Final Report

1. Executive Summary

In an increasingly competitive business landscape, retaining top talent is as crucial as acquiring it. Voluntary employee attrition disrupts operations, increases costs, and undermines workforce stability. This project addresses that challenge by developing predictive models to identify high-risk employees and uncover the drivers of attrition.

Using the IBM HR Analytics dataset, we framed attrition prediction as a binary classification problem. Our analysis integrated two modeling techniques — **Logistic Regression, Random Forest**— with robust data preprocessing and class imbalance handling using SMOTE. Among the models, logistic regression achieved the highest **ROC AUC (~0.87)** and **F1 Score (~0.85)**, while tree-based models added depth through nonlinear feature interactions.

Across all models, key predictors included **overtime, compensation metrics, job role, and employee experience.** These insights form the foundation for strategic HR interventions such as workload management, compensation realignment, and targeted retention programs.

By combining model performance with interpretability and business relevance, this solution empowers HR leadership to proactively reduce preventable attrition, improve employee satisfaction, and support long-term organizational growth.

2. Business Understanding and Objective

2.1 Business Problem:

Employee turnover is not just a human resources challenge — it's a strategic risk. For companies like IBM, voluntary attrition leads to increased recruitment costs, knowledge loss, and reduced productivity. Early identification of at-risk employees can help prevent these outcomes through timely, targeted interventions.

2.2 Business Question:

Which employees are likely to voluntarily leave the company, and what are the key factors that contribute to their decision?

2.3 Objective:

This project aims to:

1. Build a predictive model that classifies employee attrition risk using available HR data.

2. Identify key features driving employee exits.
3. Optimize model sensitivity to ensure early and accurate detection of potential leavers.
4. Translate insights into actionable HR strategies.

2.4 Success Metric:

1. **Primary Metric: Sensitivity (True Positive Rate)** – critical for flagging most potential attrition cases.
2. **Secondary Metrics: ROC AUC, Specificity, Accuracy, and F1 Score** – to evaluate overall model reliability and precision.
3. **Business Success Indicator**: Reduction in preventable attrition through data-driven HR actions.

3. Data Understanding & Visualization

3.1 Dataset Overview

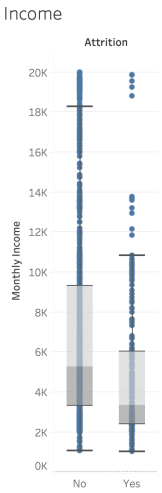
The dataset used is the IBM HR Analytics Employee Attrition dataset, containing 1,470 employee records with 36 features, including demographic, professional, and job performance attributes. The target variable is Attrition, indicating whether an employee voluntarily left the organization.

3.2 Features

To identify the factors that influence attrition (whether an employee leaves the company), we leverage several charts and a comprehensive Tableau dashboard featuring several intuitive visualizations.

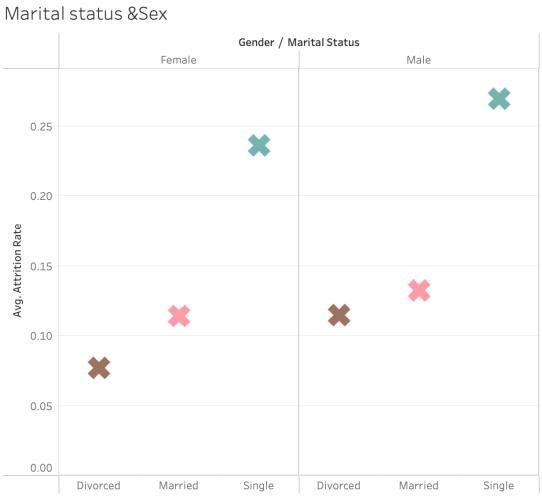
3.2.1 Monthly Income

A box-and-whisker plot of monthly income reveals that employees who departed the company earn less on average than those who remain. Specifically, the median, lower quartile (Q1), and upper quartile (Q3) for the “resigned” group all fall below the corresponding values for current employees. This suggests a clear positive association between higher income and employee retention.



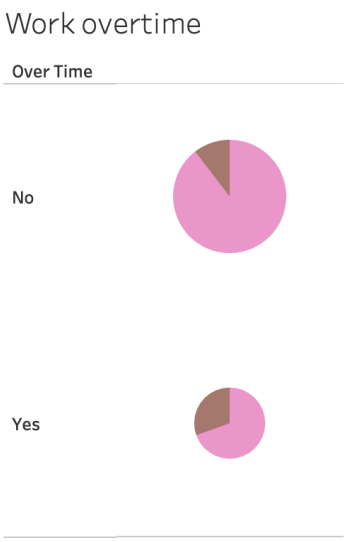
3.2.2 Marital Status & Gender

The combined gender-marital status analysis shows that male employees have a marginally higher attrition rate than female employees. More strikingly, marital status exerts a pronounced effect on turnover: divorced employees exhibit the lowest attrition, followed by married employees, while single employees display the highest attrition rate. This pattern may reflect the stabilizing influence of family or spousal support, which can discourage job changes.



3.2.3 Work overtime

A pie chart comparing overtime status indicates that employees who work overtime are significantly more likely to leave. The attrition rate among those regularly working beyond standard hours is markedly higher, suggesting that excessive overtime is a strong predictor of turnover.



3.2.4 Other factors

In addition to the variables already discussed, we examined several other features in the dashboard. As shown, employee attrition is significantly influenced by travel frequency, overtime hours, income level, job role, job involvement, and marital status. Each of these factors exhibits a clear association with the likelihood of an employee's resignation.

4. Data Preparation

4.1 Data Cleaning & Preprocessing

1. No missing values or duplicates were found.

2. Low-variance columns (Over18, StandardHours, EmployeeCount) were removed.
3. Identifier column (EmployeeNumber) was excluded, as it carried no predictive value.
4. Highly correlated features were reviewed to mitigate multicollinearity (e.g., YearsAtCompany dropped due to overlap with related tenure features)

4.2 Encoding Categorical Variables

1. Binary Encoding: e.g., Attrition, OverTime → 1 for "Yes", 0 for "No".
2. Ordinal Encoding: e.g., BusinessTravel ranked based on frequency.
3. One-Hot Encoding: Applied to nominal features like JobRole, Department, MaritalStatus.

4.3 Feature Scaling & Transformation

1. **Log Transformation:** Applied to skewed features (e.g., MonthlyIncome, TotalWorkingYears) using log(x + 1).
2. **Z-score Normalization:** Standardized numerical features to ensure uniform scale across models.

4.4 Label Distribution

The target variable Attrition is moderately imbalanced, with approximately 84% of records labeled as "No" (1,233 instances) and 16% labeled as "Yes" (237 instances).

Given the moderate imbalance, no resampling was applied during preprocessing.

5. Modeling and Results

5.1 Model Selection

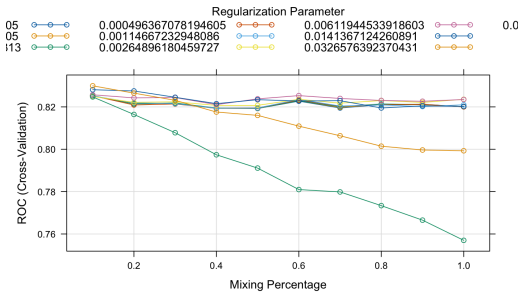
To capture both linear and nonlinear relationships, we applied two supervised classification models:

1. Logistic Regression (with L1/L2 regularization for interpretability and simplicity)
2. Random Forest (robust to noise and captures feature interactions)

5.2 Hyperparameter Tuning

Although the class distribution was moderately imbalanced (≈ 84% No, 16% Yes), we tested model performance both:

1. **Logistic Regression**: Tuned alpha = 0.1 and lambda ≈ 0.0327



2. **Random Forest**: ntrees = 100, k-fold value = 10, nodesize = 3

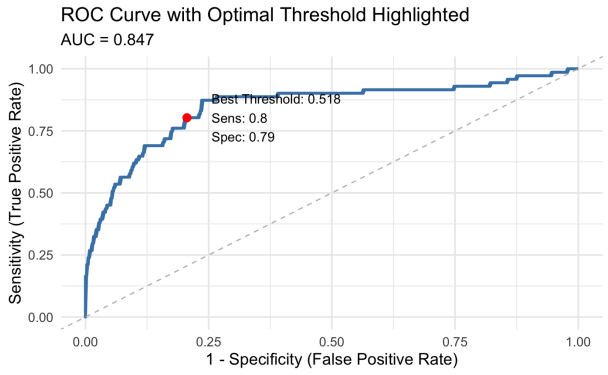
5.3 Class Imbalance Strategy

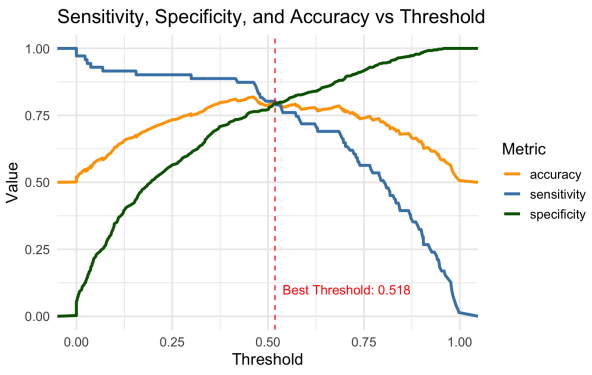
Used SMOTE during cross-validation to balance the dataset and avoid bias toward the majority class.

5.4 Threshold Optimization

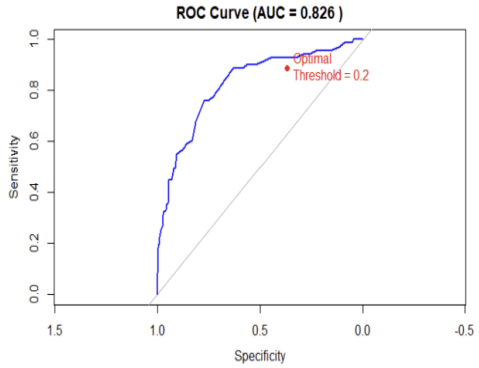
We prioritize sensitivity>0.8 and maximize the specificity to get the optimized threshold.

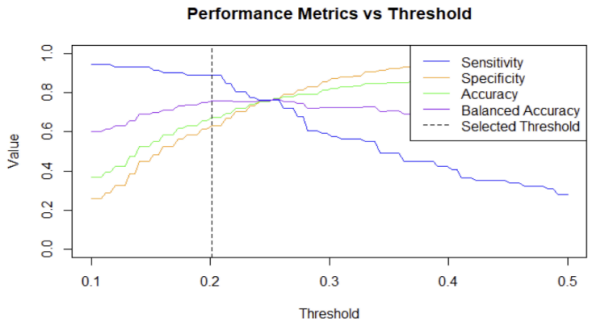
Logistic Regression: Threshold 0.518





Random Forest: Threshold 0.20





5.5 Performance Summary

	Model	ROC AUC	Optimized Threshold	Sensitivity	Specificity	Accuracy	F1 Score

1	Logistic Regression	0.849	0.52	80%	79%	79%	0.847
2	Random Forest	0.826	0.20	88.7%	63%	67%	0.467

5.7 Key Findings

1. Basic Random Forest vs CV+SMOTE RF
    a. ntrees, nodesize, class_weights parameters do not have significant impact on performance metrics. Only the cut-off(threshold) leads to apparent change in result but will cause overfitting problems.
    b. showing similar performance metrics in terms of statistics like Sensitivity, AUC, Specificity and Accuracy.
    c. Basic RF model shows significant overfitting problems (23.9% Difference between train_data and test_data) and The probability distribution of the votes is skewed to the right. It means the model's overconfidence in the prediction given its overfitting.
    d. RF with CV+SMOTE dramatically reduced the overfitting problems (down to 0.1%) and the probability distribution is bimodal distribution, showing an appropriate uncertainty for ambiguous cases. It indicate this model
    e. Conclusion: The CV+SMOTE Random Forest model provides a significantly more reliable foundation for attrition prediction than the basic Random Forest model, despite showing similar surface-level metrics.
2. Logistic Regression Model
    a. We built a basic Logistic Regression Model with 10-fold cross validation training, gained the roc-auc **0.8185**, in-sample accuracy 0.797, out-sample accuracy 0.775. This showed that the basic logistic model performed quite well and was not overfitting.
    b. We optimized the model with L1/L2 regularization for interpretability and simplicity, got the Tuned alpha = 0.1 and lambda ≈ 0.0327, and improved the roc-auc to **0.8493**, and gained accuracy 0.777
    c. We prioritize sensitivity>0.8 and maximize the specificity to get the optimized threshold as 0.52

6. Feature Importance and Predictive Insights

6.1 Common Drivers Across Models

Across logistic regression, random forest, several key factors consistently emerged as strong predictors of attrition:

1. **OverTime**: Most influential across both models; linked to burnout.
2. **Compensation Metrics**: MonthlyIncome, HourlyRate, and StockOptionLevel indicated dissatisfaction or market-driven exits.
3. **JobRole & Department**: Certain roles (e.g., Sales, Research) showed higher attrition risk.

4. **MaritalStatus (Single)**: Single employees tended to leave more frequently.
5. **DistanceFromHome**: Long commutes were associated with higher attrition.
6. **JobInvolvement:** Lower engagement increased departure likelihood.
7. **TotalWorkingYears & NumCompaniesWorked**: Career stage and prior job mobility also played a role.

6.2 Logistic Regression Insights

```
Call:
glm(formula = Attrition ~ ., family = "binomial", data = train_data)

Coefficients: (4 not defined because of singularities)
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  3.624e+00  1.818e+00   1.993 0.046292 *
Age                         -7.278e-03  1.649e-02  -0.441 0.659037
BusinessTravel               1.192e+00  2.170e-01   5.494 3.93e-08 ***
DailyRate                   -9.679e-02  1.070e-01  -0.905 0.365647
DistanceFromHome             4.363e-01  1.143e-01   3.818 0.000135 ***
Education                   -1.772e-02  1.082e-01  -0.164 0.869888
EnvironmentSatisfaction     -4.665e-01  1.012e-01  -4.608 4.06e-06 ***
Gender                       5.277e-01  2.279e-01   2.315 0.020590 *
HourlyRate                   3.502e-02  1.086e-01   0.322 0.747095
JobInvolvement              -4.911e-01  1.516e-01  -3.240 0.001195 **
JobSatisfaction             -4.085e-01  9.834e-02  -4.154 3.26e-05 ***
MonthlyIncome               -1.066e-01  2.484e-01  -0.429 0.668018
MonthlyRate                 -5.250e-02  1.079e-01  -0.487 0.626444
NumCompaniesWorked           3.944e-01  1.288e-01   3.062 0.002197 **
OverTime                     1.922e+00  2.355e-01   8.162 3.28e-16 ***
PercentSalaryHike           -1.189e-01  1.607e-01  -0.739 0.459609
PerformanceRating            2.989e-03  4.494e-01   0.007 0.994693
RelationshipSatisfaction    -2.899e-01  1.062e-01  -2.730 0.006326 **
StockOptionLevel            -1.963e-01  1.875e-01  -1.047 0.295231
TotalWorkingYears           -5.522e-01  2.090e-01  -2.643 0.008228 **
TrainingTimesLastYear       -2.363e-01  1.126e-01  -2.098 0.035938 *
WorkLifeBalance             -4.198e-01  1.536e-01  -2.732 0.006287 **
YearsInCurrentRole          -4.096e-01  1.691e-01  -2.422 0.015422 *
YearsSinceLastPromotion      5.172e-01  1.461e-01   3.541 0.000399 ***
```

	Overall <dbl>
OverTime	100.00000
BusinessTravel	65.19800
EnvironmentSatisfaction	59.70571
MaritalStatusMarried	45.98821
MaritalStatusDivorced	42.56741
YearsSinceLastPromotion	41.20357
DistanceFromHome	41.11202
EducationFieldLife.Sciences	40.59151
JobSatisfaction	39.98713
NumCompaniesWorked	37.31926

We used glm in R to see whether the logistic coefficients were positive or not and to see whether variables were significant. And we used varimp() to output the variables importance

Feature coefficients showed the following top predictors:

1. OverTime (strongest positive influence)
2. BusinessTravel
3. MaritalStatus
4. Promotion
5. distance from home

6.3 Random Forest Model Insights

Importance score showed the following top features:

1. Overtime (Highest important score)
2. MartialStatusSingle
3. StockOptionLevel
4. MonthlyIncome

7. Business Recommendations

7.1 Compensation Strategy

1. **Review pay structures** to ensure market competitiveness, especially for high-risk roles (Sales, R&D).
2. **Enhance stock option plans** and performance-based incentives, addressing factors like StockOptionLevel and MonthlyIncome.
3. Ensure **fair, regular salary hikes** to reward tenure and reduce dissatisfaction.

7.2 Work-Life Balance Initiatives

1. **Reduce excessive overtime** through resource balancing and workload audits.

2. **Support remote or hybrid work** options for employees with long commutes.
3. Reassess **business travel requirements**, offering flexibility for frequent travelers.

7.3 Career Development and Engagement

1. Introduce **career advancement paths** tailored to employees with stagnating tenure (TotalWorkingYears, YearsAtCompany).
2. Foster **employee engagemen**t through personalized development plans and recognition programs.
3. Improve JobInvolvement via better role alignment and manager-employee fit.

7.4 Targeted Retention Programs

1. Develop **support strategies for single employees**, who show higher attrition trends.
2. Create **role-specific interventions** (e.g., Sales & Research teams), such as mentoring and learning initiatives.
3. Focus on **mid-career employees** at risk of leaving for better opportunities.

8. Implementation Plan

8.1 Short-Term Actions (0–3 Months)

1. Conduct compensation benchmarking to address salary gaps.
2. Implement overtime tracking tools and initiate workload audits.
3. Launch quick-pulse surveys to gauge job satisfaction and involvement.

8.2 Medium-Term Actions (3–9 Months)

1. Develop career development tracks for high-risk roles and tenure groups.
2. Revise stock option/incentive structures for better alignment with retention goals.
3. Introduce flexible work arrangements, especially for employees with long commutes or heavy travel.

8.3 Long-Term Actions (9–18 Months)

1. Integrate predictive models into HR dashboards for quarterly attrition risk monitoring.
2. Establish continuous feedback loops for employee experience and environment improvement.
3. Refine model thresholds and features using live HR data and attrition outcomes

9. Deployment and Ethical Considerations

9.1 Deployment Plan

1. **Embed predictive models** into IBM's internal HR analytics platforms for quarterly evaluations.
2. **Display attrition risk scores** in HR dashboards with key drivers for each case.
3. Integrate into **manager workflows** for proactive intervention planning.

9.2 Ethical Risks

1. **False Positives**: Flagging loyal employees may cause mistrust or unnecessary HR action.
2. **Bias in Features**: Attributes like marital status or age may encode systemic biases.

3. **Morale Impact**: Employees labeled as “at risk” may feel unfairly targeted or demotivated.

9.3 Mitigation Strategies

1. **Human-in-the-loop**: Use model outputs as decision support, not automated judgments.
2. **Segmented Thresholding**: Calibrate thresholds by department or role to reduce overgeneralization.
3. **Explainability**: Leverage SHAP/LIME to make predictions interpretable for HR teams.
4. **Bias Audits**: Regularly evaluate performance across age, gender, and other protected groups.
5. **Ethical Framing**: Present predictions as risk probabilities, not fixed conclusions.

10. Limitations and Future Work

10.1 Current Limitations

1. **Synthetic Dataset**: The IBM dataset is publicly available and likely anonymized, limiting real-world generalizability.
2. **Cross-sectional Snapshot**: Data captures a single point in time, missing temporal dynamics of attrition.
3. **Manual Feature Engineering**: Preprocessing steps were not automated for production-scale deployment.
4. **Residual Class Imbalance**: Even with SMOTE, false positives remain a challenge in real scenarios.
5. **Model Simplicity**: Logistic regression, while interpretable, may not capture all nonlinear interactions.

10.2 Future Enhancements

1. **Ensemble Models:** Experiment with XGBoost, LightGBM, or stacked ensembles for improved performance.
2. **Time-Series Modeling:** Track employee behavior over time for dynamic risk scoring.
3. **Explainable AI (XAI):** Integrate SHAP/LIME-based dashboards for HR transparency.
4. **Bias Mitigation**: Formalize fairness evaluations across protected attributes.
5. **Real-Time Integration:** Build end-to-end pipelines that connect live HR databases with automated model inference and alerts.