

Final Project

Alyson Longworth

2024-05-3

Introduction

The growing use and occurrence of artificial intelligence in our daily lives constitute significant issues that demand acknowledgement since they may have far-reaching consequences for our society. In particular, the use of artificial intelligence in the criminal justice system is currently on the rise to enhance predictive policing. Although artificial intelligence may improve police effectiveness when used in predictive policing, some drawbacks must be considered. Its inevitable errors and potential biases may sway our decisions and raise concerns about the morality of this approach. Because the improper use of these algorithms can have detrimental impacts on individual lives and communities across the country, we must continue to be aware of, look into, and critique the tactics being used.

Analysis of Methods

To distribute resources as effectively as possible to achieve optimum policing efficiency, statistical modeling techniques are often implemented to anticipate the location and time of future crimes in the criminal justice system. Statistical models are widely practiced in predictive policing to forecast crime trends, including when and where crimes are likely to occur. These models, which look for causal relationships, include Bayesian techniques (Bayes Theorem), Point Process models (Poisson processes) and Support Vector Machines and often use the Euclidean distance metric as a way of interpreting the location of the data.

The Euclidean distance metric is often the most common and simplistic metric used for computing the distance between crimes and locations with highly concentrated crimes. While being relatively easy to apply, the use of this metric has its drawbacks. Although favorable in high-crime regions and periods when it is advantageous to achieve speedy correlations, the use of this metric can lead to biased data and potentially leave a biased foundation when creating predictive algorithms. Due to the fact that this metric fully utilizes an often discriminatory proxy, zip code, this metric often risks grouping individuals based on race or socioeconomic standing.

$$EuclideanDistance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Bayes Theorem can be used to predict the probability of a crime occurring in a specific location during a specific time based on past crime trends in data. By using the locations of previous crimes, the criminal justice system is able to predict where future crimes will occur to designate resources to those areas as a preventative measure.

$$BayesTheorem = P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$$

Point Process models will provide the crime incidence rate at a certain location and time, assuming the crimes follow a Poisson distribution. However, because this model implies a Poisson distribution, which assumes independence among the data, it often fails to acknowledge the usual fluctuations in criminal behavior and crime trends.

Support Vector Machines, however, look for a hyperplane that divides the data points of previous crimes. This approach helps decide when and where to allocate the most resources by grouping each point according to whether or not a crime will happen in a certain place at a certain time. Support vector machines, though, commonly run the danger of overfitting a random population of data points by being overly particular with only a specific set of data.

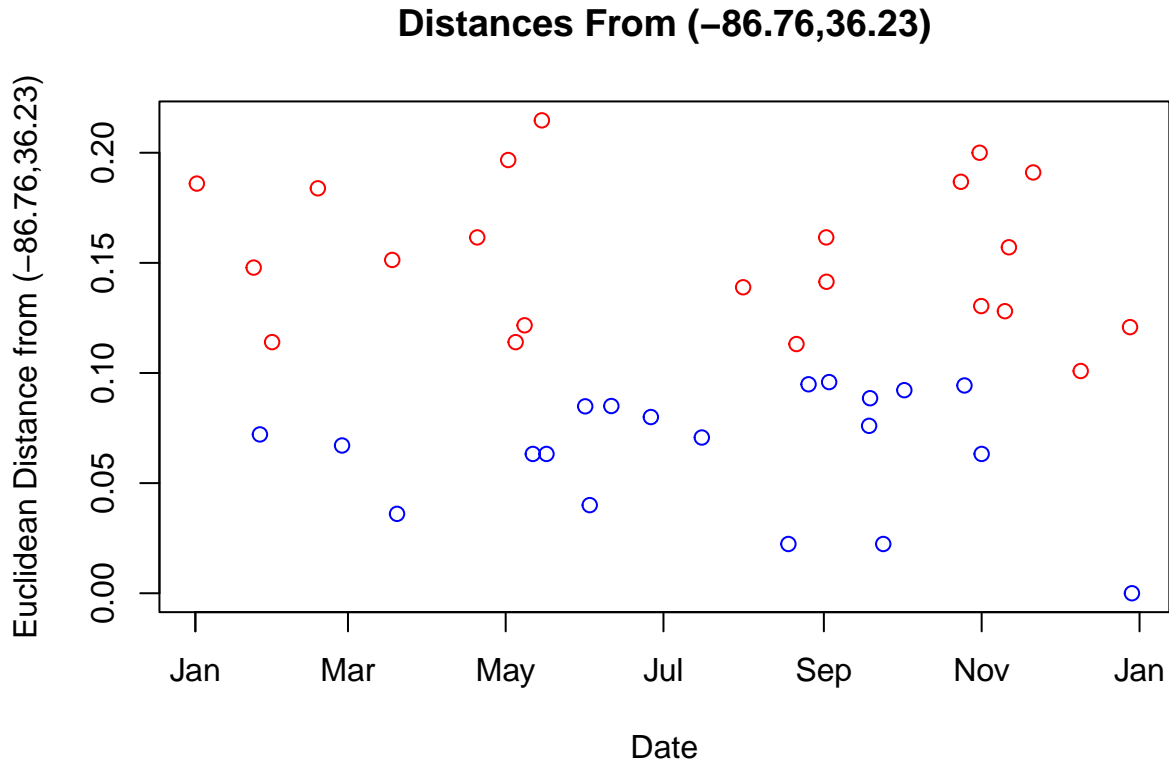
In order to analyze how these methods may be transferred for real data, I have chosen to examine the Crime Data package in R from the Open Crime Database. In particular, I will be focusing on the crimes committed in Nashville, TN during the year of 2022, and specifically, the likelihood of a motor vehicle theft occurring near the location of the recorded motor vehicle theft on December 29, 2022 at the coordinates (-86.76, 36.23). I decided to remove the recorded crime on December 29, 2022 from the dataset and examine the likelihood of this event occurring in the same location by applying Bayes Theorem, Poisson Process, and Support Vector Machines, each using the Euclidean distance metric to compute the distances from each observation to the location of the event recorded on December 29, 2022. By applying this to each of these methods previously discussed, I will be able to determine whether or not these methods provide consistent results and accuracy with one another considering my previous knowledge that this event indeed did occur. First, I subsetting the “crimedata” dataset provided into several smaller datasets: “nashville_crime_data” including the columns *uid*, *city_name*, *offense_type*, *offense_group*, *date_single*, *longitude*, *latitude* and only the crimes committed in Nashville, TN, “nashville_motor_vehicle_theft” including the columns *uid*, *city_name*, *offense_type*, *offense_group*, *date_single*, *longitude*, *latitude* and only the motor vehicle thefts committed in Nashville, and “motor_vehicle_thefts” including the columns *uid*, *city_name*, *offense_type*, *offense_group*, *date_single*, *longitude*, *latitude* and only the motor vehicle thefts. I then found the Euclidean distance from each observation to one another in each of the datasets “nashville_motor_vehicle_theft”, “motor_vehicle_thefts”, and “nashville_crime_data” in the matrices: “euclidean_distance_matrix”, “euclidean_distance_matrix_mvt”, and “euclidean_distance_matrix_nash”. Next, I combined “nashville_crime_data” and “euclidean_distance_matrix_nash” to view all the Nashville crimes alongside the Euclidean distances from each observation. I then combined “nashville_motor_vehicle_theft” and “euclidean_distance_matrix” to view all the Nashville motor vehicle thefts alongside the Euclidean distances from each observation. Below, I have shown a preview of the combined dataset created which shows all the Nashville motor vehicle thefts alongside their euclidean distances from the recorded crime on December 29, 2022.

##		date_single	longitude	latitude	
## 1	2022-01-01 13:45:00	-86.650	36.080	0.18601075	
## 2	2022-01-23 13:00:00	-86.818	36.094	0.14785128	
## 3	2022-01-25 23:00:00	-86.820	36.190	0.07211103	
## 4	2022-01-30 16:00:00	-86.850	36.160	0.11401754	
## 5	2022-02-17 08:30:00	-86.590	36.160	0.18384776	
## 6	2022-02-26 17:00:00	-86.790	36.170	0.06708204	

I examined the question: if a crime occurs 51.92% of the time within a 0.1 Euclidean distance of (-86.76, 36.23) and a motor vehicle theft occurs 5.24% of the time in the year of 2022, what is the probability that a crime occurs within a 0.1 Euclidean distance of (-86.76, 36.23) given that it was a motor vehicle theft? This question proposes one that a criminal justice system may interpret in reality in order to determine how likely it is for a specific location to witness a specific crime and whether or not to send resources beforehand as a precautionary measure. Using Bayes Theorem and the subsetting datasets previously discussed, I discovered that there was a 46.34% chance that a crime will occur within a 0.1 Euclidean distance of (-86.76, 36.23) given that it was a motor vehicle theft. Considering that there were only 41 recorded motor vehicle thefts all year in all of Nashville, and this result recognizes a nearly equally likely chance for this crime to occur in this location, it can be determined that there is a relatively high chance of this crime occurring in this location. Due to our prior knowledge that this crime did indeed occur in real life, one of two things can be noted: either this method provides a relatively accurate way of prediction or that due to the common fluctuation of crimes in general, this method provides a result that may raise questions of biases and inaccurately sway our decisions for predictive policing.

