San Francisco's Opioid Crisis and Drug Problem and Effects on Public Safety

Kartikeya Shukla (ks5173)
Apurva Bhargava (ab8687)

Problem and Motivation

The War on Drugs is a phase used to refer to a government-led initiative that aims to stop illegal drug use, distribution, and trade by increasing and enforcing penalties for offenders. The movement started in the 1970s and is still evolving today. Consequently, numerous US states are experiencing an opioid crisis in recent times. There is an ongoing debate, the opioid crisis is the product of Mexican and Central American migration - rather than the deregulation of Big Pharma and the failures of a private health care system. Consequently, at this instance, San Francisco is facing a major drug problem and opioid crisis. San Francisco (SF) has a long history of pushing the envelope on progressive public health solutions, including medical cannabis and needle exchange, before either was legal or broadly embraced. It is so out of proportion, that California passed a bill allowing SF to open Safe Injection Sites (SIS).

Safe injection sites (SIS):
Safe injection sites are medically supervised facilities designed to provide a hygienic and stress-free environment in which individuals are able to consume illicit recreational drugs intravenously and reduce nuisance from public drug use. They are part of a harm reduction approach towards drug problems. North America's first SIS site opened in the Downtown Eastside (DTES) neighborhood of Vancouver in 2003.

Primary Goals

- Correlation between types of crime and neighborhoods from 2003 to 2018
- Correlation between opioid trends and neighborhoods from 2003 to 2018.
- Prediction of the type/category of crime based on spatial and temporal features provided.

Potential Questions Answered

- Identification of potential neighborhoods for installing **SIS (Safe Injection Sites)** for San Francisco's Government.
- Comparison of opioid trends across different neighborhoods— finding top 5 neighborhoods where meth use is most prevalent etc. by analysis of narcotics related crimes.

- Comparison of types of crimes across different neighborhoods, for example, what are the top 5 neighborhoods with high possibilities of an assault?
- Co-occurrence of certain types of crimes

Data

The data is collected from the San Francisco police department's database. It is historical data regarding crimes from Jan 2003 to May 2018. The dataset has 13 columns and 2215024 rows. The attributes and descriptions are given in Table 1.

Table 1: San Francisco Crime Dataset

Column name	Definition	Type	Scale
IncidntNum	Incident Number: The number issued on the report, sometimes interchangeably referred to as the Case Number	long integer	Continuous
Category	Incident Category: A category mapped on to the Incident Number used in statistics and reporting. Mappings provided by the Crime Analysis Unit of the Police Department.	string/text	Categorical
Description	Incident Description: The description of the incident that corresponds with the Incident Number. These are generally self-explanatory.	string/text	Continuous
DayofWeek	The day of the week the incident occurred	string	Categorical
Date	The date the incident occurred	DateTime	Continuous
Time	The time the incident occurred	DateTime	Continuous
PdDistrict	The Police District reflecting current boundaries (boundaries changed in 2015). These are entered by officers and not based on the point. One can refer to them as "county" names	string/text	Categorical

Resolution	The resolution of the incident at the time of the report. Types: - Cite or Arrest Adult - Cite or Arrest Juvenile - Exceptional Adult - Exceptional Juvenile - Open or Active - Unfounded Note: once a report is filed the resolution does not change on the filed report later. Updates to a case will be issued later as Supplemental reports if there’s a status change.	string/text	Categorical
Address	Incident Address: One or more street names that intersect closest to the original incident separated by a forward slash (\)	string/text	Continuous
X	The longitude coordinate in WGS84, the spatial reference is EPSG: 4326	longitude	Continuous
Y	The latitude coordinate in WGS84, the spatial reference is EPSG: 4326	latitude	Continuous
Location	The point geometry used for mapping features in the open data portal platform. Latitude and Longitude are provided separately as well as a convenience	Point type object	Continuous
PdId	Precinct ID at which precinct was the incident reported	Long integer	Categorical

Methodology

- Perform Data profiling using frequentist statistics, and detect outliers. For example, EDA/Visualizations, null analysis. Semantic profiling to identify homogeneous columns— to eliminate extraneous features
- Create cluster-maps between crime type/category and neighborhoods— perform data normalization/standardization as necessary. Cluster-maps (i.e. unsupervised learning, will help us to find the correlation between different neighborhoods and type of crime)

- Prediction using XGBoost, CatBoost, Naive Bayes and Random Forest classifier with the response/target variable as the category/type of crime, and predictors as the spatial-temporal columns. Hyper-parameter tuning using k-folds cross-validation
- The reason for using the above the above algorithms because we have a classification task at hand, and the above algorithms are pretty standard for such tasks
- Pre-process data to filter out crimes that involved Drugs/Narcotics. Perform Step 1. on this subset again. Perform aggregations as necessary to get granular information i.e. Narcotics based crimes categorized by types of drugs i.e. opioids, marijuana, etc
- Create cluster-maps between different types of drugs and neighborhoods. Normalize/standardize as required

Analysis Approach

1. We counted the occurrences, for each category of crime and plotted it. Since the distribution was skewed, we normalized it by taking the log. Figure 1 shows the normalized crime category distribution.

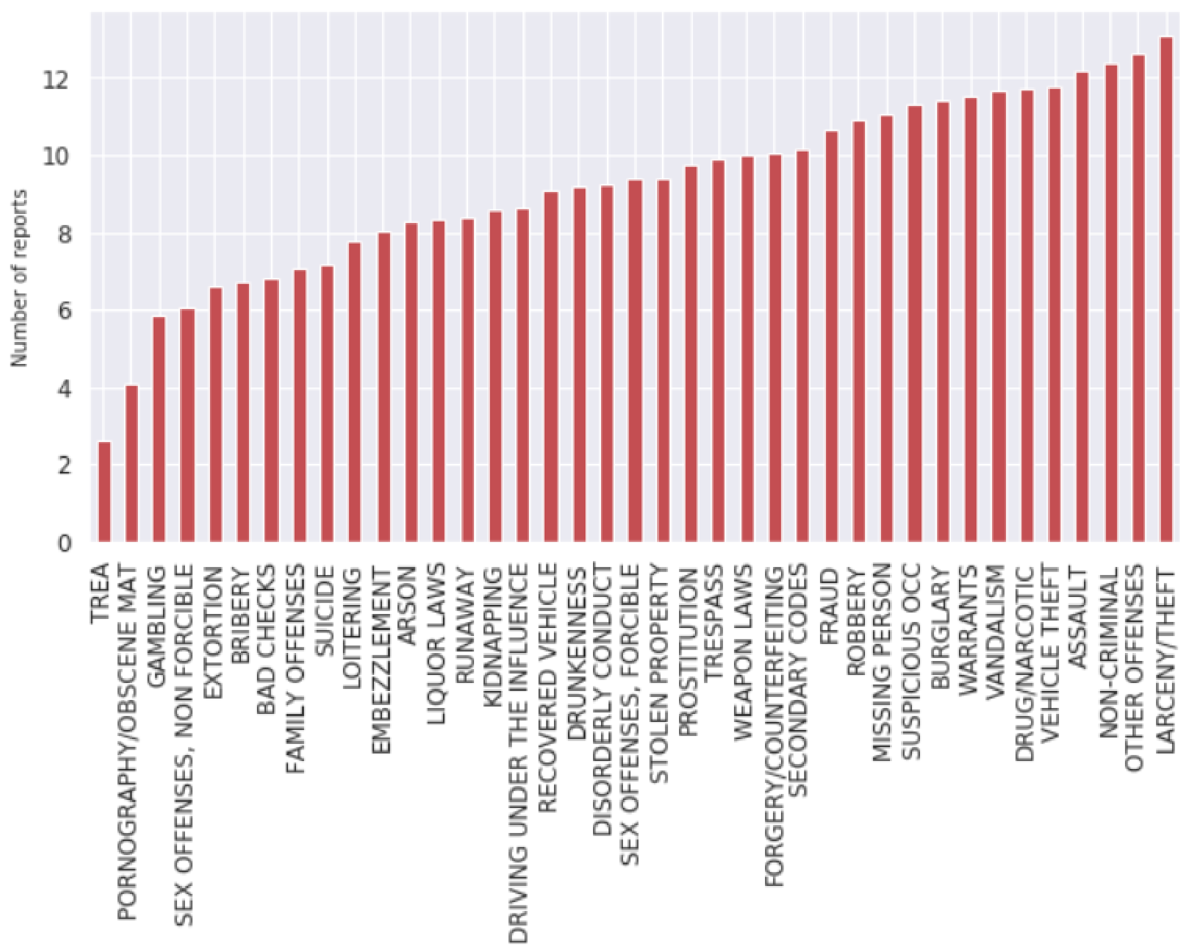

Figure 1

2. There were 915 distinct crime descriptions, and the descriptions determine whether the crime was narcotics related or not. So, we counted occurrences for each crime description and filtered those that were below 97th percentile and kept the rest for creating the cluster maps.

3. We created a cluster-map, to explore the distributions of different types (i.e. of crimes across each PdDistrict (i.e., Police District)). Again, since this distribution was skewed, it affected our model (Figure 2). One can observe that Grand Theft Auto is an outlier besides that we gain no information, thus normalization was required.

