

Employee Retention Project

Project Context

Our client is HR department at a large software company.

- They are rolling out a new initiative that they call "Proactive Retention."
- The idea is to use data to predict whether and employee is likely to leave.
- Once these employees are identified, HR can be more proactive in reaching out to them before it's too late.
- For this initiative, they only care about **permanent (non-temp) employees**.

Current Solution

Currently, their employee retention process is very retroactive. Once an employee leaves, he or she takes an "exit interview" and shares reasons for leaving. HR then tries to learn insights from that interview and make changes around the company accordingly.

This suffers from 3 main problems:

- The first problem with this approach is that it's too haphazard. The quality of insight gained from an interview depends heavily on the skill of the interviewer.
- The second problem is that they can't systematically aggregate insights across all employees who have left.
- The third problem is that they can't be proactive because they are using exit interviews to drive policy changes.

Our Role

The HR department has hired us as data science consultants. They want to supplement their exit interviews with a more proactive approach.

- They've asked their business intelligence analysts to provide us a dataset of past employees and their status (still employed or already left).
- Our task is to build a classification model using that dataset.

- Because our solution will be complementary to the existing one, and because there's no precedent, we do not have a quantifiable win condition. Just build the best model possible.

Problem Specifics

Deliverable: **Predictive Machine Learning API**

Machine learning task: **Classification**

Target variable: **Status (Employed/Left)**

Data Dictionary

For this project:

- The dataset has **14249** observations for past/present employees.
- The observations span 12 different departments.
- Each observation includes the employee's current employment status.

We have the following features:

Target variable

- `'status'` – Current employment status (Employed / Left)

Administrative information

- `'department'` – Department employees belong(ed) to
- `'salary'` – Salary level relative to rest of their department
- `'tenure'` – Number of years at the company
- `'recently_promoted'` – Was the employee promoted in the last 3 years?

Workload information

- `'n_projects'` – Number of projects employee is staffed on
- `'avg_monthly_hrs'` – Average number of hours worked per month

Mutual evaluation information

- `'satisfaction'` – Score for employee's satisfaction with the company (higher is better)
- `'last_evaluation'` – Score for most recent evaluation of employee (higher is better)
- `'filed_complaint'` – Has the employee filed a formal complaint in the last 3 years?