Coursework 3 - Predicting Emergency Crime and Safety Incidences in London (Draft)

Name

email

Name

email

Name

email

Name

email

Name

email

Introduction

Imagine, a police force that could respond and de-escalate incidences before anyone is hurt, fire brigades that ensure minimal property damage and less lives lost to traffic accidents. We believe that this goal is attainable, if emergency response teams were able to predict incidences and thus act to prevent them or minimize their impact. This project aims at providing a tool to make such predictions.

Objectives

* Predict the probability of crime occurring at a given time and location in London.
* Fire
* Accident
* Develop an interface through which emergency response teams can access these predictions.

Data Collection

*Crime Data*

The data used for the crime predictions was obtained from three sources:

London Stop and Search Dataset [1]

This is the main dataset used for the predictions and is a dataset provided by the British Home Office. It consists of stop and search reports from 2014 to mid 2017 and has the following field shown in Table 1 below



Table 1 Stop and search fields

LSOA Atlas [2]

This dataset was provided by the greater London authority and consists of a list of all of the Lower Super Output Areas (LSOAs) in London along with demographic information for several years. However, we used it as a list of all the LSOA codes in London so we could join the crime dataset, to the post codes data.

London LSOA API [3]

As the third data source we used an API provided by Chris Bell on his website doogal.co.uk. The API makes use of OS, Royal Mail and National Statistics data to provide statistics on all of the post codes within the specified LSOA code, including the coordinates, population and number of households within each post code. We iterated the LSOA Atlas to query the API and construct a postcode dataset.

Data Processing

*Crime Data*

The objectives of data processing for the crime data was to clean, join the datasets and to create a set of labels to use for training a machine learning algorithm.

*Cleaning the Stop and Search Dataset*

Cleaning of the stop and search dataset involved removing missing values. Figure 1 below shows the distribution of missing values among the columns.

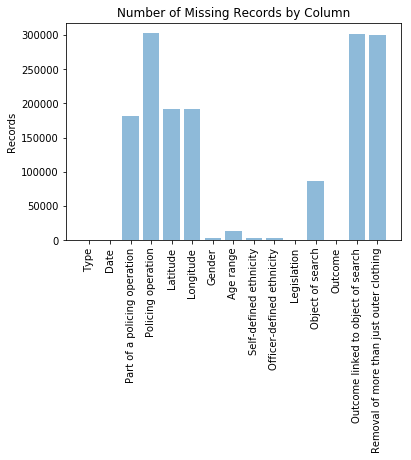


Figure 1 Distribution of missing records for stop and search dataset

To clean the stop and search dataset, we first removed the completely empty columns and removed the age, gender and race columns. We also removed the rows without coordinates, however this had the unfortunate side effect of eliminating all of the thefts from the dataset. Finally, the datetime string field was tokenised.

*Cleaning the Post Code dataset*

Cleaning the post code dataset involved removing the post codes that were no longer in use as they were replaced by newer post codes already in the dataset.

*Joining the datasets*

The population and the number of households was added to the stop and search dataset. These were taken from the post code dataset, rather than the LSOA dataset, because the post code dataset gave these features for more precise areas. To add them to the data set, we found the post codes within 0.001 latitude and longitude of the location of the stop and search and took the mean of the population and number of households for them. This was done for each data point in the stop and search dataset. The rationale behind this was to add information regarding the number of people living within an area of a crime as we wanted to investigate whether this impacts the volume or types of crime.

*Creating Labels*

The type of crime committed was inferred from the object of search, the legislation, outcome related to search. Code was written, based on combinations of these fields to generate the type of crime committed, if any at all. The type of crime committed was then used as the label.

After the pre-processing of the datasets, the final set of features were as follows:

Longitude, Latitude, Hour, Minute, Weekday, Month, Day of Month, Average Population, Average Households and Label.

The longitude and latitude were selected as we wanted to investigate how crime differs from area to area. We also wanted to investigate what times throughout the day are crimes likely to occur. We also suspected that crime would be more prevalent during the weekend, and during holidays and so the weekday, month and day of month were selected. Finally, we believed that crime would be more prevalent in more densely populated areas and thus we included the average population and average households to confirm.