Predicting Environmental Waste Violations for Large Quantity Generators in New York

Theophile Gervet, Kaila Gilbert, Juyong Kim, Ruohan Li

School of Computer Science, Machine Learning Department and
Heinz College of Information Systems and Public Policy
Carnegie Mellon University
Pittsburgh, PA, United States

May 2020

Link to GitHub repo: https://github.com/dssg/mlpolicylab_sp20_e2.git

Table of Contents

Executive Summary

New York possesses 22% of all large quantity waste generators in the United States. Few of these facilities have ever been inspected by state and federal regulatory bodies. The goal of this project is to assist the EPA in its yearly inspection planning process by understanding which facilities are at the highest risk of a violation based on historical, demographic, and site-specific information. This information would optimize the allocation of scarce public resources for public safety. We applied an equity, efficiency, and effectiveness framework to guide our analysis. The current project only explores equity and efficiency but recommends looking more into effectiveness in future work.

We formalize the problem as time-series binary classification, where the goal is to predict whether a facility will commit at least one violation in a given year. Upon running a grid of 282 model and parameter combinations — including random forest, KNN, logistic regression, and Adaptive Boosting — we found that random forests with large numbers of trees and large depth tended to have the highest performance. We selected the model with the highest precision as well as the highest minimum precision (minimax) at 3%, representing the ability of the EPA to mobilize resources. These models perform well on both our equity and efficiency metrics.

Our central recommendation is to validate these findings using a field trial. Following a successful outcome, we recommend the use of either our highest mean precision or best minimax model to help select facilities for inspection. We also recommend inspecting unexplored facilities in general to increase the performance of future models. We also recommend acquiring demographic info at the block group level to detect variations within counties. While we only explored large quantity generators in this project, we recommend conducting similar analyses on different types of waste management facilities and different types of violations.

Background and Problem Statement

The Federal Environment Protection Agency (EPA) and the New York State Department of Environmental Conservation (NYSDEC) oversee the enforcement of federal and state hazardous waste management regulations in New York. This work is essential; improperly managed hazardous waste poses a serious threat to human health and the environment. The EPA and NYSDEC inspect waste management facilities to detect violations, enforce compliance, and penalize violators. Such interventions mitigate the adverse effects of existing violations and deter further violations.

While New York possesses 22% of all large quantity waste generators in the United States, few have ever been inspected. Because the EPA and NYSDEC can only inspect 3% of the largest waste generators each year, they must be strategic in their selection process. We hypothesize that the current process is driven by heuristics (like the age of the facility or time since the last examination), randomization, or a response to complaints. While the incidence of devastating environmental events is rare, public agencies likely do not allocate their resources optimally, letting threats go undetected to the detriment of affected

communities and the environment. Data-driven, risk-informed inspection planning will help uncover existing violations, deter future violations, and ultimately safeguard public health.

Related Work

Many entities have tried to predict or mitigate adverse impacts of pollution and waste violations, including the EPA.[1] In most cases, there are substantial challenges. Due to greater demands on scarce public resources, there has been an increased focus on using "risk-informed" decision-making to prioritize which sites, projects, or activities are selected and investigated by environmental agencies.[2] While models have been used to anticipate the frequency of expected environmental violations and forecast resources[3] [4], few studies attempt to identify high risk locations given sparse historical information. Similar risk-based prediction tasks exist in other contexts (e.g. corporate fraud detection or predictive policing), but few cases have been conducted within an environmental compliance context.

Likewise, the task of making predictions with relatively few instances of labeled data has puzzled researchers for decades. Approaches like generative-based methods, graph-based methods, and co-training are increasingly popular ways that scientists have found that may improve predictive performance.[5] The current study hopes to apply a number of new and old supervised machine-learning classifiers in order to train models on labeled information (e.g. facilities which have ever been inspected) for the purposes of predicting on unlabeled entities (e.g. facilities which have never been inspected).

Problem Formulation and Solution Overview

As we will see in the exploratory analysis section, large quantity generators (LQGs) of waste represent 60% of inspections and mobilize most public resources. In this project, we restrict our attention to these facilities. We formalize the problem as time-series binary classification. Since the EPA and NYSDEC determine which facilities to inspect via annual planning, we predict at the same interval. Multiple inspections can be conducted on a single facility within a year — some detecting a violation and others not. We label a facility as a positive in a given year if at least one inspection detects a violation. For each year, we detect a violation using data from previous years aggregated over a fixed time-window. We later

[1] Borsuk, Mark E., Craig A. Stow, and Kenneth H. Reckhow. "Predicting the Frequency of Water Quality Standard Violations: A Probabilistic Approach for TMDL Development." Environmental Science & Technology 36, no. 10 (2002): 2109–15.
[2] Trimble, David C. "ENVIRONMENTAL LIABILITIES: DOE Would Benefit from Incorporating Risk-Informed Decision-Making into Its Cleanup Policy." GAO Reports, October 18, 2019, 54.
[3] Sharma, P., M. Khare, and S.P. Chakrabarti. "Application of Extreme Value Theory for Predicting Violations of Air Quality Standards for an Urban Road Intersection." Transportation Research Part D: Transport and Environment. Pergamon, April 22, 1999.
[4] Borsuk, Mark E., Craig A. Stow, and Kenneth H. Reckhow. "Predicting the Frequency of Water Quality Standard Violations: A Probabilistic Approach for TMDL Development." Environmental Science & Technology 36, no. 10 (2002): 2109–15. https://doi.org/10.1021/es011246m.
[5] Ling, Charles X., Jun Du, and Zhi-Hua Zhou. "When Does Co-Training Work in Real Data?" Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science, 2009, 596–603. https://doi.org/10.1007/978-3-642-01307-2_58.

restrict inspections to exclude those that were derivative (e.g., follow-up inspections or those related to the compliance or enforcement of previous inspections).

For every LQG active in New York in 2017 (the last year in our data), for every year from 2009 to 2014, we predict the probability that it will commit at least one violation in a given year. We restricted our attention to this period because of data availability: information on the types of waste generated and ACS data at the county level are only available during this period in our dataset. Public agencies should be able to run our model today as they may have access to more up-to-date waste reports and ACS data.

Data Description and Exploratory Analysis

Our analysis relies on the following datasets:

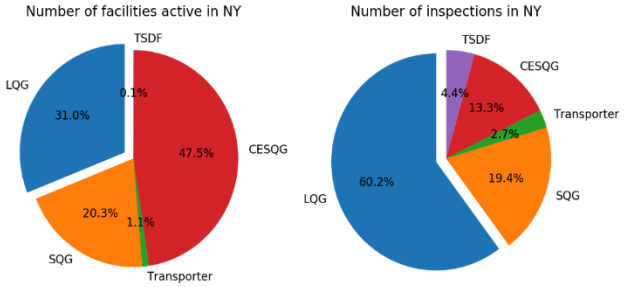
- **Resource Conservation and Recovery Act (RCRA):** Maintains data about facilities (status, type, industry), evaluations, violations, and enforcement of penalties from 1980 to 2017.
- **NYSDEC**: Contains details on the activities of Large Quantity Generators (LQGs), including the types and quantities of waste handled and the nature and volume of waste produced and received from 2006 to 2014.
- **American Community Surveys (ACS)**: Contains information on households by geographical regions (e.g., population density, income) at the zip code level, beginning in 2009.

Facility Types and Activities: A Deep Dive

According to the RCRA dataset, the state of NY manages a disproportionately large share of hazardous waste management facilities: 7% of total facilities and 22% of the largest waste generators. There are three types of waste management facilities:

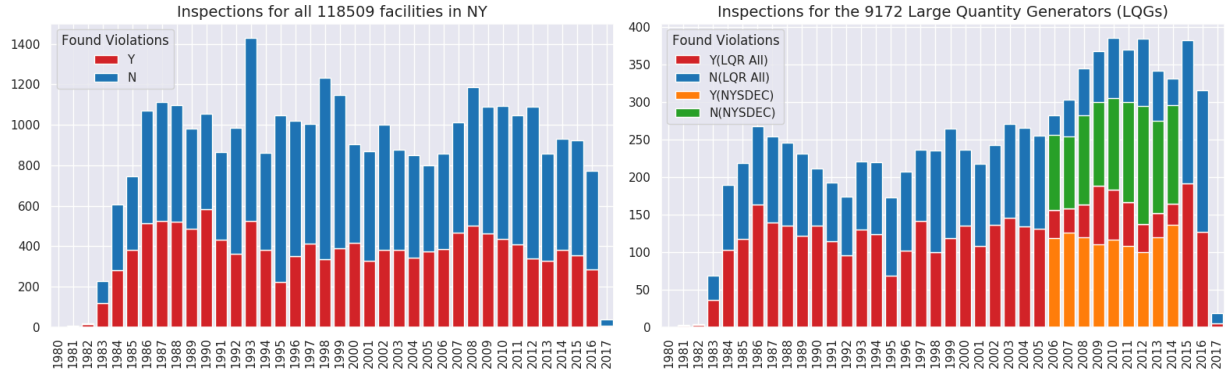
- Generators: large quantity generators (LQGs), small quantity generators (SQGs), and very small (conditionally exempt) quantity generators (CESQGs)
- Transporters
- Treatment, storage, and disposal facilities (TSDFs)

While LQGs represent only about a third of facilities, they compose 60% of evaluations. Since the inspection of LQGs mobilizes most public resources, and the NYSDEC dataset is restricted to LQGs, we restricted our analysis to LQGs.



Trends Over Time

From 1980 to 2017, public agencies have inspected roughly a thousand facilities and 300 LQGs each year. This is about 3% of active facilities and 3% of LQGs. These numbers provide an estimate of the regulatory resources available. Of these violations, NYSDEC conducted about 90% of evaluation, with the rest evaluated by the EPA. When reviewing the outcome of inspections, we noted that on average, 43% of total facilities and 51% of LQGs resulted in a violation in a given year. Breakdowns by violation outcome and by governing body are found in the figure below.



The Pipeline: Creating a Solution

The figure below shows the overall pipeline of the project. Each cylinder of the diagram represents databases, and each box of the diagram represents data (or a model) that the pipeline produces.

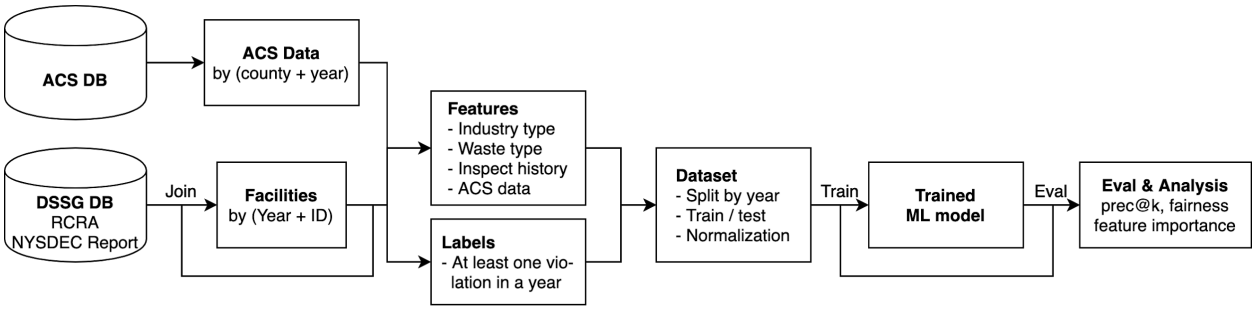


Figure: E2 Project Pipeline

The Cohort

We extract the list of the facility identifiers (RCRA ID) for LQGs active 2017. Regarding time, we focus on 2009 to 2014, where all types of data are available. Each row is a facility-year pair.

Generating Labels and Features

Given the list of RCRA IDs and the years of interest, we retrieve the inspection and violation records of the facilities. We use inspection outcomes as labels (at least one violation in a given year) and extract features from the inspection/violation history. We obtain the industry types (NAICS codes) from the

RCRA database, the waste codes of the waste generated by the facility from the NYSDEC database, and geographical features from ACS data at the county level. We normalize ACS features to zero mean and unit variance, due to the sensitivity of our classifiers to the scale of features.

Feature Set

We hypothesize that models with features that generalize to non-inspected facilities will have more predictive power on such facilities. Ideally, this would solve leakage problems by finding non-historical characteristics. Our model uses 193 continuous and categorical variables in total. Except for the four variables that belong to the Inspection History group, all other variables generalize to facilities that have never been inspected before. In total, we have four feature groups, as summarized in the table below. A full list of all features we used can be found in the Appendix.

Feature Group	Generalizable to never inspected facilities	Examples
Inspection History	No	Inspection/violation binary flag last year, number of inspections/violations in the past 5 years
Waste Codes	Yes	Type and quantity of waste generated
NAICS Codes	Yes	Industry sectors
ACS Features	Yes	Population density, median income, poverty status, heating and fuel consumption, structure year, etc

Models

Baseline. Our baseline model is a random selection model predicting at the base rate of violations (48%). Under random selection, the proportion of labeled data among the top 3% rankings is 3%.

Classifiers*.* We selected four of the most popular binary classifiers for our prediction task: Logistic Regression, Random Forest, K Nearest Neighbors, and (tree-based) Adaptive Boosting. We also tested different combinations of feature groups: for example, with or without waste codes. In total, our grid includes 282 different models (the full list of models is in the Appendix).

Evaluation

Train-Test Splits*.* Every year from 2010 to 2014, we evaluate a model trained on previous years (starting in 2009). A figure of these splits can be found in the Appendix.

Performance Metrics. Because only 3% of LQGs (around 300) are inspected every year, we use precision at the top 3% as our primary evaluation metric. This aligns with historical inspection patterns and our understanding of available resources.

Dealing with Missing Labels

Handler ID	Evaluation Year	Violation Flag	Predicted Probability
NY*********	2013	1	1
NY*********	2013	NA	0.95
NY*********	2013	0	0.91
NY*********	2013	NA	0.89

We train our model on labeled data (i.e., inspected facilities) but make predictions on all active facilities in each test set. Among the top 3% rankings on the test set, many facilities have missing labels, as illustrated in the table. We define our precision at top 3% to be calculated by Labeled True Positives / (Labeled True Positives + Labeled False Positives). The statistical significance of this precision depends on the amount of labeled data among the top 3%. We report the proportion of labeled data in our analysis.

Model Selection

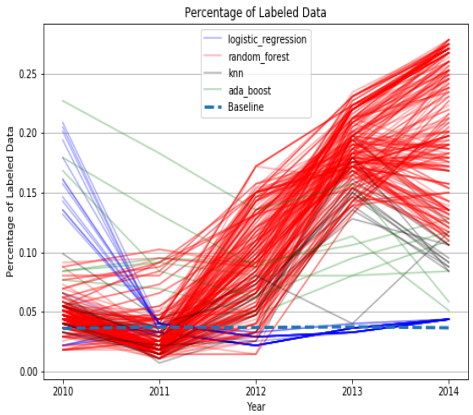
We evaluate our models on both efficiency and equity. We explore equity and fairness considerations later in this paper when choosing among our top performers. (An implication of this is that we first emphasize efficiency, and then use equity as a something of a tiebreaker). In terms of efficiency, we chose two performance metrics:

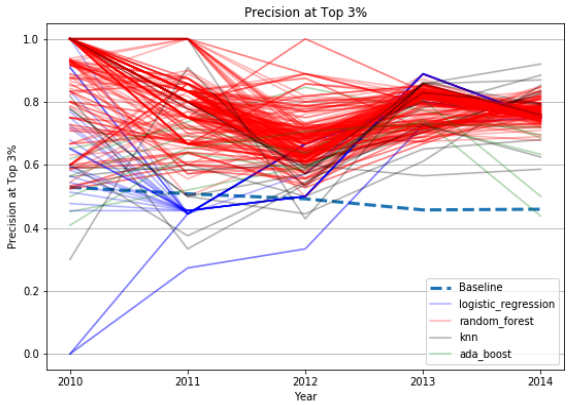
- **Highest mean precision**: we select the model with the highest average precision at 3% from 2010 to 2014.
- **Minimax precision**: we select the model with the highest minimum precision at 3% from 2010 to 2014. This metric is more conservative.

Link to GitHub repo: https://github.com/dssg/mlpolicylab_sp20_e2.git

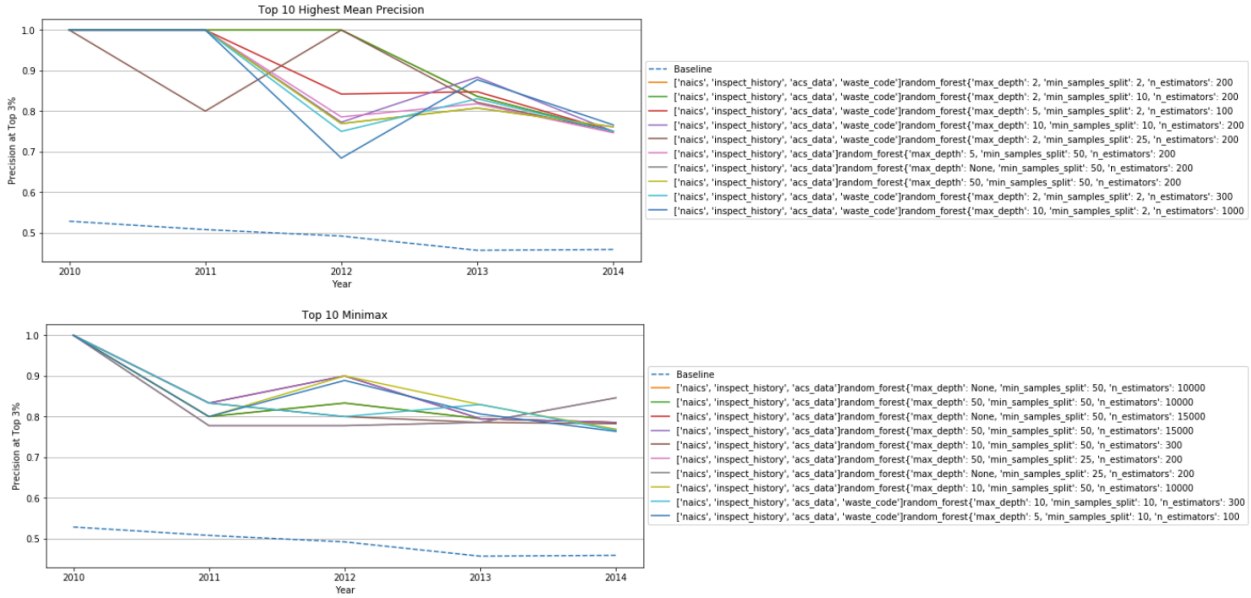
Model Evaluation

As shown in the plots below, most machine learning models outperform the random baseline model by a large margin, especially the Random Forest classifiers labeled in red. Somewhat surprisingly, the facilities most likely to violate according to our models have large proportions of labeled data. We infer that the inspection process of the EPA is not uniformly random, and our models capture this. For random forest classifiers, there is an increasing trend in the proportion of labeled data from 2012 to 2014. This might indicate that EPA improved their selection strategy in recent years. This could also represent an underlying issue with the prediction task (discussed in Limitations).





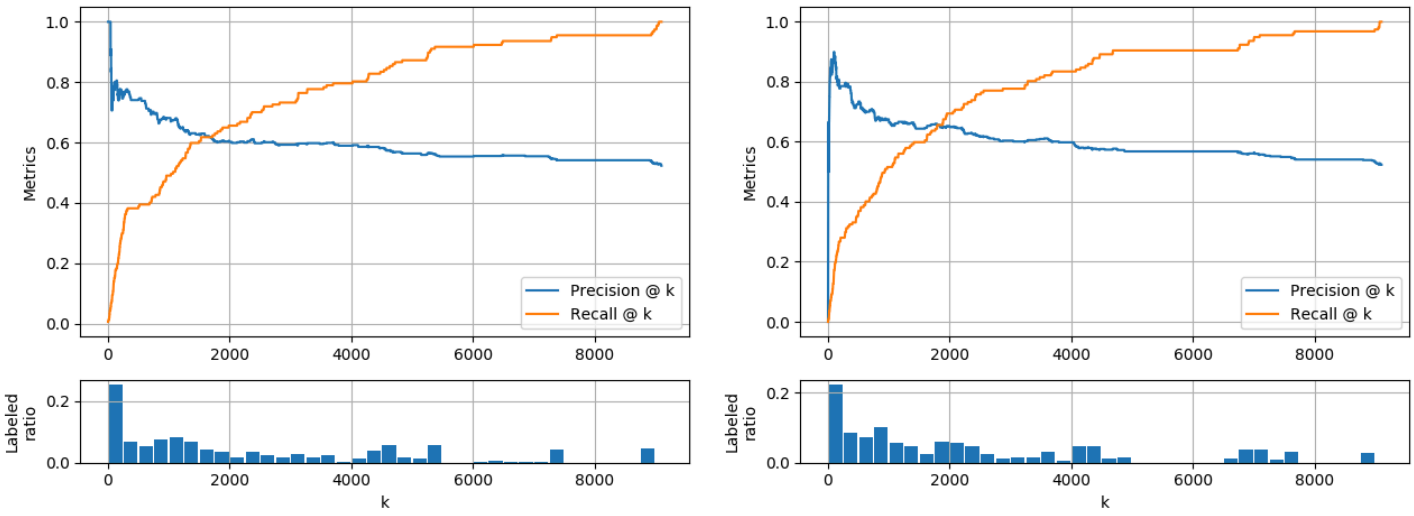
We present the top 10 models according to each metric below. We note that, in general, the minimax models offer less erratic behavior than the highest precision models. This may be of interest to decision makers when choosing which types of model to explore for recommendations. From each category (highest mean precision and minimax), we selected the top candidates for further analysis to see how they compared. We ultimately believe the decision maker can choose which to use, based on risk preferences.



The precision and recall at k curves for the two best models are shown below, along with the density of labeled data for the two best models. The models are trained on years 2009-2013 and evaluated on the last year (2014). The proportion of labeled data is computed by splitting test predictions by 250 from the top, (1~250, 251~500, and so on), and counting the number of labeled examples in each block.

Model 1: Highest Mean Prec @ 3%

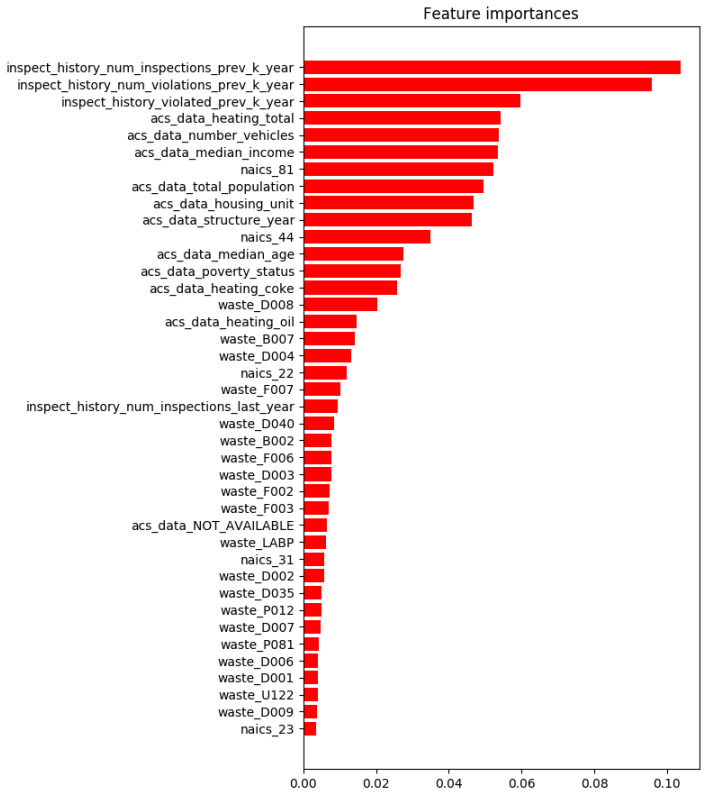
Model 2: Minimax Prec @ 3%



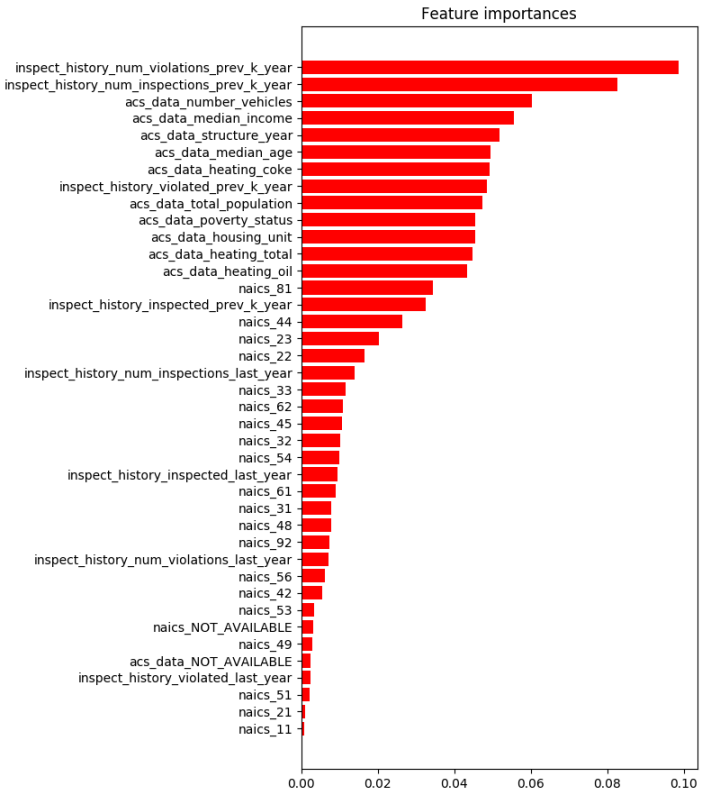
As we can see in the figures, both models have high precision at 3% (276 facilities). Since both models have a higher ratio of the labeled data, the precision at 3% is computed on many examples (>50 facilities) and has statistical significance. Decision-maker risk preferences may guide the final selection of either model.

Feature Importance and Crosstabs

For global interpretability, we plot feature importances for the two best models, up to the 40 most important features. For the highest mean precision model, the numbers of inspections and violations within the last 5 years ranked the top features and the ACS features, and then waste codes follow. Some of the NAICS codes also have significant importance. Such preference is also shown in the minimax model, but the minimax model does not have the waste code in its input features, so NAICS codes follow the ACS data.



Model 1: Highest Mean Prec @ 3%



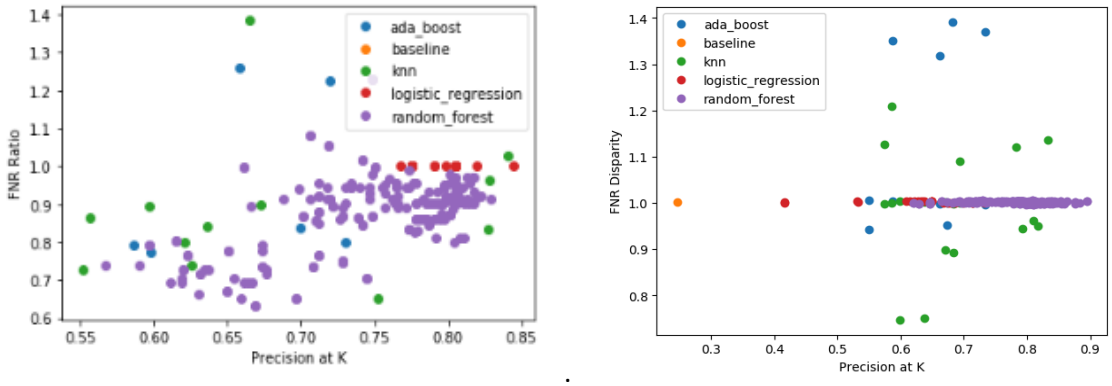
Model 2: Minimax Prec @ 3%

As another global interpretation measure of the best models, we compare the average values of the top 5 most important features on two groups: the top 3% and the remaining 97%. The results show that if a facility has been inspected or committed a violation before, it is a strong signal of future violations. Other signals include having less heating, more vehicles, more income, and old buildings in the county where the facility is located. It is unclear while lower consumption of heat is associated with high risk, but this could be driven by outliers. This may elicit further study.

Model 1: Highest Mean Prec @ 3%				Model 2: Minimax Prec @ 3%		
Features	**Top 3%**	**Bottom 97%**	**Rank**	**Features**	**Top 3%**	**Bottom 97%**
total inspections in 5 years	2.613	0.103	**1**	total violations in 5 years	1.691	0.075
total violations in 5 years	2.588	0.047	**2**	total inspections in 5 years	1.706	0.131
violated in 5 years	1.480	0.041	**3**	number vehicles (ACS)	309354.0	195816.4
heating total (ACS)	352463.2	504660.6	**4**	median income (ACS)	75936.3	57577.3
number vehicles (ACS)	306937.9	195891.1	**5**	structure year (ACS)	1944.55	1948.38

Fairness Audit

Because the project aims to affect policy, our team wanted to ensure that the decisions guided by this project were equitable. Since EPA regulations ultimately aim to protect human health, we wanted to highlight vulnerable populations. We protect zip codes ("communities") with above-average poverty rates. We reasoned these areas might possess fewer political or economic resources to ensure compliance. About one-third of counties in New York fit this description. (Drilling down, however, we later uncovered that these areas possessed relatively few facilities —about 7%.) We initially selected FNR disparity at k as our central equity metric; we emphasize the harm to communities that might come from failing to detect violations. While these metrics provide a proof of concept, we discuss later how alternative methods may be more informative at smaller values of k and illustrate those results.



FNR Disparity at 50% threshold vs FNR Disparity at k% (final metric)

The above plots from our final test evaluation year (2014) show performance on our two metrics of interest (precision at 3% and FNR disparity) as the axes. These represent tradeoffs between efficiency and equity. In each, there appears to be an assortment of outputs that can communicate the tradeoff between efficiency and equity. Most of the superior models (at the far right in the middle) were random forests, although several KNN and logistic regressions occasionally surfaced. While this only shows 2014 results, we did run the same analysis for our top minimax models.

Selecting a Fair Model

When analyzing our highest precision models on FNR, we observed that all five random forests had an FNR disparity ratio of just less than one. This demonstrates reasonable fairness. We would aim for a model that failed to inspect violations at reasonable levels equally across the board. Because our minimax scored 0.945 and our highest mean precision scored 0.928, we initially felt that either choice could be used as the best model, since there were few gross disparities. Disparities much less than one would actually discriminate against wealthier areas, which would represent unfairness in the opposite direction, potentially hurting public buy-in. Yet, disparities of substantially greater than one would mean

our vulnerable populations were being exposed to potentially more harmful violations, with perhaps less ability to react.

Discussion of Results

Random Forests as Dynamic High Performers. In general, we discovered that random forests run over large grids are very high performers. When we initially ran a small grid of 50 or so models, there were a variety of candidates of different model types that performed well. However, by increasing the depth of our trees as well as the number of estimators, while varying sample splits, random forests soon crowded out all other models as the top performers.

Feature Groups Affect Stability. We also noticed that our top two models used different feature groups and individual features when making their predictions. The top mean precision model used waste codes as a key predictor of risk, while the minimax model disregarded waste codes altogether and instead emphasized demographic characteristics. Perhaps these fixed features provided more stable estimates over time than those which vary year to year by type and quantity.

***Urban vs Rural Signifiers*.** Our feature importances picked up that denser, higher population areas are more likely to violate. (These areas might also have older buildings.) We believe in these cases that our models are mostly just finding the locations of waste management facilities, since most facilities are within these highly developed, industrial areas. This also has ramifications for our choice of a protected minority group.

Policy Recommendations

This project seeks to inform policy in a way that can improve the quality of life. A plan for a field trial is included below, followed by general policy recommendations that would follow a successful trial. Caveats and limitations are addressed later in the paper.

External Validation: A Field Trial

While our top models appear to perform well on our test data, these should be validated in a field trial. Field trials allow results to be tested in a real world setting and be evaluated based on actual performance.

Power and Minimum Necessary Sample Size. First, we must determine the necessary sample size that can provide us with sufficient power to conclude our model is useful. A common formula to calculate necessary sample size for binary outcomes is below:[6]

$$n = \frac{(Z_{\alpha/2}+Z_{\beta})^2 \times \{p_1(1-p_1)+p_2(1-p_2)\}}{(p_1-p_2)^2}$$

For our purposes, we will calculate Zα to be 1.65, corresponding with a one-sided t-test at 5% error. This test will be one-sided because we are only concerned if our model improves upon performance. We will

[6] Source: NIH, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2876926/

aim for 80% power, a common heuristic. We will test if p1, corresponding to the random baseline of 48% prevalence, is lower in a statistically significant way, from p2 at an estimated 0.79 (lowest minimax observed). This corresponds to 68 total individuals minimum for a power of 0.80 and 92 total individuals for a power of 0.90. Because there are some 300 facilities to select, we recommend at least a third of these be part of a field trial, although this can theoretically be expanded to include all 300 or so facilities (omitting pre-scheduled follow up visits).

Null Hypothesis: $p2 \leq p1$ where p is the proportion of violations found in either the random baseline (p1) or the model's recommendation (p2).

Implementation. Due to the above analysis, we recommend the EPA use at least third of its available facility inspections to conduct a field trial for this model. Half of these field samples would form a control group. These facilities would be selected at random from all possible LQGs. The second half would form our treatment group, composed of high-risk facilities flagged by our model (either in order of importance or at random from our top 3% of facilities). Then, following a year by which these facilities are inspected, the performance of these two selection methods can be compared. We expect that our model uncovered more violations than the random sample at a statistically significant level, indicating that we were more efficient. We might also see that our model is more equitable in terms of the communities it protects. This design serves two purposes. First, it would allow us to validate our model's findings. Secondly, it would encourage the EPA to explore unlabeled facilities in general, which is inherently useful. Pending a successful field trial, a few recommendations would result:

Other Recommendations

Use our top models to prioritize facilities at highest risk of violating. Our direct goal in the execution of this project is to provide the EPA with lists of facilities (LQGs) to inspect each year. The top 3% of probabilities from our best models can be used to guide this selection process. We recommend understanding similarities and differences within these model's recommendations and perhaps using a blend of risky facilities from the top models, especially where facilities overlap.

Invest generally in exploring facilities with unknown activity. As mentioned earlier, most large quantity generators have no inspection history with the EPA. Even if our best models were not immediately deployed, it is valuable to understand behaviors and characteristics for the 86% facilities with unknown activity, which will improve model performance.

Limitations, Caveats, and Future Work

Our analysis has several limitations which will be discussed at length below. Most relate to labeling issues, proxy variables, and the need to secure more (better) data and features.

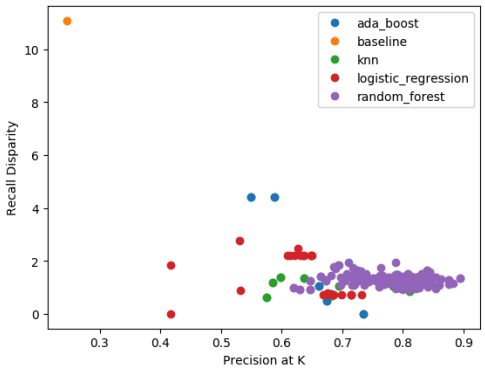
Caveats

***Zip code as a proxy for neighborhood statistics*.** The current analysis leans heavily on ACS data that is provided at the county level. Yet, there may be huge differences across this wide geographic area that

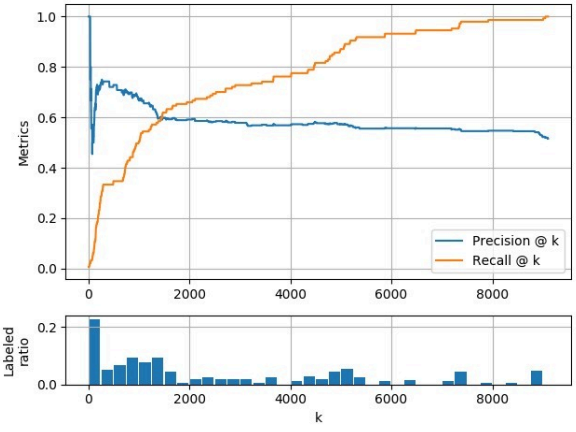
conceals more nuanced analyses. We would recommend geocoding these addresses in order to merge on smaller geographic units of analysis (e.g. block group level) to paint a clearer picture of variations even across zip codes.

***Revisiting protected groups*.** Our "high poverty" counties rarely ever had large quantity generators of waste. Therefore, there is some worry that we are trying to protect areas that are essentially rural or suburban from harm. Another way of thinking about this is that maybe denser areas need more focus, since these violations could potentially impact diverse populations in larger orders of magnitude than violations elsewhere. Finding more granular differences between these areas (e.g. matching to smaller units of geospatial analysis with ACS) could help us find vulnerable populations within these dense population areas, rather than ignoring the wealthier but heterogeneous areas for more sparse spaces.

Revisiting the fairness metric. Our models rarely exhibited strong FNR disparities, but this may be due to the fact that our top k% captures so few positive instances that FNR is close to 1 for both groups. (By definition, most positive instances will not be included at k%, making them false negatives). To investigate, we also examined recall disparity. Our team did not fully redo the analysis but noted that our top models had a recall disparity of around 1.30 on average. While not completely fair, these significantly outperformed the baseline. When evaluating the new results, we found an ideal model choice in a random forest (max_depth:5, min_sample splits:50, n_estimators:300), which managed a precision of 0.85, an FNR disparity o 0.998, a recall disparity of 0.963, and a minimum precision of 0.76 (competitive to the best minimax). Full output is in the appendix, but this finding underscores that random forests as versatile performers, although KNN models also perform well on recall.



Unequal Distribution of Labels. In the late stages of this project, our team discovered a bug in the labeling process. We believe some types of violations were not filtered out, and these could be heavily influenced by follow ups. While we had insufficient time to rerun the full analysis, we wanted to analyze performance on a revised dataset to determine its top 5 models. When we fixed this issue, we saw that labels were only slightly affected, and precision at 3% merely dipped by a few percentage points, although the best model did change from our first choice to our fourth choice. This underscores a central challenge to our prediction task.



The Difficulty of Predicting P(V) vs. P(V|Inspection). We sought to build a machine learning model that can predict the probability of violation. However, the paradox is that we know the evaluation result only after the EPA inspects the facility. Since our models also include the inspection history features, we suspect that often, we are actually predicting the probability of violation given that a facility is inspected. This could also be one of the reasons why the inspection records are not uniformly distributed among the ranking generated by our models.

Future Work

Develop an understanding of the "worst impact" violations. While beyond the scope of this current analysis, our team acknowledges that certain violations may be more urgent or more influential on human health than others. As we noted in our exploratory analysis phase, many of the most common violations are related to administrative issues and lack of documentation. While this is surely important (and may be related in fact to severe violations if fraud or carelessness drives the administrative offense), it may be important to prioritize violations which may directly impact human health (spillage of toxic waste or improper treatment). A list of these types of violations and their frequencies are in the appendix.

Add more generalizable features. As indicated by our feature importance plots, our model feels most confident in ascribing riskiness to facilities which have been inspected before. We recommend including a number of different geospatial, demographic, and business-specific features that may help explore other drivers beyond EPA's previous selection process.

***Create different models for different facility types over different time spans*.** The scope of this analysis is admittedly narrow in order to develop a good foundation. However, expanding modeling to non-LQGS and larger time frames will enable the EPA to widen its understanding and decision-making ability.

