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.

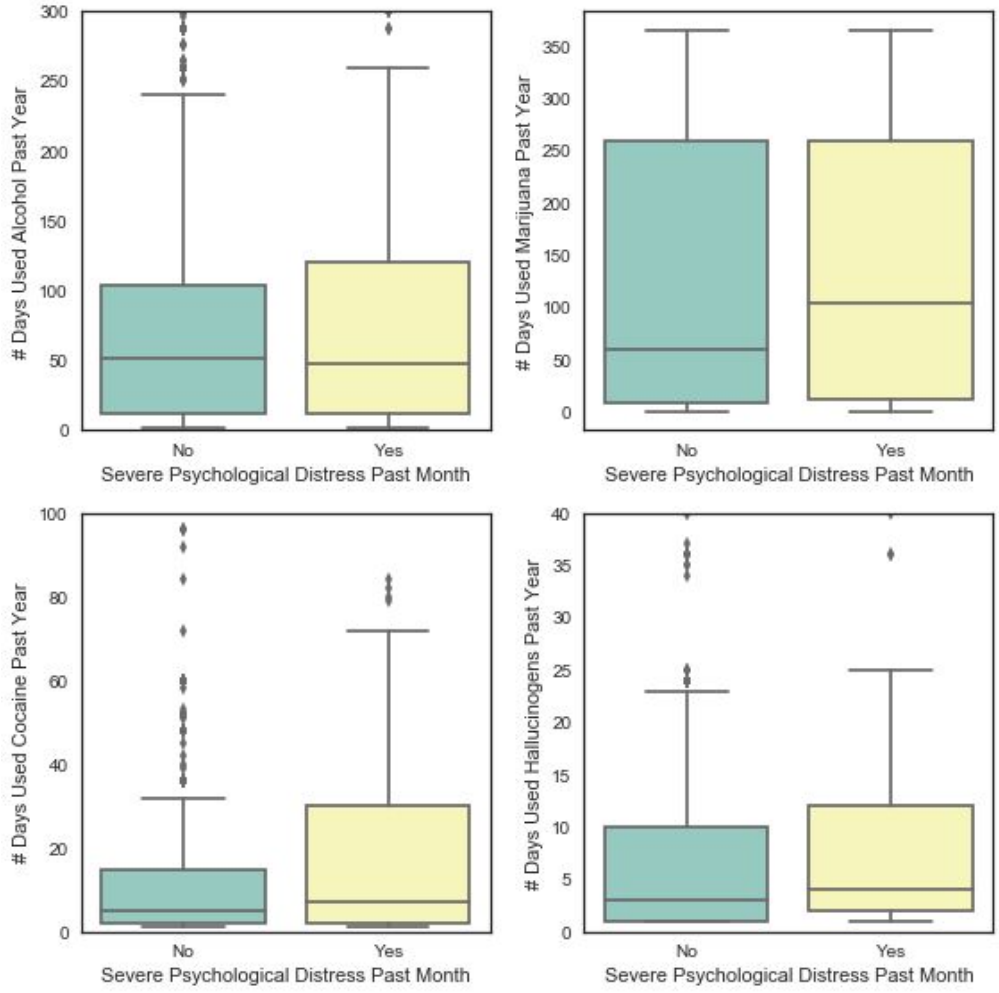
Data Cleaning

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

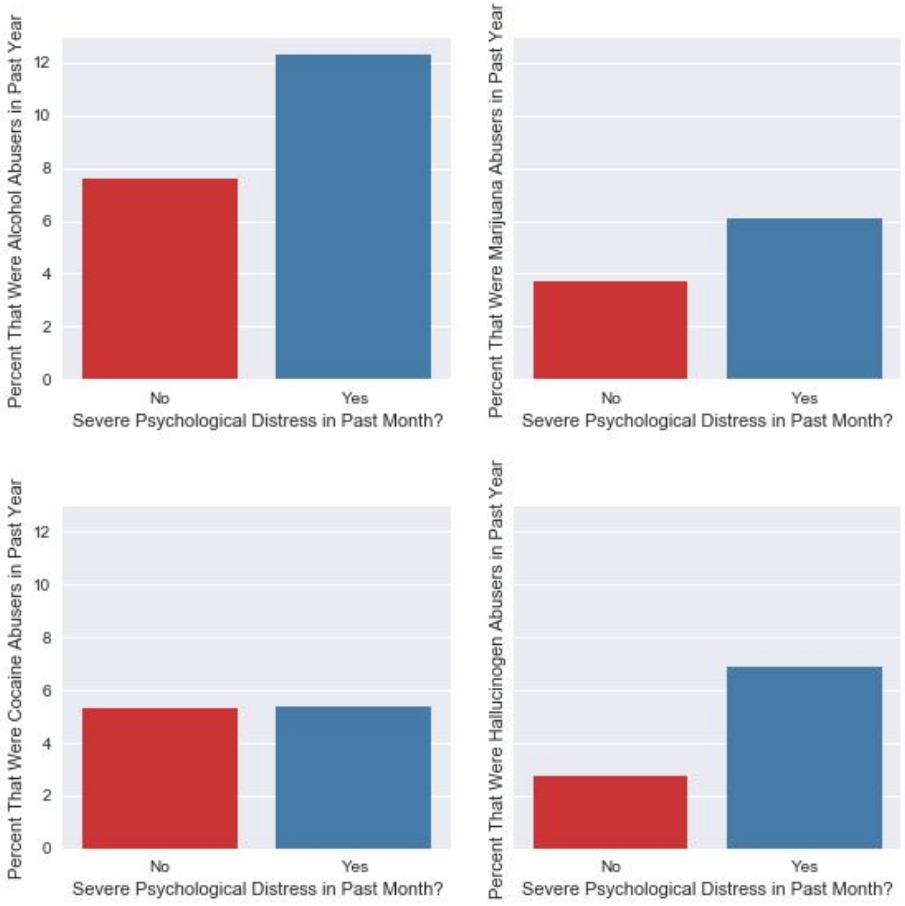
Here, we will create some visualizations to try and investigate any trends and/or patterns that may be useful to look into during our analysis. Keep in mind, our goal is to find out if excessive substance abuse has a correlation with development of mental illnesses and psychological distress. Whether 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)



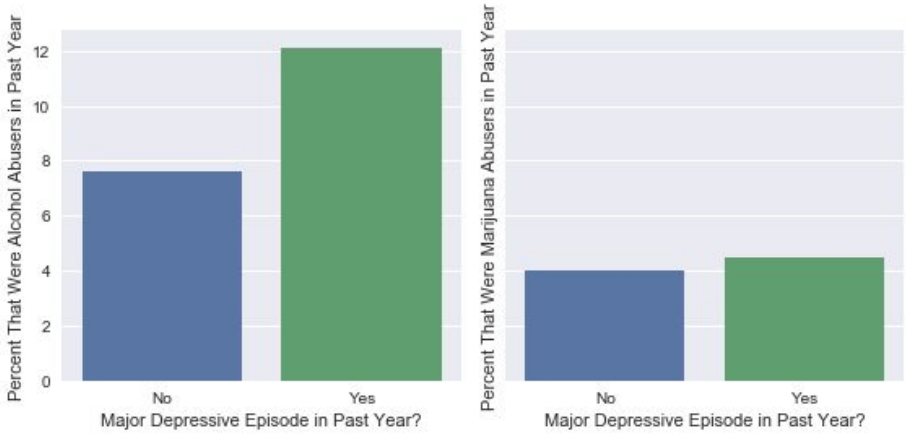
Here, we took the subsets of people with and without severe psychological distress in the past month, and compared the percent of them that were abusers of the substance. The results are shown in the above barplot.

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.

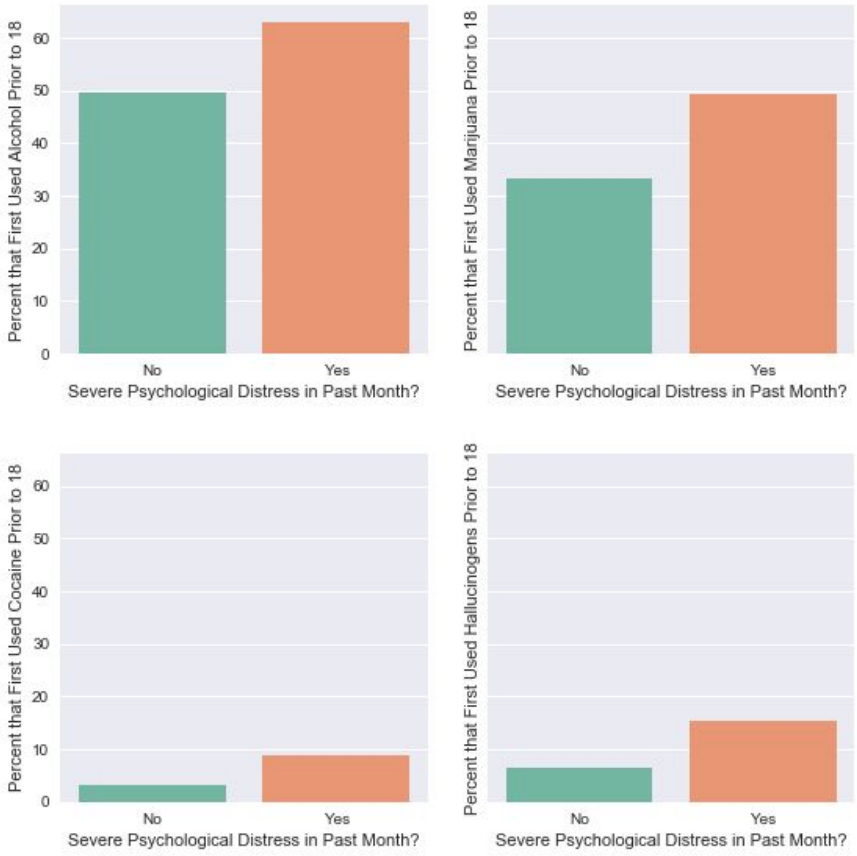
Major Depressive Episodes (Past Year) and Alcohol/Marijuana Abuse (Past Year)



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 use a pipeline in which we first scale the data so that each column has a mean of zero and a variance of one. 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 GridSearchCV (we will use 5-fold cross validation).

Here are the results we get on our training data:

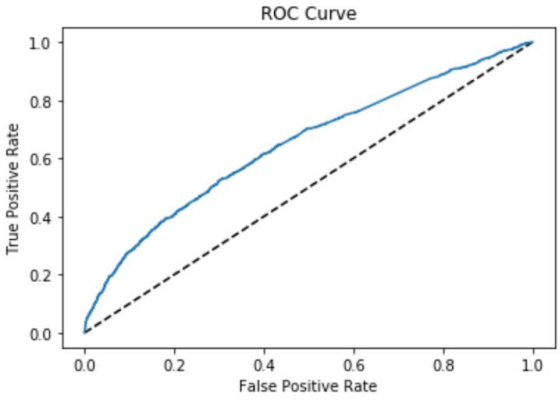
```
Tuned Logistic Regression Parameters: {'logistic__C': 0.0007196856730011522, 'logistic__class_weig
ht': 'balanced'}
Results


Accuracy Score: 0.676646706587
Classification Report:
             precision    recall  f1-score   support

          0       0.92      0.69      0.79      9099
          1       0.17      0.53      0.26      1088

avg / total       0.84      0.68      0.74     10187

Confusion Matrix:
 [[6318 2781]
 [ 513  575]]
AUC Score: 0.646936799778
```



We chose the regularization parameter C such that the area under the ROC curve is maximized. We also chose our class weights for the logistic model to be balanced: this will help fix our problem of unbalanced data (about 90% of individuals have not demonstrated severe psychological distress in the past month. 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 68%, 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.

