Applied Data Science Capstone

The Battle of Neighborhoods

Segmenting Madrid

1. Introduction
2. Data
3. Methodology
4. Results
5. Discussion
6. Conclusion

Introduction

"La Milonguita" is a chain of restaurants located in Buenos Aires, Argentina, it offers typical foods of that country. Argentine cuisine has influences of Spain an Italy, and has incorporated Indigenous components also. The owners of this restaurants wants to expand their chain to Europe and they are very interested in establish a new restaurant in Madrid, Spain.

Madrid is a big city with more than 3 millions inhabitants and a metropolitan area population of about 6.5 millions. It's composed by more than one hundred neighborhoods, with different characteristics. In order to get directions about where to establish their first restaurant in Madrid, the owners of "La Milonguita" has asked us to do an study of different places of the city.

Marketing consultants has determined that we must pay special attention to entertainment services, because they are considered attractive to potential customers. We also must take into account other food services offerings that cuold be considered in some cases rivalry or, in other cases, complementary to an Argentinian Restaurante.

Our project will consist in obtain information of the neighborhoods and make recomendations about best places to start the restaurante chain expansion.
That information and recomendations must be reported to "La Milonguita" owners in a clear and simple way.

Data

Required data will be gathered from:

- Madrid information, including districts and neighborhoods, can be obtained from Wikipedia: Madrid (https://en.wikipedia.org/wiki/Madrid)
    - In order to make recomendations about where to establish the restaurant, Madrid segmentation will be made based on different neighborhoods.
    - Full list of neighborhoods can be obtained from Wikipedia Madrid information, but only their names. They must be geolocated ir order to use Foursquare services for obtaining venues.

- For geolocation of neighborhoods *Python geocorder* will be used.
    - Every neighborhood otained will be geolocated using python geocoder package, using neighnorhood name plus city and country.
    - Geocoder returns latitude and longitude information for every neighborhood center, then it will be used as main Foursquare input.

- In order to obtain venues and their categories we will use FOURSQUARE (https://foursquare.com/)
    - Using services provided by Foursquare we can obtain venues for every neighborhood. Such services requires as input geolocalization, it means the latitude and longitude obtained in previously described step.

Methodology

A Jupyter Notebook will be developed in order to process data and segment the neighborhoods. Following steps will be implemented:

1. **Build neighborhoods list**
   A list of districts is obtained from Madrid Wikipedia page. That list contains the names of the neighborhoods for avery district.
   As output a dataset containing a list of *"district,neighborhood"* is build.

1. **Neighborhoods geolocation**
   Every element in the neighborhoods dataset is geolocated using *Python Geolocator* and two columns are updated containing latitude and longitude for each disctrict,neighborhood.
   The geolocator service has some problems, many times gives time out error, for this reason in this step the information obtained is saved in a text file in CSV format.
   Therefore this step can be run many times, invoking geolocator only for missing data (timed out errors in previous executions). After various executions all the neighborhoods geolocation is obtained and we can use the text file.

1. **Venues compilation**
   As next step Foursquare services are used for obtaining venues for every neighborhood. The output is a new dataset with many records for every neighborhood containing the venues found for everyone of them.
   A free Foursquare service with limited count of calls is used. In order to minimize the usage of Foursquare, the information is saved in a text (CSV) file. It's supposed that the information gathered doesn't change in short period of time (some hours). When the analysis must be continued for long period (many hours or next day) just deleting the generated text file will force to call Foursquae services again, refreshing the information.

1. **Neighborhoods segmentation**
   The problem in hand is a case of unsupervised segmentation and, from the possible machine learning algorithms, K-means was choosen.

   - Taking in account that the venues information obtained from Foursquare is categorical, it must be previously processed in order to be handled by K-means algorithm. For this _"pandas.get*dummies"* is used for dummies variables.
   - The list of dummy variables obtained are then grouped as features of every neighborhood.
   - After executing K-means algorithm the "Elbow Curve" it's ploted in order to obtain the *best K*. Analyzing the change in the slope of the curve, it's determined that K=10 is a good value.
   - K-means algorithm is executed.
   - Next step is build the segmentation dataframe, composed of the top venues for every neighborhood plus a segment label determined by K-means.

1. **Segments analysis**
   Every segment is printed individually, were different charateristics can be observed for each group.
   Next section describes the results.

Results

As a result of segmenting Madrid, ten clusters where defined.
Following their characteristics are shown:

- **Cluster 1**
  It's a big cluster where Spanish Restaurants together with "Tapas" Restaurants (typical of Spain) and Bars has a dominant presence. Although there are many other interesting places, like Theaters and pub in example, too many rivalry exists, therefore it doesn't appear to be the best option for an starter business.

- **Cluster 2**
  It's a medium sized cluster with various restaurants and shopping places, but no other attractions places. It doesn't result interesting.

- **Cluster 3, 4 and 5**
  Clustering algorithm produces just one record for each one. Not much useful information.

- **Cluster 6**
  This medium sized cluster results very interesting. There are Spanish Restaurants, Fast Foods and Pizza. And also Theaters and Soccer Fields can be found. Soccer is very related to Argentina, because many Argentinian soccer stars are currently playing in Spanish League.
  This is definitively a very good finding.

- **Cluster 7**
  This cluster has various kind of restaurants, Spanish, tapas, Chinese and Greek. It has also some pub and bars.

- **Cluster 8**
  Only one record in this cluster. Not enough information for making a decision.

- **Cluster 9**
  This cluster has an interesting variety of restaurants, Spanish, Chinese, French, Greek, fast food. Additionally there are bars and theaters. It results interesting to consider for establishing an Argentinian Restaurant.

- **Cluster 10**
  This cluster has many shopping places and some restaurants and entertainments, like pubs and bars.

Discussion

The objective of this project is found places in Madrid City for establishing the first Argentinian Restaurant of the chain in such city.

Main requirements are that other kind of restaurants exists and also entertainments for potential customers.
Applying a given machine learning clustering algorithm was possible to segment neighborhoods based on their venues and, most important, found a group of them that have high potential.

As a result we are in condition of present our recomendations to the owner of the restaurant chain based on concrete data.

Conclusion

We have gathered data from trusted sources and a known and strong methodology has been applied for processing

A group of eleven neighborhood has been selected from more that on hundred that Madrid has.

In such neighborhoods there are Spanish Restaurants, Fast Foods and Pizza. And also Theaters and Soccer Fields can be found.

We consider that in one of them will be able to start the company the next endeavor.

- Arganzuela,Legazpi
- Tetuán,Berruguete
- Fuencarral-El Pardo,Barrio del Pilar
- Fuencarral-El Pardo,Valverde
- Carabanchel,Buenavista
- Carabanchel,Abrantes
- Puente de Vallecas,Portazgo
- Moratalaz,Horcajo
- Hortaleza,Pinar del Rey
- Villa de Vallecas,Santa Eugenia
- Barajas,Corralejos