Predicting Employee Turnover Using Ensemble Machine Learning

A Machine Learning Approach to Reducing Turnover

Saad Ahmed Qadeer

Abstract

Executive Summary: This project analyzed employee data to predict turnover and identify key factors contributing to employee attrition. We developed and compared multiple predictive models including logistic regression, decision trees, and random forest classifiers. The analysis revealed that employee departures are primarily driven by excessive workload, lack of recognition, and inadequate work-life balance. The final random forest model achieved **96.2% accuracy** with strong predictive performance across all metrics, providing actionable insights for HR interventions.

Contents

1 Project Overview

1.1 Business Problem

The HR department recognized a critical need to understand the factors contributing to employee turnover and to develop a predictive model to identify at-risk employees. High turnover rates result in significant costs related to recruitment, training, and lost productivity. By identifying patterns in employee departures, the company can implement targeted retention strategies.

1.2 Objectives

The primary objectives of this analysis were:

1. Analyze employee data to identify patterns and factors associated with turnover.
2. Develop predictive models to accurately forecast which employees are likely to leave.
3. Provide actionable recommendations based on model insights to improve employee retention.

2 Data Understanding

2.1 Dataset Overview

The dataset contained 14,999 employee records with 10 variables capturing various aspects of employment. The variables are described below:

Variable	Description
satisfaction_level	Self-reported job satisfaction (0-1 scale)
last_evaluation	Most recent performance review score (0-1 scale)
number_project	Count of projects assigned to the employee
average_montly_hours	Typical hours worked per month
time_spend_company	Tenure in years
Work_accident	Whether the employee had a workplace accident (Binary)
promotion_last_5years	Recent promotion status (Binary)
department	Employee's department
salary	Compensation level (low, medium, high)
left	Target variable: Whether the employee left (Binary)

Table 1: Dataset Variable Dictionary

2.2 Data Cleaning and Preparation

We performed several data cleaning steps to ensure data quality:

- **Missing Values:** The dataset had no missing values.
- **Duplicate Records:** We identified 3,008 duplicate rows (20% of the data). Given that these represented identical responses across all 10 variables, they were deemed highly unlikely to be legitimate entries and were removed.
- **Outliers:** Analysis revealed outliers in the tenure variable. These were retained as they represented legitimate cases of long-tenured employees.
- **Column Renaming:** Standardized column names to `snake_case` format (e.g., `time_spend_company` became `tenure`).

3 Exploratory Data Analysis

3.1 Descriptive Statistics

Initial statistical analysis revealed several important patterns:

- The overall turnover rate was **16.6%** (1,991 out of 11,991 employees after removing duplicates).
- Average satisfaction level was 0.61, indicating moderate satisfaction.
- Mean monthly working hours was 201 hours, significantly above a typical 40-hour work week (approx. 173 hours/month).

- Only **2.1%** of employees received promotions in the last 5 years.
- Average tenure was 3.5 years.

