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

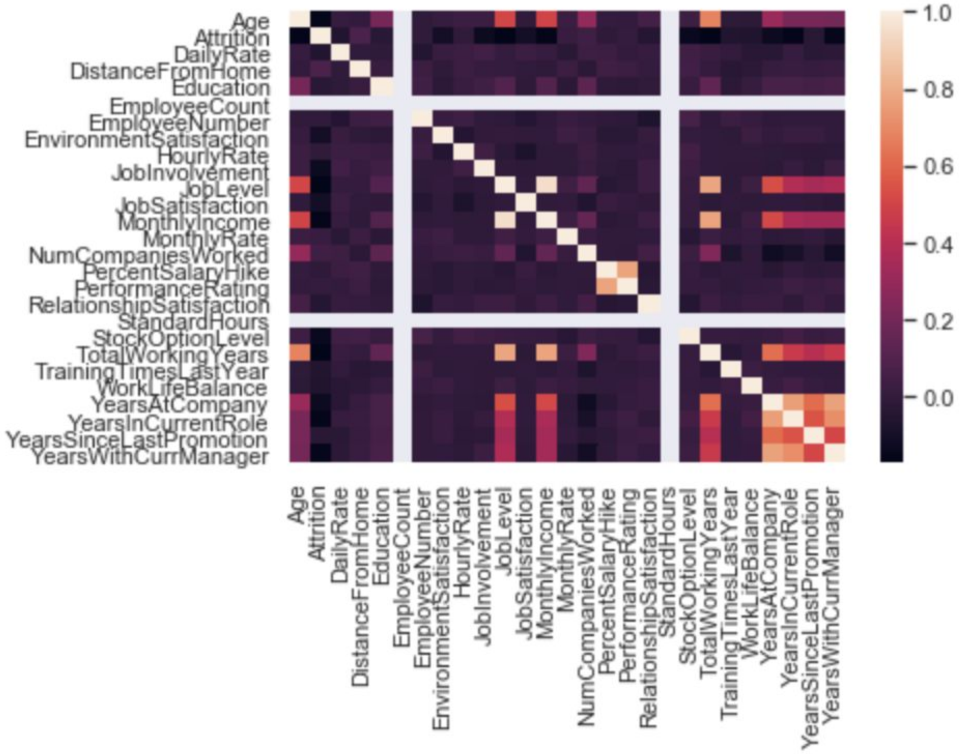
The dataset used in this analysis is IBM HR Analytics Employee Attrition & Performance. The dataset is available to download from Kaggle. This dataset contains 35 variables that may contribute to employee attrition including, work-life balance, years at the company, age and relationship status. The dataset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

Data Cleaning & Pre-Processing

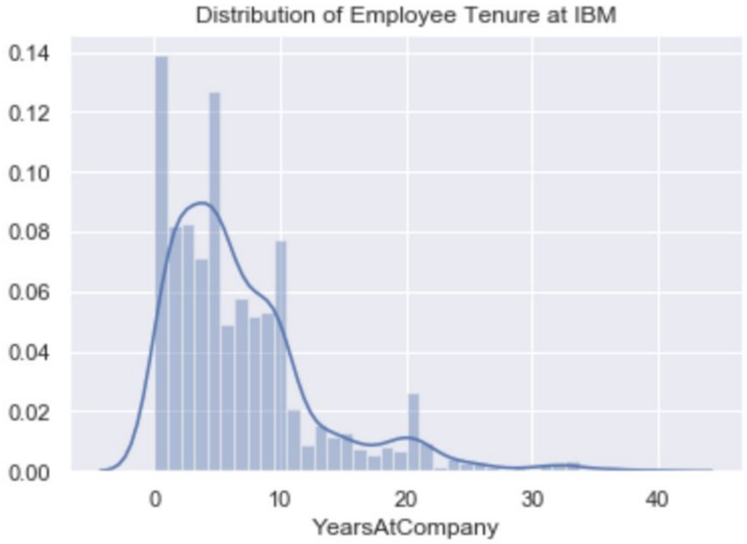
The data set was already extremely clean with no missing values. Converting the target variable (Attrition) into a binary variable (0/1).

Exploratory Data Analysis:

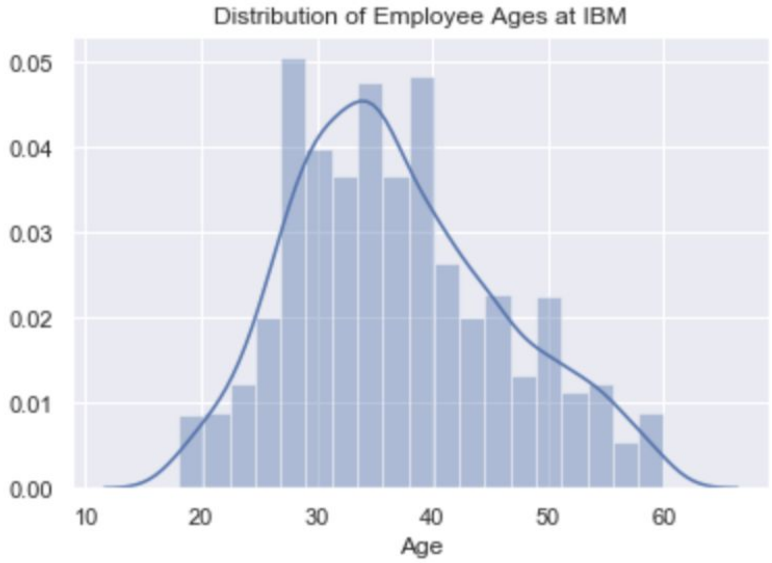
For the exploratory data analysis, I created a correlation plot to see if any of the variables were correlated with each other. I found that very few variables were correlated. This correlation plot shows that years at the company, years in the current role, years since last promotion and years with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



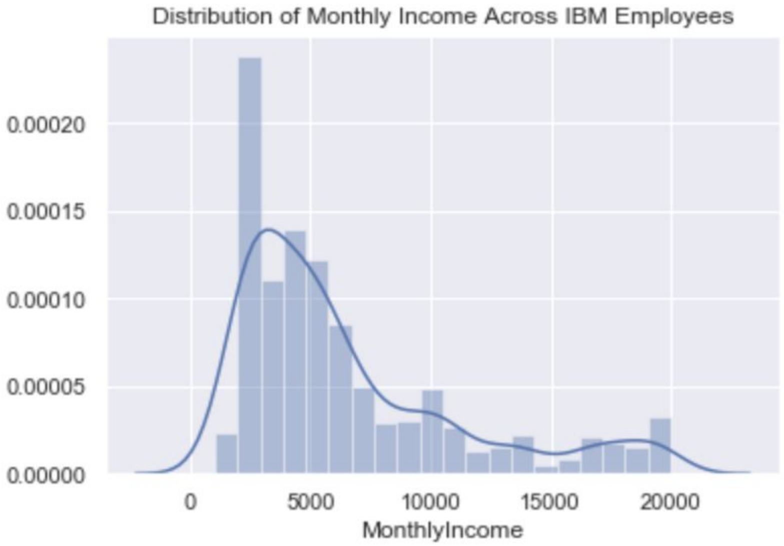
I then plotted the distribution of variables that I thought would be significant in predicting attrition such as age, employee tenure, and monthly income.



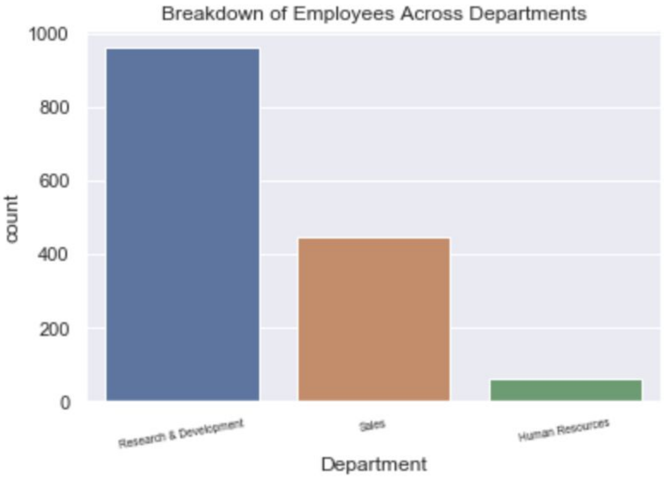
This graph indicates that the average time spent at IBM is approximately 3 years.



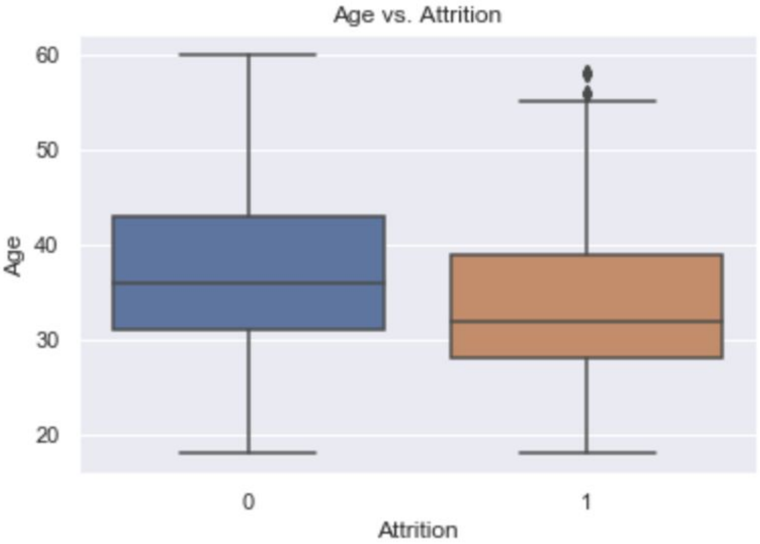
This graph indicates that the average age of Employees at IBM is approximately 35 and is slightly right-skewed indicating that the majority of employees at IBM are young.



