Predicting Food Inspection Outcomes in Chicago

Luke Farewell, Jake Gober, Sam Green, Jeremy Welborn

Computer Science 109a, Harvard University

Objectives

This research project explores modeling outcomes of health inspections of food establishments in the city of Chicago. We had the following goals:

- Design a model that can be portably applied to new data.
- Define a reasonable cost function.
- Successfully predict inspection outcomes to optimize the social costs of inspections

Introduction

The City of Chicago's Department of Public Health is responsible for inspecting upwards of 15,000 establishments that sell food within the city limits. Previous work published by the City of Chicago's advanced analytics team showed that using historical inspection data and a variety of other information helped predict establishments that were at the highest risk of critical violations. That previous work is already several years old. In this project, we aimed to replicate the results of the previous study on an updated dataset and, where possible, to include new, potentially more predictive data to optimize the inspection process further.

This project is an example of data science applied for the public good. Citizens are less likely to be exposed to critically unsafe food establishments, while inspectors are able to work more effectively.

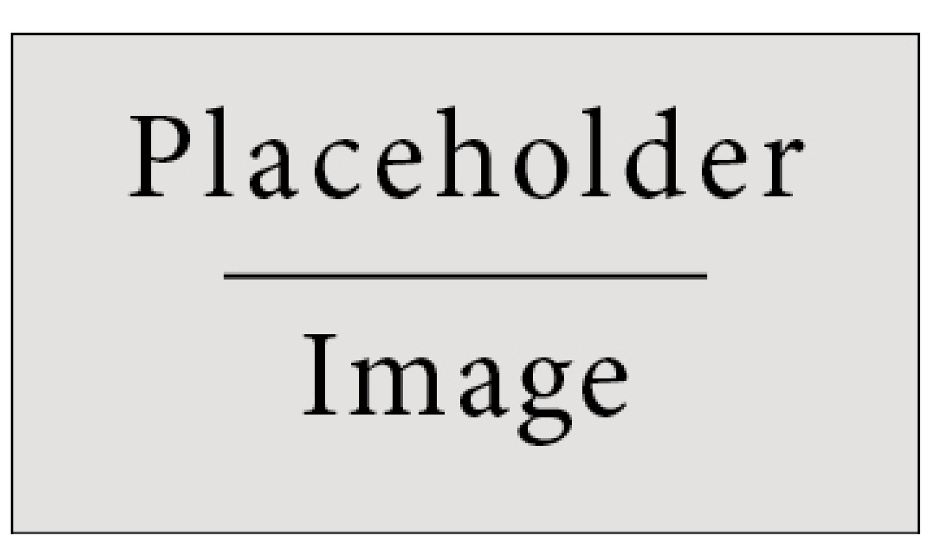

Figure 1: Figure caption

Spatial Predictors

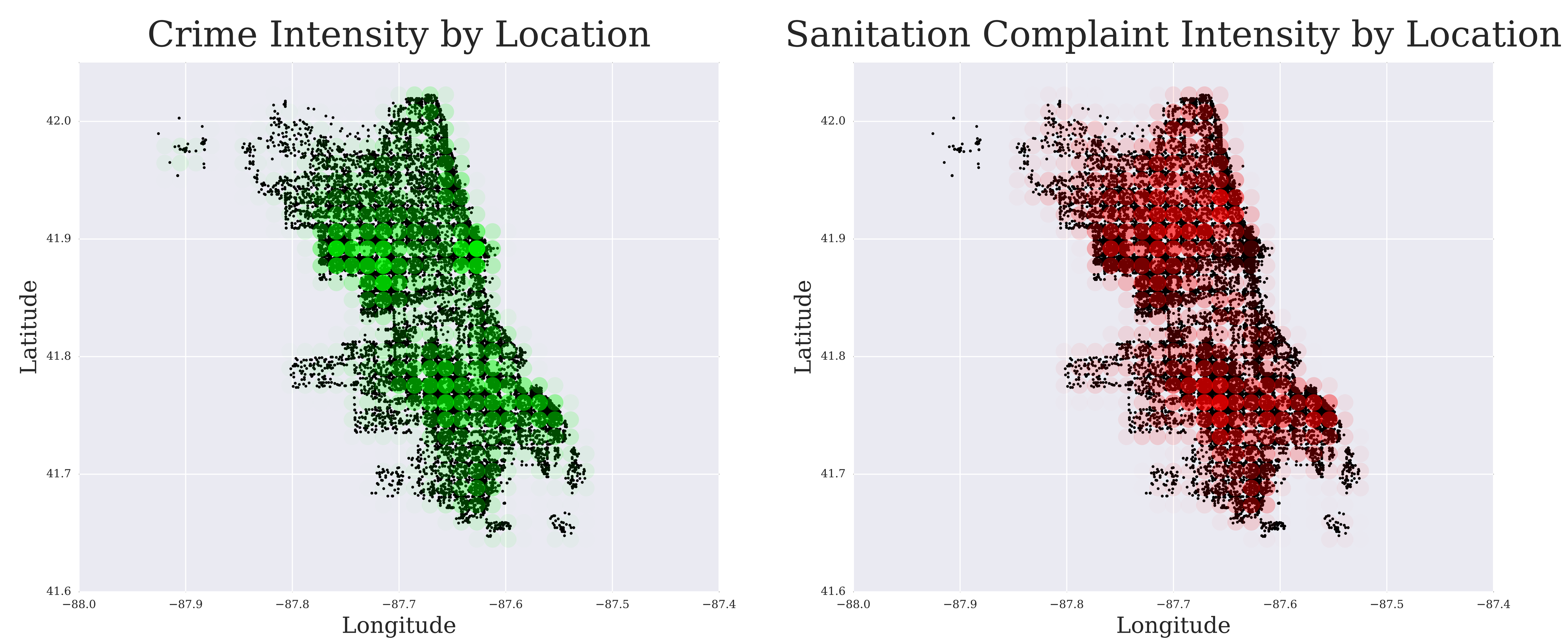

Figure 2: Crime (left) and Sanitation Complaint (Right) by Location. Darker shades represent a larger fraction of observations.

Important Result

Lorem ipsum dolor **sit amet**, consectetur adipiscing elit. Sed commodo molestie porta. Sed ultrices scelerisque sapien ac commodo. Donec ut volutpat elit. ‘

Model Scoring

Choosing a scoring methodology to measure our model performance was one of the most important decisions we made.

Standard accuracy scoring for our classifiers ignores the nuances of the setting. In practice, the model needs not to predict outcomes, but to rank possible inspection choices. The accuracy of the probability distribution assigned to each observation is more important than the actual classification.

So that we can measure *rank accuracy*, we selected **Log Loss** as our scoring function.

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

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

Results

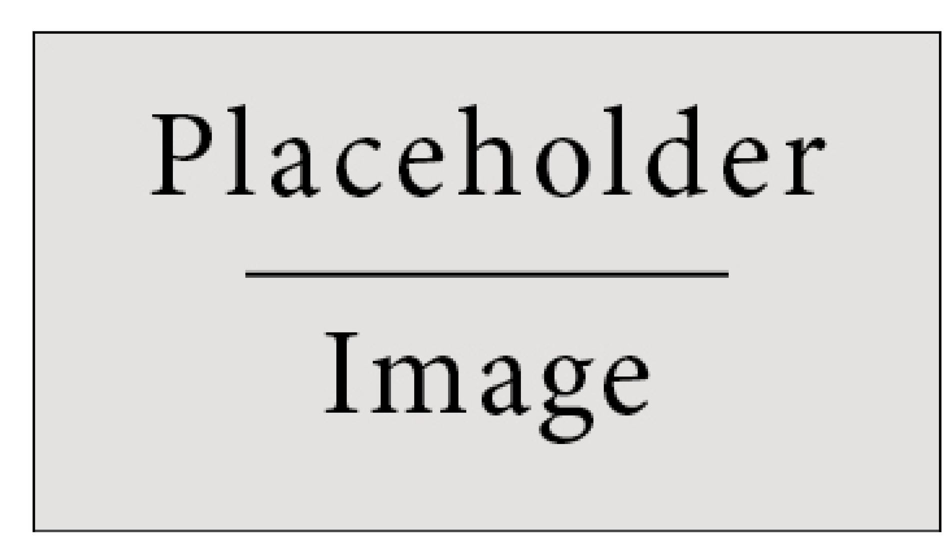

Figure 3: Figure caption

Nunc tempus venenatis facilisis. Curabitur suscipit consequat eros non porttitor. Sed a massa dolor, id ornare enim:

Treatments	Response 1	Response 2
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table 1: Table caption

Conclusion

Nunc tempus venenatis facilisis. **Curabitur suscipit** consequat eros non porttitor. Sed a massa dolor, id ornare enim. Fusce quis massa dictum tortor **tincidunt mattis**. Donec quam est, lobortis quis pretium at, laoreet scelerisque lacus. Nam quis odio enim, in molestie libero. Vivamus cursus mi at *nulla elementum sollicitudin*.

Future Work and Open Questions

Maecenas ultricies feugiat velit non mattis. Fusce tempus arcu id ligula varius dictum.

- Curabitur pellentesque dignissim
- Eu facilisis est tempus quis
- Duis porta consequat lorem

References

Acknowledgements

Nam mollis tristique neque eu luctus. Suspendisse rutrum congue nisi sed convallis. Aenean id neque dolor. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

Contact Information

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

HARVARD
JOHN A. PAULSON
SCHOOL OF ENGINEERING
AND APPLIED SCIENCES