Analysis of crime in London

Braulio Grana Gutiérrez
bgg@kth.se

Adrián Ramírez del Río
adrianra@kth.se

January 14, 2017

Contents

List of Figures

Abstract

Nowadays, cities are producing a huge amount of data from several sources: traffic, air pollution, hospitals and energy among others. In the biggest and more advanced cities this data is even publicly available via APIs (Application Programming Interfaces). We believe that all those data can be put to good use to analyze the great problems that european cities are facing nowadays.

With that objective we have decided to analyze the problem of criminality as it is one of the most common problems in big cities and one of the most impactful in citizens' lives. In order to do that we are gathering data from official sources in London to identify correlations and patterns that can help us find the factors that are most relevant for each kind of crime. Notice that correlation does not imply causality so we are not claiming the factors we find to be causes of high crime rates, but if we were to find certain patterns of correlations repeating over the years of data it could give hints about where the legislators should look for the source of the problems.

Lastly, we are doing this research with the hope that it will become a useful source of information for politicians when trying to know where to put their efforts in order to reduce the problems of criminality in big cities. As we hope for the information we generate to actually be used, we are being extremely careful not to draw conclusions that may lead to generate hatred against certain groups of people living in the cities.

1 Introduction

Currently there is a huge demand for extracting useful information generated by modern societies both on the internet and in the physical world. This demand does not come from the private sector only as more and more government institutions are starting their own Data Science projects to take advantage of the tremendous amount of data they already manage.

Among these government institutions, city halls are among the most eager as they manage large amounts of data, especially in big capital cities such as Stockholm, London or Berlin. Cities nowadays gather data through millions of monitoring devices spread across their territory and administrative processes. The data collected ranges from environmental, such as air or water pollution, to more day-to-day affairs like unemployment rates, birth rates or energy consumption. These data are publicly available for many european cities coming from official government sources, although some of them still do not have it available through a digital access point such as an API.

Using all these data, we plan to study the relevant factors for criminality in London, as it has a huge database of datasets from a large range of years. Having such diverse amount of data we will be able to find relevant factors for high crime rates and, depending on the granularity of the city's data we could relate these factors to specific types of crimes.

We do not state direct cause-effect conclusions since finding correlations is not the same as finding causes, but if we find repeating patterns associated to high criminality rates, the information could be of great use for future legislators looking to take action into this matter.

2 Theoretical framework

It has been hard to find literature related to the proposed topic, which make us think it is interesting and effort should be put in order to study it further. The majority of the found bibliography focus either on predicting crime data by means of machine learning [2] and statistical learning [3] or on finding specific crime patterns that can help crime analysts to identify series of related crimes [4].

It comes to our minds that there is a lack in the understanding of why crime is present and what factors should be looked at in order to understand and prevent it. Thus our focus has been analyzing the data with the methods for crime prediction shown in previous work [1] (or some similar predictive methods) and further examine the rsulting predictive models in order to find important variables that they focus on when performing the predictive task.

3 Research questions

We state the hypothesis that higher crime rates have associated factors which depend on the type of crime, and that these factor can be identified by means of statistical methods and modelling.

4 Method

To test our hypothesis, we have performed a quantitative analysis based on previously collected data (secondary analysis) where we have applied visualization techniques and statistical modelling. To do so, we carried out a classification analysis to predict which boroughs of London have had criminalities

above the average and in which years. Invastigating later the importance that the model had given to each input variable, we identified the relevant factors on high criminality for different types of crimes in London between the years 2005 and 2014.

We have followed the CRISP-DM methodology [5] which is specifically designed for data mining processes in the industrial environment. In our special case, where we are not interested in a commercial advantage of the obtained results (we are only performing a research study) some of the phases have been accordingly modified. The "Business Understanding" phase corresponds to the understanding of the theoretical framework and literature. For the "Evaluation" phase, we just draw conclusions from the results, and finally, no "Deployment" phase has been necessary.

Note that this is an iterative methodology, and as so we have performed several passes over the data and the generated results.

5 Results and Analysis

5.1 Data gathering and preprocessing

We have gathered data from official sources (APIs) from London open data portal. We have selected and cleaned the following datasets between 2005 and 2014 as part of our analysis:

- Employment/Unemployment rates
- Births by birthplaces of mother
- Births and fertility rates
- Population estimates

- Average income of taxpayers
- Number and density of dwellings
- Carbon emissions
- Business demographics
- Crime data
- Job seeker allowances

The dataset were initially extremely different between each other in terms of structure, time periods, periodicity and format. We used several Python and Bash scripts to clean and preprocess the dataset to a point where they could be merged together using the Python programming language. Lastly, we normalized the population data to a per 100k basis since it is the most common measure for crime rates.

	Borough	Year	Murder	Wounding/GBH	Assault With Injury	Common Assault	Offensive Weapon	Harassment	Other Violence	Violence Against The Person	...	Mean Tax Payer Income (£)	Median Tax Payer Income (£)	Total Job seekers
271	Barking and Dagenham	2006	4	143	2441	1191	153	1190	166	5288	...	20600	17900	16250
16	Hounslow	2014	0	566	1311	1496	67	1570	218	5228	...	35800	23700	13900
280	Hounslow	2006	4	131	2396	1581	151	1704	467	6434	...	26900	18800	13830
191	Enfield	2009	6	261	1669	1117	181	794	205	4233	...	20800	15900	34560
300	Waltham Forest	2005	3	155	1702	2358	313	1444	266	6241	...	20800	16600	23330

Figure 1: Merged dataset

Once we collected, preprocessed, cleaned and merged the data from all the datasets above we proceeded to perform a descriptive analysis by means of visualization of our data to give the reader a better understanding of the dataset's structure and some variables.

The overall structure has a year and a borough per row and then all the different variables that we have gathered, this way we can perform a classi-

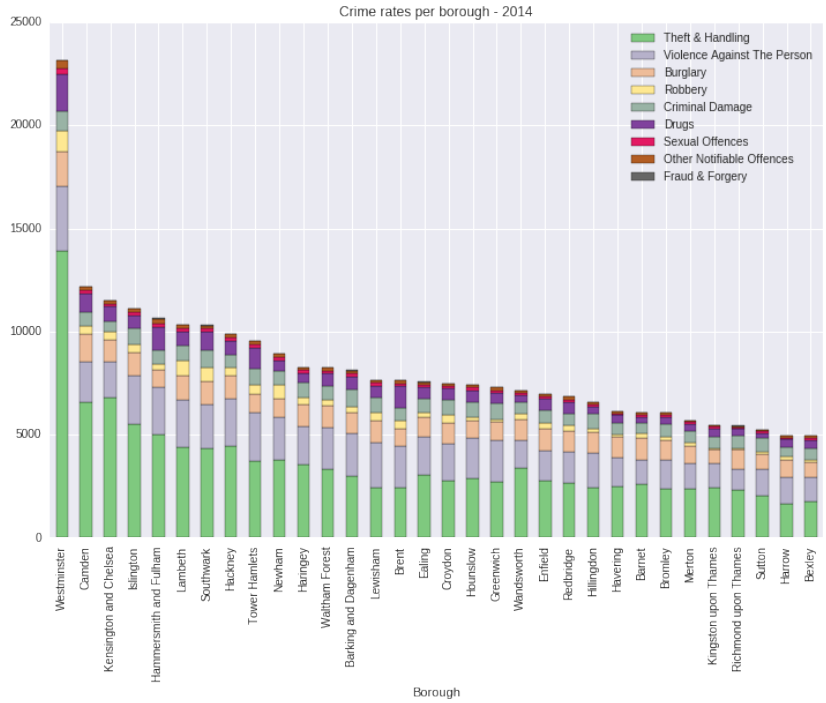

Figure 2: Crime rates over boroughs in 2014

fication analysis of the city over the course of a decade. We have performed some simple exploratory analysis.

So, in total, we are dealing with 9 crime categories, 32 subcategories, across 32 boroughs for 10 years in a yearly basis. Therefore, we have 320 observations in our final dataset.

5.2 Modeling

For the modeling part we applied the same process to the 9 crime categories:

1. Categorize observations using a binary categorization divided by a threshold chosen as the overall mean of the 320 observations since we

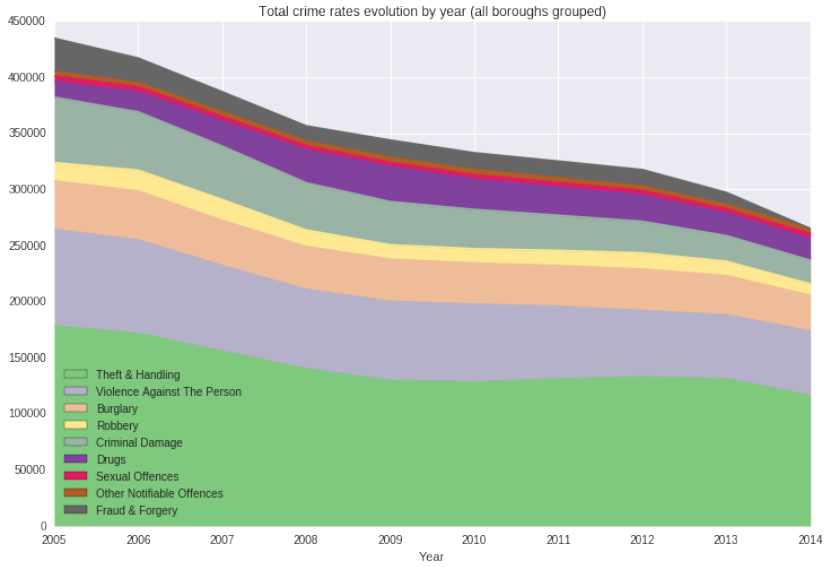

Figure 3: Crime evolution per year in all boroughs

did not find any official threshold values for high crime rates.

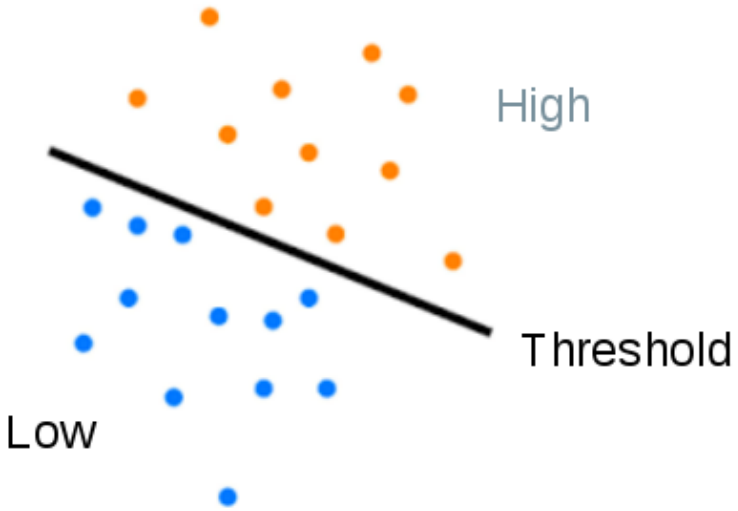

Figure 4: Binary class threshold

2. Divide data in train/test: In our case we used a 75% training and 25% testing ratio. The sampling was done using a technique called stratified random sampling, that allowed us to get proportional samples from each of the types of crime.

Figure 5: Stratified random sampling example

3. Train the model using a random forest classifier [6] with Python's machine learning library scikit-learn. For this part we allowed bootstraping and we used all the features (variables) in our dataset. We also weighted observations based on class imbalance. Note this algorithm is sensitive to a series of hyperparameters (impurity function, size of the tree, etc). For simplicity, we just took them as the default values provided by scikit-learn which tend to work fairly well. The more general way to select these parameters would be to apply some kind of space search with cross validation, but it was left out of the scope

of our research since the majority of the models had a good enough performance.

4. Evaluate the model by trying it against the testing dataset and extracting useful measures like its confusion matrix.

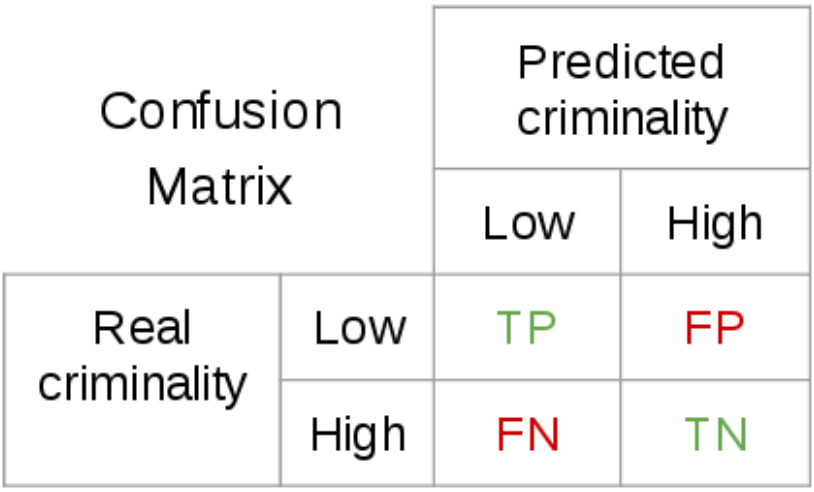

Figure 6: Confusion matrix

5. Retrain the model with the same parameters on the whole dataset as this allows better calculation of feature importance.

