Mariah Maldonado
December 16, 2024

Exploring Models and Features for Predicting Employee Attrition at IBM

Introduction to Problem and Data

Problem Statement

Employee attrition is a critical challenge for organizations, impacting productivity, profitability, and operational efficiency. IBM has outlined several key concerns that arise from employee attrition, including the substantial costs associated with training new employees, the loss of experienced staff, and the resulting declines in productivity and profit margins (Swaminathan and Hagarty). Understanding the factors driving attrition is crucial not only for large corporations like IBM but also for smaller organizations that may need more resources or historical data to address such issues effectively.

Key business questions include identifying the factors contributing most to attrition and determining actionable employee retention strategies. The insights from such analyses can guide similar organizations in mitigating attrition and developing strategies to sustain their workforce. By addressing these questions and leveraging data-driven approaches, companies can establish frameworks to enhance employee retention and long-term organizational success.

Dataset Description

The data for this project is sourced from Kaggle in CSV format ("employee-attrition-aif360"). It provides comprehensive HR analytics data on employees who stay and leave. The dataset contains 1470 rows of individual IBM employees and 35 columns representing their different demographic and work-related features, which can help predict employee attrition.

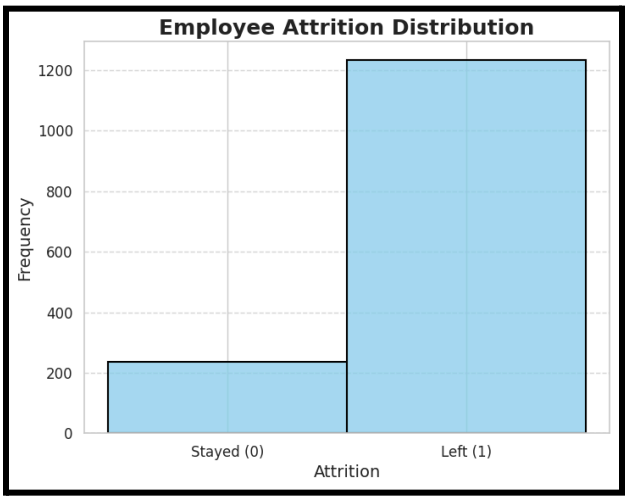
Data Pre-Processing and Preliminary Examination

With over 35 different features in the dataset, the data required light cleaning of features in the pre-processing stage. Too many features would likely lead to an overfit model. The simpler the model, the better the feasibility and research reproducibility. Therefore, this analysis will focus on work-related features with implications for where companies could intervene and reduce attrition risk rather than demographic features. Thus, the revised dataset will only include the following features: *Attrition, Business Travel, Department, Distance From Home, Environment Satisfaction, JobInvolvement, Job Level, Job Role, Job Satisfaction, Monthly Income, Over Time, Percent Salary Hike, Performance Rating, Stock Option Level, Total Working Years, Training Times Last Year, Work-Life Balance, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Current Manager.*

The data provided is complete with no null values, easing data processing.

Exploratory Data Analysis

The preliminary examination showed a clear imbalance between the number of employees who stay and those who leave. 237 of the 1470 employees listed left the company, meaning 84% of employees stayed and 16% percent left.



The table below shows descriptive statistics for all numerical variables used in this analysis. These statistics are essential later when considering the importance of features and actionable ideas for solving employee attrition.

	mean	median	std	min	max
DistanceFromHome	9.192517	7.000000	8.106864	1.000000	29.000000
EnvironmentSatisfaction	2.721769	3.000000	1.093082	1.000000	4.000000
JobInvolvement	2.729932	3.000000	0.711561	1.000000	4.000000
JobLevel	2.063946	2.000000	1.106940	1.000000	5.000000
JobSatisfaction	2.728571	3.000000	1.102846	1.000000	4.000000
MonthlyIncome	6502.931293	4919.000000	4707.956783	1009.000000	19999.000000
PercentSalaryHike	15.209524	14.000000	3.659938	11.000000	25.000000
PerformanceRating	3.153741	3.000000	0.360824	3.000000	4.000000
StockOptionLevel	0.793878	1.000000	0.852077	0.000000	3.000000
TotalWorkingYears	11.279592	10.000000	7.780782	0.000000	40.000000
TrainingTimesLastYear	2.799320	3.000000	1.289271	0.000000	6.000000
WorkLifeBalance	2.761224	3.000000	0.706476	1.000000	4.000000
YearsAtCompany	7.008163	5.000000	6.126525	0.000000	40.000000
YearsInCurrentRole	4.229252	3.000000	3.623137	0.000000	18.000000
YearsSinceLastPromotion	2.187755	1.000000	3.222430	0.000000	15.000000
YearsWithCurrManager	4.123129	3.000000	3.568136	0.000000	17.000000

