Machine Learning

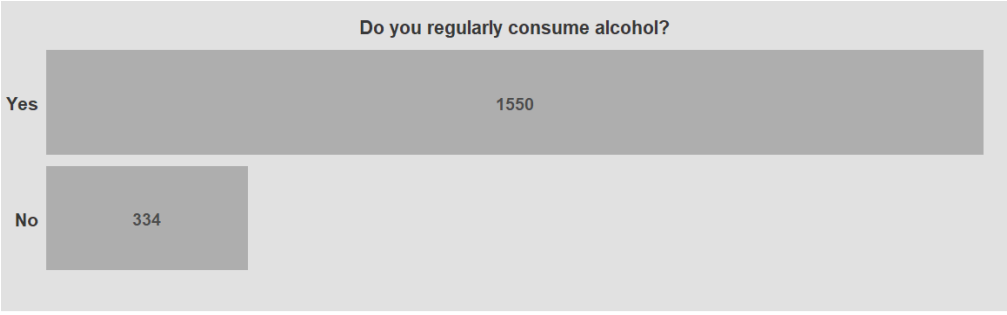
10-10-25

PSYC - 559

> Where did it all go wrong...?

- **(Thought for 20 seconds):** Good Question!

-



Simon K.
10-10-25
PSYC-559

```
    title = "Do you regularly consume alcohol?",
    x = NULL,
    y = "Count"
  ) +
  scale_fill_manual(values = c("No" = "gray60", "Yes" = "gray60")) +
  scale_y_discrete(expand = c(0, 0)) +
  theme_minimal(base_size = 14) +
  theme(
    panel.background = element_rect(fill = "gray85", color = NA),
    plot.background  = element_rect(fill = "gray85", color = NA),
    legend.position = "none",
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    axis.text.y = element_text(face = "bold", size = 16, hjust = 1, color = "black"),
    axis.text.x = element_text(size = 12, color = "black", vjust = 18.5),
    axis.title.x = element_text(size = 13, face = "bold", vjust = -10),
    plot.title = element_text(face = "bold", hjust = 0.5, vjust = -2.5, color = "black"),
    plot.margin = margin(t = 5, r = 20, b = 5, l = 5)
```

Abstract

Given data about an individual's psychological measures and drug consumption habits, could such data be modeled to predict usage of a specific drug by an individual? Using data comprising multiple dimensions of psychological measurement, as well as reported drug usage statistics, four regressional models were constructed attempting to predict both cannabis usage, as well as alcohol usage amongst the sample respondents. Regression methods used included both Elastic Net regression, as well as Lasso regression for training of models, a fourth Multiple Linear regression model was also constructed and analysed to allow for three individual models targeting cannabis. The results of these models provide insight into the ways in which

human social data can effectively be used to predict some specific substance usage, as well as how it may not provide as accurate of results with certain substances such as alcohol. The three models predicting cannabis did show accuracy in their predictive ability of >80%, while the model predicting alcohol provided results of significantly less confidence and accuracy.

Introduction

The human brain is by far the most complex *thing* that we know of, within the extent of that which we do currently know. Furthermore, the ability to map out every individual neural connection at the scale needed to understand the physical intricacies of cognition has not yet been fully realized. Thus, the only ways in which psychologists can make meaningful inferences regarding the inner-workings of an individual's mind remain constrained to external observation alone.

The ability to accurately predict the way that an individual may act emotionally has always existed as a property of how much information regarding the individual's prior actions is available for consideration. Fortunately, this is the reality

which the field of psychology has faced for a great deal of time, and frameworks for measuring an individual's emotion and cognition have become robust. While directly mapping an individual's brain is not yet feasible, attempting to bootstrap a prediction of emotions and actions from an individual's observed psychological data has increasingly entered the realm of mainstream accessibility.

Given the comprehensive data provided by psychological assessments, paired with a 21st century capability to produce large prediction models, it has become increasingly feasible to develop predictions for specific human actions and emotions. Of foremost relevance to this is consideration to an individual's drug usage habits. Since commonly the motivation to consume drugs derives from the psychological/physiological influence of the drugs, and conversely, an individual's psychological characteristics influence their respective inclination to drug usage; emotional characteristics and drug usage habits are inherently intertwined metrics.

Ultimately, since usage of some drugs provides the potential of significantly adverse side effects to health, as

well as potentially long-term addiction cycles, it is relevant to consider the possibility of establishing models which may predict an individual's inclination towards addiction or usage of a particular drug. If it could be possible to establish a model which may predict whether an individual possesses a predisposition to usage of a particular drug based on their psychological profile, it would be possible to offer the individual preventative measures or inform them to a greater degree about the concern such that they may better avoid problems resulting from substance abuse.

Methods

The dataset used for the training of the following models was sourced from a research publication which similarly intended to address the feasibility of predicting an individual's drug usage risks (and one which I admittedly did not read prior to or proceeding the creation of my models). The original authors sourced this psychological profile data using an online survey taken by **1885 respondents** which addressed categories including:

- Big Five personality traits

- Neuroticism
- Extraversion
- Openness to Experience
- Agreeableness
- Conscientiousness

Furthermore, metrics regarding individuals levels of **Impulsivity**, **Sensation Seeking**, and other demographic context such as education level, age, gender, country of residence, and ethnicity.

Next, each individual was prompted regarding their usage of 18 different types of drugs: alcohol, amphetamines, amyl-nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron). The fictitious drug category served to identify over-responders, however, for the purposes of the models constructed in this paper, this metric is disregarded. Each individual was asked to categorize their usage of each drug based on the last time which they consumed the drug, if at all. Thus, the 7 categories for each drug are as

follows: never used the drug, used it over a decade ago, used in the last decade, used in last year, used in last month, used in last week, or used in the last day.

Since each respondent is provided questions concerning a wide range of profile data as well as drug usage, this dataset benefits from particularly comprehensive interactions across categories, with the potential for many related predictors, as well as predicting both continuous and categorical targets. For this particular application, I chose to construct 4 different model approaches for predicting a binary factor classifying drug usage for a particular drug as user or non-user. I facilitated this by binarizing the target drug category:

```
24  #Binary target
25  dataPack$Cannabis[dataPack$Cannabis %in% c("CL0","CL1","CL2","CL3","CL4")] <- 0
26  dataPack$Cannabis[dataPack$Cannabis %in% c("CL5","CL6")] <- 1
```

The categorization is thus: **CL0** - **CL4** becomes **0** denoting a **non-user**, as this individual has not consumed the drug within the last week, and **CL5** and **CL6**, denoting consumption within the last week and last day respectively, results in classification as **1** (**drug user**). However, this binary categorization is only applied to the target variable, thereby maintaining the extra

context provided by the continuous nature of the drug usage data, while allowing for a dichotomous outcome prediction. For the 4 models created in this paper, 1 targets **alcohol** use, and the other 3 target **cannabis** use.

For all other drug usage data not including the target, I re-assigned the categorical identifiers as continuous 0 through 6:

```
#Global Replacement for CL identifiers
dataPack[dataPack == "CL0"] <- 0
dataPack[dataPack == "CL1"] <- 1
dataPack[dataPack == "CL2"] <- 2
dataPack[dataPack == "CL3"] <- 3
dataPack[dataPack == "CL4"] <- 4
dataPack[dataPack == "CL5"] <- 5
dataPack[dataPack == "CL6"] <- 6
```

Furthermore, and **important to this dataset**, the authors published the data entirely in a quantified format, this is *interesting* due to the nature of the entirety of the original data being *categorical*. What this provides is an interesting dichotomy of options with how to approach classifying and considering the data, of either treating all the data as “real,” or reclassifying all of the data back into its categorical format. I did both of these approaches; the second method took *way* longer and I **did not** end up using any of the models I made with that approach.

Going forward with treating all the quantified values as real, rather than re-categorizing, and after classifying all drug usage responses as numeric, it is then possible to establish the entire dataset as numeric such that all of the available data and metrics can be standardized and considered across the models.

```
#Fix any possible issues with the canna-binari-zation
dataPack[c("Alcohol", "Amphetamines", "Amyl-Nitrite",
           "Benzodiazepine","Caffeine", "Cannabis",
           "Chocolate", "Cocaine", "Crack", "Ecstasy",
           "Heroin", "Ketamine", "Legal Highs", "LSD",
           "Methadone", "Mushrooms", "Nicotine", "Fictitious Drug",
           "Volatile Substance Abuse")] <- lapply(dataPack[c("Alcohol", "Amphetamines", "Amyl-Nitrite",
                                                            "Benzodiazepine","Caffeine", "Cannabis",
                                                            "Chocolate", "Cocaine", "Crack", "Ecstasy",
                                                            "Heroin", "Ketamine", "Legal Highs", "LSD",
                                                            "Methadone", "Mushrooms", "Nicotine", "Fictitious Drug",
                                                            "Volatile Substance Abuse")], as.numeric)
```