Appendix

NAICS Key

The following codes may be useful to understanding the industries discovered to be highly correlated with violations, as discovered by our feature importance models.

Sector	Description
11	Agriculture, Forestry, Fishing and Hunting (not covered in economic census)
21	Mining, Quarrying, and Oil and Gas Extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and Warehousing
51	Information
52	Finance and Insurance
53	Real Estate and Rental and Leasing
54	Professional, Scientific, and Technical Services
55	Management of Companies and Enterprises
56	Administrative and Support and Waste Management and Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, and Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration (not covered in economic census)

Waste Codes

A list of all waste codes and their corresponding materials and designations can be found here: https://essr.umd.edu/epa-hazardous-waste-codes

List of Features

Data Source		Feature Name	Type (categorical, continuous, binary, etc.)	Imputation method
rcra		naics	categorical	
rcra		inspected_last_year	binary	
rcra		violated_last_year	binary	
rcra		num_inspections_last_k_years	continuous	
rcra		num_violations_last_k_years	continuous	
nysdec		waste_code	categorical	
ACS_data		heating_coke	continuous	0
ACS_data		heating_oil	categorical	0
ACS_data		heating_total	continuous	0
ACS_data		housing_unit	continuous	0
ACS_data		median_age	continuous	0
ACS_data		median_income	continuous	0
ACS_data		number_vehicles	continuous	0
ACS_data		poverty_status	continuous	0
ACS_data		structure_year	continuous	0
ACS_data		total_population	continuous	0

Model grid

We run a large grid with a total number of 282 different models.
We have 5 feature groups in total, including NAICS code, inspection history, ACS data and waste code. On top of different classifiers, we have two groups of machine learning models with different feature groups combination, one contains NAICS code, inspection history, ACS data, the other contains inspection history, ACS data and waste code.
For each feature group combination, we train 141 different models as summarized below.

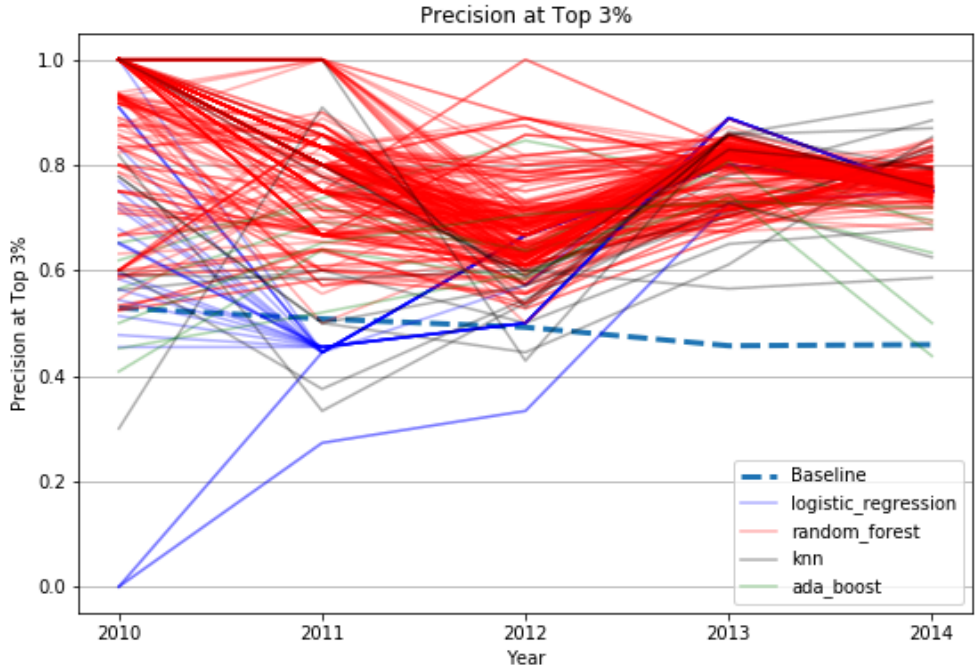
- Logistic Regression (12 models)
    - Solver: saga
    - Maximum Iterations: 5000
    - Penalty: L1, L2 (for both penalty terms, we used ‘saga’ as the solver)
    - Regularization parameter: 0.001, 0.01, 0.1, 1, 2, 10
- Random Forest (120 models)
    - Number of features to consider when looking for the best split: square root of the total number of features

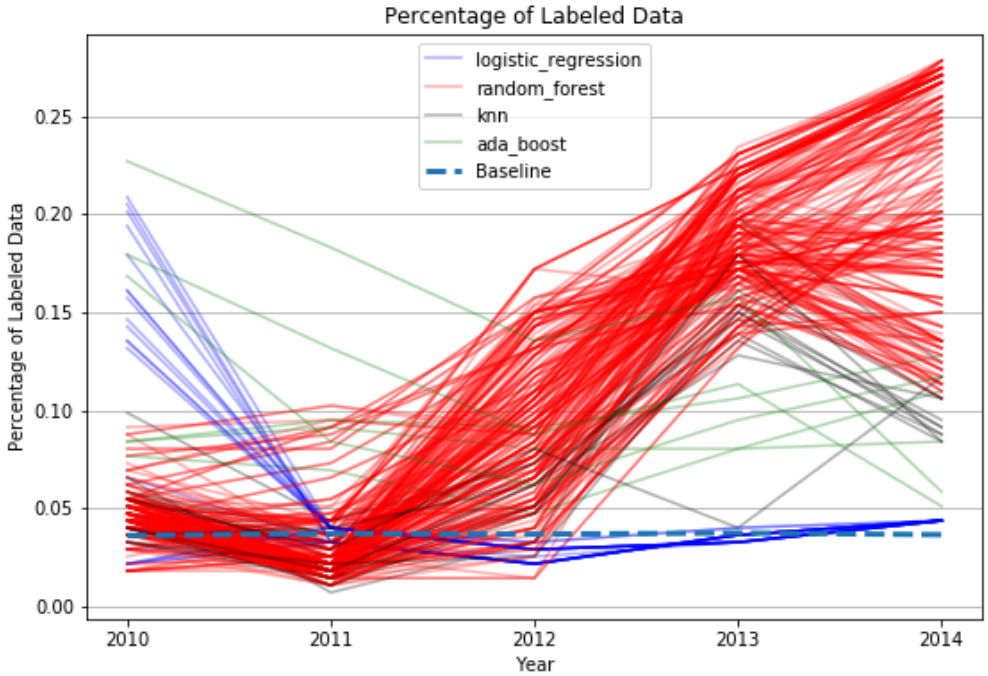
    - Maximum depth of trees: 2, 5, 10, 50, None
    - Number of trees: 100, 200, 300, 1000, 10000, 15000
    - Minimum number of samples required to split an internal node: 2, 10, 25, 50
- K Nearest Neighbor (6 models)
    - Number of neighbors: 5, 10, 15, 20, 25, 30
- Adaptive Boosting (3 models)
    - Number of trees: 50, 100, 200

Train and Test Splits

	Train Set					**Test Set**				
Train-Test Pair ID	Start date for rows	End date for rows	Interval b/w dates for rows	Start date for labels	End date for labels	Start date for rows	End date for rows	Interval b/w dates for rows	Start date for labels	End date for labels
1	2009-01-01	2009-12-31	1 year	2010-01-01	2009-12-31	2010-01-01	2010-12-31	1 year	2010-01-01	2010-12-31
2	2009-01-01	2010-12-31	2 years	2010-01-01	2010-12-31	2011-01-01	2011-12-31	1 year	2011-01-01	2011-12-31
3	2009-01-01	2011-12-31	3 years	2010-01-01	2011-12-31	2012-01-01	2012-12-31	1 year	2012-01-01	2012-12-31
4	2009-01-01	2012-12-31	4 years	2010-01-01	2012-12-31	2013-01-01	2013-12-31	1 year	2013-01-01	2013-12-31
5	2009-01-01	2013-12-31	5 years	2010-01-01	2013-12-31	2014-01-01	2014-12-31	1 year	2014-01-01	2014-12-31

Temporal graph of our primary evaluation metrics (Precision at top 3% and Percentage of Labeled data)


Precision at Top 3%
Precision at Top 3%
1.0
0.8
0.6
0.4
0.2
0.0
2010
2011
2012
2013
2014
Year
Baseline
logistic_regression
random_forest
knn
ada_boost


Percentage of Labeled Data
Percentage of Labeled Data
0.25
0.20
0.15
0.10
0.05
0.00
2010
2011
2012
2013
2014
Year
logistic_regression
random_forest
knn
ada_boost
Baseline

Criteria used to select top models

We use two strategies to select our top models:

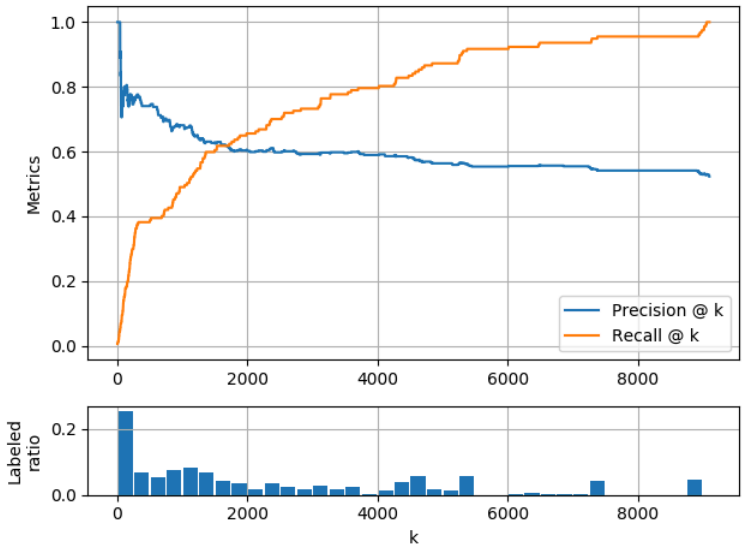
- Highest mean precision over years
- Minimax rule: select the models with the highest minimum precision over 2010 to 2014

Top Models

Here, we attach the details (features and model parameters) of the top 4 models of highest mean precision at 3% and the best minimax precision at 3%. We also provide the analysis results of these models for reference. In the tables, top_mean means the

1) Model 1 (top-1 mean precision model)

- Features: NAICS, inspection history, ACS data, waste code
- Model: random forest (max_depth: 2, min_samples_split: 10, n_estimator: 200)
- Precision/Recall@k graph with the ratio of labeled examples



List of feature importance of all features & Cross-tabs for 10 most different features

feature	importance	top_mean	bottom_mean
inspect_history_num_inspections_prev_k_year	0.103752432	2.612570178	0.103316281
inspect_history_num_violations_prev_k_year	0.095776715	2.587928746	0.047370873
inspect_history_violated_prev_k_year	0.059719783	1.47985348	0.041336353
ACS_data_heating_total	0.054279858	352463.2234	504660.5934
ACS_data_number_vehicles	0.053633113	306937.8571	195891.0685
ACS_data_median_income	0.05354779	70659.98901	57740.29932
naics_81	0.052222338	0.142857143	0.055719139
ACS_data_total_population	0.049424611	1006107.949	1341683.996
ACS_data_housing_unit	0.046802004	391664.0147	557946.898
ACS_data_structure_year	0.046300827	1959.520147	1947.922084
naics_44	0.034969606	0.153846154	0.070328426
ACS_data_median_age	0.027621946	39.49157509	37.7000906
ACS_data_poverty_status	0.026811561	109889.6044	240491.589

