Capstone Project 1: Milestone Report

This project depicts a fully-featured end-to-end Data Science workflow aimed at building models predictive of employee attrition (whether or not an employee leaves a company, in this case: IBM).

The tech industry currently has the highest employee turnover rate at 13.2%. Although big tech companies like Google provide great benefits packages and amenities at the office, employees stay at google for an average of 1.1 years. I am hoping that this analysis of employee attrition will help companies look into other variables that may cause employees to leave the company like job involvement, and environment satisfaction. My client would be the HR department at IBM. This analysis would help them look at other factors that may contribute to employee attrition. Is there another way the tech industry can retain employees? Are there other factors that play into employee attrition? Does income really matter when it comes to employee attrition? Do certain departments have a higher attrition value? I think these are important questions to ask because the tech industry is a growing industry with a lot of young people as employees. I think its important to find ways to keep employee attrition low in this growing industry.

Data Acquisition

The dataset used in this analysis is IBM HR Analytics Employee Attrition & Performance. The dataset is available to download from Kaggle. This dataset contains 35 variables that may contribute to employee attrition including, work-life balance, years at the company, age and relationship status. The dataset contains approximately 1470 entries. I read the csv with pd.read_csv() into a pandas data frame.

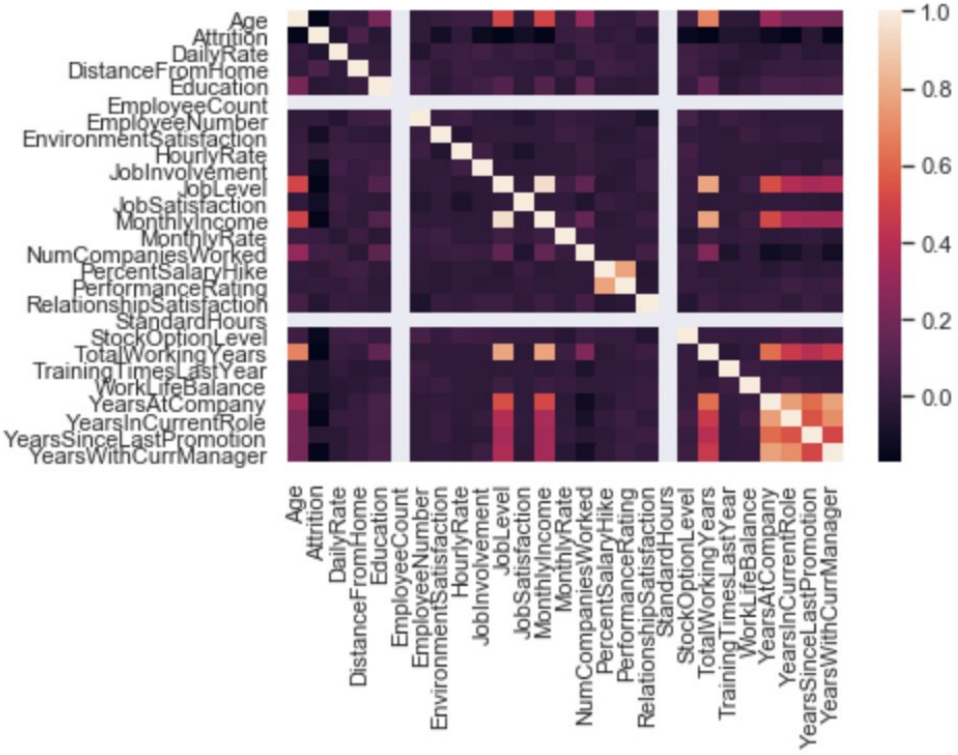
Data Cleaning & Pre-Processing

The data set was already extremely clean with no missing values. Converting the target variable (Attrition) into a binary variable (0/1).

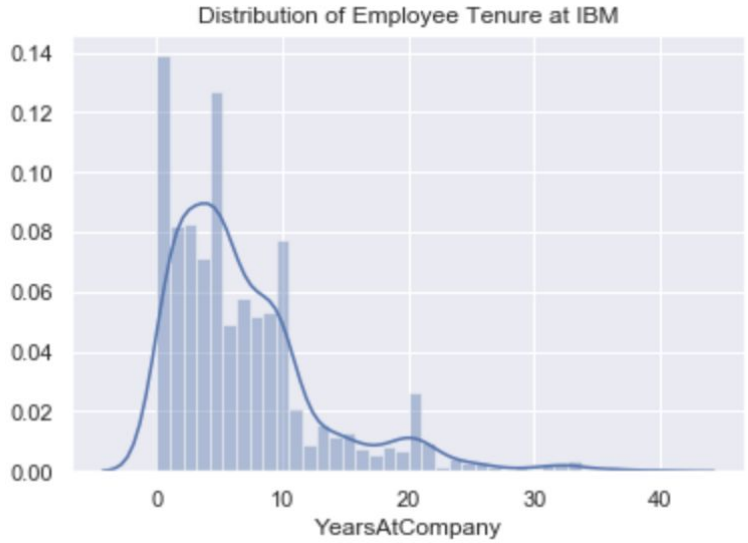
Exploratory Data Analysis:

For the exploratory data analysis, I created a correlation plot to see if any of the variables were correlated with each other. I found that very few variables were correlated. This correlation plot shows that years at the company, years in the current role, years since last promotion and years with the current manager are all correlated

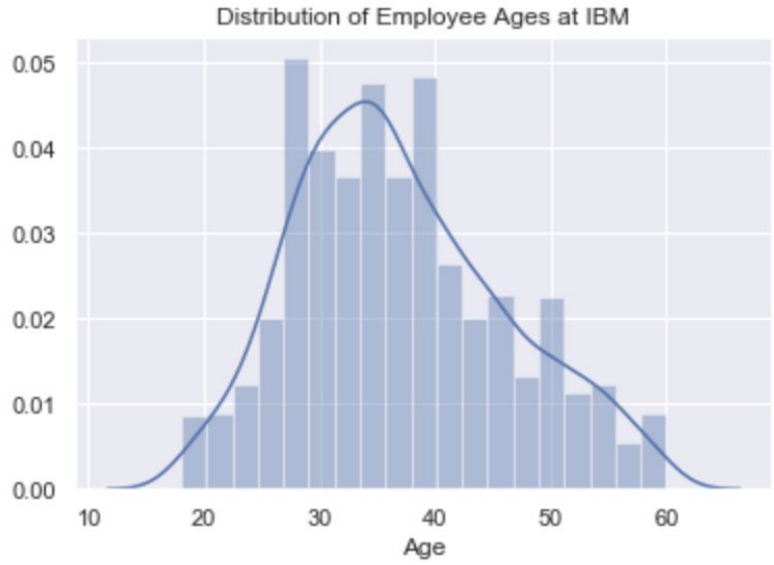
with each other. Job level and monthly income are also correlated with these four variables. Age is correlated with job level, monthly income, total working years.



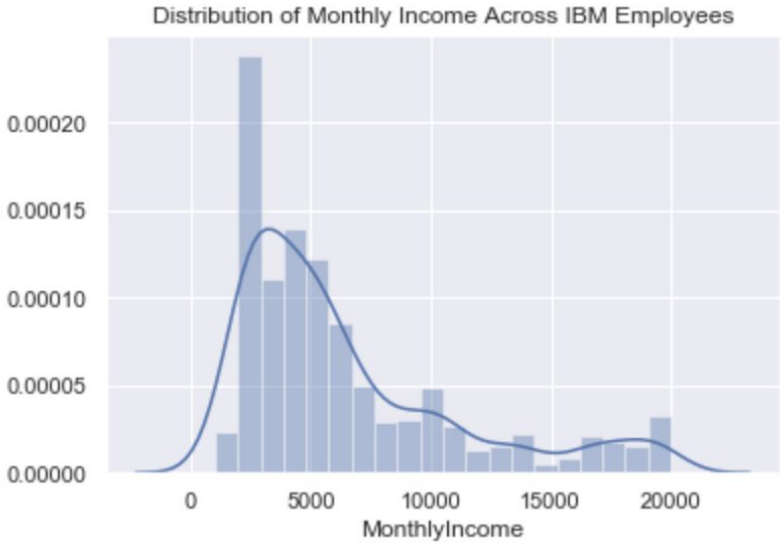
I then plotted the distribution of variables that I thought would be significant in predicting attrition such as age, employee tenure, and monthly income.



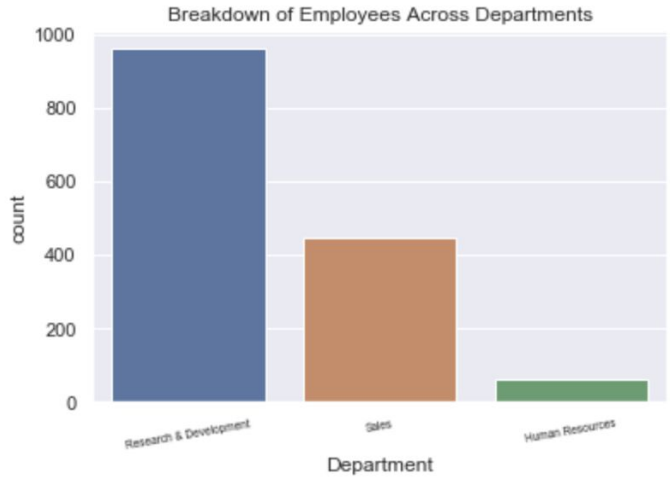
This graph indicates that the average time spent at IBM is approximately 3 years.



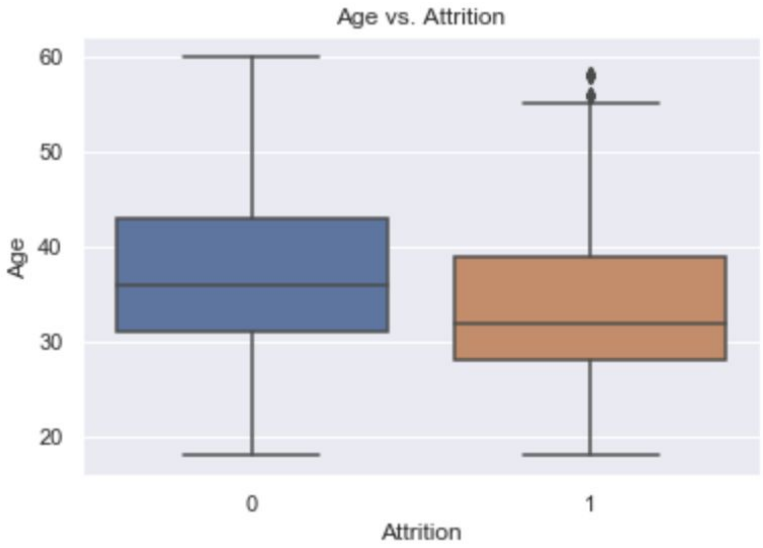
This graph indicates that the average age of Employees at IBM is approximately 35 and is slightly right-skewed indicating that the majority of employees at IBM are young.



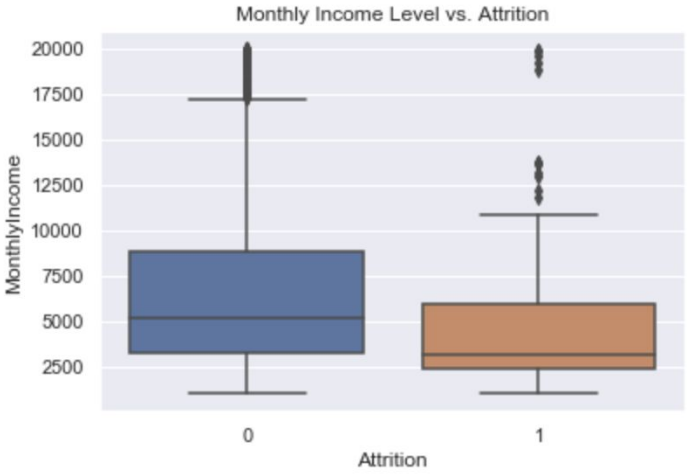
This graph indicates that the average monthly income is less than 5000 and the graph is right-skewed indicating that the majority of the employees make 5000 or less.



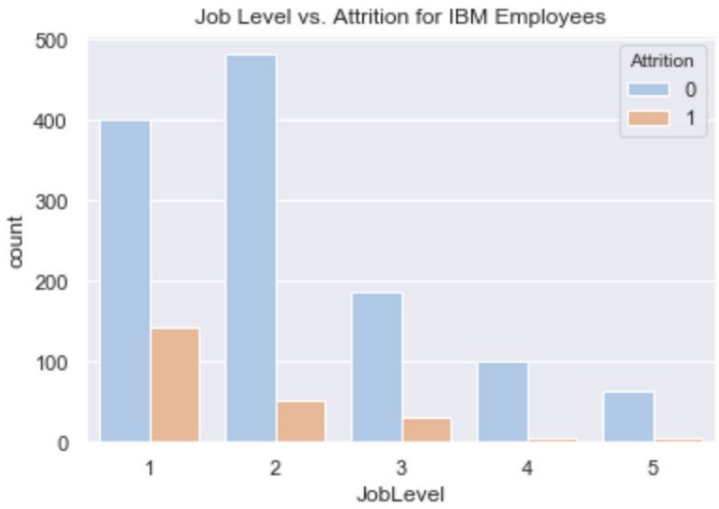
Around 900 employees work in Research & Development and approximately 400 employees work in Sales with around 70 employees working in Human Resources.



