EPA1

Final Project Presentation

Mike Chen, Wenyu Huang, Carly Jones

The Problem

The New York State Department of Environmental Conservation only has the resources to inspect **700** out of **thousands** of hazardous waste handlers in New York State every year.

So who should be inspected?

Our Goals

We can **reduce the number of serious hazardous waste incidents each year in New York State** by helping NYSDEC prioritize its inspection efforts.

We can also:

1. Minimize incidents in **low income communities**
2. Distribute the **burden of inspections and increase confidence** in the inspection process by balancing inspections between facilities with and without inspection history

Data

We will focus on **active Large Quantity Generator (LQG) handlers** inspected by the NYSDEC on behalf of the federal Environmental Protection Agency (EPA).

Available data sources include:

- **Handler activity** data from NYSDEC
- **Waste** data from NYSDEC
- **Inspection history** data from the EPA
- **Enforcement and compliance** data from the EPA
- **Water and air pollution** data from the EPA
- **Household income** data from the US Census Bureau

Analytical Formulation

On **January 1st** of each year,

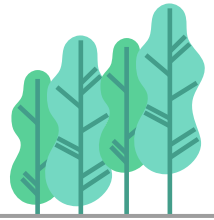
for all of the **LQGs** that were active in the previous year in **New York State**,

can we identify the **700** LQGs

that are **most likely to be the subject of formal enforcement**

so that NYSDEC can **prioritize its inspection efforts** ?

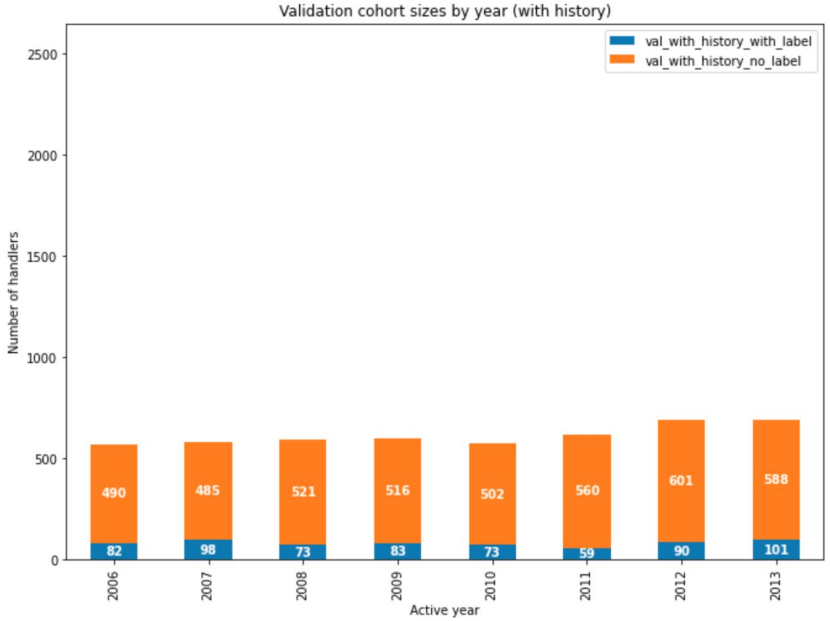
Modeling Choices

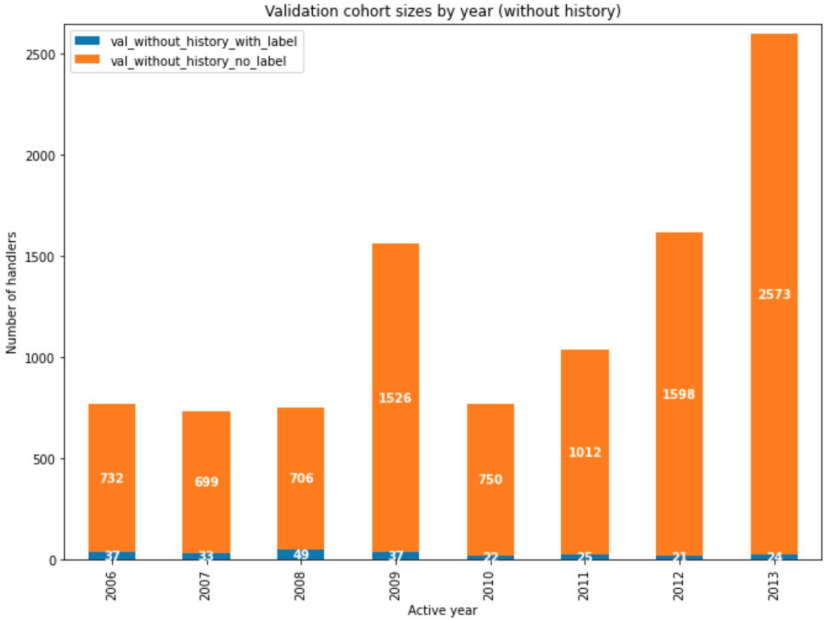


Cohort (Rows)	Unique Large Quantity Generators (LQGs) that were active between **January 1st and December 31st** of **the previous year** in New York State
Label	Likely to be the subject of one or more **formal enforcement activity(/ies)** in the **subsequent (1) year**
Features (Columns)	***Inspection history, violation history,*** facility information, waste information, industry information, parent company information, and geographic information
Model Types	**Binary classification (yes or no)** logistic regression, decision tree, boosted decision tree, random forest Commonsense baseline: **rank LQGs by number of previous violations**

To make best use of past violation data, we train TWO models (with history vs. without history) with different set of features.

Modeling Choice #1



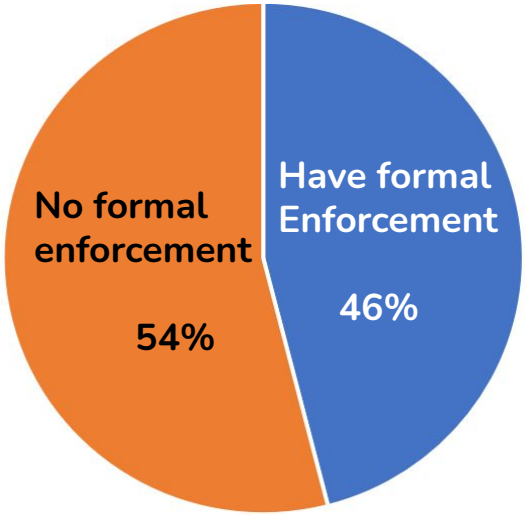


Modeling Choice #2

The model predicts whether the LQG will have formal enforcements, instead of violations only.

Example: for LQGs active in NY in 2013 and inspected in 2014:

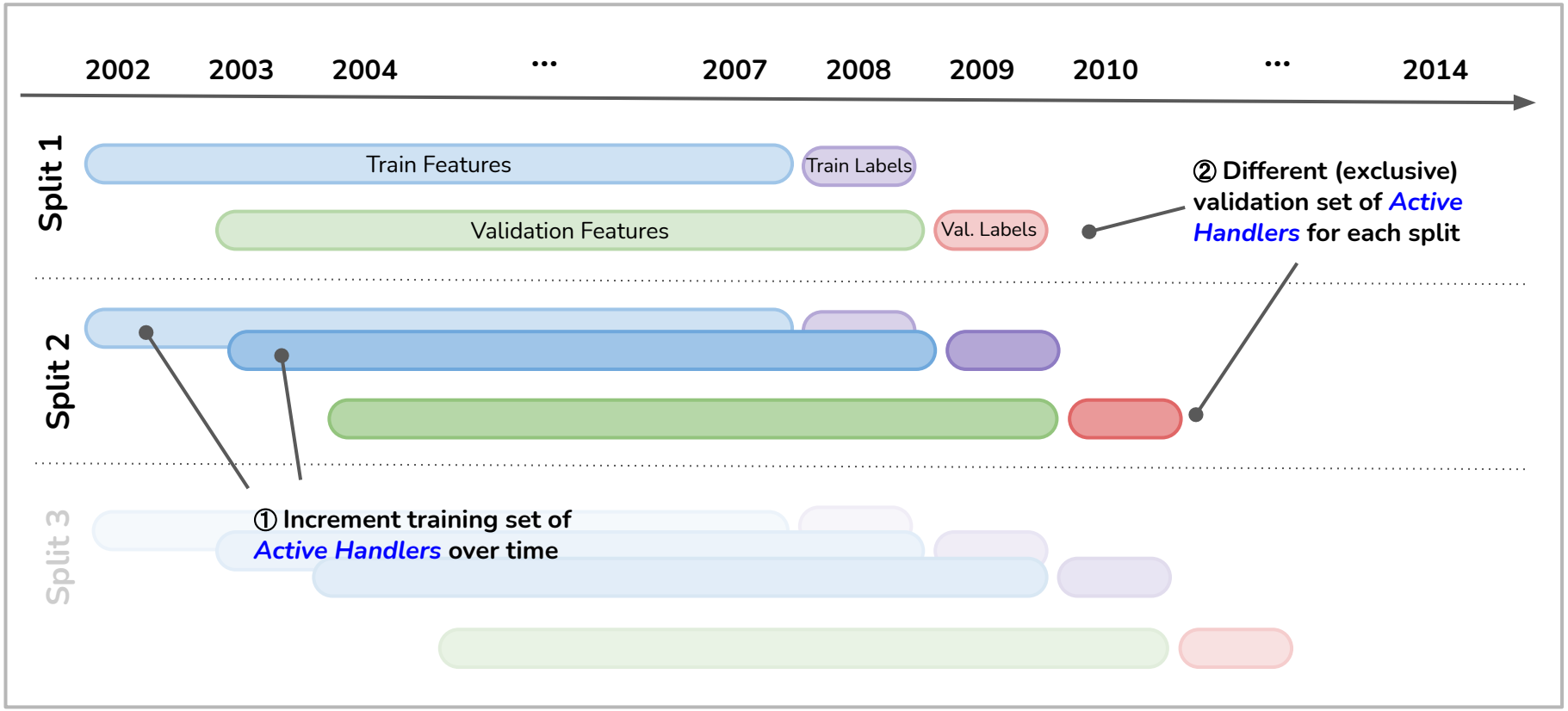
- **46%** have formal enforcements
- **100%** have violations



Metrics

Metric Name	Meaning	Priority
Precision vs Population	Efficiency of catching handlers with **Formal Enforcement** (21% total population ≈ 700 facilities) 	High
Recall	Coverage of handlers with **Formal Enforcement** 	Medium
Support	Coverage of **inspected** handlers (explore vs. exploit) 	Medium

Train and Validation Splits



Results: Model Selection Criteria

With History Cohort Models

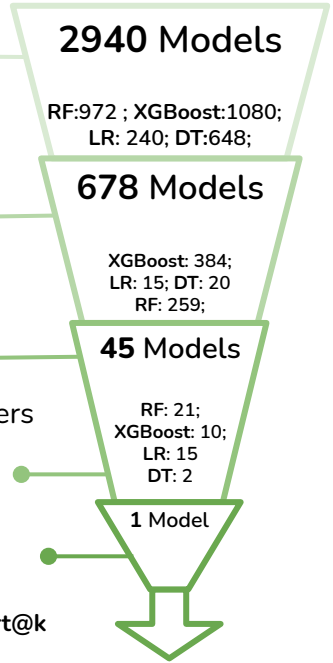
① Filter by **Avg. Support** (**confidence**, cutoff = **0.3**)

② Order by **Avg. Precision** (**stability**), **Precision in 2014,2013** etc. (**recency**)

③ Filter models with **same** performance and **similar** hyperparameters

④ Select **Top** Models

⑤ **Holistic Evaluation** (based on **Precision-Recall-Support@k** overtime)



LR_36

Without History Cohort Models

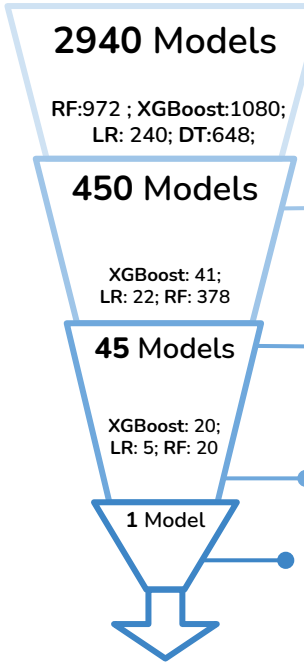
① Filter by **Avg. Support** (**confidence**, cutoff = **0.12**)

② Order by **Avg. Precision** (**stability**), **Precision in 2014, 2013** etc. (**recency**)

③ Filter models with **same** performance and **similar** hyperparameters

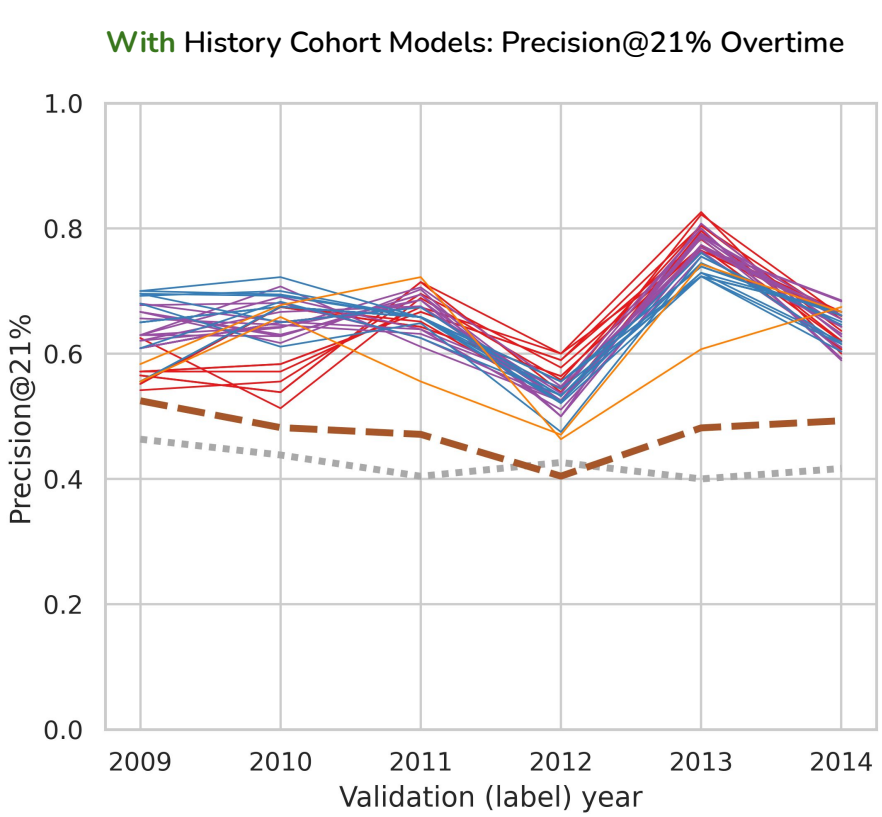
④ Select **Top** Models

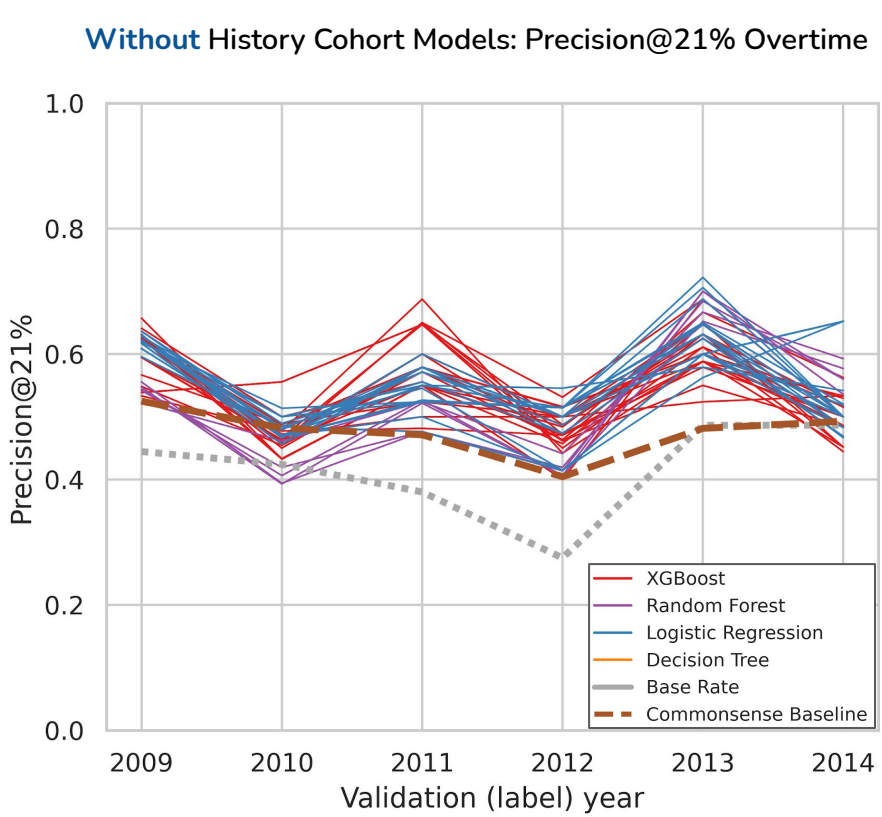
⑤ **Holistic Evaluation** (based on **Precision-Recall-Support@k** overtime)



XGBoost_97

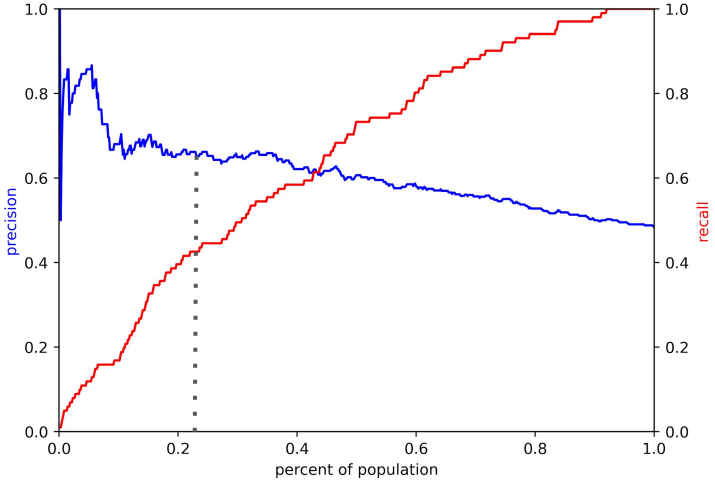
Results: Top Performing Models





With History Model with Best Overall Performance (LR)

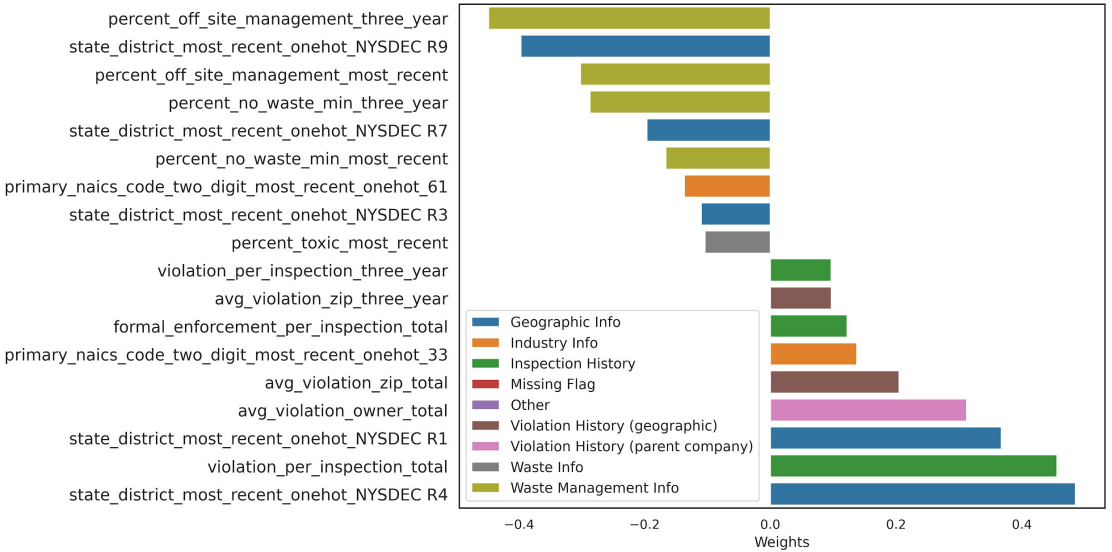
PRK Curve (2014 Val. Set)



