Predicting the Success of City of Chicago Food Inspections

41204 Machine Learning - Kolar

Alex Bue, Christian Gregorich, Conrad Liu, Ujjwal Sehrawat

Table of Contents

Introduction	*Page 3*
Data	*Page 5*
Methods	*Page 7*
Results	*Page 8*
Conclusion	*Page 10*

Inspections are a key piece of Chicago's food safety strategy for restaurants and other establishments

Introduction

- Each year, Chicago establishments that serve food are subject to inspection to ensure continued compliance with City ordinances and regulations
- In addition to recurring inspections, restaurants may also be inspected in response to a complaint
- Chicago has conducted more than 250,000 inspections of more than 40,000 unique businesses since 2010
- Inspections have seven possible outcomes, or "results"
    - Pass, pass with conditions, fail, not ready, no entry, business not located, and out of business

The team's goal was to use machine learning techniques to predict inspection outcomes

1

PREDICTION

Gain intelligence into which types of businesses are most likely to fail food safety inspections

2

IMPROVE TARGETING

Use machine learning techniques to build an accurate model for predicting inspection failure

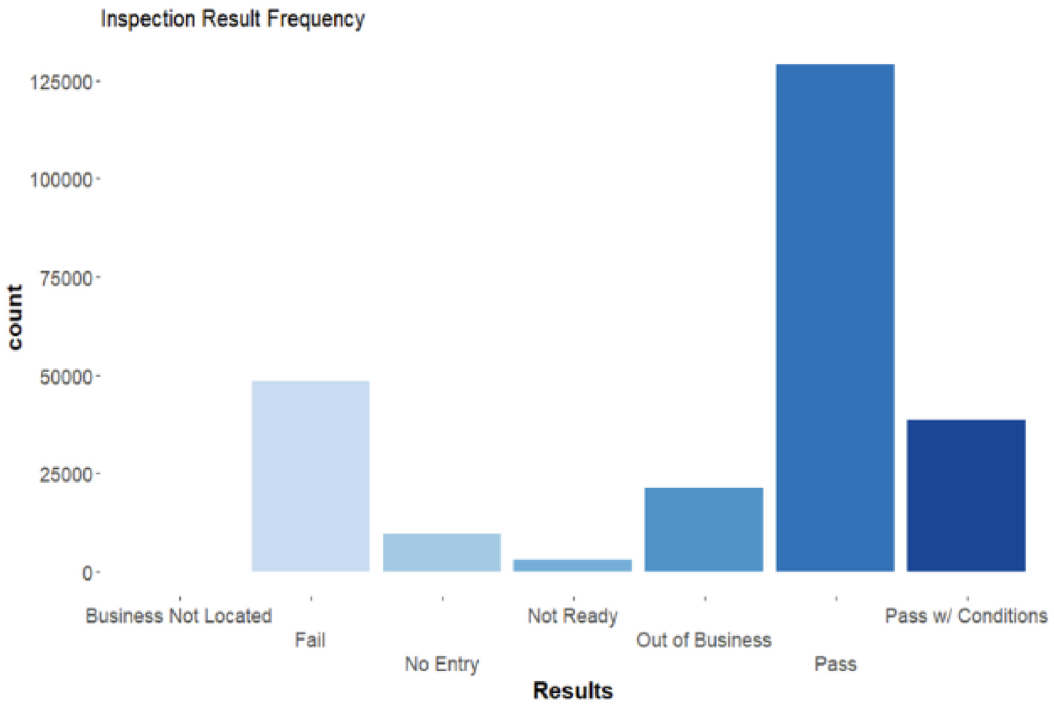
3

WORK TOWARD KEY OUTCOMES

Improve public health and optimize utilization of limited inspector resources

Thankfully, a large majority of inspections end with passing outcomes; however, a significant portion do end in failure

Inspection Outcomes



Inspections dataset contains over 250,000 observations

Over half of inspections end with a passing outcome without any conditions

Our project came with significant data integrity issues that required time and judgment to resolve

Data Integrity Issues

Idiosyncratic Data

Lack of Cleanliness and Standardization

Lack of Cleanliness and Standardization

Example Solution

Creation of New “Facility Type” Categories

Example Solution

Manual Cleaning of “City” and “State” data based on best judgment

We augmented the inspections dataset with several other City or external data sources

Augmentation Strategy	Commentary
Temporal Dependence	• *Used inspections dataset to build in "pseudo time series" features such as how many inspections an establishment had previously experienced, how many of these inspections they passed or failed across different inspection types, prior number of violations, and previous inspection result*
Other City of Chicago Data	• *Joined other City of Chicago public datasets to the inspections dataset to improve predictive power* • *City data included information on public safety, demographics, business locations, and community activity*
External Data	• *Leveraged external data, such as the Zillow Home Value Index to add further features tied to the location surrounding inspected establishments*

We chose to build models using four different machine learning techniques

We started with our project priorities

1. Optimize for predictive power and accuracy
2. Accommodate computational and operational restrictions
3. Demonstrate understanding of several methods

...and decided to move forward with 4 ML techniques

1. ***Model 1***
   Logistic Regression
2. ***Model 2***
   Random Forests
3. ***Model 3***
   Boosted Trees
4. ***Model 4***
   Neural Networks

Results: Confusion Matrices

Model 1: Logistic Regression

	Y	N
Y	18,492	15,093
N	2,990	6,602

Model 2: Random Forests

	Y	N
Y	21,671	11,914
N	3,398	6,194

Model 3: Boosted Trees

	Y	N
Y	21,588	11,997
N	3,318	6,274

Model 4: Neutral Networks

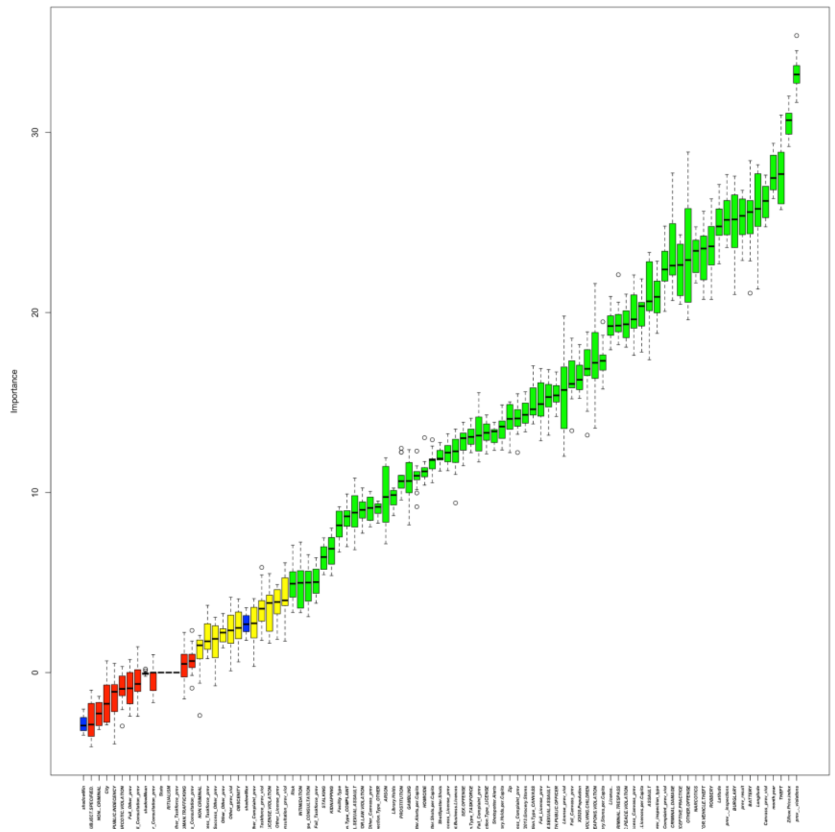
	Y	N
Y	20,939	12,646
N	3,636	5,956

Key Takeaways

- Random Forests and Boosted Trees are best with accuracy of 64.5%.
- Random forests have balanced false positive and false negative rates of 35.5%.
- Boosted Trees are better at predicting negative results with false negative rate of 34.6%.

Results of Boruta feature selection largely indicate that our feature augmentation efforts were successful

Boruta Algorithm Results



Key Takeaways

- Past failures seem to be the most useful feature for predicting future failures across models
- Added features such as Zillow Home Price Index and crime data proved to perform well
- Zip code was important in all models, indicating that there are likely differences in behavior by neighborhood that explain inspection failures
- Overall, the features we engineered or added to the dataset were very important

CONCLUSION

How did we do?

The team's goal was to use machine learning techniques to predict inspection outcomes

1

PREDICTION

Our best model (Boosted Trees) has 64.5% accuracy with, importantly, fewer false negatives than Random Forests.

2

IMPROVE TARGETING

Boosted Trees model has a lift value of 2.25 over random targeting in top decile of observations.

3

WORK TOWARD KEY OUTCOMES

- $6M in FY21 budget for food inspections
- Just 24% of total observations failed inspections.
- If our model improves efficiency by even 10%:
    - That's $600,000 in fiscal savings!