Predicting Violent and Nonviolent Crime based on Location

Andrew Lee and Martin Yim
CS 591 K1 Tools & Techniques for Data Mining and Applications

Introduction

Using open Boston government datasets, we attempted to make models that can accurately represent crime in the city of Boston. By understanding patterns of criminal and unlawful behavior in urban environments, these predictions can have many implications on urban, social and public policy discussions.

Goals

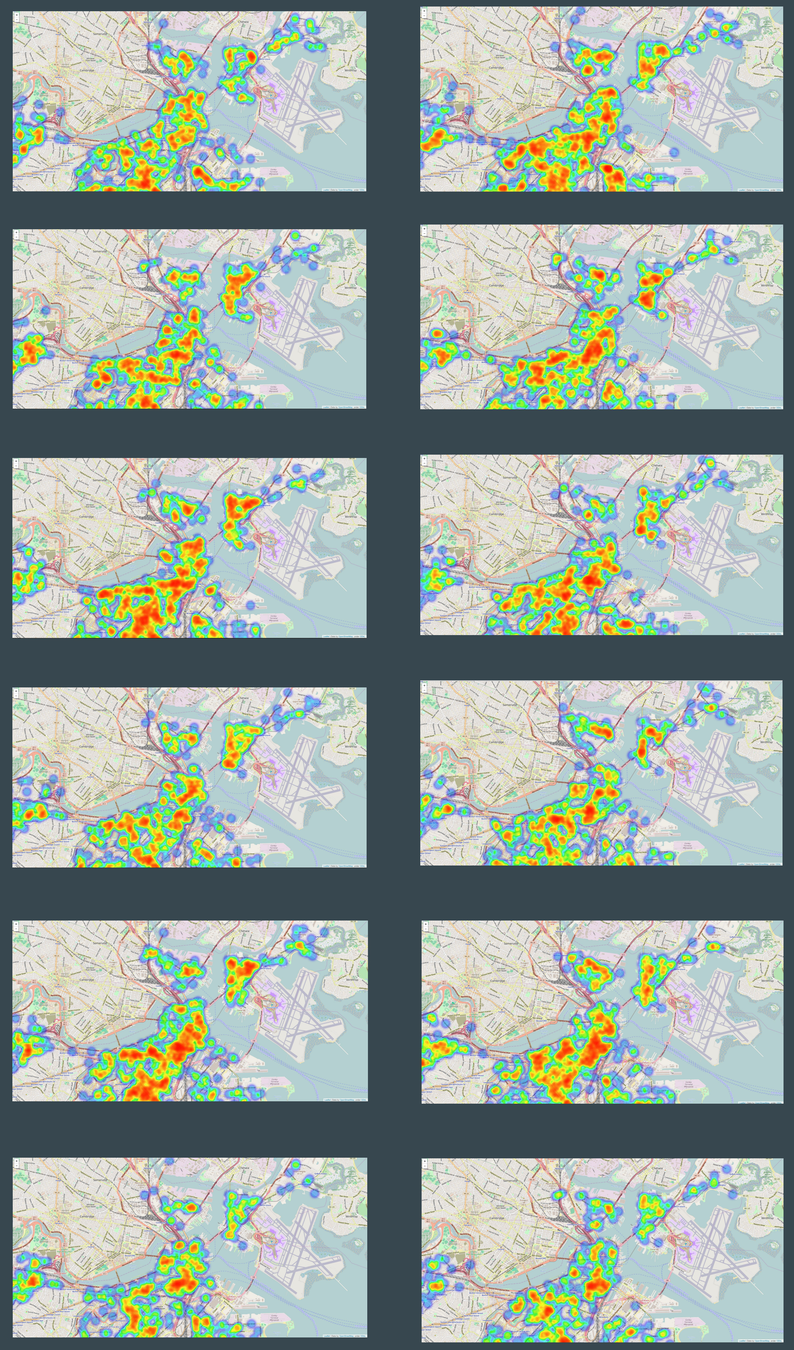
We hoped to use location and distance of crimes from places of potential significance to create a classifier to predict where a crime is more likely to be violent than nonviolent. Our secondary objective was to examine trends of violent crime location through visual exploration.

The Data

We pulled data from multiple datasets on the data.cityofboston.gov website. Our datasets include criminal incidents and locations of hospitals, police stations, establishments with liquor permits and schools. Locations were given as latitude and longitude in the data.

In the criminal incidents dataset we categorized crimes as violent and nonviolent based on the attribute “WEAPONTYPE”; where “unarmed” was nonviolent and all else (knife, firearm, fists, etc.) were considered violent. Year and month of each incident were also used in the analysis.

Data Sources
Criminal Incidents: https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports/7cdf-6fgx
Hospital Locations: https://data.cityofboston.gov/Public-Health/Hospital-Locations/46f7-2snz
Liquor Permits: https://data.cityofboston.gov/Permitting/Issued-Permits-ALL-TYPES-/2hre-tvqe
Police Stations: https://data.cityofboston.gov/Public-Safety/Boston-Police-District-Stations/23yb-cufe



Heat maps read from left to right and top down. Each heatmap represents three months of violent crime. The date range is from July 2012 to April 2015

Logit Regression Results

Dep. Variable:	y	No. Observations:	126539
Model:	Logit	Df Residuals:	126534
Method:	MLE	Df Model:	4
Date:	Wed, 27 Apr 2016	Pseudo R-squ.:	0.002584
Time:	00:28:41	Log-Likelihood:	-73454.
converged:	True	LL-Null:	-73644.
		LLR p-value:	4.356e-81

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
x1	0.0779	0.017	4.457	0.000	0.044	0.112
x2	0.2759	0.034	8.152	0.000	0.210	0.342
x3	0.0698	0.007	10.480	0.000	0.057	0.083
x4	-0.4003	0.038	-10.648	0.000	-0.474	-0.327
x5	-0.0006	8.17e-06	-67.944	0.000	-0.001	-0.001

x1 = Proximity to nearest Police Station
x2 = Proximity to nearest Establishment with Liquor Permit
x3 = Proximity to nearest Hospital
x4 = Proximity to nearest Public School
x5 = Year of Incident

Conclusion

After running a logistic regression based on our data we found that each of the predictors were statistically significant based on their p-values and confidence intervals with 95% confidence. However, the coefficients of each predictor were small and the explanatory power of our classifier is weak. This suggests that although our predictors may play a role in predicting if a crime will be violent based on location that role is easily overshadowed by other variables that we did not include in our model.

Omitted variables bias may explain why our predictors were significant but our model had little predictive power. Further analysis with other data such as local employment statistics or local cost of living may yield in a stronger model.