NYC Restaurant Inspection Risk Classification

Goal: Predict restaurant risk category based on inspection data

Possible Approaches

Grade
or
Action-based

Static
or
Time-based

Assumption: Risk category (Low, Medium, High) is related to the **grade** (A, B, C)

Details of an inspection ⇨ predict the risk

Beatriz Pereira Santos

Understanding the data

- Each row corresponds to a violation (or lack of) identified in a restaurant on a specific date.
- Therefore, an inspection often corresponds to several rows.
- An inspection where no violation was found corresponds to a single row.
- The dataset contains records for several different types of inspections. Not all are gradable.
- Each gradable inspection is associated with a score and a corresponding grade.
- Possible grades: A, B, C, N, Z, P

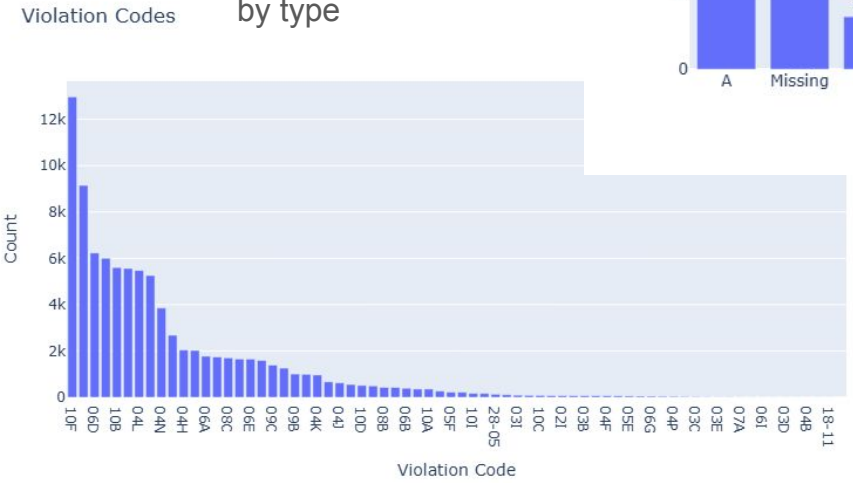
Selection of relevant data:

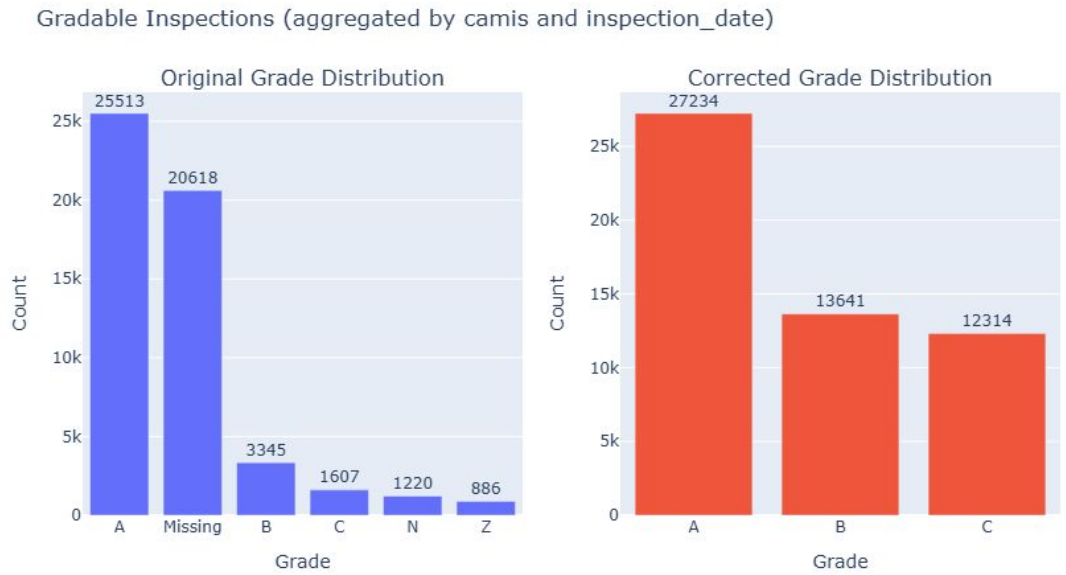
- removal of entries related to restaurants no yet graded
- removal of entries from the covid period
- selection of gradable inspections

Gradable Inspections

Independently of the type of gradable inspection, they are focused on the same scoring parameters

