# Big Apple BiteSafe

## NYC Restaurant Critical Hygiene Violation Prediction

Karl Liu, Tracy Hu, Isa Wang

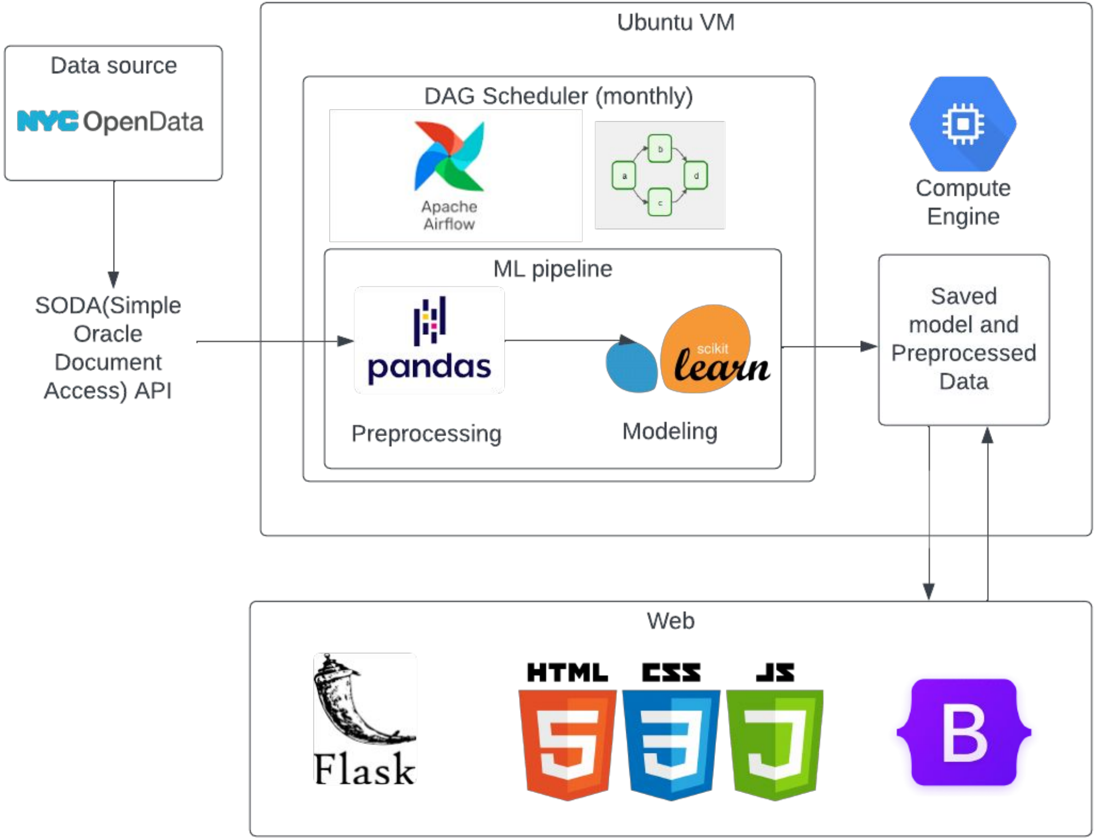
# Overview

- Goal
  - Forecasting the if there will be critical violations at your preferred dining establishment
- Novelty
  - **Untapped Information Source**: No existing channel provides customers with real-time insights into restaurant hygiene, filling a critical information gap for diners.
  - **Emphasis on Health**: Prioritizing restaurant hygiene directly aligns with the growing trend of healthy eating, ensuring meals are not only nutritious but also safe.
  - **Potential Restaurant Advisory**: Beyond informing customers, this platform can serve as an advisory tool for restaurants, guiding them towards maintaining high hygiene standards.
- Challenges
  - **Data Quality and Completeness**: Open datasets can sometimes have missing values, inconsistencies, or inaccuracies.
  - **Predictive Complexity**: The myriad of factors influencing a restaurant's hygiene can make predictions inherently difficult.
  - **Absence of Prior Models or Systems**: Venturing into an area where there are no previous models or systems to draw insights from brings inherent challenges.

# Method

- Methodology
    - Data Extraction
        - Pulling data from  NYC Open Data through API endpoint - looping to circumvent limit of 50,000 records per request
    - Data Analysis
        - Class distribution
        - Confusion matrix
    - Data Processing
        - Feature selection
        - Feature encoding (One-Hot Encoding\Label Encoding)
        - SMOTE-sampling to handle class imbalance
        - Remove outliers & Handle missing data
- Algorithms
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - Gradient Boosting/LGBM
    - XGBoost

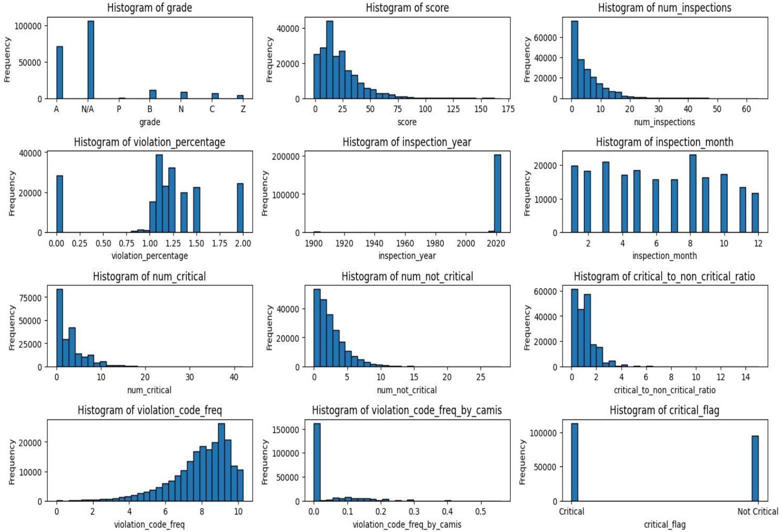
# System


Data source
NYC OpenData
SODA(Simple Oracle Document Access) API
Ubuntu VM
DAG Scheduler (monthly)
Apache Airflow
a
b
c
d
ML pipeline
pandas
Preprocessing
scikit learn
Modeling
Compute Engine
Saved model and Preprocessed Data
Web
Flask
HTML
CSS
JS
5
3
J
B

# Data

- Source
  - NYC Open Data: DOHMH New York City Restaurant Inspection Results
- Volume
  - The dataset comprises a substantial amount of data with 207,000 rows and 27 columns(and even more columns after feature encoding). This large volume indicates comprehensive coverage of restaurant inspections across NYC.
- Velocity
  - The dataset exhibits a high velocity as it is updated daily. This frequent update ensures real-time or near-real-time insights into restaurant hygiene.
- Variety
  - Numerical Columns: Quantitative data such as inspection scores or number of violations.
  - Categorical Columns: Defined categories like violation types, restaurant categories, or boroughs.
  - String Columns: Textual information, potentially including restaurant names, violation descriptions, or addresses.

# Data Processing and Analysis



- Formatting
  - strip(): Removed any leading/trailing whitespace
  - Split & Format inspection date into separate columns
    - ‘Inspection_year’
    - ‘Inspection_month’
    - ‘Inspection_date”
- Processing
  - Filled empty data: Information from related columns / median
  - Apply LabelEncoder to convert all categoricals to numericals
  - Removed outliers for numeric columns
  - SMOTE Sampling:
    - `Critical            113436`
    - `Not Critical         95018`
  - Dataset Split: Latest Record for Testing (11%), Historical Data for Training (89%);
  - Process target variable: ‘critical_flag”

# Feature Engineering - Augmenting

- 10 Augmented Features

- Cumulative Metrics Calculation
  - **‘num_inspections’**: Cumulative total of inspections.
  - **‘cumulative_violations’**: Cumulative sum of violations.
  - **‘critical_to_non_critical_ratio’**
- Violation Analysis
  - **‘violation_percentage’:** Percentage of inspections with any violation.
  - **‘violation_code_freq’**: Frequency of each violation code.
  - **‘violation_code_freq_by_camis’**: Frequency of violation codes by each restaurant.
  - **‘percentage_critical’**: Percentage of critical violations.
  - **‘percentage_not_critical’**: Percentage of non-critical violations.
- Historical Data Utilization
  - **‘prev_critical’**: Critical violations from the previous inspection of this restaurant.
  - **‘prev_violation_code’**: Violation codes from the previous inspection of this restaurant.

