Forecasting restaurants with critical violations in Chicago

Tom Schenk Jr. (City of Chicago), Gene Leynes (City of Chicago), and Aakash Solanki (City of Chicago), Stephen Collins (Allstate Insurance), Gavin Smart (Allstate Insurance), Ben Albright, and David Crippin (Allstate Insurance)

May 15, 2015

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 describes a predictive model designed to identify the presence of a critical violation in a particular food establishment. The goal of this model is to prioritize inspections by likelihood in order to identify the riskiest restaurants earlier, thereby reducing the length of exposure of risky restaurants to patrons. Critical violations were identified approximately 7.44 days earlier over a 60 day period compared to current operations in the out-of-sample test.

1 Introduction

In 2014 the Chicago Department of Public Health inspected performed over 20,000 inspections at nearly 13,000 food establishments across Chicago with fewer than three dozen inspectors. The majority of these food inspections were routine inspections that don't uncover serious problems, but some of these inspections uncovered issues that affect the health and safety of the patrons who visit these establishments. Traditionally, prioritizing these inspections is a largely manual task that relies on a combination of administrative processes and personal expertise.

The model set forth in this paper can help with the prioritization of scheduled, saving time and money as well as making the city's food safer. The model utilizes several data sources and through advanced modeling techniques the model provides additional insight into an establishment's current actual risk based on real-time data.

This paper is organized as follows: Section 1 provides an introduction and background to describe the current process and scope of the problem. Section 2 describes data that has been collected by the research team for this project and how that data was combined. Section 3 describes the model. Section 4 describes the model evaluation. Section 5 contains details of the model results from the experiment. Finally, the Summary section concludes with a brief summary of results and information regarding the ongoing project.

Ultimately, we find that a data-driven model can help inspectors discover critical violations earlier than the current "Business As Usual" (BAU) process. On average, critical violations would have been discovered 7.44 days earlier over the two-month test period. The first half of the experiment yielded 25.0% higher successful inspections. Beginning in 2015, CDPH has begun to use this analytical model to prioritize canvas inspections. Risk 1 and risk 2 Food Establishments will still undergo annual inspections; however, these restaurants with the highest likelihood of the most serious issues will be prioritized.

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 open source statistical software R, and all the necessary data to replicate these results is 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. This document is available in the same aforementioned repository.

CDPH mainly conducts three different types of inspections; regular canvass inspections, new license inspections, and inspections in response to a complaint. Currently regular canvas visits are the only type of inspection included in the model.

Before a food establishment opens their doors CDPH conducts an initial new license inspection. New license inspections are coordinated with Business Affairs and Consumer Protection (BACP), who grants food establishment licenses to new establishments. Each establishment must pass this initial inspection before it is allowed to serve food to patrons. Establishments often 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. New license inspections are not used in the model because we believe that they are not characteristic of normal inspections, and because they occur when a new business applies for a license, and are therefore not schedule.

The majority of the food inspections are regular canvass visits, which must be done on a regular basis to check the quality of sanitary conditions. The frequency of canvass inspections is driven by the risk level of the facility. Risk 1 establishments are ideally inspected two times a year.

If a restaurant fails the inspection, because of violations / citations, the inspector will return at a later date to see if the situation has been remediated. These reinspections have a very high pass rate, and are not included in the model. Only the initial canvass inspection is included.

The third type of inspection occurs when 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 inspection if it does seem the illness and restaurant can be linked together.

A breakdown of the inspection types in 2014:

Inspection_Type	2014 Count
Canvass	12,229
Canvass Re-Inspection	2,284
License	2,418
License Re-Inspection	708
Complaint	1,653
Short Form Complaint	644
Complaint Re-Inspection	689
Other	172

Uniquely, CDPH also encourages submissions through the Foodborne Chicago program. Foodborne's Machine learning algorithms scan Twitter looking for individuals complaining or indicating potential food poisoning 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, a sizeable task to complete regular canvas inspections remains. 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 risk levels are determined by food handling practices required for each establishment. Restaurants and other establishments are generally categorized as risk 1 if they directly handle ingredients, prepare food, or if

they cool or heat food. The lowest risk establishments generally consist of prepackaged and non-perishable food.

The Risk level also drives frequency of inspection. Risk 1 facilities are inspected more frequently, with a target of two annual inspections; risk 2 establishments are inspected at least once a year; and risk 3 establishments are inspected once every other year.

Although Risk levels help prioritize inspections by focusing on higher risk establishments, in 2014 [66%] of the food establishment licenses were categorized as risk 1. The high proportion of risk 1 establishments means there is still a substantial queue to be inspected.

A summary of the risk types appears below:

Risk Category	Count of Licenses
Risk 1 (High)	8,510
Risk 2 (Medium)	3,128
Risk 3 (Low)	1,264
TOTAL	12,902

2 Data

Thanks to a long-standing progressive stance on information technology and data collection, the City of Chicago has world class data in terms of quality, availability, and accessibility. Data availability was a key factor in the ability to build this model. Multiple sources of data and variables were tested and used in the development of the model, including information about previous food inspections, information about business licenses, information about associated business licenses, and events around each food establishment, such as 311 complaints, crime, and weather.

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 maintained and 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 online data portal every morning. Thus, the rawest form of data available to the research team was the same 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. All of the data that was used in the model came from the data portal, except the weather data and sanitarian data.

2.1 Historical data

For modeling purposes the food inspection history formed the basis of our analysis. Each regular canvass inspection was used as an observation in the model. Some filtering rules were applied. We filtered the inspections to only include Retail Food Establishments, which excluded establishments such as schools and hospitals. Their inspection schedules follow a different planning process, and we also believe that these establishments have different risk characteristics that are not generalizable across the entire population. We also excluded inspection records that didn't result in an inspection because the business was closed.

Each historical inspection record contained an "Inspection Key" that made it possible to uniquely identify and group inspections. Also, each record contained a license number which could later be tied back to the original business license.

Several model variables were created based on the food inspection history, including: time since last inspection, if this inspection was their first canvass inspection, and (not *currently* used) the facility type.

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 14 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 refrigerated 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, however serious and minor violations during an inspection can also culminate to a failed outcome.

We included the results of the previous inspection in the model, which was one of the most important variables. We used previous serious and critical violations in the model, looking back one period, but we did not use minor violations as an indicator.

Food inspection history is combined with business license data published by BACP. Any food seller must not only be licensed by the City but also obtain licenses for other activities, such as cigarette sales and liquor licenses before they can start selling such items. The license data provides other information about the business, including when certain licenses were first obtained–an approximation for the age of business which allows for calculating age of the food establishment at inspection.

Food inspection and Business license datasets are matched based on business names and business addresses. License description information found in the BACP data is also added to the analysis-ready data set. It lists whether a food establishment is a retail food establishment or a tobaco retail over the counter. The BACP data also consists of information whether the license is active or not, when was the application for obtaining a license to do business was filed among other information.

The BACP dataset is important in the sense that in combination with the Food Inspection dataset it curates a full list of all restaurants in the city on which predictions can be made.

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.

3 Model Development

The final variables used in the model are displayed below:

Variable Name (Literal)	Variable Description
`Inspectorblue`	Indicator variable for Sanitarian Cluster 1
`Inspectorbrown`	Indicator variable for Sanitarian Cluster 2
`Inspectorgreen`	Indicator variable for Sanitarian Cluster 3
`Inspectororange`	Indicator variable for Sanitarian Cluster 4
`Inspectorpurple`	Indicator variable for Sanitarian Cluster 5

Variable Name (Literal)	Variable Description
`Inspectoryellow`	Indicator variable for Sanitarian Cluster 6
`pastCritical`	Indicates any previous critical violations (last visit)
`pastSerious`	Indicates any previous serious violations (last visit)
`timeSinceLast`	Elapsed time since previous inspection
`ageAtInspection`	Age of business license at the time of inspection
`consumption_on_premises_incidental_activity`	Presence of a license for consumption / incidental activity
`tobacco_retail_over_counter`	Presence of an additional license for tobacco sales
`temperatureMax`	The daily high temperature on the day of inspection
`heat_burglary`	Local intensity of recent burglaries
`heat_sanitation`	Local intensity of recent sanitation complaints
`heat_garbage`	Local intensity of recent garbage cart requests

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, Hastie, and Tibshirani 2010; Simon et al. 2011), whose glmnet library for R was used to provide estimates. The (Venables and Ripley 2002) was also used in the analysis.

3.1 Significant Variables

Several variables were found to be significant when building the model. Many variables were tested, some were discarded, and as time goes on many may find their way back into the model.

The past performance of food inspections was one of the leading indicators of current likelihood for a critical violation. Both critical and serious violations from the previous inspection were used as a predictor of future performance. In effect, past performance predicted 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 was also a significant variable. The longer that it had been since an inspection the more likely the sanitarian was to find at least one critical violation. However, restaurants' scores decreased 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.

Environmental characteristics were also indicators of future performance. Trends in weather, nearby reports of burglary, and complaints about sanitation and garbage are all significant variables in the model. 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 several types of complaints such as overflowing garbage cans, food left outside, or litter.

The largest source of influence in the model was which sanitarian performed the inspection. We included the effect of the individual sanitarian, but we also anonymized the data to protect the individual identity of the sanitarian. Initial results with individual sanitarian coefficients were used to group the sanitarians into clusters. Those clusters were then arbitrarily assigned color code names. This masking had very little effect on the model performance, but makes it very difficult to tell which sanitarian performed which inspection.

A summary of the model coefficient is presented below:

Variables	Coefficients
Inspectorblue	0.950
Inspectorbrown	-1.306
Inspectorgreen	-0.244
Inspectororange	0.202
Inspectorpurple	1.555
Inspectoryellow	-0.697
pastSerious	0.302
pastCritical	0.427
timeSinceLast	0.097
ageAtInspection	-0.164
consumption_on_premises_incidental_activity	0.411
tobacco_retail_over_counter	0.171
temperatureMax	0.005
heat_burglary	0.002
heat_sanitation	0.002
heat_garbage	-0.004

Table 4: Table of Coefficients

4 Evaluation

After developing the statistical model to predict critical violations, the research team evaluates whether the model could optimize food inspection processes. 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.

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. 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 to reduce any incidental correlation between the two periods. 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 2014-09-02 to 2014-10-31 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.

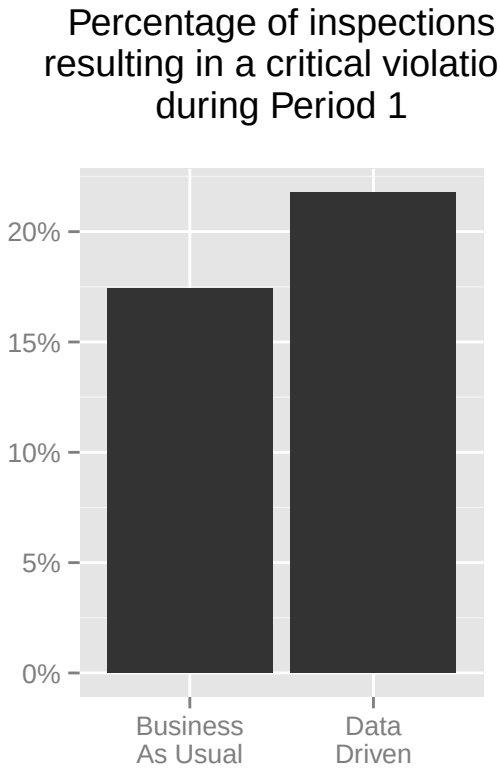
Meanwhile, we calculate the point predictions for each establishment using the training data from 2011 through 2014. The training data does not include the evaluation period so not to provide additional feedback from the evaluation period. We sort the establishments that were inspected during the evaluation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

