San Francisco Crime Classification

Jia Huang

[1] Xi'an Shiyou University, China

Introduction

San Francisco was infamous for housing some of the world's most notorious criminals on the inescapable island of Alcatraz. Today, the city is known more for its tech scene than its criminal past. From Sunset to SOMA, and Marina to Excelsior, this project analyzes 12 years of crime reports from across all of San Francisco's neighborhoods to create a model that predicts the category of crime that occurred, given time and location.

The Dataset is in a tabular form and includes chronological, geographical and text data and contains incidents derived from the SFPD Crime Incident Reporting system.

Data Visualization is the main method designed in this project. Through data visualization, the data set we will present will finally be visualized and the understanding of the data will be more direct.

The main task of this project is to make a prediction of the type, time and place of crime in San Francisco. In this article, we will give an overview of the whole project from data sets, eigenvalues, feature item selection, modeling and conclusion.

The Dataset

The data ranges from *1/1/2003 to 5/13/2015* creating a training dataset with nine features and 878,049 samples.

- *Dates - timestamp of the crime incident*
- *Category - category of the crime incident. (This is our target variable.)* and
- *Descript - detailed description of the crime incident*
- *DayOfWeek - the day of the week*
- *PdDistrict - the name of the Police Department District*
- *Resolution - The resolution of the crime incident*
- *Address - the approximate street address of the crime incident*
- *X - Longitude*
- *Y - Latitude*

Feature Item

The dataset contains 2,323 duplicates that are meaningless and should be deleted.We will also use coordinates to calculate the distribution of data points on the map of San Francisco.At the same time, 67 incorrect locations were found.

Datas	datetime64
Category	object
Descript	object
DayOfWeek	object
PdDistrict	object
Resolution	object
Address	object
x	float64
Y	float64

Dates & Day of the week These variables are distributed uniformly between 1/1/2003 to 5/13/2015 (and Monday to Sunday) and split between the training and the testing dataset as mentioned before. We did not notice any anomalies on these variables. The median frequency of incidents is 389 per day with a standard deviation of 48.51.

Per Week Also, there is no significant deviation of incidents frequency throughout the week. Thus we do not expect this variable to play a significant role in the prediction.

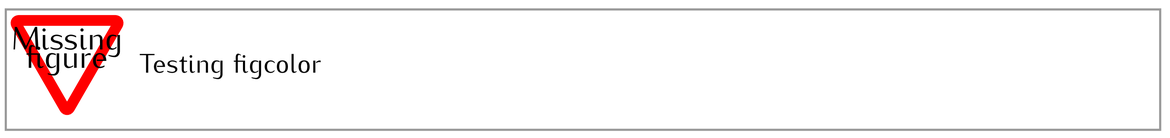
Feature Item

Category There are 39 discrete categories that the police department file the incidents with the most common being Larceny/Theft (19.91%), Non/Criminal (10.50%), and Assault(8.77%).

Police District There are significant differences between the different districts of the City with the Southern district having the most incidents (17.87%) followed by Mission (13.67%) and Northern (12.00%).

X - Longitude Y - Latitude We have tested that the coordinates belong inside the boundaries of the city. Although longitude does not contain any outliers, latitude includes some 90o values which correspond to the North Pole.

Address as a text field, requires advanced techniques to use it for the prediction. Instead in this project, we will use it to extract if the incident has happened on the road or in a building block.



Testing figcolor

Data Visualization

Based on the Projects statement, we need to predict the probability of each type of crime based on time and location. That being said, we present two diagrams to visualize the importance of these variables. The first one presents the geographic density of 9 random crime categories. We can see that although the epicenter of most of the crimes resides on the northeast of the city, each crime has a different density on the rest of the city. This fact is a reliable indication that the location (coordinates / Police District) will be a significant factor for the analysis and the forecasting.

The diagram presents the average number of incidents per hour for five of the crimes' categories. It is evident that different crimes have different frequency during different times of the day. Some examples are that prostitution picks during the evening and all through the night, Gambling incidents start late at night until the morning and Burglary picks early in the morning until the afternoon. As before these are sharp pieces of evidence that the time parameters will have a significant role also.

Testing figcolor

New Orleans Pelicans on FT%

Testing figcolor

New Orleans Pelicans on FTA

Conclusion

Problem Definition To examine the specific problem, we will apply a full Data Science life cycle composed of the six steps.
In this paper, we did not complete the design of this project in [illegible] with the whole journey, but through the main data [illegible] and analyze the data set provided visual[illegible]

Acknowledgement

- International Cooperation Project (Y7Z0511101) of IIE, Chinese Academy of Sciences