Tim Hannifan, Alec MacMillen, Quinn Underriner
Group 12
Project Progress Report
5/21/2019

In our initial assessment, we thought that all large hazardous waste treatment, storage and disposal facilities (TSDFs) were being inspected every two years, as per federal law. This, however, is not the case. Budgetary/resource constraints at the EPA prevent full coverage, which strengthens the case for using a priority-targeting model to improve the agency's return on their limited resources.

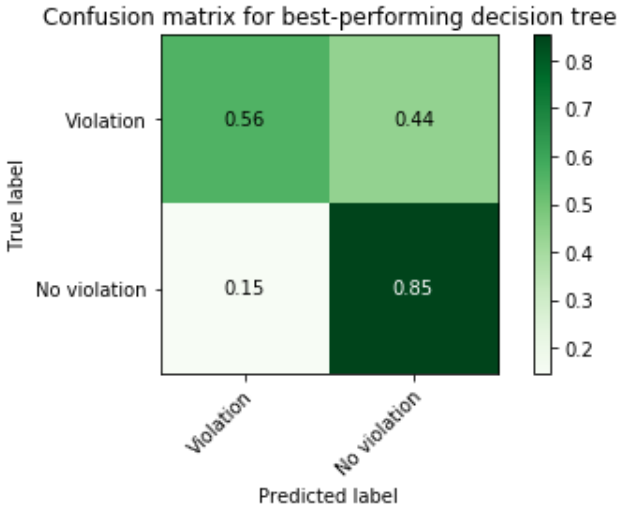
With the potential for further cuts in the EPA's 2020 budget -- perhaps as high as 30% -- we need to consider the possibility of a drastic reduction in resources. This uncertainty has implications on the thresholds we will use for assessing model precision. Our baseline assumption will be that current inspection rates (percent of facilities inspected per year) continue in the future, but we should also optimize for lower levels as well.

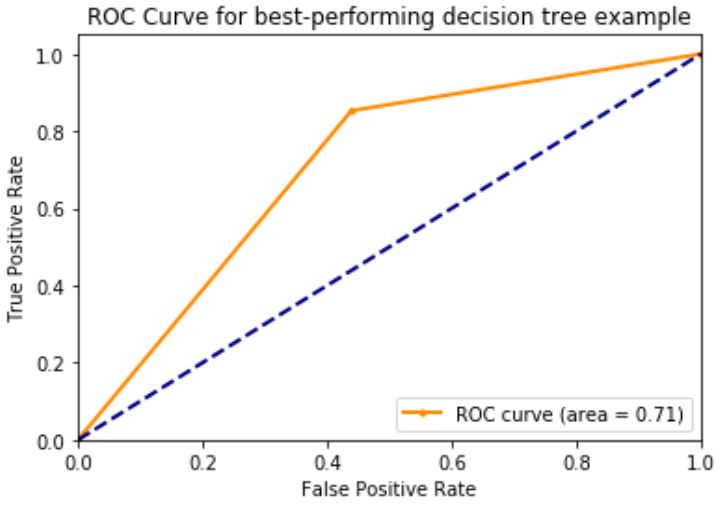
More analysis is needed to determine the role of federal and state agencies in the inspection process. Early analysis shows that state inspections account for half of all activity in a given year. How these inspections are delegated between state and local authorities is yet to be determined, as well as the overall effectiveness of state vs federal inspectors.

The aims of the project have not changed: we will produce a model that identifies likely violators (weighted by those 'high-value' targets based on the size of the facility and potential cost to public health from their non-compliance). One consideration that has been added is the population density near TSDFs. The assumption here is that those facilities closer to high density populations are more costly in terms of public health. Evidence will need to be found to support this thesis; it could be the case that the other metrics are more relevant, like proximity to water sources.

Our preliminary analysis focused on using shallow decision trees to identify important features (i.e., split features that would result in high levels of information gain). For ease of analysis, we generated features from the RCRA_EVALUATIONS dataset - future analyses will involve merging on other datasets in order to generate a broader, more diverse set of features. This preliminary step allowed us to identify qualities of evaluations, including the type and location, that were particularly useful in partitioning evaluations into "found violation" and "no violation" categories.

We then selected the top 20 most heavily-weighted features from our shallow decision tree models and ran them through several decision tree models using an arbitrary train-test date split to get a sense of how a basic model performed in predicting found violation/no violation. For illustrative purposes, we inspected the properties of the model that was "best-performing" under one metric (AUC). This model's confusion matrix and ROC curve are pictured below:





The confusion matrix and ROC curve demonstrate that our simple model has relatively high recall but low precision, which makes sense, as few evaluations result in a found violation. This suggests that our future models should prioritize maximizing F1 score so that we increase precision (thereby decreasing ‘false positives’ that will result in evaluating agencies wasting inspection resources on non-violators) without dramatically reducing the overall number of violating facilities found. We hope to achieve this by generating more detailed features from the other RCRA datasets - our highest priorities for features to add are past violations and industry of the inspected facility. The confusion matrix indicates that our naive models are in particular mistakenly labelling “violation” evaluations as “no violation,” which is a result we’ll remain mindful of as we move forward.