**Capstone Project – The Battle of Neighborhoods**

Diep, Ly Bao Long

1. **Introduction/ Business Problem**
2. **Background**

* It is undeniable that Canada is one of the worth-living countries all over the world because of having diverse ethnicity and offering a whole host of great opportunities for immigrants to develop. Significantly, Toronto and Vancouver have gained the reputation for being the two main economic cities in Ontario province and it is evident that these cities have a high standard of living and employment rate, making it very attractive and commonplace to inhabit down the road. Since Toronto and Vancouver are famous metropolitan areas, there has been a bunch of comprehensive analyses and surveys about them, and they will not be on the agenda of this project. As regards the ideal living areas, I will pay my attention to the other cities in Canada, particularly London city which is part of Ontario province and one of the potential candidate cities. Dissimilarly, London is a city with a moderate area of about 420.6 km2 in the southern part of Ontario. More intriguing, London is not densely populated, and it is estimated that the total population is approximately 404,699 in 2017 compared to 2.93 million in Toronto in 2017, which will contribute to reducing the severe competition in job markets. Moreover, the economic resources and wealth in London primarily derive from medical research, education, insurance, and information technology which demonstrates that London also has good social welfare to attract newcomers, and sufficiently creates a whole host of employment opportunities with prospective promotions to nourish and foster an abundant workforce. Thanks to these good social and living factors, London can be considered to meet several standard living conditions at least. The overview about London has been listed out, but another issue which is also far more important to deal with is which neighborhood in London is the best to settle down and initiate a business.

1. **Problem**

* **Problem definition:** Accommodation and start-up problem in London, Ontario.
* **Circumstance:** In this capstone, as an upcoming data scientist, I will reach out to stakeholders who are looking for a living area which is convenient and well-located with good public amenities and service in London, Ontario. Specifically, we are going to explore the neighborhoods and specify the number of ideal residential areas in London. What’s more, the stakeholders also have no clue for what kind of effective business patterns they should initiate, so they need some handy suggestions for some potential businesses in the recommended areas.
* **Assumption:** It is obvious that there are a number of influential factors such as the stakeholder's affordability for the cost of living and the price of the real estate in the desired areas. Since this capstone primarily focus on using Foursquare to discover the most fantastic living area in this capstone, it will be assumed that these factors are in the available budget of the stakeholders.
* A data scientist will manipulate the power of data to generate the most feasible and promising neighborhoods based on the listed above criteria. It will be expected that the upsides and downsides will be also comprehensively laid out so that the best deliverables can be used to help the stakeholders make their final decision.

**\*\* Agenda \*\***

* What neighborhood is the best to settle down in London, Ontario?
* How far is it from the living area to its surrounded venues?
* What kind of potential business pattern should be recommended?

1. **Target Audience**

* This project will significantly draw the attention of the potential stakeholders who have a desire to settle down in a second homeland and run their own business in a residential area with good living conditions. To be more precise, the stakeholders can be tenants, new immigrants, real estate companies, international settlers, home investors, etc.

1. **Data acquisition and cleaning**
2. **Data Collection**

* Based on the defined circumstance, there will be a number of factors that will have impacts on the decisions and analyses:
* The number of neighborhoods that need to be taken a look at in London, Ontario.
* The distance from the living areas to the other venues within the scope of neighborhood.
* The number of available business patterns within the neighborhood.
* As listed above, the following data sources will be needed to generate the required information:
* The location and coordinates of each neighborhood in London, Ontario will be scraped from [**http://www.geonames.org/postalcode-search.html?q=london&country=CA&adminCode1=ON&fbclid=IwAR2XipWkuSm3F9YSjjVvFqp7SfYCPl9\_XaxiehoPnn-7XmsjtnJBrbKh31g**](http://www.geonames.org/postalcode-search.html?q=london&country=CA&adminCode1=ON&fbclid=IwAR2XipWkuSm3F9YSjjVvFqp7SfYCPl9_XaxiehoPnn-7XmsjtnJBrbKh31g) by using **Pandas function/ BeautifulSoup**.
* The number of venues and their categories within each neighborhood are extracted by using **FourSquare API**.
* **The extracted venue categories** can be used as a **separated dataset** to build the **recommender system** for the suggestion of business patterns.

1. **Data Refinement and Formatting.**

* Since the required data scraping from the available dataset on webpage is originally raw, some data transformation and refinement will have to be carried out for the easily visual analysis and exploration.
* The dataset used in this analysis should be in tabular form and consist of a number of following columns: **Postal Code, Borough, Neighborhood, Latitude, Longitude.**

**Sample table:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Postal Code** | **Borough** | **Neighborhood** | **Latitude** | **Longitude** |
| 1 |  |  |  |  |  |
| n |  |  |  |  |  |

* Each row contains different data values corresponding to each column.

**Illustration for the original dataset**

* Table

  Description automatically generatedAs shown by the data table above, each individual row is filled with distinct data, but it is not in the desired form which is mentioned above. More significantly, the “Place” column seems to be messy because one labelled row contains not only the combination of borough and its neighborhood enclosed in parenthesis but also its latitude and longitude right in the unlabelled row below. The “Code” columns will be converted into the “Postal Code” column and the rest of the columns are not requisite and will be eliminated.
* **BeautifulSoup** library will be the efficient and easy-to-use tool in scraping this dataset. The rows containing the combination of borough and neighborhood will be separately derived and converted into two distinct columns (Borough and Neighborhood). Similarly, the coordinate values in the unlabelled rows will be also separated into two numeric columns (Latitude, Longitude).
* Finishing filtering out the above dataset, we have a complete refined data table, and it is ready to be used in the next stage of analytic process.

**Table

Description automatically generated\*\*NOTE\*\*:** There are **17 obtained neighborhoods** in London, Ontario in total.

1. **Methodology**

* Data visualization, data exploration, data analysis and any machine learning methods utilizing for making recommendation will be comprehensively performed in this section.
* Since the required dataset for the neighborhoods in London have been thoroughly gathered **(postal code, borough, neighborhood, latitude, and longitude)**, the exploration of venues within **the radius of 1000 meters** from the default coordinate of neighborhoods can be carried out by using **Foursquare API.**
* Firstly, to make the analytic process more insightful, the **Folium package** will be utilized for the visualization of the geographic positions of all the neighborhoods in London.
* The second step of the analysis will be **the calculation and exploration of** **the distribution of venue categories** across different areas of London. The analysis will mainly focus on the most promising areas and within those create different clusters of neighborhoods that meet some basic requirements established in the discussion with stakeholders:
  1. Considering the areas with **\*\*high density in the scatter of venues and the diversity of venue categories**\*\*.
  2. Presenting a geographic map of all such locations and creating clusters (applying \***\*k-means clustering\*\*** of Machine Learning) of those locations to identify the candidate neighborhoods based on their similarity and dissimilarity and search for optimal areas.
* In the next step, the recommendation system for the trustworthy candidates will be built based on the exploratory venues within different neighborhoods in London. In addition, the use of the Toronto dataset as an additional reference is to make a comparison between the venue categories in both cities so that the business patterns which have never been exploited in London can be taken into consideration for the recommendation.
* Finally, the optimal solution will be ready to handle the stakeholder's problems.
  1. **Data Visualization**

Map of neighborhoods in London

Map

Description automatically generated

* The default geographic coordinates of London city are 42.9836747, -81.2496068.
* Glancing at the above map, it is remarkable that the distribution of neighborhoods becomes denser and denser when it gradually exposes to the central area of London. According to this observation, it is predicted that there will be more exploratory venues within the closer neighborhoods rather than the ones which are located in or near the outskirt of London city.
  1. **Data Exploration**
* Foursquare application:
* Since the main purpose of this project is to optimize the useful application of Foursquare API which is one of the most handy and powerful Application Programming Interface tools for the geographic exploration, the utility of Foursquare will be fully exploited in the exploratory process. Now, the number of local venues in each individual neighborhood will be completely investigating.
* The neighborhood investigation only occurs within the limited radius which is set to 1000 meters from the center of the neighborhood.
* To get the venue data of each neighborhood from Foursquare, the unique client information (ID, Access Token) and the available coordinate (latitude, longitude) of each neighborhood are requisite, but the more technical part of how to create and use Foursquare API will not be mentioned here, only the final result after the neighborhood exploration. It is remarkable that the Foursquare will basically make use of the latitude and longitude of the designated area to detect the approachable local venues within the limited radius.
* At the end of the neighborhood detection, the expected data’s features that have to be captured for the later analysis will be the name of venues, the venue latitude, the venue longitude, and the venue categories.
* Illustration for the exploration:

Text, letter

Description automatically generated

* Table

  Description automatically generatedIllustration for the expected result:
* Observably, the Foursquare has done 17 neighborhood explorations based on 17 neighborhoods and after the neighborhood investigation, there are 346 rows containing the exploratory venues of different neighborhoods and 8 columns containing the venues’ corresponding features. However, this result table is messy, and it will be very arduous and time-consuming to accumulate and examine the total number of venues that one neighborhood has as there is a whole host of exploratory venues in one distinct neighborhood.
* To have a more visual overview after the exploration, each venue will be grouped by their corresponding Postal Code and Neighborhood.
* Categorical table

Table

Description automatically generated

* The above data table actually gives a more profound visualization for the neighborhood exploration. Applying coding will give the summation result of 120 unique venue categories which are approachable in London.
* Looking at the illustration above, there are a number of significant points:
  1. **Missing neighborhoods/ Outliers.**
* There are only 15 neighborhoods with approachable local venues left after the exploratory process while 17 neighborhoods have been examined in the neighborhood exploration. There are two missing neighborhoods, which are N6M, N6P.Graphical user interface, text

  Description automatically generated
* It is evident that there is no approachable venue that can be detected within the limited radius from the neighborhoods with the postal code N6M and N6P. These two missing neighborhoods can be considered as outliers from the others because of not providing the expected information and meeting the criteria mentioned in this project and are eliminated from the candidate list because of the inadequate standard for the next analysis.
* Only 15 remaining candidate areas will be taken into consideration and it is expected that the most promising living areas in London, Ontario can be discovered at the end.
  1. **Neighborhood with sparse venue density**
* Some neighborhoods are extremely poor in the number of approachable local venues, which are N6E, N5V, N5W, N5X, N6H, N6L, and N6N. Nevertheless, it is advantageous that these neighborhoods can be potentially got rid of the candidate list in the long run due to being unqualified, so the best ones will be unearthed.
  1. **Data Analysis**
* In this section, the analytic techniques, mathematic calculations, and machine learning algorithms served to perform statistic testing for the analysis will be comprehensively elaborated. To be more specific, the key methods include the one-hot coding technique, K-Means clustering (ML), and 2D Cartesian distance calculation.
* **One-hot coding**
* It is sometimes unknown that the “One-hot Coding” is one of the most efficient and easy-to-use techniques which can perform the conversion between the multiple categories or sorts which are in string values and categorical values which are in numeric format (1 if the category is available, otherwise 0).
* As indicated by the categorical table above, there are numerous venue categories detected in one neighborhood, so the application of “One-hot Coding” method will give an overall observation of venue categories which are accessible in a neighborhood.

Table

Description automatically generated

* The “One-hot Coding” table depicts 346 rows standing for different available categorical values of each venue and 122 columns will be the categorical features.
* At this point, the inquiry of how frequently the presence of a venue category will be in a neighborhood will be what have to be resolved. In other words, the frequency/ frequency distribution scores of each venue category need to be computed. The calculation of frequency score is fundamentally the mean/ average of the total categorical values of one venue category divided by the total number of venue categories in one neighborhood. This mean ratio indicates the frequent proportion (the quantity/ the distribution) of one venue category within its neighborhood. The greater the ratio is, the greater number of one venue category in a neighborhood.
* Frequency/ Frequency distribution score.

Chart

Description automatically generated with medium confidence

* Table

  Description automatically generatedTable

  Description automatically generatedLooking at the first row of the venue (row 0), the venue category with “Rental Car Location” has the highest frequency distribution score, demonstrating that there will be lots of venues that provide rental car services in the area of YXU / North and East Argyle / East Huron Heights. It is significant that the lowest score is equal to 0, implying that there is no lawyer office located in the neighborhood with the postal code, N5V. In general, the area with the postal code, N5V, has only 4 approachable venue categories, which reveals the minor number in the distribution of venue categories in this area.
* Table

  Description automatically generatedThoroughly observing the above the frequency distribution score of all venue categories of the neighborhoods, some remarkable points are found:
* A number of potential candidate areas with a variety of venue categories.
* Unqualified neighborhoods which are hardly diverse in the distribution of venue categories.
* The top popular venue categories which are based on the frequency distribution score sorted in the descending order.
* Subsequently, only the most 10 commonplace venues categories are extracted for the candidate examination. This is a sub-step before getting into the more mathematic analysis.

Table

Description automatically generated

**Personal observation:** it is discernible that several neighborhoods have a diverse distribution in venue categories, but it is also easy to notice that some areas which contain NaN values (unavailable data) are sparse in the number of accessible venue categories such as N6L, N6N, N6H, N5V, N5W, and N5X, which do not meet the required criteria. Undoubtedly, these unqualified neighborhoods can be potentially got rid of from the candidate list later. To be more exact, a clustering method, K-Means, which is an algorithm of Machine Learning will be taken into practice to identify the potential candidate.

* **K-Means clustering**
* K-Means is an unsupervised learning method of machine learning. The primary purpose of utilizing K-Means is to partition the neighborhoods into a number of k clusters based on the similarity and dissimilarity so that the trustworthy candidates can be detected. However, the flaw of K-Means is that it requires a number of k initial random centroids (clusters) for the features of the given dataset, but it is very challenging to get the best k number for the optimal results.
* Fortunately, the elbow method can solve all of this problem. Firstly, the estimated interval for the k initial clusters will be set from 1 to 11 for the elbow testing. A visual graph will be needed in this circumstance.

Elbow diagram

Chart, line chart

Description automatically generated

* In the elbow diagram above, the horizontal axis represents the number of k from 1 to 11, and the vertical axis indicates the Sum of Squared Error (SSE). Observing the sketched plot above, it is discernible that the Sum of Squared Error keeps decreasing while the K number is increasing. The point at which the drew line is greatly bent is equal to 4. This is the elbow of the line at the point with k = 4. This also interpret the functionality of how the elbow method. The dataset containing potential candidate areas will be divided into 4 clusters.
* According to the code,
* **2D Cartesian distance calculation**
  1. **Recommendation**

1. **Result and Discussion**
2. **Conclusion**

**Reference**