# Prediction of Cannabis Consumption from Demographics and Personality

HarvardX PH125.9x Data Science Capstone

*Charles Mégnin*

*10/8/2019*

## 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 the opportunity to healthcare professionals 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 original dataset 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.

In the original dataset, drug use is separated between 'Never used', 'Used over a decade ago', '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: 'Never Used' (the original predictor) and 'Used' (the combination of the others, representing people having used cannabis at least once in their lifetime).

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.

The data consists of 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 therefore consists of one Class (Cannabis consumption labeled 'Used') 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.

This report consists of two main sections:

- In the first part, after performing minor data engineering, we explore, bin, and analyze the dataset.
- In the second part, we move on to the modeling phase:
- After applying a Recursive feature elimination algotirhm to seek and discard predictors that do not contribute significantly to the outcome, we build models based on the following 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

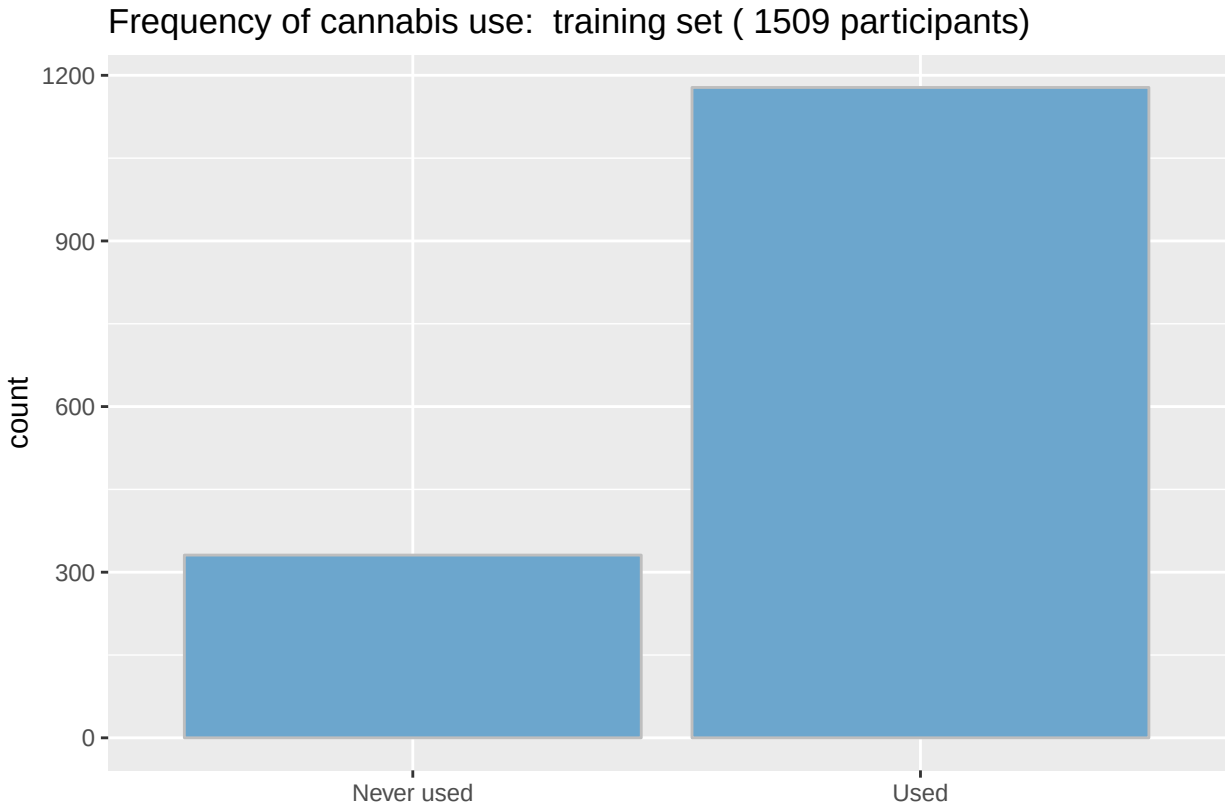
## A: Data engineering

- All predictors were already normalized and centered in the original dataset.
- We construct the 'Used' class to separate 'Never used' participants (0) from the others (1).
- We then partition the data between training (80% / df.train) and test sets (20% / df.test) preserving the distribution of the Cannabis class.

## B: Data exploration

- There are 0 NAs in the dataset.

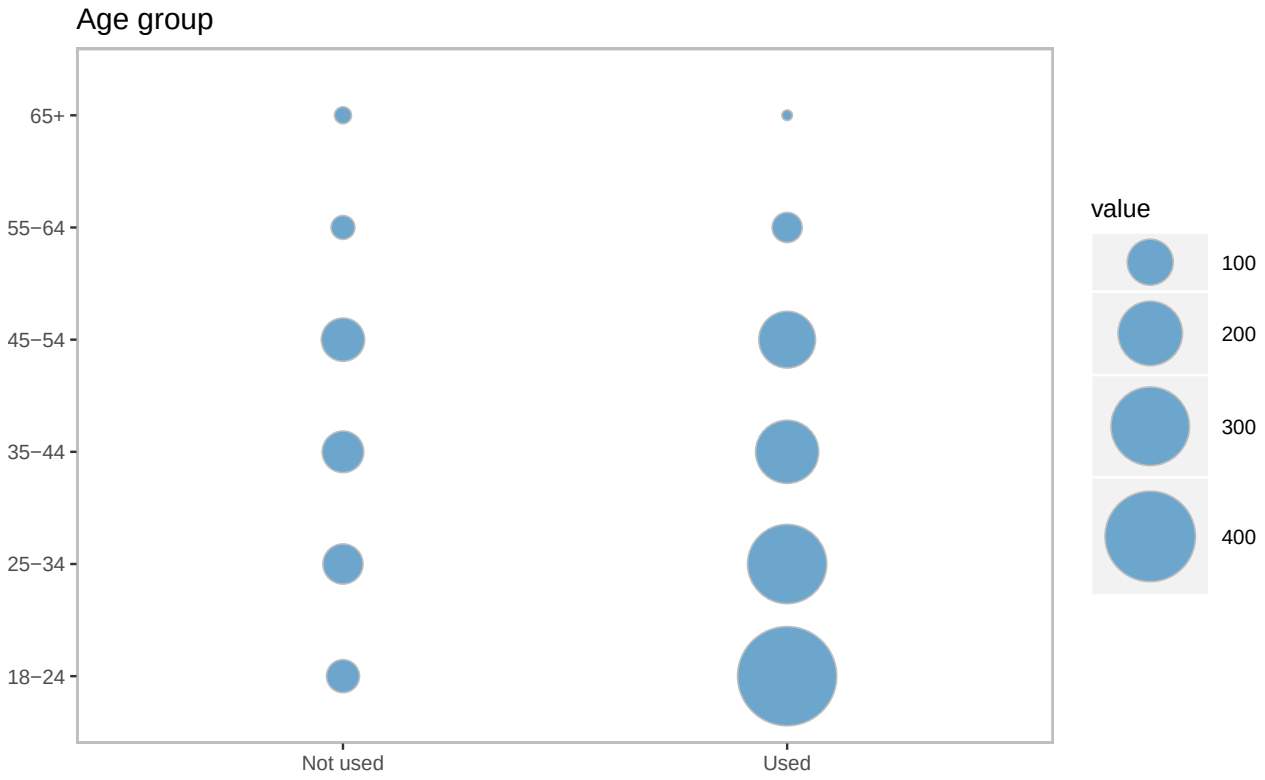
### Class distribution

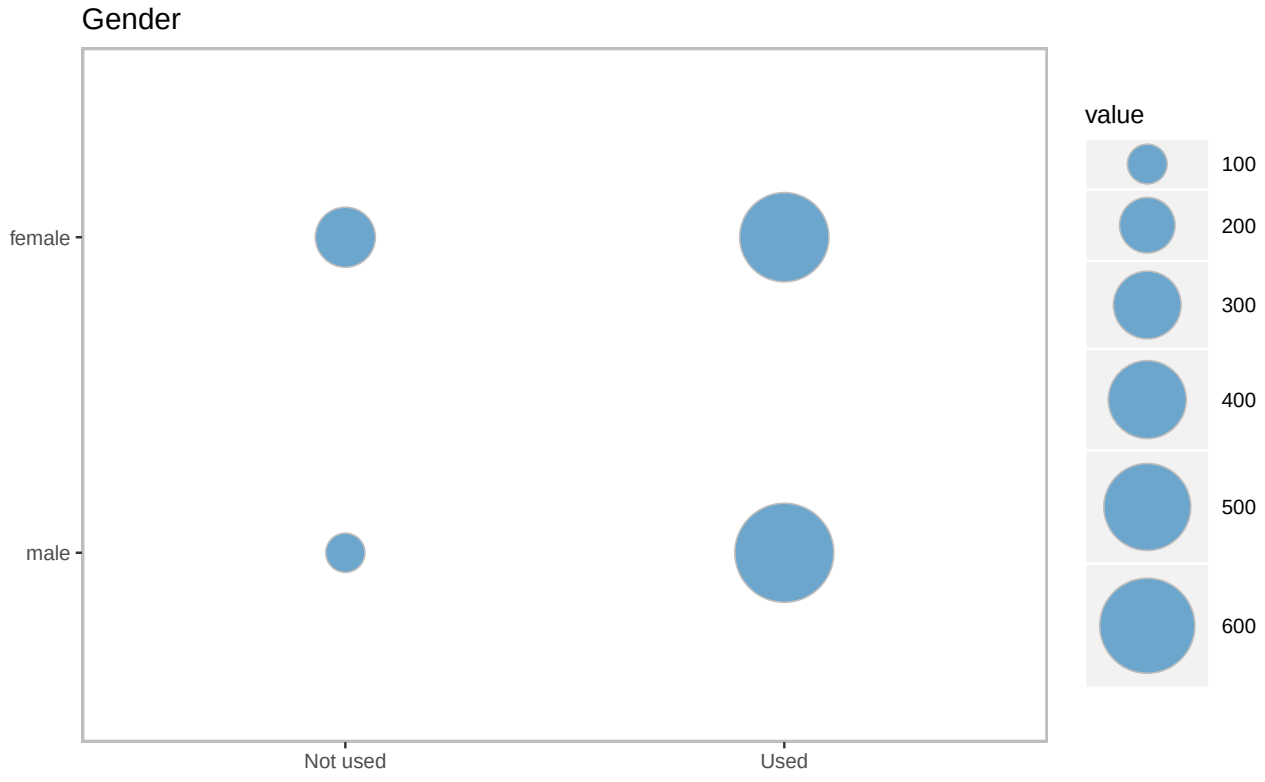


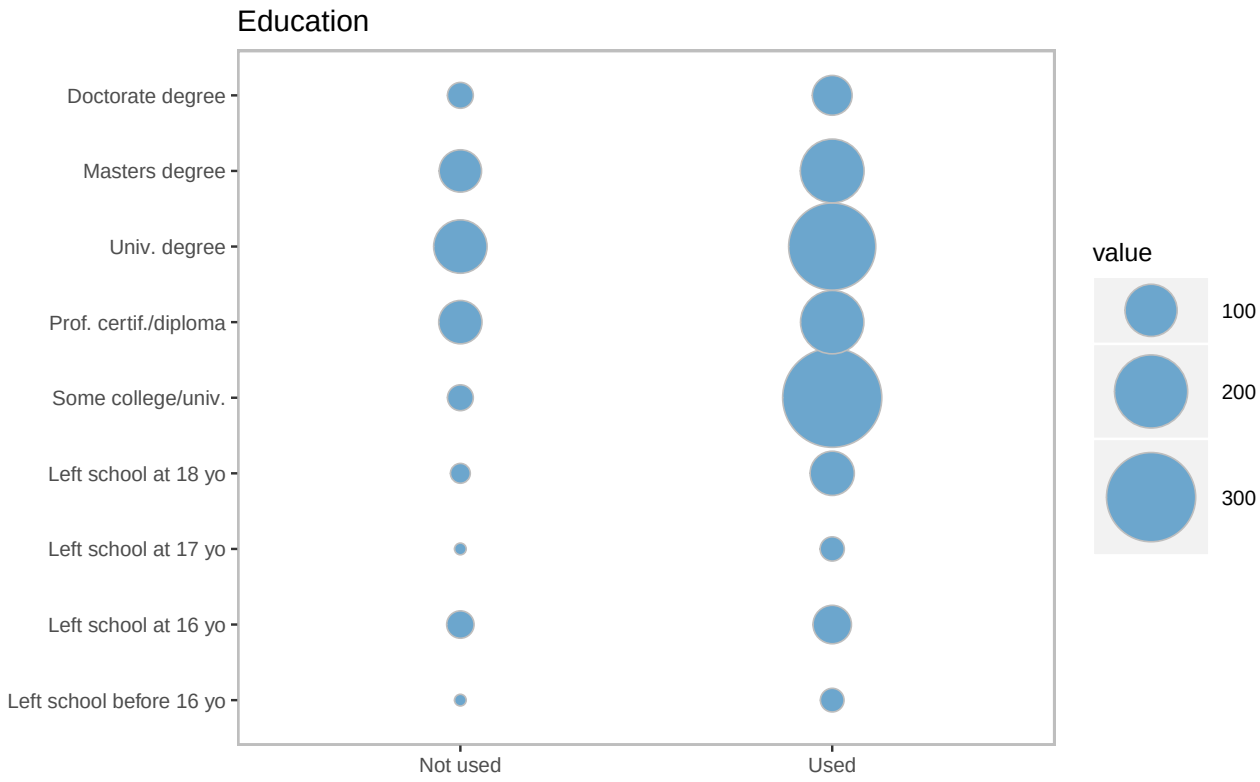

The training set consists of 1178 participants having used cannabis at some pioint in the past and 331 participants who haven't for a user-to-non-user ratio of ratio of 1:3.5589124.

