Crime Type Prediction

I. Introduction

For centuries, crime has been considered random because it is based on human behavior; even now, it encompasses too many variables for current machine learning models to accurately forecast. While minor success has been achieved by increasing police presence in at risk neighborhoods and patrols during certain hours, how we combat crime has gone largely unchanged. The last decade has seen many advances in both computational power and machine learning models, which, combined with exponentially growing datasets, are changing industries. In this study, we discuss the early results of a crime prediction model developed using the data from the San Diego County Sheriffs Office. First, we analyze how certain features like geographical location, population size, time of day, zone type (residential, commercial, etc), distance to nearest streetlamp, and neighborhood correlate to types of crime. Second, we discuss several data mining techniques we used to find meaning in our data, such as K-means clustering and predictive models like support vector machines. Finally, we select the best model for predicting type and severity of crime given multiple features.

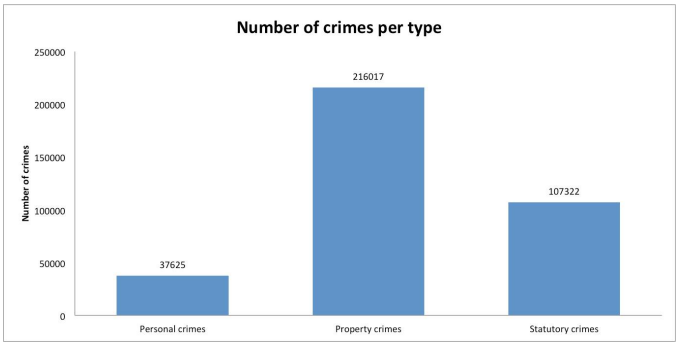
II. Exploratory Analysis

Our dataset came from the San Diego Sheriff's Office [3] and represents crimes committed in San Diego County from 2008 to 2012. This data set contains over 790,000 crimes organized into 14 distinct crime types. See appendix A for the complete listing. The data was collected by police officers in the field after each crime was committed. The data is formatted in CSV files, in which crimes are described by type, location, date and time, streetlamp distance in centimeters, whether or not it was committed during night, ASR zone (zone code where the crime was committed, e.g. industrial, residential, commercial), and community, amongst other attributes. An example crime report with all its attributes can be seen below.

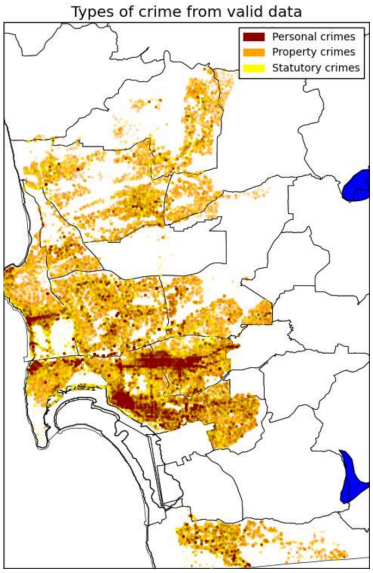
Field	Example	Field	Example
gctype	cns/segment	community	SanDOW
month	8	date	2011-08-10
is night	0	council	San003
year	2011	id	NONE
city	SndSAN	asr zone	6
lon	-117.148	comm pop	31759
type	DRUGS/ALC	week	32
gcquality	65	address	400 Block 17th
lat	32.7099	coun pop	147116
day	4239	desc	PUBLIC INTOXICATION
hour	11	segment$_i d$	3143
time	11:41:00	lampdist	1000
nbrhood	SanEAS	dow	3

The data includes the type of crime committed and we further categorized these into three distinct groups of severity. To group these crimes into three categories we used a common distinction between the nature of the crime. First, personal crimes are offenses against the person. These are crimes that result in physical or mental harm to another person and are viewed as the highest severity. Second, property crimes are offenses against property. These are crimes that do not necessarily involve harm to another person. Instead, they involve an interference with another persons right to use or enjoy their property. We view these as the second most severe. The final category we distinguished is statutory crimes, a violation of a specific state or federal statute and do not fall into the above two categories. These are viewed as the least severe crime in our data-set. A list of the types of crime and how we categorized them in our data can be found in appendix A and B below.

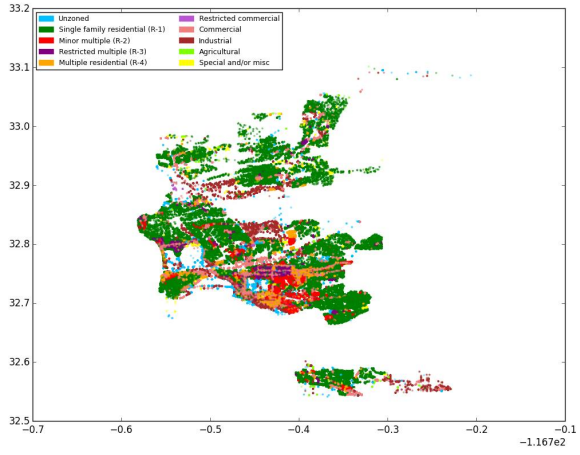
In order to choose the correct model to predict type of crime, we explored our data to identify possible trends. In our 360,964 valid crimes, the most dominant crime type was property crimes, of which there were 216,017 (approximately 59.844% of our valid crimes). Statutory was the second most common, with 107,322 crimes (approximately 29.732%), and personal crimes were the least common, with 37,625 crimes (approximately 10.423%).



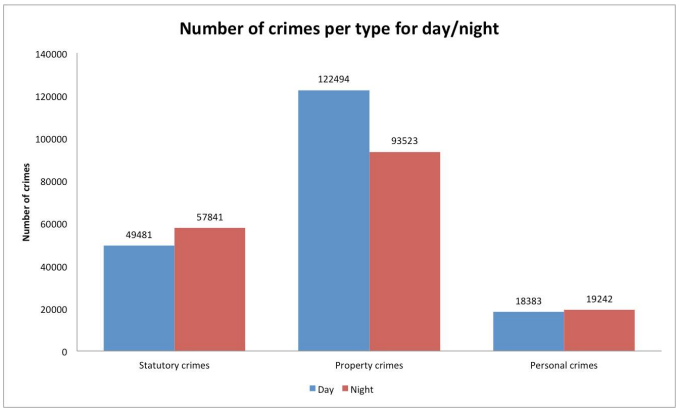
First, we plotted the geographical data on a map of San Diego to see if there were any visual patterns to where crimes were occurring. We didnt notice any clear trends but we noticed personal (most severe) crimes tended to occur in centers of densely populated areas like downtown.



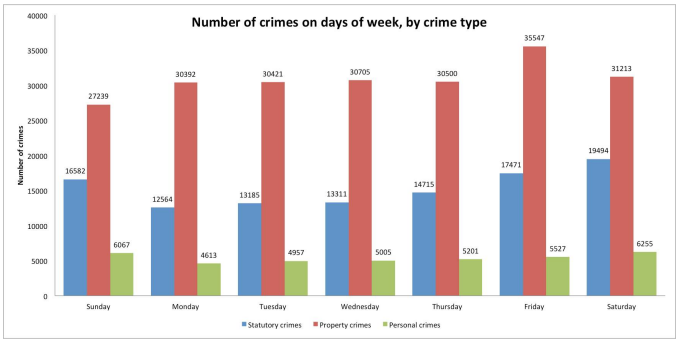
Next, we grouped crimes by the ASR zone. To do this, we simply created buckets of ASR codes and placed each crime committed inside its respective bucket. We graphed the data to see if there were any clear patterns with crime type and ASR zones; statutory crimes tended to be most common in unzoned and industrial areas (relatively), and property crimes tended to take place more often in residential and commercial areas.



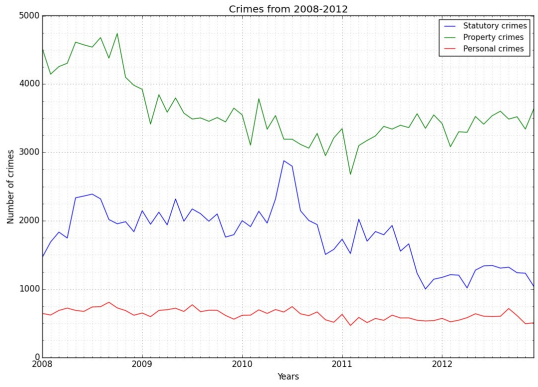
We also examined whether the day/night cycle had any visible patterns in our data. We grouped each crime by either daytime or nighttime and examined the number in each category. We noticed there were more day crimes, which shocked us initially. However, property crimes are the most common type of crime in San Diego, and most people are home at night, so property crimes would most likely take place during the day, while people are at work, away from their property and belongings. In our 360,964 crimes, 190,358 were day crimes (52.736%) while 170,606 were night crimes (47.264%).

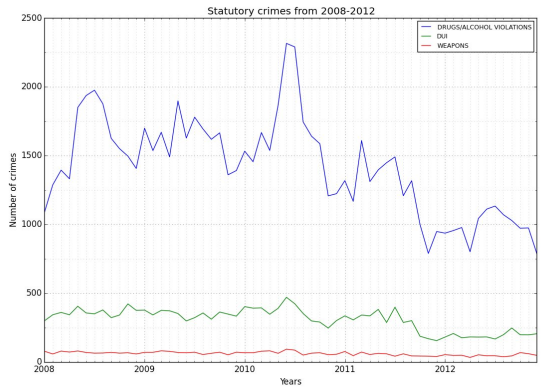


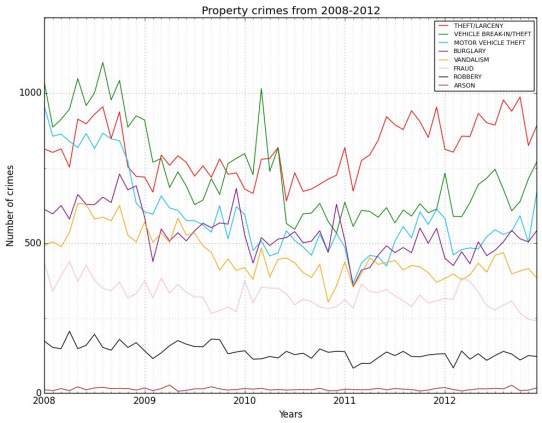
Among the attributes that we investigated from the data set was ”day of week”, a numeric representation of a day (e.g. Monday is 1, etc). Thinking that this might be a good feature to use, we created a graph, charting the frequencies of types of crimes for each day of the week. After analyzing this, we found that there does not appear to be a correlation between day of week and crime type. This can be seen in the graph below.

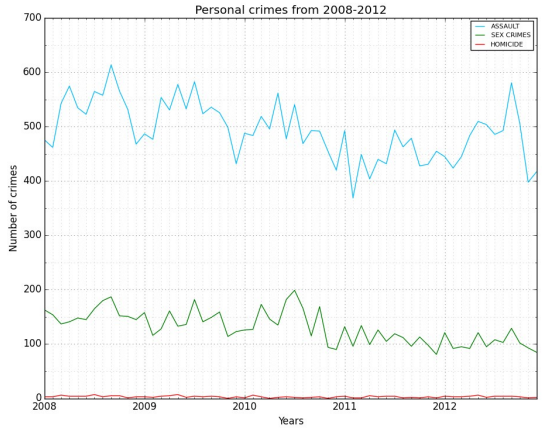


Another attribute we examined was the month and year; we graphed these against the different types of crimes to see if any trends would appear or if there were any months that specifically had a higher type of a certain crime. As seen in the line graphs below, we found no strong correlation.









These initial results gave us confidence that a predictive model was possible. There were several potential correlations between specific crimes and location, zoning types, and times of day had possible associations with type of crime committed as well.

