Ethical Data Analysis of Crime in Communities

https://github.com/ben-walczak/Ethical-Community-Crime-Data-Analysis

Shion Guha

Assistant Professor

Director of Data Science

Graduate Program

Marquette University

Milwaukee, WI

Benjamin Walczak

Undergraduate Student

Bachelors of Computer

Science and Data Science

Marquette University

Milwaukee, WI

# Abstract

Law enforcement agencies benefit greatly from the prediction of crime within communities across the United States of America. Using this information, they determine how to allocate resources to minimize and prevent crime. The prediction of crime within communities can often lead to discriminatory results by unfairly representing communities of underprivileged ethnic backgrounds. An analysis of ethical practices is explored within the realm of prediction of crime within communities to prevent further discrimination.

# Author Keywords

Communities, Crime, Ethics, Data Science

**Literature Review**

The Mythos of Model Interpretability examines many types of models and the trade-offs between bias and algorithmic transparency. Some of the many models analyzed are deep networks, linear models, etc. Deep neural networks have the benefit that they do not alter features pre-model training. However, linear models suffer from the need to normalize the data in a variety of ways to support better results. Altering the original data through normalization and other means has the potential to fog the meaning behind the original feature. Although the case can be made to include transparency in all models, many are against the push to include transparency. Take for example, the healthcare industry, “As a concrete example, the short-term goal of building trust with doctors by developing transparent models might clash with the longer-term goal of improving health care. We should be careful when giving up predictive power, that the desire for transparency is justified and isn’t simply a concession to institutional biases against new methods.” [6] Often a post-hoc interpretation of models can be misled by the confirmation bias. This belief will have the programmer design the algorithm in the context that they believe it should work reinforcing their beliefs. [6]

Bias in Computer Systems analyzes bias in early computer systems. Although the paper was written with a focus on computer systems, many of the concepts can be adopted into data science practices. The three main types of bias identified by Friedman are preexisting, technical, and emergent biases. Preexisting bias has its roots in individuals or institutions either on purpose or unintentionally. This bias can sneak through the design process of computer systems, often emerging as unfair results much further down the process. Emergent bias surfaces during the context of use. This type of bias emerges from how the design of the system interacts with the outside world. In general, Friedman defines bias as discrimination of an individual or group in favor of another due to the denial of an opportunity or good due to unreasonable grounds. [5]

“The best explanation of a simple model is the model itself...” [7] However, sometimes models can be complex and difficult to decipher. To counter this, many methods have been devised to help explain models. One of the more popular methods to do this is additive feature attribution methods. This works by summing the effect of each feature, which approximates the output of the original model. In this paper, SHAP (SHappley Additive exPlanation) Values are proposed as a method of measuring feature importance. To predict SHAP values algorithmic game theory is applied to reward each feature on how it contributes to the total output. This algorithm is closely designed to Shapley regression. The benefits of using SHAP values are that it provides accurate and interpretable results. It is intuitive to understand that each feature contributes to the output of the model differently, and thus with measured effects, one could understand how each feature effects the final output of the model. [7]

The Los Angeles Police Department analyzes a group of neighborhoods in and around LA. General statistics of each area are recorded and used to predict crime. Some of the statistics captured are population, gender, ethnicity, poverty levels, etc. According to the regression of number of crimes and population, the two variables had a strong positive relationship. “… density, unemployment rate, and poverty rate are also correlated…” [8] On average, males are victims of crimes more than women; however, women are much more likely to be victims of theft and kidnapping. Although people of white descent have a higher number of victims, people of black descent are more likely to be victims of crime. Overall, population size, poverty rate, and unemployment were the highest predictors of crime. [8]

**Data Collection and the Dataset**

The Dataset is the result of combining two different datasets, one from the Census Bureau regarding community and household information, the other from the FBI Uniform Crime Reporting division. Although both agencies operate slightly differently, they both have very respectful standards for data collection and data privacy. The dataset was found within the UCI Machine Learning Repository. It was uploaded by Michael Redmond of La Salle University.

