Crime Prediction using Businesses and Housing Values in San Francisco

James Jung Lee, Joel Kek, Yik Lun Lee

Introduction

Predictive policing is the idea of using technology and data analytics to proactively predict and preempt crime, hence allowing police departments to do their work in more intelligent ways. In the Police Chief Magazine, Los Angeles Police Department's Chief of Detectives Charlie Beck describes predictive policing as a combination of "directed, information-based patrolling, rapid response supported by fact-based prepositioning of assets, and proactive, intelligence-based tactics, strategy, and policy". [1] With police departments nationwide facing budget cuts and manpower shortages, the idea of using existing police resources more effectively has taken off rapidly. [2] While the effectiveness of predictive policing has yet to be rigorously proven, efforts have been bearing fruit so far. For example, by using predictive policing software in the Foothill neighborhood, the LAPD managed to reduce property crimes there by 12%, as compared to a 0.5% rise in surrounding neighborhoods. [3]

In our study, we attempt to develop a predictive model where we can input certain conditions about an area in San Francisco, i.e business density, housing values, etc, and try to predict the overall crime rate of that area. It is important to note that instead of using real-time data, we attempt to use more general, static factors of a region. This is useful for police departments that want to attempt predictive policing but lack access to real-time data. Even for police departments that have access to real-time data, this research allows them to understand what indicators might be useful in determining optimal allocations of resources in the future.

Method

Model

Our aim is to predict the number of crime incidents in a certain area given a certain features specific to the area. We treat this as a binary classification problem, where for the target variable we define "high crime rate" as the occurrence of a threshold number of crimes over some period of time (in our data, over 12 months). The types of crime included in our data are aggravated assault, battery, burglary, carrying a concealed weapon, child abuse, public nuisance, conspiracy, courtesy report, credit card theft, suspended or revoked drivers license, theft, rape, forgery of prescription, lost property, inflicting injury on cohabitee, malicious mischief, vandalism, missing juvenile, parole violation, violation of

restraining order, transportation of marijuana, traffic violation, threats against life, tampering with vehicle, stolen vehicle, suspicious towards child or female, robber, sales of controlled substance and resisting arrest. As an initial starting point, we looked at a region of San Francisco around the SOMA district, roughly 1.5km^2 in area, shown in Fig. 1.

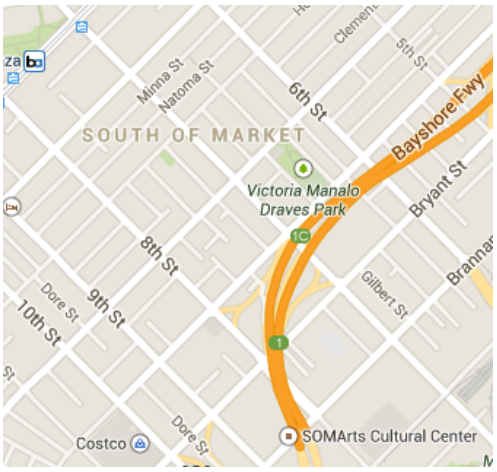

Figure 1: Area of San Francisco used for initial testing of our model.

We divide the region into square grids of area 200 m x 200 m , or 3-4 blocks across, each considered a training example. The features we considered are the number of businesses within the area and the average rental pricing, we also specifically look at the number of restaurants in the area.

Data collection

The training data is extracted from the data sets downloaded from the data.gov website, which include crime information from the San Francisco Police Department's Central Database Incident System (CABLE), businesses actively registered with the San Francisco Office of the Treasurer & Tax Collector, and rental listings on Padmapper.

Preliminary Results

Figure 2 plots the density of businesses inside a 200m x 200 m square versus number of crimes that have occurred in the past 12 months in that area. We can already begin to see a correlation between business density of an area and the number of crimes that occur in an area. We will classify using a support vector machine (SVM), and we first find a cutoff to determine which classifies as high crime versus what classifies as low crime. We used a cutoff of 60 crimes or less in the area to give an even split between high and low crime.

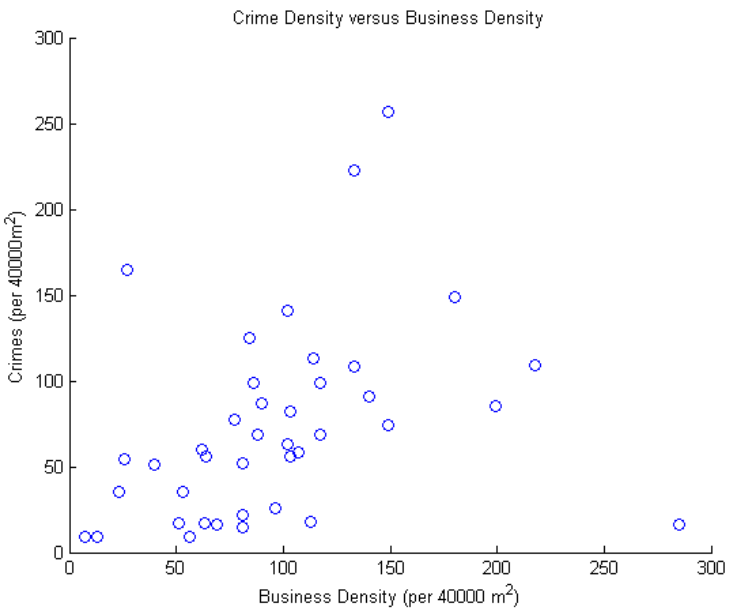

Figure 2: Scatter plot of crime rate vs business density over 200m x200m squares in the SoMa region plotted in Fig. 1.

Training of SVM and analysis

Figure 3 plots the SVM obtained using Matlab's SVM library and with a Gaussian kernel. In this example the feature size was reduced to only 2, the total business density and the average rental price. The plot shows the SVM trained on the entire data

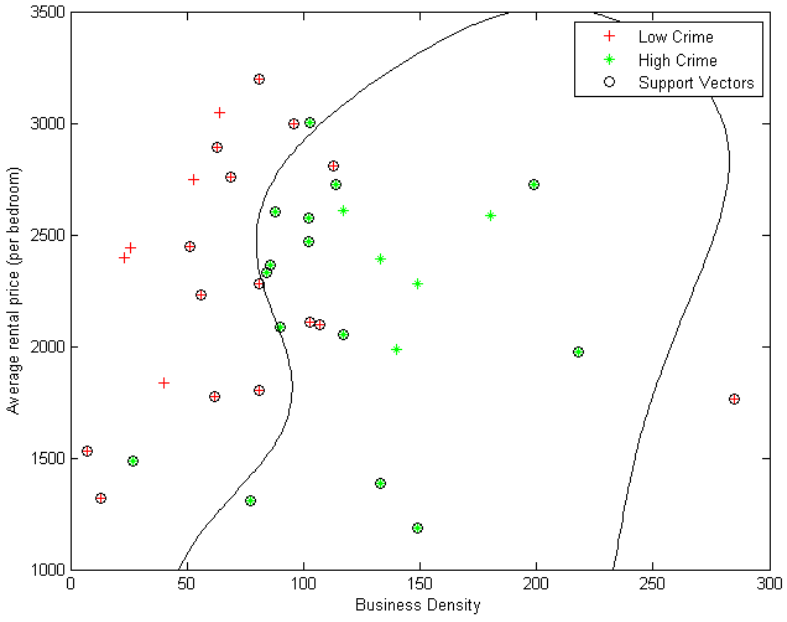

Figure 3: SVM trained on data from SOMA, using 2 features: business density and average rental price

We see that generally, places with fewer businesses had lower crime rates, as expected from Fig. 1. However, interestingly enough, rental rates seem to be only marginally useful in predicting crime rates. To test the robustness of our SVM we used leave one out cross validation (LOOCV). Our LOOCV error was 25% using just these two features. Our training error on the entire set was 5/40 = 12.5%. We see that the two values are somewhat close, and thus we seem to be making an okay tradeoff between bias and variance, although the overall error is quite high.

