**Reducing Employee Attrition through Predictive Analytics**

**Improving Employee Retention with Machine Learning**

Vishnu Lal Raveendran Nair

Centennial College

Business Analytics and Insights & 301440285

David Parent, Bilal Hasanzadah

August 12, 2025

**Contents**

[Executive Summary 4](#_Toc205580892)

[0.1. Executive Introduction 4](#_Toc205580893)

[0.2. Executive Objective 4](#_Toc205580894)

[0.3. Executive Model Description 4](#_Toc205580895)

[0.4. Executive Recommendations 5](#_Toc205580896)

[Introduction 7](#_Toc205580897)

[1.0 Background 7](#_Toc205580898)

[2.0 Problem Statement 7](#_Toc205580899)

[3.0 Objectives & Measurement 7](#_Toc205580900)

[4.0 Assumptions and Limitations 8](#_Toc205580901)

[Data Sources 8](#_Toc205580902)

[5.0 Data Set Introduction 8](#_Toc205580903)

[6.0 Exclusions 10](#_Toc205580904)

[7.0 Data Dictionary (Selected Fields) 11](#_Toc205580905)

[Data Exploration 14](#_Toc205580906)

[8.0. Data Exploration Techniques 14](#_Toc205580907)

[8.1 Key Insights from Data Exploration 40](#_Toc205580908)

[9.0 Data Cleansing 42](#_Toc205580909)

[10.0 Summary 42](#_Toc205580910)

[Data Preparation and Feature Engineering 43](#_Toc205580911)

[11.0 Data Preparation Needs 43](#_Toc205580912)

[11.1 Summary of Transformed & Excluded Fields 44](#_Toc205580913)

[11.2 Handling Class Imbalance 44](#_Toc205580914)

[12.0 Feature Engineering 45](#_Toc205580915)

[12.1 Notes on Feature Design 46](#_Toc205580916)

[Model Exploration 46](#_Toc205580917)

[13.0 Modeling Approach / Introduction 46](#_Toc205580918)

[14.0 Tree-Based Models 48](#_Toc205580919)

[14.1 Decision Tree (Untuned, No Rebalancing) 49](#_Toc205580920)

[14.2 Decision Tree (Tuned + SMOTE + AUC Criterion) 49](#_Toc205580921)

[14.3 Random Forest (SMOTE) 51](#_Toc205580922)

[14.4 XGBoost (Tuned + SMOTE) 54](#_Toc205580923)

[15.0 Regression-Based Models 55](#_Toc205580924)

[15.1 Logistic Regression (Class Weight = Balanced) 56](#_Toc205580925)

[15.2 Logistic Regression (Stepwise AIC) 59](#_Toc205580926)

[16.0 Neural Network Models 60](#_Toc205580927)

[16.1 Neural Network (Class Weight Only) 61](#_Toc205580928)

[16.2 Neural Network (Tuned + Class Weights) 61](#_Toc205580929)

[17.0 Model Comparison Summary 62](#_Toc205580930)

[Model Recommendation 63](#_Toc205580931)

[18.0 Model Selection 63](#_Toc205580932)

[19.0 Model Theory 64](#_Toc205580933)

[19.1 Model Assumptions and Limitations 64](#_Toc205580934)

[20.0 Model Sensitivity to Key Drivers 65](#_Toc205580935)

[21.0 Additional Models to Address Business Objectives 67](#_Toc205580936)

[Conclusion and Recommendations 68](#_Toc205580937)

[22.0 Impacts on Business Problem 68](#_Toc205580938)

[23.0 Recommended Next Steps for HR Leadership 69](#_Toc205580939)

[24.0 References 71](#_Toc205580940)

# Executive Summary

## Executive Introduction

Voluntary employee attrition remains a persistent challenge, with significant costs tied to replacement, productivity loss, and disruption to team cohesion. This project uses predictive analytics to move beyond historical reporting and into forward-looking risk identification, enabling HR to proactively engage employees most at risk of leaving. By combining internal HR data with market compensation benchmarks and advanced machine learning techniques, the analysis delivers actionable insights rooted in evidence.

## Executive Objective

The core objective of this initiative was to develop an accurate, interpretable model that predicts which employees are at the highest risk of voluntary attrition. The goal was not just technical performance, but practical usability for HR decision-making, with a target to reduce attrition by 5–10% through timely interventions. Key questions answered:

* Who is likely to leave?
* Why are they at risk?
* What can HR do about it?

## Executive Model Description

Eight machine learning models were tested across three categories (tree-based, regression-based, and neural networks), each evaluated with four class-balancing techniques.

Logistic Regression with Class Weighting was selected for deployment based on its superior balance of accuracy (ROC AUC = 0.8362) and interpretability (F1 Score = 0.5202). SHAP values and odds ratios provided HR-friendly explanations of top drivers.

**Top Predictors of Attrition:**

* OverTime – Doubles attrition risk
* Low Salary – Under-market pay linked to higher turnover
* Low RoleTenureRatio – Employees in new or stagnant roles more likely to leave
* Low Job Satisfaction & Work-Life Balance – Early warning indicators of disengagement
* These predictors align with Gallup, SHRM, and Work Institute research on why employees leave — reinforcing both the model’s validity and relevance to current HR challenges.

## Executive Recommendations

Based on the model’s findings, five strategic actions are recommended to reduce attrition risk and improve employee retention:

1. **Reduce Career Stagnation** *(Odds Ratio = 2.00)*
   * Flag employees with **≥5 years in the same role** and **low RoleTenureRatio**
   * Conduct stay interviews and offer targeted career mobility pathways (lateral moves, stretch assignments, cross-functional exposure)
2. **Target Compensation Gaps** *(Odds Ratio = 0.41)*
   * Use **CompaRatio** to identify below-market earners, prioritizing those under **75%** and between **75–90%**
   * Focus on high performers and “High Risk” model flags for salary adjustments or retention bonuses
3. **Manage Overtime-Linked Burnout** *(Odds Ratio = 2.09)*
   * Audit teams with sustained overtime
   * Apply workload caps and redistribute tasks where possible
   * For overtime employees with low CompaRatio, consider immediate pay equity adjustments
4. **Act on Engagement Signals**
   * Leverage **engagement survey results** for Job Involvement, Job Satisfaction, Work-Life Balance, and Environment Satisfaction
   * Flag low scorers as priority retention cases and monitor sudden drops in engagement metrics as early warning signs
5. **Tailor Actions for Travel-Linked Risk** *(Odds Ratio = 1.86)*
   * **Non-Sales:** Audit frequent travelers and set travel balance guidelines or flexible/remote options
   * **Sales:** Review and benchmark incentive programs to ensure competitiveness and reward sustainable performance
6. **Enhance Performance Data Granularity**
   * Revise the performance rating system from the current 3 and 4 scale to a 0–5 numeric scale.
   * This will allow HR to better differentiate truly top performers from average performers and target retention strategies more precisely.

# Introduction

## 1.0 Background

Voluntary employee attrition continues to challenge organizations, increasing costs, disrupting teams, and draining institutional knowledge. In today’s competitive talent market, reducing unwanted turnover has become a strategic HR priority. This project uses predictive analytics to move from reactive attrition reporting to proactive retention planning, helping the organization identify and support at-risk employees before they decide to leave.

## 2.0 Problem Statement

Despite established HR strategies, certain employee segments are experiencing higher-than-expected attrition. Traditional dashboards reveal past exits but offer little predictive power. This project aims to fill that gap by developing a forward-looking model that not only identifies likely leavers but also explains why enabling HR to prioritize resources and interventions effectively.

## 3.0 Objectives & Measurement

**Objective:**  
To build an accurate and interpretable machine learning model that predicts voluntary attrition and enables a 5–10% reduction through timely, targeted HR actions.

**Key Deliverables:**

* Predictive model trained on internal HR and external compensation data
* Evaluation using F1 Score and AUC for balanced performance
* SHAP-based feature interpretation for business usability
* Practical HR recommendations aligned to top attrition drivers

**Success Criteria:**  
A model that balances accuracy with interpretability and supports HR in taking proactive actions that align with business goals.

## 4.0 Assumptions and Limitations

* All terminations are assumed to be voluntary due to lack of exit type data.
* Hire and termination dates were unavailable, limiting the ability to analyze seasonality and tenure effects.
* The model reflects historical patterns and does not account for sudden external changes.
* Unstructured drivers such as job satisfaction or personal motivation are not captured.
* Compensation benchmarks are static and may lag real-time shifts.

# Data Sources

## 5.0 Data Set Introduction

The dataset used in this project is a modified version of the IBM HR Analytics Employee Attrition dataset, containing detailed records of 1,470 employees. To improve predictive power and relevance for HR decision-making, the original dataset was enriched with:

* External market compensation benchmarks for salary comparisons
* HR-engineered features derived from tenure, promotion, and income data
* Binary target variable Attrition (Yes/No), treated as voluntary due to missing exit type details
* The combined dataset includes demographic, job-related, and compensation variables across multiple departments and roles.

A screenshot of a computer

AI-generated content may be incorrect.

A close-up of a computer screen

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A close-up of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.A close-up of a white background

AI-generated content may be incorrect.

## 6.0 Exclusions

The dataset did not contain any missing values, so no records were removed due to null or incomplete fields.

However, several columns were excluded based on their lack of predictive value or redundancy after feature engineering. The exclusions were made to simplify the modeling dataset and avoid data leakage. These included:

|  |  |
| --- | --- |
| Column Name | Reason for Exclusion |
| EmployeeNumber | Unique identifier; does not carry predictive value |
| EmployeeCount | Constant value (always 1); provides no variability |
| Over18 | Constant value (“Y” for all employees); non-informative |
| StandardHours | Constant value (80 hours for all); not useful for prediction |
| Market Median (USD) | Used only for calculating CompaRatio; dropped after derived feature was created |
| Role & Level | Intermediate lookup column used for compensation mapping; no longer needed |
| JobRoleLevel | Support column for market benchmark mapping; removed after feature generation |

A screenshot of a computer program

AI-generated content may be incorrect.

Additionally:

* No time-based exclusions (e.g., seasonal or cohort analysis) were applied due to the absence of HireDate and TerminationDate.
* All attrition cases were assumed to be voluntary, as the dataset lacks a termination type or exit reason field.

## 7.0 Data Dictionary (Selected Fields)

Below is a representative list of features used across models. Additional engineered and encoded features were included in the final modeling sets:

|  |  |  |
| --- | --- | --- |
| Feature Name | Description | Data Type |
| Age | Age of the employee (in years) | Numeric |
| Attrition | Target variable indicating if the employee left (Yes/No) | Categorical |
| BusinessTravel | Frequency of business travel | Categorical |
| DailyRate | Daily pay rate of the employee | Numeric |
| Department | Department in which the employee works (e.g., Sales, R&D, HR) | Categorical |
| DistanceFromHome | Distance between home and work (in kilometers) | Numeric |
| Education | Education level (1 = Below College, 5 = Doctorate) | Ordinal Numeric |
| EducationField | Field of study (e.g., Life Sciences, Technical Degree, Other) | Categorical |
| EnvironmentSatisfaction | Satisfaction with the work environment (1 = Low, 4 = Very High) | Ordinal Numeric |
| Gender | Gender of the employee | Categorical |
| HourlyRate | Hourly wage | Numeric |
| JobInvolvement | Perceived involvement in the job (1 = Low, 4 = Very High) | Ordinal Numeric |
| JobLevel | Level within the company hierarchy | Ordinal Numeric |
| JobRole | Employee's specific role (e.g., Sales Executive, Research Scientist) | Categorical |
| JobSatisfaction | Job satisfaction level (1 = Low, 4 = Very High) | Ordinal Numeric |
| MaritalStatus | Marital status (e.g., Single, Married, Divorced) | Categorical |
| MonthlyIncome | Gross monthly salary | Numeric |
| MonthlyRate | Monthly pay rate | Numeric |
| NumCompaniesWorked | Number of companies the employee has worked for | Numeric |
| OverTime | Whether the employee works overtime (Yes/No) | Categorical |
| PercentSalaryHike | Percent increase in salary during the last appraisal | Numeric |
| PerformanceRating | Performance rating (1 = Low, 4 = Excellent) | Ordinal Numeric |
| RelationshipSatisfaction | Satisfaction with workplace relationships (1 = Low, 4 = Very High) | Ordinal Numeric |
| StockOptionLevel | Level of stock options granted (0–3) | Ordinal Numeric |
| TotalWorkingYears | Total years of professional experience | Numeric |
| TrainingTimesLastYear | Number of training sessions attended in the past year | Numeric |
| WorkLifeBalance | Work-life balance rating (1 = Bad, 4 = Excellent) | Ordinal Numeric |
| YearsAtCompany | Total years at the current company | Numeric |
| YearsInCurrentRole | Years in the current job role | Numeric |
| YearsSinceLastPromotion | Years since the employee was last promoted | Numeric |
| YearsWithCurrManager | Years working under the current manager | Numeric |
| Compa Ratio | Derived: MonthlyIncome / Market Median for Role | Numeric |
| PromotionStagnation | Derived: YearsSinceLastPromotion / YearsAtCompany | Numeric |
| RoleTenureRatio | Derived: YearsInCurrentRole / YearsAtCompany | Numeric |
| JobHoppingCategory | Derived: Binned frequency of job changes (Stable, Moderate, etc.) | Categorical |
| PromotionFlag | Derived: 1 if promoted recently, else 0 | Binary (0/1) |

# Data Exploration

## 8.0. Data Exploration Techniques

A structured exploratory data analysis (EDA) was conducted to uncover patterns, assess data quality, and identify variables most associated with attrition. Techniques used include:

* Descriptive statistics: Summary stats for all numeric variables

A screenshot of a graph

AI-generated content may be incorrect.

* Distribution analysis: Histograms for numeric variables.

A graph of age distribution

AI-generated content may be incorrect.

A graph of a distribution of daily rate

AI-generated content may be incorrect.

A graph of a number of people

AI-generated content may be incorrect.

A graph of a distribution of hours

AI-generated content may be incorrect.

A graph of a distribution of income

AI-generated content may be incorrect.

A graph of a distribution of monthly rate

AI-generated content may be incorrect.

A graph of a number of companies

AI-generated content may be incorrect.

A graph of a number of salary hikers

AI-generated content may be incorrect.

A graph of a number of years

AI-generated content may be incorrect.

A graph of a distribution of training

AI-generated content may be incorrect.

A graph of a number of years

AI-generated content may be incorrect.

A graph of a distribution of a number of years

AI-generated content may be incorrect.

A graph of a number of years

AI-generated content may be incorrect.

A graph of a distribution of a company

AI-generated content may be incorrect.

A diagram of a distribution of compa ratio

AI-generated content may be incorrect.

A graph of a distribution of stagnation

AI-generated content may be incorrect.

A graph of a distribution of the number of people

AI-generated content may be incorrect.

* Ordinal columns were relabeled for better interpretability (e.g., JobSatisfaction = "Low", "High", etc.).
* Categorical distributions: Count plots for all object/ordinal variables

A graph of a graph of travel

AI-generated content may be incorrect.

A graph of a graph with text

AI-generated content may be incorrect.

A graph of a graph of education

AI-generated content may be incorrect.

A graph of a graph of education

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

A graph showing a number of people

AI-generated content may be incorrect.

A graph of a graph of a number of people

AI-generated content may be incorrect.

A graph of a number of people

AI-generated content may be incorrect.

A graph of a number of people

AI-generated content may be incorrect.

A graph of a number of different colored bars

AI-generated content may be incorrect.

A graph of a number of individuals

AI-generated content may be incorrect.

A graph with a number of squares

AI-generated content may be incorrect.

A graph with a bar and text

AI-generated content may be incorrect.A graph of a number of colored bars

AI-generated content may be incorrect.

A graph of a number of different colored bars

AI-generated content may be incorrect.

A graph of a bar chart

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

A graph of a number of promotion

AI-generated content may be incorrect.

* Target class balance: Class distribution of Attrition (pie chart + table)

A screenshot of a computer screen

AI-generated content may be incorrect.

A pie chart with a blue and red circle

AI-generated content may be incorrect.

* Outlier detection: Tukey method to quantify outlier counts per variable

A diagram of a box plot

AI-generated content may be incorrect.A graph with a box and text

AI-generated content may be incorrect.A graph with a blue square

AI-generated content may be incorrect.A graph with numbers and a box

AI-generated content may be incorrect.A graph showing a number of income

AI-generated content may be incorrect.A blue rectangular box plot with numbers

AI-generated content may be incorrect.A graph with a blue square

AI-generated content may be incorrect.A graph with a bar and a number of numbers

AI-generated content may be incorrect.A graph with a bar and a number of numbers

AI-generated content may be incorrect.A graph of a training

AI-generated content may be incorrect.A graph with numbers and a line

AI-generated content may be incorrect.A graph with a blue square

AI-generated content may be incorrect.A graph with numbers and dots

AI-generated content may be incorrect.A graph with a blue square

AI-generated content may be incorrect.A diagram of a box plot

AI-generated content may be incorrect.A diagram of a graph

AI-generated content may be incorrect.A graph with a box and text

AI-generated content may be incorrect.

**Correlation analysis:**

* Pearson correlation heatmaps before and after feature reduction

A screenshot of a graph

AI-generated content may be incorrect.

* Point biserial correlation for numeric variables vs. binary attrition

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* Chi-square tests for categorical variables vs. attrition

A screenshot of a computer

AI-generated content may be incorrect.

