Urban Crime Rate Prediction

Shiyi Zhao

Project URL: https://github.com/userID

**ABSTRACT**

Crime prediction is an important application of big data for law enforcement especially in large cities. For this project, the goal is to develop a framework for crime prediction using diverse types of data available. I formulated the problem as a binary classification applied logistic regression, decision tree and gradient boosting tree method. Finally, the accuracy of the model reached 74% , which exceeds the baseline method, and I think the goal was achieved.

# INTRODUCTION

Urbanization creates a lot of social problems. One of these problems inherent in all cities of the world is crime. Police databases accumulate a large amount of data that could be analyzed in order to reduce crime rates. The analysis of criminal activity and prediction of number of crimes remains one of the most interesting problem for researchers. Specifically, given a location, time of day (say, late evening), and day of week (say, weekday), and auxiliary data about the location, predict whether a crime will occur. If successful, the prediction model can then be applied to all locations.

The goal is using machine learning models can help to make accurate prediction in the urban crime domain.

The plan is to apply classification method on a combination of location data, time data and auxiliary data indicators to improve the prediction.

Collect and analyze the crime data, state senate districts data and business improvement districts data in order to test the hypothesis.

Crime data and other auxiliary data is collected from Denver's government website. The data is numerous and complex. We need to take time to understand and search the data and learn pandas and matplotlib to preprocess and analyze data.

From the model evaluation results, the best f1 score of the model reached 0.85, and we achieve the goal to predict urban crime.

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

Crime dataset includes criminal offenses in the City and County of Denver for the previous five calendar years plus the current year to date. The underlying format is csv. And download from

<https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-crime>.

State Senate Districts dataset contains information on where the Senate districts in Colorado are, the names of the representatives that serve each district, and the party affiliation of each representative. The underlying format is csv. And download from

<https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-state-senate-districts>

Business Improvement Districts dataset approves the petition by ordinance, organizing the BID and defining the boundaries. The underlying format is csv. And download from

<https://www.denvergov.org/opendata/dataset/city-and-county-of-denver-business-improvement-districts>

Merge data from all dataset above based on same DISTRICT\_ID attribute.

|  |  |  |
| --- | --- | --- |
| **Attribute name** | **Type** | **Description** |
| INCIDENT\_ID | int64 | ID of incident |
| OFFENSE\_ID | int64 | ID of offense |
| OFFENSE\_CODE | int64 | Code of offense |
| OFFENSE\_CODE\_EXTENSION | int64 | Code extension of offense |
| OFFENSE\_TYPE\_ID | object | Type id of offense |
| OFFENSE\_CATEGORY\_ID | object | Category id of offense |
| FIRST\_OCCURRENCE\_DATE | object | Date of first occurrence |
| LAST\_OCCURRENCE\_DATE | object | Date of last occurrence |
| REPORTED\_DATE | object | Date of reported |
| INCIDENT\_ADDRESS | object | Address of incident |
| GEO\_X | float64 | Geography X |
| GEO\_Y | float64 | Geography Y |
| GEO\_LON | float64 | Geographic longitude |
| GEO\_LAT | float64 | Geographic latitude |
| DISTRICT\_ID | int64 | Id of district |
| PRECINCT\_ID | int64 | Id of precinct |
| NEIGHBORHOOD\_ID | object | Id of neighborhood |
| IS\_CRIME | int64 | 1 for crime and 0 for not |
| IS\_TRAFFIC | int64 | 0 for traffic and 1 for not |

**Table 1**: Attributes of the data acquired from crime data (source)

It is spatial data (crime for Denver). We need to count missing values , count the size of the raw data, discard a lot of irrelevant features and drop redundant columns.

|  |  |
| --- | --- |
| Number of observations | 463,462 (~67 MB) |
| Number of attributes | 19 columns |
| % missing values | 68.5% for LAST\_OCCURRENCE  \_DATE, 9.5% for INCIDENT  \_ADDRESS, 0.9% for GEO\_X, GEO\_Y, GEO\_LON, GEO\_LAT |
| % irrelevant features | 15 |
| % redundant column | 1 |

**Table 2**: Summary statistics of the raw data from crime (source).

Count missing values and drop the columns with missing value since we don’t need these features to build model. Process FIRST\_OCCURRENCE\_DATE object type to datetime type, use 24-hour clock and create date feature. Apply discretization method on the date attribute create timeOfDay feature judging the time. Create dayOfWeek feature where Monday is 0 and Sunday is 6. Integrate party and business improvement district auxiliary data. Pick the appropriate features to create train set. One hot object features. Rename one hot features from district name. Create X (features) and y (crime label).

