**7-2 Project Two Part 1**

Kayla Sacks

Southern New Hampshire University

DAT-430 Leverage Data for Org Results

Professor Aubrey Condor

October 20, 2024

## Introduction

There are multiple steps to be taken when analyzing the HR attrition data set to identify why employees are leaving their jobs and what changes can be made to reduce the number that are leaving. We will start with using exploratory data analysis (EDA) to analyze the data set to gain a better understanding of it as well as find patterns and relationships through both statistical and visual techniques (Mahadevan, 2022). EDA is a key step to gain insight and create a hypothesis to answer our questions.

After setting the baseline, the feature selection will be the next crucial step, as choosing the correct features will help us find the answer to our questions. Once the features are chosen, a logistic regression model and random forest model will be created to visualize the data and show variation in predictability. Finally, this paper will determine if our method has brought us suitable answers to our questions and solutions to our problem.

## Baseline and Feature Engineering

For the exploratory data analysis (EDA), the libraries that will be used are loaded and the HR Attrition Data is imported so that we can begin to analyze the data. Viewing the information and first 5 rows of the set can help give us some perspective on the data set. These steps can be seen in Figure 1.

To not have to use different methods of correlation, we will convert the categorial variables to numeric by factorizing them and showing the index of each variable that has been changed as shown in Figure 2.Since attrition is the variable that is being investigated, it will be the first variable to look at by viewing the counts for each value as well as the percentage of how many entries are in each value, which can be seen in Figure 3.

A screenshot of a computer

Description automatically generated

Figure 1 First steps of EDA

A computer screen shot of a code

Description automatically generated

Figure 2 Convert categorical data to numeric using factorization

A screenshot of a graph

Description automatically generated

Figure 3 Show Attrition value counts.

The first thing that is learned from this EDA looking at Figure 3 is that 16% of employees have quit.

After converting the data, we can check the statistical information for each column, for missing values, and duplicate entries. This is a crucial step to ensure the data set is clean and accurate.

A screenshot of a computer

Description automatically generated

Figure 4 Checking column statistics, missing data, and duplicate entries

Now that we know information about the data set and that there are no null or duplicate entries, we can move on to our question of what variables influence attrition by checking the correlation between the variables. One of the easiest ways to check correlation is by using a heatmap. From the map below (Figure 5), we can begin to answer our questions of why employees are leaving their jobs. It is important to note that since there are a high number of variables, the heat map was created to only show the correlation to our dependent variable of Attrition.

From the heatmap we can see that several variables show slight correlation, but since correlation from 0 to .19 is very weak (The BMJ, 2019), we can assume they do not have a significant effect on attrition. We see that there are 4 variables that have a moderate to very strong negative correlation with one variable having a weak positive correlation. From this we have learned that there are at least 5 variables that could be linked to why employees are leaving their jobs.

A graph of a number of people

Description automatically generated with medium confidence

Figure 5 Correlation Heatmap

Most of the features shown on the heatmap have a very week correlation with Attrition, so features with a correlation between -0.2 and 0.2 will be dropped so we can focus on the those that are showing higher correlation. This leaves the features of JobSatisfaction, OverTime, PerformanceRating, WorkLifeBalance, and YearsWithCurrManager. Figure 6 shows the features that were dropped and the new heat map with the features that will be used.

A screenshot of a computer

Description automatically generated

Figure 6 Removal of irrelevant features and new heatmap

Figure 7 shows attrition graphed in relation to each of these features to gain a better understanding of their correlation.

A screenshot of a computer screen

Description automatically generated

Figure 7 Line graph between attrition and each feature kept

Now that features are chosen and there is an understanding of their relation to attrition, the next step will be to test the variables by creating models.

## Models

### Logistic Regression

The first model that is used is a logistic regression model. The data was split into training and testing sets before creating the regression model. Once the model was built, a confusion matrix and classification report were created to check the accuracy of the model. This process can be seen in Figure 8.

A screenshot of a computer

Description automatically generated

Figure 8 Logistic Regression model, confusion matrix and classification report.

### Random Forest

The next model will be a random forest. After importing the necessary libraries, the data was again split into training and testing sets using a different random state from the first model. I used these sets to create a random forest model with 100 trees. The accuracy of the model and confusion matrix can be seen in Figure 9. The first decision tree from the random forest model can be seen below.

A screenshot of a computer

Description automatically generated

Figure 9 Random Forest Model

A screenshot of a computer

Description automatically generated

Figure 10 Visualization of first decision tree in the random forest.

## Outcome Assessment

The question that needs to be answered throughout this analysis is why employees are leaving their jobs with the goal of finding patterns within the data that could help us gain insights into the problem, its causes, and effective solutions. The logistic regression model accuracy was 100% with the random forest model accuracy at 74.2% using the variables of JobSatisfaction, OverTime, PerformanceRating, WorkLifeBalance, and YearsWithCurrManager.

Based on the information gained from this analysis, there are a few practical solutions that can be implemented to reduce the number of employees quitting their jobs which are:

1. Reducing the amount of overtime
2. Provide different work schedules that could fit employee life schedules to improve work life balance
3. Improve the work environment and provide encouragement and benefits to employees to increase job satisfaction.
4. Provide additional training and constructive feedback to employees to increase performance ratings.

Although these methods might not be perfect, based on this analysis, they should help to reduce the number of employees leaving their jobs.

## References

Mahadevan, M. (2022, July 31). *Step-by-Step Exploratory Data Analysis (EDA) using Python -*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/>

The BMJ. (2019). *Correlation and regression* . Bmj.com. <https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression>