Hidden Places: An Analysis of DC Restaurant Violations and Probable Indicators

Date: 15 May 2017

Authors: Vamsi Varma Kunaparaju; Nuria McGrath; Anya Mityushina; Rachel Sawyer

Context Food establishments that handle food are subject to food and health inspections conducted by the government. These inspections help preserve food safety. Critical violations, if left uncorrected, are more likely than any other inspection violation to directly contribute to food contamination, illness, or environmental health hazards.

Objective To determine which attributes associated with food and health inspections and the inspected restaurants/food establishments are good indicators of critical health violation occurrences.

Hypothesis A restaurant's Yelp rating and the type of inspection conducted on the restaurant are strong indicators for the number of critical health violations.

Data District of Columbia (DC) Department of Health food and health safety inspection data was used for this assessment. Restaurant data collected from Yelp as well as zip code based-data were also used.

Models The following models were used: 1) linear regression model (LM); 2) generalized linear model (GLM); 3) random forest model; and 4) gradient boosting model. These models were constructed and run using the program R.

Main Outcome Measures Significance and strength of variables were evaluated as follows: p-value (for LM and GLM); mean decrease in accuracy and mean decrease in node impurity (random forest model); and variable's gain and chi-square test statistic (gradient boosting model). Model accuracy was based on predicting the correct critical violation number category.

Results All the assessment independent variables used were considered significant. The strongest variables varied depending on the model, though inspection type and review count on Yelp were strong indicators for all models. Model accuracy ranged from approximately 46% (gradient boosting) to approximately 86 % (random forest model).

Conclusions Inspection type is a strong indicator for predicting critical health violation occurrences. Yelp data are strong indicators; however, Yelp rating is not the strongest Yelp indicator. Zip code associated data does not seem to be influential.

Preserving food safety is an important aspect of the government. The department of health requires that food safety inspections are conducted. Inspections ensure that meat and poultry products are safe, wholesome, and correctly labeled and packaged.[1] Almost any establishment that handles food is subject to a food inspection; from school cafeterias and restaurants to caterers and the beloved mobile vendors. The frequency of an inspection is largely based on the risk level of an establishment. Critical violations, if left uncorrected, are more likely than other violations to directly contribute to food contamination, illness, or environmental health hazard, i.e., food temperature.[2]

OBJECTIVE

The objective of this assessment is to determine which attributes associated with food and health inspections and the inspected restaurants/food establishments are good indicators of critical health violation occurrences.

For this assessment, both inspection attributes (e.g., type, or month) as well as restaurant attributes (e.g., location, or food category) were considered.

It is hypothesized that the restaurant's Yelp rating coupled with the inspection type conducted on a restaurant will be strong indicators of the number of critical health violation occurrences.

METHODS

Data Sources

Open data resources were utilized for this assessment. A data warehouse was created from five datasets. The datasets were matched by a restaurant's Yelp name (i.e., the Yelp ID) and by zip code. The datasets and sources are:

1. Restaurant Inspection Data: DC Department of Health Online Food Facilities Inspection
2. Property and Violent Crime Zip code Data: Urban Institute

1https://www.foodsafety.gov/keep/government/inspections/

2https://doh.dc.gov/service/understanding-food-establishment-inspections

3. Income: IRS Zip code Data
4. Restaurant Yelp Data: Yelp
5. Median Home Value: Zillow Data.

DC Restaurant Dataset Statistics

The DC Restaurant Inspection Data includes inspections conducted by the DC Department of Health between 2010 and March 2015. This dataset has 25,109 inspections conducted on 2,514 food establishments. See Figure 1 for the locations of the inspections. The sample mean value of critical health violations is 1.9. The sample median is 1. Approximately 55% of these inspections resulted in 2 or more critical violations. A total of 46,460 critical violations occurred of which approximately 90% resulted from inspections with 2 or more critical violations.

