Predicting Negligent and Intentional Homicide Instances in the United States and Identifying Impactful Factors

William Denny

Abstract

Homicide data between 1980 and 2014 displays noticeable trends in the nature of homicides. Specifically, the disparity between negligent homicide and intentional homicide opens avenues to question why this disparity exists and can we make predictions based on different factors surrounding the classifications. Utilizing machine learning and the logistic regression model we can attempt to predict instances of negligent versus intentional homicide and determine the factors with the most impact on the outcome. The process consists of testing the features through random forest classifier, then take these features collectively to run through a logistic regression test in order to retrieve the prediction accuracies between the two potential results, intentional or negligent. The results show the ability to predict intentional homicide instances yet not being able to reliably determine negligent homicide instances.

1 Introduction

Simple negligence is defined as, “the absence of due care, where the act or omission of act by a person under duty to use due care, which exhibits a lack of care for the safety of others which a reasonably careful person would have exercised” (United States v. Riley, Article 134, 2003)[3]. The disparity between negligent and intentional homicide could be purely subjective and down to legislation. However, if we can predict whether or not an instance of homicide was negligent based on the factors surrounding the instance contained in the dataset, we can determine certain trends that pertain to cases of negligent homicide. To accomplish the goal, we will be using data from 1980 to 2014 which contains features such as victim and perpetrator, race, sex, and age, as well as weapon type, relationship, agency name, agency type, and other time and location features. In this research we develop the features through a random forest classifier (RFC) for each column, then using the selected features through a final logistic regression test to make the predictions with all the columns, while checking the receiver operating characteristic curve

(ROC). To ensure the model learned from the experiment we check the area under the ROC curve (AUC) as different column features are added. This way we can gauge the significance of some features to help understand which features are the most influential in the predictions. A final ROC test will be run to determine the influential features not tested with the above method.

2 Related Work

Not much has been done in relation to negligent homicide and intentional homicide. With this data set there was research done on gun violence, murder solve rates, and age of serial killers [8]. Some other research has been done on how ambient climate affects intentional homicide [9]. The Murder Accountability Project was the reason the dataset was created and did research on how homicide is under reported in America [10]. Negligent homicide appears to be very subjective and is on a case-to-case basis which would explain how no predictive analysis has been done.

3 Dataset and Features

The dataset was created for the Murder Accountability Project, founded by Thomas Hargrove. Data was taken from the FBI's Supplementary Homicide Report and Freedom of Information Act data [1]. There was a lot of cleaning necessary in order to help optimize the model and utilize accurate data. To begin the dataset contained a list of unsolved and solved crimes, to account for this all "unsolved crimes" were removed as they need to be solved for the prediction. A record ID column contained indexes for each instance, this was unnecessary as the panda's library automatically indexes the dataset. A column containing the agency code was unneeded as we already had the names of the agencies which display the same data. Victim ethnicity and Perpetrator ethnicity were dropped as they contained an unknown value for over half of the entries another option to solve this would be to delete the unknowns but since they account for over half of the dataset the column elimination was a better option, see figure 2. There were some data entry issues with the age of perpetrators as the column was initially listed as objects, so to run the tests on the data these needed to be cast to floats. Some entries in the perpetrator age column had values of 0 to 2, these were clerical errors in the data set [1], these instances

were replaced with the mean perpetrator age to ensure we use as much data as possible, picking the number 3 as the cutoff was decided based on a story of a 3-year-old in North Carolina who murdered her sibling [2]. Victim age also had some data that was mislabeled with an age of 998, the same process was done to remedy this issue as well. This will cause the distribution graphs to have a spike at the average, however it still assists the model to learn, and the distribution is not affected negatively. The biggest issue with making predictions on the dataset for these two values is that negligent homicide accounts for a very minimal portion of all the homicide instances, as seen in figure 1. To assist the model 5 columns needed to be scaled, as shown in figures 3 - 6.

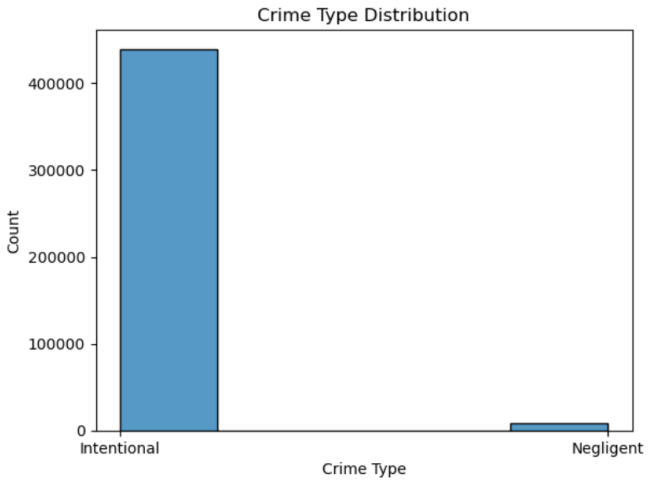

Figure 1: Crime Type Distribution

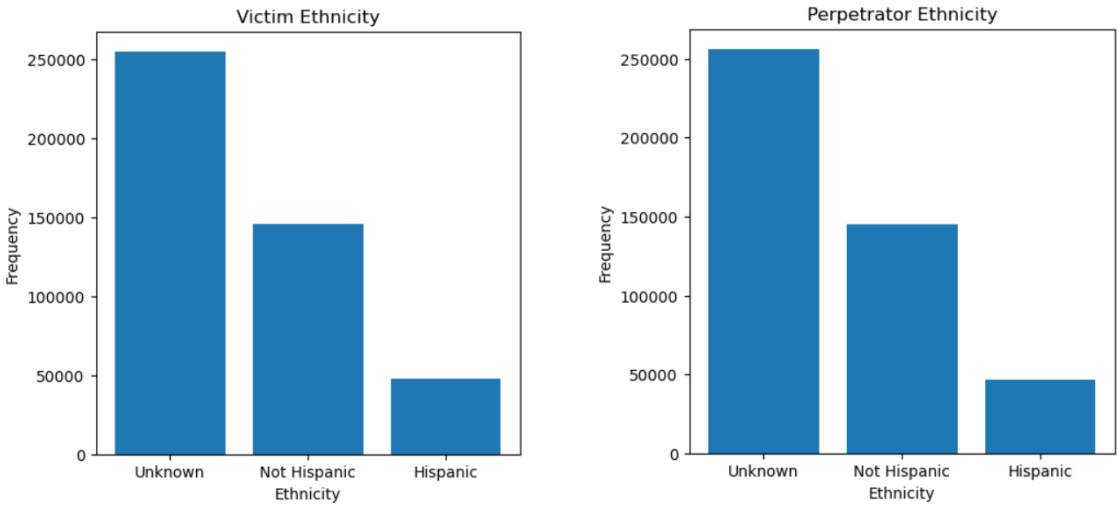

Figure 2: Perpetrator and Victim Ethnicity

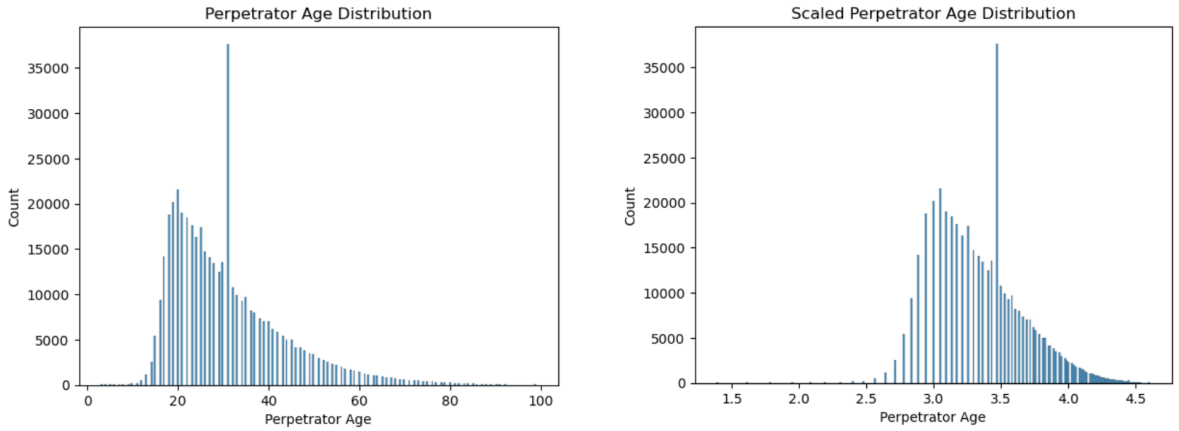

Figure 3: Perpetrator Age Distributions

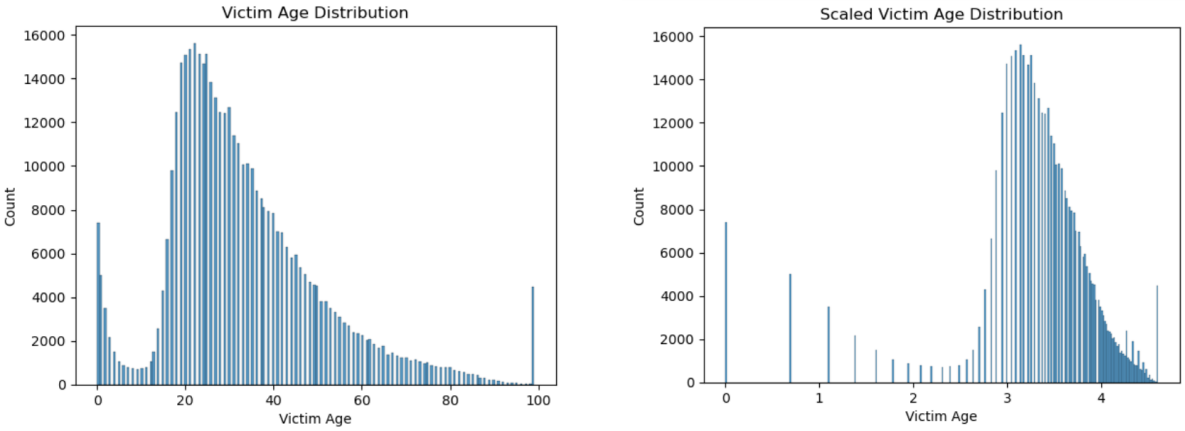

Figure 4: Victim Age Distributions

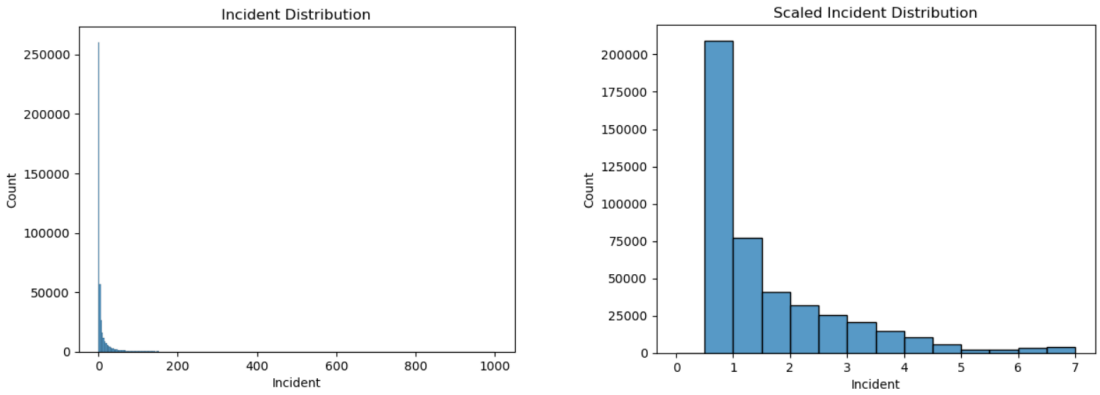

Figure 5: Incident Distributions

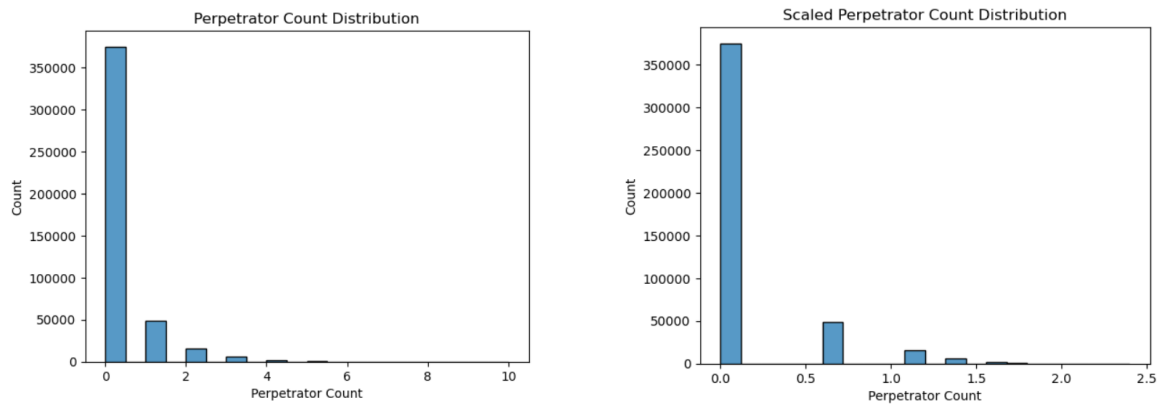

Figure 6: Perpetrator Count Distributions

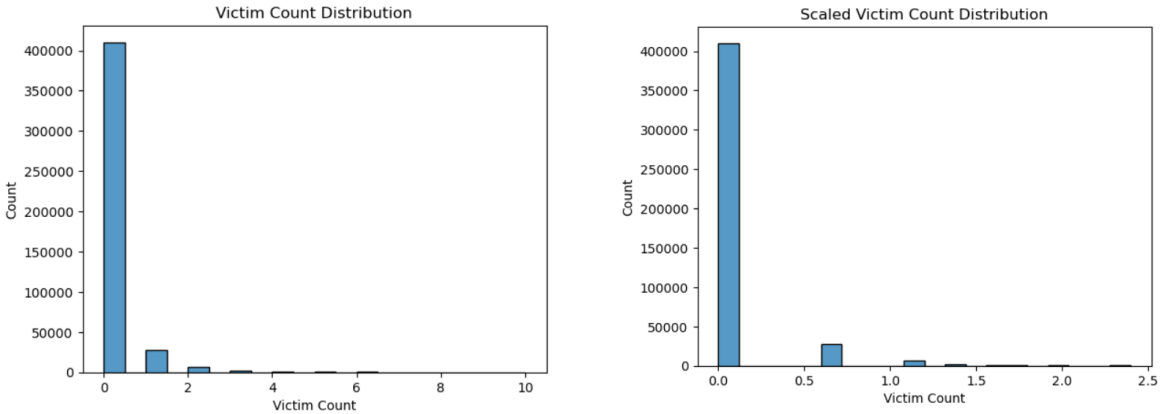

Figure 7: Victim Count Distributions

4 Methods

4.1 Overview

The approach to this model was based on a couple factors about the dataset. The number of negligent homicide numbers being significantly smaller than the intentional homicide numbers means that accuracy scores will be inaccurate. To ensure that both negligent and intentional homicide are being predicted accurately we will be looking at confusion matrices. Another factor

stems from the data primarily consisting of objects, this means we will have to use one hot encoding to run the tests, which separates object data into different columns and uses ones and zeros to display when they are present or not. Utilizing one hot encoding brings in another issue, the city and agency name columns contain thousands of unique values, with one hot encoding this can take a very long time. To resolve the column size, we will only be using the top 50 cities and agencies with the highest counts of negligent homicide in order to assist the negligent homicide prediction due to the small number of instances. The next step is to take the cleaned data and run the RFC to gather the higher accuracy features, then run a logistic regression test on the cumulative features that were selected. During the process of testing, AUC scores will be collected in order to determine the impact the feature had on the model. Finally, to get the features that were not tested by the above method since they were numeric we will run a final RFC test with a higher threshold to see which features emerge.

4.2 Selecting Features

The features being selected for the tests are a result of the RFC using 50 estimators which represent the number of classification trees being used. The random state will remain constant in order to check how the addition of different features affects the output. Finally, the threshold being used will be the median in order to get the top 50% of test scores from the classifier. This RFC will be run before adding the additional features. The initial run will be used as a baseline to check the impact of the added features. A final RFC run will be conducted with all features added.

4.3 Logistic Regression Test

After the RFC tests a logistic regression test will be used to predict the outcome by using 70% of the instances of the chosen features to train and test the model using the remaining 30%. The output from these tests will be displayed as a confusion matrix due to the small sample size of the negligent homicide instances. The baseline and final test run will be the only two collected.

4.4 AUC Scores

The AUC score represents the area under the ROC curve. This shows how much the model is learning from the training when it is performed. This data ensures that the model is not just

guessing randomly and is utilizing the data to make informed predictions. Each AUC score will be kept from each individual logistic regression test to see which column features were the most beneficial.

4.5 Finding Remaining Features

To get the remaining features, a final RFC test will be conducted with a high threshold to see which features emerge. This could not be done with all of the data as some of the columns contained object data which will be split up after the one hot encoding. This means in order to get an idea of how all of the data in that column affected the model we needed to do the AUC analysis mentioned previously. Columns that contain numeric data can be identified through the final RFC test and then can have AUC values tested to determine their impact as well.

5 Results and Discussion

5.1 Important Column Features

The ROC curve displays which column features had the greatest impact on the model. Figure 8 shows that from the initial curve, all of the features added increased the learning for the model. The three most impactful column features containing object data were weapon type, relationship, and state, with the initial AUC score being the worst, as seen in figure 9. The nature of negligent homicide requiring negligence from a person who is of duty to exercise care [2] explains how the relationship and weapon type would be good indicators for prediction. The state's importance comes down to the disparity between negligent and intentional homicide, as the dataset shows there are significantly more intentional homicides than negligent. However, states such as California have low negligent homicide numbers relative to their intentional homicide numbers, and Florida has the opposite characteristic with high negligent homicide numbers relative to their intentional homicide numbers, as show in figure 9. State law is the main difference that explains the disparities for Florida and California. Florida penal code 782.07 lists counts of aggravated manslaughter on elderly, children, and uniformed officers, fighter fighters, or paramedics with culpable negligence[4]. This means that on top of involuntary manslaughter there are instances of aggravated manslaughter which can be the result of negligence. However, in California abuse of the elderly and child neglect are not listed under manslaughter satutes [5, 6] and both involuntary and voluntary manslaughter charges do not include negligence, only vehicular

manslaughter pertains to negligence in California[7]. Due to Florida having specific statutes outlining manslaughter which include culpable negligence may be an influencing factor to the high negligent manslaughter numbers in Florida. California not listing manslaughter laws with negligence and having separate crimes for child neglect and elderly abuse explains how the negligent homicide numbers are lower in relation to the intentional homicide numbers.

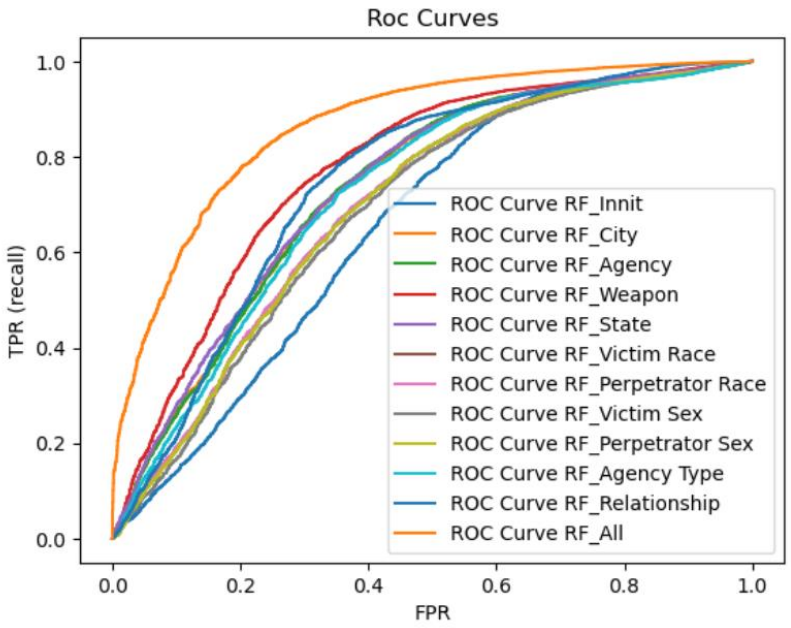

Figure 8: ROC Curves

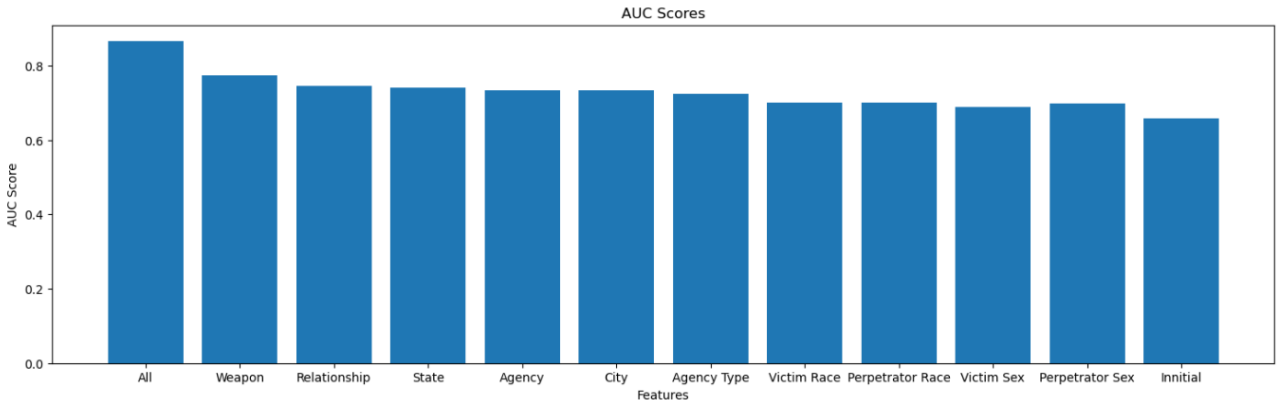

Figure 9: AUC Values

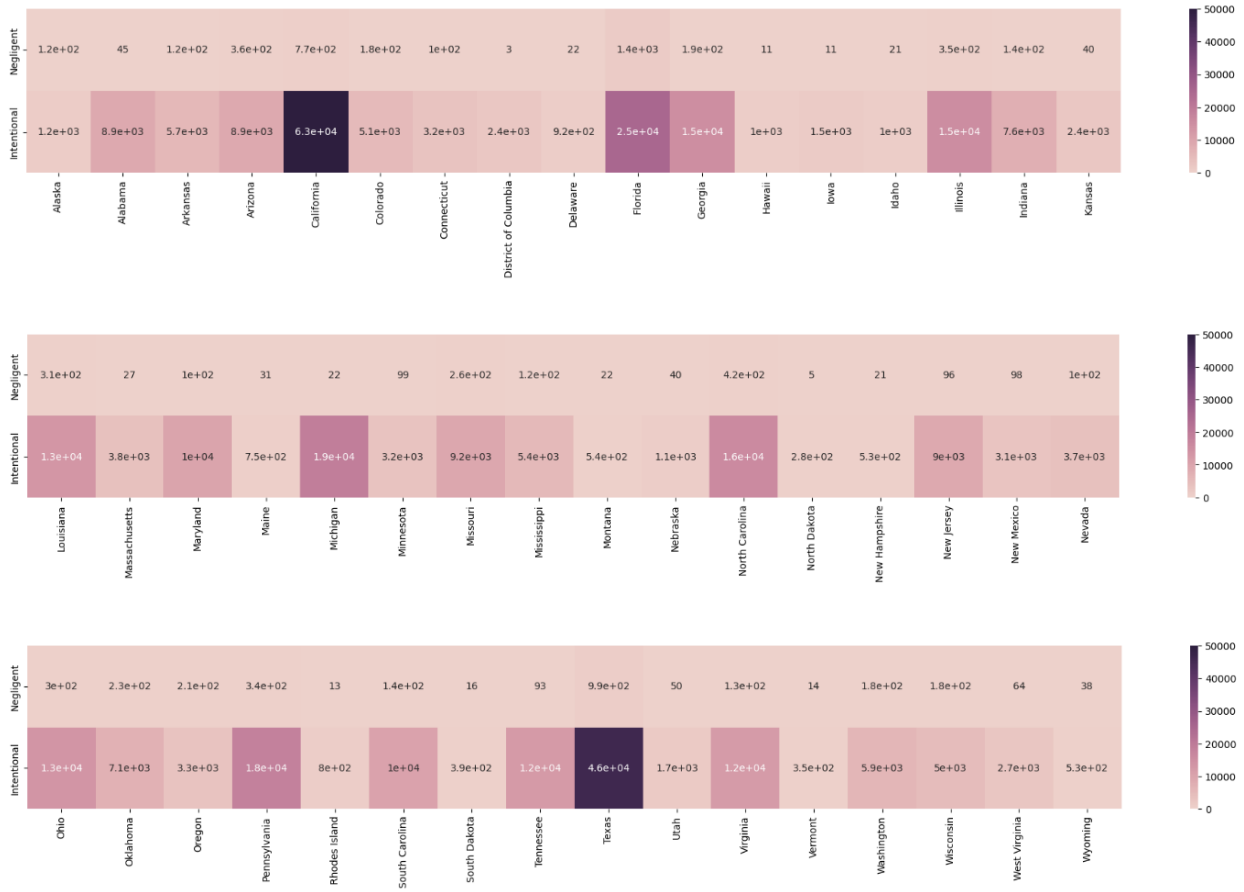

Figure 10: State Negligent and Intentional Homicide Disparity

5.3 Important Remaining Features

With testing the column features importance there still remains the rest of the features that were used. A final RFC test with a threshold of .015, which is high for this dataset, resulted in year, month, incident, victim age, perpetrator age, and relationship_friend being the overall most impactful features, as seen in figure 12. The “relationship_friend” feature is a result of one hot encoding the relationship column, more interestingly its position in this list of features is surprising as it is not as frequent as other relations, as seen in figure 11. Looking at the data, out of over 400,000 instances of intentional homicide only around 20,000 had a relationship of friend, among the 9,000 instances of negligent homicide around 2,000 of them had the relationship of friend. This could explain how a relationship of friend was a good indicator for negligent homicide as it accounted for significantly more of the instances than in the intentional homicide instances. The remaining features all originate from columns containing numerical data

so they will be more inclined to be an outcome from this test. This does not mean these features are not important, all of these features were in the baseline test and since they encapsulate an entire column of data, they are more incline to be more impactful.

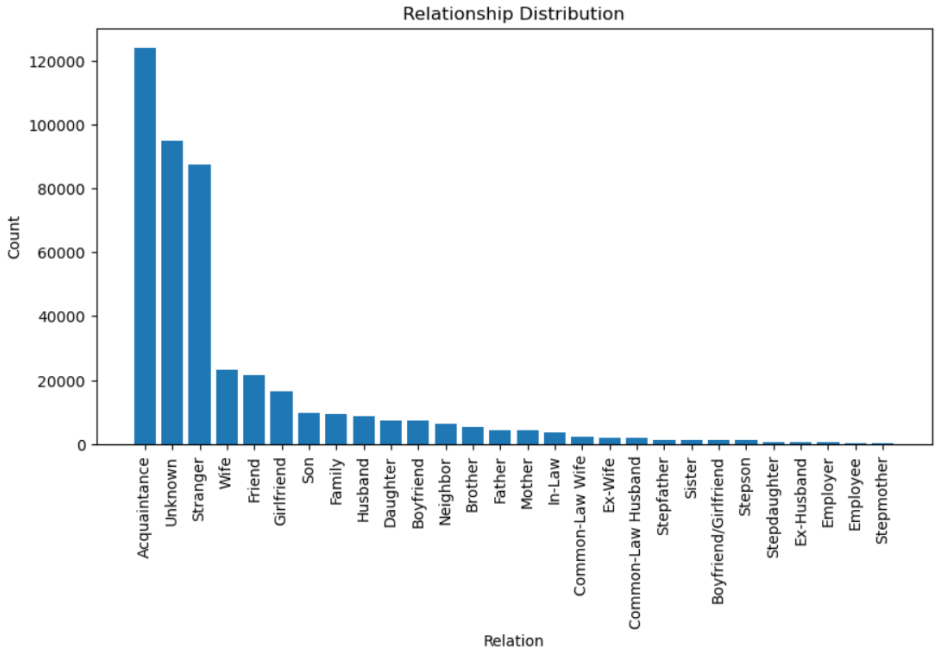

Figure 11: Perpetrator Victim Relationship

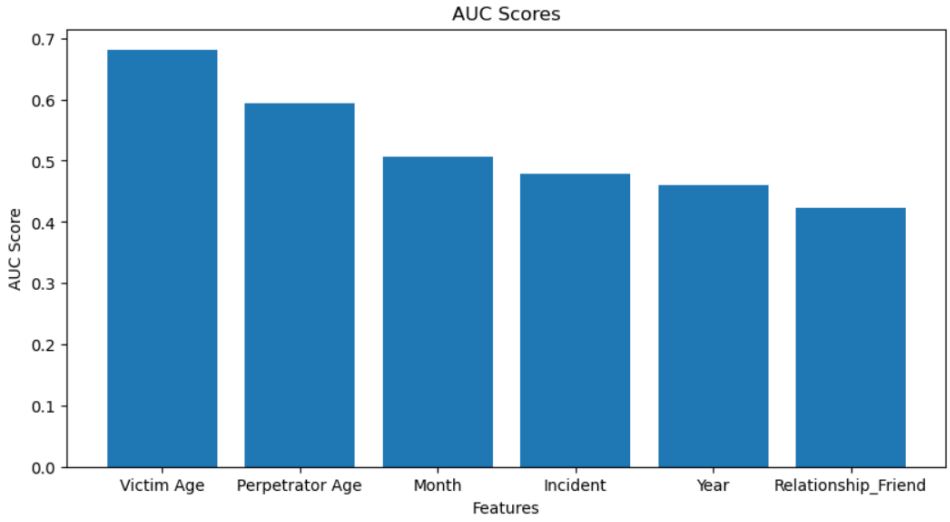

Figure 12: AUC Scores Overall Best

5.2 Prediction Scores

The baseline confusion matrix reveals that the baseline data does not allow for negligent homicide instances to be predicted at all, yet it was able to correctly predict instances of intentional homicide well, as seen in figure 12. This mainly comes down to there being significantly more data for intentional homicide as opposed to negligent. Luckily, with more features added we were able to predict some of the negligent instances shown in figure 12. Some of the features added were targeted at solving the issue of predicting the negligent instances better as losing a few predictions on the intentional homicide was not as worrying since the accuracy is still high. Features such as city and agency name were used to select 50 instances with the highest counts of negligent homicide in an effort to solve this issue. The final result is not surprising due to the disparity between negligent and intentional homicide numbers; however, the model did learn with the limited data available.

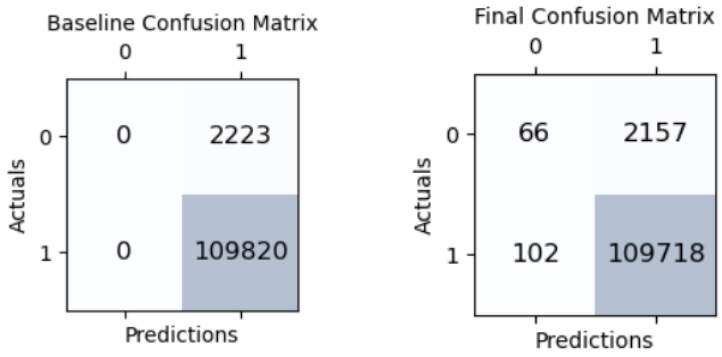

Figure 12: Confusion Matrix Predictions

6 Conclusion and Future Work

After optimizing the model with feature selection and running the features through a logistic regression test. The results showed that although we can accurately predict instances of intentional homicide, we cannot do the same for negligent homicide with the data contained in this dataset. Prediction accuracies for negligent being about 3% and for intentional being around 99.9%. In the future it would be more promising to utilize more negligent homicide data. Due to

the nature of negligent homicide in legislation[3], a perpetrator occupation feature could prove to be beneficial for those in fields which they are required to care for the well-being of others.

References

[1] Murder Accountability Project (2015). *Homicide Reports, 1980-2014*.

https://www.kaggle.com/datasets/murderaccountability/homicide-reports

[2] *The Charlotte news. (Charlotte, N.C.) 1890-1914, April 13, 1906, Image 7*. (1906). *1906/04/13*, 7.

[3]*CORE CRIMINAL LAW SUBJECTS:* (2003) *Crimes: Article 134—Negligent Homicide*.
https://www.armfor.uscourts.gov/digest/IIIA85.htm

[4] *Chapter 782—2011 Florida Statutes—The Florida Senate* (2011).

https://www.flsenate.gov/Laws/Statutes/2011/Chapter782/All

[5] *Elder Abuse Laws (Criminal)*. (2012, January 11). State of California - Department of Justice - Office of the Attorney General.

https://oag.ca.gov/dmfea/laws/crim_elder

[6] *California Legislative Information* (1982). Law Section.
https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?sectionNum=270.&lawCode=PEN

[7] *California Legislative Information* (1872). Law Section.
https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=PEN&sectionNum=192

[8] ritvikmath. (2017). *Gun Violence in America: An Analysis*.

http://ritvikmath.com/Gun_Violence_in_USA/#data

[9] Xu, R., Xiong, X., Abramson, M. J., Li, S., & Guo, Y. (2020). Ambient temperature and intentional homicide: A multi-city case-crossover study in the US. *Environment International*, *143*, 105992.
https://doi.org/10.1016/j.envint.2020.105992

[10] Hargrove, Thomas *Murder* (2015). *Murder Accountability Project*.

https://www.murderdata.org/