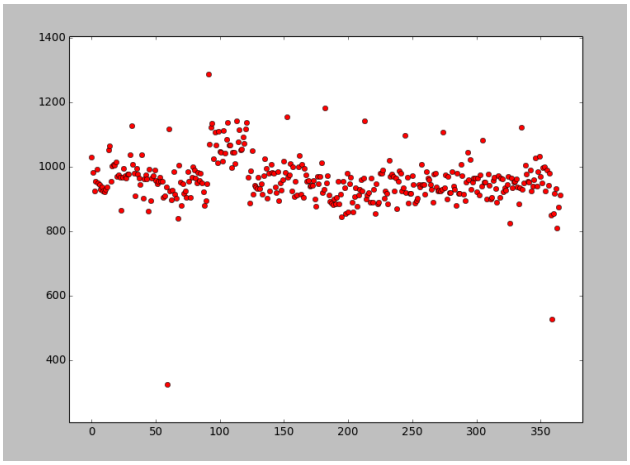
SANDAG Criminal Database
Assignment 2

Steve Morlan
A10453912

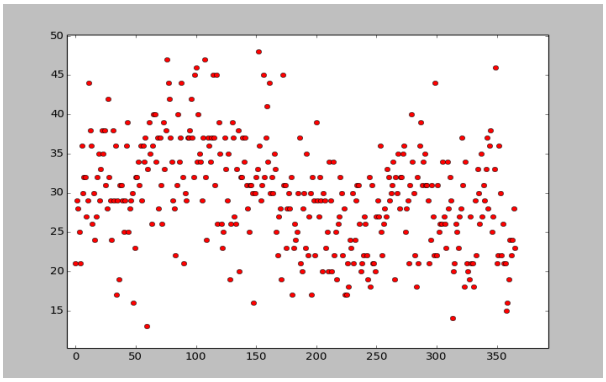
David Hart
A11027280

Ilia Shumailov
A91500840

The following document represents the statistics, problems, predictions and conclusions reached by our team for the data set we identified. It was our desire to work with the SANDAG criminal database. We intended to make predictions based on types of crimes, locations, and victims. The following plot represents theft over the year. Our assumption was that we would see spikes during the holidays, tax season, etc. While we saw spikes during tax season, we did not believe this one trend was enough to justify use of the data set. So we sought out further correlations.

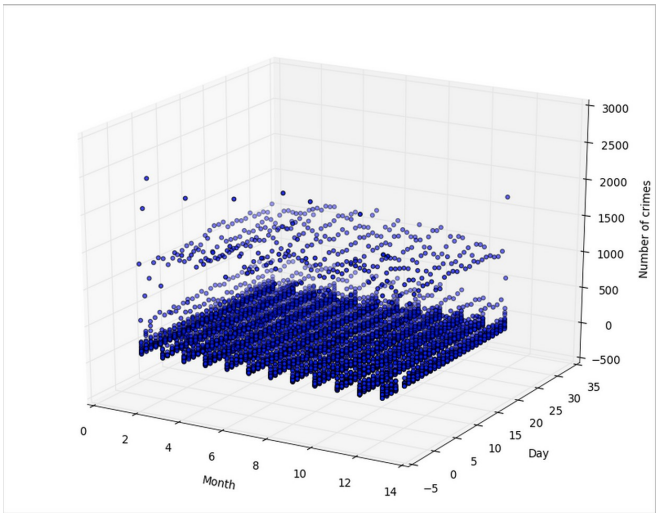


This graph represents the number of thefts in San Diego county on the nth day of the year. The 100th day of the year is April 10th, the middle of tax season.

