Brooke Nichols

Fracking Effects: a study on seismic activity and hydraulic fracturing wells in the Permian Basin

During my time working at an oil and gas company I learned the basics of hydraulic fracturing, which is an oil and gas well development process that involves injecting water, sand, and chemicals into rock via the well to create new fractures to increase oil production, that has risen in popularity in the past few decades. What has also seen an increase in the last decade is seismic activity in the Permian Basin. This influenced my capstone choice to focus on the possible relationship between well fracking, during the time the well is being fracked, and the increase in seismic activity felt by Permian Basin residents in the last few years. If so, is there a way to use the features of fracking to predict the magnitude of this seismic activity.

The oil and gas industry collects a lot of data. Where to drill, how to do it efficiently, how to get the most oil out of a well, or what to put down the well in order to increase production? With each of these questions there is a large amount of data to help the industry optimize their choices. With the use of AI and ML the decision making can be improved, the wells production can be optimized and hopefully the impact to the environment can be limited. Both those in the industry and those outside of the industry have been using AI and ML to improve the processes used and to track the environmental effects.

The data I gathered for my look into the relationship between fracking and seismic activity came from Frac Focus, IRIS, Tex Net, and USGS. Frac Focus is the national hydraulic fracturing chemical disclosure registry for the United States, has data on frac wells going back to 2001, includes information on how, when, where, and who fracked a well, and who supplied the chemicals. I pulled their entire dataset, which I limited to the Permian Basin because Texas has the largest number of fracking wells in the US and the Permian Basin is the heart of the oilfield in Texas. The frac data was read in as a csv, starting with (4734056, 39) and after cleaning ended with (569472, 15). My seismic data came from three separate places to combine into the seismic_master. The Incorporated Research Institutions for Seismology is a consortium of US universities that acquire, manage, and distribute seismological data and is a nonprofit organization. The data from IRIS was a txt that started with (1731659, 13) and ended with (547, 5). Tex Net was started in 2017 after it was established by the legislature, run by the Bureau of Economic Geology to help locate and determine origins of earthquakes in Texas and if that cause is human activity they seek to prevent it from occurring in the future. The Tex Net csv had (9175, 30), which got down to (213, 5). The United States Geological Survey tracks recent and historic earthquakes using their earthquake hazards page. The USGS was the smallest csv with only (421, 22) and it ended with (338, 5). These data sources were used to pull information on earthquakes in Texas over a 2.5 magnitude as this is the threshold at which a human can start to feel the activity and the hydraulic fracturing process naturally causes activity under a 1. The final seismic_master combination ended with (953, 5).

In order to get the frac data into shape I had to start by dropping the unnecessary columns then limit my area to Texas before limiting to the Permian Basin using latitude and longitude. I had to limit to Texas first in order to not pick up any New Mexico wells with the latitude and longitude limiting. Then after eliminating the nulls and duplicates I renamed the lat and lon columns for the eventual master combination and split the date column. Finally, the biggest thing I had to do to this data was attempt to uniform the ingredient names since this was a free form filled out by a number of people. This is where I lost the most number of rows, but made it easier to work with on my computer. For the seismic_master I went the same steps with all three data sets to make sure they would combine correctly. First I limited the data to Texas and the Permian Basin then made sure the magnitude was greater than a 2.5. After that I dropped irrelevant columns, checked and cleared null values or duplicates then changed the names of columns for combination. After that I split the date column into date and time then reordered the columns. Once I had the frac_master and seismic_master I did a master combination giving me (430500, 20). First I made sure that the seismic activity occurred within one degree of the latitude or longitude of the location of the well. Then I used the date columns to make sure the activity occurred while the well was being fracked. Then after dropping any possible duplicates that have cropped up I binarized the federal well column to make it numbers instead of a boolean.

After the master combination a new notebook was started and the modeling data set was limited to the water based and non water based carrier fluid in gallons, percentage of the ingredients in the non water fluid in percent by mass, the mass of the ingredient in pounds, depth of the well, and federal well status. I started with a simple look at the correlations testing on unscaled data comparing the features against the target; however, there were not any positive relationships revealed. Then I moved on stats linear regression modeling after scaling the data in order to look at p-values for my chosen features to assess their possible predictive power. After running the stats model all of the p-values were under the significance threshold of .05 except the federal well status indicating they would have some predictive power on the target variable; however, the R^2 value for the model was only ~1% meaning the predictive power for these variables on this target are negligible. I decided to test each feature against the target individually to see if one feature specifically was carrying the predictive power. This was the case as the only feature to continue to hold the ~1% R^2 score was the total base water volume, or the total gallons of water pumped down the well in the carrier fluid during the frac job. To see if I got similar results I moved into using scikit learn predictive regression modeling. I have a continuous target and the R^2 score for the linear model was not high, so I chose to look at non-linear regression models . After train/test splitting my data and one hot encoding the ingredient name feature I created a grid search to go through two models, Ridge Regression and Decision Tree Regressor to determine the optimal model for this data. After running the grid the best fit for the data would be the Decision Tree. It improved the R^2 score to 12% on the test set, so it explains the y-variance a little better than the linear stats model.

There is still quite a lot of seismic activity that cannot be explained by these features after regression modeling. Using the current dataset and features I can not say they affect the seismic activity that has been occurring in the Permian Basin. I would like to be able to run the entire US dataset and expand the features to see how it affects the target variable. I would also like to expand dates to include a period of time outside of the fracking period to see if fracking somehow affects seismic activity after the frack is done. If the features can be optimized to see how fracking affects seismic activity then the process can be further refined to limit those effects as well improve well production.