Predicting Food Safety Violations in Toronto Restaurants

Prepared by David Horan

Introduction

The City of Toronto's DineSafe Inspection and Disclosure System monitors and reports on food safety at Toronto's eating and drinking establishments (City of Toronto, 2016). This project was intended to identify predictive variables for food safety violations using regression and classification techniques in R. Some of the variables that were explored included:

- Geography: Are there specific 'hotspots' for food safety violations in the City of Toronto?
- Establishment Type (eg. Food Take Out, Restaurant, Mobile Cart)
- Frequency of DineSafe inspections

This research was designed to identify establishments that are more prone to food safety violations and therefore pose greater health risk. This could improve the deployment of DineSafe inspectors and increase public knowledge of food safety risk factors.

Literature Review

The most recent published study by Toronto Public Health on foodborne illness is dated 2009 (Toronto Public Health, 2009). It stated that the incidence of foodborne illness between 2003 and 2007 was about 30% lower than it was between 1998 and 2002. Toronto Public Health noted that this decrease coincided with the introduction of the DineSafe restaurant inspection and disclosure program, but did not conclude that the program was responsible for the reduction in cases.

There have been a number of prior studies to determine the predictive value of food inspection scores in relation to foodborne illness outbreaks:

- A 2001 study (Cruz, 2001) compared the inspection reports of restaurants with illness outbreaks in 1995 with a control group of randomly selected restaurants that had no reported outbreaks. It found that the inspection ratings did not predict outbreaks. However, restaurants with illness outbreaks were three times more likely than controls to have been cited for the presence of insects. Larger seating capacity was also found to be a variable associated with outbreaks.
- In 2004, an extensive study was published by the US Centers for Disease Control (Jones, 2004) based on data from over 167,000 restaurant inspections in the United States. It found that the "mean scores of restaurants experiencing foodborne disease outbreaks did not differ from restaurants with no reported outbreaks," and attributed this to a lack of uniformity of restaurant inspections. Similar to the Cruz study, Jones found two critical violations (proper

storage of toxic items and good handwashing / hygiene practices) were associated with a higher risk of illness outbreak.

- In 2012, a study was performed on an outbreak that resulted in more than 1,900 illnesses across 11 U.S. states (Petran, 2012). This study found no overall association between the food inspection scores at the outbreak restaurants before and after the outbreak; however certain food safety violations were more common at those sites.

Due to the limited availability of recent data on outbreaks of foodborne illness in Toronto, this study will focus on identifying predictors of "Significant" and "Crucial" health violations during inspections. Based on the findings of the above studies, it is reasonable to view these violations as potential indicators of illness risk.

Please refer to the 'References' section at the end of the document for details of the above-cited works.

Datasets

This project was primarily built on the following two datasets, which were sourced from the City of Toronto's Open Data Catalogue:

1. Dinesafe database (City of Toronto, n.d.):

Number of Rows: 88,159
Number of Unique Establishments: 15,362
Number of Unique Inspections: 55,419
Count of Infractions by Severity:
C- Crucial: 2,114; M – Minor: 31,394; NA – Not Applicable: 3,693; S – Significant: 20,341

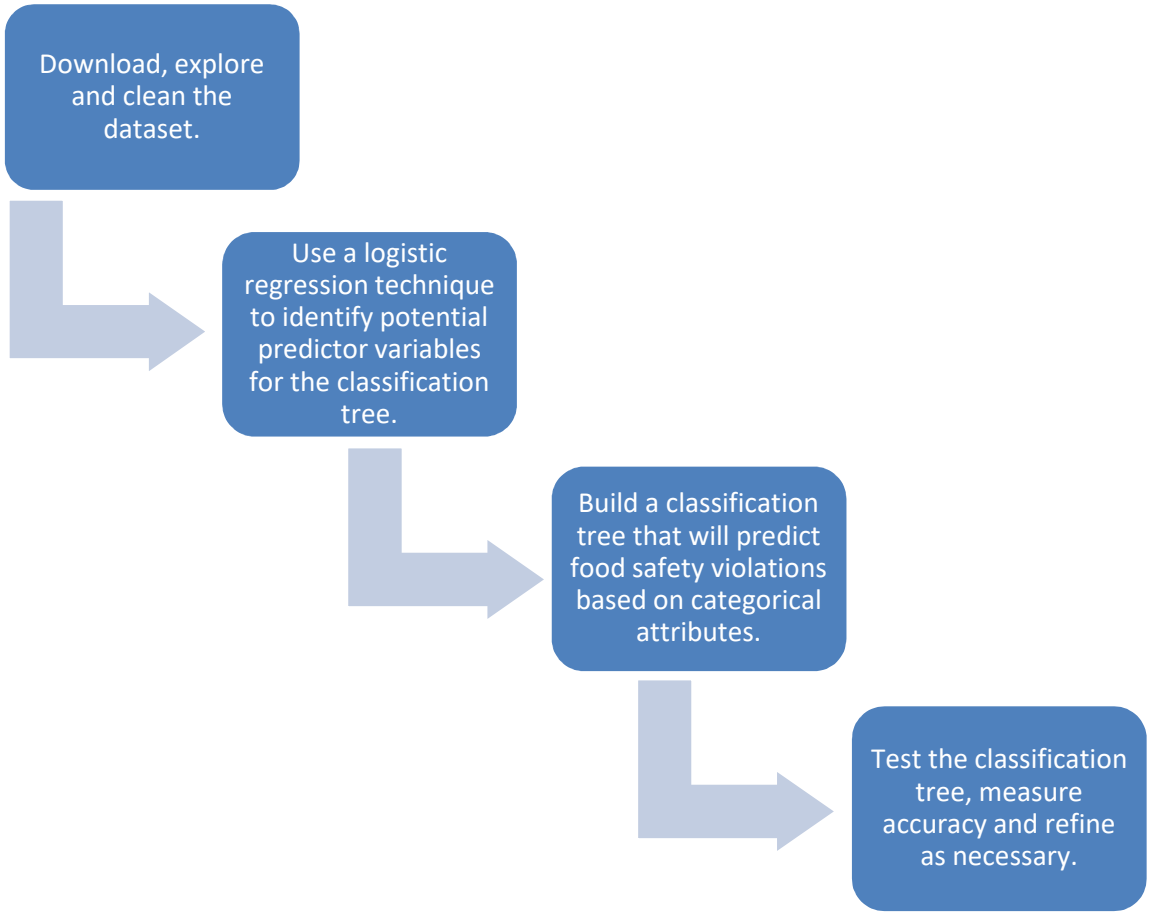
Attribute	Description	Used?
ROW_ID	Represents the Row Number	No
ESTABLISHMENT_ID	Unique identifier for an establishment	No
INSPECTION_ID	Unique identifier for each Inspection	Yes
ESTABLISHMENT_NAME	Business name of the establishment	Yes
ESTABLISHMENTTYPE	Establishment type ie restaurant, mobile cart	Yes
ESTABLISHMENT_ADDRESS	Municipal address of the establishment	Yes
ESTABLISHMENT_STATUS	Pass, Conditional Pass, Closed	No
MINIMUM_INSPECTIONS_PERYEAR	Every eating and drinking establishment in the City of Toronto receives a minimum of 1, 2, or 3 inspections each year depending on the specific type of establishment, the food preparation processes, volume and type of food served and other related criteria	Yes
INFRACTION_DETAILS	Description of the Infraction	No
INSPECTION_DATE	Calendar date the inspection was conducted	Yes
SEVERITY	Level of the infraction, i.e. S – Significant, M – Minor, C – Crucial	Yes
ACTION	Enforcement activity based on the infractions noted during a food safety inspection	No
COURT_OUTCOME	The registered court decision resulting from the issuance of a ticket or summons for outstanding infractions to the Health Protection and Promotion Act	No
AMOUNT_FINED	Fine determined in a court outcome	No

2. Address Points (Municipal) – Toronto One Address Repository (City of Toronto, n.d.)

A second dataset was also retrieved from the City of Toronto Open Data Catalogue. This dataset was solely used of the purpose of matching street addresses in the inspection file to the former municipalities of the City of Toronto (Scarborough, North York, etc).

Approach

The following is a high-level description of the approach to this project:



Download, explore and clean the dataset

(*Please refer to the code in Steps 1 - 12 of the accompanying technical report*)

The DineSafe dataset was downloaded in XML format, and parsed using the XML package in R. The number of columns (14) and descriptions were verified against the data dictionary published by the City of Toronto to ensure that the file was parsed correctly.

Exploration of the dataset revealed:

- One record may represent a single inspection (with a unique inspection_ID); however, there are sometimes multiple records per inspection, representing multiple infractions identified during a single inspection.
- The majority of establishments in this dataset are simply classified as 'Restaurant', with a large number of smaller categories. To mitigate this, a number of small categories (e.g. Child Care, Schools, Retirement Homes, Hospitals, etc) were consolidated into a group called 'Institutions.'
- The Amount_Fined attribute is entirely NULL.

Use a logistic regression technique to identify potential predictor variables for the classification tree.

(*Please refer to the code in Steps 13 - 14 of the accompanying technical report*)

The variables selected for this analysis were primarily categorical in nature (eg. Municipality, Establishment Type), therefore a dummy variable was created for each category.

The GLM linear model function in R was used to generate the model.

GLMULTI was also used to test a number of different logistic regressions. The results of the GLMULTI iterations were used to identify variables that would be input into the classification model.

Build and prune a classification tree that will predict food safety violations based on categorical attributes.

(*Please refer to the code in Steps 15 – 16 of the accompanying technical report)*

Before building the tree, the dataset from prior step was divided into a training set (approx. 70% of the observations) and a test set (approx. 30% of the observations).

First, an attempt was made to build a classification tree in R using the 'rpart' package.

A second attempt was made using the 'tree' package in R.

Finally, the 'randomForest' package in R was used.

The outcome of these techniques will be discussed in the Results section.

Test the classification tree, measure accuracy and refine as necessary.

The effectiveness of the tree was tested using the test set created in the previous step, and the accuracy of the results was measured using a Confusion Matrix.

Results

A visual exploration of the data suggested a positive correlation between Significant and Crucial violations and number of inspections per year (chart 1, below), as well as varying results in former municipalities of Toronto (chart 2, below).

Chart 1:

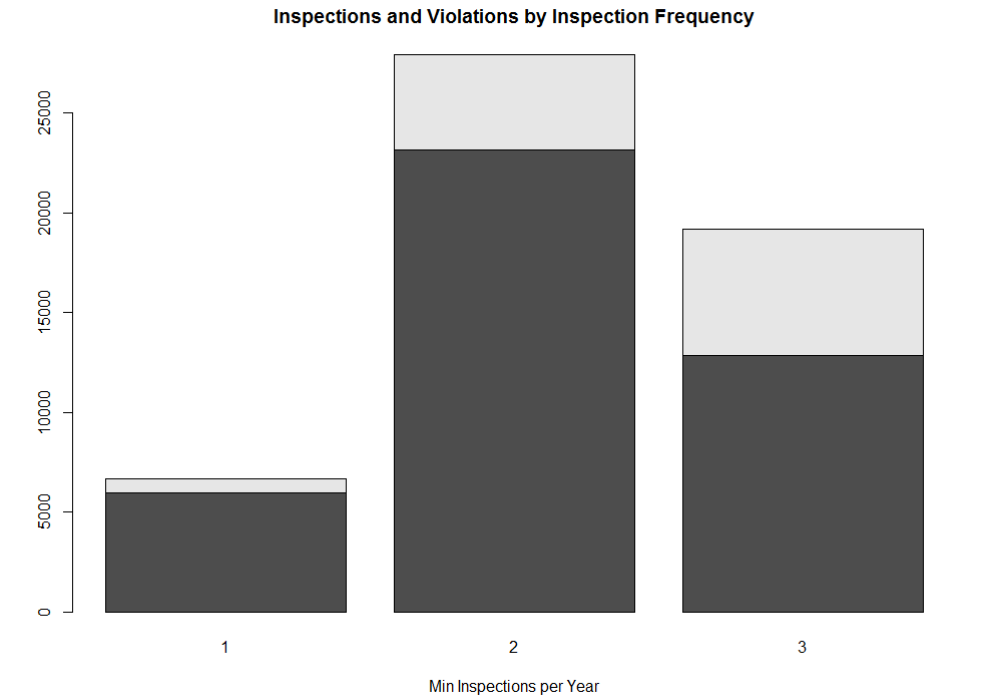
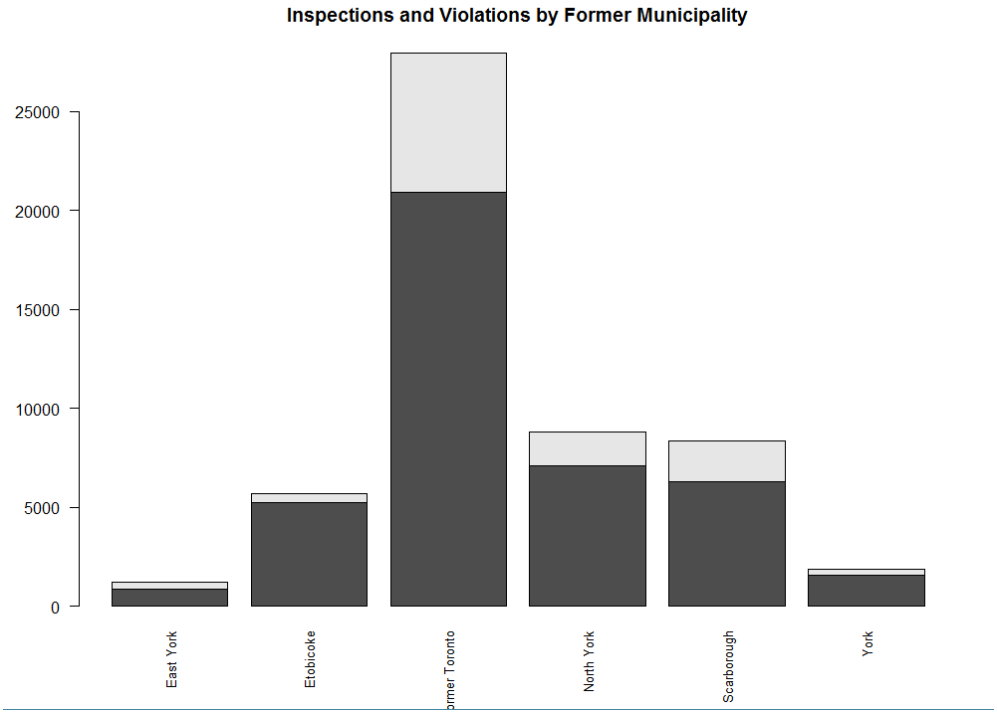


Chart 2:



A logistic regression was run to identify likely predictor variables for the classification tree. The regression indicated a number of possible variables (please refer to step 14 of the accompanying technical report for output).

These variables were used to formulate a decision tree that could be used to predict Significant and Crucial violations.

- Construction of a Classification Tree was attempted using two different packages (RPART and TREE), but neither generated a meaningful output.
- A Random Forest was attempted, but again was unsuccessful. The model yielded "False" predictions 100% of the time, or about a 22% False Negative ratio. Refer to Table 1 below for the output.

Table 1:

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE  TRUE
     FALSE 14658  4169
     TRUE      0     0

               Accuracy : 0.7786
                 95% CI : (0.7726, 0.7845)
    No Information Rate : 0.7786
    P-Value [Acc > NIR] : 0.5042

                  Kappa : 0
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.7786
         Neg Pred Value :    NaN
             Prevalence : 0.7786
         Detection Rate : 0.7786
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : FALSE
```

This result may be due to the lack of dispersion in the predictor variables. For example, nearly 50% of the observations were at foodservice locations in the former City of Toronto. Also, roughly 50% of the observations were at Restaurant locations (vs. Food Courts, Institutional foodservice, etc). Finally, 52% of the inspections were at locations inspected twice per year.

Another possible explanation for this outcome is the relatively small percentage of Significant and Crucial violations overall (approximately 22% of observations). While this would not qualify as a 'rare event,' the predictive accuracy of the Classification Tree must be very high to account for the fact that a naïve prediction of "False" will be accurate roughly 78% of the time.

Conclusions

Clearly, the model developed in this project cannot be used as-is to predict food safety violations. However, it is a starting point upon which further enhancements can be made. For example:

- The analysis could perhaps be improved in future by using a 'biased sample' technique to overcome the lack of dispersion in the independent variables.
- Additional variables, from different sources and with potentially greater predictive power, could be added to the model.

This analysis also illustrated the difficulty of predicting food safety violations, and therefore the importance of continued regular inspections and vigilance at all types of facilitates.

For the author, this project offered a valuable learning experience on the challenge of modelling with real-world categorical data.

References

City of Toronto. (2016). *Dinesafe - Toronto Public Health*. Retrieved from http://www.toronto.ca/health/dinesafe/index.htm#inspectionResult

City of Toronto. (n.d.). *Address Points (Municipal) - Toronto One Address Repository*. Retrieved from http://opendata.toronto.ca/gcc/address_points_wgs84.zip

City of Toronto. (n.d.). *DineSafe - Health - Data Catalogue*. Retrieved from http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=b54a5f9cd70bb210VgnVCM1000003dd60f89RCRD

Cruz, M. D. (2001). An Assessment of the Ability of Routine Restaurant Inspections to Predict Food-Borne Outbreaks in Miami-Dade County Florida. *American Journal of Public Health*, 91:821-823.

Jones, T. B. (2004). Restaurant Inspection Scores and Foodborne Dieases. *Emerging Infectious Diseases*, 10: 688-692.

Petran, R. B. (2012). Using a Theoretical Predictive Tool for the Analysis of Recent Health Department Inspections at Outbreak Restaurants and Relation of This Information to Foodborne Illness Likelihood. *Journal of Food Protection*, 11: 2016-2027.

Toronto Public Health. (2009, April). *Foodborne Illness in Toronto.* Retrieved from http://www.toronto.ca/health/dinesafe/pdf/staffreport_april15_2009_appx_a.pdf