Predictive Modeling of Employee Termination

Taylor Callahan

Department of Data Science, Bellevue University

DSC 550: Data Mining

Dr. Brett Werner

March 3, 2023

Introduction

Most companies around the world are internally managed by a human resources department. While the human resources department is often responsible for tracking metrics such as diversity, inclusion, relationships, collaboration, compensation, etc., one of their largest responsibilities involves retention. This problem has become one of increased interest in the last few years, as an economic trend of employees voluntarily resigning from their jobs in record numbers—often referred to at the Great Resignation—has hit companies with a force. To retain employees, the human resources department considers factors such as quarterly survey results, manager scoring, productivity, etc., however, one factor that may have been overlooked involves absences.

Especially in companies where paid time off is not paid out when an employee leaves, it may be expected that employees will take more time off prior to voluntarily terminating their position. Therefore, this analysis will examine time off data across the first third of 2022 to identify a trend in time off patterns prior to termination. This data will be considered alongside other potentially influential factors, such as the age of employee, the department in which the employee works, how long they have been in the position, and their annual salary. By the end of this analysis, the goal is to be able to rank employees based on their retention risk. A high-risk employee would have a higher ranking, while a lower risk employee would have a lower ranking.

This analysis may be fundamental in helping human resources departments identify employees who may be considering voluntary termination. With these risk factors in mind, human resources experts can attempt to reach out to employees who may be unhappy in their

job. Not only might this increase productivity due to a happier workforce, it also could function to increase retention across the board.

Data for this analysis represents a real company and was collected internally before individuals and field names were anonymized. Along with anonymization efforts, additional terminations were added randomly to allow for a large enough selection of data.

Summary

Before attempting to build a model for predicting termination, some initial exploratory data analysis (EDA) and data preparation was performed. To prepare the data, a few unnecessary fields were dropped, types were changed, re-indexing was performed, and useful calculated fields were added for averaging absences across the year and months. Along with these basic transformations, dummy variables were created for the categorical features to allow for accurate modeling without the assumption of an ordering across levels. This was performed prior to principal component analysis (PCA), which reduced the number of features to 14 from 26. One final preparatory step completed was oversampling. Due to the much larger sample of individuals who have not terminated from the company, there was an imbalance in the response. This imbalance was managed via oversampling, in which additional rows were inserted containing the minority sample of terminated individuals. With properly cleaned data, further analysis and modeling was then possible.

There were two main goals during initial EDA: to examine termination and absences over time, and to examine termination across the levels of various features, including department, generation, time in job, and salary group, as these fields were first identified as having a higher probability of impacting termination.

First examining absences and termination across time, Figure 1 below illustrates the correlation between the two variables. While there is clear deviation between the two lines, both time off count and termination count appear to follow a similar trend as they grow throughout the year and drop from July to August. This plot contributed to the understanding of the data by suggesting a positive relationship between absences and termination.

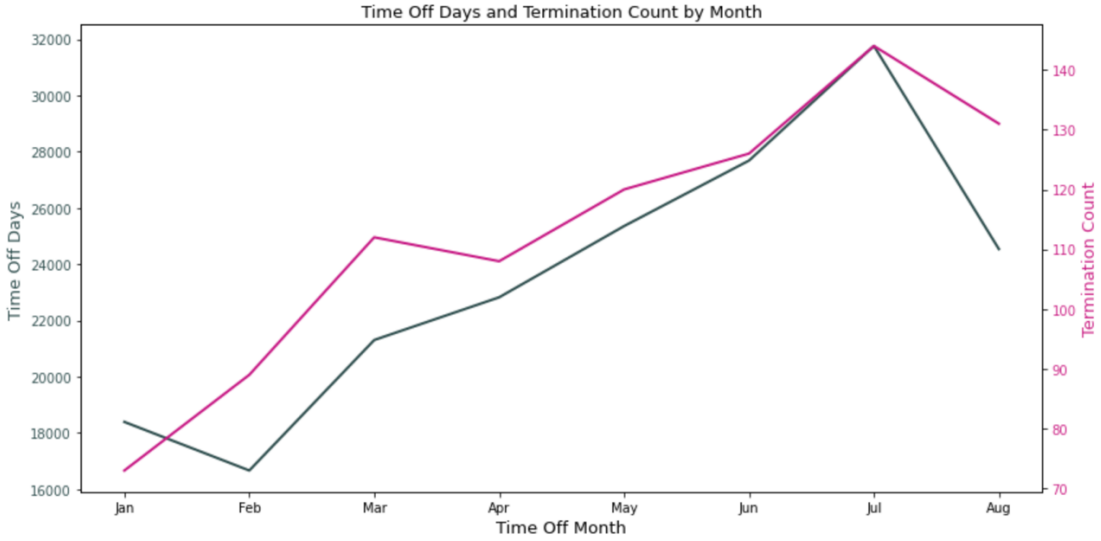

Figure 1. Time Off Days and Termination Count by Month

The above plot was further enhanced by the addition of Figure 2 below, which illustrates the average time off among terminated and active employees by month. Even more evident in this plot than the last, those individuals who terminated within the year take, on average, more time off than those who do not terminate within the year. Both plots contribute to our initial assumption in conducting this analysis.

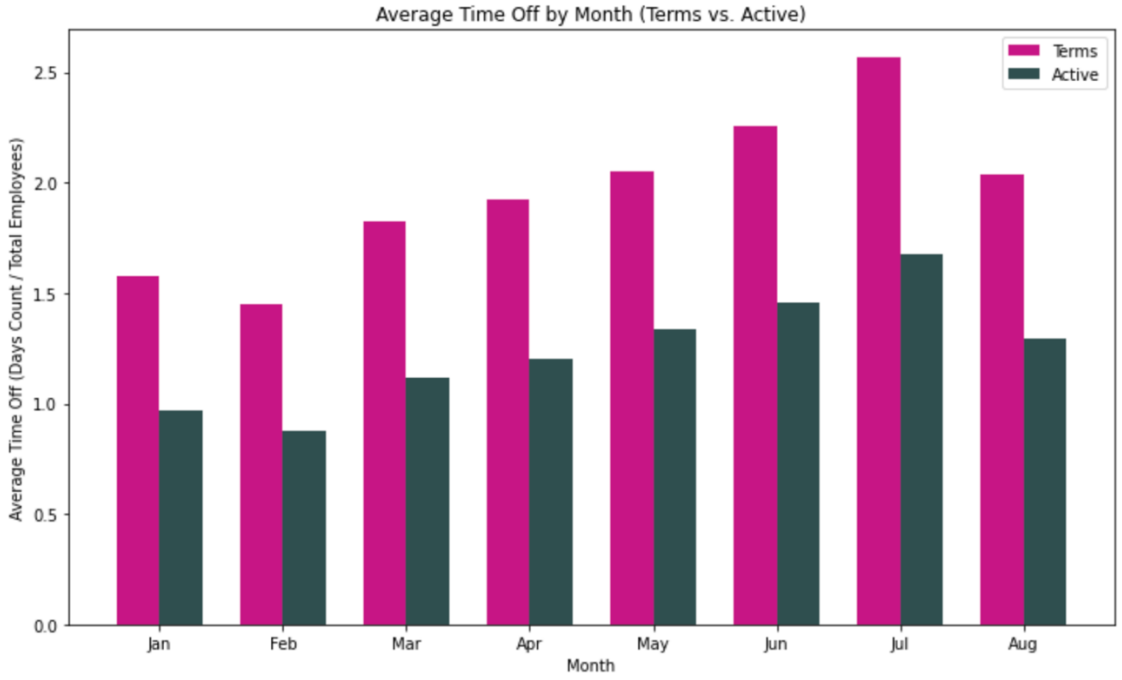

Figure 2. Average Time Off by Month (Terms vs. Active)

While four more plots were generated on the features of interest, these will not be included in this discussion due to a lack of substance. These additional plots examined show how department, generation, time in job, and salary group impacted termination. The plots suggest little to no relationship between these features and termination. Therefore, after our initial round of EDA, there appears to be a clear relationship between absences and termination, while not being as evident of a relationship between termination and the remaining features.

