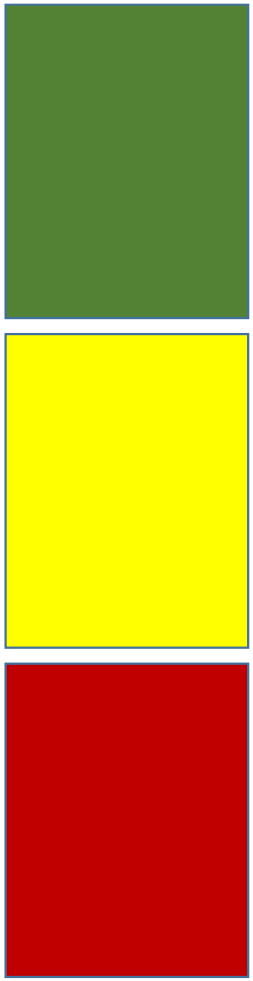


Predicting Food Safety Violations in Toronto Restaurants

Prepared by David Horan

Ryerson University CKME 136

Presented Dec 6th, 2016

Project Overview

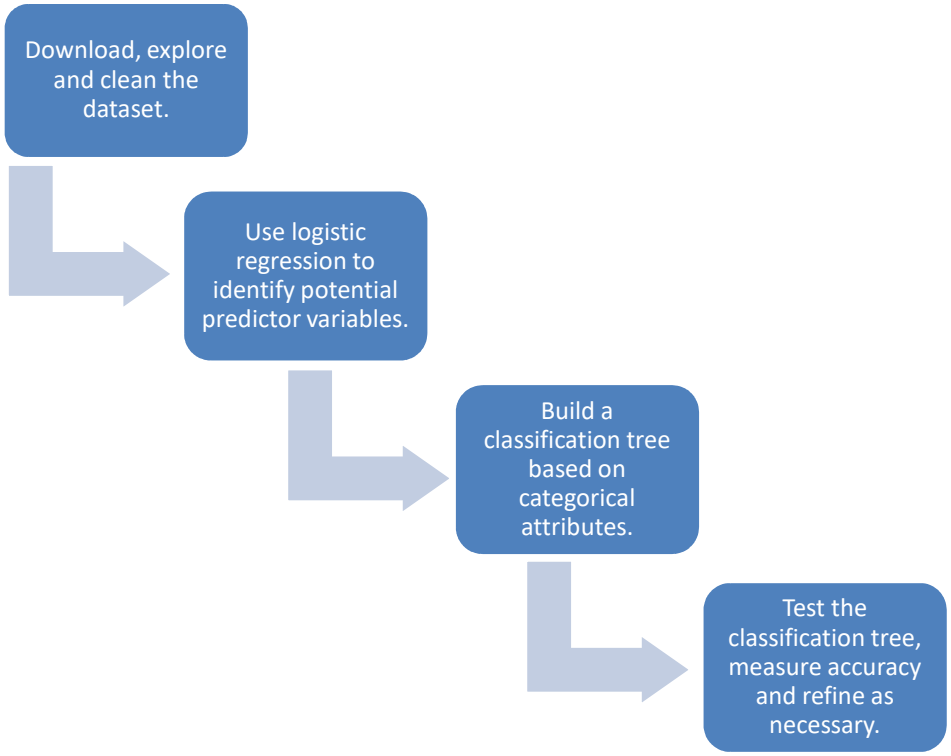
- City of Toronto DineSafe Inspection and Disclosure System monitors and reports on food safety at Toronto eating establishments

- Project Objective: Identify predictive variables for food safety violations using regression and classification techniques in R

- Purpose:
  i. Identify establishments that are more prone to food safety violations and therefore pose greater health risk
  ii. Optimize deployment of inspectors from Toronto Public Health

High-level Approach

Process



Data exploration

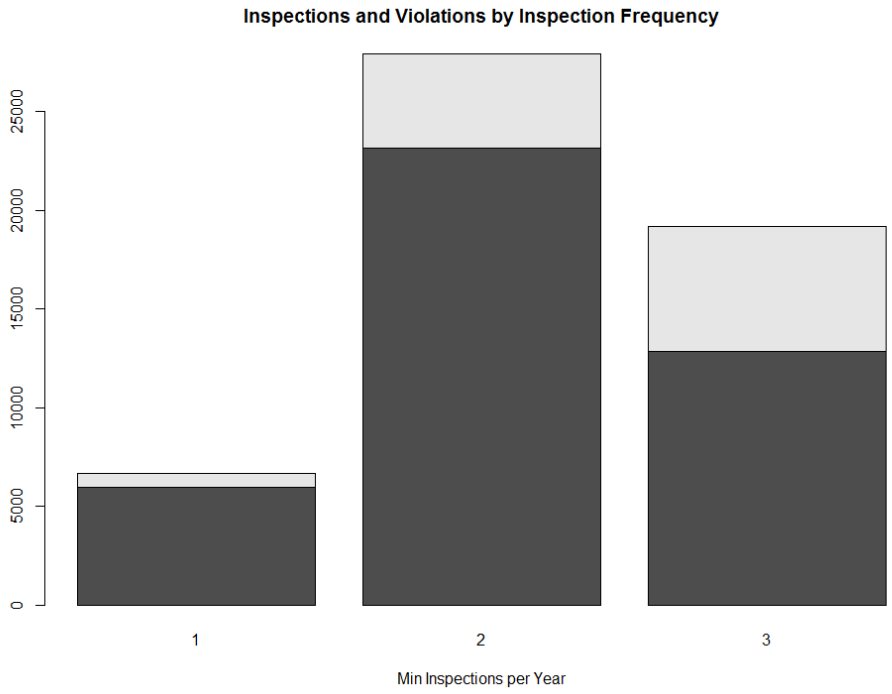
Top 10 Establishment Types

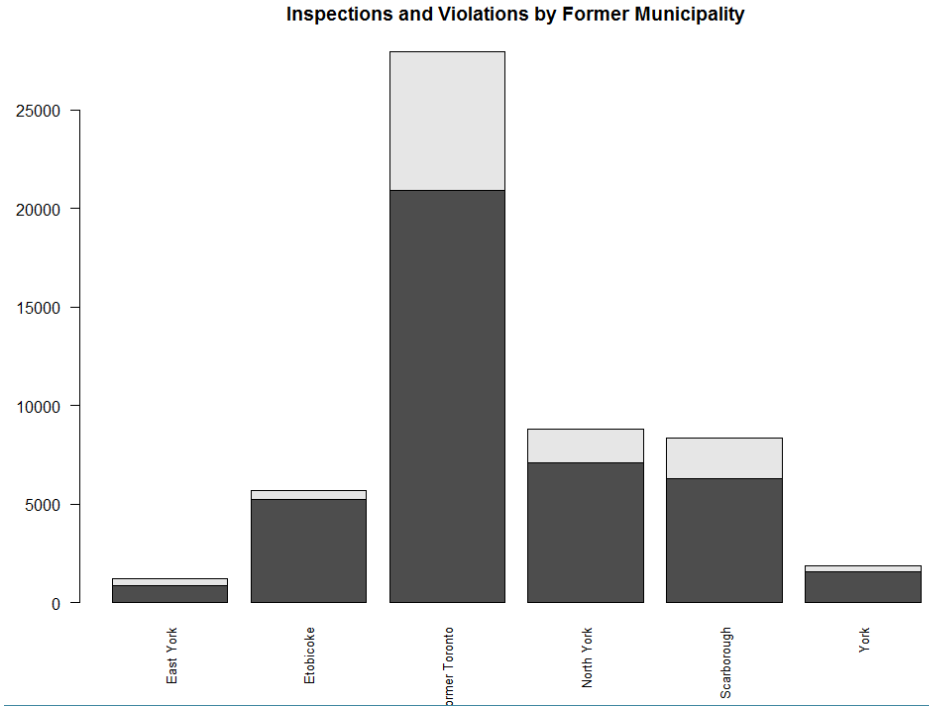
```
##
##                          Restaurant                         Food Take Out
##                               27816                                 8369
## Food Store (Convenience / Variety)                       Food Court Vendor
##                               3360                                 2080
##                          Supermarket                   Child Care - Catered
##                               1818                                 1811
##       Child Care - Food Preparation                              Bakery
##                               1596                                 1448
##                         Butcher Shop                  Food Processing Plant
##                                620                                  591
```

Top 20 Chains

```
##
##          TIM HORTONS              SUBWAY          PIZZA PIZZA
##                  939                 859                 370
##           MCDONALD'S           PIZZA NOVA    STARBUCKS COFFEE
##                  301                 181                 178
##          TIM HORTON'S           SECOND CUP           STARBUCKS
##                  173                 152                 137
##        BOOSTER JUICE                 KFC               METRO
##                  124                 119                 115
##         SWISS CHALET             FRESHII      DOMINO'S PIZZA
##                  114                 110                 108
## AROMA ESPRESSO BAR            MR. SUB        THAI EXPRESS
##                  106                 102                 101
##           PIZZA HUT SHOPPERS DRUG MART
##                   99                  95
```

Correlations between attributes and violations





Logistic regressions

- After data exploration, created a preliminary list of variables
- Focus areas were:
  i. Establishment type
  ii. Region of Toronto
  iii. Inspection Frequency
- Created dummy variables for categories, used GLMULTI to choose best fitted model

```
##
## Call:
## glm(formula = Sev_Cru ~ 1 + TYPE_RESTAURANT + TYPE_INSTITUTION +
##     TYPE_TAKEOUT + TYPE_FOODCOURT + MUN_FMR_TORONTO + MUN_SCARBOROUGH +
##     MUN_ETOBICOKE + INSP_2 + INSP_3, family = binomial("logit"),
##     data = inspect_work1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0548  -0.7055  -0.6109  -0.3271   2.7168
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.33474    0.04664 -50.056  < 2e-16 ***
## TYPE_RESTAURANTTRUE  0.42226    0.02896  14.580  < 2e-16 ***
## TYPE_INSTITUTIONTRUE -0.30690    0.06117  -5.018 5.23e-07 ***
## TYPE_TAKEOUTTRUE     0.29347    0.03721   7.887 3.09e-15 ***
## TYPE_FOODCOURTTRUE   0.44012    0.05727   7.685 1.52e-14 ***
## MUN_FMR_TORONTOTRUE  0.10142    0.02811   3.607  0.00031 ***
## MUN_SCARBOROUGHTRUE  0.32027    0.03530   9.072  < 2e-16 ***
## MUN_ETOBICOKETRUE   -1.02359    0.05573 -18.367  < 2e-16 ***
## INSP_2TRUE           0.45706    0.04437  10.302  < 2e-16 ***
## INSP_3TRUE           1.27890    0.04475  28.576  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

If a tree falls in an empty café, does anybody hear?

- Unfortunately, the project did not produce a useful tree
- Possible causes:

i. Lack of Dispersion in the predictor variables

ii. Relatively small percentage of violations

- Remedies to be considered:

i. Using a 'biased sample' technique

ii. Testing additional variables from different sources

```
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
```

Questions?

Thank you.