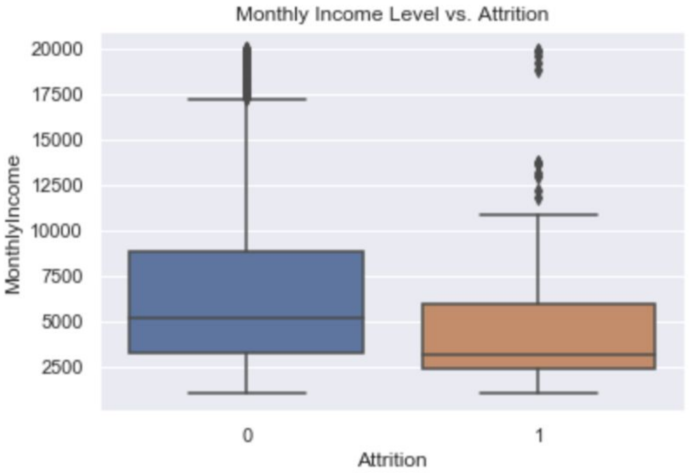
This graph indicates that the average monthly income is less than 5000 and the graph is right-skewed indicating that the majority of the employees make 5000 or less.



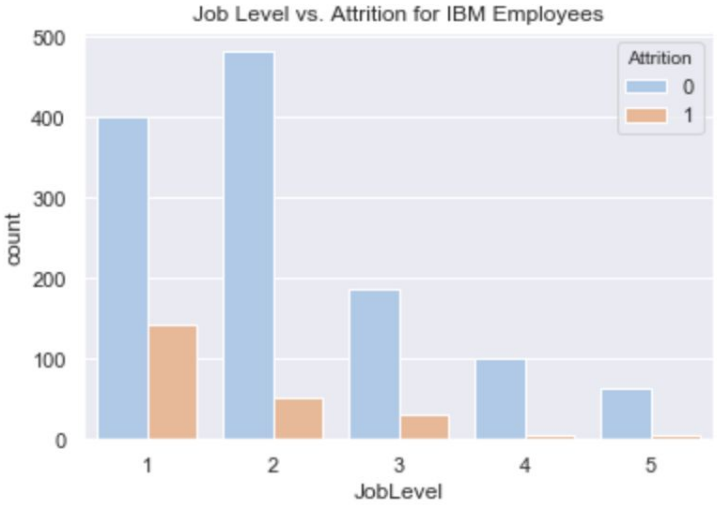
Around 900 employees work in Research & Development and approximately 400 employees work in Sales with around 70 employees working in Human Resources.



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.

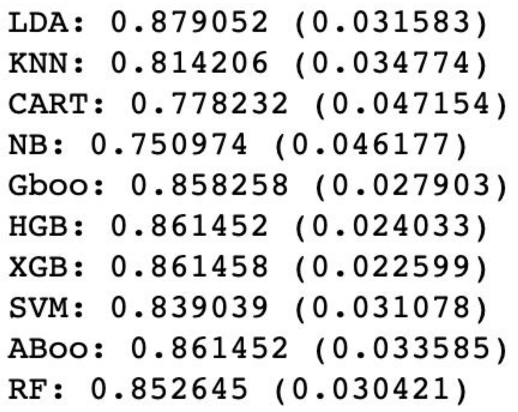


This graph shows that entry-level employees are more likely to leave the company than employees at a more senior level.

