Classification of Victim Race for Homicides in the United States

Akshay Reddy
MS in Computer Science
Indiana University, Bloomington
reddyak@iu.edu

Niketh Shetty
MS in Data Science
Indiana University, Bloomington
nhshetty@umail.iu.edu

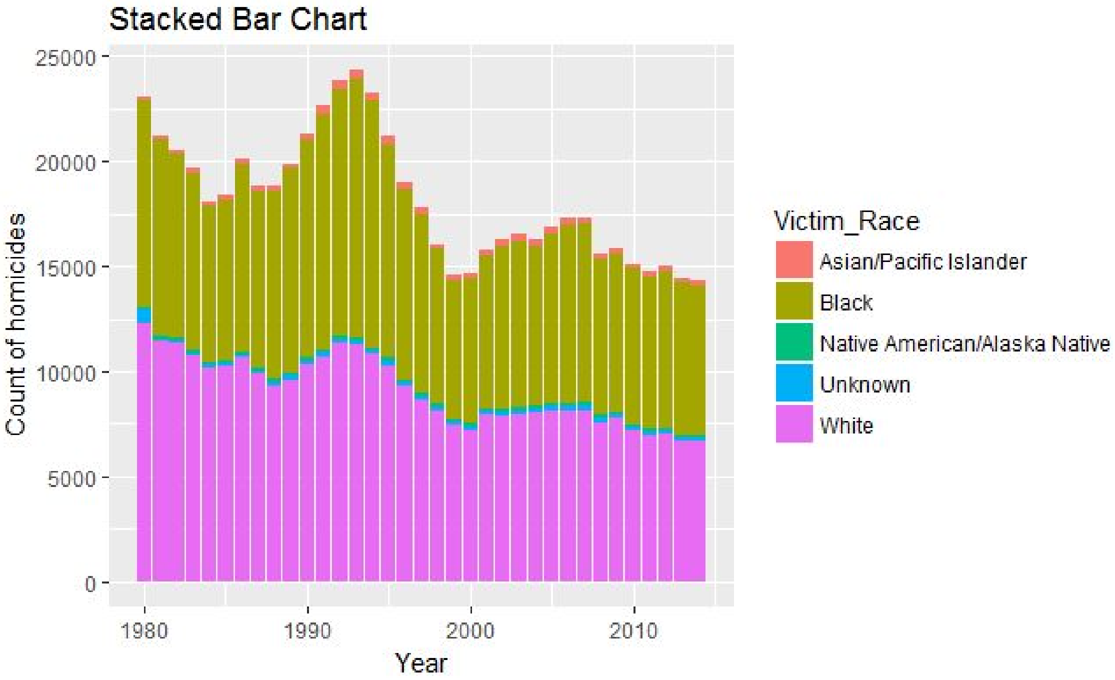
Abstract:

Can the race of a homicide victim be predicted by classification? This project attempts to classify the race of a homicide victim based on certain predictors. The goal of the project is to identify the factors that lead to homicides for each race. We use decision tree classification to build a model that classifies each case into a particular race. Our results tell us that the location, victim age and victim sex are the most important factors when attempting to predict the race of the victim.

Visualisation and Nature of the Dataset:

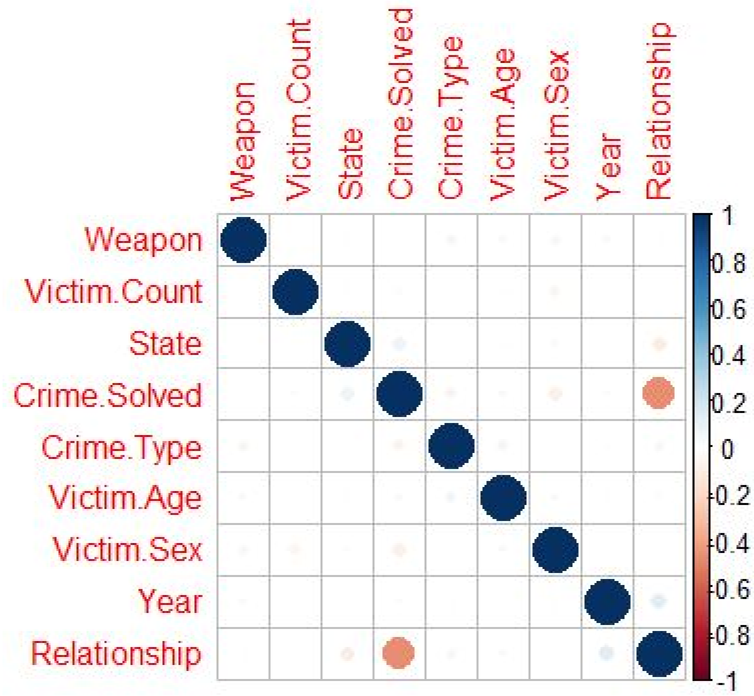
The dataset we use in this project is the Homicide Reports, 1980 - 2014 dataset available on the Kaggle[1] website. This is a database of the homicides in the United States from 1980 - 2014, compiled by the Murder Accountability Project, sourced from the FBI's Supplementary Homicide Report as well as the Freedom of Information Act. There are 638,455 rows and 24 columns. The columns (or attributes) are: Record ID, Agency Code, Agency Name, Agency Type, City, State, Year, Month, Incident, Crime Type, Crime Solved, Victim Sex, Victim Age, Victim Race, Victim Ethnicity, Perpetrator Sex, Perpetrator Age, Perpetrator Race, Perpetrator Ethnicity, Relationship, Weapon, Victim Count, Perpetrator Count and Record Source.

Victim Race is the attribute we are predicting. There are five classes: Black, White, Asian/Pacific Islander, Native American/Alaska Native and Unknown. The following image shows us the counts for each class, per year. To obtain this image, run the file "Stacked_Bar.R".



We considered 8 out of the remaining 23 attributes for our classification. Many of the attributes could not be used because they had no logical relation to the race of the victim. We decided not to consider the Record ID, Agency Code, Agency Name, Agency Type, Incident and Record Source for this reason. We decided not to consider any details of the perpetrator (Perpetrator Sex, Perpetrator Age, Perpetrator Race, Perpetrator Ethnicity and Perpetrator Count) because our goal for this classification is to predict the race of people most likely to be homicide victims to aid in its prevention and the perpetrator only comes into play after if a homicide has occurred and not before it. We decided not to consider City because many states have cities with the same name and that would cause problems for the classifier. Similarly, we decided not to consider the Month attribute either.

We then decided to check for any redundancy in the remaining attributes but checking their correlation. We used the ‘corrplot’ library in R to plot the correlation of between each of the attributes. We obtained the following figure:



The attributes Relationship and Crime Solved are the only attributes that are correlated. We ran the classifier with both attributes, without the Relationship attribute, and without the Crime Solved attribute. We obtained the best balance of performance and accuracy when we dropped the Relationship attribute.

The last bit of preprocessing we performed was to convert the data into numeric form to make the classification smoother.

Approach:

We attempted to classify the data using different classifying methods with the intention of using the classifier with the most favorable balance of accuracy and speed. We zeroed in on three classifying methods: Naive Bayes Classifier, Support Vector Machine and Decision Tree Classifier.

We first attempted to classify the data using the Naive Bayes Classifier. Using the naiveBayes() function from the ‘e1071’ library, we built a model on a training set and then tested it. We obtained accuracies of 55.1% on the training data and 54.9% on the testing data.

We then attempted to classify the data using a Support Vector Machine (SVM). We collected at random sample of 1000 rows from the dataset. We split the sample into training data and testing data. A model was then built on the training data using the svm() function that is a part of the ‘e1071’ library. This model was then tuned to obtain maximum accuracy. Upon testing the tuned model, we obtained accuracies of 58.8% on the training data and 44.1% on the testing data. Another drawback of modeling with SVM was that it was very time consuming, especially for a such a large dataset.

We then attempted to classify the data using a Decision Tree Classifier. The dataset was split on a 60/40 basis into training and testing data respectively. We used the C5.0() function from the ‘C50’ library to build a model on the training set. When tested, this model gave us an accuracy of 72.2% on the training data and 68.8% on the testing data. This classifier was also the fastest in to build the model, taking just 4.3 seconds. This informed us that the Decision Tree Classifier was the best method to classify the data.

Outcome and Interpretation:

The confusion matrix for the model when tested on the training data is:

	Asian/ Pacific Islander	Black	Native American/ Alaska Native	Unknown	White
Asian/ Pacific Islander	326	1220	7	1	4392
Black	16	119346	24	67	60265
Native American/ Alaska Native	7	456	251	1	2082
Unknown	13	1138	7	778	2167
White	113	45983	119	106	144187

The confusion matrix for the model when tested on the test data is:

	Asian/ Pacific Islander	Black	Native American/ Alaska Native	Unknown	White
Asian/ Pacific Islander	151	871	8	5	2909
Black	21	77162	19	53	42926
Native American/ Alaska Native	3	289	115	3	1360
Unknown	6	748	6	437	1376
White	146	33282	117	100	93269

Majority of the population of the United states is either White or Black. Naturally, the victim's race is more likely to be one of the two. But if we look at the above tables, another observation that stands out is that the number of black homicide victims is disproportionately large when considering the fact that Black people make up less than 15% of the U.S. population[1].

A summary of the decision tree gives us the following information about the importance of each attribute:

100.00%	State
100.00%	Victim.Sex
99.10%	Victim.Age
74.71%	Weapon
72.32%	Crime.Type
68.77%	Victim.Count
54.31%	Year
38.18%	Crime.Solved

This tells us that the state, age and sex of the victim are the most important factors to consider when predicting the race of the victim. In other words, the race of the victim depends very strongly on these attributes.

References:

[1] https://www.kaggle.com/murderaccountability/homicide-reports

[2] https://en.wikipedia.org/wiki/Race_and_ethnicity_in_the_United_States#Racial_makeup_of_the_U.S._population