Chris Le

Final Report: Cracking the Attrition Code: Predicting IBM Employee Attrition

1. Problem Statement

Employees are one of the most important assets of an organization as they are essential to providing the goods and services that the organization offers. The employee workforce is the foundation of a strong, successful, and long-running company.

Many employers face the challenge of employee attrition. Employee attrition refers to the loss of employees from an organization due to voluntary or involuntary reasons. It can be measured as the rate at which employees leave a company's workforce and are not immediately replaced over a specific period. Employee attrition can be a significant issue for organizations, as it can lead to disruptions to organizational activities, costly capital expenditures used to hire and select new candidates to fill vacancies, cost incurred to train new employees, time costs required to adjust to new changes, loss of knowledge due to losing experienced employees, and reduction in profits due to loss of productivity. Attrition is an inevitable part of any business, but it is important to minimize employee attrition, and determine potential causes so that effective countermeasures can be applied.

The purpose of this project is to analyze the data obtained from the Human Resources Department of IBM, determine the factors that influence employees to leave IBM, and build a prediction model to predict which employees will become attrition.

This project will help IBM's HR Department and employers by providing an in-depth analysis as to identify what types of employees are choosing to leave, and determining which employees are at risk to leave next. Predicting employee attrition before it happens will allow organization employers and managers to develop strategies to minimize employee attrition rates and motivate employees to stay in their jobs. This project can also help determine the underlying factors and insights important for employee retention.

2. Dataset

This dataset was retrieved from Kaggle, which sourced the data IBM HR. 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 Descriptions

- **Target Variable:** 'Attrition'
- **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'

3. Data Wrangling

The initial raw IBM HR dataset contained 1470 rows with 35 columns. The final shape of the IBM HR dataset is 1470 rows with 31 columns.

The dataset was checked for duplicates and missing values, but none were found. Four columns were dropped from the dataset. 'EmployeeCount' and 'EmployeeNumber' were dropped because they were irrelevant features to our analysis. 'StandardHours' and 'Over18' were features with only one unique value for all employees, so these features would not influence the analysis since everyone was over 18 years of age and worked 80 standard hours.

Initial review of the distribution of numerical features using histograms showed that many variables were tail-heavy. For example, ‘DistanceFromHome’, ‘MonthlyIncome’, and ‘YearsAtCompany’ were right-skewed. Boxplots of select features found that ‘YearsInCurrentRole', 'YearsSinceLastPromotion', and 'YearsWithCurrManager' had potential outliers. However, outlier values were kept in data because they may help with model prediction in terms of providing valuable information on whether time at company or long-term stability impacts a person's behavior and attitude towards attrition.

4. Exploratory Data Analysis (EDA)

The Distribution of Attrition

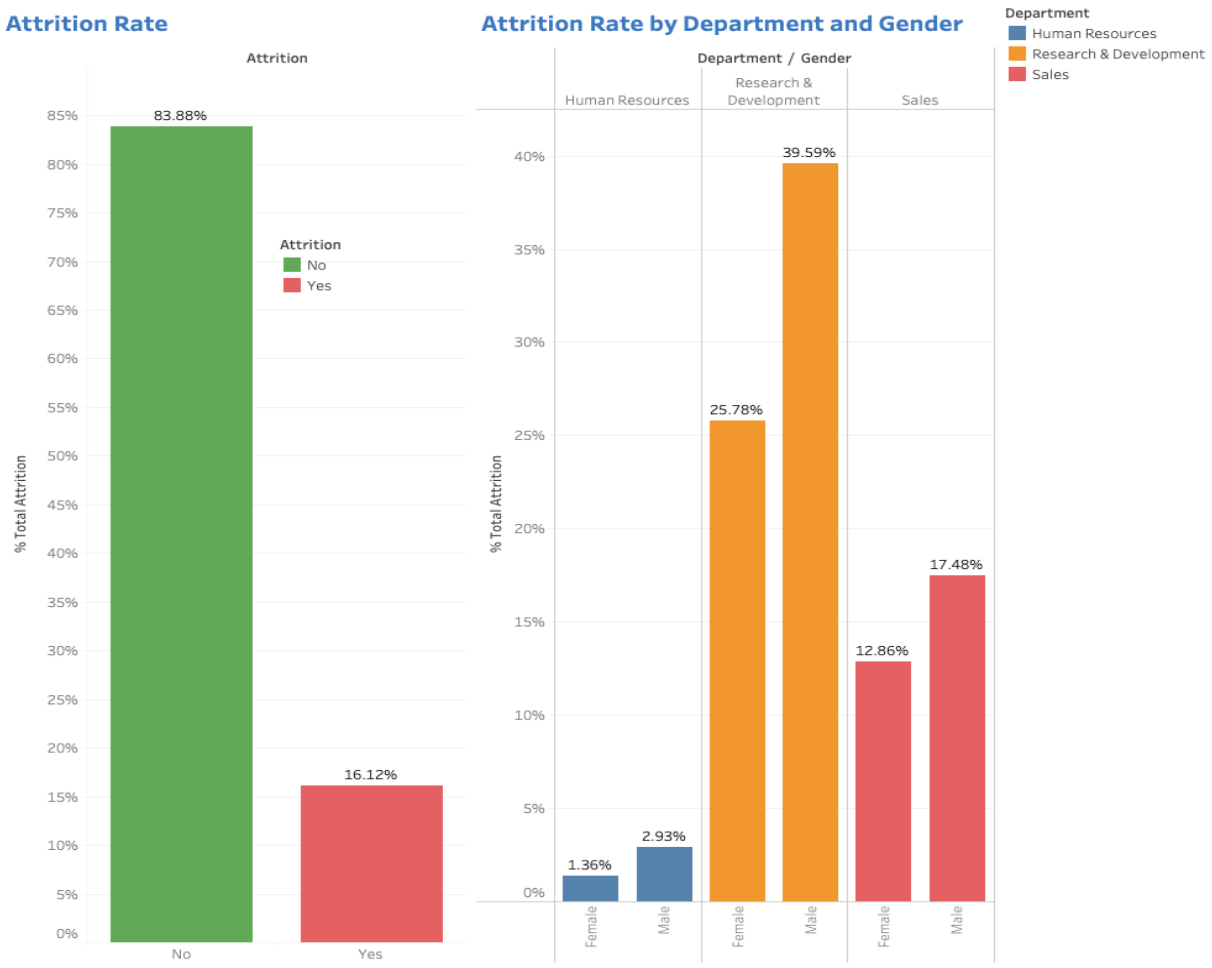

Figure 1: Bar chart showing overall Attrition Rate and Attrition Rate by Department and Gender

Figure 1 shows that the overall attrition rate was 16.12%, meaning that 237 out of 1470 employees left the company. This illustrates an imbalanced dataset with 1233 (83.88%) employees who did not attrition compared to 237 (16.12%) employees who did attrition from IBM. When breaking down the attrition by gender, we can see that regardless of department males were more likely to attrition than females. Most employees who left the company were from the Research & Development or Sales departments. The Research & Development had the highest rate of attrition overall, but this may be partly attributed to the department having more employees in total.

