Project Report 2- Predictive Modeling in Employee Attrition

IBM HR Analytics Employee Attrition Case Study

Dec 5th, 2025

Name	Email
Pouya Khazaeli Pour (Team Manager)	pkhazaelipour1@student.gsu.edu
Hiba Ahsan	hahsan1@student.gsu.edu
Mourya Bekkanur Dinesh	mbekkanurdinesh1@student.gsu.edu
Tran Le	tle167@student.gsu.edu
Sumanth Pulagante	spulagante1@student.gsu.edu

Statement of Academic Honesty

The following code represents our own work. We have neither received nor given inappropriate assistance. We have not copied or modified code from any source other than the course webpage or the course textbook. We recognize that any unauthorized assistance or plagiarism will be handled in accordance with Georgia State University's Academic Honesty Policy and the policies of this course. We recognize that our work is based on an assignment created by the Institute for Insight at Georgia State University. Any publishing or posting of source code for this project is strictly prohibited unless you have written consent from the Institute for Insight at Georgia State University.

Motivation and Problem Statement

Employee attrition carries both operational and financial consequences for organizations. When an employee resigns, the company loses not only one person but also their experience, department familiarity, and productivity. The process of hiring, onboarding, and training a replacement can take months, during which team performance, morale, and organizational continuity are affected. Research indicates that turnover costs can range from 50% to 200% of the employee's annual salary, depending on role complexity and training requirements. Beyond cost, frequent turnover can signal deeper issues such as cultural misalignment, unmanaged workload, or lack of career development opportunities, making retention a core strategic priority rather than a routine HR function.

Organizations should be willing to invest in solving attrition because preventing even a small number of resignations can generate significant financial return. A predictive model that identifies high-risk employees early enables targeted, cost-effective retention actions such as performance conversations, compensation adjustments, role realignment, or leadership support. These interventions are considerably less costly than losing and replacing high-value employees. Additionally, predictive analytics empowers leaders to shift from reactive crisis management to strategic workforce planning, improving employee experience, engagement, and trust. In competitive labor markets, businesses with lower turnover gain a measurable

advantage, reduced disruption, higher productivity, and greater long-term organizational stability.

The problem is that organizations lack the ability to accurately predict which employees are most likely to leave and what factors contribute most strongly to their departure. Without reliable predictive insight, companies remain reactive, resulting in preventable turnover, unnecessary hiring costs, and operational disruption. This project aims to build a machine-learning-based attrition prediction model that identifies at-risk employees, highlights key drivers of resignation, and supports proactive retention strategies.

Solution and High- Level Approach

Our solution is to build a predictive machine learning model designed to convert raw HR data into actionable attrition risk insights. Rather than relying on intuition or exit interviews (which happen too late), this approach utilizes historical data to proactively identify patterns associated with resignation. The solution involves cleaning demographic and job data, transforming it for machine analysis, and training supervised classification algorithms to predict a binary outcome whether an employee will leave or stay.

This data-driven approach is the optimal solution because it eliminates subjectivity and scales effectively. Human analysis often misses non-linear relationships between variables—such as how "Distance from Home" interacts with "Job Role"—whereas machine learning models excel at detecting these complex patterns. Furthermore, by implementing specific techniques to handle "imbalanced data" (the fact that fewer people leave than stay), our approach ensures the system focuses specifically on identifying the minority "leaver" group, rather than just achieving high accuracy by guessing everyone will stay.

How It Works, Architecture and Methodology

Our approach is organized as a pipeline that follows four main stages. First, Data Ingestion and Understanding involve loading the IBM HR Analytics dataset (1,470 employees) and performing descriptive analysis to understand feature distributions. Second, Data Preparation & Feature Engineering focuses on cleaning the data, this includes removing constant columns (like Employee Count), encoding categorical variables (changing Gender to numeric values), and scaling numerical features using a Standard Scaler to ensure features like Monthly Income do not disproportionately dominate the model.

The process begins with data ingestion and initial exploration, where the IBM HR Attrition dataset is loaded and examined to understand its structure, variable types, and patterns. The dataset consists of both numerical and categorical variables such as Age, Monthly Income, Years at Company, Job Role, Business Travel, and Over Time, along with a binary target field indicating whether an employee stayed or left the organization. During this stage, descriptive analytics and visualization techniques help uncover meaningful behavioral patterns for example, employees who eventually left tended to be younger, earned less, reported lower satisfaction

levels, and had shorter tenure with their current manager. This early understanding provides an analytical foundation and reinforces the relevance of certain data relationships before modeling begins. Following initial exploration, the dataset undergoes data quality assessment and preprocessing. While no missing values were detected in the source dataset, a formal missing-data strategy is defined to maintain robustness in case of future deployment or extended datasets. Numerical variables would be imputed using the mean or median depending on their distribution, while categorical variables would be handled using the mode. Notably, missing values would never be imputed for the target field, as doing so could artificially bias the model; instead, such rows would be removed. Outliers discovered in variables such as income and tenure are intentionally retained, since these represent legitimate employee groups often more senior or specialized not data errors. Removing them would risk disproportionately altering distributional characteristics and weakening the model's ability to learn realistic workforce patterns.

In parallel, categorical and numerical features are transformed to ensure compatibility with machine learning algorithms. Binary variables such as Gender and Over Time are label-encoded, ordinal features such as Job Satisfaction and Work Life Balance are encoded using ranked numeric scales, and standardization is applied to all continuous variables using a **Standard Scaler**. This ensures that variables with large numeric ranges, especially Monthly Income, do not dominate the model during training. The dataset is also affected by class imbalance, with only 16.1% of employees categorized as attrition cases. To address this, a stratified train- test split preserves proportional class representation during model evaluation, and class weights are applied during training to ensure the model treats attrition cases as equally important to retained employees rather than defaulting to the dominant class. Once the data has been processed, model development begins through a supervised binary classification framework. Since only 16.1% of employees in the dataset are attritions, a standard model would be biased toward the majority. To counter this, we utilized stratify and weighted balance techniques to effectively minimize the influence of the imbalance dataset. After, multiple algorithms are trained including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost to compare performance across linear, rule-based, and ensemble modeling philosophies. Each model is trained on 80% of the data, with the remaining 20% used for independent evaluation. The outputs include both predicted class labels and probability scores, enabling the construction of confusion matrices to analyze how models perform across true positives, false positives, false negatives, and true negatives. This perspective is critical, as attrition prediction is not purely about accuracy, but about ensuring the model can correctly identify employees at risk of leaving. In the improvement stage, we then used Hyperparameter and SMOTE techniques to see if the outcome scores can get better. We tested the techniques through all 5 different models to see if result change. Still Linear regression is the best model after all since there was consistent high accuracy, F-1 and AUC scores overall.

Evaluation, Outcomes, and Discussion

We evaluated the models using a suite of metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. While Recall (catching leavers) is critical, we also analyzed F1-Score (the harmonic mean of Precision and Recall) to ensure the model wasn't simply flagging everyone as a risk.

Model Strategy	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Hyperparameter Tuning (LogReg)	0.5986	0.2680	0.8723	0.4100	0.8047
Baseline LogReg	0.7449	0.3627	0.7872	0.4966	0.8350
SMOTE + Tuning (LogReg)	**0.7687**	**0.3871**	**0.7660**	**0.5143**	**0.8245**
SVM	0.8129	0.4412	0.6383	0.5217	0.8080
Decision Tree	0.7687	0.3478	0.5106	0.4138	0.7020
XGBoost	0.8537	0.5714	0.3404	0.4267	0.7800
Random Forest	0.8401	0.5000	0.0851	0.1455	0.7690

Table 1: Consolidated Performance Metrics

After analyzing the trade-offs, Logistic Regression with SMOTE + Hyperparameter Tuning is the best performing model. While the standard "Hyperparameter Tuning" model achieved the highest raw Recall (87.23%), it suffered from very low Precision (26.8%) and poor Accuracy (59.8%). This means it generated a massive number of "False Positives," effectively spamming HR with alerts for employees who were not actually at risk. This would waste resources and potentially damage employee morale through unnecessary interventions.

The **SMOTE + Tuning** strategy, however, offers the optimal business balance. It achieves a robust **Recall of 76.60%**, meaning it still identifies nearly 80% of all potential leavers. Crucially, it does so with significantly higher **Precision (38.71%)** and **Accuracy (76.87%)** than the tuning-only approach. It also achieved the highest **F1-Score (0.5143)** among the logistic regression variants. We also revalidated all our five models with SMOTE to address imbalance and still the logistic regression gave us the highest recall and overall better model due to having consistent high accuracy, F-1 and AUC scores.

A machine learning attrition model may stop performing well when real-world conditions change and the patterns it originally learned no longer reflect employee behavior for example, economic shifts (like COVID-19 or inflation), new organizational policies,

changes in workforce demographics, or updated job structures such as remote/ hybrid work. To keep the model relevant, new descriptive features could include variables like remote-work status, flexibility rating, employee engagement survey scores, promotion wait time, inflation-adjusted salary competitiveness, internal mobility activity, and manager turnover. These additional features help capture new drivers of employee behavior that older data may not have represented.

The SMOTE + Tuning model is the superior choice because it is operationally efficient. It captures most at-risk employees without overwhelming the organization with false alarms. It demonstrates that addressing class imbalance via synthetic sampling (SMOTE) provides a more stable and reliable predictive tool than simple weight adjustments alone. The analysis confirms that "Over Time," "Total Working Years," and "JobLevel" are strong indicators of turnover, and this model effectively utilizes those signals to provide actionable warnings.