Classifying Drug Use in America

Maxwell Waun

NetID: A52842951

Project Category: General Machine Learning

Final Report

1 Introduction

Drug use has long been a topic of conversation within American politics. From Prohibition in the 1920s to the recent wave of state legalization of marijuana, America has been presented with the issue of drug use and have tried different attempt to combat it. Using logistic regression, K-Nearest Neighbors, Shrinkage Methods, Best Subset Selection and Tree based classification, this project aims to develop a model that predicts the gender of each observation listed in the data. These models will use prior drug use, impulsiveness and other features to help predict the person's age, which is a factor of whether the person is above the age of 35 or not.

2 Related Work

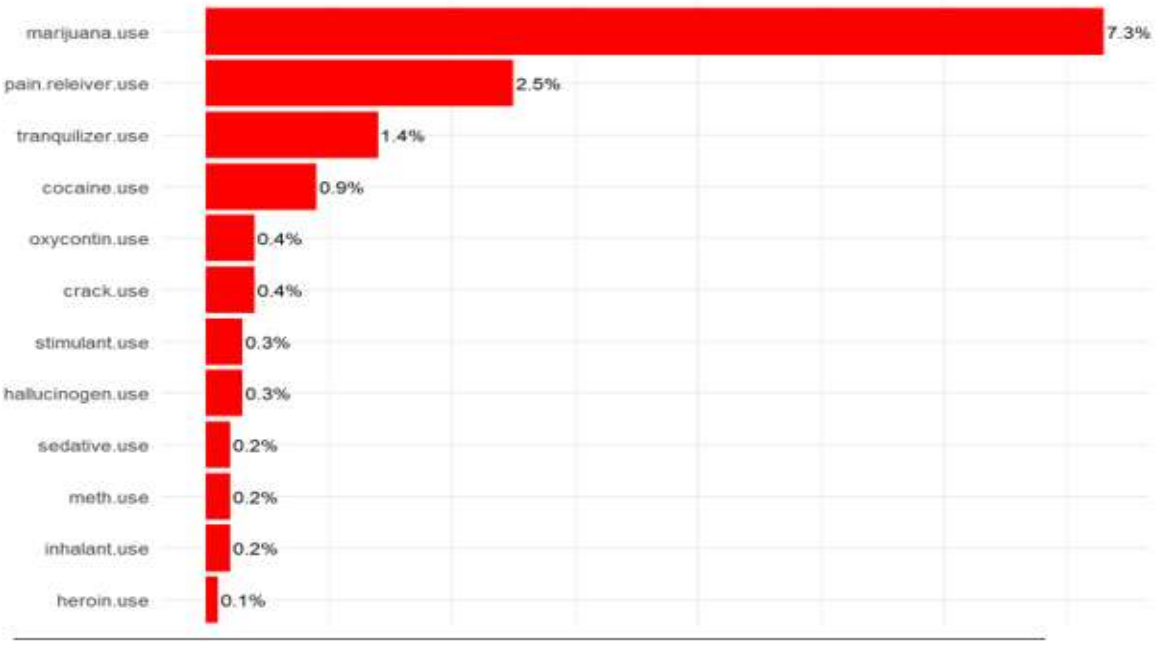
Existing work regarding drug consumption reports on 'baby boomers' and their rates of drug use compared to younger generations. This analysis was done by Anna Maria Barry-Jester[1]. In the scope of Jester's analysis, a baby boomer is someone born roughly between 1946 and 1965. Thus, baby boomers were categorized as being between 50 and 65 years old. The data used

goal summary.

tested for whether baby boomers used the following drugs: marijuana, pain relievers, tranquilizers, cocaine, crack, OxyContin, stimulants, hallucinogens, sedatives, inhalants, meth and heroin[1]. Jester thought it would be best to exclude alcohol from the graph (Figure 1) since alcohol use among boomers was 67% and would subsequently dwarf the other drugs in question. With alcohol excluded, it was determined that the marijuana and non-prescribed pain relievers were the top two drugs most used in the past year by boomers. However, it was found that boomers not only had the lowest rates of drug use compared to all the other age groups, but that they also use less of each individual drug as well.

