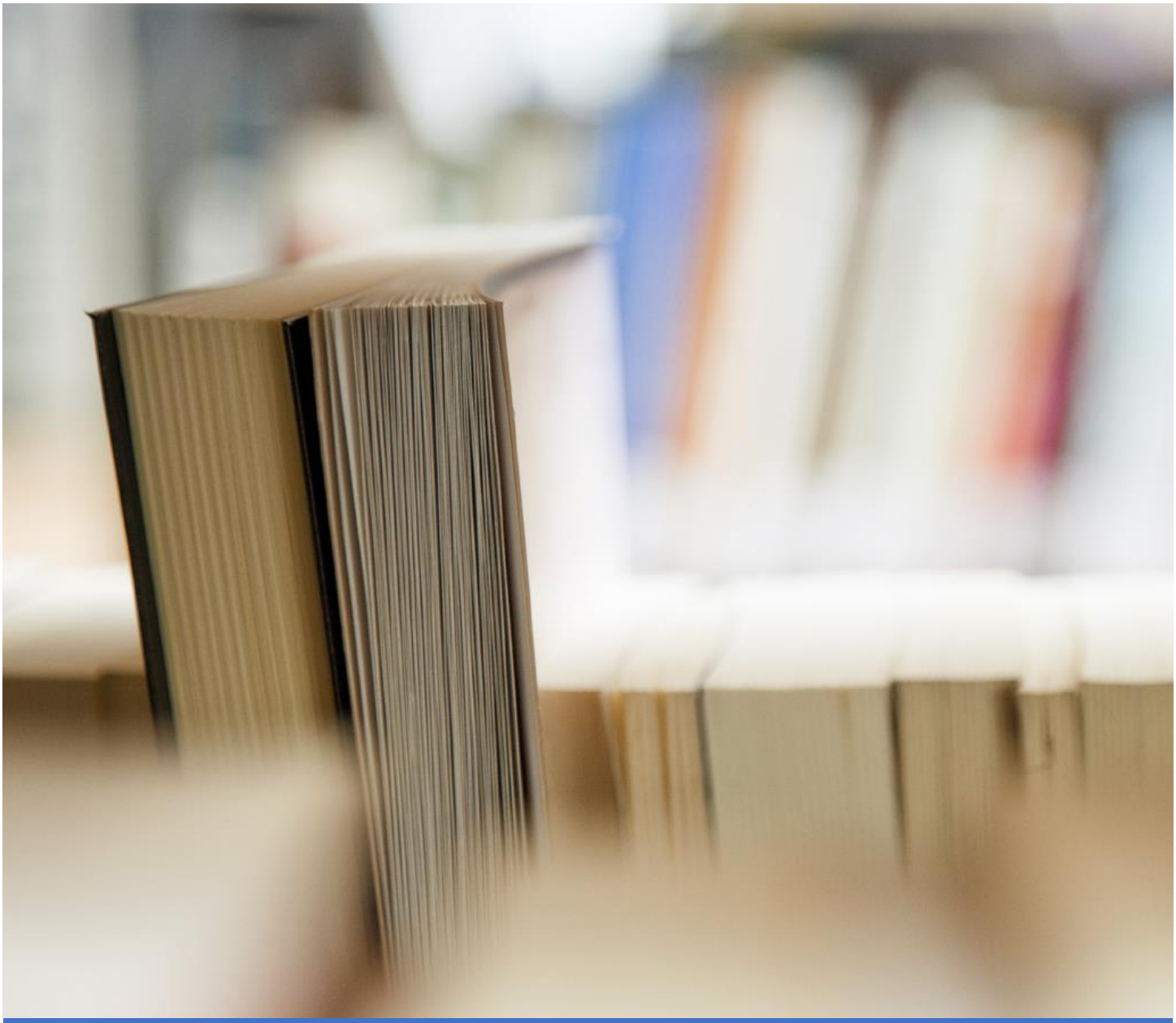


# Data Mining Final Report

## Predicting the Factors of a Drug User

Tim J Ridgway, Louis Hasley, Connor Powers, Nicholas Walsh, Jason Shanker

Data Mining

**Introduction:**

Opioid usage and whether it can lead to addiction is an ongoing debate. Many people believe that the use of prescription medicine can be an introductory drug to becoming addicted to illegal drugs. For example, an estimated 4 - 6% of people who misuse their opioid prescription transition into Heroin, and an estimated 80% of people who have used Heroin had previously misused opioids (Opioid Overdose Crisis). With this existing correlation, our group decided to do some research and analysis on the predicting variables of an opioid drug-user.

Our dataset, found on UCI Machinery Repository, is a classification data problem of supervised learning that uses 11 predictor variables to classify each individual into either a "User" or a "Non-User". The predictors are: age, gender, education, country, ethnicity, neuroticism score, extraversion score, openness to experiences score, agreeableness score, conscientiousness score, and impulsiveness score. These factors will help solve our overall business problems, which are, "What are some of the main indicators of a drug user? How can we use this information to lower opioid drug use?"

Our research was conducted using seven models: decision tree, logistic regression, random forest, SVM, neural network, gradient-boost, and ada-boost, however only our top 4 performing models are included in this report. We hope that our findings could be useful to government policy makers, hospitals, social workers, and other people affected by drug users in order to help reduce the overall drug use, specifically opioid addiction.

**Data Preparation:**

After we had identified a business problem, discovered a dataset, and identified the data mining problem, our preparation and cleaning of the dataset began. Fortunately, there were no missing values to deal with. The original dataset had many drug categories, and each was listed as either a user or a non-user based on the 11 predictor variables we chose. To specify our research for the opioid crisis, we reduced the dataset to only opioid drugs. We created a new target variable of "User/Non-User" that classified an individual as a "User" if they used any of the opioid drugs within the past year, and a "Non-User" only if they were non-users for all opioid drugs in the dataset.

**Data Definitions:**

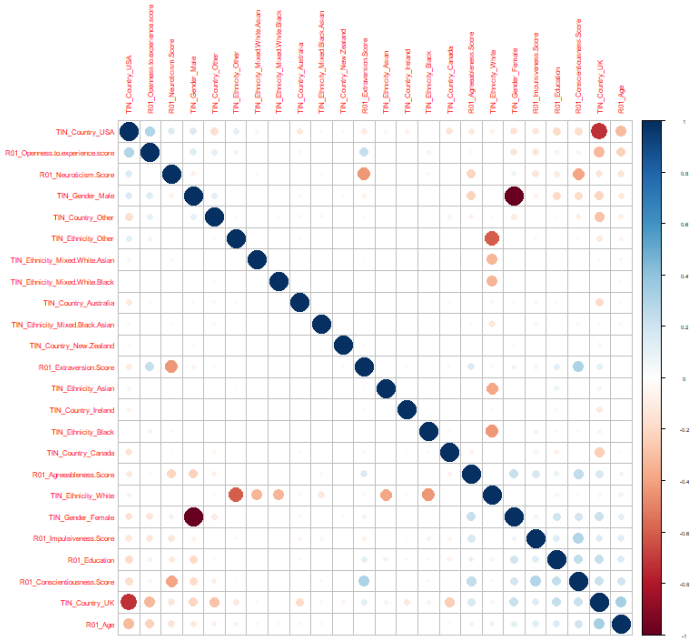
*Neuroticism*: People with a high score are more likely to be moody with feelings of anxiety, anger, frustration, depression, mood, and loneliness. People with low scores are emotionally stable and calm.

*Extraversion*: People with a high score are energetic, talkative, and tend to seek stimulation in the company of others.

*Openness to Experience*: People with high a score tend to seek fulfillment through arts and creativity while people with low scores seek fulfillment through perseverance.

*Agreeableness*: People with a high score are more compassionate, cooperative, and trusting of others. People with a low score are competitive or challenging as they typically are argumentative.
*Conscientiousness*: People with a high score indicate that a person is stubborn and focused, while low scores indicate flexibility and spontaneity.
*Impulsiveness*: People with a high score indicate acting spontaneously without thinking or planning and low scores indicate thought and planning before acting.

**Exploratory Analysis:**



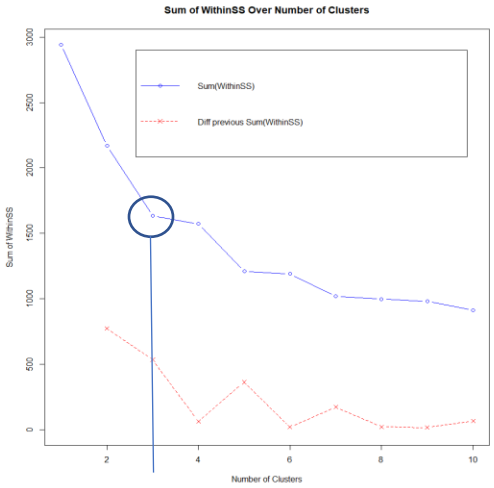

Looking at the graph, there are not a lot of strongly correlated relationships. However, a few boxes stand out. The correlation of Country_UK and Country_USA are stongly negative. We believe the correlation is high because the two countries make up almost 75% of the dataset and a person cannot be both, making it a strong negative correlation. The same intuition applies to the negative correlations between the different ethnicities and genders. These correlations do not provide much insight as the strong relationship is about mutual exclusion.

There are some interesting relationships in the given correlation plot. Responders from the United States had the highest correlation to their openness to experience. It is a positive relationship, meaning the United States responders were more open to experiences. The United States also has a negative correlation with age, making their responders younger. The opposite correlation is found in the UK responders as the correlation between UK and openness to experience is negative and high. UK responders also have a positive correlation with age, making their country's responders older. The

opposite correlations between the United States and UK in age may describe the different trends in their respective openness to experience.

Some score indicators have a correlation that gives insight into the responders. Extraversion score (outgoing, social, “life of the party”) and Neuroticism Score (more likely to experience feelings such as anxiety, jealousy, etc.) are negatively correlated. This could mean responders that score high in being outgoing feel less stress and do not care as much on others perception of them. Neuroticism Score is also strongly correlated negatively with Conscientiousness (wanting to succeed in their current position or career). The relationship between Conscientiousness and Extraversion is positively correlated. This describes the dichotomy of having social skills to succeed at a job. Conscientiousness also shares the same correlation with agreeableness and impulsiveness. Impulsiveness (acting on a whim) has a strong, positive correlation with conscientiousness. This means people who want to succeed have a tendency to try new things at a higher rate.

**Clusters:**




```
Cluster centers:

    R01_Age R01_Education TIN_Gender_Female TIN_Gender_Male TIN_Country_Australia TIN_Country_Canada TIN_Country_Ireland
1 0.3398229     0.6255731                 1               0           0.00000000         0.00000000          0.00000000
2 0.2392031     0.5279678                 1               0           0.06583072         0.12852665          0.02821317
3 0.2510911     0.5079547                 0               1           0.03499470         0.04878049          0.01166490
  TIN_Country_New.Zealand TIN_Country_Other TIN_Country_UK TIN_Country_USA TIN_Ethnicity_Asian TIN_Ethnicity_Black
1             0.000000000        0.00000000       1.0000000       0.0000000          0.01765650         0.022471910
2             0.003134796        0.11285266       0.0000000       0.6614420          0.01253918         0.009404389
3             0.004241782        0.08695652       0.4464475       0.3669141          0.01166490         0.016967126
  TIN_Ethnicity_Mixed.Black.Asian TIN_Ethnicity_Mixed.White.Asian TIN_Ethnicity_Mixed.White.Black TIN_Ethnicity_Other
1                     0.000000000                     0.009630819                     0.014446228          0.01926164
2                     0.006269592                     0.015673981                     0.006269592          0.05329154
3                     0.001060445                     0.009544008                     0.009544008          0.03605514
  TIN_Ethnicity_White R01_Neuroticism.Score R01_Extraversion.Score R01_Openness.to.experience.score
1           0.9165329             0.4921415              0.5698981                       0.5358480
2           0.8965517             0.5500914              0.5345192                       0.6651863
3           0.9151644             0.4849770              0.5386569                       0.6293449
  R01_Agreeableness.Score R01_Conscientiousness.Score R01_Impulsiveness.Score
1               0.6851926                  0.6393793               0.7483441
2               0.6480538                  0.5595611               0.6845566
3               0.6135118                  0.5513559               0.6975659
```

Cluster 1: Oldest cluster with highest education, Females from only UK, most Extraversion, least open to experience, agreeable, high conscientiousness and impulsive