ACS_data_heating_coke	0.025728073	280.8205128	444.211325
waste_D008	0.020291997	0.457875458	0.264212911
ACS_data_heating_oil	0.014658181	29764.43223	49535.7436
waste_B007	0.014048651	0.029304029	0.029558324
waste_D004	0.01306805	0.087912088	0.020838052
naics_22	0.011839076	0.032967033	0.40792752
waste_F007	0.010264643	0.058608059	0.001585504
inspect_history_num_inspections_last_year	0.009321016	0.776532626	0.020082669
waste_D040	0.008550318	0.043956044	0.007587769
waste_B002	0.007812721	0.010989011	0.013476784
waste_F006	0.007668717	0.128205128	0.008607022
waste_D003	0.007659718	0.274725275	0.034541336
waste_F002	0.007285937	0.179487179	0.026613817
waste_F003	0.007026828	0.798534799	0.066930917
ACS_data_NOT_AVAILABLE	0.006455847	0	0.002151755
waste_LABP	0.006164233	0.080586081	0.009399773
naics_31	0.005801053	0.021978022	0.004983012
waste_D002	0.00561818	1.296703297	0.181540204
waste_D035	0.005100425	0.336996337	0.039297848
waste_P012	0.004902724	0.003663004	0.003624009
waste_D007	0.004697842	0.472527473	0.08391846
waste_P081	0.004234947	0.098901099	0.0398641
waste_D006	0.004051478	0.197802198	0.029445074
waste_D001	0.003984605	2.311355311	0.36387316
waste_U122	0.003947953	0.051282051	0.006228766
waste_D009	0.00375936	0.278388278	0.070328426
naics_23	0.003444492	0.043956044	0.244960362
waste_D018	0.003309302	0.19047619	0.034994337
waste_D029	0.002661219	0	0.000906002
waste_U089	0.00265974	0.003663004	0.000339751
naics_32	0.002382593	0.274725275	0.035447339
waste_B004	0.002320399	0.018315018	0.011211778
inspect_history_inspected_prev_k_year	0.002286337	1.695970696	0.137259343
waste_P001	0.002238032	0.227106227	0.080407701
waste_F039	0.002227963	0	0.003624009
waste_D010	0.002154988	0.157509158	0.043261608
naics_92	0.002090316	0.025641026	0.042468856
waste_D005	0.001904951	0.238095238	0.039071348
waste_D039	0.001897555	0.06959707	0.009173273
naics_33	0.001841947	0.366300366	0.04801812
waste_RARE	0.001786752	0.135531136	0.017100793
waste_P030	0.001786671	0.029304029	0.00407701
waste_P042	0.001612021	0.025641026	0.006681767
waste_U112	0.001554454	0.014652015	0.001245753
naics_42	0.001521245	0.054945055	0.018686297
inspect_history_num_violations_last_year	0.001502722	0.758358601	0.007869387
waste_U058	0.001441981	0.021978022	0.011325028
waste_U044	0.001360344	0.029304029	0.006568516
waste_U226	0.001234228	0	0.000679502
waste_F005	0.001232585	0.468864469	0.040543601
waste_D019	0.001212838	0.010989011	0.00385051
waste_D026	0.001202555	0.087912088	0.010645527
waste_D043	0.001074617	0	0.002265006
waste_D022	0.001069688	0.062271062	0.014156285

waste_U059	0.001042635	0.007326007	0.001698754
naics_56	0.001024377	0.032967033	0.034314836
waste_U077	0.001022353	0	0.000453001
waste_U075	0.000998019	0	0.000226501
waste_U080	0.000920208	0.018315018	0.003057758
waste_P003	0.000824768	0.003663004	0.000792752
waste_P075	0.00074743	0.131868132	0.054926387
waste_D027	0.000728973	0.007326007	0.001698754
waste_F008	0.000709132	0.014652015	0.000792752
waste_D034	0.000705475	0	0.000792752
waste_U052	0.000686083	0.010989011	0.000339751
waste_U072	0.000668972	0	0.000566251
naics_72	0.000656896	0	0.001925255
waste_D011	0.000575483	0.355311355	0.093657984
waste_P022	0.000519702	0.003663004	0.000792752
naics_49	0.000488568	0.007326007	0.005436014
waste_U151	0.000481286	0.010989011	0.001359003
waste_F009	0.000464564	0.021978022	0.000566251
waste_P119	0.000462526	0	0.00011325
waste_P098	0.000418651	0.018315018	0.00396376
waste_U035	0.000403399	0.025641026	0.010305776
naics_54	0.000338114	0.095238095	0.011211778
naics_NOT_AVAILABLE	0.000336376	0.007326007	0.011778029
inspect_history_inspected_last_year	0.000322529	0.344322344	0.023669309
naics_61	0.000318321	0.054945055	0.046998867
waste_B005	0.000309049	0.003663004	0.001585504
waste_U010	0.000307848	0.007326007	0.005436014
waste_U159	0.000290665	0.010989011	0.001698754
waste_U002	0.000281306	0.073260073	0.010305776
waste_U150	0.000257731	0	0.002491506
waste_B006	0.000254534	0	0.001698754
waste_U220	0.000191331	0.014652015	0.002038505
waste_U211	0.000179283	0	0.000792752
waste_F004	0.000169845	0	0.001019253
waste_U201	7.59E-05	0	0.000679502
waste_D021	0	0.007326007	0.002491506
waste_D013	0	0.003663004	0.001019253
waste_D016	0	0.043956044	0.010192525
waste_U404	0	0.007326007	0.000566251
waste_D012	0	0	0.000453001
waste_B003	0	0	0.002378256
waste_B001	0	0	0.000906002
waste_D024	0	0.047619048	0.018006795
inspect_history_violated_last_year	0	0.322344322	0.006115515
naics_71	0	0.003663004	0.004530011
naics_62	0	0.065934066	0.014835787
naics_55	0	0	0.000679502
naics_53	0	0.021978022	0.022310306
naics_52	0	0.003663004	0.000906002
naics_51	0	0.021978022	0.004643262
naics_48	0	0.047619048	0.066477916
naics_45	0	0.04029304	0.015402039
naics_21	0	0	0.001812005
waste_D023	0	0.007326007	0.000566251

waste_F027	0	0	0.000453001
waste_D025	0	0.003663004	0.001245753
waste_U133	0	0	0.000566251
waste_U161	0	0	0.000566251
waste_U160	0	0.003663004	0.000566251
waste_U154	0	0.043956044	0.00385051
waste_U144	0	0.003663004	0.000679502
waste_U135	0	0.007326007	0.000566251
waste_U134	0	0.003663004	0.000679502
waste_U132	0	0	0.000679502
waste_U165	0	0	0.001472254
waste_U129	0	0.007326007	0.00419026
waste_U123	0	0.021978022	0.001132503
waste_U117	0	0.003663004	0.000566251
waste_U115	0	0	0.00011325
waste_U108	0	0.003663004	0.001019253
waste_U103	0	0	0.000226501
waste_U162	0	0.003663004	0.000906002
waste_U188	0	0.054945055	0.00419026
waste_U069	0	0	0.000339751
waste_U223	0	0	0.000226501
waste_U248	0	0	0.00011325
waste_U246	0	0	0.000339751
waste_U240	0	0	0.000226501
waste_U239	0	0.036630037	0.001812005
waste_U236	0	0.003663004	0.000339751
waste_U228	0	0	0.000906002
waste_U219	0	0	0.00011325
waste_U196	0	0.018315018	0.000566251
waste_U218	0	0	0.00011325
waste_U213	0	0.003663004	0.001472254
waste_U210	0	0.007326007	0.001132503
waste_U206	0	0	0.000906002
waste_U205	0	0	0.001019253
waste_U202	0	0.003663004	0.00011325
waste_U070	0	0	0.000453001
waste_U056	0	0	0.000679502
waste_D028	0	0.021978022	0.005436014
waste_P005	0	0	0.000226501
waste_P046	0	0.003663004	0.000792752
waste_P029	0	0.010989011	0.000453001
waste_P028	0	0.010989011	0
waste_P014	0	0.003663004	0.000566251
waste_P010	0	0	0.000566251
waste_P008	0	0	0.000566251
waste_U279	0	0.007326007	0.001925255
waste_P076	0	0	0.000226501
waste_F019	0	0.003663004	0.000792752
waste_F001	0	0.054945055	0.009060023
waste_D038	0	0.043956044	0.006908267
waste_D036	0	0.007326007	0.000453001
waste_D033	0	0	0.001132503
waste_D032	0	0	0.000226501
waste_P048	0	0.003663004	0.000339751

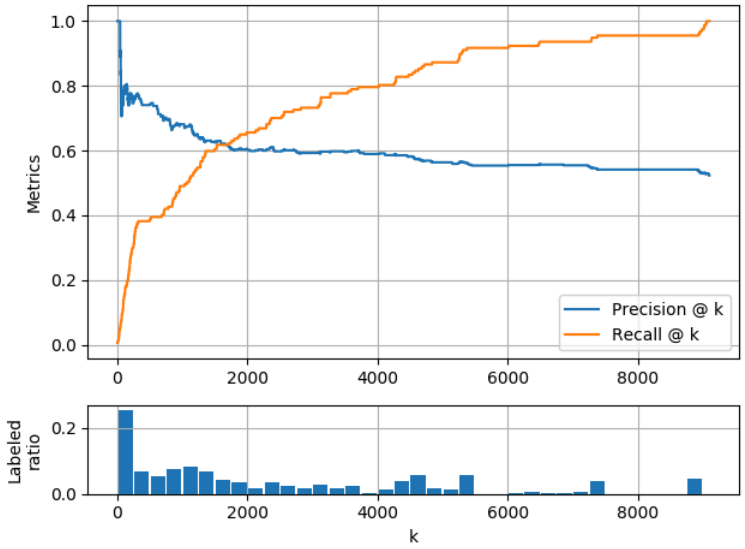
waste_P077	0	0.003663004	0.000566251
waste_U037	0	0.003663004	0.000906002
waste_U003	0	0.007326007	0.001132503
waste_U034	0	0.003663004	0.000566251
waste_U031	0	0.003663004	0.001132503
waste_U019	0	0.007326007	0.000906002
waste_U015	0	0.003663004	0.000566251
waste_U012	0	0	0.000226501
waste_U007	0	0	0.001132503
waste_U001	0	0.003663004	0.001019253
waste_P087	0	0.007326007	0.001925255
waste_P204	0	0.003663004	0.000566251
waste_P188	0	0.073260073	0.037599094
waste_P120	0	0.003663004	0.000566251
waste_P108	0	0	0.000226501
waste_P106	0	0.003663004	0.001245753
waste_P105	0	0.010989011	0.00396376
naics_11	0	0.003663004	0.00407701

FNR Disparity: 0.928

2) Model 2 (top-2 mean precision model)

- Features: NAICS, inspection history, ACS data, waste code
- Model: random forest (max_depth: 2, min_samples_split: 2, n_estimator: 200)

- Precision/Recall@k graph with the ratio of labeled examples



List of feature importance of all features & Cross-tabs for 10 most different features

feature	importance	top_mean	bottom_mean
inspect_history_num_inspections_prev_k_year	0.103752432	2.612570178	0.103316281
inspect_history_num_violations_prev_k_year	0.095776715	2.587928746	0.047370873
inspect_history_violated_prev_k_year	0.059719783	1.47985348	0.041336353
ACS_data_heating_total	0.054279858	352463.2234	504660.5934

ACS_data_number_vehicles	0.053633113	306937.8571	195891.0685
ACS_data_median_income	0.05354779	70659.98901	57740.29932
naics_81	0.052222338	0.142857143	0.055719139
ACS_data_total_population	0.049424611	1006107.949	1341683.996
ACS_data_housing_unit	0.046802004	391664.0147	557946.898
ACS_data_structure_year	0.046185127	1959.520147	1947.922084
naics_44	0.034969606	0.153846154	0.070328426
ACS_data_median_age	0.028683918	39.49157509	37.7000906
ACS_data_poverty_status	0.026811561	109889.6044	240491.589
ACS_data_heating_coke	0.025728073	280.8205128	444.211325
waste_D008	0.020291997	0.457875458	0.264212911
ACS_data_heating_oil	0.014658181	29764.43223	49535.7436
waste_B007	0.014048651	0.029304029	0.029558324
waste_D004	0.01306805	0.087912088	0.020838052
naics_22	0.011839076	0.032967033	0.40792752
waste_F007	0.010264643	0.058608059	0.001585504
inspect_history_num_inspections_last_year	0.009321016	0.776532626	0.020082669
waste_D040	0.007818445	0.043956044	0.007587769
waste_B002	0.007812721	0.010989011	0.013476784
waste_F006	0.007668717	0.128205128	0.008607022
waste_D003	0.007659718	0.274725275	0.034541336
waste_F002	0.007285937	0.179487179	0.026613817
waste_F003	0.007183465	0.798534799	0.066930917
ACS_data_NOT_AVAILABLE	0.006455847	0	0.002151755
waste_LABP	0.006164233	0.080586081	0.009399773
naics_31	0.005801053	0.021978022	0.004983012
waste_D002	0.00561818	1.296703297	0.181540204
waste_P012	0.004902724	0.003663004	0.003624009
waste_D035	0.004770326	0.336996337	0.039297848
waste_D007	0.004697842	0.472527473	0.08391846
waste_P081	0.004234947	0.098901099	0.0398641
waste_D006	0.004051478	0.197802198	0.029445074
waste_D001	0.003984605	2.311355311	0.36387316
waste_U122	0.003907016	0.051282051	0.006228766
waste_D009	0.00375936	0.278388278	0.070328426
naics_23	0.003444492	0.043956044	0.244960362
waste_D018	0.003309302	0.19047619	0.034994337
waste_D029	0.002661219	0	0.000906002
waste_U089	0.00265974	0.003663004	0.000339751
naics_32	0.002382593	0.274725275	0.035447339
waste_B004	0.002320399	0.018315018	0.011211778
inspect_history_inspected_prev_k_year	0.002286337	1.695970696	0.137259343
waste_P001	0.002238032	0.227106227	0.080407701
waste_F039	0.002227963	0	0.003624009
waste_D010	0.002154988	0.157509158	0.043261608
naics_92	0.002090316	0.025641026	0.042468856
waste_D005	0.001904951	0.238095238	0.039071348
waste_D039	0.001897555	0.06959707	0.009173273
naics_33	0.001841947	0.366300366	0.04801812
waste_RARE	0.001786752	0.135531136	0.017100793
waste_P030	0.001786671	0.029304029	0.00407701
waste_P042	0.001612021	0.025641026	0.006681767
waste_U112	0.001554454	0.014652015	0.001245753
naics_42	0.001521245	0.054945055	0.018686297

inspect_history_num_violations_last_year	0.001502722	0.758358601	0.007869387
waste_U058	0.001441981	0.021978022	0.011325028
waste_U044	0.001360344	0.029304029	0.006568516
waste_U226	0.001234228	0	0.000679502
waste_F005	0.001232585	0.468864469	0.040543601
waste_D019	0.001212838	0.010989011	0.00385051
waste_D026	0.001202555	0.087912088	0.010645527
waste_D043	0.001074617	0	0.002265006
waste_D022	0.001069688	0.062271062	0.014156285
waste_U059	0.001042635	0.007326007	0.001698754
naics_56	0.001024377	0.032967033	0.034314836
waste_U077	0.001022353	0	0.000453001
waste_U075	0.000998019	0	0.000226501
waste_U080	0.000920208	0.018315018	0.003057758
waste_P003	0.000824768	0.003663004	0.000792752
waste_P075	0.00074743	0.131868132	0.054926387
waste_D027	0.000728973	0.007326007	0.001698754
waste_F008	0.000709132	0.014652015	0.000792752
waste_D034	0.000705475	0	0.000792752
waste_U052	0.000686083	0.010989011	0.000339751
waste_U072	0.000668972	0	0.000566251
naics_72	0.000656896	0	0.001925255
waste_D011	0.000575483	0.355311355	0.093657984
waste_P022	0.000519702	0.003663004	0.000792752
naics_49	0.000488568	0.007326007	0.005436014
waste_U151	0.000481286	0.010989011	0.001359003
waste_F009	0.000464564	0.021978022	0.000566251
waste_P119	0.000462526	0	0.00011325
waste_P098	0.000418651	0.018315018	0.00396376
waste_U035	0.000403399	0.025641026	0.010305776
naics_54	0.000338114	0.095238095	0.011211778
naics_NOT_AVAILABLE	0.000336376	0.007326007	0.011778029
inspect_history_inspected_last_year	0.000322529	0.344322344	0.023669309
naics_61	0.000318321	0.054945055	0.046998867
waste_B005	0.000309049	0.003663004	0.001585504
waste_U010	0.000307848	0.007326007	0.005436014
waste_U159	0.000290665	0.010989011	0.001698754
waste_U002	0.000281306	0.073260073	0.010305776
waste_U150	0.000257731	0	0.002491506
waste_B006	0.000254534	0	0.001698754
waste_U220	0.000191331	0.014652015	0.002038505
waste_U211	0.000179283	0	0.000792752
waste_F004	0.000169845	0	0.001019253
waste_U201	7.59E-05	0	0.000679502
waste_D021	0	0.007326007	0.002491506
waste_D013	0	0.003663004	0.001019253
waste_D016	0	0.043956044	0.010192525
waste_U404	0	0.007326007	0.000566251
waste_D012	0	0	0.000453001
waste_B003	0	0	0.002378256
waste_B001	0	0	0.000906002
waste_D024	0	0.047619048	0.018006795
inspect_history_violated_last_year	0	0.322344322	0.006115515
naics_71	0	0.003663004	0.004530011

naics_62	0	0.065934066	0.014835787
naics_55	0	0	0.000679502
naics_53	0	0.021978022	0.022310306
naics_52	0	0.003663004	0.000906002
naics_51	0	0.021978022	0.004643262
naics_48	0	0.047619048	0.066477916
naics_45	0	0.04029304	0.015402039
naics_21	0	0	0.001812005
waste_D023	0	0.007326007	0.000566251
waste_F027	0	0	0.000453001
waste_D025	0	0.003663004	0.001245753
waste_U133	0	0	0.000566251
waste_U161	0	0	0.000566251
waste_U160	0	0.003663004	0.000566251
waste_U154	0	0.043956044	0.00385051
waste_U144	0	0.003663004	0.000679502
waste_U135	0	0.007326007	0.000566251
waste_U134	0	0.003663004	0.000679502
waste_U132	0	0	0.000679502
waste_U165	0	0	0.001472254
waste_U129	0	0.007326007	0.00419026
waste_U123	0	0.021978022	0.001132503
waste_U117	0	0.003663004	0.000566251
waste_U115	0	0	0.00011325
waste_U108	0	0.003663004	0.001019253
waste_U103	0	0	0.000226501
waste_U162	0	0.003663004	0.000906002
waste_U188	0	0.054945055	0.00419026
waste_U069	0	0	0.000339751
waste_U223	0	0	0.000226501
waste_U248	0	0	0.00011325
waste_U246	0	0	0.000339751
waste_U240	0	0	0.000226501
waste_U239	0	0.036630037	0.001812005
waste_U236	0	0.003663004	0.000339751
waste_U228	0	0	0.000906002
waste_U219	0	0	0.00011325
waste_U196	0	0.018315018	0.000566251
waste_U218	0	0	0.00011325
waste_U213	0	0.003663004	0.001472254
waste_U210	0	0.007326007	0.001132503
waste_U206	0	0	0.000906002
waste_U205	0	0	0.001019253
waste_U202	0	0.003663004	0.00011325
waste_U070	0	0	0.000453001
waste_U056	0	0	0.000679502
waste_D028	0	0.021978022	0.005436014
waste_P005	0	0	0.000226501
waste_P046	0	0.003663004	0.000792752
waste_P029	0	0.010989011	0.000453001
waste_P028	0	0.010989011	0
waste_P014	0	0.003663004	0.000566251
waste_P010	0	0	0.000566251
waste_P008	0	0	0.000566251

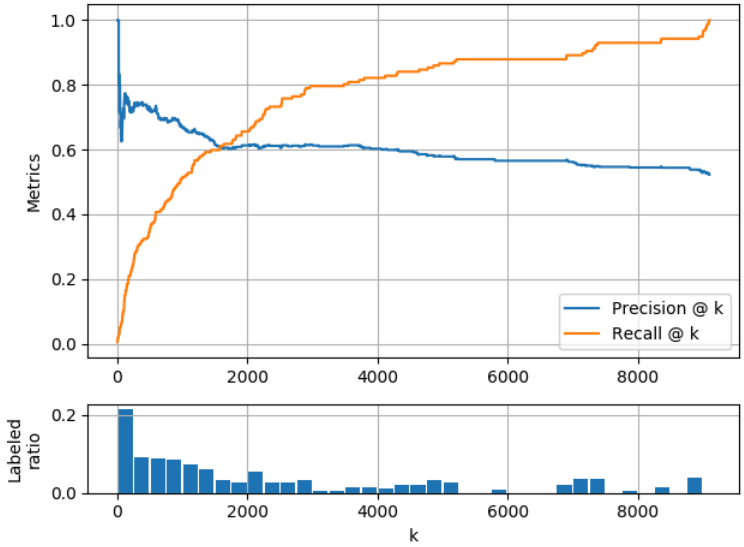
waste_U279	0	0.007326007	0.001925255
waste_P076	0	0	0.000226501
waste_F019	0	0.003663004	0.000792752
waste_F001	0	0.054945055	0.009060023
waste_D038	0	0.043956044	0.006908267
waste_D036	0	0.007326007	0.000453001
waste_D033	0	0	0.001132503
waste_D032	0	0	0.000226501
waste_P048	0	0.003663004	0.000339751
waste_P077	0	0.003663004	0.000566251
waste_U037	0	0.003663004	0.000906002
waste_U003	0	0.007326007	0.001132503
waste_U034	0	0.003663004	0.000566251
waste_U031	0	0.003663004	0.001132503
waste_U019	0	0.007326007	0.000906002
waste_U015	0	0.003663004	0.000566251
waste_U012	0	0	0.000226501
waste_U007	0	0	0.001132503
waste_U001	0	0.003663004	0.001019253
waste_P087	0	0.007326007	0.001925255
waste_P204	0	0.003663004	0.000566251
waste_P188	0	0.073260073	0.037599094
waste_P120	0	0.003663004	0.000566251
waste_P108	0	0	0.000226501
waste_P106	0	0.003663004	0.001245753
waste_P105	0	0.010989011	0.00396376
naics_11	0	0.003663004	0.00407701

FNR Disparity: 0.928

3) Model 3 (top-3 mean precision model)

- Features: NAICS, inspection history, ACS data, waste code
- Model: random forest (max_depth: 5, min_samples_split: 2, n_estimator: 100)

- Precision/Recall@k graph with the ratio of labeled examples



List of feature importance of all features & Cross-tabs for 10 most different features

feature	importance	top_mean	bottom_mean
inspect_history_num_violations_prev_k_year	0.091785706	2.121893993	0.061779422
inspect_history_num_inspections_prev_k_year	0.065645665	2.142016633	0.117864539
inspect_history_violated_prev_k_year	0.065064828	1.282051282	0.047451869
ACS_data_housing_unit	0.048420699	333169.2308	559755.4003
ACS_data_number_vehicles	0.04477905	267020.7875	197125.1976
ACS_data_heating_total	0.043846538	300328.8168	506272.4499
ACS_data_heating_coke	0.041970956	257.989011	444.917214
ACS_data_structure_year	0.041301391	1952.32967	1948.144394
ACS_data_total_population	0.039264159	856072.1392	1346322.702
naics_81	0.034693901	0.282051282	0.051415629
ACS_data_median_income	0.034277446	68432.85714	57809.15629
ACS_data_median_age	0.02633331	39.65274725	37.69510759
ACS_data_poverty_status	0.024725703	91730.46886	241053.021
naics_44	0.020540656	0.289377289	0.066138165
waste_D004	0.016999833	0.029304029	0.022650057
ACS_data_heating_oil	0.016045247	25707.94139	49661.15946
waste_D001	0.015351279	2.131868132	0.369422424
inspect_history_num_inspections_last_year	0.01506669	0.6536324	0.023882416
waste_D008	0.014142288	0.3003663	0.269082673
waste_D007	0.011524722	0.388278388	0.086523216
waste_D002	0.011132094	1.036630037	0.189580974
waste_F002	0.010540175	0.153846154	0.027406569
naics_22	0.01020766	0.018315018	0.408380521
inspect_history_inspected_prev_k_year	0.010144579	1.457875458	0.144620612
waste_B007	0.008745301	0.029304029	0.029558324
waste_F007	0.008550606	0.043956044	0.002038505
waste_F003	0.008305205	0.677655678	0.070668177
waste_D035	0.007668829	0.241758242	0.042242356
naics_23	0.007199001	0.032967033	0.245300113
naics_45	0.006671963	0.029304029	0.015741789
waste_D010	0.00620224	0.238095238	0.040770102
waste_D003	0.006034636	0.205128205	0.036693092
naics_92	0.006027777	0.018315018	0.042695357

waste_D009	0.005934015	0.351648352	0.06806342
waste_D006	0.00583763	0.102564103	0.032389581
waste_D040	0.005807014	0.032967033	0.00792752
waste_B002	0.005682426	0.010989011	0.013476784
inspect_history_num_violations_last_year	0.00560019	0.643075378	0.011433636
waste_F005	0.005460553	0.391941392	0.042921857
waste_F006	0.005294722	0.102564103	0.009399773
naics_31	0.00527582	0.021978022	0.004983012
waste_U122	0.005193034	0.076923077	0.005436014
naics_56	0.004720347	0.014652015	0.034881087
waste_P081	0.00437826	0.252747253	0.035107588
waste_D005	0.004230384	0.142857143	0.042015855
inspect_history_inspected_last_year	0.004141615	0.3003663	0.025028313
naics_33	0.003826741	0.32967033	0.049150623
waste_LABP	0.003425293	0.073260073	0.009626274
waste_D018	0.003420317	0.15018315	0.036240091
waste_U010	0.003321753	0	0.005662514
waste_D019	0.003154543	0.007326007	0.00396376
waste_D039	0.003010852	0.054945055	0.009626274
ACS_data_NOT_AVAILABLE	0.002908836	0.003663004	0.002038505
waste_B004	0.002885212	0.010989011	0.011438279
waste_D011	0.002857927	0.567765568	0.087089468
waste_U059	0.002817861	0	0.001925255
waste_RARE	0.002801777	0.113553114	0.017780294
naics_32	0.002777715	0.234432234	0.036693092
waste_U052	0.002645132	0.010989011	0.000339751
naics_48	0.002491823	0.021978022	0.067270668
waste_F039	0.002376338	0	0.003624009
waste_U058	0.002304526	0.051282051	0.010419026
naics_54	0.002272085	0.080586081	0.011664779
waste_P030	0.002230311	0.018315018	0.004416761
naics_49	0.002179445	0	0.005662514
waste_U044	0.002074605	0.003663004	0.007361268
waste_D022	0.001956069	0.04029304	0.014835787
waste_P098	0.00190074	0.010989011	0.00419026
waste_P001	0.001899049	0.512820513	0.071574179
waste_U035	0.001896856	0.051282051	0.009513024
waste_P012	0.001812055	0	0.003737259
waste_D038	0.001665994	0.014652015	0.00781427
naics_61	0.001585601	0.047619048	0.047225368
waste_D026	0.001565019	0.113553114	0.009852775
waste_U123	0.001514595	0.010989011	0.001472254
waste_P075	0.00148399	0.271062271	0.050622877
naics_62	0.001367408	0.054945055	0.015175538
waste_D024	0.001366213	0.098901099	0.016421291
waste_U154	0.001341426	0.029304029	0.004303511
waste_P042	0.001319064	0	0.007474519
naics_42	0.001202252	0.036630037	0.019252548
inspect_history_violated_last_year	0.001191157	0.289377289	0.007134768
naics_NOT_AVAILABLE	0.001180949	0.007326007	0.011778029
waste_U226	0.001160543	0	0.000679502
waste_U034	0.001149137	0	0.000679502
waste_U080	0.001128075	0.003663004	0.003510759
waste_F008	0.001040655	0.007326007	0.001019253

waste_F001	0.001014767	0.036630037	0.009626274
waste_U002	0.001008036	0.043956044	0.011211778
waste_F009	0.000997912	0.018315018	0.000679502
waste_B003	0.000932508	0	0.002378256
waste_P022	0.000882345	0.003663004	0.000792752
waste_U089	0.000853814	0	0.000453001
waste_U404	0.000814324	0	0.000792752
waste_U220	0.000761119	0.010989011	0.002151755
waste_P105	0.000756321	0.007326007	0.00407701
waste_U077	0.000726282	0	0.000453001
waste_U134	0.000696054	0.003663004	0.000679502
waste_U117	0.000687802	0	0.000679502
waste_U031	0.000686825	0	0.001245753
waste_U129	0.000656707	0.051282051	0.002831257
waste_U206	0.000637547	0	0.000906002
waste_U162	0.000590005	0	0.001019253
waste_U112	0.00057518	0	0.001698754
waste_U133	0.00055954	0	0.000566251
waste_P003	0.000516179	0.003663004	0.000792752
waste_U159	0.000511338	0.007326007	0.001812005
waste_P028	0.00049284	0.010989011	0
waste_U196	0.000447254	0.010989011	0.000792752
waste_U150	0.000444754	0	0.002491506
waste_D021	0.000433994	0.007326007	0.002491506
waste_P008	0.000429318	0	0.000566251
waste_D043	0.000419854	0	0.002265006
naics_51	0.000413062	0.018315018	0.004756512
waste_U072	0.000400847	0	0.000566251
waste_U213	0.000396468	0	0.001585504
waste_U201	0.000391135	0	0.000679502
waste_U007	0.000375595	0	0.001132503
waste_U202	0.000356426	0.003663004	0.00011325
waste_U075	0.000353495	0	0.000226501
waste_U210	0.00034659	0.007326007	0.001132503
waste_U001	0.00033633	0.010989011	0.000792752
waste_U160	0.000335335	0	0.000679502
waste_U161	0.000324036	0	0.000566251
waste_U151	0.000319065	0.003663004	0.001585504
waste_F019	0.00031824	0.003663004	0.000792752
waste_D023	0.000317866	0.007326007	0.000566251
waste_U135	0.00029114	0.007326007	0.000566251
waste_U165	0.000286333	0	0.001472254
waste_U205	0.000275139	0	0.001019253
waste_D029	0.000273778	0	0.000906002
waste_D013	0.000259405	0	0.001132503
waste_P108	0.000234778	0	0.000226501
waste_D033	0.000230446	0	0.001132503
waste_P087	0.00021251	0.007326007	0.001925255
waste_U228	0.000210839	0	0.000906002
waste_P005	0.000175989	0	0.000226501
waste_D012	0.00016783	0	0.000453001
waste_D027	0.000164964	0.003663004	0.001812005
waste_P046	0.000140944	0	0.000906002
waste_U103	0.000128985	0	0.000226501

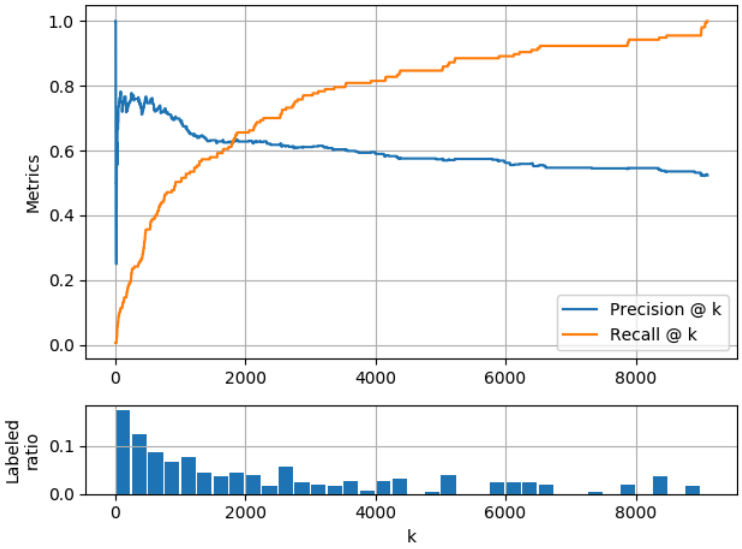
waste_U218	0.000127224	0	0.00011325
waste_U015	0.000117129	0	0.000679502
waste_P014	0.000112641	0.003663004	0.000566251
waste_U188	0.000112152	0.032967033	0.004869762
waste_P204	0.000100709	0	0.000679502
waste_P076	9.92E-05	0	0.000226501
waste_U223	9.78E-05	0	0.000226501
naics_72	8.84E-05	0	0.001925255
waste_F004	8.58E-05	0	0.001019253
waste_U070	8.17E-05	0	0.000453001
waste_U211	7.00E-05	0	0.000792752
waste_P048	3.82E-05	0.003663004	0.000339751
waste_B006	1.19E-05	0	0.001698754
naics_53	5.07E-06	0.025641026	0.022197055
waste_U069	4.26E-06	0	0.000339751
waste_D016	1.60E-06	0.003663004	0.011438279
naics_52	0	0	0.001019253
naics_55	0	0	0.000679502
naics_71	0	0	0.004643262
naics_21	0	0	0.001812005
waste_F027	0	0	0.000453001
waste_B001	0	0	0.000906002
waste_U012	0	0	0.000226501
waste_U248	0	0	0.00011325
waste_U246	0	0	0.000339751
waste_U240	0	0	0.000226501
waste_U239	0	0.036630037	0.001812005
waste_U236	0	0	0.000453001
waste_U219	0	0	0.00011325
waste_U144	0	0	0.000792752
waste_U132	0	0	0.000679502
waste_U115	0	0	0.00011325
waste_U108	0	0	0.001132503
waste_U056	0	0	0.000679502
waste_U037	0	0	0.001019253
waste_U019	0	0.003663004	0.001019253
waste_U003	0	0	0.001359003
waste_B005	0	0	0.001698754
waste_P188	0	0.245421245	0.032276331
waste_P120	0	0.003663004	0.000566251
waste_P119	0	0	0.00011325
waste_P106	0	0.003663004	0.001245753
waste_P077	0	0.003663004	0.000566251
waste_P029	0	0.010989011	0.000453001
waste_P010	0	0	0.000566251
waste_U279	0	0	0.002151755
waste_D036	0	0.003663004	0.000566251
waste_D034	0	0	0.000792752
waste_D032	0	0	0.000226501
waste_D028	0	0	0.006115515
waste_D025	0	0.003663004	0.001245753
naics_11	0	0.003663004	0.00407701

FNR Disparity: 0.923

4) Model 4 (top-4 mean precision model)

- Features: NAICS, inspection history, ACS data, waste code
- Model: random forest (max_depth: 10, min_samples_split: 10, n_estimator: 200)

- Precision/Recall@k graph with the ratio of labeled examples



List of feature importance of all features & Cross-tabs for 10 most different features

feature	importance	top_mean	bottom_mean
inspect_history_num_violations_prev_k_year	0.070564461	1.590114659	0.078220618
inspect_history_num_inspections_prev_k_year	0.069971808	1.606014069	0.134436306
inspect_history_violated_prev_k_year	0.040443239	0.978021978	0.056851642
ACS_data_structure_year	0.039240204	1950.74359	1948.193431
ACS_data_number_vehicles	0.03916816	280179.7802	196718.3567
ACS_data_median_income	0.038005803	70511.75092	57744.88245
ACS_data_heating_coke	0.036576082	252.9194139	445.0739524
ACS_data_heating_total	0.03438487	346010.9634	504860.0801
ACS_data_total_population	0.032964318	992204.3553	1342113.858
ACS_data_housing_unit	0.032811051	377659.4029	558379.8831
ACS_data_median_age	0.032479458	39.7021978	37.69357871
ACS_data_poverty_status	0.03058357	112151.293	240421.6636
ACS_data_heating_oil	0.026213062	36171.09524	49337.6667
naics_81	0.021969099	0.446886447	0.046319366
naics_44	0.017361515	0.443223443	0.061381653
waste_D001	0.016724827	2.084249084	0.370894677
waste_D008	0.016478472	0.142857143	0.273952435
inspect_history_inspected_prev_k_year	0.01631509	1.10989011	0.155379388
waste_D002	0.013542386	0.952380952	0.19218573
waste_B007	0.012604026	0.018315018	0.029898075
waste_D004	0.011720013	0.007326007	0.023329558
inspect_history_num_inspections_last_year	0.011483362	0.398459411	0.031771682
waste_D007	0.011200938	0.296703297	0.089354473
waste_F003	0.010952199	0.465201465	0.077236693
waste_F002	0.010402285	0.128205128	0.02819932

naics_23	0.009478251	0.014652015	0.245866365
naics_22	0.009098219	0.014652015	0.408493771
waste_D009	0.008521882	0.545787546	0.062061155
waste_D003	0.007975568	0.124542125	0.039184598
waste_F005	0.007616511	0.289377289	0.046092865
waste_D035	0.007180059	0.157509158	0.044847112
inspect_history_num_violations_last_year	0.007179369	0.39171721	0.019204959
waste_D005	0.007177667	0.102564103	0.043261608
naics_33	0.006768022	0.263736264	0.051189128
waste_D040	0.006757341	0.007326007	0.008720272
naics_45	0.006506514	0.014652015	0.01619479
waste_D011	0.006347747	0.923076923	0.07610419
waste_D006	0.006321127	0.036630037	0.034428086
naics_32	0.00570634	0.183150183	0.038278596
naics_56	0.005504808	0	0.035334088
waste_F007	0.005483055	0.025641026	0.002604757
waste_F006	0.0053669	0.095238095	0.009626274
waste_D018	0.005188658	0.117216117	0.037259343
naics_48	0.005143186	0.007326007	0.067723669
waste_D022	0.004944358	0.032967033	0.015062288
inspect_history_inspected_last_year	0.004731537	0.197802198	0.02819932
naics_31	0.004665238	0.018315018	0.005096263
naics_92	0.004540564	0.007326007	0.043035108
naics_54	0.004420514	0.06959707	0.01200453
waste_RARE	0.004311512	0.106227106	0.018006795
waste_B002	0.004221558	0.003663004	0.013703284
waste_U058	0.004098437	0.062271062	0.010079275
waste_D039	0.004075039	0.029304029	0.010419026
waste_LABP	0.004021242	0.054945055	0.010192525
waste_D010	0.003791165	0.344322344	0.037485844
waste_U010	0.003707412	0	0.005662514
naics_62	0.003695895	0.036630037	0.015741789
waste_B004	0.00363844	0.010989011	0.011438279
waste_P001	0.003634953	0.853479853	0.061041903
ACS_data_NOT_AVAILABLE	0.003435981	0.003663004	0.002038505
waste_F001	0.003142848	0.025641026	0.009966025
waste_U122	0.002754801	0.091575092	0.004983012
waste_P030	0.002598264	0.014652015	0.004530011
waste_U080	0.002567968	0.003663004	0.003510759
naics_61	0.002566864	0.032967033	0.047678369
waste_D038	0.002505656	0.007326007	0.00804077
waste_P075	0.002483803	0.472527473	0.044394111
naics_42	0.002375263	0.025641026	0.019592299
waste_D019	0.002361528	0.003663004	0.00407701
waste_D026	0.002313039	0.164835165	0.008267271
waste_P081	0.002284193	0.457875458	0.028765572
waste_U059	0.002215886	0	0.001925255
waste_U154	0.002115972	0.003663004	0.005096263
waste_U002	0.002107043	0.007326007	0.012344281
waste_U035	0.002082444	0.065934066	0.009060023
inspect_history_violated_last_year	0.002062167	0.19047619	0.010192525
waste_U044	0.002021724	0.003663004	0.007361268
waste_D024	0.001957906	0.161172161	0.014496036
waste_F039	0.001952214	0	0.003624009

waste_P012	0.001801458	0	0.003737259
naics_NOT_AVAILABLE	0.001747361	0.007326007	0.011778029
waste_U150	0.001707829	0.007326007	0.002265006
waste_U112	0.001566893	0	0.001698754
waste_P042	0.001512995	0	0.007474519
waste_P008	0.001434499	0	0.000566251
waste_U226	0.001423324	0	0.000679502
waste_D016	0.001422198	0.003663004	0.011438279
naics_51	0.001372318	0.010989011	0.004983012
waste_B003	0.00133002	0	0.002378256
waste_U188	0.001317762	0.036630037	0.004756512
naics_53	0.001317061	0.021978022	0.022310306
naics_49	0.001300399	0	0.005662514
waste_B006	0.0011767	0	0.001698754
waste_D021	0.001145727	0.007326007	0.002491506
waste_U089	0.00110757	0	0.000453001
waste_U123	0.001075385	0.007326007	0.001585504
waste_P105	0.00107179	0.003663004	0.00419026
waste_U129	0.001062935	0.065934066	0.002378256
waste_U239	0.001053147	0.007326007	0.002718007
waste_U220	0.00104509	0.007326007	0.002265006
waste_D028	0.000982727	0	0.006115515
waste_U134	0.000948228	0	0.000792752
waste_U031	0.000945302	0	0.001245753
waste_U075	0.000932103	0	0.000226501
waste_P087	0.000926509	0.003663004	0.002038505
waste_P014	0.000906486	0.003663004	0.000566251
waste_D043	0.000904187	0	0.002265006
waste_U135	0.000901836	0.003663004	0.000679502
waste_U151	0.000898675	0.003663004	0.001585504
waste_U117	0.000893922	0	0.000679502
waste_P098	0.000867804	0.010989011	0.00419026
waste_F008	0.000830397	0.007326007	0.001019253
waste_U007	0.000806894	0	0.001132503
waste_B005	0.000805899	0	0.001698754
waste_D029	0.000799798	0	0.000906002
waste_D033	0.000772239	0	0.001132503
waste_U201	0.000769713	0	0.000679502
waste_U404	0.000750314	0	0.000792752
waste_U206	0.000741835	0	0.000906002
waste_U034	0.000703184	0	0.000679502
waste_F009	0.000689336	0.014652015	0.000792752
waste_U165	0.000682994	0	0.001472254
waste_D034	0.000678239	0	0.000792752
waste_U161	0.000673756	0	0.000566251
waste_U228	0.000671921	0	0.000906002
waste_U052	0.000648538	0.007326007	0.000453001
waste_D027	0.000641889	0.003663004	0.001812005
waste_P188	0.000639463	0.465201465	0.025481314
waste_U202	0.000564201	0.003663004	0.00011325
waste_P046	0.000545629	0	0.000906002
waste_U159	0.00053812	0	0.002038505
waste_U003	0.000520634	0	0.001359003
waste_U162	0.0004932	0	0.001019253

