Final Project Report

Vrushali Samant
Yash Mundra
Sawyer Huang
Saturday, Dec.12 , 2020

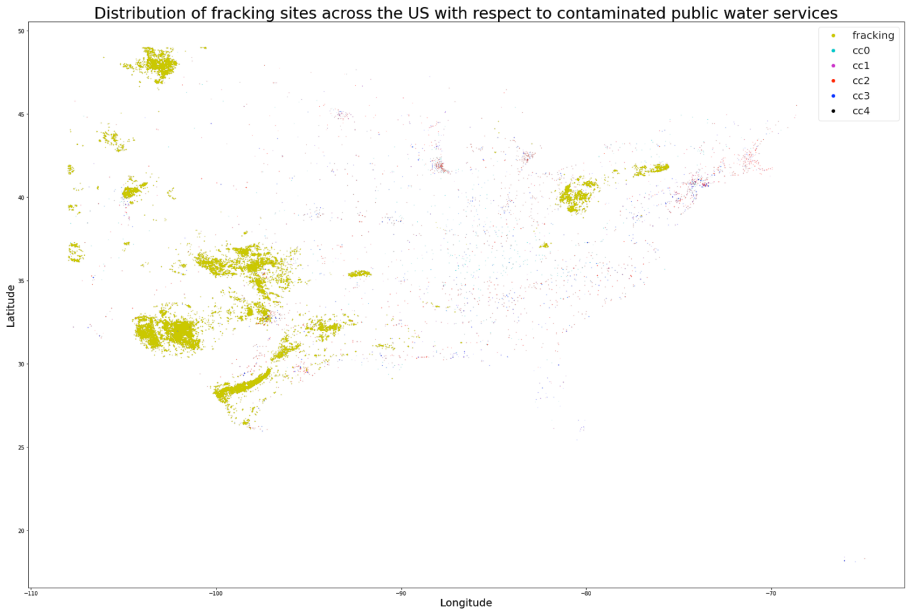
Problem Specification

Fracking, short for hydraulic fracturing, is a process for extracting natural gas by drilling thousands of feet into the ground and injecting a mixture of water and chemicals to break up layers of shale rock, allowing oil and natural gas to escape through the cracks. It has long been a contested topic in the US and is often at the root of various climate change policies that shephard the movement towards clean and natural energy sources. While a significant source of resources, fracking can often contaminate water supplies if it is not done properly, because the fracking fluid injected into rock to enable gas to be released often contains chemicals. In this report, we're aiming to predict water contamination of a public water source, given what we know about the fracking site.

Data Description

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 77% to 33% 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.

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.

Model and Analysis

1. Preliminary Linear Models

The first models we tried in our midterm report were linear regression models. We experimented with 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.

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

For our linear models, we see that a lot of our features had little correlation with the contaminant count. This likely points to a lack of a linear relationship between the features and the outcome.

2. Feedforward Neural Network

In order to reduce the scope of our outcome variable, we decided to change the task from a regression one into a classification one. We transformed the prediction variable contaminantCount into 1-6 labelled buckets based

on the raw cantimant counts. While this reduces the precision of our models in detecting the amount of contamination a water source has, we felt this was sufficient since the aim of our model was to predict whether fracking leads to increased water contamination.

The dataset was also transformed so that nominal values such as Facility Water Type, and State were one hot encoded. We decided to not use the IngredientName, Supplier and Purpose features since encoding them would have added unneeded complexity to our analysis. We also decided not to use the ZipCode feature, since one hot encoding would have created a large sparse data vector which would have likely decreased the effectiveness of the FFNN. In the end we had a data vector of length 72.

For the feedforward neural network, we used a hidden layer of size 120 with Cross Entropy loss and Relu activation. We also used a softmax after the output layer. The model was trained using Stochastic Gradient Descent with a learning rate of 0.01 and 0.9 momentum.

The model achieved a **maximum accuracy of ~54%** after training for about 30 epochs.

3. Decision Tree Classifier

The second non-linear model we explored was a decision tree. Decision Tree is a class of models which uses a tree like structure to predict the target variables by learning simple decision rules inferred from the dataset. For our decision tree, we utilized a model with a max depth of 1000, a minimum sample split of 2 and maximum leaf nodes at 2900. After we trained our model on a training set consisting of ⅔ of the dataset used for our FFNN, we used it to predict the outcomes of our remaining validation split and achieved an **accuracy of ~73%.** While the accuracy was a substantial increase from our neural network, we noticed that the training accuracy was significantly higher at 95%. This seems to indicate that our model likely overfitted,

4. Random Forest Classifier

In order to combat the overfitting we encountered with our decision tree, we decided to experiment with a Random Forest Classifier. A Random Forest Classifier utilizes a multitude of decision trees, where each tree is assigned a subset of features. The model then predicts the output variable using a majority vote from the various decision trees. Since Random Forest utilized a magnitude of decision trees, it helps prevent overfitting which is a problem we faced with our earlier decision tree. After training our model on the training set and testing it on the validation set, we achieved a validation **accuracy of ~77%** while our training accuracy reduced to ~93%.

5. KNN Classifier

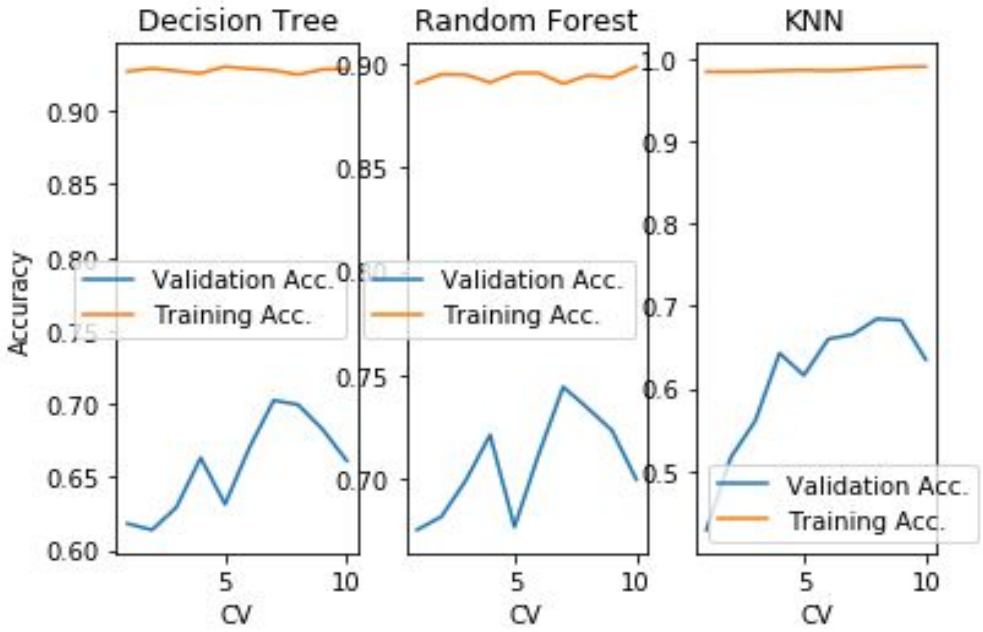
The last classification model we explored was a KNN Classifier. We thought that this model may perform well with our data because similar fracking operations will most likely lead to similar contaminants and levels of contaminants being released into the local environment. We wanted to know the ideal number of neighbors to use in our model, so we graphed training and test data accuracies from 10 different models using 1 to 10 neighbors each. Once we found that the ideal number of neighbors was 1, we trained our model on the training

set and tested it on the validation set. We achieved a validation **accuracy of ~68%** while our training accuracy was ~99%.

Further Analysis

Cross Validation

In order to test the robustness of our models, we decided to employ k-fold cross validation in order to see how our accuracies varied across each fold. In K-fold cross validation, the dataset is divided into k subsets and in each interaction, a subset is held out from training and used as the validation set. This gives us a more accuracy estimate of how our models would perform in the real world. Since the Decision Tree, Random Forest Classifier, and KNN Classifier algorithms performed the best classification, we decided to test the three models. We used a k value of 10.



Above, we can see that the Random Forest classifier has consistently higher accuracy, as well as lower overfitting. The maximum accuracy was around ~75 % while the minimum was ~ 70%. Overall, the model performed with an average accuracy between 70-75%, which is within a suitable margin.

Ablation Study

While our model performs reasonably well, we have little insight into what features contribute the most. In order to study this, we performed an ablation study where we removed one feature at a time in order to study its importance. A summary of the results are below

Feature Removed	**Accuracy**	**Feature Removed**	**Accuracy**
State	**72 %**	LatitudeFracking	47%
Water Type	48%	LongitudeFracking	47%
Latitude	55%	PercentHighAdditive	47%
Longitude	54%	PercentHFJob	47%
IsContaminated	47%	JobDuration	47%
Size	47%	TotalBaseWaterVolume	48%
TotalBaseNonWaterVolume	49%	distance_20_nearest	52%
distance_nearest	46%		

We can see that for most of the features, the accuracy significantly decreases. However, when we removed the states from the dataset, we saw a less significant drop in accuracy to 72%. This is indicative of the fact that perhaps the results are indifferent to which states the fracking sites are in. The fact that when the other features are ablated, the accuracy decreases is indicative of the fact that they are significant to the model.

Conclusion and Limitations

The results from the ablation study on our model indicate primarily that the majority of our features were good indicators for the level of contamination at any given public water source. This is beneficial as it indicates that our model is indifferent to which state it is located in. As we expected most of the features had a significant impact on the model--in other words, there is indication of a strong correlation between the location of the fracking sites public water sources and that it is possible to predict possible contamination in a public water source, given what we know about nearby fracking sites. If we were to further explore the predictability of water contamination, we might pull in additional data sources to correlate other known contributors to contamination levels: locations of landfills, farms, and sewage systems.

One major limitation that likely had a hand in our results is that there is no natural one-to-one mapping between our fracking was public water source data sets. The data set we developed our model on top of, then, was one we generated based on our observations of both input data sets--which makes it more susceptible to error.