Capstone Proposal

Sandeep Aswathnarayana
October 6, 2020

Proposal ¶

Topic: London Borough-Level Crime Analysis, Segmentation and Clustering of Neighborhoods with Feasible Recommendations for Small Businesses, Residents, Tourists, and Visitors.

A family planning to build new home, a potential business owner or contractor looking to setup their office, an international student looking for affordable housing facilities or a traveller interested in exploring the city - all have a pivotal factor in common to be taken into consideration (irrespective of the budget/affordability) before they make a decision: 'SAFETY'.

In this project, we will study in detail the segmenting, clustering and classification of London Boroughs using Metropolitan Police Service data and Foursquare Developer API. As the city grows and develops, it becomes increasingly important to examine and understand it quantitiatively. The MPS provides open data for Developers, Investors, Policy Makers, City Planners who possess an interest in answering the following questions for development and safety of residents:

- What neighbourhoods have the highest crime?
- Is population density correlated to crime level?
- Using Foursquare data, what venues are most common in different locations within the city?
- Does London Datastore provide with specific enough or thick enough data to empower decisions to be made? Or is it too aggregate to provide value in its current detail? Let's find out.

Domain Background

I have chosen to analyze, address, and predict the most frequent crime types in London boroughs and feasible recommendations for small businesses, tourists or visitors based on the characteristics of the venues or neighborhoods.

The idea for the Capstone Project is to show that when driven by venue and location data from FourSquare, backed up with open source crime data, that it is possible to present the cautious and nervous traveller with a list of attractions to visit supplementd with a graphics showing the occurance of crime in the region of the venue.

A general, broader perspective of this project:

Individuals or Businesses decide on a City of intent to move in or set up their offices (London, in this case) The Foursquare API is scrapped for the top venues in the city From this list of top venues the list is augmented with additional grographical data Based on the additional geographical data the nearby, relatively safer locations of interest are chosen The historical crime within a predetermined distance of all venues are obtained A map is presented to the to the individual, potential business owner or traveller with the selected venues and crime statistics of the area

Why is this project important

- In May 2019 the total number of offences in London was 75717 - up 6.1% from the same month last year and up 1.2% from the previous month.
- Over the last 12 months, 25 of London's 32 Boroughs have seen an increase in the number of crimes committed compared with the previous 12 months. The biggest increase was seen in Westminster (up 23.3%) while the largest fall in crime was seen in Islington (down 6.8%).
- The number of crimes on the transport network in the quarter Jul-Sept 2019/20 (9,043) was up 15.2 per cent compared with the same quarter in 2018/19 (7,853).
- The London Datastore - Greater London Authority enables us to gain an understanding of the crime volume by type by area but not specific enough to understand the distribution properties.
- Valuable questions such as, "are these crimes occuring more often in a specific area and at a certain time by a specific demographic of people?" cannot be answered nor explored due to what is reasonably assumed to be personal and private information with associated legal risks.
- There is value to the city to explore the detailed crime data using data science to predict frequency, location, timing and conditions to best allocated resources for the benefit of its citizens and it's police force. However, human behaviour is complex requiring thick profile data by individual and the conditions surrounding the event(s).

It is of utmost importance to address and meet the requirements of the following sectors or target audience:

1. Professionals relocating to London for work - The number of workers in London is projected to increase by 582,000 (up 10%) in the next 10 years. This is the equivalent to about 58,200 more jobs each year. Apprenticeship starts in London more than doubled between 2009/10 and 2010/11 and since then have maintained a fairly steady level.
2. International Students & Overseas Nationals - There were over 5 million international visitors to London in 2019 Q2, up 2.3% from the same quarter a year before. In 2018/19, there were over 240,000 new National Insurance Number (NINo) registrations from overseas nationals in London, which was 2% higher than the year before.
3. Small Businesses - With a potential opportunity to establish their footprint, small businesses benefit from the spending by local communities, travellers and visitors. Total spend by international visitors alone in London was at £12.1bn in 2018 down 10.2% on 2017. Spending in the most recent quarter (2019 Q1) was 1.8%, 6.0% higher compared with the same quarter a year previous.
4. Policy Makers - By grouping the neighbourhoods into most similar groups, the GLA and Mayors Office for Policing and Crime (MOPAC) have enabled both the police and the public to understand performance compared to similar neighbourhoods in London.
5. Law Enforcement - Police and partners now share best practice with like-for-like neighbourhoods. The Metropolitan Police Service actively utilises the Neighbourhood Confidence and Crime Comparator at neighbourhood level to identify similar areas where public confidence metrics vary the most, and respond to the challenge of MOPAC to reduce these gaps.

6. Families and Communities - The population of London in 2017 was 8,904,000 up 7% from five years ago. The population is projected to increase to 9.7 million by 2025 (an increase of 17% from the 2011 Census) and reach 10 million by 2030.
7. Housing & Real Estate - In 2019/20 the total number of GLA funded affordable housing starts in London was 12,546 accounting for 90.6% of total London housing starts. In 2018/19 London's dwelling stock saw a net increase of 35,959 dwellings compared to the year before.
8. Voluntary Crime Prevention Movements - Ex: Neighborhood Watch

Problem Statement

In this project, we will study in detail the segmenting, clustering and classification of London Boroughs using Metropolitan Police Service data and Foursquare Developer API. As the city grows and develops, it becomes increasingly important to examine and understand it quantitiatively. The MPS provides open data for Developers, Investors, Policy Makers, City Planners who possess an interest in answering the several questions of concern for development and safety of residents.

A better Police Service for London Wherever you live in the capital, and whatever your background, you should expect the same service from the Metropolitan Police Service (MPS).

Providing the best service to all Londoners is at the heart of the Police and Crime Plan and means getting the important things right: making communities safer, responding to and preventing crime, building trust and confidence, and bringing criminals to justice.

Machine Learning Task:
KMeans Clusteting Algorithm: K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are: a. The centroids of the K clusters, which can be used to label new data b. Labels for the training data (each data point is assigned to a single cluster)

Addressing this problem is feasible using KMeans since the categorical values in the final wrangled dataframe (obtained by applying data tranformation on the three datasets) are converted into numerical values using 'One Hot Encoding'.

For class label distribution, I used 'One Hot Encoding' technique to generate the features and labels in the numerical format for Clustering Analysis. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. It allows the representation of categorical data to be more expressive. I am planning to extract the top venues after performing One Hot Encoding and K-Means Clustering on the dataframe. Then, with the help of 'folium' map, I want to potentially classifify the neighborhoods based on the borough rankings obtained form the clustering results. Finally, address the target audience based on the available cluster characteristics to meet their specific requirements.

Datasets and Inputs

3.1. MPS Borough Level Crime Data:
MPS Borough Level Crime Data (https://data.london.gov.uk/dataset/recorded_crime_summary) counts the number of crimes in London at the borough-level per month, based on the crime type.

The data is available in two files for each level of geography - the most up to date data covering the last available 24 months only and one covering all historic full calendar years. To analyze the most recent patterns, I opted to explore the one with the last available 24 months.

In March 2019, the Metropolitan Police Service started to provide offences grouped by the updated Home Office crime classifications (https://www.gov.uk/government/publications/counting-rules-for-recorded-crime). This currently only covers the most recent 24 months of data.

Below is a list of the crime types covered under the new HO categories:

Major Category - Minor Category:

- Arson and Criminal Damage - Arson / Criminal Damage
- Burglary: Burglary - Business and Community / Burglary - Residential**
- Drug Offences: Drug Trafficking / Possession of Drugs
- Miscellaneous Crimes Against Society: Absconding from Lawful Custody / Bail Offences / Bigamy / Concealing an Infant Death Close to Birth / Dangerous Driving / Disclosure, Obstruction, False or Misleading State / Exploitation of Prostitution / Forgery or Use of Drug Prescription / Fraud or Forgery Associated with Driver Records / Going Equipped for Stealing / Handling Stolen Goods / Making, Supplying or Possessing Articles for use i / Obscene Publications / Offender Management Act / Other Forgery / Other Notifiable Offences / Perjury / Perverting Course of Justice / Possession of False Documents / Profitting From or Concealing Proceeds of Crime / Soliciting for Prostitution / Threat or Possession With Intent to Commit Crimina / Wildlife Crime
- Possession of Weapons: Other Firearm Offences / Possession of Firearm with Intent / Possession of Firearms Offences / Possession of Other Weapon / Possession of Article with Blade or Point
- Public Order Offences: Other Offences Against the State, or Public Order / Public Fear Alarm or Distress / Racially or Religiously Aggravated Public Fear / Violent Disorder
- Robbery: Robbery of Business Property / Robbery of Personal Property
- Sexual Offences*: Other Sexual Offences / Rape
- Theft: Bicycle Theft / Other Theft / Shoplifting / Theft from Person
- Vehicle Offences: Aggravated Vehicle Taking / Interfering with a Motor Vehicle / Theft from a Motor Vehicle / Theft or Taking of a Motor Vehicle
- Violence Against the Person: Homicide / Violence with Injury / Violence without Injury

To note:

Fraud data was transferred from individual police forces to National Action Fraud in March 2013

**Prior to April 2017, police recorded burglary offence categories were split such that dwellings (domestic burglary) and buildings other than dwellings (non-domestic burglary) were separately identifiable, where:

domestic burglary covers residential premises, including attached buildings such as garages non-domestic burglary covers non-residential premises, including businesses and public buildings, as well as non-attached buildings within the grounds of a dwelling, such as sheds and detached garages From April 2017 onwards a new classification of police recorded burglary was introduced, dividing offences into two categories of “residential” and “business and community” “Residential” burglary includes all buildings or parts of buildings that are within the boundary of, or form a part of, a dwelling and includes the dwelling itself, vacant dwellings, sheds, garages, outhouses, summer houses and any other structure that meets the definition of a building. It also includes other premises used for residential purposes such as houseboats, residential care homes and hostels.

“Business and community” burglary includes all buildings or parts of buildings that are used solely and exclusively for business purposes or are otherwise entirely outside the classification of residential burglary.

3.2. List of London boroughs:
The motive behind using this dataset is to fill the gap that our actual crime dataset lacks to address i.e., extract the key attributes or columns from the List of London boroughs (https://en.wikipedia.org/wiki/List_of_London_boroughs) dataset will help us extract and analyze the attributes including Population Density for each of the 32 boroughs and their respective Co-ordinates.

3.3. Foursquare Location Data:
The Foursquare Venues & Places Database (https://developer.foursquare.com/docs/api/venues/details) gives the full details about a venue including location, tips, and categories. We can access precise, up-to-date community-sourced venue data. Its large selection of rich and firmographic location data unlocks the potential to enhance our app or website with the ability to describe locations, analyze trends, and improve user experience.

If the venue ID given is one that has been merged into another venue, the response will show data about the other venue instead of giving you an error. User authenticated calls will also receive information about who is here now. This is a Premium endpoint with access to venue's photos, tips, hours, menu, categories, recommendations, events, stats, etc.

Using these 3 major datasets as the basis for our project, let's start leveraging its features and attributes to address our business problem.

Solution Statement & Approach

The main advantage of using K-Means Clustering is the fact that K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Based on our business problem i.e., crime analysis and predictions, knowing and having access to the most remote and unknown terrains in the neighborhoods is pivotal to the lawmakers and policy makers to be proactive in their strategy.

The post Clustering phase includes the Segemnating and Classification of clusters i.e., london neighborhoods based on the metrics including Safety Metrics, Average Monthly Crime, Cluster Numbers (for top venues and neighborhoods), Frequency aligned with each venue and other numeric results based on the given metrics and datasets. This provides a comprehensive yet simple-to-present method of providing feasible predictions and recommendations to the policy makers for any necessary changes that needs to be addressed in their existing strategy or approach. The less technical the final presentation or metrics are, the easier it is for the target audience and citizens to comprehend the final results of our model.

Benchmark Model

One of the benchmarks I have evaluated and made significant imporvements over is the 'Predicting London Crime Rates Using Machine Learning' project. This project has used the same MPS London Crime Data to try to make machine learning crime predictions.

This project attampts to make two different predictive models: crime month-by-month at the LSOA level and crime type (whether burglary, bicycle theft, arson, etc.) month-by-month at the LSOA level. LSOA is a census area containing 1,000 to 3,000 people. It also has made efforts to enrich the dataset with various open data sources, added the police station coordinates, added post codes, and inputted POIs and the LSOA statistics.

Some of the potential shortcomings in this project:

- Based on the datasets used, predictive model results, type of plugins used, I believe that this proejct only addresses the factors like 'Crime' and 'Type of crime' on a monthly basis in its final results.
- The predictive model generated here is leaned more towards the Data Analytics approach as opposed to addressing major factors influencing the crime using an algorithm.
- This project only addresses the 'obvious'. It doesn't necessarily provide insights into the reasoning and neighborhoods with feasible recommendations and suggestions to the policymakers, visitors, or small businesses.
- The correlation matrix generated is of no/little use since it addresses the already obvious fact i.e., the density of POIs correlated in part with a higher number of crimes. Or to quote the author, "I was expecting a direct relationship between the number of shops, restaurants, tourist attractions, etc. and the number of crimes committed".
- The primary objective of this analysis has been the 'volume of crime' and the 'type of crime' in its final results.

I plan to address the drawbacks or shortcomings of this above model using my approach and technique using Machine Learning.

Evaluation Metrics & Project Design

The benchmark model only generated KMeans clusters with no further reasoning or understanding of the clusters, maps, venues, geographical data, target audience, etc.

On the contrary, I am planning to produce meaningful results which are direct applications for use by small businesses, residents or tourists in the form of a deployed application on Foursqaure using Foursquare Developer API.

Also, I am using the 2 additional datasets to address the drawbacks in the MPS crime dataset. Feature Engineering is a crucial part in any ML task. I have made the best use of he features extracted from the 3 datasets to produce numerical results:

Applying K-Means Clustering to the Wrangled Crime dataframe:
The major steps in this process include:

- Data Cleaning and Transformation (say, One Hot Encoding in our case)
- Choose the initial K value and run the algorithm
- Review the results
- Ierate over several values of K

Feature Engineering:
Feature engineering is the process of using domain knowledge to choose which data metrics to input as features into a machine learning algorithm. Feature engineering plays a key role in K-means clustering; using meaningful features that capture the variability of the data is essential for the algorithm to find all of the naturally-occurring groups.

Categorical data (i.e., category labels such as Borough Name, Venue Name, Venue Type, CrimeToPop, etc.) needs to be encoded or separated in a way that can still work with the algorithm.

Feature transformations, particularly to represent rates rather than measurements, can help to normalize the data. Based on the cluster results and inferences obtained, the top 10 venues and top 5 clusters were generated by assessing the characteristics of each venue type, location, and borough.

Alternatives:

One way to address the potential shortcoming in this project is to try and experiment with the alternatives to KMeans Clusetring algorithm.

A number of alternative clustering algorithms (https://blogs.oracle.com/datascience/when-k-means-clustering-fails%3a-alternatives-for-segmenting-noisy-data) exist including DBScan, spectral clustering, and modeling with Gaussian mixtures. A dimensionality reduction technique, such as principal component analysis, can be used to separate groups of patterns in data. You can read more about alternatives to K-means in this post (https://blogs.oracle.com/datascience/when-k-means-clustering-fails%3a-alternatives-for-segmenting-noisy-data).

One possible outcome is that there are no organic clusters in the data; instead, all of the data fall along the continuous feature ranges within one single group. In this case, you may need to revisit the data features to see if different measurements need to be included or a feature transformation would better represent the variability in the data. In addition, you may want to impose categories or labels based on domain knowledge and modify your analysis approach.

For more information on K-means clustering, visit the scikit learn site (https://scikit-learn.org/stable/modules/clustering.html#k-means).

Pipeline:

The brief overview of the pipeline and methodology:

Importing necessary libraries and loading the data sets of interest Examine the crime frequency by neighbourhood Study the crime types and then pivot analysis of crime type frequency by boroughs Understand correlation between crimes and population density Perform K-Means Clustering

Analysis on venues by locations of interest based on findings from crimes and boroughs Determine the venues which are in the proximity of relatively high crime count and choose the locations of interest accordingly

In [ ]: