Employee Attrition Prediction Report

1. Executive Summary

In this project, I set out to build a model that can identify employees who are likely to leave the organisation. Employee attrition is costly and disruptive, so detecting risk early can help HR teams respond before it becomes a problem.

After exploring the data and testing several classification models, I found that overall accuracy alone was not enough. Because fewer employees leave than stay, I focused more on recall — how many actual attrition cases the model correctly identifies.

The final model selected was a tuned Random Forest with an adjusted probability threshold. This model was better at detecting employees who left, even though it produced more false positives. Based on the results, compensation, overtime, and tenure appear to be key factors associated with attrition.

2. Introduction

Employee turnover affects productivity, morale, and recruitment costs. My goal in this project was to determine whether employee characteristics could be used to predict attrition risk.

Rather than simply building a model with high accuracy, I aimed to build one that could meaningfully identify employees at risk of leaving. This required carefully comparing different models and evaluating them beyond surface-level metrics.

3. Data Overview

The dataset contains information on 1,470 employees, including demographic details, job-related information, and compensation data. The target variable, *Attrition*, indicates whether an employee left.

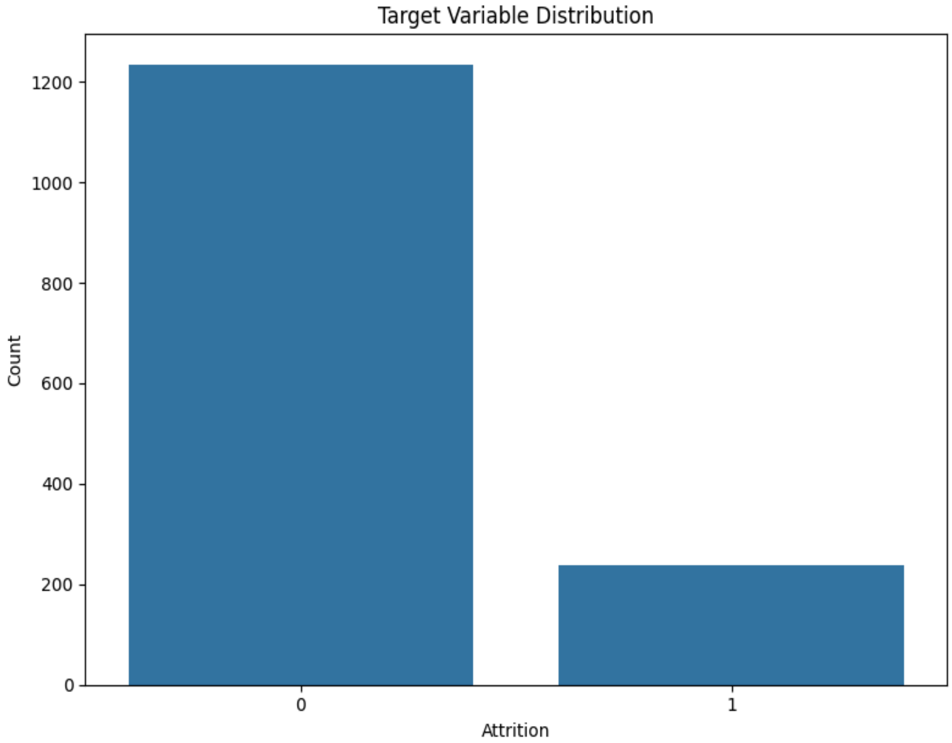

Fig.1 Distribution of Employees by Attrition Status

From the distribution above, it is clear that fewer employees left compared to those who stayed. Because of this imbalance, accuracy alone would not be a reliable measure of model performance.

4. Exploratory Insights

Before building any models, I explored the data to identify patterns that might explain attrition.