The Census Bureau collects data from a variety of sources to ensure accurate and quality data. Data is collected automatically via website access, American FactFinder (cookies made for online surveys, which are deleted after the closure of that survey), Electronic Census Bureau Surveys, etc. Using the variety of data collected, the Bureau compares and contrasts data sources to eliminate inconsistencies and anomalies within the data. Every randomly sampled participant selected for surveys are informed of every detail of both the survey they will take and the implications of that survey.

Raw data collected is confidential and protected by the federal law. Only after the data has been stripped of personally identifiable information among other things is the information made public. The Census even has a privacy statement, which says, “We do not collect personally identifiable information (name, address, e-mail address, social security number, or other personal unique identifiers) or business identifiable information on our web sites unless we specifically advise you that we are doing so.” It is key to the Census Bureau that data is made available public and accessible to all, while still respecting privacy and PII. The Bureau strictly follows ASCII and HTML formats and text describing graphics for viewable data, so that documents are accessible for those using screen reading devices.

The FBI UCR, the Uniform Crime Reporting group, similar to the Census Bureau collects their data from multiple sources. According to the UCR, more than 18,000 law enforcement agencies contribute data to the FBI in a given year. These agencies are required to follow strict data recording rules to maintain consistency and accuracy. The main use of the data is to determine funding, planning, resource allocation, and assessment of police operations in order to combat crime to the best of their ability. Therefore, it is crucial that their recorded data is accurate. However, agencies knowing this information have the potential to misconstrue numbers for additional funding. Although I am doubtful this is actually the case as I am sure there are many safeguards against this.

As hinted upon before, quality assurance is one of UCR’s top priorities. Data quality reviews often involve double checking those involved and other details via phone call, email, in person interviews, or by other means. Using these methods, the UCR can check for errors and anomalies on a timely basis.

The CSMU, crime statistics management unit, exists to ensure that UCR possesses quality data. They check data for logical consistency, reasonableness, adherence to sound estimation methodologies, adherence to monetary submission standards for stolen and recovered property, and other checks for other statistical methodologies.

All data recorded by the UCR follows the same law that the Census Bureau follows, “(e.g., Title 13, United States Code Section 9 (a)) protect against unauthorized disclosure under penalty of $250,000 and up to 5 years in prison, or both.” Therefore, all data is stripped of personally identifiable information unless the law officials request further information. Upon request one can get access to specific documents within the UCR. However, master documents also exist of general state and national crime statistics. Given discrepancies or personal information is found within the data an official request to remove the data can be made to the UCR.

Overall, both sources of data are very reliable. Both the Census Bureau and the FBI UCR follow strict guidelines that adhere closely to Deon’s general checklist for ethics of data science practices. Not all details could be found regarding the checklist, but given the strict guidelines both organizations have, the data should be sufficient to satisfy all ethical concerns.

The only concern with the given data is that the sources of data come from different dates. The Census Bureau data was recorded in 1990; whereas, the FBI UCR data was collected in 1995, five years after the Census Bureau data. Given the difference in time, discrepancies in the data will exist. Whether it is a minor or major impact the sample data will not match the population exactly. Whether this concern persists through the entire process is a question for future analysis.

**Preprocessing**

The dataset contains 147 features. 8 features were identified to be potential response variables, some of which include per capita violent crimes and per capita nonviolent crimes. For the sake of simplicity, this analysis will only focus on per capita violent crimes. Some variables were of data type string. These features included Ecommunityname and state. Although these features will not be used for the intermediate analysis due to their datatype and potential bias, these features may prove to be advantageous in geospatial analysis. Some communities were counted more than once. Given they contain the same values within each observation they may be discarded. The states seem to be distributed somewhat relative to each state’s relative population.

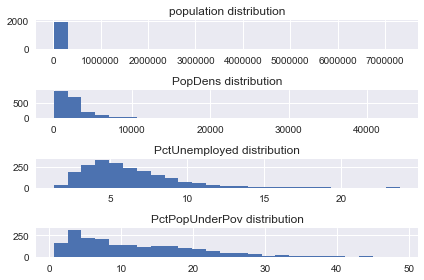
After briefly looking at the tail end of the dataset, it was discovered that missing values exist within some of the features. A further analysis of missing values was done by counting the number of observable values for each column. Most features with missing values had anywhere from around 300 observable values to around 900. Since the dataset contained over 2000 observations, any feature with less than 500 observations or less than 25% of the total observations were discarded due to lack of data. The few features with relatively 900 values were discarded due to their data type not matching the desired data type for future prediction. Note that these values are not discarded and forgotten; they may be used at a further point for geospatial analysis. After discarding those features, only a few values were missing from a few columns. These observations were simply discarded to simplify and speed up analysis.

After the removal of unnecessary features, 103 features still remained. Therefore, further feature elimination must be done. To remove more features a function via scikit learn module was used to perform chi squared test to determine the most predictive features with regard to per capita violent crimes.

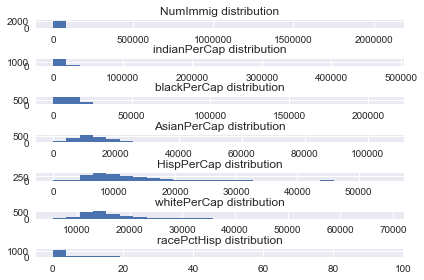
The list of top features contained 30 of the most predictive features. Some of these features included population and poverty rate, which was said to be strong predictors by the Crime and Demographics paper. [8] Other features included statistics concerning the percentage of immigrants within a town and the percentage of a given race. These features have the potential to be very biased in our final results. Therefore, multiple groups of features were created to compare how including and excluding certain features can affect the final accuracy of the model. The feature groups created for further analysis were the top 30 features, the top 30 features excluding immigration and race related features, the 4 most predictive features mentioned by the Crime and Demographics Paper, and all features for reference. [8]

**Exploratory Analysis**

The first task within exploratory analysis was to explore the distribution of features. This can be helpful to determine whether scaling or normalization is needed with respect to the model. In addition, technical bias could emerge if we do not pay attention to the distribution of features before fitting the model. After a brief look at the distributions of the features, it was discovered that the majority have right skewed or positive skew distribution. [Figure 1 & Figure 2] This means that the mode of the data occurs at lower values before the mean occurs. Some models like linear models assume normal distribution of variables. Therefore, we will need to normalize these features before fitting a linear model. Some feature distributions had a more exaggerated right skew than other feature distributions. In this case, the distribution plot appeared to only have one bar that to the far left. A further analysis via boxplots was done to better understand these features. [Figure 3] It was discovered that these particular features had outliers that distorted the distribution graph.

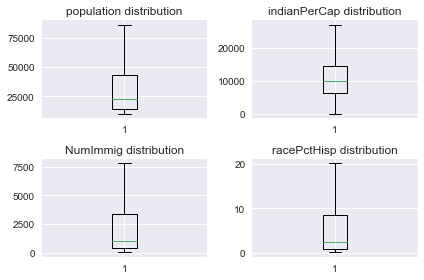


Figure



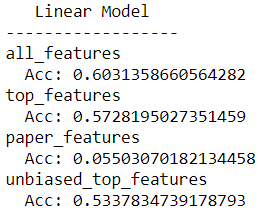
Figure

Population, Indians per capita, number of immigrants, and race percentage Hispanic were among the several features with extreme outliers. Outliers can affect models immensely. To combat this, we can either remove these outliers altogether or we can scale these features to fit the models better. Because we want the data to match the population as best as possible to not misrepresent some groups, the data will be scaled rather than the observations being removed. Understand that both scaling and the removal of observations have the potential to insert unforeseen technical bias into the model. More analysis needs to be done to understand how scaling and removal of observations can affect bias in the final model. The biggest take away from the current exploratory analysis is that normalizing and scaling should be done for the best results in the final model.



Figure

Investigating prediction accuracy and bias was also done to better understand how features affected a given model. As mentioned before, many feature groups were organized, some with biased features and some without. Using these feature groups, we can explore how unbiased or biased features can affect the model and its accuracy, thereby investigating how accuracy varies with bias. The number of violent crimes within a community will be predicted by the model via the means of regression. These feature groups will also be fitted to multiple models for a better understanding of how accuracy changes. The four models being fitted are linear regression, decision tree regressor, basic neural network, and a linear support vector machine. These models were chosen based on their popularity. Depending on the model being fitted different versions of scaling and normalizing will be used to ensure better accuracy. Further work still needs to be done to optimize hyperparameters of each model. Overall, the accuracy only varied slightly between different feature groups. When all features were fitted to the models, the models achieved their best accuracy. The accuracies of the following feature groups top features, paper features, and unbiased top features were ranked 2nd, 3rd, and 4th accordingly. Interestingly, despite the feature group unbiased features containing more features than paper features, unbiased features had the worst accuracy among all feature groups. Note that the feature group paper features only had a 5% drop off compared to feature group all features in accuracy, using only 4 features for prediction. [Figure 4]



Figure

**Critical Analysis**

The prediction of violent and non-violent crimes with regards to a given community can help law enforcement agencies can better allocate their resources to both minimize crime and prevent future crime from ever happening. As a result, more police officers will be assigned to a given community or more will patrol a given area potentially. Often it is asked, “does more policing equal less crime?” Police officers optimize their daily routes to catch petty and major crime of a community. However, more often than not, police take interest to dealing with and preventing petty crime. The Washington Post states that, “… ‘broken windows policing’ overwhelmingly targets minority communities.” Broken windows policing is the idea that stopping and preventing minor offenses will reduce major offenses. So, is the model to predict violent and non-violent crimes in a community implicitly biased just by the purpose it serves? Without a doubt, using racial statistical features of a community would introduce bias in the model by potentially targeting communities of some specific ethnic background. Therefore, further analysis of how we can achieve accurate results while excluding biased features would be beneficial to prevent further bias. Multiple models have been tested in the initial analysis for a more varied perspective of how features can affect a model. Many models are used because different models have varying levels of transparency. Linear regression and Neural Networks have different limitations and capabilities. [6] Therefore, the use of many models would benefit as you would gain many perspectives. However, there are many limitations to this. Even at its fundamental basis, transparency has many definitions, so to know how to judge a model based on its transparency is challenging. Bias can vary in its definition as well. Friedman lays out three main types of bias preexisting bias, emergent bias, and technical bias. [5] Throughout, the analysis I attempted my best to avoid technical bias. However, when choosing different methods of analysis different bias can reemerge despite the best intentions with regard to the method. For example, during the exploratory analysis there was a choice between removing observations or scaling to prevent the effect of outliers. Regardless of the choice of methodology, either method can produce bias. The original data is being changed big or small. The original data could then be diverging from its actual population.

Technical bias should not affect the implications of the model as much as discriminatory features. These features are potentially the result of preexisting bias because they contain statistical information regarding race and number of immigrants. The collection of these features could be the result of the belief that ethnic backgrounds contribute to the amount of crime in a neighborhood. Communities with overwhelming amounts of immigrants or a majority ethnic background could be unfairly judged according to these features. Because discriminatory features have the most impact on this dataset over anything else, it will be the main focus for further research. Understanding the tradeoff between prediction accuracy and bias could aid in future ethical data science projects. More ethical methodologies could be devised through the understanding of how to better handle features according to accuracy and bias. To aid in our future understanding of these features, a model will be used to explain the effect of each feature via the SHAP value. Future analysis will implement the SHAP value to explain how features affect models. [7]

**References**

1. U. S. Department of Commerce, Bureau of the Census, Census Of Population And Housing 1990 United States: Summary Tape File 1a & 3a (Computer Files),
2. U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)
3. U.S. Department of Justice, Bureau of Justice Statistics, Law Enforcement Management And Administrative Statistics (Computer File) U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan. (1992)
4. U.S. Department of Justice, Federal Bureau of Investigation, Crime in the United States (Computer File) (1995)
5. Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330-347.
6. Lipton, Z. C. (2016). The mythos of model interpretability. arXiv preprint arXiv:1606.03490.
7. Scott M. Lundberg and Su-In Lee (2017). A Unified Approach to Interpreting Model Predictions. arXiv preprint arXiv:1606.03490.
8. Bianca Cung, Yingnian Wu, Frederic Shoenberg, and Robert Gould. (2013). Crime and Demographics: An Analysis of LAPD Crime Data. eScholarship.