Coursera Capstone

IBM Applied Data Science Capstone

Opening a new Shopping Mall in Milano, Italy

By: Nestor Carmona Moreno
January 2019

Introduction

For many shoppers, going to shopping malls is a great way to relax and enjoy themselves during weekends and holidays as they offer a wide variety of “activities” such as grocery shopping, dining, shopping, watching movies or any other activity. It is safe to say that shopping malls fit at least one person in the world. For retailers, the central location and the crowds at shopping malls makes it easier for them to sell their products and/or services to respond to the increasing demand.

Finally, opening a shopping mall allows property managers to have a consistent source of rental income. Nevertheless, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, its location is one of the key aspects as it can determine the success or failure of the investment.

Business problem

The objective of this capstone project is to analyse and select the best locations in Almeria, Spain, to open a new shopping mall using Data Science methodology and Machine Learning techniques like clustering. This project aims to provide solutions to answer a business problem: In the city of Milano, Italy, if a property manager is looking to open a new shopping mall, in which neighborhood would we recommend them to open?

Target audience of this project

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in Milano, Italy. This project is timely as the city currently offers a limited amount of shopping mall in its most centric neighborhoods.

Data

To solve the problem, we will need the following data:

- List of neighborhoods in Milano, Italy: chose city for this project.
- Latitude and longitude coordinates of those neighborhoods: required in order to plot the map and get the venue data.
- Shopping mall data: required to perform clustering on the neighborhoods.

Sources of the data:

- List of neighborhoods: We will use this Wikipedia page: https://en.wikipedia.org/wiki/Category:Districts_of_Milan using web scraping techniques to extract the data with the help of Beautifulsoup packages and Python requests.
- Latitude and longitude coordinates: we will use Python GeoCoder
- Shopping mall data: we will use Foursquare API to get all the venue data and, in particular, the Shopping Mall category.

This is a project that showcases many data science skills, from web scraping to working with an API, Data Cleaning, Data Wrangling, Machine Learning and Data Visualization.

In the next section, we will present the Methodology where we will discuss the steps taken in order to answer the above business question, all the Data Analysis that we did and the Machine Learning techniques used.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Milano, Italy. Fortunately, this list if available on Wikipedia: https://en.wikipedia.org/wiki/Category:Districts_of_Milan. We will do web scraping using Python requests and BeautifulSoup packages. However, we will only get a list of names and we need geographical coordinates in the form of latitude and longitude to be able to use the Foursquare API. To get the coordinates, we will use the Geocoder package that will allow us to convert an address into geographical coordinates.

After gathering all the data, we will fill up a pandas DataFrame and then visualize the neighborhoods in a map using the Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates returned by Geocoder are correctly plotted in the city of Milano, Italy.

Next, we will use the Foursquare aPI to get the top 100 venues that are within a radius of 2 Km. To do this, we need to register a Foursquare Developer Account in order to obtain an ID and Secret Key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the

returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Shopping Mall” data, we will filter the “Shopping Mall” as venue category for the neighborhoods.

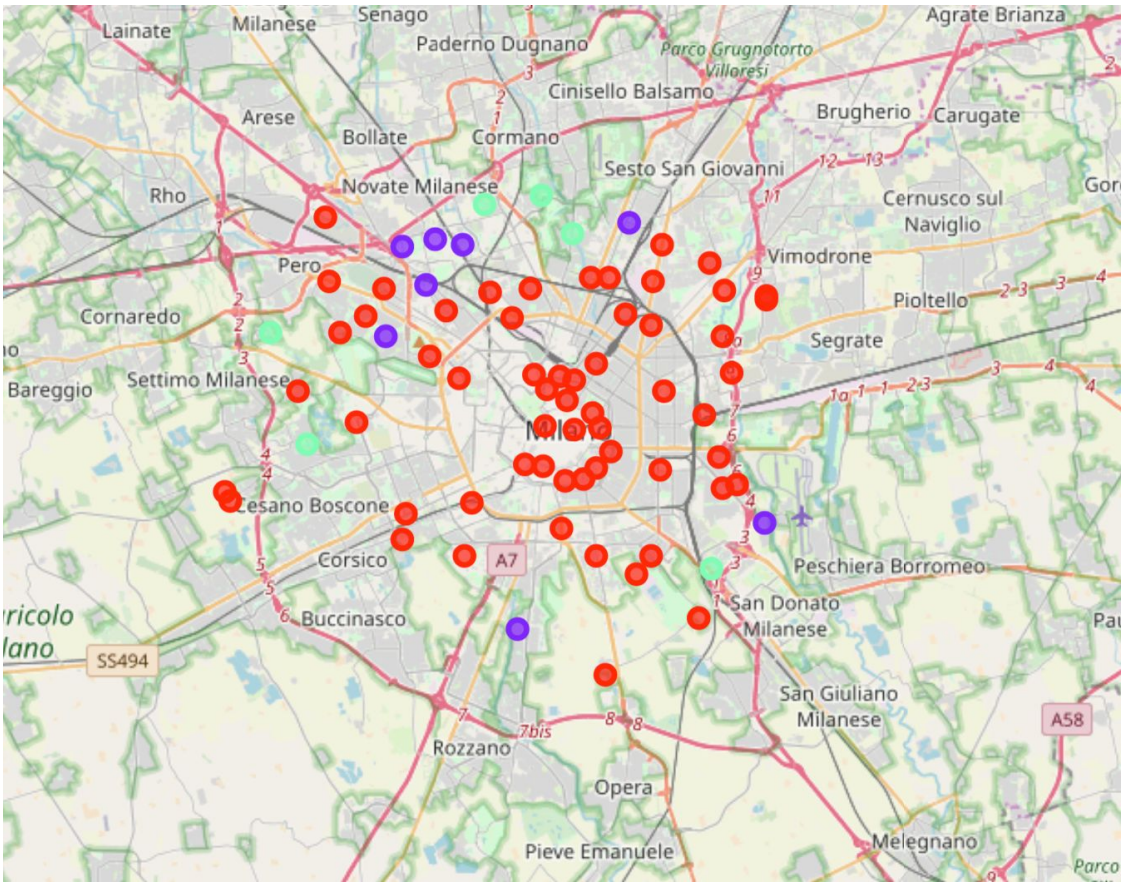
Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Shopping Mall”. The results will allow us to identify which neighbourhoods have higher concentration of shopping malls while which neighbourhoods have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new shopping malls.

Results

The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

- Cluster 0: Neighbourhoods with moderate number of shopping malls.
- Cluster 1: Neighbourhoods with low number to no existence of shopping malls
- Cluster 2: Neighbourhoods with high concentration of shopping malls

The results of the clustering are visualized in the map below with cluster 0 in red colour, cluster 1 in purple colour, and cluster 2 in mint green colour.



Discussion

Most of the shopping malls are concentrated in the suburbs of Milano with the highest number in cluster 2 and moderate number in cluster 1.

On the other hand, cluster 0 has very low number to totally no shopping mall in the neighborhoods. This represents a great opportunity and high potential areas to open a new shopping malls as there is very little to non existent competition from existing malls.

Meanwhile, shopping malls in cluster 2 are very likely suffering from intense competition due to oversupply and high concentration of shopping malls.

Therefore, this project recommends property developers to capitalize on these findings to open new shopping malls in neighborhoods in cluster 0 with little to no competition. Property developers with unique selling propositions to stand out from the

competition can also open new shopping malls in neighborhoods in cluster 1 with moderate competition. Lastly, property developers are advised to avoid neighborhoods in cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighbourhoods in cluster 0 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall