Predicting Remediation Site Duration Using Boosted Tree and Artificial Neural Net Models

Paul Trygstad
December 14, 2021

Introduction

Environmental Remediation can be a costly and time-consuming activity. In any remediation situation, many stakeholders are involved – including regulatory agents, site owners and operators, contracted consultants, and subcontractors. New York State Department of Environmental Conservation (DEC) program analysts would like to be able to predict the expected duration of remediation activities, based on historical data, to assist dedication of state resources and to help prepare site owners for what to expect.

This analysis will attempt at generating a model for predicting the duration of environmental remediation sites using models fitted by boosted tree and artificial neural net methods. Additionally, the entire modeling process will be assessed through cross validation, and the most accurate model produced will be used to fit the entire dataset, as well as make predictions for currently active sites. The data set used is "NYS Environmental Remediation Sites," Version 440, published by the State of New York, last updated 07/15/2021 (https://www.kaggle.com/new-york-state/nys-environmental-remediation-sites/metadata). This dataset contains 40216 rows of 42 variables, recording all remediation related activities for 1009 sites across the state of New York.

Data Preparation

Overview

Prior to modeling, extensive data transformation must be completed. As stated, the dataset used contains an observation for each remediation-related activity/event for each site. In order to predict total duration, information for each site must be summarized into a single observation per site. Additionally, only sites that have been designated as completed or requiring no further action will be used to create the model. Additionally, the model will be used to make predictions for sites that are currently active or preliminary information suggests contamination. There are additional designations for site status, but they are beyond the scope of this analysis.

Missing Data

After the data are filtered for site status, missing values are considered. The majority of the missing data is found in the variables containing the site owner/operator information and disposal site information (Figure 1). As this data presents little advantage to the purpose of this analysis, these variables are omitted from the working data set.

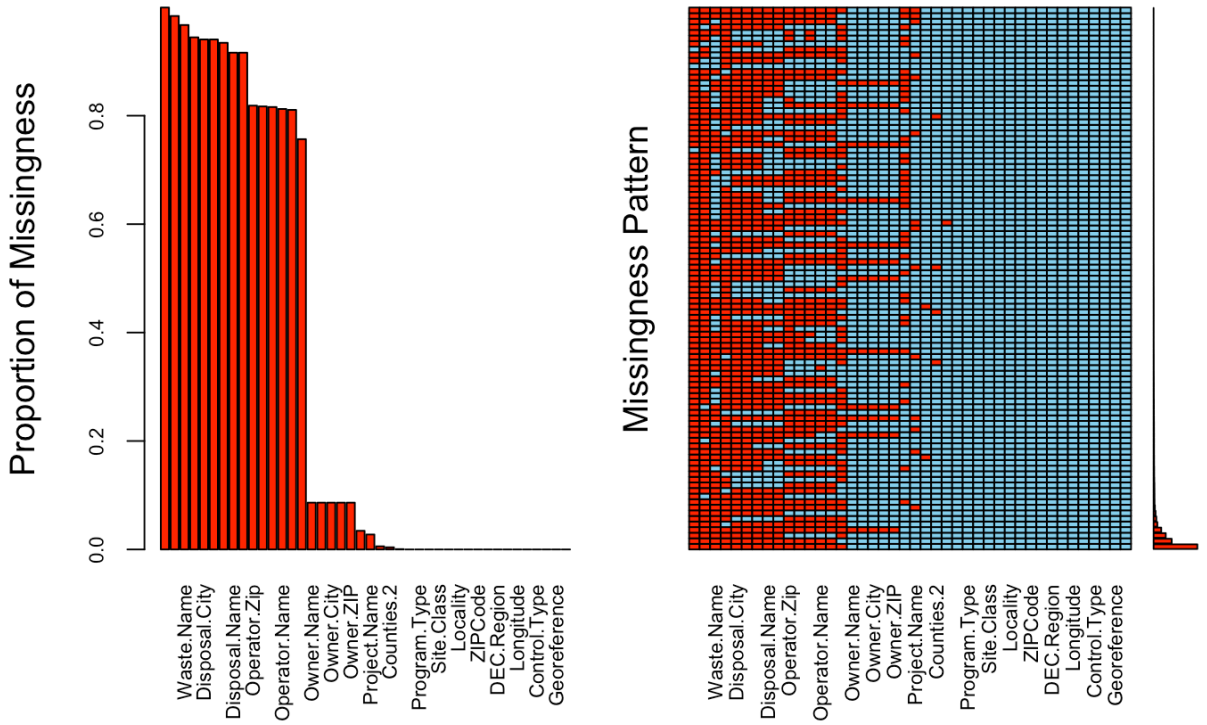

Figure 1: Proportion of Missingness in NYS Remediation Site Data

After removing the variables with highest proportion of missingness, all additional rows containing missing data are omitted from the data set. This results in a data reduction of approximately 1,000 rows. The resulting data set has 16,472 rows and 15 columns., representing 606 sites.

Site Duration

The goal of this analysis is to predict the total duration of remediation activities for a site, given its location, the actions taken at the site, and the site contaminants. To achieve this, each site must be summarized by a single observation containing the relevant information. Total site duration is calculated to begin this process.

Each row of the NYS Remediation data set contains a completion date for the remediation related activity represented by the observation. To get duration of remediation activities per site, the amount of days is calculated between the first and last recorded actions for each site. For

completed sites, this roughly represents the amount of time for the entire remedial life of the site. For active sites, this represents their to-date remedial life.

Contaminants

Each site may contain multiple contaminants. There are 159 unique contaminants contained within the working dataset, and maximum recorded at one site is 34. To record all contaminants per site, each contaminant will be one hot encoded – that is, the presence of the contaminant will be recorded in as a binary variable for each site, so that the presence of all contaminants can be represented. While this significantly expands the data, the advantage of representing each contaminant at a site is clear. To reduce the resulting number of rows, the contaminants are binned into the categories of asbestos, metals, polyaromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), petroleum constituents, organic compounds, and other inorganic chemicals. The resulting distribution of each category is summarized in Table 1.

Contaminant:	Asbestos	Metals	Organic Compounds	Other Chemical	PAHs	PCBs	Petroleum
Number of records	10	3472	4158	1159	4810	466	2397

Table 1: Distribution of contaminants after creating new categories.

Finally, the contaminants are summarized per site. For each site, a binary 0 or 1 is recorded to represent the absence of presence of each contaminant, respectively.

Control Type and Action

Similarly, control type and project phase are one hot encoded and recorded for each site. The control types present in the data are: decision document, deed restriction, environmental easement, environmental notice, and other controls. The project phase types are: site characterization, remedial investigation, remedial investigation amendment, remedial design, remedial action, IRWA, and certificate of completion.

Site Specific Information and Final Dataset

Additional information must be captured for each site, including site ID, program type, and location. This information is captured for all sites from the most recent entry for that site, to preserve the most current information for each site.
The data now represents one site per row, and duration, contaminants, and additional site-specific information is merged into one data frame. Additionally, several undiagnostic or repetitive variables are dropped. Notably, location data is preserved through the DEC Region variable, which records location based on New York State Department of Environmental Conservation

(DEC) regions. The final data frame for analysis has 606 rows and 23 variables - 19 of which are one hot encoded variables.

Variance Inflation Factor

Since most of the variables are factors, numeric correlation cannot be examined graphically. Instead, a temporary linear model is computed and used to calculate the variance inflation factor for each variable. None of the variables exhibit a variation factor raised to a multiple of the degrees of freedom that is over 5, raising no serious multicollinearity concerns.

Dividing the Data

Before model fitting, the data is divided into completed sites and active sites, and the variable of site class is dropped, since the status designation is expected to change throughout the remediation lifecycle of a site.

Hyperparameters

As a final preparatory step, hyperparameters must be considered. For boosted tree modeling, there are several hyperparameters to consider. The first is the interaction depth. For boosted regression, the maximum interaction depth to be considered is equal to the rounded down integer corresponding to the square root of the number of columns in the dataset. However, to keep this analysis computationally feasible, interaction depths of 1, 2, and 3 will be considered by our model.
Next is the number of trees. A value of 200 trees is used, as not to unduly increase variance or computational demand.
Next is the shrinkage coefficient, lambda. Typically, having more and smaller values for results in more accurate models, but it is also more computationally intensive. Here, lambda is stored as a sequence of five values from .01 to 1 in increments of 0.1. These lambda values are reasonable, but not too computationally demanding.
Finally, the n.minobsinnode hyperparameter informs when to stop growing a decision tree. For this hyperparameter, a default value of 10 is used.

For the artificial neural net model, the decay parament lambda is a sequence of eleven values from 0.01 to 1.01 in increments of 0.1. Both one and two hidden layers are considered.

Analysis

The caret package is used to conduct cross validation both in the model selection process and to assess the entire model fitting process. The data set of completed sites is first divided into training and validation sets. Then, 5-fold “inner” cross validation is used to fit the data via both methods over each hyperparameter. For each method, the model with the best fits are stored, and

the models compared. An “outer” layer of 5-fold cross validation is applied to assess the entire model fitting process.

Final Model

Of all the models produced by the 5-fold cross validation, the model with the least amount of error is a boosted tree model with a shrinkage coefficient of 0.11 and interaction depth of 3. This is selected as the final model and fit to the entire data set. The performance of the boosted tree models at different values of lambda is shown in Figure 2.

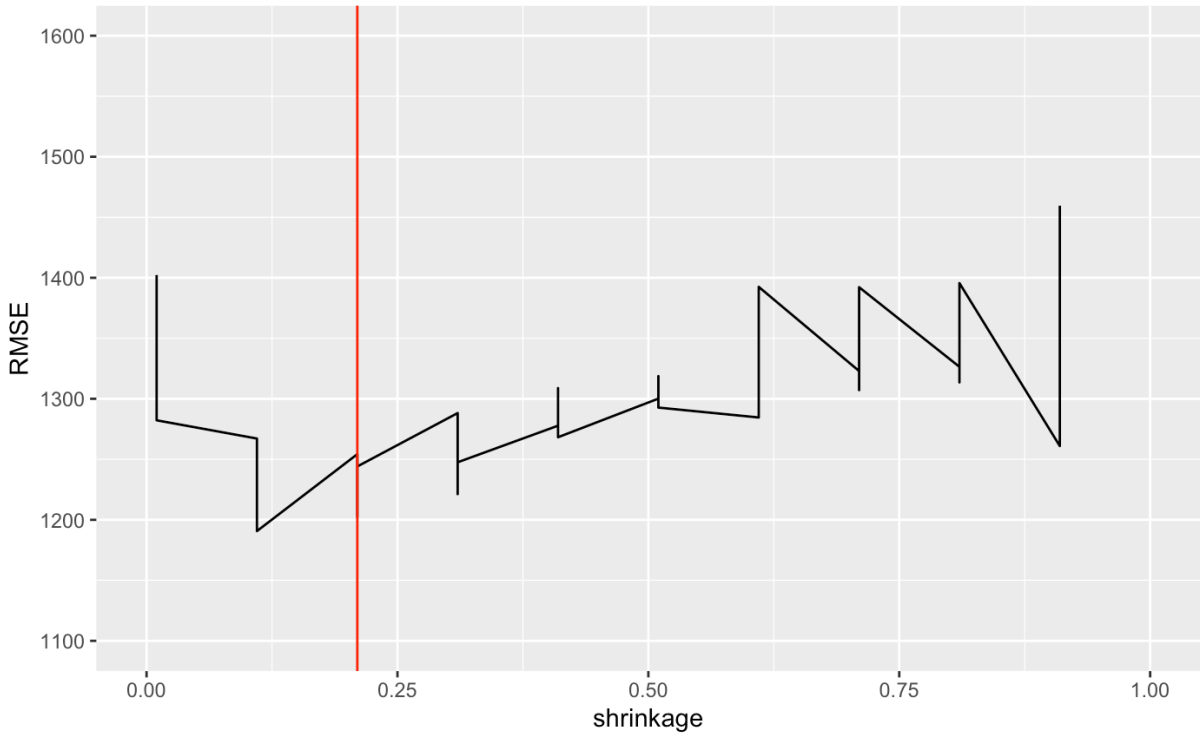

Figure 2: Performance of boosted trees models across lambda values.

The most influential variables on this model are remedial design, site characterization, state program placement, and remedial investigation. This makes sense, as characterization and investigation activities can assist in the completion of remediation activities, but also the process may ultimately lengthen the duration of a remediation activities at a site. The state program the site is put under may determine the amount of resources dedicated to the remediation at that site.

When using data mining techniques to improve management and reduce costs of environmental remediation, Farrell et al. (2007) found that the most important factors were the amount of soil excavated and the number of groundwater monitoring wells installed at a site. While the soil excavated may be more representative of remedial action rather than investigation, monitoring wells are often considered part of the characterization and investigation stages of a remediation

project. In this way Farrell et al. (2007) findings resemble the ones here. However, Farrell et al. (2007) included hydrogeologic, sociopolitical, temporal, and remedial factors that are unavailable to this analysis.

Assessment of the Model Fitting Process

A numerical assessment of the accuracy of the model-fitting process from the 5-fold cross validation can be made by examining the R squared CV measure. This comes out to ~ 0.40, so only ~ 40.0% of the variability total remediation duration can be explained by this model-fitting process.

Predictions for Active Sites

Predictions for the filtered active sites are calculated using the final model. Until the sites are completed, observed values for the duration are unknown. However, comparison to the current duration of the active sites to the predicted overall predictions reveals the inaccuracy of the model, as many of the predicted overall durations are less time than the recorded life of the remediation site.

Suggestions for Future Research

Unfortunately, this attempt did not yield a very accurate model. This is likely a limitation of the data available, and the lack of more in-depth site specific data. To increase the accuracy of this model, more comprehensive, site-specific data could be incorporated, similar to Farrell et al (2007).
Additionally, expert consultation could yield more accurate binning of contaminants.
Finally, data relevant to the stakeholders involved in the remediation efforts may help gauge the effectiveness of such efforts, provided the efficacy of specific owners/operators and consultants can be represented.

References

Farrell, D. M., Minsker, B. S., Tcheng, D., Searsmith, D., Bohn, J., & Beckman, D. (2007). Data mining to improve management and reduce costs of environmental remediation. *Journal of Hydroinformatics*, *9*(2), 107-121. (https://iwaponline.com/jh/article/9/2/107/31099/Data-mining-to-improve-management-and-reduce-costs)