Machine Learning Analysis: Predicting Employee Churn

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

University of California, Irvine

BANA 273 – Machine Learning

Professor Mingdi Xin

December 5, 2025

Abstract

Employee turnover imposes substantial organizational costs through lost expertise, productivity disruptions, and expenses tied to recruitment and training. Using the *Employee Future Prediction* dataset from Kaggle (N = 4,653 employees across Bangalore, Pune, and New Delhi), this study examines the research question: *Which employee attributes and organizational factors most strongly influence churn, and how accurately can they be used to identify at-risk employees?* Given that 65.6% of employees stay and only 34.4% churn, our analysis prioritizes recall for the churn class to minimize missed intervention opportunities.

We compare two supervised learning methods: Logistic Regression and a Decision Tree classifier. After addressing class imbalance through class weights, one-hot encoding (for logistic regression), cross-validation, and threshold tuning, an improved logistic model with L1 regularization achieves a churn recall of approximately 0.63 and identifies key predictors such as education level, payment tier, city, gender, and benching history. A pre-pruned Decision Tree with balanced class weights and a lowered decision threshold (0.35) further improves performance, reaching test-set recall near 0.79 with stable cross-validation results. The tree identifies joining year, compensation tier, and education as primary determinants, capturing nonlinear interactions that the linear model cannot fully represent.

Overall, results show that churn risk is concentrated among specific tenure–pay–education segments and that tuned tree-based models provide strong recall while remaining interpretable enough to guide targeted retention strategies and early intervention policies.

Keywords: machine learning, employee churn, logistic regression, decision tree classifier, recall, supervised learning, cross validation.

Introduction

Employee churn is a critical challenge that affects organizational stability, operational efficiency, and financial performance. Predictive modeling offers companies the opportunity to intervene early by identifying employees who are at risk of leaving before turnover occurs. In this report we analyze a publicly available Kaggle dataset titled *Employee Future Prediction* to evaluate whether machine learning methods can accurately classify churn outcomes. Our primary objective is to maximize recall (churn class) because false negatives represent missed intervention opportunities that may lead to substantial rehiring and retraining costs.

We will be focusing on logistic regression and decision tree classifier machine learning models and the model that provides the greatest improvement in recall (churn class), will be the most valuable for employers seeking to proactively retain talent. In addition to predicting churn, we also aim to understand which employee characteristics and workplace factors contribute most strongly to turnover. This leads to the central research question of our analysis: *Which employee attributes and organizational factors most strongly influence employee churn, and how accurately can they be used to identify at-risk employees?*

By answering this question, we seek both to develop a predictive framework that improves the identification of true churners and to provide actionable insights that employers and human resource departments can use to design targeted retention strategies.

Data Overview

The *Employee Future Prediction* dataset used for this project contains information on employees from three major Indian cities and is composed of 4,653 observations and nine attributes, including the dependent variable *leaveornot*. The outcome variable indicates whether

an employee will leave the organization within the next two years, where 1 represents churn and 0 represents staying.

The dataset consists primarily of structured numerical fields, and no missing values or inconsistencies were present. As a result, only minimal preprocessing was required. Several fields, such as education level and city of employment, were encoded numerically in the original dataset to facilitate model estimation. Because these variables represent categorical levels, we applied one-hot encoding for the logistic regression model to ensure appropriate interpretation and estimation. Tree-based models were able to use the original categorical encodings directly. Below is a summary of each attribute included in the dataset:

- *education*: Level of educational attainment, where 0 indicates a bachelor's degree, 1 indicates a master's degree and 2 indicates a PhD.
- *joiningyear*: The year in which the employee joined the organization.
- *city*: City of employment, where 0 indicates Bangalore, 1 indicates Pune and 2 indicates New Delhi.
- *paymenttier*: Compensation tier of the employee, where 0 indicates low pay, 1 indicates mid pay and 2 indicates high pay.
- *age*: Age of the employee.
- *male*: Gender indicator, where 1 indicates male and 0 indicates female.
- *everbenched*: Indicator of whether an employee has been placed on the bench, meaning either excluded or not assigned to a project. A value of 1 indicates yes and 0 indicates no.
- *experienceincurrentdomain*: Number of years of experience the employee has in their current professional domain

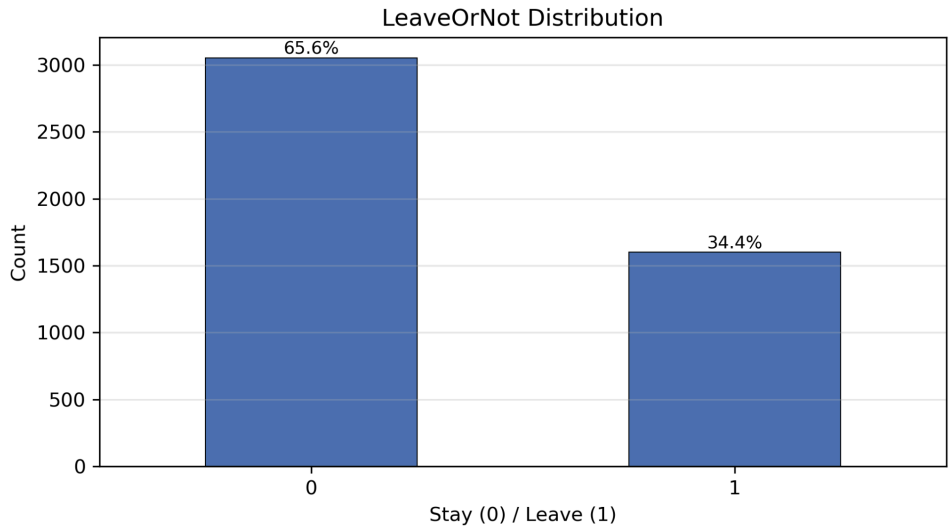
- *leaveornot*: Outcome variable indicating churn status, where 1 indicates leaving the company within two years and 0 indicates staying.

Data Description

Before building predictive models, we examined the distribution of the dependent variable and it revealed a meaningful class imbalance: 65.6% of employees stay, while only 34.4% churn (see Figure 1).

Figure 1

Distribution of employee churn outcomes.



Note. Data from Tejashvi14 (2020).

This imbalance has important implications for model evaluation. A model could achieve a 65.6% overall accuracy by predicting that everyone stays, yet such a model would entirely fail to identify individuals who are actually at risk of leaving. In the context of churn analysis, these false negatives are the most costly errors because each undetected churner represents a missed opportunity for intervention, ultimately resulting in financial losses associated with turnover, recruitment, and retraining.

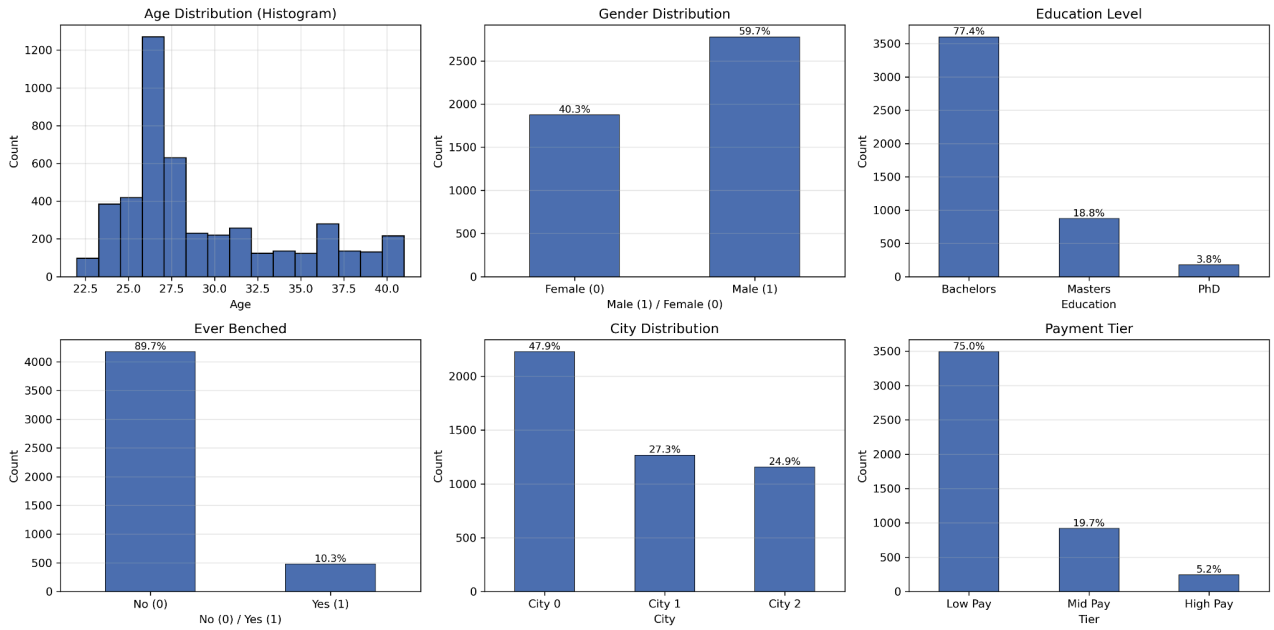
Because of this imbalance, recall for the churn class becomes our primary evaluation metric. Recall measures the proportion of true churners the model successfully identifies. A higher recall value corresponds to fewer false negatives, which is critical in a churn context because employees misclassified as staying receive no intervention. Reducing these missed cases allows organizations to intervene earlier and allocate retention resources more effectively.

Attribute Distribution

To further contextualize the modeling process, we examined how key employee attributes are distributed across the dataset (see Figure 2).

Figure 2

Summary of employee attribute distributions, including age, gender, education level, benching status, city, and payment tier.



Note. Data from Tejashvi14 (2020).

The summaries reveal several notable patterns in the workforce: the employee population is relatively young, with most individuals in their mid-20s to early 30s; males represent a larger

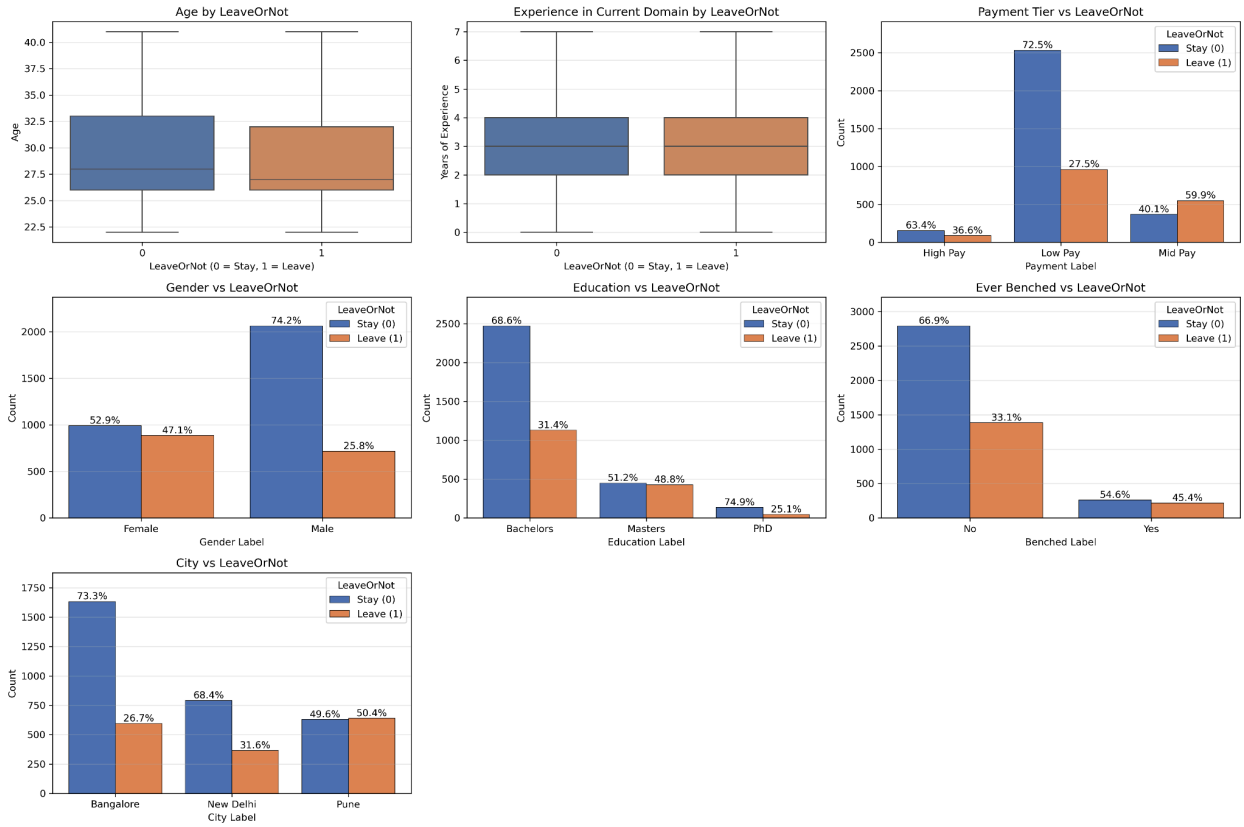
proportion of the sample, and the workforce is heavily concentrated at the bachelor's level, suggesting that many employees are early in their careers. Most employees have never been benched, although the small proportion who have may experience increased uncertainty or dissatisfaction. Employees are distributed across three major metropolitan business hubs, with Bangalore accounting for the largest share of employees, followed by Pune and New Delhi. Compensation is highly skewed toward the lowest tier, which may influence turnover risk given the well-established association between pay dissatisfaction and voluntary churn. These descriptive patterns provide an important foundation for interpreting model behavior, as they highlight structural workforce characteristics and potential sources of churn risk.

Explanatory Variable Analysis

To gain a more detailed understanding of how churn rates differ across groups or categories, we next examine how key attributes are distributed across the stay versus churn classes (see Figure 3).

Figure 3

Churn rates across employee attributes, including age, experience, payment tier, gender, education level, benching status, and city.



Note. Data from Tejashvi14 (2020).

Feature-level visualizations reveal that churn rates vary substantially across employee groups. Employees in the mid pay tier exhibit the highest churn, suggesting that dissatisfaction may arise when compensation is not low enough to be expected nor high enough to feel rewarding. Workers who have been benched also show noticeably elevated churn, indicating that periods without project assignment may reduce engagement or job security. Master's degree holders churn at higher rates than those with bachelor's or PhD degrees, and churn rates vary by region, with Pune employees leaving at the highest rate. Gender differences also emerge, with female employees showing a higher proportion of churn than male employees. Collectively, these results illustrate that churn is not randomly distributed but is concentrated in specific subgroups, a pattern that informs and motivates the predictive modeling stage.

Modeling & Analysis

To understand and predict employee churn, we applied two supervised learning methods: logistic regression and a decision tree classifier. These models were selected because they offer complementary strengths. Logistic regression provides interpretability by estimating the direction and magnitude of each predictor's association with churn, whereas the decision tree captures nonlinear relationships and interaction effects that may not be visible to a linear model.

Across all models, we used a 70/30 train-test split, a fixed random state of 42 for reproducibility, and five-fold cross-validation (CV) on the training data, with a specific focus on recall for the churn class, since reducing false negatives is the primary objective in a churn-prediction setting.

Logistic Regression Model

The first predictive method we explored was logistic regression, which models the probability that an employee will churn as a function of their characteristics. Logistic regression is based on the log-odds (logit) transformation, which expresses the relationship between the predictors and the probability of churn. Formally, the model can be written as:

$$log(\frac{P(Y=1 \mid X)}{1-P(Y=1 \mid X)}) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ \ldots \ + \beta_k X_k$$

Applying the logistic function transforms this linear expression into a predicted probability between 0 and 1:

$$P(Y = 1 \mid X) = \frac{1}{1+e^{-(\beta_o+\beta_1 X_1+\beta_2 X_2+ \ldots +\beta_k X_k)}}$$

Predictions are typically assigned using a 0.50 threshold, meaning that any employee with a predicted probability above 0.50 is classified as a churner. This formulation allows both interpretability and probabilistic decision-making. Later in the analysis, we return to these log-odds to interpret the improved model's coefficients.

Basic Logistic Regression Model

We first estimated an untuned logistic regression model using default scikit-learn settings. As expected in an imbalanced dataset, the baseline model performs poorly at identifying churners (see Figure 4).

Figure 4

Basic logistic regression performance metrics.

	Predicted Stay (0)	Predicted Churn (1)
Actual Stay (0)	828	88
Actual Churn (1)	284	196

	precision	recall	f1-score	support
Stay	0.74	0.90	0.82	916
Churn	0.69	0.41	0.51	480
accuracy			0.73	1396
macro avg	0.72	0.66	0.66	1396
weighted avg	0.73	0.73	0.71	1396

```
5-Fold CV Recall (Churn = 1): [0.397 0.42  0.321 0.308 0.438]
Mean CV Recall (Churn = 1): 0.377
```

Note. Data from Tejashvi14 (2020).

Although the basic model achieves moderate accuracy, recall for the churn class is very low (0.41), meaning many true churners were incorrectly predicted as staying. This shortcoming is common in imbalanced datasets: because approximately two-thirds of employees in the dataset remain with the company, the model defaults toward predicting the majority class. As a result, accuracy becomes a misleading indicator of performance, and recall offers a far more appropriate evaluation metric for this prediction task.

Improved Logistic Regression Model

To improve predictive performance, particularly recall for the churn class, we refined the logistic regression model using a set of targeted optimization strategies. Guided by results from GridSearchCV, we selected a model specification that incorporated L1 regularization (lasso penalty), class-weight balancing, and an increased regularization strength parameter. L1 regularization helps reduce noise in the model by shrinking weaker coefficients to zero, effectively removing uninformative predictors and enhancing interpretability. In addition, assigning the class weight parameter to "balanced" adjusts the loss function so that the model places greater emphasis on correctly identifying churners, who represent the minority class in the dataset. The solver was set to *liblinear* to support L1 regularization, and the regularization strength (C = 0.1) was chosen to allow the model to retain meaningful predictors while still applying shrinkage. The improved logistic regression model achieves notably stronger recall compared to the untuned version (see Figure 5).

Figure 5

Improved logistic regression performance metrics.

	Predicted Stay (0)	Predicted Churn (1)
Actual Stay (0)	671	245
Actual Churn (1)	179	301

	precision	recall	f1-score	support
Stay	0.79	0.73	0.76	916
Churn	0.55	0.63	0.59	480
accuracy			0.70	1396
macro avg	0.67	0.68	0.67	1396
weighted avg	0.71	0.70	0.70	1396

```
CV Recall (Churn = 1): [0.652 0.625 0.638 0.571 0.661]
Mean Recall (Churn = 1): 0.629
```

Note. Data from Tejashvi14 (2020).

Test-set recall for churn increases to approximately 0.63, indicating a substantial improvement in the model's ability to correctly identify employees who are at risk of leaving. Five-fold cross-validation on the training data further supports the model's stability: the recall scores range from roughly 0.57 to 0.65, with a mean recall of about 0.629. This consistency across folds suggests that the model generalizes well and is not overly sensitive to any particular train–test split.

Although these enhancements lead to an increase in false positives, a common and expected consequence when optimizing for recall, the trade-off is acceptable in a churn-prevention context. Missing a true churner (a false negative) is far more costly for an organization than incorrectly flagging a stable employee, making higher recall the prioritized metric. The improved logistic model therefore provides a more dependable foundation for early intervention strategies designed to retain at-risk employees.

