San Francisco Crime classification

Jia Huang

Xi’an Shiyou University
Chinese Academy of Sciences

October 11, 2020

Overview

Project Overview
Data Pre-Processing
Feature Analysis
Feature Selection
Modelling
Model Optimization
Ideas Improvement

Project Overview

- Project Background And Purpose

Data Pre-Processing

- Date Processing
- Feature Item

Feature Analysis

Feature Selection

Modelling

- Calculate the Baseline Value For The Model

Model Optimization

Ideas Improvement

Project Overview
Project Background And Purpose
Data Pre-Processing
Feature Analysis
Feature Selection
Modelling
Model Optimization
Ideas Improvement

Project Overview

Project Background And Purpose

Project Overview
Project Background And Purpose
Data Pre-Processing
Feature Analysis
Feature Selection
Modelling
Model Optimization
Ideas Improvement

Defn

- Background

  From 1934 to 1963, 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. But, with rising wealth inequality, housing shortages, and a proliferation of expensive digital toys riding BART to work, there is no scarcity of crime in the city by the bay. From Sunset to SOMA, and Marina to Excelsior, this dataset provides nearly 12 years of crime reports from across all of San Francisco's neighborhoods.

- Purpose

  predict the category of crime that occurred, given the time and location visualize the city and crimes (see Mapping and Visualizing Violent Crime for inspiration) Content.

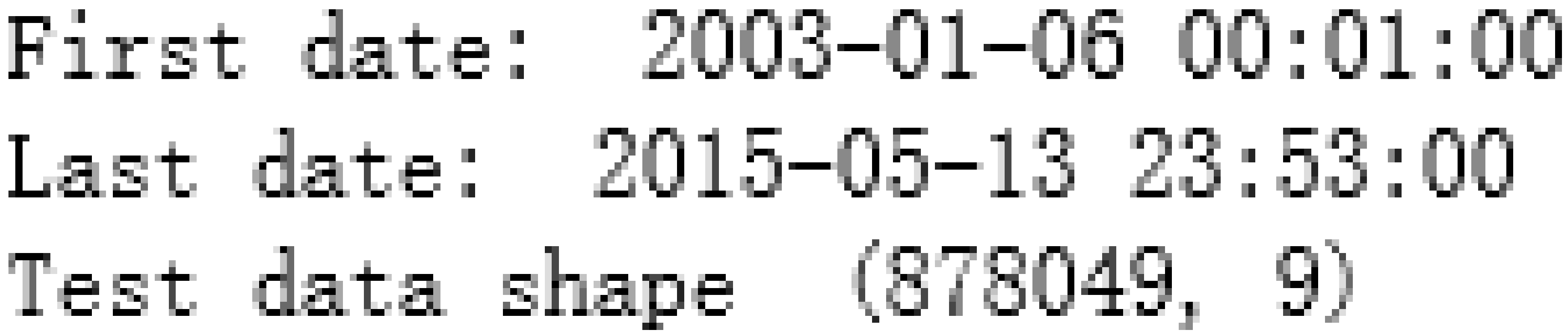
Data Pre-Processing

Date Processing

This dataset contains incidents derived from SFPD Crime Incident Reporting system. The data ranges from 1/1/2003 to 5/13/2015. The training set and test set rotate every week, meaning week 1,3,5,7... belong to test set, week 2,4,6,8 belong to training set. There are 9 variables.

Feature Item

By making a comprehensive analysis of the nine characteristic items in the data set mentioned above,we can reach the following conclusions:

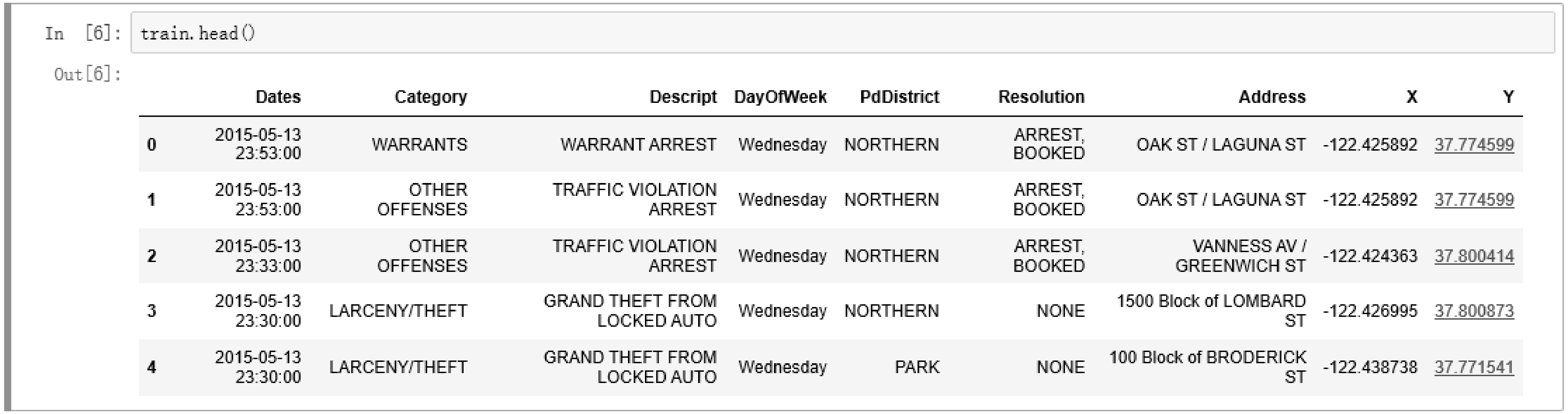


```
First date:  2003-01-06 00:01:00
Last date:  2015-05-13 23:53:00
Test data shape  (878049, 9)
```

Figure 1

The data range was from 1/1/2003 to 5/13/2015, and a data training set containing nine feature items and 87,8049 samples was created.

Features Item



```
In  [6]: train.head()
Out[6]:
```

	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

More specifically it includes the following variables.

- Date - timestamp of the crime Incident.
- Category - category of the crime incident. (This is our target variable.)
- 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 Analysis

Project Overview
Data Pre-Processing
Feature Analysis
Feature Selection
Modelling
Model Optimization
Ideas Improvement

- Statistics Were Made By Type Of ’Year’ And ’Month’
  Based on a comprehensive analysis of the data set provided by Kaggle’s website, it is clear that there are fewer crimes in summer and winter than in spring and fall.Therefore, a "seasonal" feature column can be added to the feature analysis.
- By ’DayOfWeek’ And ’Hour’ Type
  Friday saw the highest number of crimes, probably because of the American tradition of Friday parties.Sunday has the lowest crime rate.So you can add the "weekend or not" feature column.Crime was lowest in the early hours of the morning and highest at 12 o ’clock and 17 and 18 o ’clock in the evening.Therefore, the time zone can be divided and the "time zone" feature column can be added

Feature Selection

Modelling

Calculate the Baseline Value For The Model

Since this is a typical multi-classification problem, we can choose to use many kinds of algorithms, including naive Bayes, KNN, decision tree and random forest.

Model Optimization

Ideas Improvement