Cracking the Attrition Code: Predicting IBM Employee Attrition

By : Christopher Le

Understanding the Problem

- Employee Attrition – loss of employees from an organization due to voluntary or involuntary reasons.
    - Rate at which employees leave a company's workforce and are not immediately replaced over a specific time.
- Employee Attrition impacts organizations in several ways:
    - Time-consuming and costly in terms of hiring and training new employees
    - Lose experienced employees
    - Impacts profits and productivity
- **Stakeholders:** HR department, supervisors, managers
    - Develop strategies to minimize employee attrition rates and motivate employees to stay in their jobs.
    - Identify factors and insights important for employee retention

Problem Statement

- How can companies such as IBM reduce their rate of employee attrition and better manage their workforce by identifying what types of employees are leaving, and predicting employee attrition?

- **Goal:** build a supervised classification model to predict attrition of employees for the company
- **Success Criteria:** reduce employee attrition by 10% by building an attrition prediction model that can predict whether employees will attrition with at least 75% accuracy.

Dataset

Dataset was retrieved from [Kaggle]

1470 rows x 35 columns

This is a fictional data set created by IBM data scientists and its main purpose was to demonstrate the IBM Watson Analytics tool for employee attrition.

Column Description

- **Demographic Information**: 'Age', 'DistanceFromHome', 'Education', 'EducationField', 'Gender', 'MaritalStatus', 'Over18'

- **Work Characteristics:** 'BusinessTravel', 'Department', 'JobInvolvement', 'JobLevel', 'JobRole', 'OverTime', 'PerformanceRating', 'StandardHours'

- **Salary-Related:** 'DailyRate', 'HourlyRate', 'MonthlyIncome', 'MonthlyRate', 'PercentSalaryHike', 'StockOptionLevel'

- **Satisfaction:** 'EnvironmentSatisfaction', 'JobSatisfaction', 'RelationshipSatisfaction', 'WorkLifeBalance'

- **Time-Related:** 'NumCompaniesWorked', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion', 'YearsWithCurrManager

- **Other:** 'EmployeeCount', 'EmployeeNumber'

01

Data Wrangling

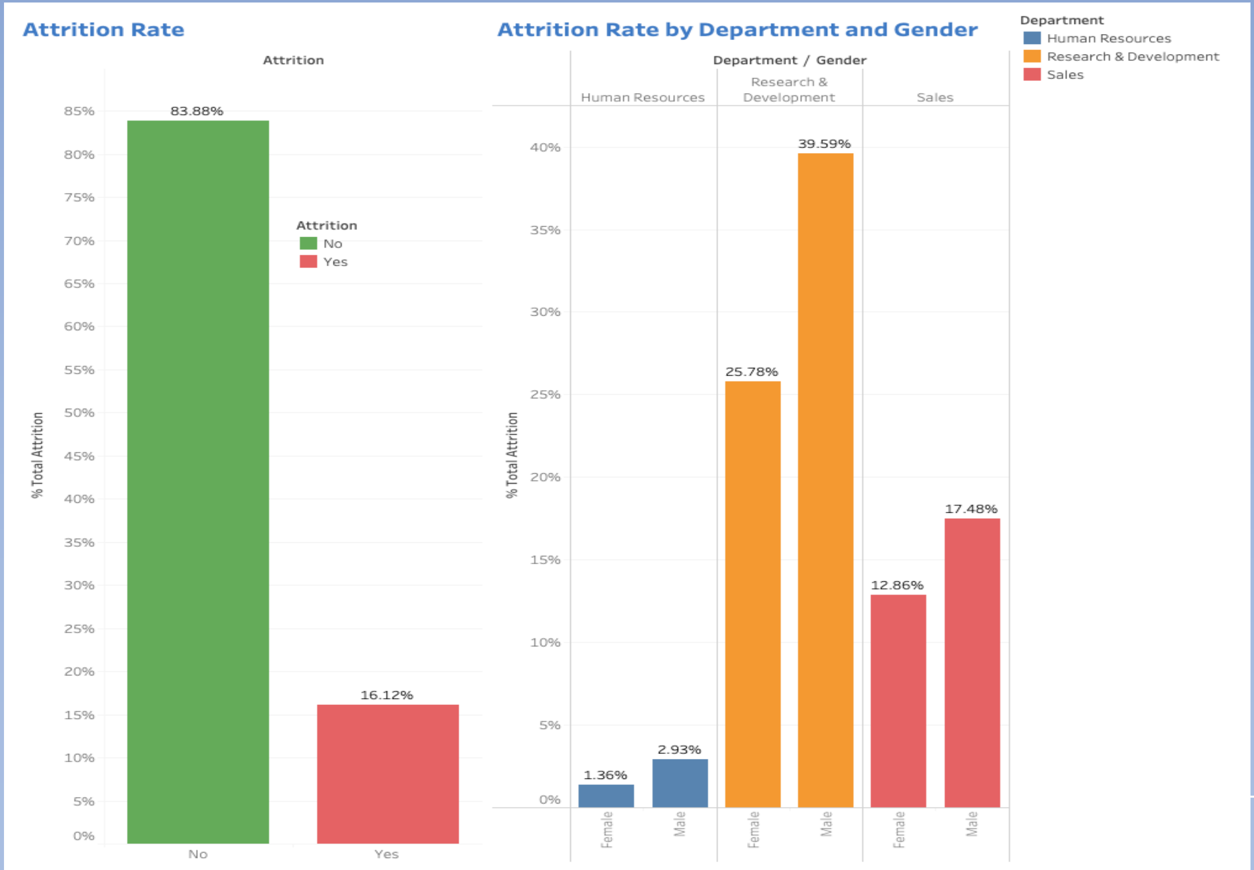
Data Wrangling

- **Initial Shape:** 1470 rows x 35 columns
- No duplicates or missing values
- Four columns were dropped:
    - 'EmployeeCount'
    - 'EmployeeNumber'
    - 'StandardHours'
    - 'Over18'
- Potential outliers:
    - 'YearsInCurrentRole', 'YearsSinceLastPromotion', and 'YearsWithCurrManager'
    - Outlier values were kept in data because they may help with model prediction
- **Final shape:** 1470 rows x 31 columns
- **Target Variable:** Attrition

02

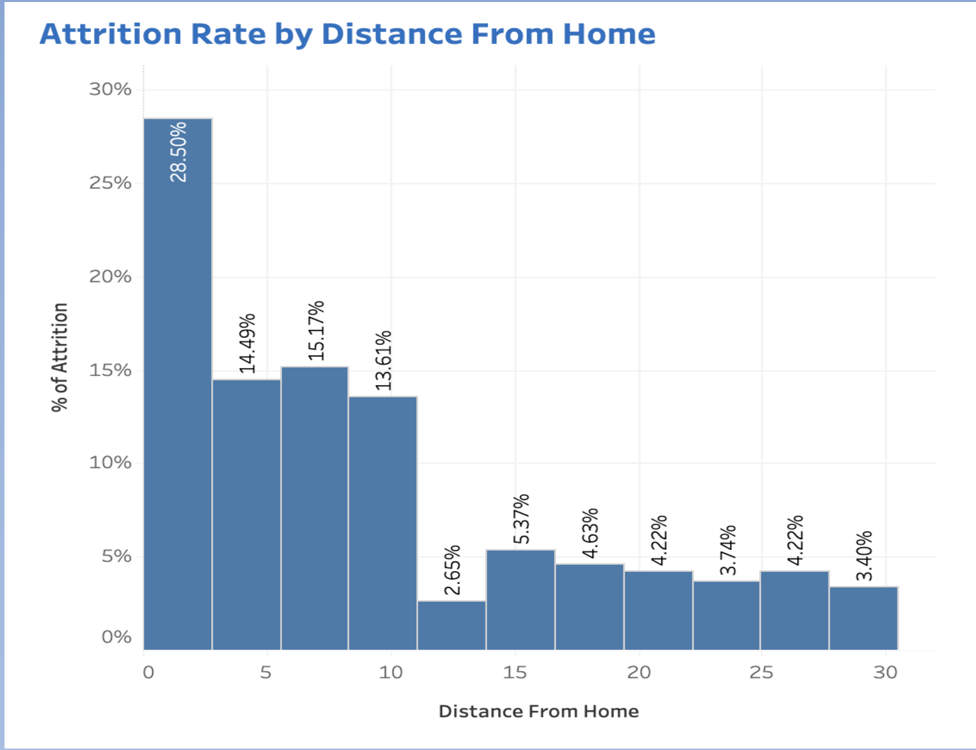
Exploratory Data Analysis

The Distribution of Attrition



- 16.12% of employees left the company overall
- Of those employees who left the company, **males** were more likely to attrition
- The **Research & Development** and **Sales** departments made up the majority of employees leaving the company
- The **Research & Development** department had the highest rate of attrition.

