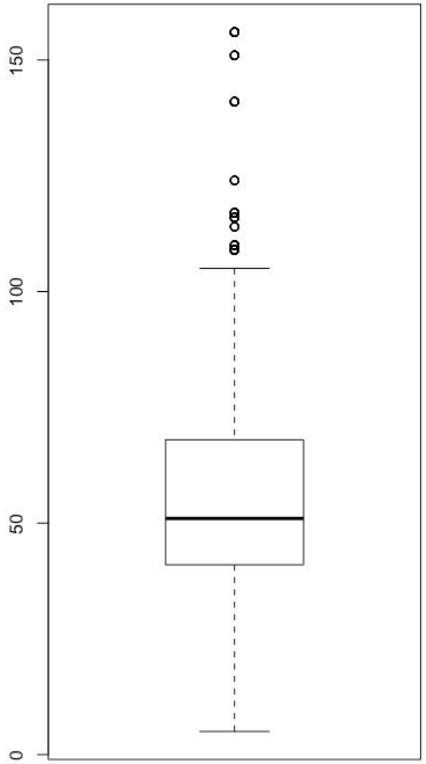
New York City and its restaurants

QMSS5052 Practicum research project: Kyi Yeung Goh

BACKGROUND AND RATIONALE

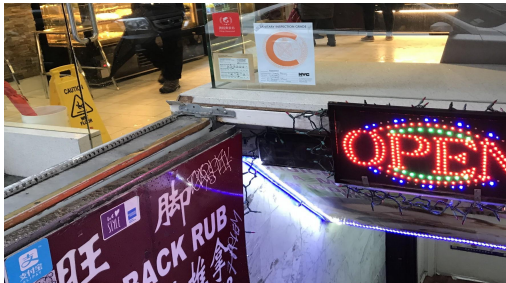


- Lack of a clear pattern as to why restaurants were closed
- Understand and predict the probability of closure based on past results
- Why does this project (hopefully) make sense?
  a. Previous literature using NYC Open Data on restaurants do not generally concern themselves with this exciting question
  b. Current analyses of the data have made assumptions that are potentially problematic
  c. Suspicious wealth of opportunity to integrate other data points to assist our analysis

Some strive to improve…

Some make the best of what they get…

Some could not Care less…



KEY METHODS

$$logit(\boldsymbol{p}) = \beta_0 + \beta_1 SCORE_i + \beta_2 VIOLATION.TYPE_i + \beta_3 log(Rate)_i + \beta_4 CRITICAL.VIOLATION_i + \beta_5 RATING_i + \beta_6 hhincome_i \ \varepsilon$$

, where p is the log probability of a restaurant being closed by health inspectors

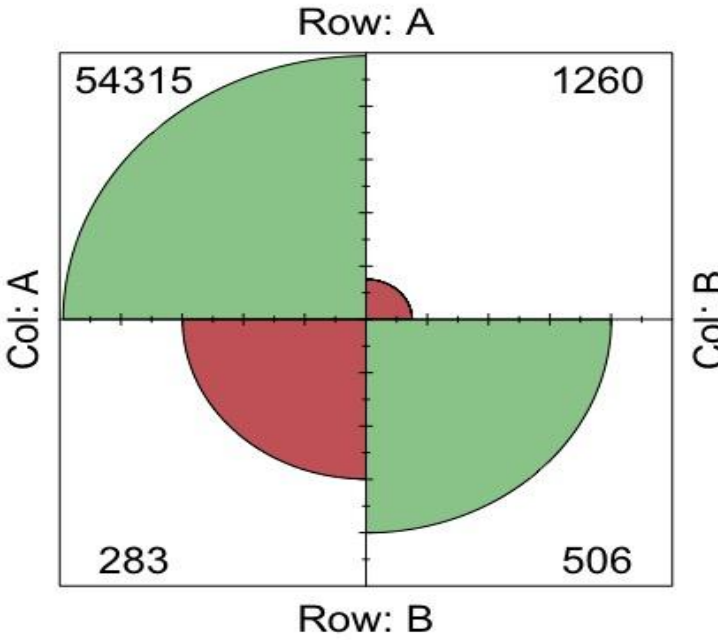
- Work primarily done in R, Excel for smaller sample sizes
- Ran a multiple logistic regression model
- Called on four different sources of data: New York Restaurants data, rat inspection dataset, census data for neighborhood income, Google Locations API*
- Utilised packages such as *stringr, fuzzyjoin, tidyverse* etc. + functions to call, clean and build data into datasets
- Used *caret* to help with logistic regression prediction

RESULTS

```
Call:
glm(formula = ACTION ~ SCORE + as.factor(violationtype) + log(Rate) +
    as.factor(criticalscore) + Rating + log(hhincome), family = "binomial",
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8376  -0.1557  -0.0957  -0.0656   3.9163

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     3.164607   0.520992   6.074 1.25e-09 ***
SCORE                           0.107701   0.001271  84.734  < 2e-16 ***
as.factor(violationtype)foodhandle -0.533399   0.050560 -10.550  < 2e-16 ***
as.factor(violationtype)misc    -1.351970   0.243418  -5.554 2.79e-08 ***
as.factor(violationtype)safety   0.204676   0.084209   2.431 0.015076 *
as.factor(violationtype)workers -0.527742   0.136510  -3.866 0.000111 ***
log(Rate)                       -0.003771   0.014936  -0.252 0.800655
as.factor(criticalscore)1        0.038083   0.062512   0.609 0.542387
Rating                          -0.074515   0.031553  -2.362 0.018196 *
log(hhincome)                   -0.853568   0.046546 -18.338  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29181  on 104679  degrees of freedom
Residual deviance: 16317  on 104670  degrees of freedom
AIC: 16337

Number of Fisher Scoring iterations: 8
```

- McFadden's Pseudo R2 is around 0.47
- Some quick interpretations:
  - Each 1 unit increase in the log(hhincome) results in around a 57.4% decrease (***) in the probability of closure holding all other Xs constant.
  - Meanwhile, each 1 unit increase (out of 5) results in around a 7.19% decrease (***) in the probability of closure holding all other factors constant.
  - Score, violations are all important predictors
  - Interestingly, critical violations are not.

RESULTS (p.2)



The model has about a 30% hit rate which is encouraging but not particularly useful if we base it on the test data.

There are, unfortunately, probably more data points that can be leveraged.

CONCLUSIONS

The most important predictor variable that determines the probability is, unsurprisingly, score given by inspectors.

Yet, it is clear that other variables matter too.

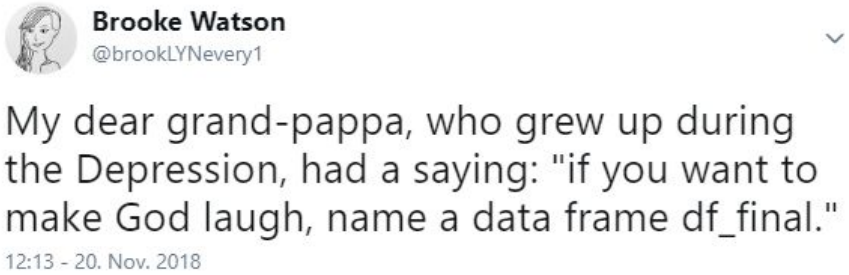
Something can also be said about the overall cleanliness of an area and the customer ratings that restaurants receive on Google.

It could explain why restaurants in particularly crowded zip codes for instance, midtown have higher-than-average closure rates.There may be a gentrifying effect in that richer areas tend to have **noticeably fewer closures**.

Further to that, customer choice also appears to be a good signal for restaurant threat to public health. It may be the case that the ambience and perceived cleanliness of the place affects review scores on Google.

LEARNING POINTS

1. Different data requires different approaches
2. There will always be a package out there that resolves your problem - if you don't try you won't know (Prof. Stack Overflow)
3. Keeping a live tab of your project to point out things that you can improve on in the future/ some interesting nuggets about the project was very useful for me
4. Always take the opportunity to have a fresh pair of eyes look at the project



BIGGEST CHALLENGES

1. Getting data structures to cooperate with me
2. Getting the data to actually contain information I want
3. Constantly re-acquainting yourself with the data + the accompanying R code
4. Knowing that there is much more that can be done but is not being done and then living with it
    a. GIS map of restaurants based on ZIP (interactive dashboard)
    b. Including more predictors/ weighting current ones to account for unequal review sizes etc.

REFERENCE EMAIL

“My follow-up question for your analysis is can you make a recommendation of ***how often restaurants should be inspected.***

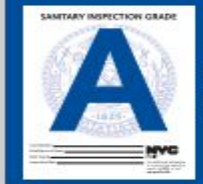
Please include one slide in your presentation next week explaining ***why this wasn't done/why it's not feasible*** *or* ***necessary to incorporate*** *it into your analysis.”*

Initial Inspection

Scoring 14+ points

Reinspection

Scoring 0–13 points



Next cycle initial inspection in 11–13 months

Scoring 0–13 points

Next cycle initial inspection in:
- 5–7 months if initial score **14–27 points**
- 3–5 months if initial score **28+ points**

Scoring 14–27 points

Either card can be posted until final grade is determined at an OATH hearing

Next cycle initial inspection in:
- 5–7 months if initial score **14–27 points**
- 3–5 months if initial score **28+ points**

Scoring 28+ points

Either card can be posted until final grade is determined at an OATH hearing

Next cycle initial inspection in 3–5 months

KEY TAKEAWAYS

Status in previous inspection	CLOSED	NOT CLOSED
% (n)	41.6 (1219)	58.4 (1714)
Some characteristics	• Mostly in Brooklyn, Manhattan • Chinese, American and Caribbean restaurants • ZIPs: 10003 (NoHo), 11237 (Bushwick)**	• Mostly in Brooklyn, Manhattan • Chinese, American and Caribbean restaurants • ZIPs: 10013 (SoHo), Sunset Park (Brooklyn)**

Perhaps,

- Consider reducing the 11-13 month window for certain ZIPs with restaurants that passed
- ** higher risk of food poisoning?

LIMITATIONS

1. Model classification can be much better.

Prediction of actual closure rate is only around ~30% as it stands as can be seen in the confusion matrix.

2. More could be integrated (Longitudinal data - prices) leading to better model fit

More data points could be integrated to provide an overall better McFadden's Pseudo-R2 as well as prediction of closure rates. One specific example is the use of location mapping to identify (i) poverty rates, (ii) percentage of subsidised housing and (iii) average household income in a predefined area around the restaurant using census data.

3. Predictions will change based on data source

The predictions are likely to change based on our data source. In this case, some replication should be done using other data sets such as Yelp, Tripadvisor amongst others.

4. Including a visual for interaction