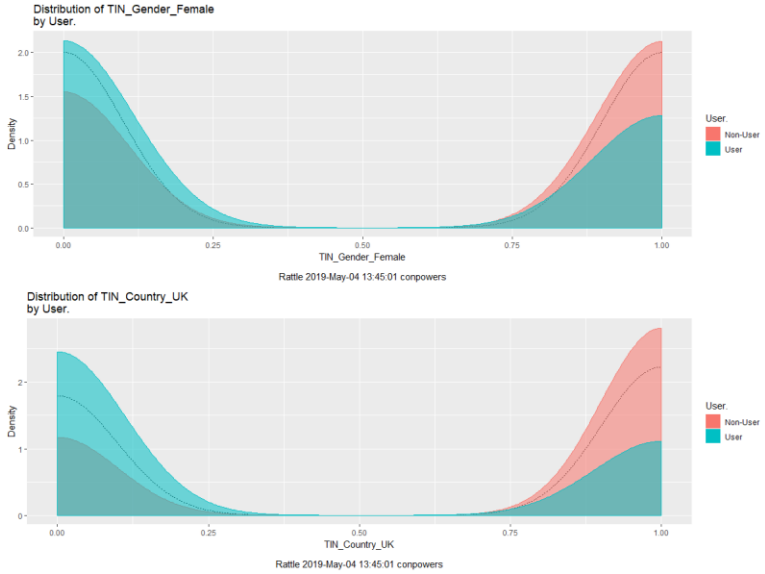
Cluster 2: Youngest cluster, Females, 66% from USA and rest from other countries (No UK), highest neuroticism and openness to experience.

Cluster 3: Least educated cluster, Males, 45% UK and 37% USA with the rest from other countries, open to experiences, lease agreeable

Psychology scores did not have a lot of variance between clusters.

**Density Plots of Clusters:**

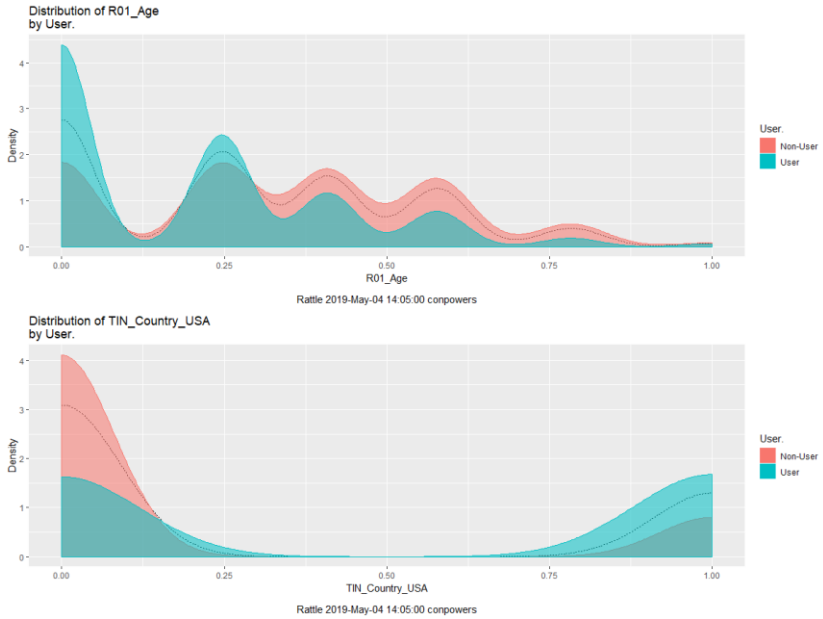
Cluster 1




Females are more likely to be non-users compared to their male counterparts.

Responders from UK are more likely to be non-users compared to other countries.

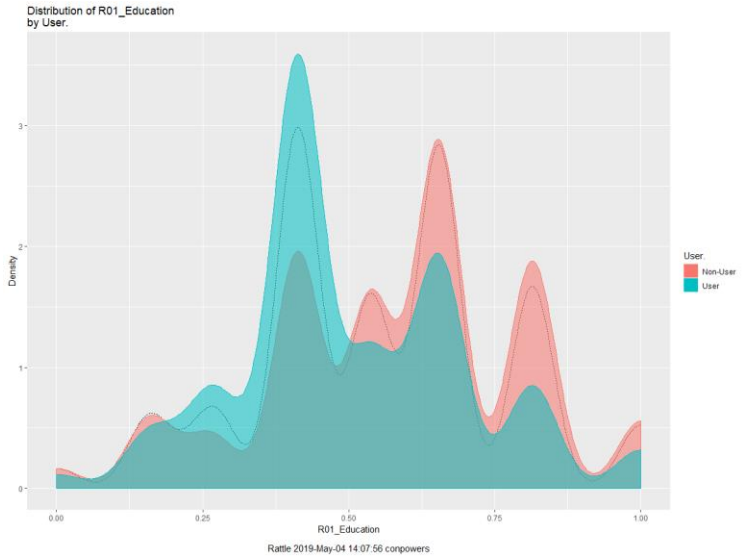
Cluster 2:



The younger the user, the more likely they are to be a user.

Responders from the United States are more likely to be users compared to other countries.

Cluster 3:




The value of 0.5 splits the responders by having a university degree or none. If a responder does not have a degree, they are more likely to be a user. The more education a responder has, the less chance they have of using opioids.

**Models**:

We chose to evaluate seven models in order to determine which is the most effective in determining predictor variables for opioid usage. The models were: SVM, ada-boost, gradient-boost, neural network, linear regression, decision tree, and random forest.

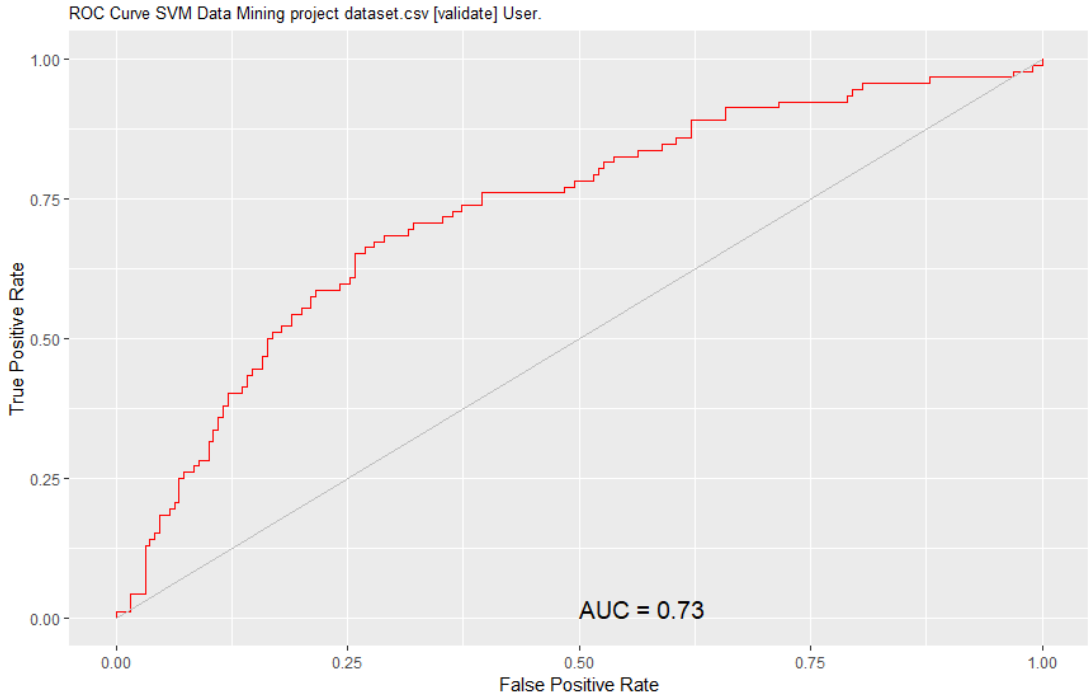
**SVM**

| Split | Seed | RBF C=15 |
|---|---|---|
| 1 | 42 | 29.5% |
| 2 | 420 | 30.2% |
| 3 | 1029 | 29.8% |
| Average | | 29.83% |
| Std. Dev. | | 0.0035 |

The SVM model with the lowest average error that we selected was the RBF kernel with a complexity of 15. The SVM model was used as a linear discriminant to classify users & non-users of opioids for our dataset. The large C we chose tells us that the SVM model fits the data as close as possible in order to minimize hinge loss. We calculated statistics from the error matrix and our most significant was accuracy of 71% from this model. Our precision and recall scores were 54% and 66%, respectively.

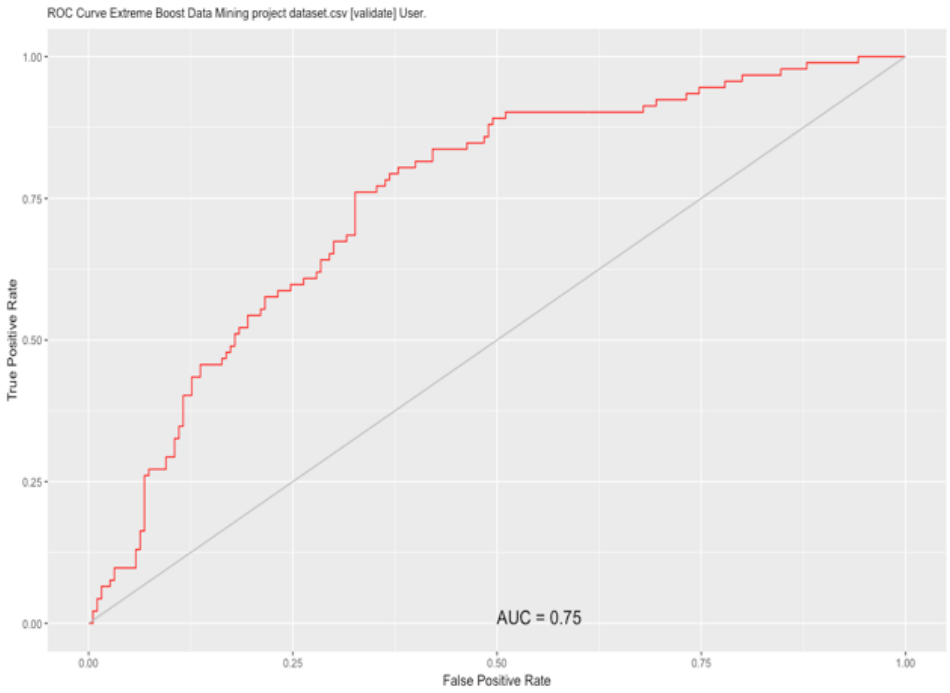
**ADA Boost**

The ADA boost model performed best with a complexity of 0.075 which resulted in an AUC of 0.7498. An AUC of 1 is an ideally perfect model, so around 75% of that for our model is a significant number. This tells us that ADA boost is a good model for predicting users and non-users of our data set.




**Gradient Boost**

The gradient model performed best with the Maximum Depth set to 6, 2 Threads, and a Learning Rate of 0.1. The results of this give an AUC of 0.7493 which indicate that this model is a good predictor of users and non-users.



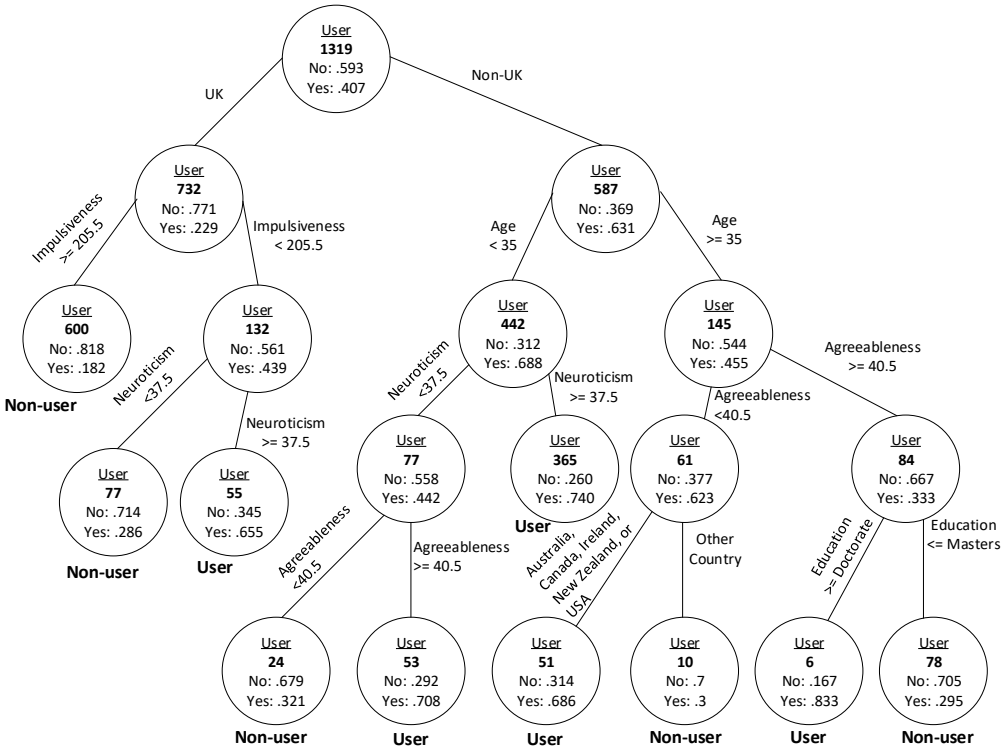

**Linear Regression:**

```
Log likelihood: -695.273 (22 df)
Null/Residual deviance difference: 392.203 (21 df)
Chi-square p-value: 0.00000000
Pseudo R-Square (optimistic): 0.52205097
```

Since our target variable was categorical, linear regression was not a significant model towards our implications. There were few visuals that were applicable to demonstrating the relevance of the model because we did not have a cost matrix that we could develop a cost curve from. However, we did run the model and go an R-square value of .522. This number is insignificant which further backs our initial expectation that linear regression would provide us with little benefit when it comes to choosing a good model of predicting an opioid user.

**Decision Tree**

After relaxing the tuning parameters to the appropriate maximum and minimums, the following table was produced a table with various model complexities and their cross-validation errors. Examining that table, we picked the complexity of 0.0074 as it had the lowest cross-validation error of 0.6648. The 0.0074 was used the complexity parameter and the other parameters remained as they are. After running this model, rattle produced the following decision tree:




This tree has 10 terminal nodes with splits based on Country, Impulsiveness, Neuroticism, Agreeableness, and Education level. Using this model, the performance of this model was evaluated on the testing set of data using an error matrix, shown below.

| **Counts** | | Predicted | |
|---|---|---|---|
| | | *User* | *Non-User* |
| Actual | *User* | 67 | 30 |
| | *Non-user* | 50 | 137 |

| **Probabilities** | | Predicted | |
|---|---|---|---|
| | | *User* | *Non-User* |
| Actual | *User* | 0.236 | 0.106 |
| | *Non-user* | 0.176 | 0.482 |

Evaluating this error matrix, we found that this model had an Accuracy of 72%, Precison of 57%, and Recall of 69%.

**Random Forest**

| Seed | # of Trees | # of Variables | AUC |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 42 | 600 | 3 | 0.7542 |
| 200 | 500 | 2 | 0.7859 |
| 15 | 600 | 2 | 0.7737 |

Using the average AUC of all cross-validation sets, the best model was 600 trees and 3 variables. The AUC of 0.7542 was the highest amongst all models with seed 42.

Evaluation of error matrix: Precision of 54%, Recall of 66%, Accuracy of 71% and a F1 Score of 60%.

**Business Insights**

With our models, we determined that impulsiveness, neuroticism, and agreeableness scores are the strongest psychological traits that predict opioid use. With this information, government officials, hospitals, and social workers can identify individuals that have low impulsiveness scores, high neuroticism scores, and low agreeableness scores as they are the most likely to be opioid users. Not only can those authoritative figures use this information, but individuals as well. If individuals are well-aware of their personality traits or have taken these and know their scores, they can take precautions when being prescribed painkillers by doctors in order to avoid being part of the 4-6% of people that transition into heroin after misusing opioids. One last insight we found was that our Precision scores were relatively low compared to the other metrics.

Although a low precision score indicates our model predicts individuals to be a User when they are a Non-User, we determined that the cost of this is significantly lower than the cost of predicting someone to be a Non-User when, in fact, they are a User. In summary, we think it is better that our model has a lower Precision than Recall score. In conclusion, we think that our models perform well for predicting opioid use based on demographics and psychological scores as "AUC of 0.7 is sought after in applied psychology and prediction of future behavior" (Rice and Harris, 2005).

**Recommendations**

To put these business insights into action, we would recommend providing doctors with this information so they can better understand their patients, especially when prescribing opioid drugs to treat pain. For doctors to realistically use this model, they would have to have these psychological scores for their patients, so we would recommend making this part of medical practice by implementing psychological tests for patients on regular basis. Finally, we would recommend identifying that have low/high scores on certain psychological test and recommend programs that improve general mental wellness.

**References**

https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis

https://www.researchgate.net/publication/7511660_Rice_ME_Harris_GTComparing_effect_sizes_in_follow-up_studies_ROC_Area_Cohen's_d_and_r_Law_Hum_Behav_29_615-620