# Serving Up Safety: Analyzing NYC Restaurant Health Grades

Project phase: Methods, Findings and Recommendations

03/25/2025
Ngawang Choega
nc87212n@pace.edu
Class Name: Practical Data Science
Program Name: MS in Data Science
Seidenberg School of Computer Science and Information Systems Pace university

# Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modelling methods
- Findings
- Business recommendations and technical next steps

# Executive summary

**Problem Statement:** NYC restaurants are graded based on health inspection results, but it is difficult to predict what grade a restaurant will receive ahead of time. This makes it challenging for restaurant owners and inspectors to take proactive steps to reduce violations and improve food safety.

**Solution:**

- The project analyzes past health inspection records to understand what factors influence restaurant grades.
- A simple predictive system is built to estimate whether a restaurant is likely to receive an A, B, or C before an inspection.
- This helps restaurants fix issues early and allows inspectors to focus on places that need the most attention.

# Project plan recap

| Deliverable | Due Date | Status |
|---|---|---|
| Data & EDA | 03/25/25 | Complete |
| Methods, Findings, and Recommendations | 04/01/25 | Complete |
| Final Presentation | 04/29/25 | In Progress |

# Data

# Data

- Key details
  - Data source: NYC Open Data – Department of Health and Mental Hygiene (DOHMH) [NYC Restaurant Inspection Results ]
  - Sample size: 274,740 inspection records.(more details)
  - Time period: September 24,2015 to March 17, 2025
- What's included
  - Restaurants that are active at the time of data collection.
  - Inspections that resulted in one or more violations, as well as inspections with no violations.
  - Restaurants with grades A, B, C or pending(P).(more details)
- What's excluded
  - New establishments that have applied for a permit but have not yet been inspected.(more)
- Assumption
  - The data provides a reliable sample of active NYC restaurants and reflects typical inspection and grading outcomes.

# Exploratory Data Analysis

# How are NYC restaurants graded?

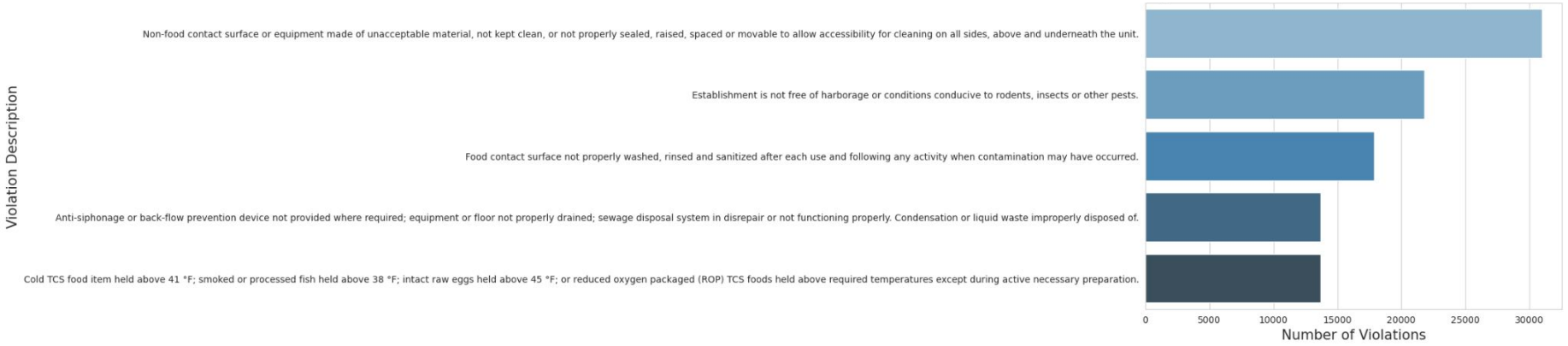

## Key Takeaways

- Over 85,000 restaurants in NYC received an **A grade**, showing overall good compliance.
- A smaller portion(less than 20,000) received **B or C grades**, indicating room for improvement.
- A large number of inspections are labeled **P (Pending)** — these have been conducted, but the official grade is not yet posted.

Data notes:

- Excluded Grades: N (not yet graded) and Z (not subject to grading) were removed from the analysis.

