Course: CS 109A Introduction to Data Science (Fall 2016)

Final Project: Predicting Food Inspection Outcomes in Chicago (Milestone #2)

Team Members: Calvin J Chiew, Angelo Kastroulis, Tim Hagmann

Teaching Fellow: Taylor Names

Literature Review

Readings:

1. Sean Thornton. Delivering Faster Results with Food Inspection Forecasting: Chicago's Analytics-Driven Plan to Prevent Foodborne Illness. http://datasmart.ash.harvard.edu/news/article/delivering-faster-results-with-food-inspection-forecasting-631.
2. City of Chicago. Food Inspection Forecasting: Optimizing Inspections with Analytics. https://chicago.github.io/food-inspections-evaluation/.

Summary:

The problem

There are more than 15,000 food establishments in Chicago but fewer than three dozen Department of Public Health inspectors to conduct sanitary inspections on them annually. On average, 15% of these establishments will have at least one critical violation on inspection. However, given the large inspection to inspector ratio, many of these errant establishments will only be discovered long after the violations have occurred, thereby exposing the public to risk of food-borne illnesses in the interim.

The idea

Instead of conducting inspections in their usual order, food establishments that are most likely to have critical violations can be forecasted and prioritized for inspections. This will optimize the food inspection process in the city allowing errant establishments to be identified more efficiently and quickly.

The approach

The City of Chicago's Department of Public Health, in collaboration with the Department of Innovation and Technology, Civic Consulting Alliance and Allstate Insurance, employed predictive data analytics to forecast establishments with high likelihood of failing inspection. The team interviewed food inspectors to understand what factors may be associated with failing an inspection. Then, leveraging on Chicago's large open data portal, they explored several data sources, including datasets on past inspections, sanitation code complaints, weather, and community and crime information, to find variables associated with food inspection outcome.

The model

The team developed a model based on random forests and wrote their code in R. They determined the following predictors to be significant in their model:

- prior history of critical violations
- three-day average high temperature
- nearby garbage and sanitation complaints
- type of facility
- nearby burglaries
- possession of a tobacco and/or incidental alcohol consumption license
- length of time since last inspection
- length of time the establishment has been operating
- inspector assigned

Each establishment could be assigned a probability of failing inspection based on the number of predictive factors met.

The evaluation

The effectiveness of this analytical mode of operation was evaluated via double-blind simulated retrodiction over a two-month pilot period. Using the data-driven order of inspections, 69% of establishments with critical violations were found in the first month, as compared to only 55% on the usual workflow. On average, establishments with violations were discovered 7.5 days earlier if forecasting was employed.

The future

The team has made the project open source allowing for other researchers to refine the current model and to improve its predictive capability. Other cities interested in adopting a similar data-optimized approach to food inspections can also use Chicago's work as a starting point.

<u>Project Vision</u>

How our project will be similar

Using the publicly available dataset of about 130,000 food inspections in Chicago since 1 January 2010, we will develop a model for predicting food inspection outcomes. Building on the intuition gained from Chicago's project, we will examine how variables contained in the dataset, as well as other factors such as day/month/neighborhood affect inspection outcomes. We will also match the food inspections to climate data from National Centers for Environmental Information and sanitation complaints data from Chicago's data portal via geohashing.

How our project will be different

We may approach this as a multiclass problem (pass, pass with conditions or fail) instead of a binary one (pass or fail). We will train different multinomial logistic regression models and code in Python. Results can be presented with map visualization (either static or d3). We will also explore whether other climate indicators (eg. precipitation) or ways of measuring temperature (eg. daily maximum, 7- or 10-day mean) makes a difference to the model. Additionally, we may consider including some of the following datasets on Chicago's open data portal:

- Rodent baiting
- Crimes
- Building permits
- Building code violations
- Vacant building locations
- Affordable rent locations
- Energy consumption
- Water temperature
- Street sweeping
- Graffiti

Other possible features considered were the type of cuisine, average menu item price and employee wage. Unfortunately, data on these variables are not readily available.