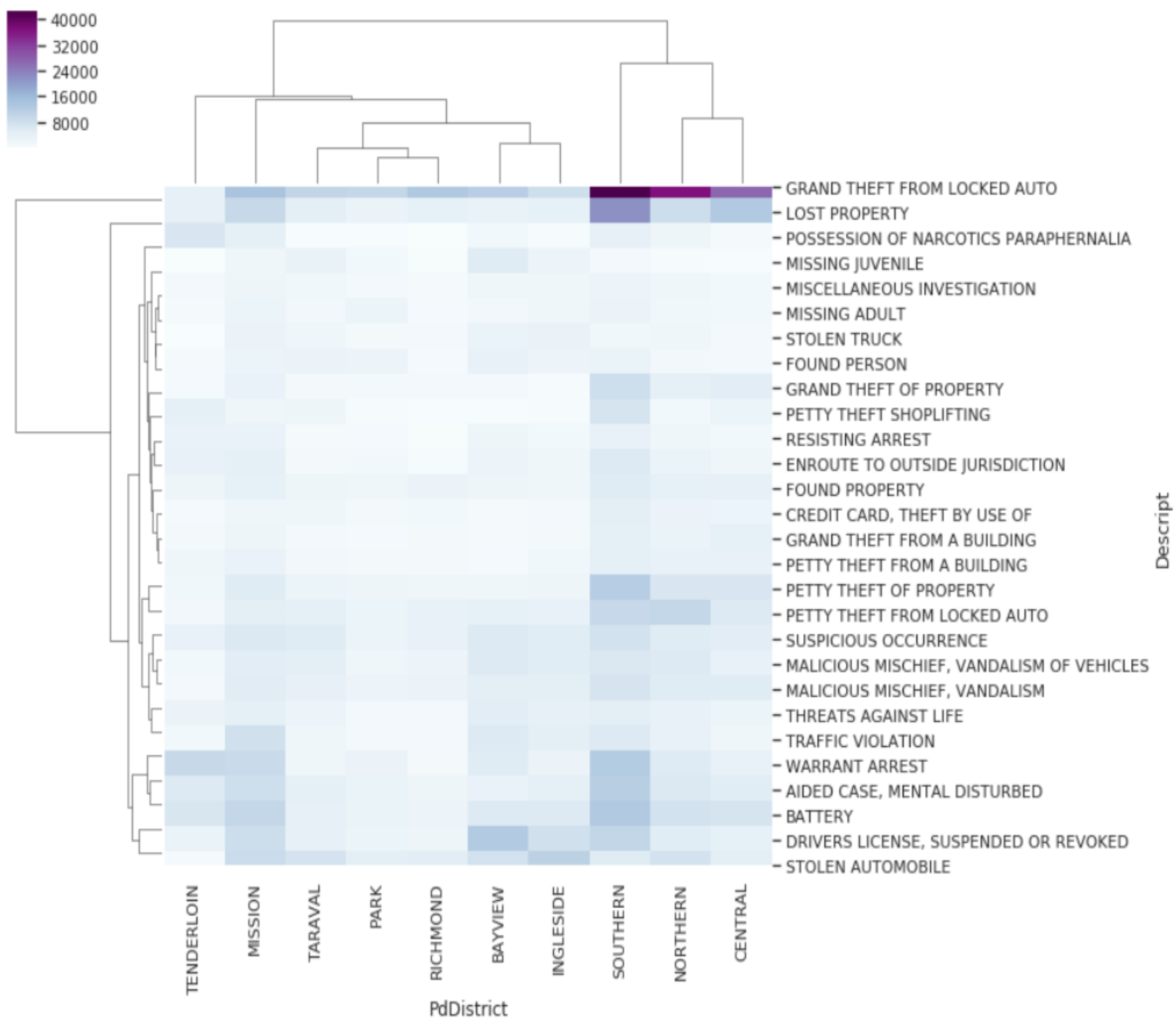

Figure 2

4. Thus normalization was performed using **min-max normalization**, since taking the log does not retain the scale as to how large/small be one feature compared to another. Figure 3 shows the normalized cluster map.

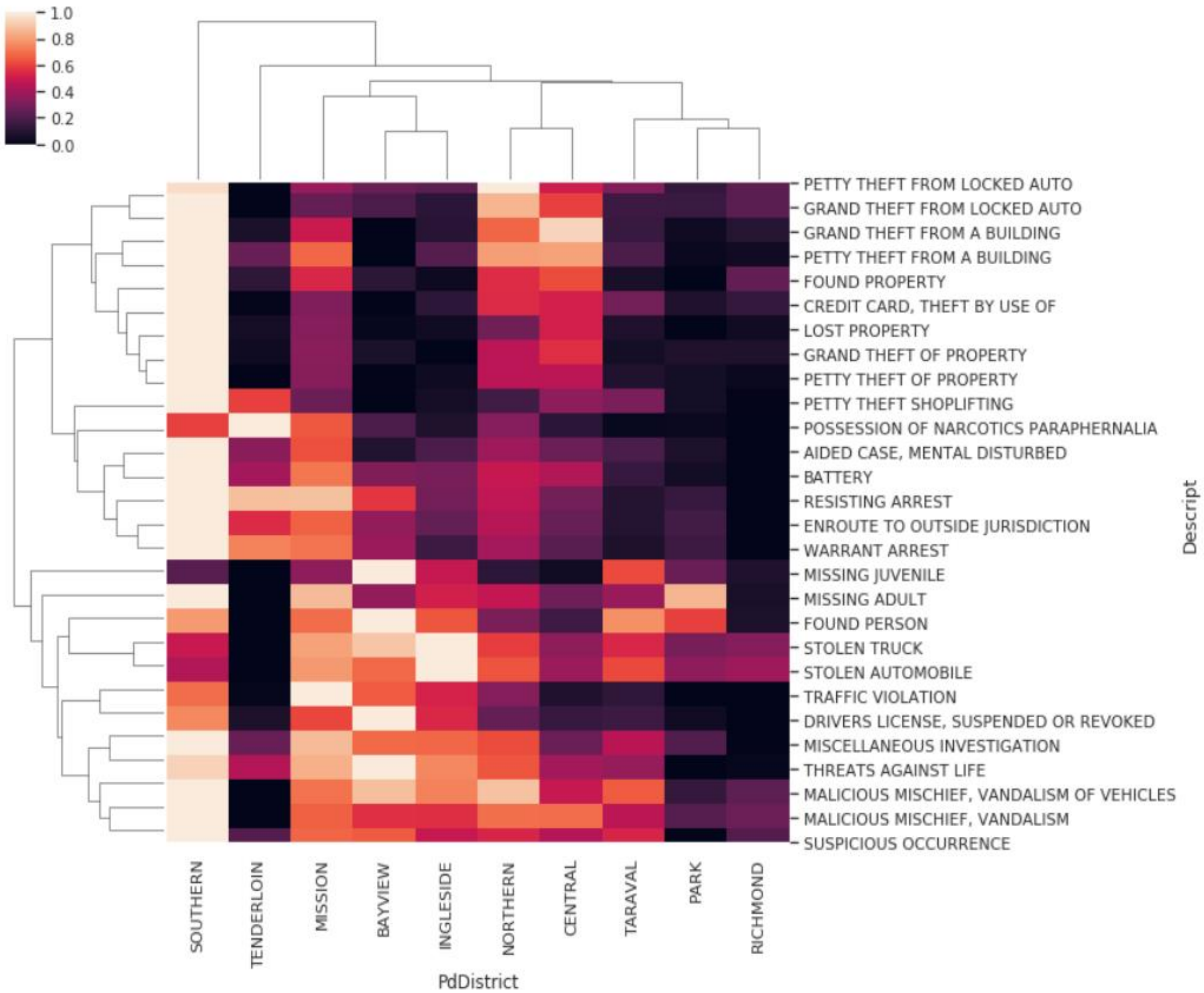

Figure 3

From figure 3, following observations are made-

- Southern: extremely high occurrences of theft, including theft from auto
- Bayview: significant occurrences of violence and threats
- Tenderloin: seems to be an outlier, with exceedingly high occurrences of possession of narcotics paraphernalia. Tenderloin, seems like a potential candidate to install SIS, although it could be a false positive (i.e. these could be due marijuana). Thus one needs to delve deeper.

5. Next, we filtered narcotics related crime using some regular expressions and string pattern matching, and counted occurrences of each distinct narcotics related description. Again the distribution was skewed, and it affected the cluster-map, as shown in Figure 4. One can observe Tenderloin is an outlier and we gain no other information.

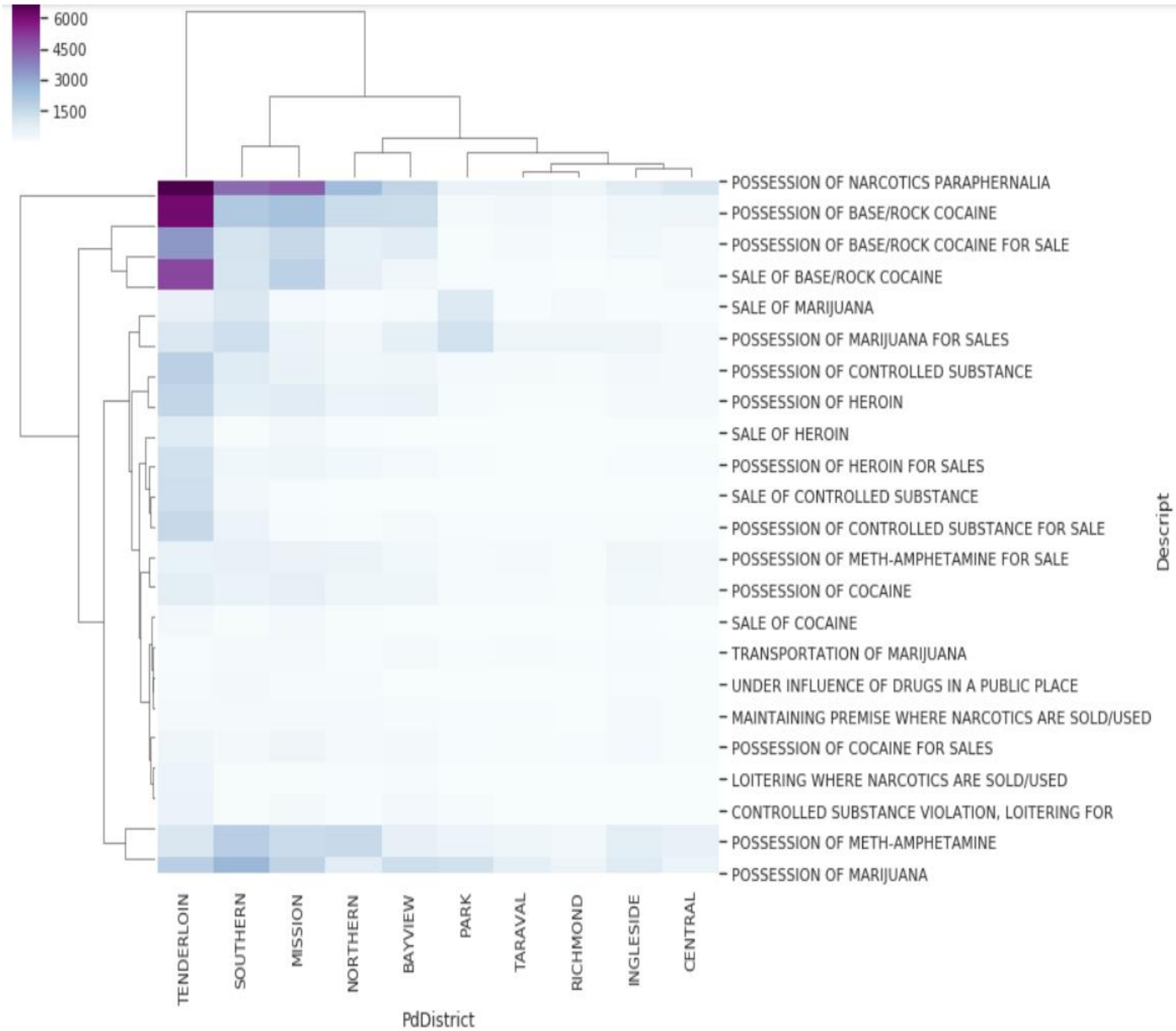

Figure 4

6. Figure 5 is normalized cluster-map which shows distribution of narcotics related crimes across each PdDistrict. From the cluster-map, we can clearly conclude that, Tenderloin, Southern, Mission and Northern are the optimal candidates for installing SIS.

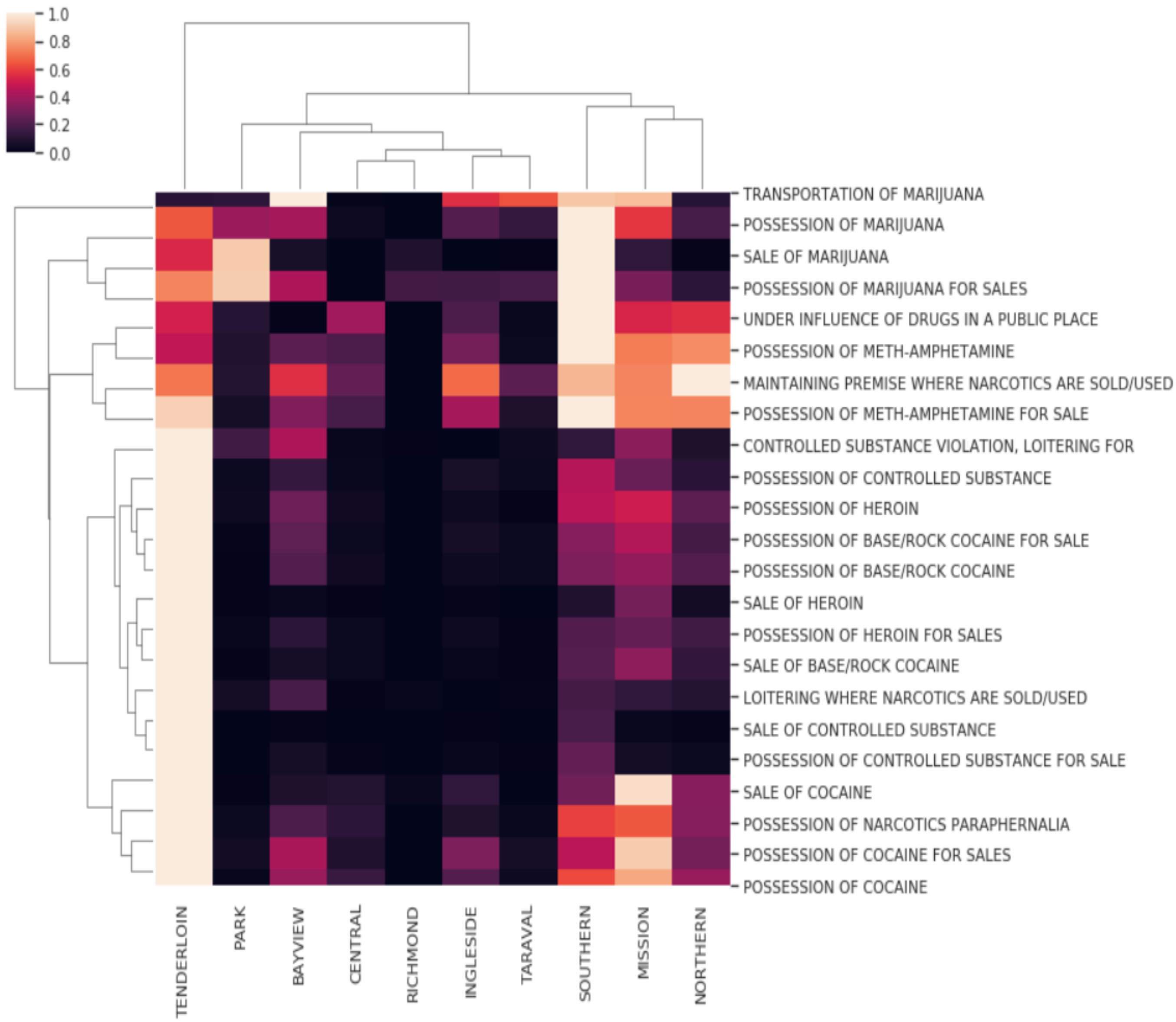

Figure 5

7. Thereafter time-series analysis was performed, to analyze opioid trends across time. First we compressed numerous narcotics related crime descriptions to create opioid groups/features (i.e. Barbiturate features, coke features, marijuana features, meth features, etc.). Then we created a 30 days window for each group, and counted the no. of occurrences for each group across this 30 day window (i.e. each month) for each month from 01/01/2003 to 05/15/2018. To remove cyclic features of the months–we indexed them from 0 to 187. Figure 6 shows a stacked histogram, to represent these trends. As can be seen, meth and heroin related incidents significantly went up.

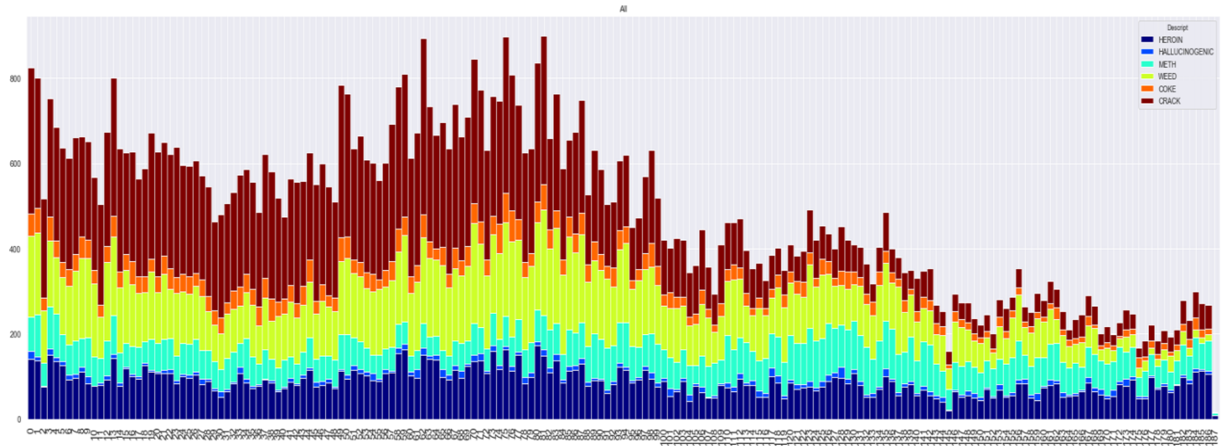

Figure 6

8. To make the trends clearer, the normalized distribution of opioid trends across the years from 2003 to 2018 is shown in Figure 7.

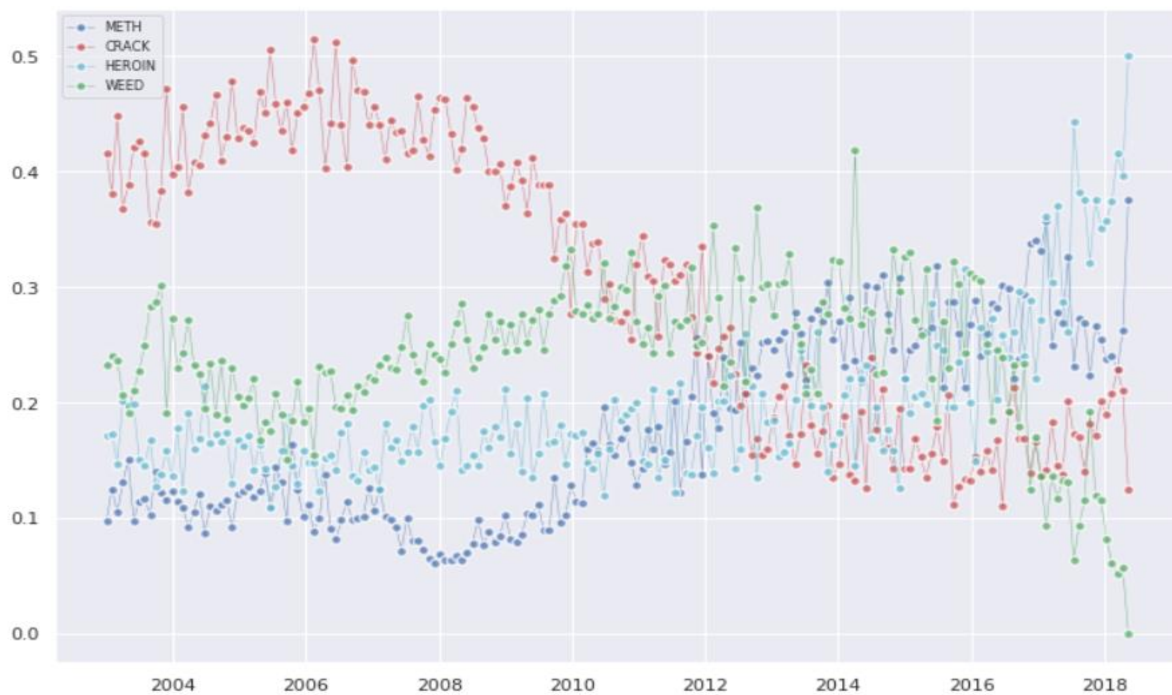

Figure 7

One can observe that crack related incidents went down over this period. Similarly, marijuana related incidents went down after it was legalized in 2016. But meth and heroin related crimes significantly shot up--this is substantial evidence to conclude it is an epidemic.

Model Selection Approach

We started off with a Binary logistic regression model, which predicted the likelihood of whether the crime was narcotics related or not, given certain geo-coordinates. This will help SF's government to allocate resources to certain areas based on the prediction. Our initial accuracy was 94% which was "suspiciously" high. Thus we checked for bias-- and we found that our target class was imbalanced--i.e. there were far more non-narcotics related crime as compared to narcotics related. Thus we oversampled our target class using SMOTE. After this our accuracy was 77%, which made sense, although AUC went up from 0.68786 to 0.69875.

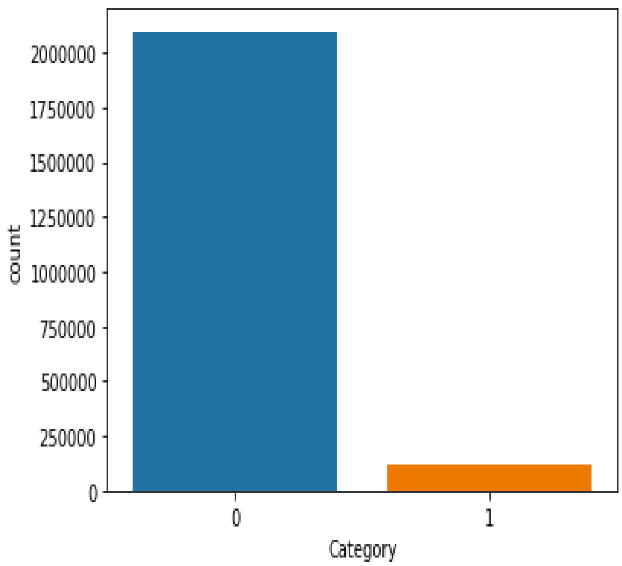

Figure 8: Target class distribution before oversampling

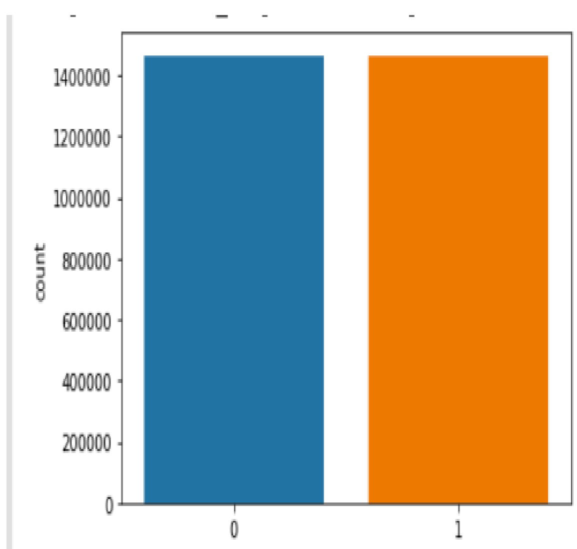

Figure 9: Target class distribution after oversampling

This imbalanced distribution can be solved by either undersampling or oversampling. The reason oversampling is chosen is because although undersampling might work, many useful data is lost during the process. These data are meaningful to the logistic regression algorithm.

The oversampling technique is SMOTE, which stands for synthetic minority oversampling technique. There are two reasons. First of all, SMOTE is a very common oversampling technique. Secondly, when oversampling a minority class in an imbalanced dataset, what could happen is the model end up learning too much of the specifics of the few examples, usually with a simple approach like randomly adding minority data. SMOTE on the other hand learns the property of the neighborhood of minority data points. This way, the model can generalize better

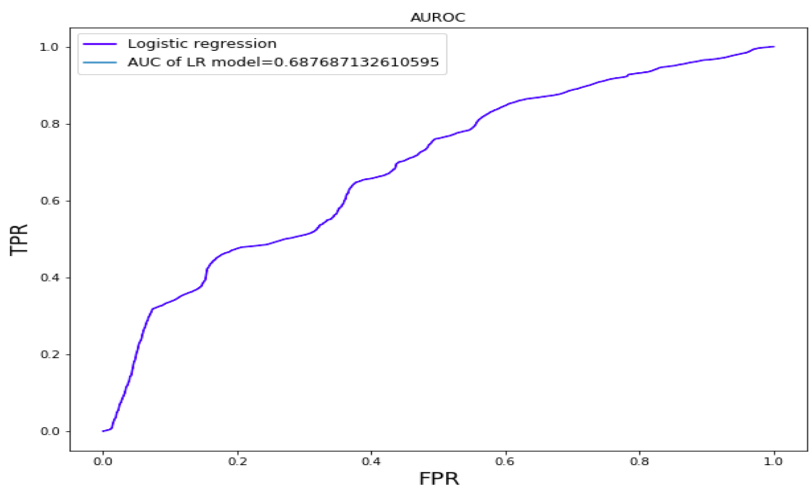

Figure 10: AUC before oversampling

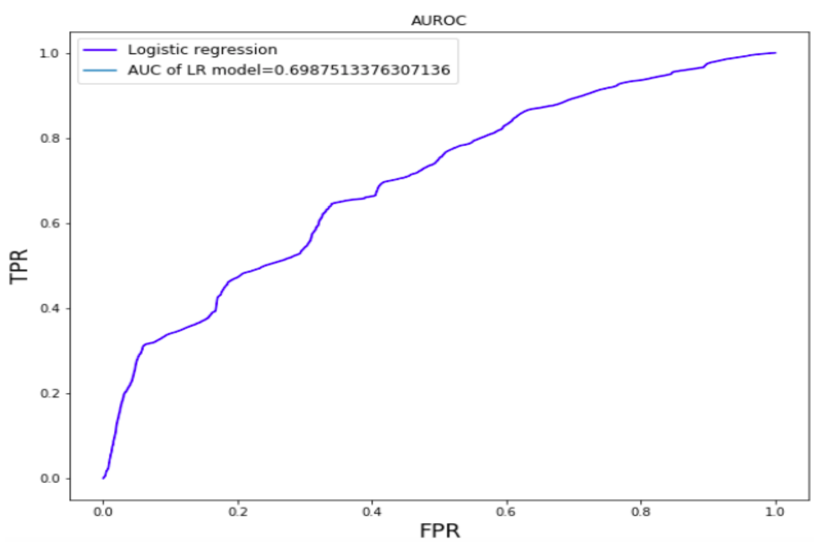

Figure 11: AUC after SMOTE

Inference: We realized that a binary logistic regression (i.e., Binary LR) wasn't adequate enough for our problem. For instance, if we used these predictions, to allocate Government resources, then we could have a very high false positive rate. Since just finding geo-coordinates/areas where non-narcotics/narcotics related crime is high is isn't enough, we needed to delve deeper- i.e., we would like to allocate more resources where there is a high rate of murder than a high rate of arson. Thus we switched to other multi-class models, such as **Multinomial Logistic Regression (Multinomial LR)**, **XGBoost**, **KNN** and **Random Forest.**

Results (Multinomial Logistic Regression, XGBoost, KNN and Random Forest)

The results of training the data using models based on other classification algorithms are listed in Table 2.

Table 2: Results

Algorithm	Accuracy	Precision	Recall	F1-Score
XGBoost	66.67 %	0.60	0.68	0.65
KNN	61.54 %	0.69	0.64	0.67
Random Forest	58.33 %	0.50	0.62	0.58
Multinomial LR	42.85%	0.33	0.43	0.36

XGBoost performed the best. The above models predict the likelihood for each category of crime, given a geo-coordinate and whether the crime was performed during day or night for a particular PdDistrict. We can use these likelihoods/probabilities and aggregate over the given geo-coordinates, to identify neighborhoods and allocate government resources accordingly.

Assumptions, Limitations and Tradeoffs

- **Assumption:** Since no one really "self"-reports whether they're using opioids or not, the only way to analyze the opioid trends was to look at crime data available from SF's police department. Since it is a "proxy" dataset, it could under-represent or over-represent our result.
- **Limitation:** What did you change from your original proposal and why? We could not implement Apriori/FP growth (i.e. association rule mining) since it requires generating frequent itemsets--and that is very computationally expensive given this huge dataset. We plan on using Spark in the future, although that is outside the scope of this course.

- **Trade Offs:**
  Feature Scaling: we used **min-max normalization** every time we normalized data, although it does not handle outliers well, on the other hand it retains the original scale. In case of log/z-score normalization outliers are handled well, but the original scale is not retained. For our analysis, it was more important to retain the scale. Thus there are always tradeoffs.
  AUC/ROC vs Precision/Recall (i.e. PR): We used PR as our evaluation metric. Since our target class is imbalanced and we wanted our models to take that into account. In the real world, the PR curve is used more since positive and negative samples are very uneven. The ROC/AUC curve does not reflect the performance of the classifier, but the PR curve can and has better interpretability. Also AUC (Area under ROC) is problematic especially if the data is imbalanced. The positive examples have relatively low rates of occurrence. Using AUC to measure the performance of the classifier, the problem is the increasing of AUC doesn't really reflect a better classifier. It's just the side-effect of too many negative examples. And since we care more about the actual prediction of true-positives rather that the "overall" performance of the model, thus PR and F-1 score seem more appropriate.

References

[1] https://www.kqed.org/news/11766169/san-francisco-fentanyl-deaths-up-almost-150
[2] https://www.sfchronicle.com/bayarea/article/Bay-Briefing-Fentanyl-epidemic-worsens-in-San-14032040.php
[3] https://www.businessinsider.com/san-franciscos-dirtiest-street-has-a-drug-market-and-piles-of-poop-2018-10
[4] https://www.sfchronicle.com/bayarea/article/California-bill-allowing-San-Francisco-safe-13589277.php
[5] https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry/data
[6] https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783/data
[7] https://data.sfgov.org/d/wkhw-cjsf
[8] https://www.quora.com/What-is-the-meaning-of-min-max-normalization
[9] https://github.com/gandalf1819/SF-Opioid-Crisis