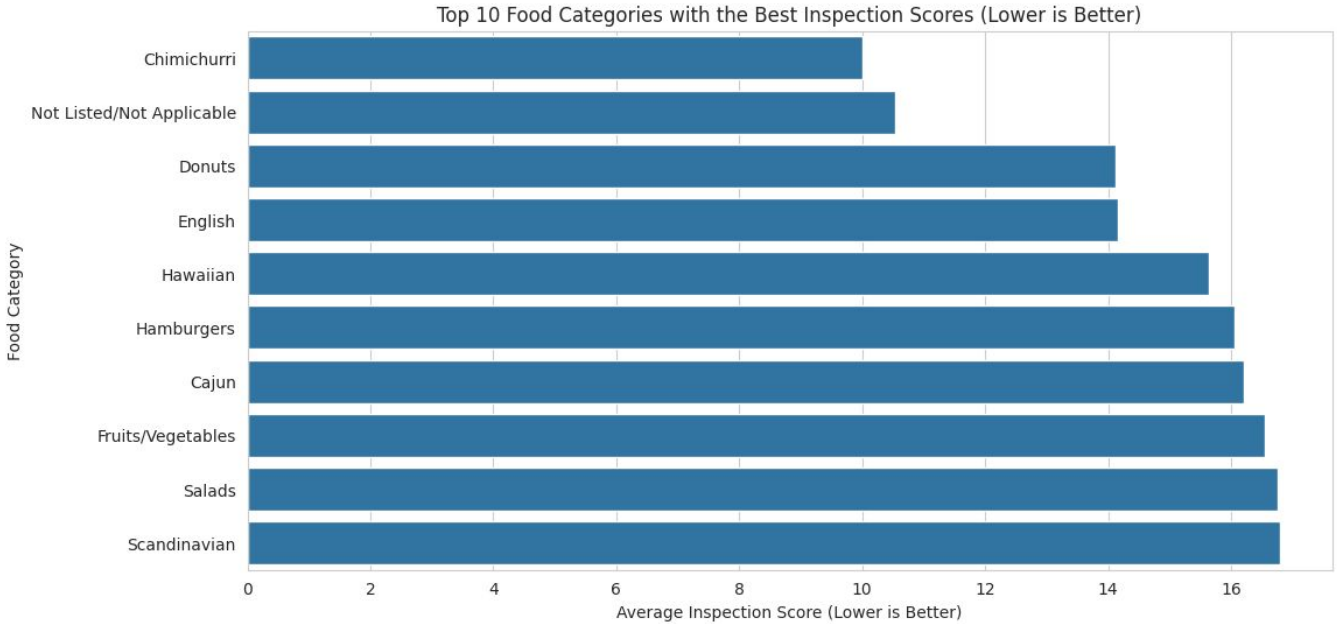
# Top 5 most common violations in NYC restaurants




## Key Takeaways

- The most common violation — **unclean or poorly maintained equipment** — appears in **over 30,000 inspections**.
- **Pest-related conditions** were cited in **more than 22,000 cases**, while **improper food surface sanitation** occurred in **around 18,000 inspections**.
- These high-frequency violations contribute directly to **point deductions** and are strong indicators of restaurants at risk of receiving lower grades.

# Which Food Categories Receive the Best Inspection Scores?




Note: Grades are based on inspection scores. See appendix for scoring breakdown.

