MICHIGAN STATE UNIVERSITY

Drug Use By Age Analysis

73

Noah Behm

Final Report

1. Introduction

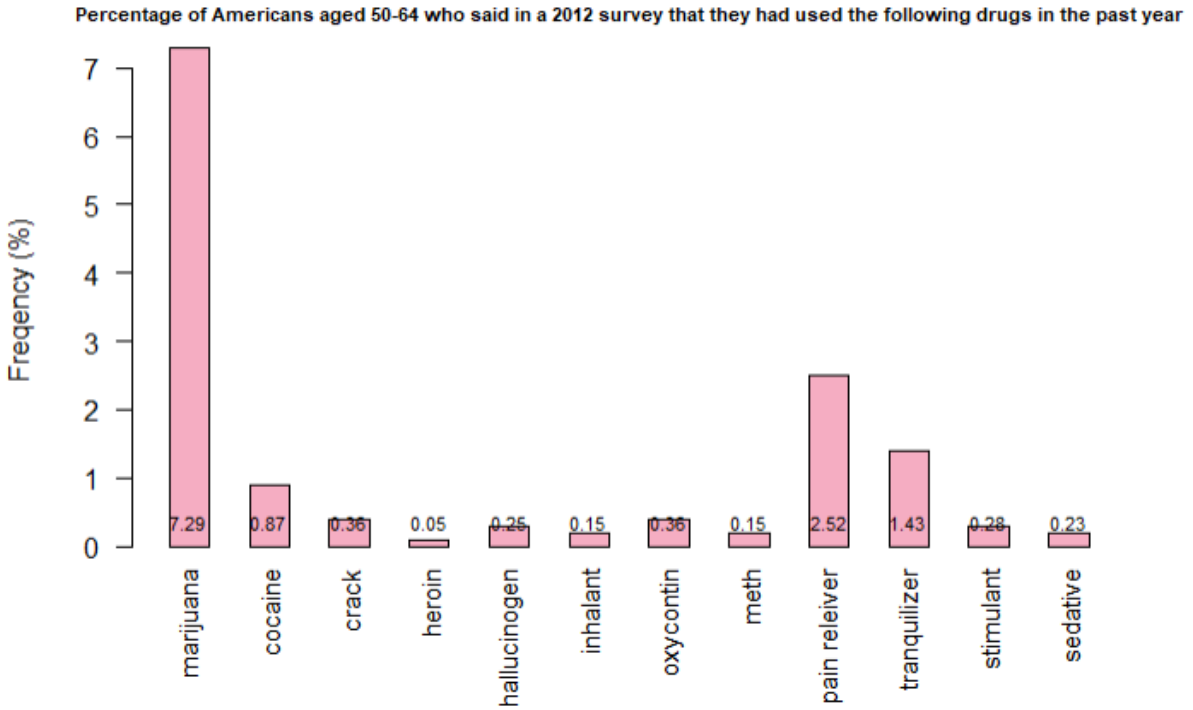
Drug use is an area of great concern for many entities - whether that be parents, the government, or schools. It is important to know who is most at risk for the abuse of drugs, what drugs people in different groups are using, and at what rates drugs are being used by people.

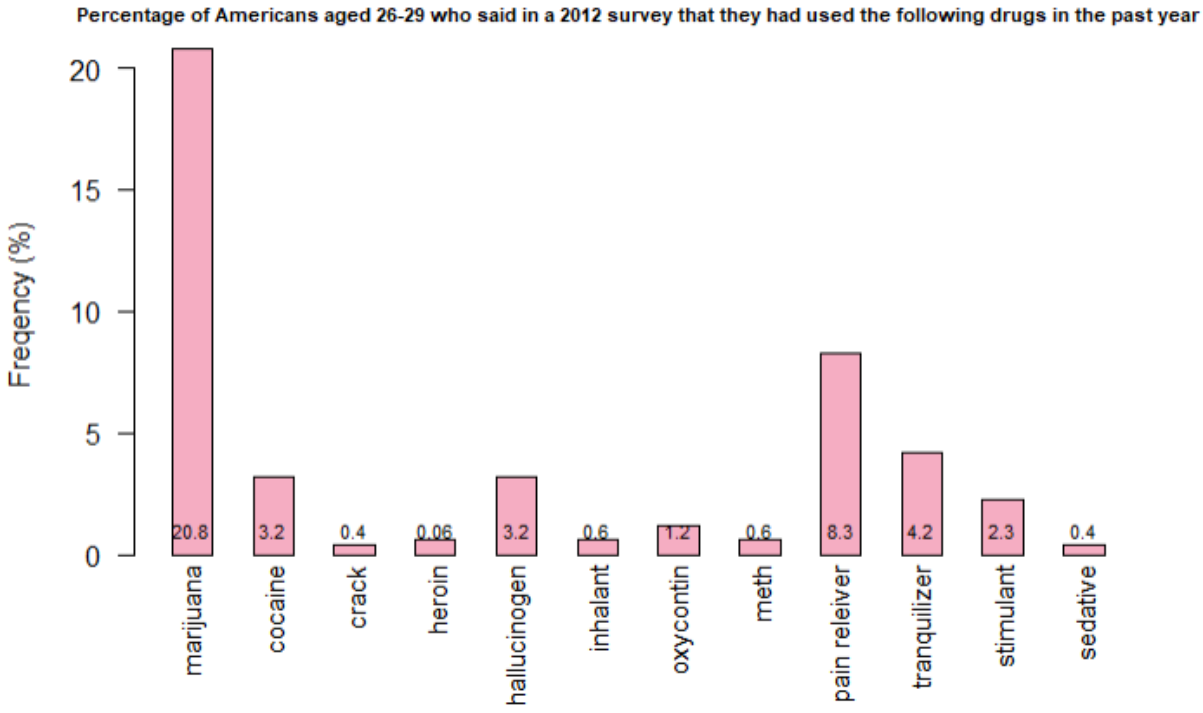
2. Problem Statement

This project aims to determine who is at risk of drug abuse, and what they are at risk of, through the use of SVM, Lasso Regression, and basic plots.

3. Related Work

In an existing article [1] on drug use by age and age groups there was a plot demonstrating the amount of people in the Boomer generation that do different drugs. Using their dataset I reproduced the graph that they had, and a similar one in order to compare the drug use of two different age groups - those aged 50-64 and those aged 26-29. Below are these graphs.





The above two graphs show the percentage of individuals in a given age group that do any of the drugs they were asked about.

By comparing the percentages of participants who do different drugs it is clear that the younger generation does more drugs - in fact the younger age group had a higher percentage of participants who said yes to doing drugs in all categories. The differences in percentages ranged

from .01 (for heroin usage) to 13.51 (for marijuana usage). Across all categories there was an average increase in drug usage of 2.61 percent. According this data it would seem that younger generation are more at risk of abusing drugs than older generations.

4. Datasets and Feature Selection

For this project I utilized two different datasets, both are cited below. The first data set I used was the same as used in the article referenced above. This data set contained information on drug use by age groups, pertaining to 12 specific drugs - seen on the bar plots above. There were 17 age ranges included, ranging from 12 years of age to 65+ years of age. The second data set I used contained information on drug related deaths, which included types of drugs, location of death, sex of specimen, age of specimen, and many similar data points. For this second dataset I focused mainly on sex, age, location of death, and race of the people represented in the set. I chose these features to try to answer questions related to my problem statement - I wanted to examine any correlation between sex, age, location of death, and race to see if any specific group of people was more at risk of drug abuse and misuse and subsequent drug related deaths. One problem with this second dataset, as you will see below, is that there were many more white specimens in the dataset than any other group. Nevertheless, to clean the data for use in a model I omitted any row of data that had NAs in them, and - since the amount of entries of racial groups other than white and black was miniscule - I eliminated any row that had somebody of a race other than white or black. I would have liked to keep these entries but there were simply so few that I thought it better to remove them. Lastly, I changed character entries, such as “white” and “Residence”, to numbers in order to use them in my models.