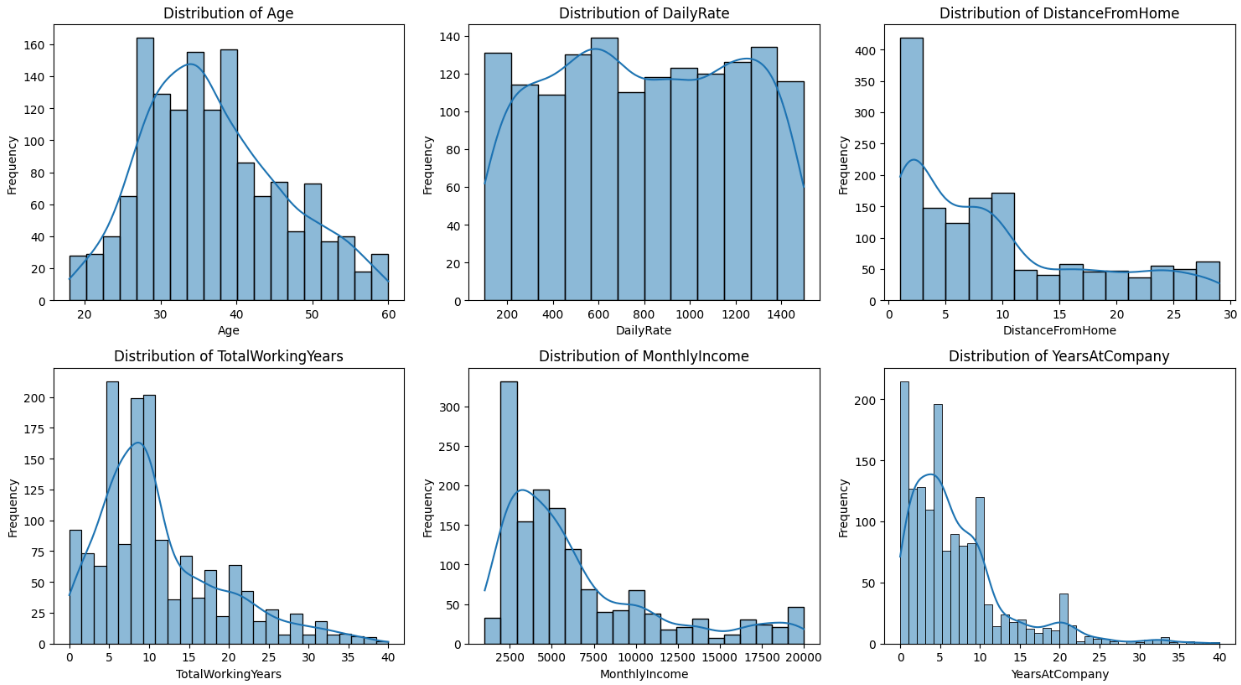

Fig.1   Distribution of Selected Numerical Features

During the exploratory analysis, I looked at how different employee characteristics were related to attrition before building the models.

Monthly income showed a noticeable pattern, as employees with lower income levels appeared more likely to leave. Age also seemed relevant, with younger employees showing higher attrition rates compared to older employees. In addition, employees who worked overtime (Overtime_Yes) were more likely to leave than those who did not.

These patterns helped me identify important variables for modelling. However, they reflect associations within the dataset and should not be interpreted as direct causes of attrition.

---

5. Modelling Approach

I treated this as a binary classification problem. The data was split into training and testing sets using an 80–20 split with stratification to preserve class balance.

I trained and compared several models:

- Logistic Regression

- Decision Tree
- Random Forest
- Gradient Boosting

Because the dataset was imbalanced, I focused primarily on recall. I also adjusted the classification threshold in the Random Forest model to improve the detection of attrition cases.

6. Model Comparison

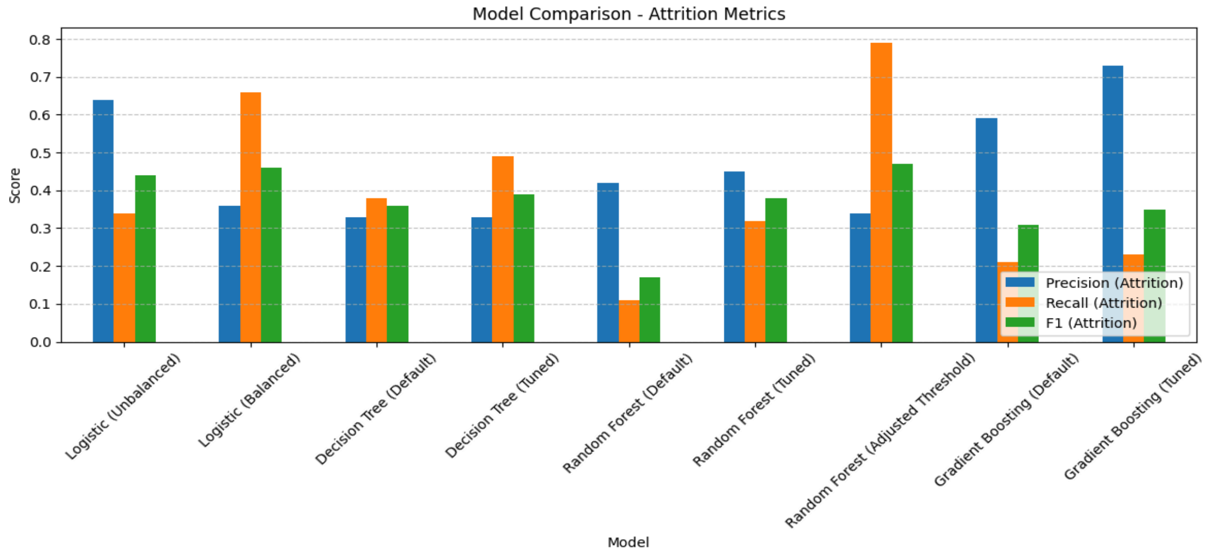

Fig.1 Comparison Between Models

In this study, I evaluated multiple models, including Logistic Regression, Decision Tree, and Random Forest. The comparison focused primarily on recall for the attrition class, as failing to identify employees at risk was considered more costly than generating false positives.

