Predictive Modeling for Environmental Protection: Hazardous Waste Management [*]

Eric Potash
University of Chicago
epotash@uchicago.edu

Jimmy Jin
University of North Carolina
Chapel Hill
jimmyjin@live.unc.edu

Maria Kamenetsky
University of Wisconsin
Madison
mkamenetsky@wisc.edu

Dean Magee
University of Melbourne
dmagee@
student.unimelb.edu.au

Paul van der Boor
McKinsey & Company
pvboor@gmail.com

Rayid Ghani
University of Chicago
rayid@uchicago.edu

ABSTRACT

The improper treatment and disposal of hazardous waste can have disastrous effects on the environment and human health. The Resource Conservation and Recovery Act (RCRA) governs hazardous waste management in the United States. To enforce its regulations, the New York State Department of Environmental Conservation (NYSDEC) inspects facilities that handle hazardous materials. However, due to resource constraints, not all facilities can be inspected each year. We worked with NYSDEC to build predictive models that use reporting, monitoring, and enforcement data to prioritize inspection resources.

The past selection of inspection sites results in non-random missing labels in both training and testing. We initially built models that ignore this selection bias. Next we modeled the selection process and incorporated it into two kinds of models: first, we reinforced that process; second, we used reweighting to try to correct its bias. Each of these three approaches produces qualitatively different predictions, though we estimate through cross-validation that each of them can significantly increase the proportion of future inspections finding violations. We propose a novel field trial design for NYSDEC to test, compare, and select from among these models.

1. INTRODUCTION

The Resource Conservation and Recovery Act (RCRA) was enacted in 1976, giving environmental agencies the authority to regulate hazardous waste "from cradle to grave." RCRA defines what constitutes hazardous waste and outlines explicit requirements for waste management. The act is intended is to minimize negative environmental impacts by reducing the total amount of waste produced and preventing spills through responsible management practices, rather than simply managing waste disposal at the end of the pipeline [3]. Over 130 million tons of hazardous waste were generated in the United States in 2013, and in the following year there were 28,000 spills, leading to 1182 deaths and nearly $85 million in property damage [2].

The New York State Department of Environmental Conservation (NYSDEC) uses a compliance inspection program to enforce RCRA regulations in New York State. It is critical to do so in order to prevent disasters such as the Kaltech Industries explosion in 2002. Kaltech Industries Group was a commercial sign manufacturer located in the basement of a building in downtown Manhattan which used hazardous chemicals for etching and cleaning. Due to improper training, labeling, and handling, incompatible chemicals were mixed and resulted in an explosion which injured thirty-six people, four of whom were admitted into intensive care [16]. There was extensive damage to the building including blow out of the elevator shaft, collapsing walls, and destruction of the center stairwell [6]. One of the causes was found to be that Kaltech had not been inspected in its last ten years of operation.

In this paper, we describe our work with NYSDEC to more effectively allocate inspection resources by creating predictive models that can identify facilities with a high likelihood of violating environmental regulations and, in turn, to better enforce RCRA, and reduce environmental damage.

The paper is structured as follows: In section 2 we describe how NYSDEC currently targets inspections and identify opportunities for a data-driven approach to the inspection targeting process. In section 3, we summarize the available data sources. In section 4, we describe our evaluation methodology in the presence of missing labels. In section 5, we describe the process we used to aggregate information and generate spatio-temporal features for all of our models. In section 6, we identify three distinct approaches to modeling in the presence of missing labels and our process and results for building models from each class and comparing them. In section 7, we present the design of a field trial for NYSDEC to test and select from among these models. Finally, section 8 identifies opportunities for future work.

[*] This work was initiated at the 2016 Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship at the University of Chicago.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

2. APPROACHES TO ENFORCEMENT

2.1 Current Approach

The bulk of environmental monitoring and enforcement resources in the United States are allocated to activities under the Clean Air Act (CAA), Clean Water Act (CAA), and RCRA. Self-reported pollution or waste management data is the primary source of compliance monitoring information on large facilities in these programs, but regulator inspections are used to confirm the accuracy of self-reported measures, and are often the only source of compliance information for smaller facilities [17].

Although legislation surrounding environmental policy is largely determined at the federal level, monitoring and enforcement responsibility in New York state falls primarily on NYSDEC, which conducts about one thousand RCRA inspections each year [7].

RCRA inspections can be triggered by citizen complaints, accidents ("for cause" inspections), or for administrative reasons ("neutral" inspections) [17]. Of the latter category, some inspections are mandatory. For example, facilities deemed significant non-compliers (SNCs) are periodically inspected until they return to compliance. Other inspections are more flexible: transportation, storage and disposal facilities (TSDFs) and large quantity generators (LQGs) are required to be inspected at least every two and five years, respectively. The remaining inspections are allotted at the region's discretion, usually according to national or regional areas of focus (e.g. facilities in the metal processing industry, concentrated animal feeding operations), geographical convenience (close to a scheduled inspection), or random selection.

The *hit rate* is defined as the proportion of inspections conducted that find a violation. In 2015, the overall NYSDEC hit rate was about 40%.

2.2 Our Approach

With guidance from NYSDEC, our work focuses on prioritizing the inspection of large quantity generators (LQGs) which have not been recently inspected. Below facilities will refer only to these facilities.

We frame the task as a binary classification problem: what is the likelihood that a given facility, if inspected in the next year, will identify a violation? Identifying more violations makes the inspection process more efficient, that is NYSDEC can find the same number of violations without conducting as many inspections, or find more violations without increasing the number of inspections. The model also provides a ranking over all facilities, meaning that if NYSDEC is able to conduct more inspections, it can do so dynamically and effectively. This approach incorporates information across regulatory programs to identify previously unidentifiable patterns of behavior associated with violation under RCRA statutes. The source code for our data and modeling pipeline available to the public.[8]

3. DATA SOURCES

The EPA collects and maintains a wealth of data across all of its enforcement programs, and RCRA inspectors routinely use the RCRAInfo database to identify likely violators based on compliance history on a case-by-case basis.

However, many RCRA facilities fall under at least one additional EPA regulatory program, and it is prohibitively difficult for inspectors to synthesize the information across all programs and all facilities. Furthermore, information is not always shared between regions: inspector knowledge and heuristics are not easily standardized across regions, and novel findings in one region are not automatically distributed to or put into practice in other regions.

The EPA's Office of Enforcement and Compliance Assurance (OECA) maintains monitoring, inspection, and enforcement data across all of its regulatory programs. It is common for a facility to be subject to regulation under multiple programs. Using the EPA's central Facility Registry (FRS) database, we linked these various sources, the most important of which are listed below.

Note that besides NYSDEC annual reports, all data is publicly available.

3.1 RCRAInfo

Characterizes facility status, regulated activities, and compliance histories on the generation of hazardous waste from large quantity generators and on waste management practices from treatment, storage, and disposal facilities. RCRAInfo contains data since 1984 on 110,000 facilities in the state of New York, of which about 32,000 are currently active. About 15,000 New York facilities have been inspection during this time, totaling approximately 68,000 inspections.

3.2 Biennial Reporting System (BRS)

Information regarding the generation, management, and final disposal of hazardous wastes regulated under RCRA for odd numbered years. It includes over 10,000 unique RCRA facilities that shipped waste between 2003 and 2013, comprising over 200,000 total shipment activities.

3.3 Integrated Compliance Information System (ICIS)

Incorporates federal enforcement and compliance (FE&C) case data on 10,000 New York RCRA facilities. Contains data related to Clean Air Act and Clean Water Act inspections such as discharge permits and monitoring reports.

3.4 NYSDEC Waste Manifests

Waste manifest data are the main supplemental dataset specific to New York State. These manifests are a part of New York State's system for tracking hazardous waste shipments to or from all generating and transporting facilities in the state. When hazardous waste is transported, stored, treated or disposed of, each agent that ships or receives data creates a manifest or receipt of this waste. These manifests include identifying data on the facilities, along with detailed data on the type and amount of waste shipped. Waste manifests are tracked via receipts from generator to transporter to disposal and at each step, there is tracking and checking of the waste and quantity, thereby monitoring waste from cradle-to-grave. Since 1990, there are manifests for over 6 million shipments generated by 95,000 unique facilities.

3.5 NYSDEC Annual Reports

As part of RCRA regulations, a biennial report is issued which summarizes (in greater detail) the information collected in the above two sources for large quantity generators (LQGs). New York State issues these reports on an annual basis, which we use to extract more detailed feature data on this subset of facilities. Since 2006, there are 20,000 annual reports available for 13,000 unique facilities.

4. EVALUATION METHODOLOGY

4.1 Temporal Cross-Validation

To evaluate our models, we use a cross-validation strategy that emulates the way in which our models would be deployed by NYSDEC. Since regions generate inspection lists at the beginning of each fiscal year, the cross validation is defined by a date t_0 and a training window dt. Facilities inspected within the dt window before t_0 are included in the training set. Therefore the training set can include multiple observations for a given facility, but at most one observation per facility per year.

The test set consists of all facilities inspected in the year following t_0. We score all facilities that were active at time t_0. Crucially, we only have outcomes for those facilities which were actually inspected.

4.2 Metrics

The model ranks all facilities eligible for inspection in a given year, and we can calculate the usual metrics such as precision, recall, and AUC by restricting to the subset of the test set for which we have labels. But those are not the metrics of interest for this problem. This is because NYSDEC inspects about 150 (or 5%) of facilities each year.

Thus in each test year we focus our attention on how a given model performs in the top k, that is the k examples predicted to be most at risk by that model. Generally we will restrict k to be less than or equal to the total number of labeled examples, i.e. inspected facilities, in that test set.

The *precision* in the top k is the proportion of labeled (inspected) examples in that set which are positive (violators). The greater the precision at the top and the more steeply sloped the precision curve, the better the model is at prioritizing likely violators.

We also scale precision on a given test set by dividing by the baseline hit rate that NYSDEC actually achieved in a given year. This allows us to aggregate precisions across test sets. This metric is called *lift*.

Thus another important metric is the *count* in the top k which we define as the number of labeled examples in the top k. Note that by definition $count(k)$ is at most k. Because of our missing labels, two models with the same precision in the top k can have differing counts. By definition

$$\text{precision}(k) = \frac{\text{count}_+(k)}{\text{count}(k)}$$

where $\text{count}_+(k)$ is the number of positive labeled examples in the top k. So for a fixed precision, the higher the count the more confident we are in that precision estimate.

We can formalize this to provide bounds on the estimate of precision in two ways. First, we can provide hard lower and upper bounds by assuming that the unlabeled examples are all negative or all positive, respectively. That is,

$$\frac{\text{count}_+(k)}{\text{count}(k)} \leq \text{precision(k)} \leq \frac{\text{count}(k) - \text{count}_-(k)}{\text{count}(k).}$$

These bounds, however, are not very informative so we prefer statistical bounds. Under the assumption that the labels in the top k follow a Bernoulli distribution, we can use a binomial proportion confidence interval. We prefer the Clopper-Pearson (exact) bounds because of the relatively small sample sizes in our context [11].

Source	Feature
RCRA	Time Since First Inspection/Violation Time Since Last Inspection/Violation Mean Time to Return to Compliance Maximum Time to Return to Compliance Number of Times Reported as SQG/LQG Number of Times Reported as Transporter
Manifests	Number of Waste Manifests Submitted Average Quantity of Waste Generated Minimum Quantity of Waste Generated Maximum Quantity of Waste Generated Types of Waste Generated
ICIS	Number of Enforcements Minimum Monetary Penalty Assessed Maximum Monetary Penalty Assessed Mean Monetary Penalty Assessed

Figure 1: Examples of features generated from various sources. These can be calculated at any spatio-temporal resolution.

5. FEATURE GENERATION

Facility violations are a spatio-temporal phenomenon, so naturally the features in our model are also spatio-temporal. The only characteristic of a facility that is static in time is its location. Most quantities of interest (e.g. the type and quantity of waste that a facility handles, how many times it has been inspected, whether it has been found in violation of RCRA regulations) can and do change over time. Therefore the majority of the features generated must be associated with both a location and a point in time.

Our approach to spatio-temporal feature generation can be defined in the following general terms. All of the data consists of events, meaning that each datum occurs at a *point* in time and space.[1] Thus given a *time period*, defined by a date (e.g. January 1st, 2016) and delta (e.g. 5 years, or all time), and a *spatial level* (e.g. facility, zip code), we can gather all observations that fall within that spatio-temporal window and apply an aggregation function (e.g. count, mean, maximum).

For example, we calculated the maximum time a facility had taken to return to compliance within the past year; the number of inspections conducted in a state over the past 5 years, and what proportion of those led to a violation; and the time since a facility was first inspected. For more examples see figure 1.

6. THREE TYPES OF MODELS

The naive approach to predicting violators is to train models on labeled (inspected) facilities. Such models are technically learning the likelihood that a facility is found in violation *conditional* on it being inspected. Denoting the random events of violation and inspection as V and I, respectively,

[1] Some data is more complex. For example, enforcement information for an inspection may not be received until well after the inspection occurs. We decompose such data into multiple events.

these models are estimating $P(V|I)$.[2]

However, if the goal of inspection targeting is to maximize the number finding violations, then the quantity of interest for each facility is of course the unconditional probability $P(V)$. As described above, the historical targeting of facilities for inspection by experts was was not at all (uniformly) random so we have

$$P(V|I) \neq P(V)$$

in general. This discrepancy between modeling $P(V|I)$ and $P(V)$ is evident in the context of our validation methodology.

We only have labels for about 2.5% of facilities in the universe in a given year. If our model performs well, then we would expect to have many labeled examples at the top only if the historical inspection targeting process did similarly well and selected similar facilities.

We should be wary wary of scoring the entire population of facilities using the $P(V|I)$ model: this model was only trained on inspected facilities. Since the facilities were not selected for inspections uniformly randomly, we are applying the model outside of the population of the training set.

There are three ways to deal with this issue. The most direct way is to simply collect additional labels for facilities that have low historic $P(I)$. However, this process is time-consuming and expensive, so we augment it with two further modeling strategies that work in opposite directions.

The first is to model the probability of violation *and* inspection, $P(V\&I)$, which, given the contribution of the probability of inspection, will be reinforcing the historical targeting process, albeit identifying new facilities. The second is to try to correct for the statistical bias of the historical targeting process by modeling the desired $P(V)$, allowing us to target the facilities with the highest probability of violation.

