Final Machine Learning Project 2017

Predicting High Priority Violators of Clean Air Act

Group Members
Andrew Yaspan, Regina Widjaya, Shambhavi Mohan, Sun-joo Lee

1. Background and Introduction

Air pollution is a significant global issue for several immediate and long-term reasons. People who live in areas in the vicinity of facilities that contribute to air pollution are exposed to unsafe levels of hazardous materials that can have serious health consequences, such as increased rates of cancer, developmental and reproductive effects among others[1]. Moreover, air pollution has long-term environmental impacts on the Earth's ability to block UV light and expel excess carbon dioxide, increasing the risk of acid rain and global warming.[2]

In order to prevent life threatening air pollution, the United States Congress originally passed the Clean Air Act in 1963, and successively amended the Act in 1970, 1977 and 1990. The Clean Air Act Risk Management Program (RMP) requires facilities to self audit for the purpose identifying hazardous substances and minimizing the risk of accidental release of these substances. Full compliance with the RMP inspection guidelines also mandates on-site or independent verification of the information presented by the facilities. Although the EPA has the capacity to monitor air pollution and inspect facilities to make sure facilities are not committing air safety violations, the reality is that it can only inspect a minority of the facilities in the US.

2-3. Related Work, Problem Formulation, and Solution Overview

There are many ways that facilities can be in violation of the Clean Air Act, but the most serious violators are designated as 'High Priority Violators' (HPV, also High Priority Violations). A facility can be designated as a HPV not only by failure of an inspection, but also by failing to obtain permits, complying to all reporting requirements, etc.[3] As policy-advisors, we are mainly concerned with HPV's that are detected by an inspection because the EPA can only conduct a limited number of inspections per year. Currently, the EPA has just two major ways of selecting facilities for inspection: 1) facilities with high residential population or 2) facilities with a history of violations or have not filled out the RMP Inspection report.[4] However, apart from this, the EPA is flexible in its selection of facilities to inspect. We are not aware of any other selection method or algorithm the EPA is relying on to pick facilities for inspection.

The problem is that serious violators may be able to continue to pollute without being detected, since the selection criteria for inspection is not robust. The issue at hand is to predict facilities that are likely to be found an HPV an inspection, so that the EPA can prioritize these facilities for inspection. We are proposing a machine learning-based approach to finding a model that predicts HPVs. This is expected to increase the detection of serious violators and

[1] https://www.epa.gov/haps/about-health-effects-fact-sheets
[2] https://www.nationalgeographic.org/encyclopedia/air-pollution/
[3] https://www.epa.gov/sites/production/files/documents/hpvmanualrevised.pdf
[4] https://www.epa.gov/sites/production/files/2013-10/documents/clean_air_guidance.pdf

help decrease air pollution. This will be especially important during a time where the funding and personnel capacity of the EPA is in question, and can very easily be reduced.[5]

4. Data Description

The EPA's Enforcement and Compliance History Online (ECHO) website provides publicly available data regarding facility-specific compliance and enforcement information. We used data from the following tables:[6]

a. **Air Full Compliance Evaluations (FCEs) and Partial Compliance Evaluations (PCEs)** (ICIS-AIR_FCES_PCES.csv) - This was a key data source containing data on all of the evaluations conducted on the facilities from 1974 to the present. A Full Compliance Evaluation (FCE) is a comprehensive evaluation to assess compliance of the facility as a whole and results in a compliance determination. For the purposes of this policy project, "facility" is used in the broadest sense of the term incorporating all regulated emission units within the facility. A Partial Compliance Evaluation (PCE) is a documented compliance evaluation conducted for the purpose of making a compliance determination and focuses on a subset of processes, regulated pollutants, regulatory requirements, or emission units at a given facility. The dataset, however, does not include the results of the evaluations.

b. **Case Files HPVs and FRVs** (ICIS-AIR_VIOLATION_HISTORY.csv) - This is another one of our main data sources. It contains the violation history, both HPV and FRV (Federally Reportable Violation) designations, of the facilities from 1975 to the present. The data is organized by date of the designation of HPV (or FRV) and also includes the end date of the designation.

c. **Air Title V Certifications** (ICIS-AIR_TITLEV_CERTS.csv) - Title V Major Facilities have the potential to generate large amounts of air pollutants like NO2, SO2, particulate matter, etc. The data specifies whether a Title V deviation occurred in the facility (FACILITY_RPT_DEVIATION_FLAG) by date and the agency that was in charge of the monitoring the facility a the time (Local control region(LCON), state, or EPA).

d. **Air Stack Tests** (ICIS-AIR_STACK_TESTS.csv) - A stack test measures the amount of pollutants being emitted and demonstrates the efficiency of a capture system. The data contains Stack Test compliance (full, partial, etc) of the facility by date and the agency (Local control region(LCON), state, or EPA) that was in charge of the monitoring. It contains some details of all the facilities that were inspected. Stack Test results can lead to HPV based on the degree/intensity of violation.

[5] https://www.nytimes.com/2017/03/15/us/politics/budget-epa-state-department-cuts.html?_r=0
[6] https://echo.epa.gov/tools/data-downloads/icis-air-download-summary#datasets

e. **ICIS-Air Formal Actions** (ICIS-AIR_FORMAL_ACTIONS.csv) - A formal action is a judicial or litigious action taken to address violations or is anything that has a penalty clock. The data contains details of the type of action taken against the facility by date fulfilled and the agency (Local control region(LCON), state, or EPA) that was in charge of the monitoring.

f. **ICIS-Air Informal Actions** (ICIS-AIR_INFORMAL_ACTIONS.csv) - An informal action curtails identification of violators, consultation, and correspondence between sources, the States, and EPA before a formal enforcement action is commenced. The data format is similar to Formal Actions data.

5. Generating Machine Learning Models

a. Methods:

We intended for a policy-maker in the EPA to use our tool at the beginning of the year to draw up its inspection plan of facilities, with a focus on facilities that are likely to be an HPV in the coming year. Thus, all of our data were aggregated by year, and we limited our analysis to data from 2000 to 2016 (2016 data was reserved for final testing of the best performing models).

Since the issue is to detect HPV facilities through inspections, the key data sources were the Air Full Compliance Evaluations and Case Files HPVs and FRVs files. The first lists all the evaluations conducted on the facilities by various authorities by date, while the second lists the HPV history of each facility. The HPV Case Files was joined into the Compliance Evaluations on the facility ID and date to determine whether the evaluation resulted in an HPV designation or not. This became the basis of the labels. The Compliance Evaluations was also joined into the HPV Case Files to filter out the HPV violations that were detected through means other than an inspection. This was used as a feature in generating models.

b. **Features:**

The characteristics of previous years' violations and deviations constituted the bulk of our features. The count of HPVs in previous years was a key feature (as opposed to the label, which is a binary). FRVs (Federally Reportable Violations), which were discluded in the label, were added back in as a feature in generating the predictive models. Other features regarding the evaluations and tests conducted on the facilities included the party that conducted the evaluation or reported the violation (i.e., State, EPA, or local government), and the type of evaluation (e.g., full/part, on/off site, emissions observation, audit, report review, etc.).

The results of more regular tests and deviations were also included as features. The result of Title V deviations were included as counts per year. Such deviations can be detected much more frequently than evaluations, so they could reveal key insights into future serious violations. Formal actions that were taken against the facility were included as categorical variables (e.g. administrative, judicial, etc.), in addition to any penalty amounts imposed on the facility. Any informal actions taken against the facility (e.g. notices, letters issued) were also included.

The facilities fall under different types of programs, which describes the regulatory areas that the facility is subject to. There are numerous programs at the national and local levels, including such programs as The Mandatory Greenhouse Gas Reporting Rule, New Source Performance Standards, Recycling & Emission Reduction Programs, State Permit Programs, etc. Any programs that applied the facility at the time in which violations were detected were included as categorical variables. In addition, the type of pollutants detected during evaluations and tests were included as features. Finally, some constant characteristics of the facility, such as the state in which the facility is in and the local control region, were included as features.

