Fast Food Retail Store Location Selection by Predictive Model

Laurence Lin

2020, June

1. Introduction

1.1 Business Problem Background

In the business field, selecting an appropriate location for new opening store is an important task when entering an unexplored market. Opening a new store near the city center or the outskirt of town may lead to huge difference on business profit, which is highly related to business's success. In order to find a suitable location for a new store, retailers and business analyst should made lots of research, including surveying local market, number of competitors, the cost to invest a new store and so on. These business analysis cost a lot of time and money, and will result in a huge loss if the new store location is decided improperly. Nowadays, since big data and machine learning technique has been more developed and maturely applied in various business field, I decide to bring up a method helps select the best retail store location by the provided geographic data and predictive model. This may help business stakeholders and retailers save time and cost to get better understanding to local market, and help with better location of new retail store.

1.2 Problem Definition

Elements that may affect a store's income including consumer behavior, consumer demographics and competitiveness of other stores. Foursquare API provides location based data such as competitors, number of other venues and other geographic data within local area. Though the check-in data provided by Foursquare API is not rich in certain location, we apply the geographic data from Foursquare API, along with retail store popularities and ratings from Google Place API as target data. Our aim is to apply location based data from Foursquare API to rank the location candidates by predicting popularity and ratings of certain location.

1.3 Interest

Retailers and investors may want to find a method to intelligently select an optimal location for a new retail store, reducing the risk of financial loss.

Analyst from other business field may also be interested if they want to start a new business of investment in an new unexplored area.

2. Data Description and Collection

2.1 Data Description

Assuming a fast food retailer wanted to open an new store in a city, and investigates the consumer potentials and competitiveness. The retailer analyzes the relationship between geographic data and profits of existing fast food store, and build a mathematical model with it. In this project, I selected Mcdonald's restaurant as the retail store to analyze. I explored the geographic data near the restaurant in New York, and build a predictive model to estimate the popularities and ratings for the new store location candidates.

Locational based service has provided informations for data scientist to analyze tasks around a certain area. Foursquare API, for example, enables us to scrape geographic data around certain venue and human activities. For retail store location selection task, business analyst considers factors such as demography, incoming flow, cuisines habit, competitiveness and lots of elements. I mine Foursquare geographic data that may be informative about the retail quality of a geographic location.

2.2 Data Collection

The Foursquare geographic data is collected around each retail store, within the area of radius 200 meters. The whole range of analysis covers an circle area with radius 10 kilometers, centroid locate in geographical center of New York(Manhattan area 40°46'26.2"N 73°58'49.9"W). For all the retail stores located in this range, I collected the geographic data around each store. The geographic data in the vicinity of retail store is assumed to be informative for the store's popularity or ratings. For example, a retail store opened near railway station implies more potential consumers, while a retail store open near lots of competitors may reduce it's consumer. As for each store's popularity and ratings, I applied Google Place API to extract the ratings and number of comments for each store as indications of the location quality. Finally, I collected 54 Mcdonald's stores with location data in New York.