Logistic Regression served as a baseline model and produced reasonable results. However, its recall was lower compared to the tree-based methods. This suggests that the relationship between the predictors and attrition may not be strictly linear, limiting the model's ability to capture more complex feature interactions.

The Decision Tree improved recall but showed less stability, which is consistent with its tendency to overfit to specific patterns in the training data. In contrast, the Random Forest model delivered more consistent and balanced performance. By combining multiple decision trees, it reduced variance and improved generalisation, which explains its stronger overall results.

I also explored class balancing and probability threshold adjustment to improve the detection of the minority class. While these adjustments increased recall, they also led to a rise in false positives, highlighting the trade-off between sensitivity and precision.

Based on overall performance and alignment with the project objective, I selected Random Forest as the final model, as it provided the most appropriate balance between recall, stability, and generalisation.

7. Final Model Performance

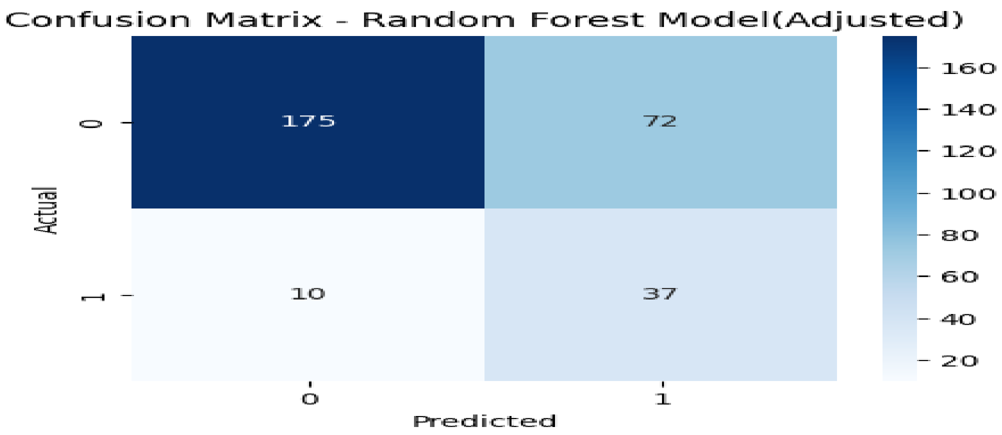

Fig.1 Confusion Matrix of Random Forest (Adjusted)

The final model correctly identified 37 out of 47 employees who left. Only 10 attrition cases were missed. Although the model generated more false positives, this trade-off was acceptable given the objective of reducing missed attrition cases.

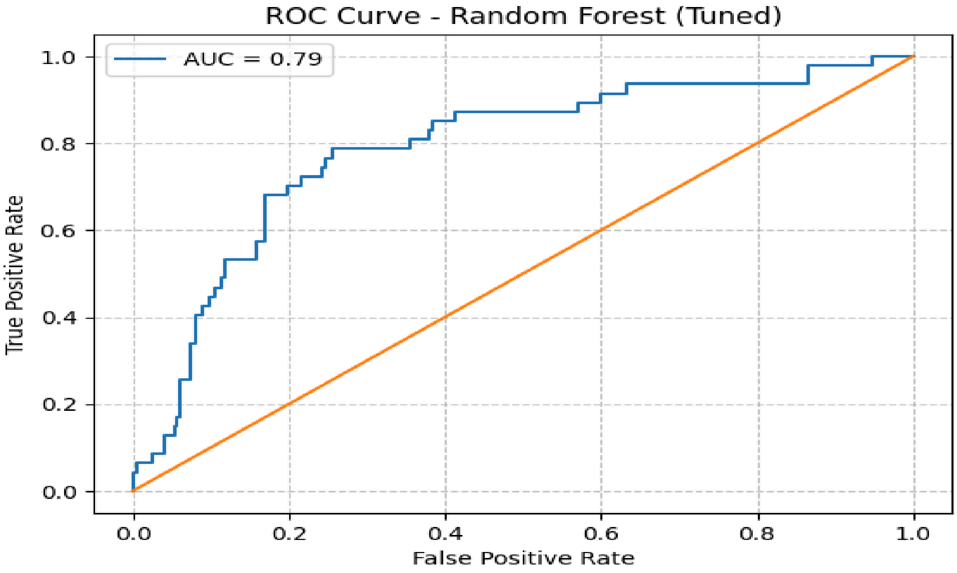

Fig.1   ROC Curve

The model achieved an AUC score of approximately 0.79, indicating that it performs reasonably well in distinguishing between employees who leave and those who stay.