Results

To better interpret the improved logistic regression model, we converted its coefficients into odds ratios, allowing us to examine how each predictor meaningfully shapes churn risk after regularization (see Figure 6).

Figure 6

Odds ratios and feature impacts from the improved logistic regression model.

	Feature	Coefficient	Odds_Ratio	Impact
0	Masters vs Bachelors	0.894	2.444	Increases churn
1	Mid Pay vs Low Pay	0.767	2.154	Increases churn
2	Pune vs Bangalore	0.461	1.585	Increases churn
3	male	-0.795	0.452	Decreases churn
4	everbenched	0.413	1.511	Increases churn
5	New Delhi vs Bangalore	-0.512	0.599	Decreases churn
6	experienceincurrentdomain	-0.031	0.970	No effect
7	age	-0.029	0.971	No effect
8	joiningyear	0.000	1.000	No effect
9	PhD vs Bachelors	0.000	1.000	No effect
10	High Pay vs Low Pay	0.000	1.000	No effect

Note. Data from Tejashvi14 (2020).

The results reveal a clear set of factors that stand out once weaker variables have been shrunk toward zero. Education emerges as one of the strongest influences: holding a Master's degree rather than a Bachelor's is associated with more than double the odds of churn, suggesting that more highly educated employees may possess greater external mobility or stronger labor market competitiveness, making them more likely to leave. Compensation patterns also play a notable role. Employees in the mid pay tier exhibit more than twice the odds of churn compared to those in the low pay tier, a finding that aligns with earlier descriptive patterns and may reflect dissatisfaction among mid-tier employees who perceive a mismatch between their contributions and compensation. Geographic differences further contribute to turnover risk. Employees located in Pune demonstrate significantly higher churn likelihood than those in Bangalore, while employees in New Delhi show substantially lower churn odds. These regional patterns likely reflect local labor market conditions, availability of alternative employment, or cost-of-living considerations that influence career mobility. Organizational signals also matter: employees who have ever been benched face markedly higher churn odds, reinforcing the earlier observation that

time spent without project assignments may be interpreted as instability or a lack of alignment within the company. In contrast, certain demographic variables show the opposite effect. Male employees display substantially lower churn likelihood than female employees, indicating gender-based patterns in turnover behavior that merit further exploration.

Other features, such as age, experience in the current domain, joining year, and having a PhD, exhibit odds ratios near one, suggesting that once other variables are controlled for, these attributes contribute little independent predictive value. The sparsity introduced by L1 regularization helps confirm that these variables do not meaningfully improve model performance. Taken together, the results illustrate how the improved logistic regression model isolates the most influential predictors of churn, emphasizing education, compensation tier, benching history, geographic location, and gender, while appropriately de-emphasizing weaker signals.

Although the model provides clear and interpretable insight into these relationships and shows improved recall, its linear structure limits its ability to capture more complex, nonlinear interactions among employee characteristics. These constraints, combined with the nonlinear patterns suggested in the exploratory analysis, motivate the transition to a decision tree classifier, which is better equipped to uncover hierarchical splits and interaction effects that may more accurately characterize the drivers of employee churn.

Decision Tree Classifier

Unlike logistic regression, which models a single linear relationship between predictors and churn, decision trees make no assumptions about linearity. Instead, they recursively split the data into increasingly homogeneous groups based on the values of the input features. To evaluate potential splits, decision trees rely on measures of impurity, in this case we used entropy, which

quantifies how mixed a node is with respect to churn outcomes. A pure node containing only churners or only non-churners has low entropy, whereas a node with an even mixture has high entropy. The model selects the split that produces the greatest reduction in entropy, known as information gain, thereby forming branches that most effectively separate churners from non-churners. In practice, this means that attributes with the strongest influence on churn tend to appear near the top of the tree, making it easy to visualize both their importance and directional impact.

Because the model learns these splits recursively, it naturally captures nonlinear patterns and complex interactions among features. For instance, a decision tree can learn that churn risk may depend jointly on factors such as pay tier, city, and benching history – relationships that may not emerge in a linear framework. This flexibility makes decision trees well suited for understanding churn behavior, where risk often arises from combinations of employee characteristics rather than any single variable acting alone.

Basic Decision Tree Classifier

The next step in our analysis was to estimate a basic decision tree classifier using default scikit-learn settings. As is typical with untuned trees, this baseline model provides a useful reference point but comes with known limitations, such as a tendency to overfit the training data and reduced generalizability to new observations (see Figure 7).

Figure 7

Performance metrics and confusion matrix for the basic Decision Tree classifier.

	Predicted Stay (0)	Predicted Churn (1)
Actual Stay (0)	814	102
Actual Churn (1)	175	305

	precision	recall	f1-score	support
Stay	0.82	0.89	0.85	916
Churn	0.75	0.64	0.69	480
accuracy			0.80	1396
macro avg	0.79	0.76	0.77	1396
weighted avg	0.80	0.80	0.80	1396

```
5-Fold CV Recall (Churn = 1): [0.696 0.701 0.634 0.629 0.688]
Mean CV Recall (Churn = 1): 0.67
```

Note. Data from Tejashvi14 (2020).

While the model achieves strong overall accuracy (0.80), its real strength relative to the untuned logistic regression model is its substantially higher recall for the churn class. Specifically, the basic tree correctly identifies 64 percent of churners, a notable improvement over the logistic model's recall of 41 percent. Cross-validation reinforces this performance: the five-fold mean recall of 0.67 suggests that the model reliably captures churn patterns across multiple resamples of the training data.

These findings indicate that even without tuning, a decision tree can uncover nonlinear relationships and interaction effects that a simple logistic model may miss. However, the presence of some variability across CV folds and the inherent risk of overfitting highlight the need for further refinement.

Pre-Pruned Decision Tree Classifier

To strengthen predictive performance, we refined the decision tree classifier using a series of targeted tuning strategies. Guided by GridSearchCV, we selected parameter values designed to balance model complexity with the goal of identifying as many true churners as possible. Specifically, we applied class_weight="balanced" so that the model places greater

emphasis on the minority churn class, and we constrained the tree with a maximum depth of 8, minimum of 10 samples per leaf, and minimum split size of 5. These pre-pruning choices help the tree avoid overfitting while still capturing meaningful nonlinear relationships in the data.

Because decision trees naturally produce probability estimates for each prediction, we further enhanced recall by lowering the decision threshold from the standard 0.50 to 0.35. This adjustment allows the model to classify borderline cases as churners, a strategy that is especially appropriate in churn prevention where the cost of missing a true churner (a false negative) is considerably higher than the cost of incorrectly flagging a stable employee (see Figure 8).

Figure 8

Performance metrics and confusion matrix for the tuned Decision Tree classifier with a lowered decision threshold (0.35).

	Predicted Stay (0)	Predicted Churn (1)
Actual Stay (0)	687	229
Actual Churn (1)	102	378

	precision	recall	f1-score	support
Stay	0.87	0.75	0.81	916
Churn	0.62	0.79	0.70	480
accuracy			0.76	1396
macro avg	0.75	0.77	0.75	1396
weighted avg	0.79	0.76	0.77	1396

```
CV Recall (Churn = 1, standard threshold): [0.728 0.763 0.75  0.692 0.701]
Mean CV Recall (Churn = 1): 0.727
```

Note. Data from Tejashvi14 (2020).

Test-set recall for churners rises to roughly 0.79, marking a substantial gain over both the untuned decision tree (0.67) and the improved logistic regression model (0.63). The confusion matrix confirms that the model correctly identifies 378 churners, while cross-validation further

validates its stability: using the standard threshold of 0.5, five-fold recall scores range from about 0.69 to 0.76, yielding a mean of 0.727. This consistency indicates that the tuned tree generalizes reliably across different training subsets.

Results

To interpret the tuned decision tree model, we examined each attributes' information gain and the structure of the tree's top-level splits (see Figure 9).

Figure 9

Information gain values for the tuned Decision Tree model, highlighting the relative importance of each predictor in separating churners from non-churners.

	Feature	Information_Gain	Impact
0	joiningyear	0.36	Key driver of churn
1	paymenttier	0.21	Key driver of churn
2	city	0.13	Secondary driver
3	education	0.12	Secondary driver
4	male	0.07	Secondary driver
5	age	0.05	Secondary driver
6	experienceincurrentdomain	0.04	Secondary driver
7	everbenched	0.02	Minimal / no impact

Note. Data from Tejashvi14 (2020).

The results reveal a clear hierarchy of predictors. Joining year emerges as the strongest driver of churn (information gain ≈ 0.36), suggesting that longer-tenured employees are substantially more likely to leave. This pattern aligns with organizational dynamics in which employees with more years of service may face stagnation or possess stronger external opportunities. Compensation tier follows as the next most influential variable (≈ 0.21), supporting earlier findings that