Attrition Rate and Distance From Home

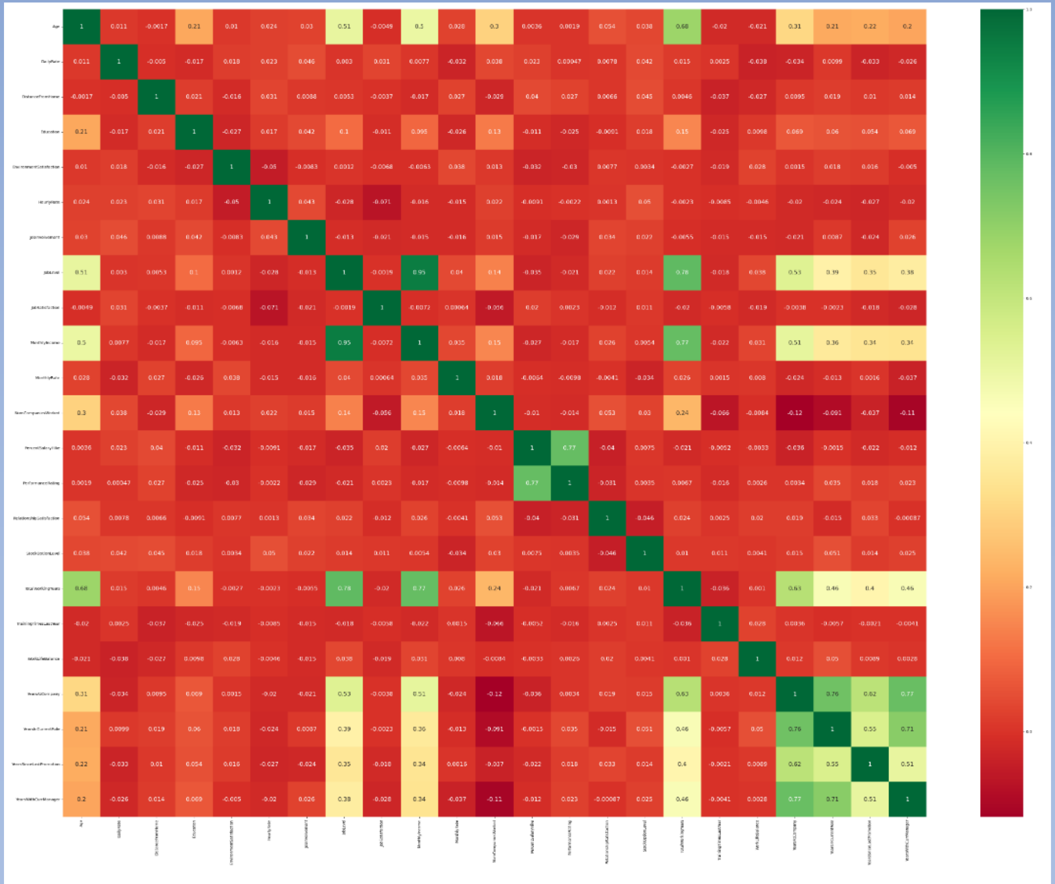


The attrition rate was higher for employees who live within 10 miles of the company.

Monthly Income and Years Worked

Average monthly income generally increased as total working years increased.

Correlation Heatmap



- Most of the features were poorly correlated with one another.

03

Modeling

Modeling Overview

Type: Supervised Learning

Binary Classification

0 – Employee stayed with company
1 – Employee attrition from company

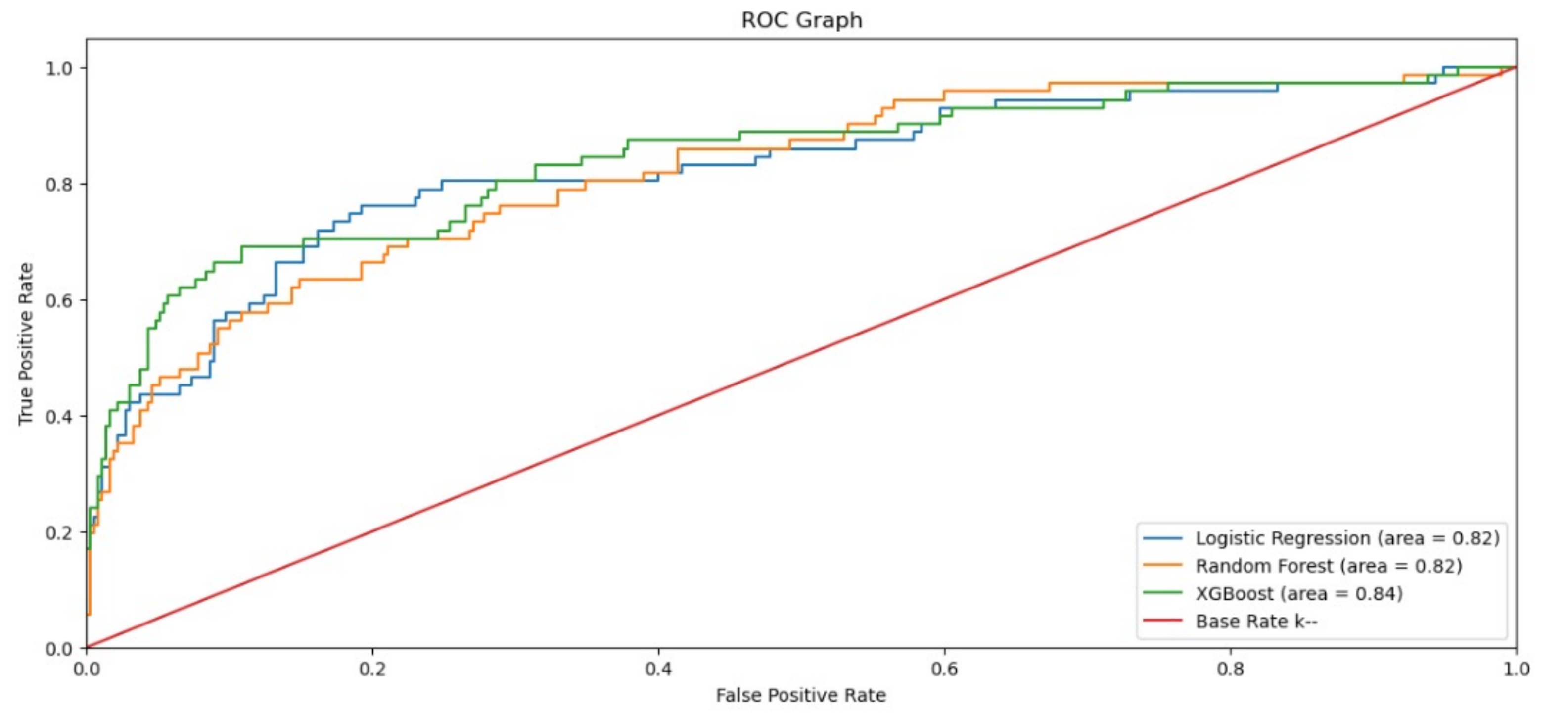
Models

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier

Model Comparison

Model	ROC-AUC	Accuracy	Weighted-Average Precision	Weighted-Average Recall	Weighted-Average F1-Score
Logistic Regression	0.8232	0.8776	0.86	0.88	0.86
Random Forest Classifier	0.8177	0.8571	0.86	0.86	0.81
XGBoost Classifier	0.8403	0.8798	0.88	0.88	0.85