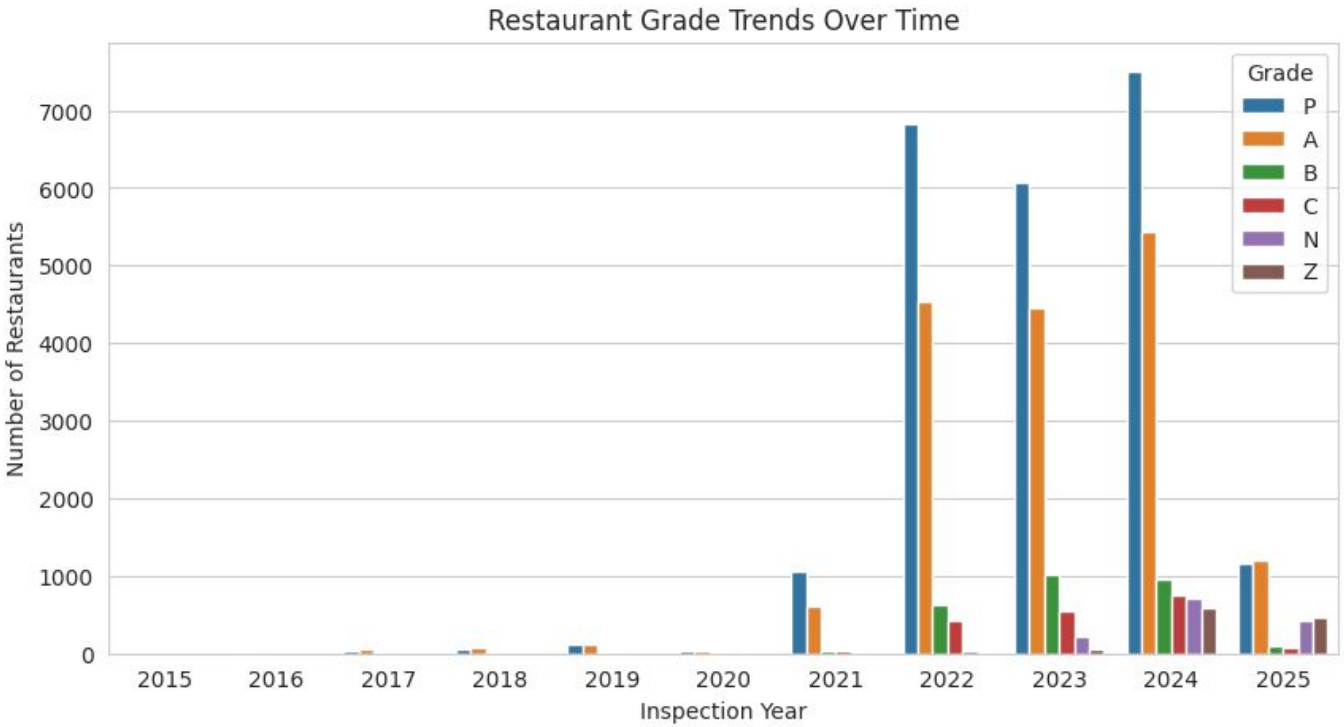
## Key Takeaways

- Food categories like Chimichurri, Donuts, and English had the lowest average violation scores, suggesting better inspection outcomes.
- Scandinavian, Salads, and Fruits/Vegetables showed higher average scores, indicating more frequent issues.

Data notes:

- “Not Listed/Not Applicable” includes restaurants that did not specify a cuisine or selected a general option on their permit — often new or small businesses.

# How have restaurants grade changed over time?




Note: Limited data appears from 2015–2020. See appendix for explanation.

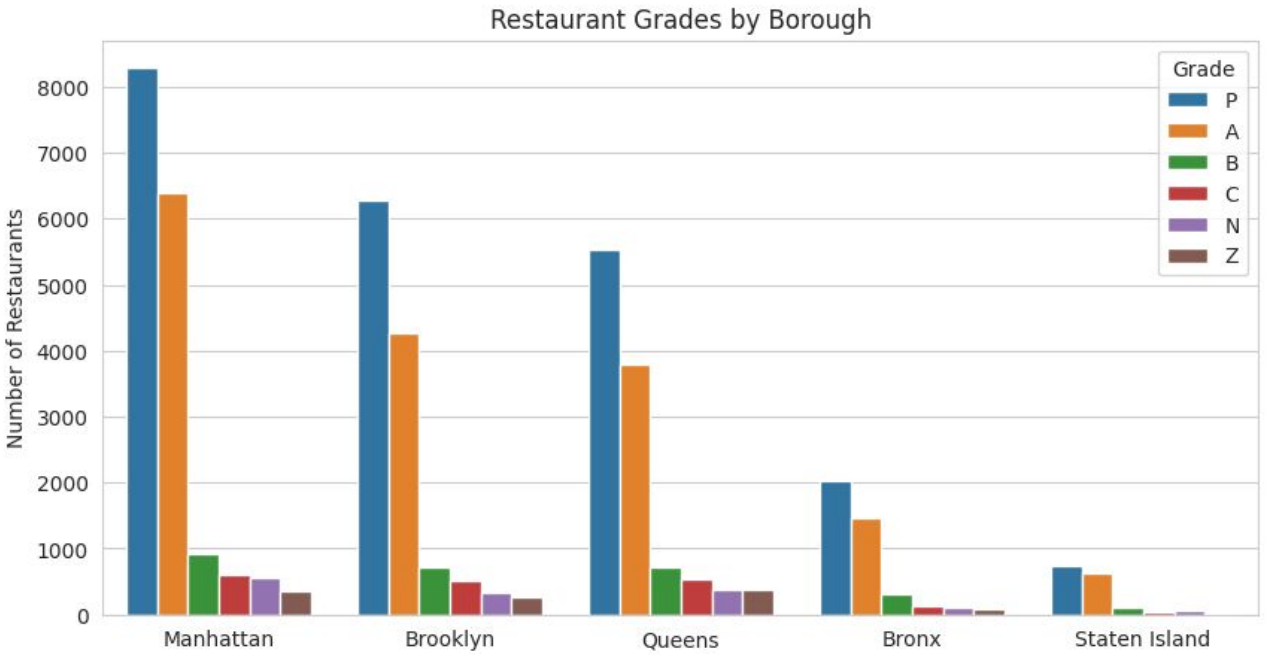
## Key Takeaways

- Inspections increased sharply after 2021, following a slowdown likely caused by the pandemic.
- A grades are the most common each year, showing consistent performance.
- Many inspections from 2022 to 2024 are still marked as pending (P).

Data notes:

- N = Not yet graded (under review)
- Z = Not subject to grading (e.g., food vendors)

# How do restaurant grades vary across NYC boroughs?




## Key Takeaways

- Pending grades (P) are common across boroughs, especially in Manhattan(~8000+)and Brooklyn(~6000+).
- Manhattan has the highest number of restaurants, with over 6,000 A grades.
- Brooklyn and Queens follow, showing strong A grade counts with some lower grades present.
- Staten Island has the fewest restaurants(less than 1000) but the highest proportion of A grades.

# Modeling Methods

# Modeling methods

- The predicted outcome -
  - Whether a restaurant will receive an **A, B or C** grade during a health inspection.
  - This helps restaurant owners fix issues early and allows health inspectors to focus on places that may need more attention.

- Features - used in building the predictive model
  - Borough
    - Different boroughs in NYC show different patterns in restaurant inspection results.
    - By including this feature, the model can capture location-based trends, helping it make more accurate predictions based on where the restaurant operates.
  - Cuisine Type
    - Some cuisines are more likely to have violations (e.g., based on how food is stored or prepared).
    - This helps the model identify patterns linked to food types.
  - Inspection Type
    - Different types of inspections (like routine vs. follow-up) can reflect different risk levels.
    - Including this helps the model better understand the situation of each inspection.

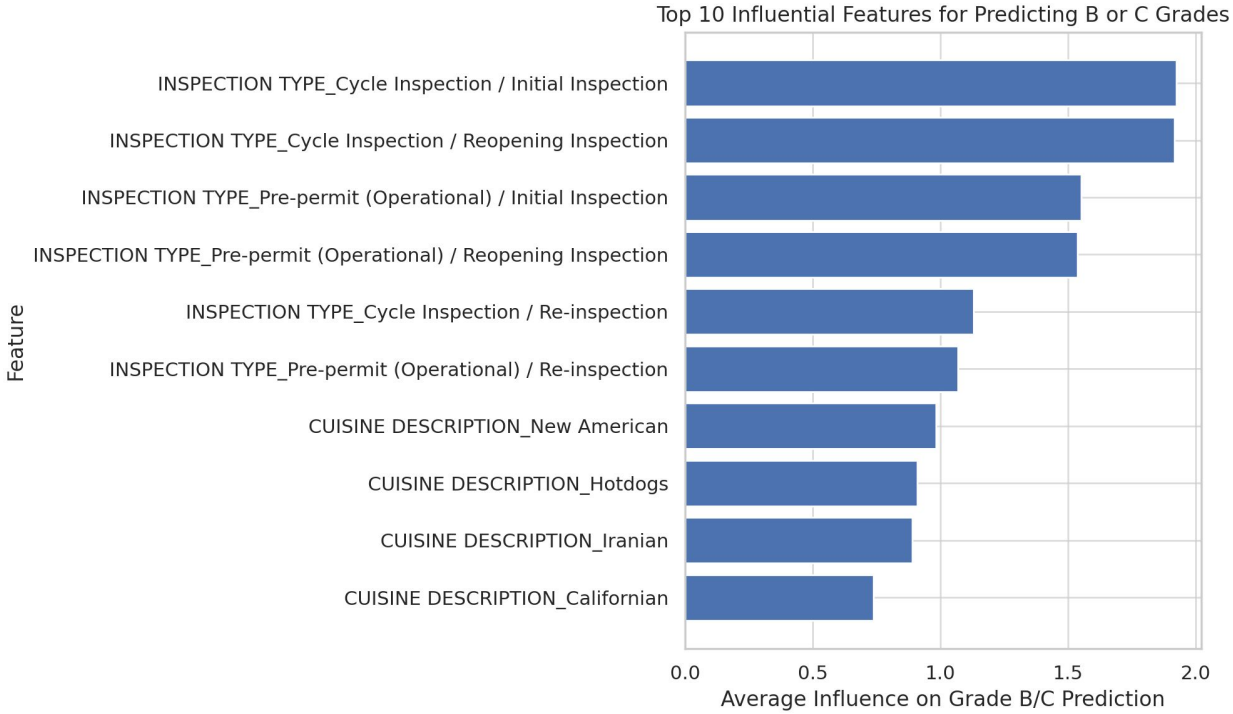
# Modeling methods

- Model type: **Logistic Regression**
  - The project uses a simple, widely trusted model called Logistic Regression to estimate whether a restaurant will receive an A, B, or C grade based on historical inspection patterns.
- Rationale behind choosing this predictive model:
  - Easy to understand
    - The model shows how each factor (like borough or cuisine) affects the prediction.
    - This helps explain why a restaurant might be at higher risk of a lower grade, making it easier for public health teams and restaurant owners to take action.
  - Fast and Efficient
    - The model delivers results quickly and doesn't require complex computing systems, making it ideal for real-world use by health departments or city agencies.
  - Proven and Practical
    - Logistic Regression is a reliable approach used in many industries for similar tasks.
    - It works especially well with structured data like inspection records, where clarity and consistency are more important than complexity.
- See appendix for a more technical explanation of how the model works.

# Findings

# Model accuracy : how well the predictions work?

- Key Takeaway:
  - The model predicts restaurant health grades with 78% accuracy using only basic information like borough, cuisine type, and inspection type — all available before an inspection.
  - Most accurate with A-grade predictions, which dominate NYC restaurant outcomes.
  - Less accurate for B and C grades, largely due to their smaller presence in the data (a known data imbalance).
  - Despite that, the model provides valuable early insight into which restaurants may need attention.
- Feature impact summary
  - Borough: Restaurants in Brooklyn and the Bronx were slightly more likely to receive B or C grades.
  - Cuisine Type: Complex food categories (like Seafood or Chinese) were more associated with non-A grades.
  - Inspection Type: Re-inspections had higher chances of resulting in B or C grades than routine checks.
- See appendix for detailed accuracy breakdown & confusion matrix.




Key takeaways:

- Certain boroughs and cuisines are more often linked with lower grades.
- Re-inspections tend to signal more risk than routine visits.
- The model learns real-world patterns from these features.

# Business Recommendations & Technical Next Steps

# Business Recommendations

1) Offer to integrate the model into the Health Department’s inspection system.Use it to flag restaurants that are less likely to receive an A, so owners can proactively improve conditions and inspectors can prioritize higher-risk visits.

2) Launch borough-specific outreach and offer preventive training or food safety workshops for high-risk cuisine categories. This can reduce violations and promote equity in food safety standards across the city.

# Technical next steps for the data science team

- Improve Prediction for Lower Grades (B & C):
  - The model performs well for A grades but struggles with B and C due to class imbalance.
  - Improving the balance of the dataset will enhance accuracy across all the grades.
- Explore advanced model
  - Testing out various other models and or even fusion of those complex model to see how it impacts the model performance and accuracy.

# Appendix

# Project Materials

- Git Repo: <link>

# Sample size details

- Dataset name: DOHMH_New_York_City_Restaurant_Inspection_Results.csv
- Original dataset: 274,740 rows
- Cleaned dataset: 193,892 rows
- Each row represents a single violation record from a restaurant inspection
    - If a restaurant had multiple violations during one inspection, it appears in multiple rows – one for each violation
    - If a restaurant had no violation, it is represented by a single row with a violation field marked accordingly .

