EPA Waste Inspection Targeting

By Tim Hannifan, Alec MacMillen and Quinn Underriner

June 12th, 2019

1 - Background and Goals

The US Environmental Protection Agency has the resources to inspect only four percent of the nation's hazardous material facilities. There are catastrophic public externalities from the improper disposal and storage of hazardous waste including cancer, genetic mutations, physiological malfunctions, physical deformations, and birth defects.[1] The public is at risk of exposure primarily through the contamination of water sources, which is a result of groundwater seepage and surface runoff.[2]

Compliance with current regulations is enforced through three primary methods: mandatory two-year inspections (which don't always occur), self-auditing (which is highly unlikely for a serious violator), and targeted inspections by state and federal regulatory agencies. On average, three out of ten state and federal inspections result in a violation. While this may indicate that most facilities are compliant, we take a more pessimistic view: the EPA's limited resources are being wasted inspecting facilities that are likely non-violators.

The goal of our analysis to create a targeted list of hazardous waste facilities that are likely to be non-compliant in the next year. The intention of this analysis is to

[1] EPA Health and Ecological Hazards Caused by Hazardous Substances https://www.epa.gov/emergency-response/health-and-ecological-hazards-caused-hazardous-substances

[2] C.C. Barton, S.C. Schmitz, in Information Resources in Toxicology (Fourth Edition), 2009

increase the effectiveness of the EPA and, in turn, put greater pressure on facilities that place the public's health at risk.

2 - Related work

Previous work has been conducted on this topic by Data Science for Social Good at the University of Chicago. While specific details about the precision achieved in their models is unavailable, their proof of concept confirmed the predictive power of the methods employed in this project. Their analysis differs in terms of complexity and artful feature engineering; the authors included proxies for violation severity as well as underreported waste generation by including biennial reporting data on the amounts self-reported by facilities.

3 - Problem formulation and overview of solution

The EPA is operating under budget and time constraints that make it difficult to evaluate the activities of all hazardous waste transport, storage, and disposal facilities. We seek to build a predictive model that will help the agency target inspections at facilities that are likely to be found in violation of RCRA regulations, so that there will be fewer resources wasted on unnecessary evaluations and fewer actual violations that go unchecked.

The RCRA_EVALUATIONS dataset contains a *violation_found* field that indicates whether or not an evaluation resulted in a violation being found ("Y" for yes, "N" for no, and "U" for undetermined). We convert this column into a binary dummy variable that takes a value of 1 for “Y” and 0 for “N.” To be conservative, we assume that evaluations with a value of “U” for *violation_found* are actually violations.

We now plan to solve a classification problem using supervised learning methods. Our target variable *violation* (the binary dummy indicator) contains the class label, and the variables about the facility’s location, industry, purpose, prior inspection history, etc. will become features in predictive models that will attempt to guess whether a given evaluation event is likely to result in an enforcement being found.

It’s important to note that we only have data on facilities that are actually inspected, so what we are really predicting is the likelihood a given facility will be in violation *conditional on being inspected*. To capture this nuance in our problem formulation, we will construct an artificial “planning date” of January 1st of each year, on which date we assume the EPA “plans” what facilities it will evaluate over the course of the entire upcoming year. Under this formulation, the prediction horizon is 1 year, which means we will keep a 1-year gap between our training and testing dates for each split. Our goal is to produce a list of prioritized facilities for targeted inspection. As we assume the EPA’s inspection capacity is capped at 10%, we will prioritize precision at 10% as our metric of choice, although we will also test precision at 5% and 15% to account for a potential range of inspection capacities.

Therefore, our final problem formulation statement is as follows: *From the list of “planned” inspections (i.e. the list of facilities the EPA will evaluate in the coming year), which facilities are most likely to be found in violation of the RCRA?*

4 - Data Description

Data was sourced from the EPA's Enforcement and Compliance History Online (ECHO) system. This database includes records from 1978 to present documenting hazardous waste inspections, violations, enforcement actions, and facility information.