Converting all values to numeric, as well as both the binary classification of the target variable and the modification of drug usage categories into a continuous response metric was conducted prior to data splitting. Next, since all 4 models are attempting to predict a binary, dichotomous outcome of user or non-user, the target variable is re-assigned as a factor to allow for compatibility with model training functions.

```
#Target variable as factor
data_train$Cannabis <- factor(data_train$Cannabis, levels = c(1,0), labels = c("Yes","No"))
data_test$Cannabis  <- factor(data_test$Cannabis,  levels = c(1,0), labels = c("Yes","No"))
```

After this stage in the process, the approach to training the models diverge. The four models I chose to create include 2

Elastic Net models attempting to predict alcohol user from non-alcohol user, and cannabis user from cannabis non-user respectively, as well as 1 **Lasso Regression** model and 1 **Multiple Linear Regression** model attempting to predict cannabis user from non-user. The reason that I ultimately decided to focus on developing Elastic Net models is due to the nature of it being one of the more adaptive regression methods we have discussed so far, as well as its strength in addressing highly co-related features in the data.

Furthermore, Elastic Net methods can also perform well with data that suffers from multicollinearity, a likely problem with many of the emotion based metrics in this data. Ultimately, by combining the benefits of both the Lasso and Ridge approaches, Elastic Net provided the most successful model performance out of the various approaches that I tested. I chose to construct one of the Elastic Net models to try and predict alcohol use, and the other to predict cannabis use, as these categories provided widely varied prediction outcomes due to a **large difference** in how frequently reported the use of alcohol was over cannabis amongst the data population (82% of the sample

population reported regularly consuming alcohol). Since alcohol did prove to be a less reliable prediction category overall, I chose to conduct the 3rd Lasso regression method targeting cannabis.

The recipe stage that I implemented for the regression models is as follows:

```
#Recipe
blueprint <- recipe(Alcohol ~ ., data = data_train) %>%
  step_impute_median(all_numeric_predictors()) %>% #I likely don't need this, no missing data
  step_impute_mode(all_nominal_predictors()) %>% #I likely don't need this, no missing data
  step_zv(all_predictors()) %>% #Remove zero variance features
  step_corr(all_numeric_predictors(), threshold = 0.9) %>% #Remove highly correlated features
  step_YeoJohnson(all_numeric_predictors()) %>% #Apply normal distribution to data
  step_dummy(all_nominal_predictors(), one_hot = TRUE) %>% #This might not be necessary but it makes sure numeric
  step_center(all_numeric_predictors()) %>% #Standardize data
  step_scale(all_numeric_predictors()) %>% #Standardize data
  step_interact(
    terms = ~ Neuroticism:Extraversion +
      Impulsivity:`Sensation Seeking` +
      Conscientiousness:Impulsivity +
      Age:starts_with("Gender_")
  ) %>% #Create interaction features
  step_pca(all_numeric_predictors(), threshold = 0.95) #Attempts to combine colinearity
```

The first two steps in the recipe are truthfully not necessary in this particular approach, as the original data actually happens to not contain any missing data at all. The next step involves removing any zero variance features from the data, involving data columns providing little variance and thus minimal predictive value. The next step attempts to reduce multicolinearity by removing any highly correlated features, followed by a step which attempts to normally distribute the data. The dummy encoding step modifies any categorical features,

of which the data should not have at this stage, and thus this step *should* be redundant. The next two steps for center and scale are modifiers which further standardize/normalize the scale of the numeric values in the data. The **Step Interact** preprocessing stage serves to establish new combination features from conditional relationships that may exist in the data. Partially, this step also involves some interaction with the dummy coding, with which I am uncertain the extent of how involved it is in the final process. The inclusion of conditional feature combinations does seem to improve model results across tests. The last step attempts to prevent multicolinearity by further combining any highly related features.

Furthermore, a *very* time consuming process involved determining why it consistently proved to be so difficult to accurately predict alcohol usage. Ultimately, I came to the realization that alcohol is a particularly difficult classification, as alcohol usage is so prevalent amongst the general population and overlaps with most other predictors. A last attempt to derive some extra accuracy from the alcohol model was through **SMOTE** (Synthetic Minority Oversampling

Technique) sampling to attempt to balance the ratio of drinker to non-drinker within the training data (not on testing data to avoid data leakage). I am not able to provide any directly quantified metrics on if this improved outcomes, but it is none-the-less accompanying us on our thermally-inefficient journey limit testing a 10 year old CPU.

To evaluate model performance I used a confusion matrix to derive descriptive statistics about the models performance, and compared the models across their metrics of accuracy, sensitivity, specificity, precision, kappa, and P-Value.

Simon K.
10-10-25
PSYC-559

Results

Model 1: Lasso Regression Targeting Cannabis Use

```
          Reference
Prediction Yes  No
       Yes 161  83
       No   34 288

               Accuracy : 0.7933
                 95% CI : (0.7575, 0.8259)
    No Information Rate : 0.6555
    P-Value [Acc > NIR] : 4.524e-13

                  Kappa : 0.5681

 Mcnemar's Test P-Value : 9.097e-06

            Sensitivity : 0.8256
            Specificity : 0.7763
         Pos Pred Value : 0.6598
         Neg Pred Value : 0.8944
             Prevalence : 0.3445
         Detection Rate : 0.2845
   Detection Prevalence : 0.4311
      Balanced Accuracy : 0.8010

       'Positive' Class : Yes
```

This model predicts cannabis usage **successfully** at a rate of 0.7933, or about **79%**. Furthermore, the model maintained a sensitivity level at 0.825 or **83%**, denoting correct positive predictions when the actual value is positive. Similarly, with consideration to specificity, the model correctly identified **78%** of true negative values. With regard to positive predicted and negative predicted values at **66%** and **89%** respectively, there does seem to be an imbalance in the direction that the model accuracy biases, such that it makes more mistakes when classifying false positives than when classifying false

negatives. The Kappa value of **0.57** is sub-optimal but still greater than random chance. Ultimately, the resulting P-Values denote significance, and the high sensitivity and accuracy suggest that this model does perform well at predicting cannabis usage given the individuals psychological profile data.

Model 2: Elastic Net Regression Targeting Cannabis Use

```
          Reference
Prediction Yes  No
       Yes 162  85
       No   33 286

               Accuracy : 0.7915
                 95% CI : (0.7557, 0.8243)
    No Information Rate : 0.6555
    P-Value [Acc > NIR] : 9.141e-13

                  Kappa : 0.5659

 Mcnemar's Test P-Value : 2.667e-06

            Sensitivity : 0.8308
            Specificity : 0.7709
         Pos Pred Value : 0.6559
         Neg Pred Value : 0.8966
             Prevalence : 0.3445
         Detection Rate : 0.2862
   Detection Prevalence : 0.4364
      Balanced Accuracy : 0.8008

       'Positive' Class : Yes
```

Interestingly, this model appears to perform extremely similarly to the level of accuracy derived from the Lasso regression. It is interesting because this model took over 10 minutes to train over each iteration, and the Lasso model took about 20 seconds each time. It required a significant amount of actual time to train the different versions of this Elastic Net regression model, just for it to actually underperform the

accuracy of the Lasso model by **0.18%** and achieve almost identical results across the other metrics, with a slight improvement in the model sensitivity.

Model 3: Elastic Net Regression Targeting Alcohol Use

```
          Reference
Prediction Yes  No
       Yes 234  86
       No  145 101

               Accuracy : 0.5919
                 95% CI : (0.5501, 0.6327)
    No Information Rate : 0.6696
    P-Value [Acc > NIR] : 0.9999545

                  Kappa : 0.1459

 Mcnemar's Test P-Value : 0.0001356

            Sensitivity : 0.6174
            Specificity : 0.5401
         Pos Pred Value : 0.7312
         Neg Pred Value : 0.4106
             Prevalence : 0.6696
         Detection Rate : 0.4134
   Detection Prevalence : 0.5654
      Balanced Accuracy : 0.5788

       'Positive' Class : Yes
```

This model produces an accuracy rate of 0.59, or correctly predicts an alcohol user from a non-user **59%** of the time. This initially seems as though it would be slightly better than chance, given that the value is greater than 50%, however, since the alcohol data appears so widely across the survey respondents, and thus the yes category makes up a disproportionate amount of the data entries, a rate of 59% actually performs worse than if the model had simply predicted yes for all possible outcomes.

The only true strength that this model has when attempting to predict the alcohol user category is in the yes predictions, of which it quickly becomes apparent this is again a result of the yes category making up a disproportionate amount of the results. Ultimately, the Kappa of **0.145** denotes that this model does not provide much of actual value beyond guessing at random. The issue is not necessarily the process used to train this model, this model performed poorly due to the alcohol category suffering from significant imbalance in the data.

