Modeling San Francisco Arrest Rates by Area and Crime Type

Graham Smith
Katherine Olson
Tyler Gordon
STA 141C, Spring 2017

Abstract

We were interested in whether or not different police departments were biased as regards to certain types of crime. To this end, we built three different models and compared them that used type of crime, police department and zipcode to predict whether or not an incident would result in punitive action. We discovered that, at least with regards to crime overall police department/zipcode was not very predictive, implying that bias between departments is not a significant factor.

Data Preparation

After importing the police department incidents data set and the ZIP Code boundary shapefiles from the Census Bureau, geopandas was used to map the coordinates of each incident to the ZIP Code it occurred in. We are working with 2,063,699 observations from 2003 to 2017.

To simplify prediction we converted the problem to a binary prediction. We considered the following resolution categories in the data set as resulting in an action: 'ARREST, BOOKED', 'ARREST, CITED', 'JUVENILE BOOKED', 'PROSECUTED BY OUTSIDE AGENCY', 'JUVENILE CITED', 'JUVENILE DIVERTED', and 'PROSECUTED FOR LESSER OFFENSE'. All other resolution categories were considered as no action.

Logistic Regression

One of our approaches to modelling was binary logistic regression via the R package glm. Binary logistic regression produces a type of linear model that generates probabilities for a "success" in the response variable based on the values of the predictor variables. In this case, logistic regression was used to predict the probability that an action was taken in response to a crime report. To convert these predicted probabilities into binary predictions (action taken or no action taken), a threshold of

0.5 was applied to the probabilities, with all probabilities over 0.5 resulting in a prediction of action taken.

The model used 80% of the observations for training; the remaining 20% were used to assess the model's accuracy. Prediction accuracy was highest with category of crime, police district, and ZIP code as predictors, at 84.76256%. However, the prediction accuracy of the model using the category of crime as the only predictor was very similar, at 84.74633%. This is a difference of only 00.016233%, which indicates that the category of crime contributes the most to the prediction and that the location of the crime is not very useful to the logistic regression model.

Support Vector Machine

We are predicting action taken (true or false) using three variables that are category of crime (39 types), police department district (11 districts), and zipcode (29). Support vector machine works by plotting each row of the data in a multidimensional space. For this problem, there are three dimensions. Each point plotted represents a feature vector. Then, the hyperplane is found that best separates the classes of data and a prediction is made.

Attempting to run this method on all of our about two million observations would take over a day, so we decided to run this method on 10% of the data. Results can be seen below for each of the 5 trials and the average accuracy for them all. We used the SVC function in the sklearn package in Python.

SVM With Predictors: Category, District, and Zipcode		
Trial	**Accuracy**	**Time**
1	0.846368173669	674.075945
2	0.843678829287	665.093293
3	0.846828511896	675.954412
4	0.843339632699	714.033406
5	0.845035615642	696.17842

SVM With Predictors: Category, District, and Zipcode	
Total Number of Trials	**Average Accuracy**
5	0.84505015263

Running SVM with each predictor on its own gives the following results:

SVM With Each Predictor Alone	
Trial	**Accuracy**
Just category of crime	0.848887919756
Just police department district	0.690386199545
Just zipcode	0.674104763289

This shows how dependent the original model is on the category of crime.

Random Forest

Random forests are an ensemble learning method for classification (in our case), regression and other tasks, that operate by constructing a multitude of decision trees outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. This is done to correct for decision trees' habit of overfitting. I found that police district and zipcode do not add much to the predictive strength of the model For comparison, while the accuracy rate with all three predictors was roughly 85% with a log-loss of .3565, when crime category was removed those numbers dropped to 72% and .5768. However, the accuracy rate was still marginally higher with all three predictors than it was with either category/zipcode or category/police department by themselves. These findings track with what was seen in the other models - that crime category is overwhelmingly more important to predicting punitive action than either of the other factors, but that including them in the model is nonetheless marginally more effective.

Conclusion

Method	Accuracy
Logistic Regression	0.8476256
Support Vector Machine	0.8450502
Random Forest	0.8528589

All three methods achieve a similar accuracy. While unimportance of location data to accuracy indicates little disparity in treatment of same crimes in different locations in the city, this may be in part because there are relatively few predictors. That being said, it might be worth followup research on individual types of crime more likely to be affected by bias (say, domestic or sexual violence for example).