The table below shows the descriptive statistics for all categorical variables used in this analysis, which is also crucial for feature importance and strategizing later on.

	count	unique	top	freq	Top	Top Frequency
BusinessTravel	1470	3	Travel_Rarely	1043	Travel_Rarely	1043
Department	1470	3	Research & Development	961	Research & Development	961
JobRole	1470	9	Sales Executive	326	Sales Executive	326
OverTime	1470	2	No	1054	No	1054

Modeling and Interpretations

Baseline Model and Performance Metrics

A baseline model is needed to evaluate the success of each of the three models by comparing its performance metrics, including accuracy, precision, recall, and F1 score. The baseline model is a probability-based prediction model where the predictions are made based on the observed class distribution. In this dataset, 84% of employees have Attrition = 0 (no attrition), and 16% have Attrition = 1 (yes attrition); thus, the model randomly assigns predictions following these probabilities. This baseline model provides a more balanced baseline than predicting only the majority class (predicting that each employee has Attrition = 0 and does not leave the company, as it is the highest probability). It reflects the actual class distribution in the data and helps evaluate whether more advanced models significantly improve prediction performance.

As mentioned, four performance metrics are used to compare the performance of the models ("Classification"):

- Accuracy: "the proportion of all classifications that were correct, whether positive or negative."
- Precision: "the proportion of all the model's positive classifications that are actually positive."
- Recall: "the proportion of all actual positives that were classified correctly as positives."
- F1: "the harmonic mean (a kind of average) of precision and recall."

The most important metric to consider in the employee attrition problem is recall, also known as the True Positive Rate, which measures how many of the employees who leave are correctly identified. Predicting true positives is essential because employee retention is costly and strategically important. Losing employees results in high costs (e.g., training, recruitment, knowledge gaps), and identifying who will leave allows companies to take proactive retention actions (e.g., bonuses, better work conditions). Additionally, because the data set is imbalanced, with 84% of employees staying and 16% leaving, recall is essential for the minority class (those who leave), as other metrics can be misleading. The risk of prioritizing true positives is the

possibility of falsely classifying employees as likely to leave (false positives), wasting resources on retention efforts for employees who would have stayed. However, missing an at-risk employee (false negative) may have higher consequences than incorrectly targeting an employee for retention (false positive), as incorrectly targeting an employee for retention would likely just improve that employee's quality of work, possibly increasing performance and likelihood of staying in the long term.

The baseline model's performance metrics are as follows:

- Accuracy: 0.75
- Precision: 0.14
- **Recall: 0.18**
- F1 Score: 0.16

With a recall score of 0.18, the baseline model correctly identifies only 18% of employees that are actually leaving the company. For the following models to add value, they must receive recall scores greater than 0.18.

First Model: Logistic Regression

The first model used to analyze employee attrition is a Logistic Regression model. It is a popular choice for analyzing binary outcomes (Yes/No) because it provides probabilities that an instance belongs to a particular class. This can be useful in predicting whether a specific individual will stay or leave based on both categorical and numerical features.

The Logistic Regression Model's performance metrics are as follows:

- Accuracy: 0.87
- Precision: 0.52
- **Recall: 0.31**
- F1 Score: 0.39

The Logistic Regression Model correctly identifies 31% of employees who are actually leaving the company. This is 13 percentage points higher than the baseline model, suggesting the model adds value.

An essential section of this analysis is the importance these models place on different features, as they provide actionable insights on what companies can do to reduce attrition risk. Feature importance for the Logistic Regression model is assessed by examining the absolute values of the model's coefficients. Larger absolute coefficients indicate a stronger relationship with the target variable (employee attrition). Positive coefficients suggest that higher values of the feature increase the likelihood of attrition, while negative coefficients imply the opposite. The most

important 5 features in predicting employee attrition according to the Logistic Regression Model are as follows:

1. **Years In Current Role** *(Importance: 0.62)*: This feature measures how long an employee has been in their current role. A higher value suggests greater job stability, and the model indicates that employees with fewer years in the role are more likely to leave.
2. **Job Role [Research Director]** *(Importance: 0.55)*: This feature represents whether an employee holds the position of Research Director. The importance value indicates that this role is associated with a higher likelihood of attrition.
3. **Stock Option Level** *(Importance: 0.53)*: This feature tracks the level of stock options granted to an employee. Higher stock options are associated with lower attrition, suggesting that employees with more financial incentives are less likely to leave.
4. **Years Since Last Promotion** *(Importance: 0.47)*: This feature measures the time since an employee was last promoted. A longer duration without a promotion is linked to higher attrition, as employees may feel stagnant or undervalued.
5. **Over Time [No]** *(Importance: 0.44)*: This feature indicates whether an employee works overtime. Employees who do not work overtime are likelier to leave, which may reflect lower engagement or work-life balance issues.

