Michelle Rice

DSC-550

Term Project Milestone 5

June 4, 2021

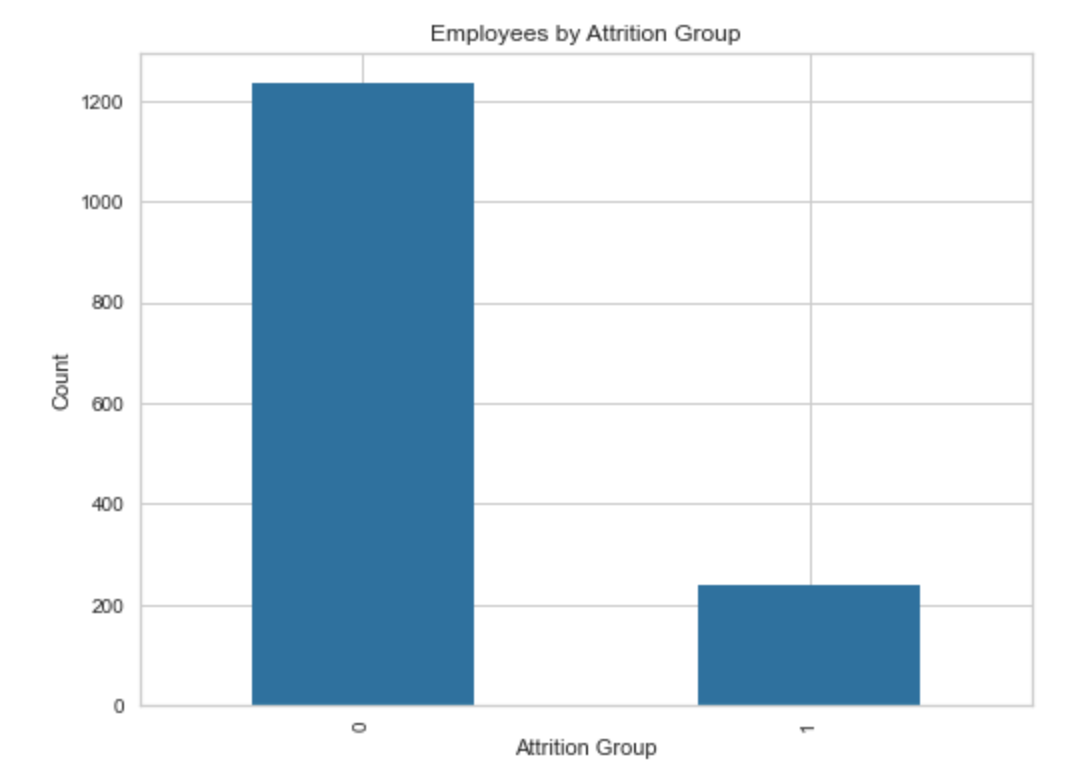
Employee turnover is a problem that every organization faces. The process of hiring and training new employees is costly, from a monetary standpoint as well as loss of production. Employers often survey employees or try to find other ways to assess happiness and to try to understand how to better retain employees. This is the topic that I have chosen to explore for my project.

I used a Kaggle dataset, <https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>, which contains 35 columns and approximately 1500 records of various data on employees. My goal is to find correlations between employee attrition and the other information in the dataset so that we can find ways to predict which employees will leave (or stay) and therefore what actions can be taken to reduce factors that might cause employees to leave and add more factors that would encourage retention.

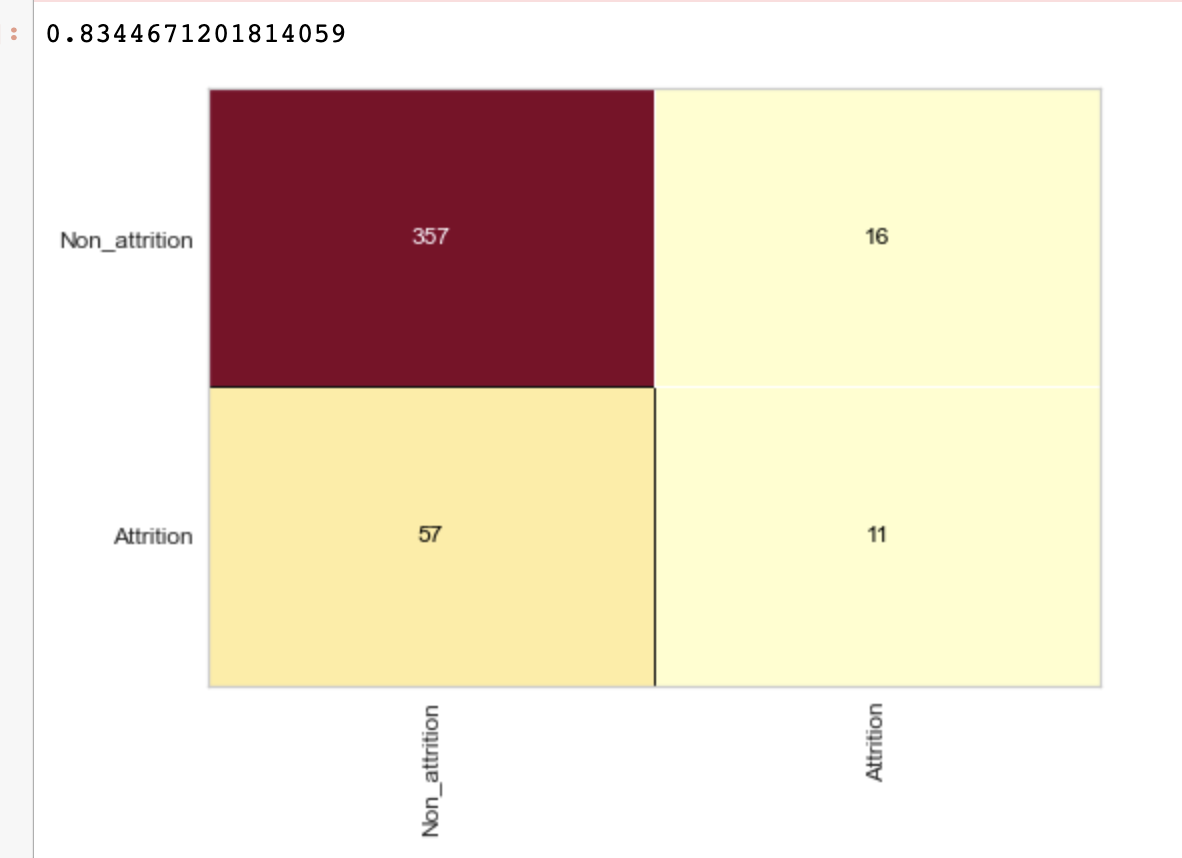
The original columns in the dataset included: Age, Attrition, Business Travel, Daily Rate, Department, Distance from home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Num Companies Worked, Over 18, Overtime, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years at Company, Years in Current Role, Years Since Last Promotion, Years with Current Manager.

The first step I took was feature reduction and data cleanup. Several columns seemed to have the same information just recorded differently, such as hourly rate, daily rate and monthly rate, monthly income, so I kept Hourly Rate and removed the rest. Employee Count, Standard Hours and Over 18 had the same data for all records and therefore had no value. I removed Employee Number as I do not see that in any way impacting or predicting attrition. I was not clear how or what Job Involvement was measuring so I removed that information. Finally, I removed data related to personal happiness outside of work such as relationship satisfaction and marital status. Along with feature reduction, I did some general cleaning of the data, checking for and removing any duplicates, changing categorical values like to numeric.

