Final Report - The Data Deities

Adam Ford, Allan Juarez, Carter Wunsch, Joseph Strobel, Patrick Wenzel

Introduction

It seems like we frequently hear about people being killed by the police in the news. Since the US government doesn't keep a comprehensive record of these killings (Viner et al., 2015), The Guardian started a project called The Counted. This is a project where the public can submit information about anyone who has died at the hands of the police and then The Guardian will verify the information sent. Once the information is verified, it will be added to the dataset and this way there is a list keeping track of people killed by law enforcement. This dataset includes the deceased's name, age, gender, race/ethnicity, cause of death, if they were armed, address, law enforcement agency they were killed by, and other information from the 2015 5-year American Community Survey. While there is the completed 2015 dataset on The Counted now, this dataset was made June 3, 2015, so we are not working with a full year's worth of data. From this dataset, we wanted to answer two questions:

- Can you predict if the person killed by the police is white based on several of the given predictors?
- Can you predict if the person was armed based on several of the given predictors?

To start off our exploratory data analysis (EDA), we first wanted to look at the demographic of the races/ethnicities in our dataset.

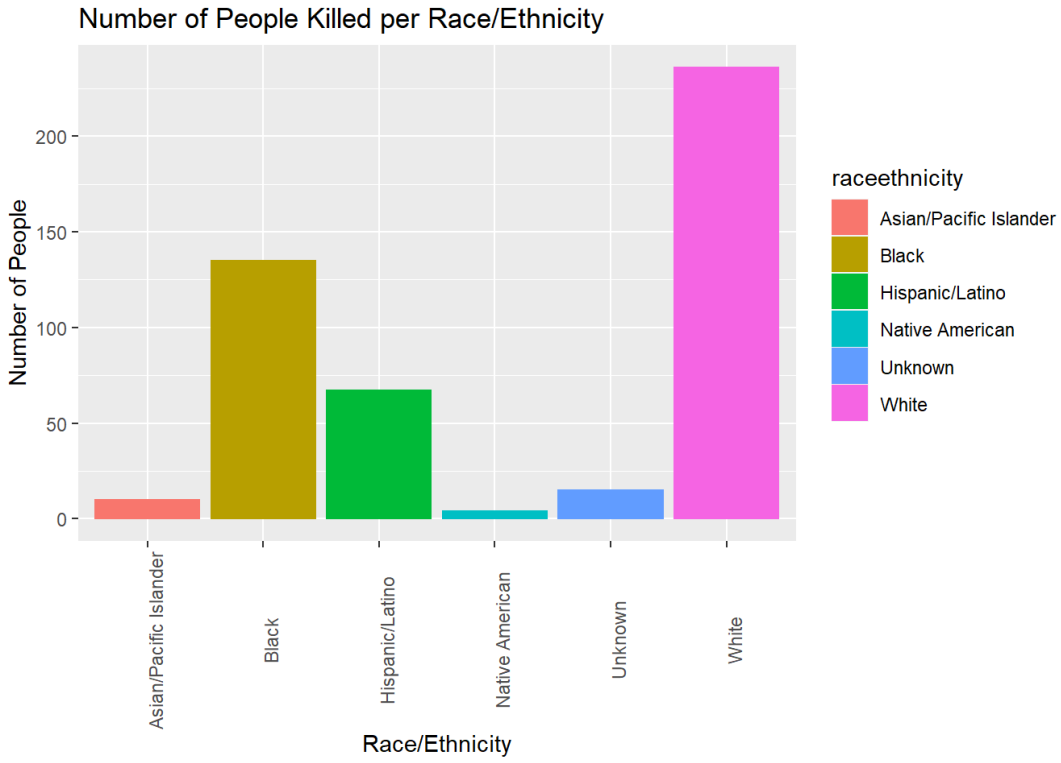


Figure 1: Race/ethnicity distribution of dataset

Looking at this figure, you can see that over half of the people in this dataset are white with the next highest number of people being black. One thing we thought that was important to note is that although there are more white people overall in the dataset, white people also make up a much higher percentage of the U.S. population so this graph is showing that black people were being killed at a higher rate per capita than white people from January 2015 to June 2, 2015.

For answering our second question, we wanted to explore the distribution of the armed classification. You can see this in the graph below:

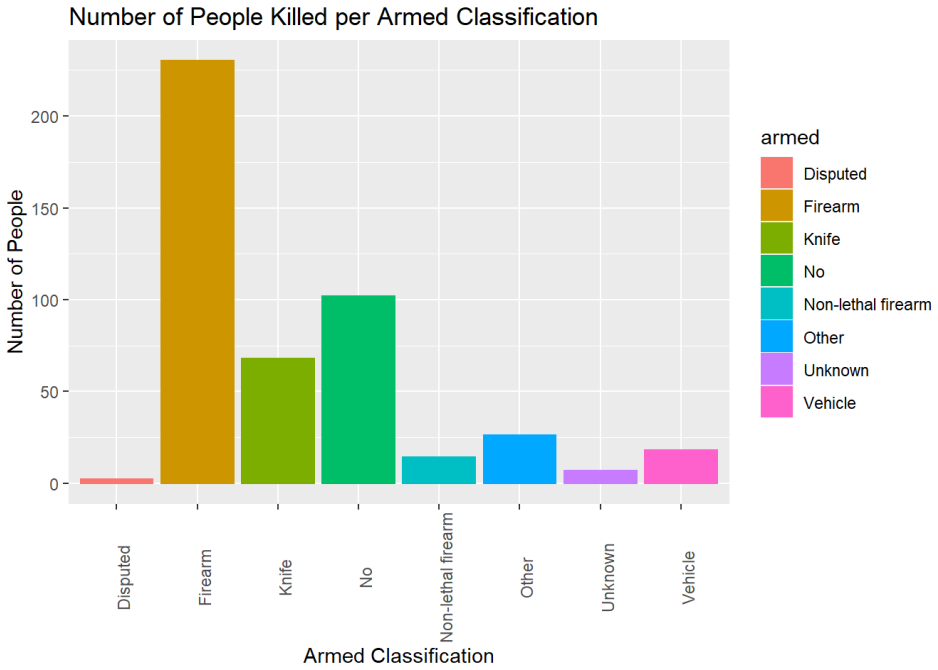


Figure 2: Arms classification distribution of dataset

What stood out to us was not how many were classified as having a firearm when they died, but that the second highest number of victims weren't armed at all. One thing to note is that according to The Guardian, the classification of if a victim was armed or not comes from law enforcement agencies where that is determined by the perception of the officer if the weapon/object would be used against them.

Based on these two variables, we think that they should help us make a decent model when answering our questions. We did, however, wanted to explore one more variable which was age.

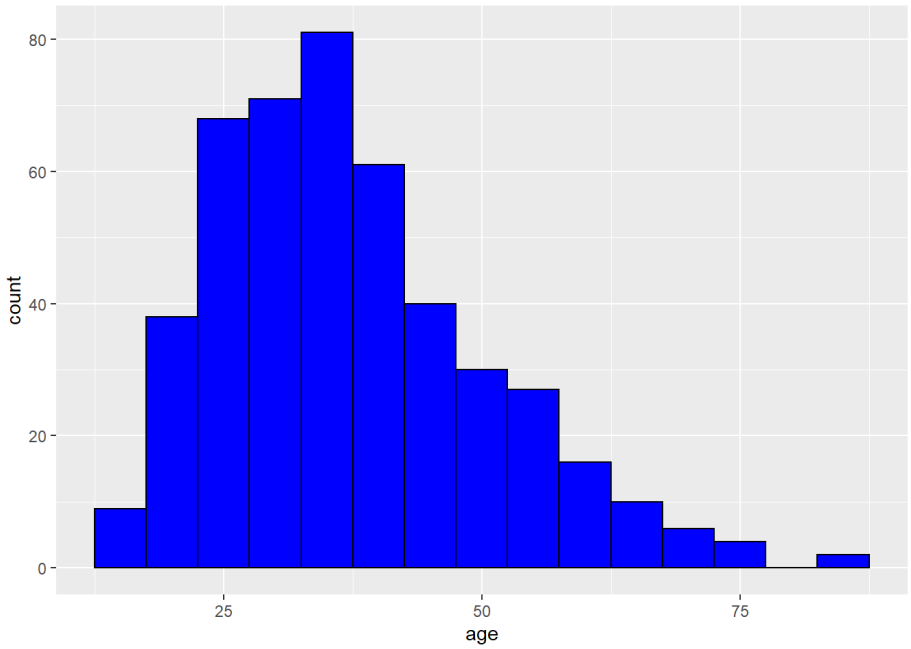


Figure 3: Age distribution of dataset

We think age will be helpful in making our models as well because of the distribution of the variable.