Figure 1: % of boomers who used drugs in the last year [Age = 50yr-65yr)

Data from Substance Abuse and Mental Health Data Archive (SAMHDA)



Jester concludes that boomers use less drugs less frequently because they use them for different reasons than younger generations. Rather than consuming drugs in social settings or to simply 'get high', Jester argues that they instead use it as a coping mechanism for the ailments that come with old age: social isolation, chronic pain, loss or even the effects of a disease such as Alzheimer's. No models appear to be built by Jester to answer any further questions. Therefore, multiple models will be built to determine whether someone's drug use can accurately predict their respective age group. By doing this, we are expanding Jester's question by including observations of people 35 years and older.

[1]https://github.com/fivethirtyeight/data/tree/master/drug-use-by-age

3 Dataset

The data source used is UCI Machine Learning's Drug Consumption dataset[2]. Within this dataset there are 1,884 observations with 32 features each. To recreate the data in Jester's analysis, the relevant columns were selected, and observations were filtered by their age values. For this dataset, 45 years – 64 years is the closest age group available that reflects similarly to Figure 1. Figure 2 displays the percentage of those aged 45-64 who used each drug in the last year.

Figure 2: % of boomers who used drugs in the last year [Age = 45yr–64yr]

Data from UCI Machine Learning Repository

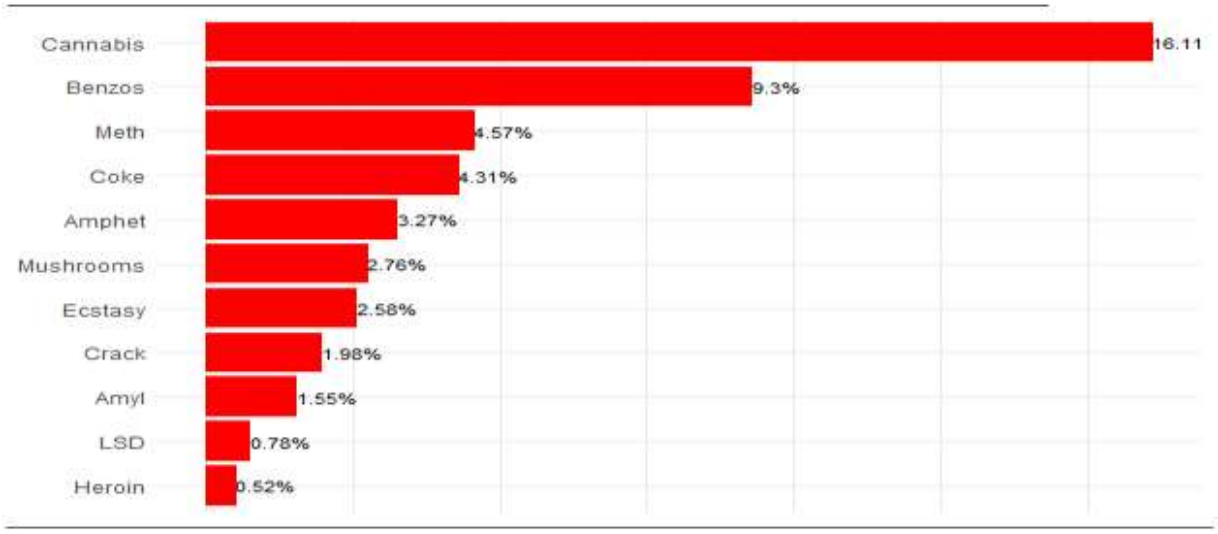


Figure 2 follows a similar trend as figure 1, where marijuana is the most used drug among boomers followed by non-prescribed medication. The 2016 data shows that boomers have increased their drug use since 2012, resulting in their use rates for each drug to increase by over a factor of two. This could be since the bounds for what is considered a boomer is larger in figure 2 than in figure 1. Another reason could be that drugs like marijuana started becoming legal in 2016 in a select few states, resulting in a higher rate of use among boomers. The models built off this data will aim to further discover the relation between drug use and age as this analysis only appears to scratch the surface of the question at hand. However, before models could be built, the data needed to be preprocessed. Preprocessing this data first required renaming the column names, followed by recoding the values of each observation from nonsensical doubles stored as characters into their respective values. Then, before converting everything to numeric, a copy of the dataset was saved in all characters so that there was a descriptive dataset on hand with the same indexes as the numeric dataset. The numeric dataset was then normalized for future models.

