Drug Use Level Prediction

Hu Tang

Imperial College London

The contents of this work and the associated code are my own unless otherwise stated.

Introduction

In this project, we will use a data set that contains:

Predictors:

- individuals background : agegroup, gender, education, country, ethnicity
- scores for personality traits : neuroticism, extraversion, opentoexperience, agreeableness, conscientiousness, impulsiveness, sensation
- legal substances: caffeine, chocolate, nicotine, alcohol

Outcomes:

- Severity : score of the severity of drug consumption.
- Uselevel : ”high” or ”low” use level of drug consumption.

We will conduct a data analysis on this data set:

- Conduct a exploratory data analysis, have a sense of the data-set’s story and explore any underlying correlation.
- Use two machine learning methods to predict Uselevel based on predictors, and record accuracy.
- Analyze on which predictors are important in predicting drug use level.

Exploratory Data Analysis

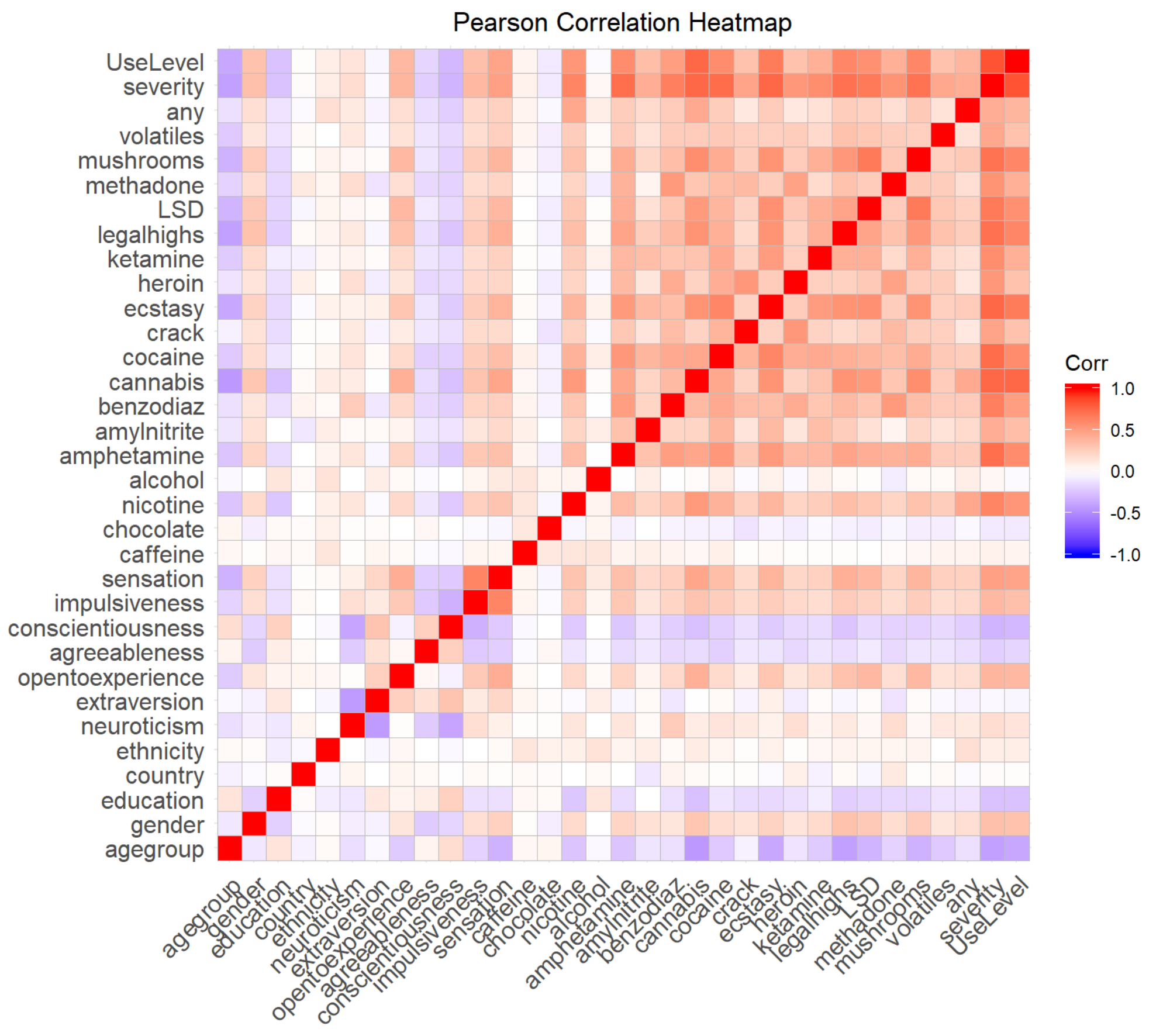


Figure 1: The plot shows that severity and use level(the last two columns on the right) both have strong positive correlations with the illegal drugs(since both severity and uselevel are constructed from them) and nicotine, sensation, impulsiveness, opentoexperience and gender. Also, strong negative correlations with conscientiousness, agreeableness, education and agegroup. Nearly no correlation with alcohol, extraversion and country.

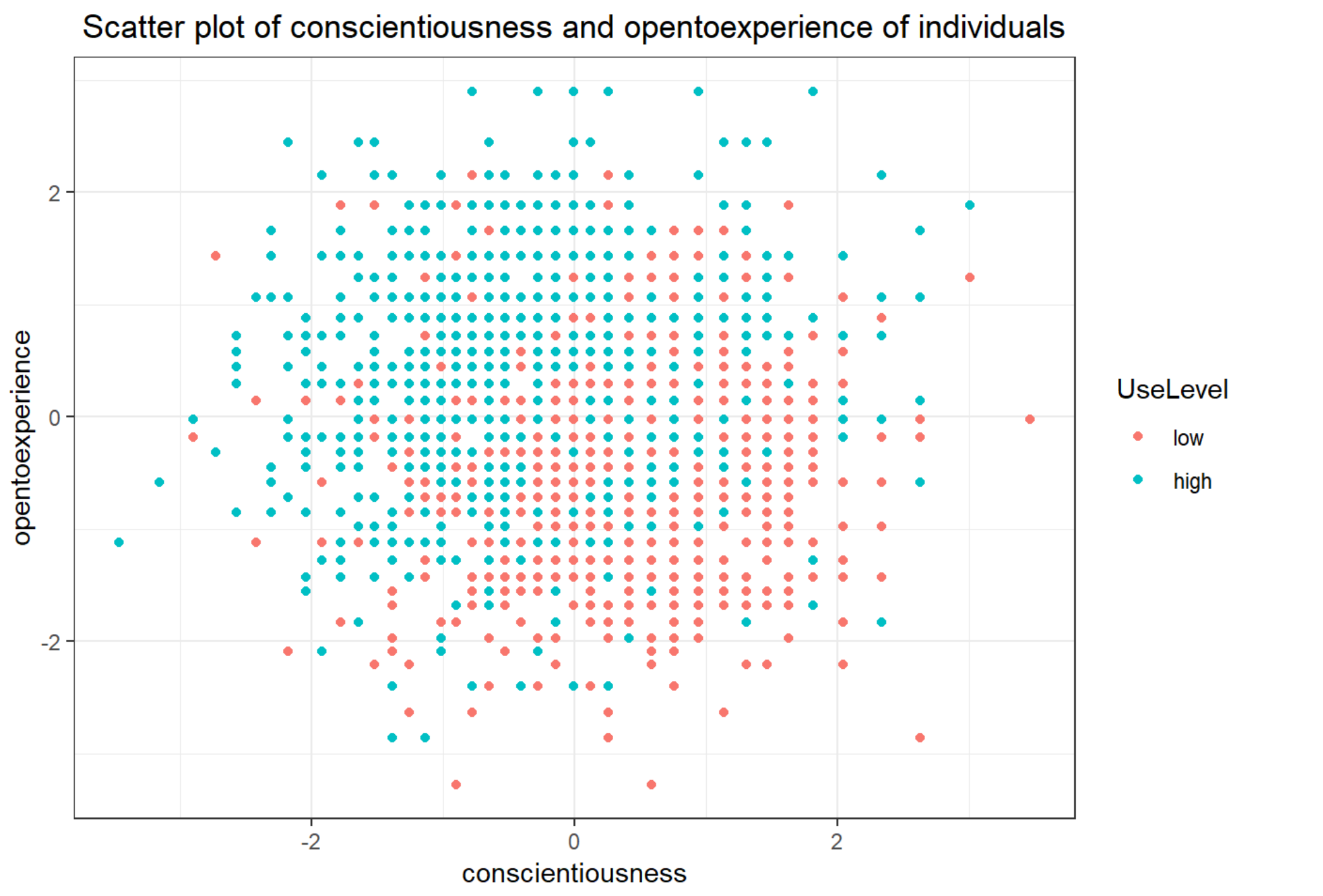


Figure 2: We can see that there is a boundary that separate low and high use levels: opentoexperience and use level has a strong positive correlation, while conscientiousness and use level has a strong negative correlation.

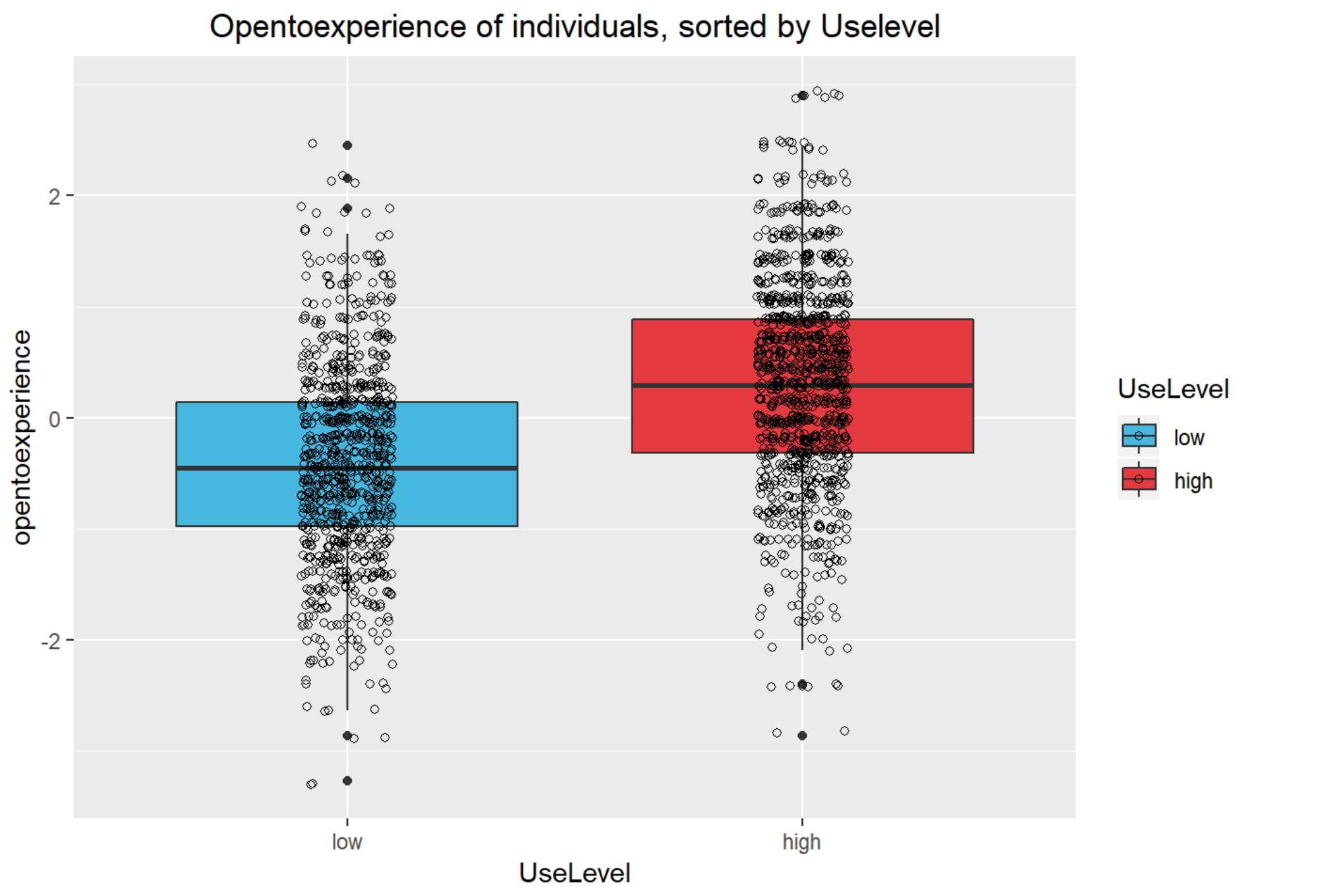


Figure 3: The plot illustrates that use level and opentoexpereince has a strong positive correlation.

