Drug Use Data Prediction

Kyle McCollum
MSUID: A55779233

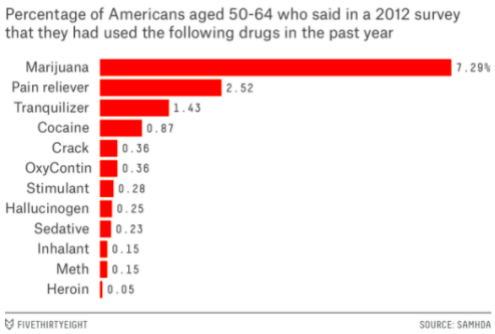
Final Report

1 Introduction

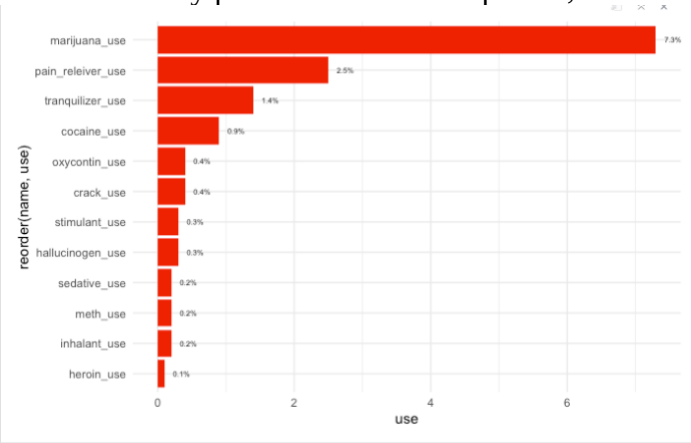
Drug use is a problem all around the world. Nowadays drugs can be acquired rather easily, and the use of drugs has significantly risen as of recent. I think part of the increase in drug use can be related to the blowing up of social media. Superstars all over, especially musicians like rappers, have drugs in their music videos and constantly post on their social media page using drugs. For this project, I am looking to make a prediction based on education, to see if levels of education relate to increased or decreased drug use, using multiple methods including, ridge regression, tree based methods and others.

2 Related Work

Using the drug dataset from GitHub that was provided initially for this project, I set out to regenerate the results of that report. First, I went to the blog page that was linked in the GitHub readme file and spotted the graph that we were supposed to replicate. Here is the plot shown in the blog from GitHub.



Then here is my plot that I made to replicate,



It is not perfect compared to the other plot, but it is pretty close.

3 Dataset

The data set I will be using is the drug consumption data set from UCI machine learning repository. It has 1879 entries and 32 features. For preprocessing on this data set I had to name all of the columns, set the age groups, set the education levels, name the countries listed in the data set, and then list the ethnicities. The data set only had numbers listed for all of the columns, so that had to be changed first. Then, I added a character version of the data to understand the numbers within the columns a little more (I did not end up using this version of the data for my report though).

4 Methods and Results

4.1 Ridge Regression

- I chose to use ridge regression because it was one of the first methods we learned through R this semester and I wanted to go back and revisit it. In this case ridge regression performed quite well on the data set. The test error rate ended up being equal to 25.6 %. (It was initially 25.5% then I reran the code and it switched to 12.6 %).

```{r}
ridge2_pred <- predict(ridge2, s = ridge2$lambda.min, newx = x[test, ])

mean((ridge2_pred - y.test)^2) # Ridge Test Error Rate
```

```
[1] 12.60833
```

4.2 Lasso

- Followed by ridge, I decided my next method would be lasso because I could use the same type of train and test values since they almost go hand in hand. With my method lasso ended up performing very poorly. The test error rate ended up being equal to 51 % (It was initially 51% then when I reran the code it switched to 61.5 %).

```r
lasso_pred <- predict(lasso, s = lasso$lambda.min, newx = x[test, ])
mean((lasso_pred - y.test)^2)
```

```
[1] 61.50303
```

4.3 Tree

- After I ran ridge and lasso I decided to dip into decision trees. First, I fitted the classification tree. The results I obtained were not the greatest, I got a test error rate of 48.4%

```r
tree.drugs.pred <- predict(prune_drugs, test, type = "class")
mean(tree.drugs.pred != test$Education)
```

```
[1] 0.4843424
```

4.4 Bagging

- The next step to decision trees is bagging. I thought this would perform the best out of all the methods that I chose to do for this project, but I was wrong. This model had a test error rate of 48.9%. That was slightly worse than the initial classification tree.

```r
bag.drugs.pred <- predict(bag.drugs, test)
bag.error <- (mean(bag.drugs.pred != test$Education))
bag.error
```

```
[1] 0.4885177
```

4.5 Random Forest
- The last decision tree I decided to run for this data was random forest. Once again in my initial prediction I thought that this and bagging would give me my best test error's, but I was proven wrong. This method performed the best out of all of the decision tree methods, but it didn't perform the best overall. The test error I ended up getting from random forest was 47.2%.

```{r}
rf.drugs.pred <- predict(rf.drugs, test)
rf.error <- mean(rf.drugs.pred != test$Education)
rf.error
```

```
[1] 0.4718163
```

5 Discussion and Conclusions

My results were not ideal by any means, but the lowest test error rate that I got from my methods happened to be ridge regression. My original thinking was that one of the tree methods such as bagging or random forest would give me the best results, but my thoughts were incorrect. Ridge had a lower test error rate compared to any of the tree types of methods that I included in my report. For all of this code I followed the book directly, so I thought I would get better accuracy for my results.

References

GitHub:
https://github.com/fivethirtyeight/data/tree/master/drug-use-by-age

Plot for Related work:
https://fivethirtyeight.com/features/how-baby-boomers-get-high/

Data set I used for my report:
https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29

Book:
https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf