Crime Rate Prediction of Chicago

Hang Zhao
Project URL:
https://github.com/hangzhao618/
cse482proj_Crime-Rate-Prediction.git

ABSTRACT

Chicago is one of the largest cities in the United States, and it is also one of the cities that has a large crime rate. Big data can sufficiently predict the crime rate. First, this project is to analyze the crime rate of Chicago at different times, like a weekday, weekend, morning of the day, and afternoon of the day. Second, this project utilizes the same way of analyzing the crime rate of the City of Chicago as analyzing the crime rate of each community in Chicago. It also adds the data of population density to analyze the correlation of population density and the crime rate within the community. Finally, this project will analyze individual crime situation of each community, such as the percentage of different type of crime, the rate of crime in the morning, afternoon and evening, and the rate of weekday and weekend.

1. INTRODUCTION

Crime has always been a problem that plagues people, reducing crime rates is a very important thing. In order to effectively control crime, only to understand what factors cause the high crime rate. Chicago is the third-largest city in the U.S., and it is one of a few big U.S. cities that has struggled with gun violence in recent years. The crime rate of Chicago is higher than the US average [4]. To predict the occurrence of crime can help people avoid danger.

This project is to analyze the crime rate of Chicago at different times, like a weekday, weekend, morning of the day, and afternoon of the day. Then using regression method on a combination of each time and crime rate, and evaluate which type is better to predict the crime occurrence. The project utilizes the same way of analyzing the crime rate of the City of Chicago as analyzing the crime rate of each community in Chicago. It also adds the data of population density to analyze the correlation of population density and the crime rate within the community. Finally, this project will analyze individual crime situation of each community, such as the percentage of different type of crime, the rate of crime in the morning, afternoon and evening, and the rate of weekday and weekend. It will also use the regression method to analyze which attribution is the best one to predict the crime occurrence.

The three datasets used in this project are the Chicago Crime data, Chicago Community Area Census data, and the Chicago Community Areas data. The Chicago Crime data is the main data in this project, this data set contains all Chicago crime information, like occur data, location, type of the crime, and so on. This project will use those data to analysis the crime rate of Chicago, the crime rate of each community in Chicago, and the crime rate for different time frames. The census data includes all the population of the community, but only two data are used in the project, one is the "total population for each community" and the other is the "community code". The last data set is community areas data, which contains all community area information, such as community code, community name, geometric information, and so on.

There are more than 6,000,000 crime recorded in Chicago crime dataset. It will take a long time to preprocess those data, and sometime the program will shut down when process the dataset. To separate those data into small part is the best way to process data and can avoid the program shut down.

In this project, it will analyze the rate of each type of crime in Chicago, and find which type of crime is the highest one in Chicago.

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

2. DATA

Chicago crime datasets was download from City of Chicago in the CSV format [3]. There is include 22 attributes for each record, like ID, Case Number, Date, Block, UCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, and Location, below is a crime record from crime datasets:

> [10000092, 'HY189866', '03/18/2015 07:44:00 PM', '047XX W OHIO ST', '041A', 'BATTERY', 'AGGRAVATED: HANDGUN', 'STREET', False, False, 1111 , 11.0, 28.0, 25.0, '04B', 1144606.0, 1903566.0, 2015, '02/10/2018 03:50:01 PM', 41.891398861, -87.74438456700001, '(41.891398861, -87.744384567)']

There are mission two attributes in the crime dataset, one is the community name and the other is the population of each community, so we need other two datasets to get information about community name and population. The community area boundaries in Chicago datasets was download from City of Chicago in the CSV format [2], there is include 10 attributes for each record, below is a crime record from crime datasets:

> ['MULTIPOLYGON()', 0, 0, 0, 0, 1, 'ROGERS PARK', 1, 51259902.4506, 34052.3975757]

The last datasets is 2010 Census Data Summarized to Chicago Community Areas [1], it was download from City of Chicago, there is include 129 attributes for each row, but only 3 attributes will be use in this project, Table 1 show the attribute name that used in the project and an example for each attribute.

Geog	GeogKey	Total Population
Rogers Park	1	54991

Table 1: Census Data for Chicago Community

The above describes the three datasets that this project will use. The crime dataset is the main datasets in this project. We join the other two database into crime datasets base on attributes: 'Community Area', 'GeogKey' and 'AREA_NUMBE'.

Crime datasets include all crime record of Chicago from 2001 to 2018. It has some repeat attributes that must be discarded. The crime datasets include 6132365 crimes after removing some record that missing community information.