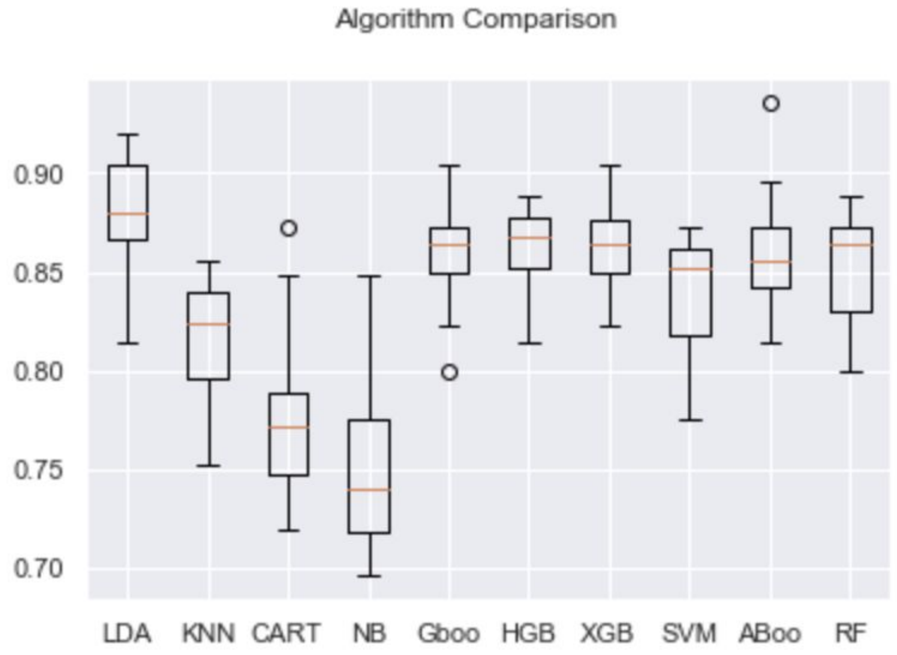
Feature Engineering:

I then continued my analysis by creating some variables that I think would be important when considering attrition. These variables include far_distance which is for employees with a long distance from the place of work, young_no_love which is for employees that are still very young and have no relationship commitments and migrating_worker which is for employees that move around companies often.

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)
```

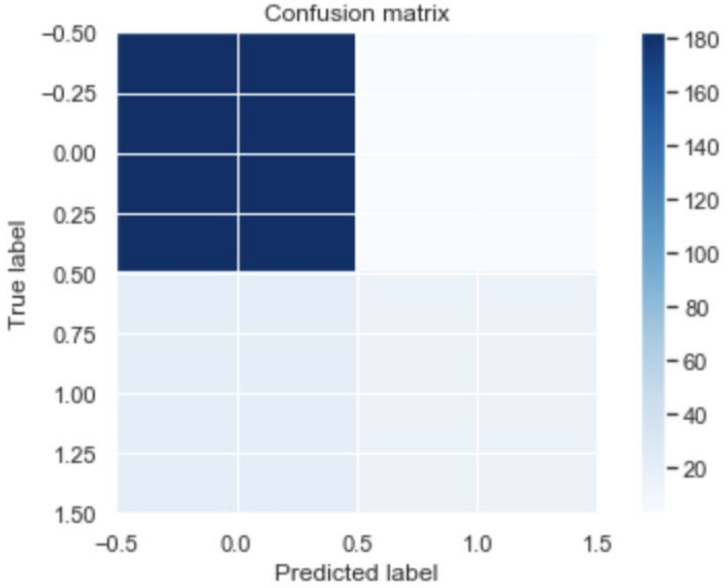


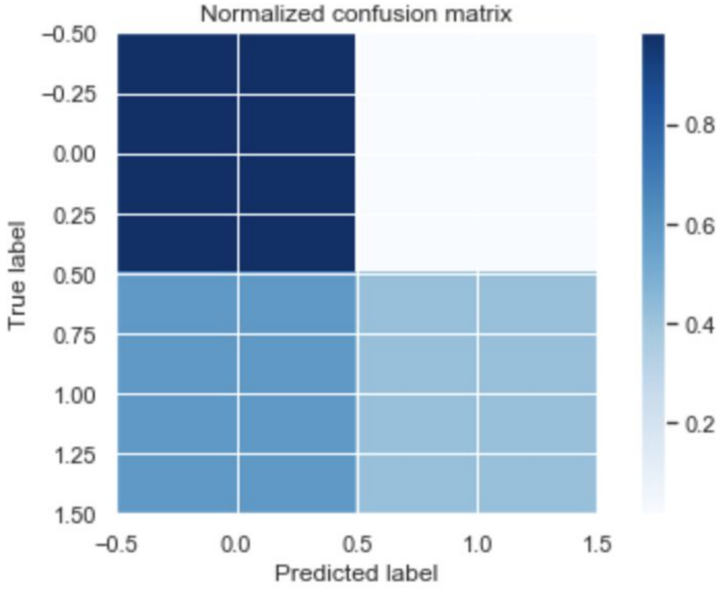
From the algorithm comparison, it looks like the Linear Discriminant Analysis (0.879 and 0.031583 std), Hist Gradient Boosting Classifier (0.861452 and 0.024033 std), and XGB Classifier(0.861458 and 0.022599 std) should be studied more.

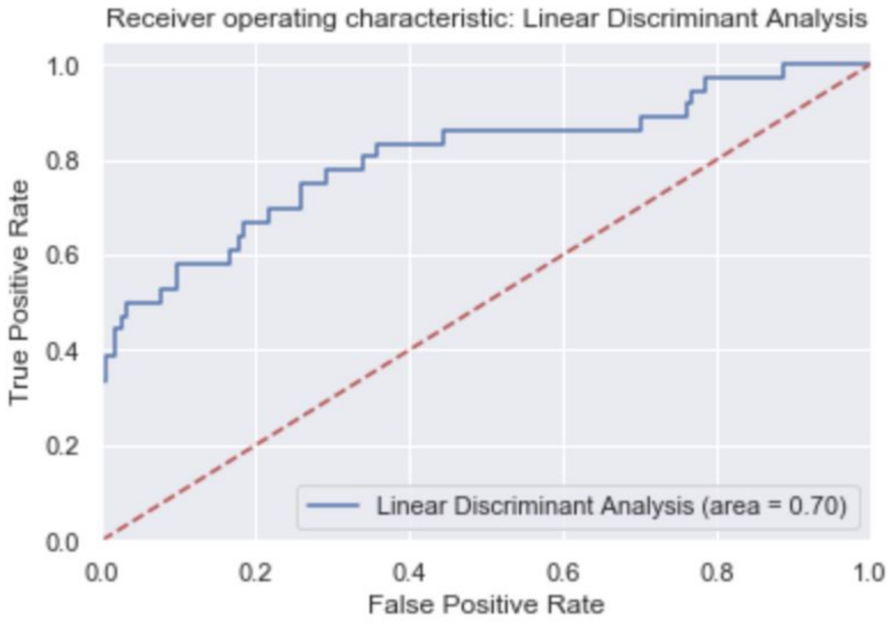
The Hist Gradient Boosting Classifier and the XGB Classifier have less variance so that might mean that they are more confident in their predictions.

Confusion Matrix and ROC curve for Linear Discriminant Analysis

```
Confusion matrix, without normalization
[[182   3]
 [ 21  15]]
Normalized confusion matrix
[[0.98 0.02]
 [0.58 0.42]]
```



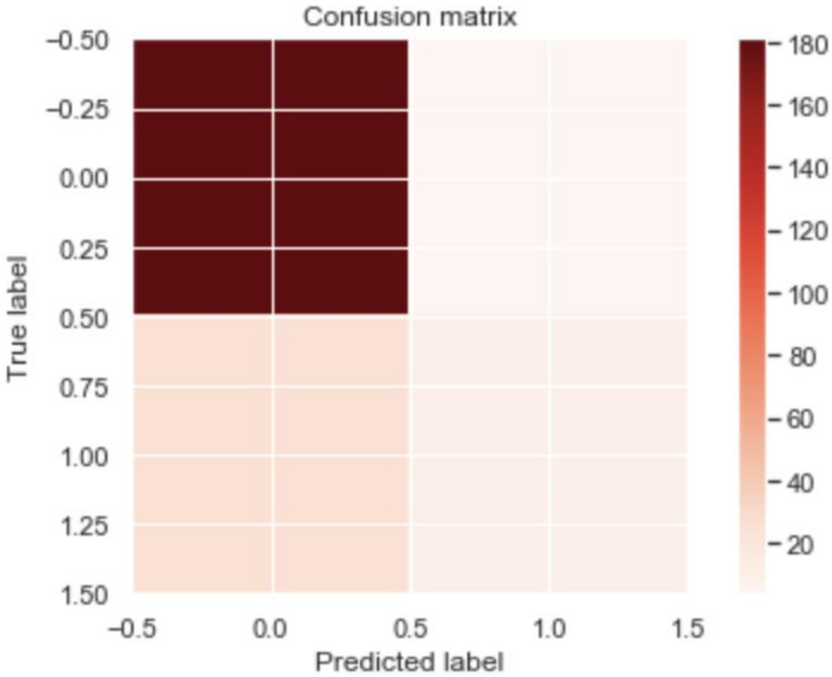


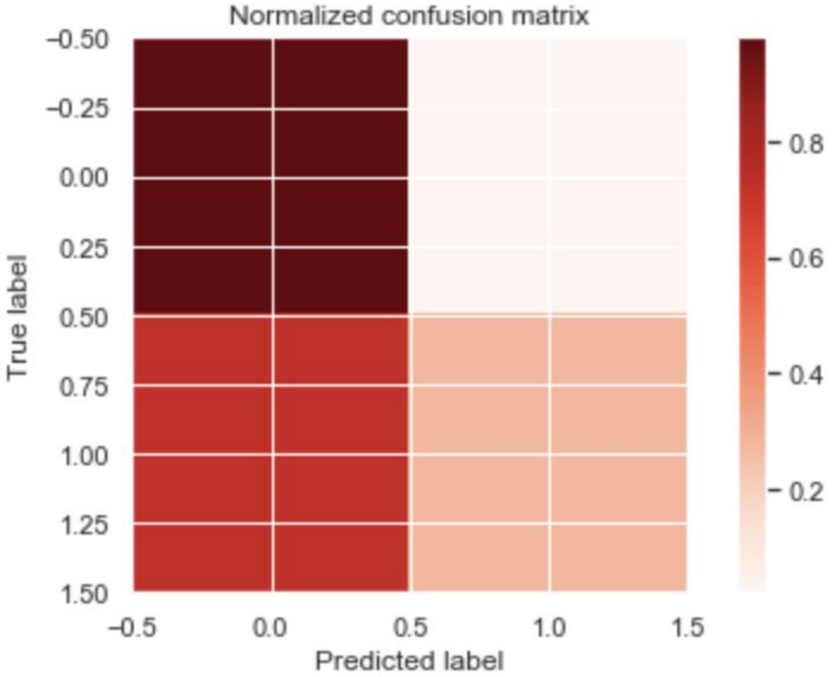


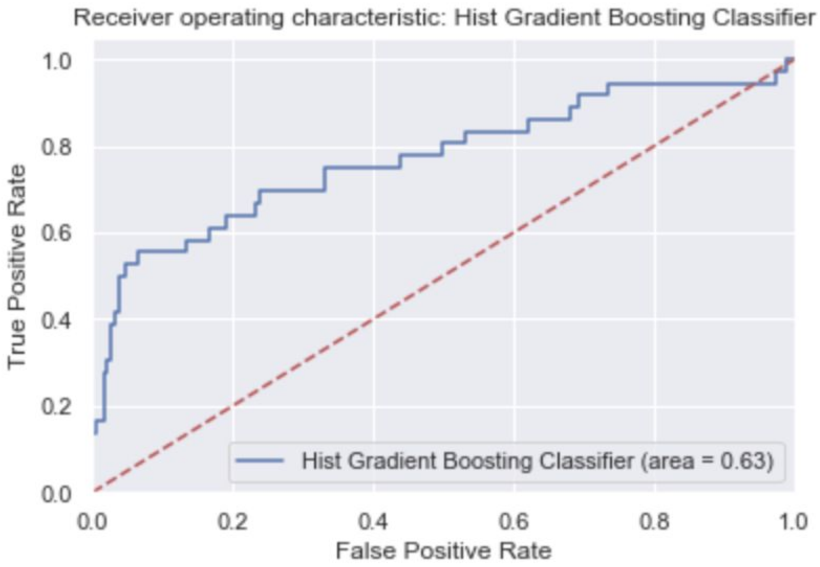
The confusion matrix for the Linear Discriminant Analysis shows that 98 percent of all actual zeros were predicted correctly. Most of the data were predicted accurately with 24 data points predicted incorrectly. The ROC curve shows an area of .7.

Confusion Matrix and ROC curve for Hist Gradient Boosting Classifier