The remainder of this section is organized as follows: 6.1 presents our model selection process in-depth for $P(V|I)$ models; sections 6.2 and 6.3 explain the theory and practice of the $P(V\&I)$ and $P(V)$ models; 6.4 describes a comparison of the best of each of these three types of models.

6.1 Conditional Probability of Violation

We trained a variety of standard binary classification models including regularized logistic regression, SVM and random forest and performed a grid search over both model hyper-parameters and feature generation parameters.

Figure 2 shows the precision for the best model of each class in the year 2015. We found the different model algorithms to perform comparably with no single model dominating in all cross-validation years. This could be because we have incorporated much of the complexity of the task in our features and all models are able to take advantage of them. See figure 3 for a comparison across years. On average our lift in the top (equal to the approximate number of inspections performed by NYSDEC) was about 1.5x or a 50% improvement over the baseline.

We also looked at the recall in the top for each of these models. Gradient boosted trees and logistic regression had somewhat better recall than SVM and random forest. See figure 4.

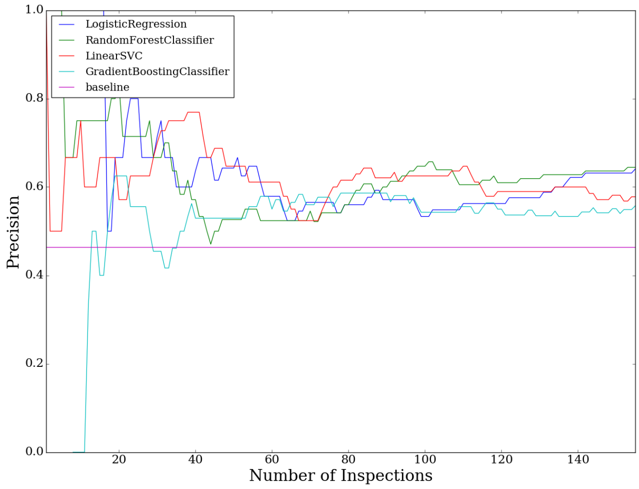


Figure 2: Precision curves at the top in 2015 for the best models of each algorithm.

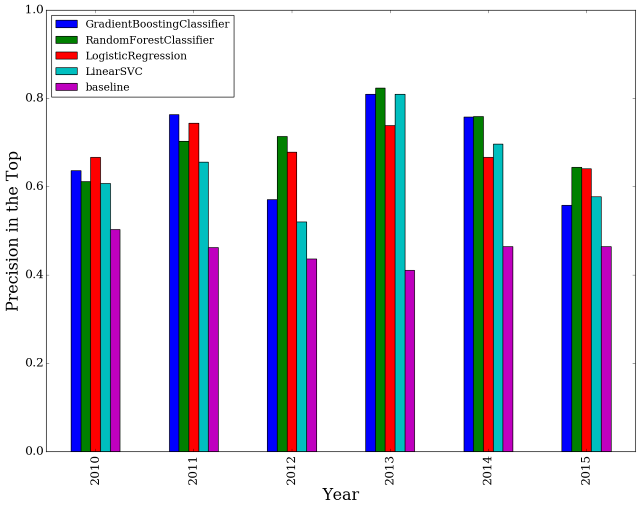


Figure 3: Precision on the top labeled examples by model algorithm by year. All models perform similarly with lift averaging about 1.5x.

[2] Models like random forest do not explicitly estimate probabilities unless they are calibrated so we use this language and notation with some caution.

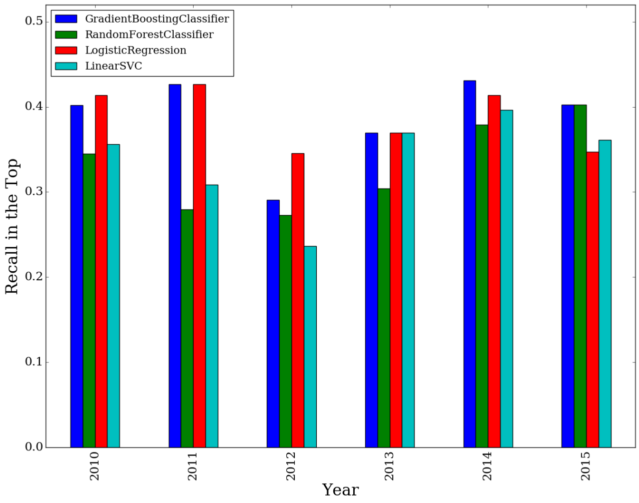


Figure 4: Recall on the top labeled examples by model algorithm by year.

When varying the number of training years we found steady improvements in performance up until about 8 years of data, after which performance plateaued. See 5. Note that in all of these models the spatiotemporal aggregation features contain information summarizing many more (in fact all) years of data for those facilities in the training set.

