Using Machine Learning to Optimize New York City's Restaurant Inspections

Matt Roth | API-222B | December 6, 2021 | Jupyter Notebook

Introduction:

Background: Restaurant inspections are a vital service provided by cities to ensure that residents are protected from food-borne illnesses and other hazards. Food safety compliance is also a critical factor in the local economy – establishments that cannot meet inspectors' standards can be forcibly shut down, or they may lose business due to unsanitary conditions or the reputational harm associated with posting low inspection grades.

Generally, the restaurant inspection process involves sending inspectors on unannounced visits to every establishment in a jurisdiction. This is a daunting logistical challenge for any local government, but the complexity is magnified in New York City, which has over 27,000 restaurants spread across 300 square miles.[i] New York City's Department of Health and Mental Hygiene (DOHMH) inspects every establishment at least once per year. Inspection grades are assigned on a scoring system which tallies points for every violation depending on the type and severity.[ii] Establishments receiving low grades are reinspected on an accelerated timeline.[iii] In total, DOHMH conducts about 45,000 inspections per year.

Motivation: By improving their inspection process, DOHMH can achieve better public health outcomes, save money, increase capacity, and help at-risk establishments stay in business. This paper explores the possibility of using machine learning to optimize restaurant inspections. If DOHMH can predict which establishments are likely to commit health violations and receive low grades, they can optimize their inspection program and achieve better outcomes for the city.

Data Description:

DOHMH publishes records from all restaurant inspections on NYC Open Data Portal, an initiative by the city government to make datasets accessible to the public.[iv] The 'Restaurant Inspection Results' dataset[v] contains a record for every violation or citation from all restaurant

inspections in the past 3 years. Establishments include all restaurants and college cafeterias that were in-business at the time the data was pulled (December 6, 2021). It is worth noting that DOHMH paused inspections during the COVID-19 pandemic, leaving a gap in the data from March 16th, 2020, to July 19th, 2021. The dataset, which contains 384,000+ rows and 26 columns, provides information in two broad categories described below: i) information on the establishment itself, and ii) inspection results and details.

i: Information on the establishment: Each establishment is provided a unique identification number, which is used in our analysis to aggregate observations. The dataset contains: 1) information about the establishment's location, including borough, address, zip code, geocoordinates, community board, council district, census tract, neighborhood tabulation area, building identification number, and the borough-block lot, and 2) details about the establishment's food and business, including name and cuisine type.

ii: Inspection results and details: Each row in the dataset corresponds to a violation from an inspection, and documents the inspection type, inspection date, the violation type, whether the violation was critical, the points tallied for the violation, and the establishment's grade from that inspection (if applicable).

Data Cleaning & Feature Engineering:

Data Cleaning: Given that our model only seeks to predict A/B/C letter grades – not any of the provisional grades (P/Z/G/N) that DOHMH assigns – we remove establishments from the dataset that have never received an A/B/C letter grade (roughly 18%). We use each establishment's most recent grade as the target variable, so we are unable to make predictions on restaurants without A/B/C grades. Next, we drop columns corresponding to building number, phone number, and the date the database was updated, as these are unlikely to be predictive of

inspection grades. Finally, we remove records with no recorded borough (0.08%) from the dataset and fill null values in all columns with zeroes.

To create our target variable, records from each establishment's most recent inspection are split out from all previous inspection data. We only include A/B/C letter grades from two inspection types: "Cycle Inspection / Re-inspection" and "Pre-Permit (Operational) / Re-inspection." Other inspection types are infeasible for prediction because they do not result in full distributions of letter grades (i.e., the other 30 inspection types either only give provisional grades or all A's). Roughly 33% of establishments do not have a qualifying target grade and are dropped from the dataset. Finally, we convert the most recent grade from three levels to two by combining B and C into a single class, "B or C." We explain the rationale for this step in the Methods & Results section.

Feature Engineering: To create our feature set, we start by rounding latitude and longitude to two decimal places to geographically bucket nearby establishments at nodes placed 1.1 km apart.[vi] We then aggregate data from all previous inspections by establishment, based on their unique ID. Our predictors include the following for each establishment: the number of inspections, the average score across all inspections, the total number of violations, the number of times an inspection resulted in no violations, the number of times an establishment was forcibly closed by DOHMH, the number of times a restaurant was reopened after a forced closure, the number of critical violations from all previous inspections, the percentage of previous grades resulting in each letter grade, the first letter grade received, the last grade received before the most recent one (i.e. before the grade we use as the target variable), a count of each violation type (105 types), the number of violations per inspection, the number of critical violations per inspection, cuisine category (a manual bucketing of cuisine types into broader

