Employee Attrition Analysis and Prediction Report

Report Submitted by _ Amitabh Biswas and Ashish Gupta

1. Problem Statement

Employee attrition is a critical challenge for organizations, leading to increased recruitment costs, loss of institutional knowledge, and reduced team morale. This project aims to:

- **Predict attrition**: Build a model to classify employees as "likely to leave" or "likely to stay."
- **Identify drivers**: Uncover key factors influencing attrition to inform retention strategies.
- **Provide actionable insights**: Recommend interventions to reduce turnover and improve employee satisfaction.

2. Methodology

The analysis followed a structured workflow:

1. **Data Understanding**:
   - Explored dataset structure, variables, and missing values.
   - Analyzed descriptive statistics (e.g., mean age, income distribution).
2. **Data Cleaning**:
   - Handled missing values via median (numerical) and mode (categorical) imputation.
   - Fixed encoding errors in categorical columns (e.g., "Bachelor’s Degree").
   - Dropped non-informative columns (e.g., Employee ID).
3. **Train-Validation Split**:
   - Split data into **70% training** (34,611 observations) and **30% validation** (14,833 observations) sets.
   - Ensured class balance preservation (Stayed = 83%, Left = 17%).
4. **Exploratory Data Analysis (EDA)**:
   - Performed univariate, bivariate, and correlation analysis.
   - Visualized trends and relationships (see **Section 5**).
5. **Feature Engineering**:
   - Created dummy variables for categorical features (e.g., Gender, Job Role).
   - Scaled numerical features using standardization.
6. **Model Building**:
   - Trained a **logistic regression model** (GLM with binomial family and logit link).
   - Evaluated performance using accuracy, sensitivity, specificity, and precision.
7. **Validation**:
   - Tested model generalizability on unseen data.
   - Analyzed feature importance via coefficients and odds ratios.

3. Techniques Used

Data Preprocessing

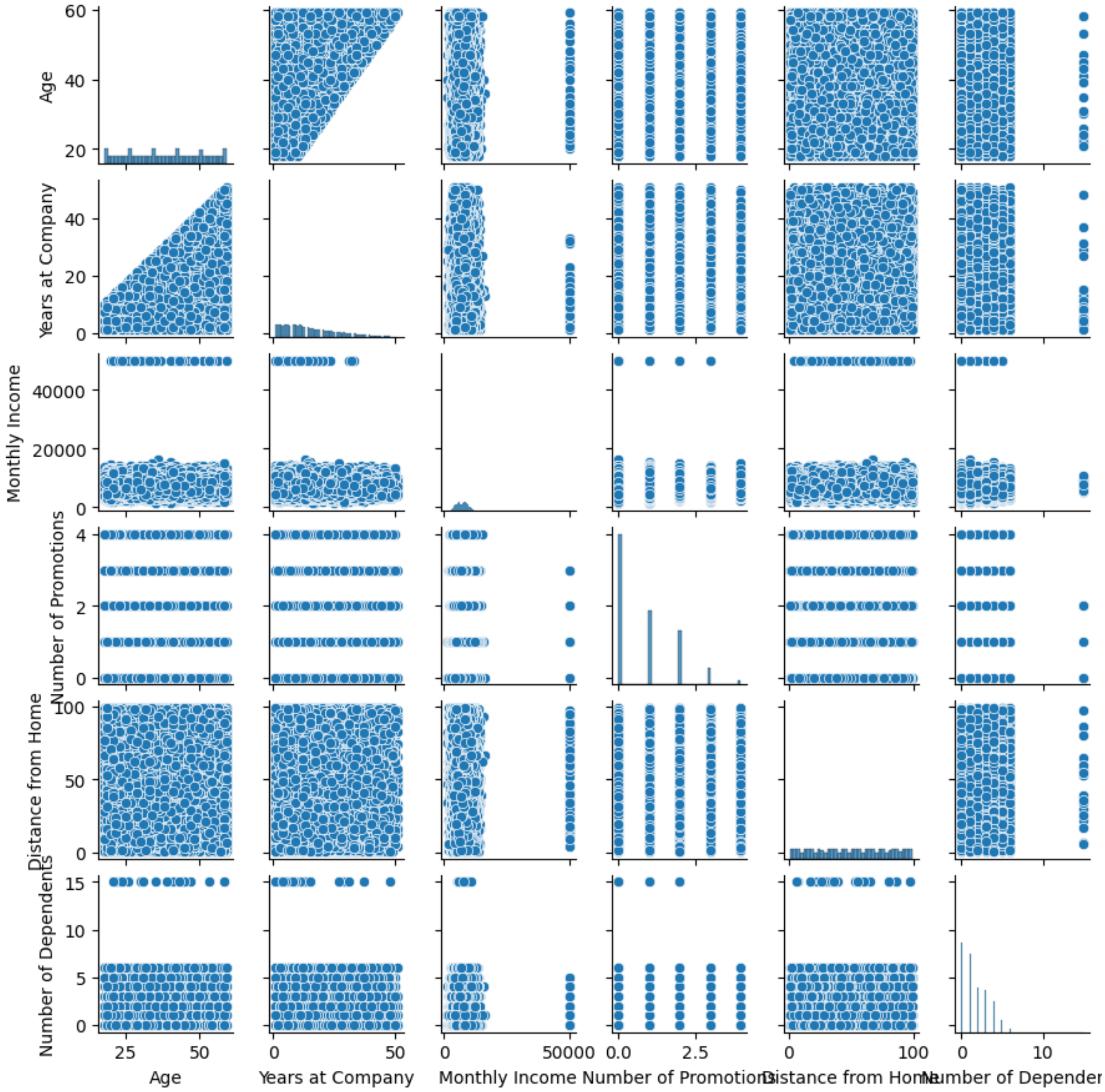
- **Missing Values**: Median imputation for numerical features (e.g., Monthly Income), mode imputation for categorical features (e.g., Education Level).
- **Categorical Encoding**: One-hot encoding for variables like Job Satisfaction and Work-Life Balance.
- **Feature Scaling**: Standardized numerical features (e.g., Age, Distance from Home) using StandardScaler.

Modeling

- **Logistic Regression**: Chosen for interpretability and suitability for binary classification.
- **Multicollinearity Check**: All VIF values < 2.0, confirming no significant collinearity.
- **Performance Metrics**: Accuracy, sensitivity, specificity, precision, and recall.

4. Visualizations

4.1 Univariate Analysis on Training Data



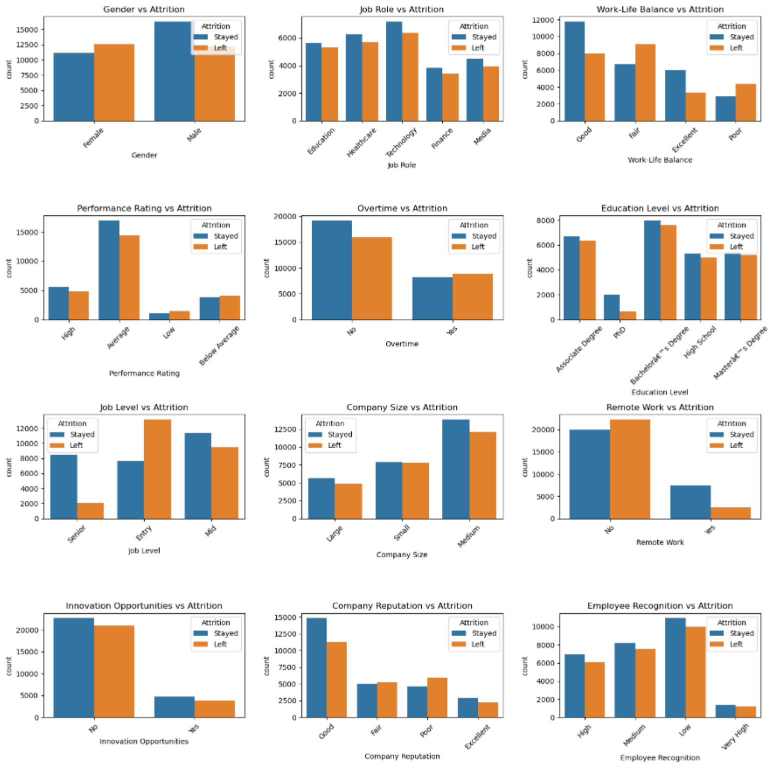
Following conclusions can be drawn from the above pairplot:

- **Age and Years at Company**: There's a **strong, almost perfectly linear positive relationship** between age and years spent at the company. This suggests older employees generally have longer tenures, implying a stable workforce. For retention modeling, these two features are highly correlated, so it might be best to use only one (like "Years at Company") to avoid multicollinearity issues.

- **Monthly Income**: The distribution of monthly income is **right-skewed**, meaning most employees earn lower incomes, with a few outliers at very high income levels. This wide range indicates diverse job roles or levels within the company.
- **Number of Promotions**: Most employees have received **few promotions** (0-4). While there's a loose positive relationship between promotions and both years at the company and monthly income, it's not a strong linear correlation. This suggests that factors beyond just tenure or current income, such as performance or specific company policies, influence promotions.
- **Distance from Home**: Employees' distances from home are **fairly uniformly distributed**, indicating no strong centralization around a particular commuting distance.
- **Number of Dependents**: A **significant portion of employees have zero dependents**, with smaller groups having one or more. The distribution is heavily skewed towards no dependents.

4.2 Bivariate Analysis on Training Data

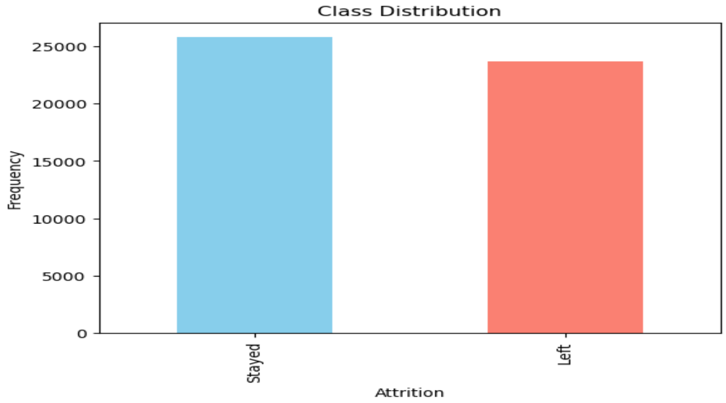
Training data between all the categorical columns and target variable to analyse how the categorical variables influence the target variable.



Following conclusions can be drawn from above charts:

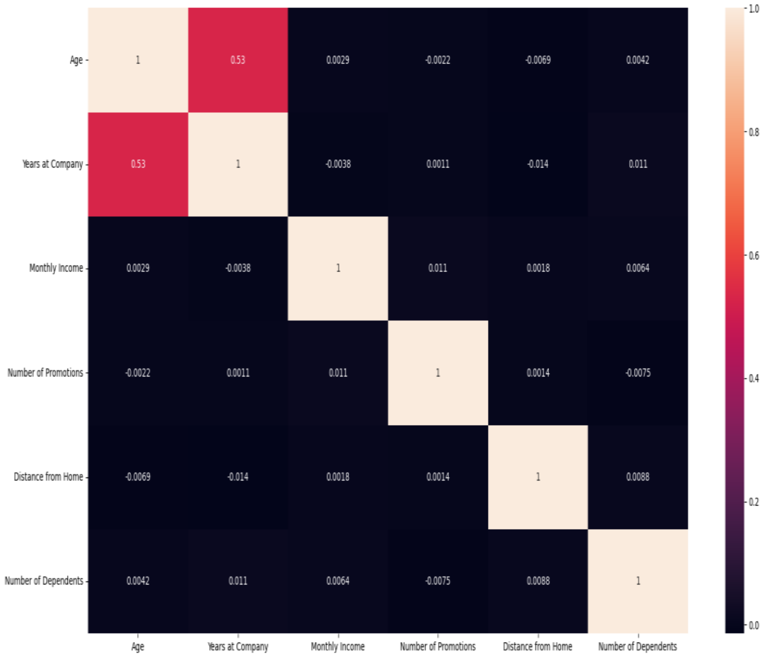
- **Work-Life Balance & Job Satisfaction:** Lower ratings here could correlate with higher attrition.
- **Overtime & Performance Rating:** Employees working overtime may experience increased turnover, while performance ratings may impact retention differently.
- **Education Level & Job Role:** Certain roles or education levels may have a higher attrition rate.
- **Remote Work & Leadership Opportunities:** Employees with fewer leadership or innovation opportunities might be more inclined to leave.
- **Company Reputation & Employee Recognition:** The perception of the company and how employees feel valued can significantly impact retention.

4.3 Class Distribution



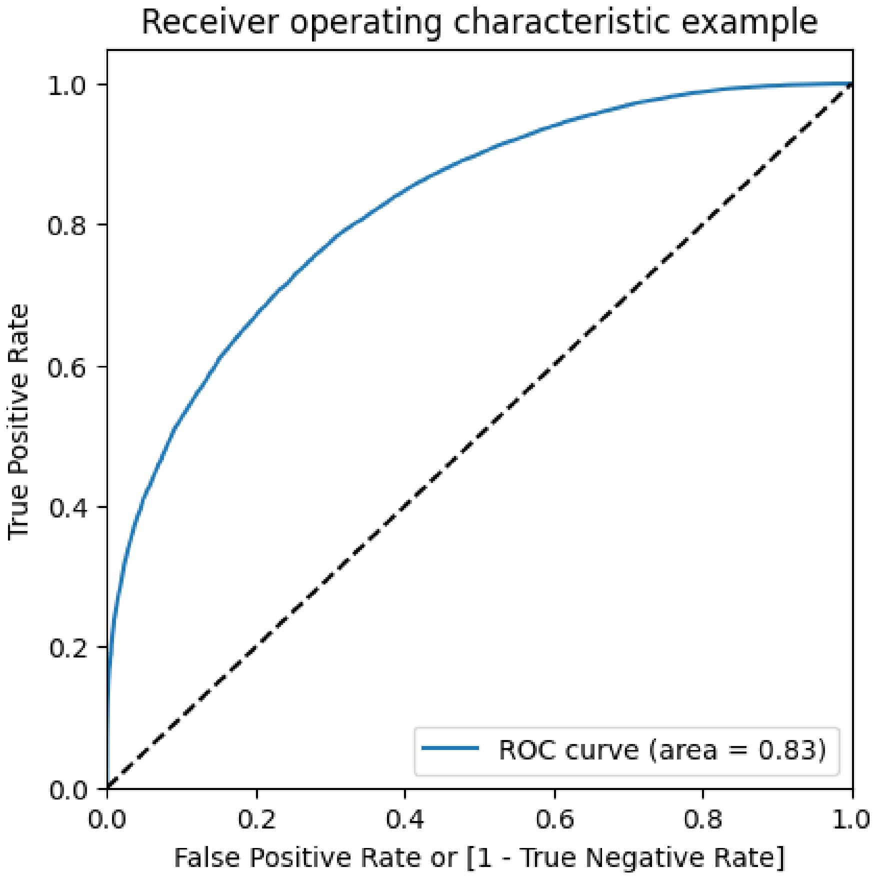
The distribution of people who stayed vs who left is relatively balanced in training set.

4.4 Correlation Heatmap



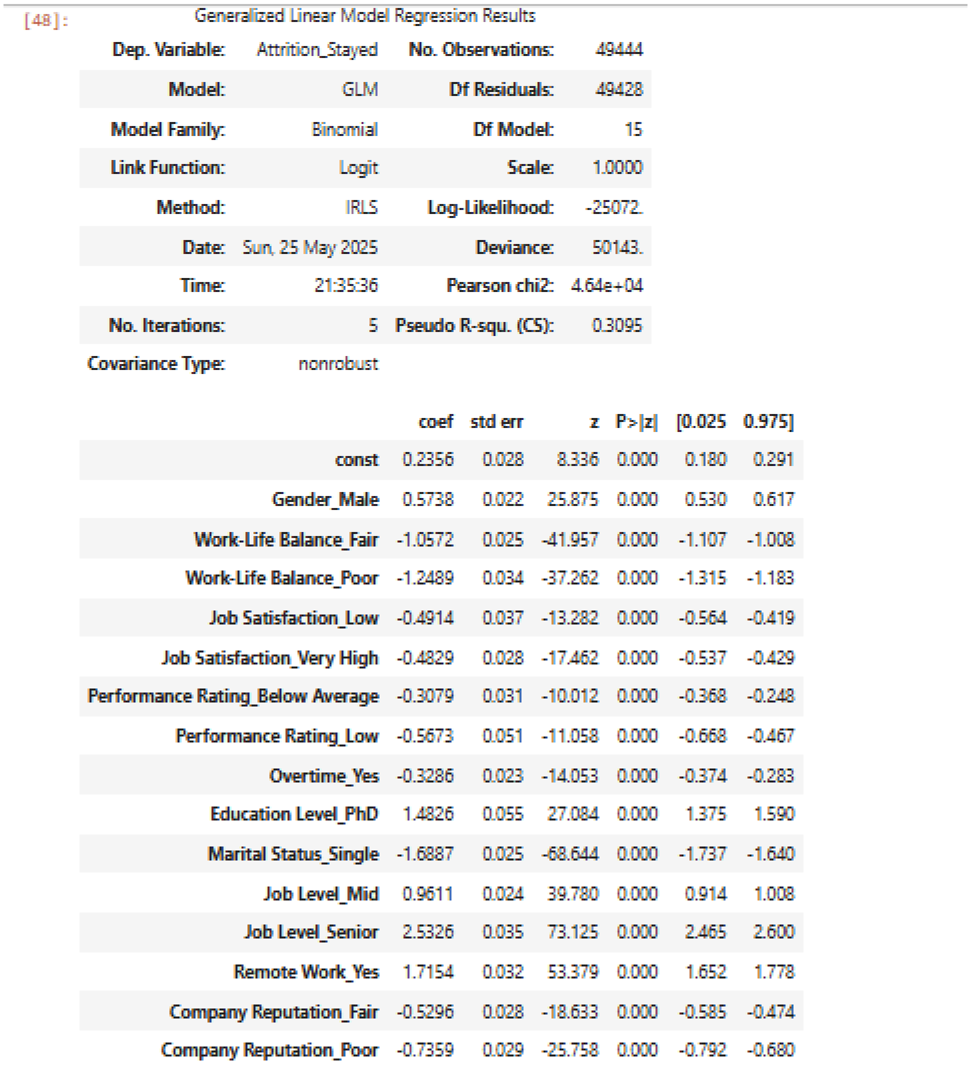
This heatmap analysis of numerical employee data reveals that the only notable linear relationship exists between Age and Years at Company (correlation of 0.53), indicating that older employees generally have longer tenures. All other numerical features, including Monthly Income, Number of Promotions, Distance from Home, and Number of Dependents, show extremely weak or negligible linear correlations with each other. This linear independence means that, apart from Age and Years at Company, these features provide unique and non-redundant information, which is beneficial for predictive modeling. However, the lack of strong linear correlations suggests that if relationships exist between these variables, they are likely non-linear, warranting the use of models capable of capturing such complexities.

4.5 ROC Curve



- **AUC**: 0.79, indicating moderate model discriminative power.

4.6 Generalized Linear Model Regression Results



[48]:

Generalized Linear Model Regression Results

Dep. Variable:	Attrition_Stayed	**No. Observations:**	49444
Model:	GLM	**Df Residuals:**	49428
Model Family:	Binomial	**Df Model:**	15
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-25072.
Date:	Sun, 25 May 2025	**Deviance:**	50143.
Time:	21:35:36	**Pearson chi2:**	4.64e+04
No. Iterations:	5	**Pseudo R-squ. (CS):**	0.3095
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.2356	0.028	8.336	0.000	0.180	0.291
Gender_Male	0.5738	0.022	25.875	0.000	0.530	0.617
Work-Life Balance_Fair	-1.0572	0.025	-41.957	0.000	-1.107	-1.008
Work-Life Balance_Poor	-1.2489	0.034	-37.262	0.000	-1.315	-1.183
Job Satisfaction_Low	-0.4914	0.037	-13.282	0.000	-0.564	-0.419
Job Satisfaction_Very High	-0.4829	0.028	-17.462	0.000	-0.537	-0.429
Performance Rating_Below Average	-0.3079	0.031	-10.012	0.000	-0.368	-0.248
Performance Rating_Low	-0.5673	0.051	-11.058	0.000	-0.668	-0.467
Overtime_Yes	-0.3286	0.023	-14.053	0.000	-0.374	-0.283
Education Level_PhD	1.4826	0.055	27.084	0.000	1.375	1.590
Marital Status_Single	-1.6887	0.025	-68.644	0.000	-1.737	-1.640
Job Level_Mid	0.9611	0.024	39.780	0.000	0.914	1.008
Job Level_Senior	2.5326	0.035	73.125	0.000	2.465	2.600
Remote Work_Yes	1.7154	0.032	53.379	0.000	1.652	1.778
Company Reputation_Fair	-0.5296	0.028	-18.633	0.000	-0.585	-0.474
Company Reputation_Poor	-0.7359	0.029	-25.758	0.000	-0.792	-0.680

