Heinz 95-845: Assessing the Performance of Machine Learning Methods at Predicting Outcomes of Food Inspections in Allegheny County

Kaila Gilbert kjgilber/andrew.cmu.edu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

Matt Samach msamach/andrew.cmu.edu
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

1. Project Details

The following project seeks to analyze five years of food inspection records discovered from the Western Pennsylvania Regional Data Center (WPRDC). The goal of this project is to compare the performance of a number of different statistical and machine learning methods at predicting the outcome of a given food inspection. This analysis will draw potential insights into food inspection outcomes while also elucidating the comparative performance of a number of popular statistical models. We expect some combination of violation frequency, violation magnitude, and customer rating will adequately predict the inspection outcome (store closure, consumer notice, or inspected permitted), although features related to geography, seasonality, and inspector-level data might also play a role. Restaurant owners and Pittsburgh consumers both have a stake in better understanding any trends related to a health code violation or store closure. This model could be used by health and safety inspectors to optimize their targeting. It could also be used for the enlightenment of the typical hungry Pittsburgher.

The data sets utilized for this study comes from WPRDC and includes information between 2014 and 2019. Two central data sets form the basis of this study: a record of inspections at food facilities and a record of violations. The inspections data set contains details for 70,000 inspections conducted at a food facility, while the violations data set contains 290,000 records of resulting violations, including a description of the type and magnitude of the violation. Both data sets are updated regularly. Due to the size of the data and the number of associated features, this data set presents a promising opportunity to see how various models could compare in predicting a certain outcome.

Our analysis will explore a number of different machine learning techniques and assess the effectiveness in predicting whether or not a given inspection will end in a facility closure. Posited methods for study include logistic regression, random forest, and neural networks. For each method, our loss criterion will be mean-squared error. In each case, we will train our model on a train set and then assess its performance on a hold-out sample for validation. For each model, necessary steps related to regularization, stochasticity, and hyper-parameters will be considered to ensure the final models do not overfit.

1.1 Our Data

- Inspection Outcome: Permitted, Consumers Alerted, Store Closure (multi-label)
- Covariates: Date, Time, Location, Consumer Rating, Year Established, Historical Info, Inspection Description, Violation Description, Magnitude of Violation, Owner demographics

1.2 Our Machine Learning Pipeline

The posited steps of our analysis are below:

- Data Collection: We will merge the inspection data with the violation data and conduct some initial data exploration to understand the nature and distribution of our variables. Due to the absence of interesting features in the original data set, we plan to join this restaurant information with other data sources. One such publicly available source is Yelp's API. By adding information related to consumer feedback (ratings), we can add new features that can paint a fuller picture of the likelihood food facility will be closed. Another source is the US Census American Community Survey, which includes yearly demographic and economic data down to the block group level. Other data sources may be incorporated as the project progresses.
- Data Processing: This process will include cleaning and merging in supplemental data, mapping/remapping inspection and violation descriptions, aggregation of violation types and quantity per inspection, and filtering out redundant or irrelevant columns.
- Feature Engineering: This step includes
  1.) One-hot encoding of dummy variables related to severity,
  2.) Creating binary indicator of closure for supplementary analysis, and
  3.) Converting address data to appropriate unit of analysis (e.g. neighborhood)
  4.) NLP methods for Yelp reviews such as word counts or sentiment analysis
- Model Formulation: During this phase, we will train a logistic regression, random forest, and neural network to our data, using methods discussed in lecture.
- Model Validation: Upon being fit to training data, each of our trained models will predict the outcomes of our validation set.
- Model Evaluation and Comparisons: We will evaluate and compare models based off performance, measured by mean squared error.
- Analysis of Outcomes, Impact, and Next Steps: Finally, we will discuss implications of both prediction findings and model performance.

1.3 Previous Analyses

At the time of this document's submission, we have no awareness of any similar analyses of inspection prediction outcomes conducted professionally or otherwise. This was somewhat confirmed by a semi-thorough Google Scholar Search.

1.4 Limitations

This analysis is subject to several limitations. Due to the availability of features, this analyses may not fully capture all the elements and factors that predict an inspection's outcome. In addition, results from this analysis may not be generalizable to other locales, as it encompasses the inspection process unique to Allegheny County. Finally, there will be a slight bias towards model simplicity, since the complexity of our methods may be somewhat limited by the capacity of our laptops.