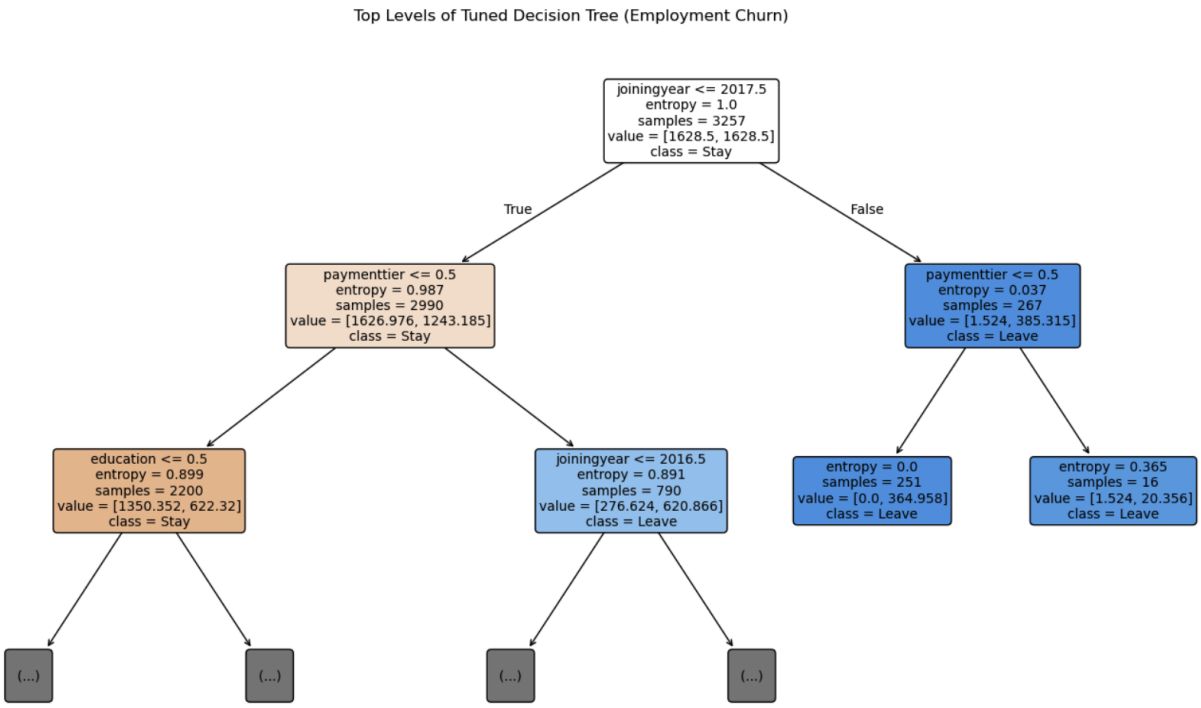
mid-paid employees face elevated churn risk and indicating that compensation structure and perceived equity meaningfully shape turnover decisions.

A second tier of predictors, which include city, education, gender, age, and experience, provides additional refinement. City-level differences (≈ 0.13) reflect regional labor-market dynamics, while education (≈ 0.12) captures mobility patterns among more highly educated employees. Gender, age, and experience contribute smaller but non-trivial improvements in classification accuracy. Notably, benching history, a strong predictor in the logistic regression model, appears minimally influential here. This contrast underscores how tree-based models prioritize variables that create strong early splits, whereas regression captures broader linear trends.

The decision tree's structure also offers further insight into how churn risk unfolds across the workforce (see Figure 10).

Figure 10

Top-level structure of the tuned Decision Tree model, showing the primary splits (joining year, payment tier, and education) used to classify employee churn.



Note. Data from Tejashvi14 (2020).

The root split is determined by joining year, confirming its role as the dominant predictor: employees who joined earlier follow a distinct risk pathway from more recent hires. For longer-tenured employees, the next major split is compensation tier, reinforcing its central influence, followed by education, which differentiates risk within this subgroup. Among newer employees (joined after 2017), the model again selects compensation tier as the next best discriminator, emphasizing its consistent impact across segments. Although city is an important predictor overall, its absence from the top layers of the tree suggests that geographic differences emerge in more granular, lower-level splits rather than shaping the broad structure of churn pathways.

Taken together, the tuned decision tree highlights a hierarchy of churn determinants centered on tenure, compensation, and education. These patterns illustrate the nonlinear interactions and conditional relationships that linear models cannot easily capture, reinforcing the

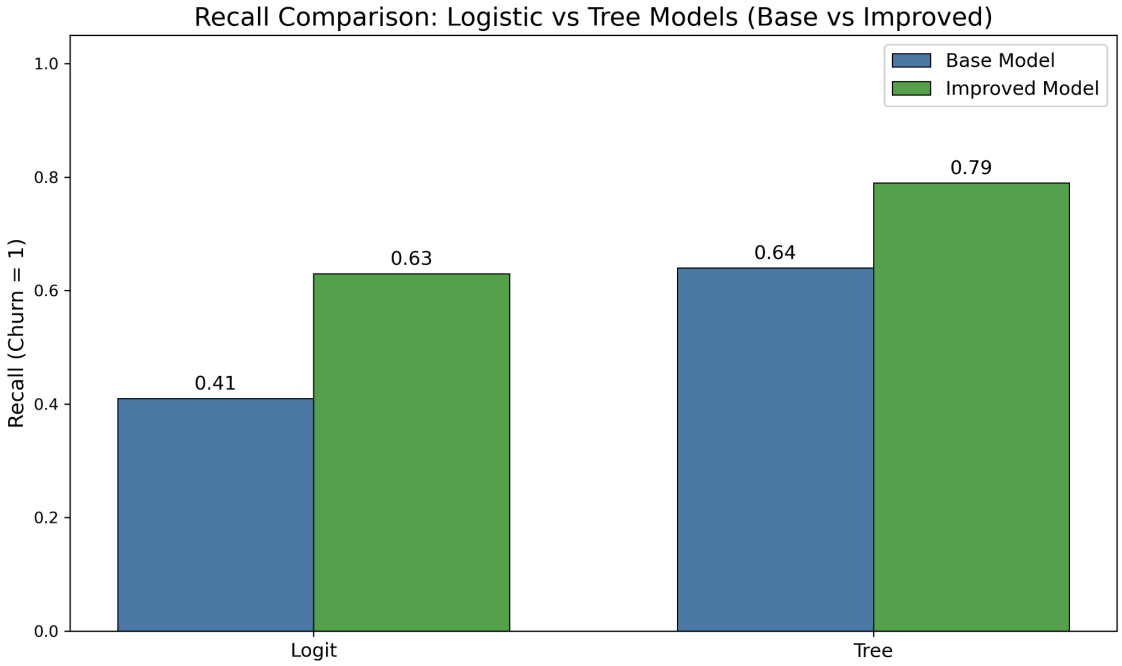
idea that churn risk arises from combinations of employee characteristics rather than any single factor acting independently.