The set of features use to create the Dataframe object for data analysis software are DISTRICT\_ID, dayOfWeek, timeOfDay\_

afternoon, timeOfDay\_morning, timeOfDay\_night, PARTY\_D, PARTY\_R, CCN\_BID, CCS\_BID, C\_BID, DD\_BID, OSG\_BID, WC\_BID. It is a classification problem, the categories are divided into two categories, 1 means crime 0 means no crime.

|  |  |  |
| --- | --- | --- |
| **Attribute name** | **Type** | **Description** |
| DISTRICT\_ID | int64 | Id of district |
| dayOfWeek | int64 | Day of week |
| timeOfDay\_afternoon | uint8 | Is it the afternoon of the day |
| timeOfDay\_morning | uint8 | Is it the morning of the day |
| timeOfDay\_night | uint8 | Is it the night of the day |
| PARTY\_D | uint8 | Democratic Party |
| PARTY\_R | uint8 | Republican Party |
| CCN\_BID | uint8 | DISTRICT\_NAME\_CHERRY CREEK NORTH BUSINESS IMPROVEMENT DISTRICT NO. 1 |
| CCS\_BID | uint8 | DISTRICT\_NAME\_CHERRY CREEK SUBAREA BUSINESS IMPROVEMENT DISTRICT |
| C\_BID | uint8 | DISTRICT\_NAME\_COLFAX BUSINESS IMPROVEMENT DISTRICT |
| DD\_BID | uint8 | DISTRICT\_NAME\_DOWNTOWN DENVER BUSINESS IMPROVEMENT DISTRICT |
| OSG\_BID | uint8 | DISTRICT\_NAME\_OLD SOUTH GAYLORD BUSINESS IMPROVEMENT DISTRICT |
| WC\_BID | uint8 | DISTRICT\_NAME\_WEST COLFAX BUSINESS IMPROVEMENT DISTRICT |

**Table 3**: Attributes of the data acquired from merged data (preprocessed)

# METHODOLOGY

Crime dataset

State Senate Districts dataset

Business Improvement Districts dataset

Data preprocess

Data collection

Data analysis

Discard irrelevant features

Drop redundant columns

Missing value

Change feature type

Discretization (one hot)

Gradient boosting tree

Cross validation

Logistic regression

Preprocess.ipynb: this is the Jupyter notebook file to preprocess raw data. The output of the script is train.csv.

Modeling.ipynb: this is the Jupyter notebook file to perform the classification task of the project.

# EXPERIMENTAL EVALUATION

## Experimental Setup

This section includes:

1. Computing platform (what operating system and hardware you use to do the experiment). Are you using AWS cluster? If so, how many nodes?

Windows 10, i7-9700, 32GB hyper ram, dual 2080ti

The baseline method is people's random guesses about whether a crime has occurred, with an accuracy rate of 50%.

Use accuracy evaluation metric to report the results.

## Experimental Results

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic regression | 74% |
| Decision tree | 74% |
| Gradient boosting tree | 74% |

**Table 4**: model accuracy

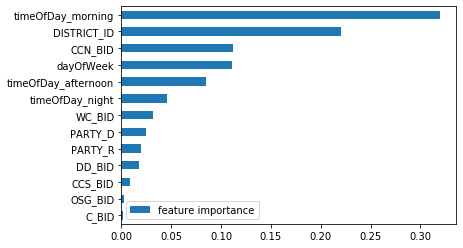


Fig 1. feature importance from gradient boosting tree model

We can find that timeOfDay\_morning and DISTRICT\_ID feature are important for making the model predictions.

Finally, we can see the best accuracy of models is 74%, the project is successful.

# CONCLUSIONS

We did data collection, data preprocess and data analysis to find out whether crimes are related to time of day and district. But the accuracy is still not high enough, we need more features to help make more accurate crime predictions.

# REFERENCES (at least 3 references)

1. V. Ingilevich, S. Ivanov. Crime rate prediction in the urban environment using social factors. Procedia Computer Science 136 (2018) 472–478.
2. S. Curtis-ham and D. Walton, “Mapping crime harm and priority locations in New Zealand : A comparison of spatial analysis methods,” Appl. Geogr., vol. 86, pp. 245–254, 2017.
3. O. K. Ha and M. A. Andresen, “Journal of Criminal Justice Unemployment and the specialization of criminal activity : A neighborhood analysis,” vol. 48, pp. 1–8, 2017.

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)?
4. Participation in the group project. How much did a team member contributes to the project.