**Predicting Induced Seismicity in the Eagle Ford Shale Play**

Despite several decades of oil and gas development in the South Texas Eagle Ford Shale Trend, there has been a relatively recent increase and severity of induced earthquakes since 2018 (Fasola et al 2019). For my Data Science Immersive capstone project, I created two models to forecast and better understand the drivers and mechanisms of this seismicity in Karnes County, TX: a SARIMAX model and Recurrent Neural Network model.

For each model, I experimented with various target representations of the seismicity along with exogenous features that represented both oil and gas activity and operator completion (frac) parameters. In testing parameters for each model, the best target to model for was Daily Average Magnitude of seismicity, and the feature that led to the highest correlation with the best model was the Average Volume of Water Injected per Lateral Horizontal Foot. In the image here, I show a possible link between the amount of water injected per foot of completed lateral and the Daily Average Magnitude of earthquakes in Karnes. While these results are preliminary, the relationship is supported by findings published in Fasola et al 2019.

**AI Radiologists: Classification of COVID-19, Viral Pneumonia, and Normal Presenting Chest X-Rays**

At the time of this project, the world at large was in the midst of the one of the largest healthcare crises since the 1918 Spanish flu. Hospitals all over the United States were undergoing extreme stress due to severe COVID-19 cases, and ensuring patients receive rapid and accurate COVID testing was paramount. Our project team strove to discover if there was a fast and reliable way to identify COVID-19 infections in hospitalized patients using chest x-rays and deep learning techniques.

To solve this problem, we focused on the classification of 3 classes of chest radiography: COVID-19 Pneumonia, Viral Pneumonia, and Normal. Harnessing the predictive power of a convolutional neural network (CNN), we produced a multi-class classification model to help radiologists diagnose if a patient is presenting with COVID-19 pneumonia, viral pneumonia, or normal status. Our final model scored approximately 93.5% on both accuracy and recall, and was deemed an overall success in that it successfully differentiated between the three classes.

**Reddit Post Classification Using Natural Language Processing**

There is overlap between the Game of Thrones fan base and the Lord of the Rings fan base, and sometimes their posts get mixed up on their respective subreddit pages. The goal is to build a classifier that can determine which subreddit a post is from given the text of each post. I used Pushshift API to scrape over 10,000 posts between each subreddit page, and cleaned the posts for missing values, emojis, and special characters in order to refine my model.

I compared both the Tfidf Vectorizer and the CountVectorizer in both a Multinomial Naïve Bayes Classifier and Random Forest Classifier to understand performance, along with utilization of stop word removal. The best performing model was the Random Forest Classifier with a Recall score of 91% and test accuracy of 93% and was successfully able to discern between posts of each subreddit page.

**Experience Section**

General Assembly: 13 week immersive course applying data collection, cleaning, exploratory analysis, modelling, data visualization, and various machine learning tools to solve real-world data science problems. Created a Naïve Bayes and Random Forest Classifier to discern between 10,000+ Reddit posts using Natural Language Processing techniques. Trained a Convolutional Neural Network to assist in diagnosing COVID-19 pneumonia presentation in chest x-rays. Capstone project utilized forecasting methods to better understand mechanism driving induced earthquakes in the Eagle Ford Shale Oil and Gas Play.