Analysis of HR Data to Improve Employee Retention

Nader Harb

Capstone Project

Part of the *Google Advanced Data Analytics* course

Business Problem

- The company Salifort Motors wants to know why its employees quit.
- Employee retention is vital to the company because employees receive regular training which is costly to the company.
- HR department conducted an employee survey, and I was tasked with determining the drivers behind employees leaving.

PACE Methodology

PLAN

ANALYZE

CONSTRUCT

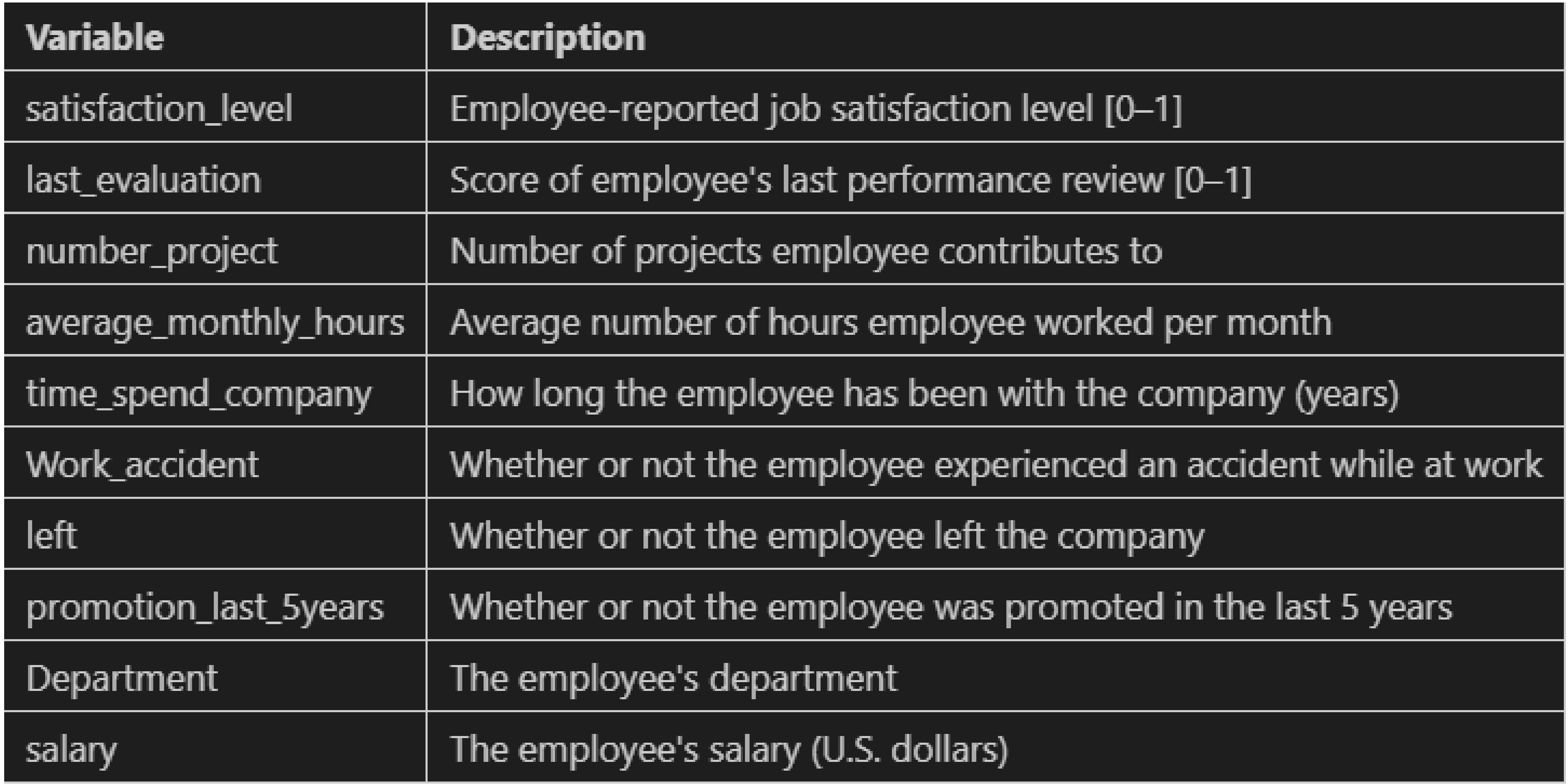
EXECUTE

Plan

- IDENTIFYING BUSINESS PROBLEM
- DEFINING OBJECTIVES
- DECIDING ON WHAT DATA TO COLLECT
- CLEANING DATA
- PERFORMING INITIAL EDA

