Machine Learning Challenge – Using AI to Validate Carbon Containment in the Illinois Basin

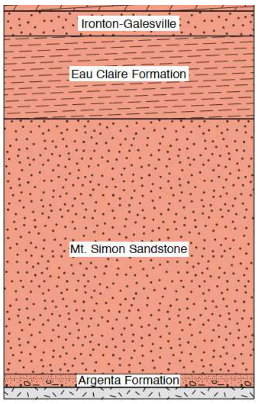
BASEM BARBARY

AGENDA

This challenge aims to use time series injection information and monitoring data on a carbon capture well to predict carbon capture well injection rates deltas.

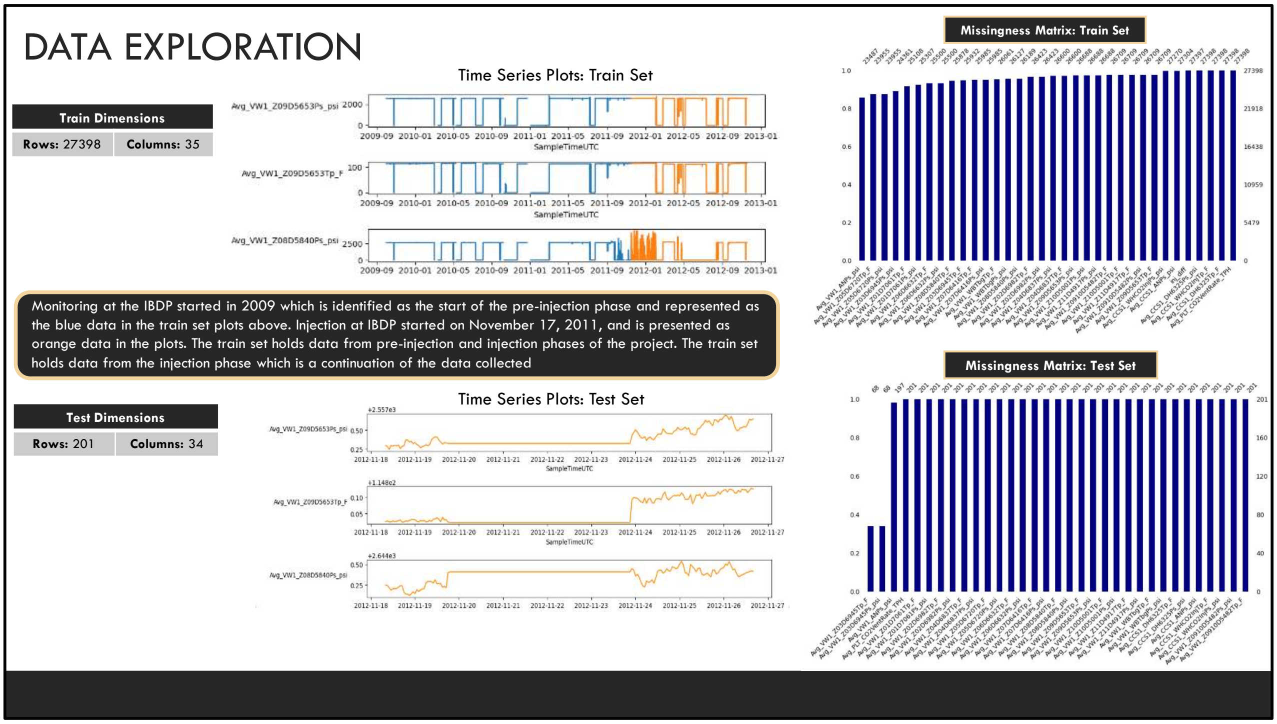
- Data Exploration
- Data Preprocessing
- Data Visualization
- Data Modeling
- Results