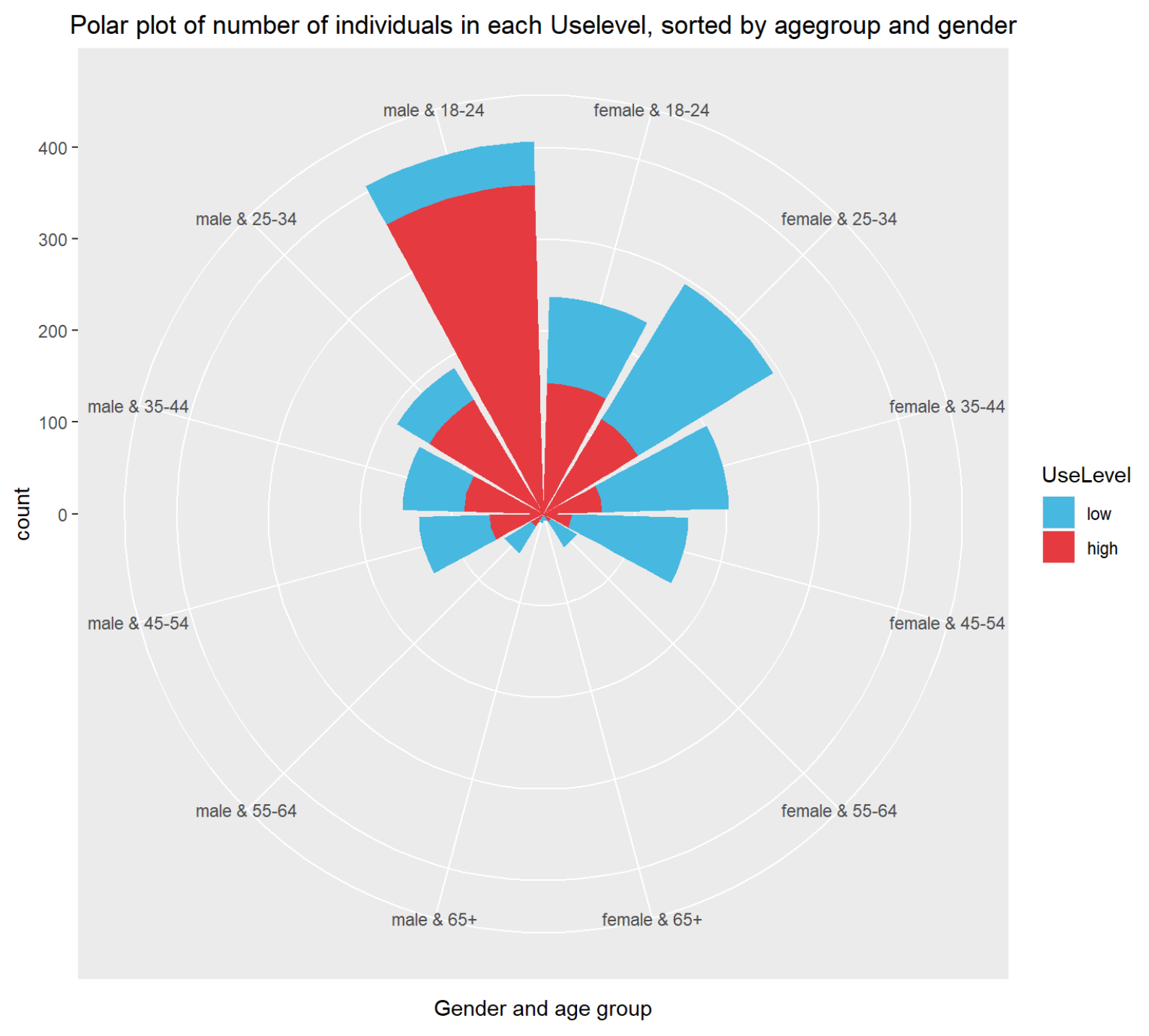


Figure 4: The polar plot shows that younger individuals are more likely to have a high use level, and female tends to have a low use level comparing to male.

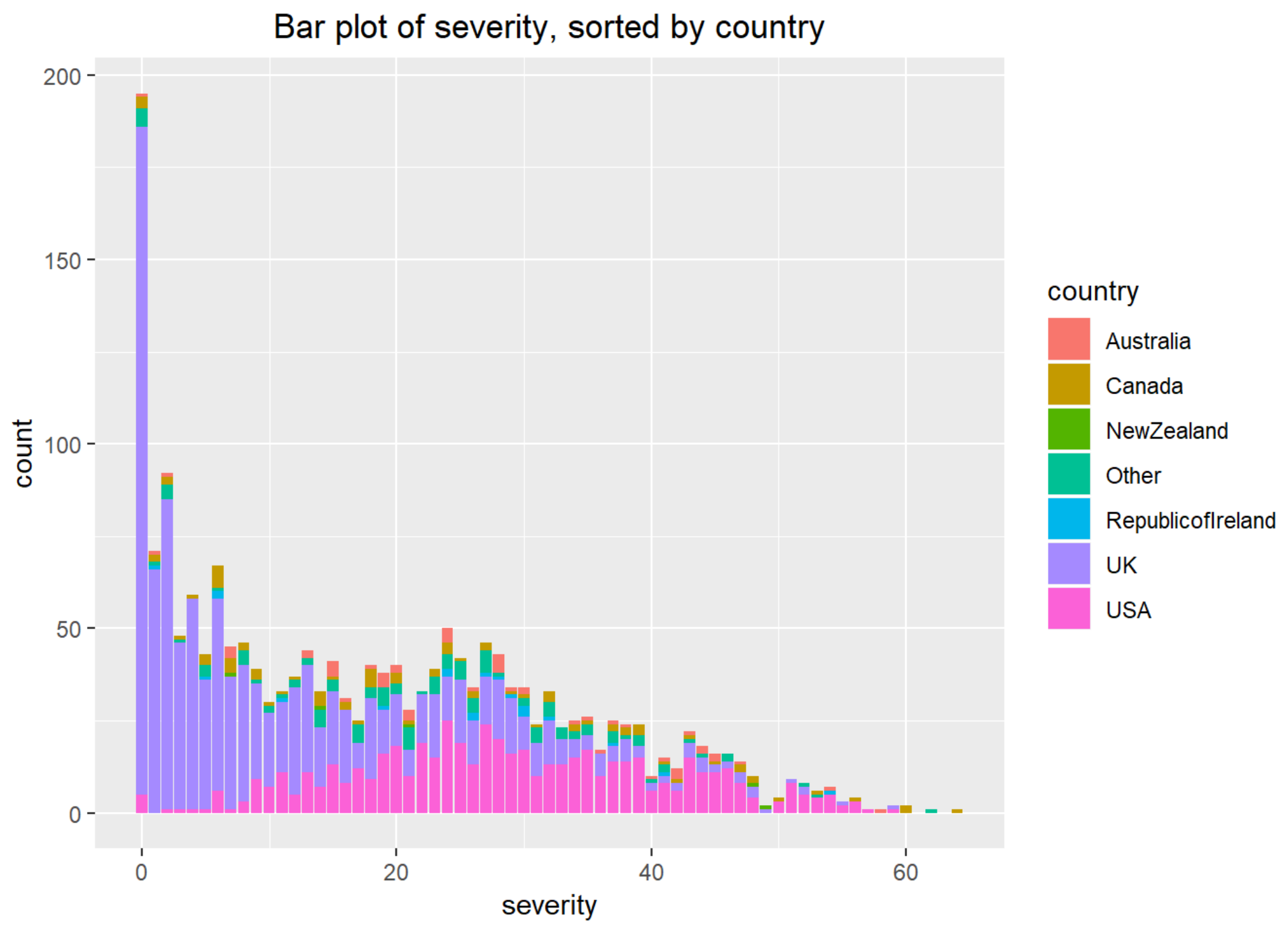


Figure 5: From the plot, we can see that individuals in UK generally have lower severity than individuals in other countries. However individuals in USA have higher severity than individuals in other countries.

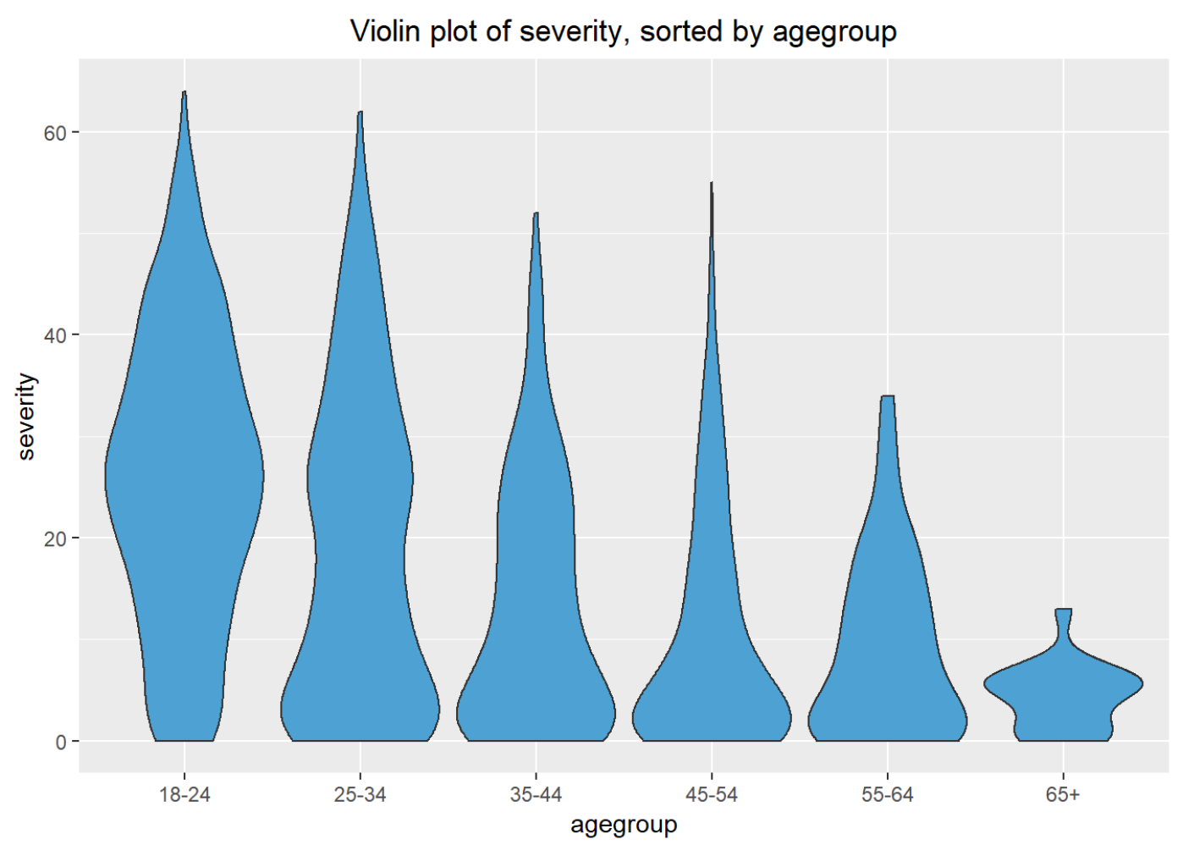


Figure 6: The plot tells us that younger people are more likely have higher severity. There is a strong negative correlation between severity and age group.

From the plots, we found that there are several predictors that are strongly correlated with use level and severity.

Machine Learning

Logistic regression model

Use logistic regression to build a classifier that predicts if an individuals substance use level will be high or low based on the predictors. ROC curve is drawn, and AUC is calculated.

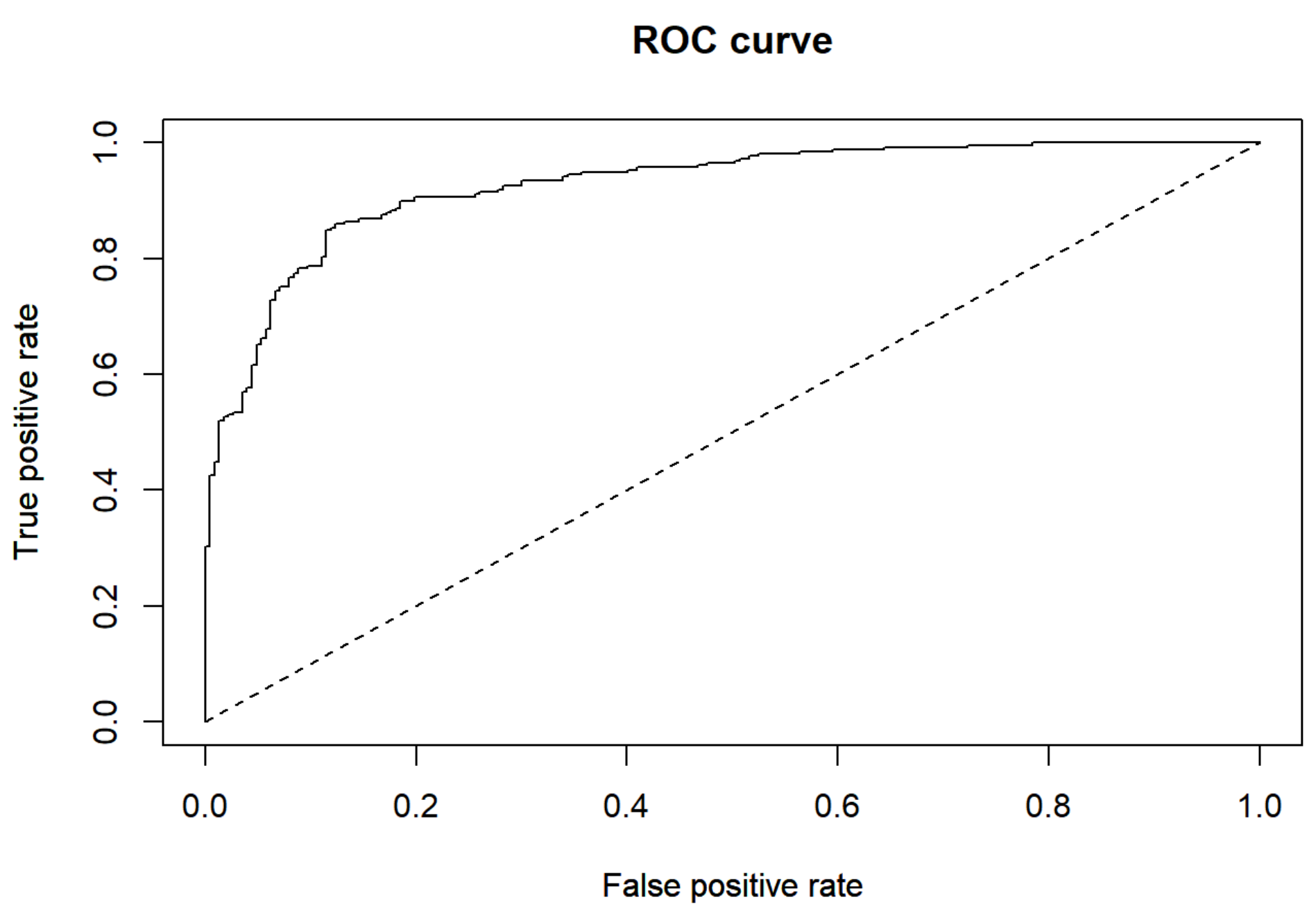
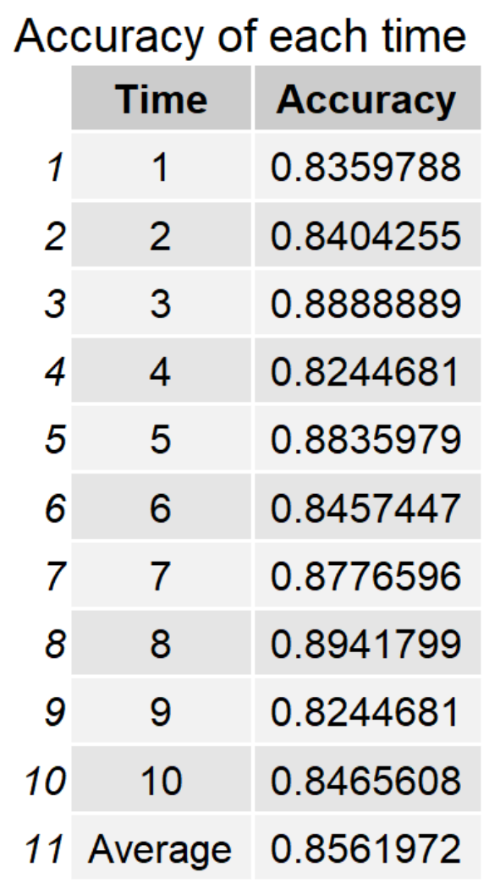


Figure 7: The ROC curve is high above $y = x$, which means it is a good classifier. Also a calculated AUC value of 0.93 also suggests that the classifier performs well.

I then used 10-fold Cross Validation to estimate the accuracy I would see if I use the model on a new data set.



Accuracy of each time

	Time	Accuracy
1	1	0.8359788
2	2	0.8404255
3	3	0.8888889
4	4	0.8244681
5	5	0.8835979
6	6	0.8457447
7	7	0.8776596
8	8	0.8941799
9	9	0.8244681
10	10	0.8465608
11	Average	0.8561972

Figure 8: If I use the model on a new data set, I will get an accuracy of 0.86

Ensemble learning

I will try another model to see if I can achieve a higher accuracy than logistic regression model: use KNN, Random Forest and SVM on training data. For each of these models, I tune the hyper parameter respectively.

Finally, I use these models to predict the test data, and ensemble the three predictions with majority vote.

Confusion matrix is then generated:

		Actual	
		High	Low
Predict	High	194	27
	Low	28	128

This time the accuracy is 0.854, approximately the same as the logistic regression model.

Predictor Importance Analysis

It is essential to say which predictors are important in predicting illegal drug use. For this study, I will discuss the predictor importance in two aspects:

Logistic regression summary

We perform significance test on predictors with significance level of 5%. By comparing the p-values of predictors with 5%, we have enough evidence to say that age group, gender, education, ethnicity, extraversion, opentoexperience, conscientiousness, sensation, chocolate, nicotine and alcohol has association with the use level. We consider them as important predictors.

Random forest variable importance

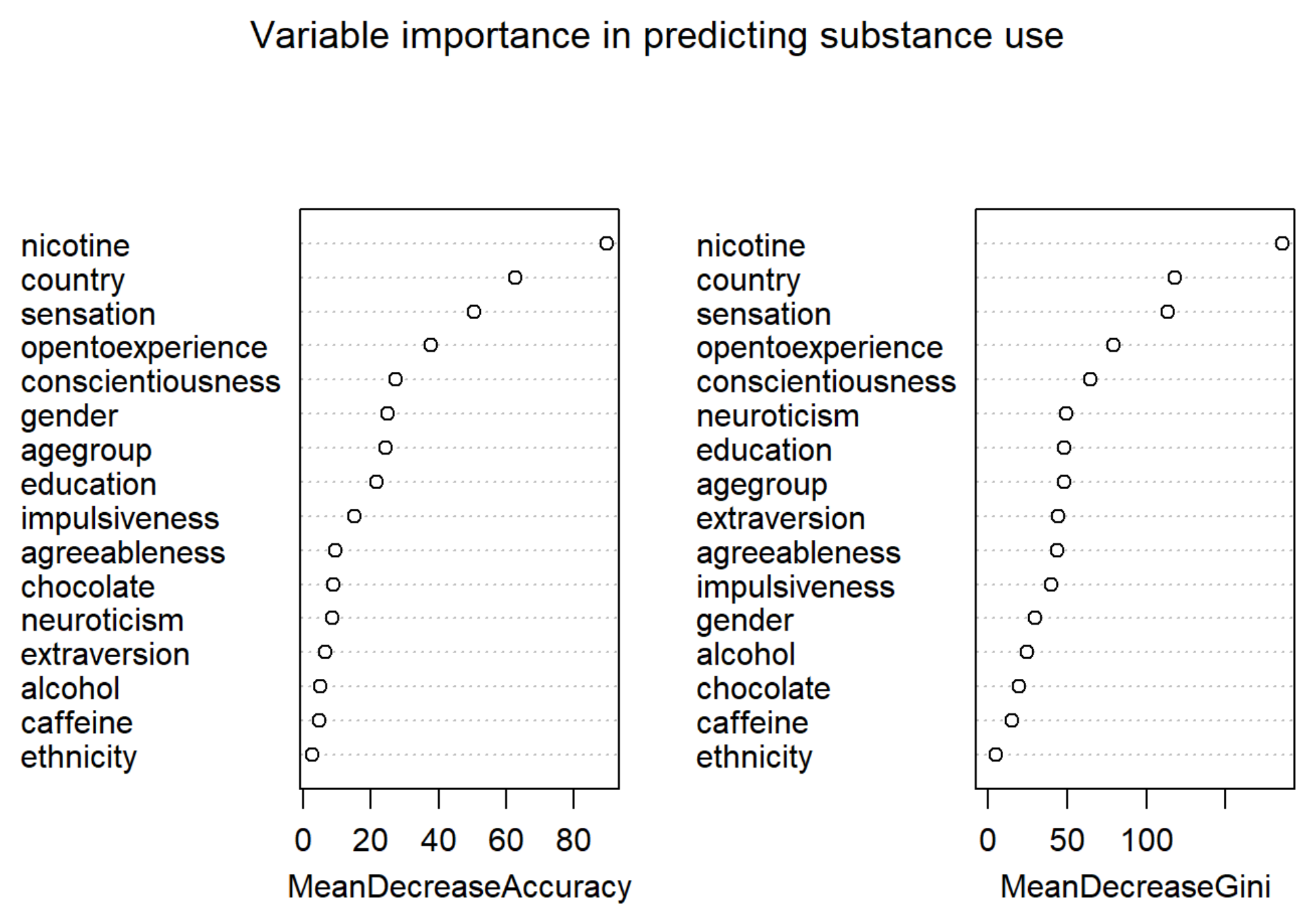


Figure 9: Mean decrease Gini are important index for variable importance. Variables with a large mean decrease in accuracy and mean decrease in Gini are more important for classification of the data. From the graph we can see that, nicotine, country, sensation, opentoexperience, conscientiousness, age group, gender, education and neuroticism are important predictors.

In conclusion, nicotine use, sensation, opentoexperience and conscientiousness personality traits , age group and gender are considered as important predictors in both aspects.

Conclusions

We found that it is accurate to predict individual’s drug use level just from individual’s background and personality traits, with over 85% accuracy. Also, nicotine use, sensation, opentoexperience and conscientiousness personality traits , age group and gender are important predictors in predicting drug use level.