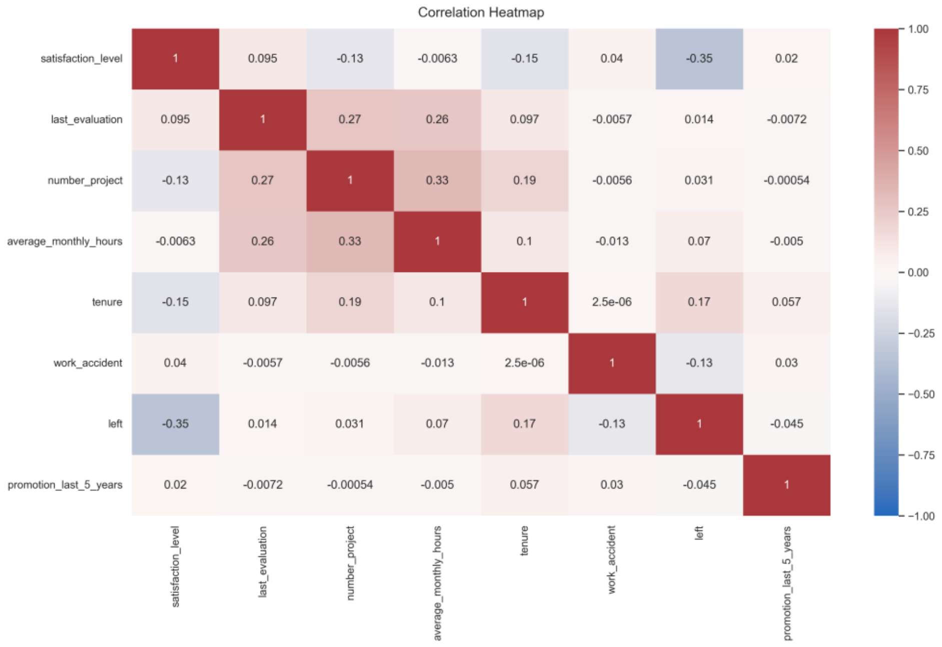
Data Distribution

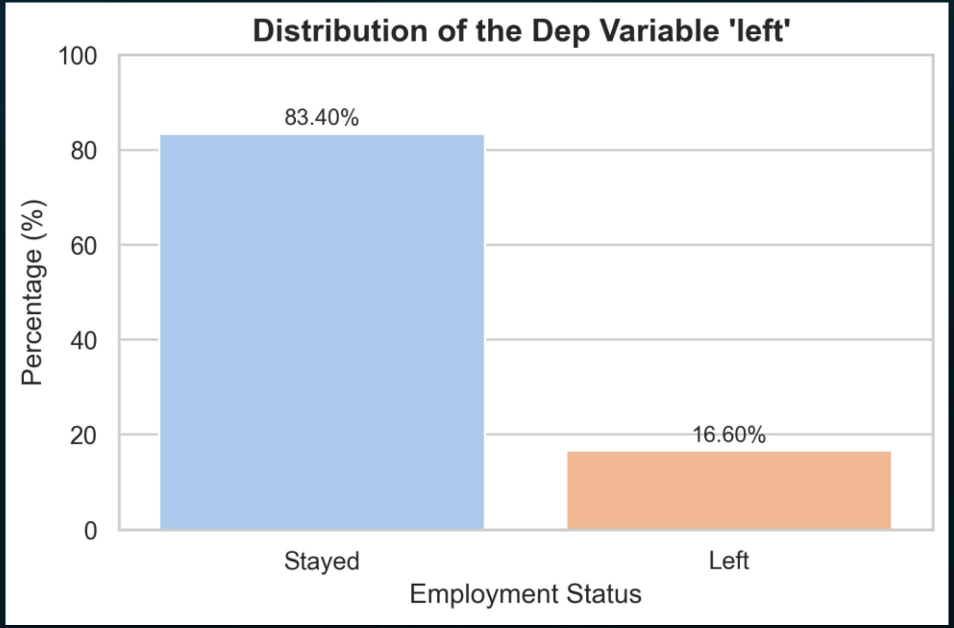
Data was retrieved from [Kaggle]



