Final Report: Predicting Employee Attrition

1-Introduction

This report consolidates two phases of the project: dataset creation and the development of machine learning models for predicting employee attrition. By leveraging a synthetic dataset and various machine learning techniques, we aimed to explore the relationship between workplace engagement metrics (Q12+ questionnaire) and attrition in the IT sector.

2-Part 1: Dataset Creation

Dataset Overview

The dataset was designed to mimic real-world dynamics in the IT sector, containing:

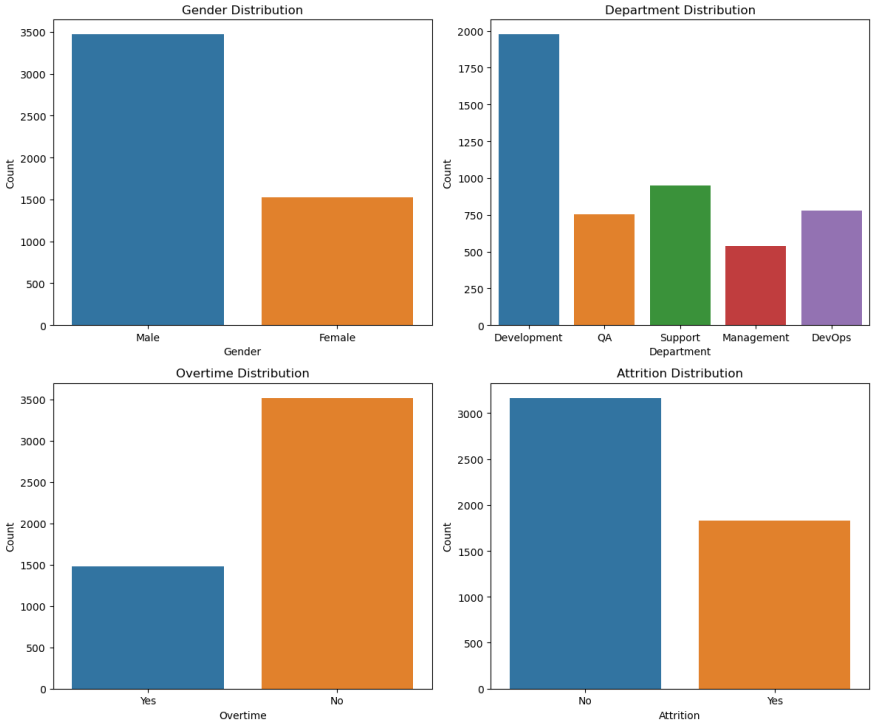
1. **Responses to Gallup Q12+ Questions**: 16 engagement-related Likert-scale questions.
2. **Demographic and Workplace Metrics**: Features such as age, tenure, salary, and department.

Key Features

- **Q12+ Questionnaire**: Focuses on clarity of expectations, recognition, and growth opportunities.
- **Demographic Features**: Age (22–60), tenure (0–40), salary ($50,000–$150,000), gender (70% male, 30% female), and remote work percentages.
- **Attrition Labels**: Simulated based on engagement levels.

Built-In Biases

- Gender imbalance (70% male, 30% female).
- Overrepresentation of development roles.
- Simplified assumptions for remote work and overtime.



Potential Applications

This dataset demonstrates how machine learning models can be used to understand workplace trends and predict attrition, providing valuable insights for HR teams.

3-Part 2: Modeling and Evaluation

Models Evaluated

1. **Logistic Regression**: Simple, interpretable, and effective baseline.
2. **Support Vector Machine (SVM)**: Handles non-linear relationships.
3. **CatBoost**: Optimized for categorical data.

Preprocessing

- **Train-Test Split**: A balanced stratification to maintain class distribution.
- **Encoding**: One-hot encoding for categorical variables.
- **Scaling**: RobustScaler for numerical features.
- **SMOTE**: Addressed class imbalance by oversampling the minority class.

Model Performance

Initial Results:

- Logistic Regression: Recall = 69%, ROC-AUC = 0.74.
- SVM: Recall = 65%, ROC-AUC = 0.72.
- CatBoost: Recall = 54%, ROC-AUC = 0.71.

Optimized Results:

- **Logistic Regression**:
    - Threshold: Adjusted to 0.3 to improve recall.
    - Final Recall: 90%, ROC-AUC: 0.76.

	Model	Precision	Recall	F1-Score	\
0	Logistic Regression (Non-Tuned)	0.57	0.69	0.62	
1	Logistic Regression (Tuned)	0.57	0.68	0.62	
2	Logistic Regression (Threshold-Adjusted)	0.48	0.90	0.63	

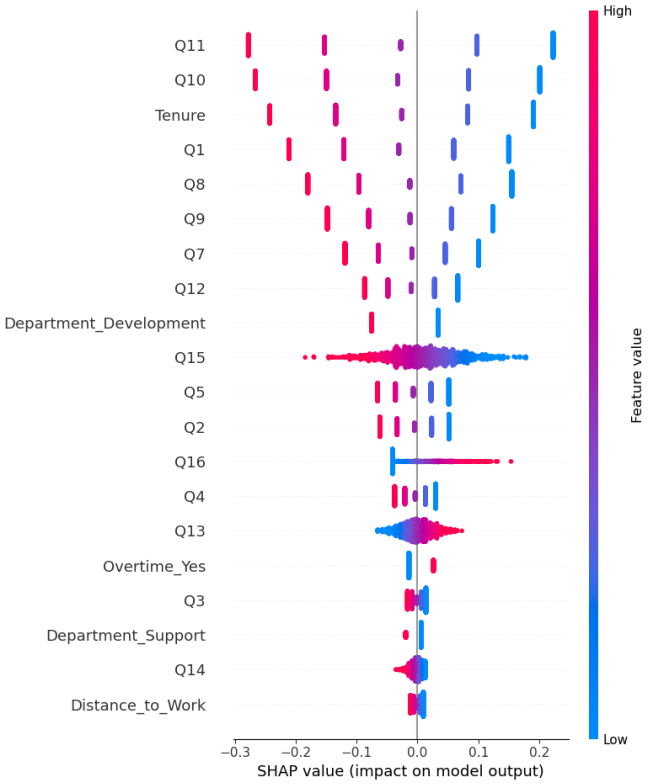
- **CatBoost**:
    - Improved Recall: 81% (threshold = 0.3).
    - Still lagged behind Logistic Regression in recall and simplicity.

	Model	Precision (Yes)	Recall (Yes)	F1-Score (Yes)	\
0	Non-Tuned CatBoost	0.58	0.54	0.56	
1	Tuned CatBoost	0.56	0.70	0.62	
2	Threshold-Adjusted CatBoost	0.49	0.81	0.61	

4-SHAP Value Analysis

Key Findings:

1. **Q11**: "In the last six months, someone at work has talked to me about my progress" emerged as the top predictor for attrition. Higher values reduced attrition probability.
2. **Q10**: "I have a best friend at work" was also significant.
3. **Tenure**: Longer tenure correlated with reduced attrition.



Actionable Insights

Regular feedback sessions (Q11) offer a cost-effective and high-impact method for retaining employees. This finding underscores the importance of engagement in HR strategies.

5-Conclusion

1- **Final Model**: Logistic Regression, with a recall of 90%, was selected for its balance between interpretability and performance.

2- **HR Implications**: Insights like the impact of Q11 feedback sessions can guide cost-effective retention strategies.

3- **Future Work**:

- Validate insights using real-world HR datasets.
- Extend analysis to other business areas like customer churn prediction.