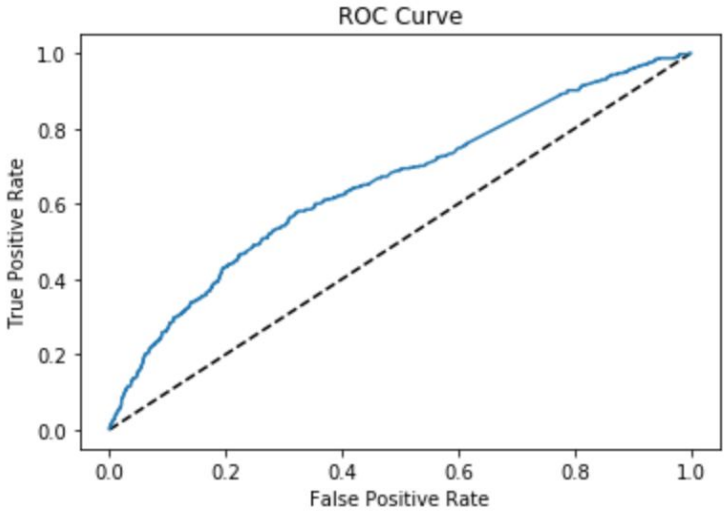
Here are the results:

```
Accuracy Score: 0.680027485112
Classification Report:
             precision    recall  f1-score   support

          0       0.93      0.70      0.80      3908
          1       0.17      0.55      0.26       458

avg / total       0.85      0.68      0.74      4366

Confusion Matrix:
 [[2718 1190]
 [ 207  251]]
AUC Score: 0.651934448651
```



We actually get slightly better results on our test data than we did our training data. Our accuracy, AUC, and average precision scores are all greater than the scores we had on 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.

Random Forest Model

Now, we will implement a random forest model to help solve our classification problem. Similar to the logistic regression model, we will use 5-fold cross-validation. However, in this model we will use RandomizedSearchCV to optimize parameters because we have a lot of hyperparameters to optimize and this will save us time. Keep in mind that RandomizedSearchCV samples certain hyperparameters from a probability distribution rather than testing all hyperparameters. As a result, we will sacrifice some accuracy for computational efficiency. Here are our hyperparameters:

```
n_trees = [10,20,50,100]
possible_max_depths = [3,5,None]
# tuning some more hyperparameters this time
param_grid = {'n_estimators':n_trees,
              'max_depth':possible_max_depths,
              "max_features":['log2','sqrt'],
              "class_weight":['balanced']}
```

As we can see, we are optimizing the number of trees in our forest, the maximum depth of our trees, and the maximum number of features we are sampling to split a node into two leaves. Like our logistic regression model, we are using balanced class weights to weight observations with a lower frequency relatively higher compared to observations with higher frequency. Here are the results from fitting our model and validating against the training data:

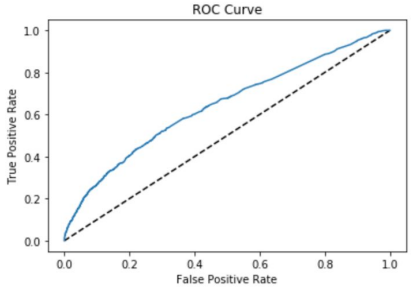
```
Tuned Random Forest Parameters: {'n_estimators': 50, 'max_features': 'log2', 'max_depth': 3, 'clas
s_weight': 'balanced'}
Results


Accuracy Score: 0.67024005864
Classification Report:
             precision    recall  f1-score   support

          0       0.93      0.69      0.79      9752
          1       0.17      0.53      0.26      1162

avg / total       0.84      0.67      0.73     10914

Confusion Matrix:
 [[6695 3057]
 [ 542  620]]
AUC Score: 0.641660027547
```



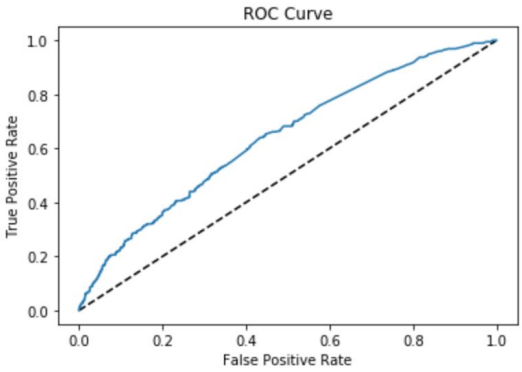
Our results are not much better than those of our logistic regression model. Here are our results on the test data:

```
Accuracy Score: 0.648804616653
Classification Report:
             precision    recall  f1-score   support

          0       0.92      0.66      0.77      3255
          1       0.16      0.53      0.24       384

avg / total       0.84      0.65      0.72      3639

Confusion Matrix:
 [[2159 1096]
 [ 182  202]]
AUC Score: 0.640178971454
```



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.

Boosting

In the final model, will test is a gradient boosting model. 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.

We will look at the results of fitting and training our model on the next page.

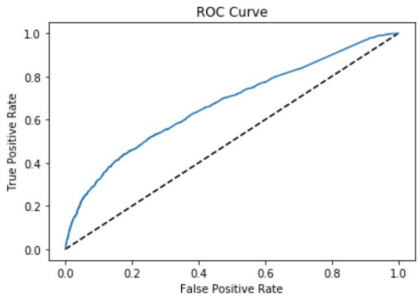
```
Tuned Gradient Boosting Parameters: {'gradient__n_estimators': 50, 'gradient__learning_rate': 0.1}
Results


Accuracy Score: 0.706157229247
Classification Report:
             precision    recall  f1-score   support

          0       0.93      0.73      0.82      9731
          1       0.19      0.53      0.28      1183

avg / total       0.85      0.71      0.76     10914

Confusion Matrix:
 [[7076 2655]
 [ 552  631]]
AUC Score: 0.672770432495
```




```
CPU times: user 9.06 s, sys: 317 ms, total: 9.38 s
Wall time: 9.31 s
```

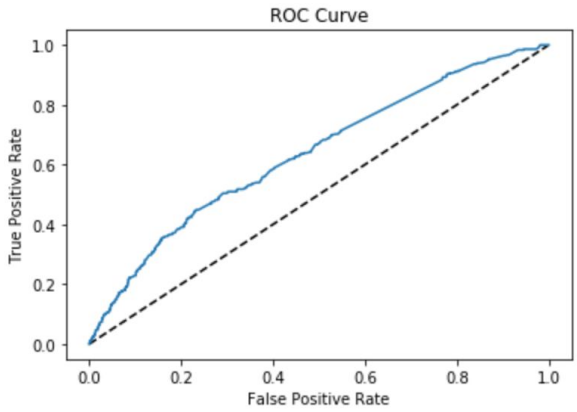
We get better results on our training data than both our logistic regression model and our random forest model. Let's see our results on the test data:

```
Accuracy Score: 0.700192360539
Classification Report:
             precision    recall  f1-score   support

          0       0.93      0.72      0.81      3276
          1       0.16      0.48      0.24       363

avg / total       0.85      0.70      0.76      3639

Confusion Matrix:
 [[2373  903]
 [ 188  175]]
AUC Score: 0.637145262145
```



As we can see, we got better a higher average precision, average recall, and accuracy score than with our other models. Moreover, our model is not overfitting. 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 gradient boosting model with previous under-sampling of the majority class. This model had an accuracy of about 70%, and it did not overfit. In order to improve some of our models, we could do a few things. For one, we may want to try using random under-sampling on our data for our logistic regression model and our random forest model. Moreover, we could have used GridSearchCV instead to test more hyperparameters in our random forest and our gradient boosting model at the expense of computational time. We could have also done more with feature selection. Nonetheless, we have a somewhat accurate model that does not overfit.