Prep and EDA\_v2

Ata, Catherine - Cortez, Angela - Foley, Amanda

5/11/2021

Introduction

Employee turnover is a major cost to organizations globally. According to Josh Bersin of Deloite Consulting, employee turnover can cost an organization 1.5x to 2x the employee’s salary([Employee Retention Now a Big Issue](https://www.linkedin.com/pulse/20130816200159-131079-employee-retention-now-a-big-issue-why-the-tide-has-turned?src=aff-lilpar&veh=aff_src.aff-lilpar_c.partners_pkw.10078_plc.Skimbit%20Ltd._pcrid.449670_learning&trk=aff_src.aff-lilpar_c.partners_pkw.10078_plc.Skimbit%20Ltd._pcrid.449670_learning&clickid=SMbXJgRkpxyLUnOwUx0Mo390UkB2duW9RX6NVc0&irgwc=1)). In addition, the financial implications of a single turnover, there are further reaching implications on the remaining employees and organizational culture, especially in firms with high turnover.

To mitigate turnover, firms can look for various indicators that may present early warning signs that employee has plans to leave, and an appropriate plan can be put into place to retain the employee. For such indicators to be identified, we have chosen several modeling techniques to test and propose models that provide insights into the indicators for turnover, and predictions for employees that are likely to leave an organization.

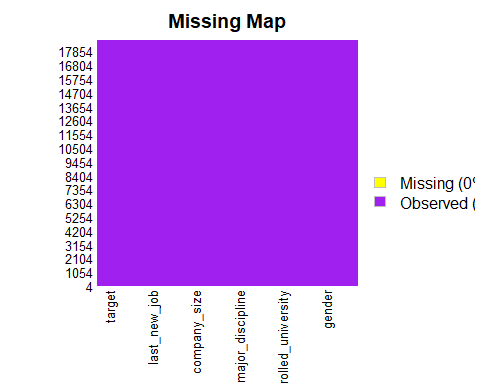
Data Preparation

The data preparation phase of any predictive modeling project is a crucial step in generating reliable and usable results. First, the team needed to source a dataset that met the criteria we set for our predictive modeling. The dataset that satisfied our requirement is the (insert link to dataset here). This dataset was chosen because it contained ample observations and features, and the features present lend themselves to the analysis of employee turnover. There are two data sets for this analysis. A training set “aug\_train” and a testing set “aut\_test”. The features and data elements will be covered more in depth in subsequent sections.

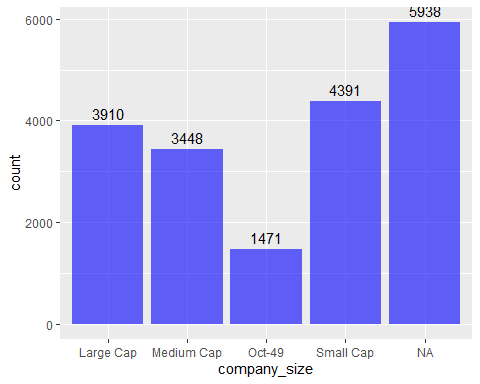
Chart

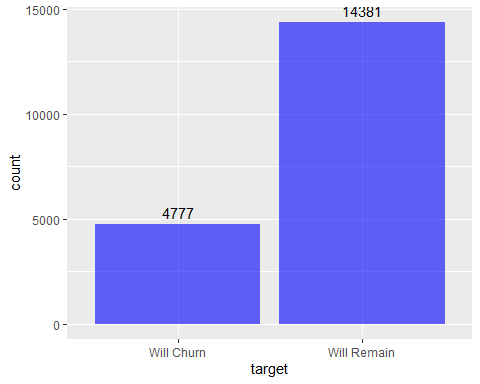
Description automatically generatedThe original dataset is comprised of 14 variables and ~19K observations and was downloaded and stored as a .csv file. After finding the dataset, we loaded the data into R so that we could began the process of assessing the completeness of the data, and to cleanse the data for use in exploratory data analysis (EDA). After the initial review of the data, we determined that there were two variables that would not be needed for our purposes, and they were removed from the dataset. These features were the “employee\_id”, which was a unique non personally identifiable identification number, and the “city\_id”, which we felt may not be relevant to the analysis.

After loading the data and removing the unnecessary features, we looked at data integrity and completeness. As can be seen in the figure above, the original dataset is missing approximately 9% of values across several of the features.

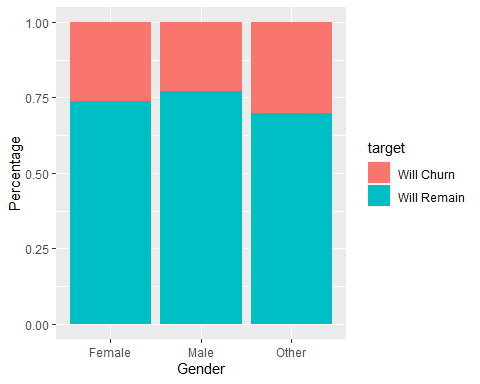
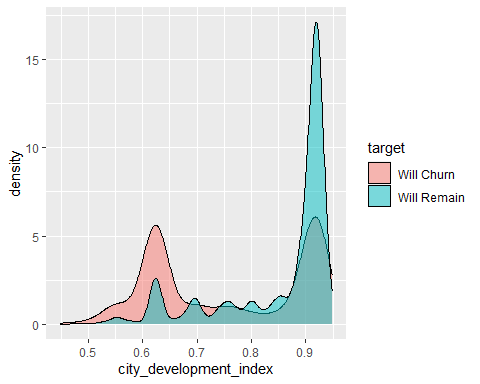
To clean up the missing values, several methods were used to either impute, fill-in, or remove incomplete records from the dataset. For data that contained multiple elements, we condensed them into categorical variables. For example, the degree type consisted of 6 different types: business degree, STEM, arts, humanities, other, no major, and NA’s. These were transformed into three categories; STEM, all other, and NA. Additional transformations include groupings by company size, years of experience and seniority level, and company type. After all the data transformations were complete, the percent After the data preparation process was completed, there were 0 missing values. The prepared data was stored in the team’s GitHub repository.

Exploratory Data Analysis

Chart, bar chart

Description automatically generatedDuring the exploratory analysis phase, the intent is was to gain more insight into dataset. This was done through evaluating summaries of the dataset, as well as creating data visualizations. In the figure to the right, we can see that a majority of the records for company size were NA, followed by Large Cap, Small Cap, and Medium Cap. In the visual below, we have more data available. The majority of the employees in the dataset meet the criteria for being considered “Senior Level”, which is defined as having between 10 and 20 years of experience, with 8682 records, approximately 45% of all records. The next group is mid-level. This group is defined as having between 4 and 9 years of experience and accounts for roughly 36% of the dataset. The smallest group for the experience feature, excluding NA’s, is the entry level employees, at 3552 records, or 19%, and are defined as have between 1 to 3 years of experience. The next categories are groupings in the feature major discipline. The decision to create the three groupings were based on overall volume in the original dataset. The largest group within the major disciplines are the STEM degrees account for roughly 76% of all data. The other categories are “other” and NA. For the feature company type, the data was condensed into three categories, Pvt. Limited, NA, and Other. After the initial EDA on the categories and their composition, we began to visualize the relationship between the features and the employee churn. In the dataset there are 14381 records that contain employees that stayed in their roles, and 4777 examples of churn.

Churn

First, we observed the churn within each gender. Overall, males had the highest total volume of churn, while other and females had the greatest percent of churn. When observing the relevant experience, we observed that the employees with no relevant experience were mostly likely to leave the organization. Additionally, employees that fell into the entry level category had the highest percent of churn amongst the level of seniority categories. We also observed that the highest percentage of churn occurs for those with a graduate level of education, and the highest churn by volume. For the city development index we observed an index around 0.9 and 0.62 have highest Churn while index at 0.9 will Remain

