**Predicting Employee Attrition Across Factors Affecting Workplace Behavior and Satisfaction**

Gillian Tatreau

Department of Data Science, Bellevue University

DSC630: Predictive Analytics

Professor Andrew Hua

12 August 2023

**Introduction**

Employee attrition is when an organization loses an employee and has an outstanding position to fill as a result; therefore, an increased attrition rate indicates that the organization is experiencing a decrease in its labor force, which impacts operations by decreasing productivity and employee morale. From an organization’s standpoint, the most influential queries regarding employee attrition become the likelihood that an employee will leave, and what factors are influential in determining whether an employee leaves the organization.

The labor force is currently experiencing major upheavals as demand for certain roles has disappeared and demand for others has increased exponentially, as employees have reprioritized their own well-being and recognized the value they bring to an organization (Wooll 2022). Many companies have recognized the need to reduce employee attrition, with many organizations focusing on employee recognition, developing flexible work environments and models, encouraging employee well-being with benefits and wellness, and contributing to career growth and employee development (Wooll 2022). Not all companies are able or willing to implement every piece of feedback from employees and must prioritize only the most crucial factors, which would make knowing the factors that influenced an employee to leave a position that much more important for improving those conditions and to decrease the likelihood of losing others due to the same reasons. The main factor for most organizations in maintaining a low employee attrition rate is the cost. The estimated average cost of hiring a new employee is $4,000 (Vasconcellos 2023). This could result in an exorbitant amount of money if employees are leaving the organization frequently. It is also estimated that it takes most employees anywhere between three and six months to feel that they are positively contributing to the organization through their role, which means that that time to get employees trained and comfortable in their role is another investment for organizations to consider (Stibitz 2015).

When we consider the cost of hiring and time dedicated to training, it becomes clear that most organizations would be interested in being able to predict the probability of an employee leaving in order to either prepare to hire a replacement or intervene in some way so that employee no longer wishes to leave. The ability to predict whether an employee is going to leave an organization and the factors that are influencing that decision would be information especially valuable to the organization’s HR department, or the person that handles employee retention, hiring, and records.

The data set being used in this study was found on Kaggle. According to the original source of the data, it is a fictional data set created by IBM data scientists (Karanth 2020). This data set contains information on both job-related and personal factors that affect an individual’s career. Some of the fields that intuitively seem most relevant to employee attrition include job performance rating, daily pay rate, distance from home, whether the employee must travel for the role, the employee’s satisfaction with their work environment, the employee’s perception of their work-life balance, stock options available to the employee, and the standard number of hours worked. This data includes a broad range of variables that could influence an employee’s decision to stay at their current position that include personal reasons (such as relationships within the workplace and at home) and workplace-specific (number of training opportunities within the past year). This data set is very convenient to answer these questions; however, since it was manufactured it would be helpful to continue to search for other data sets to corroborate this data in some way, such as data from a real company, more data on the employees’ employment history (to determine average number of years at each organization), data on employee performance that is more than just a ranking, and perhaps most importantly, a data set that includes a field for the reason why an employee left the position, for example, if they were terminated, left voluntarily for other opportunities, or retired.

**Methods/Results**

The class weights of the target variable are highly unbalanced, which requires special consideration when modeling.

A blue and white rectangular object with black text

Description automatically generated

*Figure 1: Horizontal bar chart for target variable: Attrition.*

Most of the employees travel rarely for work, reported no overtime, and received a performance rating of 3 instead of 4. Most of the employees reported 0 and 1 for stock options, and assuming the intuitive leveling of 0 being no stock options, this could possibly be a reason for attrition for some employees. Also, if a value of 1 for JobLevel is equivalent to an entry-level position, many of the employees are lower-level employees. There are high reports of job satisfaction; however, if this was collected via a survey from HR, this could be a highly biased answer.

A graph with numbers and a bar chart

Description automatically generated

*Figure 2: Histogram of Age, with the two classes of Attrition overlaid.*

The distribution of age for employees that experienced attrition is similar in distribution to those that did not, with a slight deviation at the lower age range where there appears to be a slightly higher value of attrition than is seen in the overall population.

A graph with numbers and a chart

Description automatically generatedA graph with numbers and a bar chart

Description automatically generated

*Figure 3: Histograms of YearsAtCompany (left) and YearsInCurrentRole (right) with Attrition classes overlaid.*

One possible explanation for the two spikes in the attrition class at 0 and 2 years would be that there are cycles at which attrition is most likely to occur- within the first year and again within the second year. After that, it appears from the graph that attrition becomes less common. If this trend is statistically significant in the data, it would suggest that organizations should spend the most amount of resources making sure that new hires acclimate and and to focus on other factors that prove most predictive within the first year of an employee being with the organization in order to prevent high levels of attrition at the earlier stages of their career with said organization.

A graph with numbers and a bar chart

Description automatically generated

*Figure 4: Histogram of YearsWithCurrManager, with Attrition classes overlaid.*

In this plot, we see a clearer bimodal distribution than in the overall plot for YearWithCurrManager. For the attrition class, the plot suggests the trend that attrition is perhaps more likely when management is changed. Also looking at the histogram of YearsInCurrentRole from figure 3, we see that the two variables follow a very similar distribution, which makes sense given that a new role in a company typically comes with a new manager.

There were several times during the data exploration process that the fact that the data set was created by data scientists became relevant; for example, there was no missing data or irregular values that needed to be corrected. Other than that, the data preparation process focused on the following four elements: removing irrelevant variables, converting categorical variables with only two values into 0 or 1, creating dummy variables for both the categorical and ordinal variables, and standardizing the data. The variables Over18, EmployeeCount, and StandardHours were all removed because each contained only a single value for every row. EmployeeNumber was removed because it was a unique identifier for each employee profile. In Gender, Male was recoded to 1 and Female to 0. Values of No were recoded to 0 and Yes to 1 in Overtime and Attrition. Values of 3 were recoded to 0 and 4 to 1 in PerformaceRating. All other ordinal and nominal categorical variables were encoded as dummy variables.

Determining whether an employee will leave the organization is a classification problem, with only two outcomes: either an employee left or did not. Logistic regression models, k-Nearest Neighbors, and Decision Tree models will be built, as all three algorithms excel as classification models. Logistic regression was chosen because it will not only predict the outcome, but it will predict the probability of that outcome as well. k-Nearest Neighbors models make predictions by comparing each employee to their most similar fellow employees; this has the possibility of giving insight into the type of employees that leave. Decision trees have the benefit of being highly explainable. Due to the nature of the primary question, the best evaluation metrics will be building a confusion matrix for each model built and looking at both the precision and recall scores for the models. A confusion matrix will show the number of correct predictions for each outcome as well as the incorrect predictions, which would give insight into where the model might be failing if it overpredicts one outcome incorrectly over another. Also, instead of looking only at the accuracy, precision and recall will be used to evaluate the models because these metrics will give greater insight into whether the model is overpredicting that an employee will leave or that the model is predicting too few people leave.

The first three models trained (logistic regression, k-Nearest Neighbors, and decision tree) each had their own strengths and weaknesses. The logistic regression model had the highest accuracy (88.4%), but only middling precision and recall scores (58.6% and 43.6%). The k-Nearest Neighbors model had a slightly lower accuracy (87.8%) but had a perfect precision score of 100% but a poor recall score (7.7%). The decision tree model was grossly overfit, which was demonstrated by the 100% accuracy on the training data, but which resulted in only about 76% accuracy on the test data, 16% precision and 18% recall. The logistic regression score had the highest f1 score (52.8%).

The k-Nearest Neighbor model excels at predicting when an employee will not experience attrition but does not do well predicting when an employee will experience attrition. The model that was best at predicting when an employee would experience attrition is the logistic regression model, with the highest (although not excellent) recall and f1 scores of the three models.

