Chicago Crime Prediction using Machine Learning - proposal

Katie Dums
kdums@wisc.edu

Adam Shedivy
ajshedivy@wisc.edu

Shane Spellman
sspellman3@wisc.edu

1. Introduction

Crime and violence have a direct affect on the safety, economy, the quality of life, and well-being of community residents. Accurate crime prediction and future forecasting tends can assist to enhance metropolitan safety by using computational techniques and technologies. Crime analysis in this way can help aid police organizations by processing complex information from big data in order to increase early and accurate predictions and patterns in criminal activity [9].

1.1. Project Outline

For our project, we aim to explore different machine learning techniques within the context of Crime Analysis and Predictive Forecasting. By gathering daily crime report data from the Chicago Area (2001-2021)[1], we can analyze trends and patterns overtime, detect hot-spot locations within the city, and make predictive decisions on arrest and crime type.

We plan to break up the semester into four phases: Exploratory Data Analysis/prepossessing, model building, model testing/comparison, model evaluation/fine tuning.

1.2. Related Work

Recent literature regarding crime prediction is divided into different domains. There are several studies that highlight ecological factors like education, income level, and unemployment rates [6]. There has also been studies conducted using social networks such as Twitter, Facebook, and mobile phone data [4][1].

Here are a couple former studies done using Chicago Crime data:

- Kang et al. used environmental context information to improve the prediction of models by proposing a feature-level data fusion method on deep neural networks. This study used four demographic datasets (City of Chicago Data Portal[2], American FactFinder, Weather Underground, and Google Street View) for the year 2014 and showed improved results in precision and recall [5].

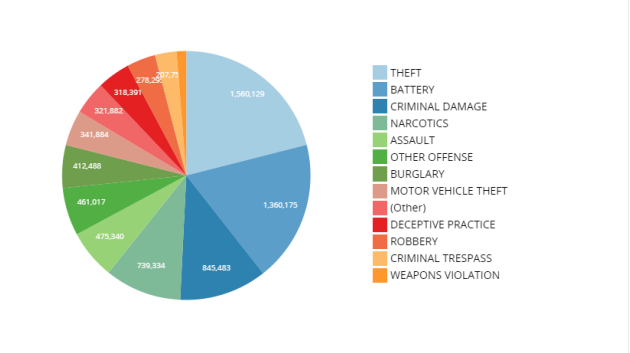

Figure 1. Pie Chart illustrating the proportion of crime types in Chicago from 2001-2021

- Christian et al. related socioeconomic and sustainable development indicators like poverty rate and unemployment toward crime by implementing Linear Regression Analysis from the year 2008 to 2012 for Chicago [2].

- There have been multiple studies that were performed by using geographical locations, meta-association rules and specific detection systems to examine the crime rate in Chicago [8].

We hope to use the recent work with crime analytics in Chicago as inspiration for developing our predictive models by focusing on using machine learning specific techniques within the scope of this course. Three major areas we plan to explore are: logistic regression, k-nearest neighbors (KNN), decision trees, and Bayes classifiers.

2. Motivation

Throughout 2020, two of the arguably largest and most talked-about events included the Covid-19 pandemic and the death of George Floyd and the subsequent Black Lives Matter protests. These events have created the perception that crime in general has risen and in particular violent

[1] For this project, we want to focus on predictive analytics using primarily Daily crime report data. During our exploratory data analysis phase, it is our goal to attempt to incorporate other data sources such as income levels, and education data.

[2] this is the same data source we are using

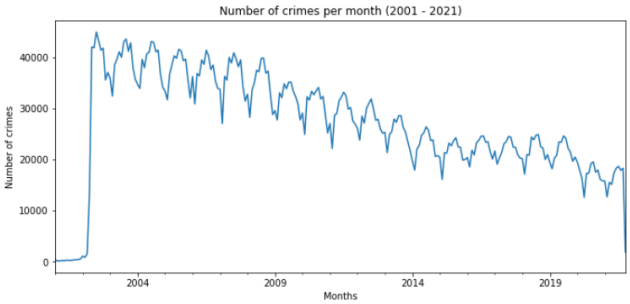

Figure 2. Time Series plot from exploratory data analysis. One initial trend we can see is the 'sine wave' like curve that increases at the beginning of the year and decreases towards the end of the year

crime. Large cities such as Chicago have become targets of criticism and referred to as a crime capital and being "worse than Afghanistan," as stated by former President Trump. As such, the perception of crime has fostered civilian fear, highlighting the importance of effective communication regarding ongoing crimes in one's neighborhood.

There are many different factors that play a role in crime rates. Various circumstances of human nature and behavior, unemployment, gender inequality, high population density, child labor, and education systems can cause an increase in violent crime. A socially sustainable community heavily relies on minimizing crime so that people can live peacefully and actively. Thus, analyzing crime report data are essential to improve the safety while maintaining sustainable development.

We hope to build a framework that can be applied to other major cities, beyond Chicago, with the goal of predicting crime for the use within a police force and also finding an effective way to quickly share the category of crime, it's location, and time with Chicago's inhabitants[7].

3. Evaluation

As mentioned in the motivation, we aim to create a model that performs well in predicting crime rates in Chicago given the quantity of crime in prior time periods. Our time-series model will be a success if it can accurately predict crime rates while considering both a possible seasonal and large-scale trend. The model will be somewhat unique in that we cannot use randomized k-fold cross validation because maintaining the temporal order of time-series data is essential[3]. We will split the data with respect to different periods of time, and use a seasonal naive model:

$$Y_{1(T+h|T)} = Y_{T+h-m(k+1)}$$

Where Y_1 is current period, h is the forecast horizon, m is seasonality, and k is the integer part of $h - 1/m$. We also look to see if the accuracy of our model can be improved from implementing exponential smoothing. We will compute the accuracy of our model predicting the quantity of all crime, as well as the quality of specific criminal acts.

4. Resources

We will be utilizing a dataset published by Chicago Data Portal titled "Crimes - 2001 to Present." The dataset includes variables such as date, primary type of crime, latitude and longitude, block, and time amongst others. We will primarily be using Python as the primary language used to code, including numpy, pandas, and scikit-learn. We also hope to find a way to update and share the information that is easily read and accessible by the public[1].

5. Contributions

We plan to split in this project evenly by assigning each member to specific tasks. Focusing on the computational aspect of the project, Adam will be responsible for handling/manipulating the data and preparing it for analysis. Shane will be responsible for creating models. Katie will be responsible for evaluating and improving model performance. In terms of the writing aspect, Adam will handle the introduction and other related information about the project. Shane will write about the methods and models created during the project. And Katie will be responsible for interpreting the models' results, as well as the discussion and conclusion. We anticipate meeting regularly and ensuring each member of the group participates evenly.

References

[1] Crimes - 2001 to present, city of chicago, data portal. *Chicago*.

[2] S. N. Christian, K. R. Majeed, and S. O. Etinosa. Application of data analytics techniques in analyzing crimes, 2018.

[3] B. Crocker. How to predict a time series part 1, Nov 2019.

[4] M. S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[5] H.-W. Kang and H.-B. Kang. Prediction of crime occurrence from multi-modal data using deep learning. *Plos One*, 12(4), 2017.

[6] L. Lochner. Education and crime. *The Economics of Education*, page 109–117, 2020.

[7] T. Monkovic and J. Asher. Why people misperceive crime trends (chicago is not the murder capital). *The New York Times*, Jun 2021.

[8] G. Rosser and T. Cheng. Improving the robustness and accuracy of crime prediction with the self-exciting point process through isotropic triggering. *Applied Spatial Analysis and Policy*, 12(1):5–25, 2016.

[9] W. Safat, S. Asghar, and S. A. Gillani. Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. *IEEE Access*, 9:70080–70094, 2021.