The Best Place for a New Juice Bar in New York

Pop-Ducheva Ina

May,2020

2. Data Acquisition and Cleaning

2.1 Data Sources

The most comprehensive dataset of location related information is the one hosted by Foursquare, which has 105 mil locations worldwide and has built its dataset on completely crowd sourced information. The dataset is updated in real-time, which means that it can be used for a continuous research of places and their characteristics.

2.2 Data Cleaning

Considering that the dataset is acquired from FourSquare it is structured,which refers to any data that resides in a fixed field within a record or file. Because we want the new juice bar to be located in the center of New York City, we will aquire all venues in a 10000m radius of the geographical coordinates of New York. In order to better facilitate the use of the dataset we will start by replacing all non-existing values with 0.

2.3 Feature selection

After the data cleaning there were 17351 venues and in the dataset, which include venue name, latitude, longitude, borough, neighborhood, address, venue category.

For each neighborhood we will calculate the total number of gyms, fitness centers, boxing gyms, climbing centers, cycle studios, dance studios, farmers markets, fountains, gardens, performing arts venues, pilates studios, pools and yoga studios as location points of interest and organize them in a dataframe. We will also clean of categories of locations which have only one entry, because those can be too specific cases. They can bring unnecessary noise results in the data processing.

Then we will apply the Standardscaler function, which standardizes features by removing the mean and scaling to unit variance. Standardization of a dataset is a common requirement for many machine learning estimators, they might behave badly if the individual features do not more or less look like standard normally distributed data. Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set.

If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

We will use this dataframe first with DBSCAN clustering to create clusters with high density where the aforementioned venues of interest are in high numbers. After we will cross-reference the map of clusters with locations of the existing juice bars in the city to determine where it is best to set up a new place.