Employee Attrition Prediction and Financial Impact Analysis

K. Pranav Suhas Reddy
G_CSE093
https://github.com/kpsr01/ml-project

May 2025

Abstract

This report presents an end-to-end solution for identifying employees at risk of attrition and forecasting the financial implications of potential turnovers. Leveraging the IBM HR Analytics dataset, I developed robust classification pipelines to flag high-risk cases and regression ensembles to project future salary commitments. My methodology integrates advanced preprocessing (SMOTE, scaling, encoding), hyperparameter optimization, and ensemble strategies. Key achievements include a tuned Voting Classifier with 86.73% accuracy and a Voting Regressor achieving an R^2 of 1.0000. The estimated expected financial exposure due to potential attrition is **$3,261,024.27** overall, with **$1,127,966.68** attributable to employees anticipated to stay.

Contents

Executive Summary

- **Objective**: Predict employee attrition and quantify the associated salary-related financial impact.
- **Data**: IBM HR Analytics dataset featuring 35 HR attributes; the target variable 'Attrition' was binarized.
- **Key Results**:
  - Best Classification – Tuned Voting Classifier: Accuracy 0.8673, ROC AUC 0.7940, Weighted F1-score 0.8709.
  - Best Regression – Tuned Voting Regressor: R^2 1.0000, RMSE 73.02, MAPE 0.0120.
- **Financial Impact**:
  - Total expected financial exposure from potential attrition: **$3,261,024.27**.
  - Expected financial exposure from the subset of employees predicted to stay: **$1,127,966.68**.
- **Recommendation**: **Prioritize retention programs** for high-value, high-risk employee groups and **utilize insights for more accurate budget forecasting** regarding salary liabilities.

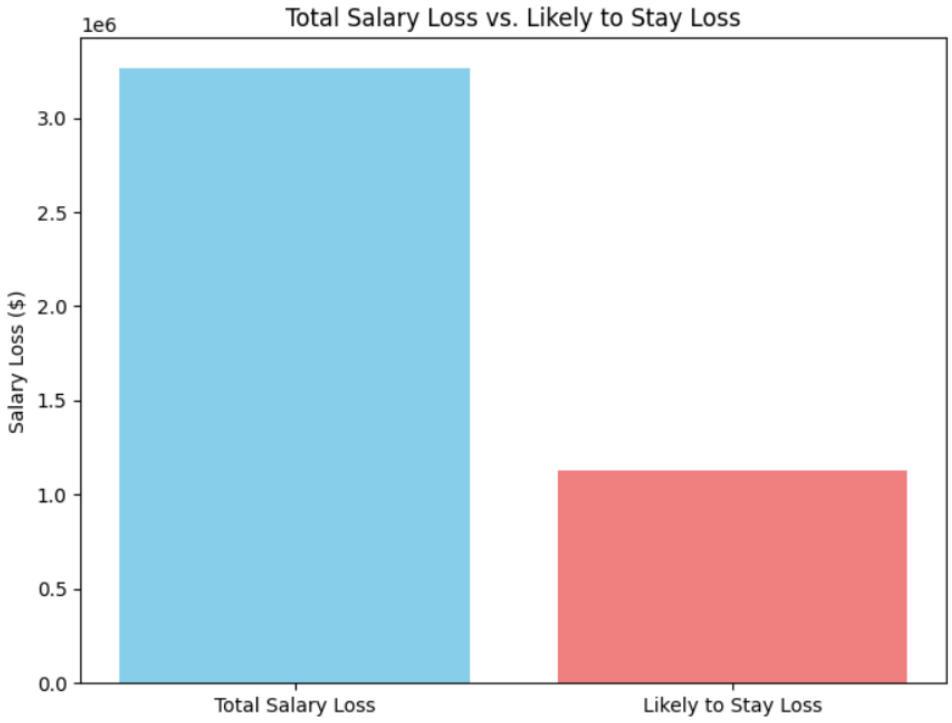

Figure 1: Total Salary Loss vs Likely to stay Loss

1 Introduction

Employee attrition significantly affects organizational performance and incurs substantial costs, including recruitment expenses, onboarding, and productivity losses. This report advances traditional HR analytics by coupling attrition risk modeling with salary projection to produce **actionable financial insights**. This dual-model framework empowers stakeholders to **identify critical retention targets** and **forecast the budgetary impacts** of employee turnover.

2 Data and Exploratory Analysis

2.1 Dataset Description

The analysis utilized the IBM HR Analytics dataset, which contains 1470 employee records across 35 attributes, encompassing demographics, job role details, and performance metrics. The primary target variable, 'Attrition', was converted to a binary format (Yes=1, No=0). Administrative fields such as 'EmployeeCount', 'EmployeeNumber', 'Over18', and 'StandardHours' were excluded from the analysis as they offered no predictive value.

2.2 Exploratory Findings

Initial summary statistics revealed a class imbalance, with approximately **16%** of employees having attrited. Exploratory data analysis (EDA) identified 'MonthlyIncome', 'JobLevel', and 'YearsAtCompany' as potentially strong predictors of attrition. Various visualizations, including correlation heatmaps (as exemplified in Figure 2), attrition distributions by department, and job satisfaction distributions, were generated to understand feature relationships and guide feature selection.

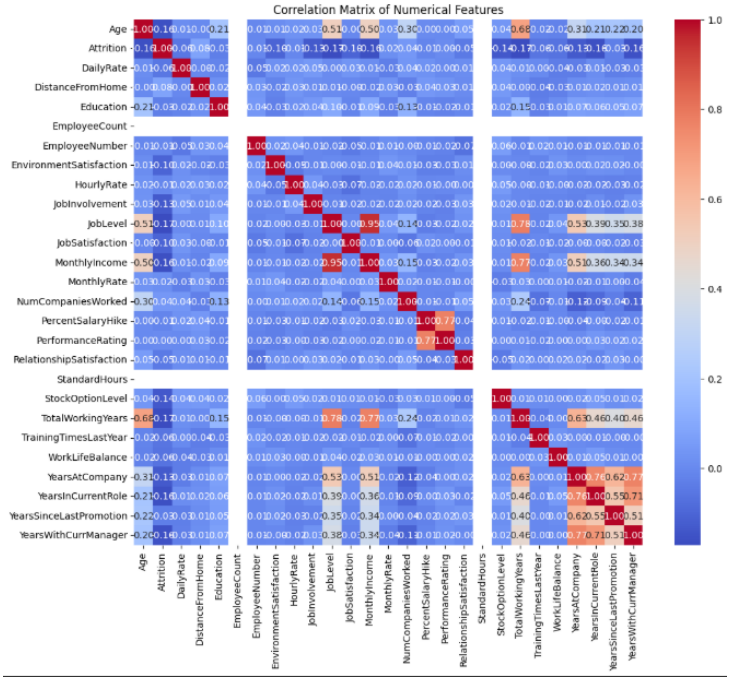

Figure 2: Example Correlation Heatmap of Key Variables.

3 Methodology

My analytical approach comprised a structured preprocessing pipeline, followed by rigorous model development for both classification and regression tasks, and an integration of these models for risk-impact assessment.

3.1 Preprocessing Pipeline

The raw data underwent several preprocessing steps to prepare it for modeling:

1. **Missing Data Imputation**: There were no missing values in the dataset.
2. **Feature Scaling**: Continuous numerical variables were standardized using 'StandardScaler' to ensure they have zero mean and unit variance.
3. **Categorical Encoding**: Nominal categorical features were transformed into a numerical format using 'OneHotEncoder'.
4. **Class Resampling**: To address the class imbalance in the 'Attrition' variable, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to synthesize new samples for the minority class (attrition=Yes).

3.2 Model Training

3.2.1 Classification Models

Several algorithms were trained to predict employee attrition: Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, XGBoost, and a Voting Classifier ensemble. Hyperparameters for each model (e.g., C values for Logistic Regression/SVM, tree depths for tree-based models) were tuned using GridSearchCV with cross-validation. The performance metrics are detailed in Table 1. The **tuned Voting Classifier** demonstrated the best overall performance.

Table 1: Classification Model Performance Metrics

Model	Accuracy	ROC AUC	Weighted F1
Logistic Regression	0.7925	0.7943	0.8109
Decision Tree	0.8163	0.6387	0.8113
SVM	0.8401	0.7817	0.8380
Random Forest	0.8367	0.7874	0.8109
XGBoost	0.8333	0.8016	0.8109
Voting (initial)	0.8605	0.8007	0.8514
Voting (tuned)	**0.8673**	**0.7940**	**0.8709**

3.2.2 Regression Models

To predict employees' future salary commitments (a simulated target for this analysis), various regression models were employed: Random Forest Regressor, Ridge Regression, Lasso Regression, Support Vector Regressor (SVR), XGBoost Regressor, and a Voting Regressor ensemble. Similar to classification, hyperparameters were optimized using

GridSearchCV with cross-validation. The **Voting Regressor** achieved a remarkably high R^2 of 1.0000 (Table 2), suggesting strong predictive power on this simulated target, likely due to the nature of the simulation.

Table 2: Regression Model Performance Metrics (Target: Simulated Next-Period Salary)

Model	RMSE	R^2	MAPE
Random Forest	120.05	0.9994	0.0087
Ridge	87.22	0.9997	0.0092
Lasso	82.42	0.9997	0.0088
SVR	5315.58	-0.1056	0.5303
XGBoost	108.17	0.9995	0.0125
Voting (initial)	72.48	0.9998	0.0063
Voting (tuned)	**73.02**	**1.0000**	**0.0120**

3.3 Retention Flagging

Employees were assigned a probability of staying (P_{stay}) based on the output of the best classification model. Those with $P_{stay} >$ **0.60** were identified as 'likely-to-stay'. This segment, comprising approximately **89%** of the workforce based on model predictions, was used for a subset financial analysis to understand potential exposure even within this lower-risk group.

