Employee Turnover Prediction Project

Ting-Chun Adam Wang

Executive Summary

Objective

Developed a predictive model for Salifort Motors to forecast employee turnover and identify key factors influencing attrition.

Data

Analyzed data from 11,991 employees, focusing on variables such as satisfaction level, number of projects, average work hours, and other relevant features.

Statistical test

Statistical tests showed significant differences between employees who stayed and those who left, providing valuable insights for targeted retention strategies.

Executive Summary (Continued)

Model

A Random Forest model was built, achieving a ROC AUC score of 96.5% and an accuracy of 96.2%.

Key Features

Top predictors of employee turnover include ‘last evaluation in 5 years’, ‘number of projects’, and ‘tenure’.

Application

The model will assist HR at Salifort Motors in developing data-driven policies to enhance employee retention and improve job satisfaction for both current and future employees.

Key initiatives include reducing project overload and initiate reward program for employees with good performance.

Introduction

Data

- Dataset from Kaggle
- Analyze using Python
- Information of 11,991 employees:

  Satisfaction Level,

  Number of Project,

  Average Work Hours,

  Department,

  Salary,

  … and more

Introduction

Exploratory Data Analysis (EDA)

Numpy, Pandas, Matplotlib, Seaborn

Statistical Test

Scipy

Machine Learning Model Selection

Sklearn

Exploratory Data Analysis

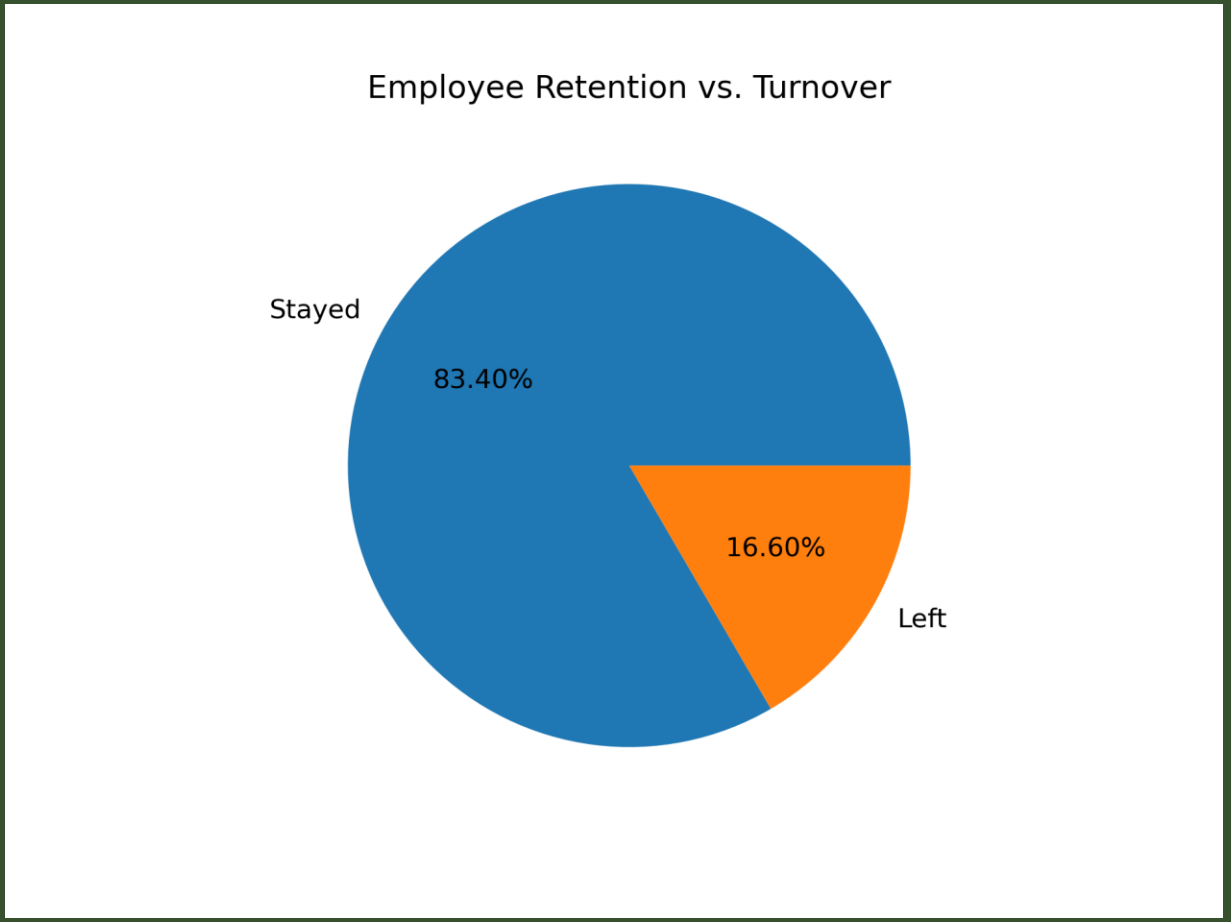
Stayed employees:

10000

Left employees:

1991

About 16.6% employee left



Exploratory Data Analysis

(-) Negative correlation:

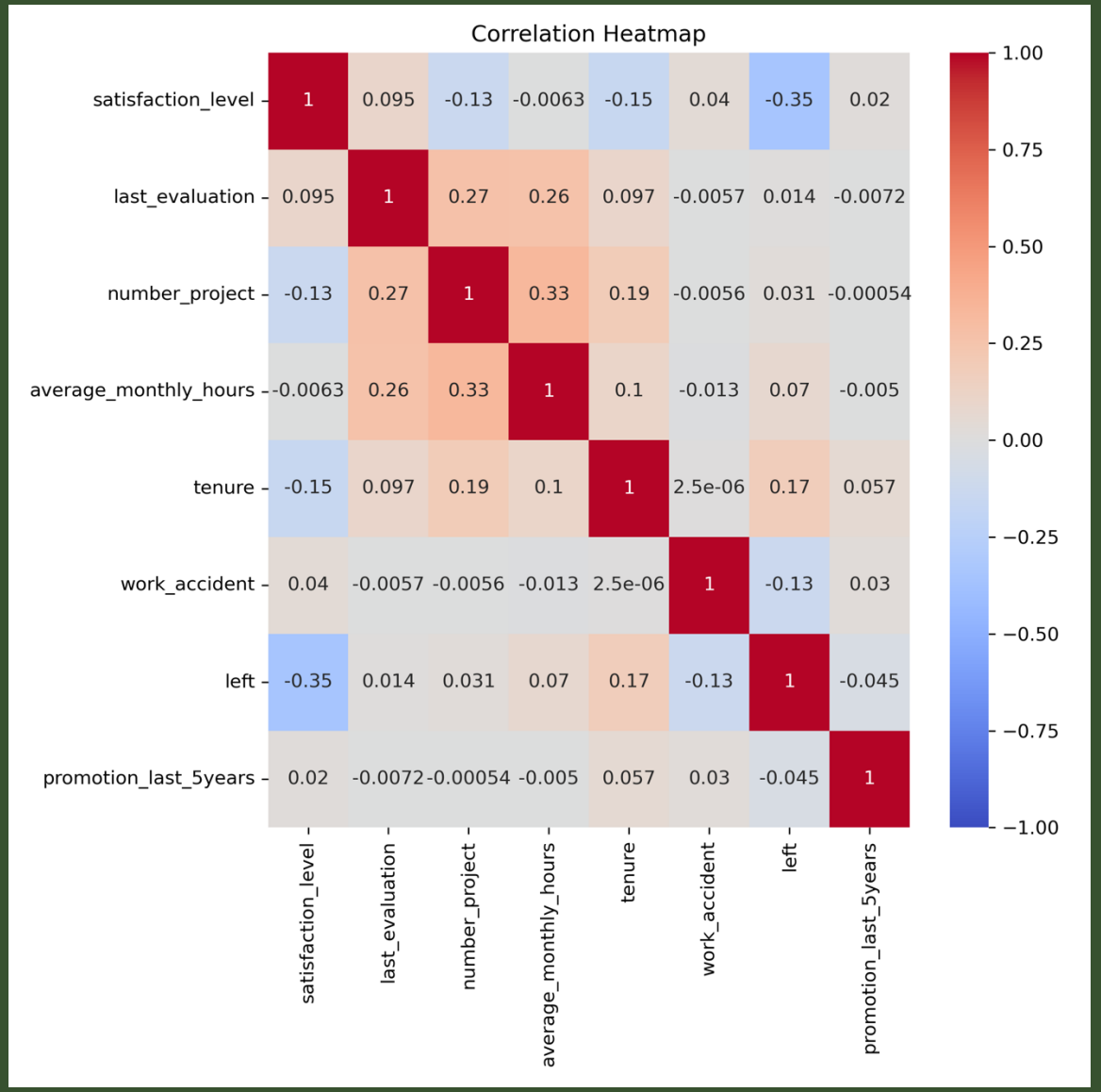
Satisfaction level & Left

(+) Positive correlation:

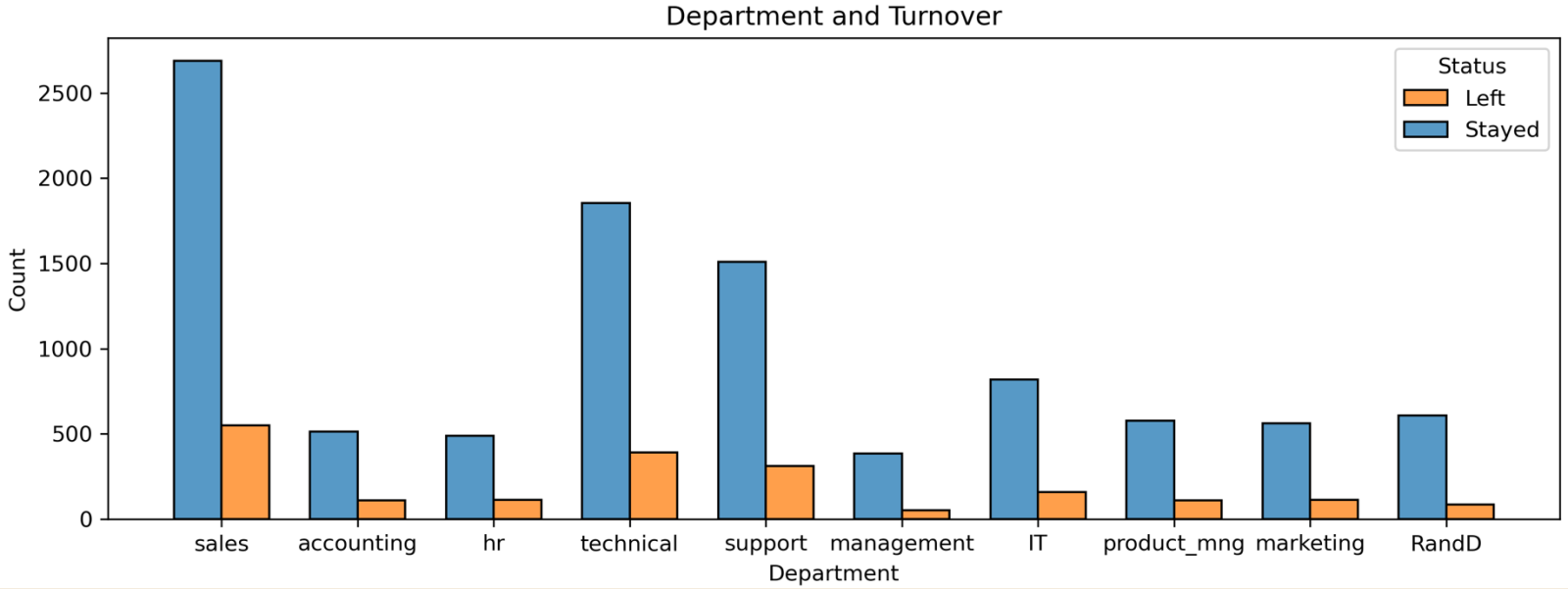
Last evaluation & Number of project,

Last evaluation & Monthly work hours,

Number of project & Monthly work hours



Exploratory Data Analysis



No significant differences in turnover rate were observed between departments.

Dept.	Stayed	Left	Left%
IT	818	158	16.19%
R & D	609	85	12.25%
Accounting	512	109	17.55%
HR	488	113	18.80%
Management	384	52	11.93%
Marketing	561	112	16.64%
Product Management	576	110	16.03%
Sales	2689	550	16.98%
Support	1509	312	17.13%
Technical	1854	390	17.38%

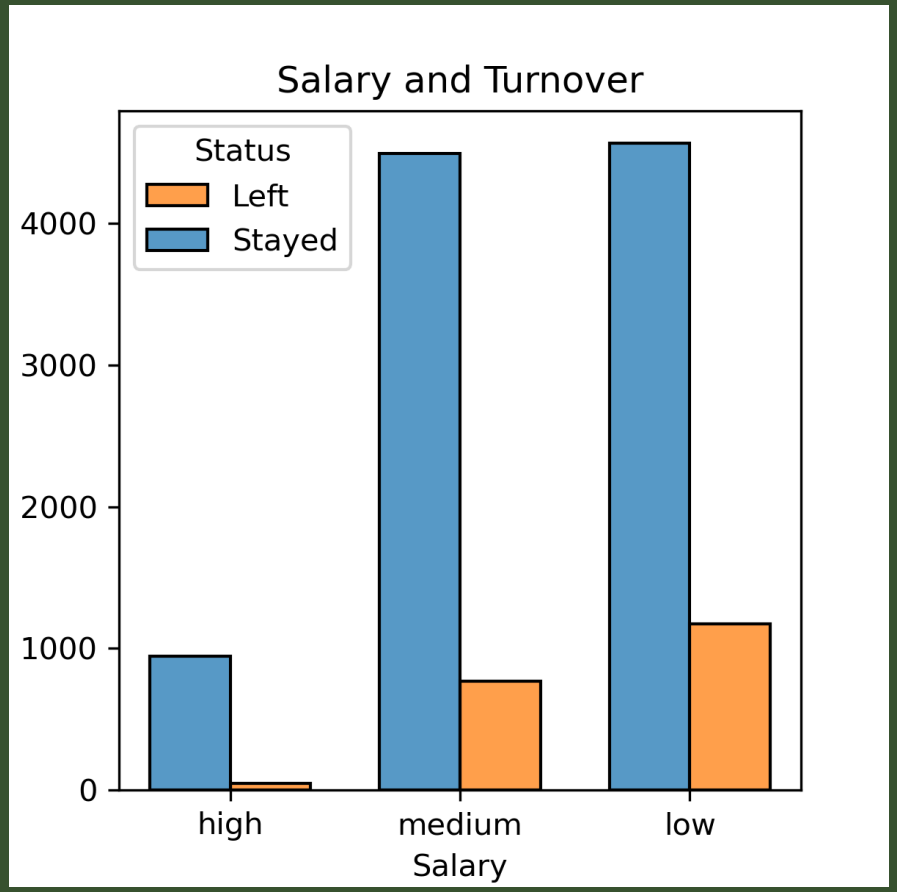
Exploratory Data Analysis

Turnover rate increases as salary decreases.

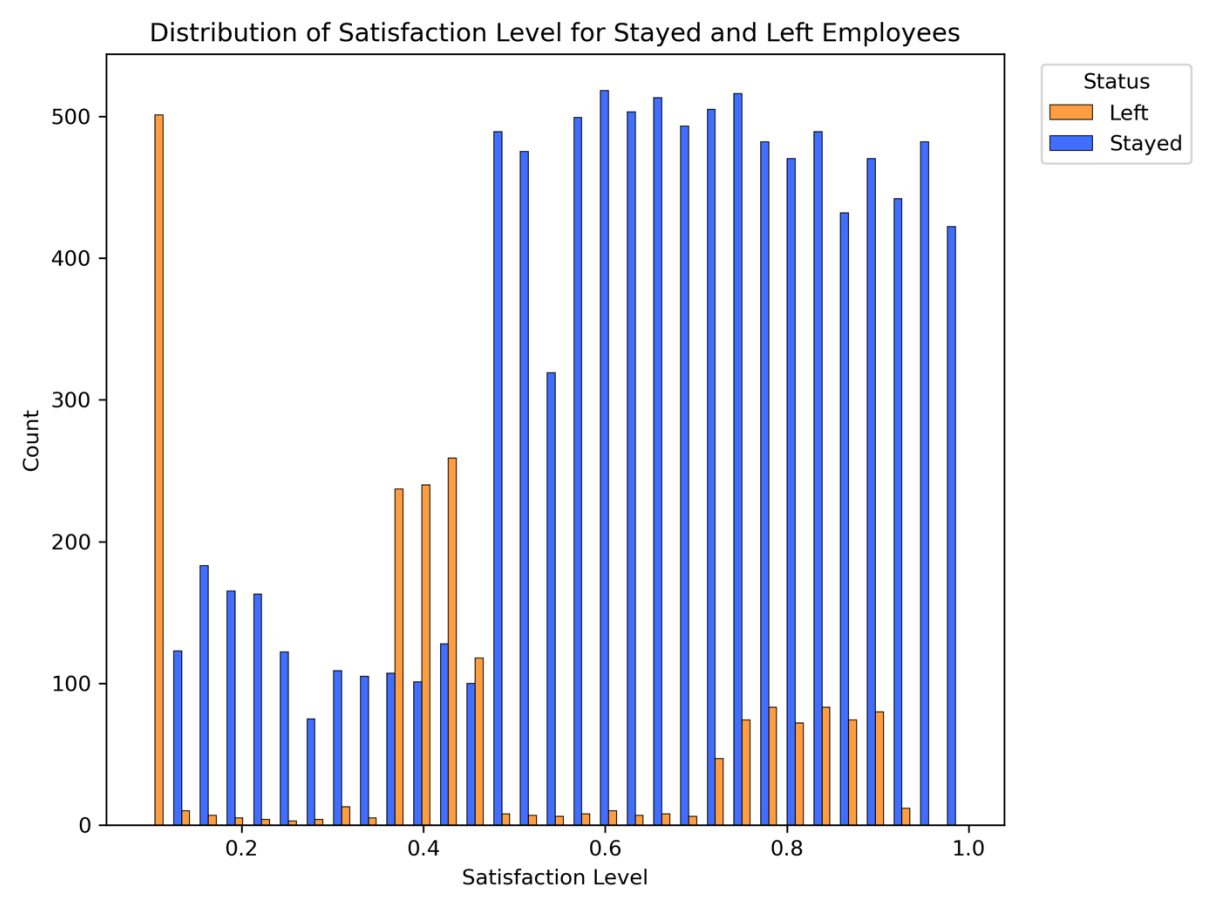
High-paid employees: Only 4.85% resigned.

Low-paid employees: Over 20% resigned.