## 8.1 Key Insights from Data Exploration

* Class Imbalance: Attrition is imbalanced — only ~16% of employees left the company.
* Skewness: MonthlyIncome, TotalWorkingYears, and DistanceFromHome show right-skewed distributions.
* Outliers: Several features, especially compensation-related ones, contain outliers. These were retained after confirming they aligned with realistic business cases.

**Highly Correlated Feature Pairs (|r| > 0.7)**

To avoid multicollinearity and retain model interpretability, the following highly correlated pairs were reviewed:

|  |  |  |
| --- | --- | --- |
| Feature 1 | Feature 2 | Correlation (r) |
| MonthlyIncome | TotalWorkingYears | 0.77 |
| YearsWithCurrManager | YearsAtCompany | 0.77 |
| YearsAtCompany | YearsInCurrentRole | 0.76 |
| YearsInCurrentRole | YearsWithCurrManager | 0.71 |

**Variable Reduction Decisions**

Several numeric features were dropped due to high redundancy, weaker association with attrition, or improved interpretability via engineered alternatives:

* **TotalWorkingYears**
  + Correlated with MonthlyIncome (r = 0.77)
  + Slightly stronger correlation with Attrition (r = -0.1711), but *MonthlyIncome* was retained due to higher actionability.
* **YearsWithCurrManager**
  + Correlated with YearsAtCompany (r = 0.77) and YearsInCurrentRole (r = 0.71)
  + Weakest link to Attrition among the three (r = -0.1562) → Dropped.
* **YearsAtCompany**
  + Correlated with both YearsInCurrentRole (r = 0.76) and YearsWithCurrManager (r = 0.77)
  + Lowest correlation with Attrition (r = -0.1344) → Dropped.
* **YearsSinceLastPromotion**
  + Replaced by derived metric *PromotionStagnation*, which provides better interpretability for stakeholders.
  + Both variables show minimal correlation to Attrition (r ≈ ±0.03).
* **NumCompaniesWorked**
  + Dropped in favor of *JobHoppingIndex* (r = 0.2259 vs. 0.0435), which captures career instability more effectively.

**Point-Biserial Correlation with Attrition**

Key variables with strongest associations:

* **JobHoppingIndex** → r = 0.2259 *(engineered feature retained)*
* **CompaRatio** → r = -0.1532 *(retained)*
* **OverTime** → Categorical variable with strong predictive signal *(retained)*

Other original numeric features showed weaker associations and were retained or removed based on overall importance, multicollinearity, and business relevance.

**Chi-Square Results:**

* Statistically significant categorical variables: BusinessTravel, OverTime, JobRole, MaritalStatus
* Non-significant ones (e.g., Gender, PerformanceRating, RelationshipSatisfaction and Education) were removed from the final dataset

## 9.0 Data Cleansing

* No missing values were found — confirmed via .isnull().sum()
* No imputation required
* Constant or low-information columns (e.g., EmployeeCount, Over18, StandardHours) were dropped
* Redundant features with high multicollinearity or weak signal were dropped based on:
* Correlation thresholds (e.g., >0.7)
* Predictive power vs. target (Attrition)
* The final dataset retained both original and engineered features for balanced representation and predictive strength.

## 10.0 Summary

The exploratory analysis confirmed that attrition is a minority class challenge (~16% attrition rate) and revealed patterns across both original and engineered variables.

Several numeric features, such as MonthlyIncome (skew = 1.37), YearsSinceLastPromotion (1.98), YearsAtCompany (1.76), TotalWorkingYears (1.12), and NumCompaniesWorked (1.03), showed substantial right skew, indicating long-tailed distributions often tied to tenure, experience, and compensation. These variables were retained with appropriate scaling or transformation where necessary. Outliers were identified but preserved after confirming they reflect valid business scenarios (e.g., high earners with long tenures).

No missing values were present. Constant-value and redundant columns were dropped, and engineered features such as CompaRatio, PromotionStagnation, and JobHoppingCategory were found to offer both interpretability and stronger statistical association with attrition.

EDA insights informed class balancing strategies and supported the simplification of the modeling dataset through variable reduction, while preserving business relevance.

# Data Preparation and Feature Engineering

## 11.0 Data Preparation Needs

The dataset was initially complete and clean, with no missing values. However, several steps were taken to prepare the data for modeling:

Removed constant and low-variance columns: Columns such as EmployeeCount, Over18, and StandardHours offered no predictive power and were excluded.

Dropped identifiers and lookup variables: Columns like EmployeeNumber, Role & Level, and Market Median (USD) were dropped after being used to generate derived features.

Retained separate datasets based on model requirements: For Logistic Regression and Neural Networks, numeric variables were scaled using StandardScaler

For tree-based models (e.g., Random Forest, XGBoost), the original unscaled dataset was retained, as these algorithms are not sensitive to magnitude or distribution.

Encoding: One-hot encoding was applied to categorical variables for all non-linear models.

Binary and ordinal variables were either label-encoded or transformed into interpretable categories before encoding (e.g., Education, JobSatisfaction).

Feature reduction: Highly correlated features were dropped using a Pearson correlation threshold of 0.7 and domain knowledge (e.g., YearsWithCurrManager, YearsAtCompany, NumCompaniesWorked).

Preference was given to engineered variables that were more interpretable (e.g., PromotionStagnation over YearsSinceLastPromotion).

## 11.1 Summary of Transformed & Excluded Fields

|  |  |
| --- | --- |
| Action | Columns Affected |
| Dropped (constant) | EmployeeCount, Over18, StandardHours |
| Dropped (identifier/lookup) | EmployeeNumber, Market Median (USD), Role & Level, JobRoleLevel |
| Dropped (correlated or weak) | YearsWithCurrManager, YearsAtCompany, NumCompaniesWorked, PerformanceRating, Gender |
| Scaled | All numeric variables for regression/NN models (excluding trees) |
| Encoded | All categorical variables (via one-hot or binary encoding) |

## 11.2 Handling Class Imbalance

The dataset was highly imbalanced, with only ~16% of records representing employees who voluntarily left the organization (Attrition = Yes). To address this, four rebalancing techniques were systematically tested for every model:

* SMOTE (Synthetic Minority Oversampling Technique)
* Random Oversampling
* Random Undersampling
* Class Weighting (using class\_weight='balanced' or equivalent)

For each model, the version that yielded the highest ROC AUC score (with F1 Score as a secondary consideration) was selected for final comparison. This ensured that class imbalance was handled rigorously and consistently across all modeling techniques.

Example: Logistic Regression and Neural Networks performed best with class weighting, while tree-based models such as Random Forest and XGBoost performed best with SMOTE.

This multi-method rebalancing approach improved model fairness, sensitivity to attrition cases, and reliability of downstream business recommendations.

## 12.0 Feature Engineering

Feature engineering was a key step in making the model more interpretable and predictive. Several new features were created based on HR business logic, compensation benchmarks, and behavioral signals:

|  |  |
| --- | --- |
| Feature Name | Description |
| CompaRatio | MonthlyIncome divided by external market median for the employee’s role |
| PromotionStagnation | YearsSinceLastPromotion divided by YearsAtCompany (0 if tenure = 0) |
| RoleTenureRatio | YearsInCurrentRole divided by YearsAtCompany (0 if tenure = 0) |
| JobHoppingCategory | Categorization based on NumCompaniesWorked and tenure: Stable, Frequent, Chronic, etc. |
| PromotionFlag | Binary flag: 1 if recently promoted, else 0 |

These features aimed to capture employee risk patterns that aren’t directly visible from the original dataset.

## 12.1 Notes on Feature Design

* CompaRatio was derived using an external benchmark table that mapped job role and level to market median salaries.
* JobHoppingCategory was binned using both absolute company count and relative frequency (NumCompaniesWorked / TotalWorkingYears).
* Some original features (like NumCompaniesWorked) were dropped after their information was captured by engineered features.
* These new variables improved both predictive power and business interpretability, especially when used in SHAP analysis.

# Model Exploration

## 13.0 Modeling Approach / Introduction

A comparative modeling strategy was employed to evaluate multiple algorithms and identify the model that best balances predictive performance with interpretability for HR decision-making. The objective was not only to accurately predict employee attrition but also to ensure that model outputs could be clearly understood and acted upon by non-technical stakeholders.

**Evaluation Metrics**  
Two key metrics were used to assess model performance:

* ROC AUC Score (Primary Metric): Measures a model’s ability to distinguish between classes across all thresholds. It was selected as the primary metric due to its robustness in imbalanced classification settings.
* F1 Score (Secondary Metric): Balances precision and recall, making it a suitable complement to AUC for evaluating models on the minority class (Attrition = Yes).

**Class Rebalancing Techniques**  
To ensure fairness and maximize performance across models, four class balancing strategies were systematically tested:

* SMOTE (Synthetic Minority Oversampling Technique)
* Random Oversampling
* Random Undersampling
* Class Weighting

Each model was evaluated using all four rebalancing methods, and the best-performing version (based on ROC AUC and F1 Score) was selected for comparison. This ensured a consistent and rigorous evaluation process.

**Model Interpretability**  
SHAP (SHapley Additive exPlanations) was used to interpret the models and identify the most influential features driving attrition risk. This allowed model outputs to be translated into actionable HR insights.

**Model Groups**  
The final set of models was grouped into three categories for structured evaluation:

* Tree-Based Models: Non-parametric models that capture non-linear splits and feature interactions
* Regression-Based Models: Interpretable, linear models grounded in statistical theory
* Neural Networks: Non-linear, flexible models capable of capturing complex relationships

A total of eight models were tested and compared across these three categories.

## 14.0 Tree-Based Models

Tree-based models operate by recursively splitting the dataset based on feature thresholds, forming a decision tree-like structure. These models—such as Decision Trees, Random Forests, and XGBoost—are powerful for capturing non-linear relationships and feature interactions.

A major advantage of tree-based algorithms is that they do not require feature scaling or transformation. They are also inherently robust to:

* Outliers, as splits are based on thresholds rather than distance-based measures.
* Multicollinearity, since trees select only the most informative features at each split and can ignore redundant ones.

**Feature Handling for Tree-Based Models**

* For tree-based models, we used the raw feature set, including both original and engineered features.
* The only preprocessing applied was one-hot encoding of categorical variables (using drop\_first=True), ensuring compatibility with algorithms that require numerical input.
* No transformations (e.g., log, normalization) or feature reductions were applied, as tree-based models are naturally suited to handle raw, high-dimensional data.

This approach ensured that the tree models could fully leverage the complexity and structure of the engineered features without the risk of distortion due to unnecessary transformations.

### 14.1 Decision Tree (Untuned, No Rebalancing)

A simple classification tree trained on the original dataset without tuning or resampling.

Pros:

* Transparent and intuitive
* Quick to train

Cons:

* Strongly biased toward majority class
* Poor generalization; overfitting risk

Performance:

* F1 Score: 0.3515
* ROC AUC: 0.6164

A screenshot of a computer program

AI-generated content may be incorrect.

### 14.2 Decision Tree (Tuned + SMOTE + AUC Criterion)

This version applied SMOTE and tuned tree depth and splitting strategy using AUC to improve balance.

Pros:

* Targeted minority class better using AUC splits
* SMOTE helped address imbalance

Cons:

* Despite tuning, model failed to classify attrition cases (F1 = 0)
* Not usable in real-world deployment

Performance:

* F1 Score: 0.0000
* ROC AUC: 0.7427

A screenshot of a computer program

AI-generated content may be incorrect.

**Tree Visualization**

A diagram of a company

AI-generated content may be incorrect.

### 14.3 Random Forest (SMOTE)

Random Forest builds an ensemble of decision trees using bagging (bootstrap aggregation) and averages their predictions.

