# Employee Retention Analytics at Salifort Motors

Dwi Budi Setyonugroho
Data Analyst
East Java, Indonesia
Phone: +62 851 8611 1556
Email: setyonugrohodwibudi@gmail.com
Website: budinugroho.com
LinkedIn: linkedin.com/in/dwibudisetyonugroho

***Abstract*—Employee attrition creates direct financial costs, disrupts business continuity, and weakens organizational knowledge retention. This project documents a complete data science workflow for analyzing and predicting employee departure at Salifort Motors, a global alternative energy vehicle company facing a turnover rate of 23.8%. Using employee survey data containing 14,999 records, the study combines business understanding, exploratory data analysis, feature engineering, predictive modeling, and action-oriented interpretation. Two modeling approaches were evaluated: Logistic Regression as a baseline and Random Forest Classification as the primary non-linear model. The selected Random Forest model achieved 99.03% accuracy, 96.36% recall on employees who left, 99.57% precision, and a 97.94% F1-score on the test set. The analysis identifies satisfaction level, tenure, number of projects, and average monthly hours as the strongest drivers of attrition. The resulting documentation provides a professional, reproducible project narrative and concludes with strategic recommendations for workforce planning, workload balancing, promotion review, and compensation alignment.**

***Index Terms*—employee retention, attrition prediction, human resources analytics, random forest, logistic regression, classification, exploratory data analysis**

## I. Introduction

Employee retention is a high-value business problem because workforce instability affects productivity, hiring costs, domain expertise, and team performance. In data-rich organizations, predictive analytics can move retention strategy from reactive intervention to proactive decision-making. This document presents a professional data science project report based on an employee retention analysis for Salifort Motors.

The core objective is to determine why employees leave and to build a model capable of identifying at-risk employees before resignation occurs. The project is framed as an end-to-end applied data science initiative, integrating business objectives with interpretable statistical analysis and machine learning outcomes.

## II. Business Problem and Objectives

Salifort Motors reported a turnover rate of approximately 23.8%, indicating substantial loss of talent and associated replacement costs. Leadership requested an analytical solution that supports both diagnosis and action.

The project objectives were defined as follows:

- Quantify the scale of employee turnover and communicate its business significance.
- Identify the organizational and behavioral variables associated with employee departure.
- Develop a predictive model to classify employees as likely to stay or leave.
- Convert findings into retention actions for leadership teams.

## III. Dataset Description

The study uses the file `HR_capstone_dataset.csv`, which contains employee survey and organizational records. The dataset includes 14,999 observations and 10 primary features, with `left` serving as the binary target variable, where 1 indicates an employee who left and 0 indicates an employee who stayed.

Key explanatory variables include satisfaction level, salary category, department, number of projects, average monthly hours, tenure, promotion history, and work accident history. The dataset supports both descriptive analysis and supervised classification.

## IV. Methodology

The project follows a structured analytics workflow aligned with standard data science practice: data understanding, cleaning, exploratory analysis, preprocessing, modeling, evaluation, and business interpretation.

### A. Data Cleaning and Preparation

Initial data preparation focused on consistency and usability. Column names were standardized to `snake_case` to improve code readability and downstream reproducibility. A review of data integrity confirmed that the dataset contained no missing values. Numeric and categorical fields were verified before model preparation.

Categorical variables required encoding before model training. Department labels were transformed using one-hot encoding to preserve nominal structure, while salary level was mapped using ordinal encoding with the order low, medium, and high. The resulting modeling table contained 18 features after preprocessing.

### B. Exploratory Data Analysis

Exploratory analysis was performed to identify patterns associated with turnover and to generate hypotheses for feature relevance. The baseline attrition rate was measured at 23.81%, confirming that nearly one in four employees had left.

Correlation analysis showed a strong negative relationship between `satisfaction_level` and employee departure, suggesting that reduced satisfaction substantially increases attrition risk. Additional bivariate analysis highlighted the role of workload and compensation. Employees who left worked more hours on average, held more projects, and were disproportionately represented in the low-salary group.

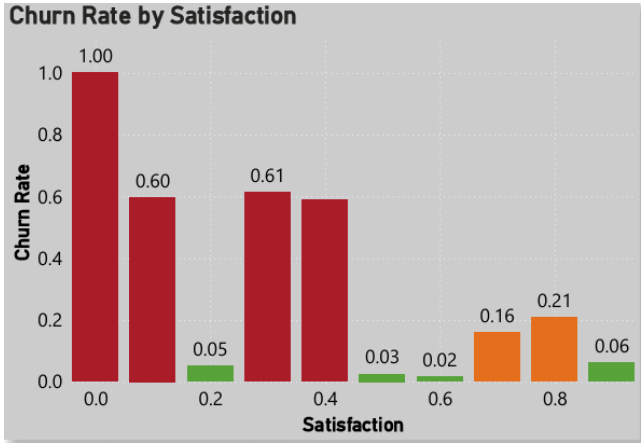


Fig. 1. Churn rate by satisfaction level, showing the inverse relationship between employee satisfaction and attrition risk.

Department-level analysis showed elevated turnover in Human Resources and Technical teams, while compensation analysis showed a large disparity between low-salary and high-salary employees. These findings informed both model design and management recommendations.

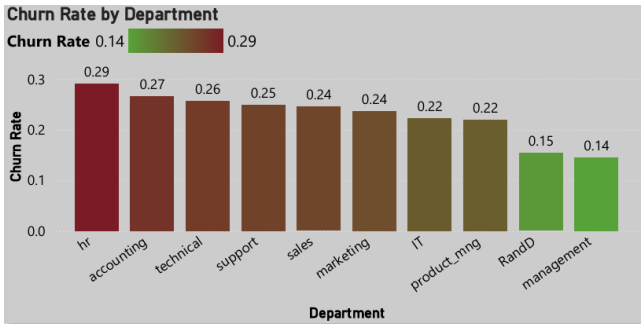


Fig. 2. Churn rate by department, highlighting the departments with the highest observed attrition.

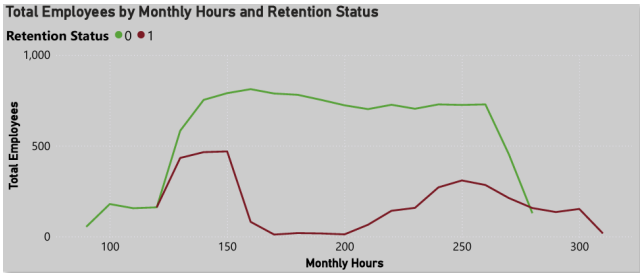


Fig. 3. Total employees by monthly hours and retention status, illustrating how working-hour patterns differ between employees who stayed and those who left.

### C. Train-Test Strategy

To ensure robust evaluation, the dataset was split into training and test subsets using an 80:20 stratified strategy. Stratification preserved the original class distribution in both subsets and reduced the risk of misleading metrics caused by class imbalance.

### D. Model Development

Two models were evaluated.

1) **Logistic Regression**: used as a transparent linear baseline for comparison.
2) **Random Forest Classifier**: selected as the primary candidate because it can model non-linear relationships, feature interactions, and mixed variable types effectively.

The baseline Logistic Regression model provided a useful benchmark but failed to recover a large share of actual leavers. The Random Forest model delivered substantially better recall while maintaining excellent overall accuracy and precision, making it more appropriate for retention intervention use cases.

## V. Results

### A. Performance Comparison

Table I summarizes the primary test-set metrics for the selected model.

TABLE I
RANDOM FOREST TEST PERFORMANCE

| Metric | Score | Interpretation |
|---|---|---|
| Accuracy | 99.03% | Correct predictions for nearly all employees |
| Recall | 96.36% | Identifies most employees who actually leave |
| Precision | 99.57% | Predicted leavers are almost always true leavers |
| F1-Score | 97.94% | Strong balance between recall and precision |

The evaluation emphasizes recall because the business cost of failing to identify an employee who is likely to leave is higher than the cost of reviewing a false positive. Under this objective, the Random Forest model is well aligned with operational HR decision support.

### B. Validation and Reliability

Five-fold cross-validation produced a mean accuracy of 98.95% with a standard deviation of 0.16%, indicating highly stable performance across folds. A departmental fairness review reported recall above 88% for all departments, suggesting that the model did not disproportionately underperform for any single organizational unit.

### C. Feature Importance

The feature importance analysis indicates that employee attrition at Salifort Motors is dominated by a small group of explanatory variables. Table II presents the strongest model drivers.

TABLE II
TOP DRIVERS OF EMPLOYEE DEPARTURE

| Feature | Importance | Business Meaning |
|---|---|---|
| Satisfaction Level | 34.9% | Low morale is the strongest attrition signal |
| Time Spent Company | 18.5% | Long tenure may indicate stagnation or burnout |
| Number of Projects | 16.6% | High workload contributes to departure risk |
| Average Monthly Hours | 13.5% | Excessive effort load is associated with churn |

The importance distribution indicates that retention risk is driven by a combination of employee sentiment, work intensity, and career progression context rather than compensation alone.

## VI. KEY ANALYTICAL INSIGHTS

Several insights emerge clearly from the analysis.

- Employees with low satisfaction scores represent the highest-risk group and should be monitored through regular engagement measurement.
- Employees carrying heavy project loads or extended working hours exhibit materially higher departure probability.
- Low-salary employees experience significantly greater turnover than high-salary employees, indicating a likely compensation effect.
- Long-tenured employees without recent advancement appear vulnerable to disengagement and attrition.