<u>Second Model: Random Forests</u>

The second model used to analyze employee attrition is the Random Forest model. This algorithm handles numerical and categorical data well, making it ideal for classification tasks like employee attrition. By combining multiple decision trees, it reduces overfitting, improves prediction stability, and naturally highlights important features.

The parameters used for this model were a maximum depth of 20 (controls the maximum number of splits each decision tree in the forest can make) and uses 50 trees. These parameters were chosen through a GridSearch to find the best parameters.

The Random Forests Model's performance metrics are as follows:

- Accuracy: 0.87
- Precision: 0.50
- **Recall: 0.10**
- F1 Score: 0.17

The Random Forests Model correctly identifies 10% of employees who are actually leaving the company. This is eight percentage points lower than the baseline model, suggesting the model provides less value than the baseline in correctly predicting true positives in employee attrition and should not be used to make assumptions about feature importance.

<u>Third Model: Naive Bayes</u>

The third model used to analyze employee attrition is the Naive Bayes Model, a probabilistic classifier that assumes feature independence and uses Bayes' Theorem to predict outcomes. It's efficient, handles both categorical and continuous data well, and performs well on imbalanced datasets, making it suitable for predicting employee attrition.

The Naive Bayes Model's performance metrics are as follows:

- Accuracy: 0.66
- Precision: 0.21
- **Recall: 0.56**
- F1 Score: 0.31

The Naive Bayes Model correctly identifies 56% of employees who are actually leaving the company. This is 38 percentage points higher than the baseline model, suggesting the model adds value.

This analysis determines feature importance in Gaussian Naive Bayes by combining two factors: the inverse of the feature's variance, which indicates its reliability (smaller variance suggests higher importance), and the mean difference across classes, which reflects its ability to distinguish between classes. The combined importance is the product of these two, capturing both the feature's discriminative power and stability. The most important 5 features in predicting employee attrition according to the Logistic Regression Model are as follows:

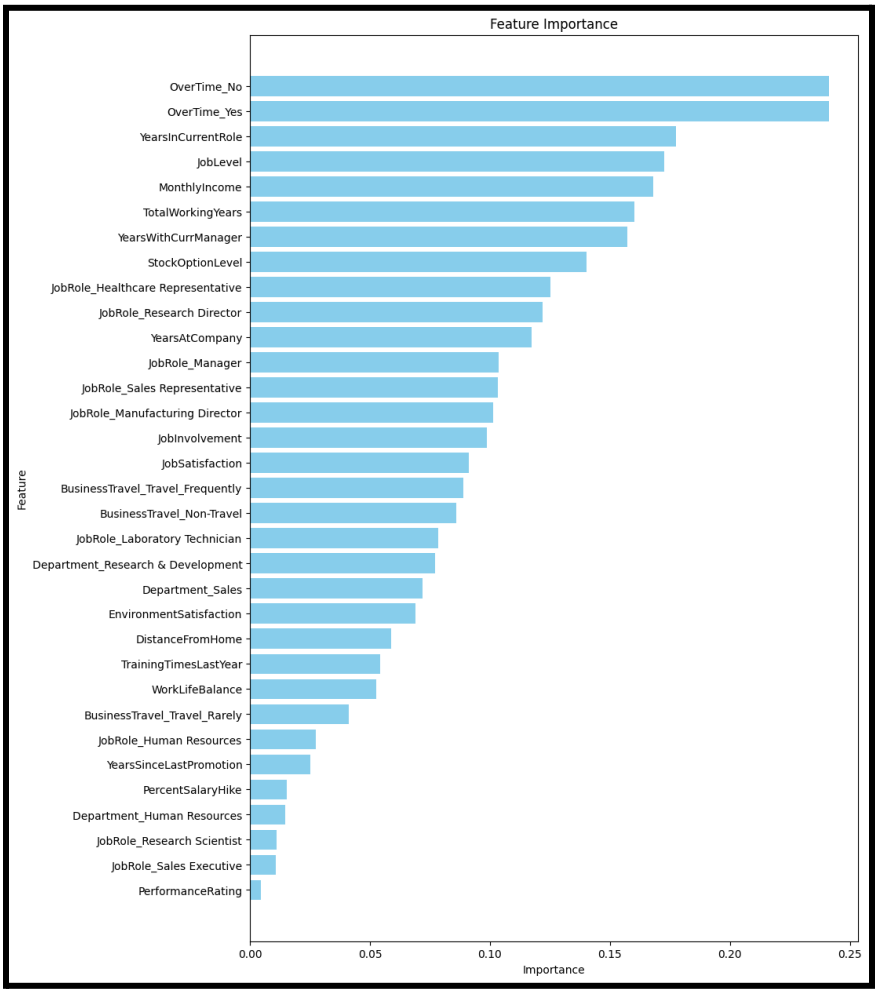
1. **Over Time [No]** *(Importance: 0.24)*: Employees who do not work overtime are among the most significant indicators of attrition. This suggests that employees without overtime might be less engaged or more likely to leave.
2. **Over Time [Yes]** *(Importance: 0.24)*: Similarly, employees who work overtime also have a high importance score, indicating that their commitment, as reflected by overtime work, is closely tied to their retention.
3. **Years In Current Role** *(Importance: 0.18)*: The number of years an employee has been in their current role is an important factor, with longer tenure potentially influencing their decision to stay or leave.
4. **Job Level** *(Importance: 0.17)*: Higher job levels are associated with a greater likelihood of staying, suggesting that more senior employees may have higher job satisfaction or better retention incentives.
5. **Monthly Income** *(Importance: 0.17)*: Income appears to play a role in retention, with higher monthly incomes potentially reducing attrition, as employees may feel more financially secure or valued.

Next Steps and Discussion

Summary of Findings

The analysis compared three models—Logistic Regression, Random Forests, and Naive Bayes—to predict employee attrition, focusing on improving recall to better identify at-risk employees. The Logistic Regression model showed a notable improvement over the baseline, achieving a recall of 0.31, correctly identifying 31% of employees who would leave. The Random Forest model exhibited a high accuracy of 0.87, but its recall of 0.10 was lower than the baseline, suggesting it was less effective in identifying attrition risk.

The Naive Bayes model outperformed both the Logistic Regression and Random Forest models, achieving the highest recall of 0.56, identifying 56% of at-risk employees. This significant improvement from the baseline (0.18 recall) underscores the Naive Bayes model's value in predicting employee attrition. Its ability to handle imbalanced data and incorporate both categorical and continuous features makes it particularly effective for this problem.



The implications of these findings are clear: while the Logistic Regression model provides valuable insights, the Naive Bayes model adds the most predictive value, particularly in identifying employees likely to leave. The top 5 features influencing the model's predictions were: Over Time (Yes/No), Years In Current Role, Job Level, Monthly Income, and Stock Option Level. Employees who do and do not work overtime, those with shorter tenure in their current role, those with lower job levels, and those with lower monthly incomes or stock option levels are most at risk of attrition. The plot to the right shows feature importance in descending order accompanying these findings, offering a visual representation of how these features contribute to the model's predictions. Based on these

factors, organizations should prioritize retention efforts for employees flagged by the Naive Bayes model.

Next Steps

Based on the top 5 features of the Naive Bayes Model, the following retention strategies could be useful in reducing employee attrition:

1. **Promote Work-Life Balance**: Both overtime (No and Yes) are linked to attrition. Offering flexibility for overtime workers and meaningful tasks for those not working overtime can help prevent burnout and disengagement.
2. **Offer Career Development**: Employees with fewer years in their current role are likelier to leave. Providing career growth opportunities and clear promotion paths could enhance engagement and retention.
3. **Implement Mentorship Programs**: Employees in lower job levels tend to leave more. Mentorship and structured career advancement programs can help improve retention at these levels.
4. **Provide Competitive Compensation**: Monthly income is a key factor in retention. Offering competitive salaries and performance-based incentives can help employees feel valued and secure.
5. **Increase Job Engagement**: Employees who do not work overtime (OverTime (No)) may be disengaged. Focusing on meaningful work and recognition can boost engagement and reduce attrition.

By focusing on these strategies, which are directly tied to the top features influencing attrition, companies can proactively create a supportive and motivating environment that helps retain valuable employees.

Improvements

In future analyses, incorporating demographic features such as age, gender, and education level could provide valuable insights into the factors influencing employee attrition. By including these variables, we could uncover additional patterns and refine the predictions made by the models, potentially improving their accuracy and interpretability. Furthermore, exploring feature engineering by examining interactions between work-related and demographic features may reveal deeper insights into attrition risk, offering a more nuanced understanding of the factors at play. Lastly, revisiting the tuning of hyperparameters, particularly for the Naive Bayes model, with an expanded feature set could further optimize performance, particularly in identifying employees at risk of leaving. These next steps help build a more comprehensive framework for predicting and addressing employee attrition, benefiting organizations in their retention strategies.

Works Cited

"Classification: Accuracy, recall, precision, and related metrics | Machine Learning." *Google for Developers*, 8 November 2024, https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall. Accessed 16 December 2024.

"employee-attrition-aif360." *Github*, 2020, https://github.com/IBM/employee-attrition-aif360. Accessed 16 December 2024.

Swaminathan, Saishruthi, and Rich Hagarty. "Data science process pipeline to solve employee attrition." *IBM Developer*, IBM, 28 March 2020, https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/. Accessed 16 December 2024.

"What Are Naïve Bayes Classifiers?" *IBM*, https://www.ibm.com/topics/naive-bayes. Accessed 16 December 2024.