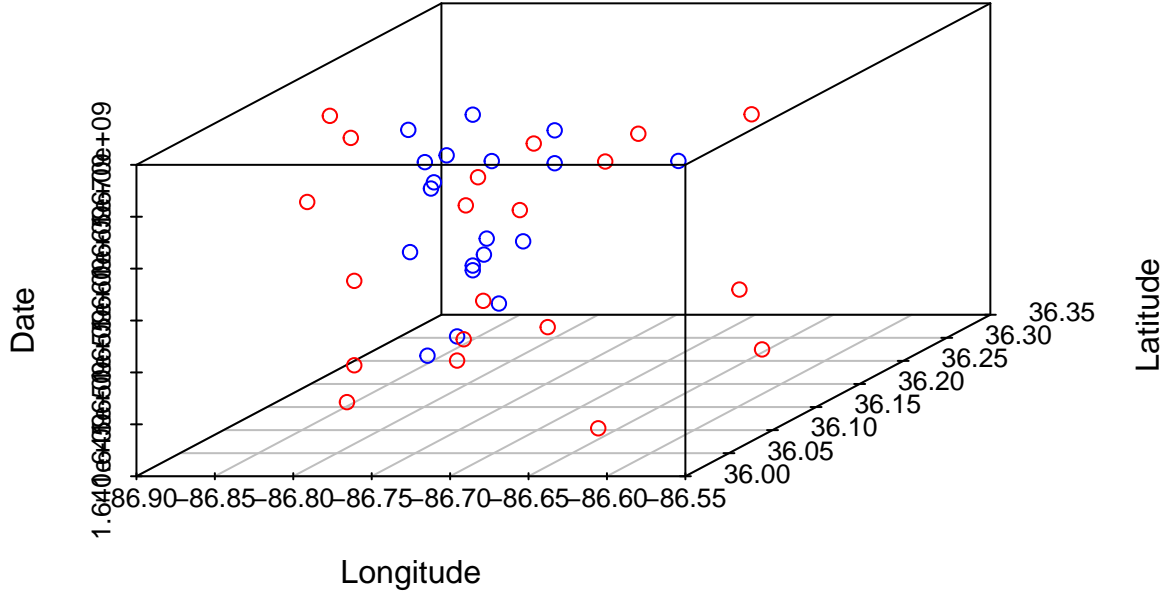
$$P(A|B) = \frac{0.487 \cdot 0.4987}{0.0524} = 0.4634$$

In light of the Poisson Process, I examined the question; if we can see that there are, on average, 1.64 occurrences of a motor vehicle theft a month from January to November and 1.58 occurrences of a motor vehicle theft a month from January to December within a 0.1 Euclidean distance of (-86.76, 36.23), what is the probability that there will be another occurrence in December within a 0.1 Euclidean distance of (-86.76, 36.23)? I used the ppois function in R, first with a lambda value of 1.64 and second with a lambda value of 1.58 to determine the probability that there will be an occurrence within a 0.1 Euclidean distance of (-86.76, 36.23) in December, first only using the data from January to November and second using data from January to December. By using the Poisson distribution and the lambda value of 1.64, it was discovered that there was a 48.78% chance that a motor vehicle theft will occur within a 0.1 Euclidean distance of (-86.76, 36.23) in December. By using the Poisson distribution and the lambda value of 1.58, it was discovered that there was a 46.86% chance that a motor vehicle theft will occur within a 0.1 Euclidean distance of (-86.76, 36.23). Comparing the probability recorded using Bayes Theorem, it can be noted that these methods predicted the likelihood of a motor vehicle theft occurring in this area similarly. However, it should also be noted that by including the month of December in the calculation of the lambda value, the probability dropped by nearly 2%. This portrays the concern previously discussed regarding the frequent fluctuation in crime data and the assumed Poisson distribution employed in this method. It can be noted that this average may easily change month-to-month, which this distribution fails to acknowledge. Although, in this event, the result may have provided a relatively accurate prediction, the use of this method risks inconsistency due to the potential fluctuations that are inevitable to occur in real life. Below displays a scatterplot of the Euclidean distances each event was from (-86.76, 36.23) during the date in which they occurred. Each point outlined in red portrays an observation that had a Euclidean distance of greater than 0.1 from the event recorded on December 29, 2022 and each point outlined in blue portrays an observation that had a Euclidean distance of 0.1 or less from the event recorded on December 29, 2022. As seen from the graph, the Euclidean distance from the event recorded on December 29, 2022 and the date of which the crime occurred is not evenly distributed. For instance, there were no crimes with a Euclidean distance of greater than 0.1 during the entire month of July, but there was a majority of events further than a Euclidean distance of 0.1 during the month of December.



In regard to the method, Support Vector Machines, I began by plotting a three dimensional graph of the observations recorded of the motor vehicle thefts in Nashville during the year of 2022. These observations were plotted with an x-axis of *longitude*, y-axis of *latitude*, and z-axis of *month*. Each point outlined in red portrays an observation that has a Euclidean distance of greater than 0.1 from the event recorded on December 29, 2022 and each point outlined in blue portrays an observation that has a Euclidean distance of 0.1 or less from the event recorded on December 29, 2022. By identifying a hyperplane that best separates the red points from the blue points using Support Vector Machines, we may be able to distinguish the locations during each month in which are most likely to experience a motor vehicle theft. With this knowledge, the criminal justice system has the potential to predict whether a certain location is likely to experience a certain crime during a certain month. However, because crimes typically fluctuate often, introducing more data or data from a different year can tremendously change the placements of each point on this graph and the hyperplane used to separate them. With this being said, this method holds the risk of overfitting to this specific dataset and may not be consistent as minor fluctuations are inevitable to occur.

Distances From $(-86.76, 36.23)$



Analysis of Normative Consideration

Although each method currently displayed relatively accurate predictions of the likelihood of a motor vehicle theft occurring on December 29, 2022 within a 0.1 Euclidean distance of $(-86.76, 36.23)$, they nonetheless have potential to fail to provide consistent and reciprocal results in the future. Because of the expectation of commonly high fluctuations in crime data, it is difficult to discover a method that will account for these fluctuations. For instance, from using Bayes Theorem to predict the likelihood of this event occurring, it was found that although this method did provide relatively reliable results, by relying solely on the locations of previous crimes, it is possible that this model may skew toward areas with previously high crimes. From examining the Poisson Process in a similar way, although this method additionally provided a relatively useful prediction, it was proven that the results of this model may easily fluctuate with changes in the data. Finally, by examining the use of Support Vector Machines in this sense, it can also be noted that because crime data fluctuates often, it can be extremely difficult to create a hyperplane that will account for these fluctuations without overfitting to a limited dataset. Clearly, each of these methods lack consistency and reciprocity and therefore, not only make it difficult to remain reliable in the future but also to hold the criminal justice system accountable for the decisions they make based on these methods. Furthermore, the application of the Euclidean distance metric to compute the distance between crimes used in each of these methods may assist in the foundation of useful, and potentially accurate results at the moment. However, the reliance on this metric has the potential to lead to the over-policing of certain areas and may exponentialize the crimes being recorded in these areas in the future. Ultimately, this risks the slippery slope of promoting systematic discrimination toward particular groups that live in these areas.

Overall, although these methods may currently be favorable in regard to accuracy provided in this study, in order to be effective in the future, they must be repeatedly updated and reviewed. Even so, the continuous use of the results of these methods have the potential to skew future data, compromising the true accuracy

these methods hold. Due to the high potential for biases toward certain areas and groups in the future, especially if these methods are not constantly checked for reliability, the use of these methods in predictive policing is immoral. Not only do these methods often lack consistency and reciprocity but also continuous reliance on these methods risk the over-policing of certain areas and a positive feedback loop in this manner. Not only does this make it difficult to remain reliable in the future but also to hold the criminal justice system accountable for the decisions they make based on these methods. From a deontological perspective and according to the second categorical imperative, treating humans as a mere means to an end is unethical. By categorizing areas depending on data in previous years rather than the current people residing in these areas and the current situations these people are in, the criminal justice system will dehumanize the situations of the groups living in these areas to the situations of the pieces of data in the past. If we are unable to prove the true accuracy of the results found by these methods that lack consistency and reciprocity to begin with, the criminal justice system risks exchanging accuracy for false efficiency. Consequently, the people living in these areas will have less opportunity to escape the previously “labeled” area they reside in and will continue to experience the positive feedback loop of over-policing. This risks prioritizing false efficiency over morality at the expense of the people of these areas’ lives and future, which proves the immorality of the use of these methods in predictive policing.

Conclusion

With that being said, biases in training data and model assumptions can lead to inaccurate predictions, which can feed into the creation of more biases. The main normative concern regarding predictive policing, therefore, is the likelihood that these inaccuracies and biases may lead to the over-policing of certain areas and the systematic discrimination against the people of these areas. These methods used in predictive policing must be addressed because of the potentially large impact they may have on the people of our society.

In appreciation to review papers, like the one cited for this project, “Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement” by Richard A. Berk, we must continue to be conscious of and investigate the methods being used to prevent the immoral use of these algorithms in our everyday lives. By understanding the ways in which these methods are used in situations like predictive policing, we can begin improving the morality of our tactics and the equality of our society in not only the criminal justice system, but everywhere.

References

Berk, Richard A. “Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement.” *Annual Review of Criminology*, vol. 4, no. 1, Jan. 2021, pp. 209–237, <https://doi.org/https://doi.org/10.1146/annurev-criminol-051520-012342> <https://doi.org/10.1146/annurev-criminol-051520-012342>