Variable	Description
satisfaction_level	Employee-reported job satisfaction level [0–1]
last_evaluation	Score of employee's last performance review [0–1]
number_project	Number of projects employee contributes to
average_monthly_hours	Average number of hours employee worked per month
time_spend_company	How long the employee has been with the company (years)
Work_accident	Whether or not the employee experienced an accident while at work
left	Whether or not the employee left the company
promotion_last_5years	Whether or not the employee was promoted in the last 5 years
Department	The employee's department
salary	The employee's salary (U.S. dollars)

Data Distribution & Correlation





Analyze

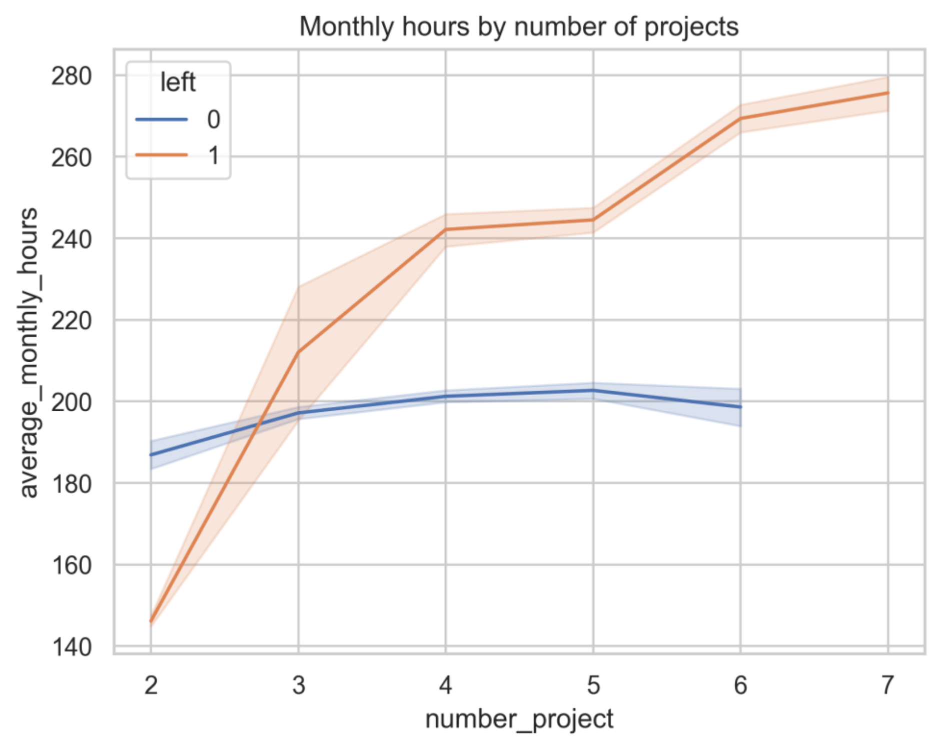
CONTINUING CLEANING FOR MODELLING

CONTINUING EDA

VISUALIZING TRENDS, INSIGHTS, AND CORRELATIONS

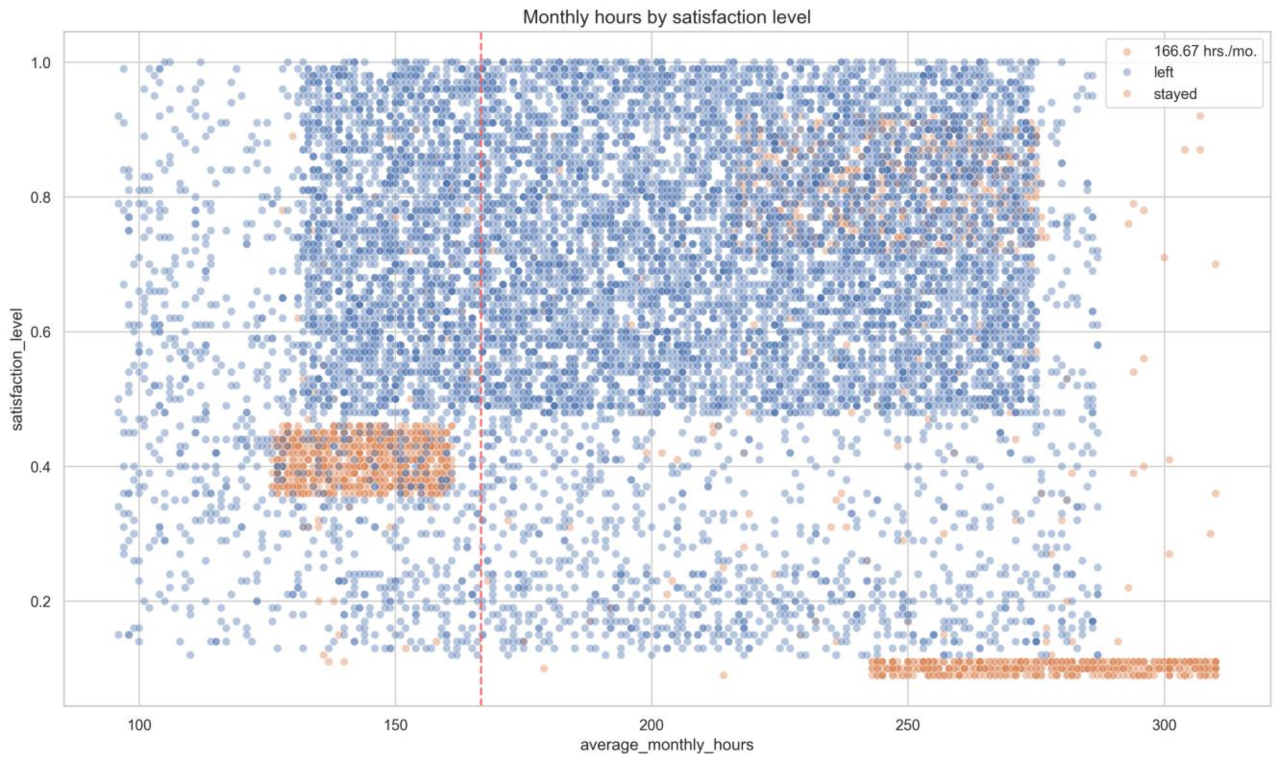
Monthly hours worked & Number of Projects

- The people who left were working more hours than the people who didn't despite taking on the same number of projects.
- The exception is that the only people who worked on 7 projects were the ones who left.
- It may be an issue of high performers vs slackers.

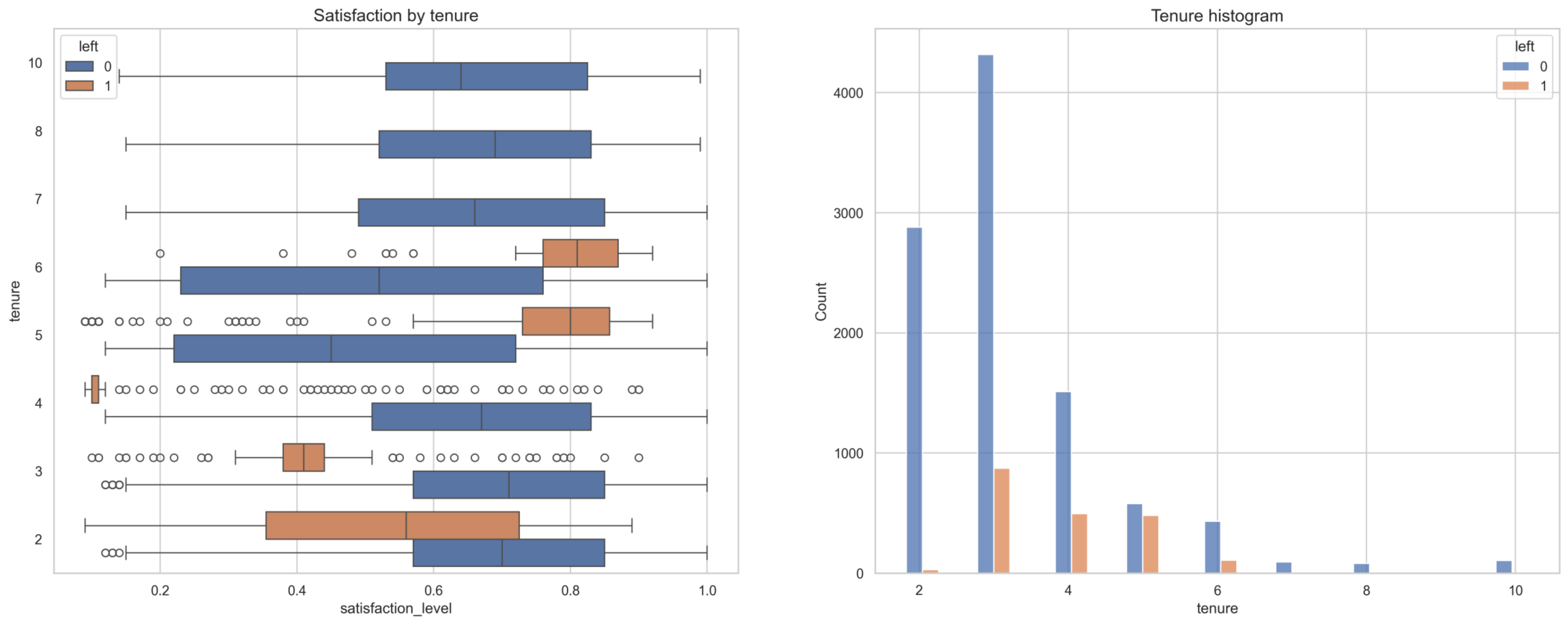


Monthly Hours Worked & Satisfaction Level

- In general, the employees of this company work more than the average 166.67 hours/month
- The employees who left can be categorized into 2 categories:
    1. Overworked employees who are dissatisfied. They may feel like they're being taken advantage of and unappreciated.
