Forecasting restaurants with critical violations in Chicago

City of Chicago

February 04, 2015

```
## Loaded glmnet 1.9-8
```

The Chicago Department of Public Health (CDPH) inspects more than 15,000 restaurants with fewer than three dozen inspectors over the course of the year. This paper develops a predictive model design to identify the likelihood any restaurant that contains critical violations. Since CDPH is obligated to inspect every food establishment, the goal of the model is to identify the riskiest restaurants earlier, thereby reducing the length of exposure of risky restaurants to patrons. In testing, the predictive model was able to identify 23 percent more violations than current operations.

1 Introduction

The Chicago Department of Public Health inspects more than 15,000 restaurants with fewer than three dozen inspectors. City of Chicago ordinance requires that most of these establishments must be inspected at least once a year. The task to inspect each restaurant, while also performing other inspections, is completed over the year.

CDPH conducts three different types of inspections. First, CDPH conducts license inspections for any new establishment with a food license prior to the establishment opening. Each establishment must pass this initial inspection before it is allowed to serve food to patrons. Second, CDPH will conduct canvas inspections; periodic inspections to check the quality of sanitary conditions. The number and frequency of inspections is driven by the type of facility and how it prepares food, inspecting the riskiest restaurants at least two times a year.

License inspections are coordinated with Business Affairs and Consumer Protection (BACP), who grants food establishment licenses to new establishments. The quantity and location of these inspections is driven by license applications, thereby, dependent on the economy, entrepreneurship, and other outside factors beyond the control of the City. These inspections can be characterized as routine, but not a guarantee to pass. Establishments fail these initial inspections because they have not yet finished setting-up equipment, such as turning-on a refrigerator, or have not finished construction. Under these circumstances, CDPH will re-inspect those establishments to ensure those conditions are passed before they are allowed to open.

Complaints are registered from residents, alderman, and referrals from hospitals. Often, these requests are driven through the City of Chicago's 311 system, which can be submitted through residents calling 311 or submitting a request through an online form. Individuals are asked to submit where they believe they contracted food poisoning, the address of the establishment, describe the symptoms and what was eaten, and when it happened. CDPH reviews the materials and may initiate a food inpsection if it does seem the illness and restaurant can be linked together.

Uniquely, CDPH also encourages submissions through the Foodborne Chicago program. Machine learning algorithms scans Twitter for individuals complaining or indicating potential food poisioning cases. These tweets are identified and a human will contact the user, providing a link and information on how they can report their complaint to CDPH. In a nine-month span, 133 food inspections were instigated from this program, where 20 percent (27 instances) of those inspections resulted in critical violations (Harris et al. 2014)

The Foodborne program and 311 system has assisted CDPH in targeting and identifying complaint-driven requests. Yet, the department has a sizeable task to complete canvas inspections. Canvas inspections occur throughout the year and are somewhat random inspections of various restaurants. The process is key to checking and enforcing consistent food safety practices throughout the city. Identifying critical issues at restaurants help rectify those issues and reduce exposure to patrons.

The work is organized by CDPH-defined "risk levels", which are divided into three categories: risk 1 (highest), risk 2, and risk 3. The risk levels are determined by food handling practices required for each establishment. Restaurants and other establishments that directly handle ingredients and prepare food, such as needs to cool or heat food, are generally categorized as risk 1. The lowest risk generally consists of prepackages and non-perishable food. Risk level also drives frequency of inspection. Risk 1 facilities are inspected more frequently, with a target of at least twice a year; risk 2 are inspected at least once a year; and risk 3 is inspected once every other year.

Risk levels do help prioritize inspections by focusing on establishments with the highest likelihood of spreading food born illnesses through categorization of food handling practices. However, of those establishments in our dataset are categorized as risk 1. The high proportion of risk 1 establishments means there is still a substantial queue to be inspected. Yet, the work is certainly achievable. Assuming 32 inspectors, each inspector would need to complete canvass inspections each working day—in addition to complaint-driven and new license inspections.

There are 42 different possible violations that can be cited by CDPH. Often, these violations are classified into three categories: critical, serious, and minor violations. Critical violations consist of 9 different violations that are most likely to create conditions for food born illnesses, such as failure to heat food to proper temperatures or to keep items properly fridgerated at the proper temperatures. Conversely, minor violations can be as simple as leaving a rag in the sink. Restaurants can fail their inspections with as little as one critical violation. Several serious and minor violations during an inspection can also lead to a failed outcome.