```
Confusion matrix, without normalization
[[181   4]
 [ 26  10]]
Normalized confusion matrix
[[0.98 0.02]
 [0.72 0.28]]
```



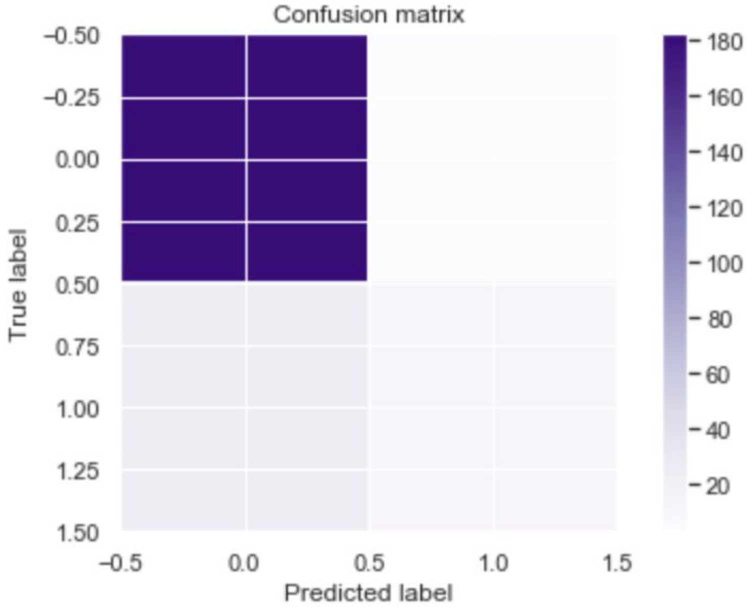


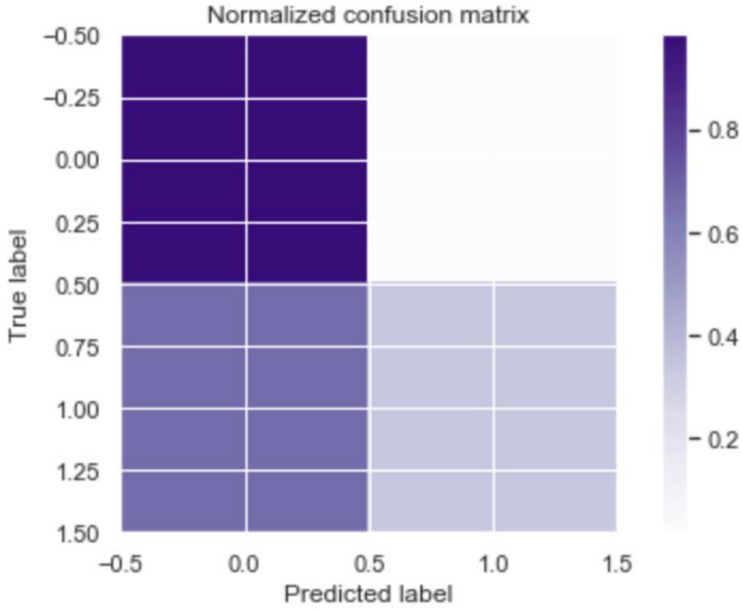


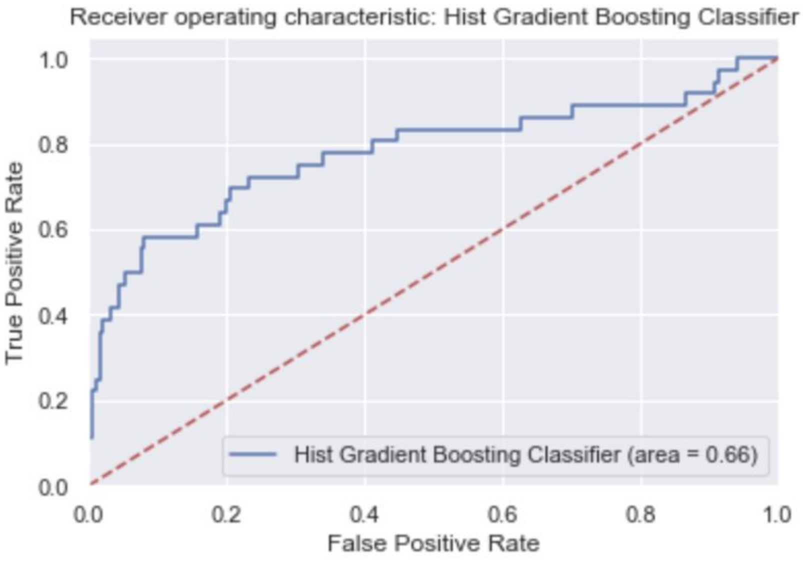
The confusion matrix for the Hist Gradient Boosting Classifier shows that 98 percent of all actual zeros were predicted correctly. There are 30 data points that were predicted incorrectly. The ROC curve shows an area of 0.63.

Confusion Matrix and ROC curve for XGB Classifier