    2. Underworked employees who are dissatisfied. They may feel like they're being taken for granted and passed up on.



Tenure & Satisfaction Level

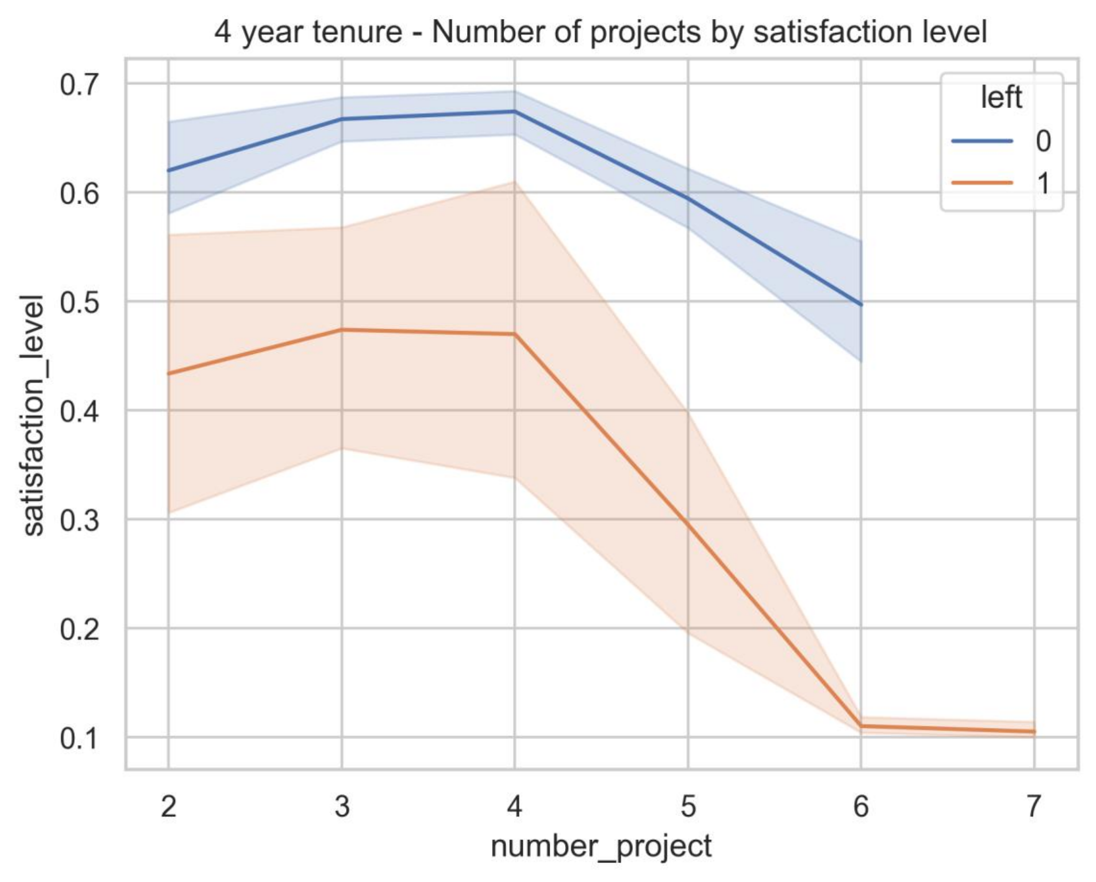


- Long-Tenured employees do not leave the company
- Some medium-Tenured employees leave the company not because of dissatisfaction
- 4-year tenure employees and under leave when they are dissatisfied

4-Year Tenure & Satisfaction Level

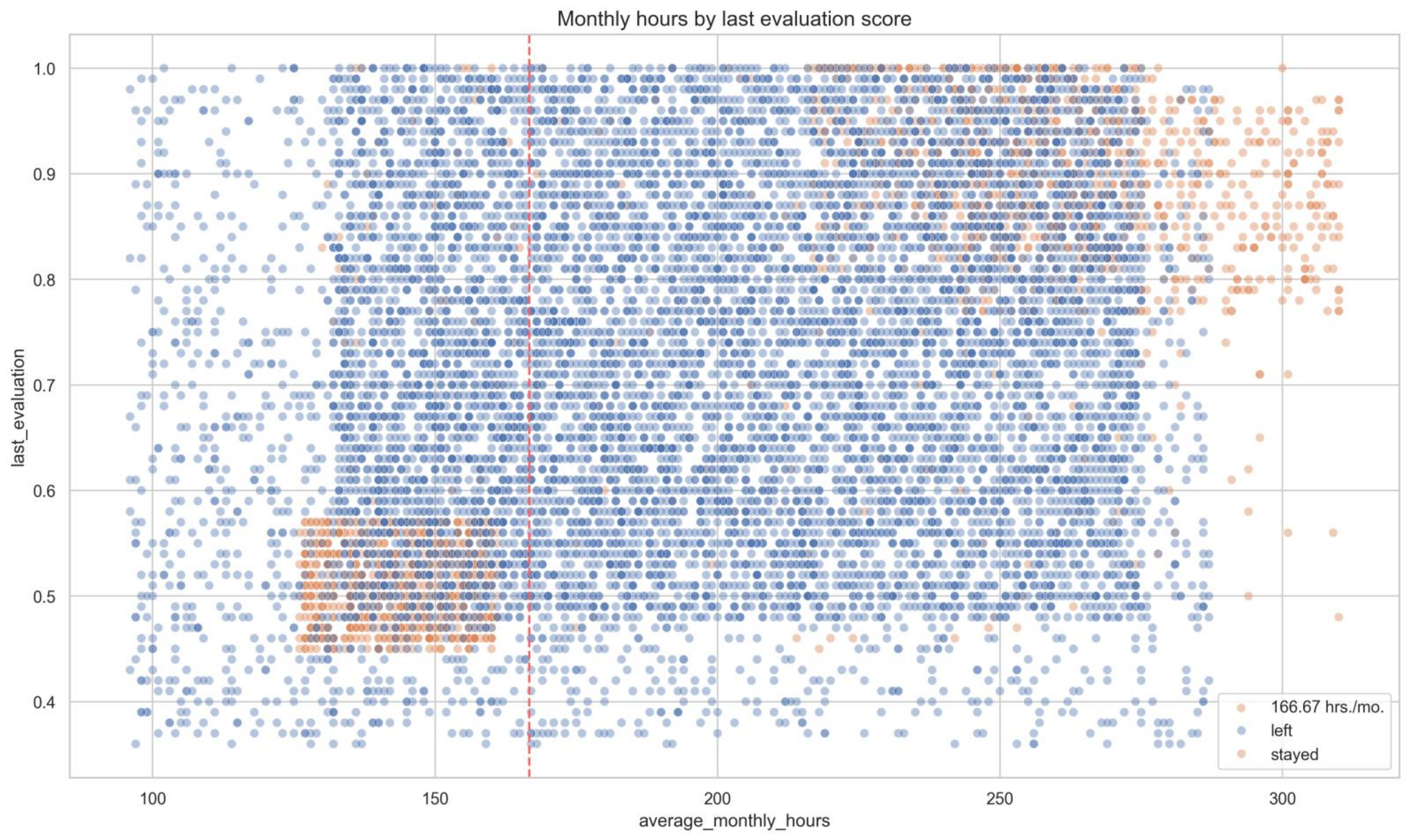
4-year tenure only:

- The employees who left are dissatisfied in general.
- Their dissatisfaction is driven by the number of projects they’re taking on past 4 projects.



Monthly Hours Worked & Evaluation Score

- There’s 2 types of employees who left:
    1. overworked employees who performed very well
    2. employees who worked slightly under the nominal monthly average of 166.67 hours with lower evaluation scores.
- There seems to be a correlation between the 2 variables.
- There isn't a high percentage of employees in the upper left quadrant of this plot; but working long hours doesn't guarantee a good evaluation score.
- Most of the employees in this company work well over 167 hours per month.



Construct & Execute

- Construct
  - Determining appropriate models
  - Constructing models
  - Confirming model assumptions
  - Evaluating results
  - Feature Engineering
- Execute
  - Interpreting model performance and results
  - Sharing actionable steps with stakeholders

Modelling

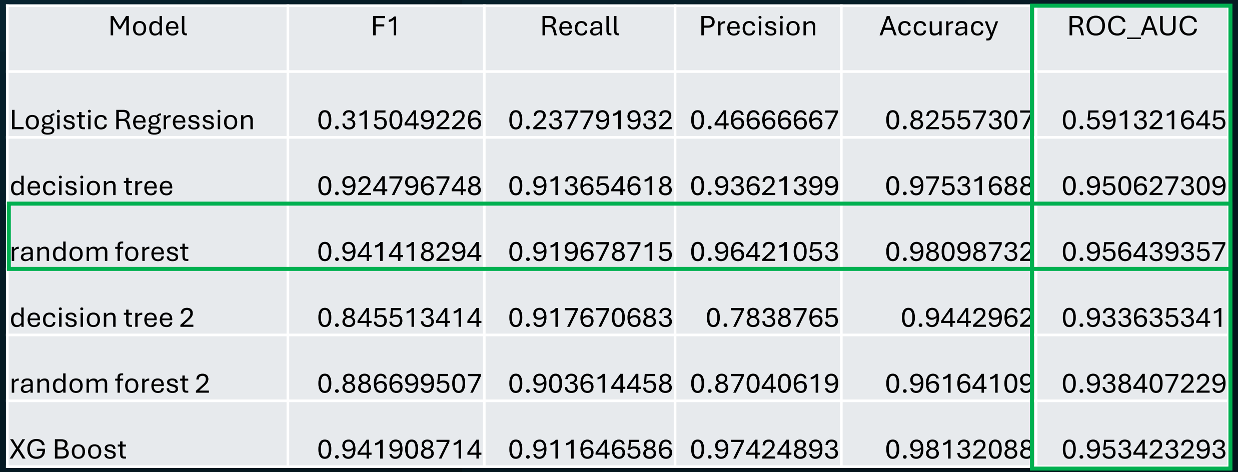
- I used GridSearchCV with the following models:
  - Logistic Regression
  - Decision Tree Classifier
  - Random Forest
  - XGBoost

- Train-Test split was 75/25

- Scoring parameter chosen was **roc_auc** because:
  - There is a class imbalance
  - The dependent variable only has 2 classes
  - The goal is to correctly identify employees who left.

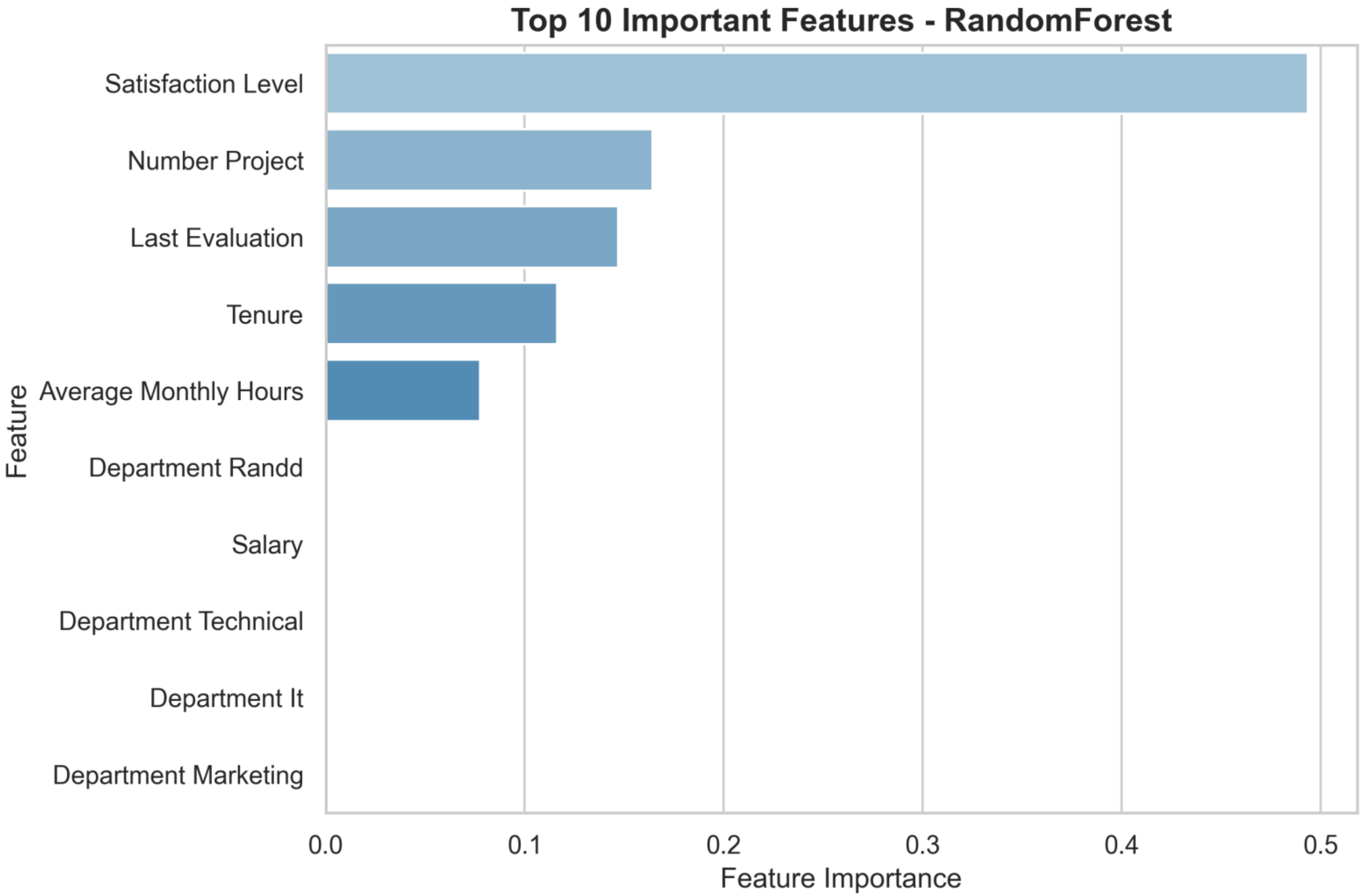
Model Results

- Feature Engineering made the models worse off
- Best model is the Random Forest
- It is the best in both the roc auc score and the recall score



Model	F1	Recall	Precision	Accuracy	ROC_AUC
Logistic Regression	0.315049226	0.237791932	0.46666667	0.82557307	0.591321645
decision tree	0.924796748	0.913654618	0.93621399	0.97531688	0.950627309
random forest	0.941418294	0.919678715	0.96421053	0.98098732	0.956439357
decision tree 2	0.845513414	0.917670683	0.7838765	0.9442962	0.933635341
random forest 2	0.886699507	0.903614458	0.87040619	0.96164109	0.938407229
XG Boost	0.941908714	0.911646586	0.97424893	0.98132088	0.953423293

Model Results



Recommendations & Next Steps

- **The models and the feature importances extracted from the models confirm that employees at the company are overworked.**
- To help reduce employee turnover, the following actions could be proposed to stakeholders:
  1. Limit the number of projects assigned to each employee to prevent burnout.
  2. Promote employees with four or more years of tenure.
  3. Reassess expectations around overtime - either provide incentives for working extra hours or avoid setting such expectations altogether.
  4. Improve communication about company policies, especially regarding overtime pay, workload expectations, and time off.
  5. Revise performance evaluation practices to ensure that high scores aren’t only achievable for those working over 200 hours per month. Implement a more balanced system that fairly rewards effort and contribution across the board.