Model Comparison



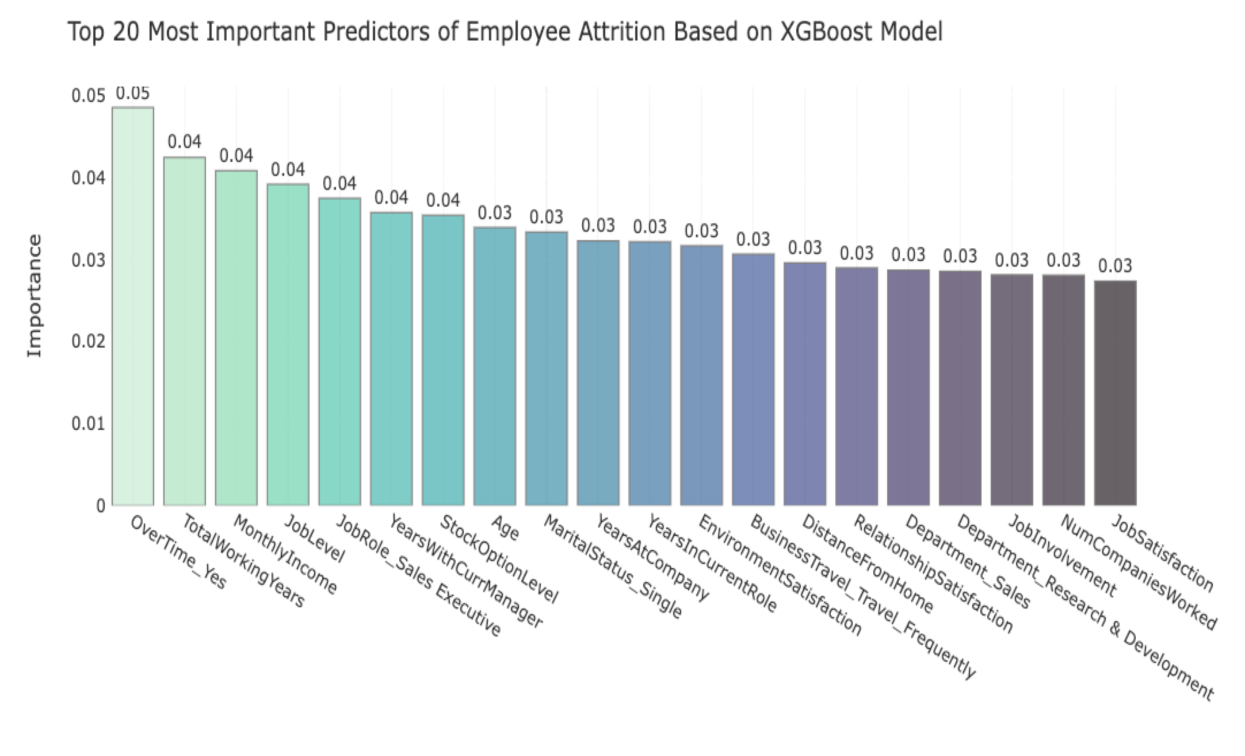
Graph visualizing ROC Curve of the different models.

Final Model Selection

XGBoost Classifier Model

- Highest AUC score: 84.03%
- Highest accuracy score: 87.98%
- Feature importance → feature selection for further analysis
- How model handled misclassifications:
  - 13% false positive
  - 12% false negatives

Analysis Results



1. OverTime_Yes
2. TotalWorkingYears
3. MonthlyIncome
4. JobLevel
5. JobRole_Sales Executive

04 Recommendations

Strategic Retention Plan

- **Address Overworking:** Implement policies and procedures to reduce excessive overtime and promote a healthy work-life balance for employees.
- **Evaluate Compensation:** Review the compensation packages offered to employees, with a particular focus on those in sales executive roles. Ensure that the compensation is competitive and provides a sufficient financial incentive for employees to remain with the company.
- **Foster Career Growth:** Offer professional development opportunities, such as training and mentorship programs, to help employees progress in their careers, and increase their job level.
- **Improve Job Satisfaction:** Conduct regular surveys to understand employees' perceptions of the company, their job roles, and the work environment. Use this information to identify areas where improvements can be made and take steps to address any concerns.
- **Offer Employee Benefits:** Evaluate the employee benefits offered by the company, including health insurance, retirement plans, and paid time off, and ensure that they are competitive with those offered by other companies in the industry.

Conclusions

Using the XGBoost Classifier Model, we were able to build a attrition prediction model that was able to predict which employees will become attrition with almost 88% accuracy.

Through this model, the HR department will be able to identify what types of employees are choosing to leave and determine which employees are at risk to leave next.

Future Work

- Obtain data on other factors that can contribute to employee attrition such as management style of supervisors/managers, or level of competition in the job market.

- More features engineering can be implemented to improve model performance by selecting only the most relevant features and removing less relevant ones so that the model can learn more efficiently and make better predictions.

- Resampling methods such as **SMOTE** (Synthetic Minority Over-sampling Technique) or **ADASYN** (Adaptive Synthetic Sampling) can be used to handle the class imbalance issue and improve model performance.

- Cross-validation methods such as k-fold can be used to get a better estimate of the model's true performance.

Thanks!

Do you have any questions?

lecm130@gmail.com
https://github.com/chrismle/Employee_Attrition

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**