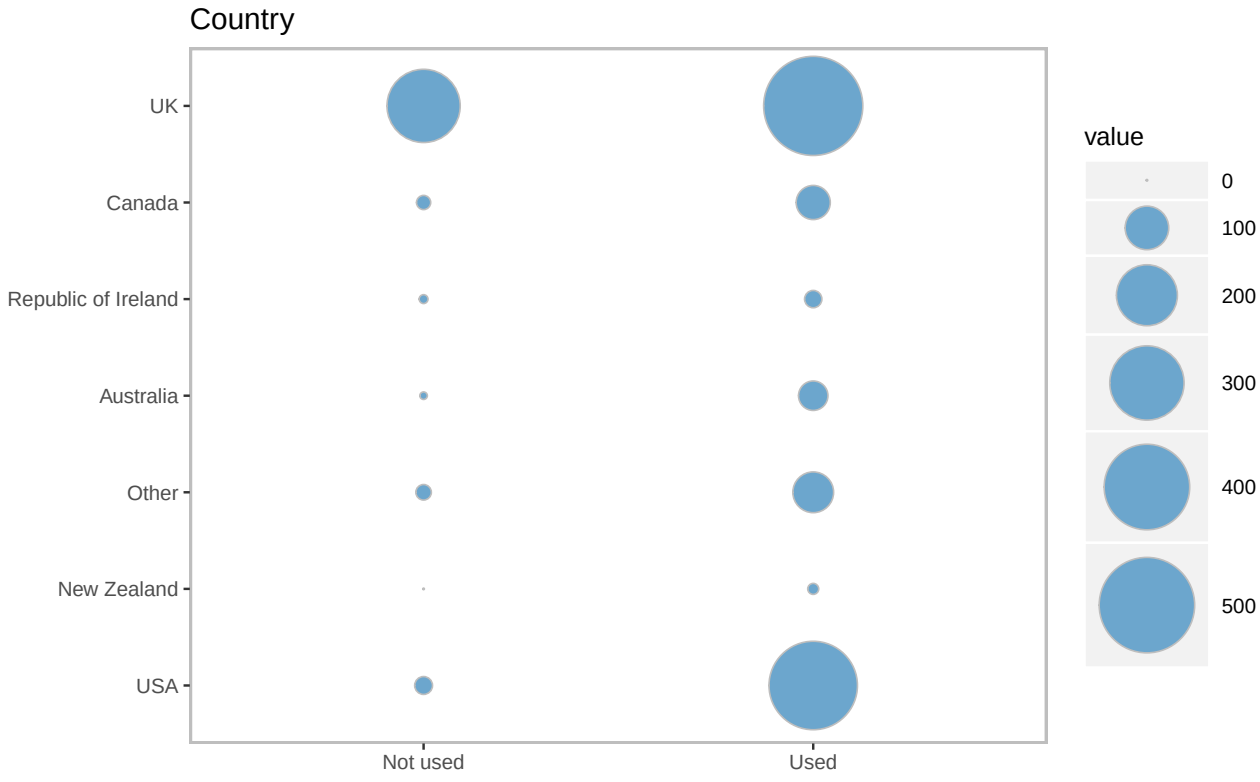
### Contingency plots (prior to binning)

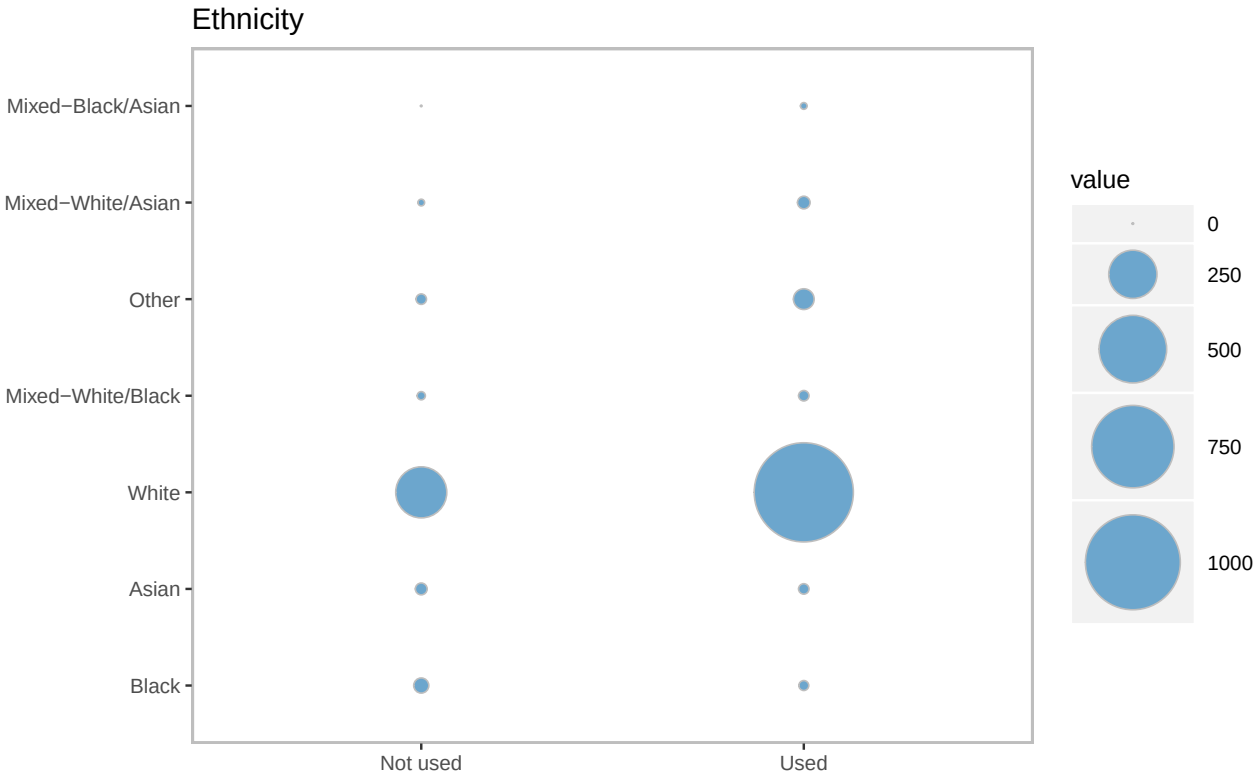
### Cannabis use by demographic group


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
100
200
300
400
500
600


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
0
100
200
300
400
500



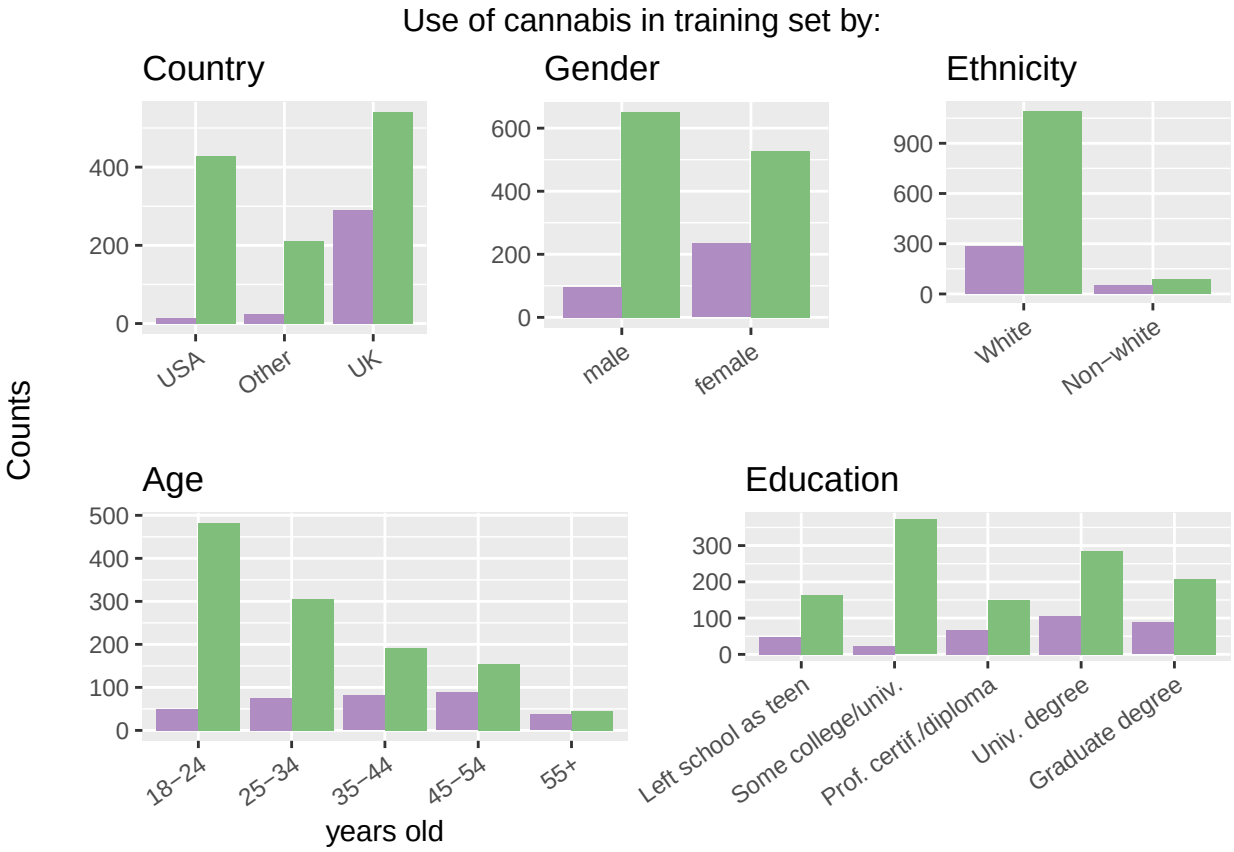

We note that the dataset of 1509 participants is dominated by young and educated white American and British participants of both sexes.

### Binning

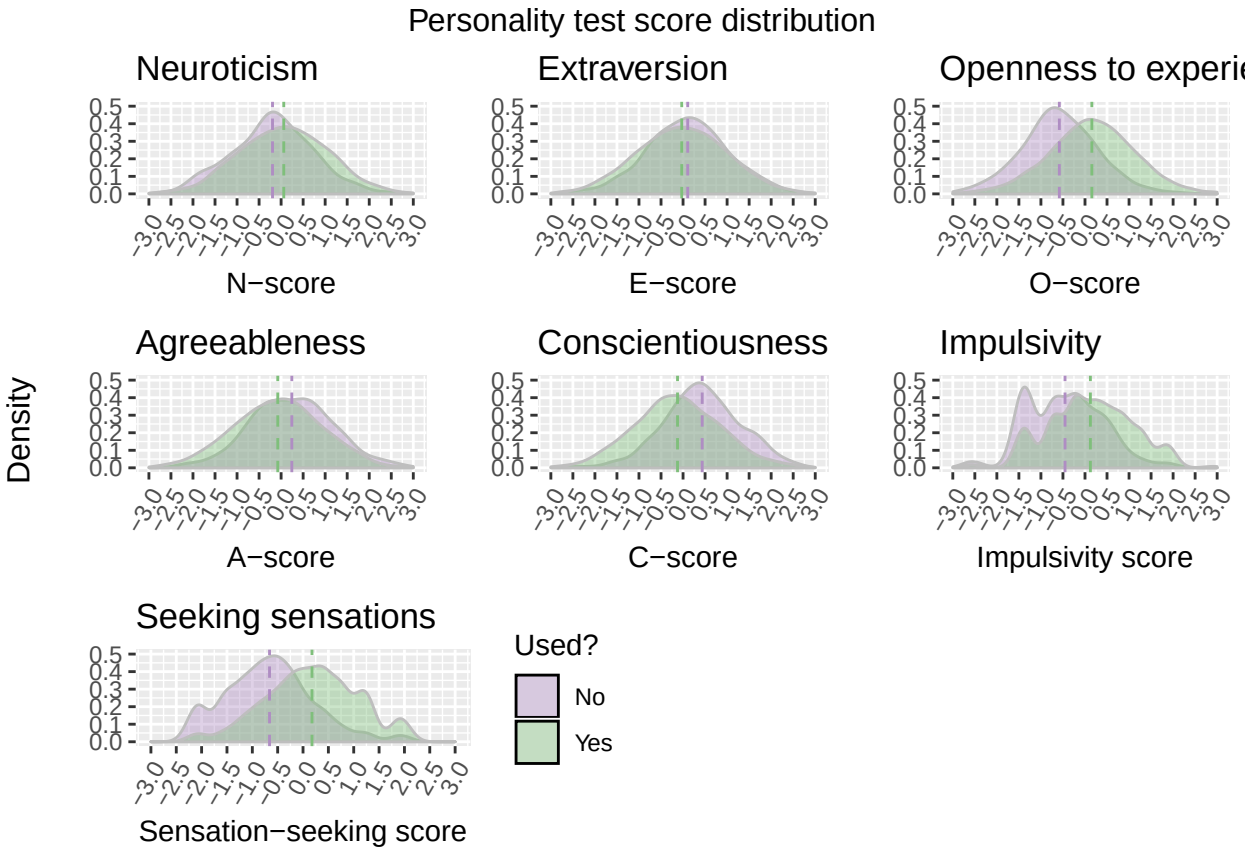
The small size of many demographic sub-groups add little valuable insight and will likely only serve to introduce variance in the analysis. At the risk of erasing behavioral differences among groups, the distribution of the dataset forces a more meaningful binning of the demographic information:

- 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”, “Graduate degree”.
- 3 groups for Country: “USA”, “UK”, “Others”.
- 2 ethnic groups: “Whites”, “Non-whites”

**Analysis of demographics**



**Personality analysis**



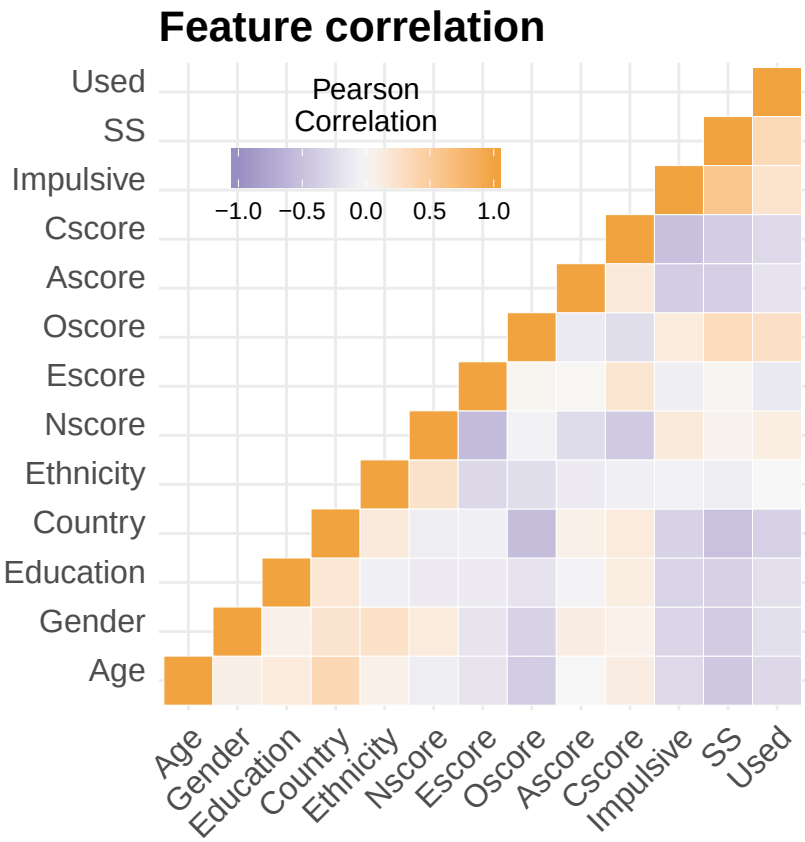

The density plots show some measure of difference between users and non-users particularly as it relates to either openness to experience, agreeableness, conscientiousness, impulsivity, and sensation-seeking. Some implications are rather entertaining, notably the notion that nice (ie: agreeable) people may be less likely to smoke weed, or conversely that exposure to pot tends to might make people less nice.

On the other hand, the distributions for users and non-users as they relate to neuroticism or extraversion are similar, suggesting that these personality traits may not impact cannabis consumption. We will examine in the modeling section below whether that is indeed the case.

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

**Feature correlation**

We examine redundancies among the 12 predictors and with the Used class:

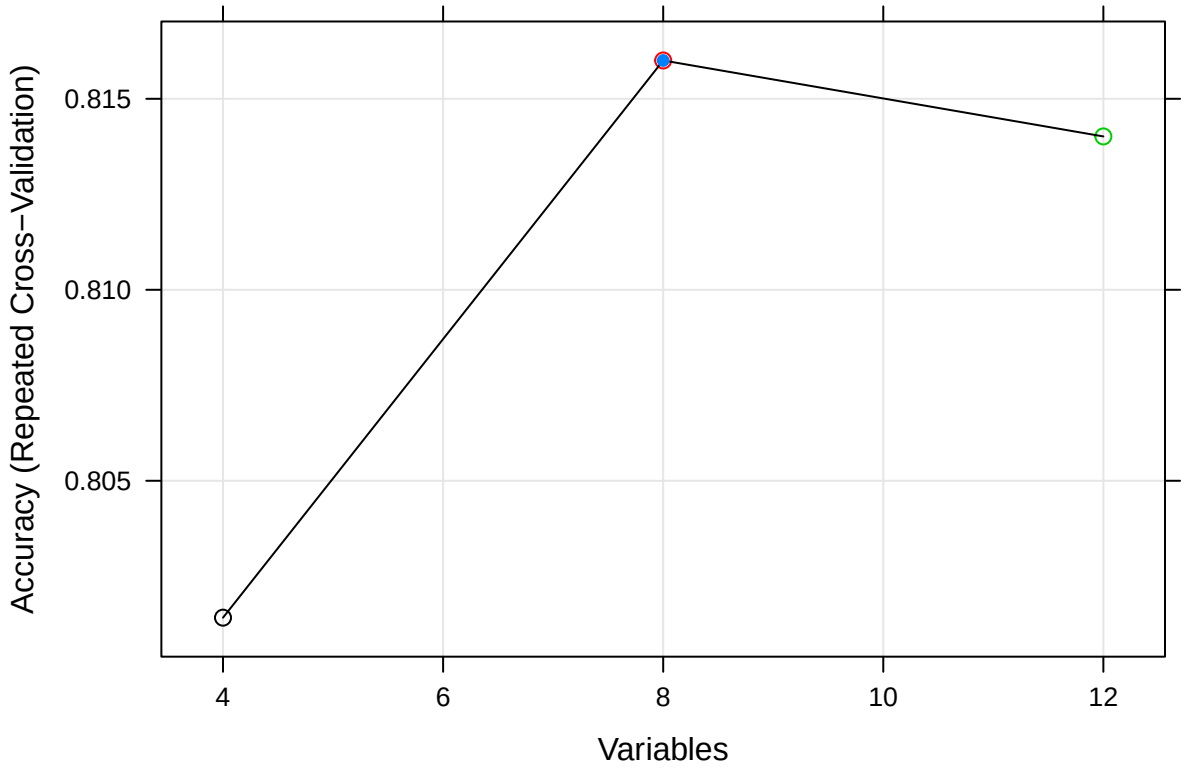

Feature correlation
Pearson Correlation
−1.0
−0.5
0.0
0.5
1.0
Used
SS
Impulsive
Cscore
Ascore
Oscore
Escore
Nscore
Ethnicity
Country
Education
Gender
Age
Age
Gender
Education
Country
Ethnicity
Nscore
Escore
Oscore
Ascore
Cscore
Impulsive
SS
Used

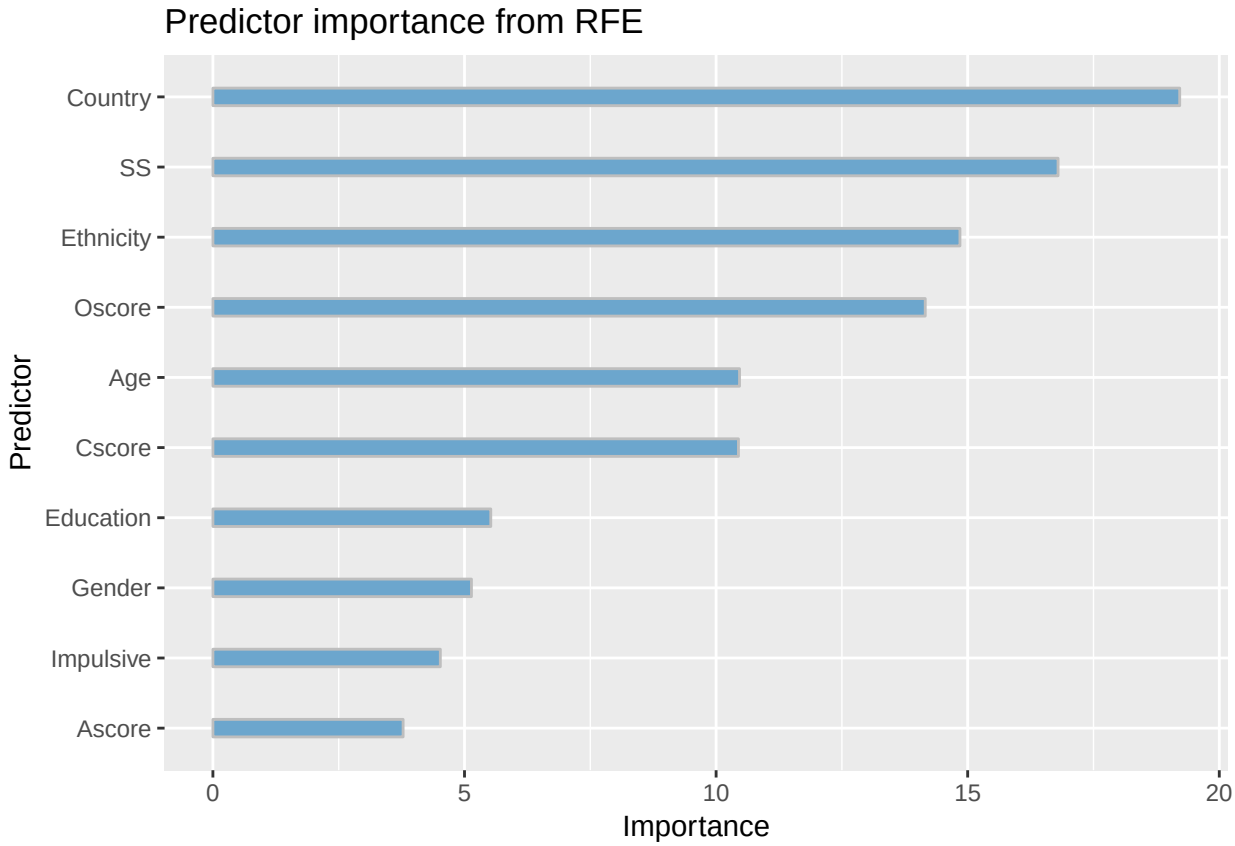
## C: Modeling

We seek a model which improves on the ratio of users to the population (78.10%). This constitutes the baseline above which predictive modeling is interesting.

### Recursive Feature elimination

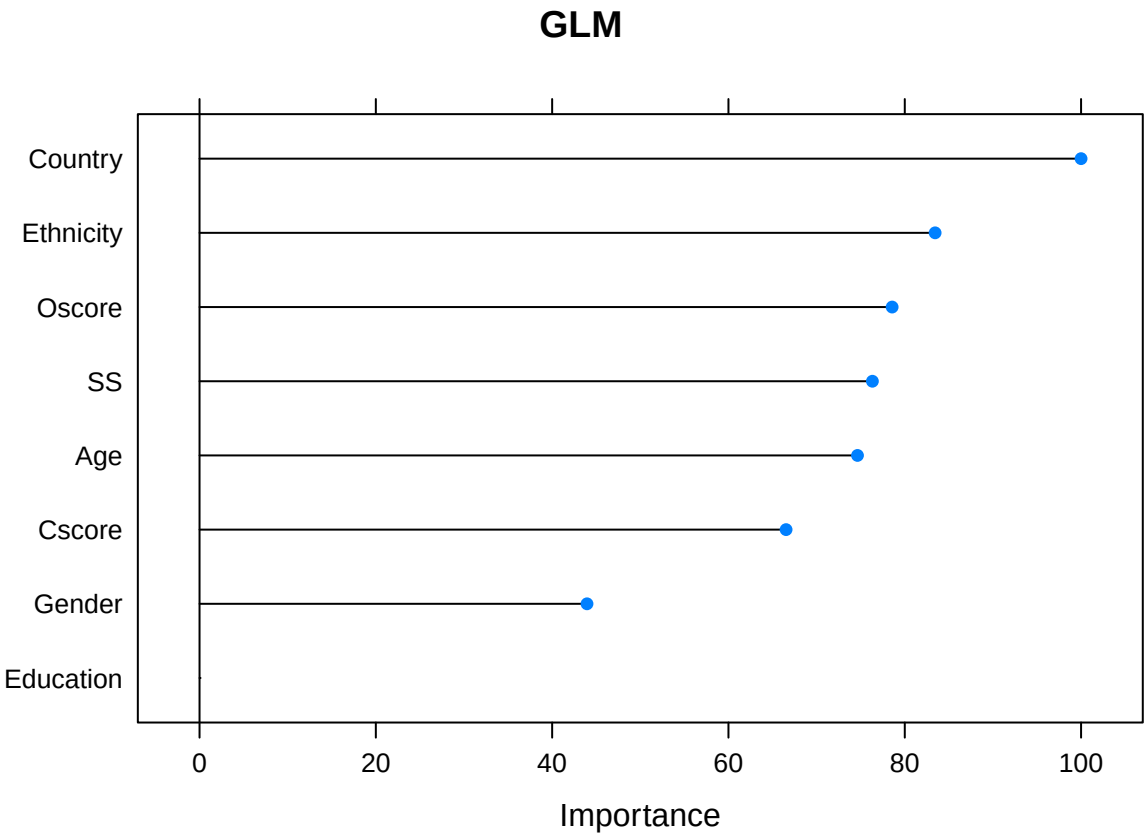
For RFE as well as subsequent modeling, we use the k-fold cross validation method which involves splitting the dataset into k subsets. The algorithm holds aside one of the subsets while the model is trained on the others. This process is repeated a predetermined number of times and the overall accuracy estimate is provided.






The comparative analysis of the contribution of each factor agrees by and large with that of the density distribution plots: among the personality trait tests, N-score and E-score contribute the least while seeking sensation and O-score contribute the most. However, the impulsivity doesn't seem to be as strong a factor as one might have expected from the density plots.

**Generalized linear model**




```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  33  27
##          1  49 267
##
##                Accuracy : 0.7979
##                  95% CI : (0.7537, 0.8373)
##     No Information Rate : 0.7819
##     P-Value [Acc > NIR] : 0.2481
##
##                   Kappa : 0.3439
##
##  Mcnemar's Test P-Value : 0.0160
##
##             Sensitivity : 0.40244
##             Specificity : 0.90816
##          Pos Pred Value : 0.55000
##          Neg Pred Value : 0.84494
##              Prevalence : 0.21809
##          Detection Rate : 0.08777
##    Detection Prevalence : 0.15957
##       Balanced Accuracy : 0.65530
##
```

```
##        'Positive' Class : 0
##
```

**Generalized linear model with penalized maximum likelihood**

**GLMnet without parameter tuning**

```
## NULL
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  24  86
##          1  58 208
##
##                Accuracy : 0.617
##                  95% CI : (0.5658, 0.6664)
##     No Information Rate : 0.7819
##     P-Value [Acc > NIR] : 1.00000
##
##                   Kappa : 1e-04
##
##  Mcnemar's Test P-Value : 0.02445
##
##             Sensitivity : 0.29268
##             Specificity : 0.70748
##          Pos Pred Value : 0.21818
##          Neg Pred Value : 0.78195
##              Prevalence : 0.21809
##          Detection Rate : 0.06383
##    Detection Prevalence : 0.29255
##       Balanced Accuracy : 0.50008
##
##        'Positive' Class : 0
##
```

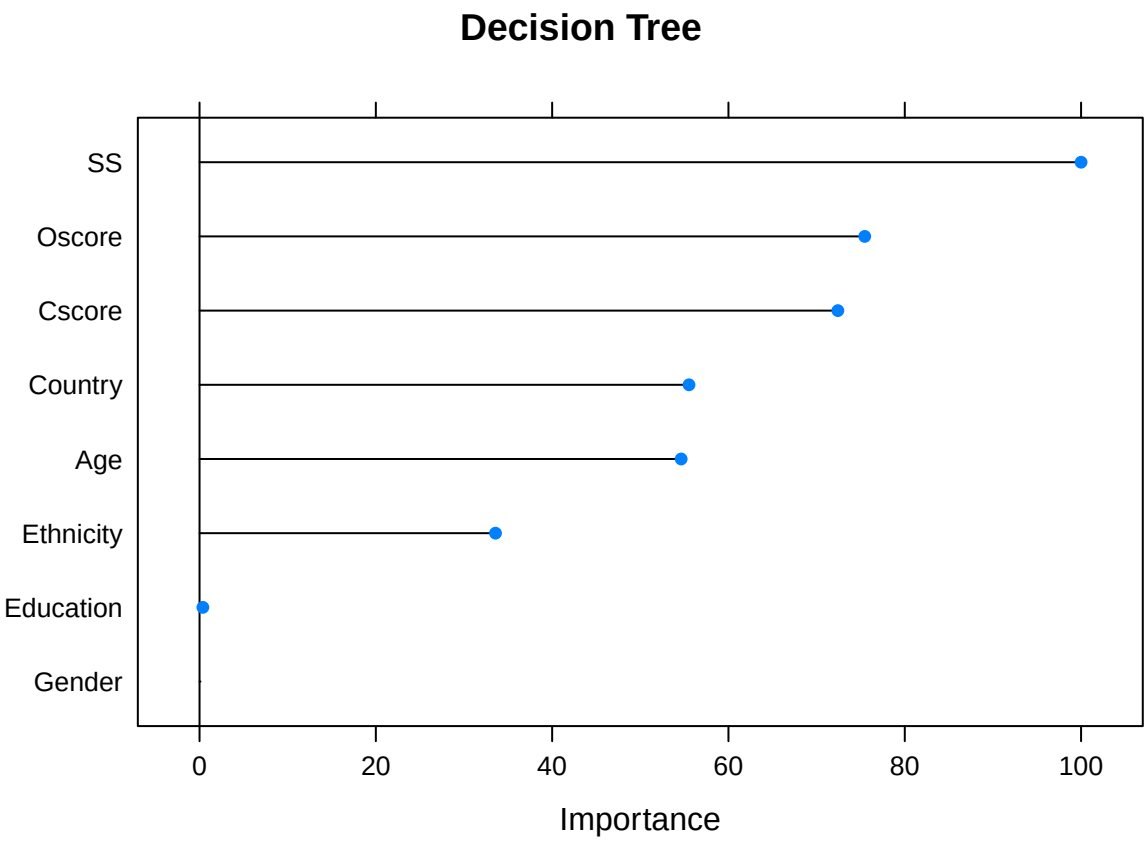
**GLMnet with parameter tuning**

```
## NULL
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  24  85
##          1  58 209
##
##                Accuracy : 0.6197
##                  95% CI : (0.5685, 0.669)
##     No Information Rate : 0.7819
##     P-Value [Acc > NIR] : 1.00000
##
##                   Kappa : 0.0032
##
```

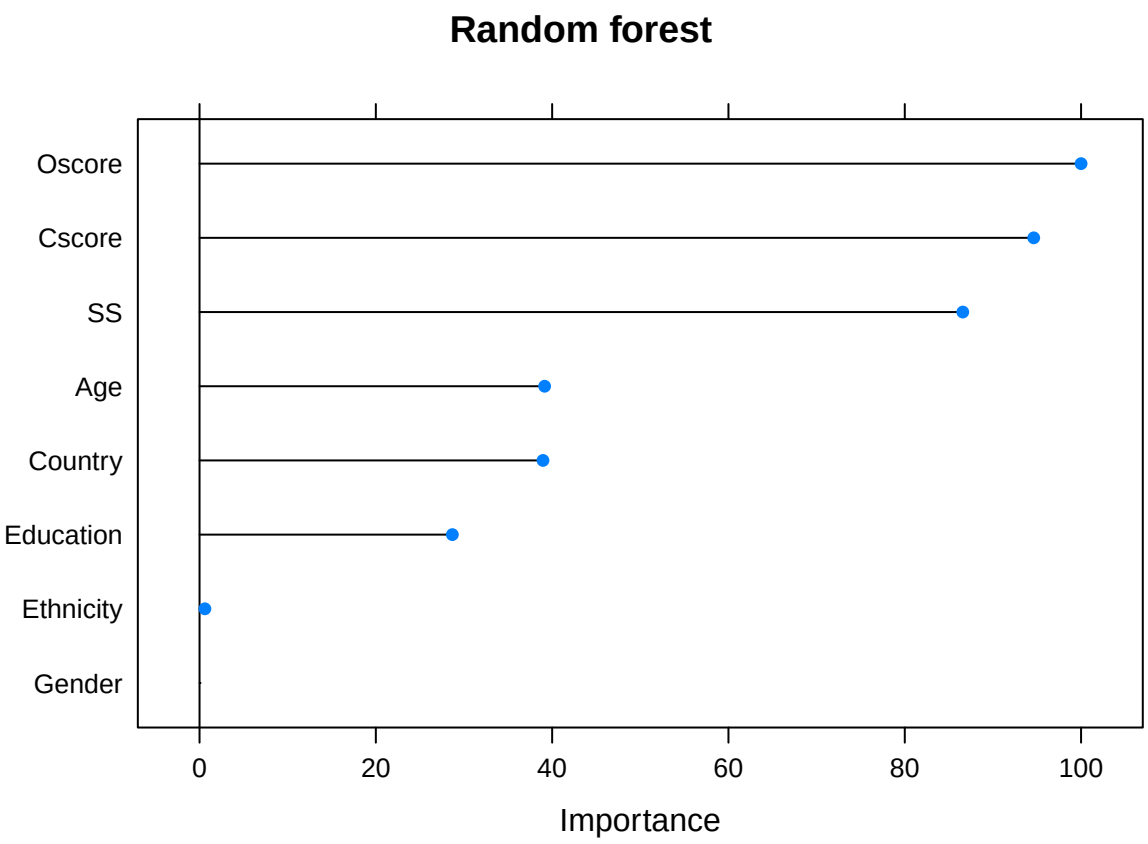
```
##  Mcnemar's Test P-Value : 0.02969
##
##             Sensitivity : 0.29268
##             Specificity : 0.71088
##          Pos Pred Value : 0.22018
##          Neg Pred Value : 0.78277
##              Prevalence : 0.21809
##          Detection Rate : 0.06383
##    Detection Prevalence : 0.28989
##       Balanced Accuracy : 0.50178
##
##        'Positive' Class : 0
##
```

**Decision trees**



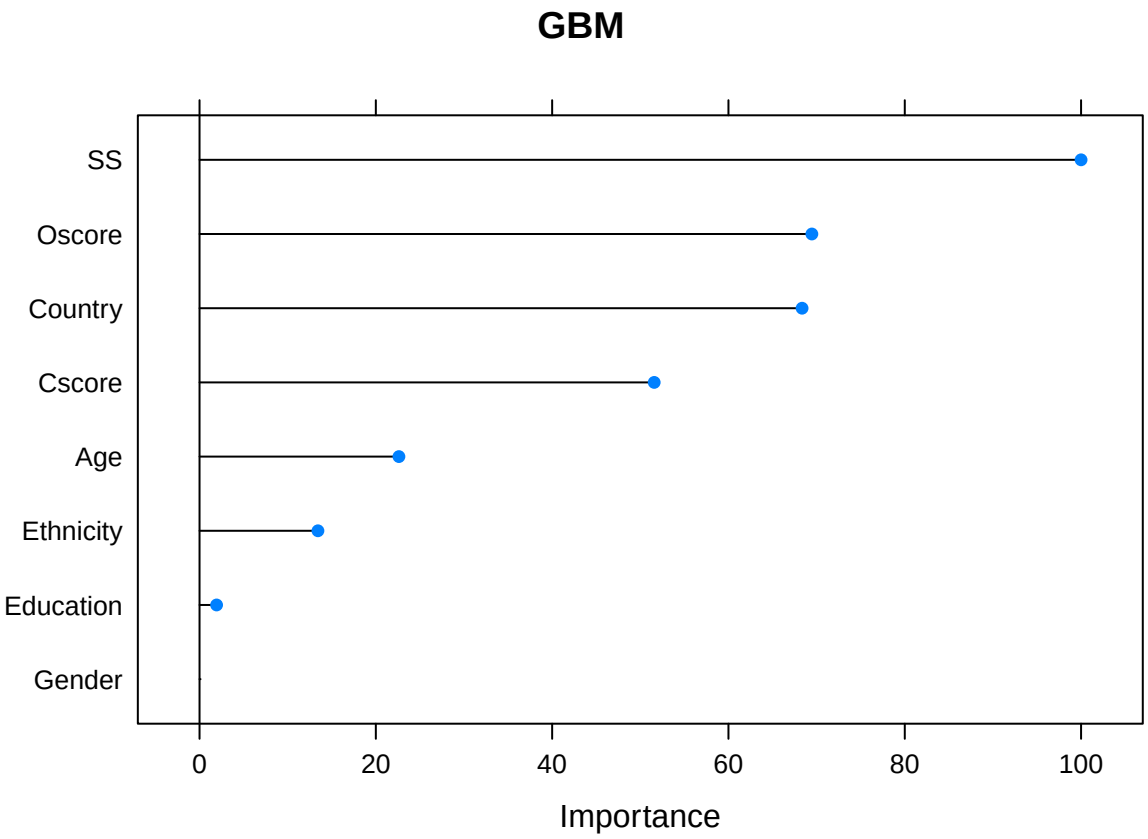

```
## NULL
```

**Random forest**



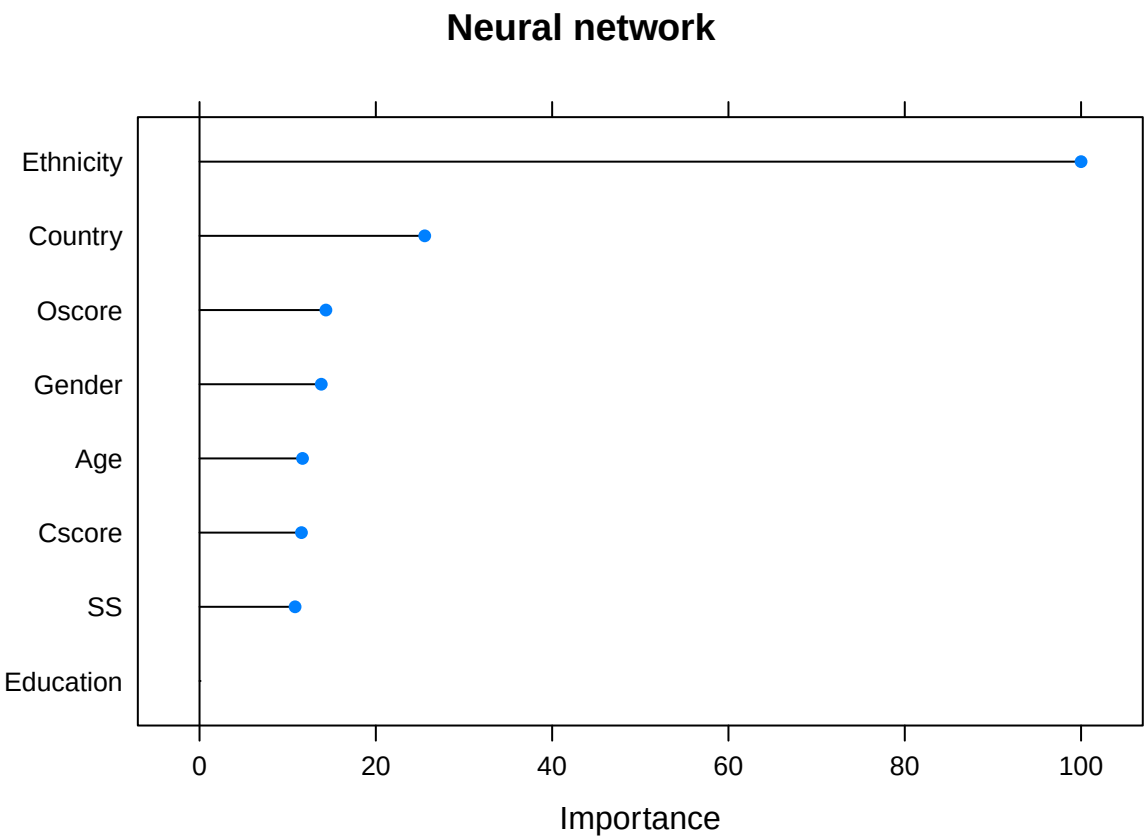

```
## NULL
```

**Stochastic Gradient Boosting**



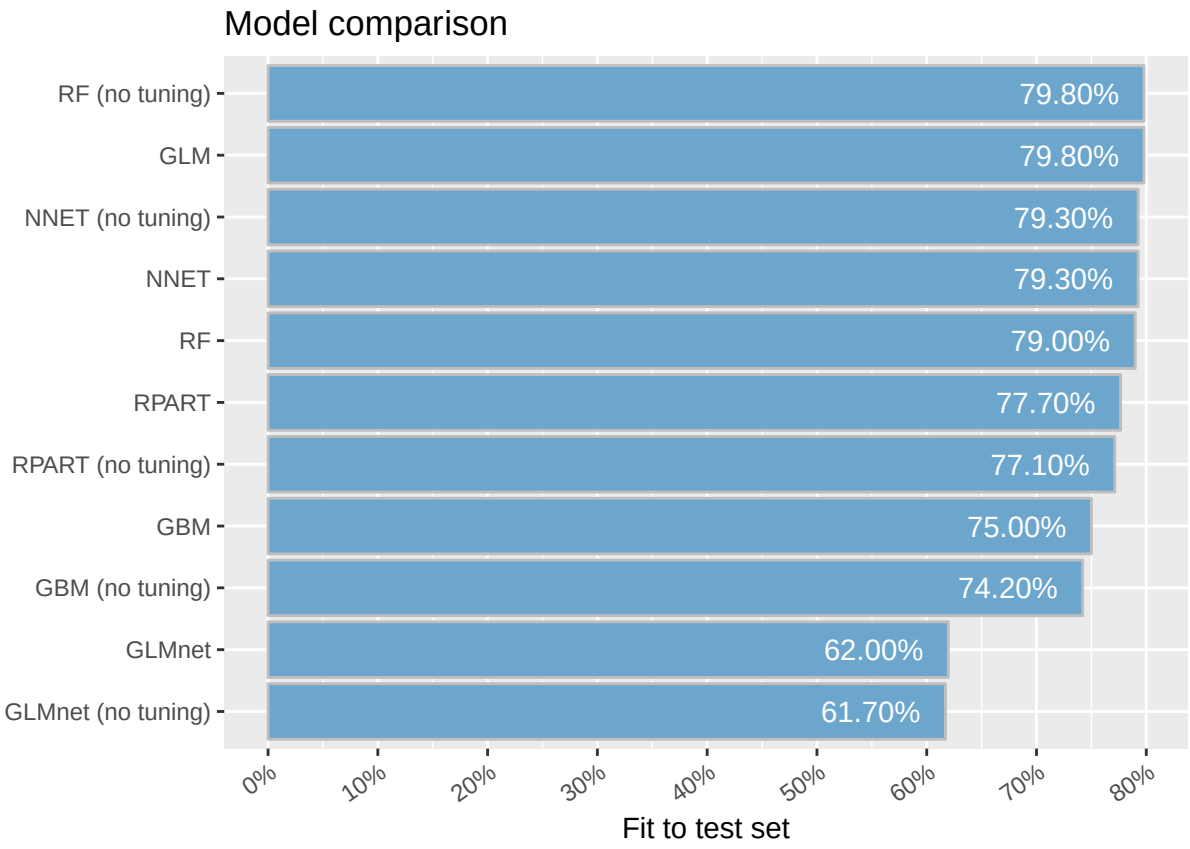

```
## NULL
```

**Neural network**




```
## NULL
```

**Model comparisons**




# Results

GBM: Little importance is given to the education variable. Improvement analysis -> what did the use of ML contribute

# Conclusion

Scope -> having used once in lifetime may not be telling. Could be interesting to bin never with over ten years ago ?