Fatal Police Shootings in the U.S.

DSC 630: Predictive Analytics

Blandon S. Lee

11/18/2022

Fatal Police Shootings in the U.S.

Introduction

Police involved shootings is a sensitive subject and an ongoing issue that keeps presenting itself with each passing year. While communities are losing members the police forces around the country are losing credibility and are faced with many difficult questions to answer. The frequency in which these types of incidents occur is staggering and leaves many to wonder what could be done to mitigate these types of police and civilian interactions.

The Washington Post began to compile data related to fatal police shooting beginning in 2015. According to the Washington Post there have been 5,000 fatal shootings where police are involved. The data compiled by the Washington Post was used in this project and contains the following variables: name, date, manner of death, armed, age, gender, race, city, state, sign of mental illness, threat level, fleeing, and body camera.

With the data I will trying to determine what the contributing factors are leading to these deadly police interactions. The project will attempt to create models that could potentially be used in predicting police shootings. I think the model could help in directing needed attention to certain locations where the shooting rates are higher. I will be providing additional visualizations along with the clustering analysis model to help in better understanding the issue.

Summary/Results

Preparation

Before beginning on the model there were some data preparations that had to be done. I had to transform some predictors from categorical or numerical. I done this so the dataset is ready for modeling and other planned analysis.

Modeling

The linear regression model was used for the variable race serving as a function for body camera, gender, and armed variables. This returned some significance where the p-value was under 0.05. The p-value for the overall model was 0.0003941 and a f-statistic of 6.088.

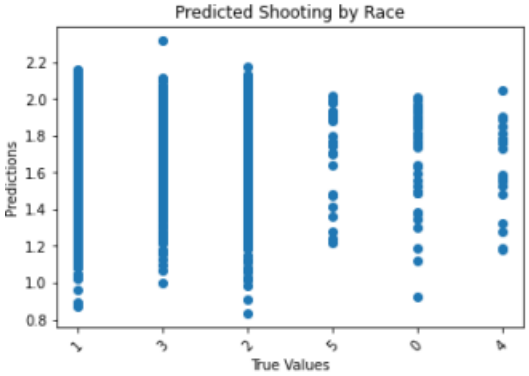
I then used a clustering algorithm as a means of better understanding which variables are the most important. I went on to utilize a k-modes clustering model to help deal with the categorical variables within the dataset. This type of model provided better matching based on the amount of matching- categories between the data point and works well with dataset that contains high volumes of categorical data and some numeric data.

Results

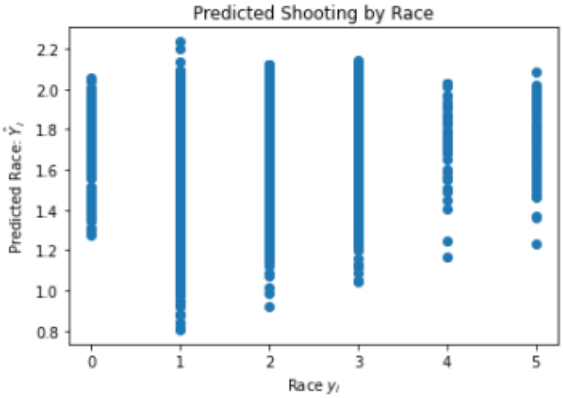
Once the linear regression model was ran using various summary statics (P-Value, R-Squared, F-Statistic) it was obvious that the model only contained some of the necessary information required to build a predictive model. I then used k-modes where each time it was ran it returned five clusters. Three of those five clusters returned white male/ 31 to 41 years old/armed while the other two returned a combination of black/hispanic/ armed.

The next model(s) used race as the target with a 60/40 test train split.

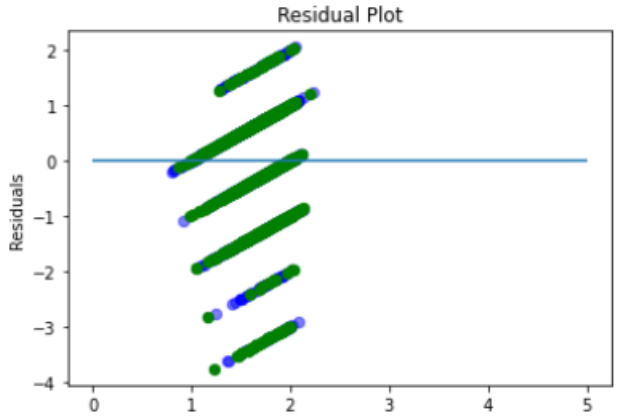
Model 1 Predictions: Unfortunately, only possessed and accuracy of 0.05% up to 0.11%. It tells me that this model is not a good fit for the data. This could be that the model is not correct or linear relations are incorrect. A different regression model such as logistic may have been the better option here.



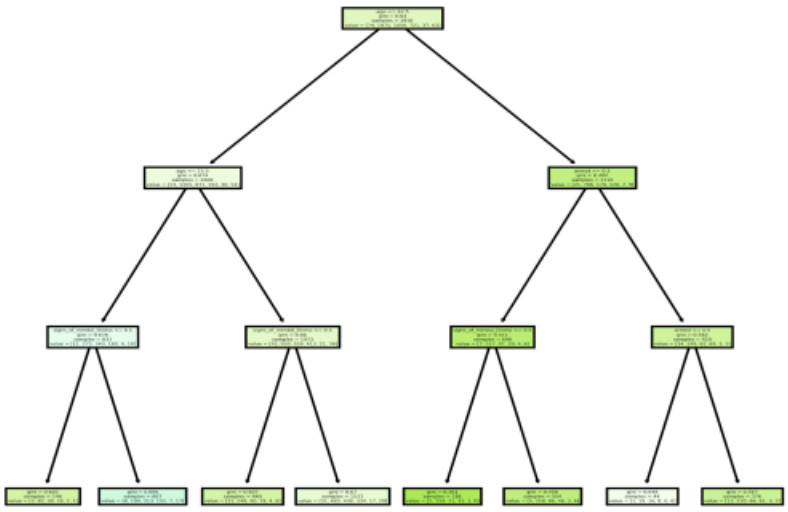
Model 2 Preditions: This one ran off the MSE rather than the accuracy. It returned an MSE of 0.81



Model 3 Residuals: Here I ran against two different variables race against type of arms. The train and test are again a 60/40 split random state is five. Here we got an MSE of 0.86. Looking below it would seem that the errors are contained in the black and white race data.



Model 4 Decision Tree: Here we see a prediction of which race a more likely to be shot by a police officer in the line of duty given certain variables. In the top node a person under 42 yrs. In the second node on the left a person under 25 yrs. Third node on the left with mental illness predicts 82 White, 30 Black, 28 Hispanic, 3 Asian, 2 Other, and 1 Native American



Conclusion

With the continued presence of fatal police shooting in the news around the country it is crucial that we continuously seek answers and work towards mitigating these occurrences. The better we understand the factors contributing to these situations the better suited we are to addressing them. In this project I was able to determine some trends between mental illness, age, and race that could potentially be useful in future research, analysis, and education. The results of this project and its processes could be used by various police departments as part of their training/department numbers to help better understand where they should focus their corrective measures. Through the modeling I was able to generate a better understanding of the causes of these types of situations. However, further analysis and additional data would be necessary to be able to consistently provide more accurate predictions.

References

1. Galarnyk, M. (2019). Understanding Decision Trees for Classification (Python). Towards Data Science. Medium. Retrieved from https://towardsdatascience.com/understanding-decision-trees-for-classification-python-96 63d683c95
2. Jeevan, M. (2018). How to run Linear regression in Python scikit-Learn. Big Data. Retrieved from https://bigdata-madesimple.com/how-to-run-linear-regression-in-python-scikit-learn
3. Kaggle. (2020). Data police shootings. Retrieved on October 31, 2020, from https://www.kaggle.com/mrmorj/data-police-shooting
4. Washington Post. (2020) Fatal Force. Retrieved on October 1, 2022, from https://www.washingtonpost.com/graphics/investigations/police-shootings-database/

Visuals

Here we can see where all the shootings have occurred and the race of the victim. Black Hispanic people are more likely to be shot in cities and white people are more likely to be shot in rural areas.