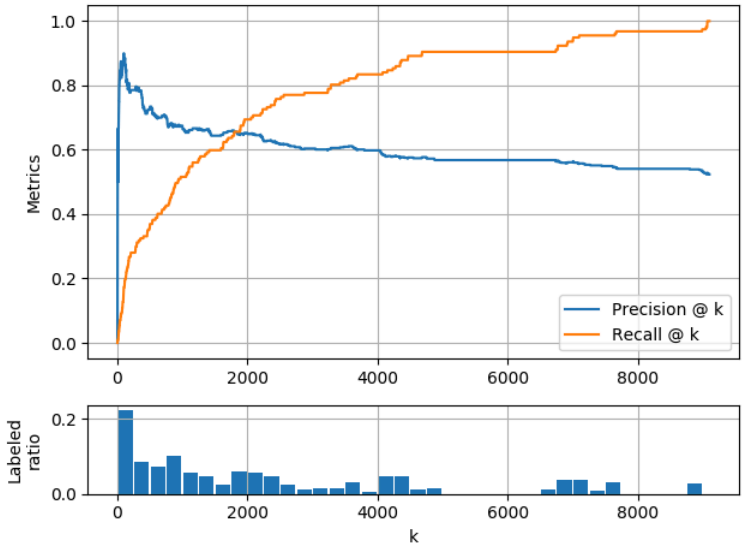
waste_P022	0.000487377	0.003663004	0.000792752
naics_21	0.00047628	0	0.001812005
waste_U133	0.000455441	0	0.000566251
waste_D012	0.000450965	0	0.000453001
waste_P028	0.000446998	0.010989011	0
waste_U213	0.000403232	0	0.001585504
waste_U218	0.000393881	0	0.00011325
waste_P204	0.000390099	0	0.000679502
waste_D032	0.000379596	0	0.000226501
waste_D036	0.000374937	0	0.000679502
waste_P005	0.000372614	0	0.000226501
waste_F004	0.000352844	0	0.001019253
waste_P106	0.000350951	0.003663004	0.001245753
waste_D013	0.000331265	0	0.001132503
waste_F019	0.000322965	0.003663004	0.000792752
waste_P108	0.000308455	0	0.000226501
waste_U279	0.00030623	0	0.002151755
waste_U196	0.000302121	0.003663004	0.001019253
waste_U248	0.00029968	0	0.00011325
waste_U144	0.000293942	0	0.000792752
waste_B001	0.000288583	0	0.000906002
waste_U001	0.00028507	0.014652015	0.000679502
waste_U019	0.000279405	0.003663004	0.001019253
waste_U077	0.000278971	0	0.000453001
waste_U070	0.000276799	0	0.000453001
waste_P077	0.000262305	0.003663004	0.000566251
waste_U072	0.000255	0	0.000566251
waste_U205	0.000235919	0	0.001019253
waste_U103	0.000227323	0	0.000226501
naics_72	0.000204924	0	0.001925255
waste_D023	0.000202288	0.007326007	0.000566251
waste_U108	0.000201883	0	0.001132503
waste_U115	0.000198611	0	0.00011325
waste_P048	0.000196325	0.003663004	0.000339751
waste_U211	0.000191029	0	0.000792752
waste_P003	0.000165328	0.003663004	0.000792752
waste_U223	0.000163891	0	0.000226501
waste_F027	0.000157725	0	0.000453001
waste_D025	0.000144905	0.003663004	0.001245753
waste_U160	0.000130143	0	0.000679502
waste_U012	0.000118198	0	0.000226501
waste_U132	0.000112935	0.007326007	0.000453001
naics_71	0.000106856	0	0.004643262
waste_U210	0.00010455	0.003663004	0.001245753
waste_U037	0.000100417	0	0.001019253
waste_P029	9.75E-05	0.007326007	0.000566251
waste_U015	8.88E-05	0	0.000679502
waste_U240	8.51E-05	0	0.000226501
waste_P119	7.80E-05	0	0.00011325
waste_U069	4.94E-05	0	0.000339751
waste_P120	4.26E-05	0.003663004	0.000566251
waste_U236	3.85E-05	0	0.000453001
waste_P010	3.80E-05	0	0.000566251
waste_U219	3.17E-05	0	0.00011325

naics_52	3.06E-05	0	0.001019253
naics_55	2.90E-05	0	0.000679502
waste_U056	2.34E-05	0	0.000679502
naics_11	2.11E-05	0.003663004	0.00407701
waste_U246	3.68E-06	0	0.000339751
waste_P076	0	0	0.000226501

FNR DIsparity: 0.945

5) Model 5 (minimax precision model)

- Features: NAICS, inspection history, ACS data, waste code
- Model: random forest (max_depth: 50, min_samples_split: 50, n_estimator: 10000)
- Precision/Recall@k graph with the ratio of labeled examples



List of feature importance of all features & Cross-tabs for 10 most different features

feature	importance	top_mean	bottom_mean
inspect_history_num_violations_prev_k_year	0.098694098	1.69056727	0.075114892
inspect_history_num_inspections_prev_k_year	0.0826482	1.705784342	0.131351676
ACS_data_number_vehicles	0.060408482	309354.011	195816.3675
ACS_data_median_income	0.055691087	75936.27473	57577.17067
ACS_data_structure_year	0.051804638	1944.549451	1948.384938
ACS_data_median_age	0.049577805	39.95750916	37.68568516
ACS_data_heating_coke	0.049247579	207.9194139	446.4652322
inspect_history_violated_prev_k_year	0.048652148	1.003663004	0.05605889
ACS_data_total_population	0.047372528	937659.8315	1343800.229
ACS_data_poverty_status	0.04557449	81634.98168	241365.1464
ACS_data_housing_unit	0.045555809	347611.3626	559308.8882
ACS_data_heating_total	0.044685203	319486.4872	505680.146
ACS_data_heating_oil	0.043324135	33973.37363	49405.61438
naics_81	0.03441821	0.388278388	0.04813137
inspect_history_inspected_prev_k_year	0.032511566	1.124542125	0.154926387
naics_44	0.026472809	0.366300366	0.063759909
naics_23	0.020233309	0.018315018	0.245753114
naics_22	0.016624781	0.021978022	0.408267271

inspect_history_num_inspections_last_year	0.014017277	0.221173083	0.037252902
naics_33	0.011493323	0.274725275	0.050849377
naics_62	0.010833171	0.062271062	0.014949037
naics_45	0.010705713	0.062271062	0.014722537
naics_32	0.010157735	0.179487179	0.038391846
naics_54	0.00984804	0.087912088	0.011438279
inspect_history_inspected_last_year	0.009478253	0.128205128	0.030351076
naics_61	0.009090284	0.047619048	0.047225368
naics_31	0.007785571	0.018315018	0.005096263
naics_48	0.007741054	0.014652015	0.067497169
naics_92	0.007413046	0.010989011	0.042921857
inspect_history_num_violations_last_year	0.007205702	0.214967892	0.024669576
naics_56	0.006259361	0.003663004	0.035220838
naics_42	0.005341121	0.025641026	0.019592299
naics_53	0.003223141	0.021978022	0.022310306
naics_NOT_AVAILABLE	0.003063647	0.010989011	0.011664779
naics_49	0.002939816	0.003663004	0.005549264
ACS_data_NOT_AVAILABLE	0.002332174	0.007326007	0.001925255
inspect_history_violated_last_year	0.002325698	0.120879121	0.012344281
naics_51	0.002042097	0.021978022	0.004643262
naics_21	0.000971623	0	0.001812005
naics_11	0.000746468	0.003663004	0.00407701
naics_52	0.000728317	0	0.001019253
naics_72	0.000519046	0	0.001925255
naics_71	0.000196826	0.003663004	0.004530011
naics_55	4.46E-05	0	0.000679502

FNR Disparity: 0.945

FNR and FNR DIsparity for Top Models (Pre-Dataset Revision)

Model Type	Max Depth	Min Sample Splits	N Estimators	Mean Precision	FNR disparity
random_forest	2	10	200	0.92	0.928
random_forest	None	50	10000	0.895	0.928
random_forest	2	50	20	0.883	0.923
random_forest	None	50	1000	0.876	0.945
Random_forest (top mini_max)	50	50	1000	0.86	0.945

A List of Most Common Violations

	violation_type_desc	frequency
1	Violations Of Reporting Requirements	3,902
2	Asbestos Requirement Violation CAA	2,423
3	Violation Of A Permit Requirement	1,872
4	Violation Of PCB Rules	1,459
5	FIFRA	1,401
6	Discharge, Emission Or Activity WO Re	1,367
7	PWS Monitoring/Reporting	1,250
8	Effluent Limit Violations,Not Otherwise	1,249
9	National Emission Standard For Hazard	1,113
10	General Facility Requirements	956
11	UST Requirements, Other Than LDAR	878
12	Violation Of A SIP, Not Otherwise Spec	857
13	Other/Miscellaneous	725
14	PWS Notification To Public	563
15	Failure To Notify	516

Select Recall Disparities

	model	min_prec	mean_pre	FNR_dispa	recall_dis	features	model_type
211	['naics', 'inspect_history', 'acs_da	0.695652	0.820999	0.997542	0.963211	['naics', 'ir	random_forest
190	['naics', 'inspect_history', 'acs_da	0.611111	0.798443	0.997411	0.963211	['naics', 'ir	random_forest
201	['naics', 'inspect_history', 'acs_da	0.761194	0.854302	0.997542	0.963211	['naics', 'ir	random_forest
194	['naics', 'inspect_history', 'acs_da	0.6875	0.807424	0.997672	0.963211	['naics', 'ir	random_forest
198	['naics', 'inspect_history', 'acs_da	0.666667	0.788411	0.997672	0.963211	['naics', 'ir	random_forest
214	['naics', 'inspect_history', 'acs_da	0.659574	0.826875	0.998771	0.977376	['naics', 'ir	random_forest
132	['naics', 'inspect_history', 'acs_da	0.52381	0.61974	0.999425	0.979167	['naics', 'ir	random_forest
108	['naics', 'inspect_history', 'acs_da	0.52381	0.61974	0.999425	0.979167	['naics', 'ir	random_forest
142	['naics', 'inspect_history', 'acs_da	0.571429	0.792921	0.944196	0.979167	['naics', 'ir	knn
143	['naics', 'inspect_history', 'acs_da	0.571429	0.792921	0.999853	0.979167	['naics', 'ir	knn
195	['naics', 'inspect_history', 'acs_da	0.666667	0.815194	0.997933	0.984615	['naics', 'ir	random_forest
197	['naics', 'inspect_history', 'acs_da	0.666667	0.825275	0.997933	0.984615	['naics', 'ir	random_forest
210	['naics', 'inspect_history', 'acs_da	0.730769	0.805037	0.997672	0.984615	['naics', 'ir	random_forest
199	['naics', 'inspect_history', 'acs_da	0.666667	0.814406	0.998064	1.006993	['naics', 'ir	random_forest
231	['naics', 'inspect_history', 'acs_da	0.666667	0.815385	0.999033	1.006993	['naics', 'ir	random_forest
64	['naics', 'inspect_history', 'acs_da	0.6	0.760322	0.999228	1.024709	['naics', 'ir	random_forest
217	['naics', 'inspect_history', 'acs_da	0.746032	0.840317	0.998325	1.030411	['naics', 'ir	random_forest
145	['naics', 'inspect_history', 'acs_da	0.6	0.782878	1.000116	1.036765	['naics', 'ir	knn
144	['naics', 'inspect_history', 'acs_da	0.6	0.782878	1.121591	1.036765	['naics', 'ir	knn

Composition of Full Dataset: Facility Types

Facility Type	Count
Large Quantity Generators (LQGs)	9,172
Small Quantity Generators (SQGs)	6,000
Conditionally Exempt Small Quantity Generators (CESQGs)	14,041
Transporters	312
Treatment, storage, and disposal facilities (TSDFs)	25