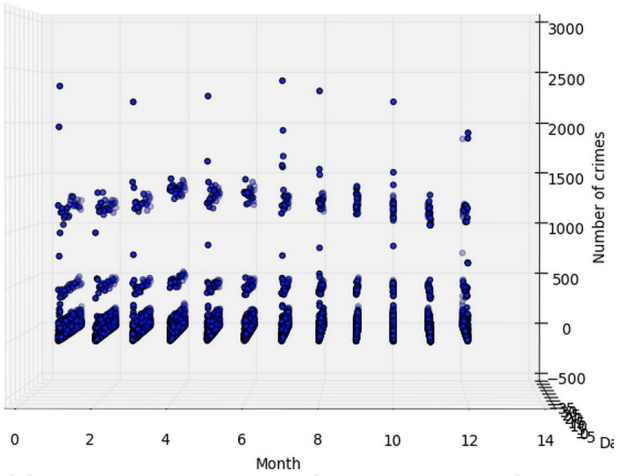


This graph represents the number of weapons related crimes on the nth day of the year.

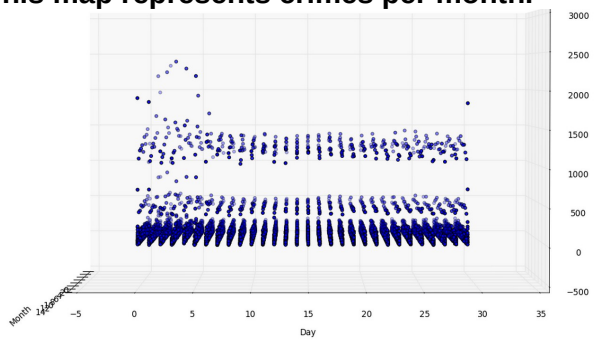
The SANDAG criminal data set consists of roughly one-million entries, where each entry is a crime that took place in San Diego county. During the exploratory phase, we identified trends in multiple features. For instance, theft rate increases during tax season, and sexual assaults increase during holidays. We also found that location has an effect are frequency of specific crimes, such as murder and armed robbery. Moreover, the first of the month seems to be higher in theft and overall crime as well. We believe this to be for the same reason that tax season has a higher frequency of theft, low income need to pay rent or other bills.



This is a 3d map of month, day, number of crimes where number of crimes is the vertical axis.



This map represents crimes per month.



This map represents crimes over the day of the month.

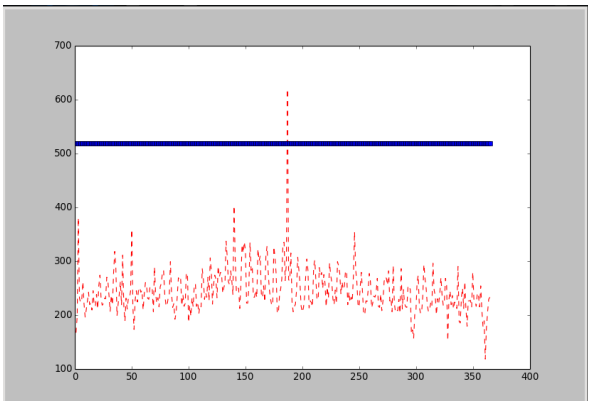
The SANDAG criminal database is formatted as follows:

- date: ISO date, in YY-MM-DD format
- year: Four digit year.
- month: Month number extracted from the date
- day: Day number, starting from Jan 1, 2000
- week: ISO week of the year
- dow: Day of week, as a number. 0 is Sunday
- time: Time, in H:MM:SS format
- hour: Hour number, extracted from the time
- is_night: 1 if time is between dusk and dawn, rounded to nearest hour. All comparisons are performed against the dusk and dawn times for the 15th of the month.
- type: Crime category, provided by SANDAG *This is the short crime type*
- address: Block address, street and city name
- segment_id: segment identifier from SANGID road network data.
- city: CPC code for the city.
- nbrhood: CPC code for the neighborhood. San Diego only.
- community: CPC code for the community planning area. San Diego only.
- comm_pop: Population of the community area, from the 2010 Census
- council: CPC code for the city council district. San Diego only.
- council_pop: Population of the council area, from the 2010 Census
- place: Census place code, for future use.
- asr_zone_: Assessor's zone code for nearest parcel.
- lampdist: Distance to nearest streetlamp in centimeters
- state: State. Always "CA"
- lat: Latitude, provided by the geocoder.
- lon: Longitude, provided by the geocoder.
- desc: Long description of incident.