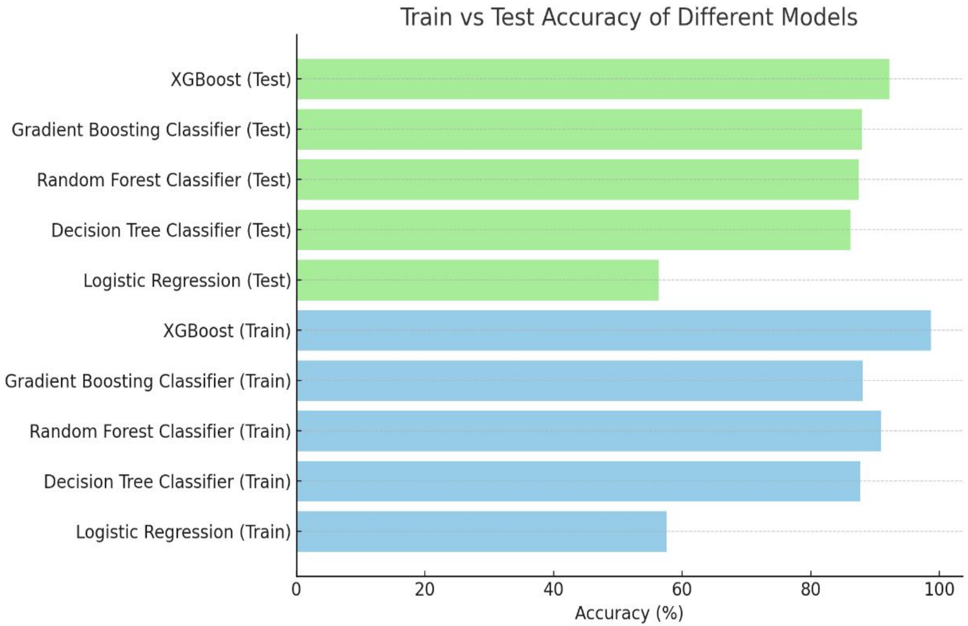
- 7 Original Features
  - **‘Score’**
  - **‘Boro’,**
  - **‘Zipcode’**
  - **‘Dba’,**
  - **‘cuisine _description’**
  - **‘Insepection_year’,**
  - **‘insepection_month’**

- = 17 Total Features
  - [208210 rows x 17 columns]

# Model Predictions

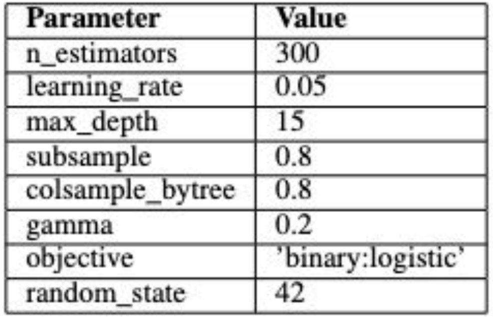
17 features -> target: ‘Critical_flag’ (‘Critical’ / ‘Not Critical’ as 1/ 0)

- Logistic Regression
  - Train Accuracy: `57.6%`
  - Test Accuracy: `56.4%`
- Decision Tree Classifier
  - Train Accuracy: `87.7%`
  - Test Accuracy: `86.2%`
- Random Forest Classifier
  - Train Accuracy: `90.9%`
  - Test Accuracy: `87.5%`
- Gradient Boosting Classifier
  - Train Accuracy: `88.0%`
  - Test Accuracy: `87.9%`
- **XGBoost**
  - **Train Accuracy: `98.6%`**
  - **Test Accuracy: `92.2%`**



# More about XGBoost

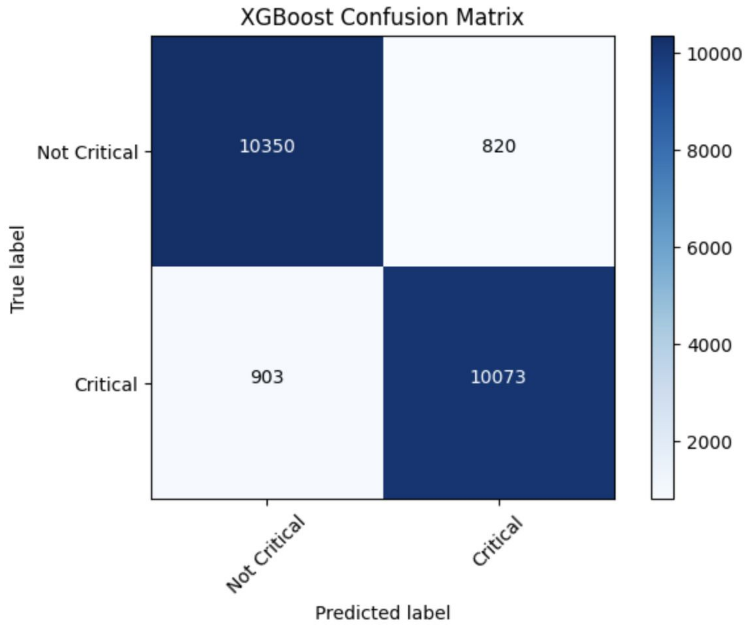
- Hyper-parameter Tuning
- 5-Folds Cross Validation with Grid Search
  - n_estimator: [100, 200, 300]
  - learning_rate: [0.1, 0.05, 0.01]
  - max_depth: [10, 15, 20]
  - subsample: [0.5, 0.8, 1.0]
  - colsample_bytree: [0.3, 0.8, 1.0]



| Parameter | Value |
|---|---|
| n_estimators | 300 |
| learning_rate | 0.05 |
| max_depth | 15 |
| subsample | 0.8 |
| colsample_bytree | 0.8 |
| gamma | 0.2 |
| objective | 'binary:logistic' |
| random_state | 42 |

TABLE 4. XGBoost Hyperparameter values

- Confusion Matrix



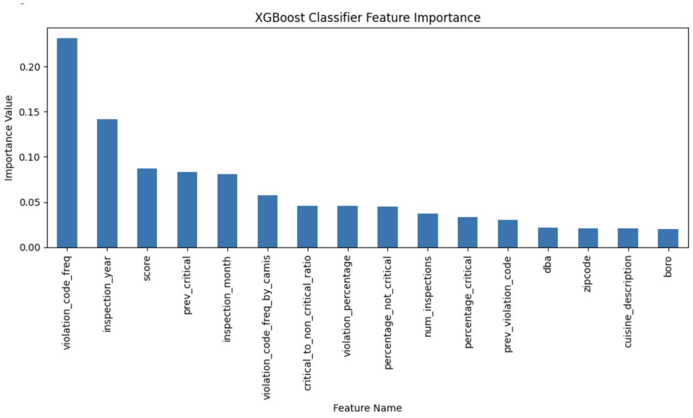
# Insights from feature importance analysis

1. Time-series Patterns: e.g. more critical violations in summers and in year 2021

2. Immediate Behavioral Change after a Critical Violation:

Previous critical -> immediate better performance for next inspection

Previous not critical -> next violation being critical is much higher than being not critical



3. Habitual Non-Compliance:

Extremely high past critical violation rate -> continued poor performance

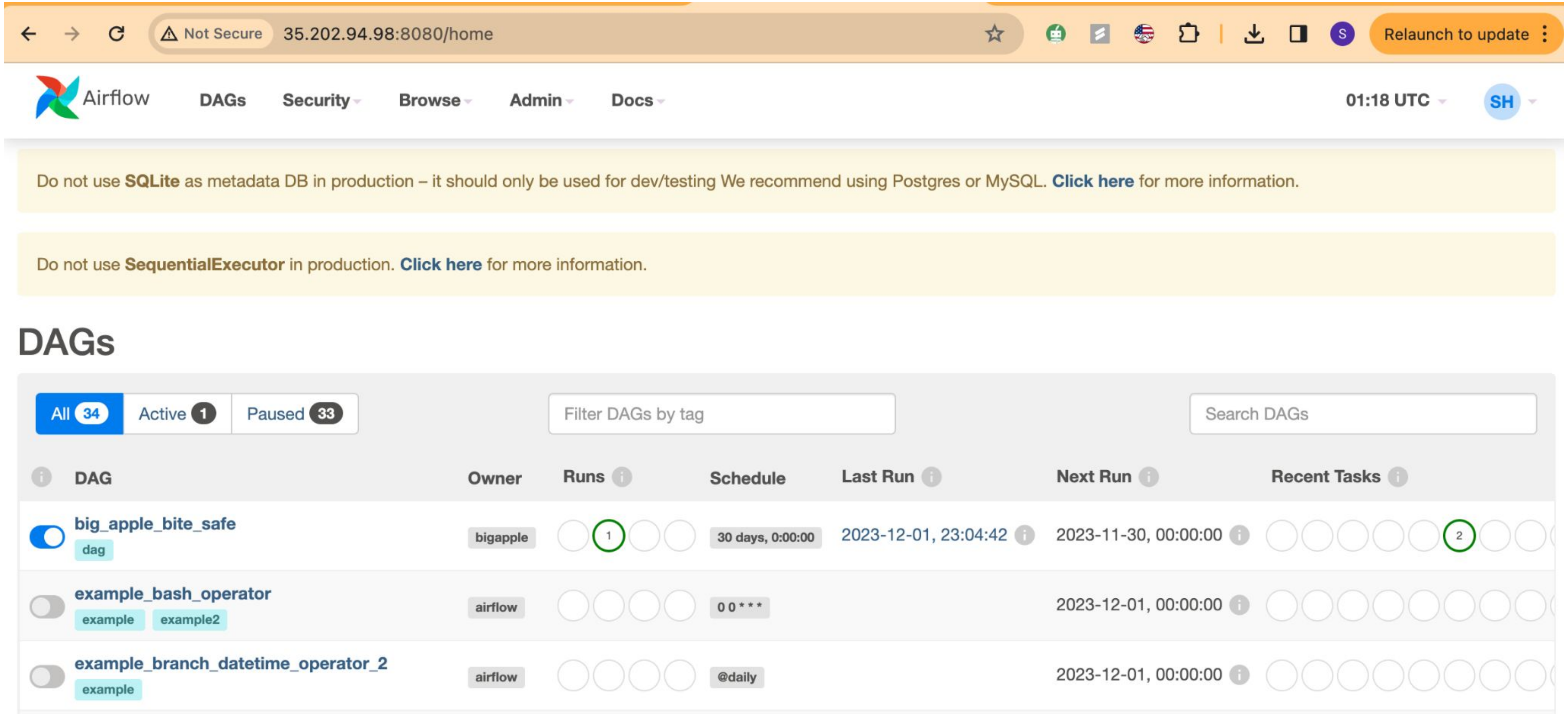
Moderately high past critical violation rate -> better performance

4. Violation Types:

make critical violations on commonly signaled violation types, or violation types rarely made before -> common issues in industry

# GCP and Airflow

Airflow: http://34.133.119.120:8080/     Web Application: http://34.133.119.120:5001/



Do not use **SQLite** as metadata DB in production – it should only be used for dev/testing We recommend using Postgres or MySQL. **Click here** for more information.

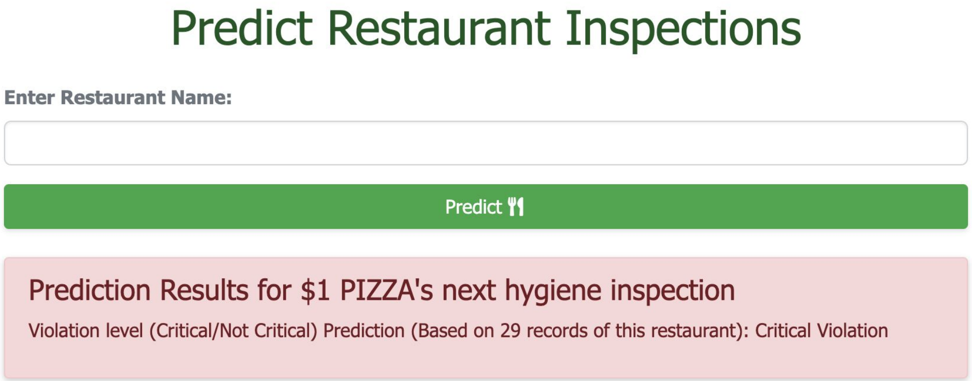
Do not use **SequentialExecutor** in production. **Click here** for more information.

## DAGs

| DAG | Owner | Runs | Schedule | Last Run | Next Run | Recent Tasks |
|---|---|---|---|---|---|---|
| big_apple_bite_safe<br>dag | bigapple | 1 | 30 days, 0:00:00 | 2023-12-01, 23:04:42 | 2023-11-30, 00:00:00 | 2 |
| example_bash_operator<br>example example2 | airflow |  | 0 0 * * * |  | 2023-12-01, 00:00:00 |  |
| example_branch_datetime_operator_2<br>example | airflow |  | @daily |  | 2023-12-01, 00:00:00 |  |

# Web Interface

- Flask
- HTML/JS/CSS
- Bootstrap JS

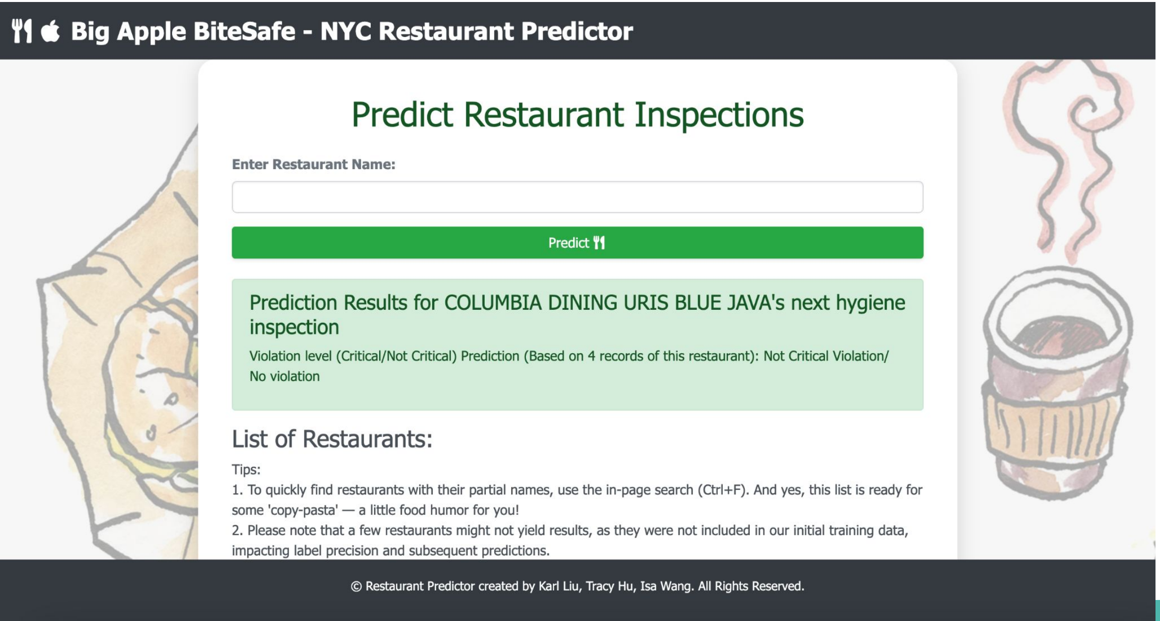


## Predict Restaurant Inspections

Enter Restaurant Name:

### Prediction Results for $1 PIZZA's next hygiene inspection

Violation level (Critical/Not Critical) Prediction (Based on 29 records of this restaurant): Critical Violation



**Big Apple BiteSafe - NYC Restaurant Predictor**

## Predict Restaurant Inspections

Enter Restaurant Name:

### Prediction Results for COLUMBIA DINING URIS BLUE JAVA's next hygiene inspection

Violation level (Critical/Not Critical) Prediction (Based on 4 records of this restaurant): Not Critical Violation/ No violation

### List of Restaurants:

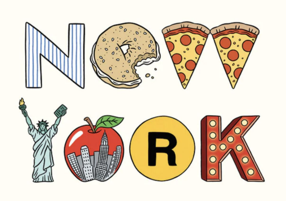
Tips:

1. To quickly find restaurants with their partial names, use the in-page search (Ctrl+F). And yes, this list is ready for some 'copy-pasta' — a little food humor for you!
2. Please note that a few restaurants might not yield results, as they were not included in our initial training data, impacting label precision and subsequent predictions.

© Restaurant Predictor created by Karl Liu, Tracy Hu, Isa Wang. All Rights Reserved.

Dynamic weighting system:

1. Numeric: 60% weight on newest record, 40% spreaded across all other records;
2. Categorical: mode



# Business Value

- A user-friendly toolkit addressing New Yorker’s daily concern and provide real-time predictions of a restaurant's hygiene conditions.
- Serve as an aggregated preemptive measure for restaurant owners and their customers, allowing restaurants to rectify potential violations before they occur, and customers to be aware of their preferred restaurants’ hygiene condition.
- Aggregated Data-Driven insights for the Department of Health and Mental Hygiene (DOHMH), potentially enabling more targeted and effective interventions with identified trends and patterns in hygiene violations.
- Scalability and Extension: With deployment on the cloud, this model can be scaled and adapted for usage in other cities

# Thank you!