ILLINOIS BASIN DECATEUR PROJECT DATA

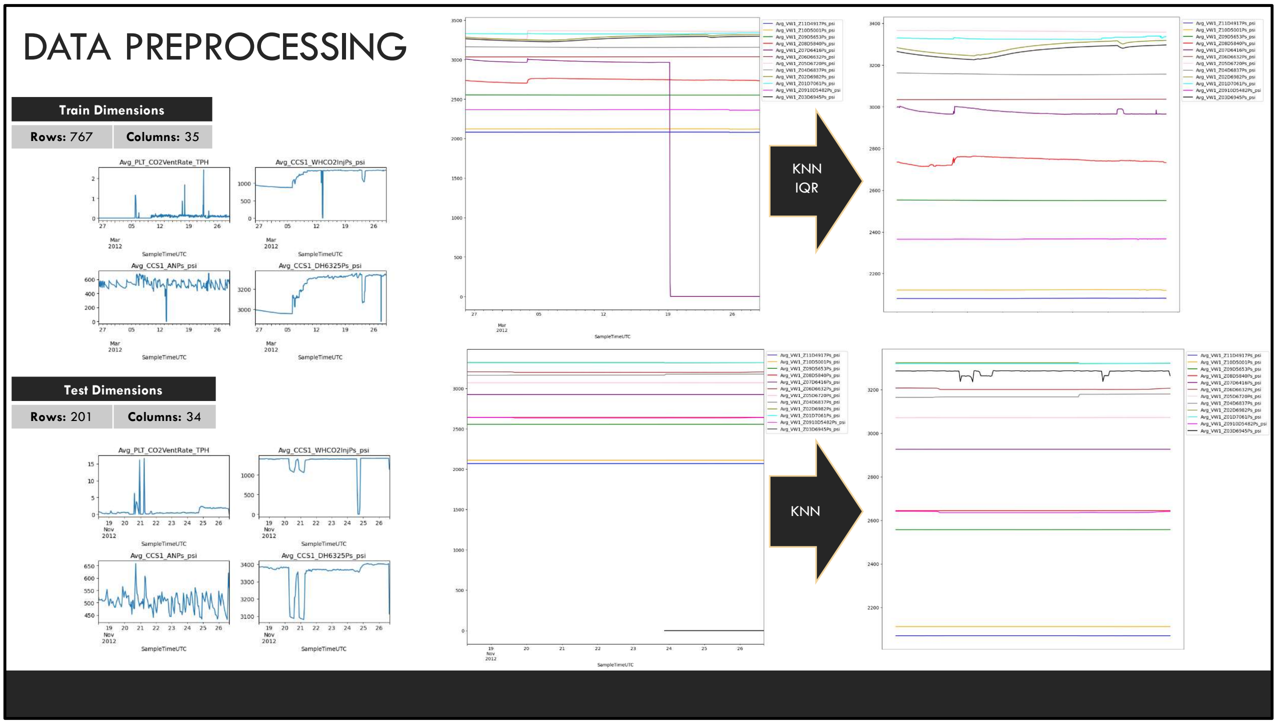


The data set holds injection data (rate, pressure, temperature) from the injector and the corresponding fiber optic DTS temperature profile from the observation well. The observation well pressure and temperature data is identified where Zone 01- Zone 09 are in Mount Simon Sandstone and Zone 0910 is just above the Mt. Simon Sandstone. Zone 10 and Zone 11 are above the Eau Claire Shale (primary seal).

	Depth (ft)
ZONE 11	4917
ZONE 10	5001
ZONE 0910	5482
ZONE 09	5653
ZONE 08	5840
ZONE 07	6416
ZONE 06	6631.7
ZONE 05	6720
ZONE 04	6837
ZONE 03	6945
ZONE 02	6982.4
ZONE 01	7061



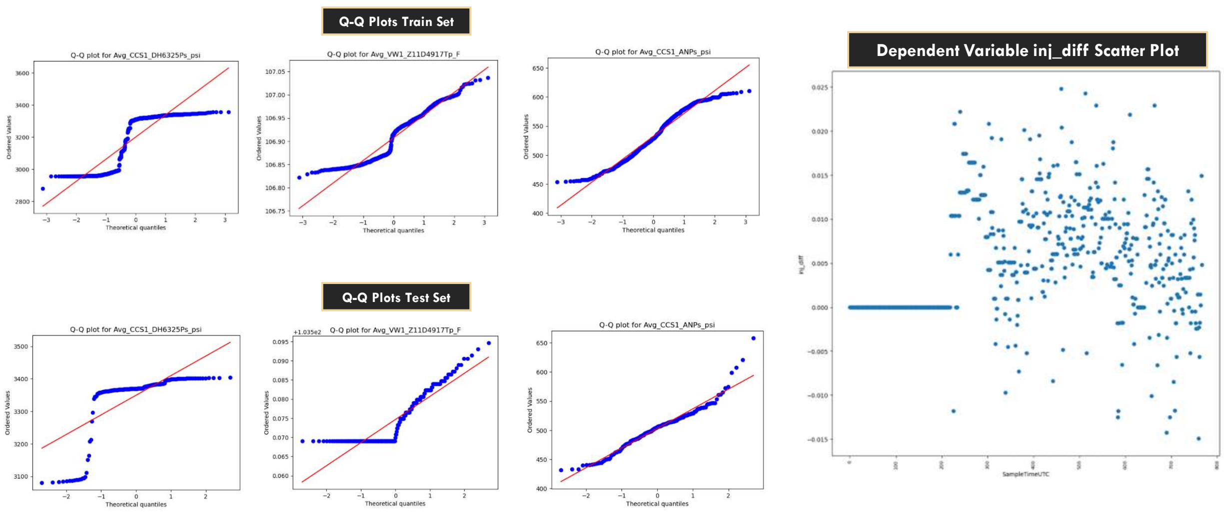
The data set at hand includes timestamps, therefore, indexing 'SampleTimeUTC' variable is necessary. 35 variables in the train and 34 variables in the test set are plotted vs. the indexed timestamps to visualize the data points. Upon observing the visualizations, it is clear that the noise observed in the zone 8 pressure variable after November 17, 2011(1) , is due to injection through the wellbore. An important observation is that there are are a few instances where pressure drops to 0 in the injection well annulus and injection well downhole during injection as well as across all zone pressures. Missing values and 0 appearances are present as seen in the visualizations. Investigating the missing values in each column beyond just the count in the dataset is important. In doing so, trends can be identified which will aid in better understanding the data set in its entirety. In the train set, 'Avg_VW1_ANPs_psi', 'Avg_VW1_Z03D6945Ps_psi', 'Avg_VW1_Z05D6720Ps_psi' and 'Avg_VW1_Z05D6720Tp_F' are missing more than 11% of their values which need to be handled accordingly. Upon further investigating the data points, it is apparent that these values are missing both at random due to 0 values continuously present before or after missing values, and systematically. It can be concluded that this is due to measurement equipment sensitivity. In the test set, the variables 'Avg_VW1_Z03D6945Ps_psi' and 'Avg_VW1_Z03D6945Tp_F' are missing 67% of values. There are several outliers present in the train set found using statistical analysis. Investigating where 0 values appear and determining if they are a worthy appearance to include in the data is necessary prior to removing outliers is important. Upon counting the 0 appearances in the train set, every variable has many counts of 0. The only variables in which 0 can appear as a data point are ‘Avg_PLT_CO2VentRate_TPH’, ‘Avg_VW1_ANPs_psi’and ‘inj_diff’. Specifically, any 0 values observed for pressures and temperatures in each zone will need to be handled. In the test set, 0 values appear in ‘Avg_VW1_Z03D6945Ps_psi’ and ‘Avg_VW1_Z03D6945Tp_F’. Filtering a portion of the train dataset will reduce the presence of outliers, 0s, and noisy data in the train set at hand and eliminate the need for aggressive outlier removal, which could potentially risk the integrity of the data set.