Methodologies and their Results:

Question 1 - Can you predict if the person killed by the police is white based on several of the given predictors?

Method 1 - Logistic Regression

An initial methodology we attempted to answer this question was Logistic Regression. However, some cleaning and adjusting was necessary before any real regression could be run. The dataset needed some rows with missing values removed or filled in with the column average. Additionally, several predictors needed to be scaled. Finally, a binary response column of *white* was added.

Two initial runs of Logistic Regression were performed to understand a baseline of the performance. These and all future tests were performed with 50-50 training and test sets. The first was utilizing the factors age, gender, and armed. This resulted in a classification rate of 61.9%. The second run utilized the interactive product of the scaled tract college graduation rate, unemployment rate, and poverty rate, which resulted in a classification rate of 60.6%. The interaction term was leveraged due to the tight relationship between the listed variables.

Before advancing we considered if either type of error was more harmful, and concluded that due to the exploratory nature of this question there is not an error that is more harmful. We hope just to maximize the overall classification rate, thus there was no need for ROC analysis.

We did however explore finding a subset of predictors that can aid in limiting the misclassification rate. We selected AIC and BIC as metrics to decide on an ideal model. The model that reduced AIC and BIC was age, armed and the interactive term of scaled tract college graduation rate, unemployment rate, and poverty rate. Another run of Logistic Regression resulted in a classification rate of 65.1%.

From Logistic Regression, we cannot predict if the victim is white based on other factors of their killings. However, we will use other methods in an attempt to find a better model to classify this.

Summary	**Predictors**	**Classification Rate**
Personal Information	age+gender+armed	61.9%
Locational Data	college*urate*pov	60.6%
Subset Selected	age+armed+college*urate*pov	65.1%

Method 2 - QDA

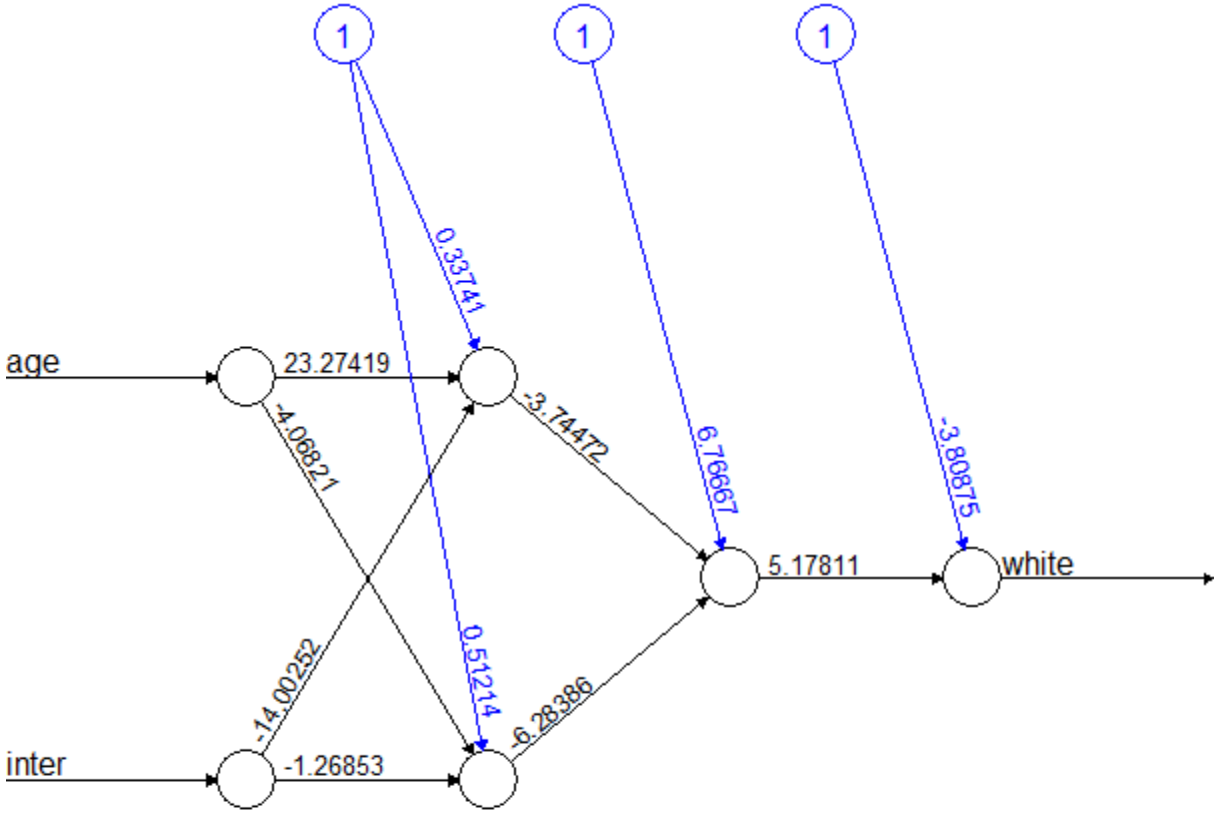
To begin exploration we began by choosing pairs of the given predictors to determine if we thought any would be particularly useful. The predictors chosen for the initial exploration were *age*, *cause*, *pov*, *college* and *armed*. Each predictor was paired with every other predictor and used in a QDA model to check for their classification rate. The two models that yielded the highest classification rates were *pov+age* and *pov+armed* yielding classification rates of 65.8% and 64.9% respectively.

After initial exploration, we performed forward stepwise selection on the same set of predictors, adding in household income well. The models with the lowest AIC were three and four. The models with the lowest BIC were two and three. The models with the highest adjusted r-squared were four and five. Lastly, the models with lowest Mallow CP were three and four. We chose model three because it had a much lower BIC than model four and competed in all respects. The predictors used in model three were *age*, *pov* and *h_income*.Using this model the classification rate was only 58.1%, which was much lower than the models generated in the initial exploration.

Method 3 - Neural Network

For our method not covered in the course, we decided to explore artificial neural networks. We did this using the neuralnet package, and we used the predictors age and the scaled interactive term of tract college graduation rate, unemployment rate, and poverty rate. ANNs are made up of an input layer, hidden layers, and an output layer. To put it simply, the neurons in the hidden layers take inputs (from “dendrites”) and pass information (through “axons”) to the next layer until the output layer is reached. Information starts at the input layer and is passed through the layers. As the information is passed, weight vectors (parameters for classification) are altered depending on the information passed. When the output layer is reached, the weight vectors

should've been altered to their proper values in order to classify the sample correctly. This idea is supposed to mimic the human brain and how it processes information.



Error: 23.461196 Steps: 4148

For the model above, we used two neurons in the first hidden layer and one neuron in the second hidden layer. We also decided to use the default logistic activation function. The idea of the activation function is that it is a differentiable function that can be used to smooth (flatten) the results outputted from the hidden layers. Although ANNs are known for their predictive accuracy, our model only had a 63.2% classification rate. This means we still cannot find a conclusive method to predict if a victim is white based on factors surrounding their killing.

Question 2 - Can you predict if the person was armed based on several of the given predictors?

Method 4 - GLM

The first method we used to answer this question was logistic regression. Because many of our variables were irrelevant to this question, we decided to select our top eight variables of interest. These variables were age, race, poverty rate, income, unemployment, cause of death,

gender, and college experience. We first trained a logistic regression model with these eight variables as the predictors. We found the misclassification rate to be 0.225.

Next, we wanted to see if we could lower this misclassification rate by eliminating unnecessary predictors. We ran forward stepwise selection on the eight-predictor model and found diagnostics to determine the best size and predictors for our model. A model size of three predictors ended up having the lowest AIC and the lowest Mallow's cp, so we chose this size as the best model size. The predictors associated with this model were age, race, and cause of death.

Finally, we went back and trained another logistic regression model with the three predictors found from stepwise selection. Our misclassification rate improved from 0.225 to 0.193. Although our model improved, we wanted to see if we could improve our prediction accuracy even further. This led us to using LDA.

Method 5 - LDA

After completing our logistic regression models, we wanted to see if GLM would perform better or worse then LDA. We first trained an LDA model using the original eight predictors we were interested in, which gave us a misclassification rate of 0.214. This prediction accuracy is better than our original GLM model, but it is slightly worse than the logistic regression model we trained with our three best predictors. Our next step was to train another LDA model, but this time we used the three best predictors found from stepwise selection. Our LDA model with the predictors age, race, and cause of death gave us our overall best misclassification rate of 0.185.

Discussion

We cannot predict if a victim of a police killing is white based on the given predictors about their killing by using Logistic Regression, QDA or a Neural Network. We deemed a maximum classification rate of 65.8% to be unsuccessful because it is not a useful level of accuracy. It could be considered a success in that it would outperform classifying every victim as white or non-white, which was at approximately 50%. Still, we do not think that slight uptick in accuracy is enough to provide any greater meaning.

Overall, this is a surprising result. We expected to be able to predict more accurately because of the way the media often portrays these killings of non-white individuals as being similar. We also expected that having data on the tract in which the victim was killed would make predicting the race possible, due to there being external trends between the tract unemployment, college graduation and poverty rates generally with the percentage of each race that live in the tract.

There is some correlation between some of the predictors like age, cause of death and race and predicting whether the person was armed or not. Using methods like LDA and GLM we found there to be about an 82% classification rate to predict this. This is high enough to be

highly efficient in guessing the response, but not quite good enough to consistently predict the correct response in the end.

The end result is a bit surprising. We expected there to be more predictors when calculating whether the person that died was armed or not, but it boiled down to only three predictors. These predictors also make sense since these are what you always hear about on the news(age, race,cause of death). The classification rate was also higher than we thought. We didn't really have any initial guess but when it was in the 80s we were pleasantly surprised by the results and tried to get the percentage to go higher but couldn't.

Regarding the dataset as a whole, we wished that mortality or result of conflict of police encounters could have been used as a response. This would imply a larger dataset including all police encounters, which is a near impossible dataset to procure. However, if it was possible we could get greater insights into what parts of an interaction (race of victim, armedness of victim, tract information, etc.) contribute the most to the fatality of a police interaction.

Conclusion

After attempting Logistic Regression, QDA and a Neural Network a maximum classification rate of 65.8% was achieved for the question, “Can you predict if the person killed by the police is white based on several of the given predictors?” This was deemed to be an unsuccessful classification rate in the context of this problem.

When using all the tools to try and see the classification rate for our question “Can you predict whether a person was armed at the time of death based on several predictors?”. We came up with a classification rate of up to 82% and deemed to be semi successful in predicting whether the person was armed.

Statement of team member contributions

Patrick Wenzel:

Did exploratory data analysis

Introduction of the report

Joseph Strobel:

Question 1, approach two - QDA and subset selection and corresponding report section

Worked on the discussion and conclusion.

Adam Ford:

Question 1, Logistic Regression - Entire Section

Question 1, Neural Network - Initial R findings and exploration

Some of Discussion and Conclusions

Allan Juarez and Carter Wunsch:

Both worked together on all these parts

Question 2, GLM

Question 2, LDA

Question 2, repeat but with forward subset selection

Question 2, Recursive Partitioning
Come up with discussion and conclusions

References:

Andrei Scheinkman, andrewflowers, and dmill. Police_killings.csv. GitHub: Fivethirtyeight, 3 June 2015. CSV.

Viner, Katharine, Lee Glendinning, and Matt Sullivan. "About the Counted: Why and How the Guardian Is Counting US Police Killings." 2 June 2015. Web. 30 Apr. 2021.

Weessies, Kathleen. "Finding Census Tract Data: About Census Tracts." 23 Feb. 2010. Web. 30 Apr. 2021.