Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Contents

Introduction

London is not only the commercial centre of the UK but also attract millions of visiting tourists every year. This cosmopolitan city has a very vibrant food scene, in that new restaurants are opening all the time but sadly at the same time many restaurants are closing down due to the fierce competition.

Now, if someone is looking to open a Chinese restaurant in London, can we use data science to help to find out whether a site is good or bad for it? This is the **business question** I set out to answer. The target audience / **stakeholders** are those who plan to open a Chinese restaurant in London, whether as a start-up or a new branch of an existing restaurant, and their investors. In addition, the landlords of the commercial properties in London may also be potential target audience.

One of the past examples of the Capstone project is about recommending where to open an Italian restaurant in Berlin, in which an assumption was that a good spot should be in "areas of Berlin that have low restaurant density, particularly those with low number of Italian restaurants". This may be a sensible assumption, but on the other hand, observation also tells us that more often restaurants are "clustered" together. For example, I once travelled to Bradford, a small city in northwest England, I was surprised to find 4 French restaurants in the same street of about 300 metres and they all looked quite busy. So the density alone may not be enough to predict whether a location is good or bad for opening a Chinese restaurant - are there any other data/indications we need in order to make a good prediction?

Data

To assess a business case, the well-known SWOT business analysis methodology tells us to look into these four aspects, namely, Strengths, Weakness, Opportunities and Threats. As strengths and weakness are internal to the business itself, here we concentrate on the opportunities and threats aspects.

High restaurant density can represent threats as it means more competition. On the other hand, it can also represent opportunities. For example, a cluster of French restaurants may have some kind of "branding" effect as this may foster a reputation of being a go-to place for French cuisines.

The other type of opportunities/threats are represented by what type of the potential customer base the restaurant location has - Is it in a high traffic area, i.e. busy and bustling with people, and whether the people are ready to spend money in restaurant?

A prediction model will need to take into account all of the above 'features' of a location in order to predict whether it is a good or bad location.

For this purpose, firstly, the datasets I am going to use to represent restaurant density are:

- For a given location, how many nearby **Chinese restaurants** and how close are they?
- For a given location, how many nearby **similar types of restaurants** and how close are they? For similar restaurants, I refer to Asian type such as Japanese, Korean, Thai, Indian, Malay, etc.
- For a given location, how many nearby **non-similar types of restaurants** and the proximity of them?

The above datasets can be obtained by using the **Foursquare API**.

Secondly, for the data about the potential customer base, there is no readily available source. However, I found two data sources I believe to be good representation of the potential customer base:

1. ATM (i.e. bank automated teller machine) location data. ATMs are installed by banks in locations where there is a high traffic of people who want to withdraw cash. Therefore, the proximity of ATMs can give a good indication of the volume of traffic and the likelihood of people with money to spend. The source of ATM I use for this project is **Overpass API (OpenStreetMap)**.
2. Public Transportation link data. The Underground and buses are two of the major public transportation links within London. **London Underground passenger counts** published by the London Transportation Department can be used as a good indicator of traffic volume.

Finally, we also need the historical data of the prediction target (i.e. how good/bad a location is for a Chinese restaurant) to train and verify the prediction model. Profitability data of existing Chinese restaurants would be the ideal historical data for this purpose but such data are not readily available. Another good representative data is the venue "checkin" statistics from Foursquare, unfortunately, it seems Foursquare does not provide the actual number of checkins (I tested this using Chinese restaurants in London). Therefore, for the purpose of this project, I use the "number of likes" in the **Foursquare venues information**.

The recap, the data I am going to use to train and verify the prediction model are:

- Data for the prediction target: "number of likes" in the **Foursquare** venues information;
- Data for the features:
    - number of nearby Chinese restaurants and average distance (**Foursquare API)**
    - number of nearby similar types of restaurants and average distance (**Foursquare API)**
    - number of nearby non-similar types of restaurants and average distance (**Foursquare API)**
    - number of nearby ATMs locations and average distance **(Overpass API - OpenStreetMap)**
    - **London Underground passenger counts.**

Firstly, I set out to investigate an area of ~6Km radius of central London. The area of interest can be changed according to stakeholder requirements by adjusting the location of the centre point and the length of the radius

London Chinatown is at the very heart of London right next to the National Gallery, therefore, I choose it as the centre point for London.

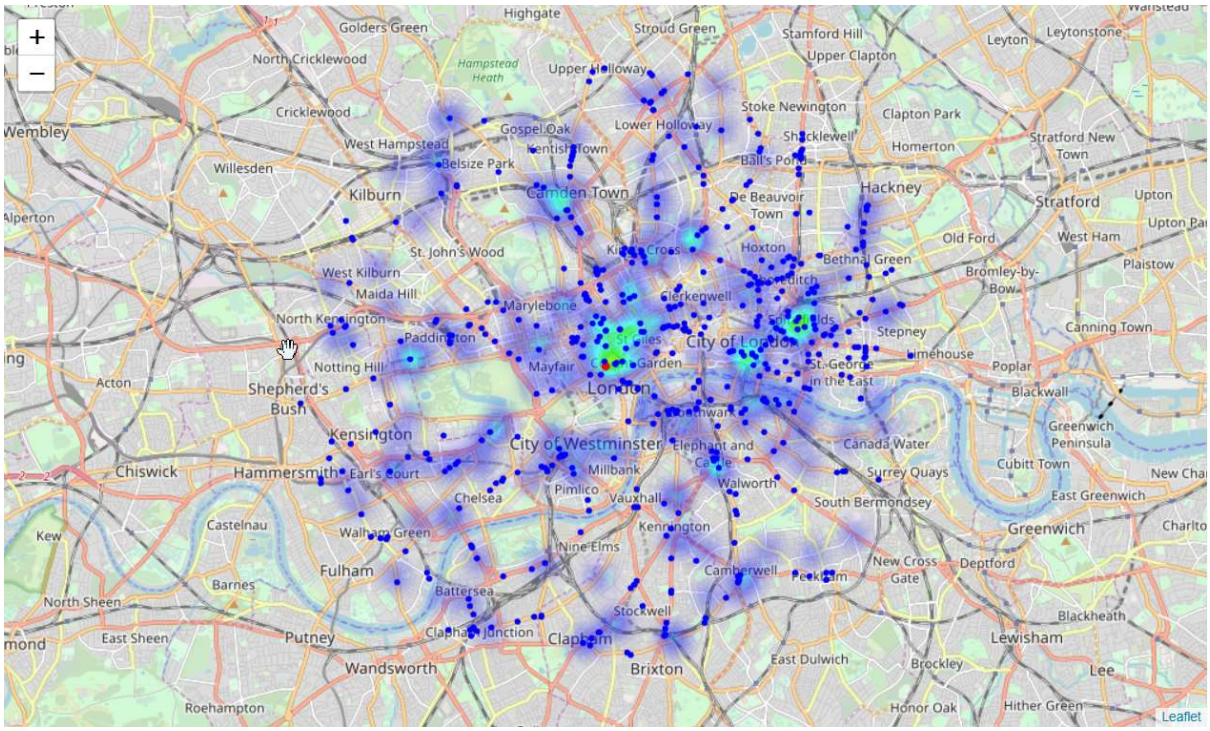
I initially used one Foursquare API call with a radius of 6Km, but it only return 100 restaurants. As 150m radius around Chinatown has already nearly 100 restaurants, an area of 6Km radius should have more than 100 restaurants!

It seems Foursquare API does not like to return large number of venues. Therefore, I'll divide the (large) area into smaller ones and make a Foursquare API call for each.

Acknowledgement

*The Python code for dividing the area is very similar to that of the example notebook provided as part of the Capstone assignment. So **a massive thanks to the original author of the notebook.** However, **the similarity stops here as I'll use a complete different inferential statistical method** to address a similar business problem.*

Using the data collected, a **Heatmap** of Chinese restaurants overlaid with ATM is created to show the density of the Chinese restaurants across the region and to see if there is any correlation with the location of the ATMs:



Methodology

Further Preparation of Datasets - Missing Data Handling

Transport for London (TfL) has provided London Underground passenger statistics by station (https://data.london.gov.uk/dataset/london-underground-performance-reports). In order to include this information in our analysis, I need to find out the latitude/longitude information of each station. This was done by using **GeoPy's geocoder**. Unfortunately, the latitude/longitude information of 3 of the stations could not be obtain from the geocoder. To fix this missing data problem, information has been extracted manually from **OpenStreetMap**'s website: https://wiki.openstreetmap.org/wiki/List_of_London_Underground_stations.

Explanatory Analysis

In order to prepare the data to be used in regression model, data collected are marshalled into a Pandas dataframe. The first 5 rows of the data are as follows:

```
chinese_df.head()
```

	ID	Name	Latitude	Lontitude	X	Y	Likes	C Count	C Avg Distance	S Count	S Avg Distance	N Count	N Avg Distance	ATM Count	ATM Avg Distance	Metro Passengers in million
0	51be1a81498e7e9475a2847c	Mama Lan	51.461582	-0.138689	698759.924396	5.705040e+06	47	0	0.000000	4	100.132016	15	90.317081	2	41.540063	4.600427
1	5743804498e6896a9ac6c3d	On Cafe	51.461160	-0.136063	698944.160516	5.705000e+06	8	0	0.000000	1	129.779787	7	130.101956	1	9.079123	0.000000
2	4fa44009e4b0cfe54c0e2cfb	Courtesan Dim Sum	51.461160	-0.111220	700669.558832	5.705068e+06	45	1	132.150789	3	84.929448	7	119.426002	0	0.000000	0.000000
3	4c0e9eb87189c92802ccd8b6	Big Fat Panda	51.464061	-0.165234	696905.658676	5.705244e+06	5	0	0.000000	0	0.000000	3	60.807068	1	58.927184	0.000000
4	4c93c6e458d4b60c7f6f2229	Ku Do	51.465477	-0.127758	699502.130768	5.705503e+06	0	0	0.000000	1	148.460659	5	88.266078	0	0.000000	2.833457

Note:
C Count = Number of nearby Chinese restaurants
C Avg Distance = Average distance of nearby Chinese restaurants
S Count = Number of nearby similar types of restaurants
S Avg Distance = Average distance of nearby similar types of restaurants
N Count = Number of nearby non-similar types of restaurants
N Avg Distance = Average distance of nearby non-similar types of restaurants
ATM Count = Number of nearby ATMs
ATM Avg Distance = Average distance of nearby ATMs

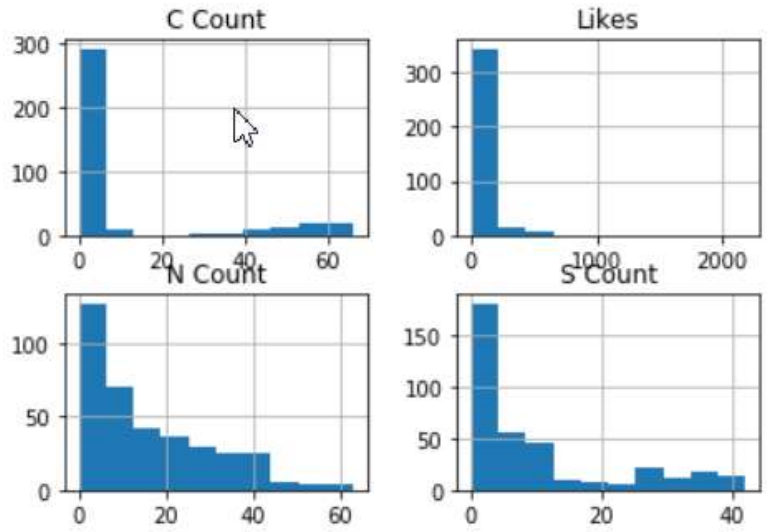
Before trying to fit the data into the model, let's do some explanatory data analysis of the features.

Let's first have a look at the descriptive statistics of our dataset:

	Likes	C Count	S Count	N Count	ATM Count	Metro Passengers in million
count	369.000000	369.000000	369.000000	369.000000	369.000000	369.000000
mean	72.810298	10.737127	9.878049	15.905149	0.615176	4.658470
std	190.129152	20.385517	11.825212	14.102659	0.948905	8.380868
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.000000	0.000000	1.000000	5.000000	0.000000	0.000000
50%	15.000000	1.000000	5.000000	11.000000	0.000000	0.000000
75%	65.000000	4.000000	12.000000	26.000000	1.000000	6.589856
max	2183.000000	66.000000	42.000000	63.000000	4.000000	63.376332

From the descriptive statistics, we can see that the number of "Likes" looks a bit odd, with 50 percentile at only 15. A quick sum on the dataset was performed and found that the number of Chinese restaurants with 10 or less Likes is 164

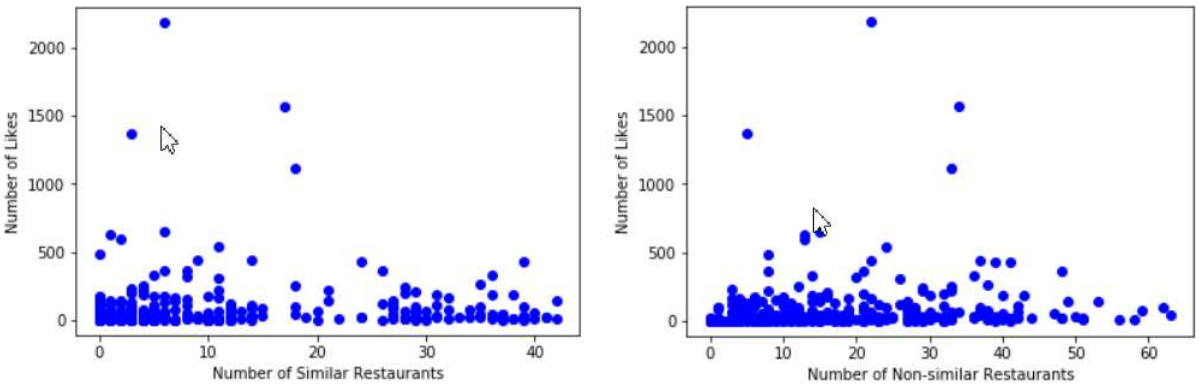
And here are the histograms of the four features, namely, the number of Likes, the number of nearby Chinese restaurants, the number of similar restaurants (as mentioned previously, similar type of restaurants refer to Japanese, Korean, Malay, Thai, Vietnamese, etc.) and the number of other types of non-similar restaurants (such as Italian, Mexican, etc.):

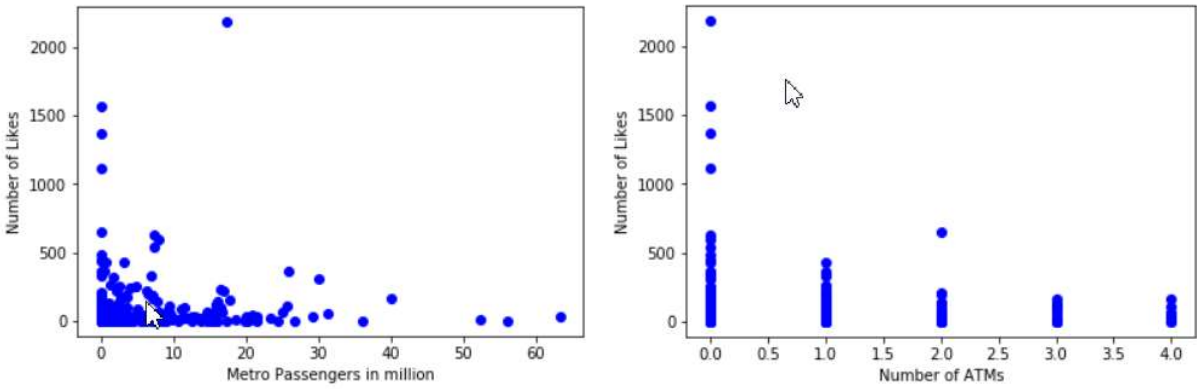


The histograms further reveal that:

- Large number of Chinese restaurants have 0 - 5 other Chinese restaurants nearby, and there are a number of Chinese restaurants have 40 - 60 other Chinese restaurants nearby. On the other hand, there are very few Chinese restaurants have 10 - 40 other Chinese restaurants nearby. That is, (in London,) Chinese restaurants exist in either very low or very high density, but not mid-density;
- Interestingly, a similar phenomenon can be observed for nearby similar type of restaurants, i.e. (in London,) Chinese restaurants are located among either very low or very high density of other similar type of restaurants, but not mid-density;
- There are large number of Chinese restaurants has very low number of “Likes”, in fact, there are 164 Chinese restaurants has only 10 or less Likes. These is rather alarming as I am going to use the number of Likes as an indicator of how well a restaurant is doing. For restaurants serving hundreds of customers every day, such low number of Likes looks to me may be because Fourquare is not popular amount London Chinese restaurant customers, hence, the number of Likes **may not be used as a reliable indicator**.

The look into the further, the number of likes as plot against each of other features:





No linearity was observed in these plots. Fitting the dataset to multiple linear regression model, the resulting variance score is -0.66, which is so low that it is obvious the model **cannot provide any reliable prediction**.

Now, back to the drawing board.

Firstly, let's look at the **business problem and the stakeholders**: in essence, I want to help stakeholder to find out whether a site is suitable for opening a Chinese restaurant.

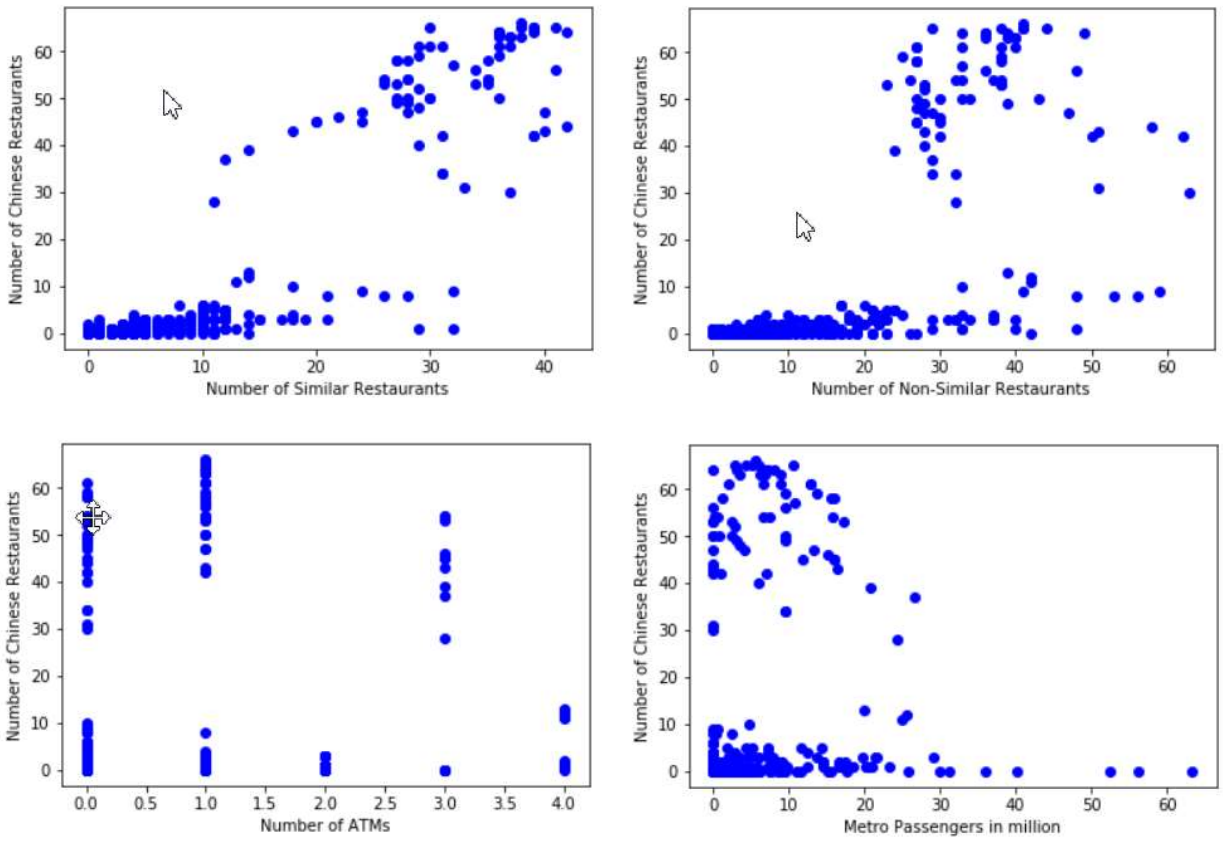
Secondly, it is found that the Likes data from Foursquare is not a good representation of how well a restaurant is doing.

Finally, let's look at what we have so far:

- Some pattern in the "density" of nearby Chinese restaurants and similar type of restaurants can be observed – they are either of high or low density but no “middle of the road”
- From the heatmap, we can also see the Chinese restaurants seem to follow the distribution of the ATM locations.

Therefore, we can try to predict the "density" of nearby Chinese restaurants using the dataset we have got. If our model can make a good prediction, we can still help the stakeholders to answer the business question. It is because, for a given location, we can actually use the model to predict how many nearby Chinese restaurants should have for that location. If in reality there are less nearby Chinese restaurants, we can say there is a good chance the area can "sustain" a new Chinese restaurant, hence, it may be a good location for opening a Chinese restaurant.

Let's now make "*the Number of Nearby Chinese Restaurants*" as the dependent variable, and plot each of the other features against it:



The Plot of Number of Similar Restaurants vs. Number of Chinese Restaurants shows some linearity.

