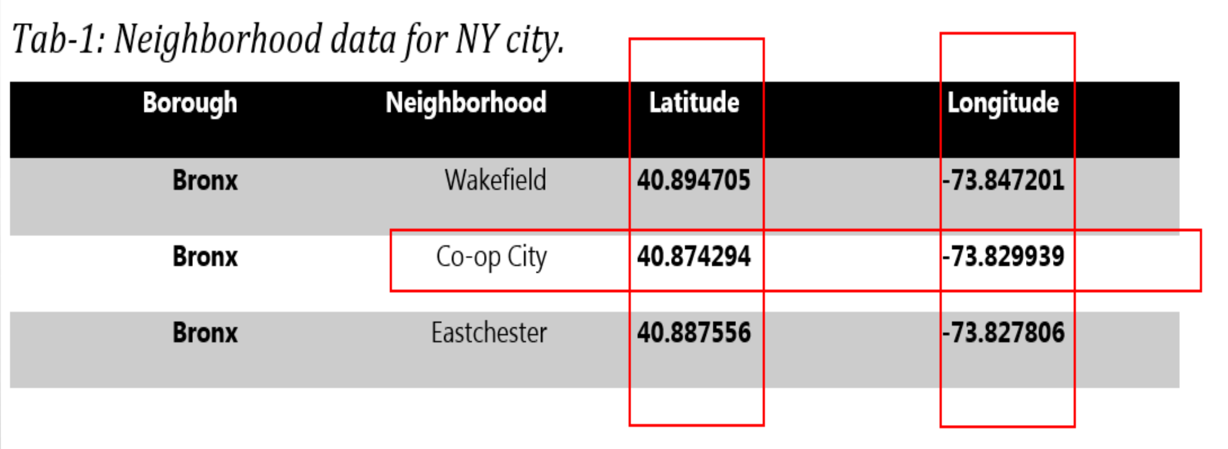
CAPSTONE PROJECT
– THE BATTLE OF NEIGHBORHOODS

Predict How Successful A New Restaurant Will Be

- **Some factors’ combination determines whether a restaurant to be finally successful**
  - Location of the restaurant often decide if you can get high quality clients
  - Food to be offered often decide if your serving cuisines attractive
  - Social network comments decide if your restaurant easier to own good reputation
  - These raw data is accessible from Foursquare and other websites through API or crawler
- **New restaurant investors expect prediction before the business opening**
  - A score will be predicted upon easy inputs such as the targeted location and planned food list
  - Keep trying the prediction service to get best combination of addresses and various menu option
  - Risk of investment on wrong choices could be greatly decreased from beginning

Feature Data Acquisition

- Rough venue information accessed from *2014 new york city neighborhood names*

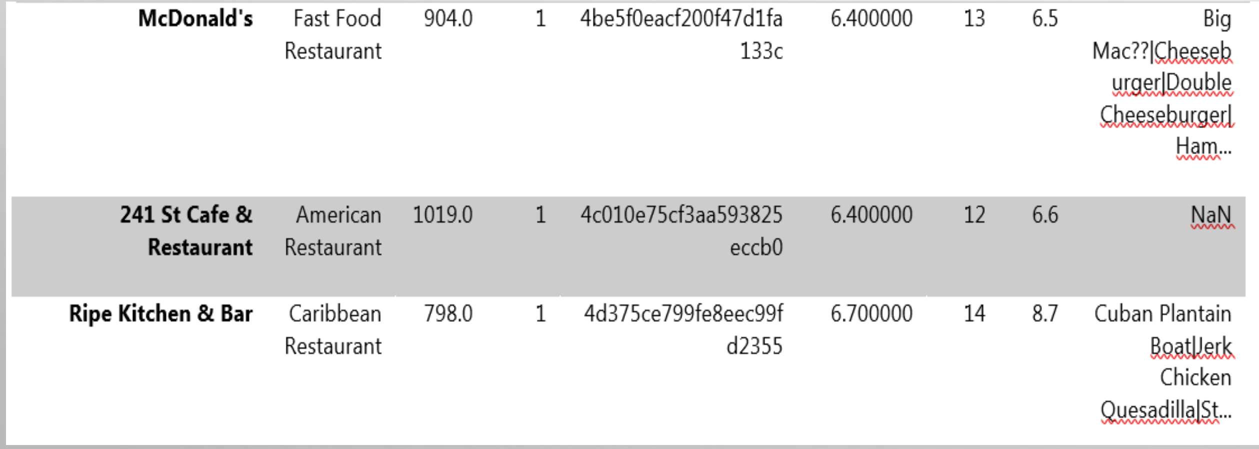


Tab-1: Neighborhood data for NY city.

