The Final Report

Wang Mingxi

Jilin University

College of Computer Science and Technology

2021-04-25

Overview

Research motivation and context

Project Objectives

- The project analyzed 12 years of crime reports from all of San Francisco's neighborhoods to create a model that can predict crime categories at a given time and place.

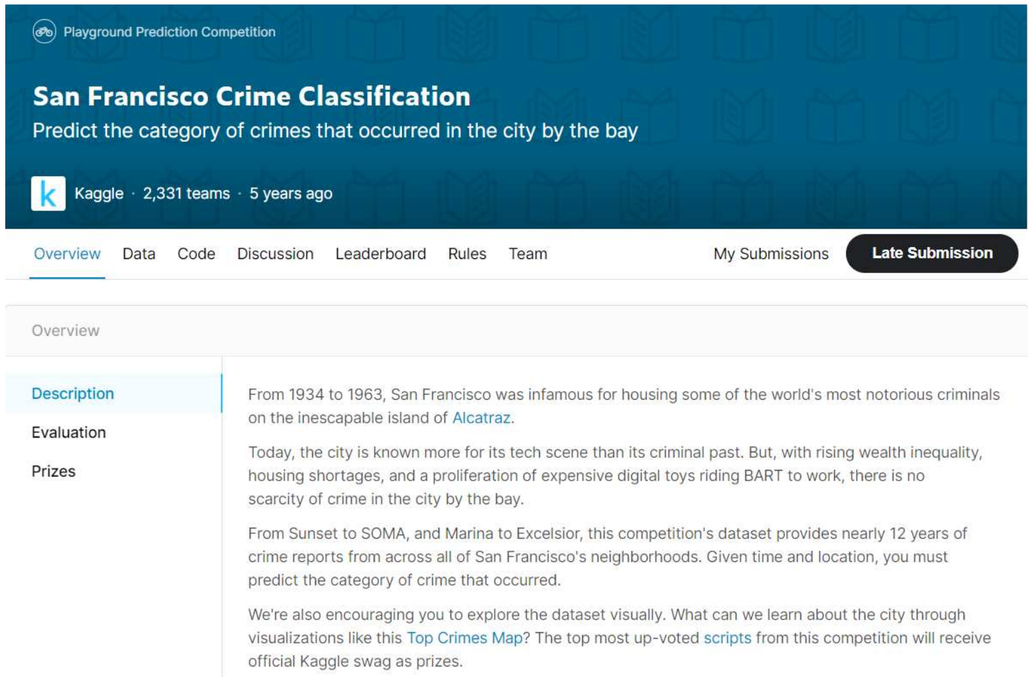

Figure 1: project preview

Project background

Crime is a kind of behavior that does great harm to the society. It brings great loss of life and property to human beings.In modern society, human beings prevent crime through strict judicial means, and research on crime prevention based on sociology, psychology and law.The objective of this project is to quantitatively analyze a data set of nearly 12 years of crime reports from all San Francisco neighborhoods by computer means and to create a model that predicts the type of crime that will occur given information such as time and location.

Research contents and methods

Evaluation metric

Multi-class logarithmic loss, which measures that the model output is a probability value between 0 and 1.

Data set preview

- 878,049 samples, a total of 9 features.

Date, category, description, day of the week, name of the police district, solution, approximate street address of the crime, longitude, latitude.

```
print('First date: ', str(train.Dates.describe()['first']))
print('Last date: ', str(train.Dates.describe()['last']))
print('Test data shape ', train.shape)
```

Figure 2: preview

First date: 2003-01-06 00:01:00

Last date: 2015-05-13 23:53:00

Test data shape (878049, 9)

Data set preview II

- train.head()

	Dates	Category	Descript	DayOfWeek	PdDistrict	Resolution	Address	X	Y
0	2015-05-13 23:53:00	WARRANTS	WARRANT ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
1	2015-05-13 23:53:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	OAK ST / LAGUNA ST	-122.425892	37.774599
2	2015-05-13 23:33:00	OTHER OFFENSES	TRAFFIC VIOLATION ARREST	Wednesday	NORTHERN	ARREST, BOOKED	VANNESS AV / GREENWICH ST	-122.424363	37.800414
3	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	NORTHERN	NONE	1500 Block of LOMBARD ST	-122.426995	37.800873
4	2015-05-13 23:30:00	LARCENY/THEFT	GRAND THEFT FROM LOCKED AUTO	Wednesday	PARK	NONE	100 Block of BRODERICK ST	-122.438738	37.771541