Attributes for each facility include location, size, type of activity, location, and designation as a previous significant violator. Inspection records included type codes identifying the nature of the violation, follow-up actions by the regulatory agency for re-inspection and certification, and financial penalties imposed on the facility.

The clean dataset used for this project identified 1,041,254 unique hazardous waste facilities and 991,196 inspections over the course of 43 years. We limited our analysis to the years 2000 to 2019 to eliminate some time-based bias in our estimates.

The data is publicly available on the EPA's website at this link, as a part of the EPA's Resource Conservation and Recovery Act Information System (RCRAInfo), a subset of ECHO.

5 - Details of solution: methods, tools, analysis, models, features

Our approach consisted of the following steps:

(1) Data cleaning and merging - merge_and_collapse.py

The first challenge was to combine the data from all six RCRA datasets into a single table, so that we had all available information associated with each individual evaluation. Using RCRA_EVALUATION as the base table, the goal was to add columns to each row so that we had more detail about the context in which an evaluation was performed to improve our prediction results. All six datasets contain an *ID_NUMBER* field that uniquely identifies facilities and makes it possible to join tables. RCRA_FACILITIES and RCRA_NAICS have time-invariant facility-specific information, so their information is easily added to RCRA_EVALUATIONS using simple left joins on *ID_NUMBER*.

For the other datasets (RCRA_VIOLATIONS, RCRA_VIOSNC_HISTORY, and RCRA_ENFORCEMENTS), we aggregated all information at the facility level for each year in the data (1984-2018) using *only* the information that occurred before a facility's evaluation date. This is important to note, because doing so allowed us to avoid using future information to predict current outcomes. For example, if an evaluation occurred at a given facility on 6/1/2000, *merge_and_collapse.py* would summarize violation, viosnc flag, and enforcement information for that facility **occurring before 1/1/2000**. That way, the data only includes information that the EPA would have access to during their theoretical "planning date" on January 1st of the year of the evaluation.

(2) Feature and split generation - generate_features.py, utils.py

The next step was to transform the built data into training and testing datasets for all 16 splits. We cleaned, imputed, dummified, and scaled data as necessary *after* splitting the data in training and test sets so that information in the testing data did not affect imputation of values in the training data. Features generated including facility location, industry, type of hazardous waste handler (e.g. generator, transporter, etc.), and information on prior violations, enforcements, and unresolved violation or significant non-complier flags. With more time, we would've liked to link the data to Census Bureau data that could help us understand how the demographics of a facility's location could affect whether or not it tended to violate RCRA standards.

(3) Model testing - run_models.py

The result of step (2) was a dictionary with split numbers as keys and a dictionary of train/test metadata and data as values. The next step was to feed this dictionary of splits to a 'magic loop' function that iterated through a host of different model types with a variety of parameters. The function trained, tested, and evaluated models' performance and output summary files of model performance by split.

