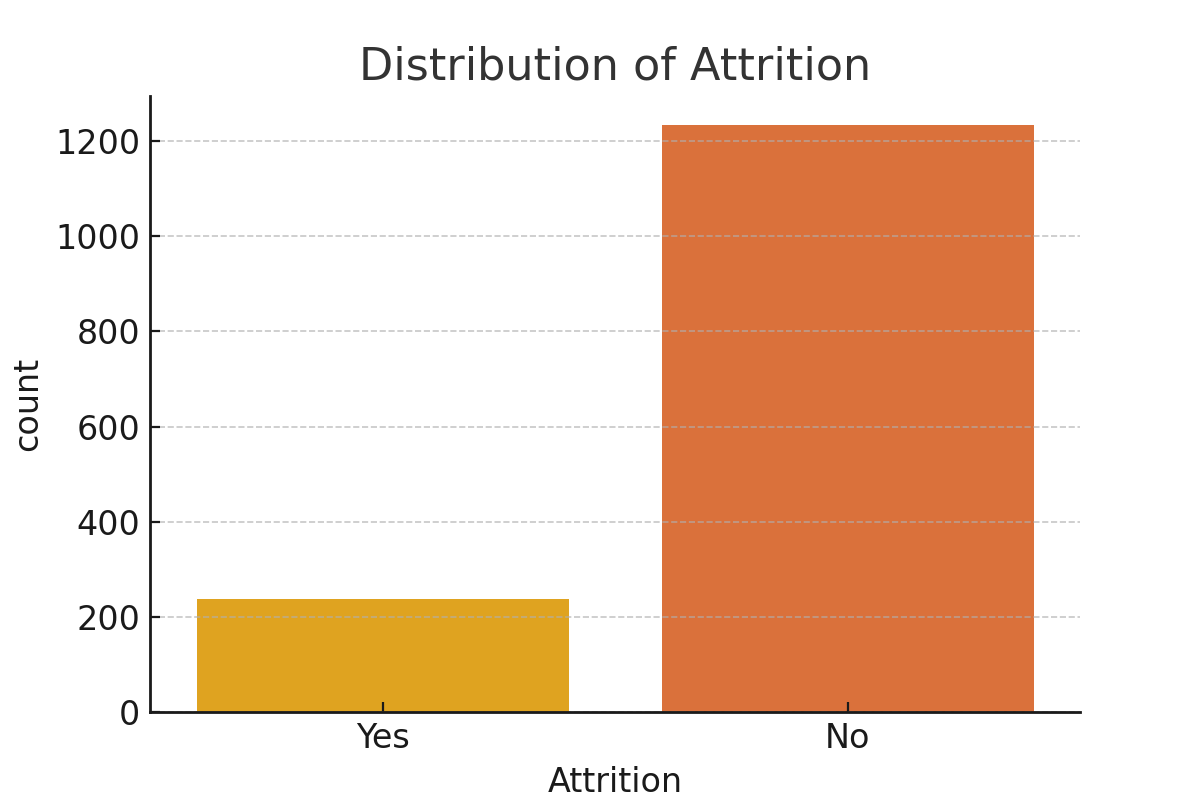
This project explores employee attrition using IBM’s HR dataset, which contains 1,470 records with a mix of demographic, job-related, and performance features. Key attributes include age, monthly income, job role, marital status, overtime status, and satisfaction scores. The target variable, Attrition, indicates whether an employee has left the company.

