Quantifying Restaurant Compliance Violations Through Machine Learning Pattern Recognition

Kyle Lyon, Lily Zeng, Somender Chaudhary,
Jenny Choi, Manu Sharma, Alexis Yang

DNSC 6317: Business Analytics Practicum

Professor Brian Murrow

9 May 2022

Table of Contents

Case Background

Deloitte is a multinational professional services company seeking new business opportunities at the municipal level of state governments. The firm has recognized that cities such as New York City (NYC) generate significant revenue from its taxation of restaurant commerce; however, its main objective is to promote the health and safety of the general populace dining in these establishments. Given the large sum of tax revenue generated by restaurants, Deloitte speculates that municipalities would consider funding agencies that can prevent food establishments from failing due to health inspection violations, thus ensuring the continued lucrativity of this tax revenue source.

Currently, the NYC Department of Health and Mental Hygiene (NYC DOHMH) conducts unannounced inspections of restaurants at least once a year. Inspectors check for compliance in food handling, food temperature, personal hygiene and vermin control. The department is responsible for regulating New York City's 24,000 restaurants.

Therefore, to demonstrate a minimal viable product, this research team sought to leverage New York City’s Open Data Initiative database to provide NYC DOHMH with analytical capabilities that enhance efficiency and reduce the cost of health inspections.

Project Objective

For scenario analysis, NYC DOHMH has contracted this team to quantify the growth of restaurant compliance violations through machine learning pattern recognition within the metropolitan city limits. A solution would enhance the department’s analytical capabilities by conducting more targeted screenings and inspections of dining establishments. The solution would also develop probabilistic risk scores for potential restaurant compliance violations and the affected restaurant patrons based on the historical patterns of restaurant inspection results paired with City Zip Code demographic data. Based on the outputs of the solution’s model(s), it should provide an estimator of the positive or negative impacts on the city’s revenue streams from compliance violations.

Problem Decomposition

The team decomposed the compliance violation quantification into three key questions: (1) Are all restaurants equally at risk? (2) How can inspectors be properly deployed? (3) Lastly, how do violations affect revenue?

These three questions guided this research team’s direction of thought and evolution of the problem decomposition as shown in Figure 1.

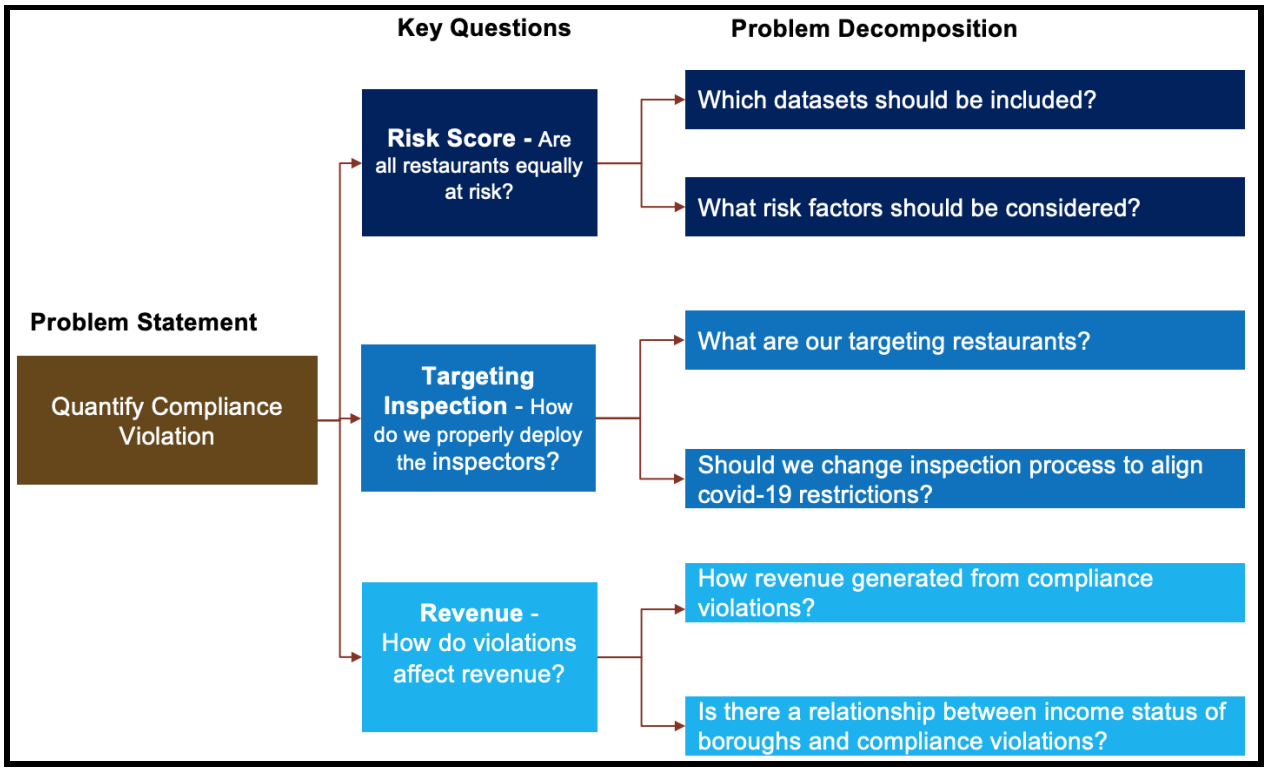


Figure 1: Problem Decomposition

The research team chose these three themes: Risk, Targeting, and Revenue because various end-users will be interested in different research areas. For instance, an elected official such as a governor would likely be most interested in how restaurant compliance violations affect revenue while the operations manager of NYC DOHMH might be more interested in the logistics of deploying health inspectors more efficiently. These themes are evident throughout our entire project whitepaper, presentation, and dashboard.

Data Wrangling and Exploration

Exploration

The team explored the New York City Restaurant inspection results provided by the Department of Health and Mental Hygiene through the NYC OpenData website.[1] The data set contains restaurant results from 2016 and beyond. The dataset contains every violation from each restaurant organized by the CAMIS number which is a unique record identification for each restaurant. The new restaurants which have not

[1] The team extracted the information towards the end of January in 2022. https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

been inspected yet were displayed with an inspection date of 1/1/1900. Any establishments that are no longer in business will not be displayed.

The dataset contained almost 400,000 rows of data from 2016 to 2021. The team wanted to focus on the results of the last 5 years and started our data exploration from 2017 and beyond. From an initial exploration of the data, the team saw a drastic reduction on the total number of inspections and restaurants inspected from 2019 to 2020 (Graphic 2). The number of restaurants inspected doubled from 2017 to 2019. This was an indicator that the restaurant business was blooming. But in 2020, the number of restaurants inspected dropped from 22,000 to a little over 8,000. This was a 63% reduction from 2019 to 2020 in the number of restaurants inspected. At the same time, the number of inspections dropped by 78% from 2019 to 2020. This drastic reduction in the number of restaurants inspected and the number of inspections was attributed to the COVID-19 global pandemic. The team decided to only focus only on the data in 2020 and 2021.

Graphic 2: The total number of restaurants inspected (left) and the total number of inspections (right) from 2017 to 2021

Data source: Inspection Results from NYC DOHMH

Wrangling

After narrowing down to the data in 2020 and 2021, we had approximately 1,400 individual restaurants or 15% of data for modeling. The team excluded the duplicate records in the data; filled in zeros for the

records without zip code values; applied the mean value of the inspection scores to the restaurants without an inspection score in 2021 to preserve the quantity of the data.

The team proceeded to join the restaurant results with other open data sources in an attempt to gain better insights on predicting the risks of restaurants. We used the following sets of data:

- NYC total population per zip code in 2019[2] – this data source provided the NYC population by zip code
- NYC median income per zip code in 2019[3] – this data source provided the NYC median income by zip code
- Fine by violation code[4] - the team extracted only the violation codes that were classified by the NYC DOHMH as critical violations

Additional data transformation was applied to the data after joining with other datasets. For example, some records of the original dataset had missing values of the zip code. This would cause any dataset that requires a join by zip code to not populate. This would include the total population and median income datasets. The team would populate the missing values with the mean of the associated dataset.

The team created a master dataset for analysis, modeling, and visualization purposes. The master dataset included the following parameters:

Table 1: Parameters derived for analysis, modeling, and visualizations

Parameters	Description	Classification Models	Clustering	Revenue
id	Unique identifier			
CAMIS	Unique identifier of restaurant			
name	Business name			
is_closed	Operation status of restaurant			
review_count	Number of reviews in Yelp	Yes		
rating		Yes		
transactions	Type of food service			
Price	Price per meal per person	Yes		
DBA	Name of restaurant			
BORO	Borough of restaurant	Yes		
ZIPCODE	Zip code of restaurant			

[2] NYC Total Population 2019 https://data.cccnewyork.org/data/map/97/total-population#84/17/3/130/62/502/4
[3] NYC Median income 2019
https://data.cccnewyork.org/data/map/1396/hard-to-count-households#1396/a/3/1656/40/a/4
[4] NYC DOHMH Violation Codes and Penalty Amounts
https://codelibrary.amlegal.com/codes/newyorkcity/latest/NYCrules/0-0-0-43593

PHONE	Phone number of restaurant			
CUISINE DESCRIPTION	Original cuisine description of restaurant			
Latitude	Latitude of restaurant location			
Longitude	Longitude of restaurant location			
MEDIAN INCOME	Median income of NYC per zip code in 2019	Yes		
TOTAL POPULATION	Total population of NYC per zip code in 2019	Yes		
INSP FREQ	Count of inspection frequency	Yes		
Crit FREQ	Count of violation code frequency	Yes		
MEDIAN SCORE	Median inspection score based on the inspection	Yes		
LATEST SCORE	Latest inspection score based on the latest inspection date			
AVG. SCORE	Average inspection scores			
BINNED_Avg.CritFine	Average fine amount into 4 equally sized bins	Yes		
TOTAL CRITICAL FINE	Total fine amount by critical violation code			
Avg.CritFine	average fine amount by critical violation code			
LATEST INSPECTION DATE	Maximum of inspection date			
LATEST INSPECTION YEAR	Maximum year of inspection year			
text_join	Joining text			
rating_min	Minimum of Yelp rating based on scale of 1-5			
rating_max	Maximum of Yelp rating based on scale of 1-5			
rating_mean	Mean of Yelp rating based on scale of 1-5	Yes		
Reviews_Flag	Flag of review			
Reviews_Flag_Count	The count of flag words in review data	Yes		

Revenue_Insp_Year	Estimated revenue per restaurant based on inspection year			
Paid_Wages	Wages paid			
Cuisine Category	Classified original cuisines into 13 new categories			

Assumptions

Although each record represents a violation, there was no way for the team to distinguish the actual citation per violation code or the penalty amount for each violation. There is a one-to-many relationship between the violation code and the citation. However, the inspection results don't distinguish the citations associated with each violation code. The team decided to remove the duplicate records within the dataset. Moreover, the dataset did not provide the violation penalty amount. The team had to assume that the maximum violation penalty was issued for each violation record.

Modeling

Classification Model Approach to Predict Risks

The team approached the risk scores in two different ways: classification and clustering. To apply classification models, the team used the latest inspection score of each restaurant as the predictive value and the parameters listed as "Yes" under the Classification Models as predictors in Table 1. In defining the risk score of each restaurant, the team followed a similar ranking approach as the NYC DOHMH. The team created an artificial class to accommodate the restaurants with extremely high latest inspection scores. Specifically, the team defined the records with the latest inspection score between 0 – 13 as Grade A; 14 – 27 as Grade B; 28 – 50 as Grade C; and values beyond 50 are Grade D. Grade A and B would be the restaurants with low risk. Grade C and D would be the restaurants with high risk.

The team had split the data as 76.6% training and 33.3% testing and performed 6 different classification models. These 7 classifications models included the following[5]:

- Naive Bayes - a family of classifiers based on a simple Bayesian model that explores the relationship between probability and possibility.
- Logistic Regression – a model with an input variable (x) and an output variable (y), which is a discrete value of either 1 (yes) or 0 (no).
- Decision Tree – the Decision Tree classifier has dataset attributes classed as nodes or branches in a tree.
- K-Nearest Neighbors – a simple classification algorithm, where K refers to the square root of the number of training records.

[5] Definition of the first 5 models from ActiveState.com
https://www.activestate.com/resources/quick-reads/how-to-classify-data-in-python/

- Support Vector Machine – a model with associated learning algorithms that analyze data for classification. Also known as Support-Vector Networks.
- Extreme Gradient Boosting[6] - is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning technique.

Classification Model Performance

Each classification model was evaluated based on the accuracy and the AUC (Area under the Curve)-ROC (Receiver Operating Characteristics). The accuracy is measured by the number of true positives (number of correct predictions that the occurrence is positive) plus true negative (number of correct predictions that the occurrence is negative) divided by the total population. Accuracy is a ratio of correctly predicted observation to the total observations. "The AUC-ROC is a performance measurement for classification problems. The ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting each class accurately".[7]

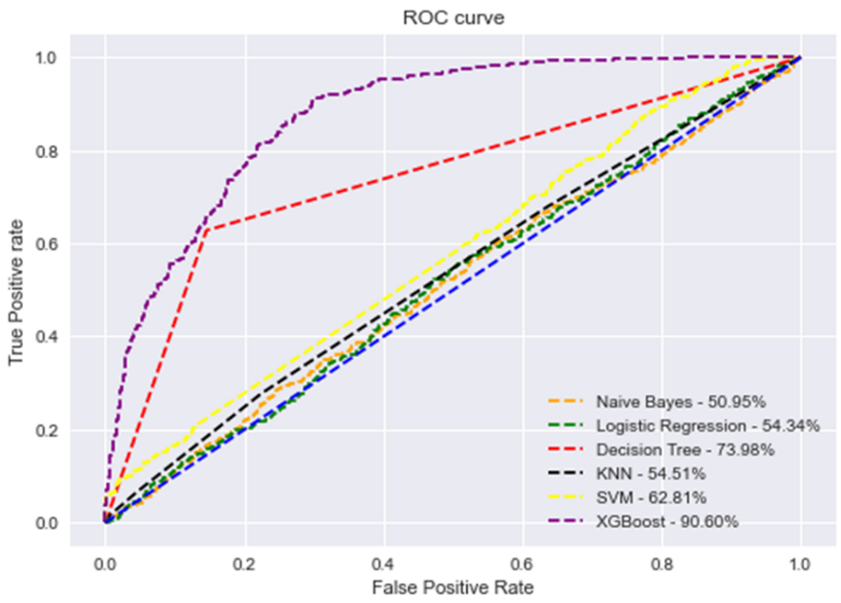
After modeling the data, we found that the Extreme Gradient Boosting had the highest accuracy measure of accuracy and AUC-ROC compared to the other 5 listed models. The Extreme Gradient Boosting model outperformed the other 5 models in both measures; the model had an Accuracy of 79.45% and the AUC-ROC of 90.98%. Reference Appendix I for details of each model performance.

Table 2: Accuracy and AUC-ROC measure results for Classification Models

Model #	**Classification Models**	**Accuracy**	**AUC-ROC**
1	Naive Bayes	65.45%	50.94%
2	Logistic Regression	65.45%	54.34%
3	Decision Tree	76.00%	73.19%
4	K-Nearest Neighbors	55.19%	54.51%
5	Support Vector Machine	67.30%	61.66%
6	Extreme Gradient Boosting	79.45%	90.98%

Graphic 3: Visual of AUC-ROC for the first 6 Classification Models

[6] Definition of Extreme Gradient Boosting: https://www.nvidia.com/en-us/glossary/data-science/xgboost/
[7] Definition of AUC-ROC: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5



Findings

Revenue Impact Findings

Total monetary loss to NYC due to inspection violation in 2020 is about $180 million and in 2021, is about $134 million. Eighty-three percent of the total loss is attributable to unemployment benefits paid out to newly unemployed workers for six months. The remaining 17 percent represents the taxable revenue lost.

On average, health inspection violations that result in restaurant closure unemploys about 3,626 NYC residents per year and decreases total tax revenue by about $87,638 per closed restaurant.

A reasonable conclusion is that health inspection violations' most significant impact on NYC's monetary bottom-line is not directly attributable to tax revenue loss, but instead the cost of supporting the households of citizens who lost their jobs due restaurant closures.

Methodology

	2017	2018	2019	2020	2021
Total NYC Tax Revenue	$54,141,527,000	$56,883,176,000	$60,298,723,500	$62,291,586,000	$64,311,769,500
Total NYC Restaurant Revenue	$24,310,000,000	$25,230,000,000	$26,120,000,000	$14,730,000,000	$17,640,000,000
NYC Tax Rate	8.875%	8.875%	8.875%	8.875%	8.875%

NYC Taxable Revenue	$2,157,512,500	$2,239,162,500	$2,318,150,000	$1,307,287,500	$1,565,550,000
Number of NYC Restaurants	26,697	27,043	23,650	13,337	15,972
Taxable Revenue per NYC Restaurant	$80,815	$82,800	$98,019	$87,211	$89,343
NYC Restaurant Jobs	317,800	317,800	317,800	179,219	214,625
NYC Restaurant Wages Paid	$9,931,900,591	$10,322,552,517	$10,700,000,000	$10,318,151,036	$10,446,901,184
NYC Taxable Revenue Lost				$30,325,875	$22,578,000
Unemployed NYC Restaurant Staff				4,157	3,095
Unemployment Benefits Paid				$149,667,740	$111,429,538
Total Monetary Loss to NYC				**$179,993,615**	**$134,007,538**
Total Monetary Loss to NYC (w/ Multiplier Effect (x2.61)				**$420,958,676**	**$313,409,093**

Figure 2: Monetary Loss Breakdown

The Total NYC Tax Revenue and Total NYC Restaurant Revenue are from the reports issued by the NYC controller's office such as NYC's annual comprehensive financial statement. The NYC tax rate is 8.875%. The NYC Taxable Restaurant Revenue is calculated by multiplying NYC Tax Rate with Total NYC Restaurant Revenue The Number of NYC Restaurants from 2016-18 are from Statista. Taxable Revenue per NYC Restaurant for 2020 and 2021 are moving average from the previous three years. The NYC Restaurant Jobs are from the restaurant association and NYC Restaurant Wages Paid are from the report from the Controller's office. NYC Taxable Revenue Lost due to inspection violation related restaurant closure is assumed to be 3% for 2020-21, which is based on previous years average restaurant closure due to inspection violation. The Unemployed NYC Restaurant Staff due to Inspection Violation is assumed to be 3% of total job loss for both 2020 and 2021. The Unemployment Benefit paid per annum to an individual with a family of four is $72,000 per annum. The Unemployment Benefit is paid for half a year. The Total Monetary Loss to NYC is a summation of NYC Taxable Restaurant Revenue Lost and Unemployment Benefits Paid. Further, the total monetary loss to NYC with a job loss multiplier is 2.61x. This effectively means that if say 100 jobs are lost in the restaurant industry then 161 additional jobs are lost in other supporting industries. Hence, the total job loss would need to be multiplied by 2.61x.

Risk Score Findings

The following findings will be based on the Extreme Gradient Boosting (XGBT) model as it had one of the best model performances compared to other classification models studied. The table below shows the count of the actual Grade in the dataset and the predicted value of each class by XGBT. The model predicts that 68% of the restaurants fell into the rank of Grade A, 23% in Grade B, 8% in Grade C, and 1% in Grade D.

Table 3: Actual vs Predicted values by XGBT

Classes	**Actual**	**Predicted**

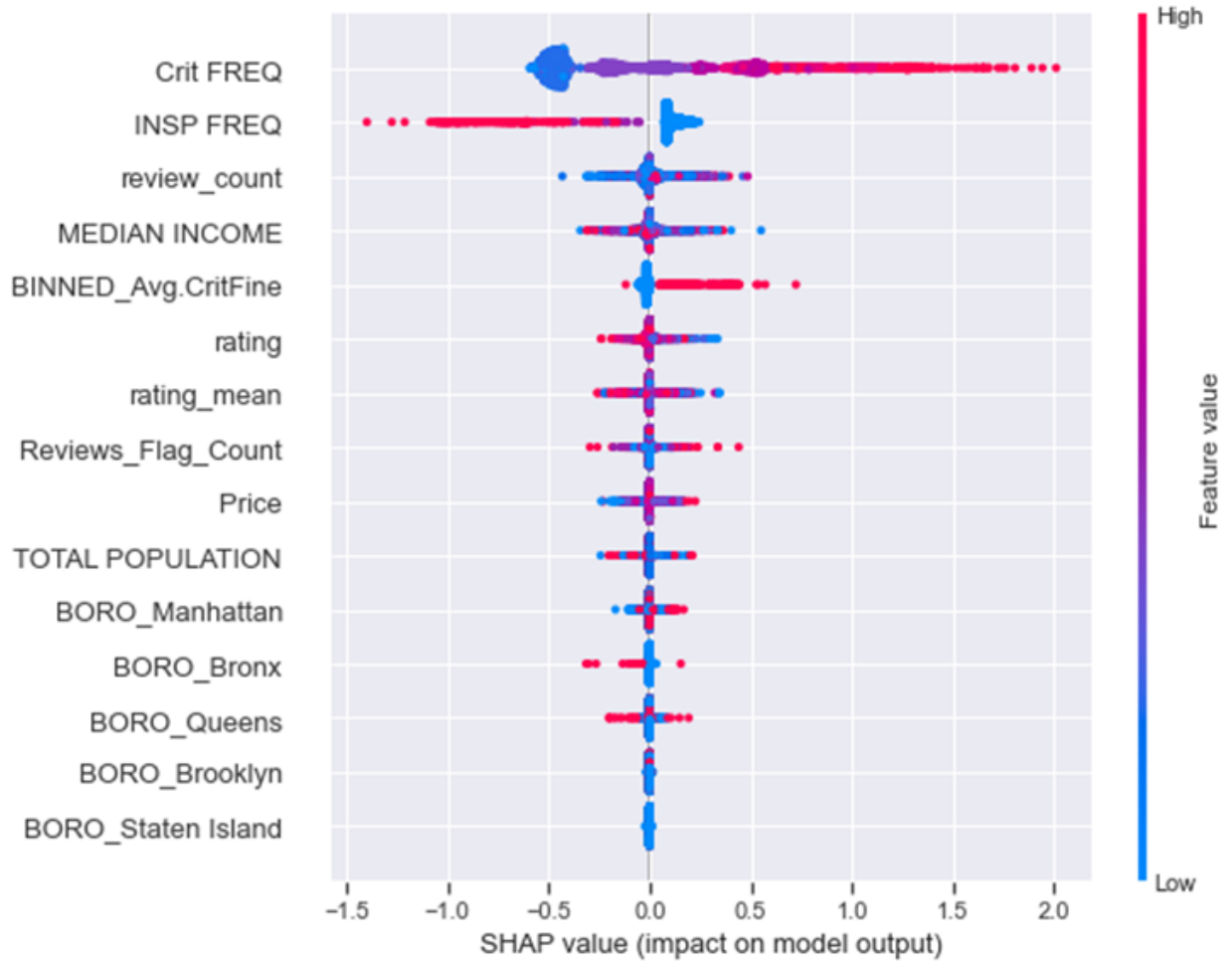
A	1519	1589
B	559	525
C	220	191
D	23	16
Total	2321	2321

To help explain the results derived from XGBT, we integrated the model results with Shapley Values. Shapley values is a method from coalitional game theory where a prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. The Shapley values will tell us how much influence is attributed to each feature and how each feature value contributed to the prediction compared to the average prediction. [8]

The graphic below shows the overall influence and relationship of the overall model for each parameter. The number of critical frequencies is the most influential predictor followed by the number of inspection frequencies. The rest of the predictors had little impact compared to the first two predictors. The color of the bars indicated the relationship between the predictor and the target value. For instance, the blue in critical frequency tells us that the lower the critical frequency, the more it impacts the predicted value to be lower. This is a positive relationship between the predictor and the expected outcome. On the other hand, the inspection frequency has a negative relationship with the expected value where the higher inspection frequency drives the expected value to be lower.

Graphic 4: Variable importance shown by SHAP values

[8] Explanation of Shapley Values: https://christophm.github.io/interpretable-ml-book/shapley.html



To visualize the impact of each predictor locally, we selected a single record out of the test set to further demonstrate the impact of our predictors per record.

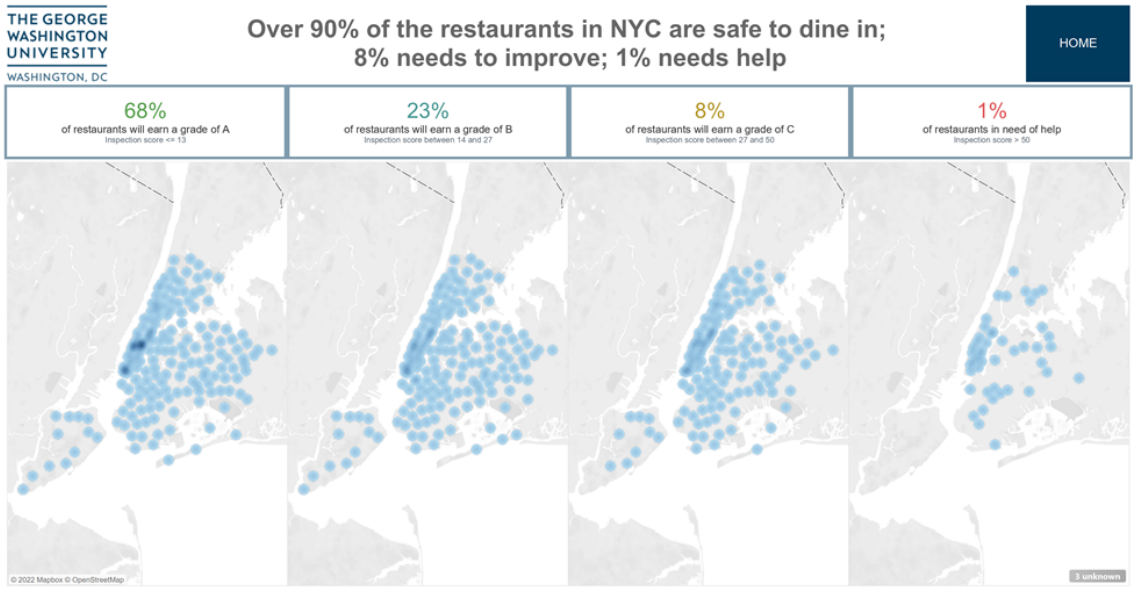
Graphic 5: Localized Predictor Influence based on Shapley Values

- The output value is the prediction for that observation (the prediction for this observation is 0 or A).
- The base value is the expected value if we don't have any predictors and only have the dependent variable. In other words, the base value is the mean of the expected value of y. The expected value of y is 0.4.
- The red and blue bars represent the impact of each variable that pushes the prediction higher are shown in red, and those pushing the prediction lower are in blue.

- Critical Violation Frequency has a negative relationship with the inspection score. In our case, we picked the 1700 observations in our data. The value of Crit FREQ of 2 is less than the average value of 2.17; therefore, it pushes the inspection score to a towards the left or a lower inspection score. As a reminder, the higher the inspection score, the higher the risk. Although this is a negative relationship, it has a positive effect on inspection scores.
- Inspection Freq has a positive relationship with the inspection score. The value 1 is lower than the average and pushes the expected value towards the right.

In conclusion, the overwhelming majority of the restaurants in New York will receive a Grade B or Grade A and less than 10% of the restaurants would fall under the high-risk category. The Staten Island restaurants do not have any restaurants that fell in the predicted Grade D category. The graphic below visualizes our conclusion and provides a density map of the restaurants per grade rank. The darker color on the map indicates a higher density of the restaurants of its respective grade per zip code.

Graphic 6: Density map by zip code with respective grades by percentage

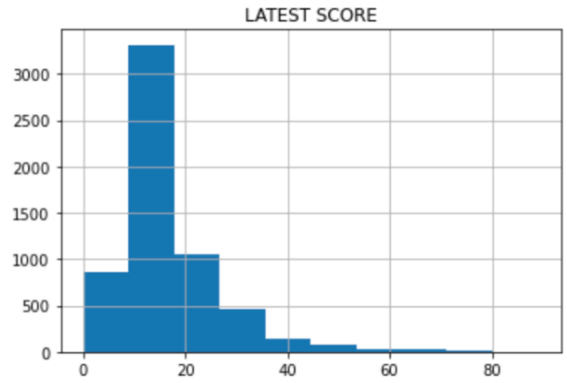


Inspection Targeting Findings

Inspection analysis was performed to help inspectors improve process efficiency and assist restaurants get better inspection grades.

Latest risk scores generated by the classification model are determined to be the risk scores for restaurants. Inspection dataset combined general inspection information, latest score (risk score) and Yelp reviews. While most restaurants are at low risk according to the risk score distribution graph below, the analysis only focuses on restaurants with grade B and C, meaning risk scores greater than 14.

Risk Score Distribution Graph



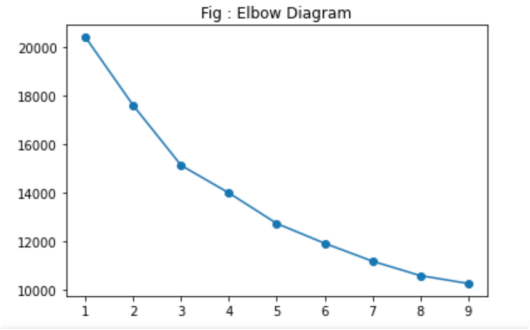
- "A" grade: 0 to 13 points for sanitary violations
- "B" grade: 14 to 27 points for sanitary violations
- "C" grade: 28 or more points for sanitary violations

Methodology

Restaurant segmentation forms a basis for the restaurant inspection process. Segmentation is often informed by clustering algorithms that attempt to group similar restaurants according to their types, location and other attributes. It allows inspectors to understand what are the factors to differentiate restaurants received Grade A and restaurants received B or below.

K-prototype and K-means are two unsupervised learning algorithms, as K-mean only works with numeric data, categorical features such as Boroughs, Restaurant Cuisine are converted to dummy variables. By comparing model errors of K-prototype and K-means method, K-prototype has a better performance as mixed data types are included in the inspection dataset.

The Elbow method was applied to determine 3 clusters are equally distributed with clear boundaries.



Recommendations

Recommendations for inspectors

Clustering model summary table provides insights that cluster 0 has the highest mean of risk score, the difference between cluster 0 among cluster 1, cluster 2 can be considered to help inspectors optimize the order of the restaurant inspection list.

According to the *cluster summary table*, *restaurant Boroughs ranking table* and *cuisine types ranking table*, restaurants in cluster 0 has the following features,

- Few yelp reviews
- Low yelp rating
- Low yelp price
- Restaurant located areas that have low median income
- Restaurant located areas that have large population
- 76% of restaurants are in Brooklyn and Queens
- 58% of restaurants are Asian, European and American cuisines

Cluster summary table:

	clusterid	CAMIS	review_count	rating	Price	ZIPCODE	MEDIAN INCOME	TOTAL POPULATION	INSP FREQ	Crit FREQ	LATEST SCORE	Reviews_Flag
		mean	mean	mean	mean	mean	mean	mean	mean	mean	mean	mean
0	0	4.671817e+07	110.345009	3.672067	1.195271	11022.069177	63541.929947	2.213801e+07	1.007005	3.042469	26.355517	0.458844
1	1	4.593820e+07	421.023370	3.738622	1.814268	10124.987700	114814.766298	1.564949e+07	1.007380	2.990160	24.793358	0.456335
2	2	4.797441e+07	165.971963	3.682243	1.196262	10818.504673	76169.116822	2.093592e+07	2.049065	6.324766	25.485981	0.509346

Cluster 0 restaurants location

	BORO	percentage
0	Bronx	0.100701
2	Manhattan	0.133975
3	Queens	0.363398
1	Brooklyn	0.401926

Cluster 0 restaurants Cuisine Category:

	Cuisine Category	percentage
3	Australian	0.000876
5	Eastern European	0.007881
0	African	0.008757
7	Mediterranean	0.008757
9	Other	0.013135
4	Caribbean	0.043783
8	Middle Eastern	0.045534
11	Sweets	0.056918
12	Universal	0.095447
10	South American	0.131349
1	American	0.160245
6	European	0.169002
2	Asian	0.258319

Recommendations for restaurants

High risk restaurants cluster 0 provide insights that restaurants have low yelp ratings. Performing Natural Language Processing on restaurants that have yelp rating less than 3 to get most frequent negative words from reviews would help restaurant owners getting better inspection grades.

According to the below *Negative words word cloud,* most words from yelp reviews relate to the dining environment. Hence, **increasing the cleaness of the dining environment** would help restaurants get good inspection grades.

Negative words word cloud

Areas Worth Future Research

One area worth future research would be to improve the model performance of the Extreme Gradient Boosting. We noticed that the model performed poorly on predicting Grade D results. Specifically, the accuracy for Grade D results were only 35%. This may be caused by the imbalanced data in the dataset itself. Applying scaling weights to each class and applying hyperparameter tuning could help improve the performance of the model.

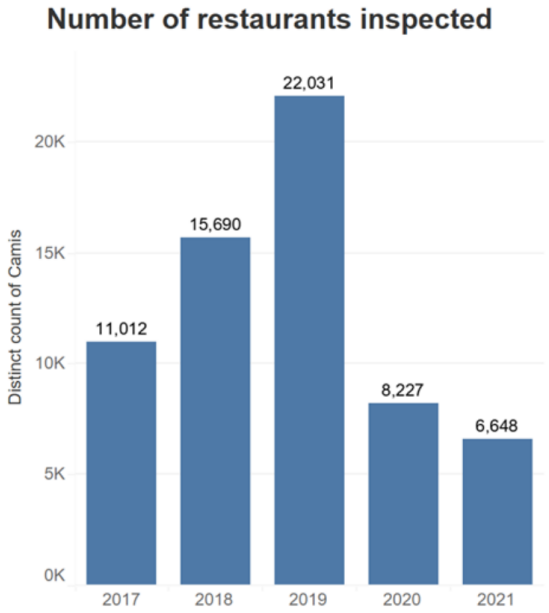
Another area of future consideration would be to include additional data into the analysis. The team only focused on analyzing the information after the global pandemic happened. It would be interesting to study the Grading differences between pre-COVID and post-COVID years. The team has the hypothesis that more restaurants were able to improve their overall grade or pass the inspection with the additional dining restrictions put in place due to the pandemic.

The following is the extension of the main modeling that predicts the latest inspection scores based on the 4 years of data using Explainable Boosting Machine techniques.

Data

Data used

4 years of the inspection results data from 2018 to 2021 was used to model as a part of our future recommendation. The team anticipates that an extended period of inspection records from 2018 to 2021 can provide a more reliable prediction since more than 2.7 times the restaurants were inspected in 2019 than in 2020. Data from 2018 to 2019 contains the restaurant inspection results pre-pandemic that can better represent the inspection performance of the entire restaurants in NYC and also can offer predictions for more restaurants.



Data source: Inspection Results from NYC DOHMH

The Yelp/ Google review data was initially included in the main modeling but excluded in this step due to the overfitting issues with the Explainable Boosting Machine method. In addition to the original dataset of the NYC restaurant inspection results, 15 additional preliminary classifiers are created which are the following: 'INSPECTION YEAR', 'LATEST SCORE', 'AVG. SCORE', 'MEDIAN SCORE', 'INSP FREQ', 'CRITICAL FREQ', 'FINE AMOUNT', 'TOTAL FINE AMOUNT', 'AVG. FINE AMOUNT', 'MEDIAN INCOME', 'TOTAL POPULATION', 'Crit FREQ', 'BINNED_Avg.CritFine', 'TOTAL CRITICAL FINE', and 'Avg.CritFine'.

Data cleaning

The data cleaning process was begun by eliminating the columns that wouldn't be used such as 'Council District', 'Census Tract', 'BIN', and 'BBL'. Then, highly correlated columns with a correlation value above 0.80 and the duplicated rows were removed. To handle missing values in the dataset, 2 types of imputation methods were applied which are replacing with 0 and the average of the particular column values. The missing values in the columns 'CRITICAL FREQ', 'FINE AMOUNT', 'AVG. FINE AMOUNT', and 'Avg.CritFine' were replaced with 0 because these restaurants performed well in the past inspections, and therefore, didn't get the critical flag marks or fined. On the contrary, the missing values in the columns 'MEDIAN INCOME' and 'TOTAL POPULATION' were replaced with the average values to stabilize the distribution of data. In the next data cleaning step, data types of columns were adjusted to numeric from categorical. Detailed explanations for each step are available in the Python notebook file for reference.

Modeling: Explainable Boosting Machine

EBM was used to model because of the interpretability of the output of the model.

Variables for modeling

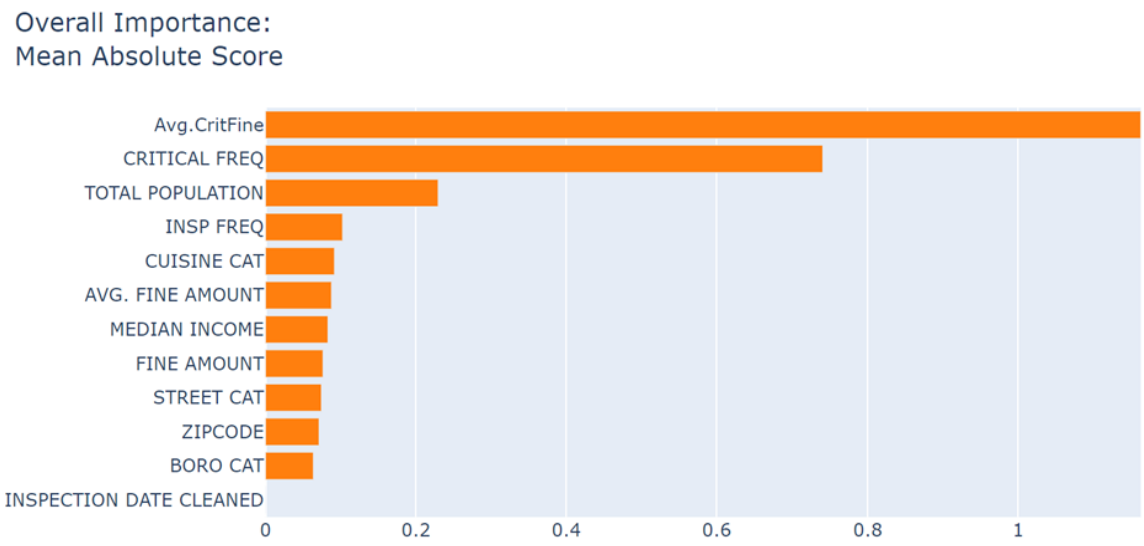
After data cleaning, 11 classifiers are used for modeling including 'ZIPCODE', 'INSP FREQ', 'CRITICAL FREQ', 'FINE AMOUNT', 'AVG. FINE AMOUNT', 'MEDIAN INCOME', 'TOTAL POPULATION', 'Avg.CritFine', 'BORO CAT', 'STREET CAT', and 'CUISINE CAT' to predict the target 'TARGET CAT' which is the discretized variable of the latest inspection score per restaurant. The subcategory of 'TARGET CAT' rages from 0 to 3. (0 = Grade A, 1 = Grade B, 2 = Grade C, and 3 = Grade D)

Fitting data

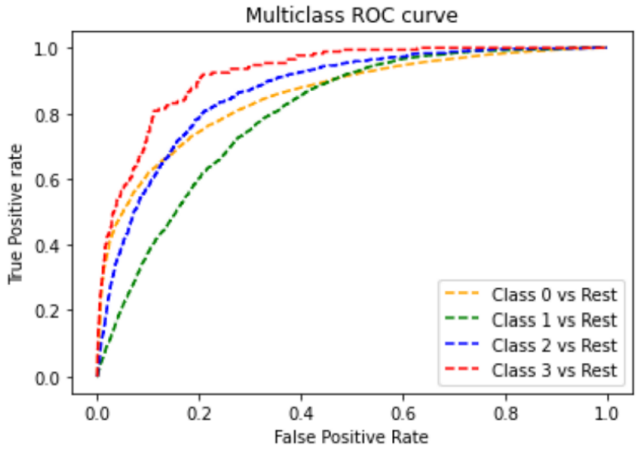
The data was sorted by the inspection date and the team used the furthest past data as the train data, the closest recent data as the test data, and in between as the validation data in the ratio of 70:15:15.

Output of the model

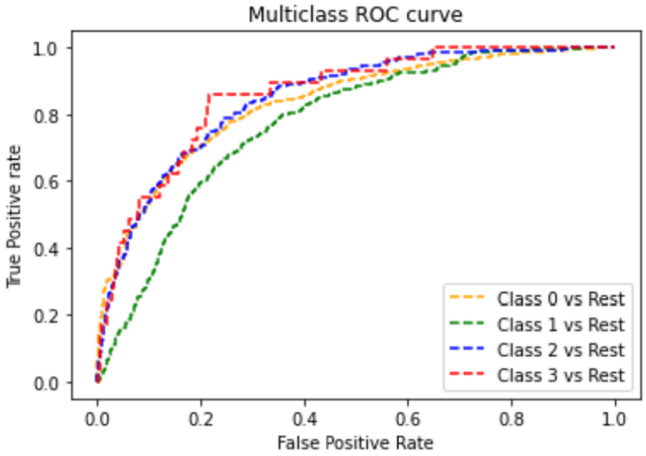
The model ranks the importance of the classifiers by the most unique predictive power in predicting the target 'TARGET CAT'. According to the result of the model, 'Avg.CritFine' and 'CRITICAL FREQ' are the top 2 important variables that offer the best prediction to the model.



Train data performance



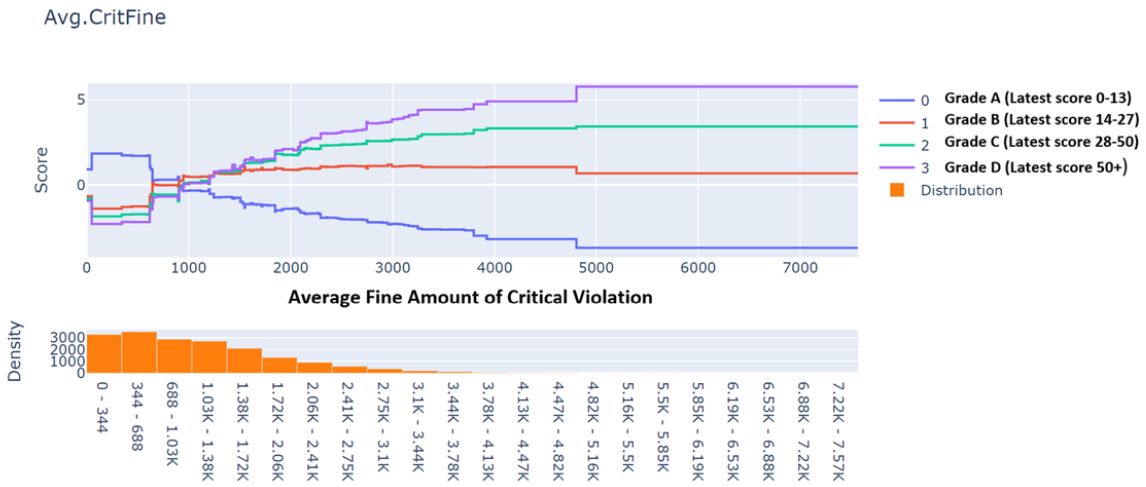
Test data performance



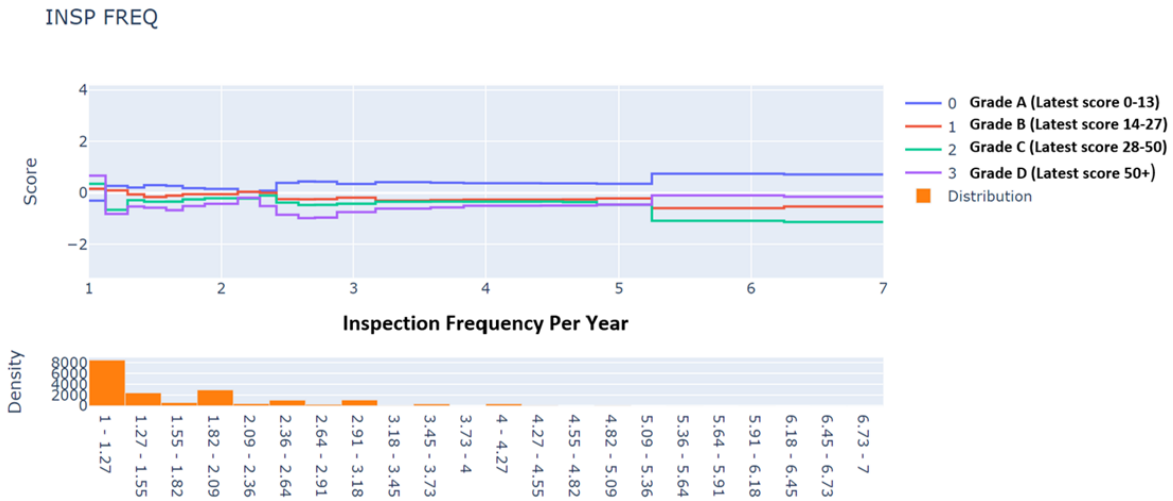
Interpretation

'Avg.CritFine' is the average fine amount of the critical violation codes per inspection. Critical violations are those most likely to contribute to foodborne illness.

The graph below shows that ‘LATEST SCORE’ and 'Avg.CritFine' are in a positive relationship. As the amount of 'Avg.CritFine' increases, the probability of getting higher inspection scores also increases.



'INSP FREQ' is the average number of inspections per restaurant from 2018 to 2021. The graph below shows that more frequent inspections per year are positively correlated to a greater probability of getting Grade A. Restaurants in this group resulted in a greater number of inspections because they performed poorly in the previous screenings that required them more frequent follow-up inspections. They ended up receiving Grade A in their latest inspections, showing an improvement over time, but inspectors should continue to consider them as potential suspects since they have historical performance patterns that involved very high inspection scores such as 99 and 120.



The team has identified the distribution of grades from the restaurants with the inspection frequency of 5 or more. All the restaurants with the inspection frequency of 6 and 7 have the latest Grade A while the distribution of grades evens out for the restaurants with the inspection frequency of 5.

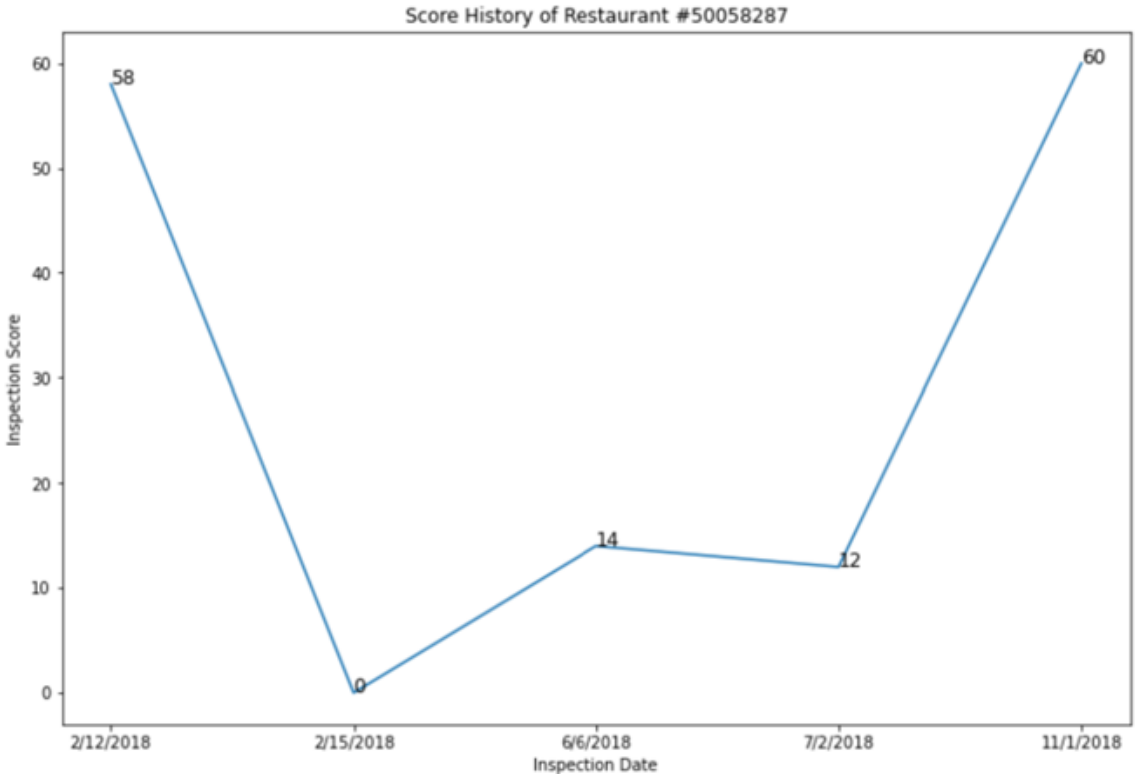
```
There are 2 restaurants that have an annual inspection frequency of 7 from 2018-2021:
A    2
B    0
C    0
D    0
Name: TARGET, dtype: int64


There are 13 restaurants that have an annual inspection frequency of 6 from 2018-2021:
A    13
B     0
C     0
D     0
Name: TARGET, dtype: int64


There are 121 restaurants that have an annual inspection frequency of 5 from 2018-2021:
A    92
B    26
C     2
D     1
Name: TARGET, dtype: int64
```

Restaurants with the combination of a higher inspection frequency and Grade D or C are the worst performers. They consistently violated compliance and didn't even show an improvement over time. Restaurants that fall into this group should be primarily targeted for a more enforced screening.

'CAMIS' is the restaurant ID number used in the dataset. 'CAMIS' #50058287 is the restaurant that has the record in the red box in the picture above and this is an example of a worse performer. This restaurant received inspections 5 times in 2018 and its historical performances were graphed below. The noticeable observations from the 5 inspections are that the 2nd inspection was conducted only 3 days after the previous screening and the inspection result is a score of 0. It's would be helpful to clarify the meaning of the score whether it's a true 0 or the score was recorded incorrectly. If the case is ture, the score indicates Grade A which typically does not require the next inspection in the next 12 months. However, the restaurant was reinspected on 6/6/2018 that is 4 months since the 2nd inspection date. On the 3rd inspection, it received a score of 14. (Grade B) and, for some reason, the subsequent screening was conducted next month on 7/2/2018. Besides, another follow-up inspection was conducted on 11/1/2018 even though the result of the 4th inspection was 12. (Grade A) In order to appropriately interpret the inspection frequency of 2018, further research and internal business insight should be obtained.



Conclusion

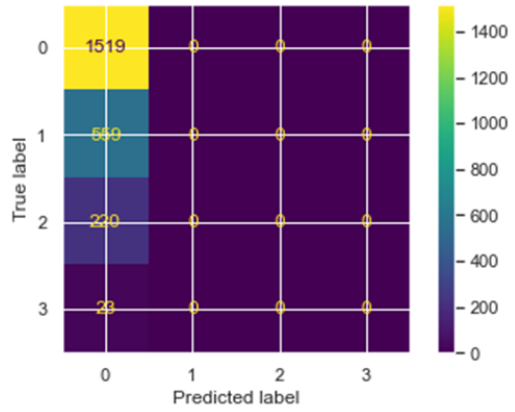
Based on the EBM model, the team could be able to discover risk factors that can lead to high inspection scores. The risk of compliance violation can be anticipated most likely by the higher average fine amount of the critical violation codes per inspection and the average number of inspections per year. In order to rank the riskiness of compliance violations, inspectors should target the restaurants first that have a tendency to have both higher average fine amounts and higher numbers of inspection frequencies with poor grades. The next target group should be either having higher average fine amounts or higher numbers of the annual inspection frequency.

APPENDIX I

Detail Model Results for Classification Models

Naïve Bayes Classification: Accuracy 65.45%; AUC: 50.94%

Confusion Matrix:Naïve Bayes Classification



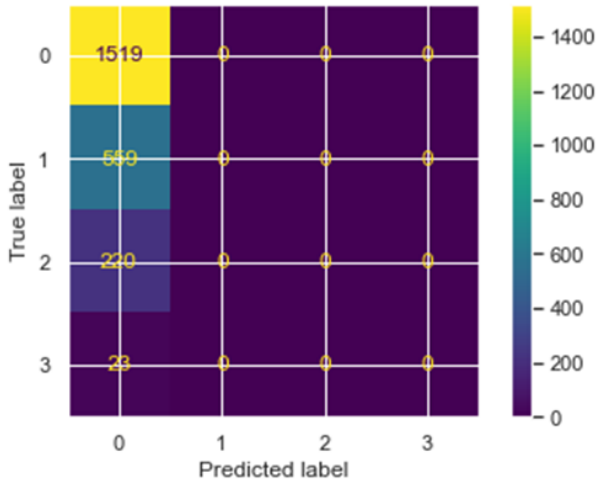
Classification Report: Naïve Bayes

	precision	recall	f1-score	support
0	0.65	1.00	0.79	1519
1	0.00	0.00	0.00	559
2	0.00	0.00	0.00	220
3	0.00	0.00	0.00	23
accuracy			0.65	2321
macro avg	0.16	0.25	0.20	2321
weighted avg	0.43	0.65	0.52	2321

AUC per Class: Naïve Bayes

Logistics Regression Classification: Accuracy 65.45%; AUC: 54.34%

Confusion Matrix: Logistics Regression Classification



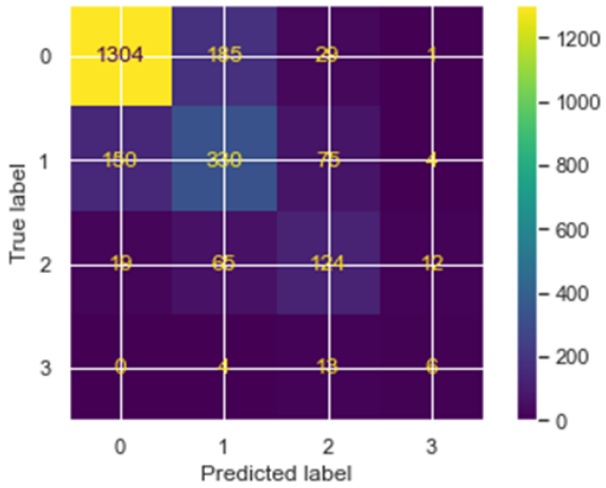
Classification Report: Logistics Regression Classification

	precision	recall	f1-score	support
0	0.65	1.00	0.79	1519
1	0.00	0.00	0.00	559
2	0.00	0.00	0.00	220
3	0.00	0.00	0.00	23
accuracy			0.65	2321
macro avg	0.16	0.25	0.20	2321
weighted avg	0.43	0.65	0.52	2321

AUC per Class: Logistics Regression Classification

Decision Tree Classifier: Accuracy 76.00%; AUC: 73.19%

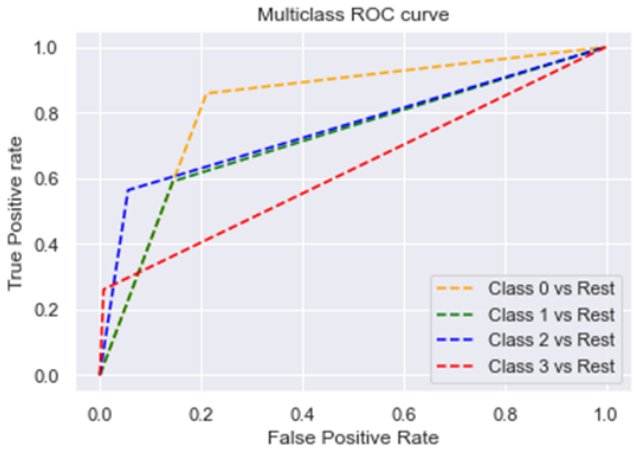
Confusion Matrix: Decision Tree Classifier



Classification Report: Decision Tree Classifier

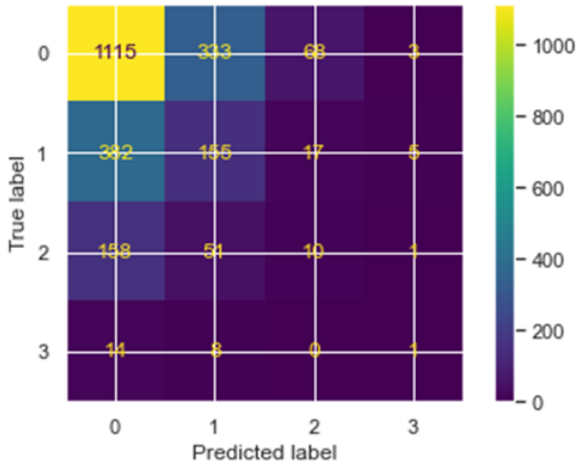
	precision	recall	f1-score	support
0	0.89	0.86	0.87	1519
1	0.57	0.59	0.58	559
2	0.51	0.56	0.54	220
3	0.26	0.26	0.26	23
accuracy			0.76	2321
macro avg	0.56	0.57	0.56	2321
weighted avg	0.77	0.76	0.76	2321

AUC per Class: Decision Tree Classifier



K-Nearest Neighbors Classifier: Accuracy 55.19%; AUC: 54.51%

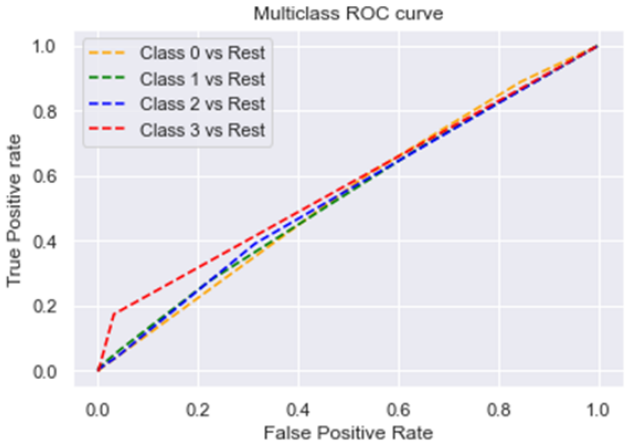
Confusion Matrix: K-Nearest Neighbors Classifier



Classification Report: K-Nearest Neighbors Classifier

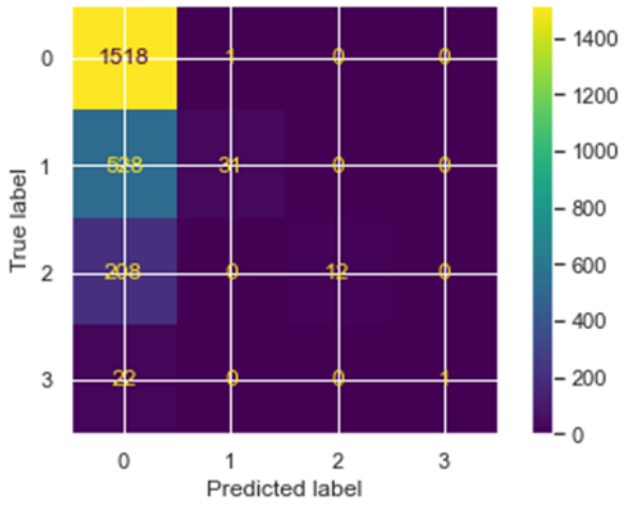
	precision	recall	f1-score	support
0	0.67	0.73	0.70	1519
1	0.28	0.28	0.28	559
2	0.11	0.05	0.06	220
3	0.10	0.04	0.06	23
accuracy			0.55	2321
macro avg	0.29	0.28	0.28	2321
weighted avg	0.52	0.55	0.53	2321

AUC per Class: K-Nearest Neighbors Classifier



Support Vector Machine: Accuracy 67.30%; AUC: 61.66%

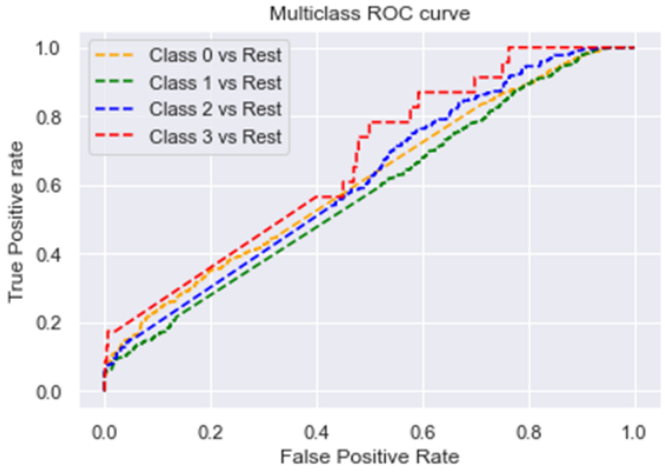
Confusion Matrix: Support Vector Machine



Classification Report: Support Vector Machine

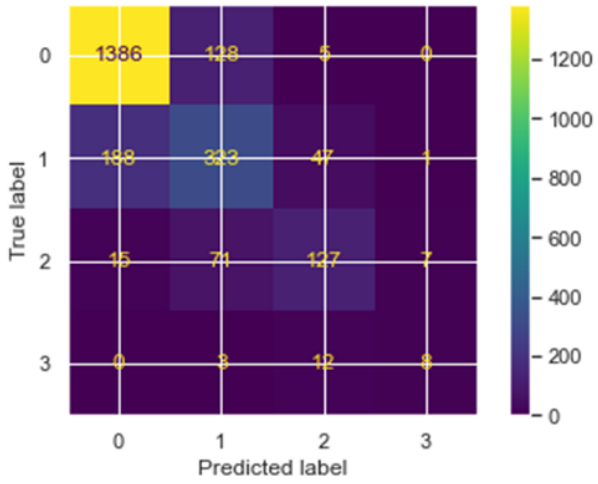
	precision	recall	f1-score	support
0	0.67	1.00	0.80	1519
1	0.97	0.06	0.10	559
2	1.00	0.05	0.10	220
3	1.00	0.04	0.08	23
accuracy			0.67	2321
macro avg	0.91	0.29	0.27	2321
weighted avg	0.77	0.67	0.56	2321

AUC per Class: Support Vector Machine



Extreme Gradient Boosting: Accuracy 79.45%; AUC: 90.98%

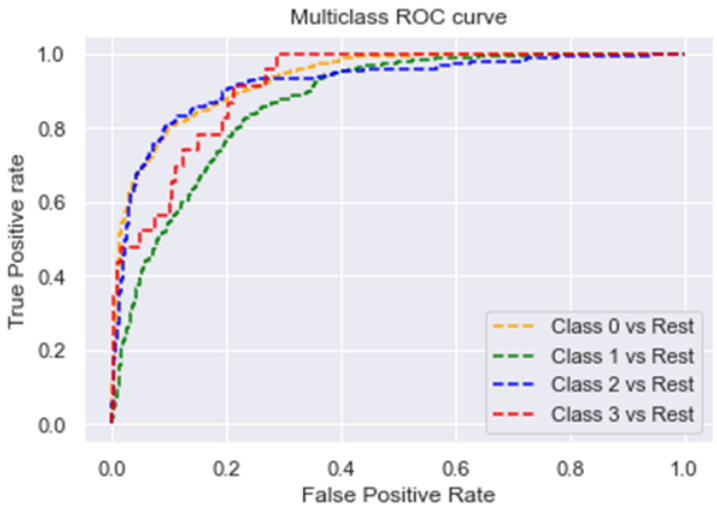
Confusion Matrix: Extreme Gradient Boosting



Classification Report: Extreme Gradient Boosting

	precision	recall	f1-score	support
0	0.87	0.91	0.89	1519
1	0.62	0.58	0.60	559
2	0.66	0.58	0.62	220
3	0.50	0.35	0.41	23
accuracy			0.79	2321
macro avg	0.66	0.60	0.63	2321
weighted avg	0.79	0.79	0.79	2321

AUC per Class: Extreme Gradient Boosting



Clustering Models - K-Prototype and K-Means Model Errors

K-Prototype error

```
# K-prototypes error
cost
```

```
[19295.49999999998,
 16530.022190110474,
 14039.5787279412,
 12680.687331237137,
 11648.18232288015,
 10756.465442917162,
 10153.262743348745,
 9590.917200370537,
 9205.70371411228]
```

K-Mean error

```
# K-means error
cluster_errors_new
```

```
[53312.23236514527,
 26815.136353383205,
 20988.597666050366,
 19487.69001371702,
 18296.972129318172,
 17474.36653729375,
 16727.314276377812,
 15986.801163888214,
 15513.112108207279]
```