III. Predictive Task

The full dataset includes 797,978 crimes; however, much of the data contains issues or is incomplete. Crime data inherently contains problems because reports are collected by officers in busy and stressful situations, and has many opportunities for error. Furthermore, while our dataset is a single body of data, it is sourced from 19 different agencies with different collection standards, which can make entries inconsistent. For example, the website claims that a missing ASR zone would be associated with the number -1, but after analyzing the data, we discovered that the value was None. In addition, our dataset includes entries that are missing one or more of the following: longitude/latitude, streetlamp distance, ASR zone, neighborhood, community population, and council population. After filtering out incomplete data, we are left with approximately 360,964 clean entries, approximately 45.235% of our original data.

Missing Data	Percentage
Coordinates	8.818
ASR zone	54.626
Neighborhood	54.689
Community	54.704
Comm Pop	54.705
Counc Pop	54.633
Lamp Dist	54.626

With our cleaned data, initially we thought it would be interesting to predict the specific type of crime (burglary, assault, arson etc), which would require a classifier with 14 classes. After some data analysis and a crude SVM based model we found this predictor to be very inaccurate (approximately 25% accuracy). Instead, we chose to predict the most likely broad type of crime (statutory, property, or personal) in order to make the task simpler. To train the predictor, we chose the longitude/latitude, ASR zone, day/night, community population, streetlamp distance, and clusters we found from k-means clustering as feature based on the patterns we seemed to perceive from our exploratory graphs.

To measure the baseline accuracy of our predictions, we first attempted to measure the exact accuracy ratio of predictions, but we eventually decided to use scikit-learns Hamming loss function, because our predictive task was multilabel. Using the Hamming loss made incorrect predictions more forgiving because it penalizes individual incorrect labels, instead of marking the entire sample wrong. To calculate the accuracy of our predictions, we subtracted the Hamming loss from 1.

IV. Predictive Model

In selecting our predictive model, we went through several iterations. Our initial model choice was a logistic regressor. Our goal was to predict the category/severity of crime and since that leaves only one of three buckets to classify, logistic regressor seemed it would be a simple model. This model gave us moderate results when using only geographical location (latitude/longitude) as a feature, so we chose it as our baseline to compare against. This simple model gave us an accuracy of .4014. After analyzing our results and data, we noticed that our data tended to have personal crimes that were very clumped together in certain areas, so we switched to using a support vector machine, which would decrease

the influence of all the clumped together easy cases, and improved our performance.

After switching to an SVM model, we divided it into two predictors, one for night crimes and one for day crimes, because we thought that certain crimes would be associated with day or night; from the charts, it seemed property crimes were more likely during the day (homeowners away from home) while statutory crimes were more likely during the night (alcohol, drug consumption). Dividing the model gave us the best performance on a feature matrix consisting of longitude, latitude, and ASR zone, which were picked because we thought they seemed to have the best correlation from looking at the map data. After examining the predictions, however, we found that the predictions were all property crimes, with no other labels. This was because the initial distribution of crimes in our training set was heavily imbalanced; over 70% of our training set was property crimes.

Once we noticed this error, we built a balanced dataset that contained equal samples of day crimes containing equal statutory/property/personal crimes and equal sample of night crimes with the same equal number of statutory/property/personal crimes. With the balance error gone, we started over on adding features, trying different ones to test which might be better than longitude and latitude.

We looked into other geographical features, like neighborhood, because we thought certain neighborhoods might be more violent or dangerous and contain more instances of certain types of crime. We assigned number labels to each of the neighborhood names in the dataset, and added this as a feature alongside day/night. These two features in the single SVM model gave us an accuracy of .4578.

Because our model had been significantly changed, we recalculated the accuracy of our baseline feature matrix (just latitude and longitude as features), which we found was .4140 for the SVM. Next, we looked at ASR zone and community population as potential features. We thought that certain types of crimes are more likely to occur in certain zones (for example, property crimes are probably more likely to occur in residential zones than agricultural zones). In addition, the number of people in the community could probably influence what kind of crimes would be committed (murder and other severe crimes are probably more likely to occur in more densely populated areas). Passing these features into the SVM gave us an accuracy of .4789.

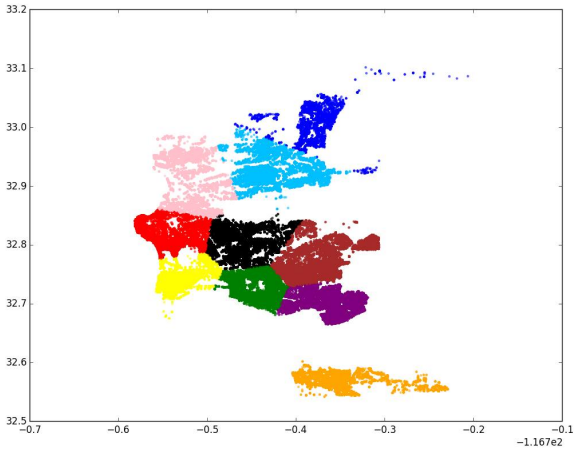
From analyzing our accuracy with just these features (without latitude and longitude), we thought latitude and longitude might not end up being useful features, as our accuracy had already improved; to test this, we put the two features back in and it ended up raising our accuracy up to .4814; we considered this and concluded the accuracy most likely rose because coordinates provide specificity that neighborhood could not convey (a certain rich household could be a primary target for robberies), but certain neighborhoods might also be able to suggest crime types (a poorer neighborhood would not have as much property crime).

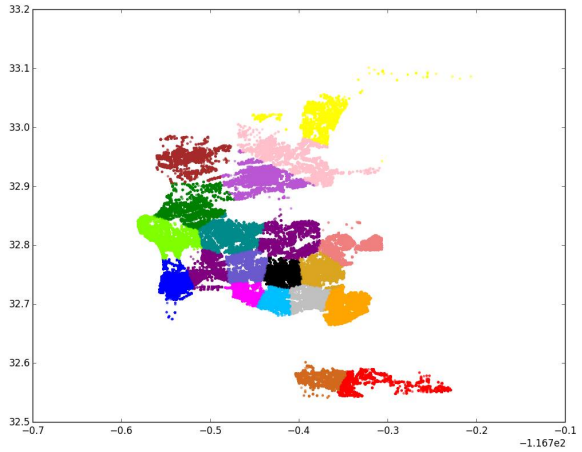
Another feature we considered after this was streetlamp distance in centimeters, which is an attribute included in all the valid crime data that we cleaned. We reasoned that it might be a useful feature because it can provide information regardless of whether or not the crime occurred at night; during the day, lamps signify proximity to a street or building, where property crimes are more likely to take place, and during the night, they tend to deter crime because they are a source of light (accuracy of .485).

As mentioned in the exploratory analysis section, we chose not to include temporal features like date or month or year because these features dont tend to correlate towards a specific type of crime (as seen in the graphs). There is no indication that a specific date or year might skew towards a specific type like property or personal crime.

At this point, we thought had gathered most of our features, so we tried a different model: a classifier using K-nearest neighbors. This algorithm classifies using a similarity measure (in our case, Minkowski distance, the default metric in scikit-learn). Because our data is mostly geographical in nature and K-nearest neighbors uses the nearest neighbors as a majority vote, we chose to use this model; it made sense to us because crimes are likely to recur in the same place or somewhere close by. After switching to scikit-learns K-nearest classifier, our model sped up significantly at the cost of a slight drop in accuracy (.4814).

Although we had access to a neighborhood field in our dataset, we wondered if we could improve performance if we made our own, as crimes do not follow strict neighborhood dividing lines. To accomplish this, we used K-means clustering to cluster the latitude and longitude points into different numbers of clusters. Other studies on crime [4] also used this approach so it seemed a proven technique. We ran the predictor with a range of different cluster numbers (from 5 clusters to 50) and we found that a community size of 20 clusters performed best; we seeded the random centroid generator in order to maintain consistency during testing, and 20 clusters gave us an accuracy of .486. This makes sense, because when the map of crimes is compared with the map of clusters, 20 clusters tends to group types of crime the most accurately. Below are maps of 10 and 20 clusters respectively, to show how they were divided.





After trying all these combinations of features we could include, we looked into other ways to improve our model. After looking into multi-labeled problems, we realized that we could represent our predictor as a vector instead of arbitrary values associated with the single label. For example, instead of labels of 0, 1, or 2 for statutory/property/personal crimes, they would instead be represented as vectors of size 3 with binary values (e.g. [0, 0, 1] for personal crimes and [0, 1, 0] for property crimes). The problem we found with the single label approach was that our predictor can only be completely right or wrong; for example, with this multi-label approach, if our model guessed [0,1,1] for personal crimes, it could be partially correct (it guessed both property and personal). This allows the Hamming loss function to be more forgiving in misclassifications as well, so our accuracy immediately went up after this change (up to .6774).

Following this mindset, we found that most of our features were categorical as well, and were better represented as binarized vectors instead of arbitrary number labels. Using arbitrary number labels could cause inaccuracy for the classifier algorithm, because it might think number labels that are close together are similar in nature. For example, adjacent ASR zone numbers have no similarities, and using binary vectors instead of integer numbers (0 - 9 for ASR) avoids this inaccuracy problem. We then changed the representation of a few of our features, namely cluster labels, community name, and ASR zone, and re-added them to our feature matrix. The addition of binarized features brought our accuracy up to .683.

Because of a radical change in the representation of both our features and our labels, we experimented more with features again and found that day/night, community population, and streetlamp distance actually harmed our accuracy. After reflection, we concluded that day/night was not useful because nighttime crimes were not indicative of what type of crime occurred; only day was more likely to indicate a property crime. Community population and streetlamp distance both only correlated to frequency of crime, and not necessarily type of crime itself. We experimented with adding a binarized form of segment ID, which is a numbered ID given to road network segments, but this made our model far too complex; the resulting feature matrix came out to be 48,000 samples by 15,290 features.

Certain features in our data set like city, state, or gctype were universal or missing in our data, so we excluded them from our predictor.

Once our accuracy had gone higher from removing these features (.6842), we began to get concerned with overfitting on the training set, so we increased the set size to 60,000 samples and created a validation set of size 30,000. In addition, we realized that we had been testing on a balanced test set, so we created another test set of 30,000 with unequal numbers of crime types. On these sets, we achieved an accuracy of .7189 on the test set and .7037 on the validation set, with an accuracy of .7798 when re-running on the training set. To confirm these results, we re-ran our baseline feature matrix (just latitude and longitude) on these sets and achieved a similar difference in accuracies between the training and test sets. This sanity check confirmed to us that we were not overfitting on the training set. Finally, we ran our original SVM on just latitude and longitude with these new sets (to check the old baseline on the new sets) and got an accuracy of .5017 on the test set. This shows that our accuracy increase of 21% was from our feature selection, and not just our choice of classification algorithm. This ended up being our final model.

V. Related Literature

Here, we will talk about some of the related machine learning studies done on crime prediction. This is a popular topic, with many papers written in a variety of journals. In order to be brief, we will discuss some of the newer literature, the techniques used, and the types of predictions achieved.

One approach to predicting crime is to take the perspective of an actual officer. We are concerned with criminal patterns, i.e. which individual committed what group of crimes. This can be used to anticipate a criminal's next target, as well as group similar crimes together to help tie an existing case to other crimes. One study, Crime Pattern Detection Using Data Mining, [4] achieved some success in this. The studys approach used clustering techniques such a k-means to help group together similar crimes. They then used semi-supervised learning for determining the best crime features. The resulting system was not capable of tying crimes together on its own, but was a useful tool for detectives nonetheless. The systems predictions helped to narrow down and focus detectives on relevant crimes and details.