The clear priority of CDPH food inspection team is to prevent foodborne illnesses, which are most likely to stem from critical violations. With a XX risk 1 establishments to inspect throughout the year, the question also extends to which establishments should be inspected first. Fortunately, the city dollects data on multiple facets of food inspections, including the outcomes of inspections, information about businesses and their activities, and events around each food establishment, such as 311 complaints, crime, and even weather. Section 2 describes data that has been collected by the research team for this project.

Section 3 describes how these data sources can be combined to show the relationship between the characteristics of a food seller with critical violations. We formulates a model to derive the likelihood for each establishment to have critical violations. This model can be used to prioritize which restaurants should be inspected first. By targeting the highest-probability restaurants, CDPH can minimize the amount of exposure restaurant patrons have to the unsanitary conditions that are most likely to lead to food born illnesses. This paper will focus on prioritizing inspections of risk 1 and risk 2 restaurants, since these must be inspected at least once a year.

After developing the statistical model to predict critical violations, the research team evaluates whether the model could optimize food inspection processes (section 4). Namely, the model is used to determine how much faster the food inspection team can discover critical violations. The team uses a simulation to compare real-life results to an alternate, data-driven arrangement.

Section 5 summarizes the results of the paper. Ultimately, we find that XX percent increase in finding critical violations at food sellers. Beginning in 2015, CDPH will begin to use this analytical model to prioritize canvas inspections. Each risk 1 and risk 2 restaurant will still undergo inspections; however, these restaurants with the highest likelihood of the most serious issues will be prioritized.

Finally, it is worth nothing that this research is an open source project. The source code of the statistical model is available on the City of Chicago food inspection project page. The statistical modeling was completed using the R statistical software, with all the necessary data available online. This paper was generated using knitr, which allows others to view the underlying calculations to generate the summaries, tables, and diagrams in this document. The document is available in the same aforementioned repository.

2 Data

The City of Chicago publishes over 600 datasets on the open data portal, including the results of food inspections from 2011 to present. The food inspection dataset includes the name of the establishment, address, risk level, inspection date, results, and a detailed list of violations found during the inspection.

At this time, the food inspection database is not hosted by the City of Chicago, instead, a file of all food inspections are sent to the City of Chicago open data team on a daily basis, which is automatically uploaded to the portal every morning. Thus, the rawest form of data available to the research team was the data available on the open data portal.

In addition, the City also publishes other relevant data on the portal, including: business licenses from 2011 to present, detailed crime data from 2011 to present, and various 311 data, including garbage and sanitation complaints.

Food inspection history is combined with business license data published by BACP. Any food seller must be licensed by the City, but must also obtain licenses for other activities, such as cigarette sales and liqour licenses. The license data provides other information about the business, including when cetain licenses were first obtained–an approximation for the age of business.

The location of the businesses are used to calculate nearby activity. Several variables were explored, but after conduting some data mining, we settled on burglaries, sanitation code complaints, and garbage cart requests. The density of each activity was calculated and stored.

Weather data was obtained from forecast.io. The data contains a significant wealth of information on not only highs, lows, and percipitation, but also on granular weather forecasts for any latitude and longitude. After several conversations with CDPH staff, we focused on the relationship of temperatures and inspections. High temperatures can lead to issues to cooling food within a food establishment, which results in a critical violation. Some empirical testing of that hypothesis helped support its inclusion.

[Table of variables]

3 Model Development

The principle question is whether we can reasonably determine the probability that a restaurant inspection will yield at least one critical violation. That is, the focus will be whether or not any critical violation is found—a binary response. We use a glmnet model to estimate the impact of

While the following form can be expressed a number of ways, the logistic form is commonly expressed as the "log-odds transformation".

$$\log = \frac{\Pr(V = 1|X = x)}{\Pr(V = 0|X = x)} = \beta_0 + \beta^T x$$

Thus, the objective function is to minimize

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} - \left[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1 + e^{(\beta_0 + x_i^T\beta)})\right] + \lambda\big[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\big]$$

A review of the methods to find this solution is provided by Friedman, Trevor, and Tibshirani (2010), whose glmnet library for R was used to provide estimates.

3.1 Significant Variables

The regression analysis shows the inspectors have a significant impact of the rate of finding critical violations. To provide anonymity, inspectors were grouped by similar performance estimated by individual regression coefficients.

Whether the restaurant had a critical violation in the past was a positive perdictor of future critical violations. Likewise, historical serious violations was also a predictor of future performance. In effect, past performance perdicted future outcomes, with those with critical violations more likely to repeat those violations than even those with, at most, serious violations.

The elapsed time since the last violation also increased the likelihood that inspectors found critical violations. However, restaurants probability of having critical violations fell over the lifespan of the restaurant. As restaurants grow older, they are less likely to have critical violations while long time-periods between inspections increased the likelihood.

Characteristics besides the restaurant itself were also indicators of future performance. Trends in weather, nearby reports of burglary, and complaints about sanitation and garbage seems to explain increases. An increase in the moving three-day average high temperature was associated with more critical violations. In conversations with inspection managers, researchers understood this to be associated with potential mechanical failures—driven by the heat—of equipment that maintained food temperature, a main source of critical violations.

Sanitation code complaints are one of the top complaints registered with the City of Chicago through its 311 system (including web and text reports). Sanitation code complaints include:

4 Evaluation

After formulating the analytical model, the the principal question for researchers turned to whether this analytical model provides more efficiency for the food inspection team. CDPH operational procedures requires the department to inspect every risk 1 and risk 2 restaurant. Therefore, the operational goal is to allow inspectors to discover critical violations earlier than their current operations (business-as-usual).

One approach for an evaluation may have also sought to determine if the predictive model could discover more restaurants with critical violations. Since CDPH is required to inspect every risk 1 and risk 2 restaurant, discovering more restaurants is not a pertinent goal. Instead, it serves a greater public interest to discover violations sooner, thereby, reducing the potential exposure of conditions that breed foodborne illnesses to the public.

4.1 Evaluation Design

The analytical model was trained on data from January 2011 through January 2014, which results were described in the previous sections.

[INSERT FIT CODE]

The researchers waited until CDPH completed food inspections in September and October 2014. This timeframe ensured significant time passed between the test period (January 2011 through January 2014) and the evaluation period. It's likely any temporal correlation would subside between the test and evaluation period. CDPH was not aware this timeframe would be used for an evaluation in order to prevent against a Hawthorne Effect or other bias. Again, to reduce any potential to bias within reason, senior management at CDPH was aware of on-going research, but sanitarians were not informed of the research. Finally, several months passed between model development and the evaluation period, reducing a perception of the evaluation period.

The evaluation period lasted two months, from to and calculate the percentage of inspections that result in critical violations in the first half of the inspections during this period. The number of violations found during this period can be considered as status quo or current mode of operation. It serves as a baseline to capture performance levels of sanitarians, namely, the proportion and rate of critical violations that are found.

Meanwhile, we calculate the point predictions for each establishment using the training data from 2011 through 2014. The training data does not include the evaluation period so not to provide additional feedback from the evaluation period. We sort the establishments that were inspected during the evaluation in descending order of predicted values, placing the highest risk restaurants at the top of the list.

We calculate the percentage of those restaurants that would be inspected in the first half if the predictive model was used. The difference between the percentage of establishments found with critical violations during this period reflects the relative gain or loss of efficiency. Finding a greater percentage of critical violations with the predictive model indicates results can be found earlier. A similiar or reduced amount indicates the predictive model provides no benefit or is less efficient, respectively.

Note that this experimental design is assumed to yield the name number of restaurants found with critical violations. Indeed, under the premise that CDPH will inspect all restaurants, researchers will presume the number of violations will remain relatively the same. The objective of the model is to find critical violations earlier throughout the year.

4.2 Results

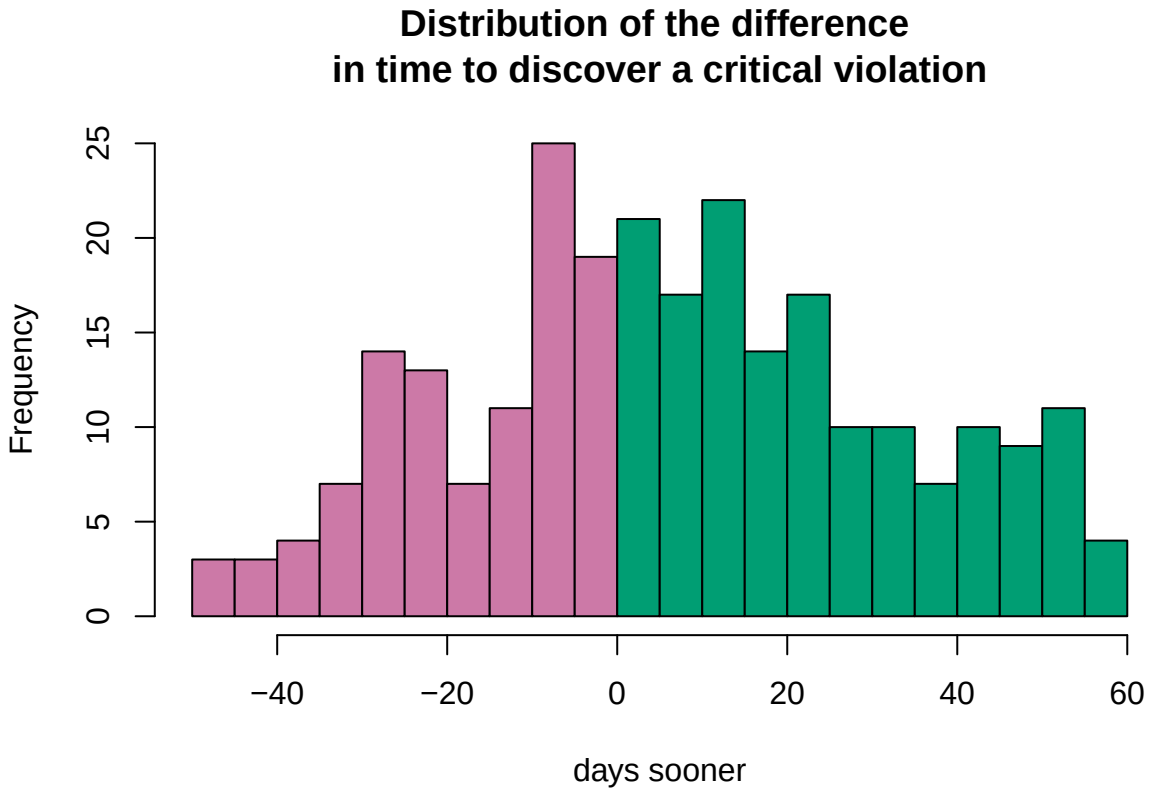
CDPH completed inspections between and . During this time, CDPH found violations, ““percent of all inspections. The rate of violations is consistent with the historical average of approximately 15 percent. While the rate of violations is slightly higher, it is close enough where we do not suspect this period is abnormal, thus, a valid comparison for our evaluation.

[Insert bar graph comparing BAU and model in first half]

On average, food establishments were identified 7.4379845 days earlier under a data-driven model. Generally, critical violations would have been found sooner under the data-driven regime. The rate of finding critical violations in the first half of the would have increased by 25 percent under the data-driven model. [Fifty-five percent] of critical food violations were found in the first half; meanwhile, under the data-driven model, [sixty-nine percent] of all of the critical violations.

While the average gain was 7 days, there was a significant range in the change. Some restaurants were identified 58 days earlier than business-as-usual. Half of the crticial violations were identified over 6 earlier while quarter of all violations were prioritized over 25 sooner. Yet, some restaurants would be prioritized lower, 99 restaurants were incorrectly prioritized lower and were found to have critical violations later–0.3837209 of the observed critical violations.

```
hist(as.numeric(time_diff),
     main = paste0("Distribution of the difference \n",
                   "in time to discover a critical violation"),
     breaks = 30,
     xlab = "days sooner",
     col = c(rep("#CC79A7", 10), rep("#009E73", 20)))
```



We conducted a t-test to measure whether the reduction in time to find a critical violation was greater than zero. Namely, the null hypothesis is the average time each food inspection was accelerated is equal to zero. The test (

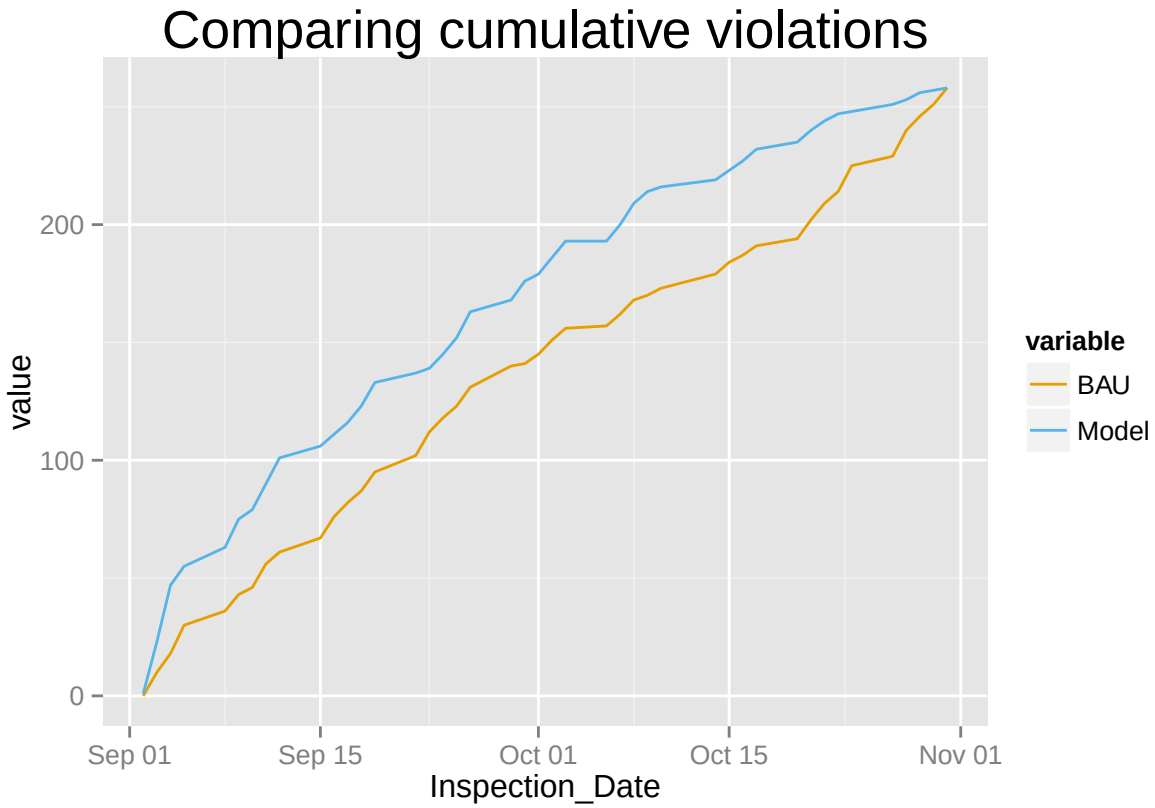
$$\sigma =$$

`25.2046494`, df = `257`) resulted in a p-value of `4.7400668`, which indicates

Below, Gini curves show the relative difference in the inspection regimes throughout the pilot. Since the first day, the data-driven model revealed more critical violations. Specifically, 111 more violations were found in the first week between September 2 and September 5 (141 under data-driven compared to 30 for business-as-usual). The cumulative number of violations found were always higher for the data-driven approach until the final day of the pilot.