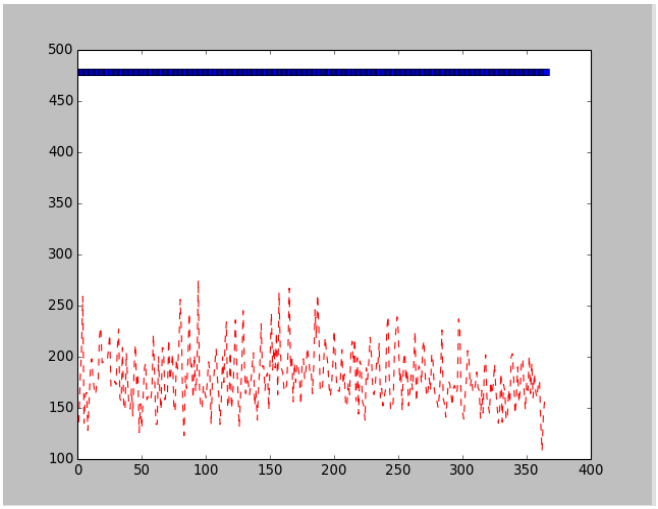
Considering the correlations identified in types of crimes and time of year, as well as the location and density of crimes, we intend to predict frequency of

crimes given a date and location. To start we will run basic analytics on the data and identify averages and trends. Following that, our baseline is to train on the first 50% of the data, then predict based on the average crimes committed throughout the years. Using the test set, we will compare our predictions and calculate the error and MSE. Following a baseline, different models using Naive Bayes and linear regression and lambdas representative of the features we identified as influential, such as time of day, absolute location, day of the year, and day of the month. For the the examples from now on we will use 2007 San Diego data for seen data and 2011 San Diego for unseen data.

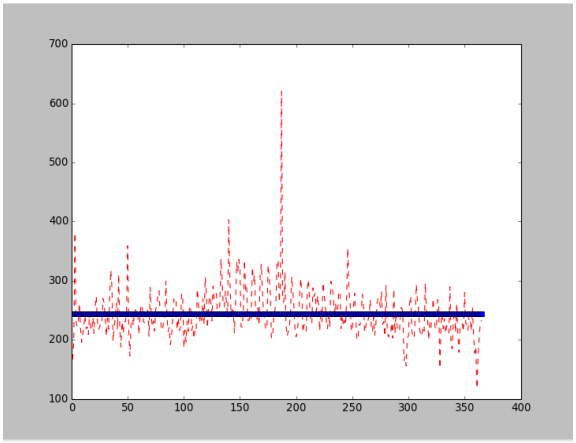
To begin, we wanted to look at a simple model to use as a baseline. The training data was organized such that we had counted the total number of crimes that occur on each day, independent of the year. From this, a global average was calculated and used as our prediction. For the seen data, the model's MSE was equal to 186,199.76. This huge number could easily be explained by observing the data. The trend could be noticed, that average number of crimes goes down and the numbers of crimes are not uniformly distributed, but they are clustered.



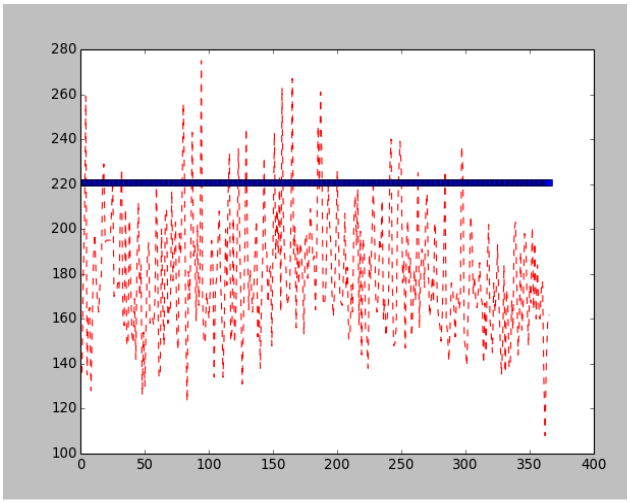
The blue line is the prediction for the seen data, the red points is the actual data. The model performs even worse on test unseen data with the MSE equal to 210304.044. Below is a figure of the model trained above run against the test data.



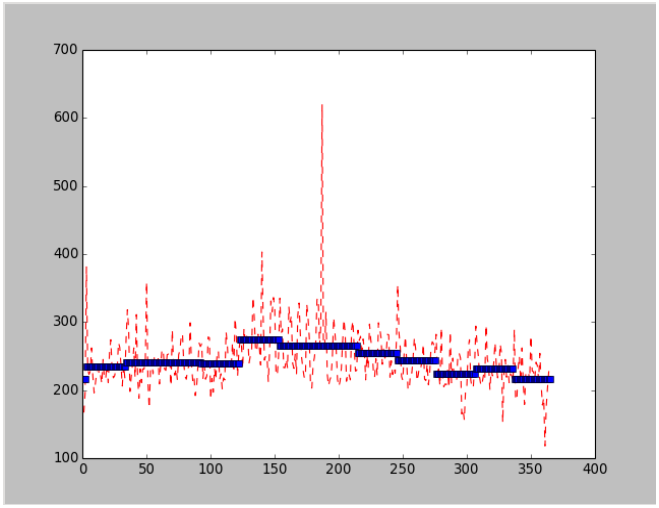
Next we figured there could be trends based on cities. So that the clustering is a city-dependant feature. That brought a massive of improvement for the model and the MSE changed to 79.173, which could be easily explained with the fact that we were classifying safe cities with a huge prediction before, whereas now being city-specific, the misclassification is very small for those places. This makes sense as the number of crimes per a specific city can vary greatly depending on the population of said city. For example, if we compared the number of crimes per day occurring in San Diego versus a smaller city like Del Mar we would expect to see much less in Del Mar than San Diego.



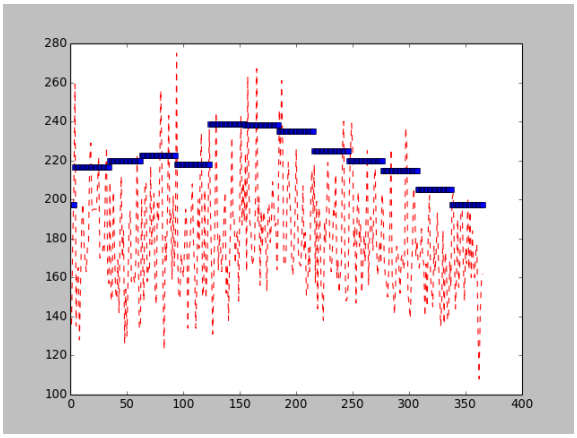
The unseen data performed much better as well. MSE changed to 182.694227. At the same time, the behavior could be connected to number of police officers in the city or even the number of lamps lighting up the city during the night.



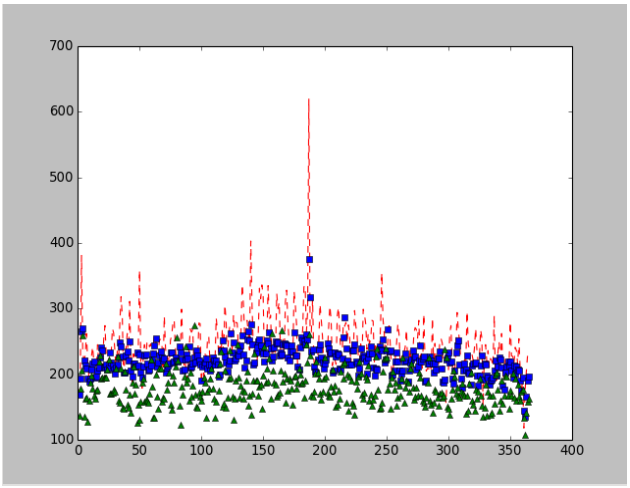
After looking at the graph we figured that the spikes on the graph could not only be city specific, but also month specific. Just the same way certain animals have specific months when they are most active. The thought was that certain times during the year could potentially affect the rate of crimes. Yet again, the MSE of this model was lowered to 68.648 and appears to fit the data much better. It seems like the crimes are most frequent during the May/June. Which could be potentially connected to popularity of California state during the summer.



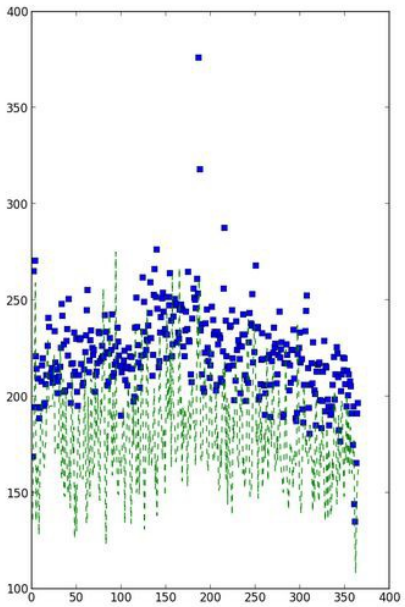
Testing this against the unseen data yielded slightly worse results with an MSEof 184.240.



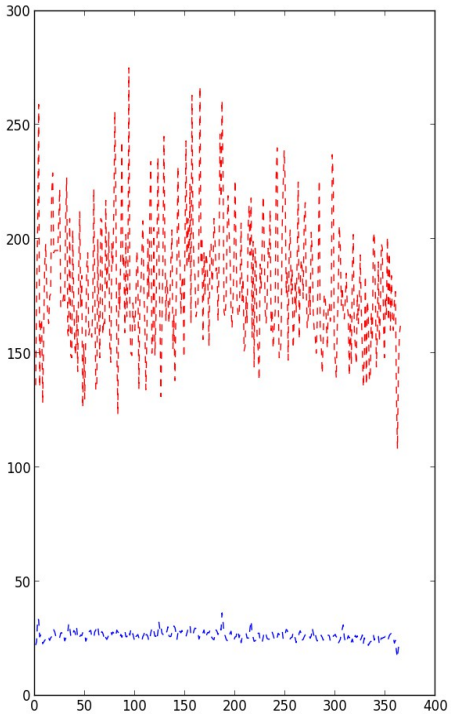
Next we figured there could be a trend based on the city as well as the actual day, so a second model was created taking this into account. As seen in the plot below, this model fit our data much, much better. Again, the blue line is the model and the red points are the data points. This MSE for the unseen data is 102.375



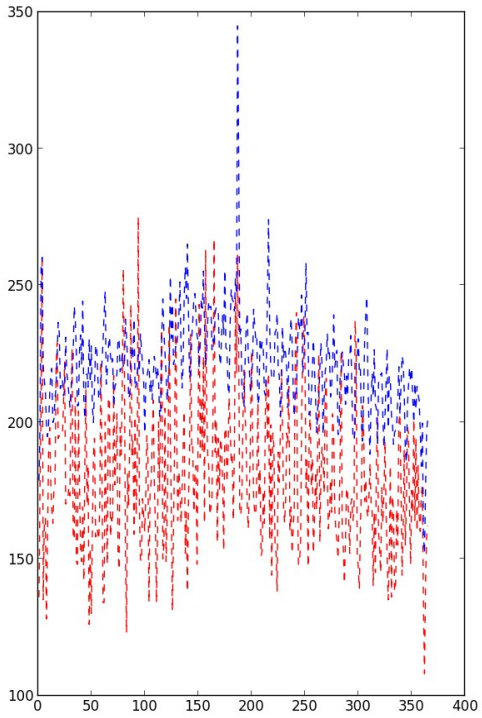
This model against the unseen data, as expected, did slightly worse than its training counterpart with an MSE of 197.043



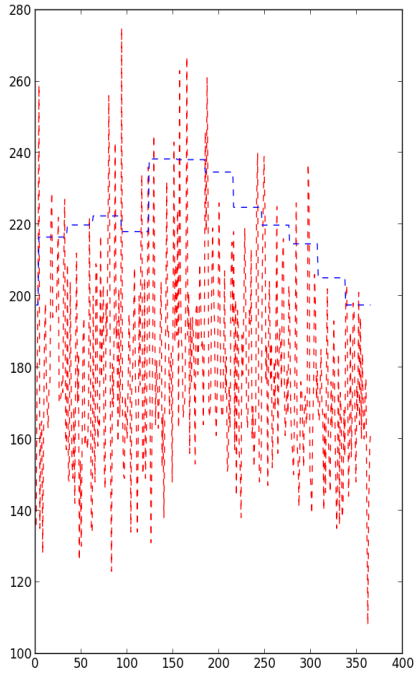
After observing such a an improvement to the baseline just by simple counting, regression analysis was used. If the feature vector was built just by using the particular day, disregarding the city, the model was performing poorly. The red line corresponds to actual data for San Diego in year 2011, the blue line is our prediction. At first, it looks surprising how regression can perform that bad, but after it becomes apparent, that if you only use the date to build the feature vector, the same feature vectors (for the same date) in different cities would have different labels. And linear regression with regularization would favor smaller value.

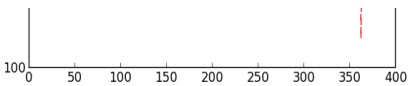


After previous experience, it became apparent that it is necessary to make use of the city in the feature. Just by training city-specifically, the following model was acquired with the MSE of 414.978780737 from 1645.025 (for San Diego 2011)



Here, it was decided to try to use month and city to build the feature vector.





This model produced MSE of 412.8820788.

Currently, the criminal data in the data set we are using is being leveraged by the Criminal Justice Research Division to provide meaningful statistics and predictions for cities in San Diego County. However, the actual predictions and statistics they predict are either trend based predictions or basic relations. In one such report, *Thirty-Five Years of Crime in the San Diego Region,* Division Director Cynthia Burke highlights the many trends that the data set included: The number of robberies reported across San Diego decreased by 11% between 2014 and 2013, with 45% of the robberies occurring in the open. Homicides increased by 6 percent between 2013 and 2014. While these statistics are useful for homebuyers, we aim to provide predictive analytics that are useful to government and private sector security and enforcement. Knowing that more crimes happened in 2014 than 2013 does not help in identifying days where more officers need to be on staff.

Similarly, IBM is focusing on making criminal predictions concerning place and time of crimes. The most notable difference is that IBM is working directly with law enforcement, which allows them to make correlations between age of offender, number of previous offenses, and statistics about the victim. While the majority of the information about the study has been made private, it seems that IBM is employing very similar predictive analysis to larger sets of data.

Despite the existence of previous work, we are unable to compare our exact findings with others due to privacy laws. The

majority of the work being leveraged in this field has been restricted from the public.

Conclusions

The assumptions for the models we made were correct and we were able to improve the performance of our naïve baseline model just by incorporating basic features. There are many things we could do to improve upon our model. With our existing dataset we would attempt to incorporate more of the features described in it such as community population, infer seasonal trends, etc. What would be more powerful, though, would be to include information from other datasets with information like politics, news, and many other ideas. Basic probabilities could even come in to play. For example if there is a large crime going on in some part of town it may be more likely that another may happen in a different area. People are probably more likely to commit a crime if they know there is less danger of getting caught.

With this model, and future improvements upon it, we could attempt to predict crime rate and various types of crimes based on location, ethnicity, population, time of day, etc. This information would be invaluable to our police officers and other men and women spending every day of their working lives trying to keep us safe. The world would be a safer place if we could predict where and when dangerous crimes would happen.

Or one other idea for the crime prediction would be to take into account the experience of the police. It seems like year-after-year number of crimes decreases and that might have correlation with police taking efficient action.

With the same crime prediction, it might be even possible to predict trends like mafia migration. For example, if predicted crime for city A is much smaller in comparison to neighbor city B, and B has increased criminal activity, that could be meaning mob expansion or migration.