The Relationship between Attrition, and Working Conditions

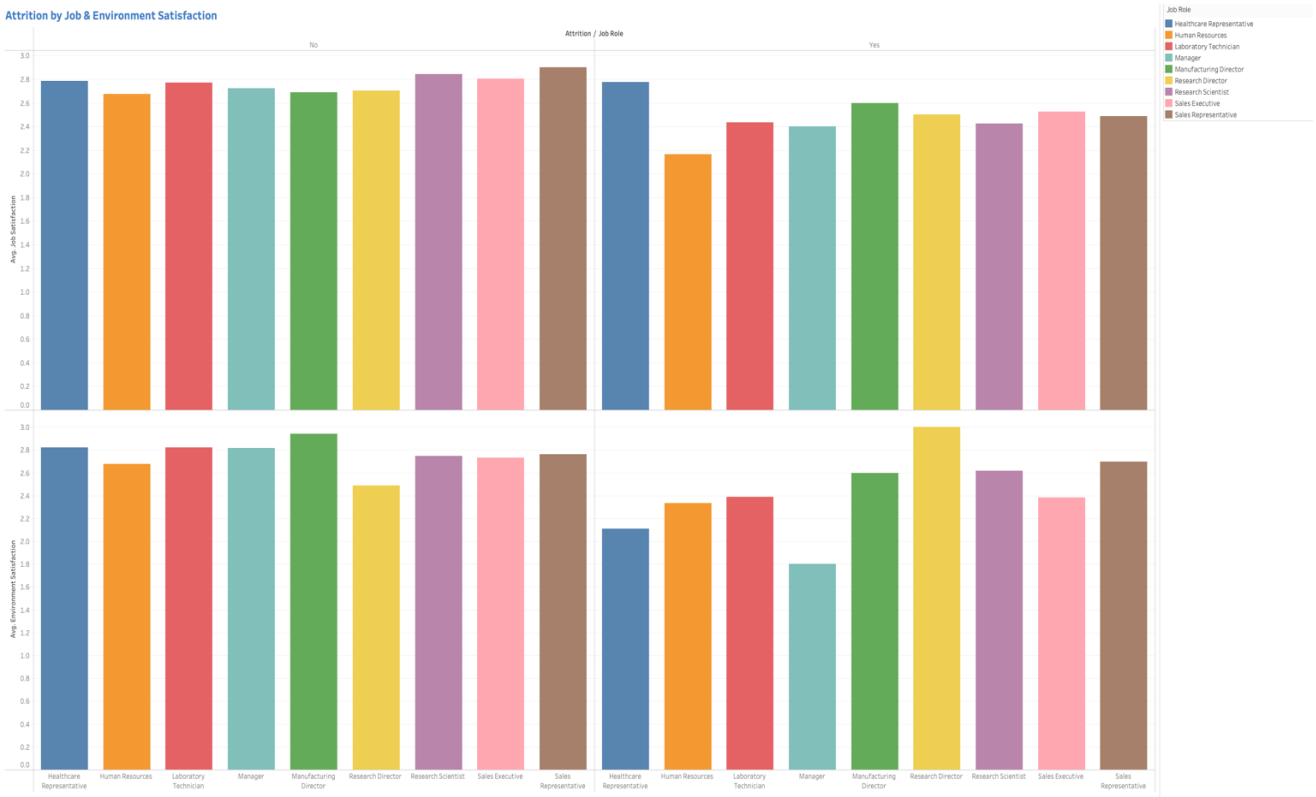

Figure 2: Attrition by Job Role, Job Satisfaction, and Environment Satisfaction

In Figure 2, we can see that employees who attrition from the company reported on average lower job satisfaction and environment satisfaction than those who stayed with the company. Furthermore, of those employees who left the company, human resources employees had the lowest average job satisfaction while managers scored lowest on average environment satisfaction.

The Relationship between Monthly Income and Years Worked

Figure 3: Line graph visualizing Monthly Income vs. Total Years Worked

We can see from the line graph in Figure 3 that total working years generally increased as the average monthly income increased. Compared to employees who stayed with the company, employees who attrition were more likely to experience significant drops in average monthly income during their later working years. This can be seen at around 19, 28, and 34 total working years.

Correlation Heatmap

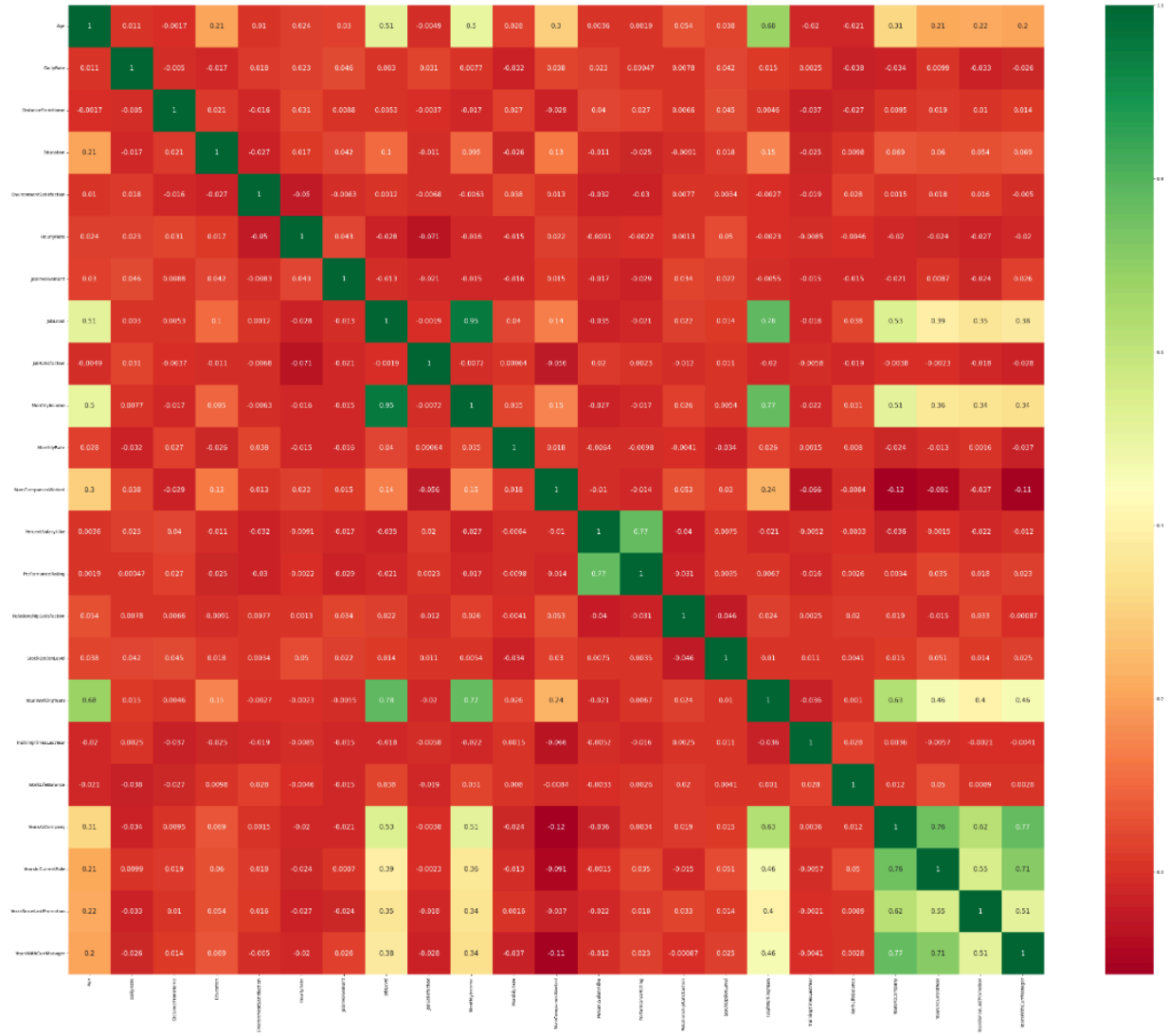

Figure 4: Heatmap representing the correlations between different features.

Analysis of feature correlation in Figure 4 using a heatmap showed that most of the features were poorly correlated with one another. However, there are a few highly correlated features. Specifically, there is a relatively strong positive correlation between: 'JobLevel' & 'MonthlyIncome', 'YearsAtCompany' & 'YearsWithCurrManager', and 'YearsInCurrentRole' & 'YearsWithCurrManager'. Furthermore, there is a negative correlation between 'YearsAtCompany' & 'NumCompaniesWorked', and 'JobSatisfaction' & 'HourlyRate'.

5. Modeling

The main goal of this project is to build a model that would predict whether an employee will attrition or stay with the company. With that in mind, supervised learning machine learning models would be ideal for this problem because it involves training a model based on a set of inputs (features) and corresponding outputs (labels, which in this case is attrition) and the goal is to learn a function that maps inputs to outputs accurately. Since our problem involves a categorical output with only two groups (Attrition – Yes, or Attrition – No), we will be building a binary classification machine learning model. To determine the optimal predictive model for employee attrition, we explored three different supervised learning models used for binary classification: Logistic Regression, Random Forest Classifier, and XGBoost Classifier.

One of the main metrics that we will use to evaluate the performance of my model is the ROC-AUC score (Receiver Operating Characteristic - Area Under the Curve), which provides a measure of how well the model can distinguish between employees who will leave and employees who will stay based on their predicted probabilities. With that said another important evaluation metric for binary classification problems is the accuracy score, which measures the proportion of correct predictions made by the model. Since there is moderate class imbalance in the dataset as employee attrition occurs less frequently than employees who stay, accuracy scores would not be ideal indicator as the main metric.

Other metrics that we will consider in the evaluation are the weighted-average precision, recall, and F1-scores. A high precision score would mean that our model makes few false positive predictions, and is good at identifying which employees will potentially attrition. This would be appropriate if the goal is to minimize instances of employees who are staying with the company, but are misclassified as leaving. On the other hand, a high recall score indicates that the classifier has a high true positive rate and thus is good at minimizing instances of employees who will attrition, but are misclassified as staying. In our case, it may be slightly more important to identify all employees who will leave, even if it means a higher number of false positive predictions, because missing employees who will attrition may result in lost resources, productivity, and replacement costs.

Logistic Regression

The first model that was implemented was the logistic regression model. The first iteration of the logistic regression model resulted in a training accuracy of 90.28% and testing accuracy of 86.62%.

After hyperparameter tuning the logistic regression model, the best hyperparameters when scoring based on ROC AUC score were: {'C': 0.1, 'solver': 'sag'}. This resulted in the testing accuracy increasing to 87.76%, but the ROC AUC score decreased from 82.57% to 82.32% which suggests the fine-tuned logistic regression model may have overfitted the data. The fine-turned logistic regression model weighted-average scores were: precision (86%), recall (88%), and f1-score (86%).

Random Forest Classifier

The random forest classifier model was the second algorithm applied. The first iteration of the random forest classifier model resulted in a training accuracy of 100% and a testing accuracy of 85.49%. While the training accuracy was excellent, the 14.5% difference in training and testing accuracy suggests that the model overfitted the data.

After hyperparameter tuning the random forest classifier model, the best hyperparameters were: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 3, 'max_depth': 10, 'bootstrap': True}. Similarly, to the logistic regression model, the finely tuned random forest classifier model had a higher testing accuracy score at 85.71%, but the ROC AUC score slightly declined from first iteration of 82.91% to 81.77%. The fine-tuned random forest model weighted-average scores were: precision (86%), recall (86%), and f1-score (81%). Based on the random forest model, the top 5 most important features were: OverTime_Yes, MonthlyIncome, Age, TotalWorkingYears, and DistanceFromHome.

XGBoost Classifier

The third model implemented was the XGBoost classifier model. The first iteration of the XGBoost classifier model resulted in a training accuracy of 100% and a testing accuracy of 85.49%.

After hyperparameter tuning the XGBoost classifier model, the best hyperparameters were: {'subsample': 0.6, 'n_estimators': 750, 'min_child_weight': 10, 'max_depth': 12, 'learning_rate': 0.02, 'gamma': 5, 'colsample_bytree': 0.6}. This resulted in 89.31% training accuracy and 87.98% testing accuracy, a 2.49% increase. The ROC AUC score also improved by 1.45% (from 85.58% to 84.03%). The fine-tuned XGBoost model weighted-average scores were: precision (88%), recall (88%), and f1-score (85%).

Final Model Selection

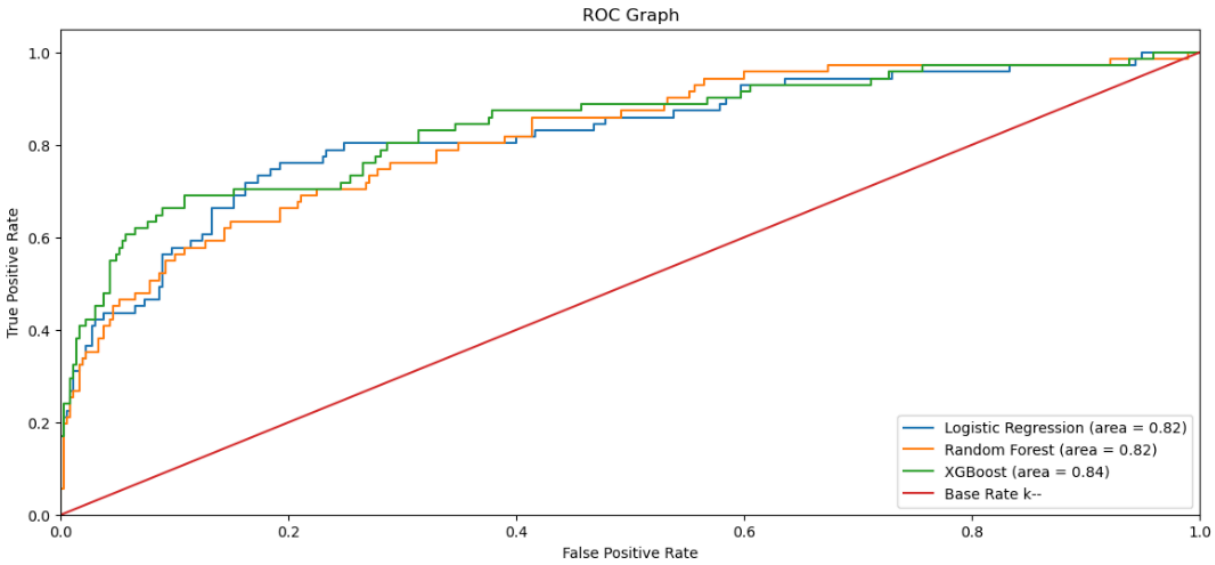

Figure 5: Graph Visualizing the ROC curve of the different models.

For the final model selection, we decided to use the XGBoost Classifier Model. As you can see from Figure 5, the XGBoost Classifier did the best on the main performance metric with an AUC score of 84.03%. It also had the highest accuracy score at 87.98% and weighted-average recall score at 88%. To see how the final model handled misclassifications, we can look at the confusion matrix. When predicting employee attrition, about 13% (3 of 24) of the time it predicted a false positive (predicting an employee will attrition, when in fact they are staying), while the false negatives (predicting an employee will stay, when they are actually leaving) were at about 12% (50 of 417).

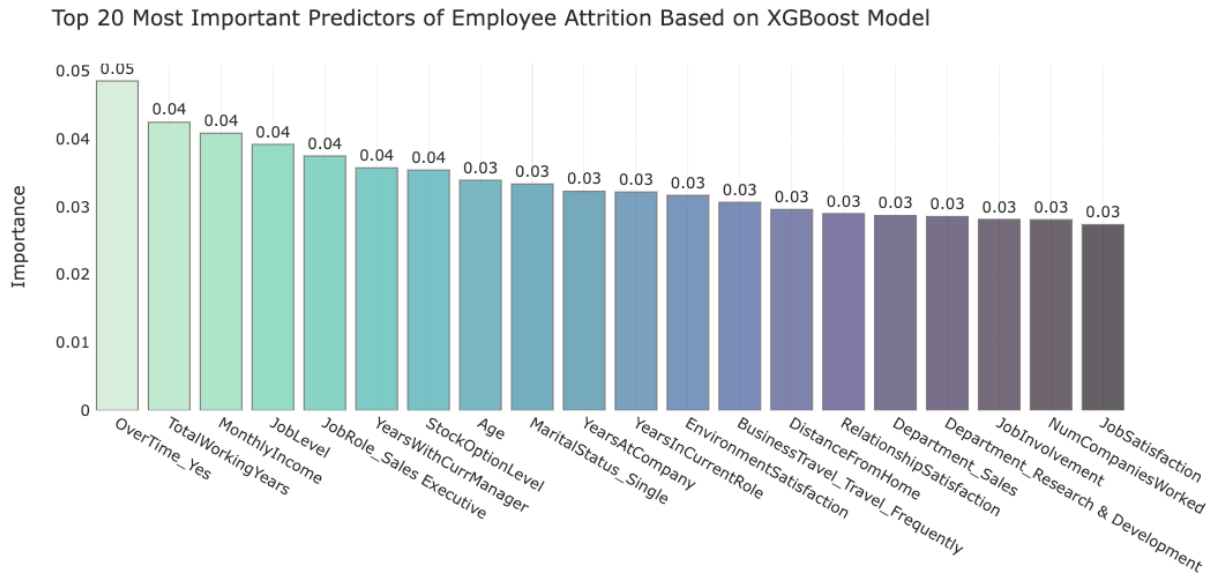

Figure 6: Chart Representing XGBoost Model's Top 20 Most Important Features

Based on the final model, the top 5 most important predictors/features of employee attrition are:

1. **OverTime_Yes**: employees who work overtime are most likely to leave the company.
2. **TotalWorkingYears**: employees with more total working years or work experience are less likely to attrition.
3. **MonthlyIncome**: employees with higher monthly income or high wages are less likely to leave the company.
4. **JobLevel**: employees with a higher job level are less likely to attrition from the company.
5. **JobRole_Sales Executive**: sales executive employees are most likely to attrition from company compared to employees with other job roles.

6. Employee Attrition Plan

We can develop a strategic retention plan based on the best indicators of employees potentially leaving the company:

- ***Address Overworking:*** Implement policies and procedures to reduce excessive overtime and promote a healthy work-life balance for employees. This could include setting clear limits on the number of hours employees are expected to work each week, providing paid time off and other incentives to encourage employees to take time off when needed, and promoting the use of flexible schedules to help employees manage their workloads.

- ***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. Consider implementing a merit-based pay structure that rewards high-performing employees.
- ***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. Provide clear pathways for employees to advance within the company and communicate expectations for promotions.
- ***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.

By addressing these key drivers of employee attrition, a company can improve employee satisfaction, increase retention, and reduce the cost and impact of high turnover rates.

7. Future Work

- Obtain more 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.