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

Group Members:

Kartikeya Shukla (ks5173)

Apurva Bhargava (ab8687)

Member responsible for uploading submissions:

Apurva Bhargava (ab8687)

Problem Description

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.
- Prediction of the types/categories of crime, based on spatial and temporal features.
- Comparison of types of crimes across different neighborhoods, for example, what are the top 5 neighborhoods with high possibilities of an assault?

Motivation for Problem

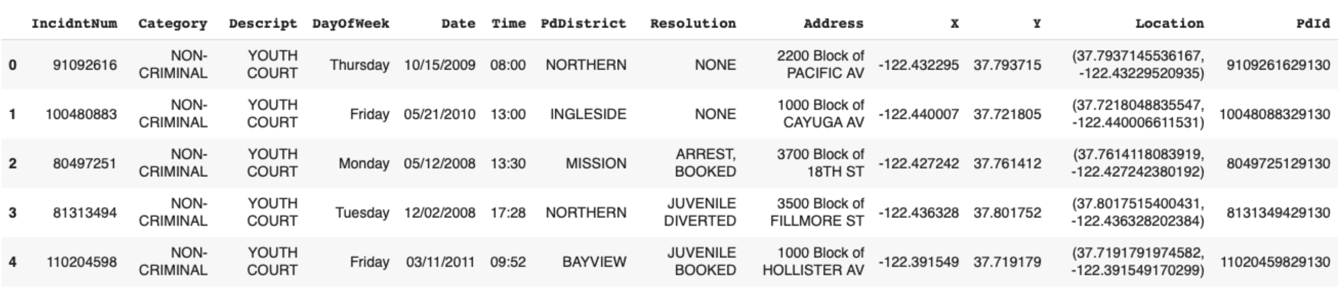
Post the government-led initiative to stop illegal drug use, distribution and trade and as a consequence of deregulation of pharmacies and public healthcare system failures, several states are facing a major drug problem and opioid crisis. In San Francisco, it is so severe that used syringes are lying out open in the streets. San Francisco has a long history of testing the limits of progressive public health solutions, including medical cannabis and needle exchange, before either was legal or broadly embraced. California has passed a bill allowing San Francisco to open 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. The opiate users can come in and consume their opiates, while the staff disposes the needles & other paraphernalia safely. SIS are part of a harm reduction approach towards drug problems. The goal of the project is to help SF's government identify potential areas to install SIS.

Assumption:
No one really self-reports whether they're using opioids or not, thus the only way to analyze these 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 the results.

Description of 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



	IncidntNum	Category	Descript	DayOfWeek	Date	Time	PdDistrict	Resolution	Address	X	Y	Location	PdId
0	91092616	NON-CRIMINAL	YOUTH COURT	Thursday	10/15/2009	08:00	NORTHERN	NONE	2200 Block of PACIFIC AV	-122.432295	37.793715	(37.7937145536167, -122.43229520935)	9109261629130
1	100480883	NON-CRIMINAL	YOUTH COURT	Friday	05/21/2010	13:00	INGLESIDE	NONE	1000 Block of CAYUGA AV	-122.440007	37.721805	(37.7218048835547, -122.440006611531)	10048088329130
2	80497251	NON-CRIMINAL	YOUTH COURT	Monday	05/12/2008	13:30	MISSION	ARREST, BOOKED	3700 Block of 18TH ST	-122.427242	37.761412	(37.7614118083919, -122.427242380192)	8049725129130
3	81313494	NON-CRIMINAL	YOUTH COURT	Tuesday	12/02/2008	17:28	NORTHERN	JUVENILE DIVERTED	3500 Block of FILLMORE ST	-122.436328	37.801752	(37.8017515400431, -122.436328202384)	8131349429130
4	110204598	NON-CRIMINAL	YOUTH COURT	Friday	03/11/2011	09:52	BAYVIEW	JUVENILE BOOKED	1000 Block of HOLLISTER AV	-122.391549	37.719179	(37.7191791974582, -122.391549170299)	11020459829130

Figure 1. Data

Tasks completed

Distribution of categories of crimes from 2003 to 2018

There were 915 different crime descriptions, we counted the number of occurrences associated with each one, and used those which were 90th percentile and above. The result was a skewed distribution (figure 2, left). After normalization by taking the log, the distribution was still skewed (figure 2, right).

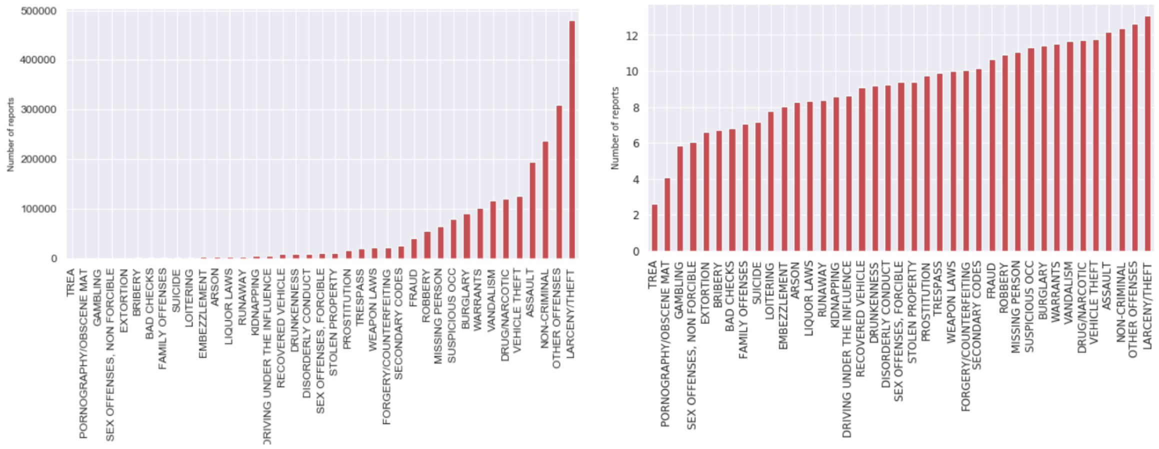


Figure 2

Cluster-maps b/w categories of crimes vs neighborhoods

By building a cluster-map for non-normalized data (figure 3, left), we don't gain any information. Grand theft auto looks like the only outlier. If we normalize by taking the log, we don't retain the exact scale (as in how large exactly is one feature as compared to another), thus we used min-max normalization to create the normalized cluster-map (figure 3, right). Although min-max normalization does not handle outliers well, thus, there are always trade-offs.

$$Value_{min-max-normalized} = \frac{Value_{actual} - minimum}{maximum - minimum}$$

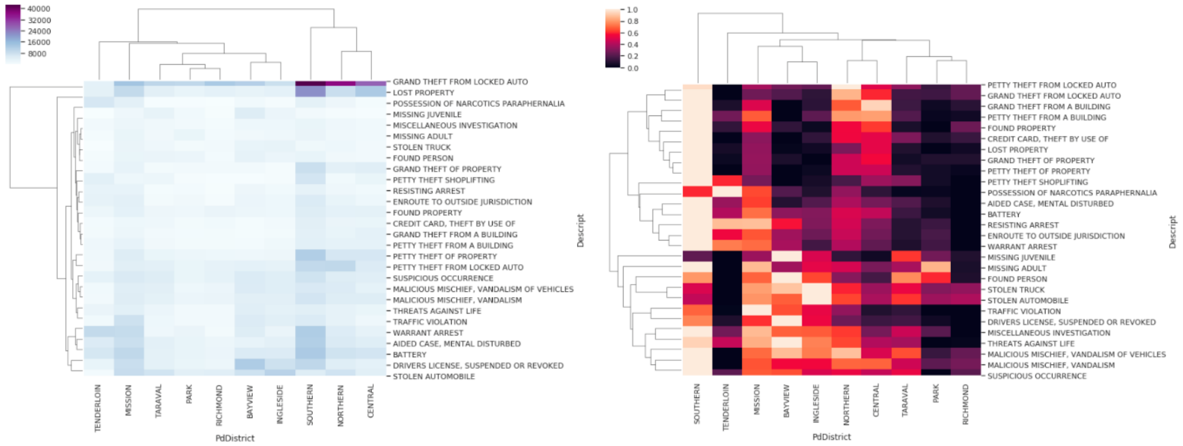


Figure 3

Cluster-maps b/w different opioids & neighborhoods

We performed pre-processing of data to get the categories. We filtered narcotics related crimes & run some regular expressions on the descriptions to get these opioid features. Again, when we built the cluster map given the raw data, we gain no information (figure 4, left). Tenderloin is just an outlier, so we standardized the data across rows (figure 4, right). And we can observe that Tenderloin is a good candidate district where a safe injection site can be installed, other good candidates would be mission, southern & northern districts.

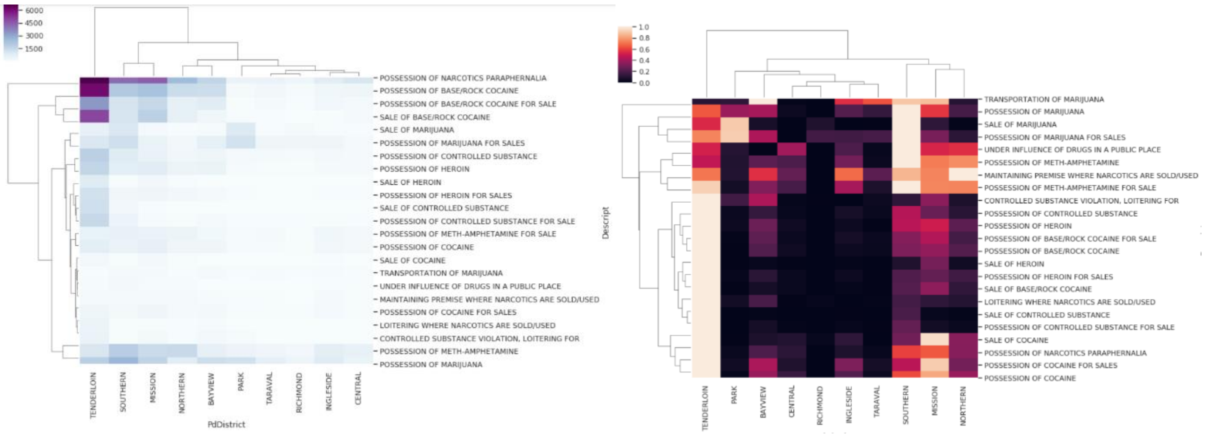


Figure 4

Specific Opioid Distributions across time

We indexed all months from 0 to 180 for each month from 2003 to 2015, & then create a stacked histogram for each month representing the opioids. It can be inferred that the crack incidents went down and meth and heroin related incidents shot up (figure 5).

