Hackathon Report:
Exploring Employee Attrition and Performance in a Corporate Environment

By

Emmanuelle Frappier

Business Requirements

The objective of this project is to help an organisation understand why employees leave and to build a model that can predict which employees are at risk of leaving. It also aims at conducting an in-depth analysis to discover patterns that could help the company better understand the causes

of employee turnover. By identifying potential leavers early, the company can take proactive measures to enhance retention and employee satisfaction, thus reducing recruitment costs

Understanding the Dataset

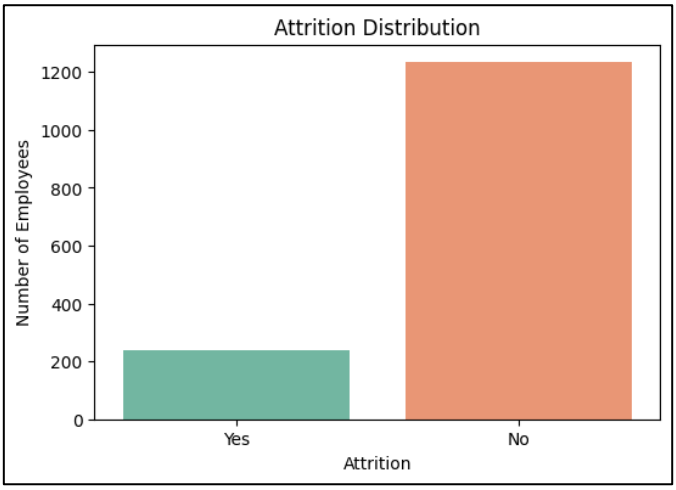
For this project, the IBM HR Analytics Employee Attrition & Performance dataset was used. The dataset contains the information such as:

- Personal Details (Age, Gender, Marital status, Distance from Home, Education)
- Job Details (Job Role, Department, Years at Company, Overtime, Job Level)
- Income (Monthly Salary, Percent Salary Hike)

The dataset consists of 1,470 rows and 35 columns. It is a relatively a small dataset containing 26 numerical features and 9 categorical features. There were no missing values in the data. The target column is Attrition, which indicates whether the employee stayed or left the company.

- Yes: The employee left the organisation
- No: The employee stayed in the organisation

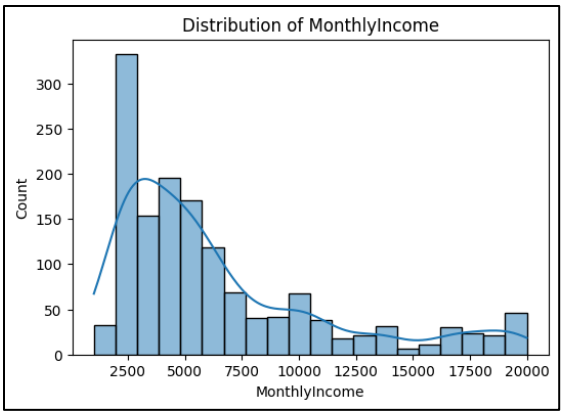
It is important to note that this dataset is imbalanced. About 84% of the employees stayed and only 16% left as shown in the figure below. This imbalance makes it harder for the model to learn patterns about leavers and predict a correct result.

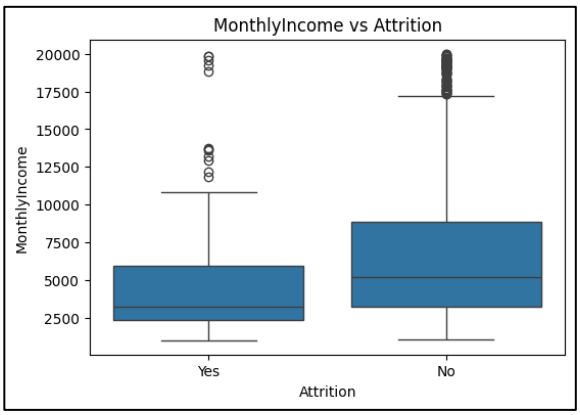


Exploratory Data Analysis and Observations

During the Exploratory Data Analysis (EDA), several important patterns were noted.