Figure 3: preview

Data cleaning

- There are 2323 duplicate items that need to be removed

train.duplicated().sum()

- Throw away samples in a range of longitude and latitude (e.g., 50)

Characteristics of the engineering

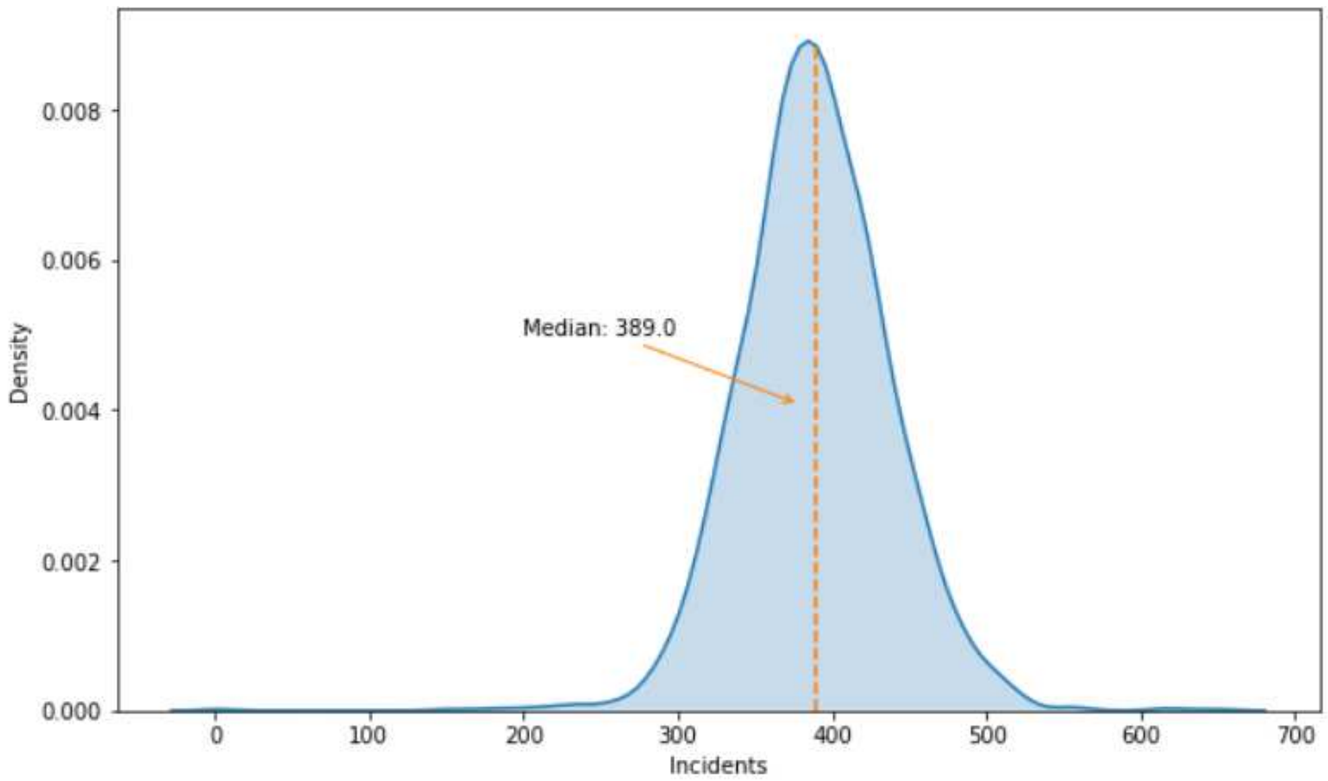

Figure 4: Distribution of number of incidents per day

Characteristics of the engineering II

Similarly, there was no significant deviation in the frequency of events over the course of the week.

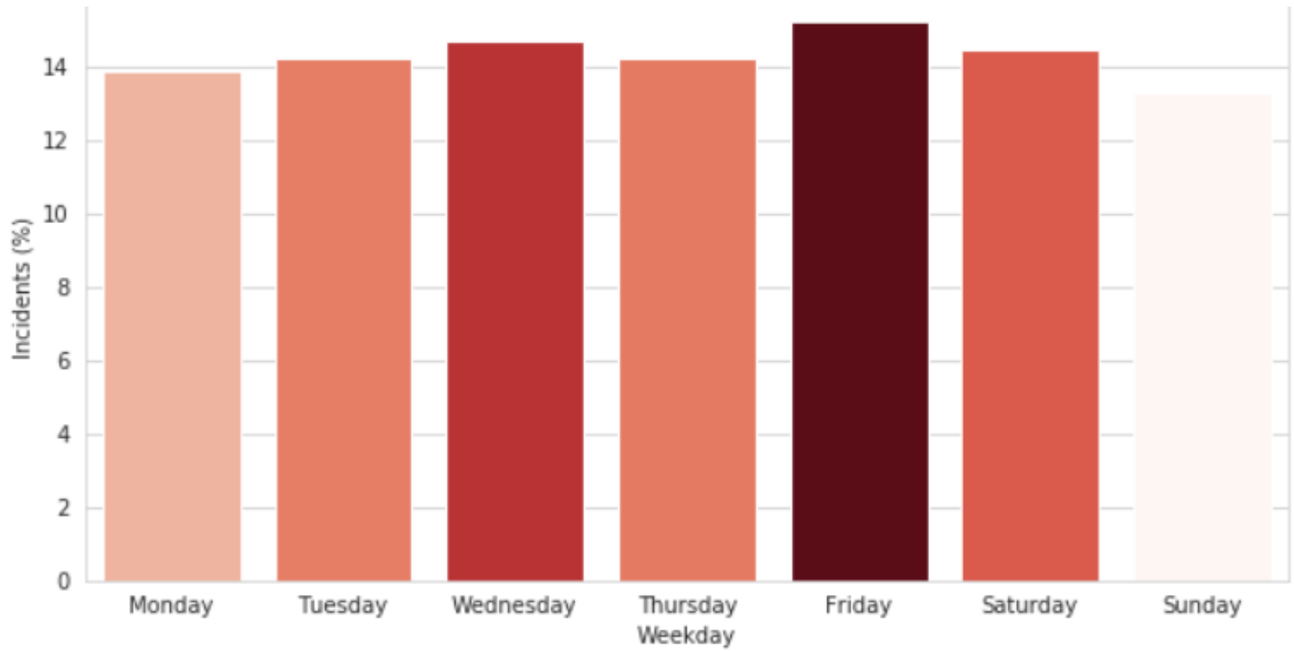

Figure 5: incidents per Weekday

Characteristics of the engineering III

A total of 39 discrete categories of incidents were recorded by police stations, the most common being theft (19.91%), non-criminal cases (10.50%) and assault (8.77%)

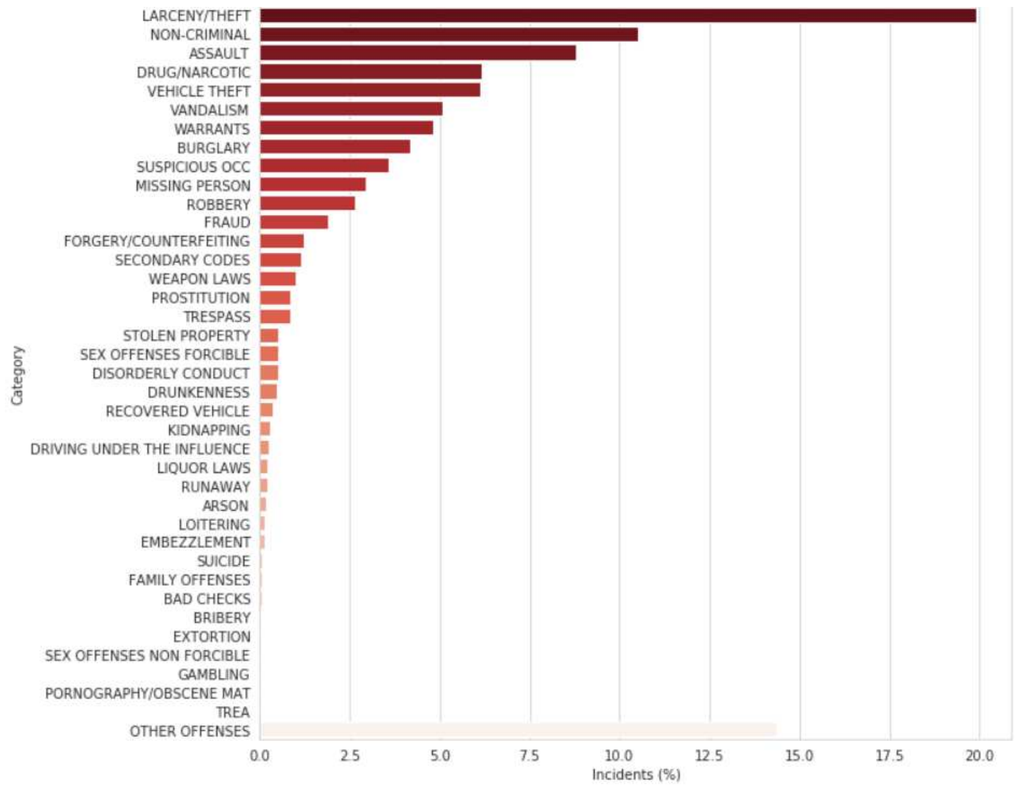

Figure 6: incidents per Crime Category

Characteristics of the engineering IV

The chart below shows the average number of incidents per hour for the five crime categories. Obviously, different crimes occur with different frequencies at different times of the day. Prostitution, for example, takes place mostly at night, gambling incidents take place from late at night until morning, and burglaries from early morning until afternoon. As before, this is clear evidence that time parameters will also play an important role.

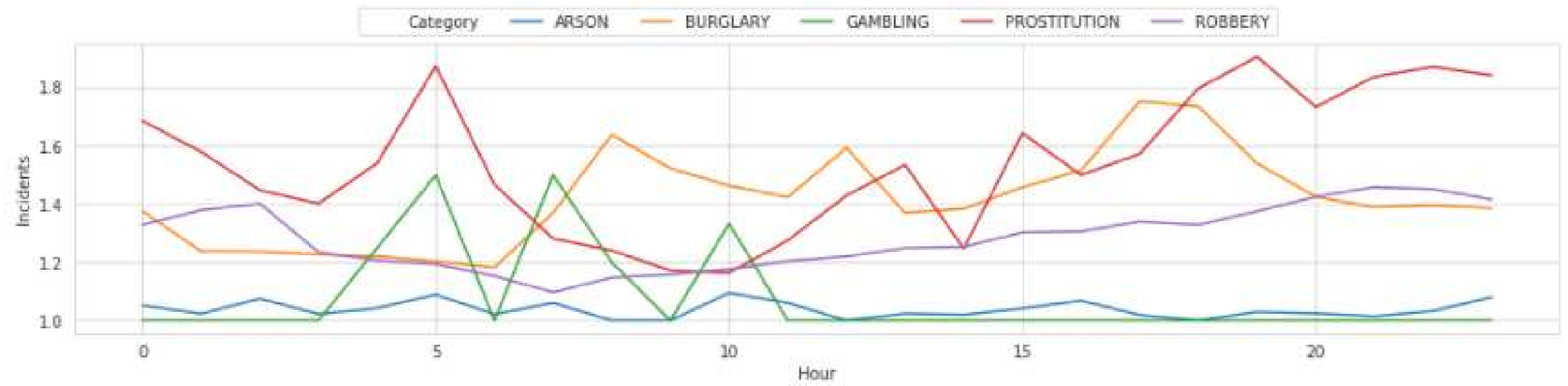

Figure 7: average number of incidents per hour

Characteristics of the engineering V

- From the "Date" field, we extracted the date, month, year, hours, minutes, days, and days since the first day.
- We extracted from the "Address" field whether the event occurred at the intersection or on a building component.

Information coding

The discrete data is converted to a number between 0 and N-1, where N is the number of different values in a list.You can think of it as the number of different values of a feature.

Training model

- Train 23 Epochs.
- Use cross validation to evaluate the quality of our model.

  After training, the model obtained:
- a cross validation score of 2.46.
- it got 2.49 on the test set.

Architecture of the model

Architecture of the model

1. Make the decision tree fit the data
2. Evaluation model
3. Add weight to incorrect samples.
4. Select the leaf nodes with the greatest incremental loss for growth.
5. Generate the tree at the node in the previous step.
6. Go to Step 2 for the cycle

Conclusion

Conclusion

Conclusion: Based on data analysis, processing and utilization, our model can effectively predict crime types.By using the multi-classification algorithm to analyze the input features and update the model parameters through iterative training, our model can be trained more and more accurately.So as to complete the task of classification prediction.

Contact Information

Wang Mingxi

College of Computer Science and Technology

Jilin University, China

MXWANG@TULIP.ACADEMY

TEAM FOR UNIVERSAL LEARNING AND INTELLIGENT PROCESSING