Another approach to predicting crime involves analyzing occurrence of crime in a given place and time. This is very similar to our approach but implements a broader classifier that classifies whether or not a crime will occur at all, regardless of type. The Crime Forecasting Using Data Mining Techniques study illustrates this approach very well [5]. This study pulled data from many police departments in the northeastern part of the United States. The research focused on using spatial and temporal data and a variety of machine learning techniques, and compared the results of each. They found success with the individual nearest neighbor algorithm (INN) which predicts

crime based on locations where crime had previously occurred. They also achieved similar results with Naive Bayes, predicting that what happened in one location is likely to recur. This is similar to how we clustered the data points into communities, and picked the community number with the best accuracy. Some of these new boundary lines gave us better results than just using the traditional neighborhood boundary lines, because sometimes certain crimes were clustered across multiple neighborhoods. The more complex algorithms dont seem to vastly improve upon a simple location based model.

VI. Conclusion

In conclusion, our final model was able to predict the type of crime with an accuracy of .7189 on the test set and .7037 on the validation set. To achieve this result we used a K-nearest neighbors classifier with longitude/latitude, binarized cluster labels, binarized community labels (from San Diego community names), and binarized ASR zones. It was able to outperform the support vector machine in speed, as the SVM took a very large amount of time to train on a feature matrix of size 60,000 by 86, and it improved accuracy (SVM had an accuracy of .4139 on unbinarized labels). Scikit-learns K-nearest neighbors classifier does not return the thetas used in the model, so there was no way we could analyze what the parameters meant. We went through many feature representations, and ended up leaving out some features we originally used because they no longer increased the accuracy of our final model; community populations, for example, were left out because we found that community populations actually only correlated with frequency of crime in general, and not with any type of crime. Similarly, we left out binarized segment ID (IDs for road network data) because it added far too much complexity to our model; it added over 11,000 dimensions to our feature matrix. Other features, like month, day, or year, provided no correlation because there were no particular times during which any type of crime was committed more often. In addition to taking attributes already present in the data, we constructed our own features like cluster labels from K-means to improve performance. Like other papers in this area of research [4], we found that K-means clustering was a key technique in achieving our final results. From frequent testing, we found that the largest contributors to our models performance were geographical location, and binarized labels on ASR zone and our clusters.

Appendix A
Types of Crime

	Type	Total
1.	DRUGS/ALCOHOL VIOLATIONS	85085
2.	THEFT/LARCENY	48583
3.	VEHICLE BREAK-IN/THEFT	43955
4.	MOTOR VEHICLE THEFT	35442
5.	BURGLARY	31975
6.	ASSAULT	29604
7	VANDALISM	27410
8.	FRAUD	19624
9.	DUI	18549
10.	ROBBERY	8238
11.	SEX CRIMES	7837
12.	WEAPONS	3688
13.	ARSON	790
14.	HOMICIDE	184

Appendix B
Categories of Crime

Personal	Property
ASSAULT	THEFT/LARCENY
SEX CRIMES.	VEHICLE BREAK-IN/THEFT
HOMICIDE	MOTOR VEHICLE THEFT
	BURGLARY
	VANDALISM
	FRAUD
	ROBBERY
	ARSON

Statutory
DRUGS/ALCOHOL VIOLATIONS
DUI
WEAPONS

REFERENCES

[1] The Omega Group, *Crime Mapping*,
Website-http://www.crimemapping.com/default.aspx

[2] Dr. David Weisburd and Dr. Cynthia Lum *Hot Spots Policing*, George Mason University, 2013.

[3] San Diego Sheriff Dept, *San Diego Region Crime Incidents*, San Diego Sheriff Dept, 2013.
Website-http://data.sandiegodata.org/

[4] Shyam Varan Nath , *Crime Pattern Detection Using Data Mining* , Oracle Corporation, 2006.

[5] Chung-Hsien Yu, Max W. Ward, Melissa Morabito, and Wei Ding, *Crime Forecasting Using Data Mining Techniques*, University of Massachusetts Boston, 2011.