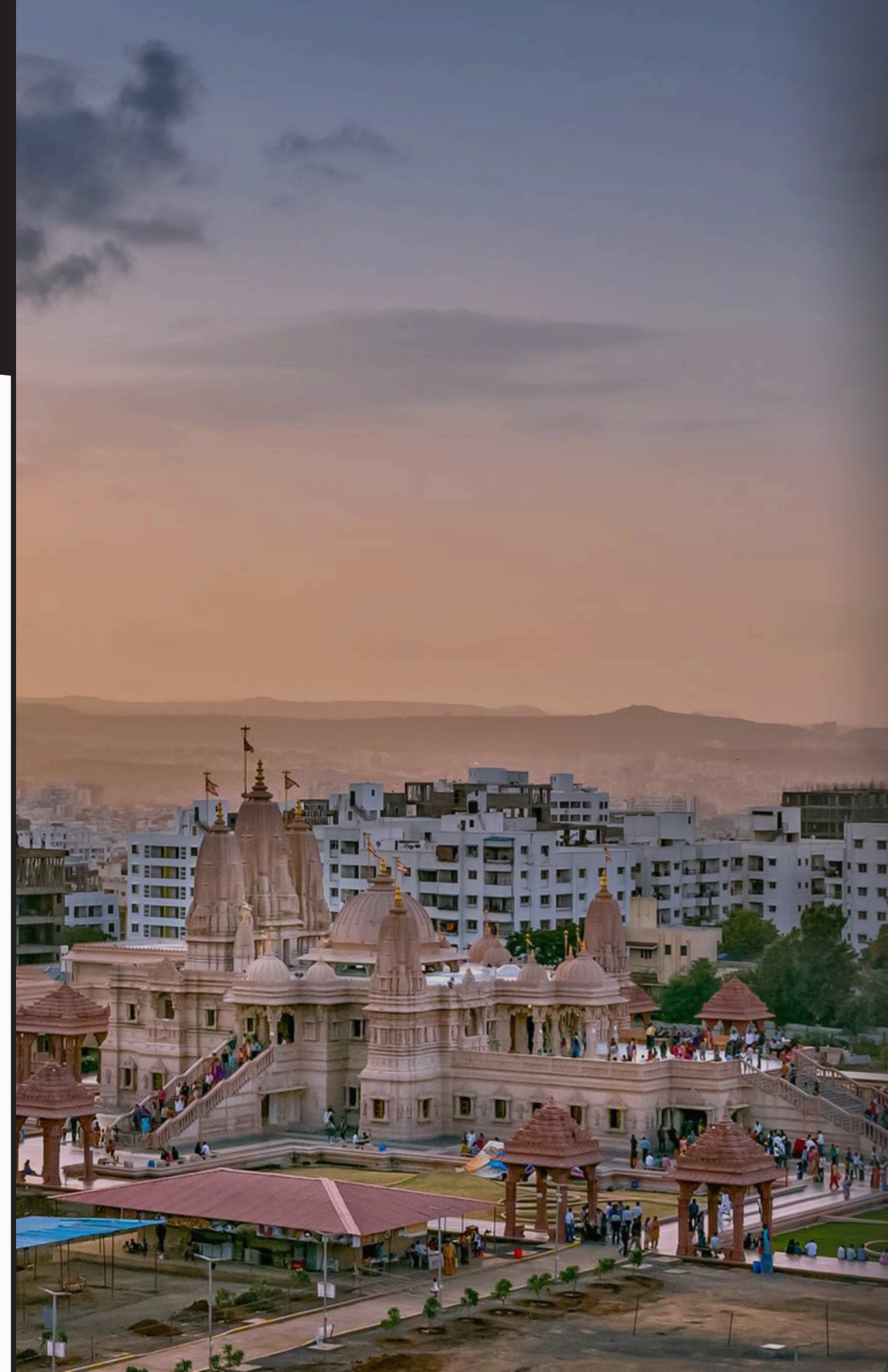
BANA 273 - Machine Learning

Predicting Employee Churn in 3 Major Indian Cities

Group 1: Allan Almaraz, Chia-Ling (Lydia) Chen , Tamara Edwin-Biayeibo, Hongde (Aldrich) Han

Overview



The Business Problem

Why does employee turnover matter?

High Costs: Replacing an employee costs 50-200% of their annual salary.

Lost Productivity: Knowledge, relationships, and momentum disappear

Downstream Effects: Team morale, client relationships, and operations suffer

The Challenge: Identify at-risk employees BEFORE they leave

Research Question

Which employee attributes and organizational factors most strongly influence churn, and how accurately can they be used to identify at-risk employees?

Identify Key Drivers: What factors predict churn?

