Google Advanced Data Analytics Capstone: Analyzing Employee Churn Risk

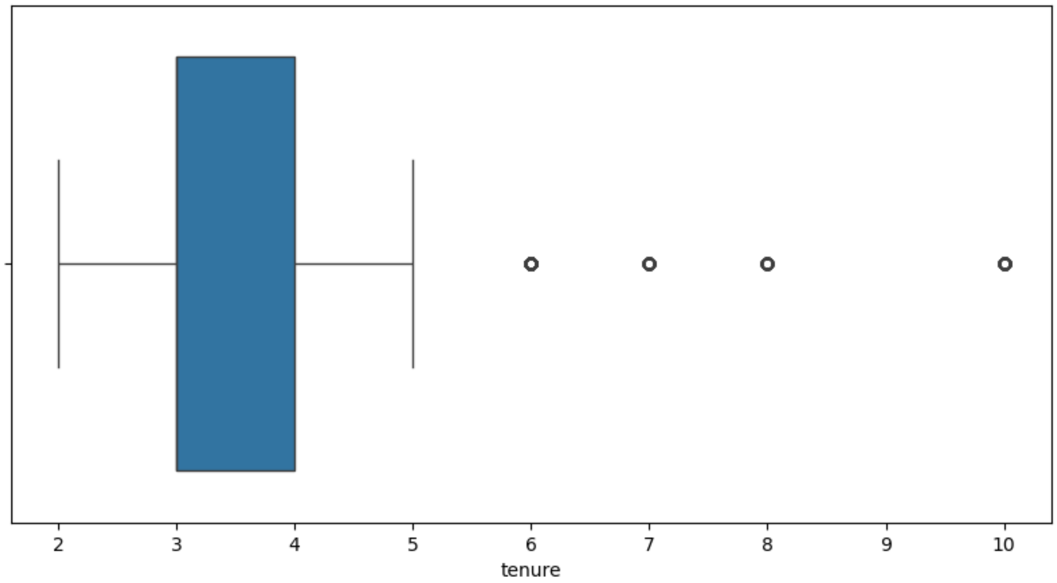
Author: Purinut Chairungrueang

Project Overview

This capstone project, part of the Google Advanced Data Analytics Professional Certificate, focuses on predicting employee churn. By analyzing key workplace factors, I developed a model to uncover critical 'red flags' in employee behavior. The objective is to provide actionable insights that explain the root causes of attrition, enabling data-driven retention strategies rather than just reactive measures.

Exploratory Data Analysis (EDA) & Key Insights

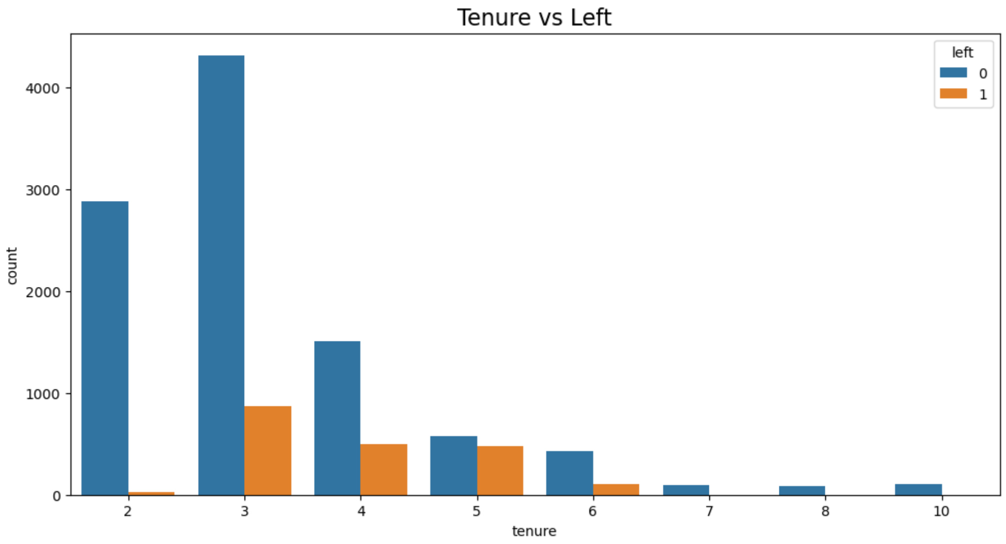
Before jumping into the predictive modeling, I conducted a thorough exploratory analysis to understand the underlying patterns. A critical step was Data Cleaning, where I identified and removed outliers in the 'Tenure' variable. By removing extreme values (employees with unusually long years of service), the analysis focuses on the core workforce where turnover trends are most actionable.



