**Capstone Project: Evaluating Food Restaurant Feasibility in London, United Kingdom using k-Means Clustering**

Joanna Cez Dela Cruz

September 7, 2020

1. **Introduction**
   1. **Background**

Food restaurants has been a booming business all over the world – from fast food chains, formal dine-in setup, and on-the-go restaurants. Many international food corporation from US and other countries are investing in London, United Kingdom and expands their network there. There are also small corporations that are willing to risk in expanding their business in Europe. In this instant, it is important to understand the market before pushing through the business expansion

* 1. **Problem**

It is notable that with the large land area of London, United Kingdom, building the first branch of a particular food restaurant is essential and critical for it to create buzz and expand more in the future. It is essential to determine which neighborhoods are the busy streets and can target the correct audience of their food restaurant. Can a business really bloom in a neighborhood where there are restaurants everywhere or should it build its own first restaurant with less competition. Moreover, we would also like to forecast on how long can a business survive in the neighborhood.

* 1. **Interest**

This project is just an overview of the capability of Data Science in aiding and guiding business decision makers on their investing and marketing strategies.

1. **Data acquisition and cleaning**
   1. **Data sources**

List of Neighbourhoods in London are sourced from here: <https://en.wikipedia.org/wiki/List_of_areas_of_London>. This will be matched with the postcode and location data sourced from here: <https://www.doogal.co.uk/AdministrativeAreas.php>. After matching the two sources, venues around neighbourhood is sourced out from Foursquare API. Demographic profiles are extracted in here: <https://data.london.gov.uk/dataset/london-borough-profiles#:~:text=The%20London%20Borough%20Profiles%20help,borough%2C%20alongside%20relevant%20comparator%20areas>.

* 1. **Data cleaning**

First, List of neighborhoods are web-scraped in the above link. It contains the location, neighborhood, post town, postcode district, etc. It consists of 532 unique locations. We then selected only location, borough, post town, and post code. Subscripts and unnecessary characters are also scraped out. Next is to split out locations with multiple boroughs. The final output in this data is Borough and its corresponding location.

Second, location data is webscraped and only filter the Latitude, longitude, and administrative area. Administrative area is also the same as borough and neighborhood. Neighborhood data and location data are merged via administrative area – resulting 524 boroughs.

To validate the location data, the neighborhood-location data is mapped over the London map.

Venues around neighborhood using Foursquare API. Neighborhood-Location data is used as input data. It will reiterate and list all available venues per neighborhood. The venues are then categorized to Bus stops, supermarkets, pub, park, etc. We will use this later to characterize the neighborhood.

Per neighborhood, all venue categories are identified. It will then be grouped together and finally identify the 1st common venue, 2nd common venue, and so on. This is now the final input for K Means Clustering.

There are several demographic variables in the source but we will only be using two-year business survival rate as the dependent variable. Average rate will be computed and another variable will be created where it will be 1 if two-year business survival rate is above average rate and 0 otherwise.

* 1. **Feature Selection**

Because of the data limitations and availability, only 2014 to 2017 are the data available. Below are the variables that will be used as explanatory variables for the average business survival rate: GLA Population Estimate 2017, Population density 2017 (per hectare), Average age 2017, Proportion of population aged 0-15 2015, Proportion of population of working age 2015, Proportion of population aged 65 and over 2015, employment rate 2015, unemployment rate 2015, gross annual pay 2015, modelled household median income estimates 2012\_2013, Number of jobs by workplace, jobs density 2015, number of active businesses 2015, crime rates per thousand population 2014\_2015, fires per thousand population 2014, Ambulance incidence per hundred population 2014, total carbon emissions 2014, Number of cars 2011 census, and Average public transport accessibility score 2014.

1. **Exploratory Data Analysis**
   1. **Foursquare API**

There are mul

* 1. **Demography Variables**

1. **Modelling**
   1. **K Means Clustering**

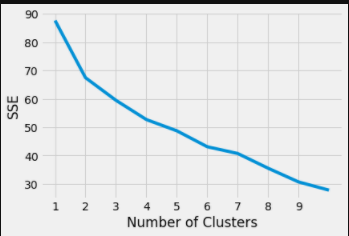
After identifying nth common venues per neighborhood, it was then used to do 10 reitaretaions of k-means clustering to identify best value of k via elbow method. It was then found that k=4 is the best value in clustering London neighborhoods. See below Elbow result.

Figure 1. Elbow method of neighborhood clustering in London, United Kingdom

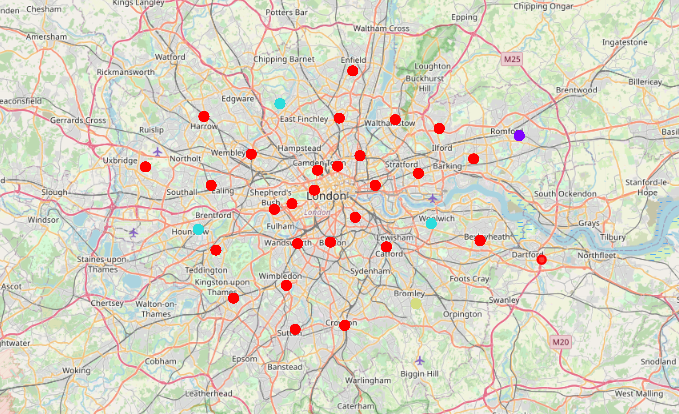
 Using k=4, neighborhoods are clustered into four groups and superimposed in he London, United Kingdom map. Each color represents a different cluster. See below K-Means cluster result:

Figure 2. Four Clusters of Neighborhood in London, United Kingdom

Cluster 1(in red) are the largest cluster in London, United Kingdom. It consists of 31 neighborhood. Below is the tabulation of common venue categories in cluster 1.

Table 1. Venue neighborhood counts in Cluster 1.

|  |  |
| --- | --- |
| **Venue** | **Neighborhoods with nearby common venues** |
| Pub | 18 |
| Park | 13 |
| Café | 13 |
| Furniture / Home Store | 13 |
| French Restaurant | 12 |
| Fried Chicken Joint | 11 |
| Fruit & Vegetable Store | 11 |
| Coffee Shop | 10 |
| Yoga Studio | 10 |
| Gaming Cafe | 9 |
| Pizza Place | 8 |
| Grocery Store | 8 |
| Indian Restaurant | 7 |
| Bus Stop | 7 |

Most common venues are Pub, stores, and restaurants. It is usually characterized as one of the busiest streets in London. This can be a good prospective when building a food restaurant.

Cluster 2 consists only one neighborhood which is in Havering. Its top three most common venues are: Park, Yoga Studio, and Food Court. It can be near Cluster 1 characteristics but is recreational-centric.

Cluster 3, on the other hand, consists of three neighborhoods which are: Hounslow, Greenwich, and Barnet. The most common venues are: Bus Stops and Stores. Although this is a good prospective also in which case there is less competitor, we will not be delving into this because of the few numbers of clusters involved.

Lastly, Cluster 4 only consist of one neighborhood which is Bromley. Its ost common venues are Soccer Stadium, Gym centers, and Park. This is a more of recreational-centric neighborhood and might not be less target market in the area.

* 1. **Logistic Regression Model in target Cluster**

From these clusters, we will only focus on Cluster 1 with the most activities and foot steps in the area. Although there are lots of competitors, it is a good area to start branding of a food restaurant. Next step is to do logistic regression.

The average two-year survival rate across London, United Kingdom is 78% while these clusters have 50% in average.

* 1. **Model score**

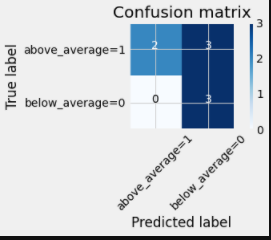
Model accuracy is only at 62% given the feature selected in the demographic data. Precision in identifying above average is greater than precision in identifying below average business survival rate. To identify more of the model score, below is the confusion matrix of the model:

Figure 3. Confusion matrix of Cluster 1 Logistic regression model where 1= above average.

It can be noted that with the True above average neighborhoods, only 2 out of 5 are correctly tagged by the model. All below average are correctly identified by the model. Although we cannot say that the model can accurately identify whether the business can survive above average in cluster 1, we can identify and use True Negatives to classify, and describe the least priority neighborhoods in London, United Kingdom.

1. **Conclusion**

It is therefore concluded that k-Means clustering is a good method in identifying cluster characteristics of neighborhood in London, United Kingdom. It is best to build a new food restaurant in Cluster 1 with Pub, Stores, and Bus Stops nearby as it is a good indicator of good business booming around the area.

It is also concluded that even with most businesses are booming in this area, it is most of the time with below average of two-year business survival rate in London, United Kingdom.

1. **Future Directions**

It is suggested to be able to get richer data for the demographic variables and do a feature selection method to further identify the best model.

It is also suggested to create a different metric as threshold decision other than business survival rate.