# Predict Crime Type In Chicago and Seattle

Yangyang Dai
Computational Social Science, University of Chicago


## Introduction

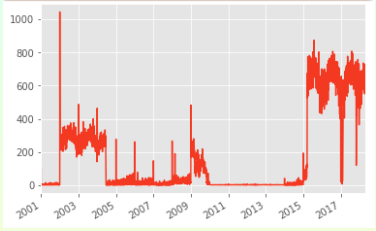
- This project focuses on predicting crime types using different machine learning methods.
- It analyses two different real-world crimes datasets for Chicago, IL and Seattle, WA and provides a comparison between the two datasets through a statistical analysis supported by several graphs.
- The project shows how to use Random Forest, XGboost methods and Neural Network Multiclass Classifier to predict potential crime types in both cities.
- To further analyze crimes' datasets, the project also uses GIS tools such as Geoda to explore the spatial and temporal aspects of the data.

## Data

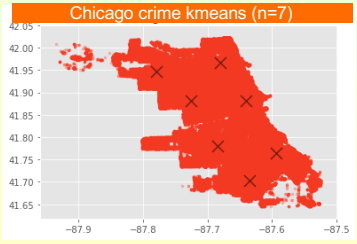
- All the categorical variables are converted to integer codes for the ease of building models. Year, Month, Day of week and Time during the day are extracted from the original time series date data. The day is put into eight time intervals: T1: 12am-3am, T2: 3am-6am, T3: 6am-9am, T4: 9am-12pm, T5:12pm-3pm, T6: 3pm-6pm, T7: 6pm-9pm, T8: 9pm-12am.
- The original Chicago crime dataset has 22 variables. The final features for the prediction are: beat_num, District, Ward, Location Description, Community Area, Year, Month, Day, time, density where the beat_num and density variables are scaled to 0 ~1 intervals.
- The original Seattle crime dataset has 18 variables. The final features for the prediction are: District/Sector, Zone/Beat, Year, Month, Day, Time.

## Chicago

Chicago crime time series from 2001-2018

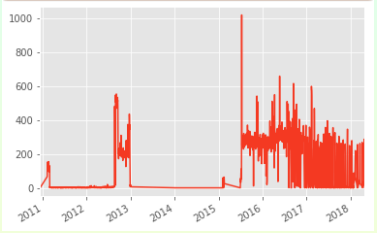


Chicago crime kmeans (n=7)

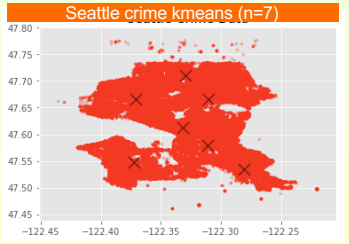


## Seattle

Seattle crime time series from 2011-2018



Seattle crime kmeans (n=7)



## Results

- Chicago vs. Seattle

Random Forest Classification Report

| | Chicago | | | | Seattle | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | f1-score | Support | Precision | Recall | f1-score | Support |
| 0 | 0.4 | 0.39 | 0.4 | 26640 | 0.34 | 0.34 | 0.34 | 5771 |
| 1 | 0.39 | 0.26 | 0.31 | 27150 | 0.4 | 0.44 | 0.42 | 8405 |
| 2 | 0.47 | 0.8 | 0.59 | 48312 | 0.35 | 0.34 | 0.35 | 6931 |
| 3 | 0 | 0 | 0 | 24402 | 0.11 | 0.06 | 0.08 | 1620 |
| avg/total | 0.35 | 0.44 | 0.38 | 126504 | 0.35 | 0.36 | 0.35 | 22727 |

Gradient Boosted Trees Classification Report

| | Chicago | | | | Seattle | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | f1-score | Support | Precision | Recall | f1-score | Support |
| 0 | 0.33 | 0.36 | 0.34 | 26640 | 0.43 | 0.29 | 0.35 | 5771 |
| 1 | 0.34 | 0.33 | 0.34 | 27150 | 0.41 | 0.72 | 0.52 | 8405 |
| 2 | 0.5 | 0.55 | 0.52 | 48312 | 0.41 | 0.25 | 0.31 | 6931 |
| 3 | 0.25 | 0.18 | 0.21 | 24402 | 0 | 0 | 0 | 1620 |
| avg/total | 0.38 | 0.39 | 0.38 | 126504 | 0.39 | 0.41 | 0.38 | 22727 |

