San Francisco Crime Classification

1. Introduction

1.1 Background

San Franscisco is the cultural, commercial and financial center of Northern California. It's city with almost 900,000 residents (2019). San Francisco has the highest salaries, disposable income and median home prices in the world. San Francisco was infamous for housing some of the world's most notorious criminals on island of Alcatraz. Today, the city is known more for its tech scene, than its criminal past. But, with rising wealth inequality housing shortgaes there is no scarcity of crime in San Francisco.

1.2 Problem

We would like to predict the category of crime occured in specific location based on coordinates and time. We will Explorer a data set of nearly 12 years of crime reports and we will create a model that predicts the category of crime.

1.3 Interest

San Francisco police would be very interested in accurate prediction.

2. Description of the data

2.1 Data source

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.

2003-01-07 07:52:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	5TH ST / SHIPLEY ST	-122.402843	37.779829
2003-01-07 04:49:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Tuesday	TENDERLOIN	ARREST, BOOKED	CYRIL MAGNIN STORTH ST / EDDY ST	-122.408495	37.784452
2003-01-07 03:52:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	OFARRELL ST / LARKIN ST	-122.417904	37.785167
2003-01-07 03:34:00	WARRANTS	WARRANT ARREST	Tuesday	NORTHERN	ARREST, BOOKED	DIVISADERO ST / LOMBARD ST	-122.442650	37.798999
2003-01-07 01:22:00	WARRANTS	WARRANT ARREST	Tuesday	SOUTHERN	ARREST, BOOKED	900 Block of MARKET ST	-122.409537	37.782691
2003-01-06 23:30:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	BAYVIEW	ARREST, BOOKED	REVERE AV / INGALLS ST	-122.384557	37.728487
2003-01-06 23:14:00	WARRANTS	WARRANT ARREST	Monday	CENTRAL	ARREST, BOOKED	BUSH ST / HYDE ST	-122.417019	37.789110
2003-01-06 22:45:00	WARRANTS	WARRANT ARREST	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:45:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421
2003-01-06 22:19:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	GEARY ST / POLK ST	-122.419740	37.785893
2003-01-06 21:54:00	WARRANTS	ENROUTE TO OUTSIDE JURISDICTION	Monday	NORTHERN	ARREST, BOOKED	SUTTER ST / POLK ST	-122.420120	37.787757

Data fields

- Dates - timestamp of the crime incident
- Category - category of the crime incident (only in train.csv). This is the target variable you are going to predict.
- Descript - detailed description of the crime incident (only in train.csv)
- DayOfWeek - the day of the week
- PdDistrict - name of the Police Department District
- Resolution - how the crime incident was resolved (only in train.csv)
- Address - the approximate street address of the crime incident
- X - Longitude
- Y - Latitude

We will use X,Y and Dates to predict crime, but we also use another columns to extract features which will help predict a category of crime more accuratly. We will extract year,Day, Hour, Minute from Dates column.

2.2 Data cleaning

There weren't missing values in dataset, but some problems were occurred.

First, train dataset had outliers in Y coordinate. It contained 67 values where Y =90. It's outside San Francisco. I decided to replace these samples by the average coordinates of the district they belong.

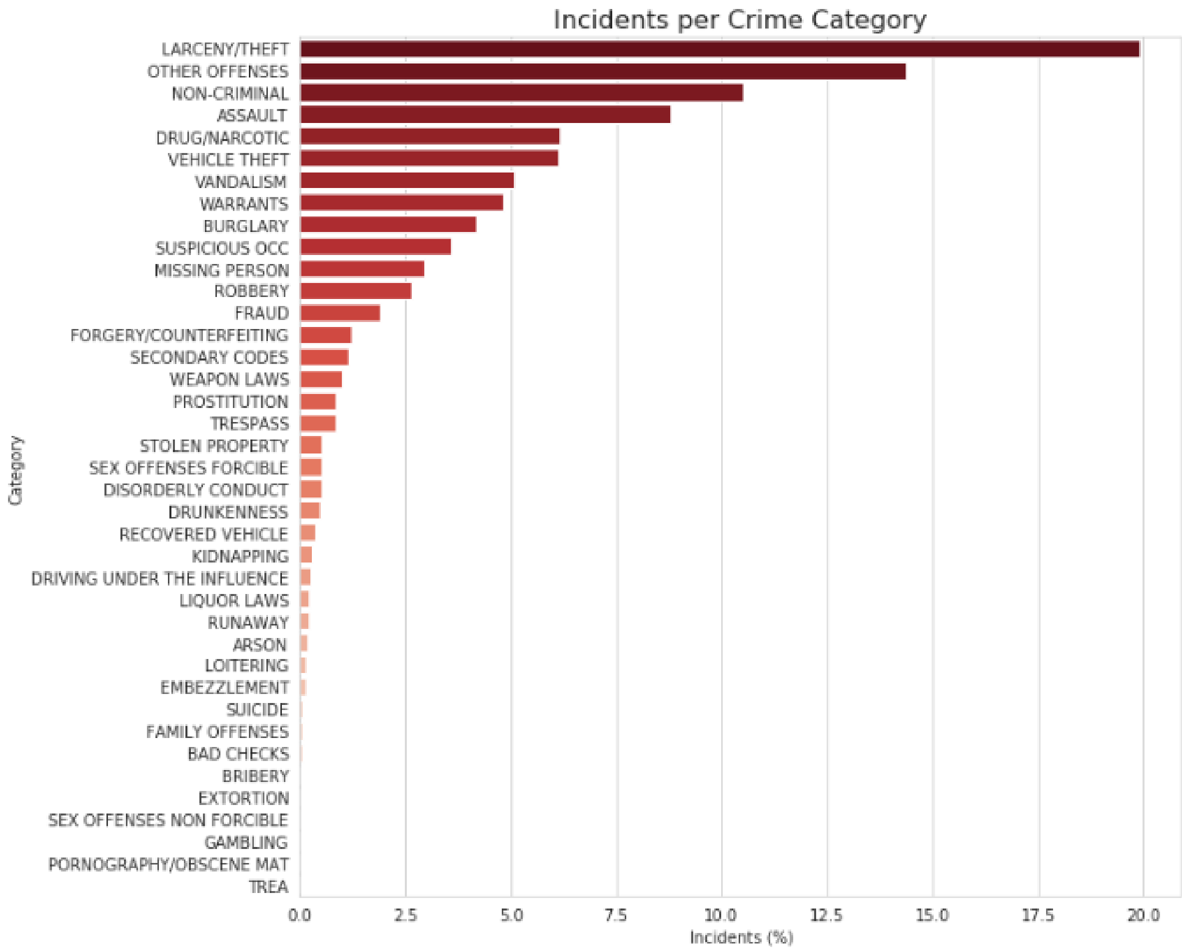
Second, dataset contained 2323 duplicated samples. I decided to drop them, because they didn't give any extra information.

After fixing these problems, I started analysis.

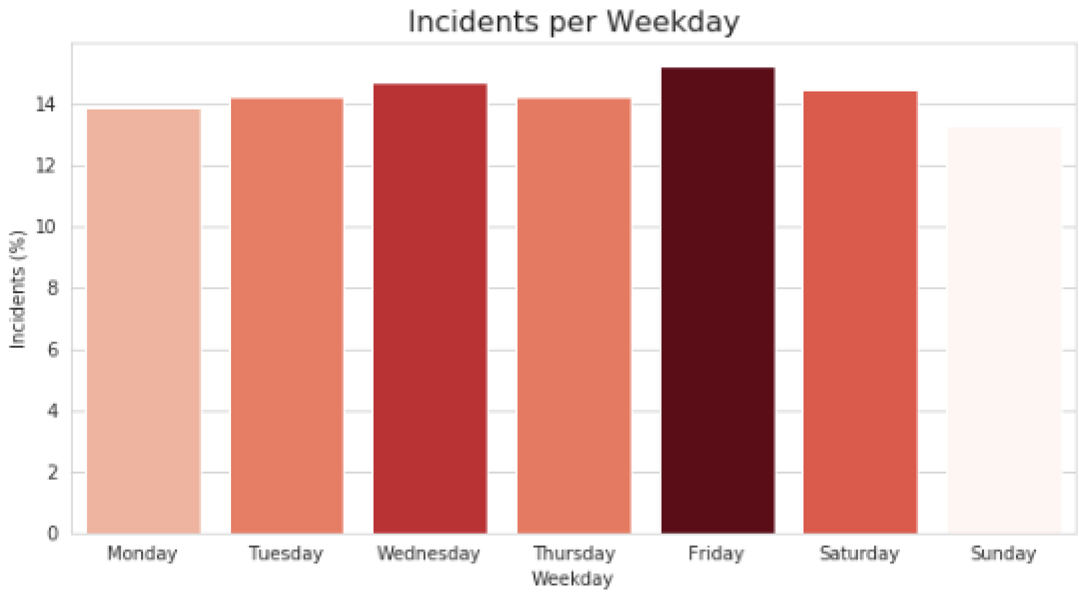
3. Exploratory data analysis

3.1. Incidents per Crime Category

There are 39 categories of crimes. LARCENY/THEFT is most common, it's about 20% of all crimes. OTHER OFFENSES is untypical category, because it isn't known what really had happened. NON-CRIMINAL is about 10,5% of all crimes and ASSAULT is about 9%. In most crimes, probably nobody died or was injured, because as we can see the most of crimes are theft.

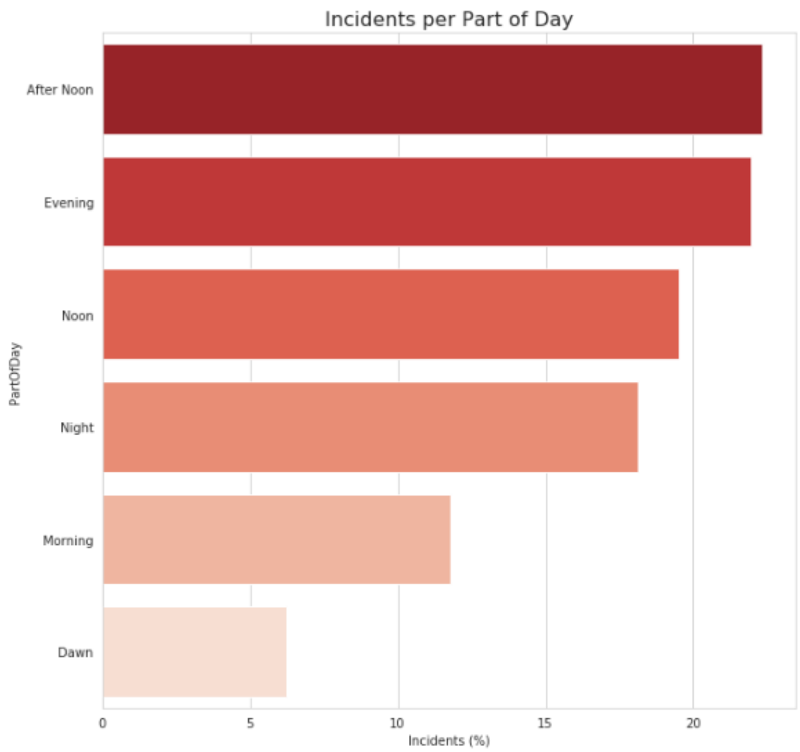


3.2 Incidents per Weekday



As we can see each day of weekday has similar incidents. The most are on Friday, but it is understandable, because this day people usually drink a lot of alcohol and partying. It's good time for thieves, because drunk people are less unconscious. Generally on Friday night there are more people on the streets. I carried out statistical analysis to examine if these weekdays are statistically equal. It was Lilliefors test performed to examine normality of the distribution of the studied variables. Saturday and Sunday haven't normal distribution. Next step was to examine homogeneity of variance, by Levene's test. At the end Kruskal test was made. Based on this test we reject hypothesis of equality in the studied groups. Doing the posthoc test it turned out that only Tuesday, Thursday, Saturday are statistically equals according number of incidents. But we didn't expect weekday to play a significant role in prediction.

3.3 Incidents per Part of Day



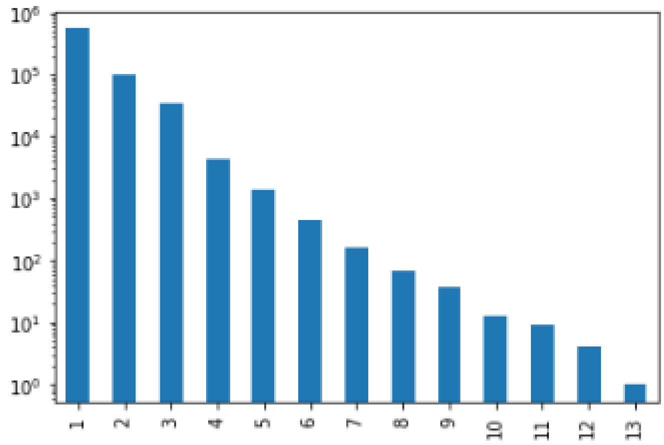
Based on date feature, part of day was extracted.

- 02:00-06:00 is **Dawn**
- 06:00-10:00 is **Morning**
- 10:00-14:00 is **Noon**
- 14:00-18:00 is **After noon**
- 18:00-22:00 is **Evening**
- 22:00-02:00 is **Night**

More than 45% incidents happened between 14:00-22:00 (After Noon - Evening). The fewest incidents were in the Morning and Dawn, because in this time, people usually are in their houses.

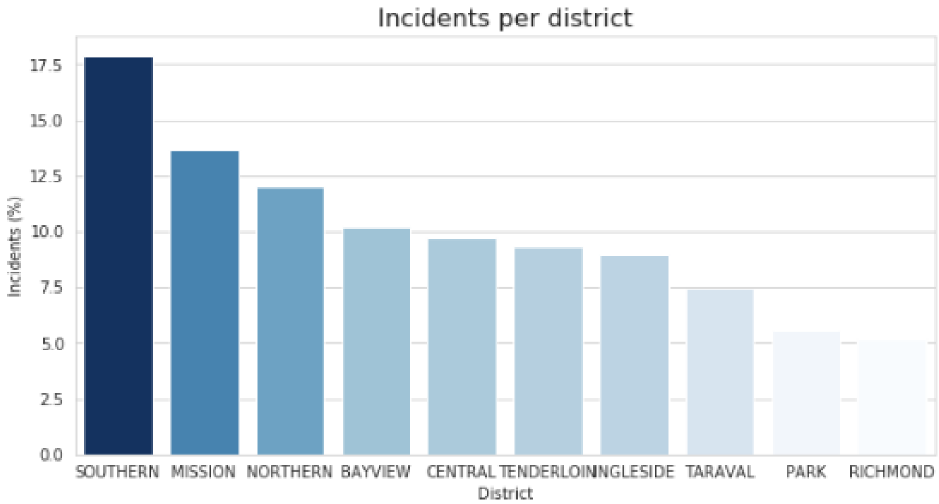
3.4 Multiple crimes?

There are about 300,000 multiply or group crimes. As we can see it occurs not so rarely. We can only assume what is it but we can't be sure based on this data.

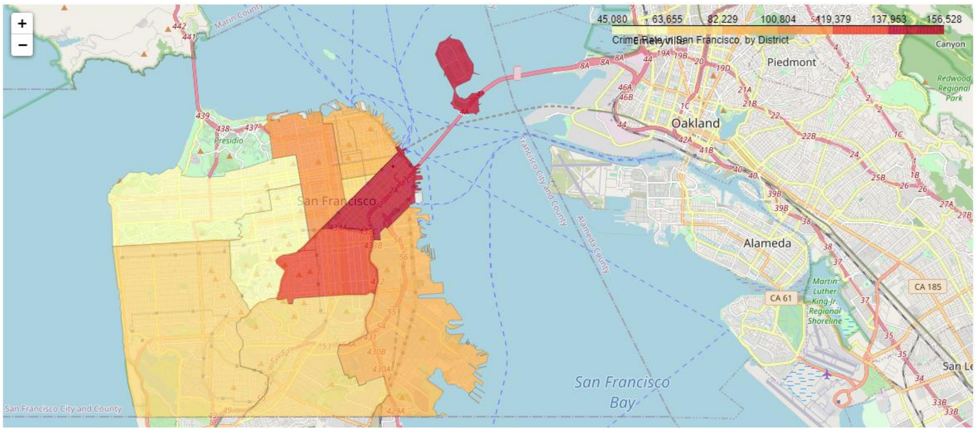


3.5 The most dangerous district

Southern disctrict has about 5% more incidents than second – Mission. There were about 17,5% commited crimes. 3 most dangerous disctricts are in neighborhood.

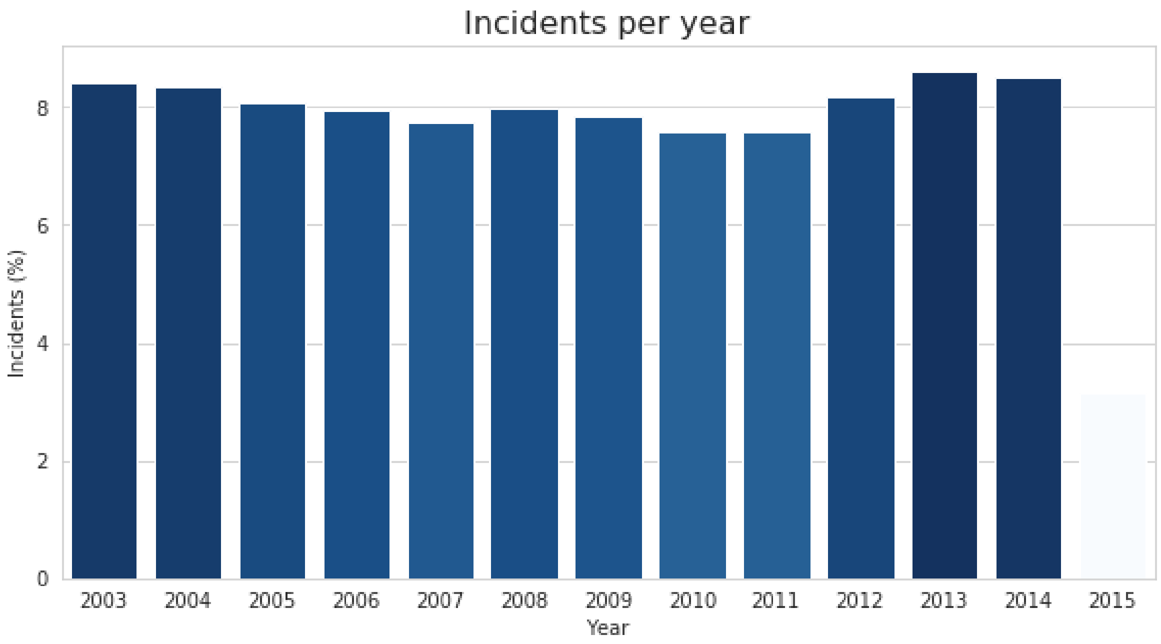


Here is map of San Francisco,



3.6 Number of crimes per year

Each year number of crimes was similar. 2015 is different, because of data ranges 1/1/2013 to 5/13/2015. You can see that crime has not decreased over the years.



4. Data Preprocessing

- 2323 duplicate values and 67 rows with wrong latitues were removed earlier
- 'year' and 'PartOfDay' features were extracted

4.1 Extracting another features

- Month, Day, Hour and Minute from 'Dates' field were extracted
- we created bolean feature “Block” if crime has taken place on a building block or not

4.2 Feature scaling

I used LabelEncoders to change string features to numerical, like PdDistrict, DayofWeek, PartOfDay.

Next step was scaling features using StandardScaler. Some classification models need it, because they could be work unproperly.

5. Modeling

I used classification models to predict category of crime. I applied Random Decision Forest, k-Nearest Neighbors and Logistic Regression. The results all had the same problems. Accuracy of prediction was very low. These results are not acceptable, but having such features we can't do too much.

Random Decision Forest performed the best (~30% accuracy). kNN and Logistic regression performed about 25%. I tuned each model to its best accuracy.

	Logistic Regression	Random Decision Forest	k-Nearest Neighbors
Accuracy(%)	25,3	30,8	24,9

6. Conlusions

In this study, I tried to predict category of crime based on given data. I identified new features and which feature is the most important. Classification models gave low accuracy, but it is hard to predict category based on such data. This can be useful for police because we categories of crime for given place. Summing up not every time data are good for some predictions.

7. Future directions

We could use better, more complicated models to predict category of crime e.g. neural networks. We could try to remove some features, perhaps they cause some noise.