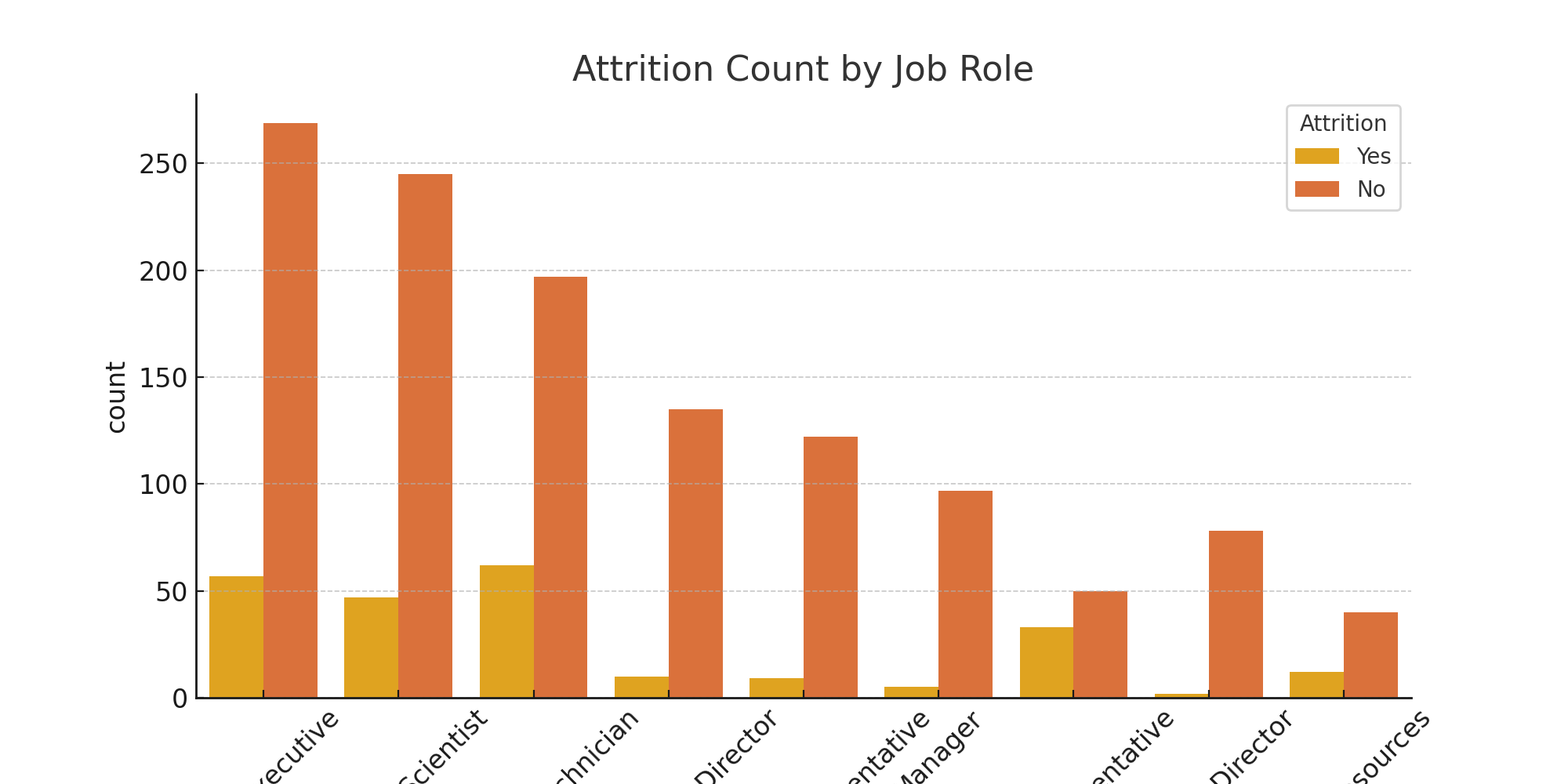
The goal of this project is to identify the main drivers of attrition and build predictive models to assess which employees may be at risk of leaving. By leveraging exploratory analysis and machine learning, the project aims to generate insights that can support HR teams in making informed, proactive retention decisions.

# Exploratory Data Analysis (EDA)

EDA was performed to identify patterns and relationships within the dataset that may influence employee attrition. The target variable, Attrition, showed a class imbalance with approximately 16% of employees marked as having left the organization. Through visualizations such as boxplots, count plots, and correlation heatmaps, several meaningful trends were uncovered.

Employees who left were generally younger, had lower monthly income, and shorter tenure at the company compared to those who stayed. Overtime status emerged as a strong differentiator those working overtime had significantly higher attrition rates. Correlation analysis of numerical features further supported these findings, showing negative relationships between attrition and factors such as income, years at company, and age. These insights helped guide feature selection and informed the modeling phase of the project.





A blue and white graph

AI-generated content may be incorrect.

# Data Cleaning and Preprocessing

The dataset was first checked for quality and completeness, and no missing values were found. We removed non-informative or redundant columns such as EmployeeCount, Over18, StandardHours, and EmployeeNumber to reduce noise. Binary variables like Attrition and OverTime were label-encoded, while multi-class categorical features such as JobRole, MaritalStatus, and Department were one-hot encoded to prepare them for modeling.

Numerical features were standardized using StandardScaler to ensure consistent scaling, particularly for algorithms like Logistic Regression. The data was then split into training and testing sets using an 80/20 **stratified split**, preserving the original attrition ratio. These preprocessing steps ensured the dataset was clean, well-structured, and ready for effective model training.

# Formulating Business Analytics Questions

1. What are the main drivers of employee attrition?

* Our analysis identified key drivers of attrition as overtime, lower monthly income, shorter tenure, and younger age, indicating that work pressure and compensation dissatisfaction significantly influence employee turnover.

2. Can we predict if an employee is likely to leave?

* Yes, by using predictive models like Logistic Regression and Random Forest, we can estimate the likelihood of attrition based on factors such as job role, income, overtime, and years at the company.

3. What differentiates employees who leave from those who stay?

* Employees who leave tend to be newer, earn less, and are more likely to work overtime, whereas those who stay typically have longer tenure, higher income, and more stable work schedules.

# Predictive Modeling

We implemented two models to predict employee attrition: Logistic Regression for its interpretability and Random Forest for its ability to capture complex patterns. Features were standardized for Logistic Regression, while Random Forest was applied to the raw data.

Both models were evaluated using Accuracy, Precision, Recall, F1 Score, and AUC. Random Forest achieved higher accuracy and AUC, making it effective overall, while Logistic Regression had better recall, making it more suitable for identifying at-risk employees. This balance highlights the trade-off between interpretability and predictive power, with each model offering unique value depending on business priorities.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score | AUC |
| Logistic Regression | 0.7483 | 0.3412 | 0.617 | 0.4394 | 0.6952 |
| Random Forest | 0.8299 | 0.3636 | 0.0851 | 0.1379 | 0.5284 |

# Insights and Answers

The analysis identified key drivers of attrition, including OverTime, Monthly Income, Years at Company, and Age. Employees who are younger, earn less, have shorter tenure, and work overtime are more likely to leave. These patterns were consistent across both EDA and model outputs.

These insights can help HR teams proactively address attrition by focusing on high-risk profiles. For instance, reducing excessive overtime or offering career development support to newer employees may improve retention. The predictive models serve as useful tools to guide such data-informed interventions.

A graph of a number of people

AI-generated content may be incorrect.

# Ethics and Interpretability

In this project, predictive models were developed to identify employees at risk of attrition using HR data. These models should be used ethically—to support, not penalize employees. For example, a high-risk prediction should prompt proactive support, not exclusion from opportunities. Sensitive variables like age or marital status must be used carefully to avoid bias or unfair outcomes.

Interpretability is essential for HR acceptance. Logistic Regression provided transparency through easily understood coefficients, while Random Forest highlighted key drivers of attrition via feature importance. Together, these models offer both predictive power and actionable insights, supporting fair, data-informed retention strategies that keep the human element at the core of decision-making.

**Appendix**

All analysis and modeling were conducted using Python in Google Colab. Below are the main code blocks used throughout the project:

* **Data Loading**: Used pandas.read\_excel() to import the dataset.
* **Exploratory Data Analysis (EDA)**:
  + Count plots and boxplots using seaborn and matplotlib
  + Correlation heatmap using .corr()
* **Data Cleaning & Preprocessing**:
  + Dropped non-informative columns: EmployeeCount, StandardHours, Over18, EmployeeNumber
  + Encoded binary fields (e.g., Attrition, OverTime)
  + One-hot encoded categorical variables using pd.get\_dummies()
  + Scaled numeric features using StandardScaler
  + Split data into training and test sets using train\_test\_split()
* **Modeling**:
  + Logistic Regression (LogisticRegression) with class\_weight='balanced'
  + Random Forest Classifier (RandomForestClassifier)
* **Evaluation Metrics**:
  + Calculated Accuracy, Precision, Recall, F1 Score, and AUC using sklearn.metrics

**B. Visualizations**

The following visualizations were generated to support EDA and insights:

* **Attrition Distribution** – Count plot of Yes vs No
* **Boxplots**:
  + Age vs Attrition
  + Monthly Income vs Attrition
  + Years at Company vs Attrition
* **Count Plot**: OverTime vs Attrition
* **Correlation Heatmap**: Numerical features with Attrition\_Flag