[2]https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29

4 Methods

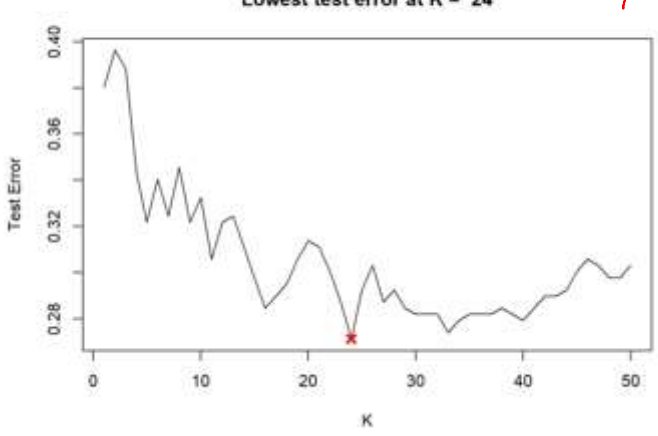
The data was split into a training and testing set. The training set contained 80% of the observations while the test set contained the other 20%. Five different models were then built to predict age group based on drug use. The models used to predict a person's age group based on their drug use were K-Nearest Neighbor (KNN), Logistic Regression, Ridge Regression, Lasso Regression and Tree based classification methods. To determine the accuracy of each model, test error averages were obtained and used to determine which model was best for the data based on which model resulted in the lowest test error.

4.1 KNN

K-Nearest Neighbors is a classification method that classifies cases based on a majority vote of its K nearest neighbors. This distance between neighbors is typically a Euclidean distance. For this project, the normalized training data was fit for different values of K ranging from 1 to 50. Then, predictions were made on the model and were compared to the actual classifications. Figure 3 shows the test error rate for the model at each K value.

Figure 3

Data from UCI Machine Learning Repository



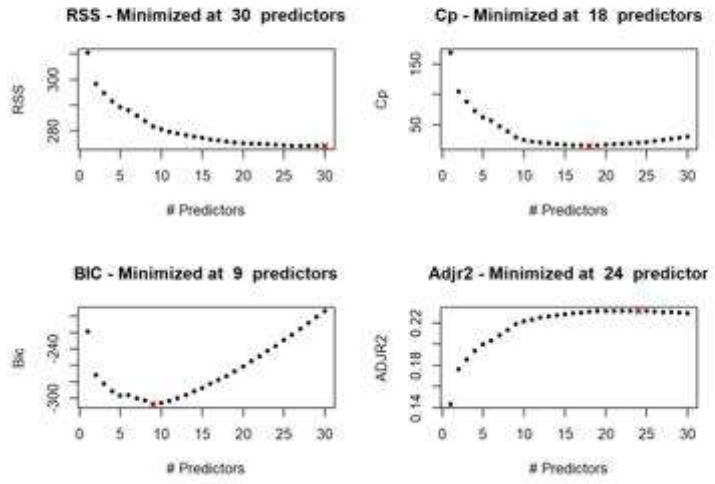
4.2 Logistic Regression

Logistic Regression was another method included in this report due to it being a binomial classification method. The model was fit to all the predictors initially before feature selection. Afterwards, best subset selection was chosen to find the best-fitting models for each number of predictors. Regression model accuracy metrics are then used to determine which model is best for the data. Figure 4 shows accuracy metrics plot for the fitted model. Using adjusted r-squared as our accuracy metric of choice, logistic regression was then fit again using the model that

corresponds to the largest adjusted r-squared value. Finally, linear discriminant analysis and quadratic discriminant analysis were ran using the optimal features found. Each of these models were then used to predict classification values for age and their test error rates were obtained to determine their accuracy.

Figure 4

Data from UCI Machine Learning Repository



4.3 Ridge Regression & Lasso

you are doing classification. How do you implement Lasso?

Ridge Regression and Lasso are similar to linear regression except they add a penalty term in order to linearly separate the data and reduce overfitting. The difference between the two methods is Lasso can reduce the coefficients to zero, where Ridge Regression can only reduce them towards zero without them ever being equal to zero. Using the leaps library[2], both regressions were fit on the normalized training data and had the same features found earlier using best subset selection. Like previous methods, predictions were obtained, classified and tested against the test data to obtain their respective test error rates.

4.4 Tree-Based Classification

Classification trees work by identifying and partitioning the space in which the data exists by creating hyperplanes. By recursively diving the space into smaller partitions and assigning them a class label, a decision tree can be made. With the tree package[3], Figure 5 is obtained by changing our target variable to a factor and fitting the classification model to the training data. Cross validation is then preformed in order to determine the optimal level of tree complexity. The corresponding values from cross validation are plotted in Figure 6.

response?

Figure 5

Data from UCI Machine Learning Repository

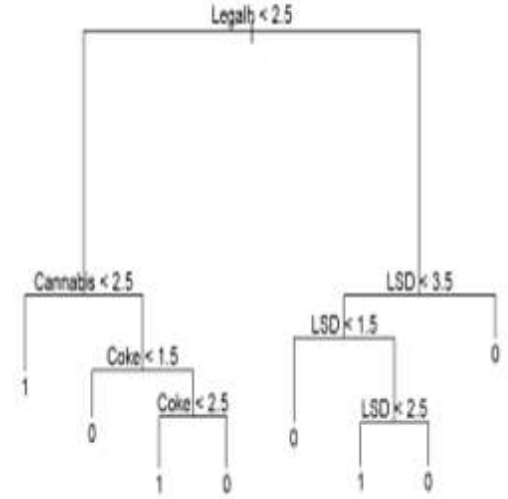
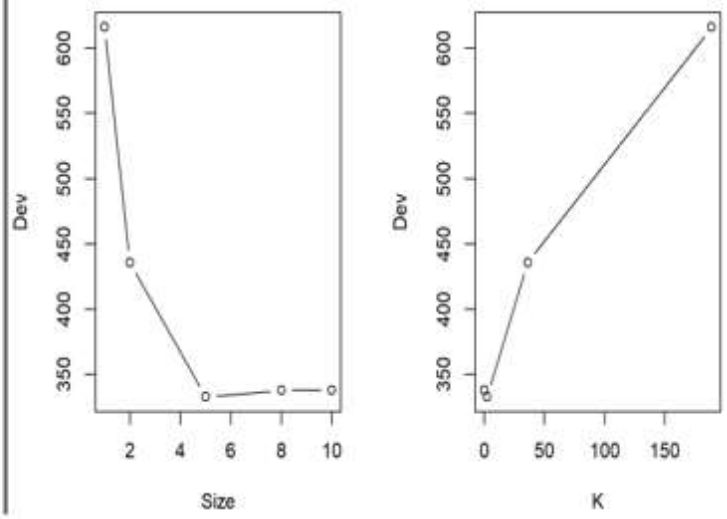


Figure 6

Data from UCI Machine Learning Repository



The 10-fold cross validation returned the most optimal tree to use, and that tree was fit and used to obtain predictions on the test set. The test error rate was then obtained after comparing the predictions to the test values. After obtaining the test error rates, Bagging and Random Forest methods were also performed on the training data. Predicted values were obtained from these fits and were compared to the test values for age to obtain a test error rate for each.

