Brandi Jones

DAT-430-T6751

Leverage Data For Org Results

August 12, 2022

7-2 Project Two

* **Establish a baseline using the HR Attrition Data set, located on the Apporto desktop in the DAT 30 folder. Include a screenshot to show your baseline.**
  + **Use exploratory data analysis (EDA) to answer the following: What did you learn from the data as it relates to the original business question(s) we are trying to answer?**

The original business question asks what metrics are most influential in predicting attrition. After using EDA techniques, I have determined that the most influential metric is Job Satisfaction. Job Satisfaction is most highly influenced by Performance Rating, which is most highly influenced by Training. I created many different visualizations outlining attrition rates compared to different features to learn about our data set, which can be seen below.

* + **Create visualizations from your EDA (using Plotly, etc.) that can be used to describe what you have learned from the data.**

Total Attritions

**Chart, pie chart

Description automatically generated**Chart, bar chart

Description automatically generated

**Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated**

**Chart

Description automatically generatedChart, bar chart

Description automatically generated**

**Chart, bar chart

Description automatically generated**

Chart, bar chart

Description automatically generatedChart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

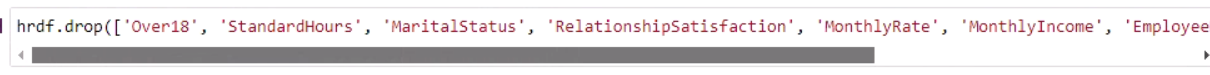
Description automatically generated

* **Select and create the appropriate features for your predictive model. Include screenshots of your engineering of features.**
  + **Which features will you use in your predictive analysis?**

The features I will be using in my predictive analysis include every column except for Over18, StandardHours, MaritalStatus, RelationshipSatisfaction, MonthlyRate, MonthlyIncome, EmployeeNumber, EmployeeCount, and Education. I want to test each other feature against Attrition to determine the metrics that influence attrition rates the most.

* + **Using Python, show how you will engineer your features to prepare your data set for your model.**

After importing the necessary libraries and datasets, I used Python to first drop the columns that I will not be using.



Next, to learn more about my data I used the describe function. This allows me to view counts, means, minimums, maximums, and more about each numerical column in my dataset.

Graphical user interface, table

Description automatically generated

To learn more about my categorical data, I first separated each categorical column into a new data frame. I can then use the list function to view my columns, and the value\_counts function to learn more about each specific column.

Text

Description automatically generated with low confidence

Text

Description automatically generated Graphical user interface, text, application

Description automatically generated

My next step was to take care of any missing values. I filled in any missing numerical values with the mean of the column, and then dropped the missing categorical values.

Graphical user interface, text, application

Description automatically generated

I also noticed that the Gender column had some observations as Male or Female, and some had M or F. I replaced M and F with Male and Female so the column would be formatted the same for all values.

Graphical user interface, application

Description automatically generated

The next step in preparing my dataset for modeling was to change my categorical values to integers for my models to run smoothly. I used the get\_dummies command to do this, and then dropped another two columns that I am not interested in.

Text

Description automatically generated with medium confidence

Table

Description automatically generated

I then used the head command to verify that my values were changed appropriately. The data is now ready to be fitted to my models.

* **Apply two predictive models to the data to show variation in predictability of the results. Include screenshots of your models.**
  + **Use the two models you selected in the Module Five assignment.**

The two models I selected in the Module Five assignment were Logistic Regression and Random Forest. I began with logistic regression, and I checked my models accuracy, ROC/AUC curve, and confusion matrix. I ran a logistic regression model on many different features, the following screenshots are for only two of the models that I ran.

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Chart, line chart

Description automatically generated

Chart

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text

Description automatically generated

Chart, line chart

Description automatically generated

Chart

Description automatically generated

After running my Logistic Regression models, I then ran my Random Forest model. I also checked the accuracy and confusion matrix of this model, and also checked for important features.

Graphical user interface, text, application, email

Description automatically generated

Chart

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Chart, histogram

Description automatically generated

After running my Random Forest model, I also ran a Logit Summary to check my coefficients.

Graphical user interface, text, application

Description automatically generated

The following is a small snippet of the summary.

Graphical user interface, table

Description automatically generated

I also plotted a trend line to compare certain features to Attrition.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart, scatter chart

Description automatically generated

* **Explain how accurately you can predict outcomes based on the data.**
  + **What is the question being asked of you?**

The question being asked of me is to determine which metrics affect attrition rates the most.

* + **How can you answer it?**

My analysis has determined that Job Satisfaction greatly impacts attrition. Using my two models, it has shown that higher training levels leads to higher performance ratings. Higher performance ratings lead to higher job satisfaction which leads to lower attrition rates.