3.2 Key Visualizations and Insights

3.2.1 Satisfaction Level vs. Tenure

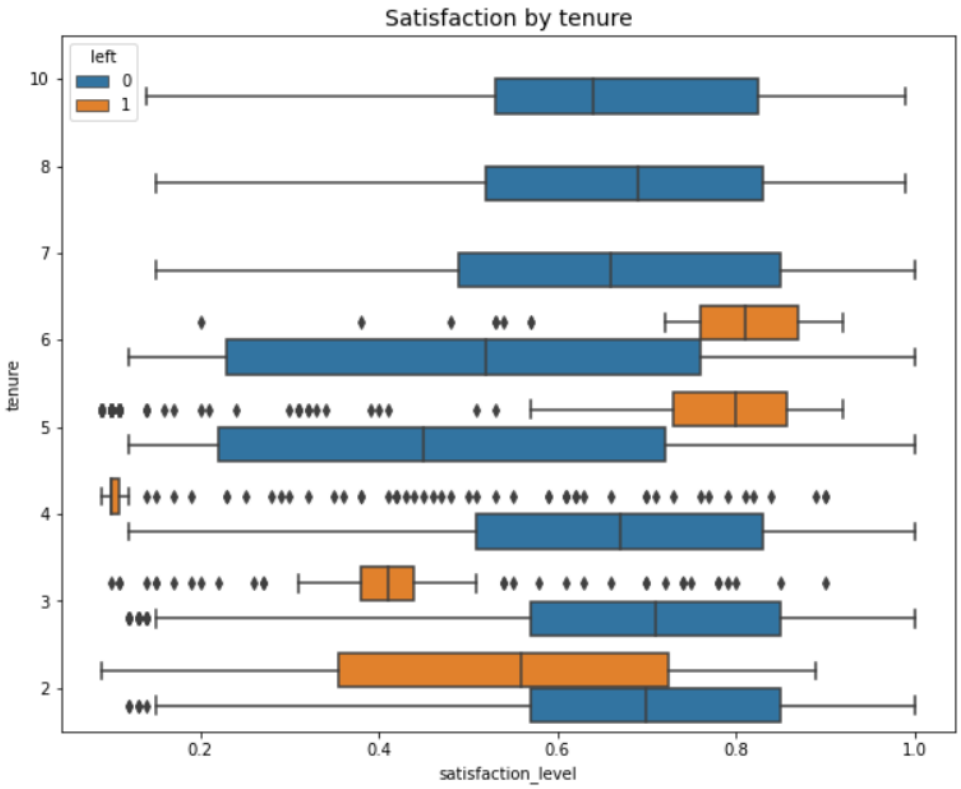

Figure 1: Satisfaction Level vs Tenure Scatterplot (colored by 'left' status).

The scatterplot analysis revealed two distinct groups of departing employees:

1. Dissatisfied employees with shorter tenures (low satisfaction, typically 2–4 years).
2. High-performing employees with medium tenure (high satisfaction, around 4–6 years).

Notably, four-year employees who left had unusually low satisfaction levels, suggesting potential issues with company policies or career progression at this specific milestone.

3.2.2 Workload and Performance

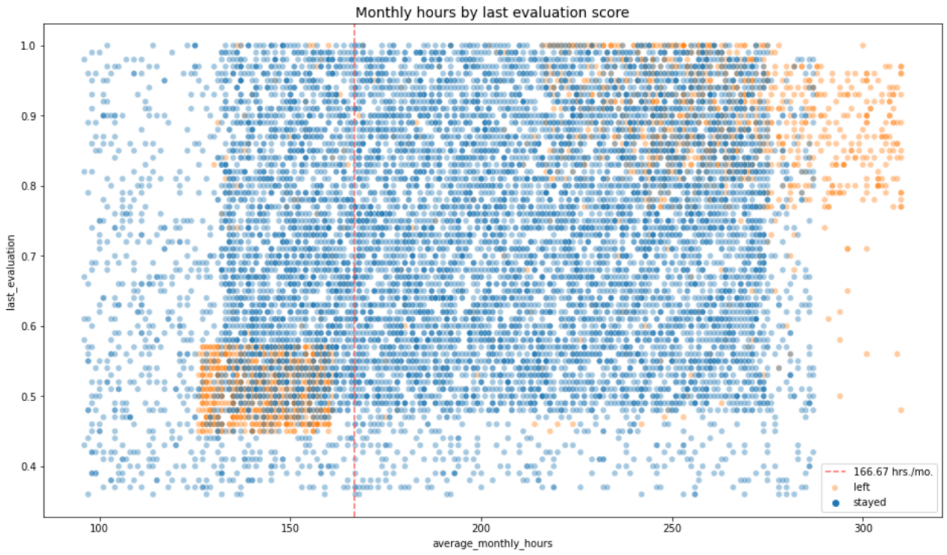

Figure 2: Average Monthly Hours vs Last Evaluation Scatterplot.

This analysis uncovered two categories of employees who left:

- **Overworked High Performers:** Employees working 240+ hours per month with excellent evaluation scores (0.75–1.0).
- **Underperformers:** Those working below average hours (140–160 hours) with lower evaluation scores.