For the crime data, dropping all attributes that are not using in this project, and splitting the attribute data into three attribute: time, day, month/day/year, then calculate the total crime record base on the three attributes. After get the number of the total crime record, then adding the target attribute named Class. If the number over the average number set the class as 'high', otherwise set it as 'low'.

Below shows all table after preprocessing all datasets.

Table 2 shows basic information about each crime record from crime datasets.

Date	Block	Primary Type	Location Description	Arrest	Community Area	Year
03/18/2015/ 07:44:00 PM	047XX W OHIO ST	BATTERY	STREET	False	25.0	2015

Table 2: Basic information about each crime record

Table 3 shows the result after process crime datasets,

Community Name	Population Rate	Time	Day	Date	Count
ROGERS PARK	0.02	Morning	Weekday	01/01/2002	0.000021

Table 3: The result after process crime datasets

Table 4 is very similar Table 3, but using the integer value to instead string value except for attribute 'Date'. And added a new attribute named 'Class' in the datasets. This attribute will set as the class when using the classification.

Community Name	Population Rate	Time	Day	Date	Count	Class
1	0.02	1	1	01/01/2002	2	low

Table 4: The result after process crime datasets (for classification method)

Table 5 shows the size of the final dataset after preprocessing.

Dataset Name	Number of rows	Number of columns	Size(Mbytes)
crime_data	6130365	7	498.1
population	77	4	0.002
community_code	77	2	0.001
com_datas	1516689	6	102.3
com_datas_int	1516689	7	45.8

Table 5: The size of the final dataset after preprocessing.

3. METHODOLOGY

All datasets were downloaded from the website and using them directly.

All data preprocessing was done in one Jupyter notebook file. First, loading all datasets into the notebook. Because the size of data is too large, I wrote a Python function to split those base on the community name. Then, I am adding the population data into the sub-datasets. Calculate the number of crimes based on a period of time. Finally, the four datasets are output in their CSV format.

After preprocessing all datasets, then I created a new Jupyter notebook file to do data analysis. First, I wrote the code to calculate the number of crimes for each type and plot it. Then apply the linear regression model to predict the crime rate, and calculate its root-mean-square error, r-square coefficient and regression coefficients of the model. Finally, using scikit learn train_test_split()function to split X and Y into training and tests sets, and setting the size of the training 20% and the remaining 80% for test. Apply a decision tree classifier to the training dataset.

Finally, give a brief summary of the the code you have written for this project. For example, you can summarize it as follows:

- Preprocess.ipynb: this is the Jupyter notebook file to convert time of event to time of day – morning, afternoon, evening; converting day of event to weekend or weekday; calculating the population rate, and set the class for each event. The output of the script four CSV format datasets.
- Modeling.ipynb: this is the Jupyter notebook file to perform the classification and regression task of the project.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

1. In this project, I used my MacBook pro to do the experiment, and the operating system is macOS 10.14.
2. In this project, the baseline the crime rate as the baseline to compare the results.
3. I used the accuracy to evaluate the results.

4.2 Experimental Results

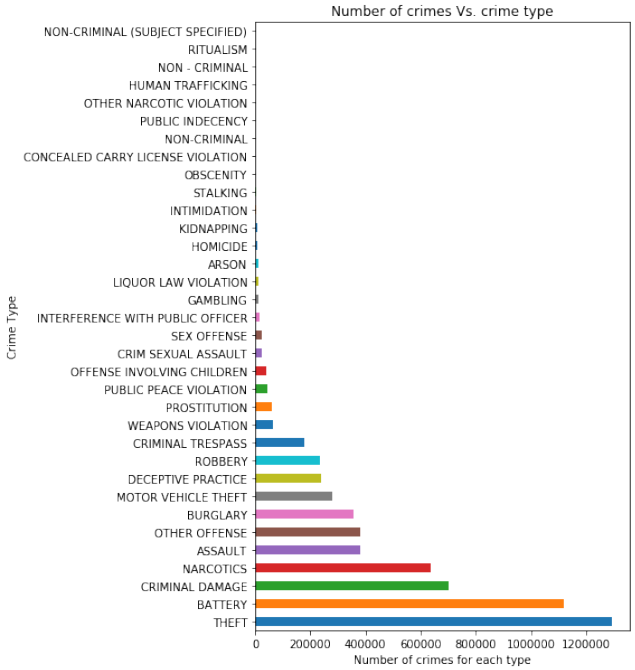

Figure 1: Number of crimes by type

Figure 1 shows the number of crime by crime type. In this plot, we easy to see that theft has the highest crime rate in Chicago.