(4) Model selection - pick_best_model.py

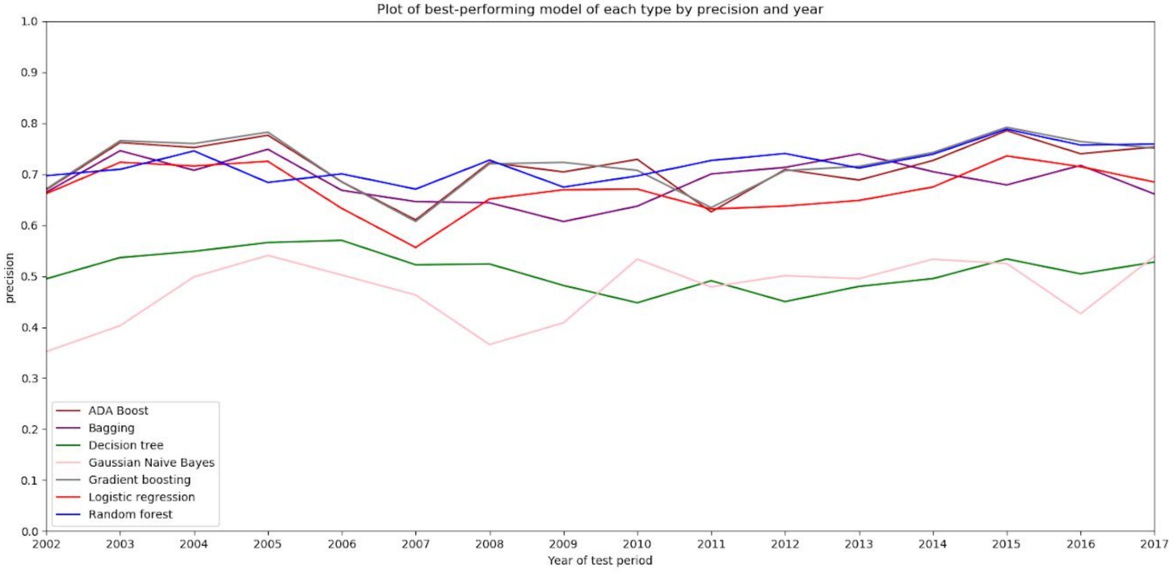
With our model performances output, the next step was to select the model that performed best with respect to precision at 10% for our most recent train-test split (split number 15, training dates 2000-2015 and testing dates 2017). We sorted all models by performance and plotted the best performing model of each type for all testing years (2002-2017). From this analysis, we determined that the best model on our most recent split was a random forest classifier using the Gini method for calculating splits, 10 estimators, and no maximum depth.

(5) Produce final list of targeted facilities - output_list.py

Finally, we used our best-performing model to produce an example list of targeted facilities for inspection that we could pass to our stakeholders (the EPA, state inspection agencies, etc.) We also produced a list of features ranked by their relative importance and plotted the precision-recall curve for our best-performing model.

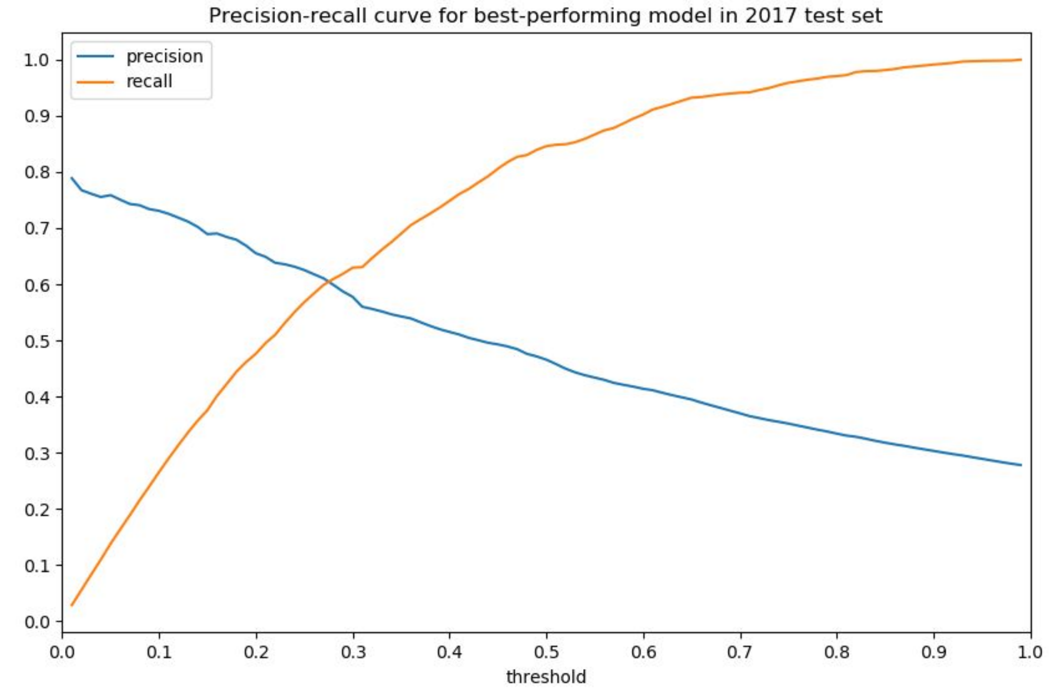
6 - Evaluation: Results and Plots

As explained in section 5, we ranked all models by their performance on precision at 10% across the range of our testing years (2002-2017). That plot appears below (K-nearest neighbors and SVMs were excluded for time and computational restraints):