4.5 Gradient Boosting

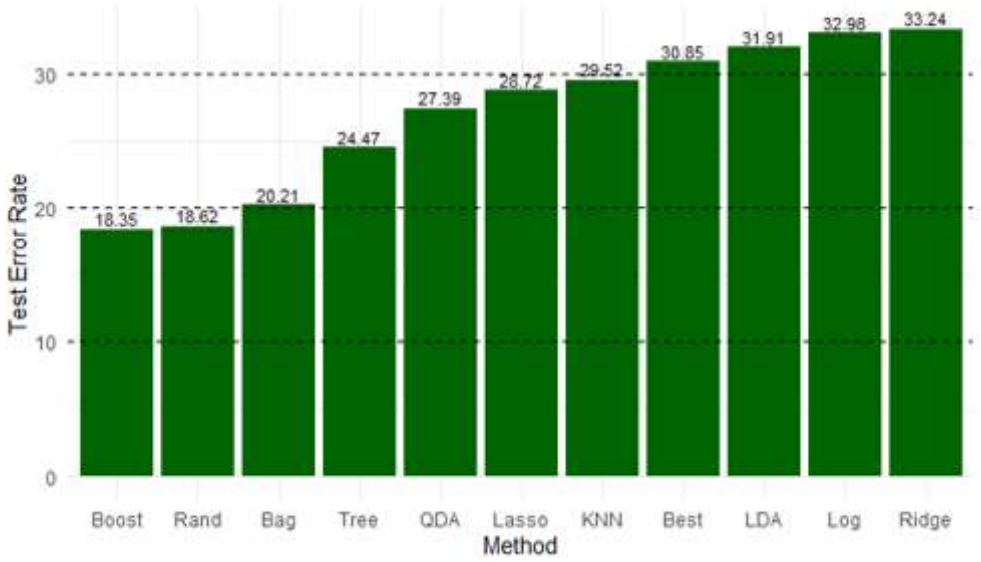
Gradient Boosting was the final method selected as it was thought to most likely outperform the previous models. It is a type of machine learning boosting that combines multiple different models together in order to minimize the overall prediction error, where each model is built with the intention of accurately predicting the cases where the previous model performed poorly. In order to use gradient boosting, the age group column first had to be extracted and converted into a binary list. Then, using the generalized boosted regression models package[4], the model was fit to all the features in the training data except for age. Then, predictions were made on the fit and compared to the test values to obtain the test error rate.

5 Results

As previously mentioned, the test error rate was used to evaluate the accuracy of each model. Each model was trained on 1,503 observations and tested on 376 observations, resulting in an 80%/20% split. Figure 7 displays the test error rate each model achieved.

Figure 7

Data from UCI Machine Learning Repository



Gradient boosting and random forest ended up having the lowest test error rates of all the models. However, ridge regression and logistic regression seemed to perform the worst. Based upon these results, the 19 predictors chosen by best subset were able to determine the age group of the user with 81.65% accuracy.

6 Discussion and Conclusion

Compared to the data used in Jester's analysis, there are many more questions that can be answered from UCI Machine Learning dataset if one takes the time to clean up the mess that it is naturally. The best subset selection process had the largest adjusted r-squared value with 18 predictors. Those predictors were gender, education, ethnicity, nscore, escore, ascore, ss, amyl, benzos, caff, cannabis, crack, ecstasy, ketamine, legalh, LSD and meth. Now that a model has been built with a relatively low test error rate, new areas of interest can be studied. For instance, now that it is known which features best determine which age group a user is part of, one could question if the personality traits of a boomer best match his peers or instead that of the younger generation given his drug use looks more like the younger generation. Future work on this project could include clustering the age groups using K-Means or instead use a support vector machine to try to answer some of the questions that the models in this project did not.

7 References

[1] Anna Maria Barry-Jester “How Baby Boomers Get High”,
https://fivethirtyeight.com/features/how-baby-boomers-get-high/

[2] Package “Regression Subset Selection”,
https://cran.r-project.org/web/packages/leaps

[3] Package ‘Classification and Regression Trees’
https://cran.r-project.org/web/packages/tree

[4] Package ‘Generalized Boosted Regression Models’,
https://cran.r-project.org/web/packages/gbm