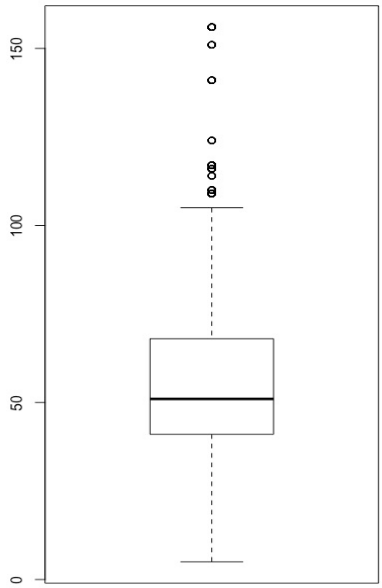
Practicum report: Modelling threats to public health from poorly-run restaurants

This report concerns itself with the question, **"Under what conditions should the New York Department of Health and Mental Hygiene close restaurants?".**



The rationale for this project is the lack of a clear pattern as to why restaurants were closed as indicated by the large variance in scores attained, indicating room for greater standardization of procedures. More specifically, scores for closed restaurants ranged from 5 to 156, with a standard deviation of over 24.2 points. As such, the aim is to create a model that can help give a probability of closure based on past results and then create an app/ portal that can help predict a result once you key in via Shiny.

This is a research question worth pursuing for three key reasons. First, previous literature using NYC Open Data on restaurants do not generally concern themselves with this question. Second, current analyses of the data have made assumptions that are potentially problematic. Some of these include using restaurant type to predict violation and assuming region to be specific enough to be representative of location effect. Third, there exists a wealth of opportunity to integrate other data points to assist our analysis. Here, we complement the restaurant dataset with Google's restaurant API and hand-coded ZIP code-level pest data from Open NYC. Our model utilises a logistic regression function to provide a log probability that a restaurant should be closed using past data of around 400,000 observations.

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

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

The most significant one is that each observation is treated as a discrete, independent observation. Although there are repeat inspections, we can treat each as independent as they provide indication for the conditions in which a restaurant is closed. Second, the model removes a significant predictor, restaurant cuisine type, due to concerns about fairness and reproducibility of bias.

The process involved cleaning strings using stringr, writing API call functions, fuzzy merges, manual imputation of data sets, test and checking of candidate models in r. Using the caret package, we split the final cleaned set into a testing and training set on a 70-30 split. A cleaned version of the dataset is provided on GitHub.

MODEL FINDINGS:

Regression Results

	Equalize	
	Model 1	Model 2
SCORE	0.104***	0.104***
	(0.001)	(0.001)
as.factor(violationtype)foodhandle	-0.537***	-0.536***
	(0.049)	(0.049)
as.factor(violationtype)misc	-1.082***	-1.080***
	(0.219)	(0.217)
as.factor(violationtype)safety	0.140*	0.143**
	(0.081)	(0.065)
as.factor(violationtype)workers	-0.704***	-0.704***
	(0.144)	(0.144)
log(Rate)	0.052***	0.052***
	(0.013)	(0.013)
as.factor(criticalscore)1	-0.004	
	(0.060)	
Rating	-0.130***	-0.130***
	(0.030)	(0.030)
Observations	106,260	106,260
Log Likelihood	-8,586.911	-8,586.913
Akaike Inf. Crit.	17,191.820	17,189.830
Note:	*p<0.1; **p<0.05; ***p<0.01	

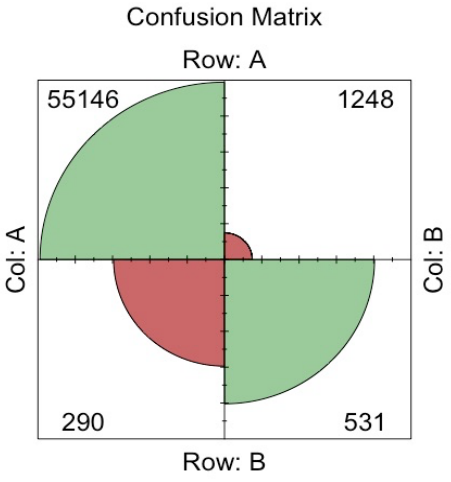
The most important predictor variable that determines the probability is, unsurprisingly, score given by inspectors. However, as previously indicated, closed restaurants have a large spread in terms of scores attained. The model, according to the pseudo-R2 McFadden's statistic, explains around 42% of the variance in our data for the outcome variable of concern.

Yet, it is clear that other variables matter too. Here, we look at the absolute value of the t-statistic for each model parameter. This technique is utilized by the *varImp* function in the caret package for general and generalized linear models. In this case, violations that fall under the category of food handling appear to be weighted more heavily than those that are miscellaneous, workers or safety where closure is concerned.

Something can also be said about the overall cleanliness of an area and the customer ratings that restaurants receive on Google. It could be that restaurants which are located in less clean areas have greater rates of closure given that pests, diseases and dirt are mobile. It could explain why restaurants in particularly crowded zip codes for instance, midtown have higher-than-average closure rates.

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.

MODEL 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.