Sequoia Data Science Case Study: Employee Turnover Prediction and Insights

# Motivation

Employee Turnover, which refers to the rate at which employees leave an organization, is an important problem in today's industry because it can have a significant impact on a company's productivity, morale, and bottom line. When employees leave an organization, it can lead to a loss of institutional knowledge, increased workload on remaining staff, and decreased overall efficiency. Additionally, high levels of turnover can damage morale and negatively impact the company's reputation, making it more difficult to attract and retain top talent. Therefore, reducing employee turnover and retaining employees is critical for businesses to maintain a stable and productive workforce and remain competitive in today's market

# Problem Statement

Meet Erika, she is a CHRO of a renowned organization. He would like to keep his employees happy and avoid their resignations. So, she asked her data scientist to come up with a model to predict who are the employees who are at flight risk. She also wants to know what the action items on her end are so she can solve for the top reasons which make the employees leave the organization.

Your task is to be that data scientist and come up with models to predict “TurnOver” and come up with contextual insights (auto generated, not manual) which can help Erika keep her employees.

## Deliverables and Points to remember

1. We prefer python, so do share the python notebook, feel free to use and openly available notebook environments like google colaboratory (<https://colab.research.google.com/>) You can upload the csv there and fire a notebook. Once done, please share the properly commented and formatted notebook (ipynb file) and exported pdf (export the same notebook as pdf).
   1. If you want to create two separate notebooks (one for predictive model and other for insight generation), it works fine with us too
2. During the interview, we would expect you to keep the colab notebook running
3. Do create a one pager note/diagram on how you would deploy this model in production, run inferences, share insights. We don’t need implementation for this architecture discussion, however low-level design decisions should be fairly clear.

*That’s it! you make the CHRO Erica happy, and we would be happy to onboard you as our data scientist, working on exciting, unique and challenging problems in HR-Tech domain, at Sequoia, a pioneer of Total People Investment and a fun place to work with.*

# Dataset

The given dataset is from various departments and employee level features as mentioned below. As you will see that this data set size is fairly small (1471 rows), Typical of HR Tech space (We have lot of variety in data (many columns from multiple sources), but as our datasets are at employee level, the number of rows are limited).

Do consider this special situation of HRTech domain while preparing your models.

We are aware that orking with a smaller dataset can present some unique challenges, such as limited variability and a higher risk of overfitting. However, it's important to remember that smaller datasets can also have some advantages, such as being easier and quicker to analyze.

To make the most out of the data you have, we would recommend that you focus on **quality over quantity** [A sequoia value]. We have tried to ensure that the given data is accurate, relevant, and representative of the population you are studying. You can also consider applying various data augmentation techniques, such as sampling or imputation, to expand the dataset's size and improve its variability.

Remember that with the right approach and tools, a small dataset can still yield meaningful and valuable insights. Don't hesitate to reach out if you need any further assistance or advice.

Data dictionary for the attached dataset is given below:

|  |  |
| --- | --- |
| Column name | Description |
| Age | The age of the employee. (Numerical) |
| TurnOver | Whether or not the employee has left the organization. (Categorical) |
| BusinessTravel | The frequency of business travel for the employee. (Categorical) |
| DailyRate | The daily rate of pay for the employee. (Numerical) |
| Department | The department the employee works in. (Categorical) |
| DistanceFromHome | The distance from home in miles for the employee. (Numerical) |
| Education | The level of education achieved by the employee. (Categorical) |
| EducationField | The field of study for the employee's education. (Categorical) |
| EmployeeCount | The total number of employees in the organization. (Numerical) |
| EmployeeNumber | A unique identifier for each employee profile. (Numerical) |
| EnvironmentSatisfaction | The employee's satisfaction with their work environment. (Categorical) |
| Gender | The gender of the employee. (Categorical) |
| HourlyRate | The hourly rate of pay for the employee. (Numerical) |
| JobInvolvement | The level of involvement required for the employee's job. (Categorical) |
| JobLevel | The job level of the employee. (Categorical) |
| JobRole | The role of the employee in the organization. (Categorical) |
| JobSatisfaction | The employee's satisfaction with their job. (Categorical) |
| MaritalStatus | The marital status of the employee. (Categorical) |
| MonthlyIncome | The monthly income of the employee. (Numerical) |
| MonthlyRate | The monthly rate of pay for the employee. (Numerical) |
| NumCompaniesWorked | The number of companies the employee has worked for. (Numerical) |
| Over18 | Whether or not the employee is over 18. (Categorical) |
| OverTime | Whether or not the employee works overtime. (Categorical) |
| PercentSalaryHike | The percentage of salary hike for the employee. (Numerical) |
| PerformanceRating | The performance rating of the employee. (Categorical) |
| RelationshipSatisfaction | The employee's satisfaction with their relationships. (Categorical) |
| StandardHours | The standard hours of work for the employee. (Numerical) |
| StockOptionLevel | The stock option level of the employee. (Numerical) |
| TotalWorkingYears | The total number of years the employee has worked. (Numerical) |
| TrainingTimesLastYear | The number of times the employee was taken for training in the last year. (Numerical) |
| WorkLifeBalance | The employee's perception of their work-life balance. (Categorical) |
| YearsAtCompany | The number of years the employee has been with the company. (Numerical) |
| YearsInCurrentRole | The number of years the employee has been in their current role. (Numerical) |
| YearsSinceLastPromotion | The number of years since the employee's last promotion. (Numerical) |
| YearsWithCurrManager | The number of years the employee has been with their current manager. (Numerical) |
|  |  |

# Good Luck!

We wanted to take a moment to wish you good luck on this data science case study. We know that this task is important to you and that you will put a lot of effort into preparing for it.

Remember to stay focused, be thorough in your analysis, and keep an open mind as you work through the case study. We have no doubt that your skills and knowledge in data science will be put to good use and that you will come up with some insightful and valuable insights.

If you encounter any challenges along the way, don't hesitate to reach out to us for help. You have a great team of colleagues and mentors who are there to support you.

Best of luck with this data science case study! We look forward to hearing from you with inspiring solution.

Team Sequoia AI