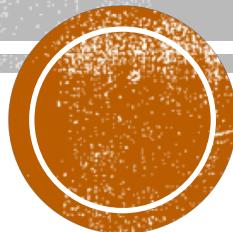


CAPSTONE PROJECT

- 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
- 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
- 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

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

McDonald's	Fast Food Restaurant	904.0	1	4be5f0eacf200f47d1fa	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 BoatJerk Chicken QuesadillaSt...

- Rough venue information accessed from 2014 new york city neighborhood names
 - https://geo.nyu.edu/catalog/nyu_2451_34572

- NY Restaurants' detail information searched through Foursquare API per above location (latitude, longitude)
 - <https://api.foursquare.com/v2/venues/search?&query=Restaurant>
 - <https://api.foursquare.com/v2/venues/{restaurantid}>
 - <https://api.foursquare.com/v2/venues/explore/ll{restaurant Latitude, Longitude}>
 - <https://api.foursquare.com/v2/venues/{recommendvenueid}>
 - <https://api.foursquare.com/v2/venues/{restaurantid}/menu>



DATA PROCESSING

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

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



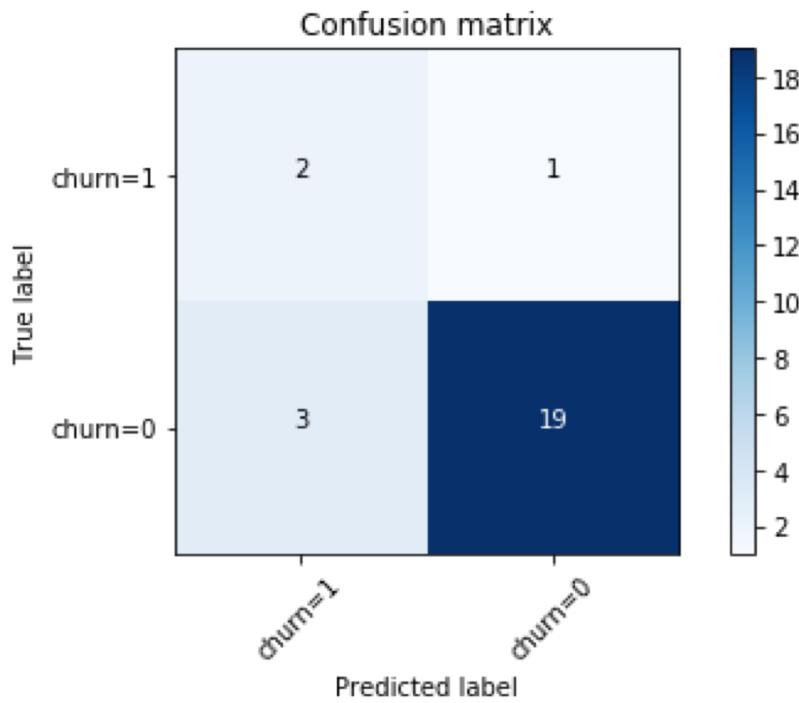
restaurant_id	lat	lng	avg_rate	nearby_rec	rating	menu	label
0578944c87392	40.898276	-73.850381	4.530947	11.0	6.5	NaN	0
200f47d1fa133c	40.902645	-73.849485	7.401281	11.0	6.5	Mac® Cheeseburger Double Cheeseburger Hamb...	0
aa593825eccb0	40.903573	-73.850228	7.411729	15.0	6.6	NaN	0
fe8eec99fd2355	40.898152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



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)
```



Jaccard index = 84%.

Accuracy of classifier through confusion matrix:

The classifier correctly predicted 19 of 22 as 0, so, it has done a good job predicting the higher risk of negative result.

Meanwhile 2 of 3 good rating prediction incorrectly to *risky*, this is bit high, Now that avoiding risk is our major goal, it is acceptable.