Model 4: Multiple Linear Regression Targeting Cannabis Use

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.9177 -0.3407 -0.1465  0.4507  1.0658

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      0.34335    0.01200  28.616  < 2e-16 ***
Extraversion                    -0.05159    0.01248  -4.132 3.82e-05 ***
`\\`Sensation Seeking\\``        0.13806    0.01703   8.107 1.18e-15 ***
`\\`Openness To Experience\\``   0.11290    0.01366   8.266 3.36e-16 ***
Impulsivity                     -0.01517    0.01613  -0.941    0.347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4354 on 1313 degrees of freedom
Multiple R-squared:  0.1628,    Adjusted R-squared:  0.1602
F-statistic: 63.81 on 4 and 1313 DF,  p-value: < 2.2e-16
```

This simple linear regression attempts to view the agreement between specifically the Cannabis target and a few of the main predicting categories. The results are not particularly impressive, however they did denote significance. The model explains 16% of the variance in target data. The results suggest that this predictive model is expectedly not as predictively strong as the prior models, however, it is still statistically significant, which could be expected from a linear regression model with only a few predictors.

Discussion

The four models performed both in ways I expected that they likely would, and also in ways that I had not considered. Functionally, the alcohol prediction model caused me some of the most issues throughout constructing and testing methods. This is due to the unique nature of the alcohol category lacking specificity due to the overencompassing nature of the appearance of positive values for reported alcohol usage in the data. Thus, despite attempting to filter and modify the data in multiple different ways using multiple different feature engineering

approaches, the alcohol data remained nearly impossible to predict for. Unfortunately, this took me a great deal of time to recognize.

However, the relative lack of performance provided by the model attempting to predict alcohol use does exist as a relative standard that could be used to compare with the more functional models. Likewise, the fact that the results deviate so significantly despite being trained on effectively the same parameters is actually an interesting result in of itself. For this reason, I chose to keep it as the third model even despite the target variable not being the same as the other three models targeting cannabis use. The implication, then, is that there exists social standards that would allow for one psychoactive substance to be so widely accepted that it flattens data and prevents meaningful predictions, and another to be accepted to a degree so much less so that a regression model can accurately predict an individual's probability of using the substance based on data regarding their psychology. Furthermore, attempting to implement feature engineering methods to try and optimize the alcohol category for predictions provided valuable experience

learning how and when to implement other forms of normalizing and sampling to address issues such as balance and skewness of the data.

The results from the Elastic Net model predicting cannabis use, and the Lasso model predicting cannabis use denote that both performed well, which I somewhat expected. I likewise expected that both methods would likely provide at least improved results from that of predicting alcohol, simply because the cannabis use category is much less widely reported across the sample population. The aspect that surprised me was the significant difference in compute requirement to train the Elastic Net model versus that of the Lasso model, despite providing results that overlap heavily. I suppose that the lasso model is able to establish a fairly accurate predictive ability within the shorter duration of training folds, and the significantly extra training instances within the Elastic Net method end up being mostly redundant and don’t provide the model much more prediction capability beyond that which the Lasso model already achieved. Fundamentally, the most effective model out of the three by far was the Lasso model, as it achieved

effectively the same level of accuracy and specificity as the Elastic Net model, but within a fraction of the same compute requirement.

I am ultimately impressed with the capability that even these relatively basic models and training methods are able to produce. Without significant computational requirements, it is possible to predict at an >80% rate of success if somebody does or does not consume a particular drug, based on individual profile data. While predicting if somebody smokes cannabis or not is relatively low as far as actual relevant value, some of the other drugs within this same dataset provide implications that can quickly be understood as a far more important prediction. The ability to predict whether an individual may have a greater probability, or even a pre-disposition for using an illicit or dangerous substance that could cause them harm could be incredibly beneficial for a great deal of people if utilized optimally. The ability to possibly predict and intervene before an individual becomes trapped in a cycle of addiction could most certainly concern the life or death of that individual; an implication which is clearly much more relevant.

However, this discussion and the models which have been investigated in this paper are certainly far from the level of complexity which would be required for establishing frameworks to attempt to save lives and prevent addiction, and likewise there are most certainly a great deal of data scientists around the world attempting to train models to do just that. While the scale and complexity of these models pales in comparison to that of the big data models trained in data centers, the concept is the same. Ultimately, the future of this area of research will involve continued development of predictive models based on the data provided through psychological testing, until the day when research may allow for analysis of specific neural pathways.

If there is a fundamental concept which can underpin the results that we see from our attempts at predicting these components of the individual's psychology, it is that humans operate through a great deal of overlapping and confounding social patterns, and given the measurement of an effectively large combination of these psychological attributes, generalizing most probable actions based on social patterns is often feasible.

Simon K.
10-10-25
PSYC-559

References

Fehrman, E., Egan, V., & Mirkes, E. (2015). Drug Consumption (Quantified) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TC7S.