- **Monthly Income**

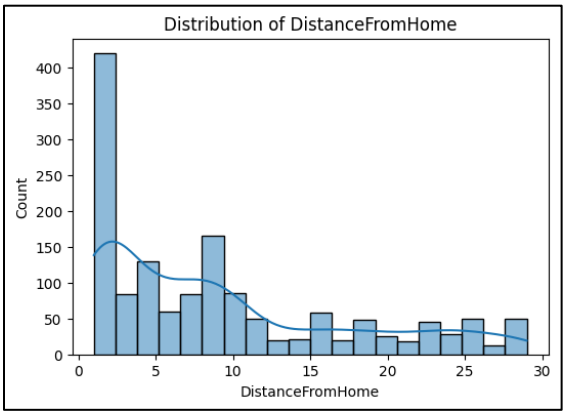


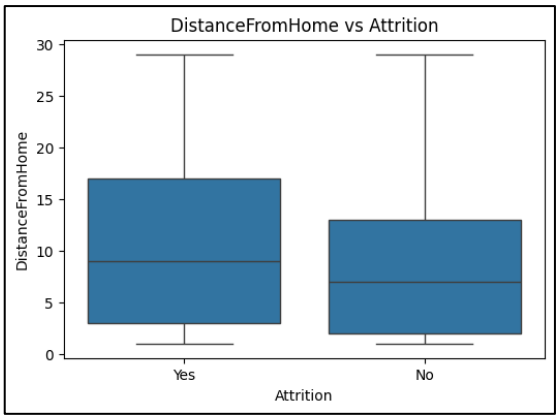


Most employees earn between 2,000 and 5,000 and very few employees earn above 15,000. When comparing attrition, employees who left have a monthly income of around 3,000 and those who stayed earn about 5,000.

This suggest that the income inequality can be associated to job role or promotion opportunities. Higher earners tend to remain within the company while those with lower salaries are prone to attrition.

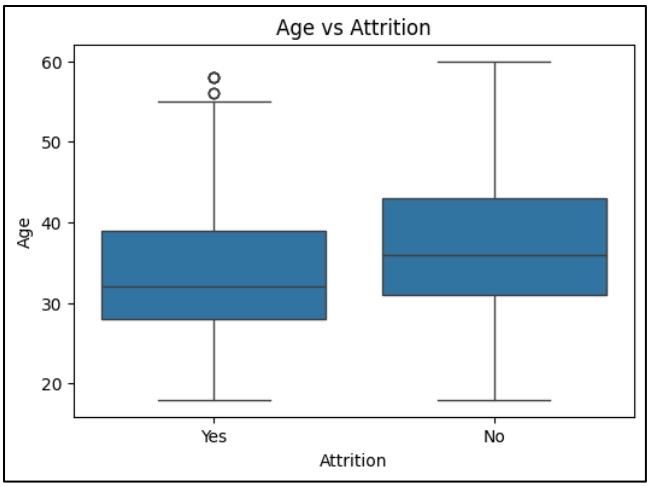
- **Distance from home**





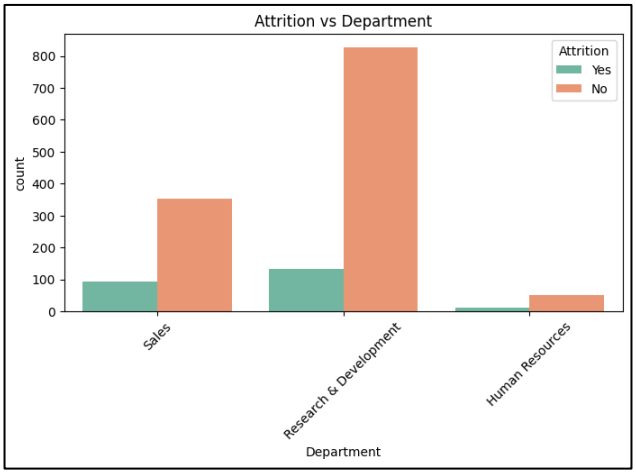
As the distance from home increases, the number of employees decreases. Proximity to the workplace appears to have an influence on attrition. Employees who live far away are more likely to leave the company.

- **Age**



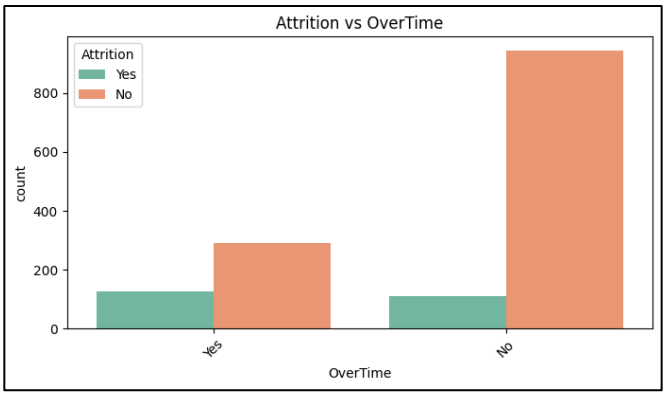
Younger employees (early 30s) are more likely to leave while older employees (late 30s) tend to remain at the company. This suggests that early career employees may be more motivated to explore new opportunities and change roles while older employees appear to be more stable.

- **Department**



The Sales and Research & Development departments show a noticeable attrition. The Human Resource department has a very low attrition but also fewer employees. Attrition might be higher in the Research & Development department due to workload and pressure. HR strategies could be tailored by department to address workload balancing and create incentives.

- **Overtime**



Employees who work overtime show a much higher rate of attrition compared to those who do not. This highlights a strong link between excessive workload and the decision to leave the company. Consistent overtime could lead to reduction in work-life balance, thus leading to attrition.

To conclude, attrition is strongly influenced by a combination of financial factors (monthly income and promotions), personal circumstances (distance from home and age), work pressures, and overtime. These observations highlight the importance of balanced compensation, career growth opportunities, and work-life balance policies to reduce turnover.

Data Cleaning and Preprocessing

The process of Data Cleaning and Preprocessing was key since the dataset contains raw information that the model could not understand. The following steps were carried out:

- Target Conversion

    - The target column Attrition was converted to binary values.

- Dropping irrelevant columns
    - Columns such as Employee Count, Employee Number and Over18 were dropped as they consist of either irrelevant information or constant values.

- Feature Identification
    - Features were separated into two categories (Numerical and Categorical).

- Encoding Categorical Features
    - The categorical features were encoded to numerical values using One Hot Encoding.

- Scaling Numerical Features
    - Continuous numerical features were standardised using the Standard Scaler.

- Splitting the data
    - The dataset was divided into training and testing sets (70% training and 30% testing).

Model Building and Observations

To predict employee attrition, several machine learning models were tested and compared. Each model was chosen to evaluate different approaches to classification. The Hyperparameter Tuning with GridSearchCV was used for Logistic Regression to better identify the potential leavers.

Model	Attrition	Accuracy	Precision	Recall	F1
Logistic Regression	Stay	0.76	0.93	0.77	0.84
	Leave	0.76	0.37	0.70	0.48
Decision Tree	Stay	0.79	0.87	0.89	0.88
	Leave	0.79	0.34	0.30	0.32

Random Forest	Stay	0.84	0.85	0.99	0.91
	Leave	0.84	0.55	0.08	0.15
Gradient Boost	Stay	0.85	0.87	0.96	0.92
	Leave	0.85	0.58	0.25	0.35
XGBoost	Stay	0.84	0.88	0.95	0.91
	Leave	0.84	0.53	0.30	0.38
Tuned Logistic Regression	Stay	0.77	0.94	0.78	0.85
	Leave	0.77	0.38	0.72	0.50

The results show that most models achieve good overall accuracy, ranging between 0.76 and 0.85. The Gradient Boost (0.85) and Random Forest (0.84) models performed best in terms of accuracy.

All models predict “Stay” very effectively, with precision and recall values generally above 0.85. However, predicting “Leave” is more difficult. The Tuned Logistic Regression performed best, with a recall of 0.72 and an F1 score of 0.50, meaning it identified more employees likely to leave compared to other models.

Gradient Boost and Random Forest are the strongest models in terms of overall accuracy. However, the Tuned Logistic Regression provides the best results for detecting employees at risk of leaving, which makes it more practical to improve employee retention.

Business Recommendations

Based on the above analysis, the Tuned Logistic Regression model is the most suitable tool for this business case. It can correctly identify about 7 out of 10 employees who are likely to leave, giving HR an early warning.

It can be concluded that younger employees and employees with lower monthly income are more likely to leave the company. Salary reviews and a good work life balance may improve the employee satisfaction and thus reducing attrition.