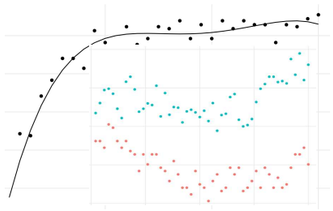


Predicting substance use in individuals

M3A50 – Project 1

Hitesh Kumar
Department of Mathematics, Imperial College London

Imperial College London

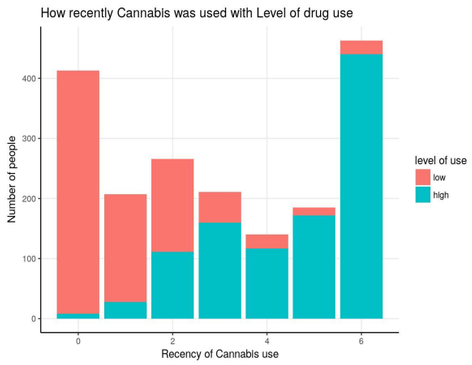
PURPOSE OF THIS INVESTIGATION

In this investigation we use a dataset provided to train and test various machine learning models. Here, we predict an individuals level of substance use with GLM's (General linear models) and GBT's (Gradient boosting trees), and finally later, we also attempt to predict an individuals age by training Neural networks.

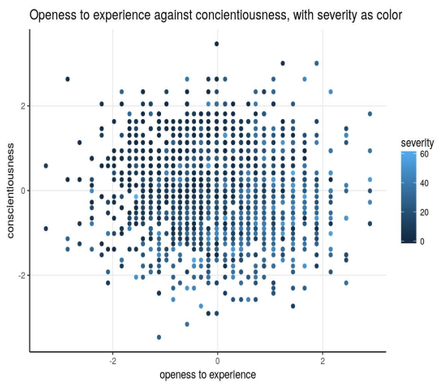
EXPLORING OUR DATA

First of all, we should look through our data to get a grasp of what paths may be most interesting to follow

One area we could investigate is how the use of "milder drugs" like cannabis indicate a high overall level of use of drugs.



Here, 0 means the individual has never used this drug, and 6 means it was used as recently as the previous day. It becomes clear now that those who have used cannabis with any recentness have clearly been using other substances too and hence have been marked as having a high level of drug use! (Perhaps cannabis is indeed a gateway drug?...)

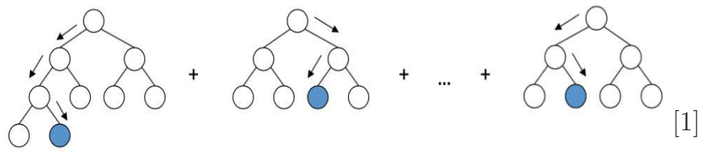


Another pattern that was interesting was the link between openness to experiences, conscientiousness and severity of drug use.

Here we see that those more willing to try new things and are less conscientious may be more likely to have severe drug use

LEVEL OF USE VIA GBT'S

Now we train Gradient Boosting Trees to predict an individuals level of substance use based on factors such as age, gender, personality and use of legal drugs such as Alcohol and Nicotine.



[1]

GBT's work by training more and more models on the mistakes of the previous ones to improve the predictive power of the combination of them. The library we use in R, XGBOOST is a particularly high performance library that uses K-fold cross validation to train the weak learners (trees) off of each other.

Using this method, we used our predictors to train the model with about 80% of our data, and used the remaining 20% to test it. We find the accuracy of our model to be almost 85%, which is quite impressive for a simple implementation of this model.

We see that our model is quite a good fit for the data we have, but how would it react to new data? Of course, more data wasn't readily available, so we used K-fold cross validation with the data we have. Once we set K = 10, it was quite easy to write code that would perform this in R.

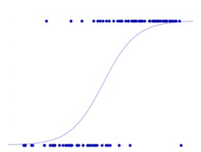
```
For i = 1 : K
    - Split data
    - Fit model
    - Predict
    - calculate accuracy
End

Find average accuracy
```

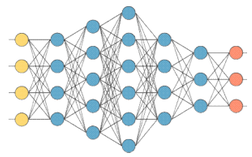
After this quick algorithm, we see that our model will react quite well to new data, showing an average accuracy of around 84.5%!

We can not only conclude that our model works well, but also now that the predictors we used are actually good at predicting substance use.

PREDICTING AGE BY SUBSTANCE USE

In the final part of our investigation, we now look at using other models to predict an individuals age based on their use of all substances .We will use a GLM, and Neural networks.





Before we train any model however, we must convert this into a regression task, by splitting our outcome into binary form. The most interesting choice perhaps was to split the data into age being less than 24, and age being greater than 24 (Quite likely splitting students from non-students too...)!

After fitting our models and making our predictions we see that as expected, the GLM performs noticeably better in this task than the Neural network, which gets a lower accuracy on average

Model	GLM	Neural net.
Accuracy	81%	~74%

We can take this assessment of the two further, and can look at the error of the two models as we increase the training data size. Here we can see that GLM's almost always perform better, and get even better with more data unlike neural networks!

WHICH PREDICTORS ARE IMPORTANT

From our EDA we see that the use of substances such as cannabis or nicotine can very strongly predict high levels of use, and personality traits such as being open to experiences and being less conscientious, or being less agreeable and more neurotic surprisingly can weakly predict more severe drug use. From our investigation, we can see that Gender and age in particular are biological traits which most strongly predict substance use, and quite surprisingly ,the use of alcohol has almost no power to predict drug use!

REFERENCES

[1] – Argozh Nikov, "gbdt attractive picture", available from: http://arogozhnikov.github.io [Accessed 16/11/17]

[2] – Open NN, "deep neural network", available from: http://www.opennn.net [Accessed 16/11/17]

[3] – Analytics Vidhya, "Logistic regression plot", available from: https://www.analyticsvidhya.com [Accessed 16/11/17]