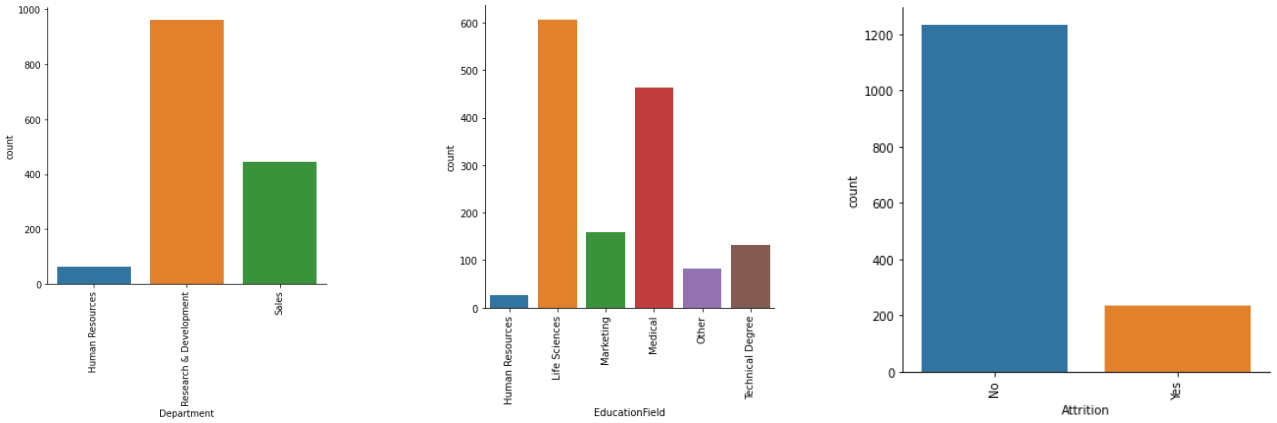
Objective & Abstract

I perform data analysis to discover characteristics of IBM employees and actionable insights in reducing employee attrition. Following exploratory data analysis where the distributions and correlations between both categorical and quantitative data are discussed, I construct two classifiers to predict employee attrition given the characteristics of an employee. The first classifier is a logistic regression and the second is a decision tree based classifier. I then interpret the results of both classifiers to discover insights and avenues for future analysis.

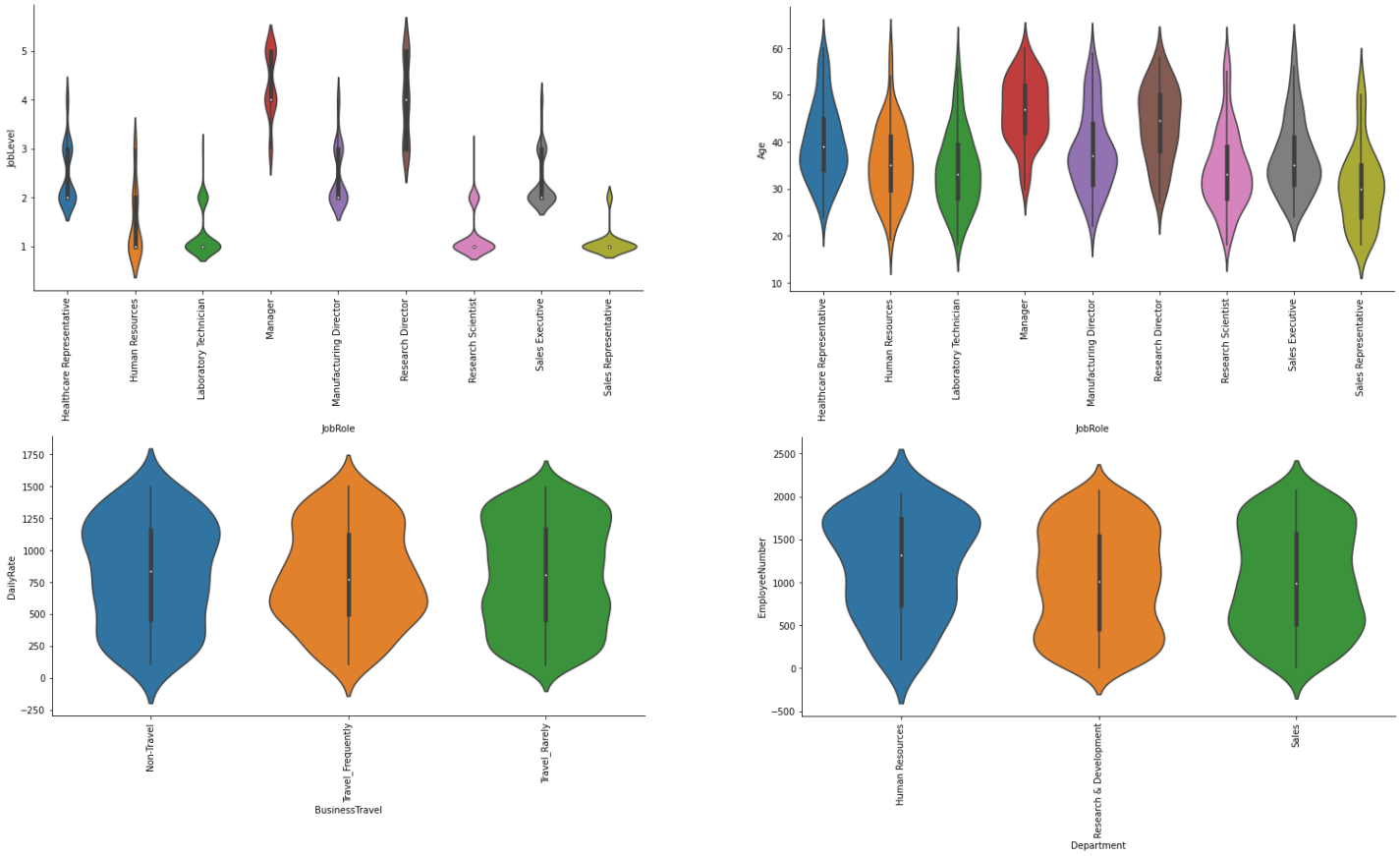
Exploratory Data Analysis: Who Are We Examining?

Starting with the categorical data, several interesting findings as observed in comparing the number of occurrences of different categories are discussed in an admittedly noncomprehensive fashion. As a rule of thumb, the more unbalanced the heights of the bars in the bar graph of a categorical variable, the more likely it is to be important to consider the effect of in any following analysis.

We will first discuss the following bar graphs from left to right. The first bar graph shows the majority of employees are in research and development. In line with this, the next graph shows the majority of employees have a background in the medical and life sciences. It may be important to consider the differences in the work done by those in R&D vs those in Sales or HR, whether as a predictor or as an explanation for attrition. The same goes for educational backgrounds: the job market for different backgrounds could be substantially different, leading to different attrition rates. The second bar graph illustrates that the employees who fall victim to attrition are in the minority, meaning that we may have to employ methods to balance the data classes before training classifiers on the data to improve classifier accuracy.



Next, we will discuss interesting findings when viewing some violin graphs. The distribution of some quantitative data drastically changes when locking the value of some categorical data columns. We will first examine some violin graphs to explore the data set we are working with.



Let us first examine the first row of graphs. The top left figure depicts violin graphs for each job role over job level. While job level probably should be treated as a categorical variable, job levels still possess an order and hence it still makes some sense to examine the data like this. For violin graphs that contain job level values on the lower side, the skew is negative, while for violin graphs that contain higher job level values/are titles typically associated more with seniority, the skew is positive. This is especially true for the graph of Managers and Research Directors, which makes sense: managers need to be above others to manage them, and directors need to be above other scientists to direct them. The span of different roles also indicators the priorities and sense of value placed on different roles within the firm: healthcare representatives tend to be valued more than sales people, for example. People may also be hopping roles: a sales executive seems like the logical next step up from a sales associate, which may be why the two distributions are almost linearly separable at a job role value of 2. In line with the top left graph are the findings in the top right graph, which plots job role against age. Here, we find the job roles that tended to have higher job levels in the top left graph are positively skewed, indicating that senior jobs tend to be occupied by those who are senior in age.

Let us now consider the bottom row of graphs. The bottom left figure depicts the daily rate of pay against how often the employee travels. Interesting to note is that the distribution of pay for employees who never travel is negatively skewed, whereas the distribution for employees that often travel is positively skewed. This indicates that the more an employee travels, the more likely they are to be paid less. This could be explained by higher level executives needing to stay in headquarters to direct people, though further investigation may be required. Finally, assuming that employee numbers are assigned over time, with lower numbers having joined the company earlier, we see that IBM has undergone a hiring surge of HR employees. Originally, I planned to remove employee numbers as a feature from the training data for my classifiers, but seeing as employee numbers may actually be a proxy for join date, I decided to leave it in the data set.

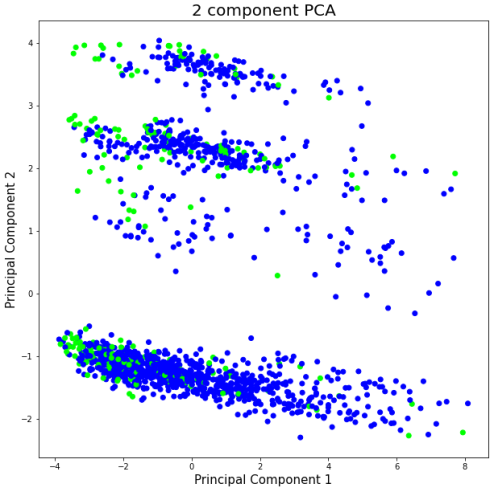
Methodology: Unpacking My Decisions

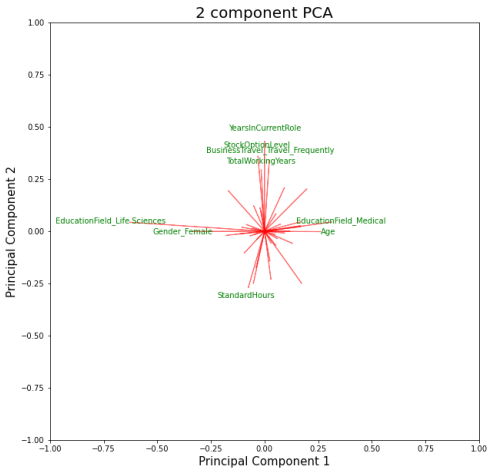
As explained before, I plan to train two classifiers: a Logistic Regression with L2 Normalization and a Random Forest Classifier. Logistic Regressions generalize well across most classification problems and hence provide a good baseline to base further inquiry upon, whereas a Random Forest Classifier can make decisions based on categorical variables more directly at decision nodes (a Logistic Regression assumes numeric input and hence may overfit our data by finding information in the ordering of our categorical variable encoding). Hence, in training a RF classifier, I hope to see whether categorical data should be treated differently in further investigation. Additionally, I perform PCA on the input data to examine if the data is linearly separable and examine the loadings of the two principal components to determine which features may be most predictive of attrition alongside the weights of the logistic regression classifier.

In terms of cleaning the data, any feature with non-numeric values has been interpreted as a categorical variable and encoded using a 1-hot encoding. An exception is the category “Over18”, which only takes one value and hence provides no information about our samples. Examining the data for empty values during EDA, I find there are no missing values, and hence I do not perform any imputation method. I split the training data 70/30, using 70 percent of the data as our training set, and 30 percent as our testing set. I split negative and positive examples separately to ensure that both the test and training set have the same proportion of both classes. Since the number of employees who have left the organization is much smaller than the number of employees who have not, any classifier I naively train on the classifier is likely to have high accuracy and low accuracy/recall. Therefore, it is necessary to balance the dataset, so I choose to do so by weighting the effect samples which display attrition have on models during training 5 to 1, as the majority class outnumbers the minority class about 5 to 1. In doing this, I hope to improve the precision and recall of our model, which would otherwise be underexposed to samples which committed attrition during training.

Principal Component Analysis

I first run PCA on the full dataset and plot the points in the principal component space, where projections of the first and second largest components are represented on the x-axis and y-axis respectively. I color the points by y-classification, where blue points represent employees that have stayed, and green points represent employees that have left the company. Furthermore, I examine the factor loadings of the first and second principal components and label features that have a loading magnitude over a specified threshold (0.3). At a lower dimensionality, the two classes are not linearly separable, though there is some higher level organization (the three streaks) in the data that merits further investigation.





Logistic Regression

To explore the effect of regression penalties on the results, we train four different classifiers with no, l2, l1 and elastic net regularization, the results of which are displayed in the figures below:

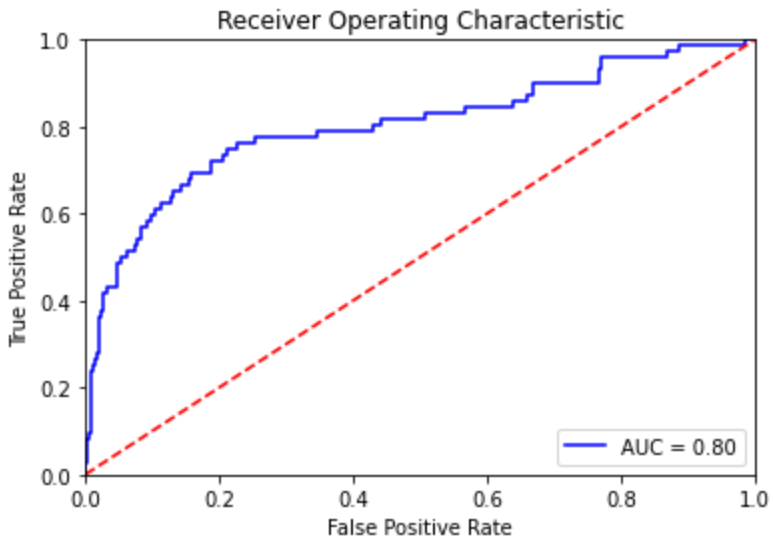
Evaluation on Training Set:

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression (L2 Penalty)	0.83	0.48	0.67	0.56	0.82
Logistic Regression (L1 Penalty)	0.84	0	0	0	0.5
Logistic Regression (Elastic Penalty, l1_ratios = [0, 0.25, .5, .75, 1])	0.16	0.16	1	0.28	0.5

Evaluation on Test Set:

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression (L2 Penalty)	0.82	0.46	0.67	0.55	0.80
Logistic Regression (L1 Penalty)	0.84	0	0	0	0.5
Logistic Regression (Elastic Penalty, l1_ratios = [0, 0.25, .5, .75, 1])	0.16	0.16	1	0.28	0.5

We examine metrics such as accuracy, precision, recall, as well as combined metrics such as F1 and Area Under Curve (AUC) of the ROC curve. You may note that Regression with the L1 penalty has higher accuracy as well as 0 for precision, recall, etc. I theorize that the Regression with the L1 penalty classifies pretty much every example as “No” for attrition value. A quick sanity check confirms this: (5/(1+6) = .833). The opposite occurs for the Elastic Penalty Logistic Regression: each example is classified as “Yes”. As L2 is the only penalty that does not naively classify all examples as one class or another, we will proceed with further analysis assuming L2 regularization.

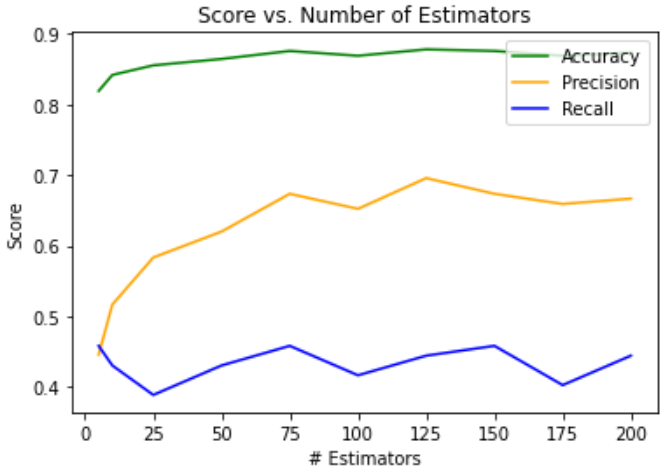


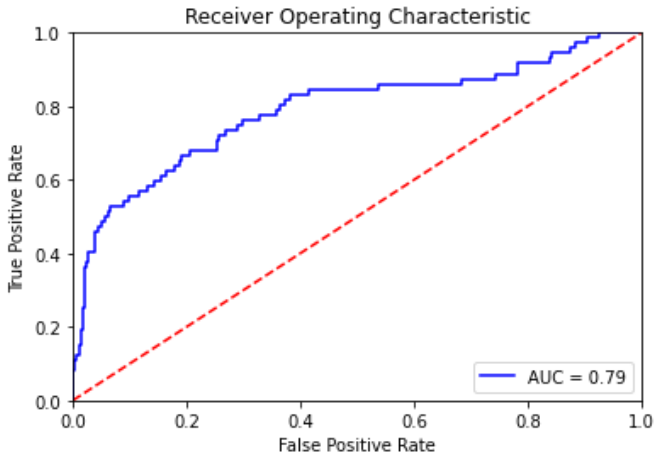
ROC Curve of Logistic Regression with L2 Penalty

Random Forest

To explore the effect of adjusting hyperparameters on model performance, I train Random Forest classifiers with various numbers of estimators including but not limited to 5, 50, 100, 150, 200. The figure below shows that after 75 estimators, the accuracy, precision and recall change only marginally. I conclude that Random Forest performs best on this dataset with 125 estimators. The results of the hyperparameter tuning are displayed below in a table and summarized in a plot.

Estimators:	Precision	Recall	F1	Acc
5	0.45	0.46	0.45	0.82
10	0.52	0.43	0.47	0.84
25	0.58	0.39	0.46	0.86
50	0.62	0.43	0.51	0.86
75	0.67	0.46	0.55	0.88
100	0.65	0.42	0.50	0.86
125	**0.69**	**0.44**	**0.54**	**0.88**
150	0.67	0.46	0.55	0.88
175	0.66	0.40	0.50	0.86
200	0.66	0.44	0.53	0.87



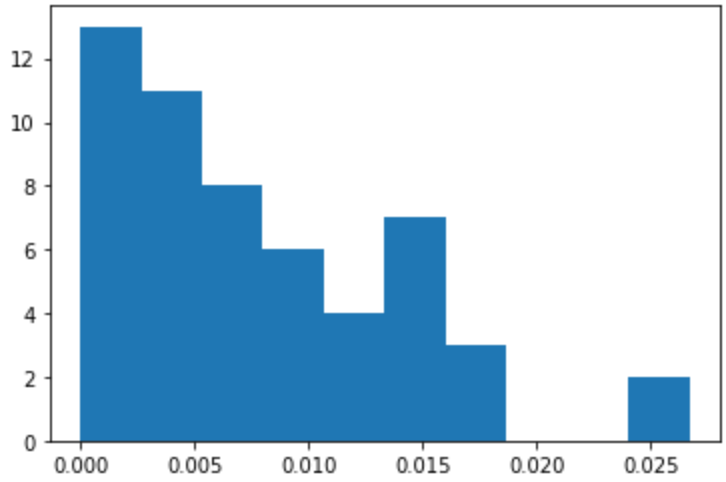


Sources of Error

Some categorical variables with numeric values (like job level) may have been left as quantitative data erroneously. This may have affected the performance of the Logistic Regression model, which may have overfit on the features by deriving information from the magnitude of the category. Additionally, it is unlikely the sklearn's implementation of Random Forest takes advantage of the limited domain of categorical variable values. In future work, more data cleaning and exploratory data analysis to transform the data (especially in investigating the 3 streaks in the PCA plot) should be done. It is likely that some transformations could make the data much more separable.

Insights and Future Work

In terms of model efficacy, the Random Forest approach is preferable to the Logistic Regression. Applying different forms of regularization seems to cause the model to collapse to a naive classifier. Given more time I would have liked to investigate if manipulating hyperparameters could stop the model from producing this erroneous behavior. Meanwhile, the Random Forest classifier was able to improve over the naive accuracy achievable by classifying every example as not likely to leave (~ 0.833) to .88 accuracy. These accuracy gains were made by correctly identifying more examples of employees who left the company, hence increasing both the recall and precision of the Random Forest as compared to the Logistic Regression classifier. Given more time, it may also be interesting to apply other aggregate classifiers such as AdaBoost, Gradient Tree Boosting, or Extreme Tree Boosting, which use more advanced techniques to train their constituent classifiers instead of randomly bucketing data.



The above is a histogram of the magnitudes of different components in the weights of the Logistic Regression. The 2 outliers on this histogram are “OverTime_Yes” and “OverTime_No” with a value of 0.02668067 and -0.02668067 respectively. I will discuss these features, alongside others, below:

Variable	Magnitude	Explanation and Recommendation
OverTime	0.026	It makes sense that employees that are forced to work overtime to complete their work may leave the firm to seek easier work. HR should investigate further into management practices with regard to scheduling work and deadlines to make sure work is being distributed fairly and in a way that minimizes overtime.
MaritalStatus_Single	0.018	Younger employees early in their career naturally have higher churn as they are not yet established at the company and hence can jump to another company with minimal repercussions. Look into incentives for staying at the company like stock options. Also look into mentorship and growth: younger people usually look for jobs where they can learn quickly.
JobInvolvement	0.016	Employees want to feel engaged and challenged by the work they do. Investigate favoritism and work distribution, and also make sure that hiring practices are tightened so that each employee is taken on only when there is work to do, to make sure there is enough work to go around.
TotalWorkingYears	0.016	Similar explanation to MartalStatus_Single: younger employees move more. Investigate if this variable is as colinear with MartialStatus_Single as the Regression suggests in the future.
StockOptionLevel	0.0159	The value for this is negative, indicating that the higher this value is, the less likely the employee is to leave. Offer employees to be paid more in stock.

In terms of features that should be looked into in the future that were not important for the Logistic Regression/were not related to other features that the Logistic Regression looked at (for example, StandardHours and OverTime) are EducationField and Gender. It may be informative to attempt to build classifiers on, say, only those in the Life_Sciences. Further investigation is recommended.