SAN FRANCISCO CRIME CLASSIFICATION

JIA HUANG

Abstract. San Francisco is notorious because some of the world's most notorious criminals live on the inescapable Island of Alcatraz.Today, the city is better known for its tech scene than its criminal history.From Sunset to SOMA, Marina to Excelsior, the project analyzed crime reports from all San Francisco neighborhoods over a 12-year period, visualizing the data ina more intuitive and concise way, and creating a model that predicted the type of crime that would occur at a given time and place.

Contents

Date: 2020-10-16.
1991 *Mathematics Subject Classification.* San Francisco Crime Classification.

1. Introduction

Crime is a social phenomenon as old as societies themselves, and although there will never be a free from crime society - just because it would need everyone in that society to think and act in the same way - societies always look for a way to minimize it and prevent it. In the modern United States history, crime rates increased after World War II, peaking from the 1970s to the early 1990s. Violent crime nearly quadrupled between 1960 and its peak in 1991. Property crime more than doubled over the same period. Since the 1990s, however, crime in the United States has declined steadily. Until recently crime prevention was studied based on strict behavioral and social methods, but the recent developments in Data Analysis have allowed a more quantitative approach in the subject. We will explore a dataset of nearly 12 years of crime reports from across all of San Francisco's neighborhoods, and we will create a model that predicts the category of crime that occurred, given the time and location.

To examine the specific problem, we will apply a full Data Science life cycle composed of the following steps:

1.Data Wrangling to audit the quality of the data and perform all the necessary actions to clean the dataset.
2.Data Exploration for understanding the variables and create intuition on the data.

3.Feature Engineering to create additional variables from the existing.

4.Data Normalization and Data Transformation for preparing the dataset for the learning algorithms (if needed).

5.Training / Testing data creation to evaluate the performance of our models and fine-tune their hyperparameters.

6.Model selection and evaluation. This will be the final goal; creating a model that predicts the probability of each type of crime based on the location and the date.

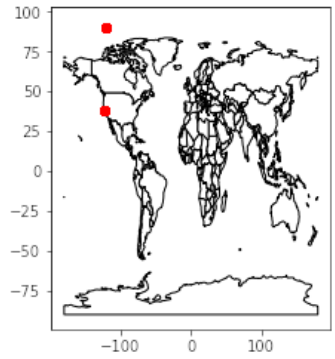
2. Data Processing

The data set is tabular, consisting of temporal, geographic, and textual data, and contains events derived from the SFPD Crime Incident Reporting system.For the given data, we first make a comprehensive analysis, and the following conclusions can be drawn:

```
First date:  2003-01-06 00:01:00
Last date:  2015-05-13 23:53:00
Test data shape  (878049, 9)
```

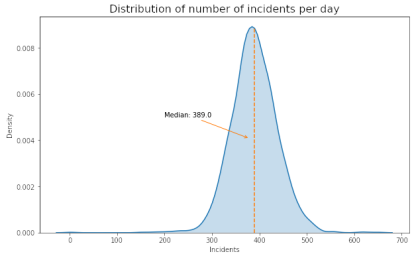
Flag.1
The data ranges from 1/1/2003 to 5/13/2015 creating a training dataset with nine features and 878,049 samples

(1) This is meaningless for our analysis and prediction, and it needs to be deleted. At the same time, according to the analysis, we can see that there are 67 locations that are wrong in the map.

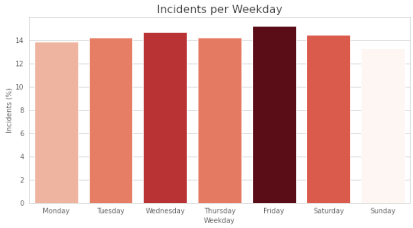


Flag.2

(2) The following two figures analyze the frequency of criminal incidents in terms of days and weeks respectively. It can be seen that there are significant differences between the two figures.



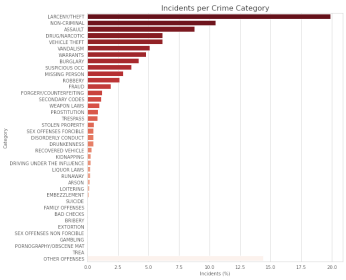
Flag.3



Flag.4

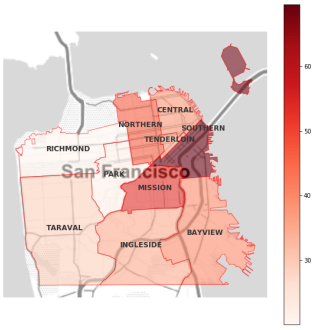
These variables are distributed uniformly between 1/1/2003 to 5/13/2015 (and Monday to Sunday) and split between the training and the testing dataset as mentioned before. We did not notice any anomalies on these variables.
Also, there is no significant deviation of incidents frequency throughout the week. Thus we do not expect this variable to play a significant role in the prediction.
The median frequency of incidents is 389 per day with a standard deviation of 48.51.