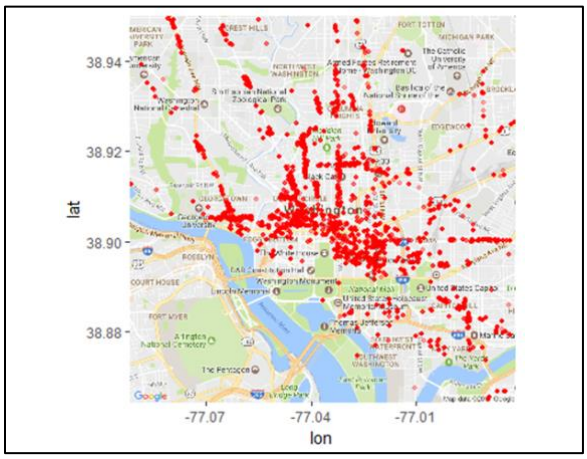

Figure 1: DC Restaurant Inspection Map

Variable Definitions

Dependent/Response Variable

The dependent or response variable is the number of critical violations. For the GLM, random forest and gradient boosting models, a binary variable is used. The binary variable is equal to 1 if the number of violations is 2 or greater, or 0 if there are less than 2 violations. The number of 2 was used to divide the violation categories based on the sample mean and sample median values, 1.9 and 1, respectively. The linear regression model predicts the number of critical violations and then sorts the value into the category it fits in.

Independent Variables

The independent variables from the Restaurant Inspection Data included:

a. Zip code: the restaurant's zip code. Derived from restaurant's address.
b. Latitude and Longitude: the restaurant's latitude and longitude coordinates.
c. Inspection Type: categorical variable with the following classes: Routine, Follow-up, HACCP, Complaint, Pre-operational, License, and Other.
d. Month: the month that the inspection was conducted during. Derived from inspection date.
e. Day of the week: the day of the week that the inspection was conducted on. Derived from inspection date.
f. AM or PM: whether inspection conducted in morning or afternoon. Derived from inspection time.

The independent variables from the Crime Data included:

g. Rate of Violent Crime per 1000 people: value for entire year of 2016 for a given zip code.
h. Rate of Property Crime per 1000 people: value for entire year of 2016 for a given zip code.

The independent variables from the IRS Data included:

i. 2014 Zip code population proportion for the following salary groups (in $1000): under 25; 25 to 50; 50 to 75; 75 to 100; 100 to 200; and over 200.
j. Mean Income for zip code in 2014. Equal to the total income divided by the number of returns.

The independent variables from the Yelp Data included:

k. Restaurant Category: categorical variable for the type of food/ establishment. E.g., pizza; café; or Asian.
l. Yelp Rating: a restaurant's average rating on a scale from 1 to 5.
m. Review Count on Yelp: the number of reviews posted for restaurant on Yelp.
n. Price: a categorical variable based on price per person range. Levels (in USD) are: $ (< 10); $$ (11-30); $$$ (31-60), and $$$$ (over 61).

The independent variable from Zillow included:

o. Median value per square foot for zip code in2016 (residential property)

Data Preparation/Cleaning

Due to the large number of observations, it was deemed acceptable to discard all observations that had any empty, NULL or NA values. This reduced dataset size from 25,109 to 20,396.

Principal Component Analysis (PCA)

Thirteen of the independent variables were converted from type list to numeric, normalized, and used in the PCA. The variables include: Zipcode, PermitID,

Latitude and Longitude, review_count, rating, price, InspectionID, Inspection Type ID, Violent and Property Crime, House and finally, Income. The scree plot shows that approximately 97% of the variance in the model can be explained by ten of the thirteen components. See Figure 2.