Gaussian Naive Bayes and decision trees performed the worst among all tested models. Among the leaders in precision at 10%, random forests, ADA boost, and gradient boosting performed best, and that performance was relatively consistent across the entire span of splits. In the final (most recent) split, a random forest model using Gini impurity, 10 estimators, and no maximum depth performed the best: it boasts precision of approximately 0.731. This means that, of all facilities flagged as being in the top 10% most likely violators, this model correctly identified true positives about 73% of the time and only resulted in false positives 27% of the time.

The preferred model's performance over a full range of thresholds can be seen below:



Recall for our best-performing model at 10% is roughly 0.265, indicating that the list of facilities output by this model would only identify 26.5% of all violating facilities in 2017. However, our relatively high precision means the EPA could be reasonably confident that it wasn't wasting inspection resources if it prioritized inspecting these facilities first.

7 - Discussion and interpretation of results

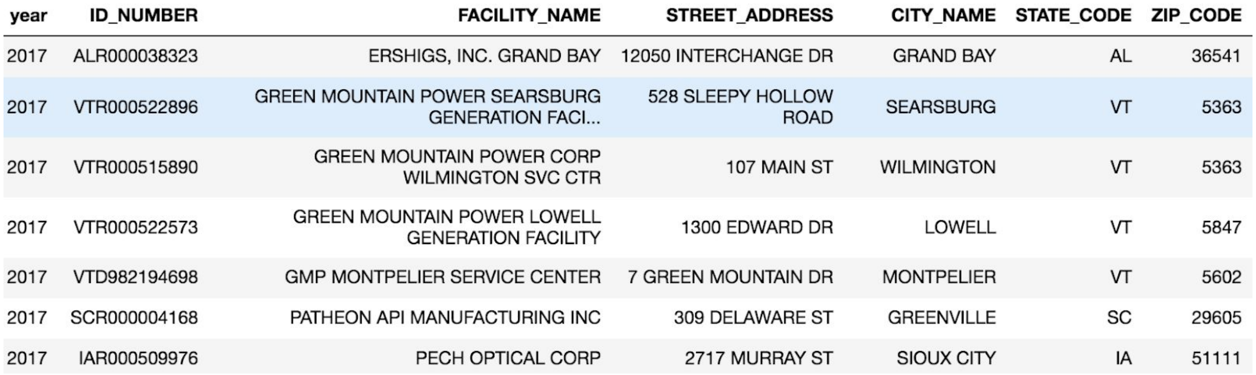
Our best-performing model on the most recent split for precision at 10% resulted in precision of 0.731, which far outstrips the baseline for split 15 of 0.275. The feature importances list that corresponds with this model's performance reveals the perhaps unsurprising result that 4 of the 6 most important features in predicting violations were binary indicators for previous violations, previous enforcement actions, previous unresolved violation flags, and previous significant non-complier flags applied to a facility, which suggests that the EPA could probably identify many violations by prioritizing inspections at facilities with any kind of previous infraction. In other words, it appears that facilities that violate the RCRA don't manage to "clean up their act" and in fact re-offend.

By far the single most important feature was the average time it takes a facility to resolve a violation (*avg_resolution_time*). While we don't have information about the direction of the correlation between *avg_resolution_time* and our target variable *violation*, it seems reasonable to hypothesize that facilities that take longer to resolve their violations may be predisposed to future offenses. Other notable important features include a binary indicator for whether the inspection is of type "compliance evaluation inspection" (*evtype_CEI*), as well as previous proposed and final monetary penalties for past violations (*pmp_amt* and *fmp_amt* - which could suggest facilities with costlier previous infractions are more likely to violate in the future).

There were three state indicator dummies that had feature weights higher than 0.01: Indiana, New Jersey, and Ohio (*state_IN, state_NJ*, and *state_OH*), suggesting that whether a facility is in one of these states has a non-trivial correlation with whether it violates RCRA standards. The most heavily-weighted industry dummy was *naics_33*, an indicator for whether a facility fell into the "manufacturing" category.

8. Policy Recommendations based on your analysis/models

Below you can see a snippet of our project deliverable in the form of a ranked list of facilities to inspect. Ultimately, these outputs could be tweaked in length, but the underlying goal is to allow the EPA to inspect fewer facilities overall and find more violations. Given the immediate uncertainty of the Trump administration's proposed 31 percent budget cut for the EPA,[3] and the long term uncertain that comes from a government openly hostile to the mission of the EPA, we are helping them to prepare for an extremely constrained budget scenario.



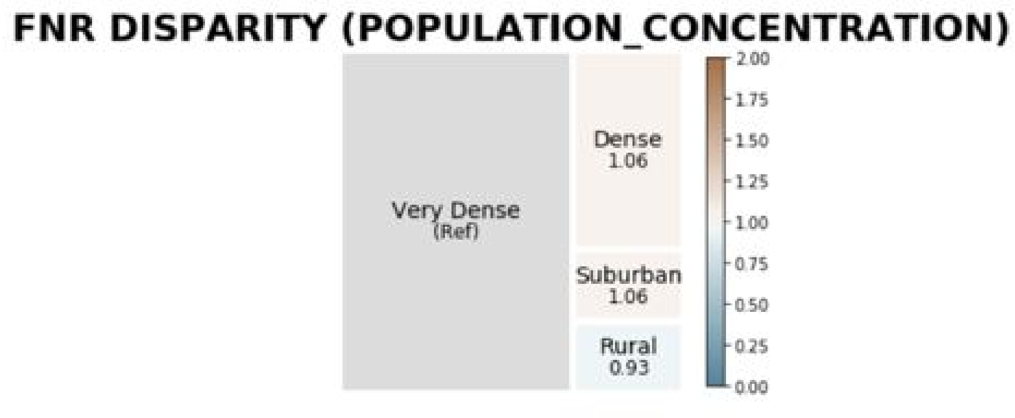
year	ID_NUMBER	FACILITY_NAME	STREET_ADDRESS	CITY_NAME	STATE_CODE	ZIP_CODE
2017	ALR000038323	ERSHIGS, INC. GRAND BAY	12050 INTERCHANGE DR	GRAND BAY	AL	36541
2017	VTR000522896	GREEN MOUNTAIN POWER SEARSBURG GENERATION FACI...	528 SLEEPY HOLLOW ROAD	SEARSBURG	VT	5363
2017	VTR000515890	GREEN MOUNTAIN POWER CORP WILMINGTON SVC CTR	107 MAIN ST	WILMINGTON	VT	5363
2017	VTR000522573	GREEN MOUNTAIN POWER LOWELL GENERATION FACILITY	1300 EDWARD DR	LOWELL	VT	5847
2017	VTD982194698	GMP MONTPELIER SERVICE CENTER	7 GREEN MOUNTAIN DR	MONTPELIER	VT	5602
2017	SCR000004168	PATHEON API MANUFACTURING INC	309 DELAWARE ST	GREENVILLE	SC	29605
2017	IAR000509976	PECH OPTICAL CORP	2717 MURRAY ST	SIOUX CITY	IA	51111

9. Ethical issues and Bias and Fairness analysis results and discussion

[3] Miranda Green, "Trump proposes slashing EPA budget by 31 percent", *The Hill,* March 11, 2019, "https://thehill.com/policy/energy-environment/433496-white-house-proposes-dramatic-cuts-to-energy-and-environment

In doing our fairness assessment with the Aequitas tool, we chose to focus on population density around a facility. We broke the population density into roughy four groups (Rural, Suburban, Dense, Very Dense), by zip code. EPA offices tend to be in larger cities (which makes sense for talent acquisition and clustering of resources), but we wanted to ensure that there wasn't an underlying bias in our dataset for inspectors to be checking places closer to urban centers more frequently or thoughougly, or that our analysis could exacerbate this problem by underpredicting violations in rural areas.