Comparison & Takeaways

Bringing both models together, a clear pattern emerges: tree-based approaches outperform logistic regression on the primary objective of maximizing recall for churners, but the two models provide complementary lenses on the problem (see Figure 11).

Figure 11

Recall comparison between baseline and improved models for Logistic Regression and Decision Tree classifiers.



Note. Data from Tejashvi14 (2020).

The untuned logistic model struggled to identify churners (recall ≈ 0.41), and even after regularization and class-weight balancing, recall rose only to about 0.63. In contrast, the basic Decision Tree already reached recall around 0.64, and the tuned, pre-pruned tree with a lowered threshold achieved recall close to 0.79. From a data-mining perspective, this illustrates how

model choice and threshold selection can dramatically change performance when the goal is to minimize false negatives, especially in imbalanced settings where accuracy alone is misleading.

The models also highlight different but overlapping stories about what drives churn. Logistic regression, which estimates global linear relationships, emphasizes benching history, education, pay tier, city, and gender as the strongest predictors. It suggests that being benched, holding a master's degree, working in Pune, being in the mid pay tier, and being female all increase churn odds, even after controlling for other variables. The decision tree, on the other hand, organizes the same variables into a hierarchy of conditional rules. It identifies joining year and payment tier as the dominant early splits, with education and city refining risk within specific subgroups. In the tree, benching plays only a minor role, which shows how a predictor can matter in a global linear sense but contribute less to the main decision pathways that best reduce entropy. Together, these perspectives suggest that churn risk is shaped both by broad structural factors (tenure, compensation, geography) and by localized experiences (benched status, education, gender) that interact differently across segments.

Business Implications

For business managers, these findings translate into several actionable insights. The tuned decision tree indicates that longer-tenured employees in specific pay tiers represent a particularly high-risk segment; targeted retention efforts, such as career-development conversations, internal mobility options, or differentiated pay adjustments, should prioritize these groups. Both models point to mid-tier compensation as a recurring pressure point, implying that salary bands and promotion pathways may need to be reviewed to reduce perceived inequities. The logistic model's strong coefficient on ever benched suggests that managers should minimize unassigned periods and proactively communicate with employees when benching is unavoidable, as these

employees may interpret bench time as exclusion, instability, or lack of fit. Finally, city and education effects indicate that retention strategies should not be one-size-fits-all: locations like Pune or highly educated employees may warrant tailored engagement, recognition, and advancement programs.

Conclusion

Taken together, the analyses conducted in this project directly address our central research question: W*hich employee attributes and organizational factors most strongly influence employee churn, and how accurately can they be used to identify at-risk employees?* Our modeling results show that churn is shaped primarily by a combination of tenure, compensation tier, education level, geographic location, gender, and—in the case of logistic regression—benching history. These factors emerged consistently across descriptive analyses, logistic regression coefficients, and decision tree splits, demonstrating that turnover risk is neither random nor evenly distributed, but concentrated among employees with specific demographic, organizational, and career characteristics.

The project also underscored several key methodological insights. Prioritizing recall fundamentally shifted how we evaluated the models, highlighting the need to detect as many churners as possible rather than maximizing overall accuracy. Addressing class imbalance through class weights, threshold tuning, and cross-validation proved essential for building models that generalize beyond the training data. Moreover, the comparison of modeling approaches showed that different algorithms offer different types of understanding: logistic regression provides interpretable effect sizes and odds ratios, while decision trees uncover nonlinear interactions and intuitive, rule-based pathways that reveal how multiple attributes combine to shape churn risk.

Ultimately, the tuned decision tree achieved the strongest predictive performance, demonstrating that employee churn can be identified with high recall when models are carefully selected, tuned, and evaluated. These findings not only answer the research question but also show how employee data can be transformed into meaningful, actionable insight. By combining interpretability with predictive accuracy, the analysis provides a practical foundation that organizations can use to identify at-risk employees early and design targeted, evidence-based retention strategies.

References

Tejashvi14. (2020). Employee Future Prediction [Dataset]. Kaggle. https://www.kaggle.com/datasets/tejashvi14/employee-future-prediction.