```
Confusion matrix, without normalization
[[182   3]
 [ 24  12]]
Normalized confusion matrix
[[0.98 0.02]
 [0.67 0.33]]
```





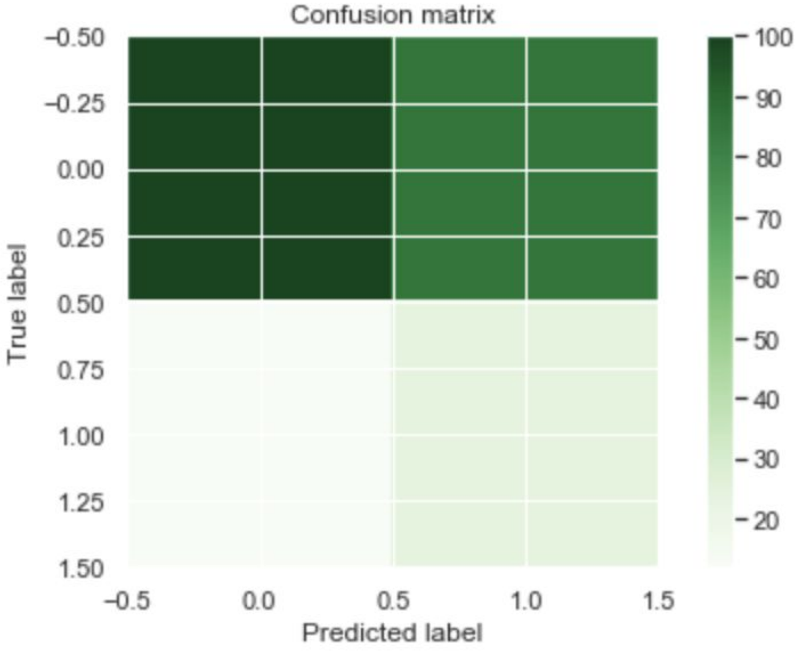


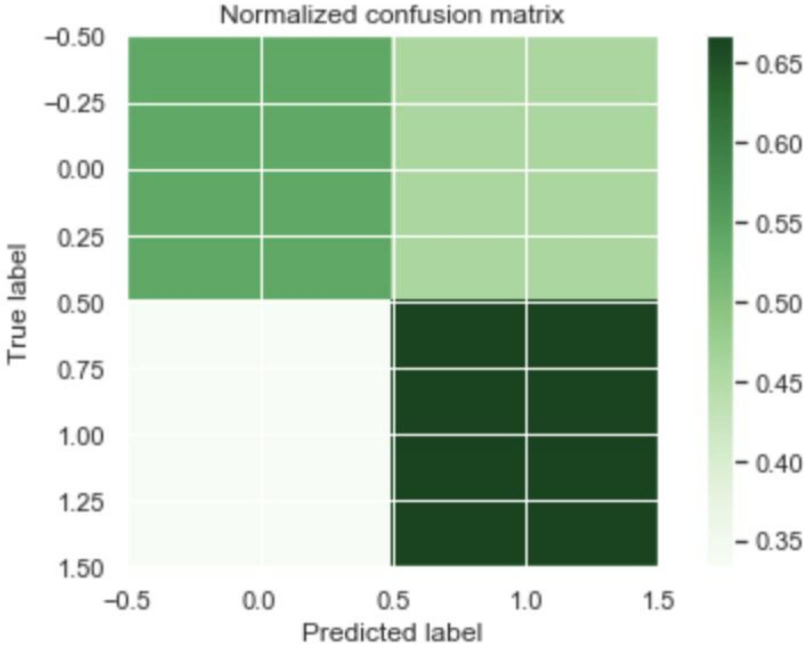
According to the ROC curves and confusion matrix of the three classifiers, Linear Discriminant Analysis performed the best. The area covered under the ROC curve was 0.70, the accuracy score was 0.879 and the confusion matrix showed that 98 % of all actual zeros were predicted correctly.

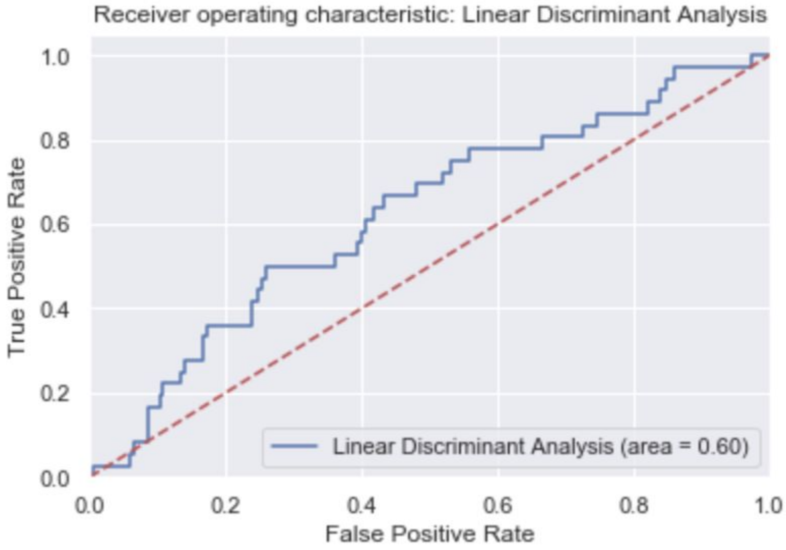
This means that we can use the LDA model to try and accurately predict attrition.

Parameter Tuning on LDA

```
Confusion matrix, without normalization
[[100  85]
 [ 12  24]]
Normalized confusion matrix
[[0.54 0.46]
 [0.33 0.67]]
```







After using Grid Search on the LDA model, the confusion matrix an ROC curve for the new model showed that it actually performed worse than the original LDA model without any parameter tuning.

Model Selection

The LDA model without any parameter tuning is currently the best model in predicting attrition.