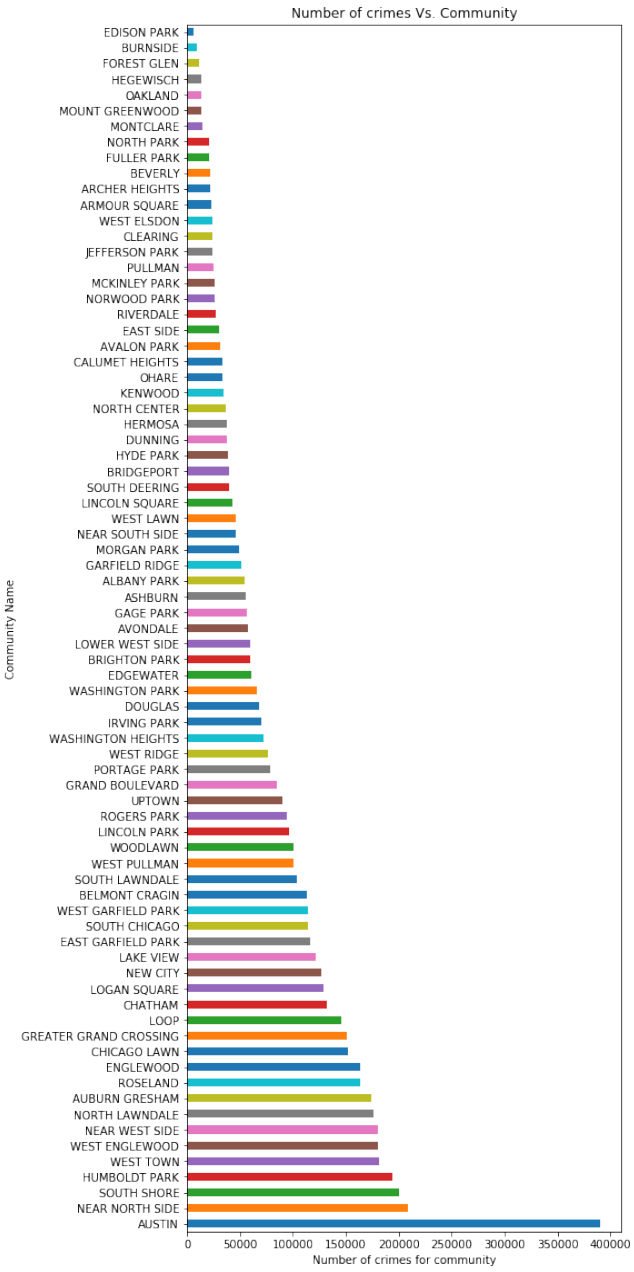

Figure 2: Number of crimes by Community

Figure 2 shows the number of crime by community. In this plot, we easy to see that Austin has the highest crime rate in Chicago.

Table 6 shows the result after apple the linear regression model.

RMSE	R-squared	Slope	Intercept
3.2576	0.2064	[2.38773950e-02 2.11455115e+02 3.25570850e-01 -5.41098504e-02]	-0.68

Table 6: The result of the linear regression model.

Table 7 shows the result after apple the classification model.

Classification Method	Accuracy
Decision Tree	0.81
Nearest Neighbor	0.807
Random Forest	0.8108

Table 7: The result of classification models

Due to the accuracy over 0.8, this project is successful.

5. CONCLUSIONS

In this project, we got the high accuracy for test case. Due to its high accuracy, this project is successful. But in this project, I only used four features to predict the crime rate. In the future, I will add more features to predict the crime rate, it may increase accuracy.

6. REFERENCES (at least 3 references)

[1] "2010 Census Data Summarized to Chicago Community Areas - Spreadsheet: 2010 Census Data Summarized to Chicago Community Areas." *CMAP Data Hub*, datahub.cmap.illinois.gov/dataset/2010-census-data-summarized-to-chicago-community-areas/resource/b30b47bf-bb0d-46b6-853b-47270fb7f626?inner_span=True.

[2] "Boundaries - Community Areas (Current) | City of Chicago | Data Portal." *Chicago*, data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6.

[3] "Crimes - 2001 to Present - Map | City of Chicago | Data Portal." *Chicago*, data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Map/c4ep-ee5m.

[4] Madhani, Aamer. "Crime down in Windy City? Chicago Just Had Its Least Violent January in 9 Years." *USA Today*, Gannett Satellite Information Network, 1 Feb. 2019, www.usatoday.com/story/news/2019/02/01/chicago-least-violent-month-2019/2742741002/.