IBM Applied Data Science Capstone

1. Introduction

This project is an implementation of the skills learned from this series of course. Though an entire data science pipeline including data collection, data cleaning, exploration data analysis, machine learning process, visualization and information delivery, we are able to solve the real business problems and present the meaningful insights derived from the data.

The topic of this project is to make a location recommendation for opening an Italian restaurant in Chicago central area. Chicago is the third-most-populous city in the United Status. As an international hub for finance, culture and technology, there are a lot of opportunities. However, starting a business is challenging especially at the beginning. The selection of location requires a comprehensive consideration of facts including population, transportation, security, rent and so on. An investigation of neighborhood environment is likely to provide meaningful insights for decision-making. Therefore, this project aiming to investigate the business structure in each neighborhood will be helpful to a business owner/ someone who is interested in opening a new business.

2. Data Preparation

1. The name of neighborhoods

Chicago is officially divided into 77 community areas. In this project, we will focus on three communities: Near North Side, Loop, and Near South side in the city center. This area containing 16 neighborhoods: https://en.wikipedia.org/wiki/Community_areas_in_Chicago. The name of these neighborhoods can be scarped from this website with the help of BeautifulSoup package.

2. The coordinates of neighborhoods

With the names of neighborhoods have been collected, the Geocoding can transfer the place names to coordinates. A data frame with the names and coordinates of the neighborhoods define the scope of the analysis.

3. The venues information for each neighborhood

In each neighborhood, there are venues in various categories. The information of these venues can be requested from the Foursquare API. In this project, we will utilize the venue name, location and categories information in the radius of 500 of each neighborhood.

After these processes, the data for analysis have been collected. It is noteworthy that the created data frame has only 14 neighborhoods. It is because a. The neighborhood “South Loop” occurs twice in the neighborhood name list; b. there is no exploration result in Foursquare for the neighborhood “Goose Island”.

In 14 neighborhoods, there are 163 different categories of venues. As for Italian restaurant which is our interest, there are 27 venues distribution in these neighborhoods. What’s more, Italian Restaurant is the most, and 2nd common venue in The Gold Coast and River North neighborhood.

4. The venues information for Toronto

Both of the Toronto and Chicago have a population of around 2.5 million. Therefore, the business structure in Toronto may have some similarity with that in Chicago.

From the assignment in Week 3, we have explored the venues information for Toronto. There are 234 categories for 38 neighborhoods. This data is stored in “Toronto.csv” and will be reused in this project.

With the distribution of venues in different neighborhood, we might be able to build a machine leaning model which takes features (the number of each category in the given neighborhood) to predict the target (the number of Italian restaurant). The detail of this process will be elaborated in the next section.

3. Methodology

3.1 Exploratory Data Analysis

To better understand our data, some EDA process have been implemented. For example, a bar plot shown in Figure 3.1 shows the number of venues in each neighborhood. Five neighborhoods: Magnificent Mile, River North, South Loop, Streeterville and The Gold Coast have the number of 100 venues is due to Foursquare limiting 100 records.

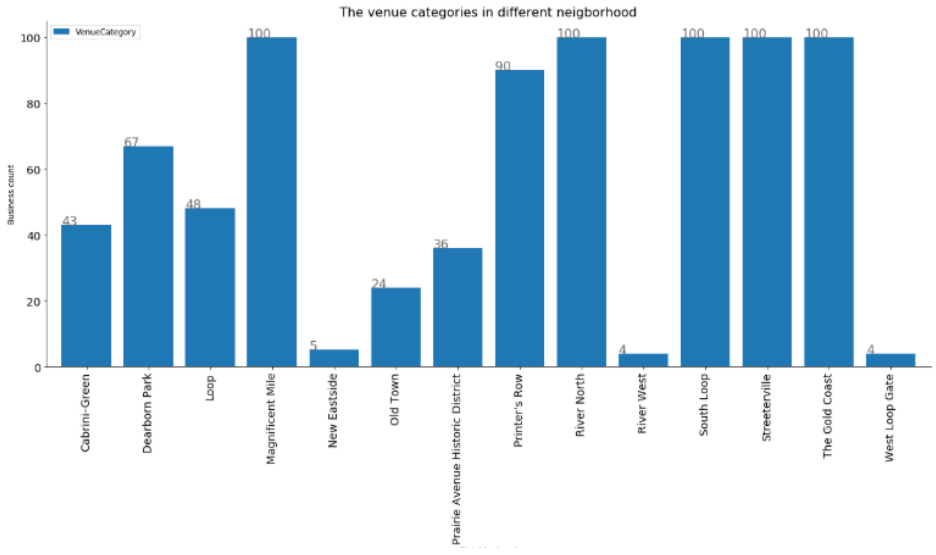

Figure 3.1 The bar plot of venues in each neighborhood

3.2 Machine Learning Process

In this project, we adopt the regression process to predict the number of Italian Restaurant according to the surrounding venue types. The selection of proper regression model follows the steps:

a. Comparison between models:
   The 5-folder cross validation is applied to Support Vector Regressor and Gradient Boosting Regressor. The SVR model shows a higher score than GBR of 0.29. Therefore, the SVR with RBF kernel is implemented in this project.
b. The selection of parameters:
   The Grid Search method is used to determine the parameter C and gamma. It has shown that the combination of C=1.0 and gamma=0.1 will give the best score for Toronto data.
c. Refine the features:
   The SVR is trained with the numbers of every categories in Toronto. If we only train with the number of venues for 47 restaurants and remove others such as parks, hotel and forests, the accuracy will slightly reduce from 0.29 to 0.28. However, the size of the features is reduced by almost 200 items.
d. Evaluate the model:
   In Figure 3.2, it can be seen that after rounding the predicted Italian Restaurant numbers, the result is similar to the actual situation. This prediction will be implemented with the test data from Chicago neighborhoods.

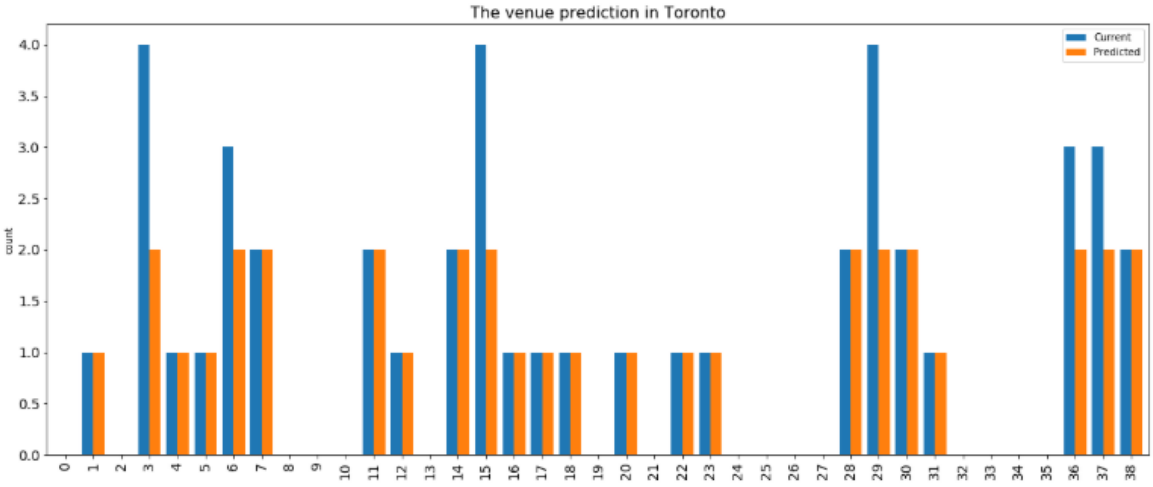

Figure 3.2 The current Italian restaurant and predicted target in Toronto

4. Results

4.1 Machine Learning Result

As the machine learning model based on SVR showing a good result in predicting the Italian restaurant number in Toronto, we apply this model to Chicago neighborhoods. A comparison of the current Italian restaurant and the predicted number in each neighborhood is shown in Figure 4.1.

