Predicting Food Inspection Outcomes in Chicago

Luke Farewell, Jake Gober, Sam Green, Jeremy Welborn

Computer Science 109a / Statistics 121a, Harvard University

Objectives

Chicago's Department of Public Health is responsible for inspecting 15,000 food establishments across the city. Our goal was to reduce the amount of time required to discover critical violations:

- Aggregate and clean useful data sources.
- Build models for probabilities of failure.
- Optimize inspection efficiency and social cost of undiscovered health risks.

Data: Sources and Cleaning

Food Inspections	**Business Activities**
Weather	**Crime Reports**
Sanitation Complaints	**Business Location**

Sources: City of Chicago Open Data Portal, NOAA.

1. Pair inspections with business information.
2. Process past inspection text to build history.
3. Bucket spatial data with a grid (see visualizations).

Exploration (I)

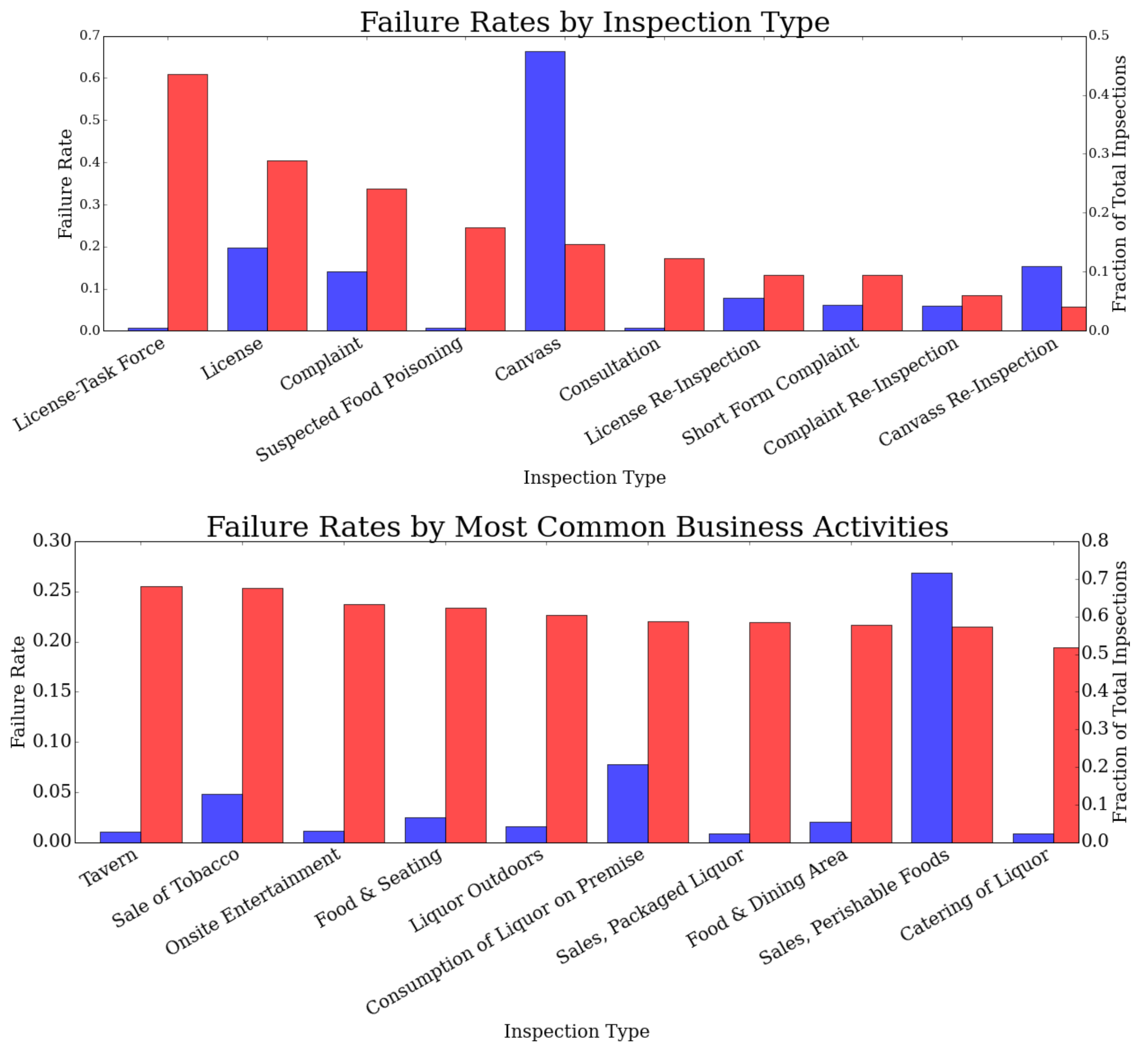

Figure 1: Inspection Results by Feature

Results (I): Unsuccessful Models

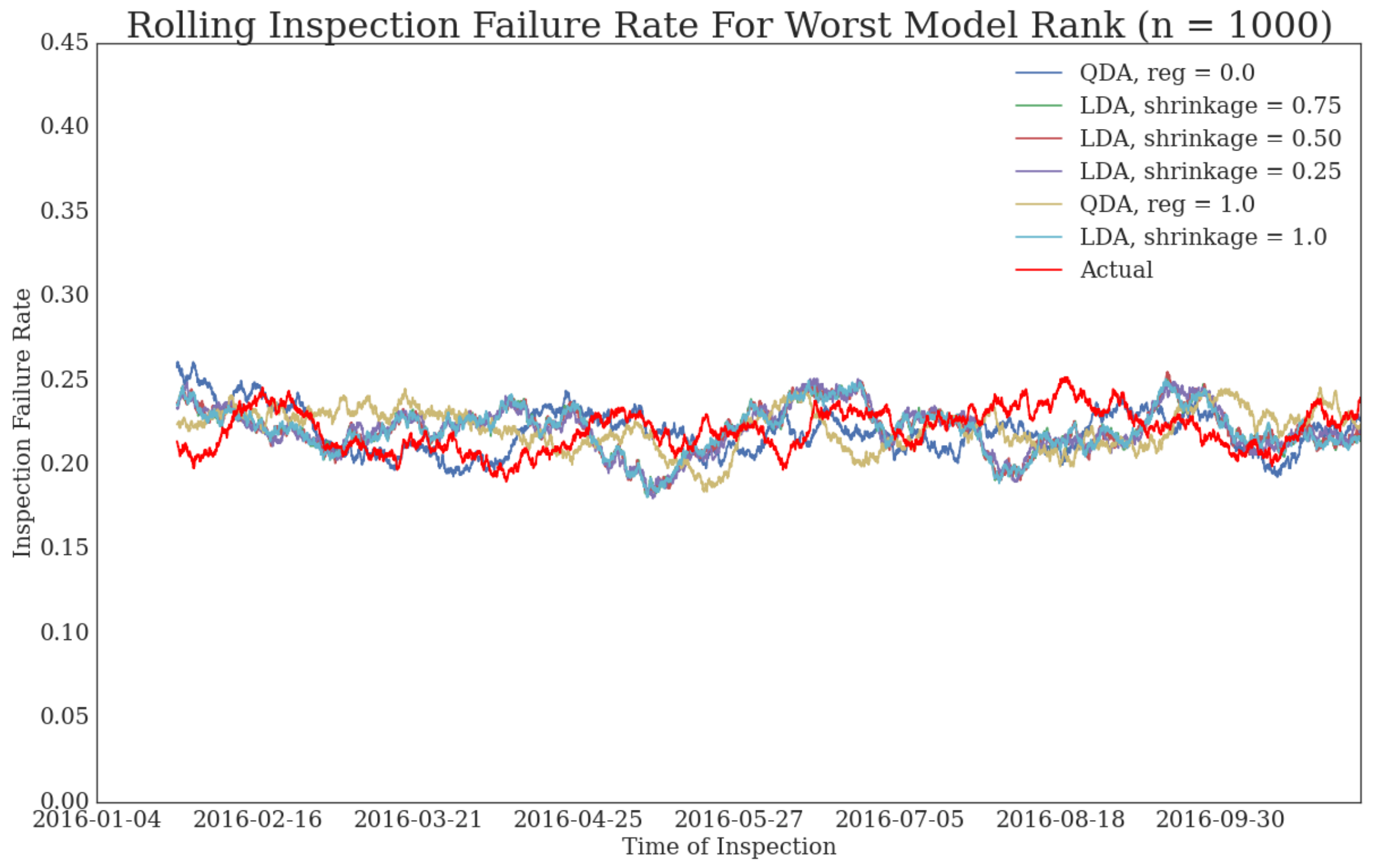

Figure 2: Rolling Failure Rate, Poorly Performing Models

Model Scoring: Log Loss

To measure *rank accuracy* and match the recommendation setting, we selected **log loss** as our scoring function.

$$-\frac{1}{n}\sum_{1}^{n}[y_i\log(p_i)-(1-y_i)\log(1-p_i)] \qquad (1)$$

for n observations, where the ith observation is of correct class $y_i \in \{0, 1\}$ which our model assigns probability p_i.

Results (II): Successful Models

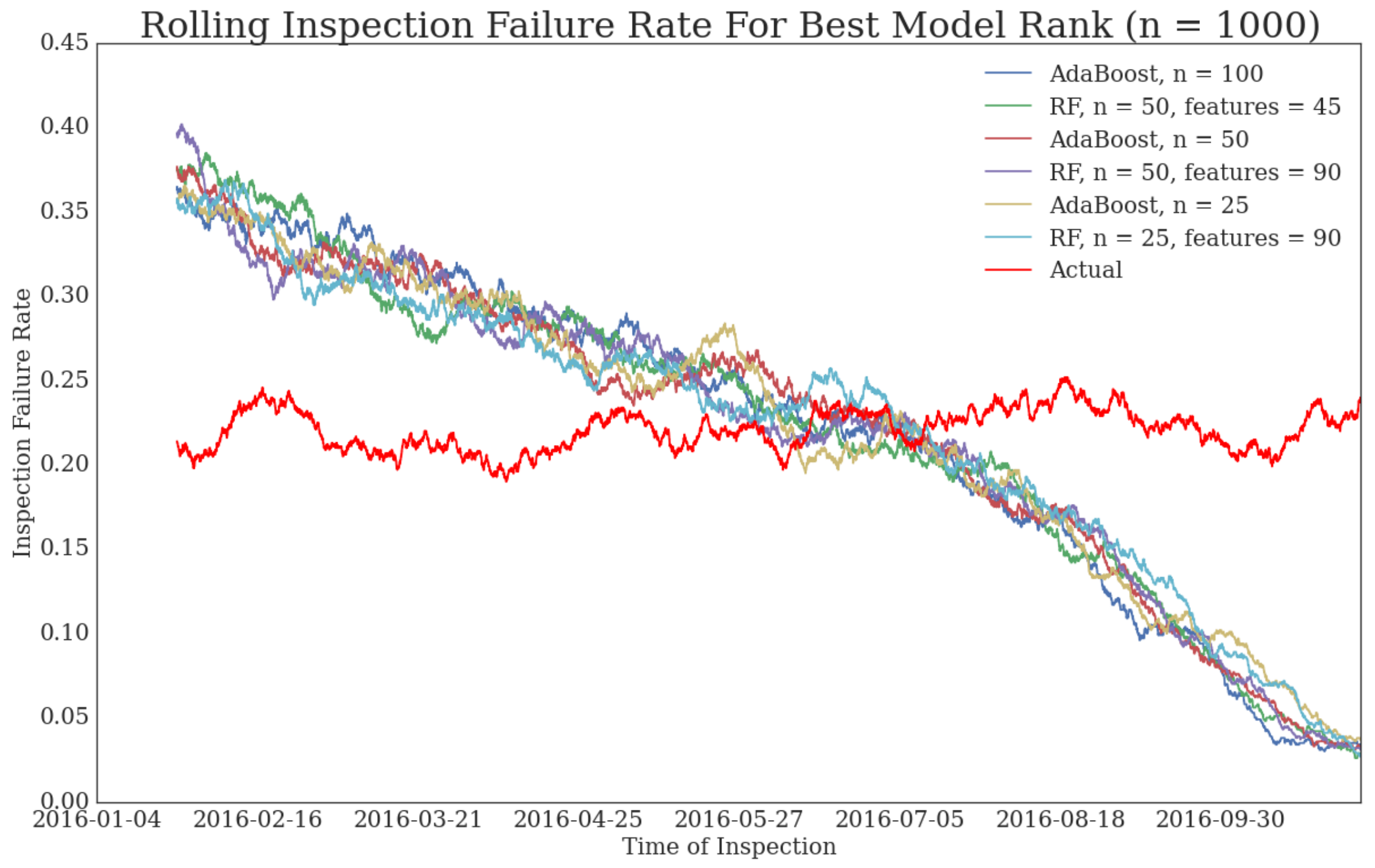

Figure 3: Rolling Failure Rate, Well Performing Models

Model Selection: Rank Quality

To choose between models, we treated test data as undated possible inspections and select the model that minimizes

$$\frac{1}{n}\sum_{i}^{n}\mathbf{1}_{y_i=1} * n_i^{\text{days\_to\_discover}} \qquad (2)$$

because the goal is to catch failures as early as possible. This quantity is the average days until true failures are discovered.

Exploration (III): Spatial Predictors & Neighborhood Dynamics

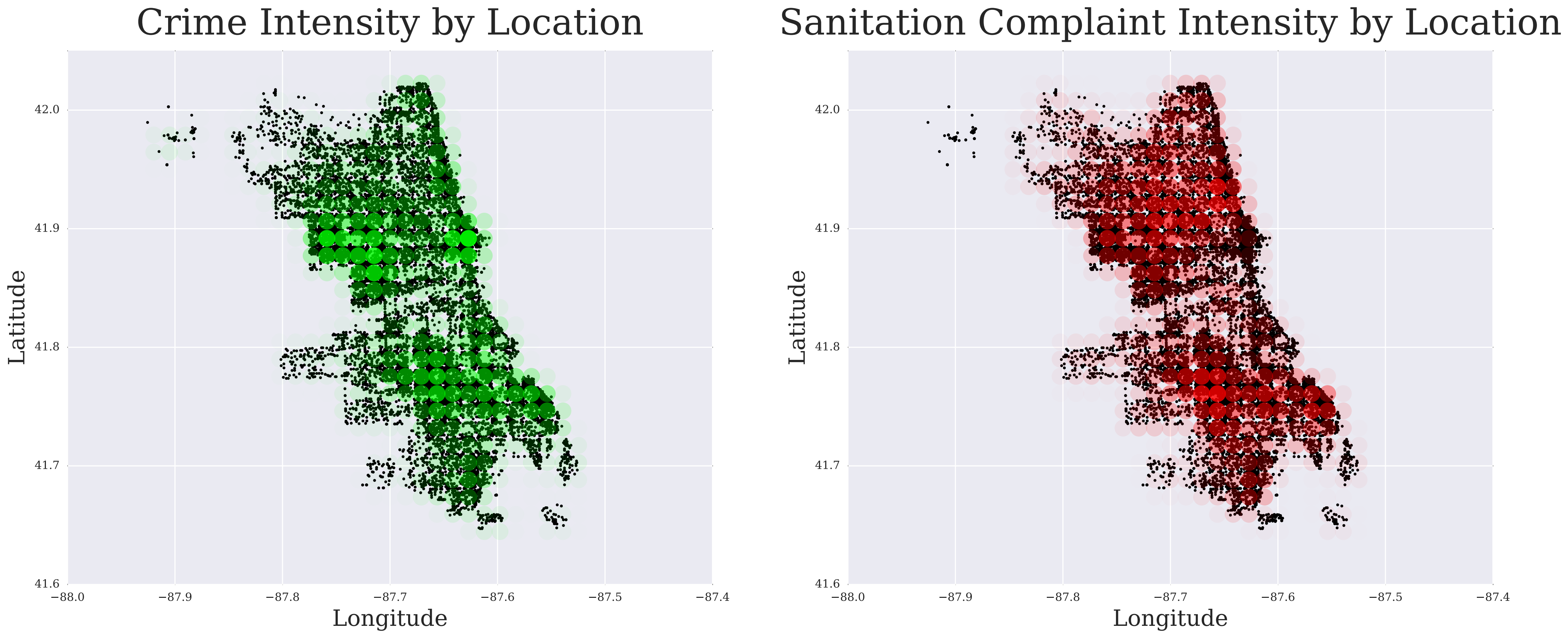

Figure 4: Crime (left) and Sanitation Complaint (Right) by Location. Darker shades represent a larger fraction of observations.

Results (III)

Based on our cross-validition process, AdaBoost was selected as the optimal model in this setting, as shown below in the barchart displaying the selection statistic:

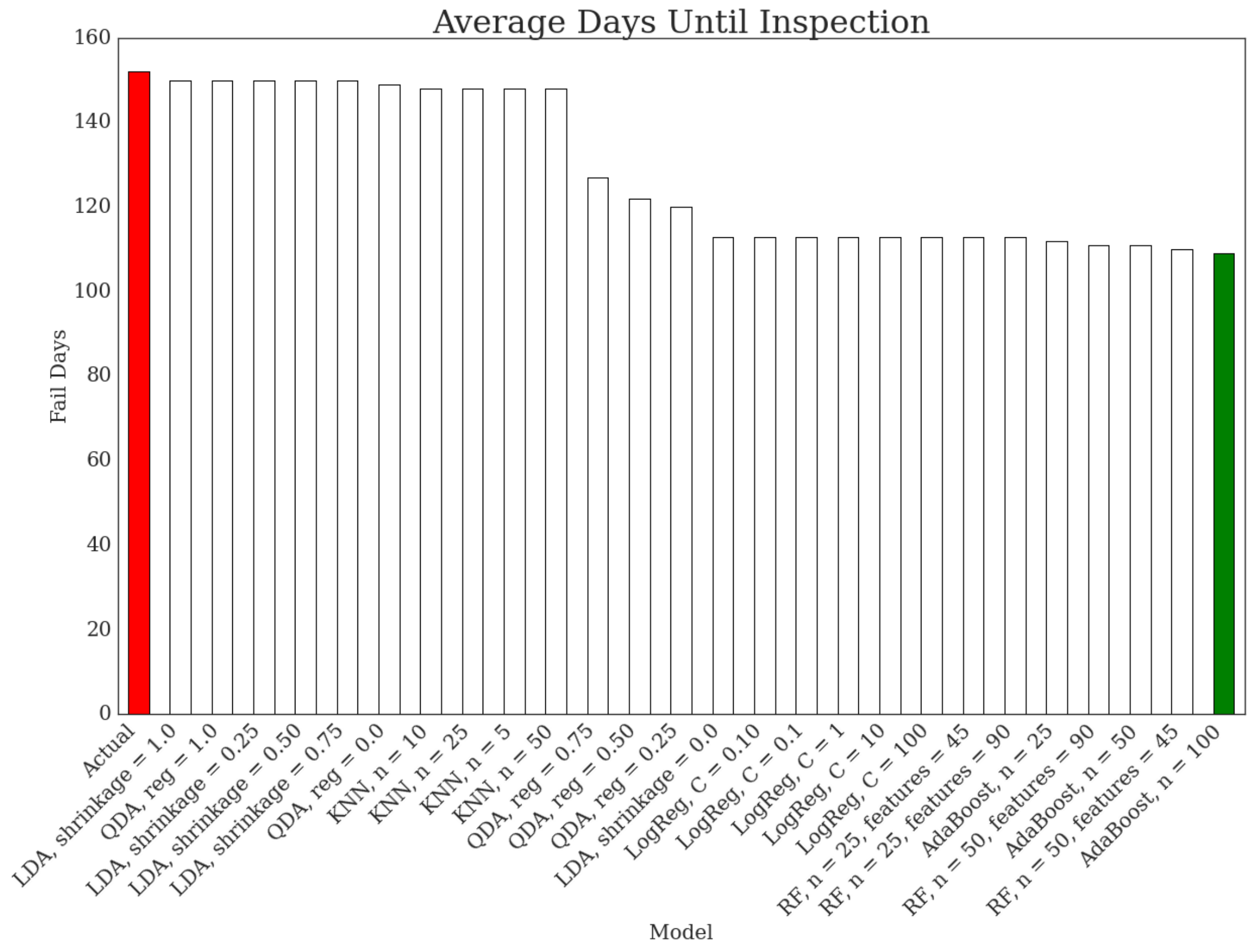

Figure 5: Average days to discovery for failing businesses.

A boosted Random Forest model is intuitive in this setting, considering that several predictors were highly non-linear in the response. All of the best performing models are bagged or boosted decision tree ensemble methods.

Next Steps

Currently the model ranks all possible inspections in the testing data, but in practice, it would be ranking inspections of all possible businesses and itself determining what inspections actually take place. The model also does not account for the gradual appearance of complaints over time.

As a result, the frequency with which the model is reset needs to be tuned, to set the d days worth of inspections generated by each ranking.

Contact Information

- Web: fggw.github.io/foodinspections/
- Code: github.com/fggw/foodinspections/
- Contact: {lfarewell, jgober, samuelgreen, jeremywelborn}@college.harvard.edu