|  |
| --- |
|  |

Amanda Loy

Southern New Hampshire University

DAT 430: Leverage Data for Org Results

Final Paper

June 18, 2023

**Establish a baseline**

A screenshot of a computer

Description automatically generated with medium confidence

To fully understand what information I'm working with I start with importing the data into the data frame. Then I want to list all of the head columns to see what information I can gain from this data frame. Looking at this information I know that the business problem I want to address is which variables impact attrition within the company. So I'm going to focus my attention on those variables that would impact attrition such as job satisfaction job involvement training times and of course attrition.

A screenshot of a computer

Description automatically generated with medium confidence

By pulling out this information I am able to see the sums the total counts and the variables that I am interested in for their impact on attrition. From this I can tell that the information is rated on a scale that will determine how I view the data

A screenshot of a computer program

Description automatically generated with low confidence

The first visualization that I pulled from the exploratory data was to see what the training times versus job satisfaction would possibly be. Above the data really isn't clear based on the bar graph that was provided it's hard to tell how job satisfaction and training hours are actually related. Therefore I created a confusion matrix on a heat map that can be viewed below. It shows that there is a strong correlation between job satisfaction and training hours. The second strongest correlation seems to be between job satisfaction and job involvement.

A screenshot of a graph

Description automatically generated with low confidence

After being exploratory data and the heat map that was created I can see that I want to focus on the variables job satisfaction and training hours in relation to attrition. I will continue to look at job involvement and see if the correlation becomes higher when against attrition

**Select and create the appropriate features for your predictive model. Include screenshots of your engineering of features.**

A screenshot of a computer

Description automatically generated with medium confidence

the features that I will be utilizing for the predictive analysis include job involvement, job satisfaction, and training times last year. Based on the heat map conducted in the exploratory analysis, I believe these will be the strongest in determining attrition and retention within the company.

A screenshot of a computer program

Description automatically generated with low confidence

Here I'm preparing the data to be entered into a decision tree by splitting the features and the target into their separate data frames. I'll begin then by doing a train test split which will split all of the data collected into various variables that will be used to train the decision tree and then test the decision tree.

A screenshot of a computer

Description automatically generated with medium confidence A screenshot of a computer program

Description automatically generated with low confidence

Above is the code that was utilized to decide how to prune the tree to find the best fit for the training data. Based on this after here we'll move to test the data utilizing the tree

**Apply two predictive models to the data to show variation in predictability of the results. Include screenshots of your models.**

A screenshot of a computer

Description automatically generated with medium confidence

Above are the results for the decision tree when tested for accuracy. It is shown to provide precise precision and accuracy within .45 of yes values are correctly determined. The model performs relatively well in predicting the "No" class (majority class) with high precision, recall, and F1-score. However, the model performs less effectively in predicting the "Yes" class (minority class) with lower precision, recall, and F1-score. The overall accuracy of the model is 86%, indicating that it correctly predicts the majority of instances in the test dataset. Considering the class imbalance, it is important to further evaluate the model's performance and consider techniques such as oversampling the minority class or adjusting the decision threshold to improve predictions for the "Yes" class.

A screenshot of a computer

Description automatically generated with low confidence A screenshot of a computer

Description automatically generated with medium confidence

The logistic regression model suggests that "JobSatisfaction," "JobInvolvement," and "TrainingTimesLastYear" have significant effects on the likelihood of "Attrition." The coefficients indicate the direction and magnitude of these effects, with negative coefficients suggesting a decrease in the odds of "Attrition" with higher values of the corresponding independent variable. However, further analysis and interpretation are needed to fully understand the practical significance of these findings and their implications for predicting attrition in the given context.

A screenshot of a computer program

Description automatically generated with medium confidence

All four predictors (const, JobSatisfaction, JobInvolvement, and TrainingTimesLastYear) are statistically significant in predicting the outcome.JobSatisfaction and JobInvolvement exhibit the highest statistical significance, followed by TrainingTimesLastYear, while the constant term (const) also contributes significantly.

Based on the logistic regression model results, which include significant coefficients for the variables JobSatisfaction, JobInvolvement, and TrainingTimesLastYear, it indicates that these variables have an impact on the likelihood of an employee experiencing attrition in the company. With this information, you can use the logistic regression model to make predictions and classify employees into categories of staying or experiencing attrition based on their JobSatisfaction, JobInvolvement, and TrainingTimesLastYear values. The model's coefficients provide insights into the direction and magnitude of the effects of these variables on attrition. By considering these coefficients, you can assess which variables have a stronger influence on attrition. For example, a higher negative coefficient for JobSatisfaction suggests that lower job satisfaction is associated with a higher likelihood of attrition.