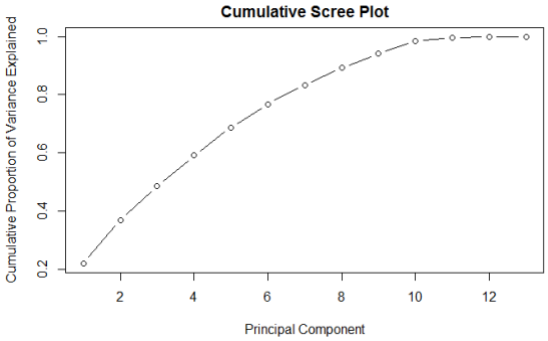

Figure 2: Cumulative PCA Scree Plot

Outcome Measures

Significance and strength of variables were evaluated as follows: p-values, model AIC, and prediction error (for LM and GLM); mean decrease in accuracy and mean decrease in node impurity (random forest model); and variable's gain and chi-square test statistic (gradient boosting model). Model accuracy was based on predicting the correct critical violation number category.

Model 1: Linear Regression

Linear regression is a statistical procedure used to predict the value of a dependent/response variable using the independent variables such that the dependent variable is equal to a linear combination of the independent variables. This model assumes that the data is normally distributed.

Data Preparation: To design a linear model, both the independent variable and the response variable must be continuous. Therefore, the variables that had multiple levels of categorical values were converted into relative continuous values (e.g. Best=3, good=2, fair=1).Also, the dataset was normalized. The dataset was randomly split to develop the training (60%) and validation (40%) sets.

Certain variables, e.g., violent crime, were not included in the model because of collinearity. The classification variable restaurant category could not be included also.

Tools: The tool R programming was used to develop/run this model.

Model Tuning: Several linear regression models were constructed and compared including: a full model (all continuous variables available); a forward model; a backward model; and a reduced model (a model with just the statistically significant variables). The model that performed the best—based on the goodness of fit (R^2 value) and root mean squared error (RMSE)—was the reduced model (R^2 = 0.638, RMSE = 2.275).

By observing the p-values, "price", "review count", and the "inspection type" seem to be influential predictors.

Model

There are seven predictors in the final model: month; day of the week; Latitude; Longitude; review count; price; inspection type code; property crime; and median house value..

Model 2: Generalized Linear Model, Binomial

The Generalized Linear Model (GLM) was used to transform the linear model into a link function, in this case binomial. Like the linear model, it is used to determine the relationship between two or more variables.

Data Preparation: In order to create the GLM model, inverse logit function transformed the continuous output from the linear predictor 'Critical Violations' into a binomial indicator with greater than or equal the average of 2 violations as '1'. Factor variables were created for 'price' and 'Critical Violation ID'.

Tools: The tool R programming was used to develop/run this model.

Model Tuning: Three models were created:

1.) The full model used the factor variables, and the variables that were not collinear. The resulting coefficient plot was created with a confidence interval of 95% on the outer line and 68% on the inner line. See Figure 3.
2.) The reduced model was based on the coefficients from the full model but using only those variables which had a p-value less than 0.1. There were NAs for the coefficients of the factor variables--this may have been because there were no observations with the specified combination of levels of the factors.
3.) The backwards model had an almost identical AIC value and prediction error as the full model.

Cross validation was conducted on the full model, which was the best model.

Model

There are seven predictors in the final model: inspection type; income; price; rating; house median value; crime property; and review count.

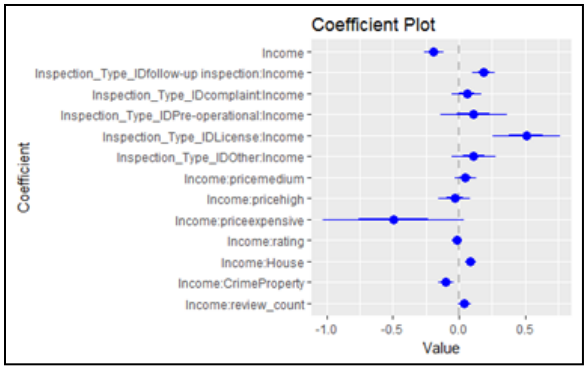

Figure 3: Coefficient Plot for GLM Full Model

Model 3: Random Forest

The random forest model, can be used for both classification and regression problems. A random forest model builds decision trees such that at each split in a tree, a random sample of m predictors is chosen as split candidates from the full set of p predictors—the split will only use one of those m predictors.

