Predicting Food Inspection Outcomes in Chicago

CS 109A Data Science Final Project, Fall 2016
Calvin J Chiew, Angelo Kastroulis & Tim Hagmann (TF: Taylor Names)

Introduction

Problem

- There are more than 15,000 food establishments in Chicago but fewer than three dozen food inspectors.
- 15% of these establishments will have at least one critical violation on inspection.
- Many of them are discovered long after the violations have occurred, thereby exposing the public to **risk of food-borne illnesses** in the interim.

Solution

- The City of Chicago developed a model to **predict establishments with critical violations** and prioritized them for inspections.
- In a pilot study, the data-optimized order of inspections identified unsafe establishments earlier than the usual workflow, by 7.5 days on average.

Chicago's Model

Their model was based on random forests with the following predictors:

- Prior history of critical/serious violations
- 3-day average high temperature
- Nearby sanitation complaints
- Nearby burglaries
- Possession of a tobacco/alcohol license
- Length of time since last inspection
- Length of business operation
- Inspector assigned

Email: calvinjchiew@mail.harvard.edu
GitHub: https://github.com/angelok1/cs109project

Data

Data Sources

- Chicago's publicly available dataset of ~130,000 **food inspections** from 1 January 2010 to present
- Requested daily **climate** data from National Centers for Environmental Information (NCEI)
- **Business licenses**, **crimes** and **sanitation** code 311 complaints data accessed from Chicago's open data portal

Data Preparation

- **Transformed** new variables, eg.
  - Days since last inspection
  - Proportion of past inspections failed
  - Most recent inspection outcome
  - 5-day rolling mean of max temperature
- **Regrouped** into broad categories:
  - Inspection type
  - Facility type
- **Mapped** to main dataset:
  - Climate data by date
  - Business licenses data by license ID
  - Crime & sanitation data by "cluster"

Data Exploration

- Past failure, inspection type and geographical "cluster" are associated with failed outcome.
- Failure rates tend to increase in the mid-year months corresponding to higher temperatures.
- There is geographic variation in crimes and sanitation complaints which may be related to outcome.
- As days since last inspection increases, proportion of failure also increases.

Fig 1: Selected Predictors

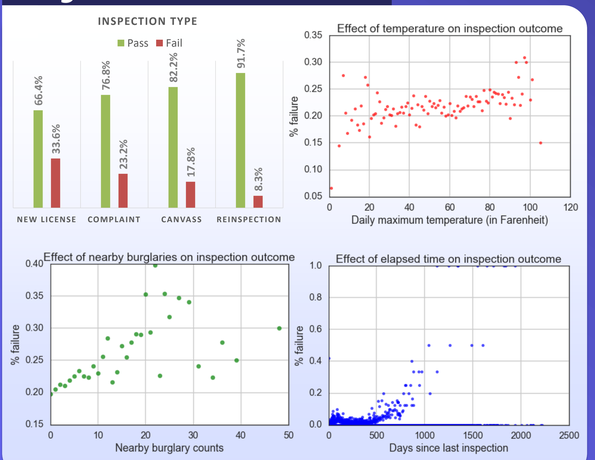
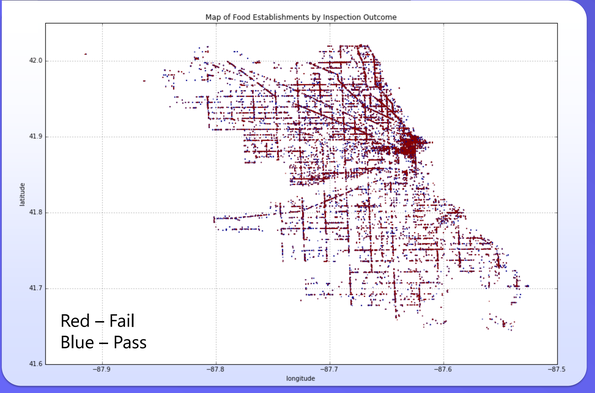


Fig 2: Geographic Location



Model

Model Fitting

- We handled missing values, applied one-hot encoding to categorical variables, and split data into training and test sets in 7:3 ratio.
- Using a binary outcome (pass/fail), we created two **baseline models**, an "all-zero" model predicting all inspections to fail, and a "random label" model predicting failures based on overall proportion.
- We explored various **classification models** and employed logistic regression as an ensemble method to "stack" all the models.
- We also tuned the parameters of our models via cross-validation. (We used a time series cross-validator to prevent information on lag predictors leaking across validation sets).

Model Evaluation

- Overall **test accuracies** of the various models are shown in Table 1.
- The by-class test accuracies reflect the **sensitivity** and **specificity** of the models, which relate to their **F1 score**. ROC curves are shown in Fig 3.
- Using an out-of-sample test set (last 2 months), we simulated predicting failures and ordering inspections.
- For each model, we scored **average number of days earlier** (or later) failures were inspected compared to the actual dates, and **proportion of failures caught in the first month**.

Table 1: Model Test Accuracies

Model	Test Accuracy (%)			F1 score
	Overall	"Pass" class (Specificity)	"Fail" class (Sensitivity)	
Baseline Models				
All zero	22.57	0.00	100.00	0.368
Random label	34.58	22.16	77.21	0.348
Candidate Models				
Logistic regression (weighted)	70.59	64.20	92.52	**0.587**
Linear discriminant analysis (LDA)	77.63	96.45	13.08	0.209
Quadratic discriminant analysis (QDA)	77.56	99.34	2.81	0.054
Random forest (RF)	**82.13**	91.82	48.91	0.553
Support vector machine (SVM)	77.88	99.58	2.22	0.043
ADA boost (ADA)	80.79	91.85	42.85	0.502
Ensemble	82.13	91.82	48.91	0.553

Fig 3: Model ROC curves

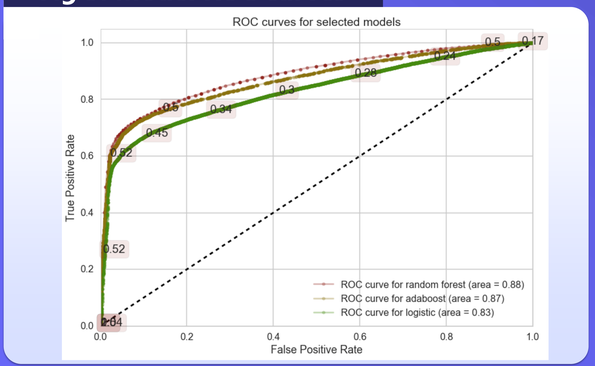


Fig 4: Feature Importance Rank