Coursera-Capstone-Project

Location recommendation for *Chinese Restaurant* in *Columbus*

Table of content

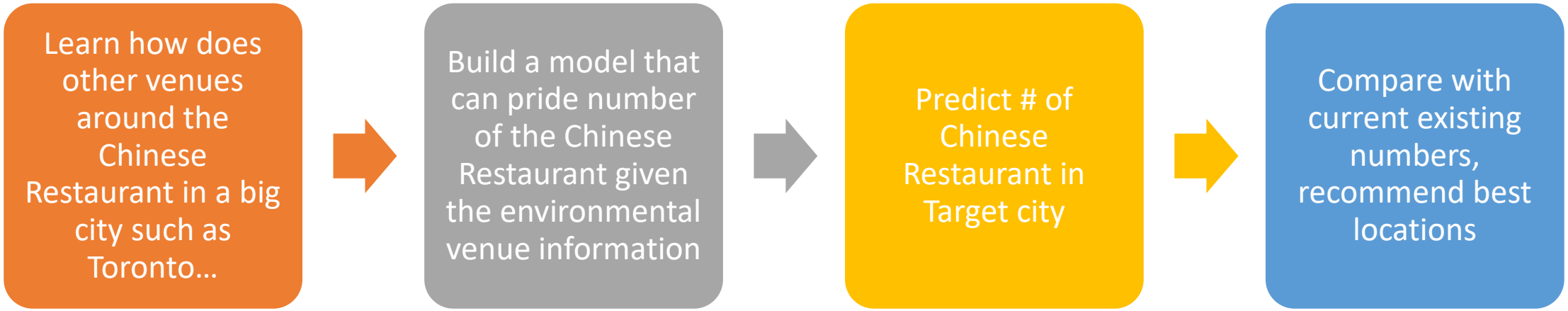
Introduction/Business Problem

1. Introduction/Business Problem

Let say if you are a buisiness manerger who want to invest a *Chinese restaruant* in your resident city. You are live in the mid-size city which has fast growth. You have to decide where or which **neighberhoods** to open the restaruant.
In order to answer this question, you have to build a model get some recommendations where to start your business.

- Therefore, we will learn a model from a mature city/metropolitan city since we believe that it is more develeped. Your city will become a metropolis some day.
- Another thing you believe is that any one of business venue does not exist alone,and "Chinese restaruant" always tends to be find with some other type of shops,because neighberhoods have "cultures" to like them both.

Solution / Methodology flowchart



Data import (Toronto)

- Neighborhood information get from Wiki

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M5A	Downtown Toronto	Regent Park
M6A	North York	Lawrence Heights
M6A	North York	Lawrence Manor
M7A	Queen's Park	Not assigned
M8A	Not assigned	Not assigned
M9A	Etobicoke	Islington Avenue

- Scrap from website and organize into Dataframe

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Rouge,Malvern
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union
2	M1E	Scarborough	Guildwood,Morningside,West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

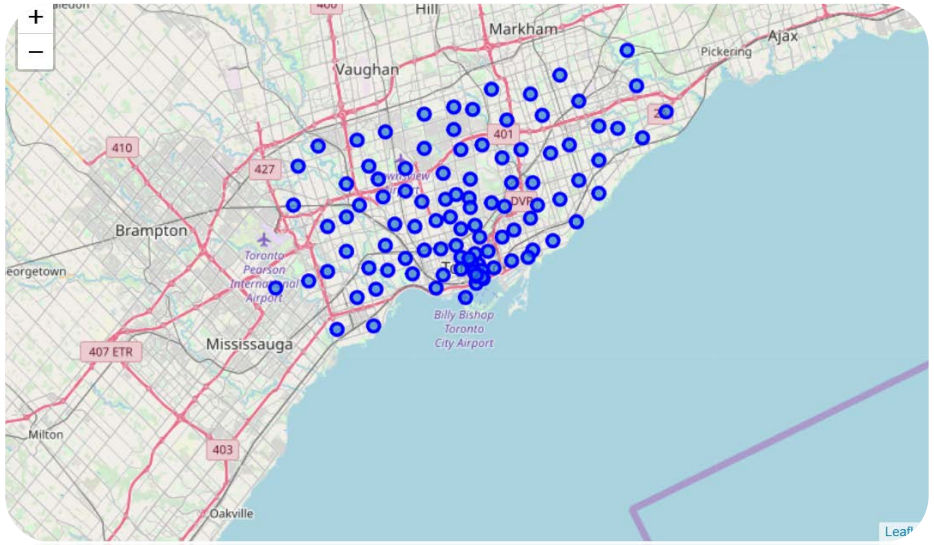
Get location information

- Use geocoder package to get location information:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

- Plot the location using folium package

Neighborhood location in Toronto on map



Get venues information

- Use **Foursquare** API, we can explore the venues around on specific location, so we could achieve venues' name and category

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rouge,Malvern	43.806686	-79.194353	African Rainforest Pavilion	43.817725	-79.183433	Zoo Exhibit
1	Rouge,Malvern	43.806686	-79.194353	Toronto Pan Am Sports Centre	43.790623	-79.193869	Athletics & Sports
2	Rouge,Malvern	43.806686	-79.194353	Toronto Zoo	43.820582	-79.181551	Zoo
3	Rouge,Malvern	43.806686	-79.194353	Polar Bear Exhibit	43.823372	-79.185145	Zoo
4	Rouge,Malvern	43.806686	-79.194353	Canadiana exhibit	43.817894	-79.193260	Zoo Exhibit

- Create one-hot encoding for each catogory

	Neighborhood	Afghan Restaurant	African Restaurant	Airport	Airport Lounge	American Restaurant	Aquarium	Art Gallery	Arts & Crafts Store	A: Restau
0	Adelaide,King,Richmond	0	0	0	0	1	1	2	2	0
1	Agincourt	0	0	0	0	1	0	0	1	1
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	0	0	0	0	1	0	0	1	2
3	Albion Gardens,Beaumond Heights,Humbergate,Jam...	0	0	0	0	0	0	0	0	3
4	Alderwood,Long Branch	0	0	0	0	1	0	0	1	1

http://www.foursquare.com

Build Model for prediction

We will use number of venues in each neighborhoods except Chinese restaurant as inputs and number of Chinese restaurant as output.

Use SVR (rbf kernel) as learning algorithm

Step 1. optimize the hyperparameter using GridSearchCV on parameter ‘gamma’ and ’C’. 5 fold cross validation is used.

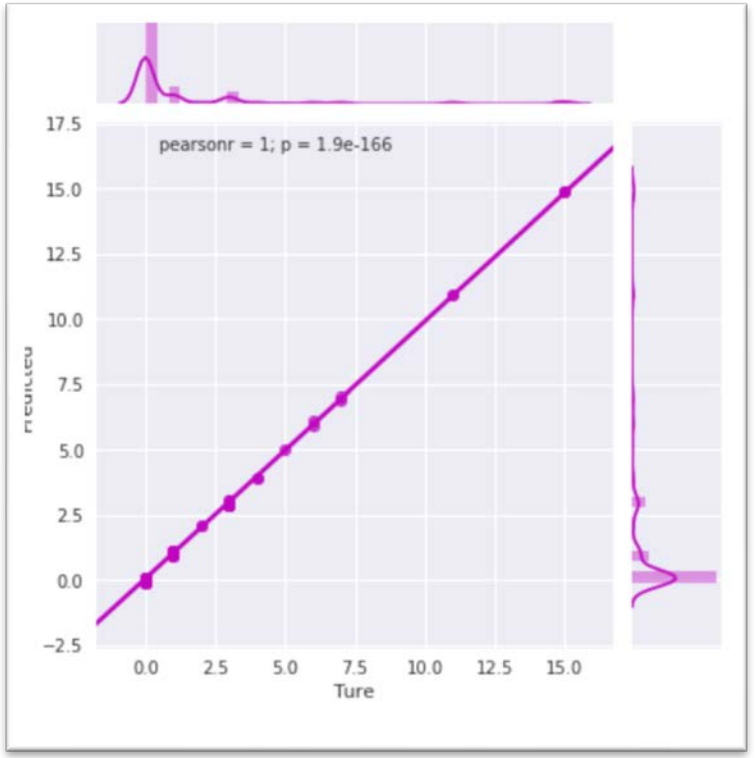
```
svr_rbf = GridSearchCV(SVR(kernel='rbf', gamma=0.1),
cv=5, param_grid={"C": [1e0, 1e1, 1e2, 1e3], "gamma":
np.logspace(-2, 2, 5)})
```

Best parameter is here

```
SVR(C=100.0, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma=0.01,
  kernel='rbf', max_iter=-1, shrinking=True, tol=0.001, verbose=False)

  Training score=0.99909028    CV score:  R2=0.40262049
```

Step 2. Train the dataset the plot prediction form the model and True value



Get information of target city

- Neighborhoods infomation is get from

http://www.city-data.com/nbmaps/neigh-Columbus-Ohio.html

And GPS location is get from

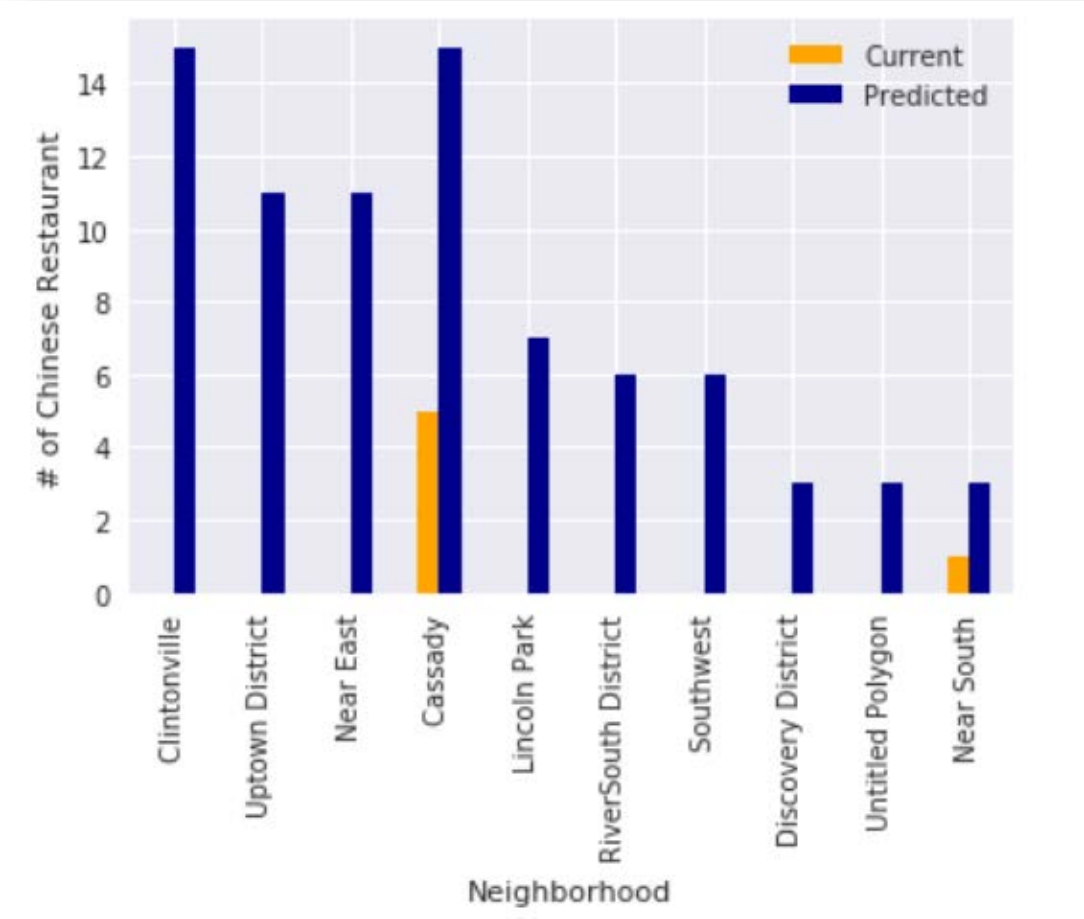
https://www.gps-coordinates.net/

	Neighborhood	Latitude	Longitude
0	Arena District	39.968959	-83.005251
1	Brewery District	39.951159	-83.001111
2	Cassady	39.998355	-82.930182
3	Clintonville	40.052178	-83.009280
4	Discovery District	39.963340	-82.996524

- Get venues information similar to Toronto

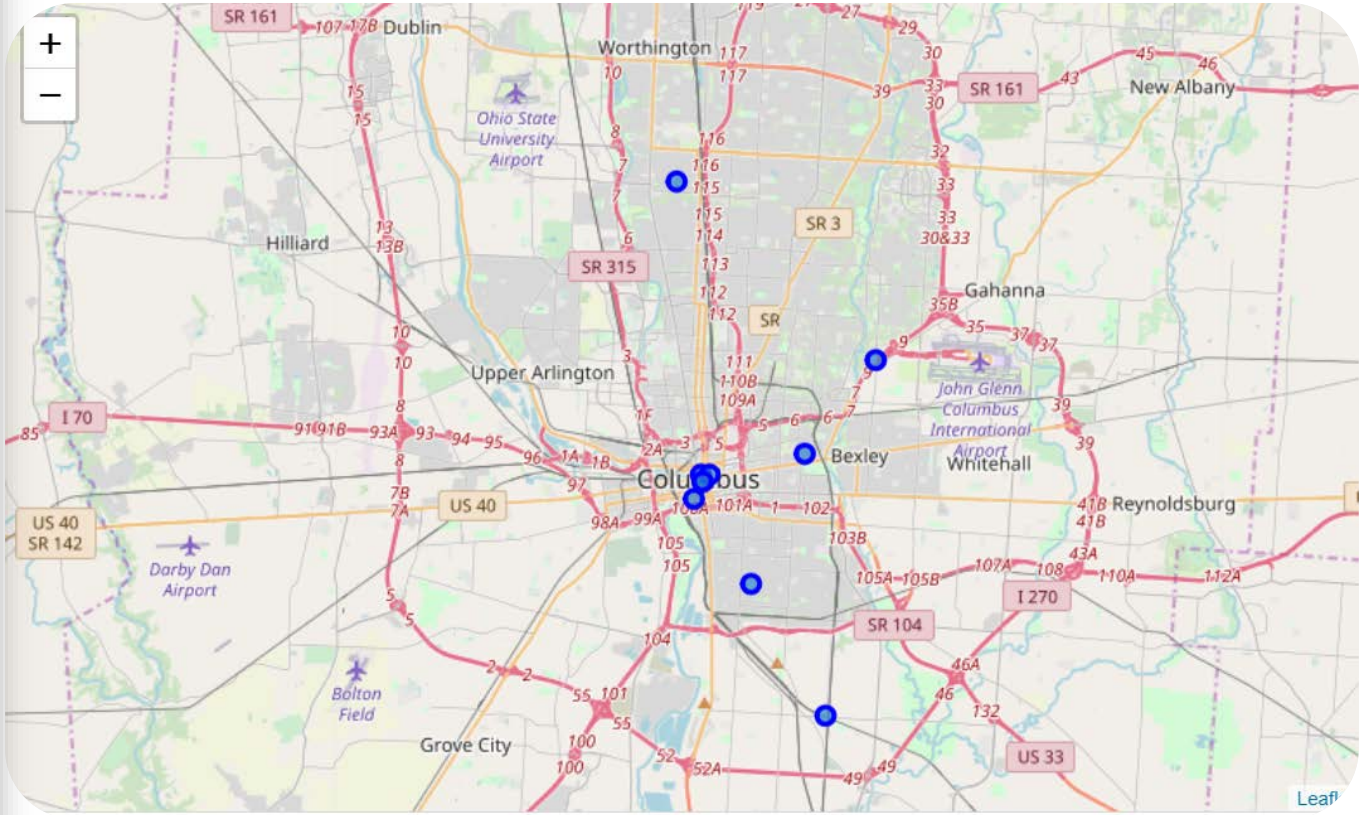
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Arena District	39.968959	-83.005251	Nationwide Arena	39.969135	-83.006098	Hockey Arena
1	Arena District	39.968959	-83.005251	North Market	39.971891	-83.004321	Market
2	Arena District	39.968959	-83.005251	Hilton Columbus Downtown	39.970826	-83.002724	Hotel
3	Arena District	39.968959	-83.005251	Jeni's Splendid Ice Creams	39.971903	-83.004322	Ice Cream Shop
4	Arena District	39.968959	-83.005251	Hot Chicken Takeover	39.971527	-83.004470	Fried Chicken Joint

Predict using trained model



Top 10 recommendations for start your business in Columbus

Top 10 recommendations for start your business in Columbus (Map)



Conclusion

- We used the Foursquare API get the venues information on given locations
- We build predictive models with SVR algorithm
- We get the top 10 recommendations of location to invest “Chinese Restaurant ” in Columbus

These things can make it better:

1. This model is built on the assumption that the target city will have a trend to grow to “big-city” like we used into model training.
2. The training dataset still very small, if we can get more data from more big cities , we can make the model better
3. Foursquare app can only give 100 venues exploration on free version, it is better to conclude all of the venues to avoid bias coming from the sampling

Acknowledgement

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