# The Effect of Drug Abuse on Mental Health

Capstone Report

Raj Singh

# Table of Contents

# Problem Statement

We still don't know how exactly certain drugs are related to development of mental health-related issues. Drugs such as alcohol, marijuana, cocaine, and hallucinogens are frequently used by college students. Use of these substances is sometimes carried into their young adult lives after they graduate. At the same time, mental health disorders (such as anxiety and depression) begin to develop for certain individuals. If we could establish a strong correlation between psychological distress and substance use, we could alert young adults that use of these substances could have deleterious effects on their mental health. This could have applications in the treatment of mental health disorders as well; people may be delighted to hear that they could alleviate their depression simply by quitting their use of drugs. On the other hand, if we discover that use of substances has little correlation with development of these disorders, we could choose to focus our treatment efforts elsewhere.

# The Data

The data comes from the Substance Abuse and Mental Health Services Administration. Specifically, it comes from the 2015 National Survey on Drug Use and Health. Because we are mostly interested in the effect that drugs have on young adults, we will focus on individuals between the ages of 18 and 25. There are several factors that the survey uses to delineate substance use, such as the number of days that the individual has used the substance in the past year. There are two dependent variables that we will focus on in particular: the survey has devised its own sophisticated metrics to indicate whether an individual is psychologically distressed, in addition to whether the individual has had a major depressive episode in the past year. The data comes with a codebook that describes how the data was collected and how certain metrics were calculated; it also clarifies the meaning of all ambiguous numbers within the dataset. Hopefully, this will make data cleaning a relatively simple process.

