

Milestone 3 – Ten Questions & Answers (Project 1)

1. Why were you interested in this business problem of employee attrition?

I currently work as a data engineer at a financial services company, and my team has been directly impacted by our coworkers resigning. Our smaller-sized team is then left with the same scope of work to complete with a smaller capacity of hours, which has been difficult in terms of workload management, mental health, etc. Given that the issue of high employee attrition has been a personal problem for me day-to-day, I wanted to explore it with the usage of data science to understand the problem from a different perspective and even see if I could learn anything to bring back to my team and my company.

2. How did you account for the imbalance in the target classes?

I used random oversampling! There is a function from the imbalanced-learn Python library, which allows for the random duplication of examples in the minority class. It is best used on a smaller dataset, and given that my dataset only had 1,470 records, I figured it would be a good fit for handling the major class imbalance of my target variable.

3. How often do you think these models should be tested on employees?

I think every quarter would be a great timespan! Usually, there is some type of change every quarter within an organization, and I think it would be a good time point for measuring employees' feelings and moods around their job and gauging employee attrition tendencies.

4. Should the dataset consider outside-of-work features that can play into employee attrition such as mental health, changing career paths, etc.?

While I think these external features could play a role in employee attrition, I also think they would be difficult to collect data on from a privacy perspective. Datasets could include a column for the reason why an employee resigned, if an employee wants to disclose it, but I think other features will be suitable for the business problem at hand if these outside-of-work features are difficult to collect.

5. Is there a certain industry that is experiencing higher attrition rates than others?

After doing some research, according to the Harvard Business Review, tech and health care industries are experiencing higher attrition rates than others. In quantitative measurements, “3.6% more health care employees quit their jobs than in the previous year, and in tech, resignations increased by 4.5%” (Cook, 2021). These industries could be ones where employees feel more impacted by the main factors mentioned in the paper and the presentation, but from this statistic, I would think they should especially become the focus for how employees in these industries can be aided.

Link to source: <https://hbr.org/2021/09/who-is-driving-the-great-resignation>

6. Why was employee job level not considered in the modeling, i.e. leadership positions vs. technical positions? Do you think this would make a difference?

After the completion of my modeling and presentation, I realized that this would be an interesting perspective to view the data from. From knowledge of an organization, there are usually less leadership positions than non-leadership positions, so it would be interesting to compare the groups and see if having a leadership responsibility leads to higher or lower attrition.

7. Could you have employed other techniques for feature reduction?

Yes! I thought about employing other techniques such as Variance Inflation Factors (VIF), the Gini Index, and others. However, I ended up following a more business knowledge approach instead for choosing top features and testing the hypothesis of their weight in the prediction.

8. Did you consider other forms of visualizations for the exploration data analysis (EDA)?

I think the forms of visualization I implemented (bar charts, histograms) were suitable for the type of data from the dataset. Since the dataset had more categorical attributes, the bar charts were especially helpful for visualizing categories and the differences between them.

9. Would you suggest that companies be transparent around this work if they choose to implement it for employees?

Yes, most definitely. I think transparent communication is key to the success of an organization, and I think if employees are aware that the leadership at a company is investigating employee attrition and putting initiatives in place to hopefully help it, then it will allow the employees as well to build trust and gratitude.

10. How does predicting whether an employee will resign (Yes/No) compare with assigning them risk scores as also recommended later in the paper?

Whether the predictor variable is categorical or numerical, it does not much matter as much as the value it brings to the business problem. I think one advantage of the numerical predictor is that it can give a more variable indicator of how likely it is for an employee to resign. The employees will be assigned a risk score, which can then be utilized to focus on the highest risk candidates which may be for better efforts anyway. Also, by helping the top population, I think it

Emily Phillips
Dr. Catherine Williams
DSC 680 – Applied Data Science
4/10/2022

will help the other employees as well since the takeaways from the most at risk will hopefully be utilized throughout the organization for all employees. However, either predictor variable will bring great awareness to the employees that are going to quit and can therefore help in forming the focus group.