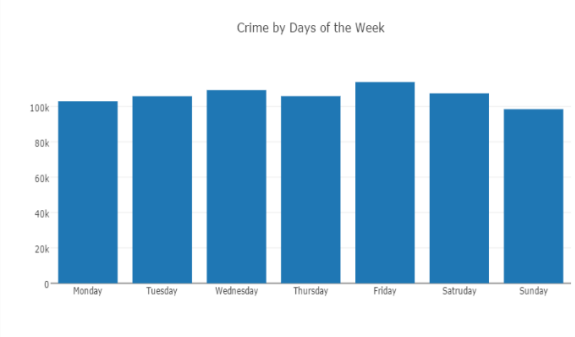
Is There a Time for Crime?

By David Thomasson (A10871310)

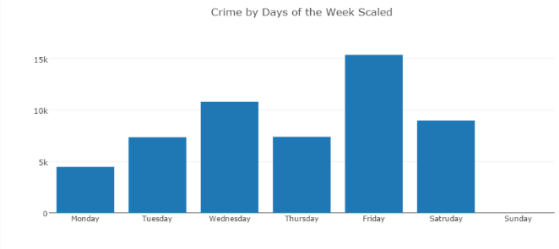
ABSTRACT: Predicting whether or not a given crime is violent or non-violent has been proven to be doable by other users on Kraggle using geographical data provided. Using Police Records from the San Francisco Area from 2003-2015 I made attempted to make a model using only temporal data. I attempted to model the data on Day of Week, Hour, Month, and Year using Logistic Regression, Support Vector Machines, and Naïve Bayes. Ultimately, the Naïve Bayes model was the most successful, and even though the model was only 10% more accurate than the baseline predictor it is still significant enough to confirm some of the intuitions about when violent crimes are most likely to occur. With a temporal model proven effective, it suggests combining it with a purely geographical model could produce even more impressive results in the future.

I INTRODUCTION:

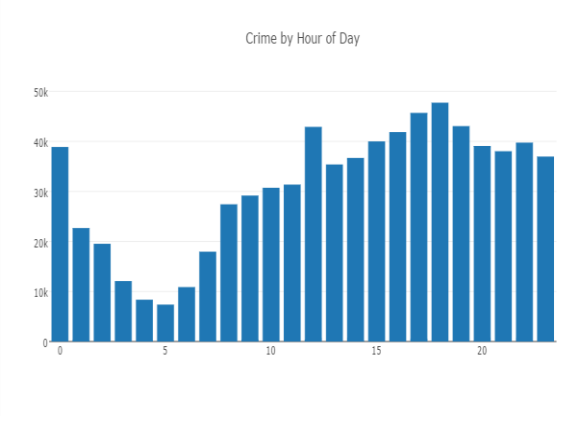
The data we will be looking at is the collection of crime reports from the San Francisco Area from 2003 – 2015 (1). In this data we are given the time, location, and category of a crime. The categories of crime for this dataset range from bad checks to assault. An interesting feature given in the data is the day of the week. I started exploring the data by examining the number of crimes that occurred by day of the week.



There is some difference between these days, so I subtracted the lowest crime rate day (Sunday) and then plotted the data again to make the differences a little more obvious.



I also examined crimes by time of day.



Looking at this data, there do seem to be definite lulls and peaks in crime activity. For example, you would expect to see very little crime on Sunday at 5am compared to Friday at 6pm.

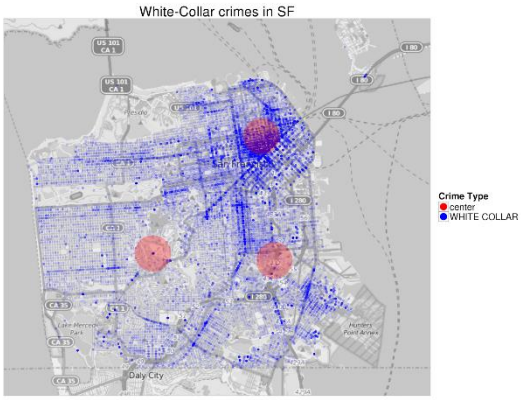
Looking further into the categories of crimes reported, I found that there are police reports that are titled “NON-CRIMINAL”. There are also reports adding “SECONDARY CODES” to the preceding report. I removed these from the data so I only look at criminal activity. There are also “WARRANTS” in the category feature. I removed these since I wanted to look at not just police activity, but where crimes were actually being committed. After these were removed I found that 84.67991572233927% of reports in the data are considered “Criminal”. I then looked at all the categories of crime and put all the ones I saw as violent or dangerous to a person who may be in the area or their belongings under the category “Dangerous Blue Collar Crimes”. They are as follows: ["VANDALISM", "LARCENY/THEFT", "STOLEN PROPERTY", "ROBBERY", "DRIVING UNDER THE INFLUENCE", "DISORDERLY CONDUCT", "LIQUOR LAWS", "VEHICLE THEFT", "ASSAULT", "KIDNAPPING", "TRESPASS", "ARSON", "SEX OFFENSES FORCIBLE"].

I found that any given report of a crime has a 50.384248078759608% probability that it will fall into this category

II REALATED WORKS

Reading the report “PREDICTIVE POLICING: The Role of Crime Forecasting in Law Enforcement Operations”(2) it states that some of the most common features used in predictive policing were “time of day, day of week, and time and day cycles” and “repeat locations”. These are represented in the data I am looking over. There are other things such as “demographic and economic data from the crime area” or “weather”. While these could be helpful, these are not included in the data I was given.

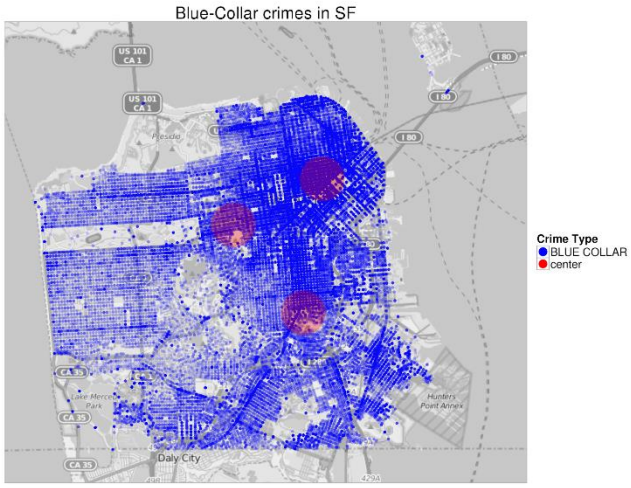
This data comes from a Kraggle competition, so there are quite a few others who are working on this data. One stand out example that was publicly available was “White-Collar vs. Blue-Collar Crime in SF”. In this script it finds clusters of white-collar and blue-collar crimes and puts it on a map of San Francisco.



I found this particular script interesting, as it differentiated between crimes by category much like I was hoping to do and it showed that where a crime was committed influenced what category it is likely to be.

III IDENTIFY A PREDICTIVE TASK

Given the fact that someone had already shown that there is a way to predict the category of a crime based on the area, I wanted to see if I could predict the category based on temporal factors rather than spatial factors. I predict that



temporal factors alone will produce an incredibly strong and accurate model, but I want to find one that is at least stronger than blindly predicting the baseline of “always a violent crime”.

If I can get anything above that baseline, there is enough evidence through related works on this dataset that location is a predictor of violent crime that I could say that adding my model based on temporal features to someone else's predictions using spatial features should improve both models.

During the time I spent exploring the data I did find that there was some correlation between days of the week and crime rates, as well as hours of the day. Given this information,

it is reasonable to try to make a model under the assumption that there is some temporal element as to when certain crimes are committed.

As for establishing a baseline, I mentioned earlier that only slightly more than 50% of the crimes in the training set were violent. Therefore it seems that a reasonable model for the baseline would be the model that predicts that any report about a crime is a report about a violent crime. Any model worth it's salt on a binary prediction task should be able to do better than 50% anyways.

The two models that I will try for this task are Logistic Regression and Naïve Bayes. The features I will be looking to include in these models are as follows: Day of the Week, Hour of the Day, Year, and Month. I would obtain these features by going over the CSV file in which the reports are provided and first removing all reports that are non-criminal. Then I would remove all reports where the police were acting on a warrant. After that, I will grab the reports "Date" and parse it and split it into "Year" and "Month" and "Hour". Then I will get the "Day of Week" field. When creating the Validation set, I found that the entries were sorted by Year, so I randomized the entry rows before training the model so that Year had the opportunity to be a helpful feature.

I will then take all of the reports that have not been culled and record whether or not the crime falls under my list of "Dangerous Blue Collar Crimes". Any report that is in this list will be marked with a "1" and all others with a "0".

IV MODELS AND TESTING

The models I tested were Logistic Regression, Naïve Bayes, and SVM. These models make sense to use since my features can all be represented in a binary fashion. An important thing to note about how I set up my features is each Year/Day/Hour was represented in the form [is hour 0, is hour 1, …,is hour 22] so it is in a binary representation. Note that hour 23, month 12, and Sundays are not included in the feature vector to avoid double counting.

The first model I tested was Logistic Regression. After splitting the data into training and validation sets I trained the model first on "Days of the Week" and "Hours of the Day".

