Predicting Cannabis Consumption from Demographics and Personality

HarvardX PH125.9x Data Science Capstone

Charles Mégnin

10/12/2019

Contents

Executive Summary

+ Introduction

Drug use is a behavior that constitutes an important factor linked to poor health, including early mortality, and which presents significant adverse consequences for the social fabric, notably with respect to criminality and family cohesion. Early detection of an individual's predisposition to drug consumption offers healthcare professionals an opportunity to short-circuit the onset of addiction.

The present study is based on a dataset that includes demographic and psychological information related to the consumption of 18 legal and illegal drugs by 1885 participants. For the purpose of this study, we choose to focus the data analysis and modeling on the use of cannabis.

+ Goal of project

The goal of this project is to assess whether an individual's consumption of cannabis can be predicted from a combination of demographic and personality data.

To do so, we build and assess the effectiveness of six machine learning classifiers and confront the results obtained with the insights provided by data exploration.

+ Dataset description

The dataset used here is found on the UCI machine learning repository. It is based the research paper by 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.," arXiv, 2015. The data was collected from 1885 English-speaking participants over 18 years of age between March 2011 and March 2012.

The original dataset includes answers to questions related to the use of alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms, nicotine and volatile substance abuse (VSA)) and one fictitious drug (Semeron) which was introduced to identify over-claimers. In the present study, we restrict our scope to the analysis of cannabis consumption.

In the original dataset, drug use is separated between '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'. For the purpose of this study, we separate the data in two groups: 'Non-Users' which is a combination of 'Never used', 'Used over a decade ago', 'Used in last decade', and 'Used' (the combination of the others, concisting of 'Used in last year', 'Used in last month', 'Used in last week' and 'Used in last day').

We create a classification that distinguishes those that have either never used cannabis or used it a decade ago or more and which we refer to as 'Non-users', from those that have used it more recently which we call 'Users'. This nomenclature is used for convenience, not for its negative connotation. We chose to bin the data this way as we find the ten-year mark to be a reasonable dividing line between those with no interest in cannabis use from those who occasionally or regularly consume it.

The features in the data can be separated in two groups of pre-normalized and centered predictors:

1. Five demographic predictors : Age, Gender, Level of education, Ethnicity, and Country of origin.
2. The results from seven scored tests administered to assess personality, specifically:

- Neuroticism (a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression);
- Extraversion (manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics);
- Openness to experience (a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests);
- Agreeableness (a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness);
- Conscientiousness (a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient);
- Impulsiveness;
- Sensation-seeking.

The working dataset in this study consists therefore of one Class (Cannabis consumption) and twelve predictors (five demographic and seven personality-related).

+ Key steps

We extract a training subset (80% of data) from the dataset for the purpose of training our model, and use the remaining 20% of the data as a test set for the purpose of evaluation. This being a classification problem, we use accuracy as the metric to assess the goodness of fit.

The analysis consists of two main sections:

1. In the first part, after performing minor data engineering (A), we bin, explore, and analyze the dataset (B).
2. In the second part, we move on to the modeling phase (C):

- After a 3-step preprocessing consisting of examining correlation among predictors, seeking low-variance factors and applying a Recursive Feature Elimination algorithm to seek and potentially discard predictors that do not contribute significantly to the outcome, we build models based on the following six popular machine learning methods:
  - Generalized linear model (glm)
  - Generalized linear model with penalized maximum likelihood (GLMnet)
  - Decision tree (rpart)
  - Random forest (rf)
  - Stochastic gradient boosting (gbm)
  - Neural network (nnet)

We compare the modeling approaches, both in terms of accuracy and coherence with the data analysis.

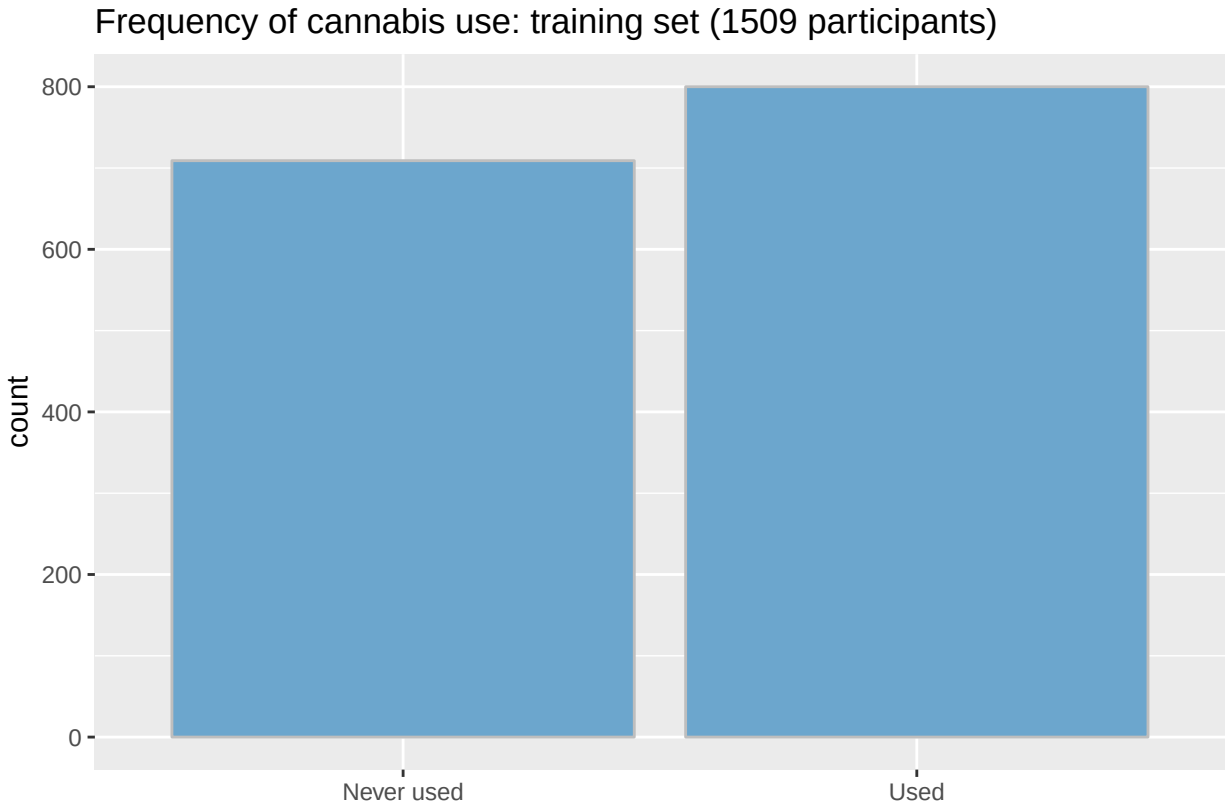
Analysis

A: Data engineering

- All predictors were already normalized and centered in the original dataset;
- We construct the 'Used' class to separate 'Non users' from 'Users';
- We partition the data between training (80% - 1509 participants) and test sets (20% - 376 participants), preserving the distribution of the Cannabis class;
- There are 0 NAs in the dataset as a whole (no interpolation or imputation to perform)

B: Data exploration

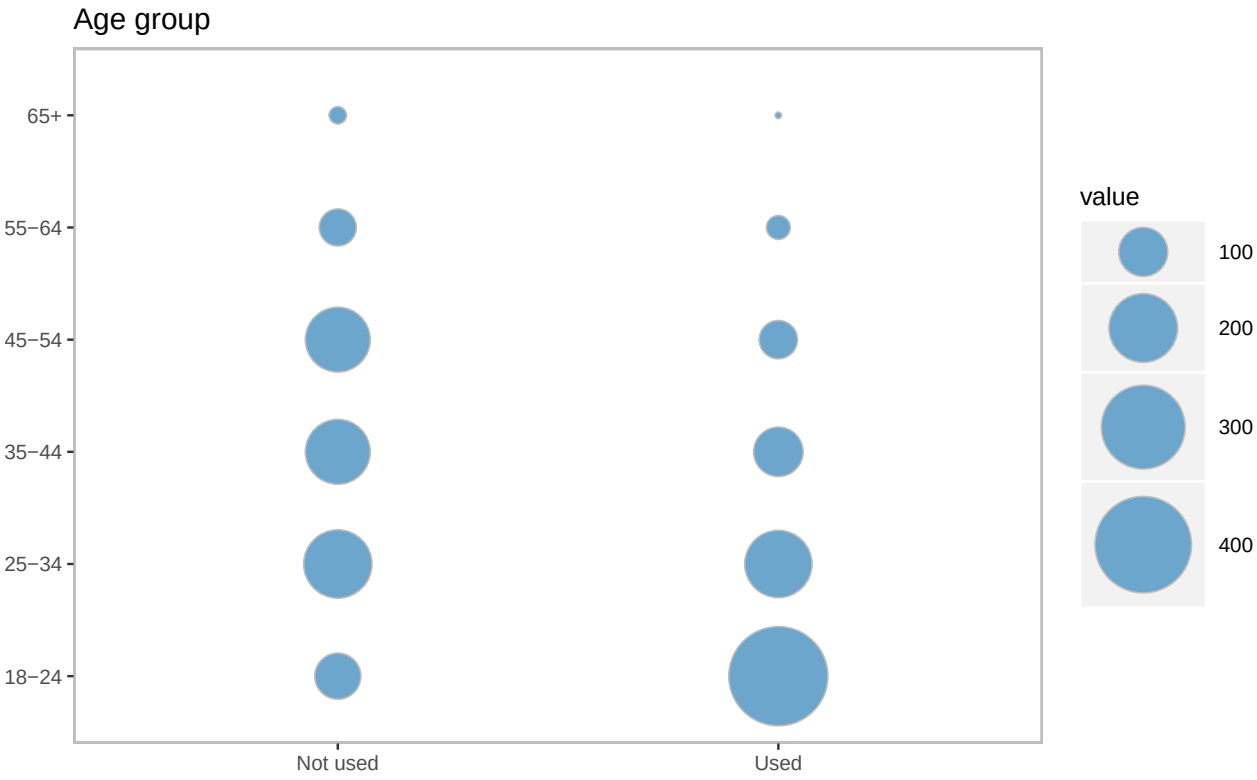
Class distribution

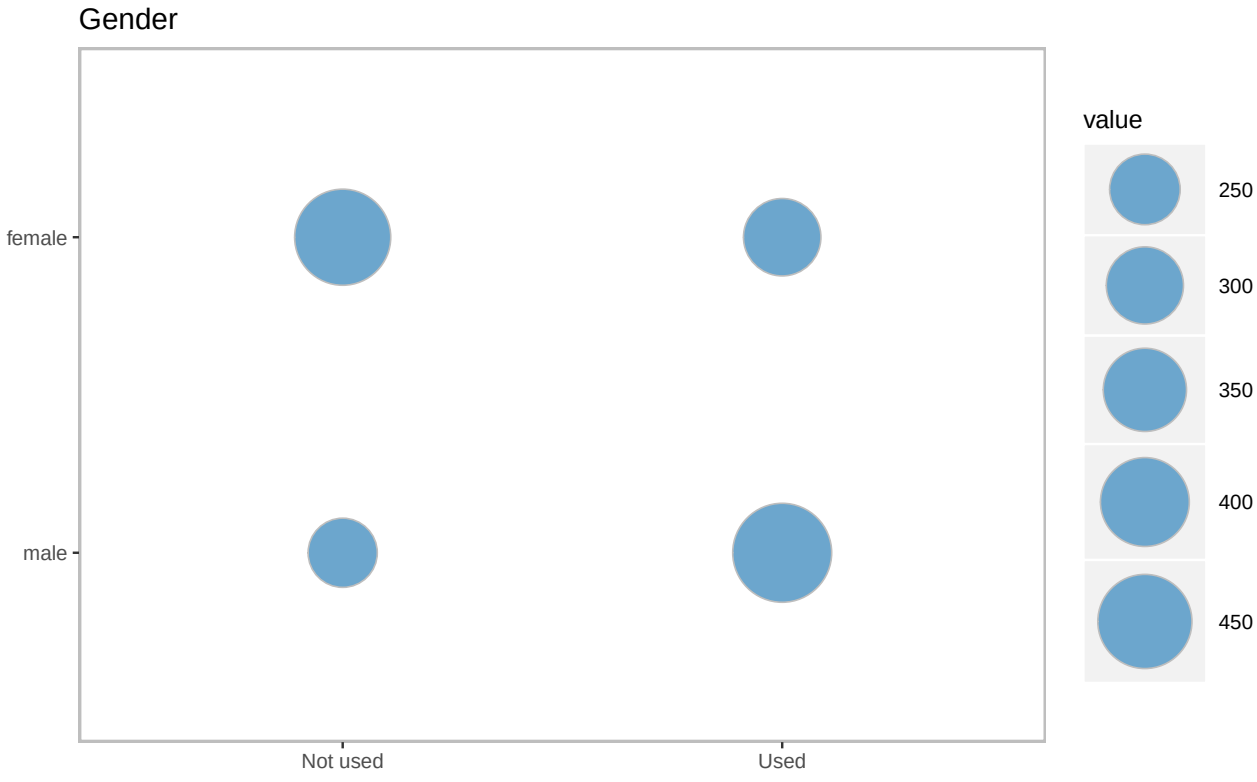


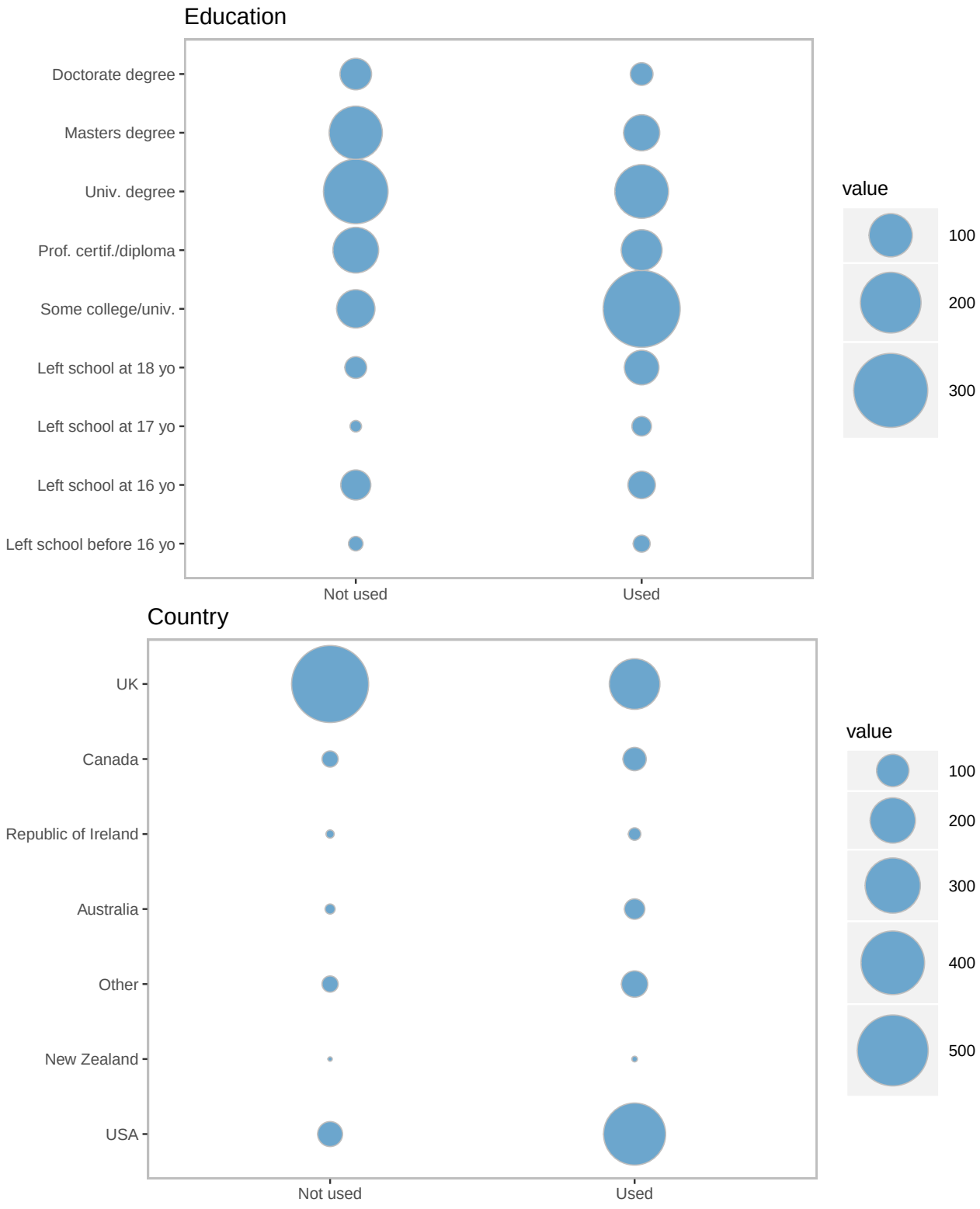
The training set of 1509 participants consists of 800 participants (53.0%) having used cannabis and 709 who never have (47.0%), for a user-to-non-user ratio of 1:1.1

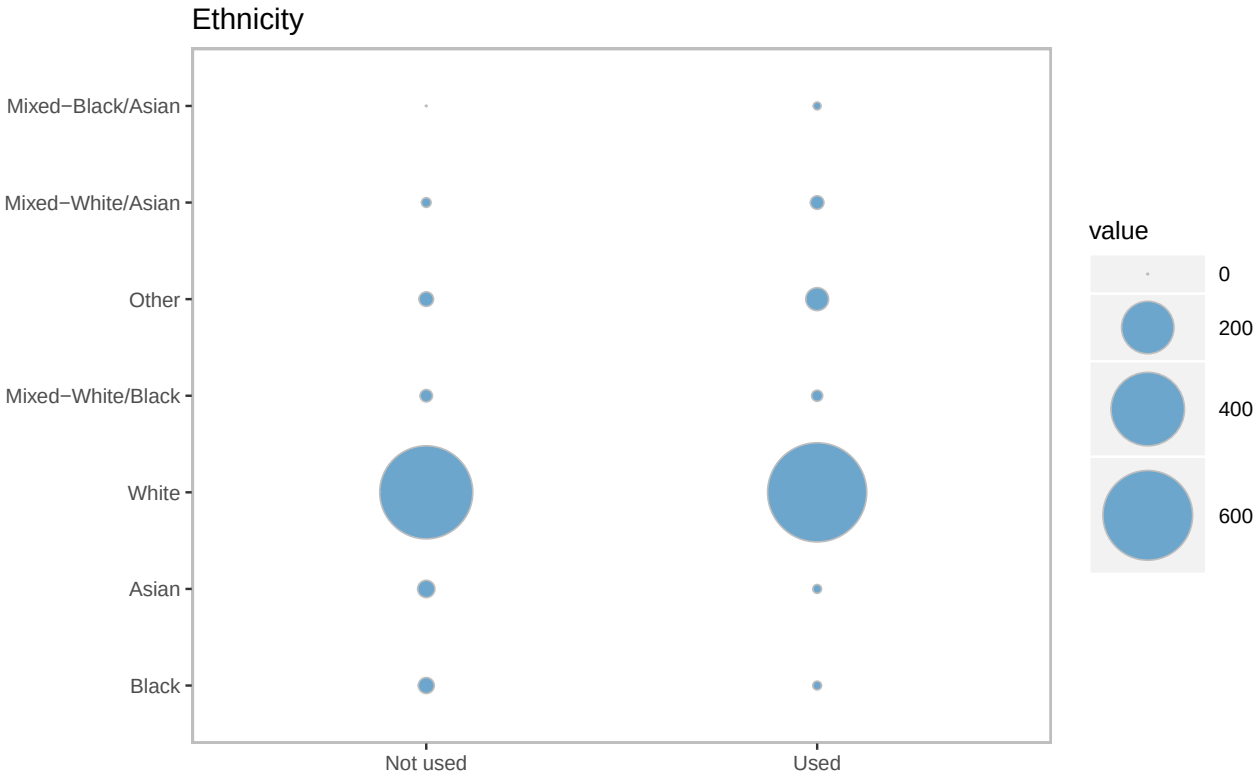
Contingency plots (prior to binning)

