STOR390 Final Paper

Jonathan Zhao

2024-12-11

Introduction

As technology continues to advance at an exponential pace, it has opened up many new solutions for addressing some of society's lasting challenges. One such challenge is combating crime, which has been impacted by the advancements in data collection and machine learning. Among these solutions is predictive policing, which aims to predict future criminal activity using historical data. One of these ideas is predictive policing, which tries to predict future crime with data on past crimes. An example of predictive policing is Geolitica(previously called PredPol), which uses historical crime data to try to predict potential hotspots where crimes are likely to occur. However, a downfall of this method is that it can't be used during an active and immediate situation. When a crime is happening in real-time, police still need to rely on traditional methods of using the bits of information gotten from the 911 call to respond to a crime. In 'Crime Prediction For Better Response Using KNN, PCA, And Random Forest', Uchenna Akujuobi highlights this issue, explaining how predictive policing does little to support the real-time decision making that police officers need to do during the event of an active crime.

The paper illustrates two good scenarios that demonstrate this downside. The first example is the situation where a victim is hiding from dangerous people and can't make a 911 call without risking their life. This causes the victim to be unable to communicate the details of their situation with the police, and the only option left is for the victim to keep hiding and hope they are not found. The second example is when a police officer receives a crime alert with little information, and goes to the crime location not knowing what to expect. This

could potentially put the police officer in a dangerous situation, or require the police officer to make quick impulse decisions which could lead to a negative outcome. In these situations, predictive policing is unable to be of any use since its main purpose is in predicting future crime, not impacting active crime. Developing a way to report and predict these crimes in real time could help police officers respond to crime calls accordingly and reach better outcomes.

This is where Akujuobi's research idea comes in. Instead of attempting to predict future crime using past data, Akujuobi's goal is to make real time predictions based on real-time information of a crime, to help police better deal with the situation. This model would entail creating a way for victims to report crimes without putting themselves in danger, such as a clicking a button on a mobile phone and a quick few words describing the situation. Next, the model would then use the information gathered to predict the type of crime being committed, and notify the police department of the possible crimes being committed. Using this information, police can make better decisions on how to respond, such as sending in specialists or additional reinforcements.

Akujuobi's research proposes training a model that would predict the type of crime at a crime scene using the information that a victim provides without needing to put the victim into danger. This model uses the Chicago Crime dataset as a base, using the following features: Location Description, Time, District, Ward, Longitude, Latitude, Primary Description(Target class) and Secondary Description(The last word).the first 5 features can be easily received from a victim's 911 call, and will be used to predict the Primary Description and the Secondary Description. These two descriptions make up the actual crime that was committed. Akujuobi then tested a variety of ML algorithms to determine the most accurate algorithm to use for this approach at crime prediction, such as random forest, decision tree, and KNN. With the highest accuracy at 88.99%, the research shows that this model has a high potential to be effective in enhancing the police's real–time decision making in the field. If put into use, it could reduce the risk that both police and victims take on during an active crime.

Analysis of Methodology

To verify the results of this research paper, I decided to replicate one of the algorithms used on the dataset. The simplest algorithm to replicate successfully would be the KNN without PCA algorithm, since it didn't require many extra steps or advanced computation. This algorithm had the lowest accuracy, it still achieved an accuracy of 74.84%, showcasing its potential to significantly enhance police officer's safety and response to an active crime.

First, I downloaded the data from the City of Chicago website. This data was updated from 2001 to 2024, so in order to verify the results, I needed to extract the data from 2001 to April 2015. To do so, I converted the Date column into a Date type, then used the dplyr package to filter the date between "2001-01-01" and "2015-04-01". This gave me all the entries between January 1st 2001 to April 1st 2015. Next, I removed all of the entries with missing data using R's built in na.omit() function. Finally, I removed duplicates from the dataset to get all the distinct rows. In order to replicate Akujuobi's model while still limiting the processing power needed for the model, I decided to do a stratified random sample. I chose to randomly sample 1000 entries from each year, from 2001 and 2015. This gave me a total of 15000 observations.

In addition, some of the column names listed by Akujuobi were not the exact names given on the dataset, so I replaced them with the corresponding column names. This included Primary Description, Secondary Description, and Time. For the first two, they were actually named Primary Type and Description. For Time, the column name is very broad, so I selected the closest match, which was Date. Below is a graph showing the features used in the KNN model:

Feature Name	Type	Description
Location Description	Categorical	Describes the location where the incident occurred.
Description	Categorical	Secondary description of the crime
Date	Timestamp	The timestamp indicating when the incident occurred.
District	Numerical	Numeric code representing the district of the incident.
Ward	Numerical	Numeric code representing the ward of the incident.
Longitude	Numerical	Longitude coordinate of the incident location.

Feature Name	Type	Description
Latitude	Numerical	Latitude coordinate of the incident location.

Next, I split the dataset into a training partition and a testing partition. Since KNN does not need to be split into a training and testing partition, I randomly sampled 150 observations out of the dataset to test the accuracy of the model. This ensures that I have plenty of observations to improve the model, while also saving time and computing power on the tests.

In the paper, Akujuobi then used KNN to predict the ward and the district of the crime, simulating a scenario with the absence of internet signal. For predicting the ward, a K value of 3 was used to achieve a prediction accuracy of 99.86%. For predicting the district, an accuracy of 99.97% was achieved from using a K value of 1. In contrast to the paper, I chose not to replicate this step in my analysis, but to instead use the pre-existing values for ward and district present in the dataset. This is because the KNN prediction had sufficiently high accuracy that the difference was negligible. The differences of that 0.2% for the ward and 0.1% for the district would not substantially affect my overall findings.

For the KNN analysis without PCA, the research used Euclidean distance as the distance metric. I replicated this by creating a function that calculates the Euclidean distance between two points. Next, I normalized the numerical features, and applied KNN on those features. Euclidean distance doesn't work for categorical features, so I decided to ignore them since the Akujuobi didn't cite any other distances used for those features. Akujuobi also fails to mention the K value that their KNN algorithm uses, so I chose a K value of 7, that provided a balance between being too sensitive to noise and losing focus on the local patterns. I looped through each observation in the testing dataset, and collected the KNN algorithm prediction and the actual classification value. Finally, I created a simple accuracy function that would compare the two dataframes and calculate the accuracy rate.

The accuracy rate I achieved with this model is 18.67%. This is substantially lower than the accuracy rate that Akujuobi states in the paper, which is 74.84%. Although I did make changes in the process of replicating the KNN algorithm, the main reason for this huge

disparity is that I was unable to replicate Akujuobi's methodology because of a severe lack of documentation.

A significant critique I have about Akujuobi's research is the lack of transparency and detail, particularly in their description of how they designed and implemented their models. To bring this to light, I will focus on the KNN model that I have tried to replicate above. Firstly, Akujuobi fails to mention whether they normalized their numerical feature values or not. Normalization plays a critical role in improving the performance and reliability of a KNN algorithm, as it equalizes the scales of the features, preventing any feature from disproportionately affecting the classification process. Although it is not that necessary if the features are already on the same scale, that is not the case for the features that the research focuses on. Akujuobi fails to mention any detail about normalization at all in their paper, which means that the results that they achieved could be invalid and potentially biased.

Secondly, the process that Akujuobi uses to implement the KNN algorithm is not explained at all, especially how they adapted KNN to work with the mix of numerical and categorical features. KNN is inherently an algorithm that relies heavily on the calculation of distance to achieve its purpose. However, most distance metrics are fundamentally designed for numerical data, not categorical data. The metric that Akujuobi mentions is Euclidean distance, which is not appropriate for categorical data because it lacks meaningful numerical relationships that can be measured with distance. Despite this, Akujuobi only mentions Euclidean distance, and does not mention any other distance metric or method used to connect the 2 categorical features to the KNN algorithm. This lack of clarification makes the model much harder to replicate, decreasing the validity and reproducibility of its results.

Finally, Akujuobi does not mention any details on what K value is used in the algorithm and why it was chosen, as well as how they dealt with potential ties in the classification. The K value of a KNN algorithm determines the number of neighbors considered when making a prediction, and can significantly impact the algorithm's accuracy and sensitivity to noise. Additionally, there is no explanation about the ties in distance that can occur when two neighbors are equidistant from the observation point, such as a method to break

the tie. The paper fails to mention any of these important details, making it much harder to replicate and validate the model's performance.

Overall, Akujuobi fails to uphold the level of transparency and credibility that is expected of a research paper, which raises major concerns about the reliability and reproducibility of their research. Although the direction and intent of the work is commendable and explained, there is a lack of evidence to support the accuracy and results of the model. The absence of key details is a fatal flaw in the reproducibility and reliability of this work, and the model should not be used out on the field.

Analysis of Normative Concerns

I believe that although there may be ethical concerns with using a predictive policing model such as Akujuobi's model in the real world, it is justified by Utilitarianism. Utilitarianism evaluates actions based on the total pleasure and pain they generate, and states that the morally just action is the action that is maximizing that pleasure and minimizing that pain. The use of predictive policing models can come with many benefits, and the issues can also be minimized, so it is the morally right course of action. To support this claim, I will first introduce the concerns with predictive policing models as a whole.

When addressing predictive policing models, it's natural to bring up the potential for ethical concerns. A common normative concern that arises with crime prediction algorithms is the possibility of algorithmic bias. Because these algorithms are based on historical data, any societal or institutional biases from old policing practices could have an effect on the predictions. For example, if certain areas or demographics were disproportionately policed, then models based off of that data could also disproportionately predict for those areas or demographics. Akujuobi's work relies on Chicago's historical crime data from the years 2001 to 2015, so if there were any biases that disproportionately affected a certain racial or socioeconomic group, then Akujuobi's model may perpetuate such biases. Even though none of the features in the model explicitly contain any information about the crime's suspect, it is still possible for the algorithm to exhibit bias. This is known as discrimination by proxy, where an algorithm uses a different, innocent looking feature as a stand-in for a certain feature to reach a decision that should not be based on that feature. An example of this is

the COMPAS algorithm, which is used to assess potential recidivism risk. Although it had no access to race, it was found that Black defendants were often predicted to be at a higher risk of recidivism than they actually were, and were also twice as likely as white defendants to be misclassified as a higher risk of violent recidivism (ProPublica, 2016). In this study, the model uses the location of the crime being committed as one of its prediction features. Many locations in Chicago may be demographically homogeneous, so using location as a predictive feature might cause the model to associate certain demographics to certain crimes, reinforcing stereotypes. If such an algorithm is used in the field, it may cause increased tensions between communities and law enforcement. These are the moral concerns that are against the use of predictive policing.

However, the same potential for biases can be said for humans as well. Humans are also susceptible to the same biases. Although we can make a conscious effort to be as impartial as possible, it is something that affects everyone. According to a research paper published in the National Library of Medicine by Mihal Emberton, "this type of unconscious bias is a human condition, not bounded by culture or place or time"(Emberton 2021). For example, if someone owns an expensive car, we would view them as rich, based on our own experiences that rich people drive expensive cars. Humans are creatures that use their ability to recognize patterns and find rules to those patterns to try and understand their social hierarchies. When a correct rule is found for an observed pattern, it brings joy and also allows them to relate to other people in their society. However, these rules can be miscalculated due to gaps of knowledge, which is what creates unconscious biases. Many of us try our best to recognize our own unconscious biases by embracing the knowledge gap, and genuinely trying to overcome it. Even still, there is no way for us to completely rid ourselves of any biases, because we are inherently pattern recognizing beings and can make mistakes. This also includes law enforcement, as they are made up of humans too. Biased based policing is a real issue that police deal with throughout their careers, and they are working to overcome errors in reasoning such as confirmation bias that can lead to racism and discrimination.

If humans naturally have this unconscious bias and try to overcome it, then why can't we use machine learning models that could have this bias but use different methods to try and

overcome/diminish that bias? From manually auditing the data, to mitigate any problematic patterns, to excluding features that could reinforce stereotypes or act as proxies, there are many different methods that can be used to overcome the biases that a model such as the one in this research paper could have. Although it may not be able to completely overrule that bias, we as humans can't completely rid ourselves of all biases either. If such a model is able to exhibit many benefits for society as a whole, then based on utilitarianism using that model is the morally correct action.