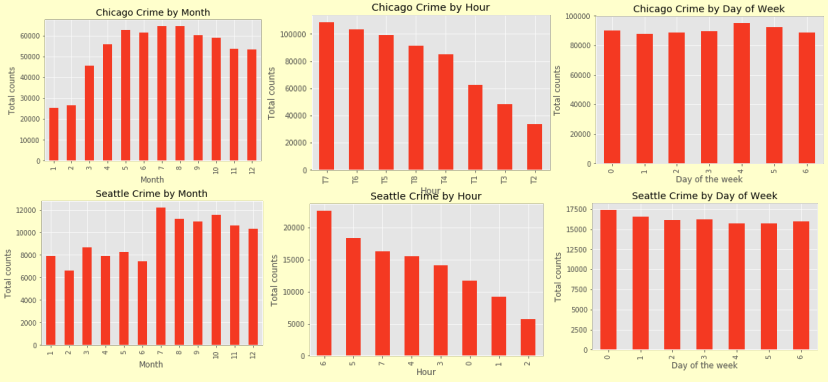
Neural Network Classification Report

| | Chicago | | | | Seattle | | | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | f1-score | Support | Precision | Recall | f1-score | Support |
| 0 | 0.35 | 0.48 | 0.4 | 26640 | 0.32 | 0.29 | 0.31 | 5771 |
| 1 | 0.36 | 0.14 | 0.2 | 27150 | 0 | 0 | 0 | 8405 |
| 2 | 0.47 | 0.77 | 0.58 | 48312 | 0.32 | 0.82 | 0.47 | 6931 |
| 3 | 0 | 0 | 0 | 24402 | 0 | 0 | 0 | 1620 |
| avg/total | 0.33 | 0.42 | 0.35 | 126504 | 0.18 | 0.32 | 0.22 | 22727 |

| | Chicago | | Seattle | |
|---|---|---|---|---|
| | Train Accuracy | Test Accuracy | Train Accuracy | Test Accuracy |
| RF | 93.69% | 39.09% | 86.29% | 35.92% |
| Xgboost | 44.45% | 44.33% | 41.23% | 41.48 |
| NN | 42.18% | 42.25% | 32.59% | 32.45% |

- The time series data shows higher recorded crime counts and their fluctuated nature from 2015 to 2017.
- The spatial distribution using kmeans (n=7) cluster of both datasets are shown on the left.
- Chicago crime types: VIOLENT CRIME, THEFT, CRIME INVOLVING WEAPONS, OTHER CRIMES
- Seattle crime types: PROPERTY CRIME,TRAFFIC RELATED INCIDENTS,VIOLENT CRIME,OTHER INCIDENTS

## Exploratory Data Analysis




- Chicago
  - Throughout the year, month July and August have relatively more crime than other months while January and February have lowest crime.
  - From the time slot graph, crimes occur most often within T7, which indicates 6pm-9pm tend to more have highest crime numbers.
  - During the week, Thursday has the highest crime counts in Chicago.
- Seattle
  - Month July and October have relatively more crime than other months while June and February have lowest crime.
  - Crime occurs most often within T7 as well, which indicates 6pm-9pm tend to more have highest crime numbers.
  - Sunday has the highest crime counts in Seattle.

## Methods

- Data Preprocessing, cleaning, reduction, integration, transformation and discretization.
- Cross validation and Split train and test data sets for two cities.
- Build models:
  - Random Forest
    Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.
  - Gradient Boosted Trees
    The Gradient boosted trees have to be built in series so that a step of gradient descent can be taken in order to minimize a loss function. The speed is a much faster grid search for optimizing hyper-parameters in model tuning.
  - Neural Network Multiclass Classification
    In the output layer, there will be N binary neurons leading to multi-class classification. The last layer of a neural network is usually a softmax function layer, which is the algebraic simplification of N logistic classifiers, normalized per class by the sum of the N-1 other logistic classifiers.
- Test the models using accuracy scores and classification report.

## Future Work

- Spatial heat map and crime rates across the cities can be further explored.
- Other socioeconomics and demographic factors can also be incorporated into machine learning models.

## References


- Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.
- Seattle police department. (https://data.seattle.gov)
- Almanie, T., Mirza, R., & Lor, E. (2015). Crime prediction based on crime types and using spatial and temporal criminal hotspots. arXiv preprint arXiv:1508.02050.