categories), and concatenations of geographic data and establishment information including (street + borough), (category + borough), (category + street + borough), (cuisine + street + borough), (category + census tract), (cuisine + census tract), and (cuisine + borough). All categorical features from the original dataset (see "i: Information on the establishment" in Data Description) are included in the final feature set. Due to the high-cardinality of our categorical features (e.g., there are 82 cuisine types, 225 census tracts, 162 zip codes, etc.), we perform target encoding, which replaces an observation's categorical level with the mean of the target variable across all other observations belonging to that category.[vii] We assign a score of 100 for an A and a score of 0 for a B or C for our calculation. In total, we use 97 features, 21 of which were present in the original dataset.

Methods & Results:

Model Selection: Given that our policy application calls for predicting which establishments will receive low inspection grades, we built a classification model with the positive class being "B or C" grades, and the negative class being "A" grades. Our criteria for model selection included 1) interpretability, 2) performance on the minority class ("B or C"), and 3) ability to handle high-dimensional data on varying scales without requiring much preprocessing. Given these requirements, early versions of our model used a Random Forest Classifier, which works by fitting multiple decision trees to separate samples drawn from the training dataset – each using different subsets of observations and features – before making predictions based on the most common result from all the component decision trees.

In defining our success criteria, we considered that an effective predictive model for DOHMH's use-case must flag as many low grades as possible while maintaining a precision-level well above chance – 23% in our case. In other words, the model's value hinges on

achieving high recall, as long as precision is kept above the point where misclassifications of high-grade recipients waste DOHMH's time or money, or cause an undue burden on restaurants.

Dealing with imbalanced classes part 1 – reframing as a binary classification problem: Initially, we attempted a multi-class prediction with each grade as a distinct class, but the large class imbalance – 77% A's, 16% B's, 6% C's – limited the model's ability to make accurate predictions. When we viewed the model's predicted class probabilities, it was clear that low and high grades could be differentiated, but it was harder to separate B's from C's. As a result, the predicted probability of "B" and "C" membership would be split across two classes, diluting each relative to the predicted probability of "A." For example, the model might predict probabilities of 40%, 30%, and 30% for "A," "B," and "C" respectively – in this case, it is more likely "not A" than "A," but the model predicts "A" because that probability is higher than "B" and "C" individually. After combining "B" and "C" into a single category, the Random Forest performed relatively well with hyperparameters tuned using a cross-validated randomized search – that model achieved 56% recall on "B or C" predictions at 80% precision.

As described earlier, however, the success of our model is contingent on achieving high recall. We determined it would be desirable to create a second model with better recall, even if it came at the price of lower precision. We once again targeted the class imbalance to improve performance.

Dealing with imbalanced classes part 2 – rebalancing classes: Even after combining "B" and "C" into a single class, the dataset was still highly imbalanced at 77% "A" and 23% "B or C." We hypothesized that the low-recall, high precision performance of the first model was attributable to the class imbalance. Specifically, we suspected that the imbalance was rewarding the model for predicting "A" frequently and failing to penalize overly selective "B or C"

predictions. We used a method known as Synthetic Minority Oversampling Technique, or SMOTE, to rebalance the dataset to have an equal number of observations in each class. SMOTE accomplishes this by iteratively selecting a random datapoint from the minority class, finding the 5 nearest neighbors, and creating a synthetic observation in that neighborhood's feature space.[viii] We fit a new Random Forest Classifier, with newly tuned hyperparameters, to the rebalanced training data, and made predictions on the imbalanced test data. The result was an improvement in recall from 56% to 76% at the expense of precision, which fell from 80% to 67%. We consider this model to be an improvement over the previous one given the importance of recall to our success criteria. Trading 13 percentage points in precision for a 20-point gain in recall allows DOHMH to flag an overwhelming majority of low-grade recipients at an acceptable level of precision – nearly 3 times better than chance. Given these promising results, we sought to further improve performance on our synthetically balanced dataset using a Voting Classifier, which compares the predictions of multiple models and assigns classes based on the most frequent or most confident prediction from the component classifiers.[ix]

Using a Voting Classifier on the synthetically balanced dataset: Our next model built off the SMOTE-balanced Random Forest by combining that model with several weak learners in a Voting Classifier. We tried including various combinations of classifiers including Logistic Regression, K-Nearest Neighbors (KNN), AdaBoost, Gradient Boost, and XGBoost before settling on just the Random Forest and KNN with a 4:1 vote-weighting. The Voting Classifier achieved a modest improvement over the standalone SMOTE Random Forest with a recall of 85% (+9% compared to the previous model) and a precision of 63% (-4%).

Replacing Random Forest with XGBoost on the imbalanced dataset: Having improved the performance of our model using the synthetically balanced training data, we

decided to revisit the original imbalanced dataset to see if we could improve minority class recall using boosting, which runs multiple classifiers sequentially and corrects for past prediction errors in future iterations. We created two models: a standard XGBoost and a weighted XGBoost, the latter of which scales the degree of error correction for the minority class in boosting.[x] We used a cross-validated randomized search to tune hyperparameters for both models, resulting in a standard model with 72% recall (+16% compared to the original imbalanced Random Forest) and 78% precision (-2%), and a weighted model with 91% recall (+35% compared to the original imbalanced Random Forest) and 63% precision (-17%). The weighted model outperforms the SMOTE Voting Classifier by 6% in recall with no sacrifice in precision. The standard model achieves a good balance between precision and recall, and provides DOHMH a strong alternative if they were to decide to place more emphasis on the model's precision.

Choosing the best model: The highest performing models were the two XGBoost Classifiers trained on the imbalanced dataset. Each was evaluated using 5-fold cross validation:

Table 1: Weighted XGBoost – Results from 5-Fold Cross Validation

	Precision	**Recall**	**F1 Score**	**Support**
Grade: A	0.97	0.87	0.92	2573
Grade: B or C	0.63	0.91	0.74	627

Accuracy: 0.87
ROC AUC: 0.95
F1 Macro Score: 0.82

Table 2: Standard XGBoost – Results from 5-Fold Cross Validation

	Precision	**Recall**	**F1 Score**	**Support**
Grade: A	0.93	0.95	0.94	2573
Grade: B or C	0.78	0.72	0.75	627

Accuracy: 0.89
ROC AUC: 0.94
F1 Macro Score: 0.82

The weighted XGBoost Classifier, which flags 91% of low grades at 63% precision, is likely to be preferable for DOHMH's use-case. The standard XGBoost model presents a viable alternative if DOHMH were to decide to prioritize precision – for example if misclassifications became expensive due to high-cost interventions targeted at establishments that were incorrectly flagged by the model.

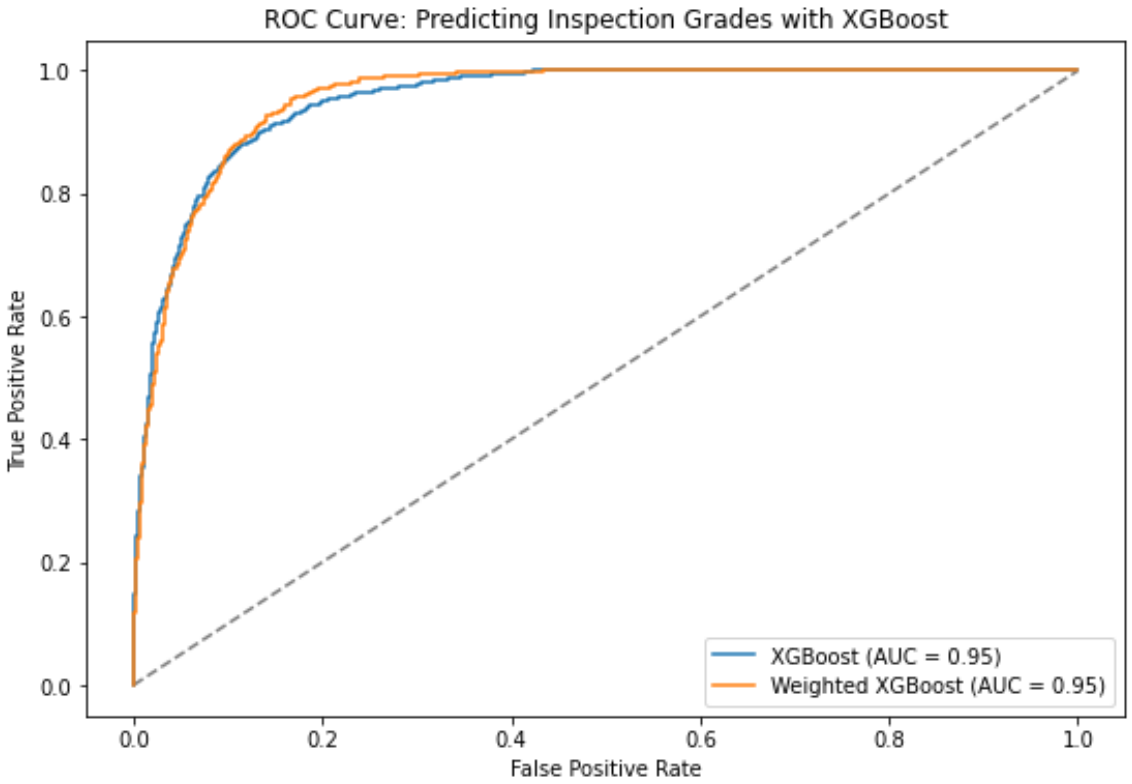

Figure 1: Comparing ROC curves illustrates the tradeoff between models, despite their equal AUC scores

Evaluating bias and fairness: DOHMH's use of machine learning to predict inspection grades could create an undesirable equity-efficiency tradeoff. For example, a model that makes perfect predictions based only on the restaurant owner's race would streamline DOHMH's operations, but it would be objectionable on ethical grounds as it might cause or reinforce discrimination. Any model that determines how a government interacts with its constituents should not be a black box – this is why interpretability was our top criterion for model selection.

XGBoost allows for examination of feature importances, giving us insight into how much weight each feature receives in the model's decision-making. Location, cuisine type, and concatenations of location & cuisine were the strongest predictors in our models. This finding may be problematic and should be investigated in greater depth by city officials who understand New York City's unique context before DOHMH operationalizes the model.

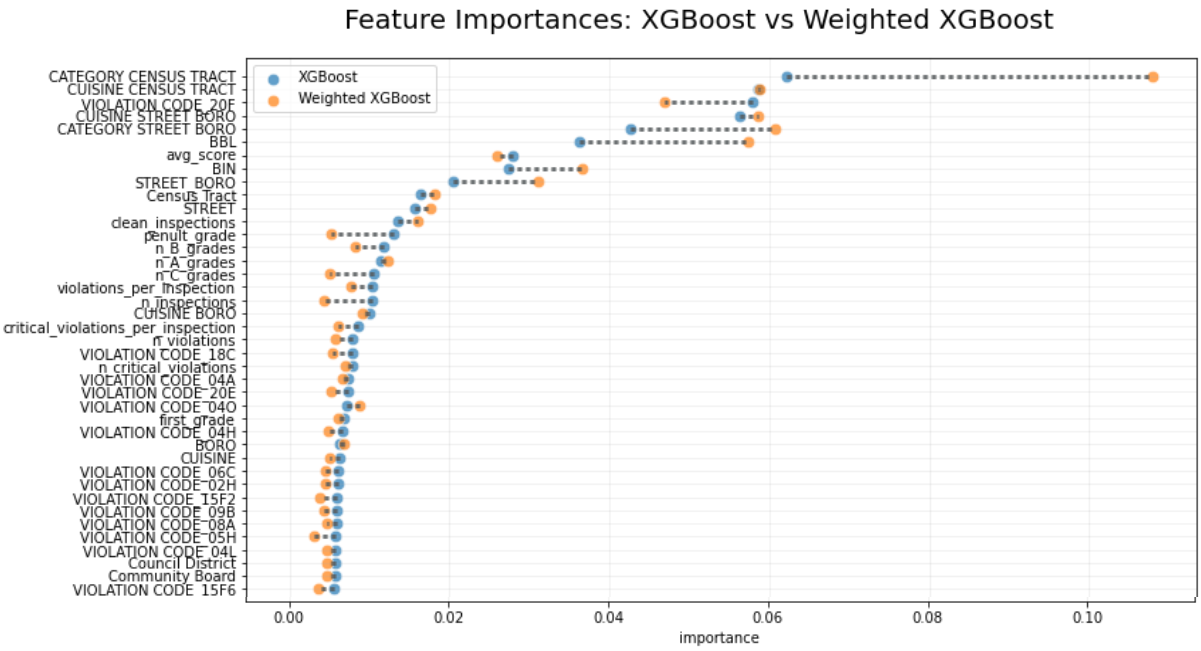

Figure 2: Comparing the 40 most important features for each XGBoost model reveals potential concerns

There are several options to mitigate algorithmic bias if the model's decision-making process were found to be objectionable. First, problematic predictors could be removed from the feature set. It is possible that the other features would provide enough predictive power for the model to retain its value. Alternatively, we could leave all features in the model but downrank problematic predictors, forcing them to be less central to the model's predictions. DOHMH could also take bias and fairness into account in their implementation plan so that human

decision-makers would be empowered to use their judgement and knowledge of the local context to override the model's predictions and audit its impact.

Expectations for out of sample performance: There are a few notable limitations to our model's predictions on out-of-sample observations. First, it can only make predictions on establishments with a prior inspection history. Our model provides no value to DOHMH for the subset of restaurants that are new or have otherwise never been inspected. Second, our model's out-of-sample performance may be hindered by the particular time period during which our training data was produced. The COVID-19 pandemic forced DOHMH to pause inspections for 16 months, which created a large discontinuity in the dataset. While that sparsity itself may affect the model's external validity, it is also possible that the pandemic affected restaurants or inspections in a way that could bias our model toward the peculiarities of 2021. For example, restaurants experiencing financial distress from the pandemic may have cut costs, leading to worse food safety compliance and lower grades. Alternatively, it is possible that inspectors graded more leniently in their first rounds after restarting the program. Or maybe sanitary conditions in restaurants deteriorated without oversight during the long layoff, necessitating harsher inspections. In any of those examples, we would expect the effects of the pandemic to influence or model's decision-making logic, which would affect its performance on new data. Finally, out of sample performance could suffer if DOHMH were to change their inspection methodology, the structure of their data tables, the categories they use for cuisine type or other features, or the types of violations they log.

Conclusion:

Machine learning proves to be effective for predicting restaurant inspection grades, giving DOHMH a new tool to identify the establishments that are most likely to pose health and safety risks to customers. DOHMH can use this information to target interventions to prevent outbreaks of food-borne illness and to help at-risk restaurants stay in business. We believe our model's performance – successfully flagging 70-90% of low grades at a precision 3x better than chance – is sufficient to use in the field. With accurate predictions, DOHMH can optimize inspection logistics and resource allocation, and improve restaurant outreach, oversight, education, and support.

Limitations: Model performance is very low when predicting B's and C's separately. It is possible that DOHMH would only find value in a model that could flag C's, given that the lowest grades present the highest risk and demand the most attention. It is also possible that our results, while promising, fall well short of the best possible predictor that could be built with better data or a stronger model. Future work should seek to incorporate other datasets and use other algorithms to achieve better results. Finally, the model can only make predictions when the inspection type is "Cycle Inspection / Re-inspection" or "Pre-Permit (Operational) / Re-inspection." We believe these are the most important inspections to predict, however, given that they're the only ones that result in assignment of all non-provisional letter grades. A new model would need to be created if a use-case arose for predicting provisional grades or other inspection types.

Implementation Steps: DOHMH would need to undertake the following steps if they were to integrate a predictive model into their inspection program: First, they would need to check that the model does not use any objectionable criteria when making predictions. If the model cannot be operationalized without significant risk of bias or discrimination, it must be

redesigned or scrapped entirely. Next, they would need to decide on what interventions they would undertake for establishments that the model flags. Options range from low-cost, low-touch strategies, such as instructional food safety pamphlets, to high-cost interventions including training programs and frequent re-inspections. If predictions are used to direct inspectors in the field, it is crucial that measures are taken to keep them blind to the model's predictions. The knowledge that an establishment was flagged by the model might bias inspectors, so DOHMH's implementation plan must account for protecting the process' integrity and neutrality. Finally, DOHMH would need to consider how to build a workflow around the model. How often would it be run? How would the results be used, and by whom? How would it be audited, and how often? Who would be responsible for technical support and maintenance?

These questions and issues are vital to consider – ultimately the model's success depends on how well it can be implemented within the operational context of city government. With thoughtful design, implementation, and oversight, we believe machine learning can offer significant improvements to how cities conduct restaurant inspections. The current approach relies on a brute-force strategy – cities send inspectors to every restaurant and determine follow-on actions depending on what they find there. Predictive modeling, on the other hand, allows cities to take targeted and/or preventative action where it is needed most. If they know which restaurants are most likely to commit health violations, they can achieve better public health outcomes and support at-risk restaurants while increasing operational efficiency and reducing costs.

References

[i] https://www1.nyc.gov/site/doh/services/restaurant-grades.page
[ii] https://www1.nyc.gov/assets/doh/downloads/pdf/rii/restaurant-grading-faq.pdf
[iii] https://www1.nyc.gov/assets/doh/downloads/pdf/rii/inspection-cycle-overview.pdf
[iv] https://opendata.cityofnewyork.us/
[v] https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j
[vi] https://gis.stackexchange.com/a/8674/7913
[vii] https://contrib.scikit-learn.org/category_encoders/targetencoder.html
[viii] https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html
[ix] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html
[x] https://machinelearningmastery.com/xgboost-for-imbalanced-classification/