Turnover Rate by Salary



Employee Turnover by Satisfaction Levels



Exploratory Data Analysis

Three Distinct Groups for Left Employees:

Satisfaction levels below 0.2

Satisfaction levels around 0.4

Satisfaction levels from 0.7 to 0.9

- Employees with higher satisfaction levels might have left for reasons other than dissatisfaction.

Exploratory Data Analysis

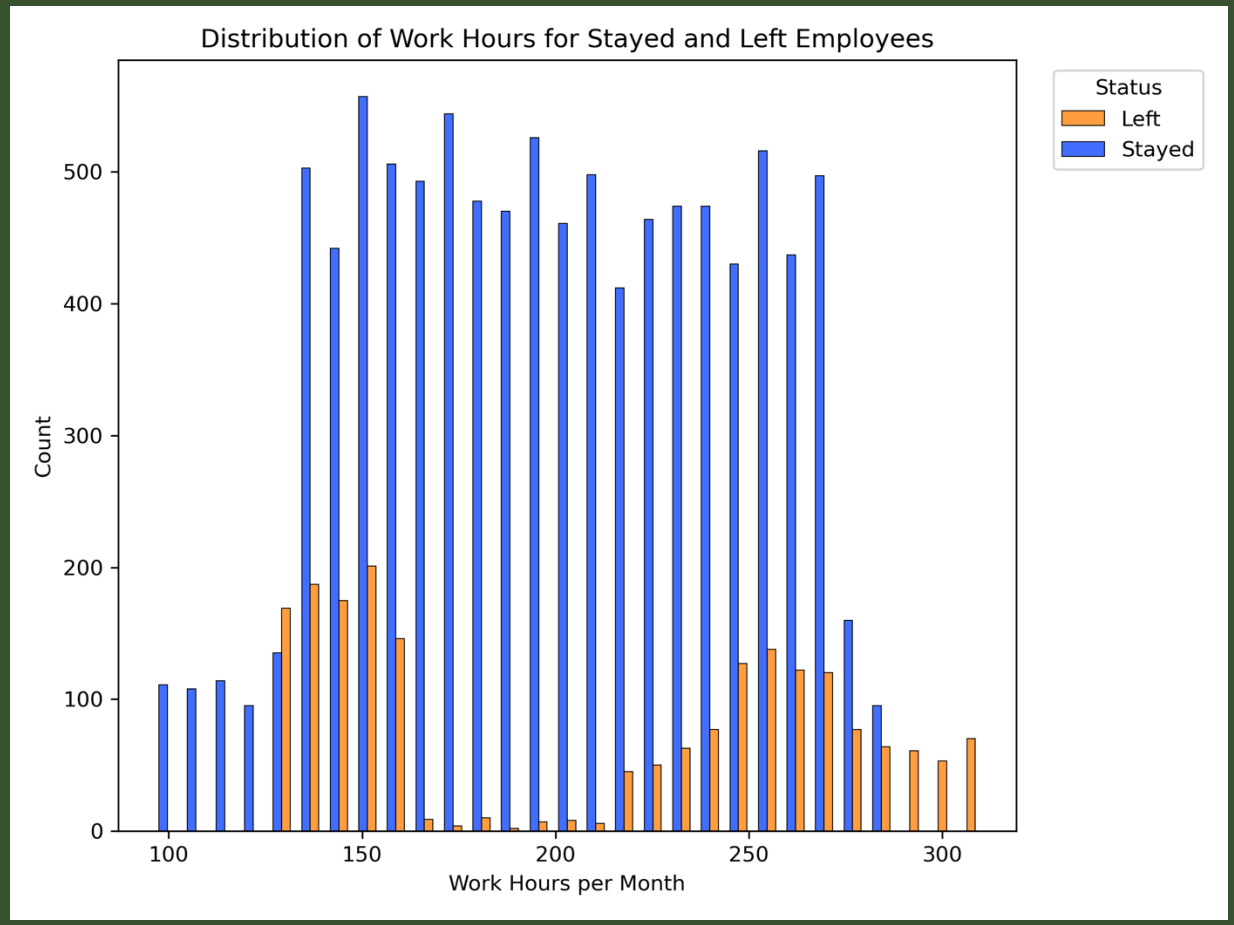
Two Distinct Groups for Left Employees:

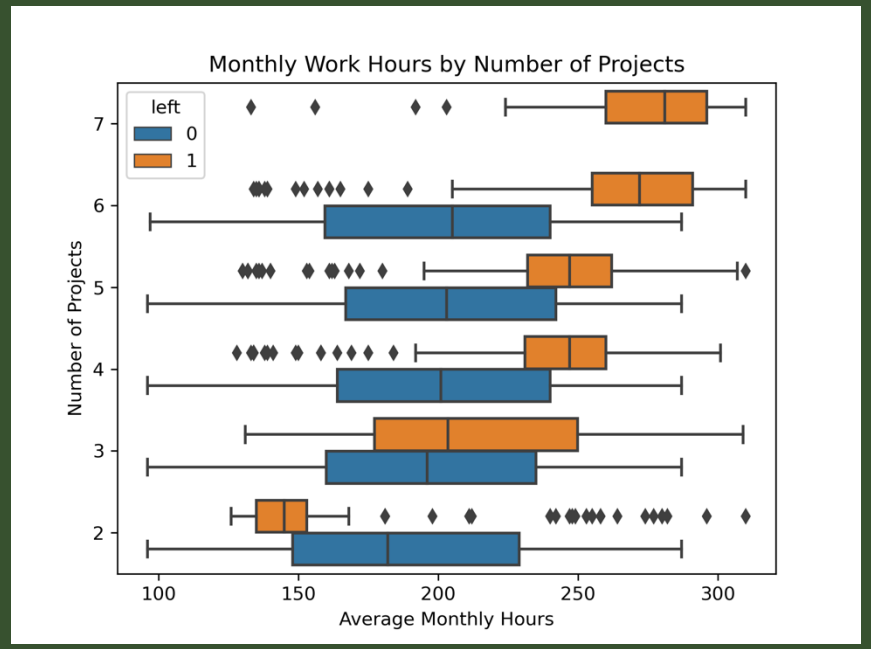
Employees working around 150 hours per month

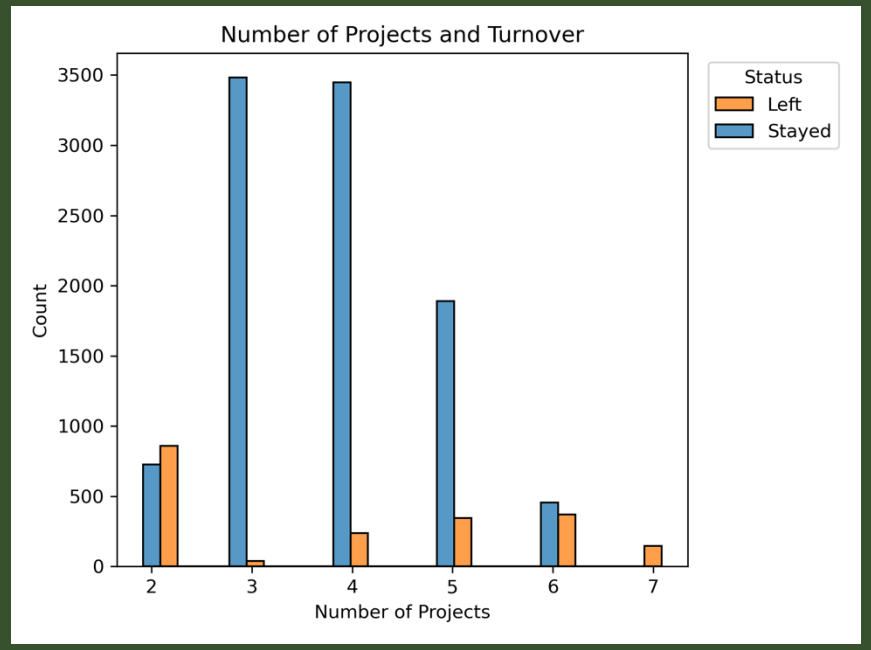
Employees working 200 to over 300 hours per month

Employees working 170-200 hours per month show better outcomes.

Turnover Rate by Work Hours







Exploratory Data Analysis

Pattern :

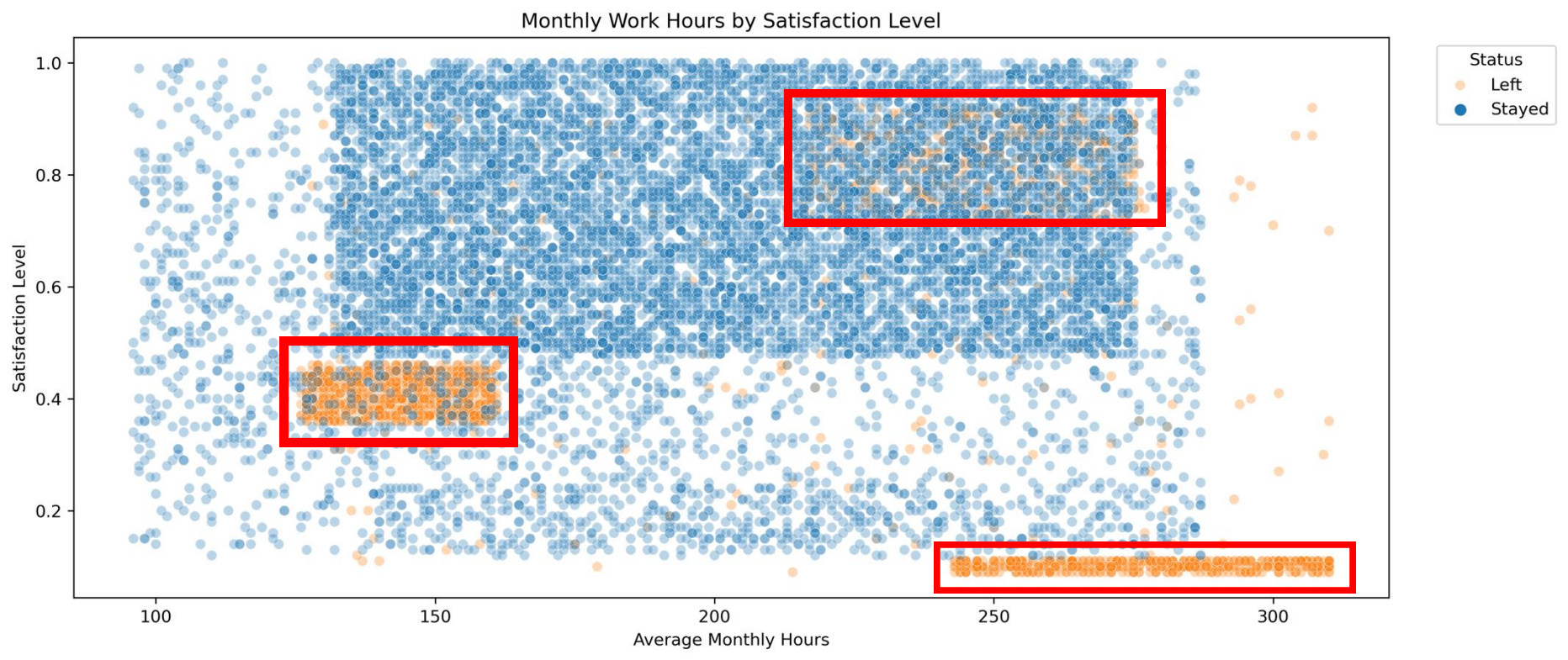
All employees handling 7 projects left the company.

Left employees with 6, 7 projects have excessive work hours.

More than half of employees working on 2 projects left.

3 or 4 projects seem to be the ideal balance.

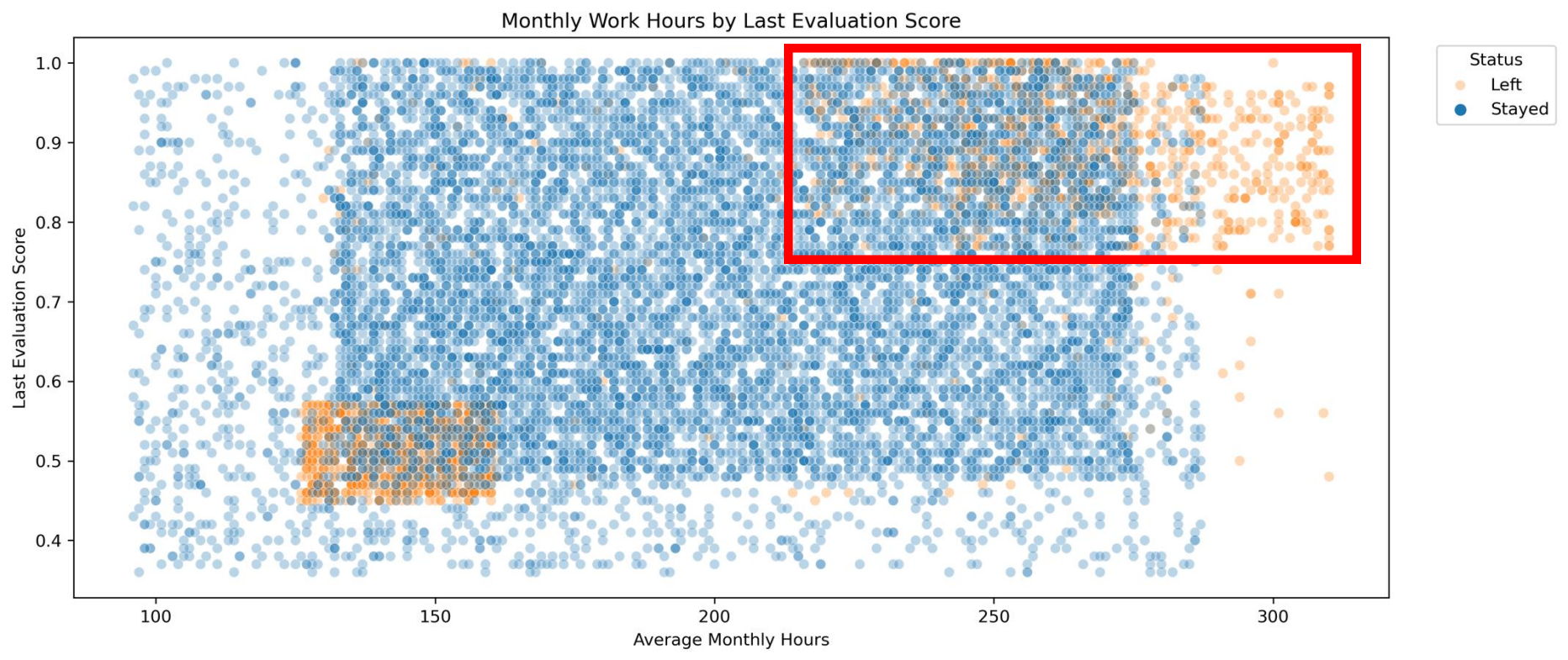
Exploratory Data Analysis



Three Distinct Groups for Left Employees:

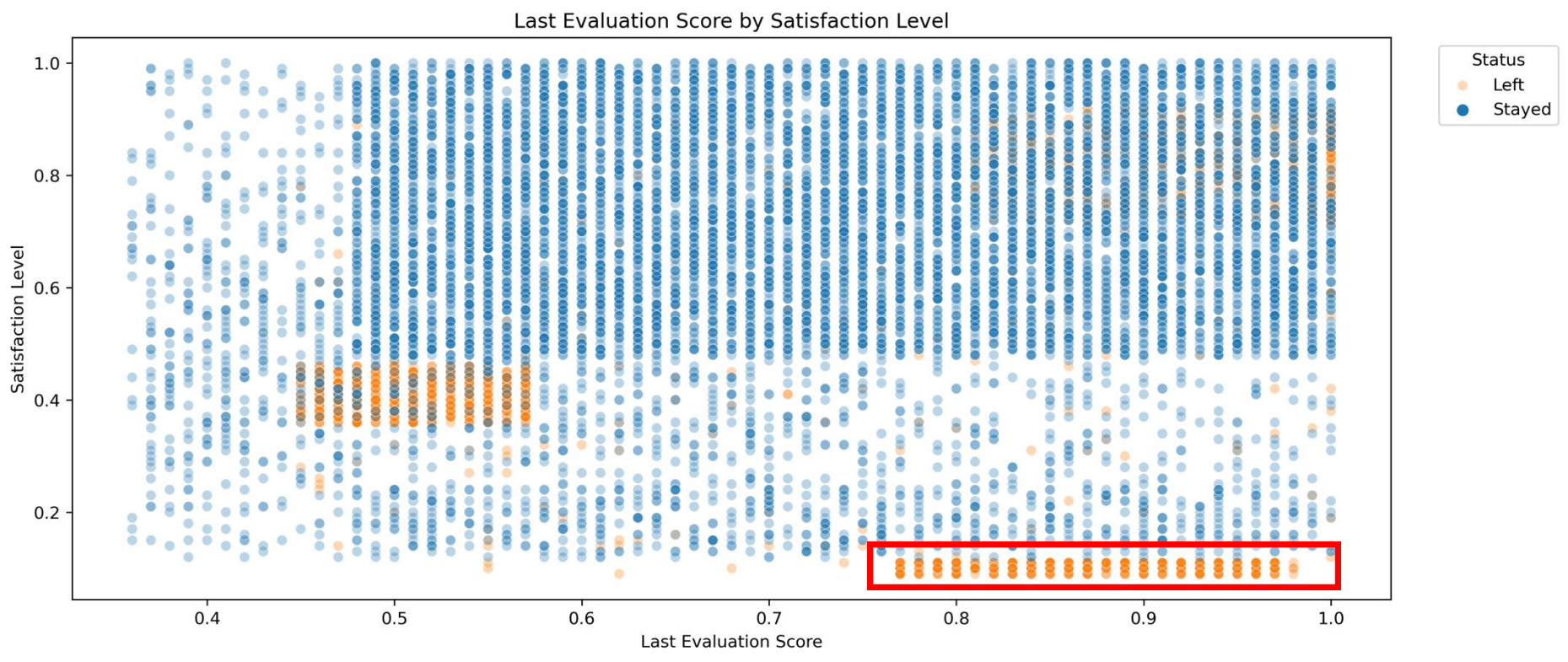
High work hours and low satisfaction level
High work hours and high satisfaction level
Low work hours and medium satisfaction level

Exploratory Data Analysis



Left employees with long work hours also received high evaluation score.

Exploratory Data Analysis



But most of they reported low satisfaction level score.