Some features that would have been useful were eventually not added because the data could not be matched with the label in a significant way. For instance, facilities that were evaluated largely did not match the facilities that were inspected for stack tests. Since we did not want to lose so many observations and cannot assume the results of stack tests that were not performed, we did not impute them and the stack test results were not included as a feature (AIR_STACK_TEST_STATUS_CODE, ENF_TYPE_CODE were excluded).

c. Models:

Our approach was to take advantage of the relative ease of fitting our feature data to a variety of different types of machine learning models. Given that the members of our group are new to the application of machine learning, and, even more, our audience is likely largely unfamiliar with it as well, we thought that the ‘cast a wide net’ approach would be helpful for our learning and for understanding which features are important, or parameters need “tweaking.” Most of the models we choose to use were ones that we learned about in class, or learned about by viewing the code of machine learning practitioners. Model selection itself proved to be a helpful learning exercise.

Logistic Regression	KNN	Decision Trees	Random Forests	Support Vector Machines
Stochastic Gradient Descent	Extra Trees Regressor	Adaboost Regression	Gradient Boost Regression	Naive Bayes

We were expecting logistic regression, and the ensemble methods to be the best performing models based on the fact that we were predicting binary outcomes of either predicted to violate or not predicted to violate. Logistic regression is a natural choice for such binary outcome predictors, since the result of the model is a score between 0 and 1 that with a given threshold can be used to predict either a violation (1), or no violation (0).

The ensemble method we were most familiar with, and thought would perform the best, was Random Forests. We believed it would be a good model for this problem for two reasons. 1) It utilizes the splitting of a training set at various depths based on an increase in entropy greater than a particular threshold or number of values on either side of a split (of decision trees) identifying important features in order of importance in combination with 2) the consideration of overfitting that an ensemble method generally handles better than pure one-type model.

We may have expected K-nearest neighbors to perform reasonably well too, however, this method is extremely prone to overfitting. We also ran Support Vector Machine models, knowing that if this model produced a high accuracy, our classifier would likely be linear.

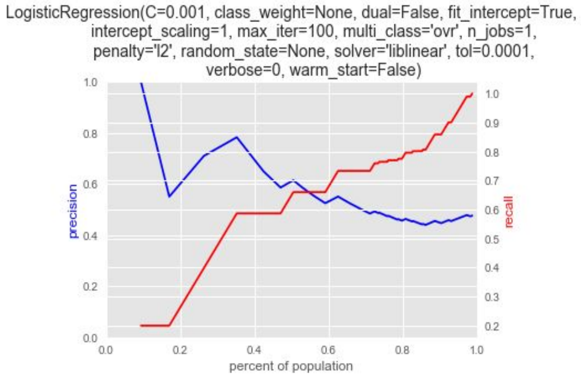
d. Analysis of Models

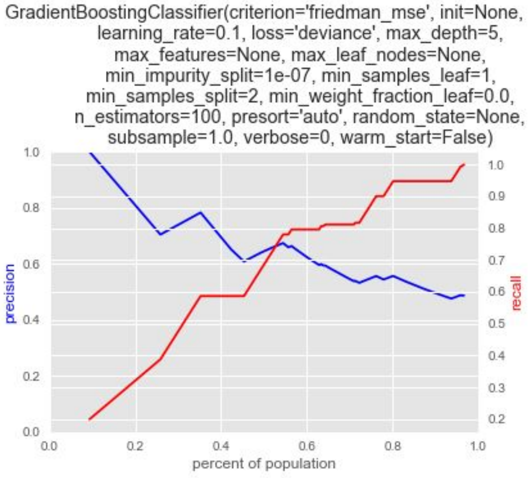
For the purposes of the policy question, we considered recall and the F1 score the most important performance metric. Since the EPA is already aware of some of the HPV facilities based on past violations, we thought it important for our model to be able to help the EPA detect additional potential HPVs that it otherwise would not have been able to detect. We also thought precision was important, so we considered the F1 score in the performance metric as well.

Based on the fact that we do not have the domain expertise to point out the presence of collinearity, in the case of logistic regression, or other types of redundant or extraneous features in our model we are pretty certain that some of the models we ran, especially the well performing one-type models, have overfit predictions. Given more time and the ability to have direct interactions with EPA inspectors or researchers, perhaps we would have been able to select only the necessary features. With that, it appears that the highest performing models across all evaluative scoring methods are ensemble methods. It seems that ensemble tree (gradient boosted and extra tree) methods outperform the other methods. This is reassuring since ensemble methods tend to do the best to control overfitting.

Further, all of our models performed best when we used the dataset from all of the previous years (year 2000 to the year previous to testing data). We think that ensemble methods and using data from the full range of years helped performance by offsetting some of the effects of having a smaller number of final data points.

We ran the final test on year 2016 data on the Gradient Boosting models as they performed the best on the recall and F1 scores. The two graphs below show how Gradient Boosting improves on the Logistic Regression model. Gradient Boosting weights the incorrect prediction points in each successive iteration, thereby enhancing a weak learner to a strong learner. The graph on the left shows a Logistic Regression model’s recall curve, which is jagged (lower half) and somewhat concave (upper half). Through Gradient Boosting at max depth 50, the recall curve was improved to more or less a neutral, straight line.

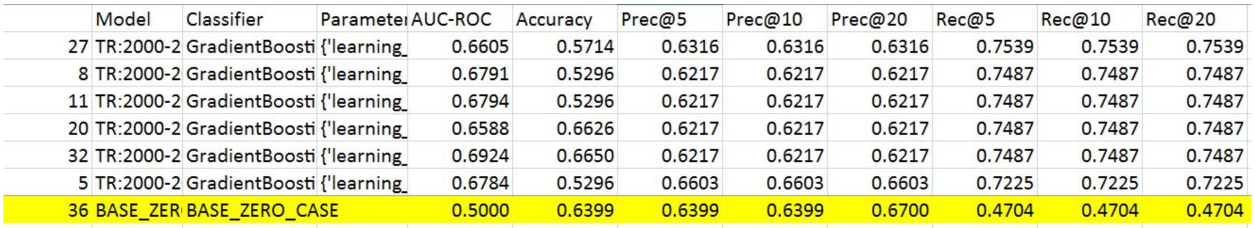




6. Evaluation

To test our models, we created test baseline cases (BASE_ZERO_CASE) where all observations were designated as 1 (= HPV). Our chosen model performs better than our baseline on recall (at 5%), meaning that it is better able to detect probable HPV facilities. It also performs better on Area Under the Curve (ROC), Accuracy, and Precision. This suggests that the facilities that we predict as HPVs are more often correct than a random designation of all 1s.

Best performing models, based on Recall at 5%



	Model	Classifier	Parametei	AUC-ROC	Accuracy	Prec@5	Prec@10	Prec@20	Rec@5	Rec@10	Rec@20
27	TR:2000-2	GradientBoosti	{'learning_	0.6605	0.5714	0.6316	0.6316	0.6316	0.7539	0.7539	0.7539
8	TR:2000-2	GradientBoosti	{'learning_	0.6791	0.5296	0.6217	0.6217	0.6217	0.7487	0.7487	0.7487
11	TR:2000-2	GradientBoosti	{'learning_	0.6794	0.5296	0.6217	0.6217	0.6217	0.7487	0.7487	0.7487
20	TR:2000-2	GradientBoosti	{'learning_	0.6588	0.6626	0.6217	0.6217	0.6217	0.7487	0.7487	0.7487
32	TR:2000-2	GradientBoosti	{'learning_	0.6924	0.6650	0.6217	0.6217	0.6217	0.7487	0.7487	0.7487
5	TR:2000-2	GradientBoosti	{'learning_	0.6784	0.5296	0.6603	0.6603	0.6603	0.7225	0.7225	0.7225
36	BASE_ZER	BASE_ZERO_CASE		0.5000	0.6399	0.6399	0.6399	0.6700	0.4704	0.4704	0.4704

7. Sample Experiment Design

To test whether our selected models are effective, we would recommend the EPA to conduct the following experiments.

A) Validation experiment:

Goal: The first experiment is designed to validate that our models are actually effective. In this experiment, the inspections will be divided into three groups.

a) Existing Internal Approach (Control Group I):

Facilities in this group will be selected for inspection based on the existing EPA selection criteria, namely based on past HPV data and inspectors' intuition.

b) Random Approach (Control Group II):

Facilities in this group will be selected for inspection based on random selection (e.g. every fifth facility will be inspected).

c) Treatment Group (Experimental Group):

Facilities in this group will be selected based on the HPV prediction of our model.

The assumption in each group is that a High Priority Violator is being inspected. We would then measure the precision of HPV detection in each of the groups. If the precision is higher in the treatment group (the rate of HPV facilities is higher in the group) than in the control group, particularly the first control group, it would show that our model is improving the EPA's inspection selection process.

B) Further Experimentation:

Goal: Further cement our confidence in the model and expand the scope of our prediction to include facilities that have yet to be inspected.

a) **Existing Internal Approach (Control Group I):**

Facilities in this group will be selected for inspection based on the existing EPA selection criteria, namely based on past HPV data and inspectors' intuition.

b) Treatment Prediction Group (Experimental Group I):

Facilities in this group will be selected based on the HPV prediction of our model.

c) Treatment Exploratory Group (Experimental Group II):

Use our (reduced) model to predict future violators from the pool of facilities that has never been inspected before.

If the model is proven to be reliable for our third sample group (Experiment Group II), the model will further benefit the EPA in targeting facilities that historically have been out of their inspection network.

9. Policy Recommendations

Since we have built a model to identify which EPA monitored facilities are likely to become a High Priority Violator of the Clean Air Act, our recommendation would be that when planning inspections for the next year to consider the binary scores we attribute to a facility based based on data collected from the previous years. This would enable the EPA to use limited resources to inspect facilities that are highly suspect and increase the probability of finding a HPV in each inspection. And, while it is always important to use inspector's time most efficiently, we would also hope that the EPA would use this information as food for thought if targeting certain facilities for air pollution prevention interventions. Perhaps it could even be used to find ways to promote more self-monitoring. Lastly, if our model is helpful, it could help the EPA advocate for more funding based on the department's effectiveness and efficiency.

Additionally, we would like to suggest that the EPA collect and format facility data in a way that is more useful for analysis. This would improve the reliability of outcome labels and models created from the data, and accessibility to the public.

10. Limitations and Caveats

Our biggest limitation was that we had to work with very segmented datasets that proved hard to merge. Unfortunately, the result of these merges were over 100,000 rows of mostly NaN values. Since this was such a large proportion of our data (roughly 90%), we had to drop these observations instead of imputing feature values.

In addition, not all facilities were inspected for HPVs each year, which resulted in dropping even more observations. Even when we took care of these issues, based on facility_id (PGM_SYS_ID) by year, there are still questions about the reliability of this data. An additional consequence of the need to merge this data in a very broad way was the limited number of observations we yielded, which is characteristic of inspection-based problems. More specifically, various tests (Stack Test or Title V certs) were not performed on all facilities. This weakened our models, since they could not build off of many data points.

Lastly, a big limitation was time. We spent a lot of time pivoting to a new project two weeks after we selected our original project. This pivot required us to find a new project, and try and understand a new dataset very quickly, and then figure out the best way to combine and aggregate this data, so that we could run our magic loop making it possible for us determine the best model to make our predictions.