To further justify the use of predictive policing, we should examine the benefits that such use will bring. The net benefits that predictive policing like Akujuobi's model can bring are numerous, and can affect society as a whole. These benefits include optimizing law enforcement resources, improving public safety, and reducing crime. In 2023 alone, there were a total of 14 million criminal offences in the US, up from around 11 million the year before(FBI 2024). With this increase in crime, law enforcement will have less resources to deal with each offense, which could hurt both the police officers who were underprepared, and the victims of the crime. Akujuobi's model could enable law enforcement to efficiently allocate their limited resources, enabling them to focus their strength on high crime areas. The model can also help them avoid unnecessary community-police confrontations by limiting the officers and resources sent to respond to minor crimes, which will help build trust with the communities. By helping law enforcement efficiently allocate their resources, predictive policing can also reduce the chances of officers being harmed on duty as well as better protect the victims of crimes. Better police responses will have a higher chance of resolving the issues at a crime scene, lowering the risks for both officers and the victims trapped at the scene, potentially saving numerous lives.

Furthermore, predictive policing models have a unique advantage because they are inherently data driven and consistent. Humans can potentially be clouded by emotions, making decisions that might not be the best course of action and could actually harm lots of people. However, algorithms do not have this problem, and only take data into account. Additionally, human biases are often unconscious and hard to overcome, while algorithmic biases can be tested for and corrected, which allows for more accurate and fairer systems overtime.

These benefits that come from predictive policing models can make society a better place overall, and severely outweigh the concerns that come with using the models.

Overall, the use of predictive policing models combined with measures to overcome potential biases is the morally correct action based on the principles of Utilitarianism because it maximizes the societal benefits while minimizing the risks. By addressing bias through multiple methods, these tools can be used for the benefit of society. Therefore, predictive policing models should be used because it is a justified and ethically correct action.

Conclusion

The aim of predictive policing is to predict future crime with data on past crimes in order to reduce crime and protect citizens. However, many of the predictive policing models can't be used in real-time, during an active and immediate situation. This is where Akujuobi's research can create a significant impact, as it could help in faster and better crime scene response for police officers in the field. Instead of attempting to predict future crime using past data, Akujuobi's goal is to make real time predictions based on real-time information of a crime, to help police better deal with the situation. This model would entail creating a way for victims to report crimes without putting themselves in danger, such as a clicking a button on a mobile phone and a quick few words describing the situation. Next, the model would then use the information gathered to predict the type of crime being committed, and notify the police department of the possible crimes being committed. With the predicted crime information, police would be better prepared to take on the crime call, enabling them to take on less unnecessary risks. The system could also help victims who are in dangerous situations to call for help without endangering themselves. Although this model is still not perfect, as the biases and issues that arise from using the model could cause potential issues, it is still a great step forward in the direction of a safer and better society. The field of combating crime with algorithmic analysis is still in its developing phase, and this research will help bring to light the many benefits that predictive policing can bring. As technology continues to advance and new models and methods are created, these issues should begin to diminish, and predictive policing will become more widely used.

References:

Akujuobi, Uchenna. Crime Prediction for Better Response Using KNN, PCA and Random Forest.

Larson,Julia, Jeff. "How We Analyzed the COMPAS Recidivism Algorithm." ProPublica, 23 May 2016, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

"Cogent | Blog | Predictive Policing Using Machine Learning (with Examples)." Www.cogentinfo.com, www.cogentinfo.com/resources/predictive-policing-using-machine-learning-with-examples.

Chicago Crime Data City of Chicago. "Crimes - 2001 to Present." City of Chicago Data Portal, https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data.

FBI 2023 Crime Statistics Federal Bureau of Investigation. "FBI Releases 2023 Crime in the Nation Statistics." FBI News, https://www.fbi.gov/news/press-releases/fbi-releases-2023-crime-in-the-nation-statistics.

FBI 2022 Crime Statistics Federal Bureau of Investigation. "FBI Releases 2022 Crime in the Nation Statistics." FBI News, https://www.fbi.gov/news/press-releases/fbi-releases-2022-crime-in-the-nation-statistics.

Unconscious Bias Article Emberton, Mihal. "The Universal Nature of Unconscious Bias." National Library of Medicine, National Center for Biotechnology Information, 19 Jan. 2022, https://pmc.ncbi.nlm.nih.gov/articles/PMC8784036/#s3.