3.4 Feature Importance

Analysis of feature importance from the ensemble classifier (Figure 3) highlighted key drivers of attrition.

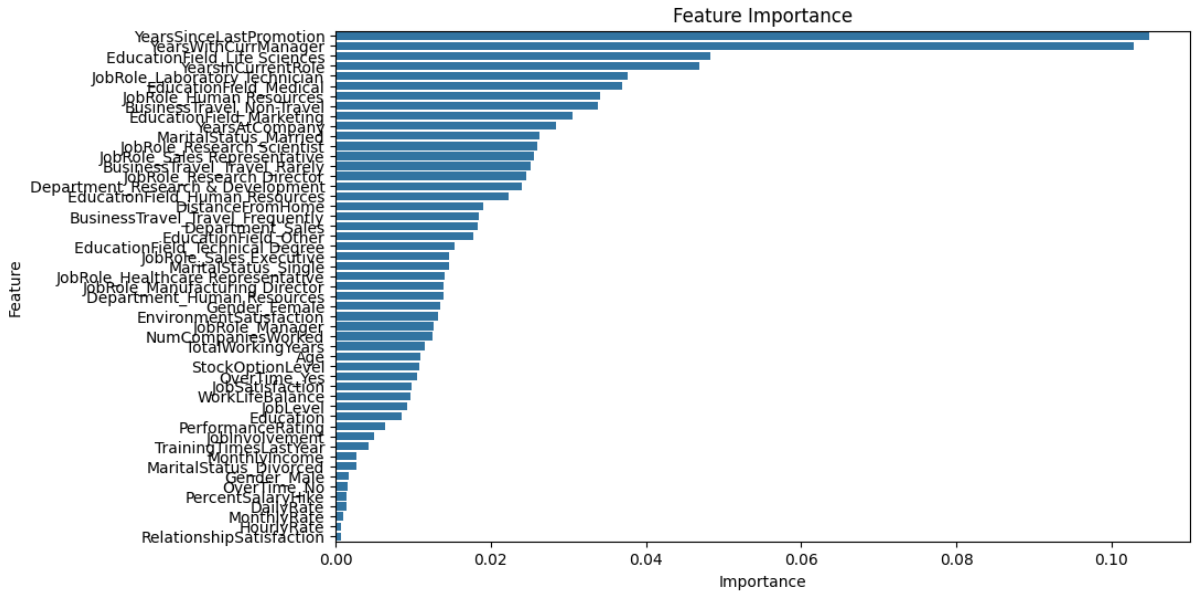

Figure 3: Top Features by Importance in Ensemble Classifier.

4 Financial Impact Analysis

The potential financial impact of employee attrition was estimated by calculating an expected loss for each employee. This was defined as the product of their probability of leaving ($P_{\text{leave},i} = 1 - P_{\text{stay},i}$) and their predicted next-period salary:

$$\text{ExpectedLoss}_i = P_{\text{leave},i} \times \text{PredictedSalary}_i$$

These individual expected losses were then aggregated to provide overall financial exposure:

- **Total expected financial exposure across all employees**: **$3,261,024.27**. This figure represents the total anticipated salary cost associated with potential turnover across the entire workforce.
- **Expected financial exposure from employees predicted to stay**: **$1,127,966.68**. This amount is the sum of expected losses for the subset of employees flagged as 'likely-to-stay' (i.e., $P_{stay} > \mathbf{0.60}$, representing **89%** of employees). It quantifies the residual financial risk even within this group.

These estimations assist HR and finance departments in forecasting budgetary risks associated with attrition and strategically allocating resources for talent retention initiatives.

5 Conclusion and Recommendations

The integrated modeling framework provides **precise employee-level attrition risk scores** and corresponding **salary cost projections**, thereby enabling a **data-driven approach** to retention strategies and financial planning. Based on my findings, I recommend the following:

- **Targeted Retention Programs**: Deploy tailored retention incentives and interventions focused on employees identified with **high predicted financial impact scores** (a combination of high P_{leave} and high PredictedSalary).
- **Refine Salary Projections**: Incorporate historical time-series salary data and macroeconomic factors, if available, to **further refine the accuracy** of the regression forecasts for future salary commitments.
- **Enhance Feature Set**: Augment the existing dataset with information from employee engagement surveys, exit interviews, and more granular performance data to **potentially improve the predictive power** of the attrition models.
- **Advanced Segmentation**: Utilize dimensionality reduction techniques (e.g., t-SNE, PCA) combined with clustering algorithms to **visualize and identify distinct segments** of at-risk employees, allowing for more nuanced HR intervention strategies.