6. Compute factor importance using the gini importance measure [7] (not to be confounded with the gini index), which orders factors according to their importance.

Here we show the results obtained for the crime class violence against the person as an example:

For this type of class we can see that Active enterprises, Domestic Emissions and Dwellings per hectare identified as important factors for the classification algorithm.

F1-Score 0.79		Predicted criminality	
		Low	High
Real criminality	Low	37	6
	High	9	28

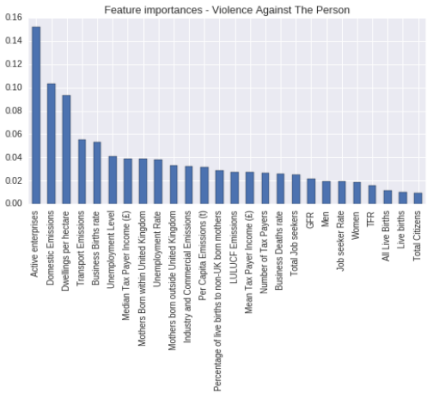

Figure 7: Results for violence against the person

The complete results for each type of crime can be found in the code. Appendix B has a link to the Github repository containing it. An important result to notice, is the differences between the importances from one type of crime to another, as we had hypothesized.

6 Discussion

In this section, we would like to discuss the moral implications of handling data coming from citizens, even when these data are just aggregates they still link to certain groups of people by their ethnicity, nationality, religion, income and many other factors. For this reason have been extremely careful not to draw wrong conclusions that may hurt the image or even the lives of the citizens we are analyzing.

This said, we have guaranteed that all our data come from official governmental sources so we can be confident that it has not been contaminated with biases by third interested parties.

7 Conclusions

After completing the work we can conclude that, as we had hypothesized, some factors seem to work better in order to explain high criminality rates, and these depend on the type of crime considered.

Also, the factors found aren't necessarily the cause of the criminality even when they are somehow related, so expert crime analysts and sociologists should look into them and try to find why they correlate. Sometimes, the considered factors are not good enough to explain the high crime rates, so the produced models should be evaluated (and some discarded) in order to avoid drawing wrong conclusions.

8 Future work

As for future work and improvements in the current work, we spotted possibilities:

- Include more data sources with factors that could be relevant for the analysis.
- Include more data sources with criminality rates of other areas or even other cities.
- Further investigate the crime analysis literature to see how criminality rates thresholds are usually applied in order to differentiate high/low criminality areas.
- Try performing the analysis with more specific types of crime.
- Tune model's hyperparameters with cross validation.

Appendices

Appendix A

Dataset links

- Employment/Unemployment rates
- Births by birthplaces of mother
- Births and fertility rates
- Population estimates
- Average income of taxpayers
- Number and density of dwellings
- Carbon emissions
- Business demographics
- Crime data
- Job seeker allowances

Appendix B

Code

Github repository: `https://github.com/AdrianRamirezRio/crimeAnalysis`

Bibliography

[1] Powers, D. M. W., "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". 2011, Journal of Machine Learning Technologies 2: 37-63.

[2] Lawrence McClendon and Natarajan Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data". 2015, Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.1.

[3] Berk, R., Sherman, L., Barnes, G., Kurtz, E., Ahlman, L., "Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning". 2009, Journal of the Royal Statistical Society: Series A (Statistics in Society) 172: 191-211.

[4] Wang, T., Rudin C., Wagner D., Sevieri R., "Learning to Detect Patterns of Crime". 2013, Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science, vol 8190.

[5] Shearer, C., "The CRISP-DM model: the new blueprint for data mining". 2000, Journal of data warehousing, 5(4), 13-22.

[6] Breiman, L., "Random forests". 2001, Machine learning, 45(1):5–32.

[7] Breiman, L., Freidman, J.H., Olsen, R.A., Stone, C.G., "Classification and regression trees". 1984, Belmont, California: Wadsworth International Group.