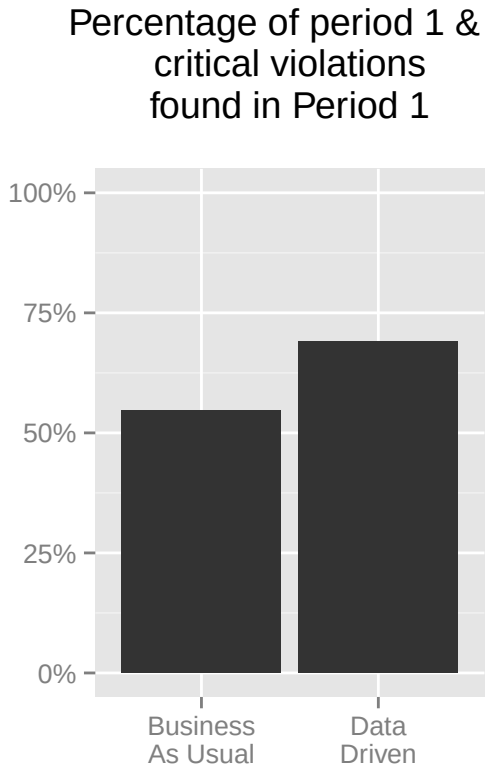
We calculate the percentage of those restaurants that would be inspected in the first half if the predictive model was used. The difference between the percentage of establishments found with critical violations during this period reflects the relative gain or loss of efficiency. Finding a greater percentage of critical violations with the predictive model indicates results can be found earlier. A similar 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.

5 Results

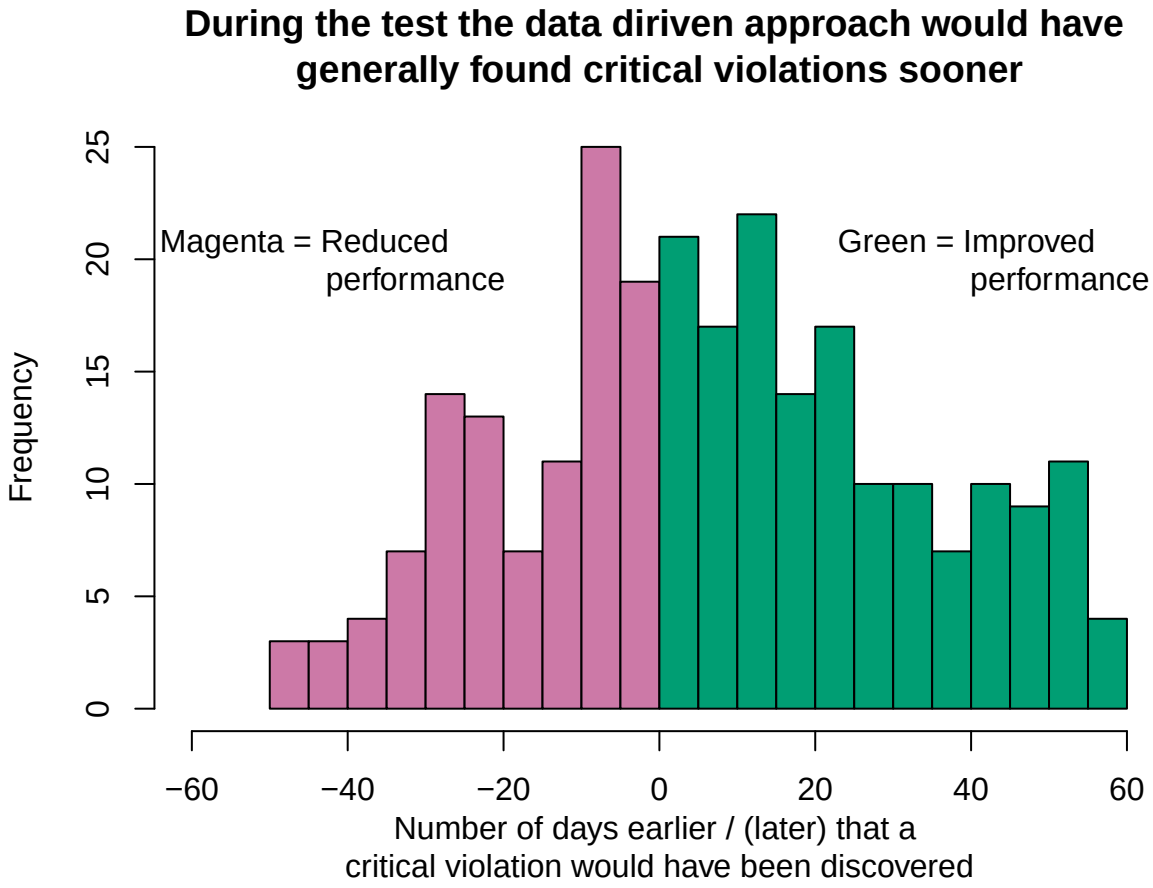
CDPH completed 1,637 inspections between 2014-09-02 and 2014-10-31. During this time, CDPH found 258 violations, 15.8%percent of all inspections. The rate of violations is consistent with the historical average of approximately 15 percent. While the rate of violations is slightly higher, it is close enough where we do not suspect this period is abnormal, thus, a valid comparison for our evaluation.





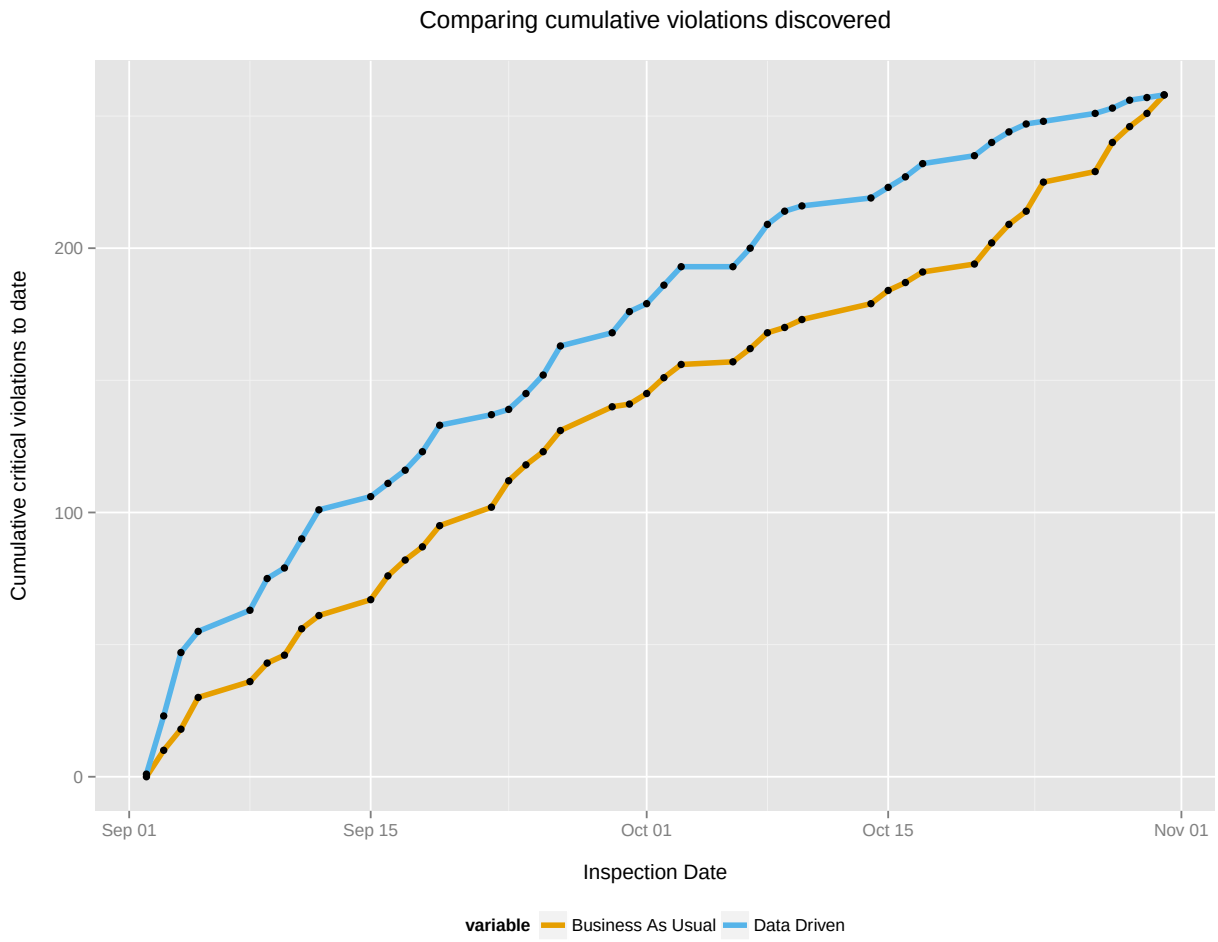
On average, food establishments were identified 7.44 days earlier under a data-driven model. Generally, critical violations would have been found sooner under the data-driven regime. The rate of finding critical violations in the first half of the would have increased by 25.0% under the data-driven model. % of critical food violations were found in the first half; meanwhile, under the data-driven model, % of all of the critical violations (an increase of 26.2%).