no need to segregate by type





Grade correction

so that each inspection is associated with its corresponding grade

Beatriz Pereira Santos

Data Preprocessing for ML model

Train/test split: 80/20%

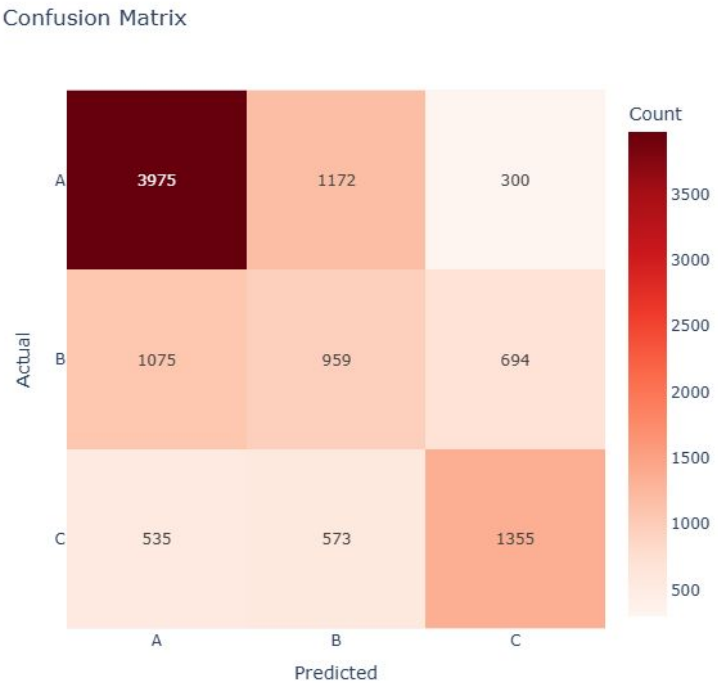
Features (different combinations were tested):

- violation_code (OHE of each violation code)
  - violation codes that appeared less than 10 times in the dataset were aggregate as OTHER
- boro (OHE)
- month (OHE)
- violation_reported (bool)
- nr_critical_violations
- nr_not_critical_violations

Target:

- risk_category

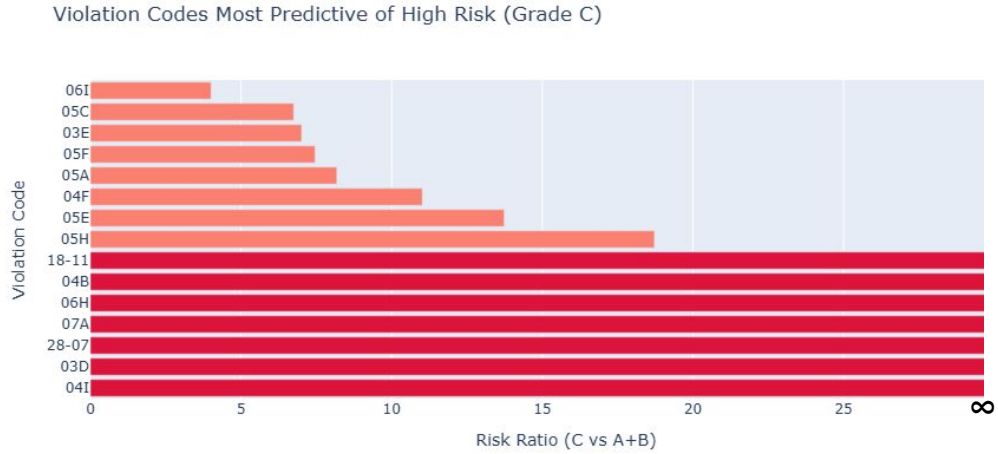
Model: Random Forest (selected using *RandomizedSearchCV*)



Macro F1-score: 0.55 Accuracy: 0.59

Insights

- Queens has the highest proportion of High-risk restaurants (26%) and State Island the lowest (18%)
- Certain violation codes are more predictive of higher risk
- some violation codes have only been recorded in high risk inspections



Beatriz Pereira Santos

GenAI: Automatic Generation of Monthly Alerts

mistral-small-2506

Prompt + Borough + Percentage increase in high-risk + Top violation codes

FOR IMMEDIATE RELEASE

PUBLIC HEALTH DEPARTMENT ALERT: INCREASE IN HIGH-RISK RESTAURANT VIOLATIONS IN MANHATTAN

The New York City Public Health Department is issuing an alert regarding a significant increase in high-risk restaurant violations in Manhattan. Recent data indicates a **34.8% rise** in inspections associated with high-risk violations, underscoring the need for heightened vigilance among restaurant operators and the public.

- **The most common violations identified include:**
- **Non-food contact surfaces or equipment** made of unacceptable material, not kept clean, or improperly sealed, raised, spaced, or movable to allow accessibility for cleaning on all sides.
- **Cold TCS (Time/Temperature Control for Safety) food items** held above 41°F; smoked or processed fish held above 38°F; intact raw eggs held above 45°F; or reduced oxygen packaged (ROP) TCS foods held above required temperatures except during active necessary preparation.
- **Establishments not free of harborage** or conditions conducive to rodents, insects, or other pests.

Restaurant operators are reminded to **adhere to strict food safety practices**, including proper sanitation, temperature control, and pest management, to protect public health. **Compliance with health codes is critical** to preventing foodborne illnesses and ensuring safe dining experiences for all New Yorkers.
Residents and business owners are encouraged to **report any observed violations or concerns** to the Public Health Department. Together, we can maintain high standards of food safety and public health in Manhattan.
For more information or to report a violation, **contact the NYC Public Health Department** at *[insert contact information]*.

Improvements

Data Preprocessing and ML

- additional handling of unbalanced data (undersampling/oversampling)
- explore different models
- extract the entire NYC Inspections database (currently working with 100k samples)

Code

- automatically combine different features
- automatically store experiment results
- optimize