CMSE 381 Final Project- Analyzing Trends in Drug Use

April 19, 2021

1 Introduction

Drug Use is a topic that has sparked much debate over the better part of the last century. The negative effects of many drugs are apparent, and many policy initiatives have been tried out to help curb drug use.

This project will attempt to better understand drug use through a combination of machine learning techniques and data exploration. The dataset used consists of 1885 observations (individuals) and 12 attributes that describe personality traits and demographic information such as age and ethnicity.

2 Problem Statement

This report asks three main questions-

1) Can we accurately predict aggregate illegal drug usage using a combination of demographic/personality data?

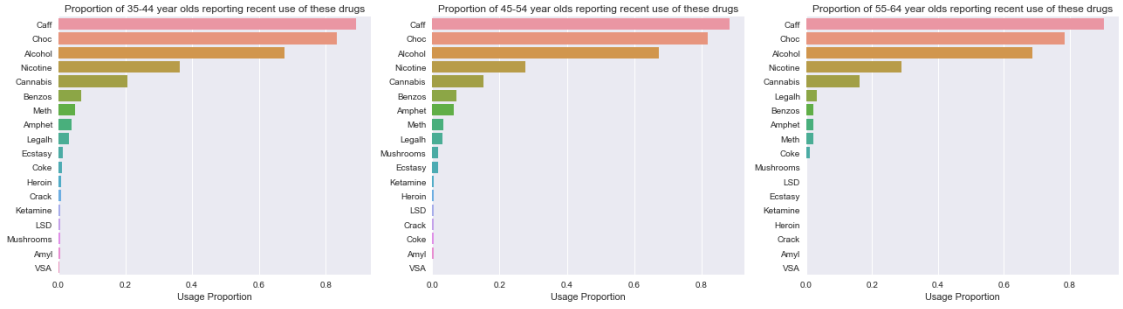
2) Can we accurately predict the recent use of of specific illegal drugs using demographic/personality data?

3) What are the common traits of drug users? What kinds of users are attracted to specific drugs?

This project will attempt to answer all three questions.

3 Related Results

First, I recreated the results from FiveThirtyEight [1] using this new dataset.



The original results consisted of a barplot of drug use for people aged 50-64. I extended these results a little, making three separate plots for the age ranges 35-44, 45-54, and 55-64. There are also "drugs" in this dataset (such as Caffeine, Alcohol and Chocolate) that are not present in the initial analysis. These drugs lead to very interesting results that make this dataset much richer than the original.

4 Modeling and Results

4.1 Feature Extraction

This data required a bit of preprocessing before it was ready for analysis. The researchers included a fake drug called "Semeron" to identify the individuals with a tendency to "over-report" their drug use. I simply filtered out the individuals who reported they used "Semeron" at any point.

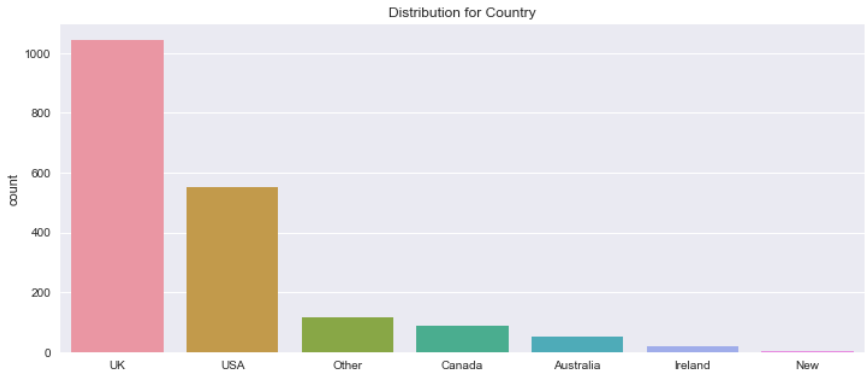
I also had to find a way to better encode the drug features. Each of the drug columns consisted of seven classes, CL0-CL6, each representing the frequency of use for the drug. For example, CL0 indicates that the individual never used the drug, while CL6 indicates that the individual used the drug in the past day.

My purpose for this project is predicting recent drug use, so I decided to transform these columns into binary features. If an individual used the drug in the past day or the past week (CL5 and CL6), they would be given a value of 1 for that column. Otherwise, they would be assigned a value of 0. This binary feature transformation also made it easier to use methods such as logistic regression.

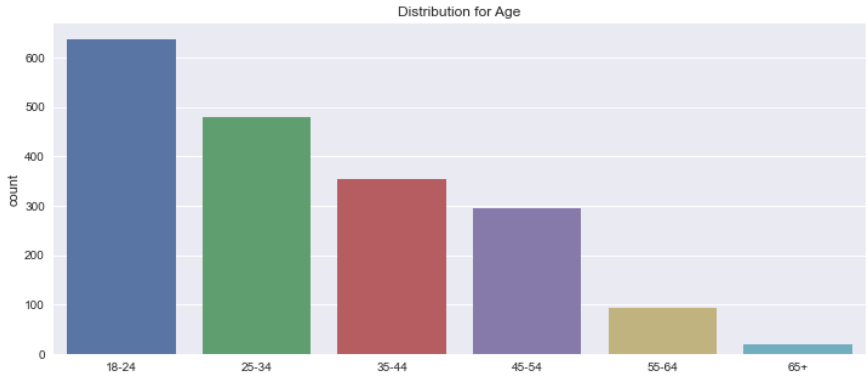
In addition, I generated a new feature to represent recent illegal drug use. This binary feature, which I called Illegal Drug Usage, is equal to 0 when the individual consumed a drug legal in the United States ("Alcohol", "Caff", "Choc", "Legalh", "Nicotine", and "Amyl" in this dataset) and 1 otherwise. I thought this would be an interesting feature that could be used to separate some of the more serious drugs from the less serious ones, while providing a single target variable that can be used in modeling.

4.2 Exploratory Data Analysis

First, before diving into modeling, I wanted to dig a little bit deeper into the data. I first plotted the distribution of the Country feature in the dataset.



As the plot above shows, this dataset samples almost exclusively from Western, English-speaking countries. The U.K, in particular, is very heavily represented. This serves as both an advantage and a disadvantage. The advantage is that cultural factors often play a huge role in drug use, and these countries are (relatively) similar culturally; this prevents hard-to-quantify cultural factors from potentially muddling the results. The disadvantage is that the dataset is less representative of the world as a whole.



This second plot is a distribution of Age variable, and it demonstrates a reasonably heavy skew towards younger respondents. This is significant because younger people tend to use drugs at higher rates, and a disproportionate representation of younger people may make it harder to make conclusions about the data that apply to the population as a whole.

4.3 Model Selection

My first stage for modeling was predicting the Illegal Drug Usage feature using all of the personality/demographic features. The three models I chose were Logistic Regression, Support Vector Classifier (SVC), and a Random Forest Classifier (RFC).

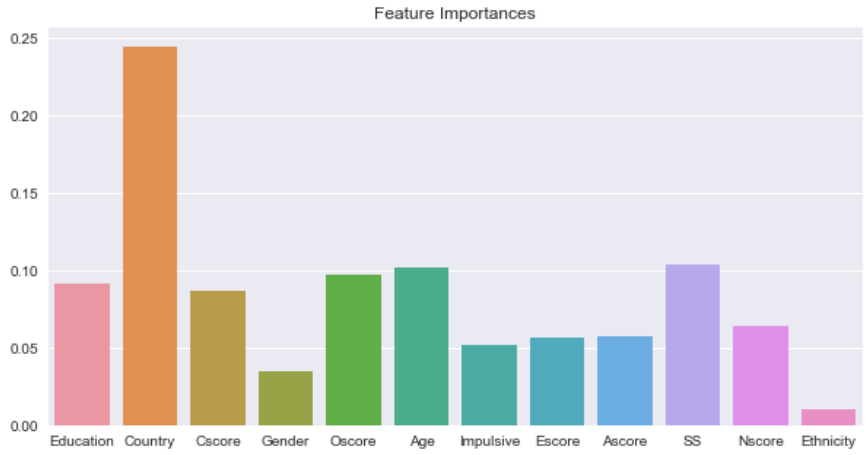
All of the optimal hyperparameters for these three models were tuned using 5-fold cross validation. The optimal C hyperparameter (the inverse of λ) came out to be 0.04641589. The optimal kernel for the SVC was a gaussain rbf, while the optimal C was equal to 1. As for the RFC, the optimal number of trees was equal to 1000, the maximum depth of each tree was equal to 16, and the maximum number of features was equal to the square root of the total number of features.

4.4 First Stage Results

I used two metrics to measure the performance of these classifiers, simple classification accuracy and f1-score. F1 score better accounts for imbalanced classes than classification accuracy, and I thought it wise to include it as a result.

As for the performance of the three classifiers- SVC came out on top with a classification accuracy of 0.812 and an f1-score of 0.792. RFC placed second with a classification accuracy of 0.804 and an f1-score of 0.779. Logistic Regression followed with a classification accuracy of 0.796 and an f1-score 0.756.

The plot below shows the importance of each feature outputted from the Random Forest Classifier. It is interesting to note that the Country feature appears to be dominant; this very well may be due to differences in drug policy between the countries in the dataset, since my Illegal Drug Usage feature is based on drugs illegal only in the United States.



4.5 Second Stage

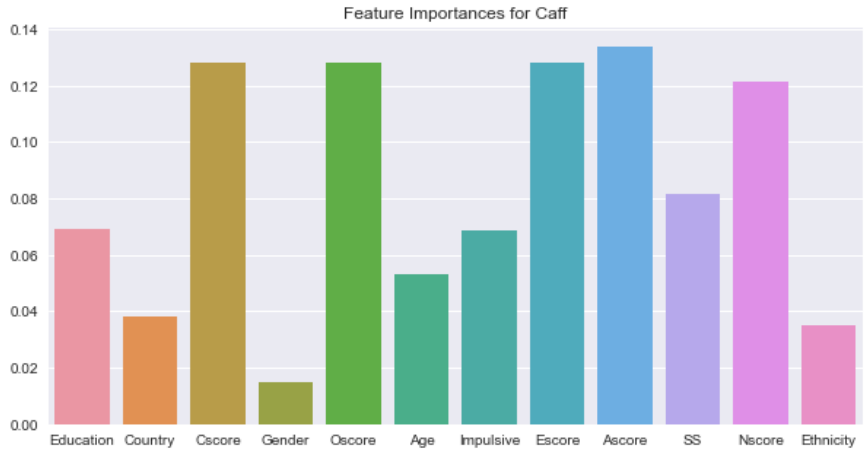
For the second stage of my modeling, I sought to answer the second question posed by this project- how accurately can we predict usage for each individual drug based on this demographic/personality data?

I started off by looping through the set of drug columns fitting the demographic/personality features to each one using cross validated Logistic Regression. It became quickly apparent that it is much harder to predict the usage of several of the individual drugs because of heavily imbalanced classes.

For many of the drugs such as Crack, Heroin, and VSA, there were very few users present in the dataset. As such, it made accurately predicting usage a difficult task that resulted in very low accuracy. Nonetheless, several of the more common drugs such as Alcohol, Cannabis, Caffeine saw f1-scores above 0.65. For the next stage, I chose only the drugs that had at least 100 recent users in the dataset and had a Logistic Regression f1-score of above 0.5. These drugs ended up being 'Alcohol', 'Caff', 'Cannabis', 'Choc', and 'Nicotine'.

4.6 Third Stage- Patterns among users

The third stage of my analysis sought to uncover patterns among the users of specific drugs. I started off by running a cross-validated Random Forest Classifiers with each of these targets to identify the most important features.

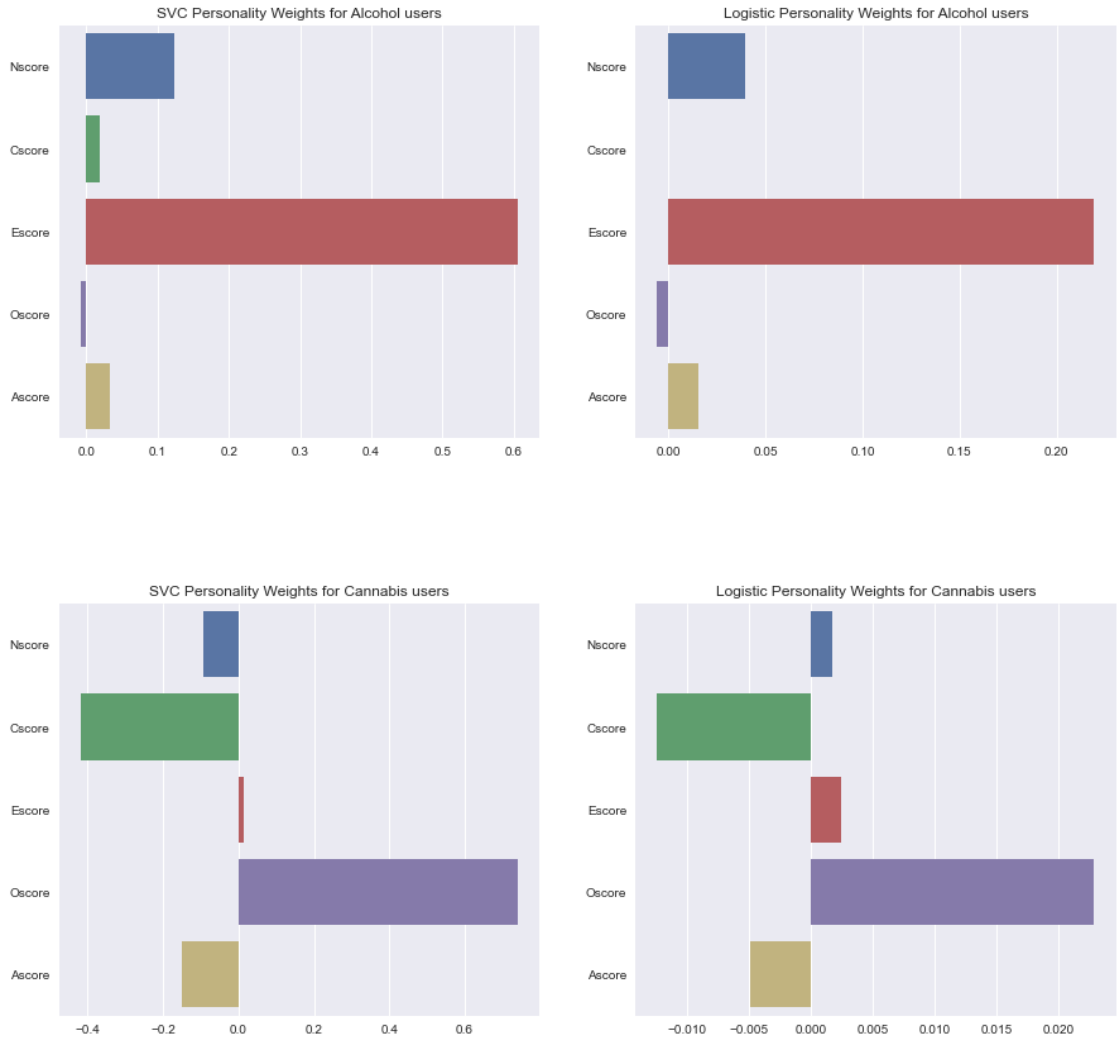


Above is the results for Caffeine. I was very pleased to discover that the top five most significant features were the personality features (Oscore, Escore, Ascore, Nscore, and Cscore). This is a trend that held up for Alcohol, Caffeine, and Chocolate as well, but notable not for Cannabis (where the Country feature was the most significant- presumably due to differences in marijuana laws across the countries).

4.6.1 Quantifying Effects of Personality Traits

I was very encouraged by the fact that the Big 5 personality traits seemed to be important predictors of drug usage for most of the commonly used drugs. For the final part of my modeling, I attempted to quantify the effects of these traits on drug usage and answer the big question- what distinguishes users of one drug from another?

I subsetted the features to only include the personality traits and ran Logistic regression as well as a Linear SVC on each of the common drugs. I then retrieved the model coefficients from both models and plotted them in an attempt to visualize the effects of each of the personality traits. Here are the results for Alcohol and Cannabis:



It is apparent that there are pretty stark differences in the coefficients between Cannabis and Alcohol users. The Escore coefficient for Alcohol is much higher than that of Cannabis, which would imply that Alcohol users are more extroverted. Conversely, the Oscore coefficient for Cannabis is higher, which would suggest that Cannabis users are more open to new experiences. These results are exactly what I was looking for when I started this project; a quantifiable way to distinguish users of different drugs using data such as personality scores.

5 Conclusion

Overall, I was pleased with the results of the project. It seems that it's possible to predict aggregate illegal drug usage based on the features given in this dataset with a reasonable degree of accuracy; the Logistic Regression, SVC and RFC models all produced f1-scores of above 75 percent and classification accuracy of above 79 percent. Predicting the usage of individual drugs was less simple due to imbalanced classes, but the major conclusion from this project is that certain personality traits are more likely to predict the use of some drugs than others.

5.1 Future Work

This analysis is by no means comprehensive. It may be prudent to further look into interactions between the use of certain drugs; for example, does the use of Cannabis make an individual more or less likely to use other drugs? It may also be interesting to find data from different parts of the world, particularly non-Western countries to see if these same trends hold.

6 References

https://fivethirtyeight.com/features/how-baby-boomers-get-high/

https://archive.ics.uci.edu/ml/datasets/Drug+consumption+

https://www.drugs.com/article/csa-schedule-1.html

https://scikit-learn.org/stable/modules/generated/sklearn.linear$_m$*odel.LogisticRegressionCV.html*

https://sklearn.org/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html