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 days earlier while quarter of all violations were prioritized over 25 days sooner. Yet, some restaurants would be prioritized lower, 99 restaurants were incorrectly prioritized lower and were found to have critical violations later–38% of the observed critical violations.



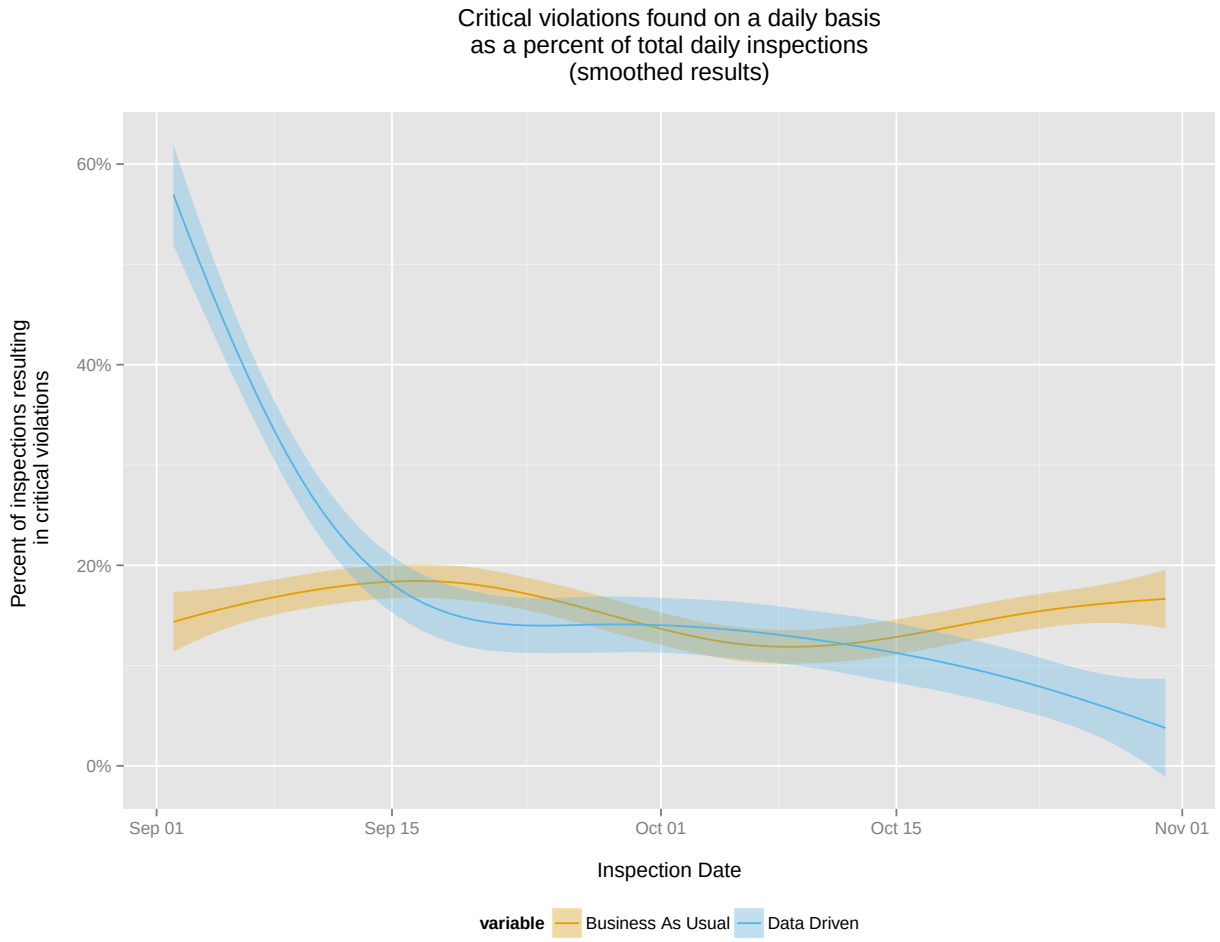
We conducted a t-test to measure whether the reduction in time to find a critical violation was greater than zero. Namely, the null hypothesis is the average time each food inspection was accelerated is equal to zero. The test ($\sigma = 25.20$, df $= 257$) resulted in a p-value of 3.542e-06, which indicates that the model is extremely likely to be significant.

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.

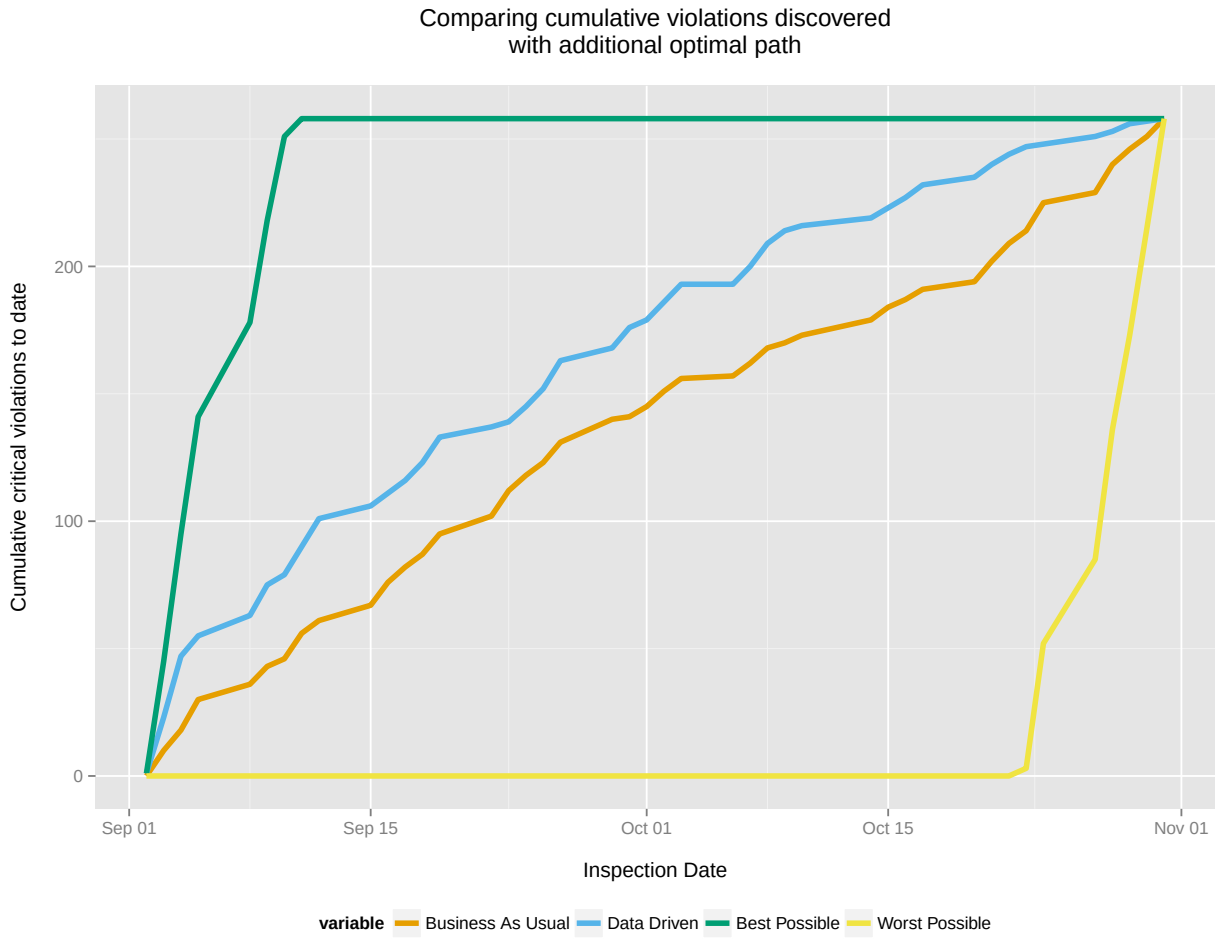


By extension, the rate of finding critical violations is higher for the initial quarter of the pilot. The rate of the violations are higher in the first portion as the analytic model correctly ranks higher-risk restaurants for earlier inspection. Business-as-usual has a more consistent rate of discovery, 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.

```
## Warning in loop_apply(n, do.ply): the matrix is either rank-deficient or
## indefinite

## Warning in loop_apply(n, do.ply): the matrix is either rank-deficient or
## indefinite

## Warning in loop_apply(n, do.ply): the matrix is either rank-deficient or
## indefinite

## Warning in loop_apply(n, do.ply): the matrix is either rank-deficient or
## indefinite
```



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



6 Summary

This model was able to reduce the timeframe to discover critical violations at Chicago's food establishments. The sanitarians proved to be significant predictors of finding critical violations, as did the outcomes of previous food inspections, the time since the last inspection, whether the establishment had an alcohol consumption and tobacco retail licenses, the average temperature, nearby burglaries and sanitation though 311. Older businesses and the number of garbage complaints through 311 were negatively related to finding critical violations.

Within a two-month window, the average time to find a critical violation was reduced by 7.44 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.

While we found inspectors are helpful in fitting and explaining the variation of restaurants, it is not practical to include inspector estimates on a daily basis. Inspectors may be reassigned based on absences, uneven workload, and other factors. Thus, the daily model removes inspectors from generating the estimates levels for each restaurant.

Likewise, weather is included, but effectively undifferentiating factor in daily estimates. Our weather model uses a single temperature across the entire city, which does not impact any region more than the other. Since we're only moving the intercept for all restaurants, the inclusion of weather only provides a more meaningful predicted value of a food establishment, but does not help change the order of inspections.

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. Because this is an open-source model, outside researchers will be able to freely reproduce, verify, and suggest improvements to the model presented in this paper. Researchers can submit suggestions and modifications by creating a pull reqest. The instructions to submitting a pull request can be found in this study's contributing guidelines.

References

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. http://www.jstatsoft.org/v33/i01/.

Harris, Jenine K., Raed Mansour, Bechara Choucair, Joe Olson, Cory Nissen, and Jay Bhatt. 2014. "Health Department Use of Social Media to Identify Foodborne Illness — Chicago, Illinois, 2013–2014." *Morbidity and Mortality Weekly Report* 63 (32): 681–85. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6332a1.htm.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13. http://www.jstatsoft.org/v39/i05/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.