Borough	Neighborhood	Latitude	Longitude
Bronx	Wakefield	**40.894705**	**-73.847201**
Bronx	Co-op City	**40.874294**	**-73.829939**
Bronx	Eastchester	**40.887556**	**-73.827806**

https://geo.nyu.edu/catalog/nyu_2451_34572

- NY Restaurants' detail information searched through Foursquare API per above location (latitude, longitude)



McDonald's	Fast Food Restaurant	904.0	1	4be5f0eacf200f47d1fa133c	6.400000	13	6.5	Big Mac??|Cheeseburger|Double Cheeseburger|Ham...
241 St Cafe & Restaurant	American Restaurant	1019.0	1	4c010e75cf3aa593825eccb0	6.400000	12	6.6	NaN
Ripe Kitchen & Bar	Caribbean Restaurant	798.0	1	4d375ce799fe8eec99fd2355	6.700000	14	8.7	Cuban Plantain Boat|Jerk Chicken Quesadilla|St...

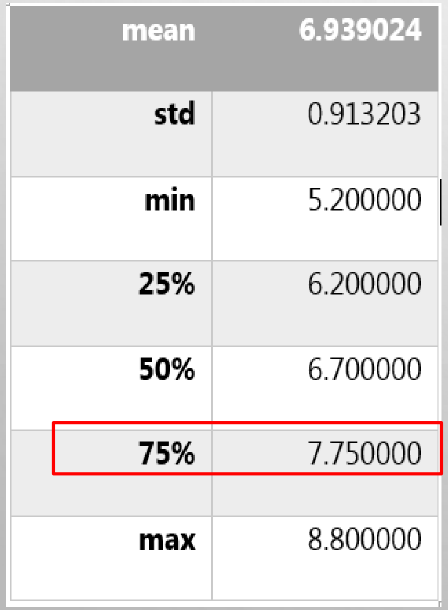
- https://api.foursquare.com/v2/venues/search?&query=Restaurant
- https://api.foursquare.com/v2/venues/{ restaurant id }
- https://api.foursquare.com/v2/venues/explore/ll{restaurant Latitude, Longitude}
- https://api.foursquare.com/v2/venues/{ recommend venue id }
- https://api.foursquare.com/v2/venues/{restaurant id}/menu

Data Preprocessing

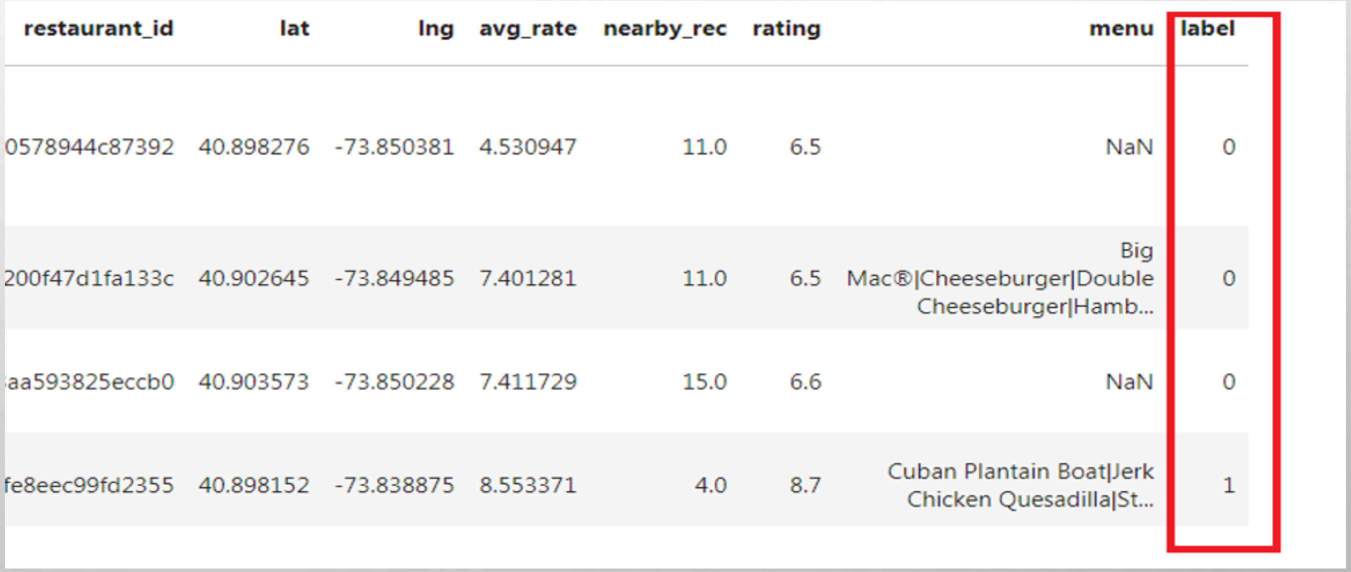
- Following features are selected as input of the model training
  - category
  - average distance to neighborhoods
  - number of nearby neighborhoods
  - average nearby rating
  - recommended nearby popular sites
  - menu item

- Label or target variable: rating

take the 75% value 7.75 as a threshold, if rating larger than 7.75, label it as "Good"(1); or else label it as "not good"(0).



mean	6.939024
std	0.913203
min	5.200000
25%	6.200000
50%	6.700000
75%	7.750000
max	8.800000

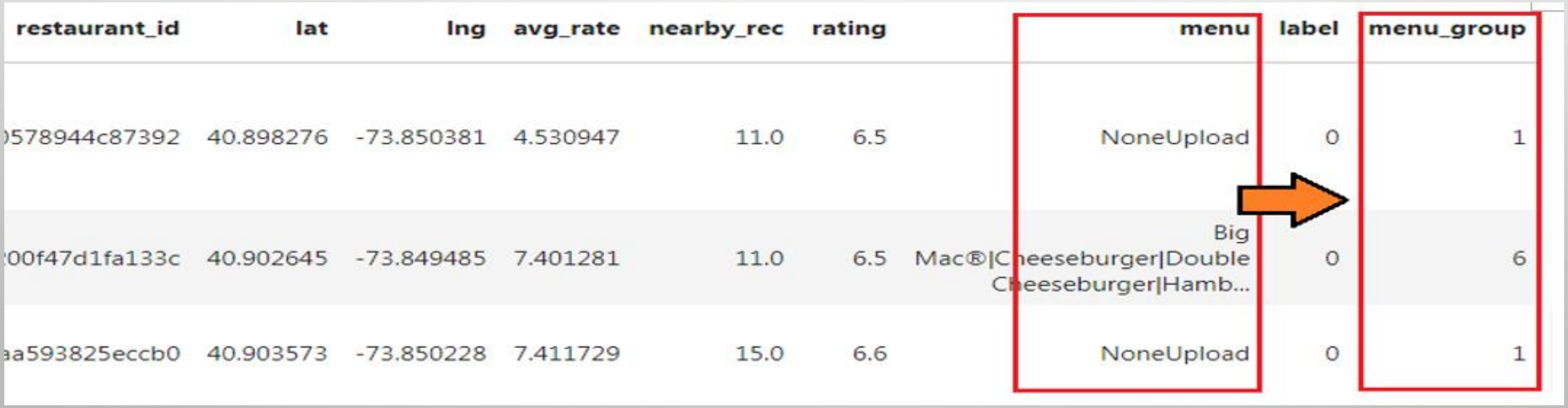


restaurant_id	lat	lng	avg_rate	nearby_rec	rating	menu	label
0578944c87392	40.898276	-73.850381	4.530947	11.0	6.5	NaN	0
200f47d1fa133c	40.902645	-73.849485	7.401281	11.0	6.5	Big Mac®|Cheeseburger|Double Cheeseburger|Hamb...	0
aa593825eccb0	40.903573	-73.850228	7.411729	15.0	6.6	NaN	0
fe8eec99fd2355	40.898152	-73.838875	8.553371	4.0	8.7	Cuban Plantain Boat|Jerk Chicken Quesadilla|St...	1

Feature Data Vectorization- Text Attribute (menu) Clustering

Need to quantify menu texts before leveraging it in classification model:

- Clean text (by NLTK lib and Re) and transform the text into a list of bow(bag of words)
  - remove punctuations
  - remove stop words
  - Tokenize it to make the string into "words "
- Turned the above menu bow into a quantified tf-idf matrix (numpy array)
- Invoked scikit-learn KMEANS model, cluster the tf-idf matrix(menus array) into groups labels

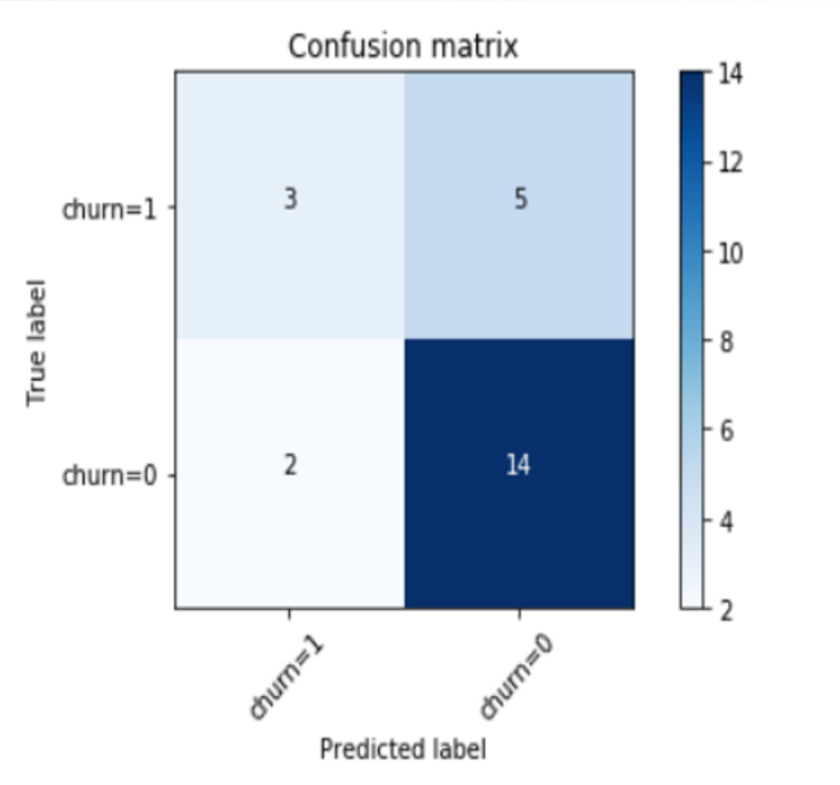


restaurant_id	lat	lng	avg_rate	nearby_rec	rating	menu	label	menu_group
578944c87392	40.898276	-73.850381	4.530947	11.0	6.5	NoneUpload	0	1
00f47d1fa133c	40.902645	-73.849485	7.401281	11.0	6.5	Big Mac®|Cheeseburger|Double Cheeseburger|Hamb...	0	6
a593825eccb0	40.903573	-73.850228	7.411729	15.0	6.6	NoneUpload	0	1

Classification Model And Evaluation

The classification models of this proposal is much straightforward, with the location and food input features, the model predicts whether these choices combined together will bring up one successful (1) or failed (0) business.

```
LR = LogisticRegression(C=0.1, solver='liblinear',class_weight={1:0.65,0:0.35}).fit(X_train,y_train)
```



Result evaluation

- *Jaccard index* = 71%.
- *Accuracy of classifier through confusion matrix:*

  The classifier correctly predicted 14 of 16 as 0, so, it has done a good job predicting the higher risk of negative result. Meanwhile 5 of 8 good rating prediction incorrectly to *risky*, this is bit high, Now that avoiding risk is our major goal, it is acceptable.

* Decision tree and SVM models are also tested, finally ignored as worse result

Conclusion And Future Directions

- I have implemented a new restaurant business success prediction algorithm based on the data accessed through Foursquare API
  - Feature selection, data preprocessing, and LR classification have been done by Pandas and Sklearn lib
  - Text vectorization is done by NLP and clustering algorithm
- Accuracy measurement of the model is acceptable, specially the risky recall rate is good. But this current version is indeed influenced by small volume of training data(as Foursquare limitation), which could have been avoided if we can use other data services, unlucky google is not for free any longer.
- Implementation can be improved by:
  - Horizontal extension by involving more features such users, and tips, etc.
  - Vertical extension by accumulating more training data over time to offset Foursquare API quota limitation for free developer