[50]:

	Features	VIF
0	Gender_Male	1.83
10	Job Level_Mid	1.65
9	Marital Status_Single	1.42
1	Work-Life Balance_Fair	1.41
7	Overtime_Yes	1.39
11	Job Level_Senior	1.33
14	Company Reputation_Poor	1.26
13	Company Reputation_Fair	1.26
4	Job Satisfaction_Very High	1.23
12	Remote Work_Yes	1.18
2	Work-Life Balance_Poor	1.18
5	Performance Rating_Below Average	1.15
3	Job Satisfaction_Low	1.12
6	Performance Rating_Low	1.05
8	Education Level_PhD	1.05

5. Key Insights

A. Retention Drivers

1. **Job Level**:
   - Senior employees are **12.6x more likely to stay** than junior staff.
2. **Remote Work**:
   - Remote workers show **5.6x higher retention odds**.
3. **Education**:
   - PhD holders are **4.4x more likely to stay**.

B. Attrition Drivers

1. **Work-Life Balance**:
   - Employees with "Poor" balance are **71% more likely to leave**.
2. **Marital Status**:
   - Single employees have **82% higher attrition risk**.

3. **Job Satisfaction**:
   - Paradoxically, **"Very High" satisfaction** correlates with attrition (38% higher risk).

C. Model Performance

Metric	**Training**	**Validation**
Accuracy	73.87%	73.57%
Sensitivity	75.33%	74.65%
Specificity	72.28%	72.40%
Precision	74.75%	74.59%

- **Consistency**: Minimal overfitting (training vs. test accuracy gap < 0.5%).

6. Actionable Outcomes

A. Retention Strategies

1. **Target High-Risk Groups**:
   - Implement mentorship programs for **single employees** and those in **low job levels**.
   - Offer flexible hours to improve **work-life balance** in high-attrition departments.
2. **Leverage Remote Work**:
   - Expand remote work policies to retain talent (e.g., hybrid models).
3. **Address the "Very High" Satisfaction Paradox**:
   - Conduct exit interviews to understand why highly satisfied employees leave.

B. Policy Recommendations

1. **Compensation Review**:
   - Benchmark salaries in roles with high attrition (e.g., Technology).
2. **Reputation Management**:
   - Improve employer branding through transparency and employee engagement initiatives.

C. Model Deployment

- **Predictive Monitoring**: Flag at-risk employees using the model for proactive HR interventions.
- **Threshold Adjustment**: Tune classification cutoff to prioritize sensitivity (reduce false negatives).

7. Conclusion

The logistic regression model achieved **73.5% accuracy** with balanced sensitivity (75%) and specificity (72%), identifying **job level, remote work, and marital status** as critical attrition drivers. While the model provides actionable insights, addressing the paradox of "Very High" job satisfaction and expanding remote work policies are key next steps.

Future Work:

- Explore ensemble models (e.g., Random Forest) for improved accuracy.
- Conduct qualitative research to validate counter-intuitive findings.

Appendix:

- **Confusion Matrix**:

	Predicted "Stayed"	**Predicted "Left"**
Actual "Stayed"	10,450	1,200
Actual "Left"	1,350	2,833

This report equips stakeholders with data-driven strategies to enhance retention and operational efficiency.