University of Waterloo

Faculty of Engineering  
Department of Management Sciences

Status Update on Exploring and Predicting Violent Crime in Chicago

University of Waterloo

200 University Ave W, Waterloo, ON N2L 3G1

Waterloo, Ontario, Canada

Prepared by  
  
Yingzi Zhang  
20515934  
4A Mechatronics Engineering

And

Xiang Li

20574900

4A Mechatronics Engineering

13 November 2018

1. Summary

* Scope of project reduced so instead of random unspecified blocks of places, we just do by community areas
* No dimension of time considered anymore (makes analysis harder)
* 12 predictors: average school rating of community area, average SSL rating of community area, total park area each community area has, number of hospitals that area has, teenage mother birth rate, infant mortality rate, proportion of Hispanic, black, white, Asian, other races, and percent of children in poverty
* Preprocessing took a lot of time to do (a lot of data was geographic data and we had to use polygon algorithms and do intersections of polygons to get e.g. total park area for each community area)
* EDA: some x’s were definitely related to the class variable (e.g. race, teen mom rate, etc.)
* Prediction:
* Clustering:
* Association:

1. Data Collection Progress

Table 1. Information on Final Chosen Datasets.

|  |  |  |
| --- | --- | --- |
| Dataset | Number of Rows | What Does a Row Represent? |
| Crimes from 2001 | 6,706,459 | Reported Crime |
| Strategic Subject List | 398,684 | Person Likely to be Involved in a Shooting |
| Chicago Public Schools - School Profile Information SY1718 | 661 | School |
| Population and Poverty Data by Chicago Community Area | 77 | Community Area |
| Parks - Chicago Park District Park Boundaries (current) | N/A (Shapely File of 597 Parks) | N/A (Shapely File of 597 Parks) |
| Boundaries - Community Areas (current) | N/A (Shapely File of 77 Community Areas) | N/A (Shapely File of 77 Community Areas) |
| Hospitals - Chicago | N/A (Shapely File of 42 Hospitals) | N/A (Shapely File of 42 Hospitals) |
| Public Health Statistics - Births to mothers aged 15-19 years old in Chicago, by year, 1999-2009 | 77 | Community Area |
| Public Health Statistics- Infant mortality in Chicago, 2005– 2009 | 77 | Community Area |

1. Explanatory Data Analysis Progress
2. Supervised and Unsupervised Learning Progress

Numeric Regression Analysis

Clustering Analysis

Association Rule Mining Analysis

Table 2. Explanatory Variables for Regression, Clustering, and Association Rule Mining

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Dataset that Variable is From | Why Variable is Chosen as Possible Predictor | Does the Variable have to be Further Transformed for Appropriate Use as a Predictor? |
| Location Description (example values are e.g. “restaurant”, “alley”, etc.) | Crimes from 2001 | Violent crimes may be more likely to be committed outside e.g. “alley” versus “restaurant” | No |
| Police District (which police district the crime occurred in) | Crimes from 2001 | Different police districts may be more effective in reducing violent crime | No |
| Ward (which ward the crime occurred in) | Crimes from 2001 | Different wards may have more violent crime than others | No |
| Community Area (which community area the crime occurred in) | Crimes from 2001 | Different community areas may have more violent crime than others | No |
| Date of Crime Occurrence | Crimes from 2001 | Temporal data can be important, as mentioned in *Relevant Studies* section | Maybe (transforming the combined date into month, year, day, week of year, holiday season, etc.) |
| Location of Crime Occurrence (latitude/longitude location of where crime occurred) | Crimes from 2001 | Geographic data can be important, as mentioned in numerous studies in the *Relevant Studies* section | No |
| Strategic Subject List Score | Strategic Subject List | If a subject with a very high probability of being involved in a shooting was arrested near a certain area, then perhaps that area may have more violent crime. | Yes (will have to be connected with the subject’s latest arrest location) |
| Which School Boundary Crime Occurred in | Chicago Public School Attendance Boundaries | Proximity to certain schools may be important, as mentioned in the *Relevant Studies* section | Yes (will have to programmatically determine in which school boundary the radius of area in question is in) |
| Nearby School’s Overall Rating | School Profile Information | Quality of education may be important, as mentioned in the *Relevant Studies* section | Yes (will have to connect this rating with the above variable in this table) |
| Amount/Size/Proximity of Nearby Parks | Chicago Park Boundaries | Amount of tree canopy area and parks may be important, as mentioned in the *Relevant Studies* section | Yes (will have to think of a way to create some score(s) that combines some or all of: the number of nearby parks, the total land area of all the nearby parks, and the proximity distance to the nearest park) |
| Most Frequent Age Range of Population in Community Area | Population and Poverty Data by Chicago Community Area | “Aging out of crime” is a prominent theory in literature [15]. | No |
| Most Frequent Ethnicity/Race of Population in Community Area | Population and Poverty Data by Chicago Community Area | Demographics may be a factor | No |
| Percent of People Below Poverty Line in Community Area | Population and Poverty Data by Chicago Community Area | Socio-economics may be a factor | No |

**Appendix**