Report on Capstone Project – The Battle of Neighborhoods

**Making Predictions on Critical Violations at Manhattan Italian Restaurants Using Foursquare and DOHMH Data**

Doc Woo, May 16, 2020

1. **Introduction**

In this work, I would use the customer ratings obtained from Foursquare and the New York City DOHMH inspection scores to make predictions on the occurrences of critical violations at Manhattan Italian restaurants.

DOHMH stands for Department of Health and Mental Hygiene. It conducts regular inspections on the restaurants to check that restaurants comply with food safety rule. It publishes the result data report and routinely updates the dataset based on the latest inspections. Only restaurants in an active status are included in the dataset. The officials give scores on the restaurant that an inspection score of 0 to 13 is an A, 14 to 27 points is a B, and 28 or more points is a C. They would issue a critical flag indicating critical or not critical violations. Critical violations are those most likely to contribute food-borne illness. Any violation would contribute to the score while a critical flag does not necessarily lead to a non-A score and vice versa.

The DOHMH report includes the latitude and longitude data of the restaurants. With that information, I can get the customer ratings from Foursquare. The customers expectedly would not rate the restaurant high if they felt it was not clean or felt ill after dinning there. So the ratings would give useful information, an extra dimension to predict/determine the likelihood of the occurrence of critical violations.

This work would focus on the Italian restaurants in Manhattan.

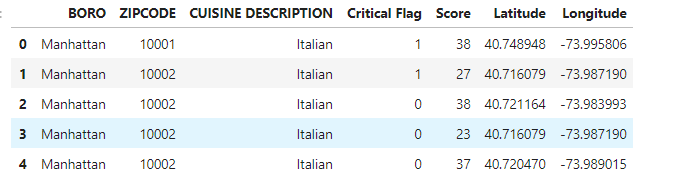
1. **Data**

In this works, two data sources were utilized. First one is the DOHMH restaurant inspection report. The report can be exported at <https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/rs6k-p7g6> . The report contains 389K rows and 26 columns. The columns include information such as Borough, Address, Zip Code, Phone, Cuisine Description, Inspection Date, Violation Code, Critical Flag, Score, Grade, Record Date, Latitude, Longitude etc. The information I would use is Borough, Zip Code, Cuisine Description, Critical Flag, Score, Latitude and Longitude. Particularly, I was interested in the data of ‘Italian’ in ‘Cuisine Description’ and ‘Manhattan’ in ‘Borough’.

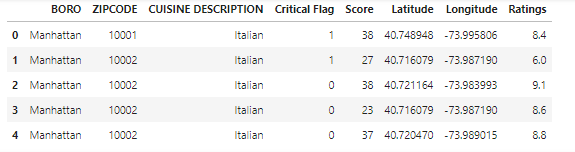
After getting the latitude and longitude, I was able to use this information and the Foursquare API to successfully collect the customer’s ratings data. I intended to query for ~300 restaurants, but due to the quota I got 50 ratings that appeared to be okay to make the modeling and predictions.

1. **Preparing and Preprocessing Data**

I established a sub-dataset from DOHMH inspection report dataset. Because the DOHMH dataset is compiled from several large administrative data systems. It contains some illogical values and some data are missing. So I removed all the records with missing values. And then used Python to do further data processing – (1) I removed all the records with score of more than 50 that might be a mistaken value or typo. (2) I recalculated score with 50 minus the original score that make this value more in line with ratings. The higher value the better. The figure below shows the first 5 records after processing:



With latitude and longitude available, I setup the query using the latitude and longitude information to find the restaurants and collect the customer’s ratings from Foursquare. Then I combined the ratings with above dataset. The new one looks like this:

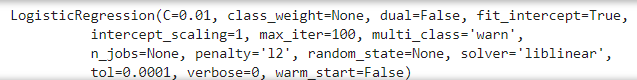


Please refer to my Python code for more details.

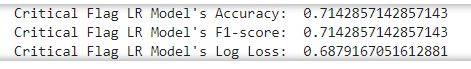
1. **Modeling with Logistic Regression Method**

With dataset ready to go, I selected logistic regression (LR) approach for the modeling. As learned from the class, the logistic regression method is good when the target field is categorical, particularly fit for binary cases.

At first, the dataset was preprocessed as the rows with 0.0 ratings were removed. After that, features selection was made including Score and Ratings. And the output to predict is Critical Flag. Then, a standard normalization step was taken for the feature dataset. Subsequently the train and test sets splitting was taken with which train set is for modeling and test set is for evaluating the developed model. The next step is to call the Logistic Regression function and utilize the train set to establish the LR model that shows below:



For the final step, I used the test set to evaluate the models. Three indexes were used: Jaccard index, F1-score and Log Loss. The results are shown below.



1. **Conclusions**

As the evaluation results show, the developed model presented satisfying Jaccard index and F1-score that coincidently present the same values. If more data can be used, the value of Log Loss would be expected to improve.

Overall, from this project I had learned more important things beyond using Python, Foursquare and other tools – I have achieved a solid idea on how to start and finish a Data Science project, thereby ultimately benefiting my preparation for becoming a data scientist.