These patterns support a retention strategy that combines operational workload control with career development and employee experience initiatives.

## VII. BUSINESS RECOMMENDATIONS

Based on the empirical results, four strategic recommendations are proposed.

### A. Workload Management

Employees managing six to seven projects show extreme attrition behavior in some tenure segments. Salifort Motors should implement workload balancing policies, establish review thresholds for excessive project allocation, and prioritize intervention in Technical and Support teams.

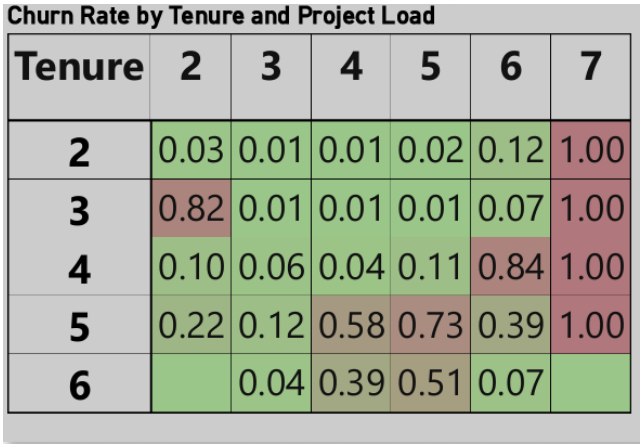


Fig. 4. Churn rate matrix by tenure and project load, highlighting high-risk workload-tenure combinations that support targeted intervention.

### B. Employee Engagement Programs

Because satisfaction is the strongest risk factor, leadership should deploy regular pulse surveys, manager coaching, and targeted intervention plans for employees whose satisfaction indicators fall below internal thresholds. This is especially important in teams with elevated historic turnover.

### C. Career Progression Review

Employees with long service and no recent promotion require formal development pathways. A structured promotion and mobility audit can reduce stagnation risk, particularly in R&D and Product-oriented functions where expertise retention is especially valuable.

### D. Compensation Equity Assessment

The high proportion of leavers in the low-salary group suggests that salary alignment is a major retention lever. A focused market review and internal equity assessment should be conducted for departments with persistent turnover concentration, especially HR and Technical teams.

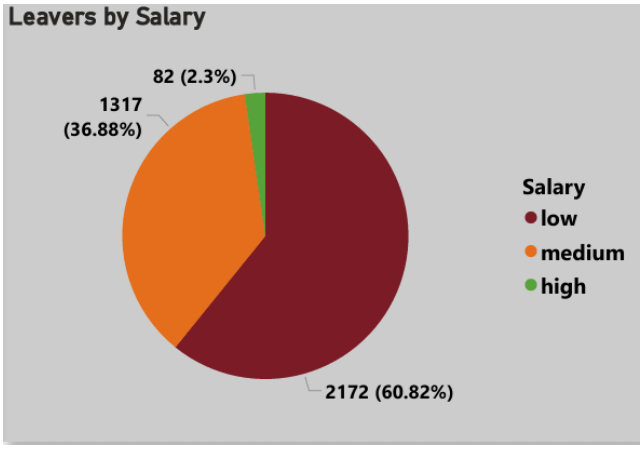


Fig. 5. Distribution of employees who left by salary level, highlighting the concentration of leavers in the low-salary group.

## VIII. Deployment and Monitoring

The project also includes a deployment pathway. The trained Random Forest model is stored as the artifact file `salifort_model.pkl`. This artifact supports integration into HR analytics workflows and dashboard pipelines. A Power BI dashboard tracks turnover by department, compensation, and workforce characteristics.

For sustainable performance, the model should be retrained on a regular schedule, such as quarterly, to address data drift and evolving workforce patterns. Monitoring should include predictive performance, class balance, feature shift, and intervention outcomes.

## IX. Reproducibility and Project Structure

The project repository is organized to support reproducibility and stakeholder handoff.

- `notebooks/project_2.ipynb`: notebook for EDA, preprocessing, and modeling.
- `models/`: serialized model artifacts and feature metadata.
- `reports/`: executive summaries and analytical outputs.
- `docs/`: supporting presentation materials.

This structure enables reproducible review of both technical implementation and business communication assets.

## X. Conclusion

This project demonstrates how a structured data science workflow can translate employee survey data into measurable business value. The final Random Forest model performs strongly on the task of identifying likely leavers, while the supporting analysis provides interpretable evidence for why attrition occurs. The most influential drivers are satisfaction, tenure, workload intensity, and compensation context.

From a leadership perspective, the project supports a shift from retrospective reporting to proactive retention management. By integrating predictive outputs with HR policy decisions, Salifort Motors can reduce turnover, improve employee experience, and better preserve organizational knowledge.