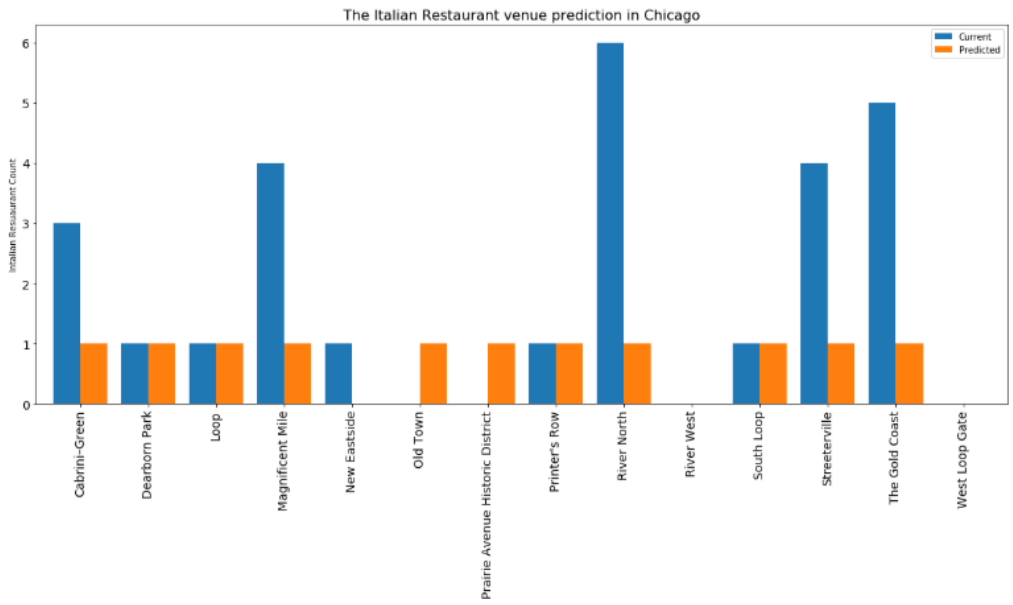

Figure 4.1 The current Italian restaurant and predicted number in Chicago

In figure 4.1, 4 out of 14 prediction is accurate. Most neighborhoods have a larger number of Italian restaurant than the predicted number. However, the "Old Town" and "Prairie Avenue Historic District" are the only two neighborhoods which does not have an Italian restaurant as prediction. There are likely to be good locations to open a new restaurant.

4.2 Visualization in Map

As it is indicated by machine learning model, we might decide our new Italian restaurant in two neighborhoods: "Old Town" or "Prairie Avenue Historic District". We visualize the existing 26 Italian restaurants in Chicago central area using small blue circles in Figure 4.2. The two potential neighborhood is marked by large red circles.

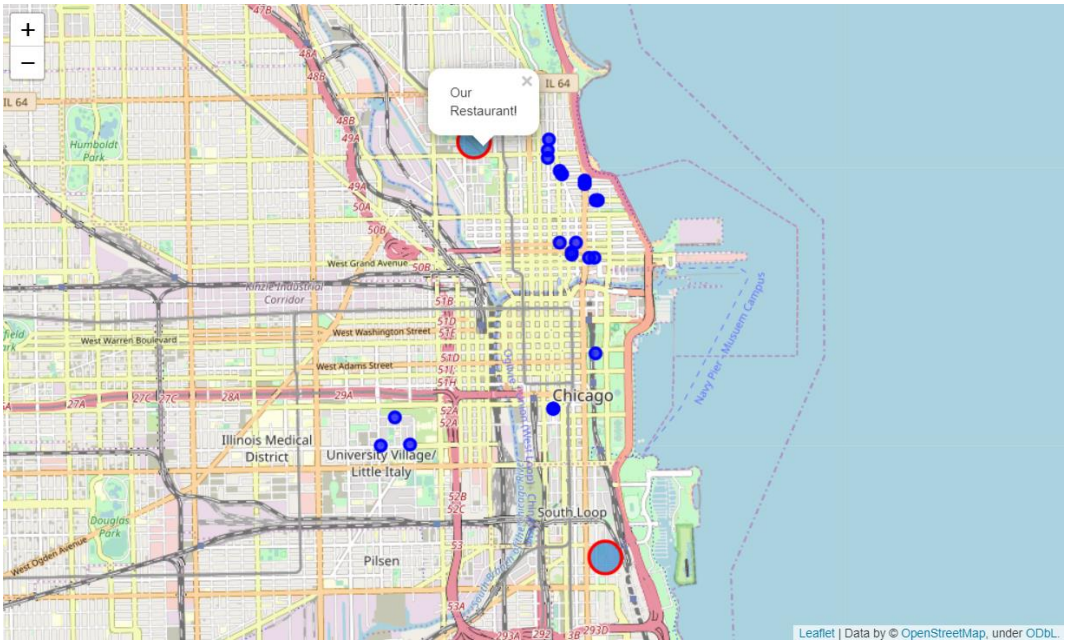

Figure 4.2 The map of existing Italian restaurant and predicted restaurant in Chicago

5. Conclusion

In this project, we have successfully predicted the number of Italian restaurants in 14 neighborhoods of Chicago. We have utilized the data provided by wiki, geocoder and Foursquare API. After data cleaning, the regression model has been employed to predict the number of business of certain category. The number of each kind of categories in certain neighborhood reflect the business environment of this area. In this case, what we are interested is the relationship between various business categories and the Italian restaurant.

The difference between this predicted number and current situation helps us to decide the neighborhood to open a new Italian restaurant. In this case, “Old Town” and “Prairie Avenue Historic District” will be a good choice.

In addition, there are some possible improvements could be made. For example, we have used the business categories of Toronto to train the machine learning model. However, there are some variations of business between different cities. A comparation of different cities can be made to improve the accuracy to predict the situation in Chicago. What’s more, although the business venues in a neighborhood can reflect the commercial environment in that area, other factors such as the rating of competitors, the public transportation, and parking may also have an influence for location selection.

6. Acknowledgement

In this project, we have to acknowledge the data science course provided by IBM powered by Coursera.