ITM885 Case Study 3: Employee Retention

In this case study, we focus on EDA and feature engineering of response variable by data aggregation and will see a well-defined problem simplifies analysis and interpretation.

We got employee data from a few companies. We have data about all employees who joined from 2011/01/24 to 2015/12/13. For each employee, we also know if they are still at the company as of 2015/12/13 or they have quit. Beside that, we have general info about the employee, such as average salary during her tenure, department, and years of experience.

"employee_retention.csv" contains comprehensive information about employees. Concretely,
employee_id: id of the employee. Unique by employee per company
company_id: company id.
dept: employee dept
seniority: number of yrs of work experience when hired
salary: avg yearly salary of the employee during her tenure within the company
join_date: when the employee joined the company, it can only be between 2011/01/24 and 2015/12/13
quit_date: when the employee left her job (if she is still employed as of 2015/12/13, this field is NA)

As said above, the goal is to predict employee retention and understand its main drivers. You can use a survival model to study the retention. But survival models are not good at point prediction and your boss may not be familiar with it. Moreover, you are analyzing the problem from the perspective of company instead of individual. In addition, there may be nonlinear effects of salary on retention. So, you can aggregate individual information as below.

Assume, for each company, that the headcount starts from zero on 2011/01/23. Estimate employee headcount, for each company, on each day, from 2011/01/24 to 2015/12/13. That is, if by 2012/03/02, 2000 people have joined company 1 and 1500 of them have already quit, then company headcount on 2012/03/02 for company 1 would be 500. You should create a table with 3 columns: day, employee_headcount, company_id. What are the main factors that drive employee churn? Do they make sense? Explain your findings. If you could add to this data set just one variable that could help explain employee churn, what would that be?