Build Predictive Model: Maximize recall (catch churners)

Dataset Overview

Employee Future Prediction Dataset (Kaggle)

4,653
Employees

3 Cities
Bangalore, Pune, New Delhi

34.4%
Churn Rate

Key Variables: Education, Joining Year, City, Payment Tier, Age, Gender, Ever Benched, Experience in Current Domain, and Churn

Why Recall Matters

Goal: Maximize Recall (Minimize False Negatives)

Better to flag someone who stays than miss someone who leaves.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Key Patterns Amongst Variables

Mid-Pay Tier = Highest Churn

Employees in mid-tier compensation show elevated turnover

Masters Degree Holders Churn More

Show significantly higher churn than BA and PhD employees

Pune Has Highest Churn Rate

Geographic differences in turnover patterns

Modeling Approach

Logistic Regression

- Interpretable Coefficients
- Clear Odds Ratios
- Linear Effects

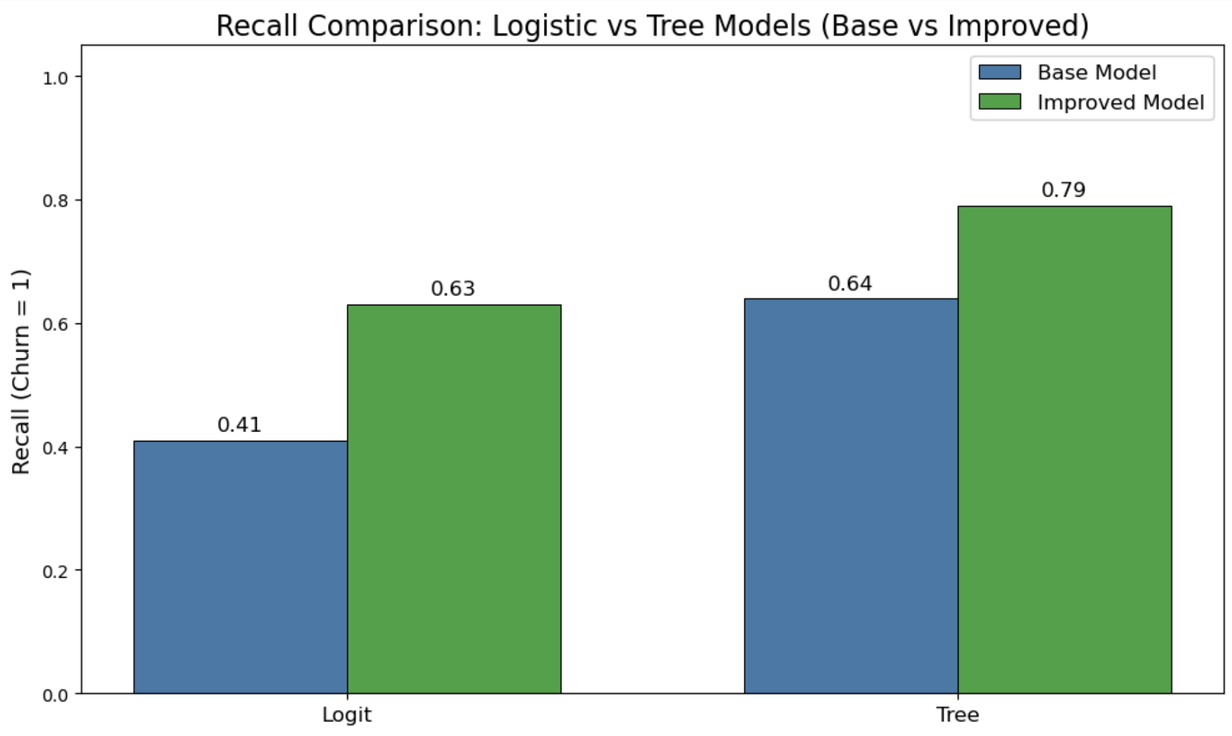
Decision Tree

- Captures Non-Linearities
- Interaction Effects
- Intuitive Rules

- Class Weight Balancing (address 65.6% / 34.4% imbalance)
- Regularization (L1 for logistic, pre-pruning for tree)
- Threshold Tuning (lowered to 0.35 for tree to boost recall)
- 5-Fold Cross-Validation (ensure stability and generalization)

Model Performance

Winner: Tuned Decision Tree, 79% Recall



What Drives Churn?

Logistic Regression Insights (Odds Ratios)

- **Masters Degree:** 2.4x odds vs Bachelors
- **Mid-Pay Tier:** 2.2x odds vs Low Pay
- **Pune Location:** 1.6x odds vs Bangalore
- **Ever Benched:** 1.5x odds vs Never Benched

What Drives Churn?

Decision Tree Insights (Information Gain)

- **Primary Drivers:** Joining Year > Payment Tier
- **Secondary Drivers:** City, Education, Gender

Churn risk is highest among long-tenured employees in specific pay/education segments. The tree captures nonlinear interactions that the linear model cannot.

Business Recommendations

Actionable Strategies for Retention

- **Target High-Risk Segments:** Focus retention efforts on long-tenured, mid-pay employees.
- **Review Compensation Structure:** Address mid-tier dissatisfaction through pay equity analysis
- **Minimize Benching Impact:** Reduce bench time or improve communication during unassigned periods
- **Location-Specific Strategies:** Develop targeted retention programs for Pune employees
- **Career Development for Masters:** Create advancement opportunities for highly educated employees

Key Takeaways

79% Recall

Churn is highly predictable with the tuned decision tree model

Clear Drivers

Tenure, compensation, and education consistently drive churn across modeling approaches

Proactive Intervention

Model enables early identification of at-risk employees

79% Recall

Reduced turnover costs through targeted retention efforts