<back to main slide>

# Additional notes on the data

- The dataset includes records for:
  - Restaurants, college cafeterias, mobile vendors, special programs,etc.
  - New establishments that applied for a permit but have not yet been inspected (marked with the date **01/01/1900**)
  - Inspections with and without violations
- Grades used in the dataset:
  - **A, B, C** — Based on inspection outcomes
  - **P** — Grade is pending after inspection
  - **N** — Not yet graded (inspection under review or appeal)
  - **Z** — Not subject to grading (e.g., mobile vendors, special programs)

🔗 Source: [NYC Health Restaurant Grades Website]

# Why is there limited data from 2015-2020?

- The dataset only includes inspection records from up to three years prior to the most recent inspection for active restaurants.
- As a result, older data (pre-2021) is mostly excluded unless the restaurant had recent inspections.
- This design ensures the dataset stays focused on currently operating restaurants with recent inspection activity.

🔗 source: DOHMH New York City Restaurant Inspection Results

<back to main slide>

# NYC restaurant grading system

- A Grade: 0–13 points ⟶ Minimal violations
- B Grade: 14–27 points ⟶ Moderate violations
- C Grade: 28 or more points ⟶ Significant violations
- P Grade: Inspection done, grade pending
- N Grade: Not yet graded (under review or appeal)
- Z Grade: Not subject to grading (e.g., food vendors)

🔗 source: NYC restaurant grading system

<back to main slide>

# Technical details : Logistic Regression model

- Outcome Variable:
  - GRADE – Multiclass target with values: A, B, or C
  - Classification type: Multinomial Logistic Regression
- Model Type & Setup:
  - Algorithm: **LogisticRegression** from sklearn.linear_model
  - Configuration:
  - multi_class='multinomial'
  - solver='lbfgs'
  - max_iter=1000
- Features used:
  - BORO (One-hot encoded)
  - CUISINE DESCRIPTION (One-hot encoded)
  - INSPECTION TYPE (One-hot encoded)

<continued on next page>

# Technical details : Logistic Regression model

- Model Training:
  - Train-Test Split: 80/20 with stratify=y to preserve grade balance
  - Loss Function: Cross-entropy
  - Probability Estimates: Model uses softmax to calculate the probability of each grade

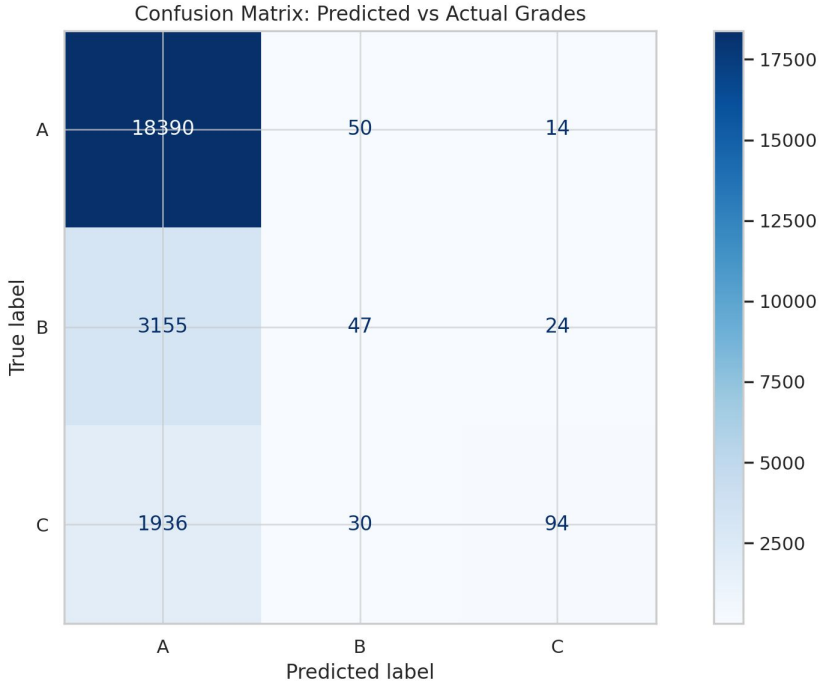
<back to main slide>

# Model Evaluation : Classification report

| Grade | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| A | 0.78 | 1.00 | 0.88 | 18,454 |
| B | 0.37 | 0.01 | 0.03 | 3,226 |
| C | 0.71 | 0.05 | 0.09 | 2,060 |
| **Accuracy** | — | — | **0.78** | 23,740 |

<back to main deck>

# Confusion matrix




**Key takeaways:**

- The model correctly predicts most A grades (bottom-left cell).
- It struggles to distinguish B and C grades (upper-right portion), due to class imbalance in the dataset.
- Shows why model is strongest as an A-grade identifier and early risk flagger.