Emily Phillips
Professor Parajulee
DSC 530
8/12/2021

Employee Attrition EDA

My hypothetical question in regards to the dataset I chose, IBM HR Analytics Employee Attrition and Performance, is: "What factors play a significant role in predicting voluntary employee attrition?". Voluntary employee attrition was the focus type for this exploratory analysis, because it pertains to when employees make the choice of leaving a company, rather than being laid off due to external reasons.

The final step in my exploratory analysis was performing logistic regression on my dataset with the 'Attrition' variable as the outcome variable (labels of 'Yes' or 'No'). A 'yes' value pertained to the employee leaving the company, and a 'no' value pertained to the opposite. I created two logistic models: one with just 'Age' as the explanatory variable and the second with the following explanatory variables: Age, Yearly Income (Salary), Years Since Last Promotion, Years in Current Role, Job Satisfaction, and Distance From Home. From personal experience, I decided that these were the optimal variables from the dataset that could help with my analysis. Pay is a huge factor in people's decisions for taking a job as it sustains their living and keeps their job competitive. I also think Age dictates how people decide the place they want to be in their lives; as people get older, they want to progress in their careers and this progression depends on what the company can provide. With the other variables, they can influence the satisfaction that people have in their careers.

The results of the logistic regression models showed that the accuracy changed very slightly from just having Age as the predictor to including the other five variables (84.5% to 84.61%); only ~.1% difference. The baseline prediction strategy was also correct the same percentage of the time for both of the models at 84.45%. Therefore, in terms of the significance of the factors I chose to include in predicting voluntary employee attrition, it seems that just using Age is just as good of a predictive measure as the other variables all together. Age by itself would allow us to detect whether an employee is likely to leave their company relatively well.

In reflecting on the exploratory data analysis of this dataset, I did not fully understand how to include any of the categorical variables in my work (ex. Job Level, Gender, Education Field, etc.). While I think some of them could have helped in my analysis, I was still getting confused on how to interpret them since we primarily focused on numerical variables.

This dataset was also quite small (1,470 rows), so I feel like even though I concluded that I could generalize the observed difference in mean yearly income to the population, the dataset itself might not even be representative enough of the population and could possibly have sampling bias. Therefore, for future work, I would try to find a more representative dataset or merge multiple datasets to pull in that information.