These boxplots show the age distribution of employees that stayed at IBM versus the employees that left. The distribution of employees that left IBM has a mean at approximately 32 years old while the mean of employees that stay is around 36 years old. This shows that younger people are more likely to leave IBM.



These boxplots show the age distribution of employees that stayed at IBM versus the employees that left. The distribution of employees that left IBM has a mean at approximately 32 years old while the mean of employees that stay is around 36 years old. This shows that younger people are more likely to leave IBM.

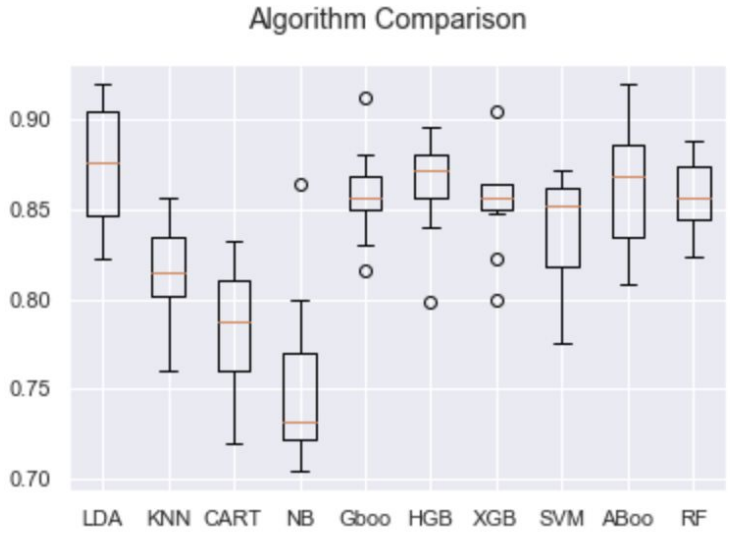


This graph shows that entry-level employees are more likely to leave the company than employees at a more senior level.

Baseline models:

There are many different machine learning models that can be used to predict employee attrition. We will start by testing a range of models and then focus on improving the high performing models. We will start by testing 10 different models with the original dataset.

```
LDA: 0.875058 (0.033246)
KNN: 0.815052 (0.025411)
CART: 0.783032 (0.037393)
NB: 0.751781 (0.046765)
Gboo: 0.858265 (0.024947)
HGB: 0.863839 (0.026924)
XGB: 0.853458 (0.025977)
SVM: 0.839039 (0.031078)
ABoo: 0.863052 (0.034974)
RF: 0.857465 (0.021273)
```



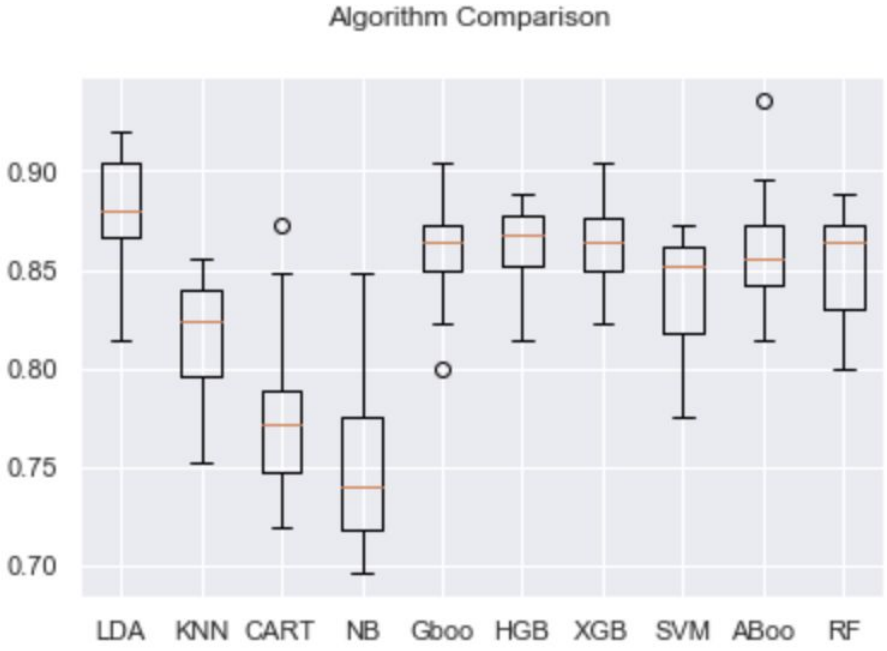
Feature Encoding

A few additional variables that might be important when predicting attrition.

- far_distance is a variable for employees that live more than 10 miles away from work.
- young_no_love is a variable for employees that are still very young and have no relationship obligations.
- migrating_worker is a variable for employees that move around companies often.
- comp_adjusted is a variable to compare the adjusted hourly wage with the monthly wage to determine how well-compensated an employee is based on time spent in the office.
- distance_comp is a variable that is a ratio of total monthly income to distance from the work location

Machine Learning:

```
LDA: 0.879052 (0.031583)
KNN: 0.814206 (0.034774)
CART: 0.778232 (0.047154)
NB: 0.750974 (0.046177)
Gboo: 0.858258 (0.027903)
HGB: 0.861452 (0.024033)
XGB: 0.861458 (0.022599)
SVM: 0.839039 (0.031078)
ABoo: 0.861452 (0.033585)
RF: 0.852645 (0.030421)
```



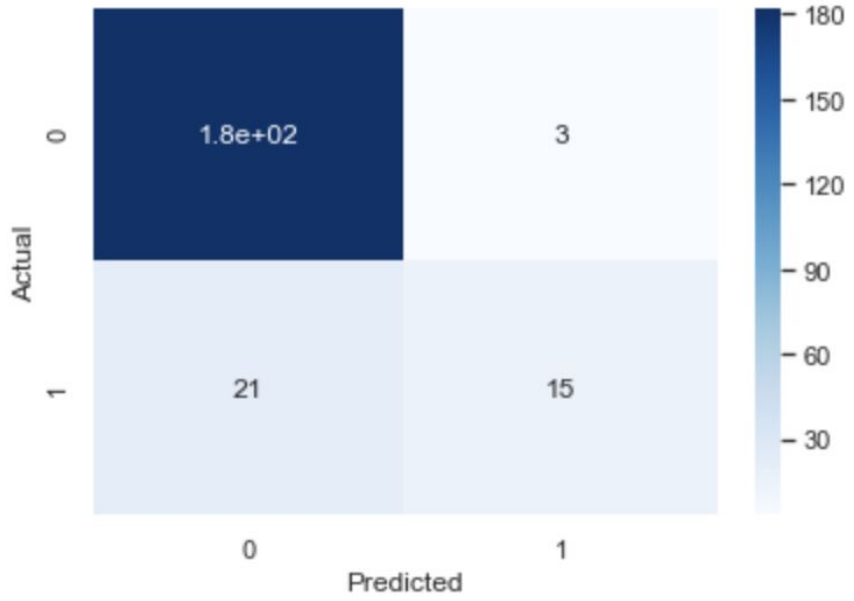
From the algorithm comparison, it looks like the Linear Discriminant Analysis (0.879 and 0.031583 std), Hist Gradient Boosting Classifier (0.861452 and 0.024033 std), and XGB Classifier(0.861458 and 0.022599 std) should be studied more.

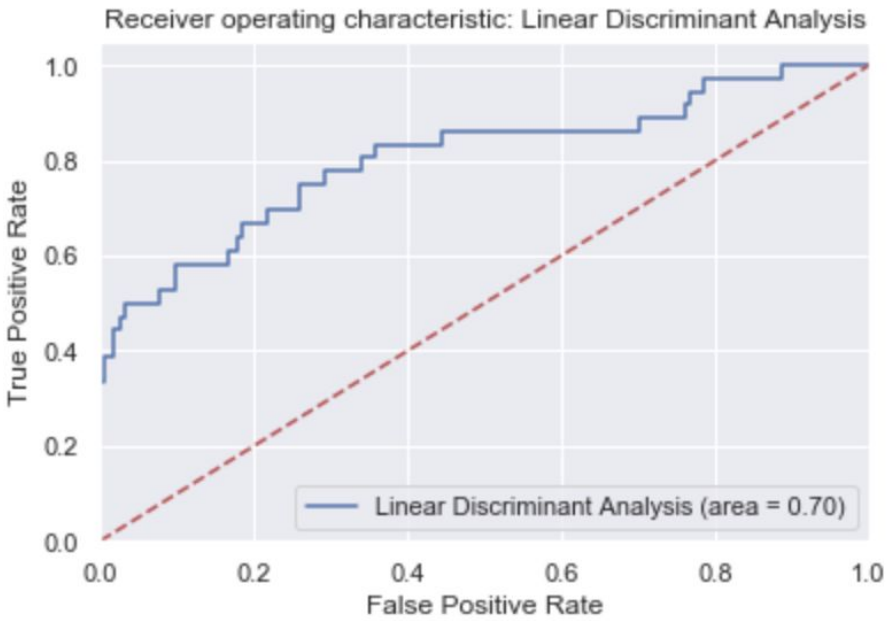
The Hist Gradient Boosting Classifier and the XGB Classifier have less variance so that might mean that they are more confident in their predictions.

I don't believe the feature engineering helped improve the models. Some models did show very slight improvement(Gboo, HGB, and Random Forest), while CART performed worse. However, I still think that the variables created are important when considering employee attrition.

Further analysis on the LDA, Hist Gradient Boost and XGB Classifiers

Confusion Matrix and ROC curve for Linear Discriminant Analysis

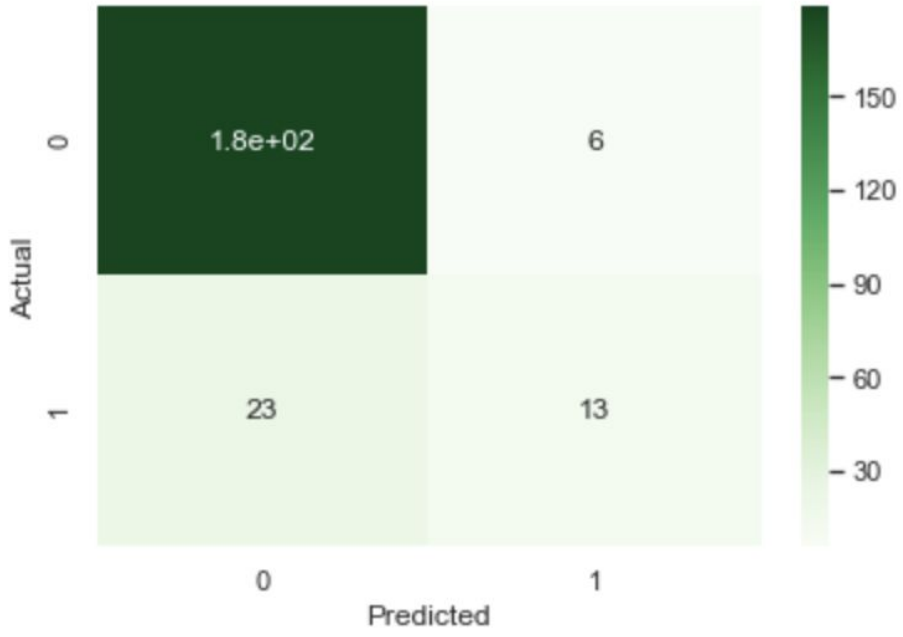


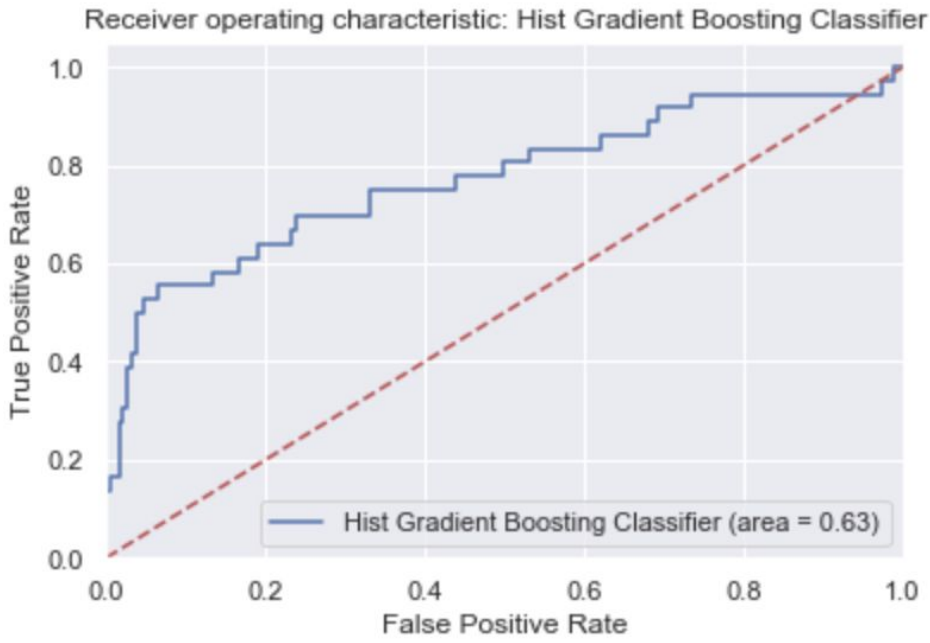


The confusion matrix for the Linear Discriminant Analysis shows that 15 of the actual ones were predicted correctly while 182 of the actual zeros were predicted correctly.

The roc curve for the LDA model shows an area of .70.

Confusion Matrix and ROC curve for Hist Gradient Boosting Classifier

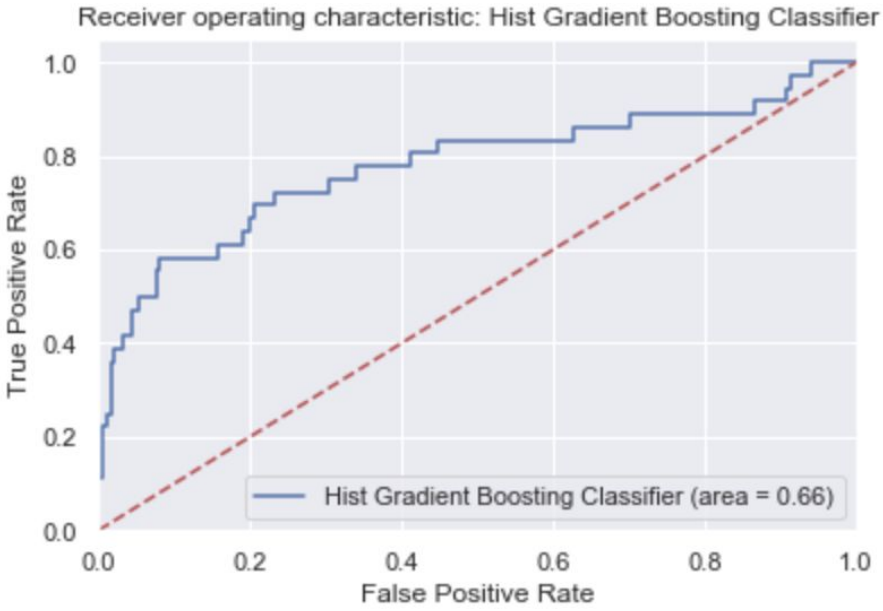




The confusion matrix for the Hist Gradient Boosting Classifier shows that 13 of the ones were predicted correctly and 182 of the zeros were predicted correctly. So far the LDA model did a better job of catching the ones.

The roc curve for the Hist Gradient Boosting Classifier shows an area of .66 which is not as good as the LDA model which has an area of .70.

Confusion Matrix and ROC curve for XGB Classifier

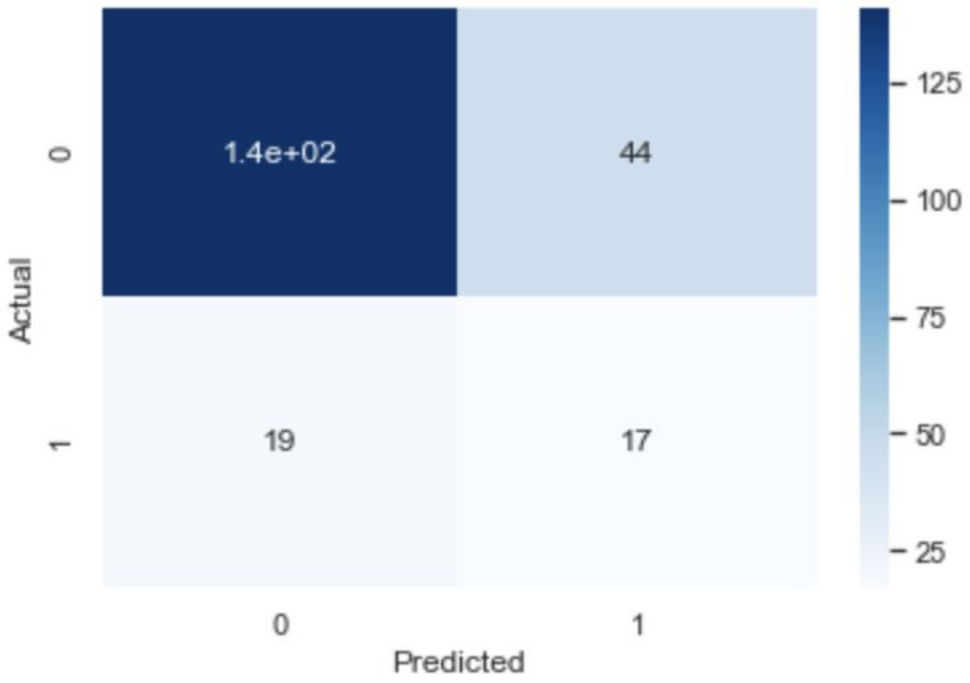


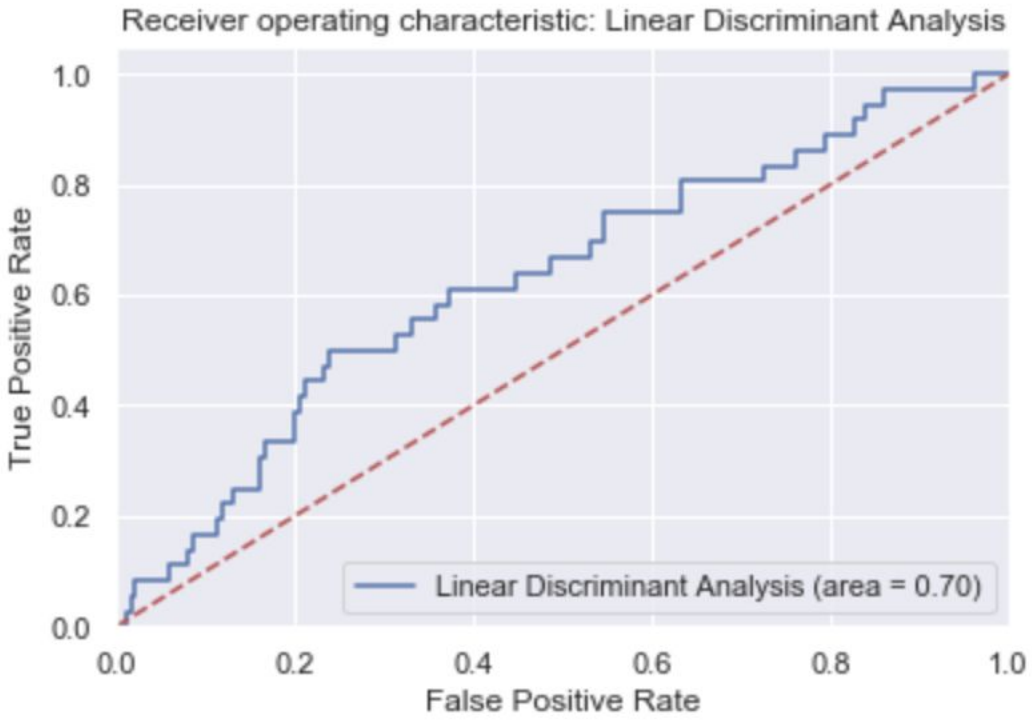
The confusion matrix for the XGB Classifier shows 13 of the ones were predicted correctly with 182 of the zeros predicted correctly. The XGB Classifier performed similarly to the Hist Boost Classifier but did not perform better than the LDA model.

The ROC curve for the XGB model shows that the area under the curve is .67.
According to the ROC curves and confusion matrix of the three classifiers, Linear Discriminant Analysis performed the best. The area covered under the ROC curve was 0.70, the accuracy score was 0.879 which means that 87% of the time we can accurately predict attrition based on the given variables in the dataset.

This means that we can use the LDA model to try and accurately predict attrition. Next, we will try and optimize the LDA model in order to better predict attrition.

Parameter Tuning on LDA


1.4e+02
44
19
17
0
1
Actual
0
1
Predicted
125
100
75
50
25


Receiver operating characteristic: Linear Discriminant Analysis
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Linear Discriminant Analysis (area = 0.70)

Model Selection¶

The roc curve shows that the model with the grid search actually performed the same as the LDA model without any parameter tuning. The LDA model with parameter tuning did a better job of capturing the 1s.

This model is better for our business case since the number of ones captures the number of people that actually left IBM. We want to make sure we can capture the most amount of 1s in order to accurately predict attrition. This means that the variables in the data set are useful to look at when determining if an employee will leave IBM, some of these variables are Age, Job Level, Monthly Income, and Total Working Years.