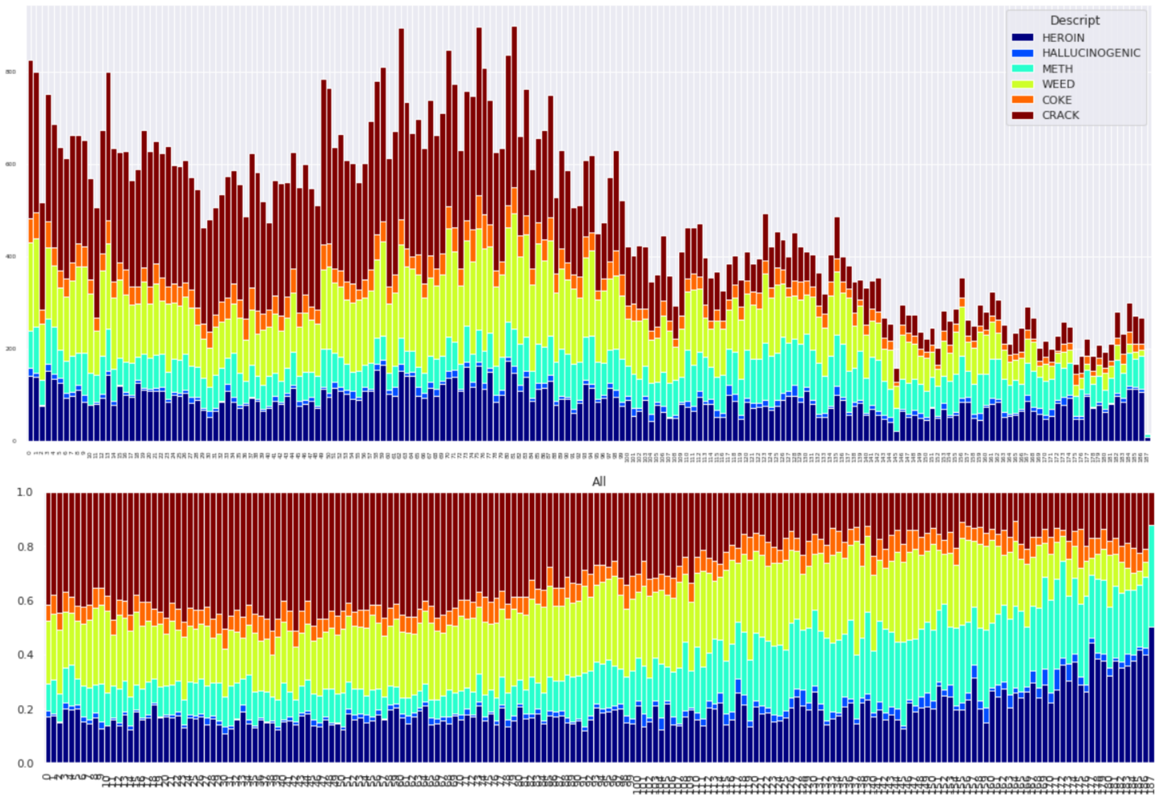


Figure 5

Opioid Distributions across years

To make the trends clearer, we analyzed opioid trends across the years and noticed that crack related crimes gradually went down; marijuana was legalized in 2016 and we can observe that marijuana related crimes went down. But meth and heroin related crimes have grown over the years considerably. From Figure 6, it could be concluded that the crisis is indeed an epidemic.

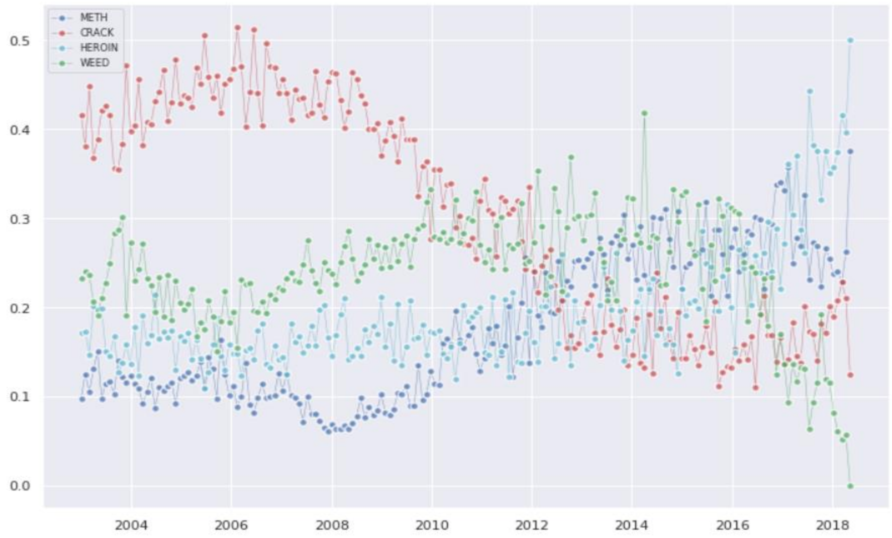


Figure 6

Classification with Logistic Regression

To run logistic regression model, the data was converted from long to wide format by adding extra district columns. The model predicted the likelihood of whether crime was narcotics related or not with 0.94 accuracy (figure 7).

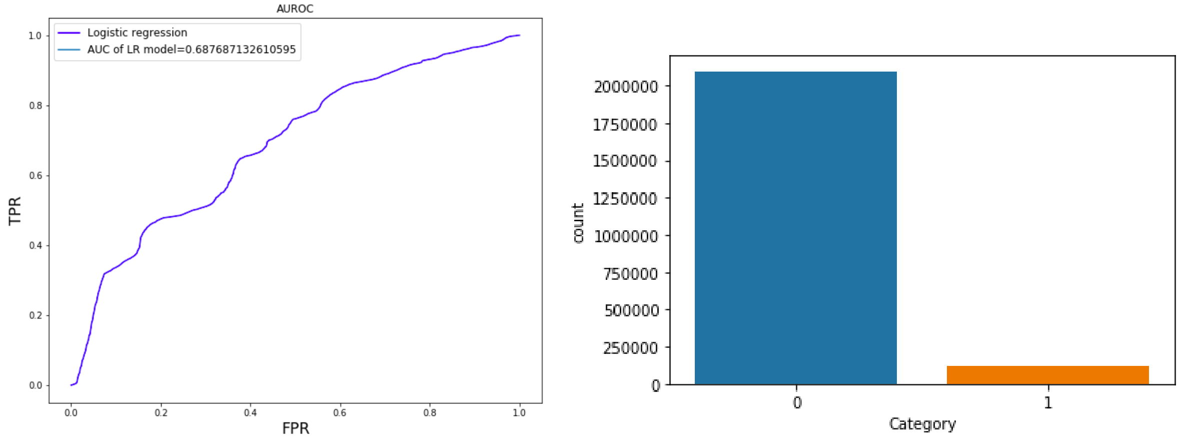
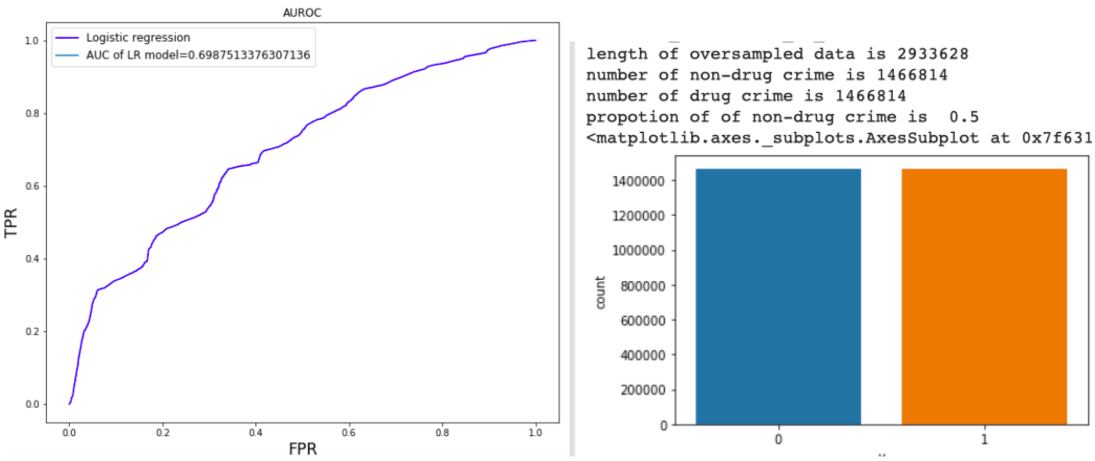


