Employee Attrition

Chan, Natural

22 February, 2022

Abstract

The goal of this project is to identify individuals that are most likely to leave their current job. We will primarily observe demographics, employment status, and household information for a designated survey group. From there, we will identify which variables that are strong indicators that an individual is searching for a new job and predict on new observations. Knowing which variables are the strongest indicators will allow employers and recruiters take appropriate action to either focus on employee attrition or target potential new employees.

1 Design

The current generation has had an increase in work apathy and burnout in the recent years. In 2021, it was reported that over 4.5 million people had resigned from their jobs in the U.S, resulting in an event that people have referred to as the Great Resignation.

With such a drastic number of people leaving their jobs, we would like to take this opportunity to identify key factors that may cause an individual to leave their current position. Ultimately, this should benefit both employers and recruiters by answering two questions:

- Which employees are at risk of leaving for another job and why?
- What people should recruiters target?

To answer this, we will attempt to create a model to predict which people are currently searching for a job based on demographics, employment information, and household information. This will allow us to also take a look at which variables are the strongest indicators for an individual that is looking for a job.

2 Data

The U.S. Bureau of Labor Statistics conducts a survey known as the National Longitudinal Survey every other year on a select group of people. We will use the survey group that began in 1997 and analyze the most recent year available, 2019. One of the questions included was whether or not an individual was looking for a job - this variable will be what we hope to predict. The survey is only conducted on the same 8984 individuals every year and they were are all around the same age so many key variables that would assist in our predictions will not be available. Since this is survey data, there are a lot of missing values, but we will want to keep these values since missing information can be an important

3 Algorithms

3.1 Data Cleaning/Preprocessing

- Identifying approximately 3.68% of people that are searching for a job
- Replaced missing values with -999 for tree based methods

3.2 Modeling

- Used F2 score for evaluation since predicting true positive cases is more important here
- Used tree based methods (Decision Tree, Random Forest, XGBoost) only due to significant number of missing values
- Created pipelines for each method to over-sample minority class while performing 5-Fold cross validation
- Created functions to tune hyper-parameters on each model

Ultimately, we found that XGBoost performed the best of the 3 after tuning hyper-parameters

- Accuracy: .8587
- Precision: .1164
- Recall: .4944
- F1: .1887
- F2: .2997

Due to the limitations of our data, we are unable to achieve a significantly high F2 score. However, from our results, we do have a few takeaways. Of all of the variables, we found that the job type and job satisfaction were the greatest indicators with the others being fairly far behind. When looking at the 10 individuals with the highest probability to be searching for a job, we found that those with the highest probability responded that they think their job is okay and that they have a non-traditional job. Ultimately, employers and recruiters should target employees by identifying which are more likely to be unsatisfied with their job and seem to be experiencing employee apathy.

4 Tools

- Pandas for data processing and cleaning
- Scikit-learn and XGBoost for modeling
- Imblearn and hypopt for creating pipelines and tuning hyper-parameters
- Tableau and matplotlib to create visualizations

5 Communication

Presentation slides, Google Sheets, and Tableau visualizations are located in my github repository.