Predicting drug use of individuals using machine learning

Overview

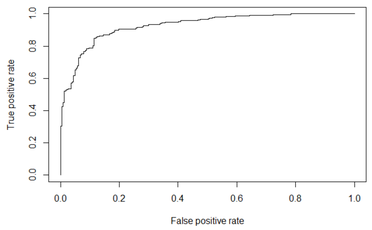
The data set provided contained 5 predictors on background, 7 with scores for personality traits, 4 on the consumption of legal substances, 14 on the consumption of illegal substances and any, severity and use level. Using machine learning techniques we can use the data to make predictions.

Data analysis

To begin with I explored the relationships between the predictors through calculating the correlation matrix shown as a heat map to the left in figure 1 and the corresponding p values to determine significance of the relationships in the data.

Predicting Use Level of a individual

Initially I created a logistic regression classifier to predict use level. Then used 10-fold cross validation to evaluate its performance on a new data set. This model performed well as can be seen from its good ROC curve below with a high area under the curve of 0.93 (1 being perfect).



Then I tried other techniques listed below to determine if we could improve the accuracy and removed predictors with a high p value as found in my data analysis which were alcohol, country and chocolate.

- Näive Bayes
- Linear Discriminant Analysis
- Mixture Discriminant Analysis
- Support Vector Machines
- Neural Networks
- K-Nearest Neighbours

I then combined these first by using the modal prediction then a neural net to weight the predictions of each classifier to produce the best classifier overall. The neural net using the predictions from all the above classifiers and the logistic regression classifier produced the best overall classifier. This technique of combining classifiers into a single classification model is known as ensemble learning. The box plots of each classifiers accuracy are shown to the left in figure 1 with the ensemble NN performing best.

Significant predictors of drug use

Through my data analysis and the calculation of correlations as shown in the heat map and their p values. For use level and severity the predictors that weren't significant at the 5% significance level were:

- Alcohol
- Extra version
- Country

Through further exploration for alcohol and extra version we could see through plotting the histograms of alcohol and extra version given use level they were distributed almost identically showing it isn't a good predictor. With country however through producing a tally table it seems it is due to the bulk of the data being from the UK or US so there isn't enough data in the data set to draw meaningful relationships. However with any drug use chocolate and country weren't significant. Chocolate was for the same reason as alcohol with use level and severity and can be seen the same way. Country wasn't significant for the same reason as explained before.

Correlation heat map

A heat map of the correlations between the columns of the initial data showing their relationships.

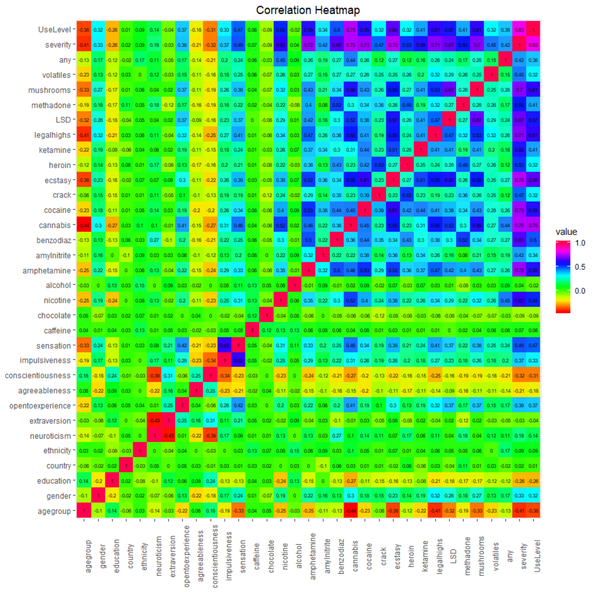


Figure: Correlation Heat Map

Accuracies of different classifiers for use level

Box plot showing the accuracies of the different classification techniques tested and of the ensembles.

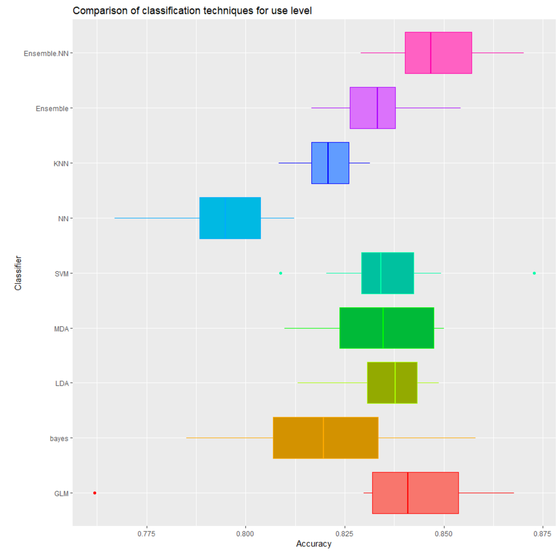


Figure: Box plot of use level classifiers

Predicting severity of a individuals drug use

I used all the predictors I used for use level however I added cannabis use as I thought it would be interesting to see if severity of use is closely linked to cannabis usage. The techniques used are as follows:

- Lasso/Ridge regression
- General linear model with the data defined as Gaussian and the basis being the products of all the predictors
- Multivariate Adaptive Regression Splines
- Neural networks

As previously these were also all included in ensemble learning techniques other than neural nets due to their high range taking the mean output and then using a neural nets to best predict severity. The techniques were compared using mean squared error. Figure 3 shows a box plot of the MSEs of the tried techniques extreme outliers for the NN were removed prior to plotting.

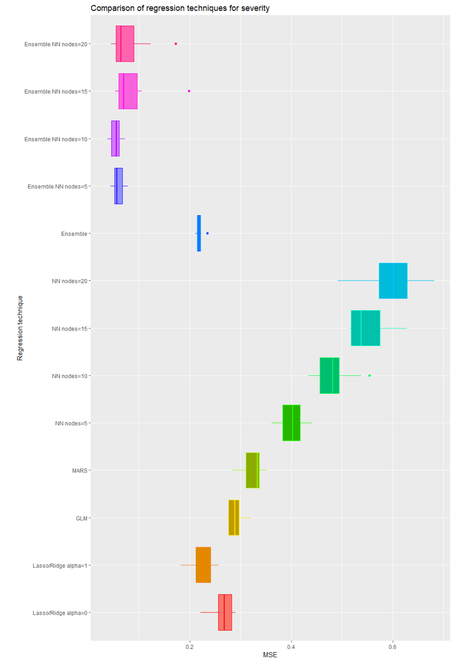


Figure: Box plot of severity regressors

The ensemble techniques used all techniques except neural nets and $\alpha = 0, 1$ to just use pure lasso and ridge regression but other values of α were also tested but not included in the ensemble. The ensemble technique using a single hidden layer with 10 nodes performed much better than the regressors on their own or by weighting them equally through calculating their mean.

Ensemble Learning

Ensemble learning combines machine learning methods together in the hope that it'll improve the performance of the classifier or regressor on the data set. Some methods such as random forests are well known ensemble learning techniques but any combination of models can be combined in a simple or complicated way. Examples of ways they can be combined are but not limited to:

- Average output
- Weighted average output
- Neural Networks

Alexander John Pinches
CID:01201653 "The contents of this work and the associated code are my own unless otherwise stated"
Imperial College London