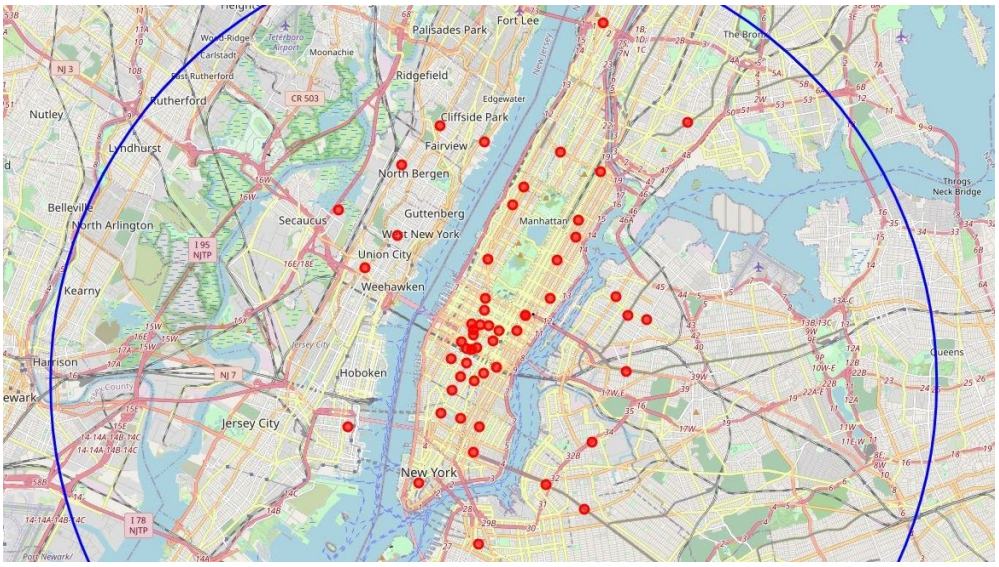

Fig. 1 Analyze area in New York center

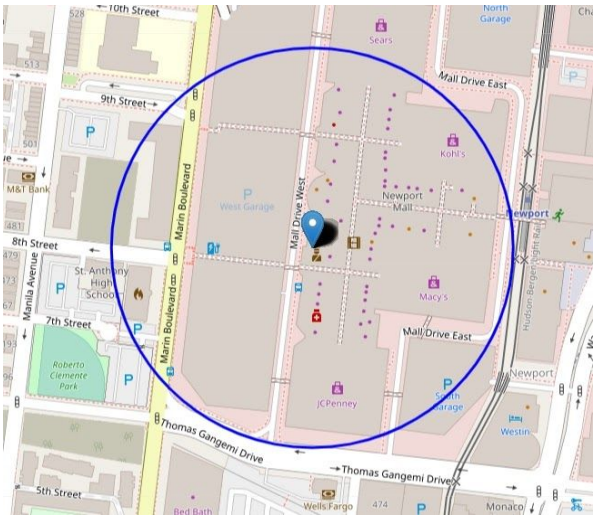

Fig. 2 Analyze area nearby retail store

The Foursquare geographic data contains informations of density, heterogeneity, competitiveness, and area popularity. Density represents the number of nearby venues. Heterogeneity calculated the neighbor entropy, indicates the diversity of category venues in the surroundings. Competitiveness is measured by number of same category venues nearby the retail store, in this case including number of fast food restaurants and other restaurants. Area popularity is calculated by number of residential venues around the store, which may imply number of residents. Given the analyze area set $\{p \in P,\ dist(p,\ l) < r\}$ where p stands for each nearby venue around location l within radius r, the geographic features is measured as follows:

Density: Number of all venues p around store location l.
$density = |\{p \in P,\ dist(p,\ l) < r\}|$

Heterogeneity: Define number of neighbor venue with type γ as $N_\gamma(l,\ r)$, total venues in the area $N(l,\ r)$, measure entropy of all place types in the area.

$neighbor\ engropy = -\sum_r \frac{N_\gamma(l,\ r)}{N(l,\ r)} * log(\frac{N_\gamma(l,\ r)}{N(l,\ r)})$, for all types γ in the area.

Competitiveness: Compute number of same type venues around the retail store, here including type *fast food restaurant* and *other food restaurants*.
$Competitiveness = -\frac{N_{\gamma,l}(l,\ r)}{N(l,\ r)}$, $N_{\gamma,\ l}(l,\ r)$ is number of venues of type γ same as location l.

Area popoularity: Assume number of residential places around location l indicates the popularity.
$Area\ popularity = N_{res}(l,\ r)$, $N_{res}(l,\ r)$ is number of residential venues around location l.

The retail quality of an geographic area is measured by popularity and ratings, we aimed to build a model that could predict high ranking locations based on the retail quality. First we assess the performance of each geographic feature based on ability to predict highest retail quality spots, observe the performance of each feature, then we combine these features to observe if the model improves.

3. Methodology and Metrics

3.1 Exploratory data analysis

3.1.1 Feature relationship between target

We apply EDA to observe the structure of the data. Though the location data size is not large, take a look of the correlation between features and target may help explore useful features within the data.
First, we observe the scatter distribution of each features with respect to the popularity:

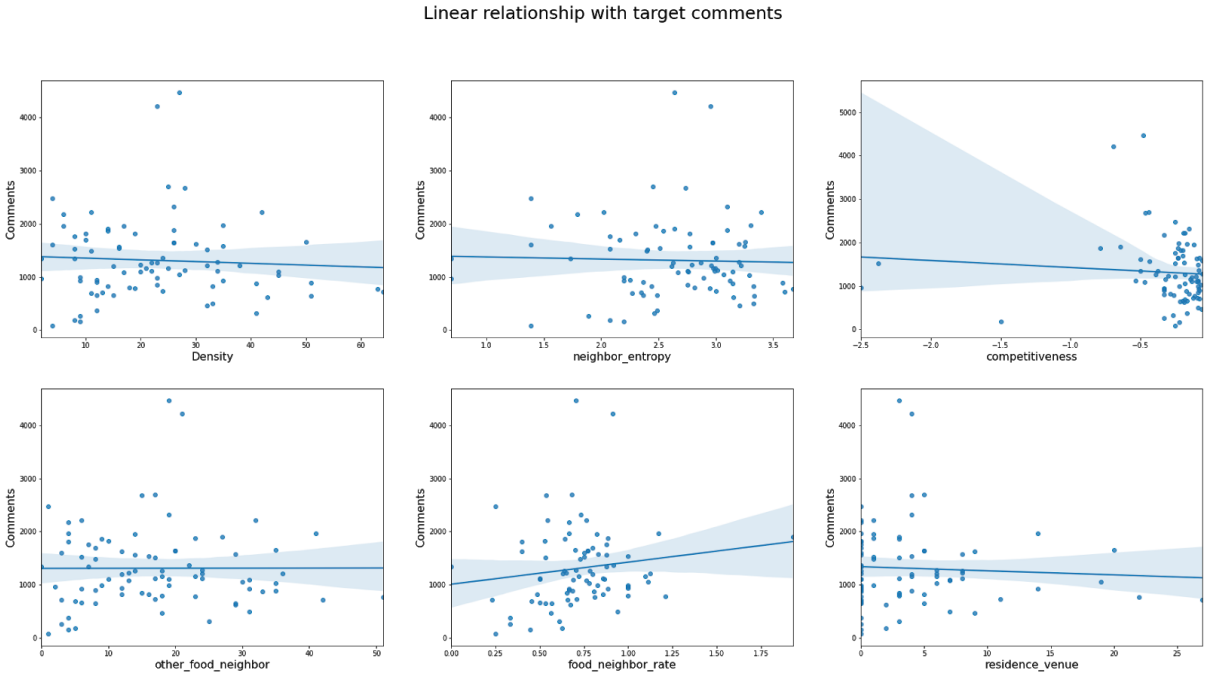

Fig.3 Scatter distribution of each features with popularity

Each feature is hardly linear relative with the popularity, thus it may be difficult to apply linear regression to model the target. Then, we observe the feature distribution with respect to retail store ratings in Fig.2. We can see that it's hard to find linear relationship between each feature and the store ratings.

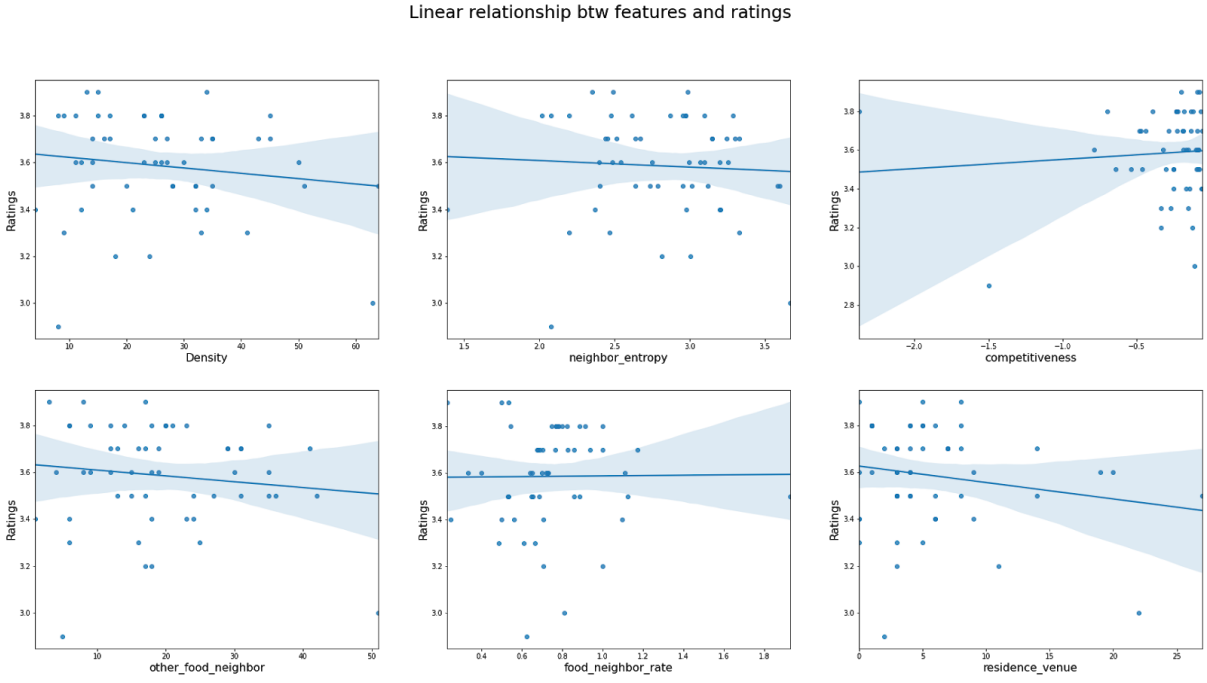

Fig.4 Feature distribution with respect to retail store ratings

However, by the above figure Fig. 1, there are some outliers observed in the data. We could remove these outliers, and the result shows that this would slightly improve the performance.

3.1.2 Relationship between features

As for the area popularity, we assume that the residential venue located in the area represents the demographics. The following figures shows that the more residential venues in the area, the higher the density is. This could support the assumption that more residents implies dense area that contains more venues.

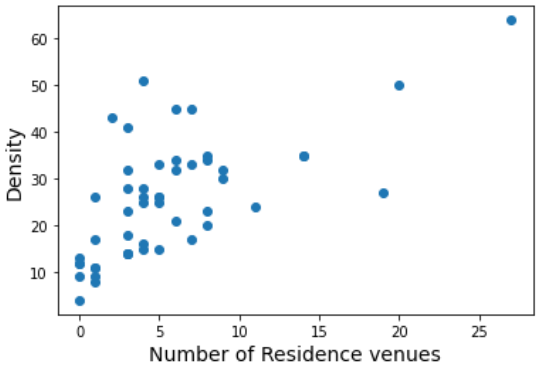

Fig 5. Residence venue vs. Density

The residential venues is collected from Foursquare API, since we couldn't get the exact demographic within the area. This area presents the circle area of 200 meters around each retail store, which we believe correlates to the store's popularity. To validate the usability of the residential venue, we observe the relationship with other features:

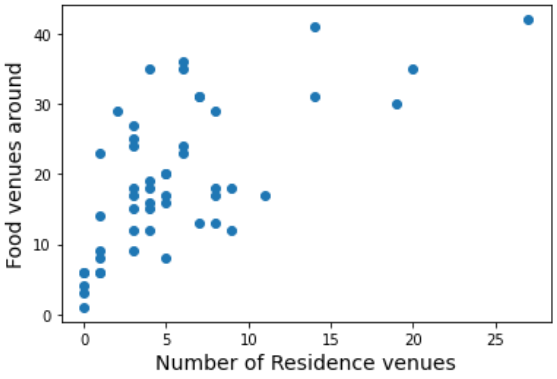

Fig. 6 Residence venue vs. Neighbor food venues

Assume residential venues indicates more residents, then the more residential venues may attract more neighbor food venues. Fig. 4 presents positive relationship between these two features. However, lots of places where less residence venue contains many or less food venues. For this, we conclude that

area with less residential venues like commercial site, differ in lots of aspects, thus have varying food venue neighbors. The data is collected in New York, which contains lots of commercial area that may not have rich food restaurants.

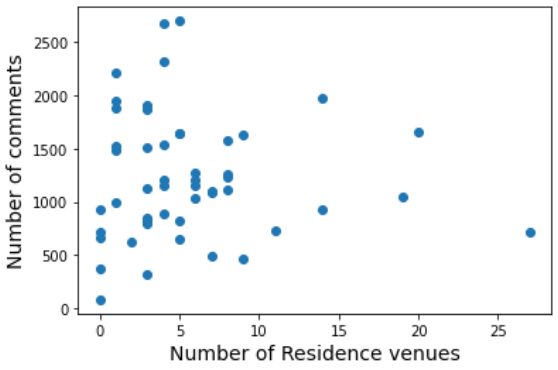

Fig. 7 Residence venue vs. Number of comments

As observe in Fig. 5, the retail stores in New York city mostly have fewer residential venue neighbors. On the other hand, residential venue may not present the tendency of retail store popularity. In an retail store located in commercial area, the popularity of the store might be high or low. Thus, residence venue might not be a helpful feature to predict the target popularity.
As for the competitiveness, we compute the feature represents the number of neighbor fast food restaurants. However, consumer might not select to go to Mcdonald's not only because of same type restaurant, but also other restaurant choices. Thus, we add an feature *total_compet_rate* representing the percentage of all type food restaurants in vicinity.

$$total\ compet\ rate = competitiveness + other\ restaurant\ competitiveness$$

Finally, we apply total 6 features including: *Density, Neighbor entropy, Competitiveness, Other Food neighbor, Other neighbor rate, Residence venue.* Before input to the model, we first normalize the features. Due to the metric of ranking problem, the output target score is assumed positive. We apply the Min-Max Scaler to normalize the features.

3.2 Metrics evaluation

To evaluate the ranking performance, I adopt the NDCG@k (Normalized Discounted Cumulative Gain) metric to compute the top-k prediction list

accuracy. This metrics represents the ability of gathering top-k highest rank item in the front of list. The calculation formula is as follows:
First we evaluate the DCG@k (Discounted Cumulative Gain):

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel(i)}-1}{log_2(i+1)}$$

$$rel(i): \text{ relevance score of item } l_i, \text{ compute by actual rank position of item.}$$

$$rel(i) = \frac{|L| - rank(l_i) + 1}{|L|}$$

Then normalized it by DCG of ideal ranked list (IDCG):

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

In this project, by experience and the number of samples, select k = 3.

3.3 Methodology

3.3.1 Cross validation training

We aimed to predict the rank of retail quality for unknown locations. However, to analyze the location data within an area, our data size is limited to the total retail stores locate in there. Due to the small size data, We use cross validation to split 12% of data for testing, and the rest for training model. We run the iteration 1000 times, and compare the result with the random baseline to observe the training performance. The random baseline is obtained by randomly sort the list, then compute the NDCG@k as random ranking performance. Overall performance of the ranking is measured by average all 1000 results of NDCG@k.

3.3.2 Individual feature performance

The comparison of the performances of each individual feature is shown below in table 1. It can be seen that feature *total_compet_rate* performs the best, followed by area venue density. We could conclude that number of restaurants within the area highly affects consumers' choice, thus have influence on the retail store popularity.

```
Individual performance result:
total_compet_rate      0.793444
Density                0.761142
competitiveness         0.76059
food_neighbor_rate     0.732563
other_food_neighbor    0.732174
neighbor_entropy       0.663222
residence_venue        0.650172
Random base line       0.557107
Name: Individual NDCG@3 metric, dtype: object
```

Table 1. Individual feature performance NDCG@3

Residence venue has poor performance on predicting the popularity, as we expected. However, neighbor entropy which performs well in other research don't get high performance. This may be discussed further, however with our poor data size of location data it may be difficult to dig deeper. This may be work on when we have larger data. For the top rank individual features, we observe the predicted retail quality by giving different values of feature to a full trained model:

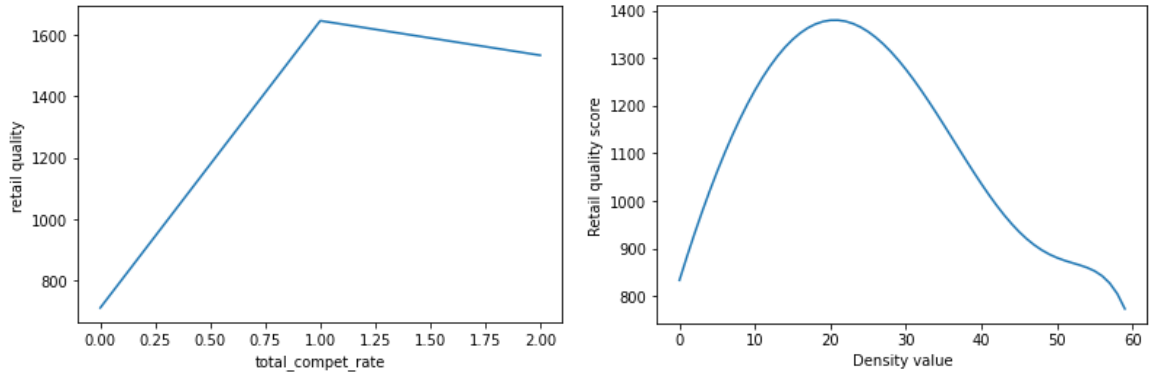

Fig. 8 Predict retail quality for different individual values

The result in Fig. 6 suggests that as the density of an area increase, the popularity of the retail store might rise; however while reaching a threshold, the total consumers decrease. In a highly crowded place, such as business center, retail store like Mcdonald's might not receive the maximum customer. Competitiveness on the other hand shows interesting result, in area where competitors is very few, the popularity tends to be small. We interpret this by that places of very few competitors has low popularity of consumers. The increasing of competitiveness indicates the potential consumers in the area, thus increase the retail store popularity. After the competitors reached a limit,

consumers in that area is satisfied by these food restaurants, then the retail store popularity starts decreasing.

3.3.3 Combined feature with supervised learning algorithm

To get better ranking prediction, we combined geographic features and apply supervised learning framework to see if performance improved. We collect the features that perform well in individual feature assessment, and apply these features to different supervised learning models. Four models are selected to compare the performance:

	NDCG@k of combine features
SVR	0.813482
Neural Net	0.812825
LR	0.789437
DecisionTree	0.767698

Table 2. Different model performance comparison

The combination of geographic features shows different performance on different models. Out of the four selected models, support vector regression performs the best, followed by neural network. For the reaction of different to the combined features, most of the model lower the performs for exploit multiple features. Only linear regression is able to import full geographic features and get fine performance. As for the other three models, total competitiveness feature gives the best result, while adding density and residence venues slightly reduce the performance. We believed that factors that influence a retail store's location includes consumer behavior, area popularity and competitiveness, eventually we choose one feature to represent total competitiveness and another to represent area popularity, the final result is shown in Fig. 7. Further discussion is in section 4.

3.3.4 Testing performance on other city

For validation, I test the ranking by the trained model on geographic data from another city, Toronto. However, the ranking performance show poor results, even worse then the random baseline. It seems that the information contains in location data for one area is not applicable for another area.

4. Results

Different combination geographic features if applied on a set of models, as opposed to our expectation, not all combination features improves the performance. It's expected that combination features may contain noises, such like density and residence venues which both implies demographic information combined together. It's proved that features contained correlated informations won't improve the predicting ability for retail store popularity.

	LR	DecisionTree	SVR	Neural Net
total_compet_rate	0.828283	0.827483	0.844368	0.84022
total_compet + residence_venue	0.786605	0.803779	0.80686	0.783583
total_compet + Density	0.799004	0.819679	0.829042	0.80687
total_compet + residence_venue + Density	0.803201	0.804203	0.798687	0.804378
total_compet + residence_venue + Density + competitiveness	0.786679	0.798467	0.795333	0.785635

Table 3. Performance of combination for different geographic features

To scrape sufficient informations of area popularity and competitiveness, we decide at least one popularity feature and one competitiveness feature is required. However, as shown in Fig. 8, none of the combined geographic features outperform that of single total competitiveness. We conjecture that this may due to the noise in our feature, and the small size of data. Local popularity which implies potential consumers should provide information that contributes to a retail store's popularity.

5. Discussion

In this project, geographic features are collected from Foursquare API and popularity of retail stores from Google Place API. The geographic features we collected contains area popularity and competitiveness, though consumer behavior is not considered. Due to the limit of the API service, retail store data collected from one city is restricted to a small size. Since prediction model trained on one city is not able to applied on another city, our data size is limited small. However, there are still interesting information we could gather from our result.

For a retail store restaurant, for example Mcdonald's, it's competitors include not only same type fast food restaurant but also other food restaurants. The performance result proves that, since total competitiveness works better than same type competitiveness feature. As for area popularity, residence venue which behaves poorly in linear regression model, could perform comparable results with other feature. This shows that number of residential venues collected from Foursquare API could represent an area's popularity in some extent. Venue density in an area somewhat overlaps the effect of residence venue. The space heterogeneity however, did not show any benefit in predicting retail store's popularity. In some previous research, it's believed that diversity of a place could imply the types of consumers, which is related to the potential consumer. This may be another factor that suffered from the small data size.
We've made assumption that number of comments on the google map represents the popularity of a retail store. Single geographic features like total competitiveness is proven the ability to predict the popularity of a retail store. The performance of combined geographic features not as good as single feature may due to coincidence of the specific dataset or the noise within the data. If there are more data to explore, we believe that plural geographic features could provide more useful information for retail store popularity.

6. Conclusion and future work

To decide location of a retail store is not an easy task, and involves a lot of considerations. To analyze the task requires huge amount of data, which is the first difficulty. Important and private data for each company is hard to collected, such as a retail store's revenue or total consumers during a time. Some developer API provides access of part of these data, yet the data we could scraped is still limited. In this project, we benefit from two API that enable us to collect the location data, and proves that data provided by them is valuable for data analysis.
Selecting the location of an opening retail store, we collected venue location data from Foursquare API as basis to predict retail store popularity collected from Google Place API. We analyze the relationship between geographic features, and proves its effectiveness to predict a retail store's popularity. However, the behavior of our result didn't improve significantly from previous

research. We believe this is due to our small data size, and we didn't take into account the consumer behavior. Consumer behavior should be analyzed while selecting retail store location. We don't want to open a Mcdonald's in an area that most of the consumers eat vegetable food. While we are analyzing an local area, incoming consumers from outside the area should be considered. How to collect these data from another API of data sources is another challenge.

For real world cases, analyst may want to select an area from another city or country. In this project, application of prediction model on another city failed. Future work may look forward to gather multiple data from different cities, and aimed to build a predictive model that could rank candidate locations for retail store across different city or countries.