Pros:

* Robust to overfitting
* Captures interactions between variables
* SHAP values revealed key drivers: OverTime, CompaRatio, PromotionStagnation

Cons:

* Less interpretable than a single tree or logistic regression
* Can be sensitive to noisy data if not tuned

Performance:

* F1 Score: 0.4086
* ROC AUC: 0.8283

A screenshot of a computer

AI-generated content may be incorrect.

The highest performing model among tree-based model is Random Forest. SHAP analysis results below:

A screenshot of a computer

AI-generated content may be incorrect.

A graph of different colored lines

AI-generated content may be incorrect.

### 14.4 XGBoost (Tuned + SMOTE)

XGBoost is a gradient boosting framework that sequentially builds trees to correct errors from previous trees. It is known for high performance and fine control.

Pros:

* Excellent classification ability
* SHAP output aligned with business logic
* SMOTE helped increase recall

Cons:

* Requires careful tuning
* More complex to explain to non-technical stakeholders

Performance:

* F1 Score: 0.4959
* ROC AUC: 0.8115

A screenshot of a computer

AI-generated content may be incorrect.

## 15.0 Regression-Based Models

Logistic Regression models the log-odds of the binary target variable (Attrition) as a linear combination of the predictor variables. It is widely used for classification problems due to its simplicity, interpretability, and effectiveness. Coefficients from the model provide insights into the direction (positive or negative) and strength of influence of each feature on the likelihood of attrition.

**Data Preparation for Logistic Regression**

To ensure optimal performance and interpretability of the logistic regression model, several preprocessing steps were performed:

**1. Handling Skewness in Numeric Features**

Skewed features can negatively affect model performance and the validity of statistical assumptions. The following steps were taken to reduce skewness:

* Step 1: Skewness was calculated for all numeric variables in the dataset.
* Step 2: Features with absolute skewness greater than 0.6 were flagged for transformation.
* Step 3: A log transformation using log2(x + 1) was applied to these variables.
* Step 4: Post-transformation skewness was recalculated. Where necessary, extreme outliers (e.g., in PromotionStagnation\_log2) were clipped at the 1st and 99th percentiles.

Result: All transformed features were successfully brought down to a skewness value of less than 0.75, thereby ensuring a more normalized distribution suitable for logistic regression.

**2. Transformed Features**

The following features were log-transformed and retained for modeling:

PromotionStagnation\_log2

PercentSalaryHike\_log2

MonthlyIncome\_log2

DistanceFromHome\_log2

YearsInCurrentRole\_log2

**3. Encoding and Splitting**

Categorical features were one-hot encoded with drop\_first=True to avoid multicollinearity from the dummy variable trap.

The dataset was then split into training and testing sets using a stratified 70-30 split to preserve the proportion of the target class (Attrition).

**4. Standardization**

A StandardScaler was applied to ensure all features had a mean of 0 and a standard deviation of 1. This is especially important for logistic regression, which assumes normally distributed predictors.

These preprocessing steps ensured the input data met the assumptions of logistic regression and supported both accurate coefficient interpretation and improved model performance.

### 15.1 Logistic Regression (Class Weight = Balanced)

The base logistic regression model applied class\_weight='balanced' to improve recall on the minority class.

A screenshot of a screen

AI-generated content may be incorrect.

Pros:

* Best overall performance
* Coefficients directly show feature impact
* SHAP confirmed alignment with business intuition

Cons:

* Assumes linear relationship between features and log-odds
* Sensitive to multicollinearity (mitigated via scaling and feature selection)

Performance:

* F1 Score: 0.5202
* ROC AUC: 0.8362

A screenshot of a computer

AI-generated content may be incorrect.

A graph with red and blue dots

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* Mean SHAP value → how much the feature contributes to predictions
* Coefficient → direction of influence (positive = increases attrition risk)
* Odds Ratio → how much the risk changes (relative to baseline)

### 15.2 Logistic Regression (Stepwise AIC)

Stepwise logistic regression reduces the model to a subset of variables using the Akaike Information Criterion (AIC) for optimal balance between simplicity and fit.

Pros:

* More compact; easier for stakeholder communication
* Removes non-informative features

Cons:

* Over-simplified: key drivers were excluded
* Very poor F1 score due to underfitting

Performance:

* F1 Score: 0.0533
* ROC AUC: 0.6866

A screenshot of a computer

AI-generated content may be incorrect.

## 16.0 Neural Network Models

Neural networks are inspired by the human brain and consist of layers of interconnected neurons. They are powerful for capturing non-linear and high-dimensional patterns but lack transparency. First, the best rebalancing technique was identified by comparing class\_weight='balanced' with three other methods—SMOTE, Random Oversampling, and Random Undersampling—based on F1 Score and ROC AUC. Then, the class\_weight='balanced' technique was implemented within a detailed neural network architecture using Keras, incorporating batch normalization, dropout layers, and early stopping to enhance generalization and prevent overfitting.

A screenshot of a computer screen

AI-generated content may be incorrect.

### 16.1 Neural Network (Class Weight Only)

A two-layer feedforward neural network trained using class weighting to address imbalance.

Pros:

* Captures complex interactions
* Class weights handled imbalance without oversampling

Cons:

* Requires careful architecture design
* Interpretation requires SHAP or DeepExplainer

Performance:

* ROC AUC: 0.7703
* F1 Score: 0.4607

A screenshot of a graph

AI-generated content may be incorrect.

### 16.2 Neural Network (Tuned + Class Weights)

This version was tuned using keras-tuner to optimize dropout, layer size, and learning rate. Class weights were retained.

Pros:

* Performance improved significantly after tuning
* SHAP revealed similar top drivers to other models

Cons:

* Less HR-friendly due to limited explainability
* Computationally heavier

Performance:

* F1 Score: 0.5098
* ROC AUC: 0.8090

A screenshot of a computer screen

AI-generated content may be incorrect.

## 17.0 Model Comparison Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model # | Model Description | F1 Score | ROC AUC | Rebalancing Strategy | Notes |
| 0 | Decision Tree (Full) | 0.3515 | 0.6164 | None | Weak baseline |
| 1 | Decision Tree (Tuned + SMOTE + AUC) | 0.0000 | 0.7427 | SMOTE + AUC criterion | Failed on positive class |
| 2 | Random Forest (SMOTE) | 0.4086 | 0.8283 | SMOTE | Strong AUC, decent interpretability |
| 3 | XGBoost (Tuned + SMOTE) | 0.4959 | 0.8115 | SMOTE | High performer; less interpretable |
| 4 | **Logistic Regression (Class Weight)** | **0.5202** | **0.8362** | **Class Weight (balanced)** | **Best overall performance and interpretability** |
| 5 | Logistic Regression (Stepwise AIC) | 0.0533 | 0.6866 | Class Weight | Poor recall; not suitable |
| 6 | Neural Network (Class Weight) | 0.4607 | 0.7703 | Class Weight | Moderate performance |
| 7 | Neural Network (Tuned + Class Weights) | 0.5098 | 0.8090 | Class Weight (tuned model) | Tuned version competitive, less interpretable |

# Model Recommendation

## 18.0 Model Selection

After evaluating eight classification models using four rebalancing strategies, Logistic Regression with class weighting was selected as the final model for deployment. While other models like Random Forest, XGBoost, and Neural Networks performed well, Logistic Regression stood out for its ability to offer both high predictive performance and clear interpretability — a key requirement for informing HR strategy.

This model achieved the highest ROC AUC score (0.8362) and best F1 score (0.5202) among all candidates. More importantly, it provides coefficients and odds ratios that are easily translatable into business insights. These attributes make it suitable not just as a statistical model, but as a decision-support tool for HR leaders to identify and engage at-risk employees.

The model's output has been validated using SHAP (SHapley Additive Explanations), reinforcing the reliability of its predictions and supporting evidence-based decision-making.

|  |  |
| --- | --- |
| Metric | Score |
| ROC AUC | 0.8362 |
| F1 Score | 0.5202 |

## 19.0 Model Theory

Logistic Regression is a linear classification algorithm that models the log-odds of an event occurring (in this case, employee attrition) as a function of multiple independent variables. Each coefficient in the model represents the change in the log-odds of attrition associated with a one-unit change in the corresponding feature, holding all else constant.

The key advantage of logistic regression lies in its interpretability:

Coefficients can be exponentiated to derive odds ratios, allowing HR professionals to understand how much more (or less) likely attrition becomes when certain factors are present.

Unlike black-box models, this approach supports transparent and accountable decision-making, which is especially important in workforce planning and ethical HR practices.

The model was trained using standardized numerical features and one-hot encoded categorical features, with class\_weight='balanced' to ensure fairness in classification across both classes.

## 19.1 Model Assumptions and Limitations

While Logistic Regression is robust and transparent, it does come with some assumptions and limitations:

**Assumptions:**

* Linearity in log-odds: Assumes a linear relationship between input variables and the log-odds of attrition.
* Independence of features: Assumes features are not highly correlated (addressed via feature reduction).
* Binary classification: Suitable for “Yes” or “No” prediction tasks but does not rank risk levels beyond that.

**Limitations:**

* No time dimension: Cannot predict when an employee will leave, only if they are likely to.
* No causal inference: Correlation is not causation — the model identifies patterns, not reasons.
* No unstructured data: It doesn’t include qualitative signals such as manager feedback or sentiment from exit interviews.
* Despite these constraints, the model remains a practical and reliable tool for real-world HR deployment.

## 20.0 Model Sensitivity to Key Drivers

Feature sensitivity was assessed using a combination of Mean Absolute SHAP values, model coefficients, and odds ratios. This triangulated approach helped ensure that the most influential predictors were not only statistically significant but also meaningful for HR action.

The following features were identified as the top drivers of attrition:

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Direction of Impact | Odds Ratio | Interpretation |
| RoleTenureRatio | Strong Negative | 0.40 | Employees who spend longer time in their role relative to tenure are less likely to leave |
| MonthlyIncome\_log2 | Strong Negative | 0.41 | Higher income reduces the likelihood of attrition |
| OverTime\_Yes | Strong Positive | 2.09 | Overtime work more than doubles the odds of attrition |
| JobInvolvement\_High | Moderate Negative | 0.52 | Highly involved employees are less likely to leave |
| JobSatisfaction\_Very High | Moderate Negative | 0.49 | High job satisfaction significantly reduces attrition risk |
| YearsInCurrentRole\_log2 | Moderate Positive | 2.00 | Longer time in the same role increases attrition odds — possibly due to stagnation |
| BusinessTravel\_Travel\_Frequently | Moderate Positive | 1.86 | Frequent business travel is associated with higher turnover |
| EnvironmentSatisfaction\_High/Very High | Moderate Negative | ~0.51–0.56 | Higher environmental satisfaction correlates with lower attrition |
| WorkLifeBalance\_High | Moderate Negative | 0.59 | Better work-life balance lowers attrition odds |
| JobHoppingCategory\_Moderate | Mild Negative | 0.63 | Moderately stable job history reduces risk slightly |
| Department\_Sales | Mild Positive | 1.57 | Employees in Sales show moderately higher attrition risk |

**Interpretation:**  
The SHAP scores validate that both organizational factors (e.g., role stagnation, travel, overtime) and individual perceptions (e.g., satisfaction, involvement) meaningfully influence attrition risk. The top driver, RoleTenureRatio, reflects employees who stay in the same role for a long time — and depending on their growth experience — may either be retained or become disengaged.

For HR, these findings suggest that career mobility, recognition, compensation, and workload management are critical levers for improving retention.

## 21.0 Additional Models to Address Business Objectives

While several advanced models such as XGBoost, Random Forest, and Tuned Neural Networks demonstrated strong predictive capabilities, Logistic Regression emerged as the top candidate based on both ROC AUC (0.8362) and F1 Score (0.5202).

In addition to its competitive performance, Logistic Regression offers a major advantage in interpretability — a critical requirement for HR decision-making. The ability to translate model outputs into clear odds ratios and business language makes it the most suitable choice for real-world deployment.

Given this strong balance of accuracy, transparency, and stakeholder usability, we are not recommending a switch to a more complex model at this stage. The other models will remain documented for reference but are not proposed for implementation, as they do not significantly outperform Logistic Regression and introduce interpretation challenges.

Should future business needs evolve — such as the inclusion of unstructured data, real-time scoring, or predictive time horizons — more advanced models can be revisited as part of a next-phase analytics roadmap.

# Conclusion and Recommendations

## 22.0 Impacts on Business Problem

The recommended Logistic Regression model meets the core business goal — enabling HR to proactively identify employees at risk of voluntary attrition while remaining transparent and easy to interpret for non-technical leaders.

* **Strong Performance:** ROC AUC = **0.8362**, F1 Score = **0.5202** — a reliable balance of accuracy and fairness for both stayers and leavers.
* **Business-Aligned Drivers:** OverTime, CompaRatio, and RoleTenureRatio directly link to HR realities like burnout, pay inequity, and career stagnation.
* **Actionable Insights:** Model coefficients and SHAP values translate into specific HR levers — career mobility, compensation reviews, workload audits.
* **Bias Mitigation:** Class weighting ensures equal attention to both classes, avoiding skewed recommendations.
* **Expected Impact:** If embedded into HR workflows, the model can help achieve a 5–10% reduction in voluntary attrition.

**Limitation:** While the model does not predict *when* an employee will leave or capture sudden external market changes, it is a practical, explainable tool to guide proactive retention strategies.

## 23.0 Recommended Next Steps for HR Leadership

These priority actions are directly tied to the top predictive drivers identified through SHAP values and odds ratios in the logistic regression model:

1. **Address Role Stagnation (Low RoleTenureRatio)**
   * **Why:** Employees new to their role or moved late in tenure are more likely to leave.
   * **Action:** Flag employees with RoleTenureRatio < 0.3 for early engagement, mentorship, and role-fit check-ins in the first 6–12 months.
2. **Target Pay Equity Gaps (Low MonthlyIncome / CompaRatio)**
   * **Why:** Under-market pay is a significant attrition driver (OR = 0.41).
   * **Action:** Review bottom quartile earners, especially those with CompaRatio < 75%; apply targeted salary adjustments or bonuses.
3. **Reduce Overtime-Driven Burnout (OverTime\_Yes)**
   * **Why:** Persistent overtime doubles attrition risk (OR = 2.09).
   * **Action:** Audit overtime-heavy teams, set workload caps, and track overtime risk alerts in HR dashboards.
4. **Strengthen Engagement Levers**
   * **Why:** High Job Involvement, Job Satisfaction, and Environment Satisfaction are protective factors.
   * **Action:** Recognize high scorers; monitor drops in survey scores as early warning signals; hold managers accountable for engagement outcomes.
5. **Tailor Retention for Sales & High-Travel Roles**
   * **Why:** Sales roles and frequent travelers face higher attrition, though causes differ.
   * **Action:**
     + **Sales:** Review and benchmark incentive programs.
     + **Non-Sales Frequent Travelers:** Set travel balance guidelines, offer wellness support, and flexibility options.
6. **Recommendation X: Improve Performance Rating Scale for Better Talent Insights**
   * **Why:** In the current dataset, the performance rating field only contains two values — 3 (Average) and 4 (High). This lack of granularity reduces the model’s ability to detect meaningful differences between employees.
   * **Action:**
     + Update the Performance Management System to capture ratings on a 0–5 numeric scale instead of broad categories.
     + Define clear performance criteria for each numeric level to ensure consistency across departments.
     + Integrate the revised scale into future attrition models to enable:
     + Identification of top-tier performers at risk
     + More targeted retention incentives for high-value talent

# 24.0 References

IBM HR Analytics Employee Attrition & Performance Dataset  
Source: IBM Watson Analytics (via Kaggle)  
URL: <https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset>

Radford Global Compensation Database  
Aon. (Accessed 2024). Radford Global Compensation Database – Market salary benchmarks used to derive CompaRatio.  
URL: <https://rewards.aon.com/radford>

SHRM (Nov 2024). Future of Talent Retention – Why Employees Leave — Highlights toxic culture, leadership issues, and work-life imbalance as top attrition drivers (pay ranks lower).  
URL: <https://www.shrm.org/executive-network/insights/future-of-talent-retention-report-why-employees-leave>

SHRM (Dec 2024). Workplace Culture Fosters Employee Retention Worldwide — Finds that strong, unbiased, and supportive cultures—with honest management—correlate with higher retention.  
URL: <https://www.shrm.org/executive-network/insights/shrm-report-workplace-culture-fosters-employee-retention>

WorkInstitute (2025). 2025 Retention Report – Employee Retention Truths — Calls career development, work-life balance, and managerial effectiveness key retention levers.  
URL: <https://info.workinstitute.com/hubfs/2025%20Retention%20Report/2025%20Retention%20Report-%20Employee%20Retention%20Truths%20in%20Todays%20Workplace.pdf>

Gallup (May 2025). State of the Global Workplace 2025 — Reports global employee engagement fell to 21%, with manager engagement dropping from 30% to 27%; stresses importance of manager training in retention.  
URL: <https://www.gallup.com/workplace/349484/state-of-the-global-workplace.aspx>

FT (May 2025). What Makes a Positive Difference to the Way Staff Feel About Work? — Cites Gallup’s finding that well-trained managers influence 70% of team engagement; links weak manager skills to higher attrition.  
URL: [https://www.ft.com/content/9e53def2‑d5b2‑4d89‑b385‑15df25c3b539](https://www.ft.com/content/9e53def2-d5b2-4d89-b385-15df25c3b539)