Understanding the relationship between drugs and the psyche

Introduction

According to a recent study, nearly 24 million people in the United States abuse illicit drugs, nearly 18 million people abuse alcohol, and in 2012 alone 22,114 people died from prescription drug overdoses. Epidemiologists in several states blame the increasing number of drug-related deaths on greater use of heroin and synthetic opioids, such as fentanyl.

Infact, fatal overdoses surpassed shooting deaths and fatal traffic accidents years ago. The negative consequences of drug abuse affect not only individuals who abuse drugs but also their families and friends, various businesses, and government resources. Although many of these effects cannot be quantified, ONDCP recently reported that in 2002, the economic cost of drug abuse to the United States was $180.9 billion.

All these facts clearly indicate that the burden in terms of costs, trauma, and influence on the nation's youth is substantial. Moreover, the effects are pervasive and carry huge costs to society as a whole. The efforts to fight drug abuse, the "War on Drugs", has been declared a failure due to its inability to understand the core reasons behind the use and spread of drugs. Drug use is not something that happens on its own in isolation. In order to start addressing this problem from the ground up, we want to look at the relationship between an individual's personality traits and the drugs he uses. We extend this to find a way to predict if a person can be a potential drug user.

Figure 1: Are certain types of people more prone to drug use?

Materials and Methods

Materials - Information about data The dataset we have used offers gives us information on personality traits, which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and

conscientiousness - also called the big five personality traits) , BIS-11 (impulsivity), and ImpSS (sensation seeking) scores, level of education, age, gender, country of residence and ethnicity. All these predictors can be very relevant and apt indicators to evaluate the risk of an individual being a drug consumer. For each user, the data provides a flag to indicate when he used a drug (one of Amphetamine, AMYL, Benzos, Cannabis, Coke, Crack, Ecstasy, Heroin, Ketamine, Legalh, LSD, Meth, Mushrooms, Semer, VSA)

We have added Alcohol, Caffeine and Nicotine use to the list of predictors as they are the three most widely consumed psychoactive substances in the world and can give us a better insight into each person's personality.

Materials - Preprocessing: Before building the models, we preprocessed the data by binarizing the response class from the seven classes in the original data ("Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day") to two classes : 'User' and 'Non User'. This reduces the problem from a multi class classification to a bi-class classification.

The dataset has a fictitious drug named 'Semeron' to weed out potentially dishonest answers on the survey used to create the dataset. We have used this column to fulfill the same purpose. Anyone who has answered in the affirmative for Semer has been classified as a 'Non User' for all other drugs.

```r
#Reading in the data
drugs_data = read.csv("drug_consumption_uci.csv")

#Have to code the different variables from their number state to categorical state

cols_recode = c("Amphet", "AMYL", "Benzos","Coke", "Cannabis", "Alcohol",
                "Crack","Ecstasy","Heroin","Ketamine" ,"Legalh","LSD",
                "Meth","Mushrooms","Semer","VSA", "Caff", "Nicotine")

drugs_data[,cols_recode] = apply(drugs_data[,cols_recode],2,recode_func)

hard_drugs = c("Amphet", "AMYL", "Benzos", "Coke", "Crack", "Heroin", "Meth", "VSA")
soft_drugs = c("Cannabis", "LSD", "Mushrooms", "Ecstasy")

drugs_data$soft_user = apply(drugs_data[,soft_drugs],1,
                             function(x){ifelse("User" %in% x, "User", "Non User")})

drugs_data$hard_user = apply(drugs_data[,hard_drugs], 1,
                             function(x){ifelse("User" %in% x, "User", "Non User")})

#transforming the hard and soft drug flags based on semer
drugs_data$soft_user = as.factor(ifelse(drugs_data$Semer == "User", "Non User", drugs_data$soft_user))

drugs_data$hard_user = as.factor(ifelse(drugs_data$Semer == "User", "Non User", drugs_data$hard_user))
```

Data - Framing the question:

Our initial idea was to group the drugs into two - soft drugs and hard drugs (Inspired by the information from this link - https://simple.wikipedia.org/wiki/Hard_and_soft_drugs). But to do so, we needed definitive proof that the underlying patterns between the drugs in a group were the same. We ran a series of t-tests to test this idea.

```r
t_function = function(column1, column2, col_no){
t.test(drugs_data[drugs_data[,column1] == 'User', col_no],
       drugs_data[drugs_data[,column2] == 'User', col_no])
}

tvalshard = vector("list", 36)
```

```
k = 1
for(i in 1:8){
  for(j in i:8){
    col1 = hard_drugs[i]
    col2 = hard_drugs[j]
    tvalshard[[k]] = lapply(7:13, t_function, column1 = col1, column2 = col2)
    for(m in 1:7){
      if(tvalshard[[k]][[m]]$p.value < 0.05){
        print(sprintf("col1 = %s, col2 = %s, column = %d",col1, col2, (m+6)))
      }
    }
    k = k+1
  }}

tvalssoft = vector("list", 11)
k = 1
for(i in 1:4){
  for(j in i:4){
    col1 = soft_drugs[i]
    col2 = soft_drugs[j]
    tvalssoft[[k]] = lapply(7:13, t_function, column1 = col1, column2 = col2)
    for(m in 1:7){
      if(tvalssoft[[k]][[m]]$p.value < 0.05){
        print(sprintf("col1 = %s, col2 = %s, column = %d",col1, col2, (m+6)))
      }
    }
    k = k+1
  }}
```

We concluded from these tests that none of the hard drugs can be grouped together. That idea will have to be shelved. On the other hand, user of LSD, Mushrooms and Ecstacy do seem to have similar personality traits on the whole, and can hence be combined.

Models

We constructed a series of models to evaluate the risk of someone consuming the 'soft' drugs in consideration (LSD, Mushrooms and Ecstasy). Rather than just look at the accuracy of each model, we were also taking a serious look at the sensitivity and specificity in the model performance results. In clinical use, identifying a potential drug user as someone who is not likely to do the same would defefat the point of the model.

```
set.seed(seed)
soft_lda = train(soft_user ~ .,
                 data = soft_trn,
                 method = "lda",
                 trControl = trainControl(method = "cv", number = 5))

lda_soft_pred = predict(soft_lda, soft_tst)
lda_soft_acc = calc_acc(actual = soft_tst$soft_user, predicted = lda_soft_pred)
lda_soft_cf = confusionMatrix(data = lda_soft_pred, reference = soft_tst$soft_user, positive = "User")
```

When modelling Heroin usage, we faced the issue of unbalanced data. The sample training datasets created all had more Non Users than Users, and hence most of the models would have obtained high accuracy even if they had classified everyone as a non user.

Table 1: Predicting soft drug usage - kNN and Random Forest stood out

Model	Accuracy	Sensitivity	Specificity
KNN	0.7854610	0.8683274	0.7031802
LDA	0.7854610	0.9252669	0.6466431
Naive Bayes	0.7925532	0.8790036	0.7067138
RF OOB	0.8067376	0.9074733	0.7067138
Glmnet	0.8085106	0.8967972	0.7208481

To combat this issue, we tried using weights in some of the models that we built. Observations that indicated a 'User' in the training data were given increased weightage in building the model.

```
#LDA
model_weights = ifelse(heroin_trn$Heroin == "User",
                       (1/table(heroin_trn$Heroin)[1]) * 0.2,
                       (1/table(heroin_trn$Heroin)[2]) * 0.8)

set.seed(42)
heroin_lda = train(Heroin ~ .,
                   data = heroin_trn,
                   method = "lda",
                   weights = model_weights,
                   trControl = trainControl(method = "cv", number = 5)
)

lda_h_pred = predict(heroin_lda, heroin_tst)
lda_h_cf = confusionMatrix(data = lda_h_pred, reference = heroin_tst$Heroin, positive = "User")
```

Results

When comparing the models built to evaluate soft drug usage, we saw that the kNN model and random forest (with OOB) gave the best combination of accuracy along with good sensitivity and specificity.

When doing the same to predict the propensity to use Heroin, we found that most of the models came up with very poor sensitivity, despite having very good accuracies. They were not predicting whether someone was a User with near enough accuracy for the model to be useful. In this case, a Naive Bayes model provided us with the best results (with a sensitivity of 67%)

Discussion

These models show strong predictive ability in personality profiles to predict drug usage. - From Table 1, we can see that soft drug usage is best predicted by both a parametric model (kNN) and a non-parametric one

Table 2: Models built for Heroin use- Most had poor sensitivity

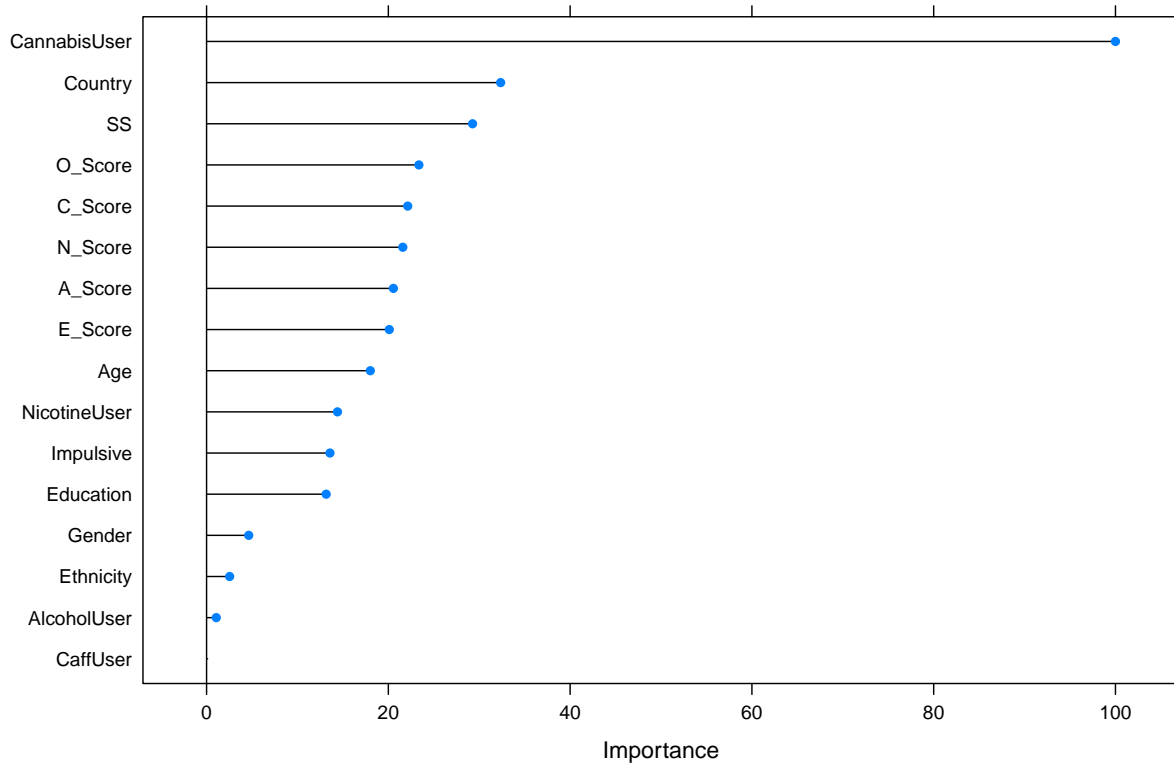| Model	Accuracy	Sensitivity	Specificity
KNN	0.8847518	0.0000000	0.9960080
LDA	0.8794326	0.0634921	0.9820359
Naive Bayes	0.7960993	0.5396825	0.8283433
RF OOB	0.8741135	0.0793651	0.9740519
GLM	0.8829787	0.0000000	0.9940120

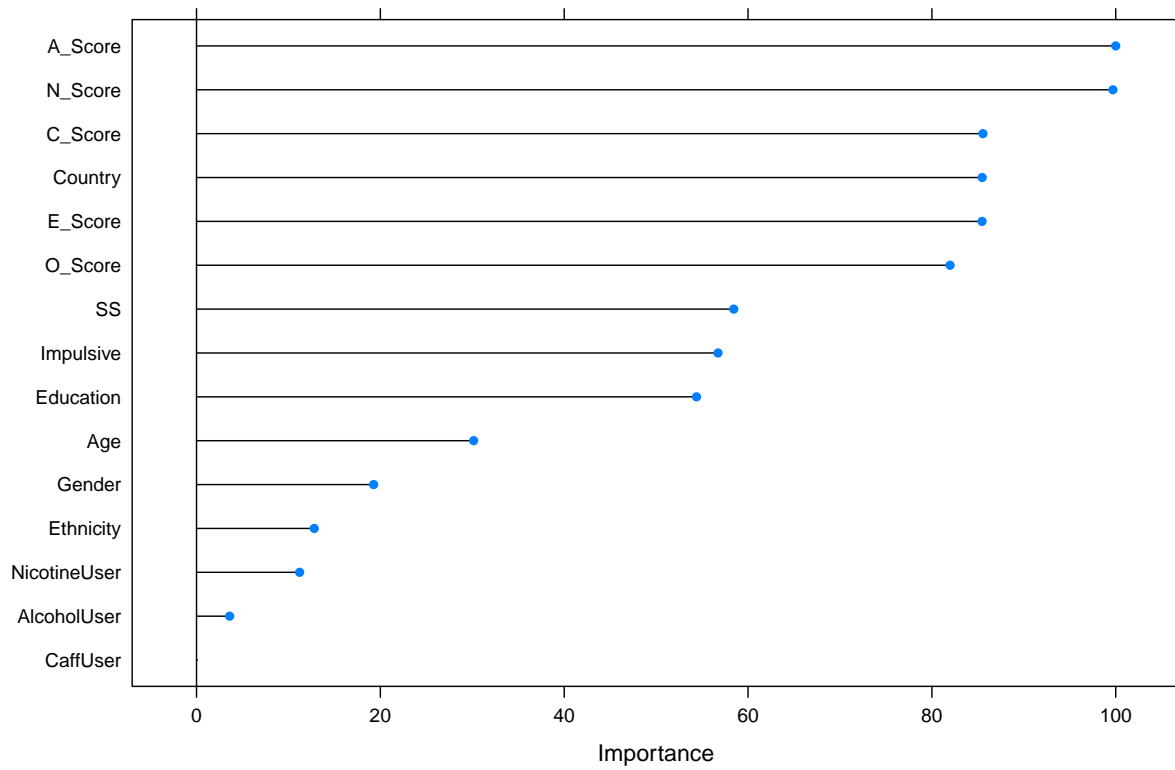Figure 2: Variable Importance for soft drug usage prediction

Figure 3: Variable Importance for heroin usage prediction

too (Random Forest). The data is not well represented by linear models, with both of these models being non-linear.

- From Table 2, the ability of the Naive Bayes models against the rest stands out. This could be due to the assumptions made by the generative model about the underlying data allowing it to fit better to the test data, which the discriminative models struggle with. The absence of a linear boundary in making these predictions was the downfall of the LDA model.

Looking at the variable importance plots generated for the random forest models built to predict soft drug and heroin usage, a number of interesting conclusions can be discussed.

- From Figure 2, we can see that usage of Cannabis plays a crucial role in your affinity towards these soft drugs. With recreational use of cannabis being legalized in a number of states in the US (and other countries worldwide), this link can prove to be an interesting starting point for a future study to explore.
- The presence of O-score higher up the list also poses some interesting questions. O-scores quantify the curiousity and open-mindedness of a subject (and are also correlated with higher intelligence). This provides further proof for the hypothesis that certain personality traits are more prone to be attracted to a certain type of drug use.
- The presence of Country on this list is questionable. The sample did not contain a significant sample of data from various countries for us to draw any confident conclusions on nationality and proclivity for drug use. Age, on the other hand, can be quite usefeul. It might point towards the need for prevention of drug use from a young age.
- From Figure 3, the absence of Cannabis usage is a conspicuous absence when it comes to predicting heroin usage. Rather, A-scores (quantifies Agreeableness) and N-scores (Neuroticism - sensitive/nervous vs. secure/confident) are very important predictors.

Table **??**

Conclusion

In exploring the relationship between the psyche and drug use, we found several useful and interesting indicators of specific drug-use while establishing the validity of using these scores as predictors. Certain types of people are definitely prone to certain types of drugs. While this by itself cannot be used to end the 'War against Drugs', they are a useful weapon if we want to dive deeper into the effects of other macroeconomic factors and their influence on people.

These results also point to a need for better emotional self-management education. If personality traits by themselves have such predictive power, it shows that with better ability to understand one's own tics and triggers, we can start the war against drugs from within, while we battle other outside forces.

References

- http://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
- E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," 2015