Data Preparation: All variables were set as either continuous or factors. As the randomForest model can only handle factors with 32 levels, the restaurant category variable had to be re-binned. The data was randomly split into a training (60%) and a test (40%) set.

Tools: The tool R programming was used to develop/run this model. The randomForest and Caret packages were used.

Model Tuning. A full model was run and variable importance values (pure node impurity) were obtained. The bottom 10 predictors were dropped: the resulting model accuracy was not affected. Then, a 10-fold cross-validation was conducted to determine the optimal combination of predictors based on minimized RMSE. The resulting predictors are the final model variables. See Figure 4 for the cross-validation plot.

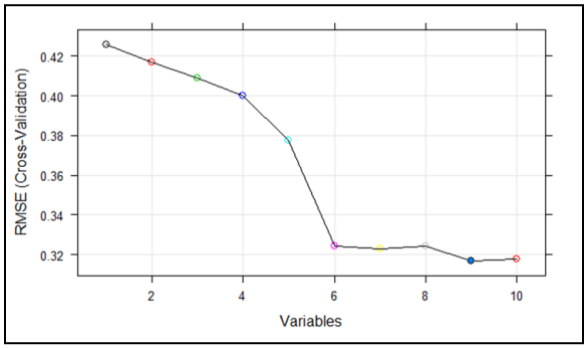

Figure 4: Cross-Validation Plot for Random Forest Model

Model

There are nine predictors in the final model (in decreasing order): inspection type; restaurant category; month; day of the week; latitude; review count; longitude, rating, and AM PM.

Model 4: Gradient Boosting

Gradient boosting is a machine learning technique that is suitable for both regression and classification problems. The model arrives at a prediction by building an ensemble of weak prediction modes. After the shallow trees are built, the model then generalizes and optimizes the results (by minimizing MSE).

Data Preparation: First, the data is subset to include only the independent variables. Second, the matrix is transformed into a sparse matrix. The sparse matrix converts an original data matrix into a binary matrix that is specific to each type of predictor. The data must then be randomly split into a training (60%) and a test (40%) set.

Tools: The tool R programming was used to develop/run this model. The xgboost and caret packages were used.

Model Tuning: The xgboost package for gradient boosting provides a model with various tuning opportunities. To prevent overfitting, model performance was monitored for "outstanding" performance.

The features that provided the most gain during the splits are identified in Figure 5. These were the predictors used for the final model.

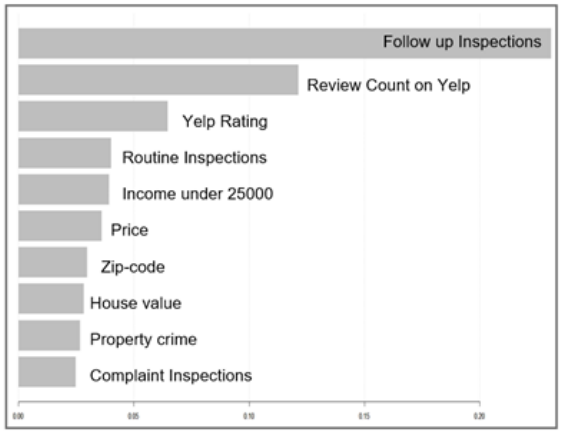

Figure 5: Variable Gain for Gradient Boosting Model

Model

There are eight predictors in the final model (in decreasing order): inspection type; review count; rating; income under 25,000; price; zip code; median house value; and property crime.

RESULTS

Model 1: Linear Regression

Figure 6 shows the actual number of critical violations against the predicted number of violations. It seems that the model consistently underestimates the response variable.