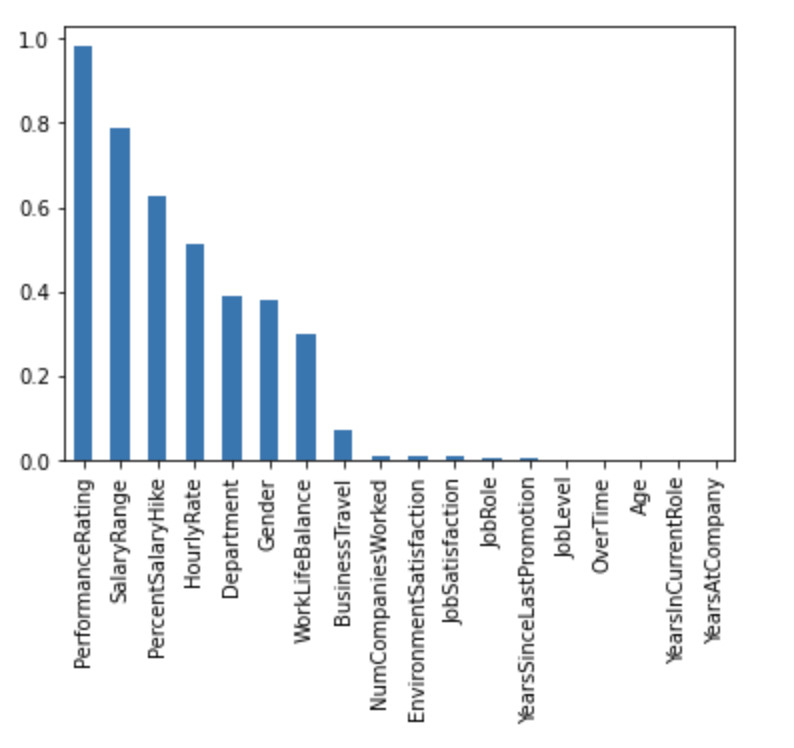
Now that the data is cleaned and in a more usable format, we can start our analysis. The information in this dataset tells us that the current attrition rate is about 16%, we have 237 individuals who have left, compared to 1233 who have not, as depicted in the graph below.



I initially ran a Logistic Regression on the data, which produced the confusion matrix below. The false positive number seems to be a bit high to me.



Since my goal is to determine what factors lead to an employee leaving the company, I ran a chi-square test to determine which variables could be said to have a relationship to attrition. When reviewing results of this test, variables with higher p-values should be considered independent of the dependent variable (Attrition). In other words, variables with lower p-values are more likely to have a correlation to attrition. The results of the test are below:



Based on the p-values from this test, I determined that some of the variables most associated with attrition are years with the company years in current role, age, overtime, job level, years since last promotion, job role and job satisfaction.

Running a logistic regression on the dataset returned an accuracy score of .849 and it also showed that it was predicting an accuracy score of more than one class, which is good. However, as was mentioned previously, the data contained 1233 people who had not left and only 237 who had left. This is considered imbalanced data which could cause bias in the dataset and therefore should be handled before completing the analysis. In order to adjust for the imbalanced data, I used the resample function to upsample the minority class. I then combined the majority class with the upsampled minority class to give the two classes an equal number of records. I used SMOTE to oversample the data.

Now that I have fixed the imbalance issue, I can look at different models and evaluate them. I used a function to compare models using the resampled dataset. The three models that I chose to compare are K Neighbors Classifier, Random Forest Classifier and Logistic Regression. The results I got were these:



Looking at the results of the model compare, Logistic Regression has the best score for each measurement except precision, which the K Neighbors Classifier model has the best score. With an accuracy of .800, this model is about 80% accurate, which is pretty good, however the precision score of .456, recall score of .378 and AUC of .637 are not great. I believe that the lower precision and recall is also related to the high number of false positives that we saw in the confusion matrix.

Based on this information, I am not ready to recommend using this model at this time. I would perhaps recommend taking a little more time on this to evaluate the features that I used in the model. I do not feel that the scores are high enough to justify basing important business decisions on it. I do think there could be value in going back and revisiting which features we are running the models against to see if we might see a stronger correlation with any of the others. I believe a larger data set could be valuable as well.