Job-a-thon: Employee Attrition Prediction

Approach

As opposed to predicting whether an employee will leave or not, which would require binary classification, the problem here required to predict when an employee would quit if they would. Standard regression-based approaches do not work with censored data (data that hasn't been recorded or doesn't exist yet eg, whether an event will occur and when it will occur but we use that fact that it hasn't occurred yet).
Hence, survival analysis (also called survival regression) seemed to be an apt choice since it deals with predicting the duration of time before a certain event might occur. If the employee isn't expected to leave now, that doesn't mean that they won't leave sometime in the future. Survival analysis can be used to predict their "survival probability" at different points of time in the future.

Data-preprocessing/feature engineering

Monthly data was given for each employee (data for each month since the employee started working). So the data had to be aggregated into one row for each employee based on their ID. This included:

- Number of months the employee has worked for until now
- Number of promotions (change in designation level)
- Average business value per month created by the employee

Categorical data, gender and education level, was converted to numerical labels (One-hot encoding could also have been used for education level). Other than that the data was mostly clean.
Another column about the employee's current status was added which had a binary variable for whether the employee had left the company (yet) or not.

Model

The Cox Proportional Hazard (CPH) Model was used. The CPH model determines the effect that a unit change in a covariate (feature) will have on an observation's survival probability. In python, CPH is included in the lifelines library. It is one of the most commonly used models in survival analysis. It requires that the correlation between the features isn't too high and assumes that the features have a constant proportional impact on the hazard rate (although the model works well in practice even if the assumption isn't completely satisfied).