Here are the key findings from the refined dataset:

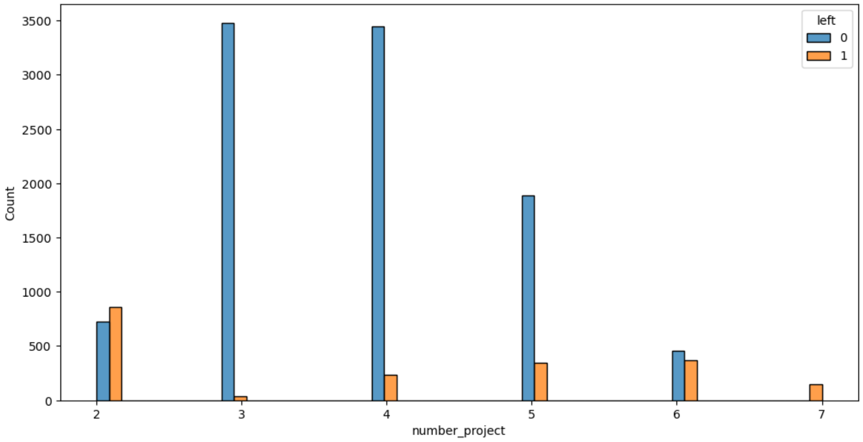
1. The Retention Danger Zone (Tenure)

- Insight: After removing outliers, the data clearly shows that employees are most likely to leave between their 3rd to 5th year with the company.
- Observation: There is a sharp spike in resignations at year 3, peaking at year 5, before dropping significantly. This suggests that the 3-5 year window is the most critical period for retention efforts.



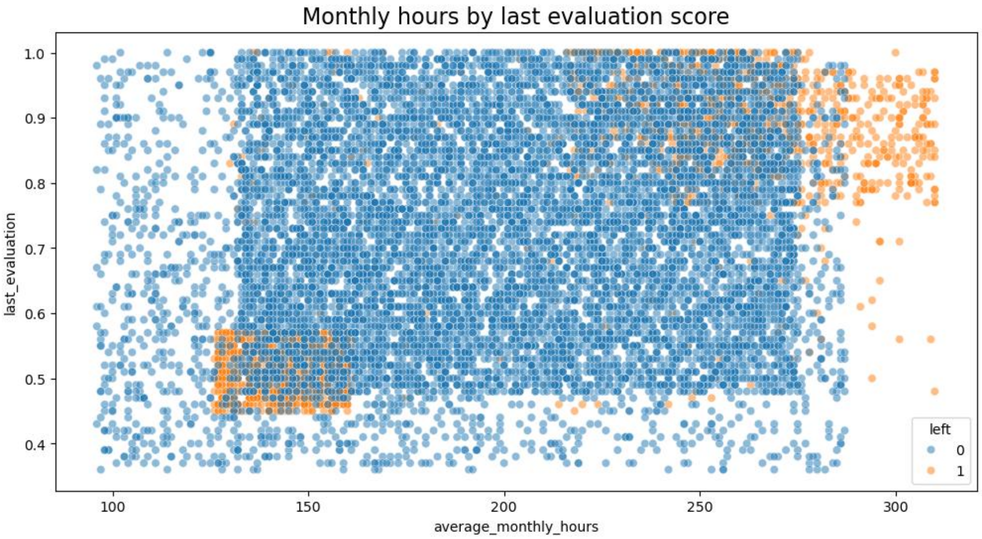
2. Workload and the U-Shaped Risk

- Insight: Both Under-worked and Over-worked employees are high-risk groups.
- Observation: Employees handling only 2 projects and those handling 6-7 projects show the highest attrition rates. The Sweet Spot appears to be 3-4 projects, where employees are most stable.