Precision@21%: 0.65 **Support@21%:** 0.437

Commonsense Precision@21%: 0.493

Most Important Features (large absolute weights)

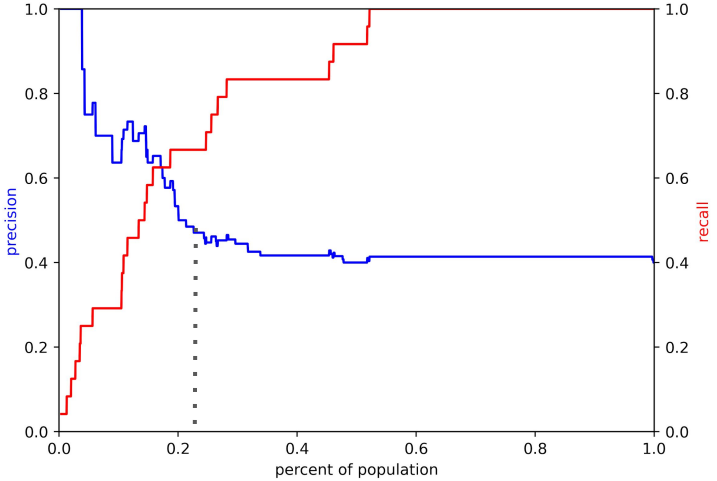


Top Features: inspection history, geographic information, waste management information

Top Feature Types: geographic information & waste management information

Without History Model with Best Overall Performance (XGBoost)

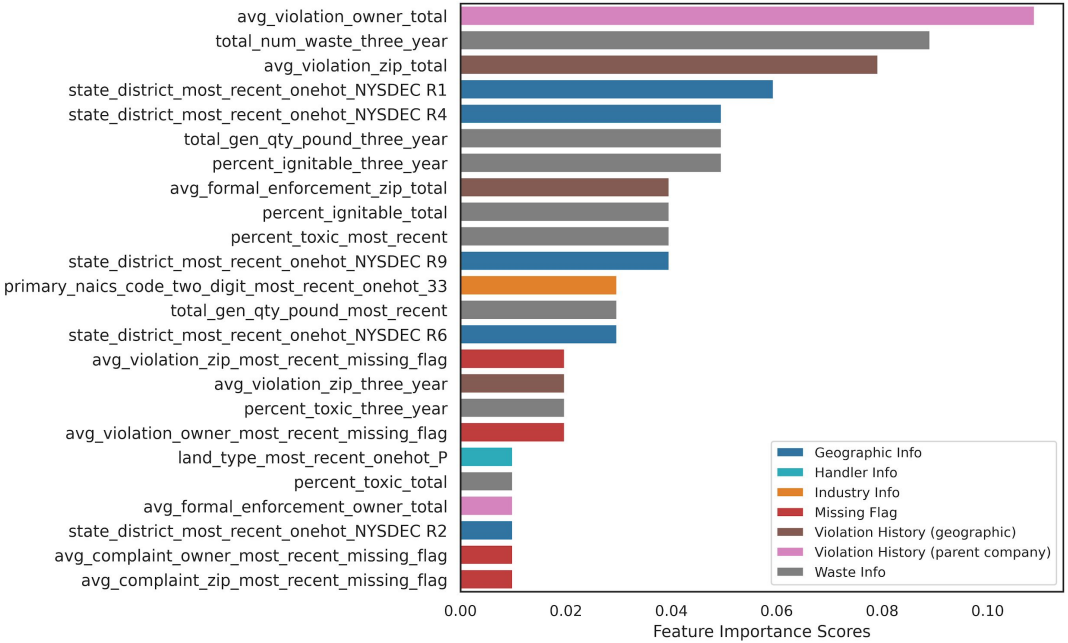
PRK Curve (2014 Val. Set)



Precision@21%: 0.5 **Support@21%:** 0.058

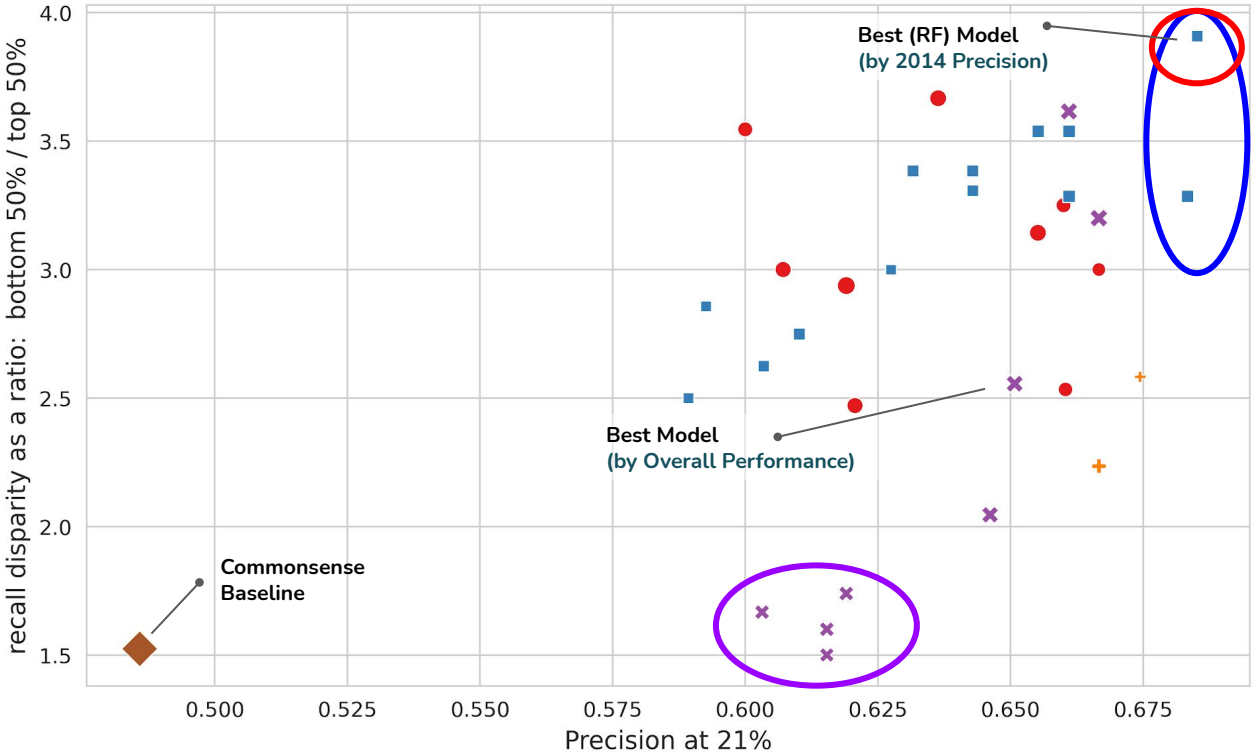
Commonsense Precision@21%: 0.493

Most Important Features (most used in making splits)



Top Features: parent company's violation history, waste information, geographic information

Bias & Fairness Audit: With History

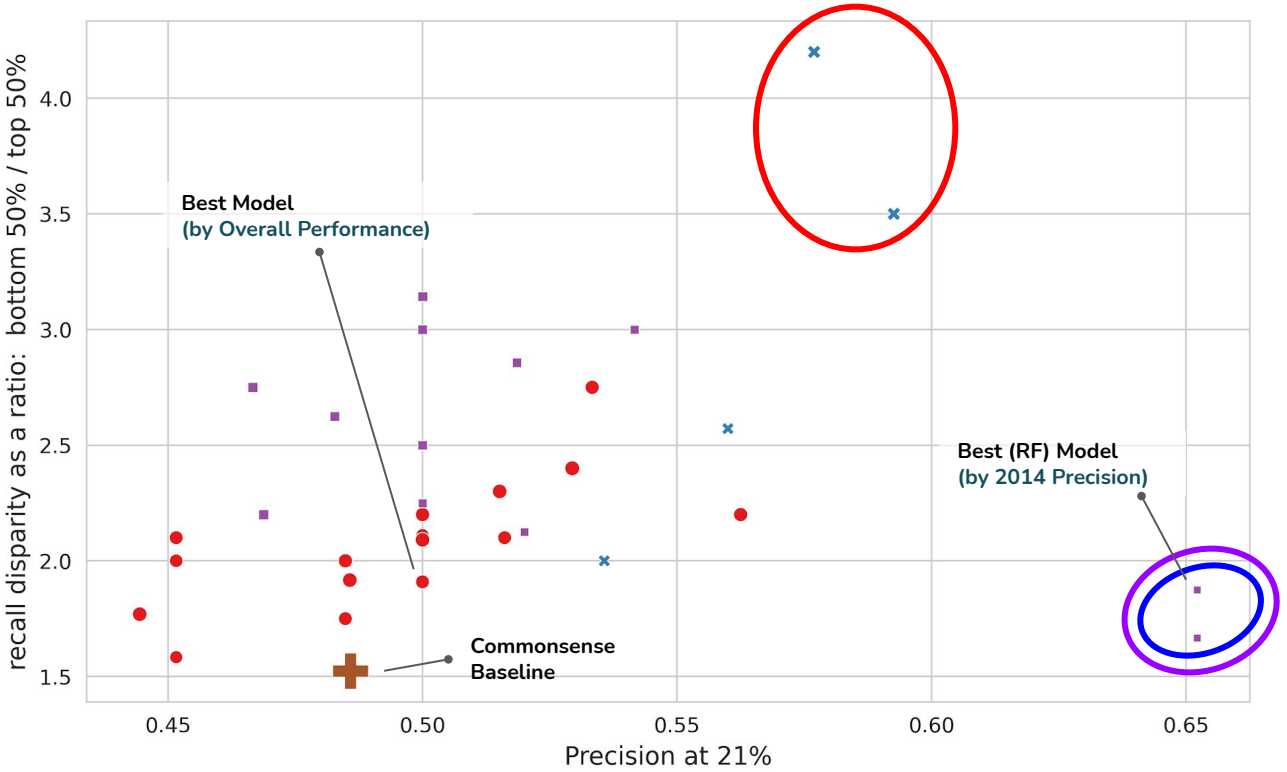


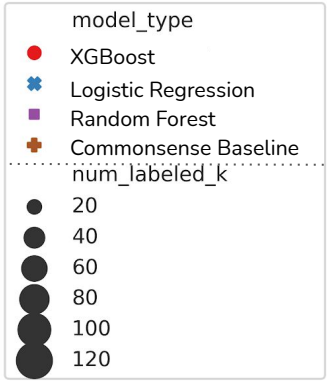
Models that maximize *efficiency*

Models that best support *equity*

Model that may be *most fair*, in a corrective sense, depending on empirical evidence of inequities

Bias & Fairness Audit: With History





Models that maximize *efficiency*

Models that best support *equity*

Model that may be *most fair*, in a corrective sense, depending on empirical evidence of inequities

Policy Recommendations

- Adjust percentage of handlers to inspect in 2 models depending on immediate goals (explore vs exploit)
- Conduct field trial(s) to improve bias & fairness analysis.
- Design interventions for large parent companies with histories of violations
- Increase incentives for waste minimization efforts
- Collect more (block level) granular geographic data

Caveats & Future Work

- Our model only focuses on LQGs but there are many **other types of handlers** that need inspection.
  - SOLUTION: Use additional datasets that contain reports of other types of handlers.
- Our metric is limited because a large portion of handlers do not have inspection results.
  - SOLUTION: Field Trial provides a chance to compare accuracy fairly.
- We need **more feature engineering**, especially for the cohort without previous inspection history.
  - e.g., inspection-related features from handlers similar in terms of location, parent company

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik** and illustrations by **Stories**

Please keep this slide for attribution

Thanks!

Do you have any questions?