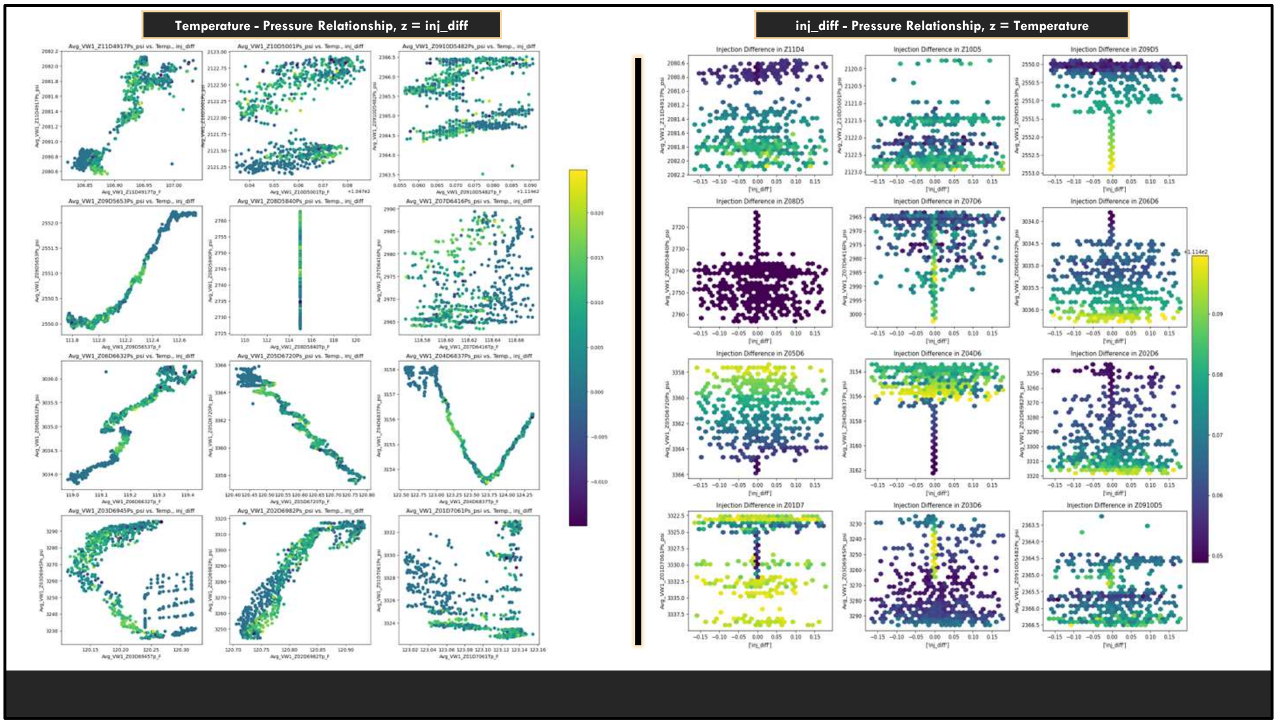
There are outliers present in the train set, observed from the statistical analysis Within the test set, the distribution is also skewed by extreme outliers. Initially, the train set is reduced in size by using a filter to exclude data prior to first day of CO2 injection into the Illinois Basin . The first 3 months of injection data will also be excluded due to the presence of noisy data in Zone 8. After experimentation, a filter on the train set to include observations between ‘2012-02-26’ and ‘2012-03-29’ best resembled the distribution and trends portrayed in the test set. It is important to maintain the integrity of the train set through out the preprocessing techniques. The train set holds 767 rows of data in the extracted time series data. Reducing the outliers in the train set with this filter will minimize the risk of introducing bias when imputing missing values. As mentioned previously, handling the presence of 0 values is important. Imputation will be used to handle missing values and inexcusable 0 appearances in the train and test set. Outlier detection methods would consider these values outliers for which handling these 0 appearances with a more aggressive approach is necessary to avoid introducing bias. In the train set, the zone pressure and temperature variables with 0 appearance is changed to NA. In the train set, the 0 appearances found in zone 3 pressure and temperature variables are also changed to NA. KNN Imputer is used to replace the missing values using the mean values of the five nearest neighbors with estimations.

The imputer is fit on the train dataset and then used to impute the missing values in both the train and test set. The univariate distribution is visualized for the train and test set. Outliers were observed in several variables which can affect the accuracy of the model built if not handled accordingly. IQR outlier detection method is used to handle these outliers using the first and third quartiles where the values outside the bounds imposed are detected and replaced with missing values. KNN imputer is then used to fit the updated train set resulting in a cleaned dataset, most importantly, without removing rows of data from the time series. It is important to note that the train and test set needs to be normalized prior to modeling to give equal weight between the features.

DATA VISUALIZATION



Q-Q plots, along with Histograms, were used to identify trends in the train set for each variable and then compared to the test set. The residuals for many of the variables are not normally distributed. Between the train set and test set, the trends in comparison are alike in behavior further verifying that the train set is resemblant of the test data set and can be used for predictions. This does not fully confirm that there is no evidence of overfitting or underfitting and further validation will be needed. It is determined however that training the data using a model that will handle and account for non-linear relationships. Visualizing ‘inj_diff’ highlights an important notion to understand within the segment of this data where there are consecutive 0 appearances in ‘inj_diff’.

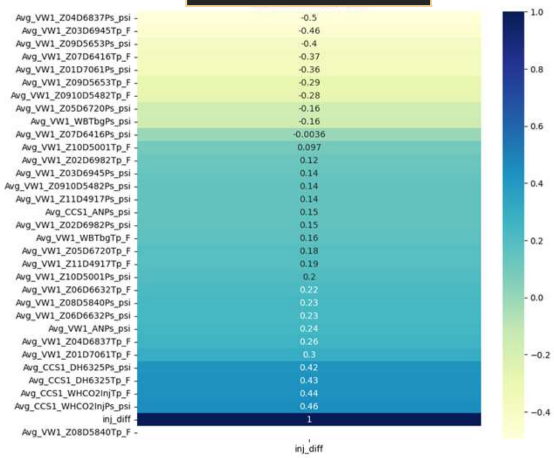


The collection of plots on the left are used to identify trends and relationships between the pressure and temperature variables each zone with a color gradient to help visualize the ‘inj_diff’. There is clustering of points observed through the teal and green colors is obvious. The general direction and distribution of each zone’s pressure and temperature trend are important to visualize and verify against that of the test set. This ensures that the train set data will more accurately fit the model and produce predictions on the test set. The collection of plots to the right identifies how temperature in each zone impacts ‘inj_diff’ in relation to pressure in each zone. Since the train set includes observations between ‘2012-02-26’ and ‘2012-03-29’, it is important to note that these trends are after the first injection period and the CO2 migrating through the reservoir.

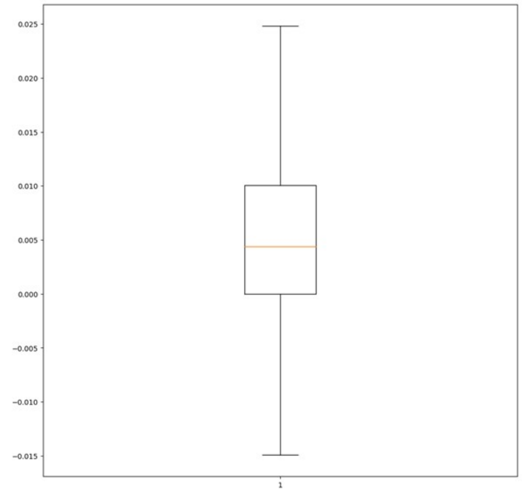
DATA MODELING

- Scale by Normalizing
- Machine Learning Model: Neural Network
- Validation Set using Train-Test Split

Test Set Correlation Matrix

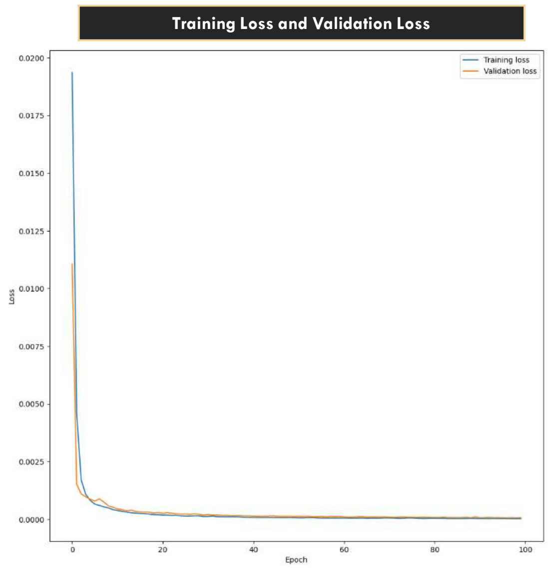


Dependent Variable inj_diff Box Plot



In the train set, there are variables that have negative correlation to the dependent variable. It is important to note that 'Avg_VW1_Z08D5840Tp_F' does not show any linear correlation. The objective of this workflow is to build a regression model to predict the continuous variable, 'inj_diff'. Neural Network modeling is selected to be used to . The train set and test set are normalized in order to scale the different magnitudes present. The neural network architecture is built using 64 neurons, 34 neurons and 1 neuron for the first, second and last layers, respectively. Rectified Linear Unit is used as the activation for the first two layers. The model is trained with a batch size of 34 and 100 epochs with 20% of the data used for validation.

RESULTS



inj_diff Predictions

	inj_diff
count	201.000000
mean	0.020770
std	0.057702
min	-0.112286
25%	-0.014073
50%	0.028427
75%	0.039984
max	0.161508

Training loss and validation loss are visualized to investigate if the model is overfitting or underfitting the data. With the parameters used to train the model, there is no evidence of overfitting or underfitting over the epochs.

Consider:

- Handling missing values and outliers for pre-injection and injection separately.
- Filter to include more injection data into train set.
- Filter to include cleaned pre-injection data.
- Including geological data and reservoir properties such as porosity and saturation of reservoir before and during injection into model.
- Redefining the goal with the data set and approaching this problem as a classification problem where the target variable, inj_diff, is split using specific thresholds to classify injection difference. This can be an opportunity to highlight the trends that will lead to migration occurring and insight against migration.

RESOURCES

(1) "Illinois Basin – Decatur Project (IBDP)." *Netl.doe.gov*, https://netl.doe.gov/coal/carbon-storage/atlas/mgsc/phase-III/ibdp#:~:text=Simon%20Sandstone%2C%20in%20Decatur%2C%20Illinois,1%2C000%20metric%20tons%20per%20day.

(2) "Illinois Basin–Decatur Project (Chapter 19) - Geophysics and Geosequestration." *Cambridge Core*, Cambridge University Press, https://www.cambridge.org/core/books/geophysics-and-geosequestration/illinois-basindecatur-project/Cn.d.81369208B12D4BF6F09CCAAFF3F6.