This produces 55.810220948895251% accuracy on the training data, which is better than the baseline of always predicting "Violent Blue Collar". On the validation set, the baseline established earlier performs very poorly (~46% accuracy) and the model with the mentioned features has 0.499712501437% accuracy on the Validation Set. This is still noticeably better than our baseline, but is still a little lackluster. After adding the feature of "Year", we see better performance on the Validation Set. With year there is an accuracy rate of 0.503212483938%. While this is a very small improvement, it does get the model over the 50% mark, which is significant. With the feature of Month added we see another increase in how well the model preform on the Validation set, increasing its accuracy to 52.02%. After doing a little more research on the season's effect on crime, I found an article by Gerhard J. Falk saying "that homicides reach a high point during Christmas week with a peak on Christmas day" so I tried to add this in to the logistic repressor by adding a feature as to if the date fell in the week of Christmas. This did improve slightly the performance on the training set, but negatively affected the performance on the validation set, so this feature was scrapped as it caused overfitting.

The next model I tried was SVM. That model made sense in the fact that I wanted to find a way to make a distinction between violent crimes and non-violent ones. However, I ran into an issue with complexity. It took forever for the model to train! After looking briefly for works regarding large data sets and SVM's I found that "SVMs are usually not chosen for large-scale data mining problems because their training complexity is highly dependent on the data set size" (4). The paper does include some information about how to make SVM's scale well with large data, but it all involved clustering, so I chose not to move forward with this model.

I then moved on to using Naïve Bayes. I figured that Bernoulli Naïve Bayes made the most sense for this as all of the features are binary features.

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

I wanted to test Gaussian Naïve Bayes as well on the data set just in case it produced better result despite my intuition that Bernoulli would

be better. It is also worth noting that I could not use features I had tried in the Logistic Regression like “does the date fall on the week of Christmas” as this feature is not independent from the month features also used in the model. After training the model with the features [Year, Month, Day of Week, and Hour] and testing it on the Validation set it predicted whether a crime was a “Dangerous Blue Collar Crime” correctly 56.917215413% of the time, which is 10% better than the baseline and about 4% better than the Logistic Regression Model. I also tried the Gaussian Naïve Bayes Model.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Its performance with the same data used in the Bernoulli was 55.148224258% accuracy, which is slightly worse than the Bernoulli model. This means my decision to use the Bernoulli model over the Gaussian model is reasonable.

V SIGNIFICANCE OF RESULTS

The results of the model that had a positive impact on dangerous crimes committed are as follows: “Saturdays and Sundays have a positive impact on Violent Crimes”, “The Hours of 1,2,3,4,18,19,20,21,22,23 have a positive impact on crime”, “January, March, and December are the months that have a positive impact on crime”, and “The years 2003, 2004, 2006, 2013,2014, and 2015 from the years 2003-2015 are the years that have a positive impact on crime”. Some interesting conclusions can be drawn from this, like while crimes may happen all day, the more dangerous ones are more likely to happen in the hours of night. This is something that our mothers have always known, but it is nice to see actual data backing up this intuition.

These models used only temporal data and were able to beat a baseline predictor to predict whether a crime is a “Dangerous Blue Collar” crime or not. While 10% above the baseline may not seem like the predicative power we would like, the importance of these results is that there is value to be gained from even simple use of temporal features.

The lack of geographic features is really what is limiting this model I believe. The fact is, is that while dangerous crimes are more likely based on certain temporal factors, crimes are not easy to predict with 100% certainty. This isn’t the Minority Report. But as we add more and

more helpful features we can push the accuracy up slightly with each feature. It is just the case that we run out of features very quickly when we limit a model to just temporal data. However, with the addition of spatial features we open ourselves up to a whole new possible set of features. We could look at other crimes occurring in the vicinity at the same time that could link multiple instances of crime together (eg. Theft -> Assault). Or we could pull additional datasets for weather in an area given the longitude and latitude of a crime report.

That means that if combined with any of the other models on this data mapping types of crimes to a geographic area (Such as other models previously mentioned like the "White-Collar vs. Blue-Collar Crime in SF" script) it should increase the power of the model. These models could even be taken a step further to include temporal/geographical data such as the weather. Hopefully, with these predictive models increasing in strength, we can do an even better job predicting violent crimes before they happen and ensuring that the innocent are out of harm's way.

VI REFERENCES:

1. Data – San Francisco Crime Classification: https://www.kaggle.com/c/sf-crime
2. Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, PREDICTIVE POLICING The Role of Crime Forecasting in Law Enforcement Operations from the Rand Safety of Justice Program.
3. "White Collar vs. Blue Collar Crime in SF" by the user: Olalekan at https://www.kaggle.com/ampaho/sf-crime/white-collar-vs-blue-collar-crime-in-sf
4. Making SVMs Scalable to Large Data Sets using Hierarchical Cluster Indexing by Hwanjo Yu, Jiong Yang, Jiawei Han, Xiaolei Li, SUBMISSION TO DATA MINING AND KNOWLEDGE DISCOVERY: AN INTERNATIONAL JOURNAL, MAY. 2005