Statistical Test

2 sample t-test

- Satisfaction Levels: Stayed > Left
- Project Load: Left > Stayed
- Work Hours: Left > Stayed
- Tenure: Left > Stayed

Test	T statistic	P-Value
Satisfaction Level	35.889	<.001
Number of Project	-2.308	0.021
Last Evaluation Score	-1.298	0.194
Average Hour Worked per month	-6.369	<.001
Tenure	-24.050	<.001

Machin Learning Model

Logistic Regression Model

- The model achieved a score of 82%.
- Despite adequate accuracy, the model underperformed in predicting which employees are likely to leave

	Precision	Recall	f1-score
Predicted would stay	0.86	0.93	0.90
Predicted would leave	0.45	0.27	0.33

Accuracy: 82%

Machine Learning Model

Decision Tree Model

Train set (75%) :

cross validation with 4 sections

Hyperparameters	Parameter with best ROC AUC
Max depth	4, 6, 8, None
Min samples leaf	1, 2, 5
Min samples split	2, 4, 6

Machine Learning Model

Decision Tree Model

Performance on test set

Model	Precision	Recall	f1	Accuracy	ROC AUC
Decision tree	93.62%	91.37%	92.48%	97.53%	95.06%

The decision tree model achieved a ROC AUC score of 95.06% on test set, which is quite good.

Machine Learning Model

Random Forest Model

Train set (75%) :

cross validation with 4 sections

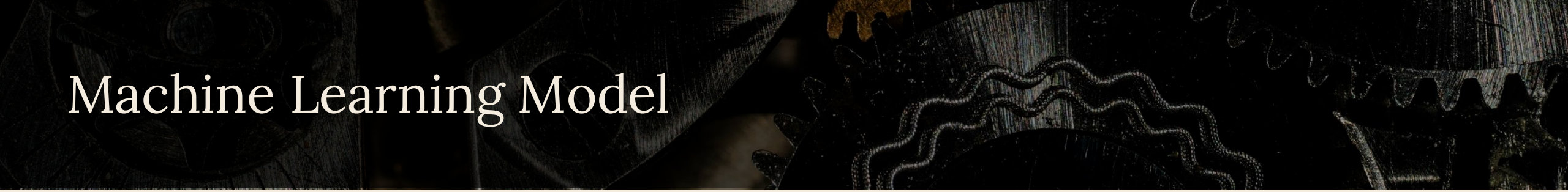
Hyperparameters	Parameter with best ROC AUC
Max depth	3, 5, None
Max features	1.0
Max samples	0.7, 1.0
Min samples leaf	1, 2, 3
Min samples split	2, 3, 4
Number of estimators	300, 500

Machine Learning Model

Random Forest Model

Model	Precision	Recall	f1	Accuracy	ROC AUC
Decision tree	93.62%	91.37%	92.48%	97.53%	95.06%
Random Forest	96.42%	91.97%	94.14%	98.10%	95.64%

The random forest model performed slightly better than the decision tree model.



Machine Learning Model

Data Leakage Risk:

The “average_monthly_hours” column might introduce data leakage.

Employees who have decided to quit or are marked for termination could be working fewer hours.

Feature Engineering:

Introduce a new column: “overworked”

Definition: Employees working over 175 hours per month.

Machine Learning Model

Decision Tree Model with Feature Engineering

Train set (75%) :

cross validation with 4 sections

Hyperparameters	Parameter with best ROC AUC
Max depth	4, 6, 8, None
Min samples leaf	1, 2, 5
Min samples split	2, 4, 6

Machine Learning Model

Decision Tree Model with Feature Engineering

Model	Precision	Recall	f1	Accuracy	ROC AUC
Decision tree	93.62%	91.37%	92.48%	97.53%	95.06%
Random Forest	96.42%	91.97%	94.14%	98.10%	95.64%
Decision tree with FE	78.39%	91.77%	84.55%	94.43%	93.36%

Although the ROC AUC score decreased, the issue of data leakage was mitigated through feature engineering.

Machine Learning Model

Random Forest Model with Feature Engineering

Train set (75%) :

cross validation with 4 sections

Hyperparameters	Parameter with best ROC AUC
Max depth	3, 5, None
Max features	1.0
Max samples	0.7, 1.0
Min samples leaf	1, 2, 3
Min samples split	2, 3, 4
Number of estimators	300, 500

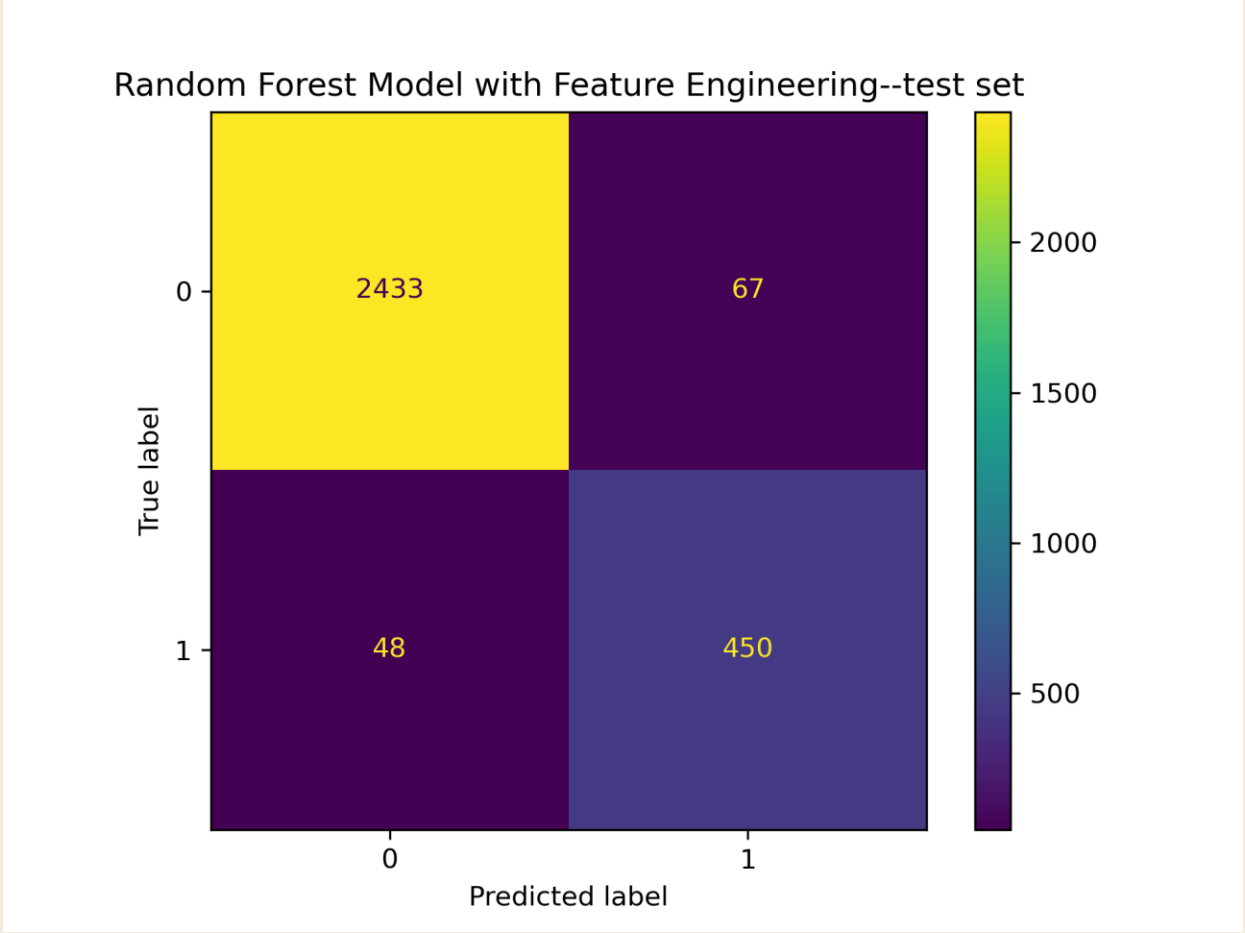
Machine Learning Model

Random Forest Model with Feature Engineering

In test set:

False Positive: 67

False Negative: 48



Machine Learning Model

Random Forest Model with Feature Engineering

Model	Precision	Recall	f1	Accuracy	ROC AUC
Decision tree	93.62%	91.37%	92.48%	97.53%	95.06%
Random Forest	96.42%	91.97%	94.14%	98.10%	95.64%
Decision tree with FE	78.39%	91.77%	84.55%	94.43%	93.36%
Random Forest with FE	87.04%	90.36%	88.67%	96.16%	93.84%

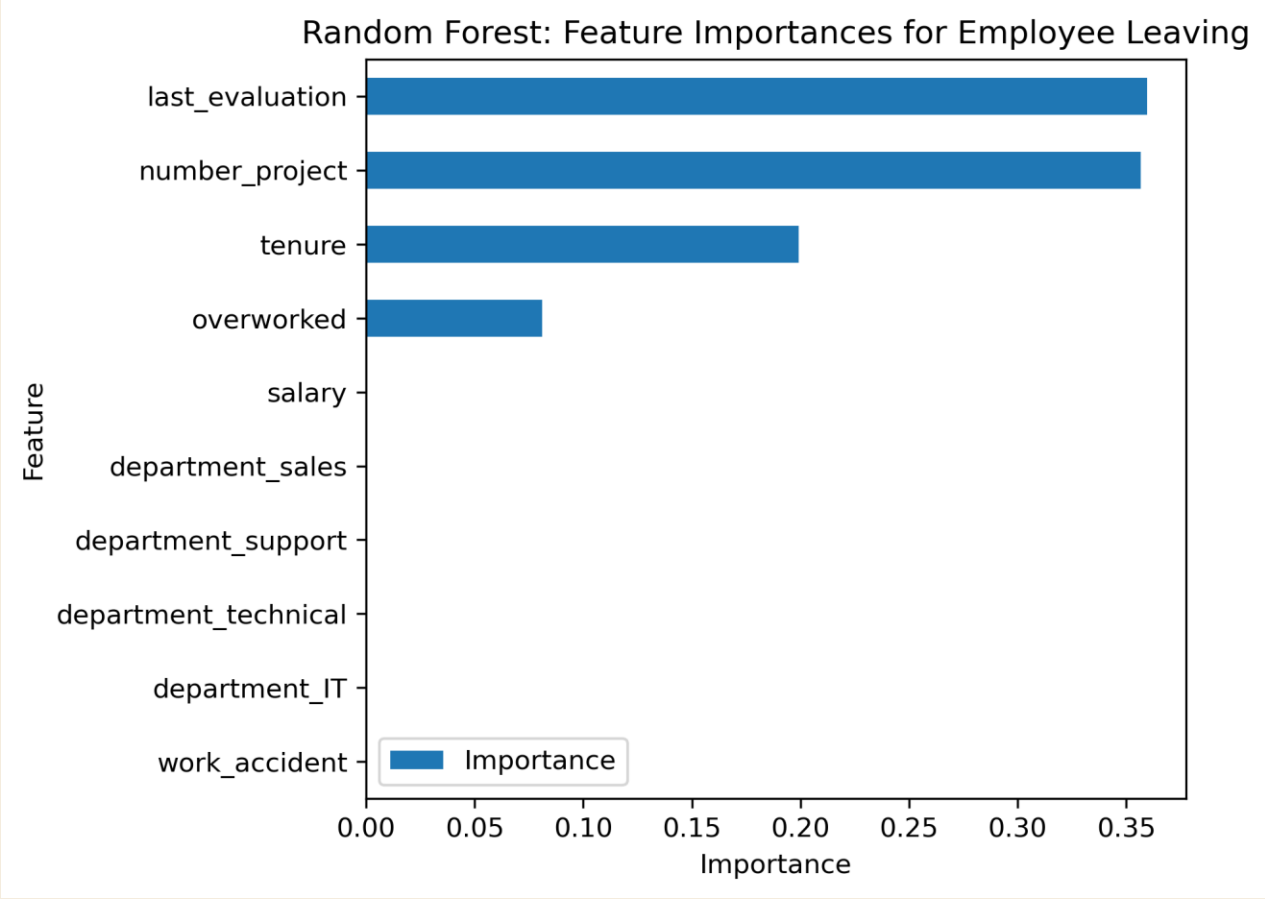
Although the ROC AUC score decreased, the issue of data leakage was mitigated through feature engineering.

Machine Learning Model

Random Forest Model with Feature Engineering

• Last evaluation score, number of project, tenure, and overworked are the most important features.

• These variables are the most helpful in predicting whether an employee will leave the company.



Conclusion

Recommendation for HR:

1. Reduce Project Overload

   Rebalance workloads to prevent employee overwork and boost productivity.

2. Recognition Program

   Create rewards to motivate employees with good performance to increase their satisfaction level.

3. Investigate Turnover

   Explore all factors beyond low satisfaction contributing to employee turnover.

4. Support Underperformers

   Offer workshops to help employees meet project expectations and improve skills.