3.2.3 Project Load and Promotions

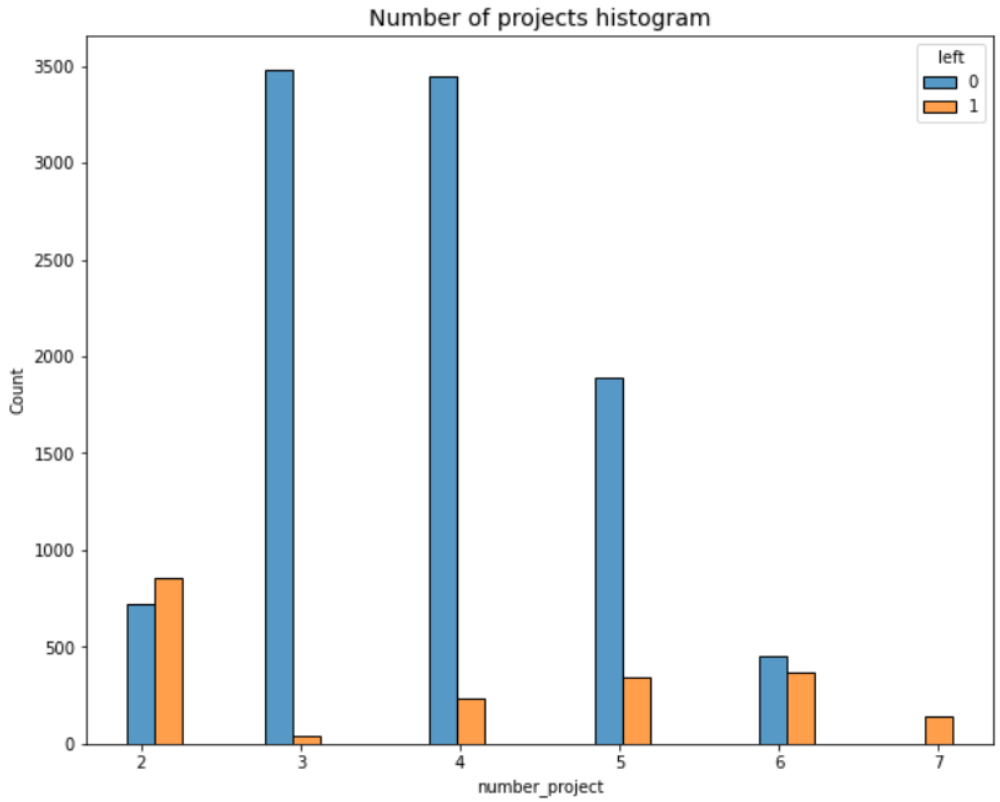

Figure 3: Distribution of Projects by Departure Status.

Employees who left were primarily those with either 2 projects (underutilized) or 6–7 projects (overworked). The ideal project load appeared to be 3–4 projects, where retention was highest.

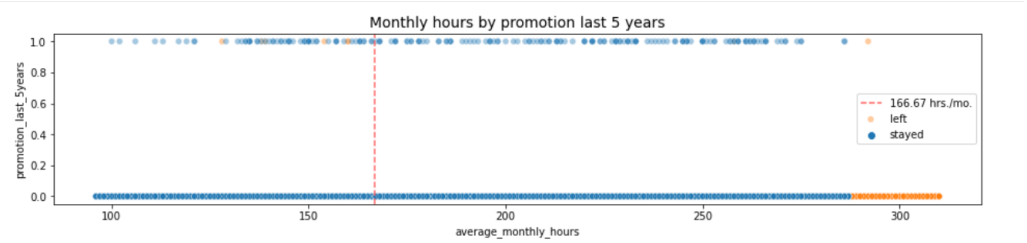

Figure 4: Promotion Status and Turnover.

Furthermore, virtually no employees who left had received recent promotions (overall promotion rate 2.1%), indicating lack of career advancement is a significant driver of attrition.

4 Model Development

4.1 Model Selection Rationale

We developed and compared three types of predictive models:

1. **Logistic Regression:** A baseline model useful for understanding linear relationships.
2. **Decision Tree:** A non-linear model capturing complex interactions and creating interpretable rules.
3. **Random Forest:** An ensemble method to improve accuracy and reduce overfitting.

4.2 Feature Engineering & Training

We created additional features such as an **Overworked Indicator** (binary flag for >175 hours/month) and used one-hot encoding for categorical variables (Department, Salary).

The data was split into 75% training (8,993 records) and 25% testing (2,998 records). We utilized `GridSearchCV` with 5-fold cross-validation to tune hyperparameters (max_depth, min_samples_leaf, etc.) for the tree-based models.

5 Model Results and Evaluation

5.1 Performance Comparison

The Random Forest model achieved the best performance across all metrics.

Metric	Logistic Regression	Decision Tree	Random Forest
Accuracy	83.0%	96.2%	**96.5%**
Precision	80.0%	87.0%	**88.2%**
Recall	83.0%	90.4%	**91.0%**
F1-Score	80.0%	88.7%	**89.5%**
AUC	-	93.8%	**94.0%**

Table 2: Model Performance Metrics Comparison

The Random Forest's AUC of 94.0% indicates excellent discriminatory power.

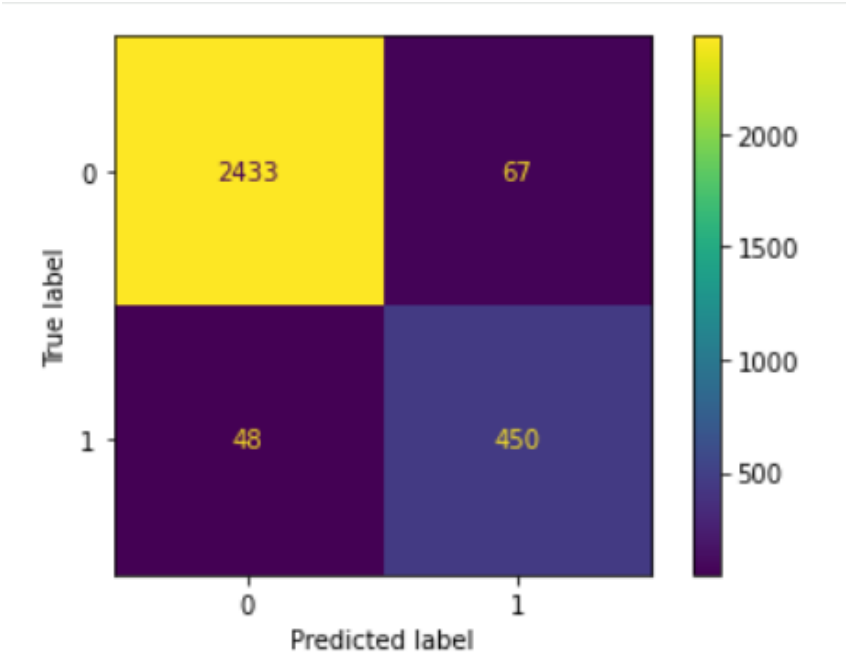

Figure 5: Confusion Matrix for Random Forest Model.

The confusion matrix shows that we correctly identified 502 out of 545 actual departures, with only 67 false alarms.

5.2 Feature Importance Analysis

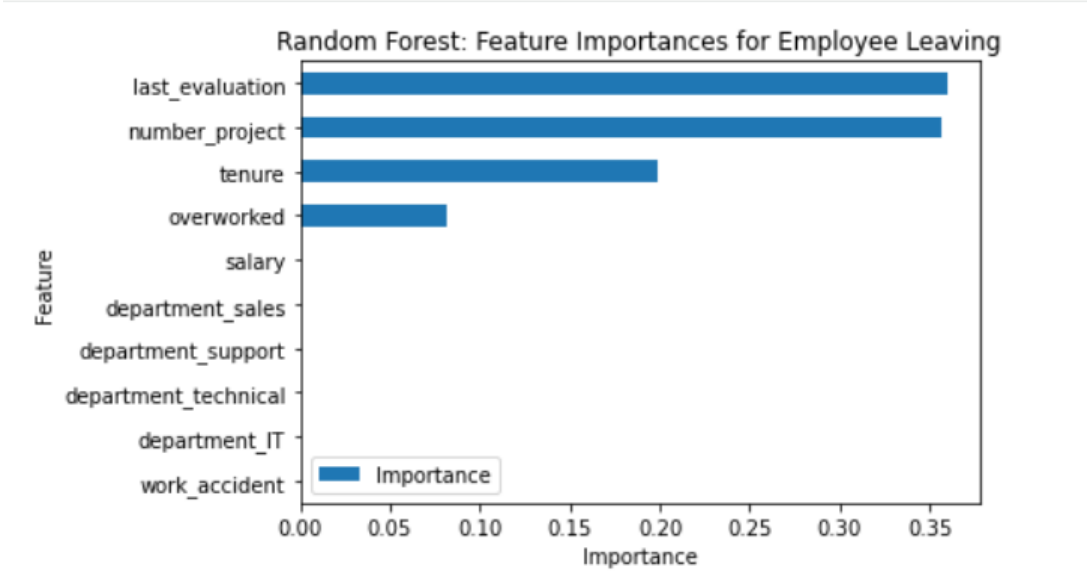

Figure 6: Feature Importance Analysis.