8. Key Drivers of Attrition

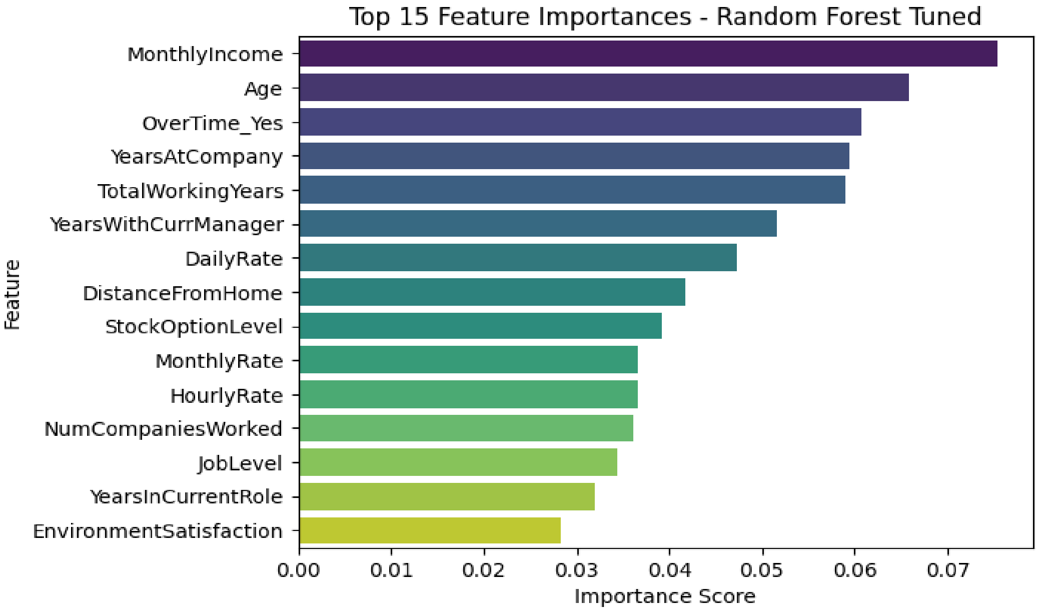

Fig.1 Feature Importance Of Model

Feature importance analysis shows that monthly income, overtime status, total working years, and tenure are among the most influential variables.

From this, I can conclude that compensation, workload, and career stage are important factors in understanding attrition risk.

9. Error Review

The final model produced 72 false positives and 10 false negatives.

Although some employees were incorrectly flagged as at risk, the model successfully reduced the number of employees who left without being identified. Given the objective of this project, given the objective of prioritising recall, this trade-off was considered acceptable.

10. Recommendations

Based on the findings, I would recommend:

- Monitoring employees who frequently work overtime.
- Reviewing compensation structures, especially for lower income bands.
- Strengthening onboarding and early-career engagement programs.
- Using the model as a support tool to guide HR interventions.

The model should inform decisions, not replace natural judgment.

11. Limitations and Future Work

This analysis is limited to the dataset used in this study. External factors such as economic conditions, organisational changes, or personal circumstances were not included, even though they may influence attrition decisions.

The dataset is also relatively small and represents a single organisational context, which limits how far the findings can be generalised. In addition, the data is cross-sectional, meaning the model cannot capture changes in employee behaviour over time.

It is also important to clarify that the model identifies patterns, not causes. For example, overtime and income were associated with attrition, but this does not mean they directly cause employees to leave.

Finally, model performance may change over time, so continuous monitoring and periodic retraining would be necessary in a real-world setting. Future work could explore more advanced validation techniques and cost-sensitive approaches to further improve reliability.

12. Conclusion

In this study, I developed and evaluated multiple classification models to predict employee attrition. After comparing their performance, Random Forest was selected as the final model because it provided the best balance between recall and overall stability.

The analysis showed that factors such as monthly income, age, and overtime were strongly associated with attrition in this dataset. While these findings provide useful predictive insights, they should not be interpreted as direct causes of employee turnover.

Overall, this project demonstrates how machine learning can be applied to HR analytics to support decision-making. However, the results should be used cautiously and complemented with organisational context and human judgement.