Crime Rate Analytics

Eunice Yoon

Project URL: https://github.com/yooneuni/CSE482\_Final\_Project.git

**ABSTRACT**

The goal is to develop a framework for crime prediction using diverse types of data available, for example historical data, weather data, traffic data, and demographic data. Specifically, given a location, time of day and day of week and auxiliary data about the location, predict the likelihood that a crime will occur. If the prediction model I make is accurate then I will be able to use the model not only for the cities I have collected data for but for all cities. I am looking at the city of Chicago and estimating the probability of battery in the area. I will be looking at if the arrest status and domestic status of each data point

# INTRODUCTION

For introduction, you need to include the following information:

1. Crime is a problem that is happening everywhere in the world. While it is impossible to stop it this project was to help predict the crime rate of Chicago. In particular information about battery. The definition of batter is the act of beating someone or something with successive blows. Battery is a serious crime that can even end with a person being arrested and charged. The idea of domestic battery is battery against a spouse, intimate partner, or family member.
2. The goal of this project was to find the probability of a battery crime in Chicago to be declared as domestic and if the person committing the crime was arrested. The plan of this project is to use multiple regression models against the data to and get the accuracy level of all those regression models. Doing this will provide a more stable accuracy level.
3. For this project data has been collected from an online website that provided information about all crime in Chicago from 2001 to the present. The data that has been filtered was for all battery data in the year 2018. While this did prove to be quite a small amount of data it was much easier to process than the original tens of thousands of data points.
4. While working on this project, there were quite a few obstacles that occurred, such as what regression parameters would be sufficient/efficient enough for this project. To overcome this challenge I tried multiple different parameters and tried to look at the one that was the best for what I was looking for.
5. Based off the results of the project I was shocked to see that half of the battery cases in the year 2018 were ruled to be domestic and that 80% of battery cases observed resulted in the offender being arrested and that there were no real correlation between if the battery was domestic and if the offender was arrested.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*CSE881-2015*, Month 1–2, 2004, City, State, Country.

Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

# DATA

In order to find the best data, I originally went to google and found a website that held multiple crime data for different cities. After finding the dataset for Chicago crime, it was seen to contain all crime information in Chicago from the year 2001 to the present which resulted in an original dataset with tens of thousands of rows of data in a downloadable CSV file. Since the file was so large it was necessary to find a way to shorten this data not only for time purposes but also for accuracy since crime from just 5 years ago is different than crime today. Before preprocessing the data some of the attributes I looked at were date, case number, primary type, description, location, location, and year.

Since this is special data I could not just look at data from all over the world so I instead focused on one area. My original plan for this project was to compare three different cities however that proved to be far too time consuming so I put more focus on the city of Chicago, especially because Chicago is a city where a lot of crime still happens there was still several data points to look at. The size of the raw data had tens of thousands of points especially because it held data from almost 20 years ago. To help find the most accurate prediction I first got rid of half of the data. When I saw that there was still some inaccuracy, I then filtered the data in half again. While this did make the accuracy slightly better I decided to keep it safe and look at one year’s worth of data.

Even though I was only looking at a year’s worth of data, there was still the matter of looking into the different attributes of each data point. For example, did I want to do one crime type or multiple crime types? Did I want to consider location as a factor of these events? Since using different types of crimes might alter the accuracy of the prediction I stuck with just using battery as the main focus of my project.

In order to not have to worry about the unnecessary data I preprocessed the data for points I only care about. For example I was not too interested in the FBI code for the information and any rows that had blank information were discarded. There were about 8 columns that I dropped from the raw data I collected in the beginning. There were also various crime rates to look at so I filtered the data to only look at crimes related to battery. When the preprocessing steps were over it was actually necessary to remove any columns that contained any strings such as the date and location of the event as that would often cause a lot of problems when I was analyzing the data.

When comparing all my datapoints one of the hardest parts was finding the best way to look at all these different datapoints. Since there are so many ways to analyze this data determining the most accurate comparisons was important for this project.

I used more regression for this problem. There were several different neighborhoods involved. For simplicity purposes I only looked at one type of crime but there were originally multiple different crimes to choose from. I looked at the information regarding arrest information and domestic information. My dataframe ended up having multiple different columns such as date, location, description, etc. In order to see the level of accuracy for these two different attributes I wanted to use more than one kind of regression model to keep in to consideration of outliers and other issues that may have come up.

My original raw data contained 1.49 GB of data. After my first run of preprocessing the data contained was 157 KB of data and the final run of preprocessing contained 7 KB of data. It is easy to see how important preprocessing was for this project.

# METHODOLOGY

I used a decision tree, k nearest neighbors and regression models in order to find the accuracy of this data. To decide the clusters I used different neighbor numbers.

To analyze the data I ended up using the resources from Exercise 8 and calculated the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for each of the attributes. The findings of each were still similar to each other despite the significant difference in regressional data the differences between these two errors were not too different from each other.

Finally, give a brief summary of the code you have written for this project.

* Collection.ipynb: this is the Jupyter notebook file that I wrote to collect crime data in Chicago
* Preprocess.ipynb: this is the Jupyter notebook file to look for the crime battery in Chicago while also looking at if the offender was arrested and if the crime was domestic.
* Modeling.ipynb: this is the Jupyter notebook file to perform the classification task of the project.

# EXPERIMENTAL EVALUATION

This section describes the experimental setup and results you obtain.

## Experimental Setup

This section should include:

1. I primarily used JupyterLab to do preprocessing and processing of all my data
2. For baseline models I used a hand calculation of how many times a column returned true out of the total number or rows. I used several different regression models in order to check different accuracies.
3. I used accuracy as my evaluation method.

## Experimental Results

Based off of the experiment run, it was found that there will be a 50% probability that the battery crime was domestic and about an 80% change that the offender would be arrested for their crime. The different regression models were only a tenth of a percent off for their corresponding characteristics showing that there is consistency and accuracy between the models. These results show that there is a slight correlation between whether the offender would be arrested for domestic battery or not. Generally it was seen that if the battery was domestic the user was not arrested for the crime showing that there was no real correlation between these two different attributes.

I was somewhat surprised by these results because I expected that if the battery was domestic then the offender would be arrested. However I was not too surprised to see that the offender was arrested 80% of the time. I do believe that my project was a success because I was able to see the probability of a battery crime to result in an arrest and whether it was domestic or not. Since my original data showed that there was a 6.8% probability of the offender being accused of battery and being arrested, I believe that my results had shown a better correlation between the two data points.

When I compared the data I found to when I calculated the probability by hand there was a large difference between the times an offender was arrested for battery to what was found in my project. Based off of just the difference of true and false, the probability of an offender being arrested was 18%. Compared to my finding of 80.5% there is a much better prediction. Again when I compare my results to my baseline data of domestic battery to my calculated predictions. When calculated by hand there was a 54% probability that the battery in question was domestic. However after my prediction model the probability became 52%. This had a 2% less accuracy however it was still similar to what was being found in my baseline prediction.

Based off of my results it was clear that my prediction model was accurate and could be used to predict whether a battery would be domestic or if the battery offender would be arrested.

# CONCLUSIONS

In conclusion, it was found that an offender would have a 61% probability of an offender accused of battery to be domestic battery and an 80.5% probability of the offender being arrested for the battery crime.

# REFERENCES (at least 3 references)

1. “Crimes - 2001 to Present | City of Chicago | Data Portal.” *Chicago Data Portal*, data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.
2. Tan, Pang-Ning Lecture 11 (Classification) Big Data Analysis Lecture.
3. Tan, Pang-Ning Lecture 12 (Market Basket Analysis) Big Data Analysis Lecture

Grading criteria

Note that the project accounts for 10% of your final grade. The project will be graded based on the following criteria:

1. Presentation - structure/organization and clarity of writing (including tables and figures).
2. Technical - Correctness and thoroughness of the analysis performed. What are the challenges faced and how well did you address them? How do you evaluate the performance of the method you'd applied to the data? How much detailed discussion you provide to explain the results you'd obtained (e.g., discussion about why the method works or didn't work on the data)?
3. Difficulty level - How large is the dataset used? How much effort you had to spend to collect, integrate, preprocess, and analyze the data? Are you implementing the project on a cluster or a single machine? What tools did you use (do you have to implement them or are you simply using existing libraries)?