Checking a site is assistive to a community, and we are intervening with a relatively large percentage of the population, so we chose to focus on False Negative Rate Parity. A false negative in our case would be incorrectly indicating that a site was not in fact be found to have a violating in a given year, when in fact it did. Our results indicate that our false negative rate disparity is thankfully quite low. Using the quartile of the highest population density as our baseline of 1.0 (with the assumption that these areas will be closest to where EPA officials currently live), we find that Rural areas actually have a lower False Negative rate at 0.93, and the two middle density areas both have roughly 1.05 the False Negative rate of the densest area. These results satisfy our concerns about parity in inspections related to population density.



10. Limitations, caveats, future work to improve on what you've done.

In the interpretation of our research it’s important to note several caveats. First, a fundamental problem in our underlying dataset is that we don’t know the actual number of facilities that have hazardous waste violations, we only know the violations that the EPA inspectors found. In our ultimate goal of reducing the public's exposure to hazardous waste we want to know the former. Relatedly, another important thing to note about our data is that we only have data on facilities that the EPA has inspected. That means, without added ability to include new facilities into our pipeline and retrain our models, a newly opened facility would never be suggested for inspection. If we don’t have a way to identify facilities on more than their name and confidently update our records that a facility has changed its name (perhaps in response to bad press related to its handling of waste) companies could dodge inspection with name changes alone.

In our dataset the violation data is at the day level (e.g., each row, which is a violation, gives the date the violation occurred). Due to the computational complexity of computing our features at this level, for this project we created the features at the year level, which causes us to lose some nuance and likely predictive signal within the more granular data. With more time and computing power, we would create features at the day level. The day level data is not without its complications as the day of a violation is not actually the day the violation was found, but rather is contingent on exogenous factors such as waiting for lab results to confirm a leak. Both in our year level data, and certainly when in the future we will move to the more granular day level, this could have caused some features timing to be misclassified.

We coded the relatively few facility inspections in the dataset that lacked an outcome to be violating facilities. We assume the conservative approach is the correct one as the alternative is potentially missing a violating facility in our dataset, and allowing for more toxic waste to escape into the environment.

In our future research our top priority would be to reformulate our prediction. Currently, we are predicting hazardous waste violations conditional on being inspected by the EPA. From a public health perspective, the ultimate goal is to get the unconditional probability of a facility having a violation. To back out this unconditional probability we would need to calculate the probability of a given facility being inspected and incorporate Bayes’ theorem. Given more time and resources, we would also include more demographic data than just population density, and be able to examine the racial and economic makeup around communities where these facilities exist and ensure equity in their inspection.

We currently treat all types of violations the same, whereas we know that in our data a violation could be a massive spill of cancerous materials into a river, or improperly disposed paint at a hardware store. We would want to weigh these outcomes differently based on their severity in our final recommendation of which facilities to audit first. We also know that, for facilities that produce large amounts of hazardous waste, there is a significant disparity between how frequently federal facilities are inspected and how frequently state facilities are inspected[4]. With more time we would break out our analysis at the agency level provide more specific recommendations for the federal government and for individual states.

[4] Parker et al., “EPA Has Not Met Statutory Requirements for Hazardous Waste Treatment, Storage and Disposal Facility Inspections, but Inspection Rates Are High”, United States Environmental Protection Agency, March 11, 2016, https://www.epa.gov/sites/production/files/2016-03/documents/20160311-16-p-0104.pdf