Project 3
Classification for Prediction Problems
Sociology 273L: Computational Social Science

1 Introduction

In this project, you will learn how to develop machine learning models for classification problems. You will be working with data drawn from the City of Chicago's Open Data Portal. Specifically, you will develop models to predict whether a business passes or fails a food inspection. Food safety is a major concern for city governments, and most regularly conduct food safety inspections to ensure that establishments maintain adequate safety standards. However, municipal governments have limited resources and cannot audit every establishment.

You have been tasked with developing an algorithm that will help the Department of Public Health prioritize establishments to audit. You have been provided with a dataset that combines various open data sets curated by the City of Chicago as part of a program to improve food inspection auditing. Your goal is to predict the **Results** column using features taken from datasets covering business information, previous inspection results, and neighborhood information.

2 Data Pre-Processing and Cleaning

The "Chicago Inspections 2011-2013.csv" dataset contains data taken from eight different datasets. The task of merging these datasets together has already been done for you. We have also provided some preliminary visualizations and dropped out several features. You may still wish to examine the dataset and drop any other features you believe will not be helpful. Also, make sure to examine the target variable and determine whether you wish to keep it as three possible categories, or if it should be recoded.

3 Fit Models

3.1 Data Splitting

You may either split your data into train/validation sets, or use cross-validation. Note, if you use train/validation sets, **do not** create a test set (you will see why soon).

3.2 Fit Models

First, do the following:

- Choose 3 different machine learning techniques. See available ones in the scikit-learn documentation
- Detail the basic logic and assumptions underlying each model, its pros/cons, and why it is a plausible choice for this problem.

Then, train your models (again either on the train set or using cross validation). Remember to use hyperparameter tuning.

3.3 Validation Metrics

How well did your model do? Report the following:

- Accuracy
- Recall
- Precision
- F1 Score

Which of these metrics would you want to prioritize when conducting predictive auditing in this context? Why?

4 Policy Simulation

4.1 Interpretable Machine Learning

Use tools like coefficient plots or feature importance plots to investigate your models. Which features contribute to your predictions? Are there any additional features you wish you could incorporate?

4.2 Prioritize Audits

Imagine that the City of Chicago only had resources to conduct 1000 food safety inspections. Generate a list of the 1000 riskiest establishments. Using your chosen metric, demonstrate how well your algorithm prioritized finding potential violations.

Then, conduct a simulation where you choose 1000 establishments to randomly audit. Again using your chosen metric, how well did random audits do? How did random auditing compare to predictive auditing?

4.3 Predict on Data with Unseen Labels

Use your favorite model to make predictions based on the features in “Chicago Inspection 2014.csv”. Note that this dataset **does not** include a “Results” column. After you make your predictions, choose a metric that you think makes the most sense from a policy perspective and explain your choice. Then, save your predictions in a file called “predictions.csv”. The instructor will check your predictions against the observed labels on the metric you chose and report your model performance back to you after you submit the assignment.

5 Discussion Questions

5.1 Why do we need metrics beyond accuracy when using machine learning in the social sciences and public policy?

5.2 Imagine that establishments learned about the algorithm being used to determine who gets audited and they started adjusting their behavior (and changing certain key features about themselves that were important for the prediction) to avoid detection. How could policymakers address this interplay between algorithmic decisionmaking and real world behavior?