Emily Phillips
Dr. Catherine Williams
DSC 680 – Applied Data Science
3/19/2022

Project 1 Proposal – Predictive Proactivity with Employee Attrition

This data science project titled 'Predictive Proactivity with Employee Attrition' aims to predict, based on an employee's characteristics, whether they will be at high risk for leaving their company. The main goals of the project will be to determine the significant predictor variables, and to give a predictive indicator of whether an employee will contribute to attrition.

Employee attrition, at the voluntary level, can have a significant negative impact on the productivity and camaraderie of a company's work environment. By being able to identify significant predictive factors and to predict the possibility of employee attrition, it will help a company prepare more proactively and work towards putting initiatives in place that lower attrition rates.

The project dataset is from Kaggle, and represents a fictional dataset created by IBM data scientists (https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset). The data came from a survey conducted by the data scientists on about 1,471 distinct employees to gather information on thirty-five employee characteristic factors such as their education level, environment satisfaction scoring, performance rating score, income, and demographics.

Before the modeling portion of the project, I plan to use exploratory data analysis methods to understand the employee attrition data. These methods will include assessing distributions and outliers, visualizing trends and designating feature importance through measures such as variance-inflation factor and the Gini index. In terms of modeling, the target variable from this dataset is the categorical 'Attrition' variable which has two unique values: 'Yes' and 'No'. I plan to use classification modeling methods to predict whether a new employee will attrite, which

consist of algorithms such as k-Nearest Neighbor, Naïve Bayes Classifier, Decision Tree Classifier, and others.

One potential ethical concern of analyzing this data is that it does not consider the other factors of why an employee may leave their company such as significant life events, organization-wide layoffs and work harassment that would lead to involuntary employee attrition. I am going to assume that the employees surveyed for this dataset are not undergoing these conditions, but it is a potential ethical concern that could create bias in the target variable. A concern that will also need to be considered is that employee attrition is normal at an organization. Employees are most likely to leave their current workplace at some point in their career, and oftentimes, it is encouraged for career development and growth.

A potential project issue is that since the dataset is of smaller size, it may be difficult to get accurate modeling results when splitting the dataset into a training set and a test set, in terms of having enough data for adequate performance and evaluation. There will also be lengthy time spent on cleansing and preparing the data for the EDA and modeling portions of the project. There are many categorical variables which will need to be numerically encoded, and it may be a challenge to ensure that they represent the original data as expected.

I plan to use credible sources that provide insights on employee attrition trends over the recent years and resources that specify primary factors that feed into attrition rates at a company, to qualify my findings. I think it will also be good to research employee attrition patterns from one job industry to another, since there can be wide differences in how different industries organize and support their employees. To list a few sources to start:

1. https://www.quantumworkplace.com/future-of-work/employee-turnover-trends
2. https://www.dailypay.com/resource-center/blog/employee-retention-rate/