Furthermore, from these plots, we can see Chinese restaurants exist in either low density or high density but nothing in-between.

Predictive Modelling

Now, let's try to fit dataset to a multiple linear regression model again but this time the dependent variable is "*the Number of Nearby Chinese Restaurants*":

- the dependent variable is the number of nearby Chinese restaurants; and
- The independent variables are
    - the number of nearby similar restaurants and their average distance,
    - the number of nearby non-similar restaurants and their average distance,
    - the number of nearby ATMs and their average distance, and
    - the annual passengers numbers of the nearby Metro stations (weighted with distance)
- Randomly selected 80% of the dataset is used for training the model.

The resulting model was then tested with remaining 20% of the dataset, the **variance score** of the resulting model is **0.87**, which is **very close to a perfect prediction (score of 1)**, it means the model can provide a very good prediction. Therefore, a multiple linear regression model is used as the predictive model for addressing our business problem.

Results

A multiple linear regression model is then derived using 100% of dataset, and here is the resulting model:

- Coefficients:
  [2.11546858 -0.03974706 -0.57464468 0.05390643 -0.64187237 0.01702939 0.18028619],

  The Coefficients are for:

  [Number of nearby similar types of restaurants
  Average distance of nearby similar types of restaurants
  Number of nearby non-similar types of restaurants
  Average distance of nearby non-similar types of restaurants
  Number of nearby ATMs
  Average distance of nearby ATMs
  Metro passenger numbers]

- Intercept: -3.53078879

The coefficient for the "Number of nearby similar types of restaurants" is the largest, which is consistent with the observation of the linearity shown in the scatter plot of the Number of Similar Restaurants vs. Number of Chinese Restaurants.

The variance score of the model is 0.85, which shows the model can be a very good predictive model to address the business question I set out to answer, i.e. whether a site is suitable for opening a Chinese restaurant, by predicting number of nearby Chinese restaurant this site would have. If the number of nearby Chinese restaurants that actually exist is larger than the predicted number, it indicates that the location is 'over-crowded" and may not be suitable for opening yet another new Chinese restaurant; on the other hand, if there are less Chinese restaurants actually exist than predicted, it indicates that location is likely to be able to 'sustain' another Chinese restaurant so the site may be suitable for opening a new Chinese restaurant.

To illustrate this, I've selected a few addresses and assume these are the sites short-listed for opening a new Chinese restaurant:

1. 119 Newington Causeway, Elephant and Castle, London SE1 6BN
2. 160 Old Street, London EC1V 9FR
3. 276 Kentish Town Rd, London NW5 2AA

Then Foursquare API is used to find out how many Chinese are nearby of each of these candidate sites. This number will then be compared with the predicted number from the multiple linear regression model just developed above. Here are the results:

1. For "119 Newington Causeway, Elephant and Castle, London SE1 6BN" - currently 0 Chinese restaurant nearby, predicted 1.74. This can be interpreted as the location may be a suitable site for opening a new Chinese restaurant.

2. For "160 Old Street, London EC1V 9FR" - currently 2 Chinese restaurants nearby, predicted 1.66. The actual number of nearby Chinese restaurant is marginally more than the predicted number, which could mean the site may or may not be suitable for opening

a new Chinese restaurant but any more new openings in the future could tip the balance further into the negative side.

3. For “276 Kentish Town Rd, London NW5 2AA” - currently 0 Chinese restaurant, predicted -1.86. This can be interpreted as the location may not be a suitable site for opening a new Chinese restaurant.

Discussion

I started with an intention of using the number “Likes” as an indicator of how well a restaurant is performed. Unfortunately, this data from Foursquare cannot be used for this purpose.

Ideally, if more reliable information can be found, it will make the inferential model much more compelling as it will directly predict how well the site will perform if it is used for a Chinese restaurant.

Although Foursquare data does include number of Checkins, which may be a good indication of how popular a restaurant is, however, Foursquare only offers such data to "authorised" users such as the owner of the establishment.

Another possible information can be the profit/loss account of the restaurants, which may be available from the Companies House filing records. However, these records would need to be obtained and interpreted manually, and companies may use accounting tactics to hide the true picture of profit/loss, which could render the information inconsistent and/or unreliable.

Conclusions

The variance score of the resulting multiple linear regression model is 0.85, a score that is very close to a perfect prediction (score of 1), it means this model can provide a very good prediction, and can to be used to address the business question set out at the start of the project.