Fatal Police Shooting

Problem Statement

Over recent years, police brutality has become a big topic of discussion in the United States. Therefore, we are analyzing data from 2000-2016 on fatalities caused by police in order to see if there are any factors that affect fatalities, especially factors such as race, age, and whether or not the suspect was armed. This is an important problem to analyze since it could possibly expose police bias towards certain groups of people. The results of our model can be used to target police bias in order to reduce unjust fatalities as well as address a public health crisis for minority groups.

Data

We found our data from : *https://data.world/awram/us-police-involved-fatalities*

The data describes the individuals who are victims of police-involved fatalities in the US from 2000 - 2016. There are 26 features in this dataset, which are:

- Id
- Name
- Age
- Gender
- Race
- Date
- City
- State
- Manner_of_death: how the victim died
- Armed: whether or not the victim was armed and the tool they had
- Mental_illness: whether or not the victim had signs for mental illness
- Flee: whether or not the victim flee

Method

Data Cleaning

Our dataset has a significant number of missing values in some of the features. These features included ‘armed’ and ‘race’ with 5,677 and 3,965 missing values, respectively. There were missing values in other columns as well, but these rows were able to be removed without issue.

In order to account for both race and armed as a factor, we are going to analyze three separate data sets. For the first dataset, we are removing the race and armed columns. For the second dataset, we are removing the armed column as well as the rows without race data, in order to analyze race as a factor. For the third dataset, we are removing the race column as well as the rows without armed data, in order to analyze armed as a determinant factor. In each DataFrame, we deleted records with missing values.

Feature Engineering
Since a lot of our features are categorical data, we had to One-Hot Encode our features to get numerical values to cluster. After one-hot encoding, we realized that the armed column had a lot of uncommon weapons that would increase dimensionality. Therefore, we decided to categorize the armed data into four categories: gun, knife, unarmed, and other.

Hot encoding creates a lot of dimensions to our data, so we did PCA to reduce them to avoid the curse of dimensionality. Prior to PCA, we scaled them first to have a mean of 0 and standard deviation of 1. Every one of our data points included the date, which we decided to separate into day, month, and year.

Challenges

One of the challenges we ran into was missing values in some of the features that we thought would be significant in the data analysis. These features included armed and race 5,677 and 3,965 missing values, respectively. In order to account for both race and armed as a factor, we are going to analyze three separate data sets. For the first dataset, we are removing the race and armed columns. For the second dataset, we are removing the armed column as well as the rows without race data, in order to analyze race as a factor. For the third dataset, we are removing the race column as well as the rows without armed data, in order to analyze armed as a factor. However, at the end of our data analysis, the splitting of the data into three sets did not help procure meaningful results so we decided to just delete all the rows that had NaNs for either of the race or armed categories.

We attempted to separate the data by city, however, there were too many features even after PCA was used to reduce the dimensions. Therefore, we decided to just split the dataset based on the state that the crime took place in. Another challenge we encountered with dimensionality was the one-hot encoding of the armed data which produced too many columns, which we solved by categorizing the armed data into gun, knife, unarmed, and others. Towards the end of the data analysis, we found that there were not many instances of other items when compared to gun, knife, and unarmed.

When clustering the data, we found that the clustering was not as discrete as we hoped it would be. Therefore, we decided to go with a more recent dataset with a smaller amount of data points to see if the clustering was clearly defined. In the end, we ended up only using data from the states of TX and CA.

Results

On a high-level, our results led us to believe that there is not a concrete correlation between police shooting and our features (i.e. race, armed, etc.) but this might have been a result of the shortcomings of our initial data used and the fact that we had to reduce the dimensionality of the data. This is further discussed in the next section. Even the slight correlations that we were able to notice after conducting further analysis on the data proved not to be useful as all of the graphs showed there to be an approximately equal correlation between fatalities and any of the other factors that we could consider (i.e. mental illness, all states, whether suspect was fleeing, etc.). So without the inclusion of more pertinent data such as the location of the shooting, the mean income of those locations and more, we find it difficult to draw resolute and data-driven conclusions that would allow us to discover particular trends for these shooting.

Next steps

If we are able to gather more recent data along with whether or not the police officer had a body-cam, we may be able to draw new conclusions from the new clustering data. Additionally, more feature engineering could have been done on some of the removed features, such as the city in which the fatality took place. Due to the dramatic increase in features because of Hot-Encoding, we decided to eliminate it from the dataset. However, it is possible that a city’s population, average income, local laws or many other factors can be used to recluster our data in the future. By adding city population data, we could calculate the percentage of fatalities caused by the police per racial group in order to identify if any racial group is being disproportionately targeted. Average income would be valuable data to add because aside from evaluating if there are any disparities between income class, we could also evaluate if there is any correlation between fatalities of those with mental illness and their income. However, this could address another problem: inaccessibility to mental health care in lower income communities. As for local laws, these pertain to whether or not the police are required to wear body cameras. We believe that police officers required to wear body cameras are less likely to unjustly kill civilians, which we would be able to analyze with the added data.