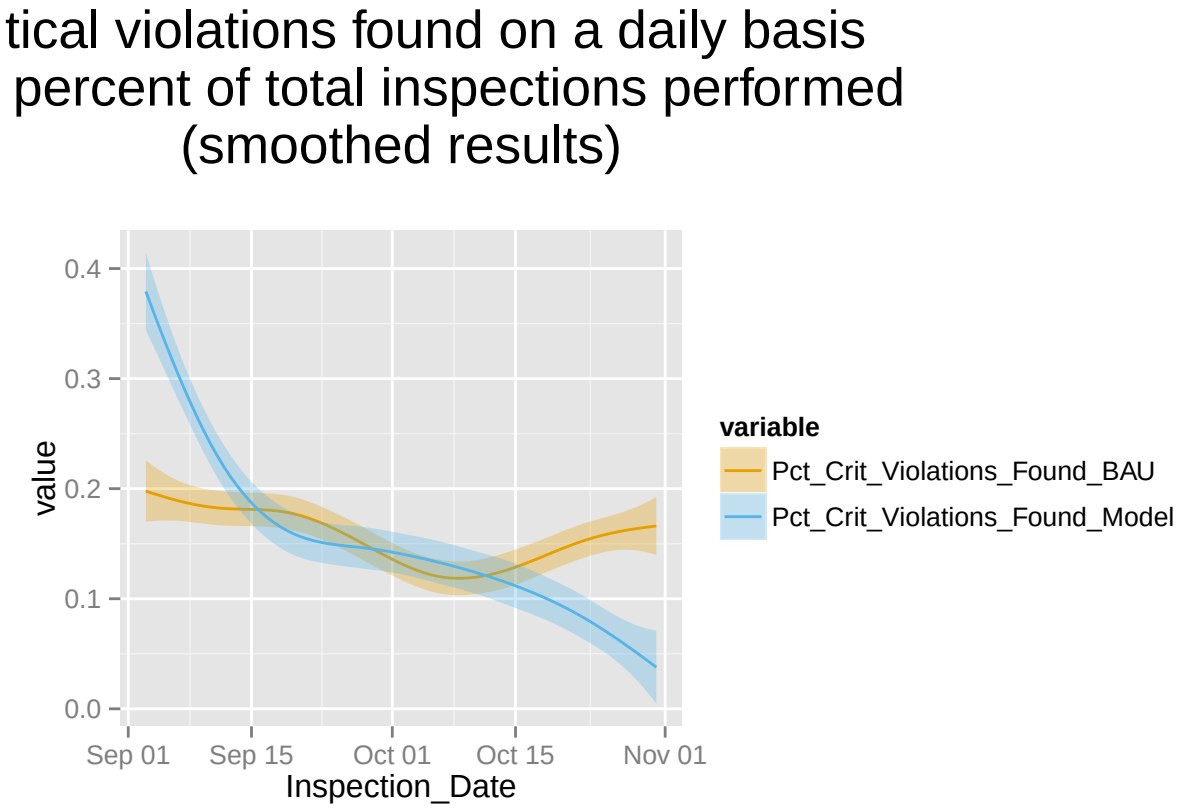
```
comp_summary_cumsum <- comp_summary[
    i = TRUE,
    j = list(Inspection_Date = Inspection_Date,
             BAU = cumsum(Crit_Violations_BAU),
             Model = cumsum(Crit_Violations_Model))]

ggplot(melt(data = comp_summary_cumsum,
            id.vars = "Inspection_Date")) +
    aes(x=Inspection_Date, y=value, colour=variable) +
    labs(title="Comparing cumulative violations") +
    geom_line()
```



By extension, the rate of finding critical violations is higher for the initial quarter of the pilot. The rate of the volations are higher in the first portion as the analytic model correctly ranks higher-risk restaurants for earlier inspection. Business-as-usual has a more consistent rate of discovery, approximately 0.2 per day and stays above 0.1 violations per day. However, whereas the data-driven model is more successful early, the rate of finding violations declines in the last quarter of the pilot.

```
ggplot(
    melt(data = comp_summary[
        i = -1,
        j = list(Pct_Crit_Violations_Found_BAU = Crit_Violations_BAU / Total_Inspections,
                 Pct_Crit_Violations_Found_Model = Crit_Violations_Model / Total_Inspections),
        keyby = Inspection_Date],
            id.vars = "Inspection_Date")) +
    aes(x=Inspection_Date, y=value, colour=variable, fill=variable) +
    labs(title=paste0('Critical violations found on a daily basis\n',
                      'as a percent of total inspections performed\n',
                      "(smoothed results)\n") ) +
    stat_smooth(method = "gam",
                formula = y ~ s(x, k=10, sp=2, bs="ps"),
                alpha=.3,
                level = .65)
```



The retrospective analysis allows us to surmise and compare to a "best case scenario", the most efficient order of restaurants to inspect based on their risk. In this case, we surmise the best case scenario is where every critical violation is found Below, a graph shows the difference between the most efficient path

5 Summary

This model was able to reduce the timeframe to discover critical violations at Chicago's food establishments. Within a two-month window, the average time to find a critical violation was reduced by `7.4379845` days, a statistically significant finding.

At times, we've explicitly assumed that finding violations is time invariant throughout the pilot phase. That is, a food establishment found with a critical violation on day 40 would have also been found to have a violation even if it was inspected earlier. Unfortunately, we do not have a method to test this assumption. However, the relatively short time window of the study helped ensure external factors, such as severe weather, had limited impact on temporal violations.

Additional data can also be used to supplement this model. Restaurant review data, such as the Google Places API or Yelp, could help supplement data on the conditions of a food establishment.