Figure 7

Oversampling SMOTE

It was observed that the high accuracy was due to the fact that the target class was imbalanced, so we used SMOTE to balance class weights. And observed that the accuracy went down to 0.77 (figure 8).

Then we realised that this model was not adequate for our problem since just telling if the crime is narcotics related or non-narcotics related does not provide a lot of information. Based on these predictions we cannot allocate police/government resources accordingly, we'll want to allocate more police patrolling where more murders occurred, or more fire/hazard staff where there is arson or the building are structurally compromised. Thus we came to the conclusion that logistic regression can't be used for a multi-class problem.



```
length of oversampled data is 2933628
number of non-drug crime is 1466814
number of drug crime is 1466814
propotion of of non-drug crime is  0.5
<matplotlib.axes._subplots.AxesSubplot at 0x7f63148ad208>
```

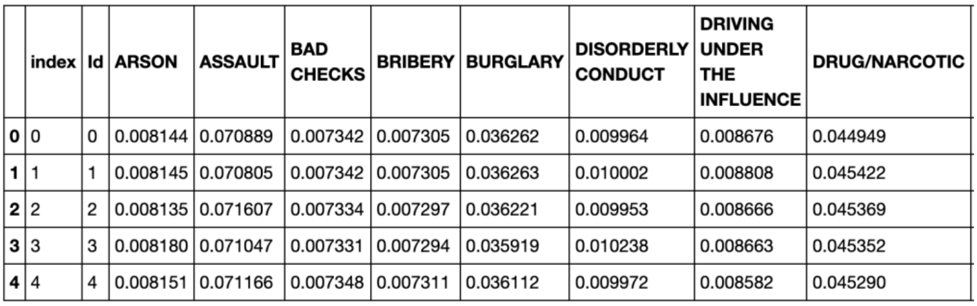
Figure 8

Feature Engineering for XGBoost, Random Forest and KNN

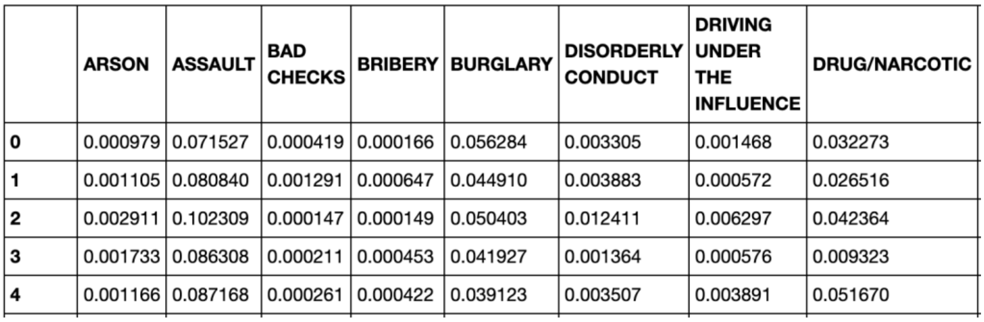
For all the three models, we did some heavy feature engineering; we discretized date & timestamps through binning to create IsDay, DayOfWeek, Month, Hour, Year. We also converted data from long to wide format, thus creating dummy columns for each PdDistrict and indicating whether crime was done during day (0) or night (1).

Predict crime category by day/night/hour across neighborhoods

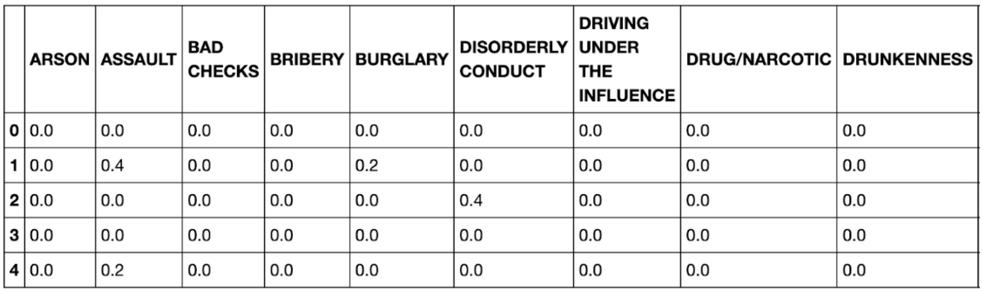
The likelihood for each category of crime given day/night & geo-coordinate was predicted using the models. We performed hyper-parameter tuning for XGBoost using 3-fold cross-validation. From probabilities, we can find the coordinates for the respective IDs and aggregate over neighborhoods to figure out what all resources are to be sent to those areas for different types of crimes (figure 9).



	index	Id	ARSON	ASSAULT	BAD CHECKS	BRIBERY	BURGLARY	DISORDERLY CONDUCT	DRIVING UNDER THE INFLUENCE	DRUG/NARCOTIC
0	0	0	0.008144	0.070889	0.007342	0.007305	0.036262	0.009964	0.008676	0.044949
1	1	1	0.008145	0.070805	0.007342	0.007305	0.036263	0.010002	0.008808	0.045422
2	2	2	0.008135	0.071607	0.007334	0.007297	0.036221	0.009953	0.008666	0.045369
3	3	3	0.008180	0.071047	0.007331	0.007294	0.035919	0.010238	0.008663	0.045352
4	4	4	0.008151	0.071166	0.007348	0.007311	0.036112	0.009972	0.008582	0.045290



	ARSON	ASSAULT	BAD CHECKS	BRIBERY	BURGLARY	DISORDERLY CONDUCT	DRIVING UNDER THE INFLUENCE	DRUG/NARCOTIC
0	0.000979	0.071527	0.000419	0.000166	0.056284	0.003305	0.001468	0.032273
1	0.001105	0.080840	0.001291	0.000647	0.044910	0.003883	0.000572	0.026516
2	0.002911	0.102309	0.000147	0.000149	0.050403	0.012411	0.006297	0.042364
3	0.001733	0.086308	0.000211	0.000453	0.041927	0.001364	0.000576	0.009323
4	0.001166	0.087168	0.000261	0.000422	0.039123	0.003507	0.003891	0.051670



	ARSON	ASSAULT	BAD CHECKS	BRIBERY	BURGLARY	DISORDERLY CONDUCT	DRIVING UNDER THE INFLUENCE	DRUG/NARCOTIC	DRUNKENNESS
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.4	0.0	0.0	0.2	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 9. XGBoost, Random Forest and KNN results

Modules used

- Data manipulation and exploration: pandas, numpy
- Visualisation: matplotlib, seaborn
- Model building: sklearn, xgboost

Tasks pending

- Association rule mining to evaluate cluster quality and the correlation of those results with our clusters.
- Comparison of our models across different metrics such as accuracy, precision, recall, etc.
- Hyperparameter tuning as required.

Links to Dataset

[1] https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports -Historical-2003/tmnf-yvry/data

[2] https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783/data

[3] https://data.sfgov.org/d/wkhw-cjsf

References

[1] https://www.kqed.org/news/11766169/san-francisco-fentanyl-deaths-upalmost-150

[2] https://www.sfchronicle.com/bayarea/article/Bay-Briefing-Fentanylepidemic-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