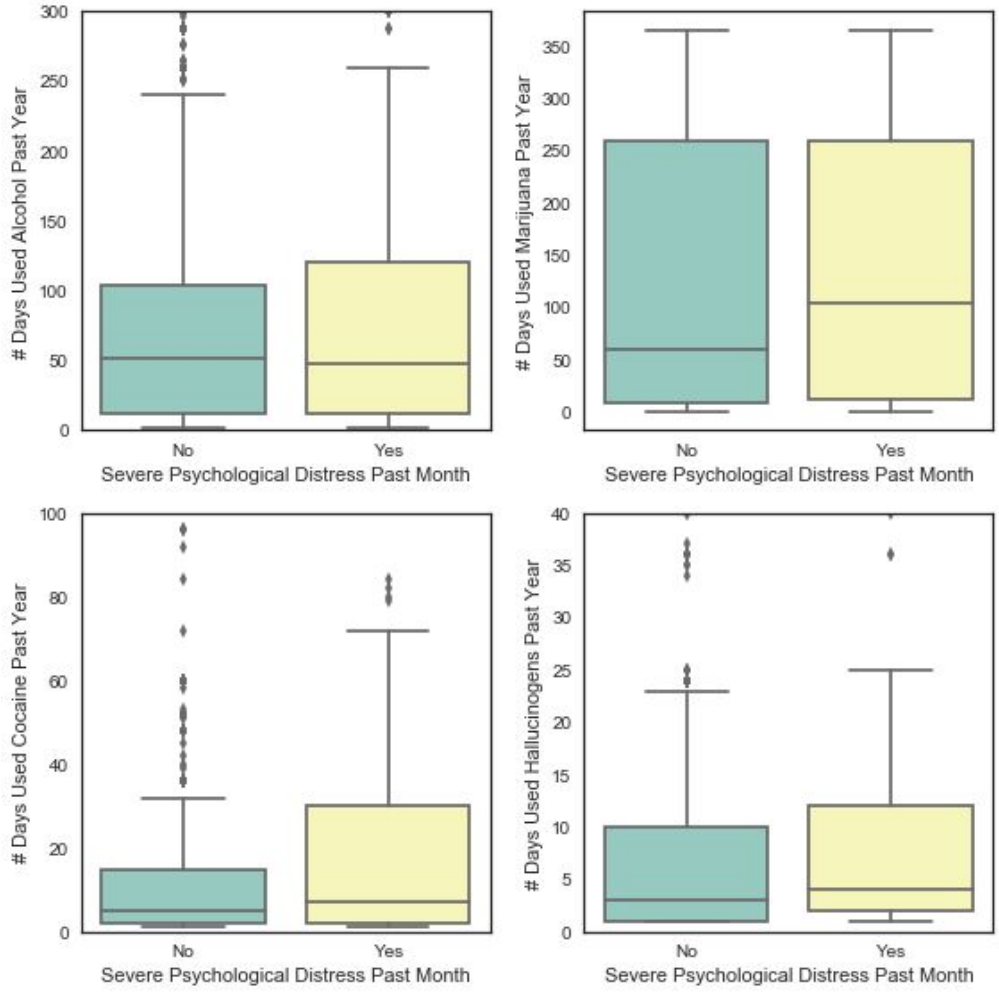
# Data Cleaning

First, we must import the data by using tabs as the delimiter. Then, we will make  a Python dictionary corresponding the column names of the .tsv file that we are interested in with the meaning of the code according to the codebook provided by the Substance Abuse and Mental Health Services Organization.We use the keys of the dictionary to boolean index the DataFrame so we get a new DataFrame with the columns that we are interested in. We then use a renaming function to rename the columns so that we can see what each column name really means. Then, we filter the age to make sure that we are only getting the observations inside the age range of 18-25. What follows was the most time intensive step thus far: we must to replace the values in the DataFrame based on their meaning according to the codebook. We have to do this for all 47 columns. Then, we inspect the numerical data for outliers that would definitely be bad data. All of the outliers that we find make sense in the context of the problem, and so there is no real need to eliminate any of the outliers. Finally, we write the cleaned data to a new .csv file. The details of these cleaning steps can be found here.

# Exploratory Data Analysis

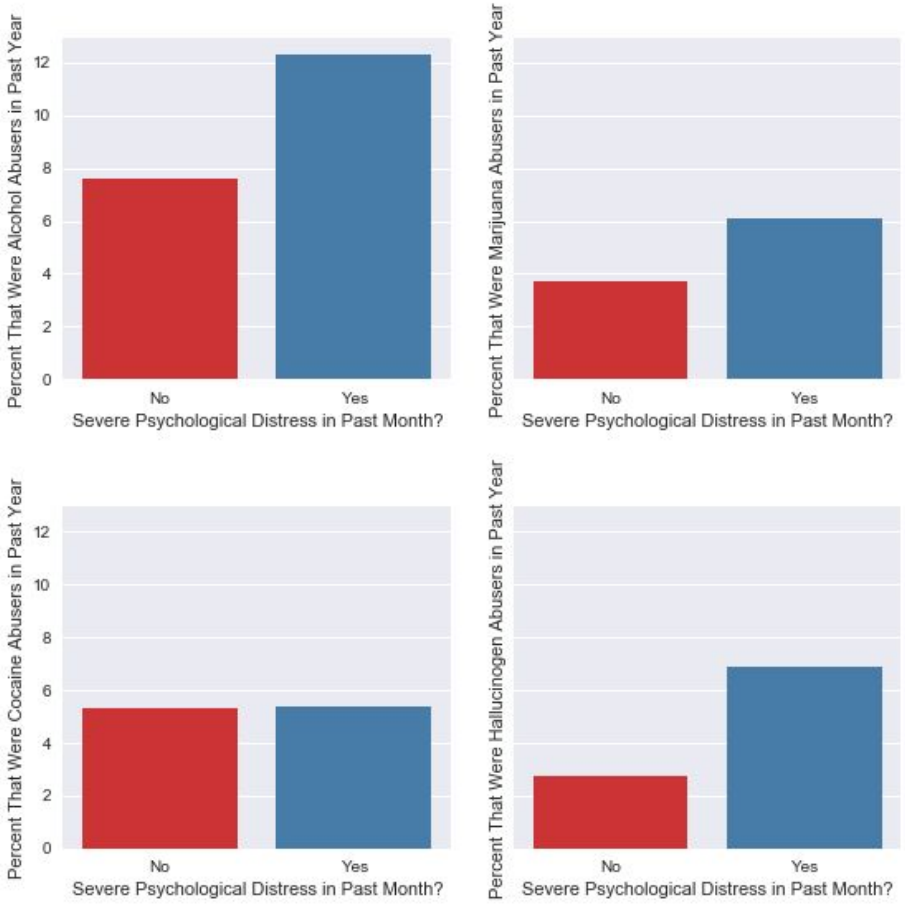
Here, we will create some visualizations to try and investigate any trends and/or patterns that may be useful to look into during our analysis. Keep in mind, our goal is to find out if excessive substance abuse has a correlation with development of mental illnesses and psychological distress. Whether one causes the other is yet to be seen, but we can certainly learn if the two are associated by exploring the data further. Once we establish this correlation, we can further investigate (perhaps using whether or not first use of the drug was at a young age) whether this drug use caused their psychological distress. Then, perhaps, we can use this as a reason to caution adolescents against using drugs. Or, if we find no definitive correlation/contribution, we can choose to focus our efforts to improve worldwide mental health elsewhere.

## Severe Psychological Distress (Past Month) and # of Days that Substance Was Used (Past Year)




As we can see from the above visualization of boxplots, marijuana is the only substance where there was a noticeably higher median '# Days Used in Past year'. Given that we did not find much correlation in the remaining substances, we will choose to move along with our analysis without doing any hypothesis testing. Much of the exploratory analysis on fields involving the number of days that the substance was used in the past month or year yielded insignificant results. Thus, we will choose not to highlight any more of these findings.

# Severe Psychological Distress (Past Month) and Substance Abuse (Past Year)



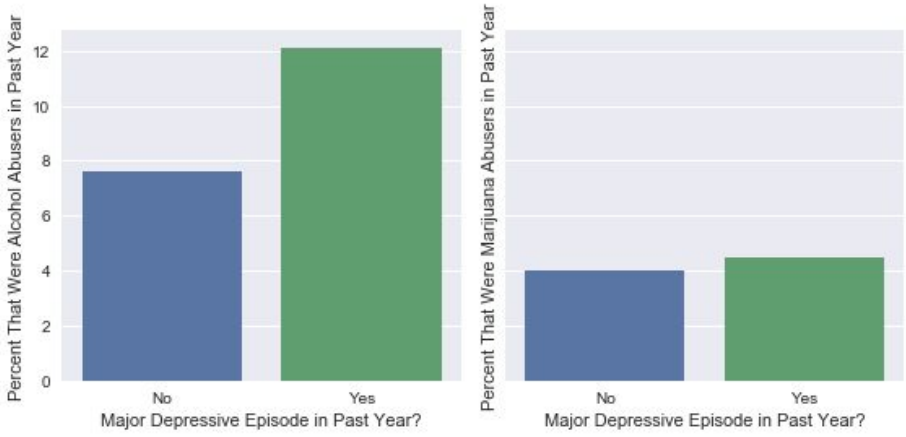

Here, we took the subsets of people with and without severe psychological distress in the past month, and compared the percent of them that were abusers of the substance. The results are shown in the above barplot.

As we can see, there is a higher percentage of abusers for every substance in the subset of people with severe psychological distress in the past month. This seems like a trend worth investigating.

So, we set up two-sample proportion z-tests for each substance. The details of performing these tests can be found here under the heading "Applying Inferential Statistics." In the cases of alcohol and marijuana abuse, we reject the null hypothesis at the $\alpha = 0.01$ level. We can conclude that the proportion of alcohol and marijuana abusers with severe psychological distress in the past month is significantly greater than that of people without severe psychological distress in the past month. With cocaine abuse, we fail to reject the null hypothesis and conclude that there is no significant difference. When performing this test on the proportion of hallucinogen abusers, we get a p-value of 0.016. So, we fail to reject the null hypothesis at the $\alpha = 0.01$ level, but we reject it at the $\alpha = 0.05$ level. There does indeed seem to be a correlation between substance abuse and severe psychological distress.

Since we got clear statistical significance with alcohol and marijuana, let's investigate how abuse of these substances relates to major depressive episodes in the past year.

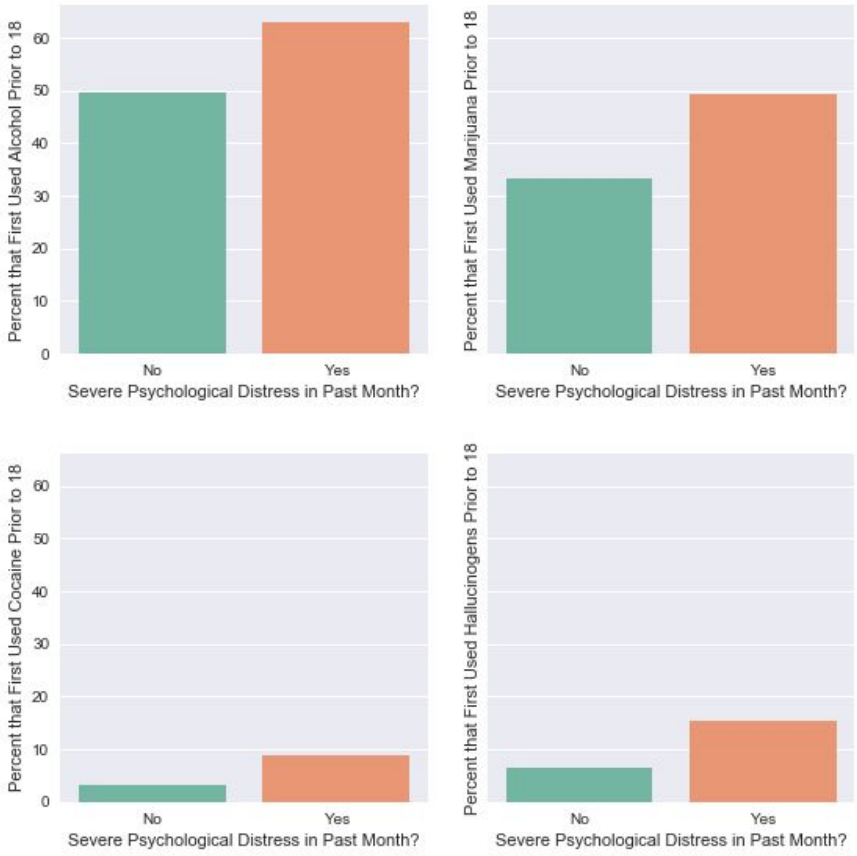
## Major Depressive Episodes (Past Year) and Alcohol/Marijuana Abuse (Past Year)




In both cases, there is a higher percentage of individuals that have abused the substance amongst those who have had a major depressive episode in the past year.

We will take the sample approach that we did earlier and perform two-sample proportion z-tests. With alcohol abuse, we reject the null hypothesis at the $\alpha = 0.01$ level and conclude that there is a significant difference between the proportions of alcohol abusers amongst those with and without MDE's in the past year. In the case of marijuana abuse, however, we fail to reject the null hypothesis. This is an interesting finding because we rejected the null hypothesis with high certainty in regards to those with severe psychological distress. The details of this analysis can be found here under the heading "Alcohol Abuse on Major Depressive Episode." Let's continue to investigate more trends in the data.

## Severe Psychological Distress (Past Month) and Trying Substance Prior to 18




As we can see from the above barplots, there is a higher percentage of individuals that have tried the substance prior to 18 amongst those with severe psychological distress in the past month.

We will use a two-sample proportion z-test again. With every substance, we reject the null hypothesis at the $\alpha = 0.01$ level: thus, We can conclude that the proportion of those who have used each substance prior to 18 amongst those who have had severe psychological distress in the past year is significantly greater than that of individuals that have not had psychological distress in the past month. The details of this analysis can be found here under the heading "Substance Use before 18 - Effect on Psychological Distress."

We are seeing that excessive substance use (as measured by a variety of metrics) does have a strong correlation with severe psychological distress.

# Modeling

We will try three different approaches to classify whether or not someone has severe psychological distress or not: logistic regression, random forest, and boosting.

Before we run any of our models, we need to do some pre-processing on the data because we have a lot of categorical variables in our dataset. So, we change all the categorical labels to numbers. The details of these steps can be found in this notebook.

## Logistic Regression Model

Here, we will use a logistic regression model to classify whether or not someone has had severe psychological distress in the past month. We will try two approaches: one with random-undersampling and one with balanced class weights. These are two methods used to deal with our class imbalance. In both cases, we will use standard scaling as well. While this will not have an impact on our results, it will make our regularization more efficient (part of regularization involves calculating norms, so this will be a less costly calculation if we have data with smaller scales). Then, we will test our parameters by using RandomizedSearchCV (we will use 5-fold cross validation).

**Logistic Regression Model with Random Under-sampling - Training Results**

**Optimized Parameters:**

| | |
|---|---|
| **C** | 0.006 |

**Accuracy Score: 69.6%**

**Classification Report:**

| | **Precision** | **Recall** |
|---|---|---|
| **0** | 0.93 | 0.72 |
| **1** | 0.18 | 0.51 |
| **avg/total** | 0.85 | 0.70 |

**Confusion Matrix:**

| | |
|---|---|
| 7003 | 2767 |
| 556 | 588 |

**AUC Score: 0.653**

We chose the regularization parameter C such that the area under the ROC curve is maximized. We have a low C value, which means that we are doing a great deal of L2 regularization. That is, we are shrinking the coefficients of many of the features. As we can see from the results, our AUC score is not particularly high, but our precision is relatively high. We don't get a poor recall score either. Our accuracy is about 70%, which falls in response to the increased recall rate.

We do not have a great model, but let's try using our model on the test data.

**Logistic Regression Model with Random Under-sampling - Testing Results**

**Accuracy Score: 68.3%**

**Classification Report:**

| | **Precision** | **Recall** |
|---|---|---|
| **0** | 0.92 | 0.71 |
| **1** | 0.17 | 0.50 |
| **avg/total** | 0.84 | 0.68 |

**Confusion Matrix:**

| | |
|---|---|
| 2283 | 954 |
| 200 | 202 |

**AUC Score: 0.640**

We only get slightly worse results on our test data than we did our training data. We may not have a highly accurate model, but we certainly get one that does not overfit. The details of implementing this model can be found here.

Now, let's look at how our logistic regression model does with balanced class weights.

**Logistic Regression with Balanced Class Weights - Training Results**

**Optimized Parameters:**

| | |
|---|---|
| **C** | 0.052 |

**Accuracy Score: 70.4%**

**Classification Report:**

| | Precision | Recall |
|---|---|---|
| **0** | 0.92 | 0.73 |
| **1** | 0.19 | 0.51 |
| **avg/total** | 0.84 | 0.70 |

**Confusion Matrix:**

| | |
|---|---|
| 7077 | 2654 |
| 577 | 606 |

**AUC Score: 0.653**

Our results are very similar to that of our model with random under-sampling. The only main difference is that we achieve the results with less regularization. Let's see our results on the test data.

**Logistic Regression with Balanced Class Weights - Testing Results**

**Accuracy Score: 70.0%**

**Classification Report:**

| | Precision | Recall |
|---|---|---|
| **0** | 0.93 | 0.72 |
| **1** | 0.16 | 0.48 |
| **avg/total** | 0.85 | 0.70 |

**Confusion Matrix:**

| | |
|---|---|
| 2369 | 907 |
| 188 | 175 |

**AUC Score: 0.647**

Across, the board, we get very similar average precision and recall scores. Random under-sampling and balanced class weights yielded nearly identical results: we may want to choose balanced class weights because it required less regularization. The details of implementing this model can be found here.

## Random Forest Model

Now, we will implement a random forest model to help solve our classification problem. A big advantage of using a random forest model is that we can view how important our features compared to others. Moreover, we can prevent overfitting by adding more trees to our forest: it is crucial that we do not overfit because we are dealing with a very delicate issue and we can not afford to have unexpected misclassifications.

Similar to the logistic regression model, we will use 5-fold cross-validation with RandomizedSearchCV. Like our previous model, we will also try out both random under-sampling and balanced trees. Here are our hyperparameters we are testing out for our random forest model:

| **Hyperparameter** | **Values Tested** |
|---|---|
| Number of Trees | 100, 250, 500, 1000, 1500, 2000 |
| Max Depth | 1, 2, 3, 4, 5, No limit on max |
| Max Features | ‘log2’, ‘sqrt’ |
| Criterion | ‘gini’, ‘entropy’ |

As we can see, we are optimizing the number of trees in our forest, the maximum depth of our trees, the maximum number of features we are sampling to split a node into two leaves, and what

the criterion is for a split. Here are the results from fitting our model and validating against the training data:

**Random Forest with Random Under-sampling - Training Results**

**Optimized Parameters:**

| | |
|---|---|
| **Number of Trees** | 1000 |
| **Criterion** | Gini |
| **Max Features** | Square Root |
| **Max Depth** | 4 |

**Accuracy Score: 60.6%**

**Classification Report:**

| | **Precision** | **Recall** |
|---|---|---|
| **0** | 0.93 | 0.61 |
| **1** | 0.15 | 0.59 |
| **avg/total** | 0.85 | 0.61 |

**Confusion Matrix:**

| | |
|---|---|
| 5935 | 3842 |
| 461 | 676 |

**AUC Score: 0.649**

Our results are quite different from those of our logistic regression model. We have a considerably lower accuracy. We have a much lower recall on 0's than we did before, but a much higher recall on 1's. That is, we were less accurate in identifying those without psychological distress but we were more accurate in identifying those with severe psychological

distress. We needed 1000 trees to get these results, which is a fair amount of trees needed to train our model. However, our max depth of each tree is 4, which is relatively low. This will offset the increase in computational time needed by having 1000 trees. The standard for the number of features randomly sampled is typically the square root of the number of features, and the default criterion for choosing splits is the Gini Index. So, there is nothing out of the ordinary with these parameters. Here are our results on the test set:

**Random Forest with Random Under-sampling - Testing Results**

**Accuracy Score: 60.0%**

**Classification Report:**

| | **Precision** | **Recall** |
|---|---|---|
| **0** | 0.93 | 0.60 |
| **1** | 0.16 | 0.63 |
| **avg/total** | 0.84 | 0.60 |

**Confusion Matrix:**

| | |
|---|---|
| 1925 | 1305 |
| 152 | 257 |

**AUC Score: 0.653**

As we can see, our results are similar to our training results. So, as with the logistic regression model, we do not have an overfit model. The details of this model can be found here.

We get very similar results when using balanced class weights, so we will not summarize these findings here. The details can be found here.

## Boosting

In the final model, we will try two boosting approaches: AdaBoost and Gradient Boosting. Keep in mind that AdaBoost focuses on the observations with the most error, and adds Decision Trees with particular weights according to how well they perform. Learners that have high performance will be given a higher weight. On the other hand, gradient boosting does not focus on specific observations: rather, the Decision Tree trains on the remaining errors of the strong learner.

Since we do not have any features that interact with class weights in this model, we will need to use a random under-sampler to make the data more balanced. That is, we will need to randomly under-sample the majority class without replacement. The hyperparameters we are optimizing are the number of learners and the learning rate. Furthermore, we will used RandomizedSearchCV again to reduce computational time.

Let's look at the results of our AdaBoost model.

**AdaBoost with Random Under-Sampling - Training Results**

**Optimized Parameters:**

| | |
|---|---|
| **Number of Trees** | 1000 |
| **Learning Rate** | 0.1 |

**Accuracy Score: 67.2%**

**Classification Report:**

| | **Precision** | **Recall** |
|---|---|---|
| **0** | 0.93 | 0.69 |
| **1** | 0.17 | 0.54 |
| **avg/total** | 0.85 | 0.67 |

**Confusion Matrix:**

| | |
|---|---|
| 6711 | 3040 |
| 530 | 633 |

**AUC Score: 0.665**

We get similar results to those of our logistic regression model: we have higher recall on 0's and lower recall on 1's. Like with our random forest, our optimized number of trees is 1000, indicating that we need a fair amount of learners to get our results. We have a learning rate of 0.1, which implies we shrink the contribution of each classifier by 0.1, which is not an insignificant amount. Thus, we need a lot of "information" but the contributions of those pieces of "information" is shrunk. Let's see our results on the test data:

**AdaBoost with Random Under-Sampling - Testing Results**

**Accuracy Score: 67.6%**

**Classification Report:**

| | **Precision** | **Recall** |
|---|---|---|
| **0** | 0.93 | 0.69 |
| **1** | 0.17 | 0.55 |
| **avg/total** | 0.85 | 0.68 |

**Confusion Matrix:**

| | |
|---|---|
| 2250 | 1006 |
| 172 | 211 |

**AUC Score: 0.646**

Our results are nearly identical to those of our training results. Our model is not overfit. The details of this model can be found here.

Let's look at the results we get from Gradient Boosting with random under-sampling.

**Gradient Boosting with Random Under-sampling - Training Results**

**Optimized Parameters:**

| | |
|---|---|
| **Number of Trees** | 500 |
| **Learning Rate** | 1 |

**Accuracy Score: 69.7%**

**Classification Report:**

| | **Precision** | **Recall** |
|---|---|---|
| **0** | 0.93 | 0.71 |
| **1** | 0.19 | 0.55 |
| **avg/total** | 0.85 | 0.70 |

**Confusion Matrix:**

| | |
|---|---|
| 6963 | 2777 |
| 529 | 645 |

**AUC Score: 0.680**

Our results are better than those of our AdaBoost training results. We have a higher recall and precision. Accuracy, as well as AUC score, is slightly higher. We require less trees, although our learning rate is slightly higher. So far, we have a more efficient (AdaBoost model took about 5

minutes to train, and the Gradient Boosting model took about 1.5 minutes to train) and higher performing solution than the AdaBoost model. Let's see the results on the test data.

**Gradient Boosting with Random Under-sampling - Testing Results**

**Accuracy Score: 67.9%**

**Classification Report:**

| | **Precision** | **Recall** |
|---|---|---|
| **0** | 0.92 | 0.70 |
| **1** | 0.16 | 0.50 |
| **avg/total** | 0.85 | 0.68 |

**Confusion Matrix:**

| | |
|---|---|
| 2287 | 980 |
| 187 | 185 |

**AUC Score: 0.621**

Unfortunately, our gradient boosting model was slightly overfit. All of our metrics were slightly worse than with our training results. So, we may want to use our AdaBoost model instead. The details of this model can be found here.

# Conclusion

In this project, we looked at how a variety of metrics related to drug use affected psychological distress. First, we did some exploratory data analysis to verify that there was, in fact, relationships we could explore. Our most interesting findings involved abuse of the substance, as well as whether or not the individual had tried the substance before 18. In both cases, we found

statistically significant results. Then we tried using three models to predict whether or not an individual had experienced severe psychological distress in the past month. Our most successful model was our logistic regression model with previous random under-sampling of the majority class. This model had an accuracy of about 70%, and it did not overfit.

Based on our exploratory data analysis and modeling, we can conclude that drug usage certainly does have an impact on whether or not an individual has had severe psychological distress in the past month. Unfortunately, our model accuracy is not particularly high yet. This leaves a lot of room for future improvement. When doing modeling, we recommend that random under-sampling is always used to deal with the imbalanced dataset. Although we technically had the best results with our logistic regression model, we achieved similar results with our boosting models. Nonetheless, our logistic regression model is faster and so it is recommended to use a logistic regression model with random under-sampling.

In order to improve some of our models, we could do a few things. we could have used GridSearchCV instead to test more hyperparameters at the expense of computational time. We could have used our random forest model to do some feature selection, and used the most important features in our both our logistic regression model and boosting model. Finally, we could have tried XGBoost, a different boosting algorithm which has been shown to typically have higher performance than other boosting algorithms. Nonetheless, we have a somewhat accurate model that does not overfit.