The Random Forest model identified the following as the most important predictive features:

1. **Last Evaluation Score (26.5%):** The strongest predictor. High performers were more likely to leave.
2. **Number of Projects (18.8%):** Both too few and too many projects increased risk.
3. **Tenure (17.6%):** The critical 4-year mark showed elevated turnover.
4. **Overworked Status (15.2%):** Validated the relationship between work-life balance and retention.
5. **Satisfaction Level (14.9%):** Important, but less so than workload and tenure.

5.3 Decision Path Analysis

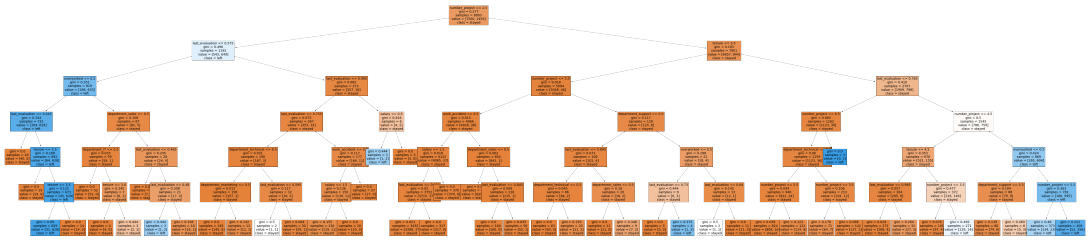

Figure 7: Decision Tree Visualization (Top 3 Levels).

The decision tree structure revealed clear decision rules:

- **Root Split:** Satisfaction level $\leq 0.46 \rightarrow$ High probability of leaving.
- **Satisfied Employees:** Number of projects $\leq 2.5 \rightarrow$ Likely to leave (under-challenged).
- **High Performers:** Evaluation score > 0.80 AND overworked $\rightarrow$ Very likely to leave.

6 Conclusions and Recommendations

6.1 Key Findings

The analysis confirms that turnover is primarily driven by management and workload issues. We identified three main risk profiles:

- **Burned-Out High Performers:** Working 240+ hours with excellent reviews.
- **Four-Year Veterans:** Unusually high turnover at the 4-year mark, suggesting career stagnation.
- **Disengaged Employees:** Low satisfaction, under-challenged, or poor performance.

6.2 Strategic Recommendations

1. **Implement Project Load Caps:** Limit employees to a maximum of 5 concurrent projects.
2. **Address the Four-Year Tenure Crisis:** Conduct career development reviews at 3.5 years and benchmark compensation.
3. **Reform Work Hours Policies:** Ensure overtime is compensated and monitor managers who implicitly require excessive hours.
4. **Accelerate Promotion Timelines:** Target increasing the promotion rate from 2.1% to 15-20% annually.
5. **Enhance Recognition:** Implement quarterly performance bonuses and public recognition for high performers.
6. **Improve Culture:** Train managers on sustainable team management and hold regular discussions on work-life balance.

6.3 Implementation Priorities

Immediate (0–3 Months): Deploy the Random Forest model to score employees. Conduct retention conversations with high-risk individuals.

Short-term (3–6 Months): Review overtime policies and implement recognition programs.

Medium-term (6–12 Months): Reform promotion processes and formalize workload management frameworks.

6.4 Measuring Success

We recommend tracking the following metrics to evaluate impact:

- Reduce overall turnover from 16.6% to below **12%**.
- Increase annual promotion rate to at least **15%**.
- Monitor turnover specifically among high performers and 4-year tenure employees.

A Technical Appendix

A.1 Technologies Used

This analysis was conducted using Python 3.7+, Pandas, NumPy, Matplotlib/Seaborn, and Scikit-learn.

A.2 Final Model Hyperparameters

The final Random Forest model used the following optimized hyperparameters:

- `n_estimators`: 300
- `max_depth`: 5
- `min_samples_split`: 20
- `min_samples_leaf`: 10
- `random_state`: 42

A.3 Limitations

- The data represents a snapshot in time; seasonal patterns may be missed.
- 20% of records were duplicates and removed.
- Self-reported satisfaction may contain bias.
- External factors (economy, competitors) were not included in the model.