**Data Science Professional - Applied Data Science Capstone – Andre Loney**

**Feature Selection for Relocation Assistance**

1. **Introduction**

One of the main challenges of permanent relocation is determining whether you'll have lifelong love for your residential area. Personally, I have lived in 3 different states (Connecticut, Florida, Maryland) and a handful of cities, but deciding where home will be for the majority of life has been extremely stressful. Thus, I am curious about the similarities and dissimilarities between the neighborhoods of the city I live in currently (Baltimore), and a few prospective cities, primarily Atlanta and Seattle. The Foursquare data allows a high-level view of the structure of the local neighborhoods and the benefits they provide.

The main objective of this project is to create the first step in designing a neural network that would provide the opportunity of feature selection to aid:

* Future prospective re-locators support in deciding which location would be ideal for permanent relocation.
* Creation of a dataset that a recommender system could use to predict landings spots.
* Further adjustment would allow realignment of the data to extend to vacationers as well.
* Individual companies offering relocation assistance will be able to use this to attract potential candidates.
* Travel sites to advertise as a feature to aid in vacation planning.

1. **Data**

The data that will be used is as follows:

* Foursquare location data
  + Neighborhoods with Venues extracted
* Various websites for Neighborhoods
  + <https://livebaltimore.com/neighborhoods/>
  + <https://arcgis.atlantaregional.com/arcgis/rest/services/OpenData/FeatureServer/196/query?where=1%3D1&outFields=NEIGHBORHO&returnGeometry=false&outSR=4326&f=json>'
  + http://clerk.ci.seattle.wa.us/~public/nmaps/neiglist.htm

The exploration of the data will include the compiling of dataframes with the neighborhoods and venues, but further we will compare and contrast the neighborhoods across the multiple cities in hopes to find a solid landing spot according to particular features.

The data was scraped and extracted from 3 very disparate websites. Originally, the data was to be extracted from Wikipedia pages, but upon further review the formatting on such pages left the neighborhoods data in difficult formats and/or variables which posed quite the dilemma. Variables within html sites are not as simple and straightforward to extract as originally anticipated since they are embedded deep in the back-end. After exploring many packages and modules, it was abandoned for various methods that would allow a more stream-lined extraction approach, as well as provide tools for my Data Science toolbelt.

The Foursquare API provides the geolocation data (coordinates). Anaconda.org provided the packages to install the required modules for proper assessment.

1. **Methodology**

As this was the initial foray into this analysis, it was crucial to get as strong of a representation of the neighborhoods as possible. Thus, exploring the government and community-curated sites for the actual names of the neighborhoods was vitally important. This provided an accurate overview of the names and locations of the neighborhoods in comparison of the cities. Nonetheless, **Baltimore** data was scraped from a website with an embedded map.

The data was parsed from html and arranged in a dataframe consisting of the neighborhoods. The neighborhoods state and/or country was appended to the end of the neighborhood to ease geolocation. The geolocation provided a long-form description of the neighborhood consisting of the name of the neighborhood, county, state, country, and coordinates (latitude/longitude). This data was then separated by neighborhood and coordinates and joined together. Outliers in the data are removed and a map with markers for each neighborhood is created through Folium.

One dataset is then created out of the compiled dataframes from Baltimore, Atlanta, and Seattle data, and markers added on a global map. A function was created via Foursquare API to get the nearby venues within each neighborhood, limited at 100 venues and a radius of 500 meters. This is passed through the full dataframe including all the neighborhoods and coordinates. A function finding the category of each venue is created and applied to the dataframe with the nearby venues for each in the corresponding neighborhoods. Next, the amount of venues per neighborhood was counted across the neighborhood and venue categories for correctness to ensure data would be viable. 394 unique categories were found at this point. Dummy variables are then created to essentially tally the amount of a particular category of a venue in each neighborhood, and these are averaged to moderate the quantity. The top 5 venues for each neighborhood are ranked and listed. Then a function for the most common venues is created and passed over the data, followed by listing the top 10 most common venues.

The data is then clustered for which, I arbitrarily chose 7 clusters (see Discussion section for explanation). Cluster labels are added, which required the dropping of any neighborhood that returned no venue data. The colored clusters are then added to the USA map to provide a visual for the clusters.

The FourSquare API served equally useful and frustrating in providing the required information. It provided plenty of location and venue data, but missed in finding all the neighborhoods.

1. **Results**

Seven clusters were derived from the provided dataset, one in particular that is quite large.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | | City Qty. | | |
| # | Description | B | A | S |
| 0 | Larger Neighborhoods with many venues. Needs to be assessed further. More variance within than between clusters | 144 | 59 | 54 |
| 1 | The parks of each of the cities. | 11 | 13 | 10 |
| 2 | Lincoln Homes Park did not report all the data. | 0 | 1 | 0 |
| 3 | Mainly residential with a convenience store only in the radius | 2 | 0 | 0 |
| 4 | Residential areas with playgrounds and small delis | 2 | 0 | 1 |
| 5 | Home services businesses, small commercial areas | 4 | 1 | 0 |
| 6 | Bars and small eateries | 9 | 0 | 1 |

The first cluster will need to be further observed. Baltimore reported a considerably higher amount of venues for the neighborhoods, which is potentially due to the density of the city. This may have a chance of skewing the data since it has so much control on the venue size. Incorporation of more cities with comparable density could help regulate the effect that Baltimore has on the rest of the data.

At this point, I do not believe the data provides enough information to suggest that Baltimore, Atlanta or Seattle are more similar to one city than any other.

1. **Discussion**

Performing this analysis in a clustering format serves as the first step in exploration of this fashion. The literature around clustering suggests that clustering is to be performed primarily on numerical data, and otherwise, on smaller datasets than the one I have created. I aimed to cluster the data into a range that was reasonable to draw some valid conclusions, but the true outcome resulted in a few smaller clusters that were reasonably paired and one vastly large cluster that needs more analysis. A potential takeaway in this space, confirms the understanding that there is much similarity between populated cities.

Normalizing the categorical data in order to find the proper quantity of clusters would be important in the future. The dataset gets quite large so trial and error on the clusters is not ideal, but for this assessment this was the methodology followed. The number was chosen to reduce the size of the second largest cluster so conclusions could be drawn. Unfortunately, this resulted in one very large cluster which will be its own explorable cluster reassignment.

Nonetheless, this analysis is the first step in creating a larger, dynamic dataset which would be capable of recommending a location to live or vacation. The next step in this analysis would be to create a recommendation system that would be capable of using the dataset. By creating a system that could take the input of a user’s weighted interests, the reflection of the acquired rating would be related to this dataset.

The bounds of the city were used to simulate the urban, high-population areas of the cities, but neighborhoods vary drastically in size and composition. This affected the clustering as some of the clusters simply did not provide enough data to be useful, while other areas inundated the dataset. I would like to explore this further but limiting the upper and lower bounds of the quantity of returned venues to strengthen the clusters.

In the age of gentrification, the raw dataset would continually need to be evaluated to ensure viability. According to the Baltimore Sun, “Baltimore is among seven U.S. cities that accounted for nearly half the country’s gentrification between 2000 and 2013, according to a new study” (Meehan, S., 2019). The other two comparison cities, were primarily less gentrified (Atlanta more so than Seattle). With creating a larger feature set for the future recommendation system, a new/old city may have an appeal to certain crowds as well. Thus, finding a metric that could note this could be useful as well.

1. **Conclusion**

Overall, the initial exploration of the Baltimore, Atlanta, and Seattle Urban areas to create a Feature Selection dataset for a neural network and recommender system could be useful for a travel or relocation site. Overall, Baltimore is more densely packed thus offering many more venues than the other cities, potentially skewing clusters. More cities added could help dilute the immensity. At this point, with the limitations of the data, it is hard to say one city is more similar to another city.