The next three models built represent different attempts to improve model performance. The logistic regression model with only the top 10 features (identified in the first logistic regression model) grossly underperformed with only a slight decrease in accuracy (a drop from 88.4% to 83.7%) but significant decreases in the precision, recall, and f1 scores. The decision tree built with these same 10 features represents an improvement from the overfit decision tree, with an accuracy of 77.9%, 24% precision, and 30.8% recall. This does not rival the original logistic regression from above (as this model has only a 27% f1 score) but it is the highest performing model type other than the logistic regression algorithm. The imbalanced class weights are addressed in the final model built, where the training data was up sampled so the class weights for Attrition would be equivalent; however, this model does not necessarily outperform the original logistic regression model as the accuracy, precision, and f1 scores are all lower in this newer model. The recall score is much higher (at 71.8%), which tells us that this model is slightly better at identifying when employees will experience attrition, but it also has a high false-positive rate as well.

**Conclusion**

Feature importance analysis was conducted on both the logistic regression and the decision tree models. Both indicate that OverTime, YearsInCurrentRole, TotalWorkingYears, and YearsSinceLastPromotion are factors that are important in whether an employee experiences attrition. The feature importance analysis based off the logistic regression model (based on the coefficient size) should be given greater consideration than the features chosen by the decision tree, as the first decision tree was greatly overfitted.

An ensemble model that combines the performance of the logistic regression model with up sampling (highest recall) and the k-Nearest Neighbors model (highest precision and lowest false positive rate) should be examined later in order to possibly create a better model for the business objective: identifying which employees are likely to experience attrition. Otherwise, the best model built in this study was the first logistic regression model, with the highest f1 score, which represents the best balance of precision and recall.

The modeling process has identified that it is possible to predict the likelihood of an employee leaving an organization and some of the most influential factors in the decision to leave. This information allows an organization to prepare for when an employee might be leaving their position, to make structural changes that might convince that individual to stay, or to prevent future attrition. There are several ethical concerns that come up with a model such as this, depending on the way an organization decides to use the information gained. The most ethical, and potentially better business, decision would be to focus on ways that attrition could be prevented by implementing structural changes when possible and creating an environment where employees feel comfortable talking to their employers when they are dissatisfied and finding solutions to rectify that. A far more insidious application of this information would be if an organization were to make hiring decisions based off whether the model predicted that a potential employee would have a high probability of leaving the company, or for an organization to preemptively terminate an individual when the model predicts that they might leave at some point in the future.

The way in which this model is deployed is paramount to the ethicality of the project. The most ethical, and likely most effective, application would be to focus only on the features that are identified as the highest predictors of attrition in the model that the organization is able to control. For example, the logistic regression model identified OverTime, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, JobSatisfaction, and EnvironmentSatisfaction as being highly predictive variables that the organization ultimately has control over and can positively impact. NumCompaniesWorked and TotalWorkingYears, while also highly predictive, are not factors the organization should attempt to control by altering their hiring criteria to filter out applicants with resumes that have work experience that would imply they might experience attrition at this company.

Also, when the model is deployed, the probability for each class should be considered more carefully than the predicted class itself because people with a probability that favors attrition could be brought in for some kind of friendly intervention that considers what that employee might be experiencing and ways that they can be better supported or accommodated in their workplace.

**References**

*Employee attrition and factors*. Kaggle. (n.d.). https://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors

Karanth, Manohar. (2020). *Tabular summary of HR analytics dataset*. [Data set]. Zenodo.

https://doi.org/10.5281/zenodo.4088439

Stibitz, S. (2015, August 12). *How to get a new employee up to speed*. Harvard Business Review. https://hbr.org/2015/05/how-to-get-a-new-employee-up-to-speed

Vasconcellos, E. (2023, February 21). *How much does it cost to hire an employee?*. Business News Daily. https://www.businessnewsdaily.com/16562-cost-of-hiring-an-employee.html

Wooll, M. (2022, January 24). *Employee attrition vs. turnover: What’s the difference?* BetterUp. https://www.betterup.com/blog/employee-attrition