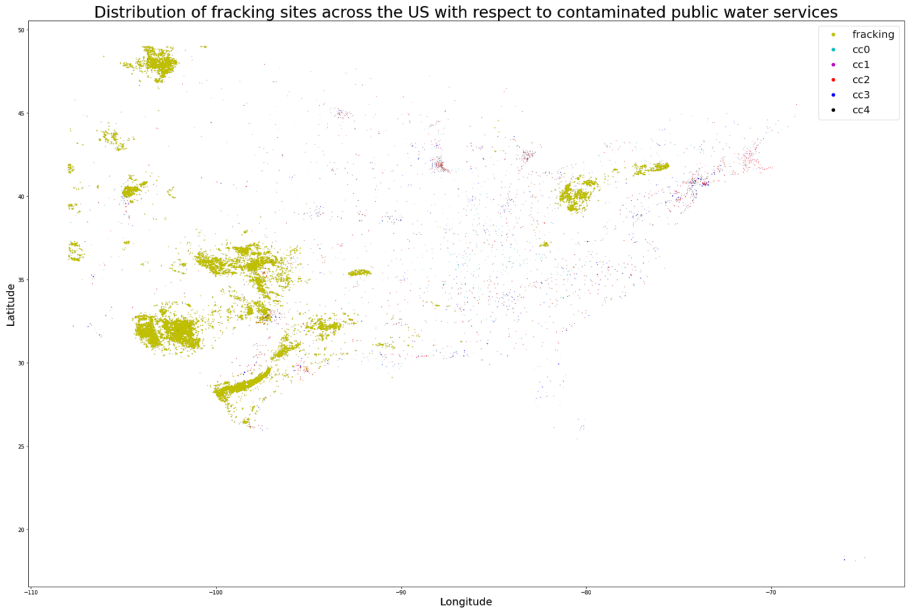
Project Midterm Report

Vrushali Samant
Yash Mundra
Sawyer Huang
Sunday, Oct. 8, 2020

Description of the Dataset

Our data consists of public water service well sampling reports taken from EPA for the years 2014, 2015, and 2016. We saw from data exploration that sampling of any well could continue over the course of several days - as a result, we grouped together samples taken in a time frame of 1 month to extract the count of unique contaminants seen for that particular round of sampling.

We initially graphed the map of the United States, layering fracking sites with a bracketized representation of the contaminant count as seen below with fracking sites in yellow and contamination ranked lowest to highest in categories cc0 to cc4:



We appear to have more contaminants in the North-East despite significantly fewer fracking sites than in Texas. There is not always a strong correlation between fracking sites and contamination simply due to the distribution of fracking sites, which are evidently more densely located in less populous areas like Alaska where public water systems are fewer and far between.

While this gave us some visual insights into the frequency and proximity of fracking with respect to levels of contamination, we still had to consider how to represent our two data sets in combination. Given the sample dates for our public water service samples, we correlated all fracking done with a job end date less than the sample collection date. A single fracking job often consisted of several entries in the FracFocus set due to subtle differences such as the fracking liquids injected, for instance. To handle this, we aggregated this by averaging out real-numbered columns and for ordinal columns, representing them as sets of unique values seen. Last but not

least, we performed a k-nearest neighbors approximation to find the nearest fracking site to the sample site, thus associating each sampling event with an aggregated entry in our fracking data set.

The resulting data set is represented by the parameters in the below:

Fracking

- ➔ The total volume of water used as a carrier fluid for the hydraulic fracturing job
- ➔ Fracking well depth
- ➔ The name of the state where the surface location of the well resides
- ➔ Names of ingredients used in fracturing job
- ➔ Names of the companies that supplied the product for the hydraulic fracturing job
- ➔ Duration of the fracturing process
- ➔ Average percentage of additive ingredients
- ➔ Reason all products were used as a set

PWS Sample

- ➔ Distance of nearest fracking site
- ➔ State to which the public water service well belongs
- ➔ Zip code of the PWS well
- ➔ Total unique contaminants found in the PWS sample
- ➔ Facility water type (groundwater, surface water, etc

Avoiding Overfitting /Underfitting

Our data set consists of around 25,229 rows, which allows for a slightly more complex model than strictly feasible with a smaller data set. Nevertheless, we plan to use regularization to penalize parameters so that our model does not overfit. We can also follow early stopping by setting some maximum number of iterations to after which our model will be considered as overfitting.