With data cleaning and EDA contributing to a thorough understanding of the data and features of interest, modeling was performed to determine the predictability of termination given the features provided. Although the size of the data made a grid search of model type non-feasible, three models were chosen for testing: K-Nearest Neighbor (KNN), Decision Tree, and Random Forest. Further, due to the presence of imbalanced data, accuracy was not chosen to be best metric for evaluating the final models, and area under the curve (AUC) and Matthew's correlation coefficient (MCC) were also considered. Although AUC is not intuitive for an end-user to interpret, it covers both sensitivity and specificity and performs well on imbalanced datasets. AUC values between 0.9-1.0 will be considered excellent, good for AUC values between 0.8-0.9, fair for AUC values 0.7-0.8, poor for AUC values between 0.6-0.7, and failed for AUC values between 0.5-0.6. MCC summarizes the confusion matrix, also performing well on imbalanced datasets. A higher MCC is considered better.

After performing a grid search of hyperparameters for each model, along with fitting each model with the chosen metrics, the KNN model was chosen as the best model for predicting termination. While both the random forest and decision tree models had an AUC around 94 and a MCC around 88, the KNN model had an AUC of 99 and a MCC of 97. Further, upon analysis of the confusion matrix, which is Figure 3 seen below, only 662 of the predicted outcomes were incorrect, while 55,620 were correct. Therefore, the KNN model had the highest performance across all three model which were tested.

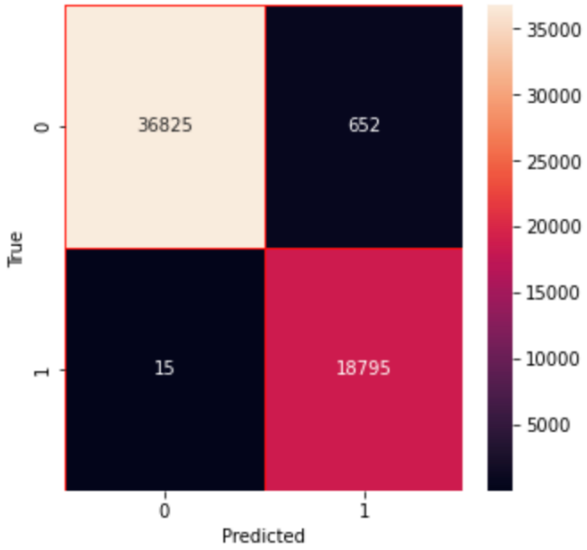

Figure 3. Confusion Matrix for KNN Model

Conclusion

After performing data cleaning, EDA, model building, and model evaluation, it was determined that absence data is a good predictor of employee termination. With additional hyperparameter tuning, the chosen KNN model may perform even better than it has already performed given the data provided and would be ready for deployment.

Although this data did perform well predictively, it would be recommended to perform oversampling at each new iteration of data ingestion. Without oversampling, each model predicted no terminations across the board, resulting in a high accuracy but low predictive effectiveness. Therefore, in the continued use and growth of this data and model, the imbalanced nature of the data would need to be continually considered.

Additional opportunities may also be present in the inclusion of more features. Although not included in this analysis, data exists regarding management hierarchy and whether or not the individual is a manager themselves. The potential for management issues, along with the

additional stress of being a manager, may impact termination and could lead to even greater predictability in a model.

In conclusion, absences can be a good predictor for employee termination, which can allow companies to keep a close watch on high-risk employees to address their concerns and improve retention. A tuned KNN model with oversampled data can provide a starting point for addressing this growing issue.