Predictive Policing Data Analysis

Chaile Grandy

University of Maryland, College Park

INST414: Section 0201

**Problem**

Predictive Policing models seek to reduce neighborhood crimes, but instead are unethical and biased modes of surveillance on people of color and low-economic communities. Predictive policing systems take in historic and cycled data to predict patterns of crime. This data is then used to surveillance certain neighborhoods or areas that are likely reported to have a large rate of crime. The data also takes in biometric information such as facial recognition or thumb prints to predict the chance of repeated crimes from past activity. However, both methods of policing (neighborhood surveillance and biometric scanning) have led to many complaints of discrimination and unlawful policing.

Predictive policing is applicable to predictive analytics, by models and the information that they digest. The data used in such models are algorithmic: the results given to police officers and departments are based on patterns within the data. If the data inputted into the model includes information relating mostly to people of color who primarily live in low economic environments (because of pattern recognition in arrests and high crime activity) results will reflect, even though not always accurate.

For example, a robbery took place in a Texas Sunglass Hut, in which police relied on their predictive facial recognition tool to identify the suspect. The model identified a man who was nowhere near the scene at the time and was in another state (California). The reason being, is that he committed a similar crime not too long ago and had lived in Texas prior to moving. Not to excuse his prior behavior, however, he was not at fault for this incident. But he was still arrested, prosecuted, and sexually assaulted as a result. Though there was no real evidence aside from the model used to identify the man, he was still falsely accused and dealt with unlawful treatment.

The datasets that are being used by predictive policing models pose a few threats, the most common being privacy risks along with racial, social, and economic discrimination. If this problem is not solved, police departments will continue to use predictive policing models as a bypass to falsely imprison individuals or suspected criminals with acts of stop-and-frisk and people of color will continue to be at risk of experiencing racial bias along with the exploitation of their privacy and security.

# Scope of work

To address this problem of predictive bias and surveillance, it is important to find ways of intervention within police departments and predictive policing models. This looks like retraining on both ends to better handle incoming data. With police departments, this would mean providing better training on how to interact with suspects, better implementation of the algorithmic assessment tools used to develop models and better training on privacy using transparency (so notifying people that they are being watched or investigated).

To improve predictive policing models, it is important to remodel the data to include only important/relevant data, identify repetitive trends or errors and adjust as needed, as well, include more diverse datasets to reduce the chances of bias.

Policing models like HUNCHLAB have already begun incorporating some of these strategies. HUNCHLAB is a predictive policing model that has dealt with complaints of the biases listed above. As a result, researchers have implemented new strategies to improve the model and reduce instances of discrimination: HUNCHLAB included only publicly recorded crimes in their data and added more randomness into their programs to reduce the level of bias in its datasets; police officers were also sent to random locations to reduce racial/economic bias and risk levels of encountering dangerous situations. So, the goal for this project would be to develop a new model that limits the rate of bias by including unbiased/diverse code, utilizes assessment tools, provides transparency with suspects, and includes random datasets to offset chances of bias.

**Evaluation Criteria**

To determine the effectiveness of this new model, an assessment tool would have to be created to measure the accuracy of the newly refined datasets that are used to build predictive models. Factors of race and poverty would have to be heavily observed to determine the risk factor of each and how likely patterns of bias are to be identified within the data. As a template, an SSL (Strategic Subject List) could be used to help refine models. SSL is a database of 300,000 individuals that are assigned a score based on their likelihood to be involved in crime activities more specifically gun violence. This at one point was a commonly used predictive policing tool used in Chicago but received a lot of complaints of bias and discrimination, so it was dismantled. However, it can still be used as an integral part to combat racial bias in predictive policing models. By, using this large dataset, it is possible to identify the racial and economic risk factors mentioned above as patterns within the data and use this information to build a new dataset that can be used to rewire models by reducing algorithmic bias.

# Data Validation & Cleanup

The goal for my data is to use the first dataset as a template on how I will analyze my dataset and the models that I will use. The other dataset contains the actual data that I will be using to observe and use to create my data visualizations. The factors I am using to determine the effects on race and socio-economic status will be incident reason, whether an arrest has been made, race categories (Black, White, Hispanic, Asian), and income by race. However, there are some inconsistencies, limitations, and missing values within the data. For instance, a lot of the data that I am using consists of categorical data and is merged from two different datasets, so running the statistical tests may present unwanted issues.

Exploratory Data Analysis

Originally, for this dataset, I was going to use a KNNClassifier, Poisson Regression model, and a Ridge Regression model to run statistical tests. The KNN model would be used to identify patterns within the data that may point to instances of racial bias. The Poisson Regression statistical test would reveal the number of arrests made per race and hopefully provide accurate results of unfair bias within the dataset. The last test (Ridge Regression model) would be used to identify the margin of error within the data and then develop more accurate calculations and visualization for the tests being used above. However, after further research and partially due to one of the empirical studies listed below, I have chosen to use a XGBoost, RF, and Linear Regression Model to accurately run tests on my data.

Refining

According to one of the empirical studies used, (*Learning predictive analytics with Python: gain practical insights into predictive modelling by implementing predictive analytics algorithms on public datasets with Python*), the author analyzed the data by utilizing XGB and RF models to accurately run their calculations and receive better results. Looking at my dataset, these two models would probably be a better option opposed to a Poisson Regression and a Ridge Regression model. Both models are designed specifically for classifying algorithmic based data. The XGB model is designed to provide results quicker and more accurately for algorithmic performance. Whereas RF is used to do both tasks of classification models and regression models. So, with my dataset, I would receive quicker results that are more accurate in predicting the rate of racial bias in the data. Additionally, I can do more with less as I do not have to include the KNNClassifier as well because the RF model consists of this. So, due to this new information, my tests along with results should change.

Challenges & Findings

Since the beginning, looking at the original dataset, there have been lot of complex challenges and findings. Since the data was merged from multiple datasets, it was difficult to ensure the predictive and targeted variables provided accurate results. Running the tests were also difficult because the original dataset includes both categorical and numerical values. So, this hindered my progress a lot and prevented me from discovering more data. Along with this, I reconsidered different models as I found within my research that some can transform datasets better than others. Overall, there were many challenges or findings along the way that slowed my progress, however, I was still able to gather insightful and accurate data relating to predictive policing models. In the future, these are factors that I would have to be mindful of and ensure to ensure the project moves smoother.

Future Work

For future improvement, running all models would be helpful to generate more accurate results and allow for more data manipulation. A lot of limitations were placed on the dataset, due to errors in running the actual models leading to potential instances of missing or inaccurate data. So, in the future learning how to run certain models and revising them to align more with the goals of the project.

In general, the data that was shown throughout the project, could be used to improve other algorithmic models. Bias was detected as it relates to race with an 80% accuracy result. So, using this data researchers and developers can find margins of error and make improvements as needed or deploy policies that will help to mitigate some of the racial biases that occur in such policing models.

Data Visualizations

Figure 1.

Include all figures in their own section, following references, footnotes, and tables. Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.

A screenshot of a graph

Description automatically generated

Figure 2.

A screenshot of a computer screen

Description automatically generated

*Figure 3.*

A graph of different colored bars

Description automatically generated with medium confidence

A graph of blue bars

Description automatically generated

A graph of a number of numbers

Description automatically generated

A graph of a bar graph

Description automatically generated

References

Diaz, A. (2021, September 13). *Data-driven policing’s threat to our constitutional rights*. Brookings. https://www.brookings.edu/articles/data-driven-policings-threat-to-our-constitutional-rights/

‌Zuvanich, A. (2024, January 23). *Man, falsely accused of robbing Sunglass Hut in Houston suing companies overuse of facial recognition software*. Houston Public Media. <https://www.houstonpublicmedia.org/articles/technology/2024/01/23/475288/man-falsely-accused-of-robbing-sunglass-hut-in-houston-suing-companies-over-use-of-facial-recognition-software/>

DaViera, A. L., Uriostegui, M., Gottlieb, A., & Onyeka, O. (Cynthia). (2023). Risk, race, and predictive policing: A critical race theory analysis of the strategic subject list. *American Journal of Community Psychology*, *73*(1-2). <https://doi.org/10.1002/ajcp.12671>

<https://www.kaggle.com/code/adiamaan/scalable-data-cleaning-and-racial-analysis/input>

<https://www.kaggle.com/datasets/center-for-policing-equity/data-science-for-good>

Kumar, A. (2016). Learning predictive analytics with Python: gain practical insights into predictive modelling by implementing predictive analytics algorithms on public datasets with Python (First edition). Packt Publishing.

Li, A., Shalaginov, M. Y., Tao, A., & Zeng, T. H. (2023). Investigation of Racial Bias in Property Crime Prediction by Machine Learning Models. 2023 International Conference on Machine Learning and Applications (ICMLA), 2253–2256. https://doi.org/10.1109/ICMLA58977.2023.00340