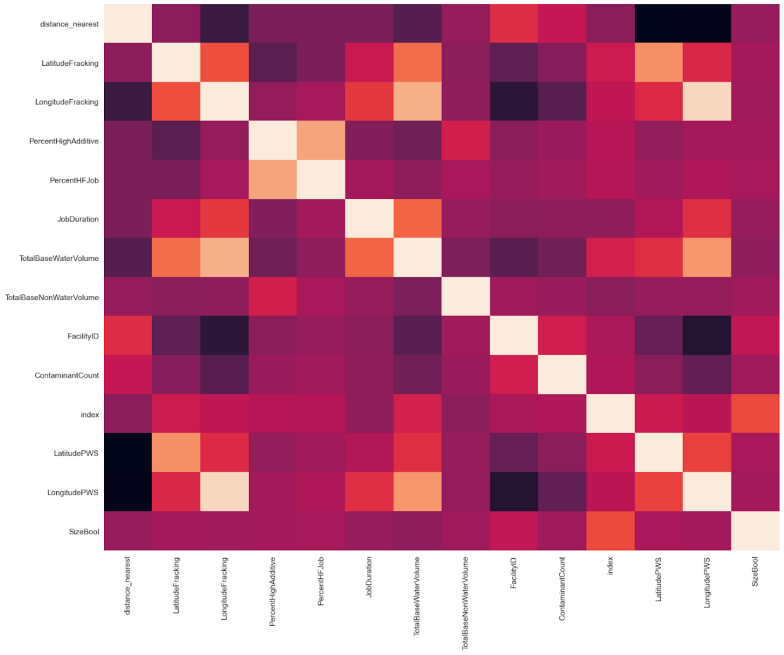
Testing Model Effectiveness

We will perform a 70% to 30% split of our input data set into training and test data sets. Our training set will be leveraged for training of our model and testing will be used for model evaluation. We considered the use of k-fold cross-validation to avoid model susceptibility to general noisiness - however, as our data set itself is large we don't see a particular need to iterate over training/testing splits and will instead leverage different loss functions to generate models impervious to noise.

Missing & Corrupted Data

We did find some corrupted data when looking at latitudes and longitudes. Though both our data sets (EPA & FracFocus) claim to be United States specific, we did find some data points outside the physical latitudinal/longitudinal borders of the continental US that also did not fall in Hawaii. We also found around ~2,000 rows with some form of missing data, taking our total number of rows down from 25,229 to 23,594.

Descriptive Statistics



The heatmap to the left is demonstrative of correlations between various real-valued columns of our data set. If we look at contaminant count, we see its strongest correlations are with FacilityID and the distance of the nearest fracking site, followed then by job duration, average percent of additives, and average volume of water used as a carrier fluid. We expect other (not included) ordinal/nominal columns like state to have relatively high correlation

Preliminary analyses/Feature selection

For our preliminary analysis, we first looked at the features with the highest correlations with contaminantCount and ran some preliminary linear regressions. We experimented with a Least Squared, Ridge and Lasso regression models. In the end they all had an approximate Mean Squared error of ~76.4 and so we kept the Linear Regression model due to its simplicity. Overall, the model had a pretty R square of 0.03, and most of the weights were close to zero. The feature with the highest weight was size followed by distance_nearest.

We also used one-hot encoding to fit the nominal features to a linear regression model. For features such as zip code, the number of unique zip codes were pretty large and so a combined model of it and other features would have been difficult to interpret.

Future Scope

From our analysis, we can see that for a lot of our features, there is little linear relation with contaminant count. So for the rest of the semester, we would look at non-linear models such as decision trees or neural networks for better analysis. We will also look to refine our feature selection and try to reduce the no. of features. This could be done by looking at the features which don't offer any statistical significance.

Appendix

Feature	Coefficient	Feature	Coefficient
JobDuration	4.88349289e-03	**distance_nearest**	1.30783066e-01
PercentHFJob	9.01647706e-02	**TotalBaseNonWaterVolume**	-5.48004344e-08
PercentHighAdditive	-9.91864264e-02	**SizeBool**	1.31312071
TotalBaseWaterVolume	-3.54816615e-08		

Testing with one-hot encoded features

- FacilityWaterType: It had four different categories. The model had a MSE of 74.8 and the maximum coefficient was off abnormally small at -41215573899553.32
- ZipCode: The model had a MSE of 5.56, and the max coefficient was 1426903606502695.8.
- State: The model has a MSE of 70.79, and the max coefficient was -8.101526