IBM HR Employee Attrition: EDA and Logistic Regression Prediction

Jane Wong

COG 105 | 12/14/2025

Introduction

Employee attrition is often a challenge for organizations and companies that want to keep company relations good, often leading to losing money, talent, and causing disruptions to workflow. For my project, I will be looking at whether attrition at IBM can be predicted using demographics, job-related factors, satisfaction levels, and performance at work. The main goal of my final project was to predict employee attrition using the variables just mentioned. Some other questions that I had in mind were:

1. Which types of employees were at higher risks of leaving IBM?
2. What are the key factors contributing to attrition?
3. What are some actionable insights to prevent attrition and for HR to know?

My analysis used exploratory data analysis (EDA) to find underlying patterns to employee data and applied a logistic regression model to create a predictive model for attrition. This project would have allowed IBM HR managers to target the different types of risk factors for those who leave, such as low job satisfaction or being too far from the workplace.

The Dataset

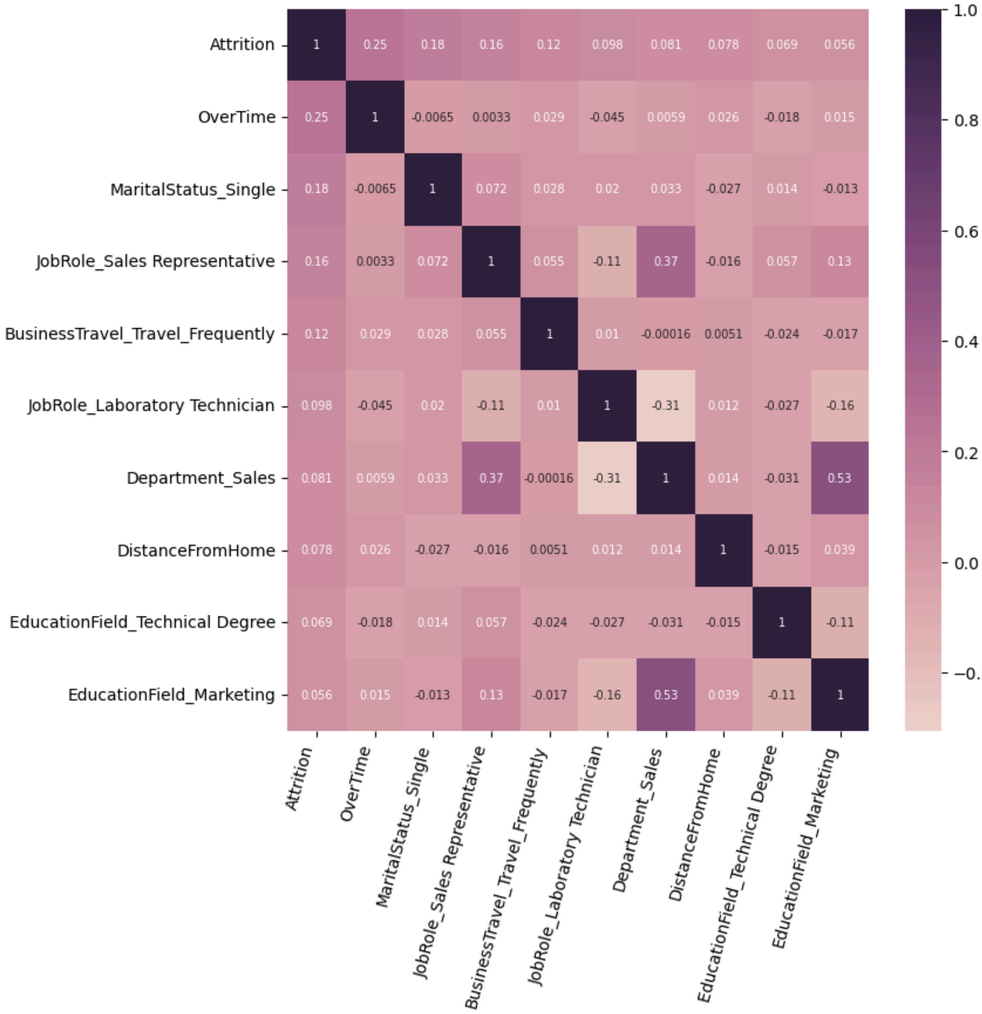
- The dataset I used contained 1470 employee records from IBM, each with 35 variables that captured demographic information, job, compensation, and satisfaction information.
    - Variables:
        - Target Variable: *Attrition* (Yes or No) – does the employee leave?
    - Quantitative Variables:
        - *Age, DailyRate, DistanceFromHome, Education, EnvironmentSatisfaction, HourlyRate, JobInvolvement, JobLevel, JobSatisfaction, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager*
    - Categorical Variables:
        - *BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, OverTime*
- Preprocessing:
    - First, I dropped variables that seemed irrelevant or did not add to attrition. These variables were: *EmployeeCount, EmployeeNumber, StandardHours, Over18.* Often, these variables were constant and did not add significantly to the findings.
    - Next, I did some binary encoding with: *Attrition, OverTime, Gender.*
    - One-hot encoding was done for categorical variables that had more than 2 variables: *BusinessTravel, Department, EducationField, JobRole, MaritalStatus.*

- Data Quality:
  - There were no missing values in the data.
  - There was some slight class imbalance with about 16% of employees leaving the company, which might affect how sensitive the model is while detecting attrition.

Results

Exploratory Data Analysis (EDA):

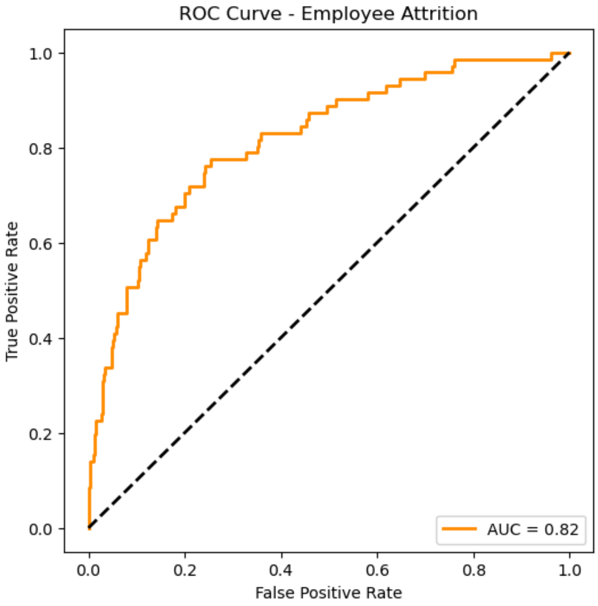
- Overall Attrition rate was 16.12% of employees who left IBM.
  - Numeric Variables: Boxplots showed that Age, Monthly Income, Distance From Home, and Number of Companies Worked were most associated with attrition.
  - Rating Variables: Job Satisfaction was the strongest predictor, with low satisfaction linking to attrition.
  - Categorical Variables: Employees that were in sales, sales representative roles, frequent travelers, or single had the highest rates of attrition.
  - Binary Values: Employees with higher rates of working overtime and males were slightly more likely to leave.
- Correlation analysis, using a heatmap that filtered the top correlations with attrition, showed that OverTime, MaritalStatus_Single, and BusinessTravel_Frequently were the most strongly correlated with attrition.



	Attrition	OverTime	MaritalStatus_Single	JobRole_Sales Representative	BusinessTravel_Travel_Frequently	JobRole_Laboratory Technician	Department_Sales	DistanceFromHome	EducationField_Technical Degree	EducationField_Marketing
Attrition	1	0.25	0.18	0.16	0.12	0.098	0.081	0.078	0.069	0.056
OverTime	0.25	1	-0.0065	0.0033	0.029	-0.045	0.0059	0.026	-0.018	0.015
MaritalStatus_Single	0.18	-0.0065	1	0.072	0.028	0.02	0.033	-0.027	0.014	-0.013
JobRole_Sales Representative	0.16	0.0033	0.072	1	0.055	-0.11	0.37	-0.016	0.057	0.13
BusinessTravel_Travel_Frequently	0.12	0.029	0.028	0.055	1	0.01	-0.00016	0.0051	-0.024	-0.017
JobRole_Laboratory Technician	0.098	-0.045	0.02	-0.11	0.01	1	-0.31	0.012	-0.027	-0.16
Department_Sales	0.081	0.0059	0.033	0.37	-0.00016	-0.31	1	0.014	-0.031	0.53
DistanceFromHome	0.078	0.026	-0.027	-0.016	0.0051	0.012	0.014	1	-0.015	0.039
EducationField_Technical Degree	0.069	-0.018	0.014	0.057	-0.024	-0.027	-0.031	-0.015	1	-0.11
EducationField_Marketing	0.056	0.015	-0.013	0.13	-0.017	-0.16	0.53	0.039	-0.11	1

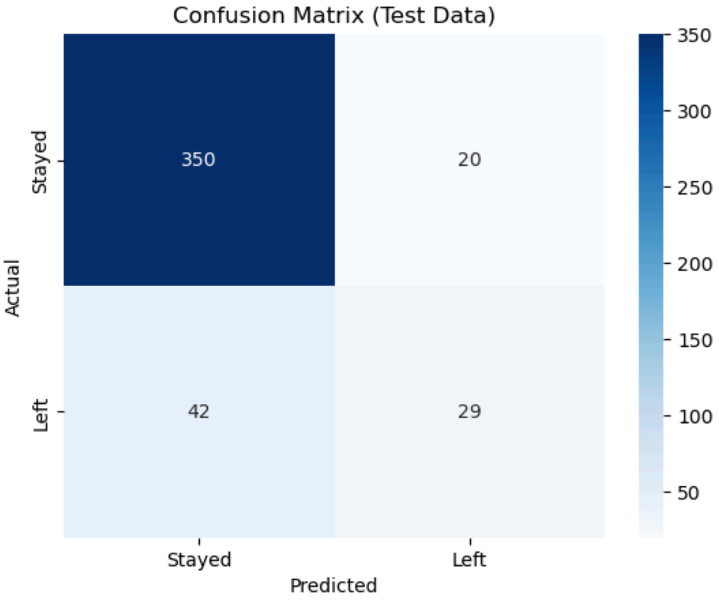
Logistic Regression Model:

- I did a Train/Test Split: 70% train, 30% test
- Before modeling, I also standardized the predictor variables by scaling.
    - Training accuracy: 89.4%
    - Test accuracy: 85.9%
    - ROC AUC: 0.82
        - This indicates good discriminate ability that the employee is able to distinguish between employees who stayed and those who left.



Confusion Matrix:

- The model performed well at predicting employees who stayed with the test data, but it had a difficult time at predicting employees who left. The recall for attrition was 0.41 (~29 of 71 employees that were correctly identified). This meant that other factors outside the scope of the data influenced employee attrition at IBM.



Regression Coefficients:

- Next, I extracted Beta coefficients with logregz.coef. Since predictor variables were standardized before modeling happened, the coefficients were comparable by magnitude or strength, which helped to see the significance of specific variables. The intercept was -2.69, which represents the baseline log-odds of attrition when standardized predictors are at a mean of 0. When converted to a probability, the baseline attrition is equal to 6.4%, which makes sense with the finding that attrition was relatively rare at a 16% rate.
- Overtime had the highest influence on attrition based on its coefficient of 0.925. Other notable variables were being single (0.716), frequent business travel (0.751) and role as a laboratory technician (0.777). Variables with a strong negative effect on attrition (people stayed at IBM) were job satisfaction (-0.526) and years with the current manager (-0.586), indicating employees with high satisfaction and stronger manager relationships were less likely to leave IBM. Variables with small influences on attrition but often accelerate leaving the company, were years since last promotion, percent salary hike, training times last year. These variables are still meaningful since they observe how careers progress at work and the investment that the company has on their employees.

Model Trust

- The model was reasonable with its prediction for overall attrition (86% accuracy, ROC AUC 0.82). However, there were some limitations where the model missed attrition for some cases because of circumstances outside of the dataset that might have affected the prediction. For future models, adding behavior data could improve the model by capturing more employee behavior to improve the recall score in predicting attrition.

Discussion + Reflection

The analysis showed that employee attrition could be predicted with workplace information and the satisfaction levels of employees. What drove attrition was overtime, low job satisfaction, being single, certain jobs, and frequent business travelling. These factors make perfect sense, especially with overtime and certain jobs (like Sales), leading to burnt out and high performance demands from employees. Some recommendations for IBM and its HR team would be to reduce how much employees work overtime. Enhancing employee engagement could improve employee satisfaction and adding programs could also support career development. However, it is also important to observe those at high risk, such as single status employees or frequent travelers, for early interventions.

Looking back on the model, it was able to successfully predict employees who stayed, with strong performance. However, attrition was more difficult to predict, likely due to other factors outside the dataset. To improve recall for attrition, future models could add behavioral and contextual or longitudinal variables like changes in manager feedback, workload trends or survey responses. This suggestion could track changes in employee behaviors over time, improving the model's prediction of attrition.

Overall, the logistic regression model and EDA provided advice and insights to employee attrition at IBM. Identifying these key factors would help HR target and reduce turnover rates, improve satisfaction levels at work, and retain talent and young hires for company growth.