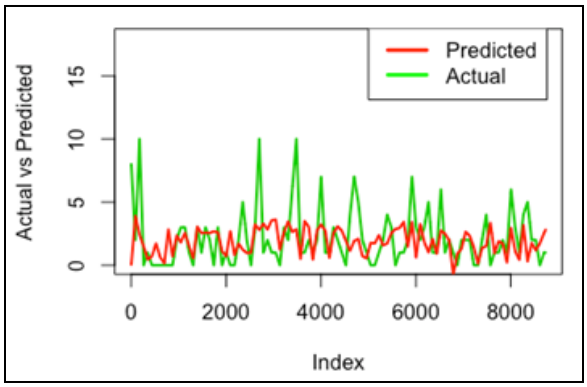

Figure 6: LM Actual vs. Predicted

Residual Analysis: Based on the trend observed in the residual plot, it seems that the response variable and independent variables are not linearly related. (See Figure 7.) Therefore, the linear model is neither the best model nor an appropriate model to be used to predict the number of critical violations.

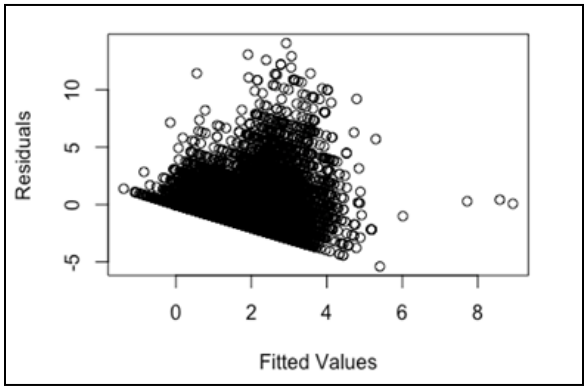

Figure 7: Residuals vs. Predicted Values

Confusion Matrix and Accuracy: The predicted values were classified into the two categories >= 2 and < 2. See Table 1 for the confusion matrix. The resulting model accuracy is 63.2%.

Table 1: LM Confusion Matrix

		Actual	
		≥ 2	< 2
Predicted	≥ 2	1370	618
	< 2	2061	3240

Model 2: Generalized Linear Model, Binomial

See Figure 8 for the Receiver Operating Characteristic (ROC) curve for the GLM model.

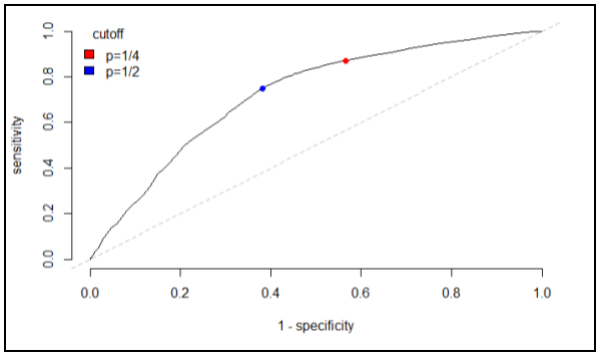

Figure 8: ROC Plot GLM

Confusion Matrix and Accuracy: See Table 2 for the confusion matrix. The resulting model accuracy is 63.2%.

Table 2: GLM Confusion Matrix

		Actual	
		≥ 2	< 2
Predicted	≥ 2	1745	685
	< 2	2527	3777

Model 3: Random Forest

Figure 9 shows the ROC curve for the random forest model.

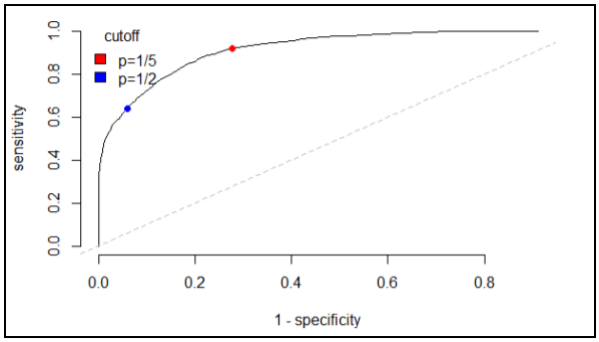

Figure 9: ROC Plot for Random Forest Model

Confusion Matrix and Accuracy: See Table 3 for the confusion matrix. The resulting model accuracy is 85.6%

Table 3: Random Forest Confusion Matrix

		Actual	
		≥ 2	< 2
Predicted	≥ 2	1464	826
	< 2	351	5518

Model 4: Gradient Boosting

Figure 10 shows the ROC curve for the gradient boosting model.

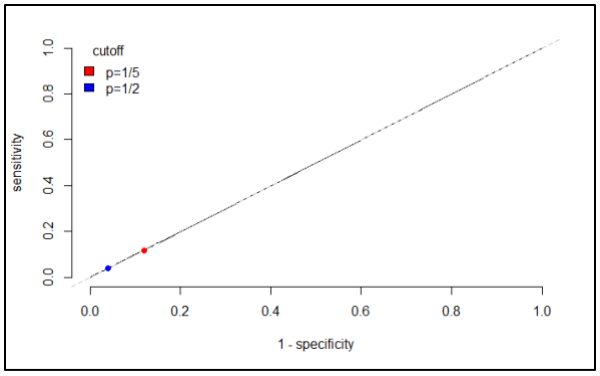

Figure 10: ROC Plot for Gradient Boosting Model

Confusion Matrix and Accuracy: See Table 4 for the confusion matrix. The resulting model accuracy is 48.7%. (This is in stark contrast to the 5.6% error for the training data. This model was clearly overfit and is very sensitive to the variable values.)

Table 4: Gradient Boosting Confusion Matrix

		Actual	
		≥ 2	< 2
Predicted	≥ 2	1950	109
	< 2	340	1841

Model Comparison

All the assessment independent variables used were considered significant. The strongest variables varied depending on the model, though inspection type and review count on Yelp were strong indicators for all models.

Model accuracy ranged from approximately 49% (gradient boosting) to approximately 86% (random forest model). By comparing the ROC curves for the GLM, random forest, and gradient boosting models, it is clear that random forest is the most appropriate model when considering both sensitivity and specificity.

The gradient boosting model is accurate at selecting variables for model predictors. Though, with binary predictions, it may be useful to utilize a logic regression as the final model. This gives you the benefit of gradient boosting for feature reduction but, the consistency of a logic regression. This also helps take out the "black box" effect from your model.

CONCLUSIONS

All four models supported the hypothesis that inspection type is a strong indicator for critical health violations that result from food safety and health inspections. Yelp data variables are strong indicators; however, Yelp rating is probably not the strongest Yelp indicator. Also, the review count on Yelp is a stronger indicator than expected. Zip code associated data does not seem to be influential.

REFERENCES

1. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
2. Department of Health (DoH), Food Safety and Hygiene Inspection Services Divions (FSHISD). https://doh.dc.gov. (2017)
3. IRS. Individual Income Tax Zip Code Data, Tax Year 2014. https://www.irs.gov. (2017)
4. MacDonald, Graham. Open Data DC, Restaurant Inspection Data. (2016) http://data.codefordc.org.
5. NeighborhoodInfo DC. Property and Violent Crime Data, Zip code Table. https://www.neighborhoodinfodc.org/ (2017)
6. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
7. Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich and Yuan Tang (2017). xgboost: Extreme Gradient Boosting. R package version 0.6-4. https://CRAN.R-project.org/ package=xgboost
8. Yelp, Inc. Yelp Fusion Documentation. www.yelp.com/developers/documentation/v3/. (2017).
9. Zillow, Inc. Median Home Value Index (ZHVI). www.zillow.com/research/data. (2017)