To evaluate the usefulness of the features in the model we looked at logistic regression coefficients, SVM margins, and random forest feature importance. See figure 6 for a list of the highest ranked features. The most surprising feature that surfaced as a highly predictive was a binary feature indicating whether the mailing address matched the physical address of a facility. Through conversations with inspectors, this seems to be indicative of large companies (with off-site headquarters), which tend to have more robust compliance and legal practices in place.

To measure the usefulness of each of the datasources mentioned in section 3, we ran models with just one of those data sources at a time. We found that, individually, the RCRAInfo investigations and NYSDEC waste manifests data are the most important. We also found that excluding the NYSDEC annual reports had no effect, so that the model effectively only uses public data.

6.2 Probability of violation and inspection

Unlike violation (V), the compound event of violation *and* inspection ($V\&I$) is observed for each facility each year. Note that negative examples under this model are facilities which were either inspected and not found in violation *or* not inspected. Thus there is no sample selection bias in training this model. However, since about 5% of facilities are inspected each year and of those only about half find violations,are imbalanced with only about 2.5% positive labels in each training set. Consequently, we use a balanced random forest, i.e. a random forest whose bootstrap samples contain equal numbers of examples from each class [14]. Three years of training data are sufficient for this model, perhaps because there are significantly more examples (about 3000) per year.

A model trained in this way performs well. Note, however,

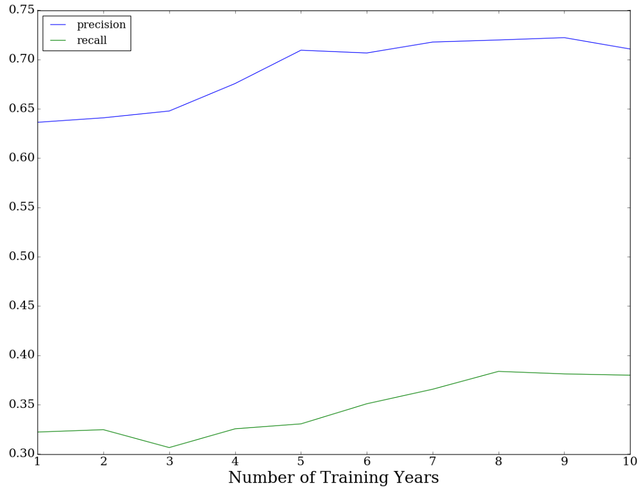


Figure 5: Precision and recall in the top for random forests with different numbers of training years, averaged over prediction years. Performance increases until about five training years.

Feature
Time since first and last handler registration
Total amount of waste shipped
Variance in amount of waste shipped over time
Proportion of inspections finding violations
NAICS industry
Time since last violation
Whether or not facility and mailing addresses are co-located
Total number of investigations

Figure 6: The most important features in a random forest model.

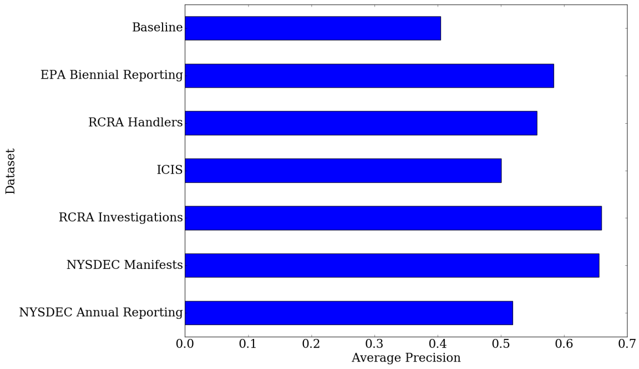


Figure 7: The average precision in the top for random forest models using features from only one dataset.

Type	$P(V\|I)$	$P(V\&I)$ direct	$P(V\&I)$ factored
Precision	.709	.585	.626
Recall	.331	.461	.464

Figure 8: Precision and recall at the top averaged over years for three random forest models. The $P(V\&I)$ models have significantly better recall at the top than the $P(V|I)$ model at the expense of precision. The factored model $P(V|I) \times P(I)$ improves in precision over directly fitting $P(V\&I)$.