We next attempted to run the SVM by including another feature, that is the number of restaurants in each area. However, doing so gave no improvement in the LOOCV error, but increased the number of support vectors and increased the training error. Hence this feature was not useful in helping to predict the crime rates. (Visualization of the data was not shown as there was no good way to show the data and separating hyperplane in 3 dimensions).

Thus we focus our analysis on the 2-feature model and try to increase the the number of training examples.We expanded our area of interest and instead looked at two locations: central area of SF (Japantown + Divisadero) and south SF. We combined these two areas to form another data set with ~1000 training examples. We now used a cutoff of 30 crimes. Training on this data set gave us a training error of 16%. The plot of this data is shown below in Fig 4.

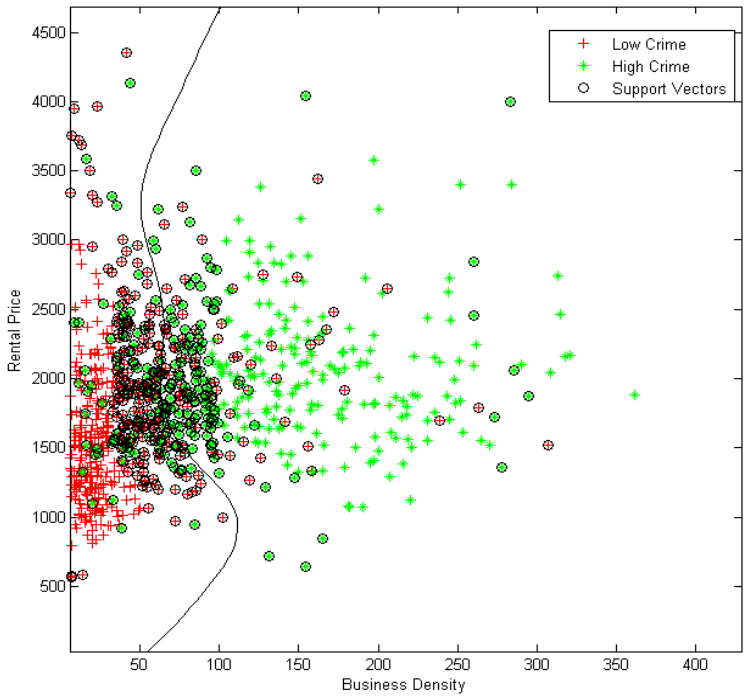

Fig. 4: SVM trained on larger dataset, including data from Central SF and Southern SF

We can see a very similar trend in this data, where it seems rental prices only have a marginal ability to predict crime. However, even the qualitative trends seem consistent with those of the smaller region. Notably, there always seems to be a small "bump" in the center where higher crime exists in areas of median rental price. This may just be variance due to the fitting. If it is true though, it may also be due to many factors: for example, reported crimes may be lower in poorer locations. However, the most important predictor seemed to be business density consistent with our smaller data set.

To obtain a more realistic idea of what the testing error of our model might be, we used the model obtained by training over the data from South San Francisco and Central San Francisco and tested it over SoMa data. Doing so gave us an error rate of 13/40 = 32.5%. Comparing with that of our training error, we see an even greater discrepancy, indicating even higher variance.

Conclusions

The results of our model demonstrate that there is indeed some predictive power (if only by a small margin) in using the aggregate data of business density and housing values. Studying specific businesses, in our case, restaurants, did not improve the predictive power of our model. Further studies that might strengthen this model include using more reliable housing data (which was not available for this project), and taking into account times that businesses stayed open and times of the crime locations. With such data we would be able to get a better model that may be of interest. For instance, one might then consider areas in which bars/clubs which are isolated and open late during the weekday nights to be more prone to crime. Furthermore, there are other factors to control for, including population density, number of visitors to the area, and other socioeconomic factors. The conclusions drawn from our data seem somewhat intuitive and natural, though perhaps a more quantitative analysis would be helpful in determining how much money should be spent in a certain area for crime enforcement.

References

[1]http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display_arch&article_id=1942&issue_id=112009
[2]http://www.economist.com/news/briefing/21582042-it-getting-easier-foresee-wrongdoing-and-spot-likely-wrongdoers-dont-even-think-about-it
[3]http://www.predpol.com/