Midterm Report: Predicting Civilian-Officer Outcomes

Lucy Du (dd483), Elina Hvirtsman (eh582), Julia Ng (jen67)
November 1, 2020

1 Problem Statement

Our project's goal is to understand what factors most influence civilian and police interaction outcomes and determine if there is evidence to suggest racial bias in Minneapolis police precincts. Given demographic and situational factors, we hope to develop a model that can predict whether the civilian will be issued a citation in the event of being stopped by police, or if the civilian will experience police force in more severe situations.

2 Dataset

The two main datasets are sourced from Open Minneapolis. The goal of the Minneapolis Open Data Portal is to encourage access to data managed by the City of Minneapolis, making content available to be freely used, modified, and shared by anyone for any purpose. The open data portal does not represent all the public information managed by the city, but it is meant to make the most frequently requested and most useful data easily accessible with the intent of adding content over time. The datasets are refreshed on a daily basis by 9:30 AM, and the website will reflect the last time the data set was updated and the total count of rows. For the purpose of this project, we use the finalized dataset published on October 24, 2020.

The Police Stop dataset spans from July 6, 2017 to October 24, 2020. It provides basic information about the civilian like race and gender. It also provides information about each civilian-officer interaction like if there was a search of the person, search of the vehicle, and the specific problem, such as suspicious person, traffic law enforcement, attempt pick-up, etc. There is also a call disposition variable that outlines the result like booking, advised, tagged, etc. Police precinct, neighborhood, and latitudinal/longitudinal data is present too. Our outcome variable of interest is citation issued, which contains either yes or no. There are 12 features and 116,165 examples in this dataset.

The Police Use of Force dataset spans from January 1, 1970 to October 23, 2020. It provides basic information about the civilian like race and sex. It also provides information about the situation like if it was a 911 call, the primary offense, and the specific problem, such as emotionally disturbed person, sound of shots fired, domestic abuse, etc. There is also a subject role variable that outlines the civilian's situation, like person in crisis, arrestee, suspect, etc. Police precinct and neighborhood data is present too. Our outcome variable of interest is force type, which contains values like bodily force, taser, chemical irritant, etc. There are 20 features and 29,787 examples in this dataset.

3 Data Cleaning

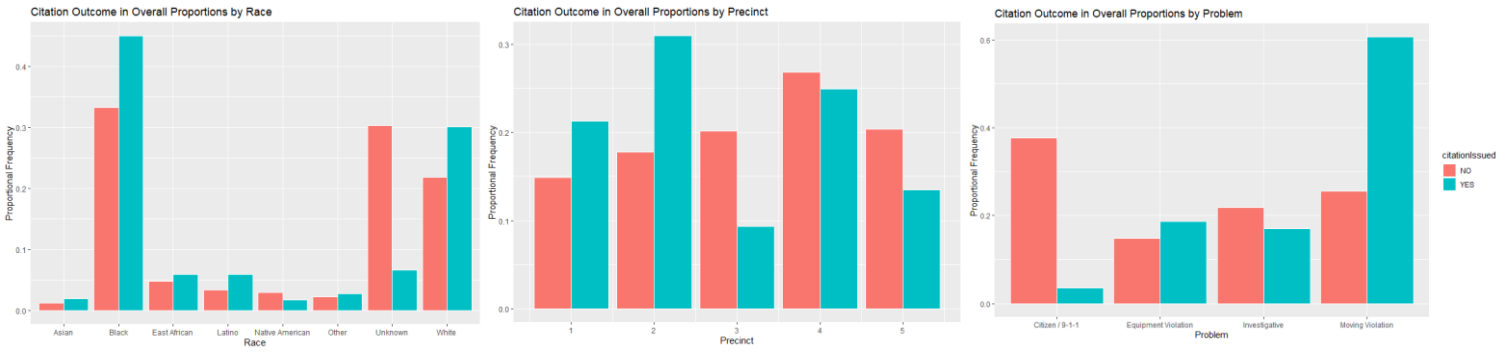
To make our datasets fit our problem statement, we dropped any observations with missing `citationIssued` or `ForceType` values. To make our dataset more approachable, we converted some nominal values into one-hot encoding. For example, the single `Race` column, which contains values like `Asian, Black, East African, Latino, Native American, Other, Unknown,` and `White`, was converted into eight columns with a 1 in the applicable column and a 0 in the others.

The datasets are comprehensive and complete with intuitive values. Histograms were created for the numeric variables like latitude and longitude to check for outliers. In the Police Stop data, there are around 700 observations with missing latitude and longitude, which are dropped given it is less than 1% of our entire dataset. Frequency tables were created for the nominal variables like race, problem, and neighborhood to check for outliers, missing, or erroneous data. Data validation was most likely employed in the data collection process as there are no known outliers in the data. There are now 115,391 examples in this dataset with around 80% no citation issued and around 20% citation issued.

In the Police Use of Force data, there are around 100 observations with missing latitude and longitude, which are dropped given it is less than 1% of our entire dataset. Frequency tables were created for the nominal variables like race, problem, and neighborhood to check for outliers, missing, or erroneous data. There are around 300 observations with non-recorded race, which are dropped given it is around 1% of our entire dataset. There are now 29,302 examples in this dataset with the following proportions for `ForceType` with `Bodily Force` the most common at around 75%.

4 Descriptive Statistics

Before modeling, we conduct preliminary descriptive analysis to understand our dataset better. For the Police Stop data, we plot frequency plots for citation outcome by race, precinct, and problem. The following plots show that Asian, Black, East African, Latino, Other, and White have relative greater citations issued than not within their own race category. Overall, Black, Unknown, and White have the absolute highest citations issued. Precincts 1 and Precinct 2 have relative greater citations issued than not within their own precinct category. Overall, Precinct 2 and Precinct 4 have the absolute highest citations issued. Finally, Moving Violation has the highest relative and absolute, and Citizen/911 has the lowest relative and absolute citations issued. This could be an important predictor in our later models.



5 Preliminary Model Fitting

6 Future Plans