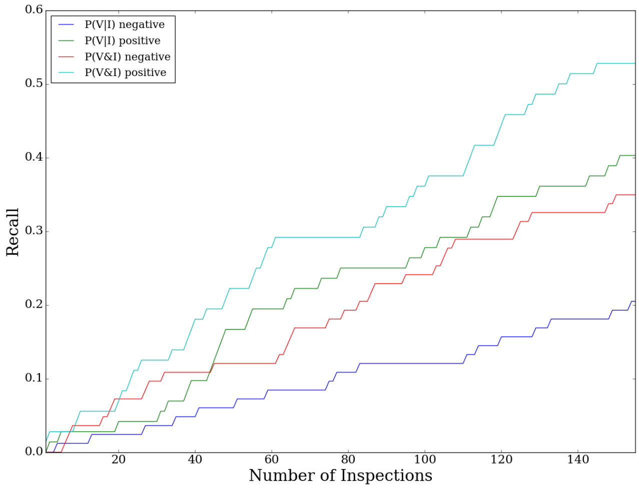


Figure 9: Positive recall counts.

that

$$P(V\&I) = P(V|I) \times P(I).$$

Thus another way of estimating $P(V\&I)$ is by factoring into two models: $P(I)$, and $P(V|I)$ which we already estimated in section 6.1. From another perspective, we are reordering the predictions of the $P(V|I)$ model according to their probability of inspection. We can also think about $P(I)$ as the degree to which we believe that our $P(V|I)$ model is applicable to any given facility.

We find that this product model is a significant improvement over the direct estimate in terms of precision while maintaining the same advantage over the original $P(V|I)$ model in terms of precision. See figure 8 for a summary. Below we will refer to this product model as the $P(V\&I)$ model.

We can decompose the labeled examples at the top into positive and plot recall curves for each class. See figure 9 for these curves. We see that $P(V\&I)$ significantly outperforms $P(V|I)$ in positive recall, with decrease in negative recall. Intuitively, multiplying by the probability of inspection has elevated positive examples into the top, without significantly changing the distribution of negative examples.

Inspectors at NYSDEC remarked that the $P(V\&I)$ predictions were more sensible than those of $P(V|I)$. This is not surprising since the former is designed to learn from their own targeting methodology.

	$P(V\|I)$	$P(V\&I)$	$P(V)$
$P(V\|I)$	1	.494	.860
$P(V\&I)$		1	.464
$P(V)$			1

Figure 10: Similarity at the top between model types, averaged over years.

Type	$P(V\|I)$	$P(V\&I)$	$P(V)$
Precision	.709	.626	.675
Recall	.331	.464	.313

Figure 11: Precision and recall at the top for each model type, averaged over years.

6.3 Probability of Violation

The fact that we only have labels for a small, non-randomly sampled subset of the population of interest is an instance of sample selection bias. The literature on sample selection bias ”correction” theory starts with the work of Heckman who developed a method of estimating and correcting for the bias in the context of linear models and under assumptions of normality [9].

Cortes, et al. present a simple and general sample bias correction framework [10]. Under the assumption that every example has a positive probability of being included in the sample, we can reweight the sampled examples to approximate the true distribution of examples. That is, assume $P(I) > 0$ over the population, where I is the probability of being included in the sample. In this case, learning a model from the entire population is equivalent, in expectation, to learning from a sample after reweighting each example x by $P(x)/P(x|I)$. By Bayes' rule this is inversely proportional to $P(I)$, which we have already modeled in section 6.2.

We implemented this reweighting by training a regularized logistic regression for $P(I)$. We did this because logistic regression naturally estimates probabilities, unlike the random forest $P(I)$ we trained in section 6.2 above. Next we inverted these probabilities and provided them as sample weights to the same random forest that we selected for $P(V|I)$ in section 6.1. See the next section for a comparison to previous models.

6.4 Inter-model Comparison

For a given prediction year and two models, we define the similarity k to be the Jacard similarity between the top k sets of the models:

$$\text{similarity}(k; M_1, M_2) = \frac{\#(\text{top}(k; M_1) \cap \text{top}(k; M_2))}{\#(\text{top}(k; M_1) \cup \text{top}(k; M_2))}$$

where $top(k; M)$ is the set of facilities ranked in the top k by model M. The similarity between two models lies between 0 (no facilities in common) and 1 (all facilities in common).

Figure 10 shows the similarity between the three model types we have developed, averaged over years. While the reweighted $P(V)$ model is quite similar to the original $P(V|I)$ model, the $P(V\&I)$ model is quite different from either.

In terms of precision and recall, we find that the $P(V)$ model suffers slightly compared to the $P(V|I)$ model. See figure 11 for raw numbers. Another perspective is to incorporate the number of missing labels in the top into the precision using binomial proportion confidence intervals described in section 4.2. Then compared to $P(V|I)$, $P(V)$ has a slightly lower precision and wider confidence interval

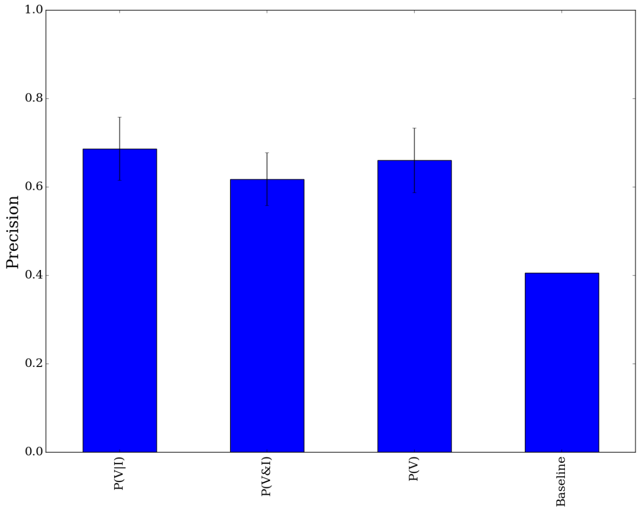
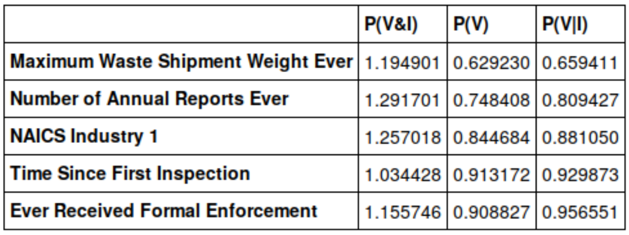


Figure 12: Precision of each model type pooled across all years. Error bars are 95% binomial proportion Clopper-Pearson confidence intervals as described in section 4.2.



	P(V&I)	P(V)	P(V|I)
Maximum Waste Shipment Weight Ever	1.194901	0.629230	0.659411
Number of Annual Reports Ever	1.291701	0.748408	0.809427
NAICS Industry 1	1.257018	0.844684	0.881050
Time Since First Inspection	1.034428	0.913172	0.929873
Ever Received Formal Enforcement	1.155746	0.908827	0.956551

Figure 13: For each model type and each feature, the table contains the ratio of the mean of that feature on the top set for that model, and the mean in the population.

while $P(V\&I)$ has a substantially lower precision but narrower confidence interval. See figure 12. Figure 13 shows crosstabs for select features between the top sets of each model. These features were selected by fitting a multiclass decision tree with examples equal to the union of top sets across the three model types across all years. The outcome was the subset of model types for which the example appears in the top predictions.

We found that, compared to the population, the facilities selected by the $P(V\&I)$ model were more likely: to have higher maximum waste generation weight; have more annual reports recorded; belong to NAICS industry 1 (agriculture, forestry, fishing and hunting); have received their first inspection earlier; and to have ever been subject to an enforcement. As mentioned above, the facilities selected by $P(V|I)$ and $P(V)$ were similar; one of the few features that discerns them is that $P(V)$ selections are even less likely to have had an enforcement.

These observations support the intuition that the $P(V)$ model is selecting larger and more typical facilities for investigation, while the $P(V)$ model is selecting more novel facilities.

7. IMPLEMENTATION

We are working with NYSDEC on a preliminary validation of these methods. In the long term, a rigorous validation of our models requires a statistical framework for comparing methods for targeting inspections.

7.1 Field Trials for Targeted Inspections

By a *targeting method* we mean a function which, given a list X of entities (facilities) and a number k of inspections to perform, selects a subset of X of size k. Let $y(X')$ denote the number of violations in a subset X' of X.

Let E_k and M_k denote the lists of k facilities chosen by an expert and a model (or two different models), respectively. Then we wish to know whether there are more violations on the model list than on the expert list. That is, we will test the alternative hypothesis

$$H_1 : y(M_k) > y(E_k) \qquad (1)$$

against the null hypothesis

$$H_0 : y(M_k) = y(E_k). \qquad (2)$$

If we were able to perform (up to) $2k$ inspections instead of k inspections, we could inspect the union of the two lists and directly compare $y(E_k)$ and $y(M_k)$. However, those resources are not available, so we need to test the hypothesis statistically.

It is tempting to test the hypothesis using a randomized controlled trial design in which we randomly partition X into equal halves X' and X'', apply one targeting method to each half, and then perform $k/2$ inspections from each half, thus observing $E_{k/2}(X')$ and $M_k(X'')$.

However, in this context the unit to which the treatment (targeting method) is applied is *all facilities* (X' or X'') rather than a single facility, making the sample size 1 instead of $N/2$. Furthermore, the quantities observed in this here would be $y(E_{k/2}(X'))$ and $y(M_{k/2}(X''))$, which can be shown to be biased estimators of the quantities of interest $y(E_k(X))$ and $y(M_k(X))$, respectively.

To remedy these issues we will use the following study design which can be shown to provide an unbiased test of the above hypothesis.

1. Experts select E_k from X, model selects M_k from X.
2. Randomly sample k facilities from $(E_k \cup M_k)$ and inspect them.
3. Conduct the hypothesis test comparing the means (proportions of violators) of the two groups.

We plan to conduct a field trial following this design during the NYSDEC 2017 fiscal year.

8. FUTURE WORK

The outcome modeled in this work has been restricted to *whether* a violation occurred. As noted in section 2, not all violations are equally severe. Regulators are more interested in identifying violations that will lead to a formal civil or criminal enforcement than those that lead to an informal administrative enforcement. We have done some preliminary modeling of this outcome; however, it is significantly more complex in two important ways. First, the outcome is not known at the time of inspection. The average time between inspection and a formal enforcement action outcome is over one year, with some taking over ten years, and no maximum time window after which a label is guaranteed.

Second, the baseline of formal enforcement actions among all inspections is 5% so the classes are significantly imbalanced. We plan to explore this outcome as well as more specific violation types in future work.

Another direction for future research is a better understanding of the relationship between the immediate goal of maximizing the number of inspections finding violations and the comprehensive goal of minimizing violations at large– especially those that are the most damaging to human and environmental health. Inspections of a facility, and enforcement actions against it, serve in the short term to prevent that facility from violating, but in the long term the inspections process aims to deter all facilities, including the uninspected, from violating. A better understanding of these two effects is essential to achieving the broader goals of the hazardous waste management [15].

The setting in which there is a large population of examples but only a small number of labels has connections to other areas of research. We are exploring techniques from semi-supervised learning to improve our model by using both labeled and unlabeled data [12]. However, in addition to improving the classifier, we have the further goal of improving our estimates of its performance in the top. Another related area is active learning[13], which optimizes queries for labels on unlabeled data. In our setting, this means inspecting a facility not necessarily because it has a high likelihood of violation (positive label) but because it would be informative to improve the model.

9. CONCLUSION

We used insights from a variety of data sources to inform an innovative, data-driven approach to targeting inspections of facilities that generate, transport, and dispose of hazardous waste. Our approach has the potential to both increase the number of inspections finding violations and address the statistical biases present with data gathered through expert targeting. We have initiated a framework for evaluating machine learning models in the presence of missing labels, which is by definition commonplace in resource allocation problems. We anticipate that this project can serve as a prototype for predictive analytics projects in other regulatory programs.

10. ACKNOWLEDGMENTS

This work was done as part of the Eric & Wendy Schmidt Data Science for Social Good Fellowship through the University of Chicago. We thank our collaborators at NYSDEC including Paul John, Hassan Hussein, Mike Lintner, Sharif Rahman, and Stephen Zahn.

11. REFERENCES

[1] Edmondson, Brad (2001). "Environmental Affairs in New York State: An Overview." New York State Archives.

[2] Right-to-Know Network: National Response Center Spills and Accidents Database. `<http://www.rtknet.org/db/erns/>`. Accessed: 2016-01-10.

[3] "5 Years of RCRA: Building on Our Past to Protect Our Future." `<http://nepis.epa.gov/Exe/ZyPDF.cgi/10000MAO.PDF>`. Accessed: 2016-02-05.

[4] "Compliance Monitoring Strategy for the RCRA Subtitle C Program." `<http://www.epa.gov/sites/production/files/2013-11/documents/rcracms.pdf>`. Accessed: 2015-11-23.

[5] Giles, Cynthia. "Next Generation Compliance." The Environmental Forum: September/October 2013. `http://www.epa.gov/sites/production/files/2014-09/documents/giles-next-gen-article-forum-eli-sept-oct-2013.pdf`

[6] Kaltech Industries Building Explosion. U.S. Chemical Safety and Hazard Investigationn Booard. Investigation Digest. April 25, 2002. *http://msc.fema.gov/portal*

[7] EPA Envirofacts Database. `<http://www3.epa.gov/enviro/>`. Accessed: 2016-02-08.

[8] RCRA. GitHub Repository. `<https://github.com/dssg/rcra`. Accessed: 2016-02-08.

[9] Heckman, J. J. (1979). "Sample Selection Bias as a Specification Error." Econometrica 47: 153.

[10] Cortes, Corinna; Mohri, Mehryar; Riley, Michael; Rostamizadeh, Afshin (2008). "Sample Selection Bias Correction Theory." Algorithmic Learning Theory 5254: 38-53.

[11] Clopper, C.; Pearson, E. S. (1934). "The use of confidence or fiducial limits illustrated in the case of the binomial". Biometrika. 26: 404âĂŞ413.

[12] Zhu, Xiaojin and Goldberg, Andrew B (2009). "Introduction to Semi-supervised Learning." Synthesis Lectures on Artificial Intelligence and Machine Learning, Vol 3: 1-130.

[13] Settles, Burr (2012). "Active Learning." Synthesis Lectures on Artificial Intelligence and Machine Learning.

[14] Chen, Chao, Andy Liaw, and Leo Breiman. "Using Random Forest to Learn Imbalanced Data." University of California, Berkeley (2004).

[15] Harcourt, Bernard. "Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age." University of Chicago Press, 2008.

[16] Chemical Safety Board's (CSB) Findings in New York Chemical Waste-Mixing Incident. Leadership ViTS Meeting, March 6, 2006.

[17] Shimshack, Jay P. "The Economics of Environmental Monitoring and Enforcement." Annu. Rev. Resour. Econ. 2014. 6:339-60.