Recommending neighborhoods to open food stores

Gabriel Santos

July 26, 2020

1. Introduction

The city of Sao Paulo is the biggest metropolitan area of the South Hemisphere and is the biggest, the most populated and the richest city in Brazil. This city has a lot of italians, spanish and japanese immigrants and there is a great diversification in terms of gastronomy with about 75 different ethnicities. We can go in a luxury restaurant or in a place to snack and drink, at each corner there is a lot of options to support the demand of people in the city.

With the growth of people in the city it's a nice spot to open food stores, but where should you open your store in this city ? The city is huge and has a lot of boroughs with different categories of venues. Search the best place manually it's not the best option to take, so we can use the data and the power of the machine learning to help us finding the best place in the city.

2. Data

This project will rely on public data from "CEP Lá" and Foursquare.

The *Dataset 1* is the treated data from the .csv file generated by the .txt file available at http://cep.la/CEP-dados-2018-latin1.zip that contains all the data from postal codes, cities, boroughs and neighborhoods. I uploaded the csv file inside the jupyter notebook and I inserted to code transforming into a dataframe. With the dataframe I limited to use only data from São Paulo/SP and to use only data from 9 boroughs (Consolação, Bela Vista, Sé, República, Bom Retiro, Brás, Liberdade, Santa Cecília, Cambuci) in the center of the city, the heart of the city. After that I rename the columns. So far this dataset is large so I select randomly 40 neighborhoods in those boroughs and we will analyze this neighborhoods.

I used the "Geocoder" package with "arcgis_geocoder" to obtain the latitude and longitude of the needed locations. Running a loop in the dataset 1 it's possible to obtain the latitude and longitude for each neighborhood and

insert in the dataframe. The dataframe now have these columns: Postcode, Borough, Neighborhood, Latitude, Longitude.

The *Dataset 2* it's the final result applying the Foursquare API in the dataset 1 pushing the nearby venues with the categories from each neighborhood 100 meters around and inserting this venues in the dataframe. The dataframe now have these columns: Postcode, Postcode Latitude, Postcode Longitude, Venue, Venue Latitude, Venue Longitude, Venue Category.