Predicting Food Inspection Outcomes of Chicago Restaurants

Most cities perform food inspections on an annual basis, dividing all restaurants to be inspected between a relatively small number of inspectors. Often times, due to their annual schedule, these inspections are performed long after violations have occurred and thus, put the general public at greater risk. If restaurants with greater probabilities of a violation were inspected more hastily, the spread foodborne illness could be reduced. The city of Chicago (among other cities) has turned to data analytics to tackle this problem and has made historical inspection data publically available. Can one use techniques from machine learning to analyze historical data of food inspections, and predict the inspection outcome given the restaurant and time?

Milestones:

1. Project Selection: Form teams of 3 or 4 and select a project from the provided list.

2. Literature Study: Go through the following resources for background on the project and write a half to 1 page summary for each one:

- http://datasmart.ash.harvard.edu/news/article/delivering-faster-results-with-food-inspection-forecasting-631

A previous project on this topic:

- https://chicago.github.io/food-inspections-evaluation/

3. Data Exploration and Cleaning: The primary source of data for this project is a publicly available repository of around 334,000 restaurant inspections in the Chicago area since 2010:

- https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/data

In addition, one can also make use of weather data recorded during this period:

- http://www.ncdc.noaa.gov/cdo-web/datasets

and data about sanitation code complaints in the Chicago area:

- https://data.cityofchicago.org/Service-Requests/311-Service-Requests-Sanitation-Code-Complaints/me59-5fac

Perform the following exploration steps:

- Decide on a suitable database to store the data, and on a computing resource to process the data.

- Visualize changes in inspection outcome during different days of the week, month of the year; observe common patterns (if any).

- Visualize changes in inspection outcomes during different weather conditions and seasons; observe common patterns (if any).
- Visualize changes in inspection outcome for differing facility types, neighborhood, and risk levels.
- Check for correlations (if any) between inspection outcome and nearby sanitation code complaints
- Remove records with spurious entries (e.g. with invalid geographic coordinates, inspection outcomes that are neither **pass**, **pass with conditions**, nor **fail**)

4. I5. Proposal:

Propose methodologies and ideas to be implemented, tested and interpreted for your final project.

Implement Baselines:

- Decide on the *performance metric* to evaluate prediction.
- *Feature extraction*: Extract a set of basic features from restaurant inspections data and from the weather data. Decide on a suitable method to convert location attributes to numerical values (e.g. consider using geohashing)
- Implement the following baseline techniques:
  - *Simple averaging*: Predict inspection outcome for a location and time of week/month/year by simply averaging the outcome probabilities at the same location and time from previous weeks/months/years.
  - *Multinomial logistic regression*: Train a multinomial logistic regression model on features extracted from past restaurant inspections.
  - *Additional features*: Train a multiclass logistic regression model with additional features from the weather data, as well as nearby sanitation code complaints; observe changes in performance.