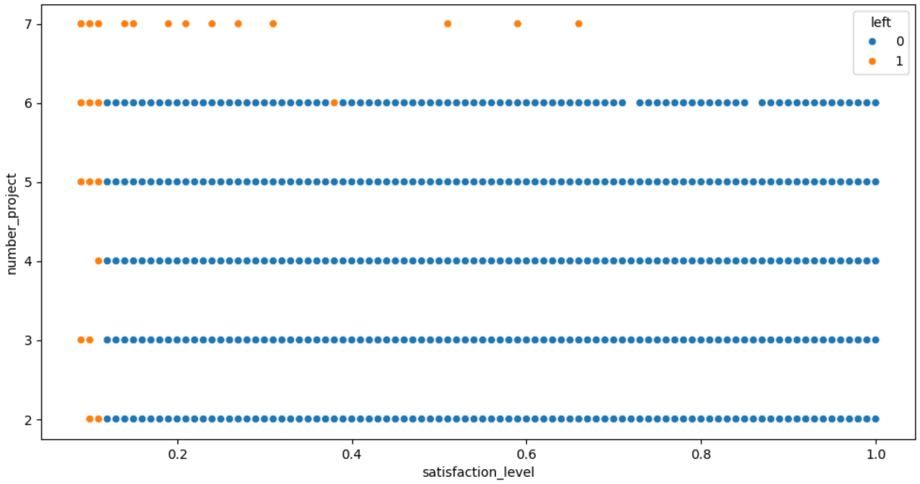
3. The Burnout of High Performers

- Insight: High performance does not guarantee retention.
- Observation: A concerning cluster of employees who left had High Evaluation scores (>0.8) but also High Monthly Hours (>250). This suggests that the company is losing its top talent due to burnouts, rather than lack of performance.

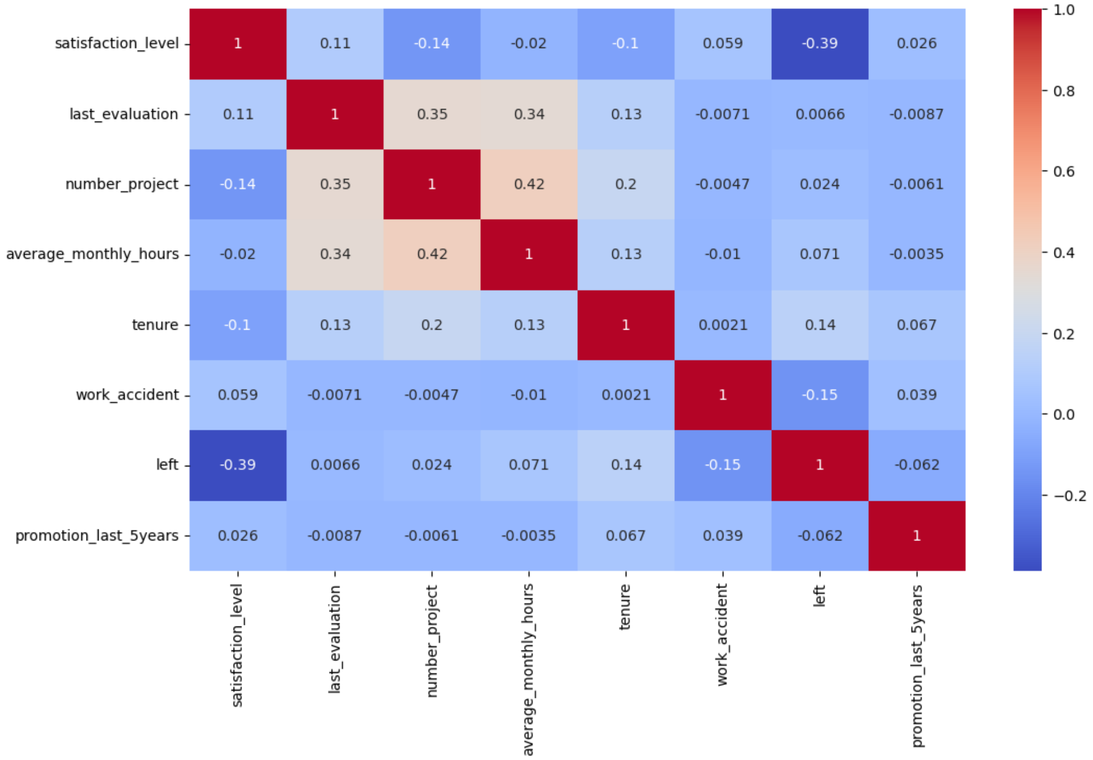


4. Satisfaction as the Primary Indicator

- Insight: Satisfaction level remains the strongest individual predictor of churn.
- Observation: There is a distinct threshold; employees with a satisfaction score below 0.4 are significantly more likely to leave, regardless of their department or salary level.



5. Correlation Matrix: Identifying the Key Drivers



Baseline Modeling and Initial Results (Before Feature Engineering)

In this phase, I established a performance baseline using the original dataset features (including satisfaction_level). To ensure a comprehensive analysis, I selected three distinct machine learning algorithms to compare their effectiveness in predicting employee attrition.

1. Model Selection

- **Logistic Regression:** Used as a simple, linear baseline to understand the basic relationships between features.
- **Decision Tree:** Selected for its ability to capture non-linear patterns and provide high interpretability.
- **Random Forest:** An ensemble method used to improve prediction stability and handle complex data interactions more effectively than a single tree.

2. Methodology

The dataset was split into Training (75%) and Testing (25%) sets. All models were trained on the same training data, and their performance was evaluated using the hold-out test set to ensure an unbiased comparison.

3. Baseline Performance Metrics

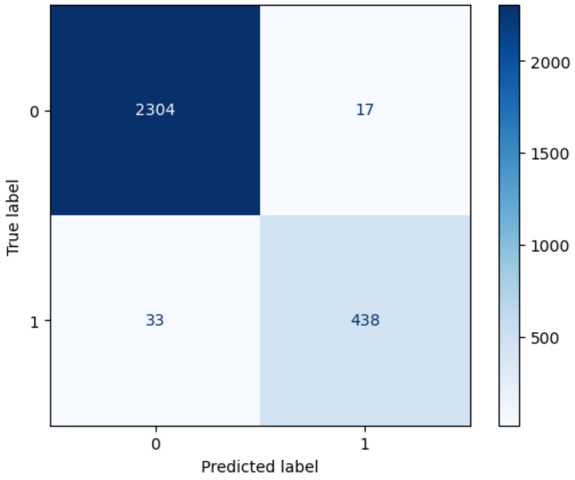
Using the get_scores function, the initial results are summarized below:

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	81.90%	0.44	0.26	0.33	-
Decision Tree	98.20%	0.96	0.93	0.95	0.96
Random Forest	**98.50%**	**0.98**	**0.93**	**0.96**	**0.96**

4. Model Diagnostics (Confusion Matrix & Feature Importance)

To gain deeper insights, I analyzed the Confusion Matrix and Feature Importance for the tree-based models:

- Confusion Matrix: Both Decision Tree and Random Forest showed exceptional ability in identifying "at-risk" employees with very few errors.



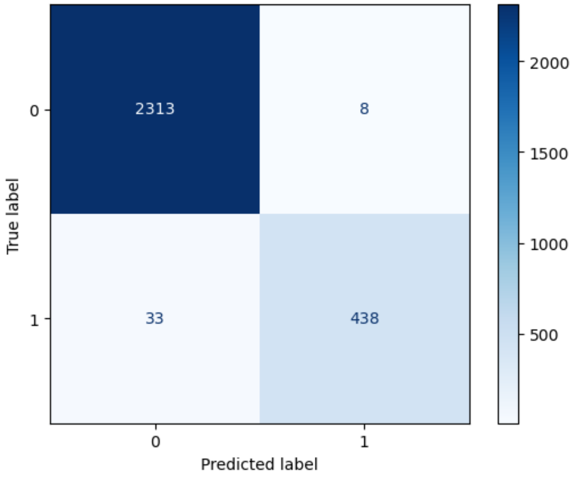
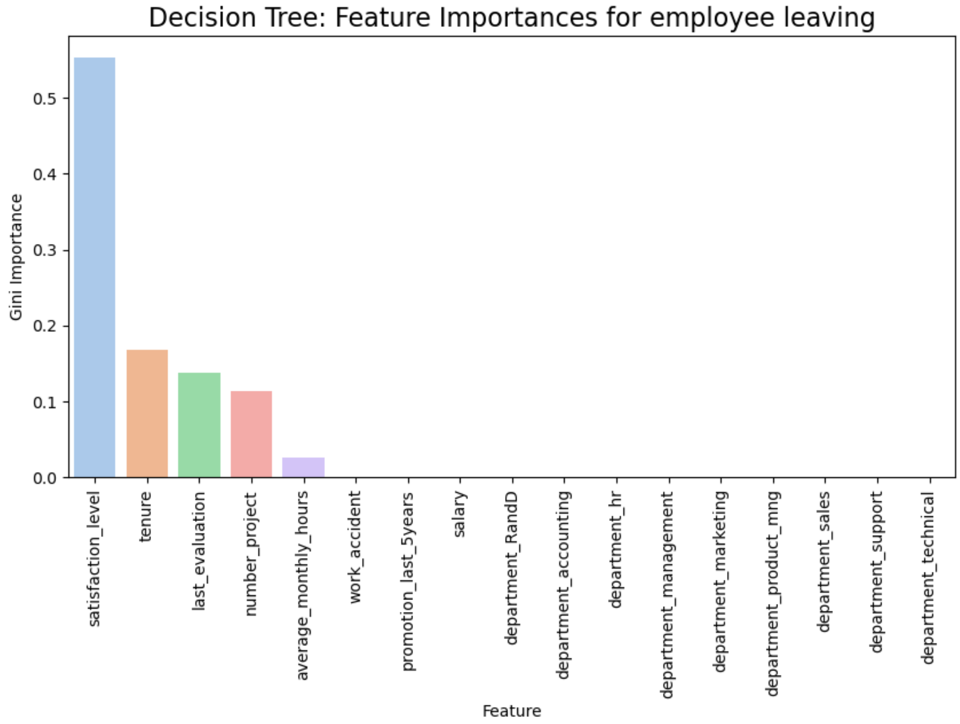
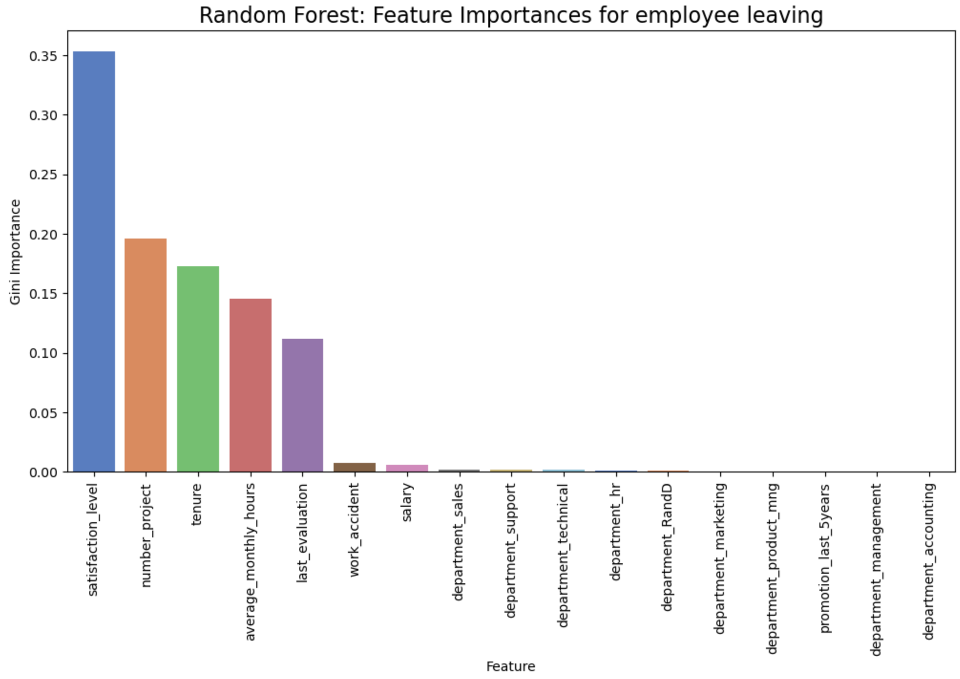


Figure 1: Decision Tree - Confusion Matrix

Figure 2: Random Forest - Confusion Matrix

- Feature Importance: The baseline analysis revealed that satisfaction_level was the most influential predictor by a significant margin, followed by working hours and tenure.





Advanced Feature Engineering & Model Refinement

After evaluating the baseline models, I identified a key limitation: the high reliance on subjective survey data (satisfaction_level). To create a more robust and proactive tool for HR, I shifted the focus towards objective behavioral metrics.

1. The Strategic Shift: Dropping Subjective Metrics

While satisfaction_level is a strong predictor, it has several real-world drawbacks:

- Survey Bias: Employees may not always provide honest feedback if they fear repercussions.
- Lagging Indicator: Satisfaction surveys are conducted periodically, whereas behavioral data (hours, projects) is updated in real-time.
- Goal: By removing this feature, the model learns to predict attrition based on observable patterns rather than self-reported feelings.

2. Engineering the Burnout Index

To compensate for the removal of satisfaction scores, I developed a new feature to capture the High-Performance, High-Stress phenomenon:

- Feature Name: burnout_index
- Calculation: $last\_evaluation * average\_monthly\_hours$
- Rationale: This index identifies employees who are performing at a high level but are also working extreme hours. These "high-performers" are often the most valuable assets to a company but are also the most prone to sudden resignation due to exhaustion.

3. Workload Optimization (hours_per_project)

I also calculated the average hours spent per project to identify:

- Overwhelmed Employees: Those with high hours but too many projects.
- Under-utilized Employees: Those with very few projects, which often leads to disengagement and eventual departure.

Final Model Evaluation (Post-Feature Engineering)

In this final phase, the models were refined by removing the subjective satisfaction_level and incorporating engineered features like the Burnout Index. This transition ensures that the model remains a proactive tool based on objective behavioral data.

1. Training & Cross-Validation Results

To ensure the model’s stability, I utilized 4-fold cross-validation (CV=4). This method confirmed that the model performs consistently across different segments of the dataset, providing a reliable foundation for real-world predictions.

2. Final Performance on Test Set (Random Forest 2)

Metric	Score	Interpretation
Accuracy	**97.26%**	High overall correctness in predicting employee status.
Precision	**0.9407**	94% of predicted leavers actually leave.
Recall	**0.8916**	The model successfully captures ~89% of all actual attrition cases.
F1-Score	**0.9155**	Excellent balance between Precision and Recall.
AUC-ROC	**0.9402**	Strong ability to distinguish between "Stay" and "Left" classes.

Conclusion & Strategic Recommendations

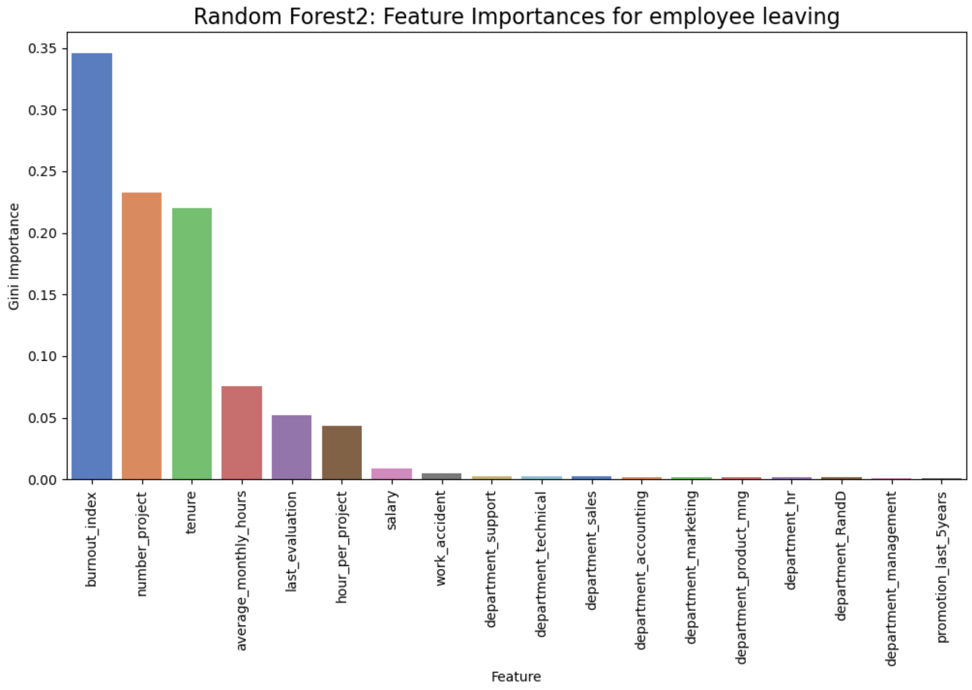
1. Project Summary

Achieved a 91.5% F1-score by using an objective 'Burnout Index' instead of subjective survey data. This confirms that behavioral patterns like workload intensity and tenure are sufficient to accurately predict employee attrition.

2. Key Predictors of Attrition

The final model identifies three primary risk factors:

1. High Burnout Index: Top-performing employees working excessive monthly hours (>250 hours) are the most likely to resign suddenly.
2. Critical Tenure (3-5 Years): Employees in this mid-career window are at a peak "transition point" and require higher engagement.
3. Project Overload/Underload: Employees assigned to either too few (2) or too many (6+) projects show a higher tendency to leave.



3. Strategic Recommendations

- Workload Rebalancing: Use the Burnout Index to trigger capacity audits and redistribute tasks, ensuring top talent is not over-leveraged.
- Retention Focus (Year 3-5): Conduct Stay Interviews at the 3-year mark to discuss career growth and re-engage talent before the 5-year exit window.
- Optimal Project Load: Maintain a sweet spot of 3-4 projects per employee to balance engagement and prevent exhaustion.