5. Methods

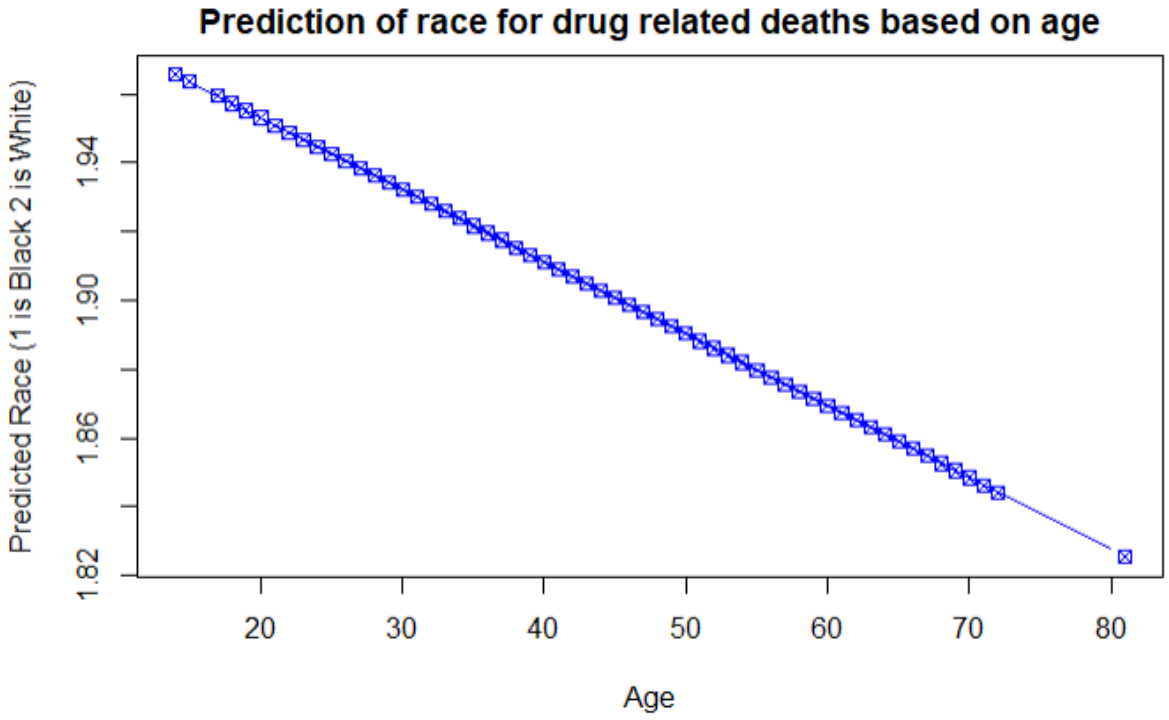
To investigate the questions I had I used glmnet to produce a lasso regression model using different features, namely sex, location, and age, with race as the classification variable to determine if any groups were more susceptible to drug related deaths. I also used an SVM model to investigate the same question

6. Results

As I mentioned in the datasets section, I ran into the problem of having many more white entries than black, so I believe that this affected my models very negatively. For example, my SVM model predicted every single person the be white based on sex, location of death, and age. My lasso regression model ran into the same problem and had r squared values less than .025 for all of the models that I produced. The models I produced are as follows: age as a feature predicting race, location of death as a feature predicting race, and sex age and location of death as features predicting race. Below are graphs representing what I found.

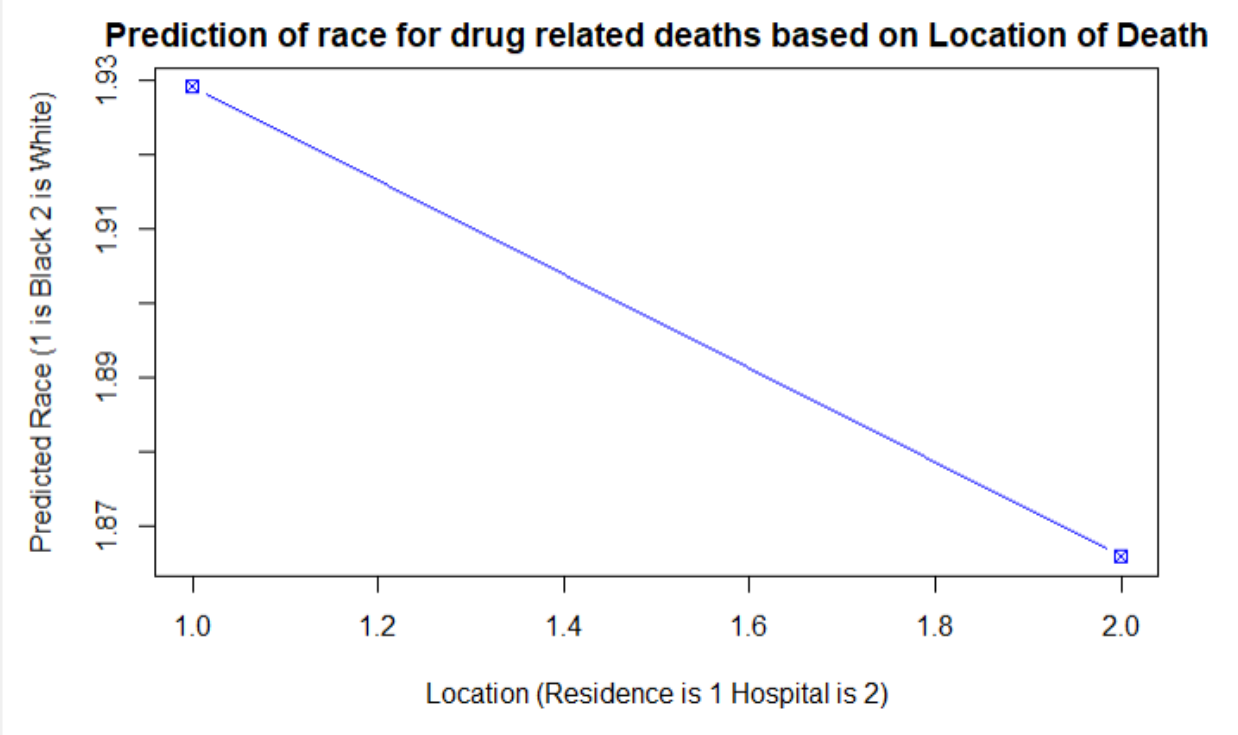
```
Call:  glmnet(x = x_train, y = y_train, alpha = 1, lambda = best_lam)

  Df %Dev Lambda
1  1 1.42   0.01
```




```
Call:  glmnet(x = x_train, y = y_train, alpha = 0.5, lambda = best_lam)

  Df %Dev Lambda
1  1 1.49   0.01
[1] 0.01017961
```



From the graphs that I've included we can see that it appears that more white people died in their residences than in the hospital, and that younger people who died tended to be white. However, I will reiterate that I don't believe that these values are accurate because the "most black" that was predicted was around 1.82 when black was supposed to be represented as 1. With that said, I'm not sure that any meaningful conclusions can be drawn from this analysis, and furthermore a more representative data set would be needed to carry out work that would adequately represent the reality of the situation.

7. Discussion

Since the dataset included so many more white people than black people, it is hard to say that this analysis is accurate. I think that one of two things occurred here. Either many more white people die due to drug related circumstances and the model predicted this correctly, or the dataset just did not have enough entries from black drug related deaths. I believe that the skew towards white entries made the model biased toward predicting people as white. Further analysis would have to be done on the subject, preferably with a more representative dataset in order to gather any meaningful insights into this very important topic.

8. Citations

https://fivethirtyeight.com/features/how-baby-boomers-get-high/
https://data.world/fivethirtyeight/drug-use-by-age

https://www.kaggle.com/muhakabartay/accidental-drug-related-deaths-20122018