Cannabis use by demographic group


Age group
65+
55−64
45−54
35−44
25−34
18−24
Not used
Used
value
100
200
300
400


Gender
female
male
Not used
Used
value
250
300
350
400
450


Education
Doctorate degree
Masters degree
Univ. degree
Prof. certif./diploma
Some college/univ.
Left school at 18 yo
Left school at 17 yo
Left school at 16 yo
Left school before 16 yo
Not used
Used
value
100
200
300
Country
UK
Canada
Republic of Ireland
Australia
Other
New Zealand
USA
Not used
Used
value
100
200
300
400
500



The dataset is dominated by young and educated white American and British participants of both sexes.

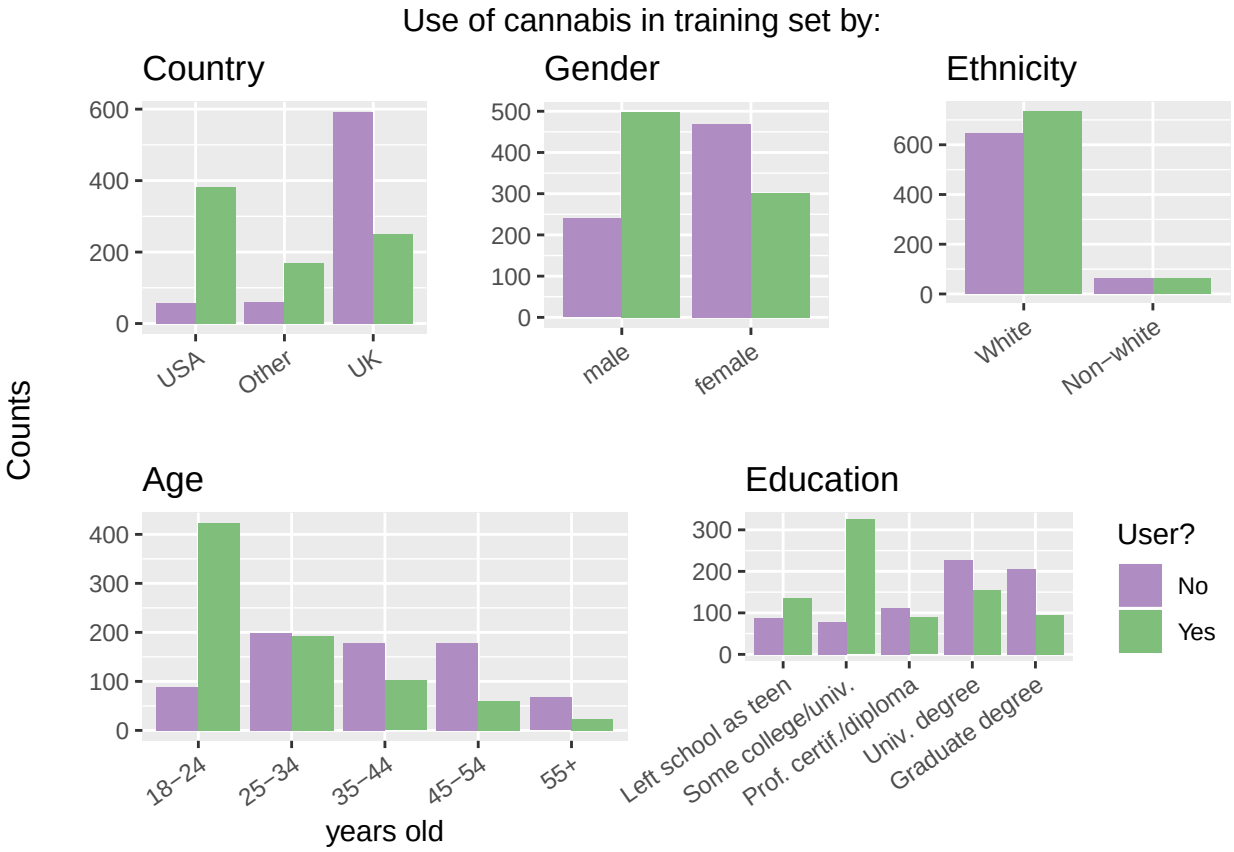
Binning

The small size of many demographic sub-groups add little valuable insight and will likely only serve to introduce variance in the analysis. At the risk of erasing behavioral differences among groups, the distribution of the dataset forces a more meaningful binning of the demographic information. In particular and in addition to the small individual population sizes, we can think of no rational reason to distinguish among those that left school before 16, at 16, at 17 or at 18 and lump these in a "Left school as a teen" group.

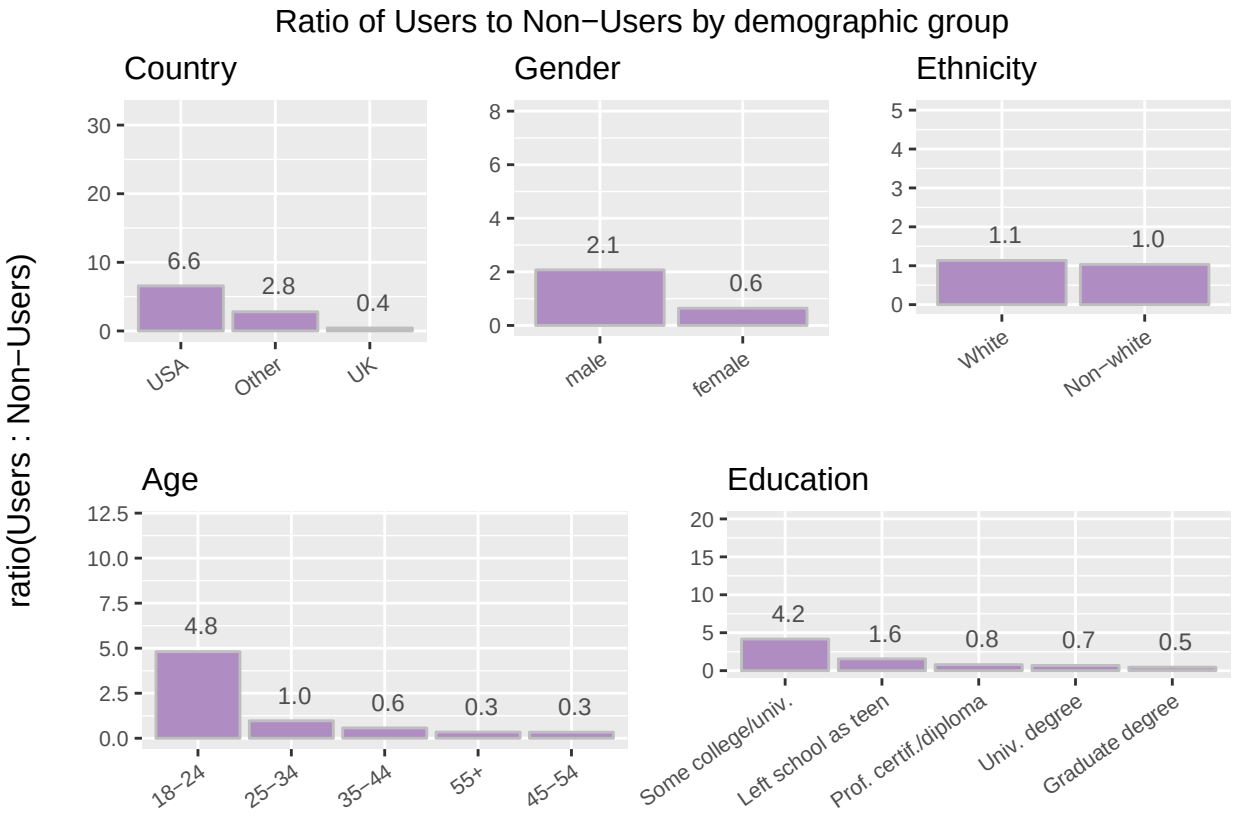
Visual inspection of the balloon plots above provide a straightforward path to binning the data as:

- 5 age groups: "18-24", "25-34", "35-44", "45-54", "55+"
- 5 groups for Education: "Left school as a teen", "Some college", "Professional certificate", "University degree", "Masters degree" and "Doctorate".
- 3 groups for Country: "USA", "UK", "Others".
- 2 ethnic groups: "Whites", "Non-whites".

Analysis of demographics



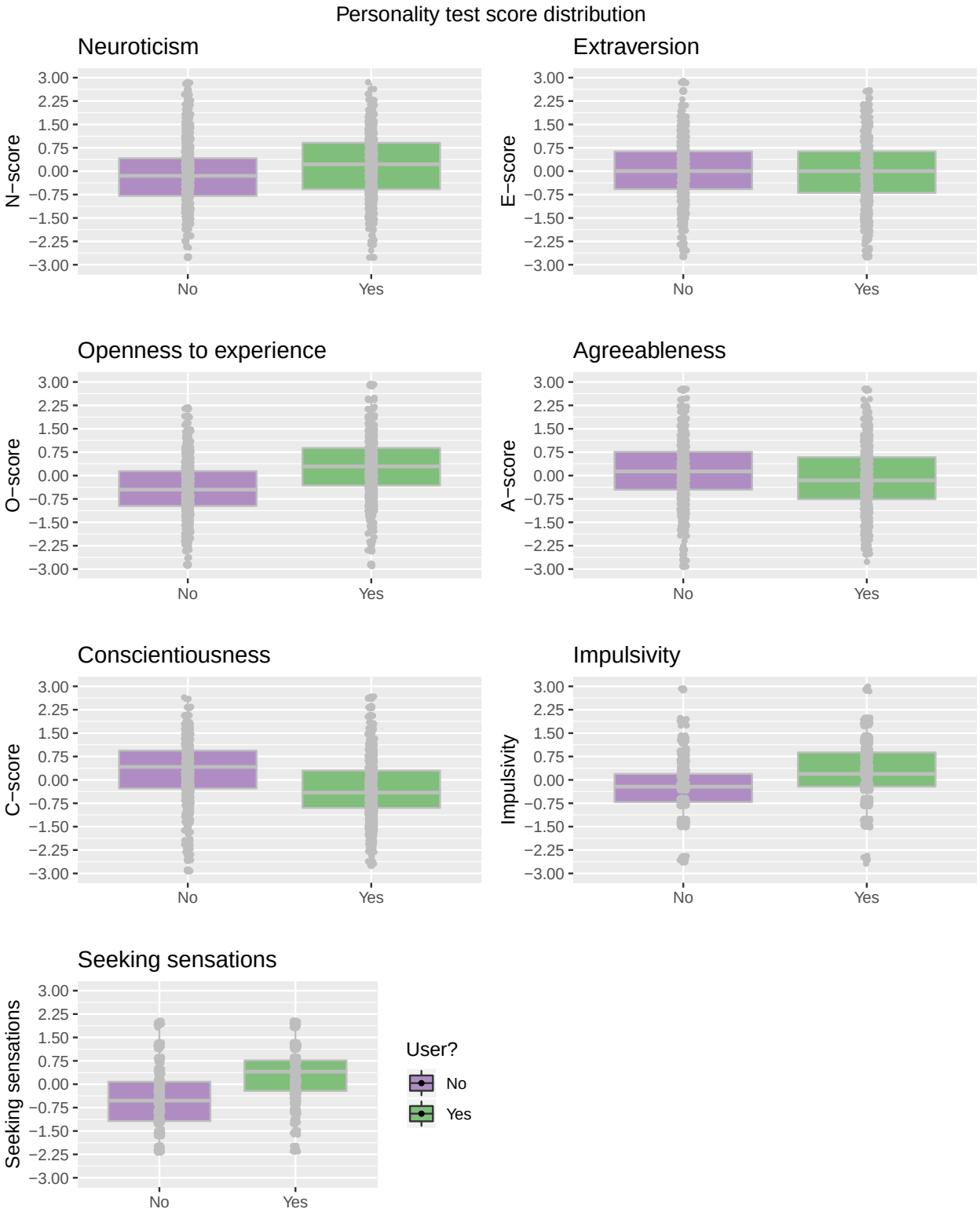
Users outnumber non-users in many demographic sub-group, particularly for men, in the USA, among 18-24 year olds and for those that didn't complete a college degree. In the modeling phase, we therefore expect the corresponding predictors (country, age group, education and gender) to have significant weight.

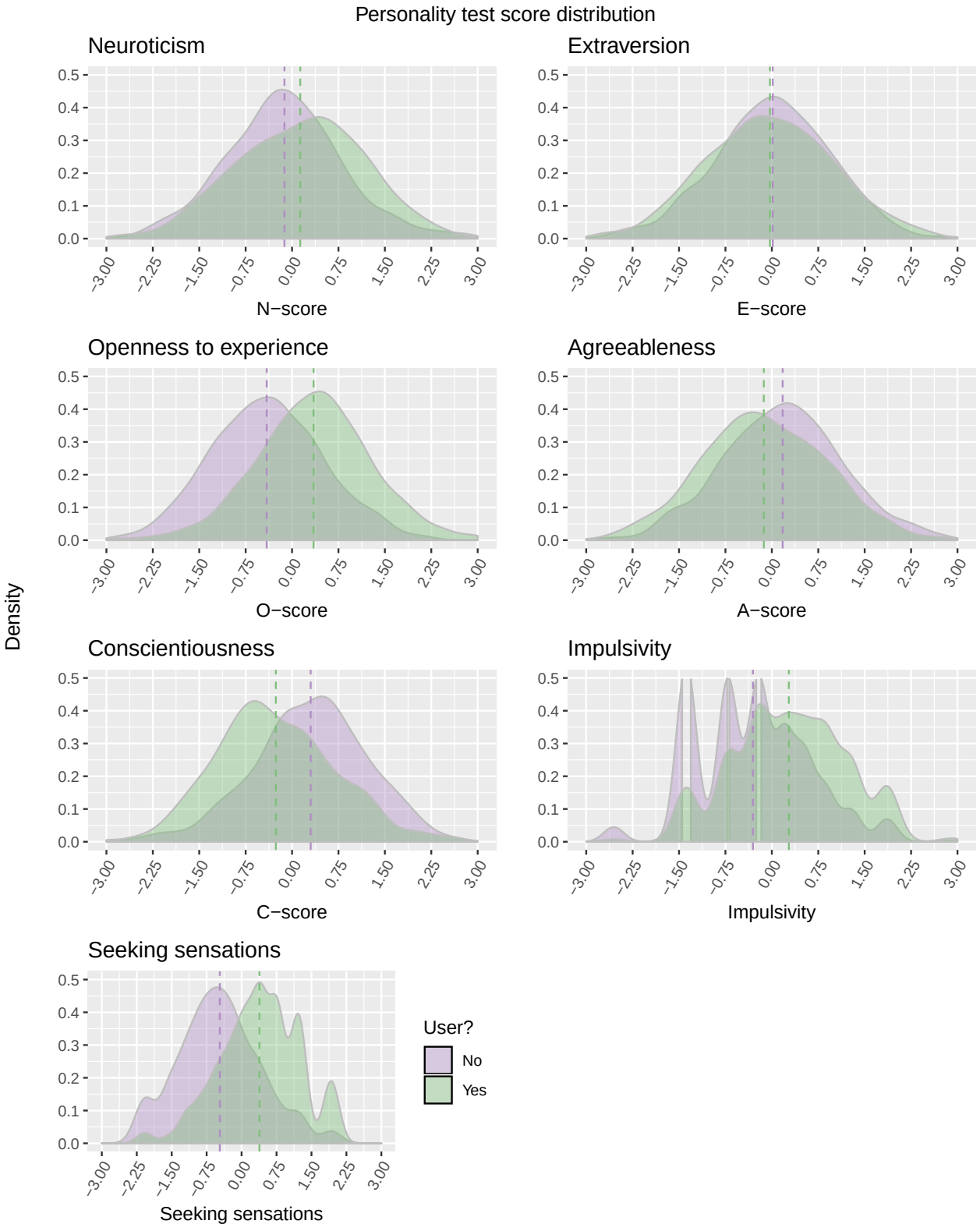


For better readability of the differences among them, we use the ratio of users to non-users as index of consumption for each demographic sub-group.

- Americans consume more cannabis (6.6) than all other nationalities, and particularly the British (0.4) who, along with females (0.6), people over 34 years of age and people with degrees, are among the lowest using group with more non-consumers than consumers.
- Men consume more (2.1) than women (0.6)
- Whites consume about the same as other ethnic groups (1.1) and we do not expect ethnicity to be a significant factor. (1.0)
- Those having only had some college/university are the highest users in the education sub-group (4.2)
- Finally, we observe a steady decline of cannabis use with age from 18-24 year olds (4.8) to participants over 55 (0.3) This points to a generational phenomenon (for instance, roughly five times as many 25-34 year-olds abstain when compared to 18-24 years olds, even though they have had ten additional years to experiment).

Personality analysis





For each test-score distribution, after checking for normality (Shapiro-Wilk test) we test the hypothesis that there is no difference between Users and Non-Users. We use the Student t-test when the data is normally distributed and the Mann-Whitney-Wilcoxon test otherwise.

- Neuroticism: Shapiro p-values = [Not Used: 0.00, Used: 0.04] < 0.05: the data are not normally

distributed. Wilcox p-value = 0.00 < 0.05: the 'User' and 'Non-User' population means are not identical.

- Extraversion: Shapiro p-values = [Not Used: 0.08, Used: 0.27] > 0.05: the data are normally distributed. t-test p-value = 0.37 > 0.05: the 'User' and 'Non-User' population means are identical. For the variance, t-test p-value = 0.37 > 0.05: the 'User' and 'Non-User' population variances are also identical.
- Openness to experience: Shapiro p-values = [Not Used: 0.11, Used: 0.00] > 0.05: users are not normally distributed (but non-users are). Wilcox p-value = 0.00 < 0.05: the 'User' and 'Non-User' population means are not identical.
- Agreeableness: Shapiro p-values = [Not Used: 0.12, Used: 0.19] > 0.05: the data are normally distributed. t-test p-value = 0.00 < 0.05: the 'User' and 'Non-User' population means are not identical.
- Conscientiousness: Shapiro p-values = [Not Used: 0.00, Used: 0.00] < 0.05: the data are not normally distributed. Wilcox p-value = 0.00 < 0.05: the 'User' and 'Non-User' population means are not identical.
- Impulsivity: Shapiro p-values = [Not Used: 0.00, Used: 0.00] < 0.05: the data are not normally distributed. Wilcox p-value = 0.00 < 0.05: the 'User' and 'Non-User' population means are not identical.
- Seeking sensations: Shapiro p-values = [Not Used: 0.00, Used: 0.00] < 0.05: the data are not normally distributed. Wilcox p-value = 0.00 < 0.05: the 'User' and 'Non-User' population means are not identical.

In summary:

Feature	p.value	User_NonUser
Neuroticism	0.000	Different
Extraversion (means)	0.371	Identical
Extraversion (variances)	0.374	Identical
Openness to experience	0.000	Different
Agreeableness	0.000	Different
Conscientiousness	0.000	Different
Impulsivity	0.000	Different
Sensation-seeking	0.000	Different

While significant overlap is observed, most personality-related density plots show significant differences in the mean between users and non-users. particularly as it relates to openness to experience, agreeableness, conscientiousness, impulsivity, and seeking sensations. The observations are by and large consistent with intuition when it comes to openness to experience, impulsivity and seeking-sensations.

Some implications are rather entertaining, notably the notion that nice (Agreeable) people may be less likely to smoke cannabis, or conversely that exposure to pot might make people less nice. Likewise, either conscientious people tend to not use cannabis, or cannabis smoking tends to make people less meticulous.

On the other hand, the distribution means and variances for users and non-users as they relate to extraversion are statistically identical, suggesting that this personality trait may not impact cannabis consumption. The modeling section below examines whether this observation is consistent with the results derived from machine learning.

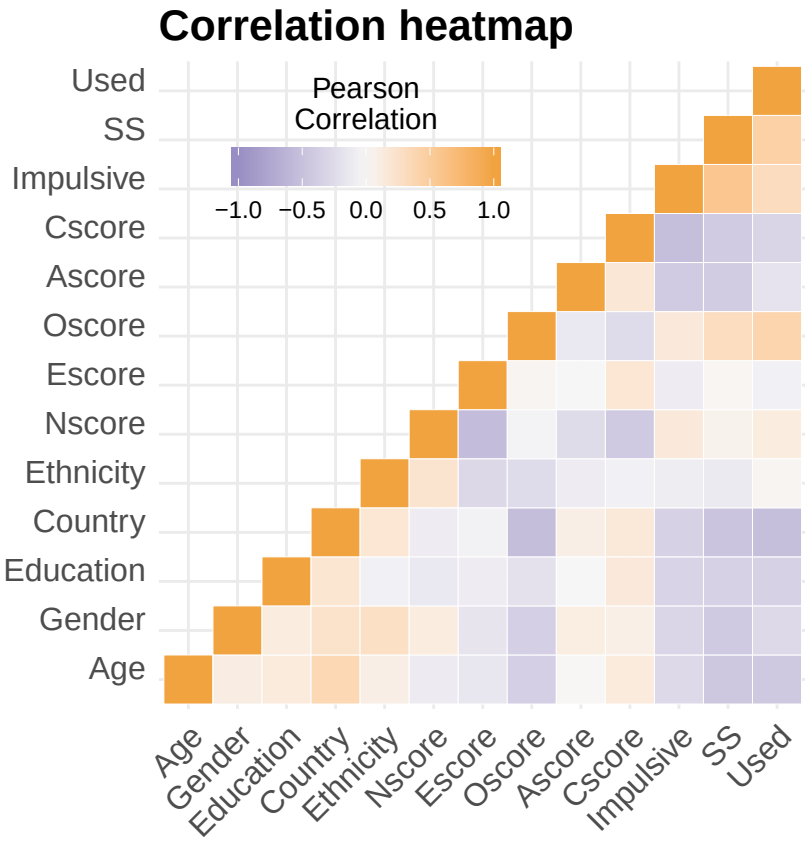
C: Modeling

We seek a model which improves on the ratio of users to the population (53.0%). This naive approach constitutes the baseline above which predictive modeling is interesting.

Pre-processing

Feature correlation

We examine the most contributing cells to the total Chi-square score redundancies among the 12 predictors and with the Used class by examining the Pearson χ^2 residuals.



With a maximum correlation of 0.57 none of the features are strongly correlated among each other or with the used class.

Besides the goodness of fit to the test data, the demographic and personality-related observations above will guide the assessment of the models we derive.

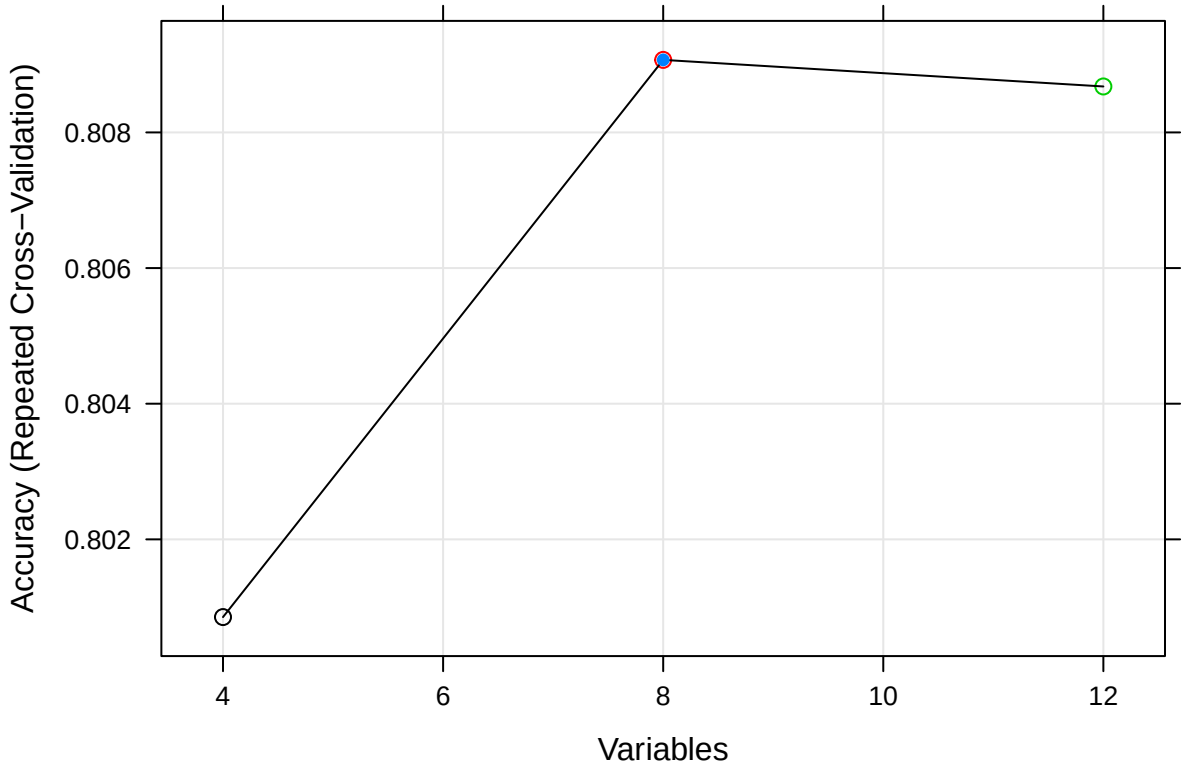
Low-variance analysis

There are 0 variables that meet the low-variance removal threshold.

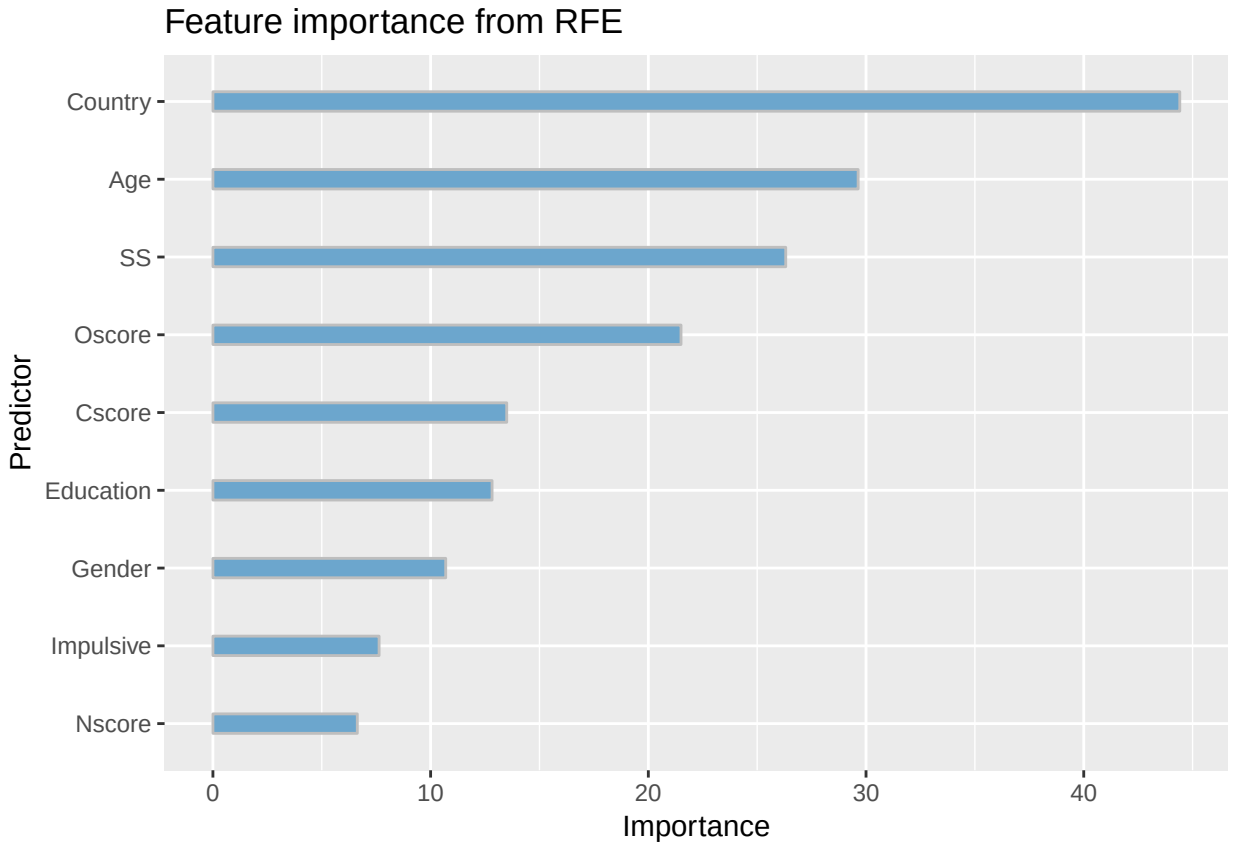
Recursive Feature elimination

For RFE as well as subsequent modeling, we use the k-fold cross validation method which involves splitting the dataset into k subsets. The algorithm holds aside one of the subsets while the model is trained on the

others. This process is repeated a predetermined number of times and the overall accuracy estimate is provided.



After the RFE, we retain 8 features: Country, Age, SS, Oscore, Cscore, Education, Gender, Impulsive



The comparative analysis of the contribution of each feature agrees by and large with that of the density distribution: among the personality trait tests, the E-score contributes the least, as expected from the results of the t-test aboce. Seeking sensation and O-score contribute the most, as expected. We did not expect ethnicity to be much of a contributor and, although not eliminated by the RFE, this factor is by far the least significant.

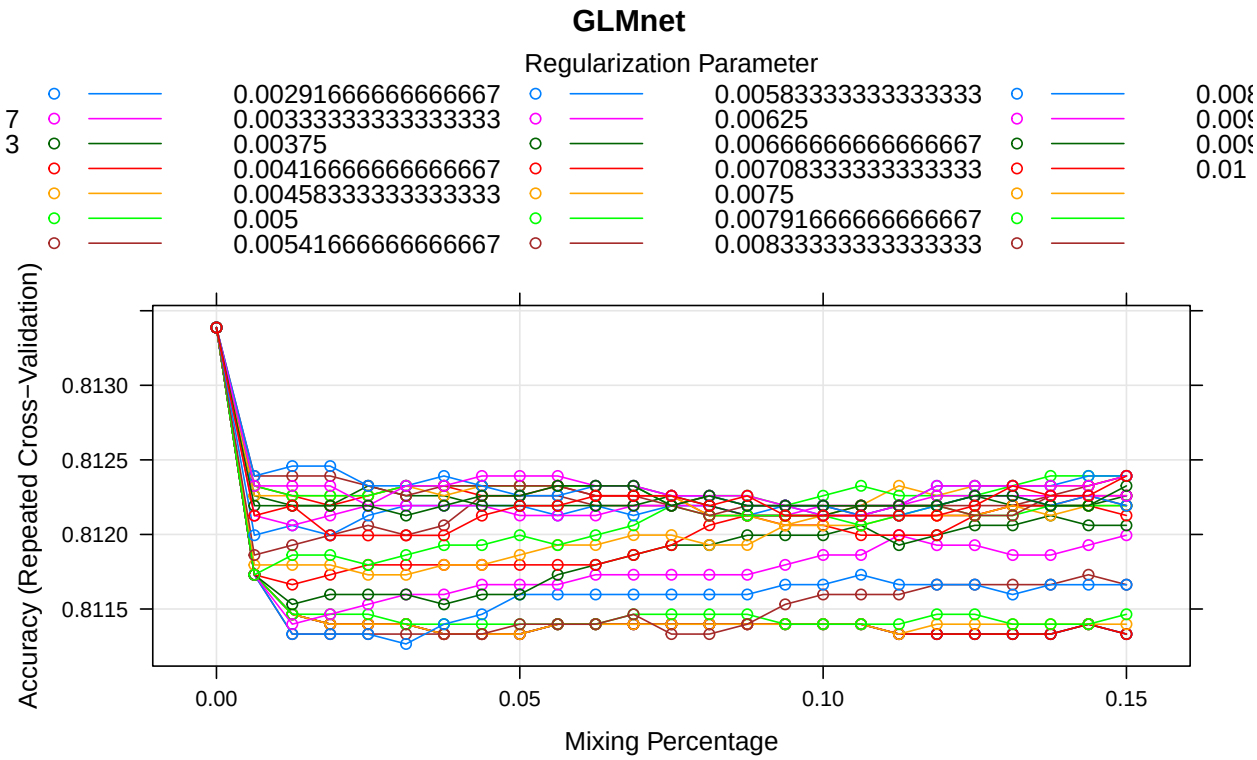
Modeling

Our approach consists, for each of the six methods used, in starting off by training an unoptimized model which is in turn used as the starting point before determining a set of optimal tuning parameters by cross-validation.

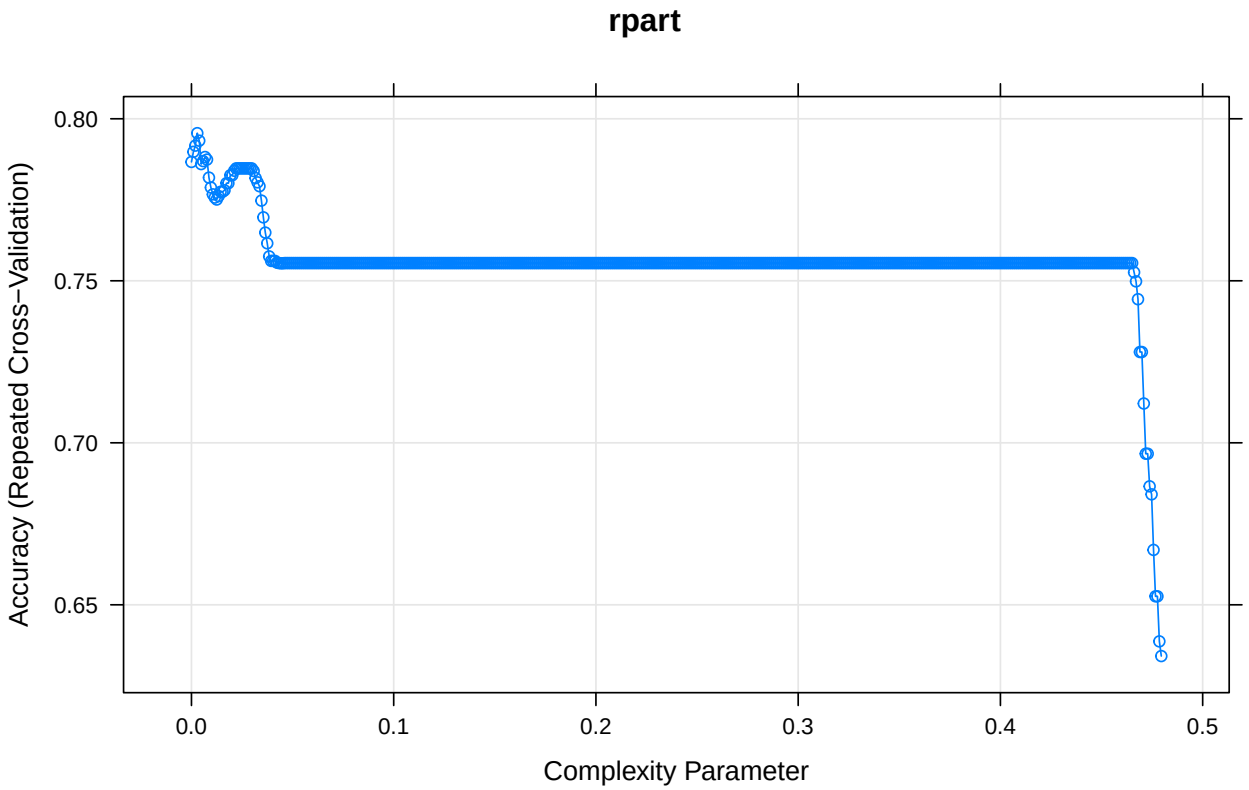
GLMnet

Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elasticnet penalty at a grid of values for the regularization parameter lambda

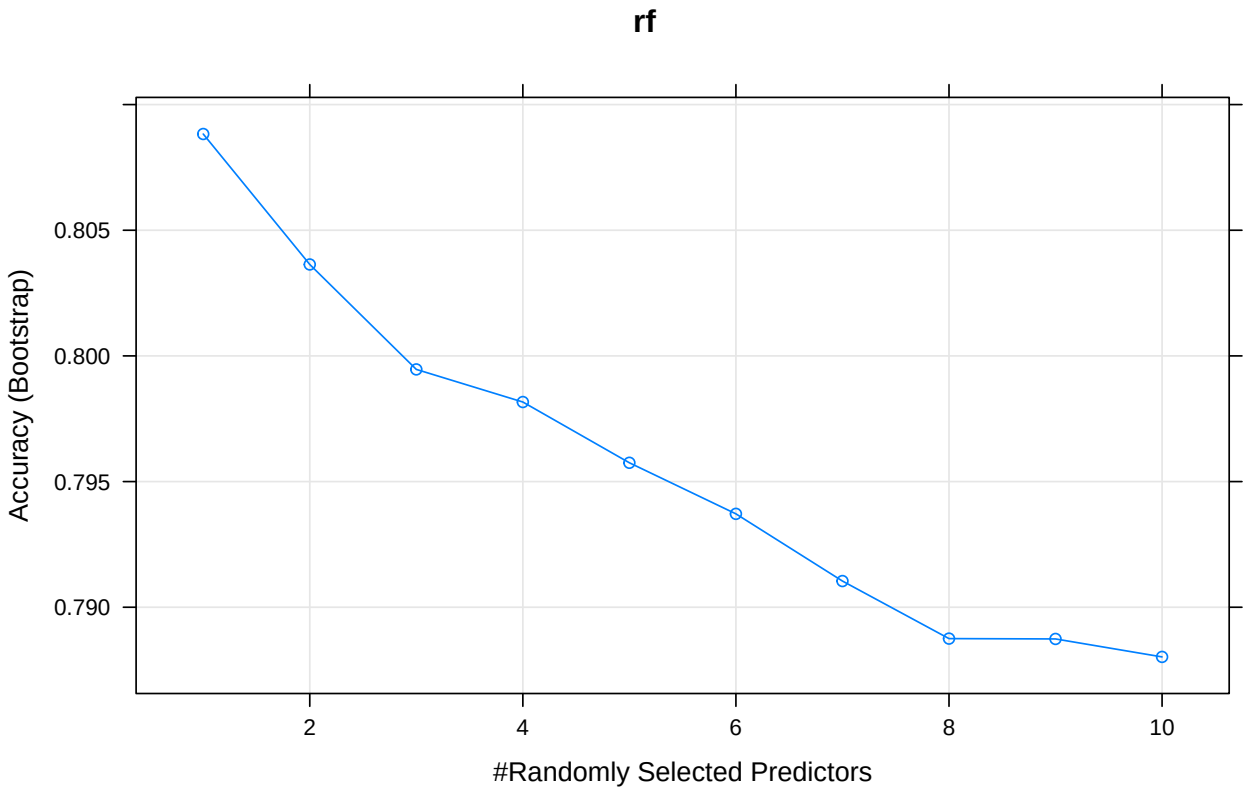
Optimization of parameters for the glmnet model:



Decision trees

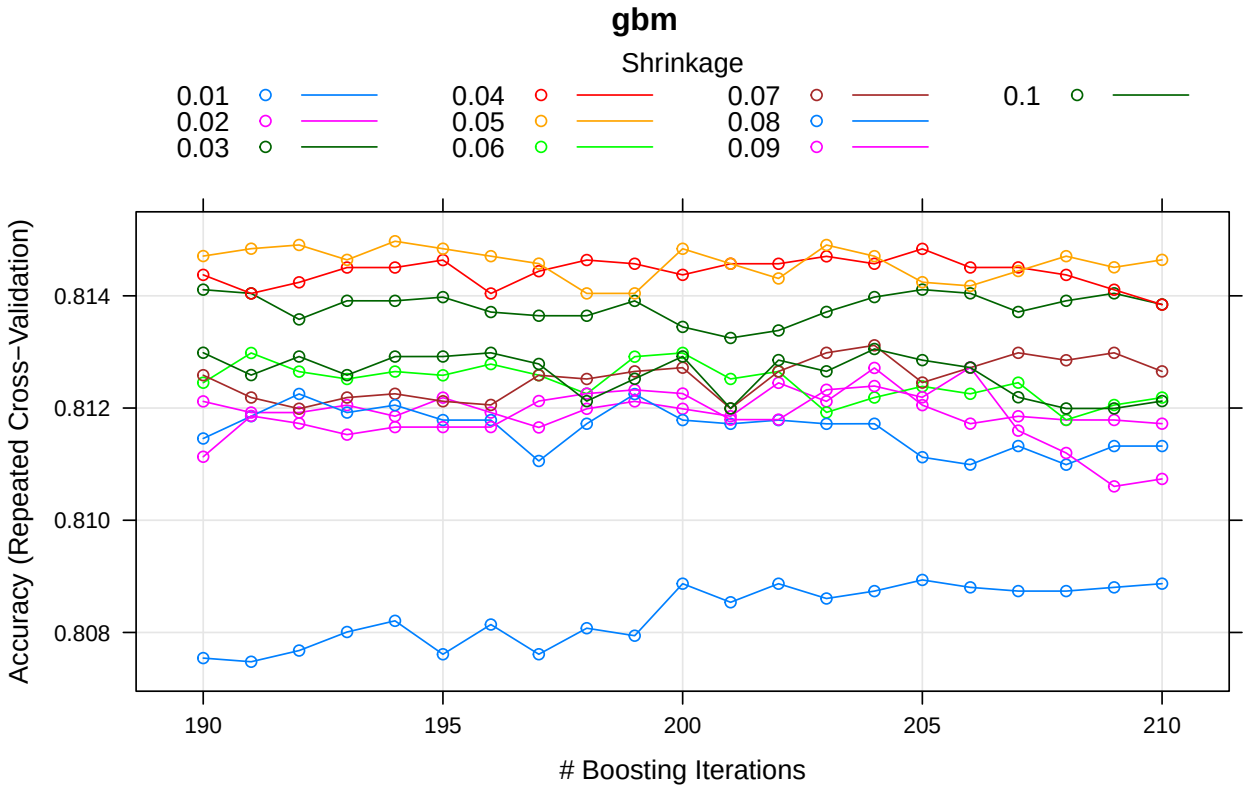


Random forest



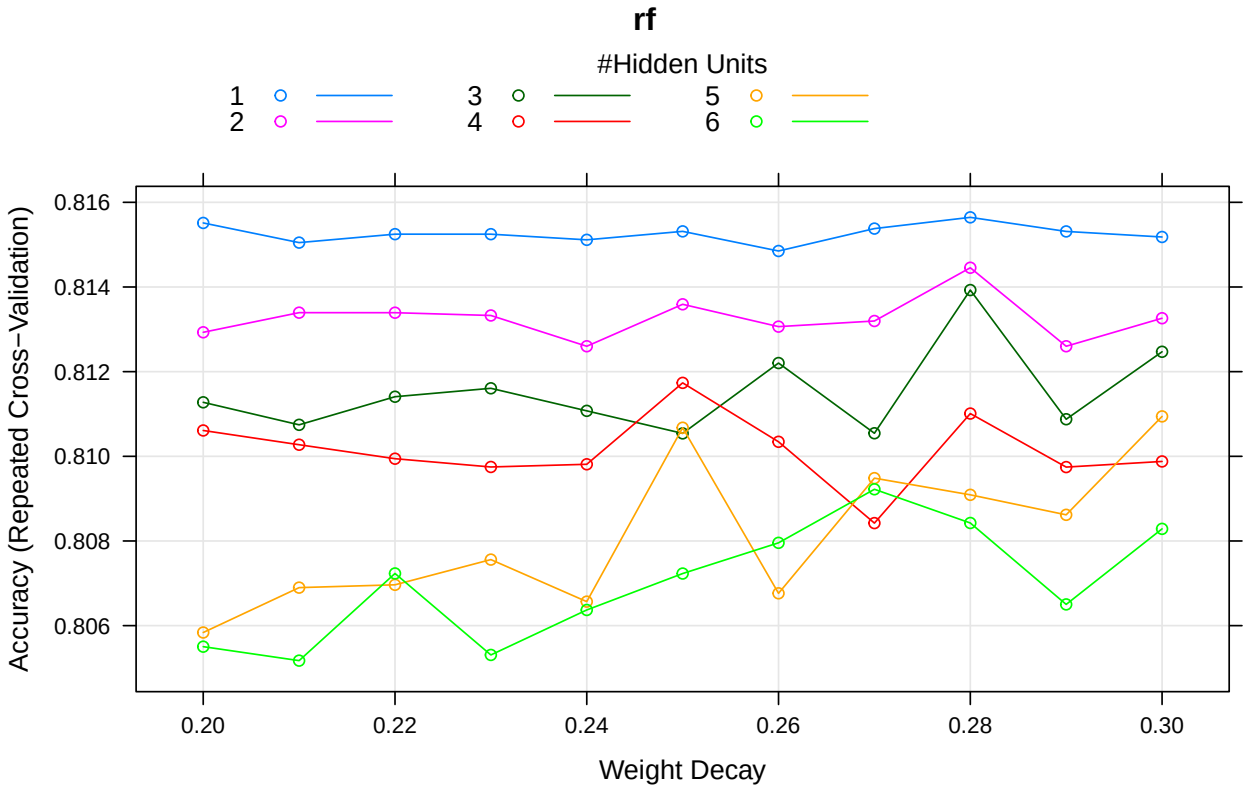
Stochastic gradient boosting

Optimal values of shrinkage and boosting iterations for the gbm model:

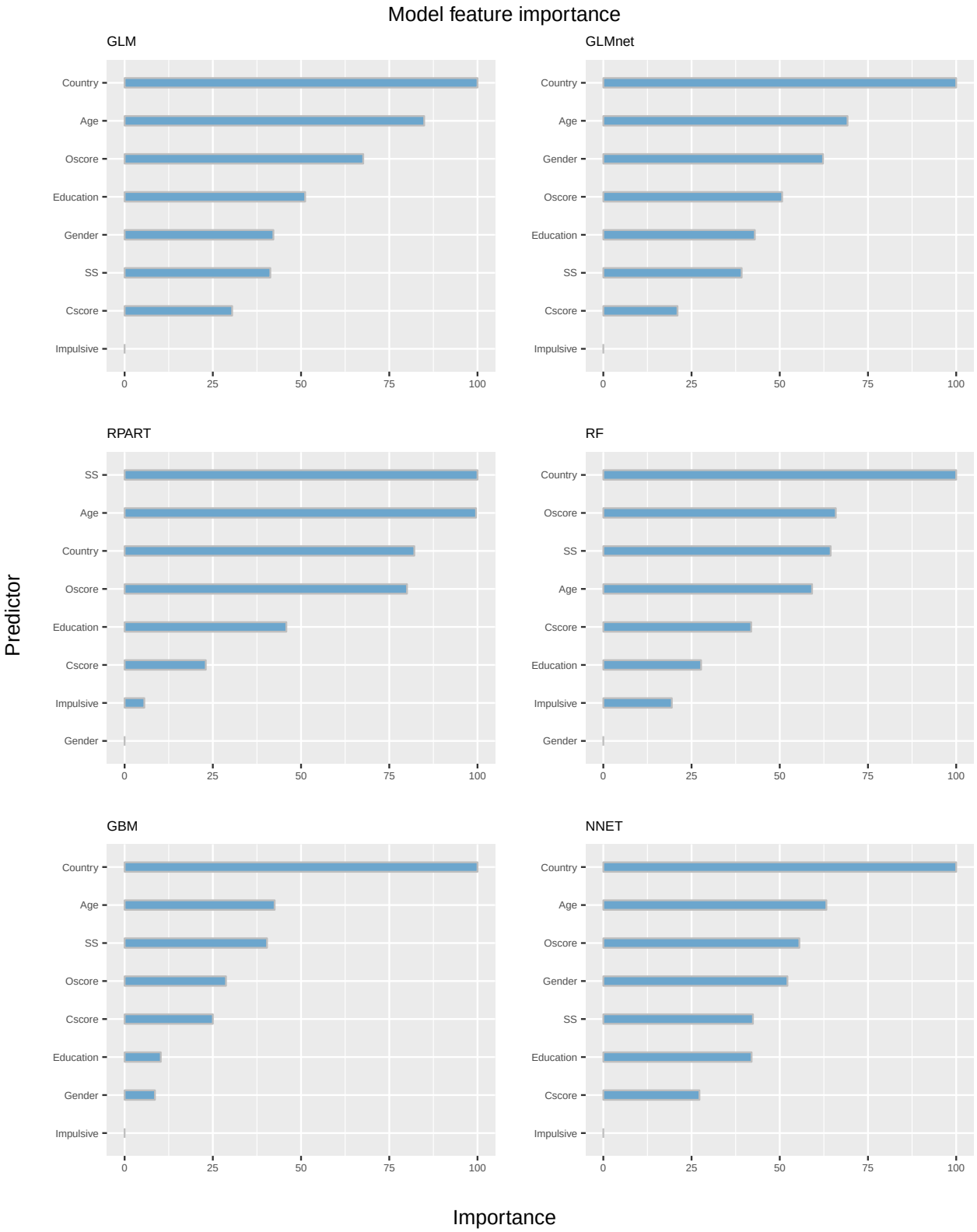


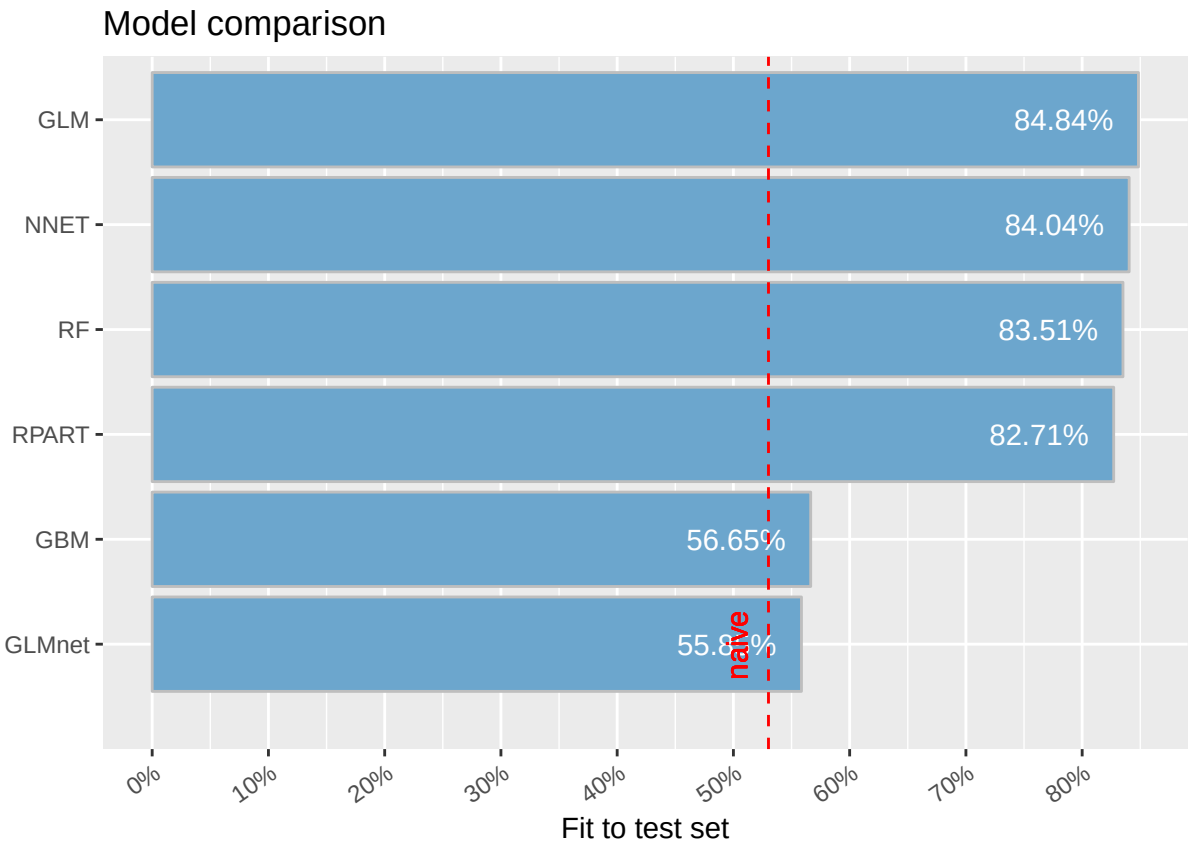
Neural network

Optimal values of decay and number of hidden units for the neural network:



Model comparisons




Model comparison
GLM
NNET
RF
RPART
GBM
GLMnet
84.84%
84.04%
83.51%
82.71%
56.65%
naive
0%
10%
20%
30%
40%
50%
60%
70%
80%
Fit to test set

Results

None of the modeling approaches used provided an improvement over generalized linear regression (84.8% accuracy). Besides offering the highest accuracy, the importance plot related to the glm model is by and large consistent with results from the data exploration:

- Country of origin, Age, Openness to experiment and Education are the factors contributing most to the accuracy.
- Edu ation and Sensation-seeking have a significant importance
- Ethnicity, N-scores and E-scores and do not contribute much to the metric.
- However the low impact of Impulsivity is at variance with the observations from data exploration.

The confusion matrix for the GLM model has a sensitivity of (0.82) and specificity of (0.87):

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 146  26
##          1  31 173
##
##                Accuracy : 0.8484
##                  95% CI : (0.8081, 0.8831)
##     No Information Rate : 0.5293
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6953
##
##  Mcnemar's Test P-Value : 0.5962
##
##             Sensitivity : 0.8249
##             Specificity : 0.8693
##          Pos Pred Value : 0.8488
##          Neg Pred Value : 0.8480
##              Prevalence : 0.4707
##          Detection Rate : 0.3883
##    Detection Prevalence : 0.4574
##       Balanced Accuracy : 0.8471
##
##        'Positive' Class : 0
##
```

The next best model (84.0% accuracy) is obtained with a neural network. It suggests that demographic factors have more impact on cannabis use than personality (top three importance factors: country of origin, ethnicity, and age). Given the statistical similarity between the two ethnic groups discussed in the Personality analysis section, this model is somewhat unconvincing.

The confusion matrix for the neural network model is :

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 143  26
##          1  34 173
##
##                Accuracy : 0.8404
```

```
##                  95% CI : (0.7994, 0.876)
##     No Information Rate : 0.5293
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.679
##
##  Mcnemar's Test P-Value : 0.3662
##
##             Sensitivity : 0.8079
##             Specificity : 0.8693
##          Pos Pred Value : 0.8462
##          Neg Pred Value : 0.8357
##              Prevalence : 0.4707
##          Detection Rate : 0.3803
##    Detection Prevalence : 0.4495
##       Balanced Accuracy : 0.8386
##
##        'Positive' Class : 0
##
```

With 0 (Non-user) as the 'positive' class, the 3-point decrease in sensitivity (0.81) indicates a drop in this model's ability to predict Non-users.

While less accurate, the random forest model (83.5% accuracy) gives a preponderant importance to country of origin, sensation-seeking trait, age and openness to experiment). It also gives no importance to ethnicity and E-score and less than expected to gender.

The confusion matrix for the random forest model is :

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 143  28
##          1  34 171
##
##                Accuracy : 0.8351
##                  95% CI : (0.7937, 0.8712)
##     No Information Rate : 0.5293
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.6685
##
##  Mcnemar's Test P-Value : 0.5254
##
##             Sensitivity : 0.8079
##             Specificity : 0.8593
##          Pos Pred Value : 0.8363
##          Neg Pred Value : 0.8341
##              Prevalence : 0.4707
##          Detection Rate : 0.3803
##    Detection Prevalence : 0.4548
##       Balanced Accuracy : 0.8336
##
##        'Positive' Class : 0
##
```

With RF also, the decrease in sensitivity (0.81) indicates a drop in this model's ability to predict Non-users.

For this dataset, the optimized GLM offers the weakest modeling technique (55.9% accuracy), performing less well than the naive approach (53.0% accuracy).

Conclusion

With demographic factors being more predictive than personality, the modeling suggests that cannabis consumption is first and foremost a cultural phenomenon.

Generalized linear modeling offers the highest improvement (31.8%) over the naive approach and the neural network (31.0%). While both values are similar,the GLM model agrees better with the results from the exploration and statistical analysis of the data.

In a previous run, we had labelled non-users only those that had never used cannabis and users all the others. In that case, machine learning offered only a modest improvement (81.1% for the neural network compared to 78% with the naive approach) We felt this may be due to the choice of classification and that a potentially more insightful analysis would take into account frequency of use (this information was nevertheless not available in the dataset).