(3)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%).



Flag.5

(4)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%).



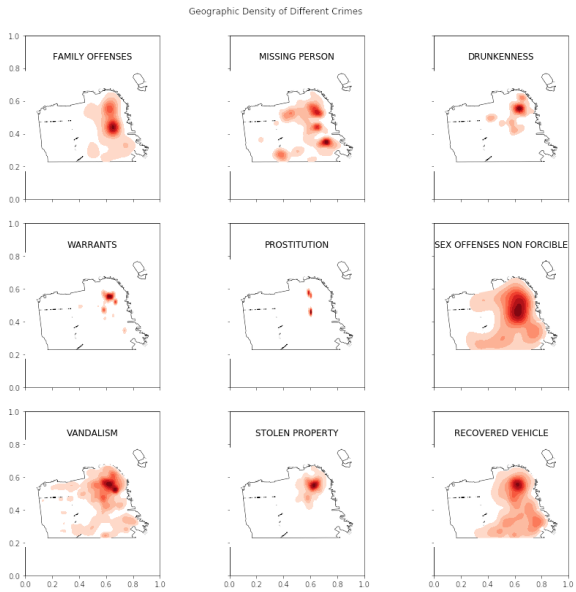
Flag.5

(5)Address
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.
(6)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 ludes some 90o values which correspond to the North Pole.

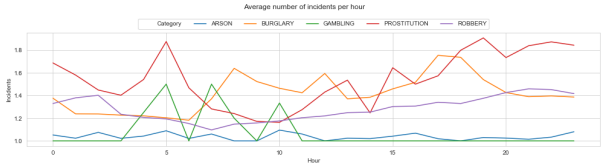
According to the nine features and meanings provided by the data set, we analyze them from each Angle and get a different result.Similarly, different perspectives will make our analysis of data more comprehensive and the results more reliable.

3. Exploratory Visualization

IBased on the Project's 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 second 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.



Flag.5

4. Methodology

4.1 Feature Engineering Then, we created additional features. More specifically:

- From the 'Dates' field, we extracted the Day, the Month, the Year, the Hour, the Minute, the Weekday, and the number of days since the first day in the data.
- From the 'Address' field we extracted if the incident has taken place in a crossroad or on a building block.

4.2 Feature Selection

the feature engineering described above, we ended up with 11 features. To identify if any of them increased the complexity of the model without adding significant gain to the model, we used the method of Permutation Importance.

The idea is that the importance of a feature can be measured by looking at how much the loss decreases when a feature is not available. To do that we can remove each feature from the dataset, re-train the estimator and check the impact.

Doing this would require re-training an estimator for each feature, which can be computationally intensive. Instead, we can replace it with noise by shuffle values for a feature.

The implementation of the above technique showed that there is no need for any feature removal since all of them have a positive impact in the dataset.(1) The update rate of cost function is the cost function of mean square error. We can clearly see that after 10 times of updating, the loss value of his cost function remains around 8000.

Out[19]:

Weight	Feature
0.0579 ± 0.0012	Minute
0.0470 ± 0.0008	Y
0.0355 ± 0.0008	X
0.0179 ± 0.0002	Block
0.0176 ± 0.0008	n_days
0.0138 ± 0.0009	Hour
0.0129 ± 0.0007	PdDistrict
0.0108 ± 0.0004	Year
0.0028 ± 0.0002	Month
0.0017 ± 0.0004	Day
0.0014 ± 0.0003	DayOfWeek

5. Algorithms and Techniques

The concrete problem is a typical multi-class classification problem, and there are several algorithms to solve this problem.First, we evaluated several suitable algorithms, from linear models (stochastic gradient descent), nearest neighbors (K nearest neighbors), set methods (random forest and AdaBoost), and enhancement algorithms (XGBoost and LIghtGBM), using basic feature engineering and default parameters to assess whether they have a clear lead.

LightGBM is a decision tree boosting algorithm uses histogram-based algorithms which bucket continuous feature (attribute) values into discrete bins. This technique speeds up training and reduces memory usage. In layman terms the algorithm works like this:

Fit a decision tree to the data
Evaluate the model
Increase the weight to the incorrect samples.
Choose the leaf with max delta loss to grow.
Grow the tree.
Go to step 2

6. Conclusions

This paper mainly abstracts an outline of a crime type and analysis project in San Francisco, and gives the steps required for data processing of a project.Provides a general idea of the shape of a project.At the same time, it also gives some data visualization, which helps us understand the data more intuitively.

List of Todos

(A. 1) School of Computer Science,, Xi'an Shiyou University, Shaanxi 710065, China
Email address, A. 1: xxx@tulip.academy