
Air Liquide
ML Challenge - CO2 containment prediction at Illinois Basin Decatur Project (IBDP)
Ayman Mhamdi
Luis Pinto
Amina Talipova
THIS DOCUMENT IS PUBLIC
©Getty

Agenda

1. Challenge project overview
2. Data
3. Machine learning solution
4. Results
5. Discussion & future steps
6. Conclusion

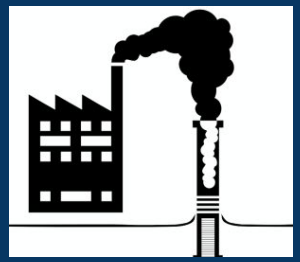
1. Project overview

Two wells:

1. CO2 Injection well;
2. CO2 Verification well:

CO2 is injected deep underground in the Mount Simon Sandstone formation beneath Decatur, Illinois, a Deep Saline reservoir

1

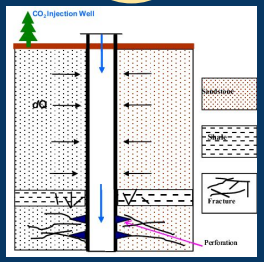


Captured from the ethanol production facility at Archer Daniels Midland Company.

2

Dehydrated & compressed to 1,400 psi for injection into a well (7,000 ft deep)

3



Dehydrated & compressed to 1,400 psi for injection into a well (7,000 ft deep at 5-7 thousand psi)

1. Project overview

Challenge goal

By using metadata, predict the CO2 injection rate delta (first diff)

Storage capacity

1M tCO2 store target

The IBDP is one of the first projects of similar scope in the United States to complete the goal of a 1-million-ton injection of CO2 for geological sequestration

Analyzed period

2009-2013 injection period

Train data set includes ~27K observations and 33 features. Injection period: CO2 was injected at a rate of ~1000 tons per day from November 2011 until November 2014.

Terms

- **7025-7050 ft injection depth;**
- **~1000 tpd injection rate**

The injection pressure rate was kept well below the fracture propagation pressure, as required by the regulatory framework for Underground Injection Control and determined by a step-rate test.

2. Data

2.1 Data analysis:

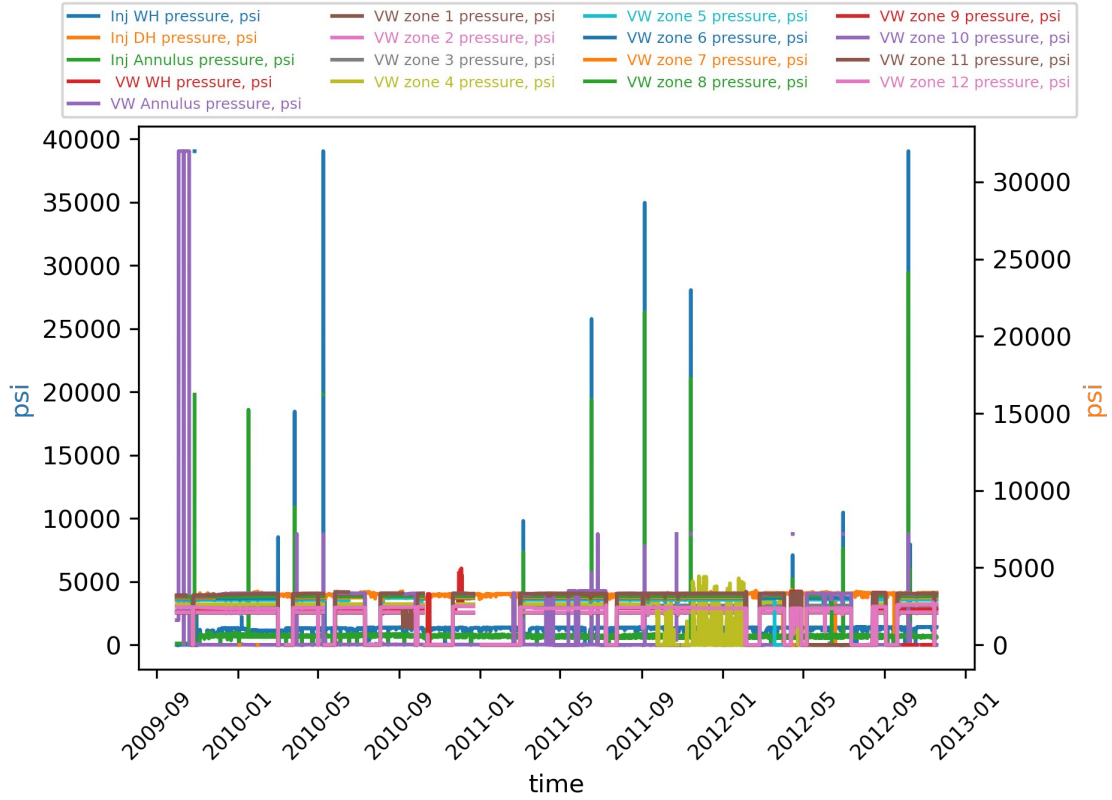
- Initial analysis, visualization;
- Understanding the datasets;

2.2 Data pre-processing:

- Missing data analysis and imputation;

2.3 Feature engineering

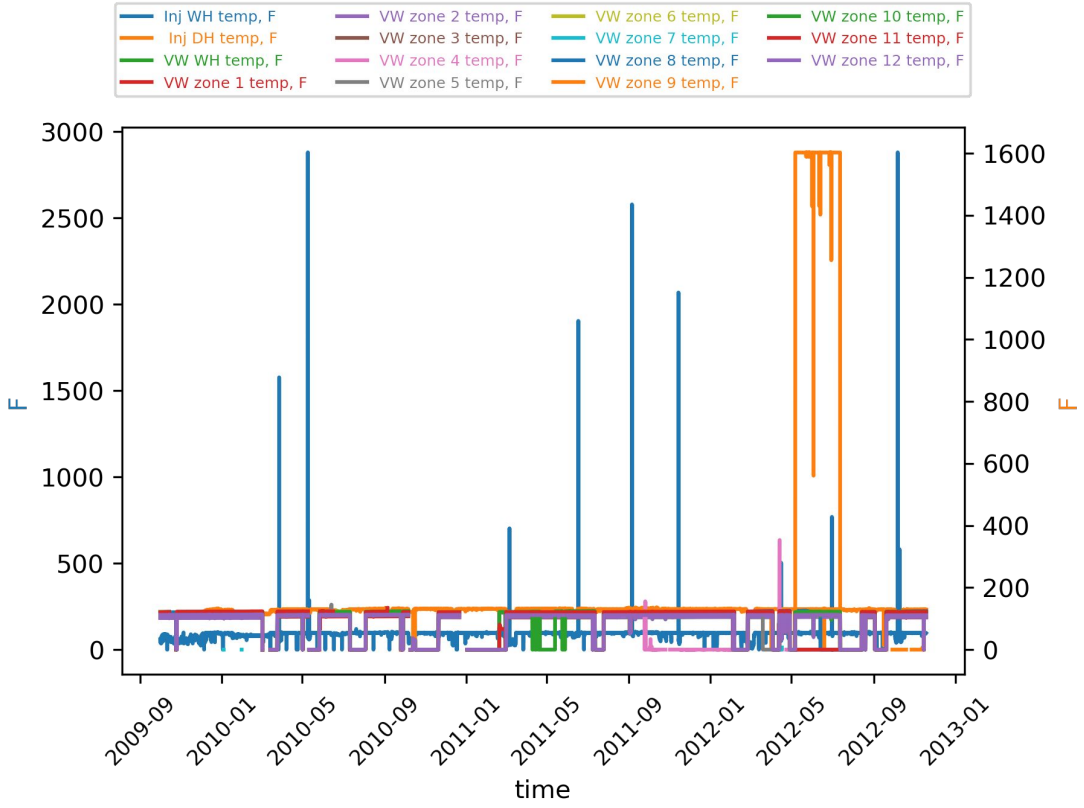
2.1. Data analysis



Pressure series:

1. **Injection Well pressures:**
   - Well Head
   - Down Hole
   - Annulus
2. **Verification Well pressures:**
   - Well Head
   - Annulus
   - 12 Additional time series in different zones (Pressure in zones 1- 12)

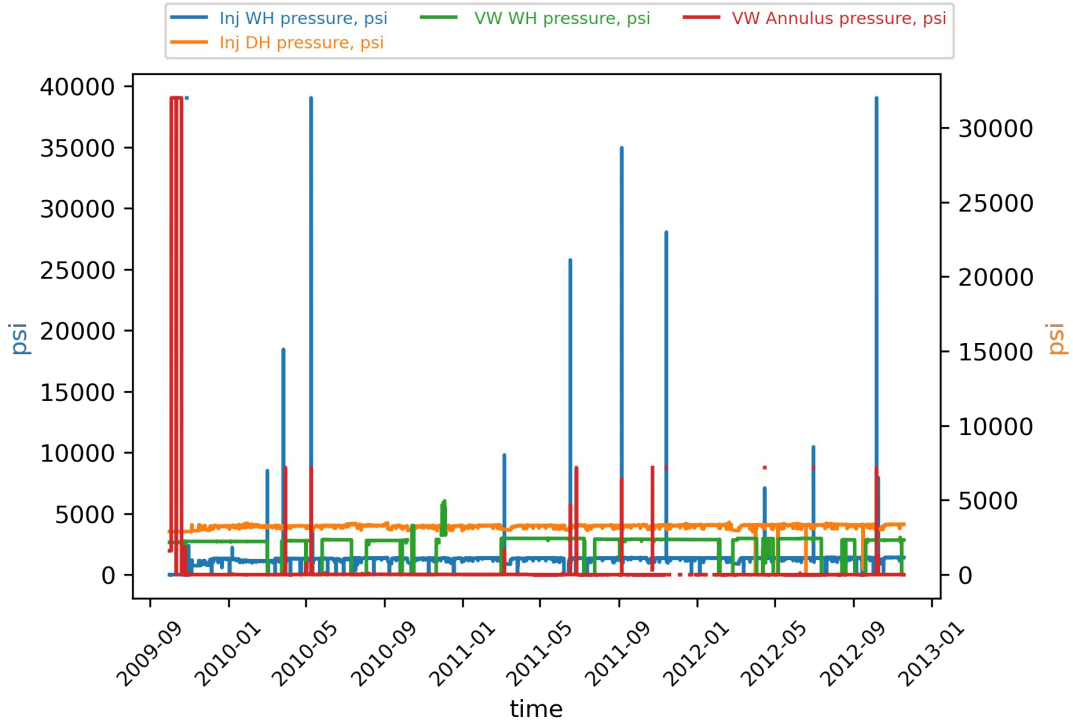
2.1. Data analysis



Temperature series:

1. **Injection Well pressures:**
   - Well Head
   - Down Hole
2. **Verification Well pressures:**
   - Well Head
   - 12 Additional time series in different zones (Temperatures in zones 1-12)

2.1. Data analysis



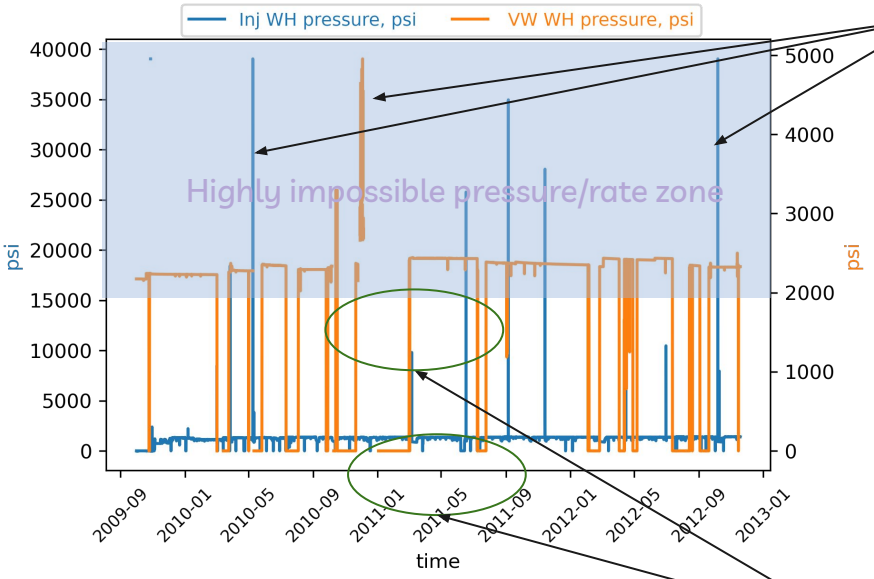
- A lot of outliers and non-physically meaning data points (e.g. Pbh =0 psi) - e.g. VW pressures on the left Figure.
- Strong correlation (actually physically linked time series - pressure at different depths)
- In the dataset, there exist periods with absent information for which imputation techniques cannot be applied to accurately estimate the missing values.

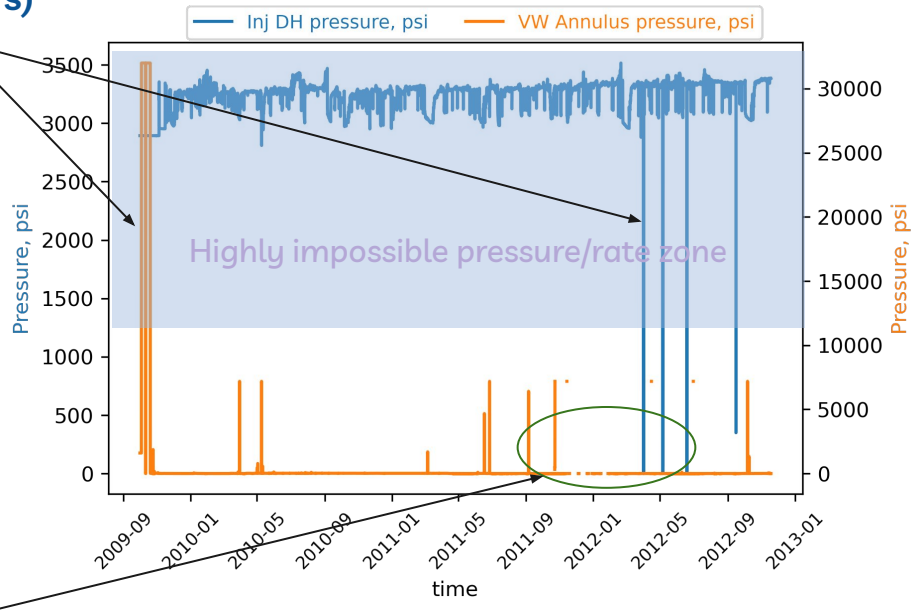
2.1. Data analysis

Well Head pressure

Abnormally high pressure (outliers)

Down Hole pressure



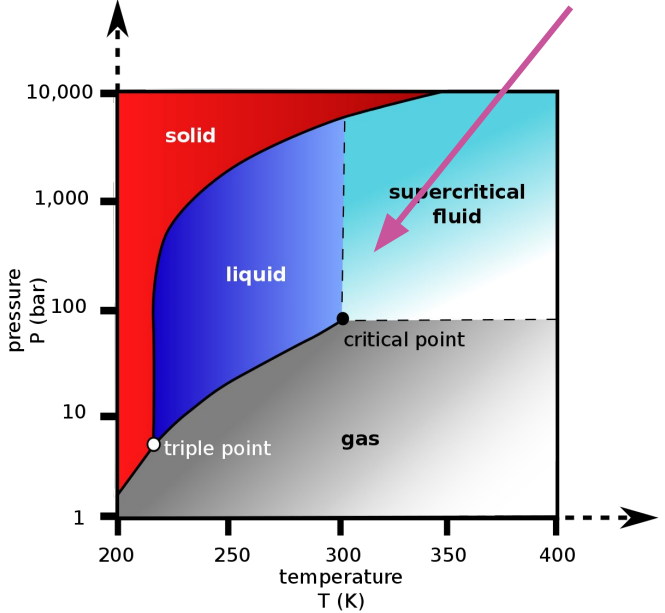


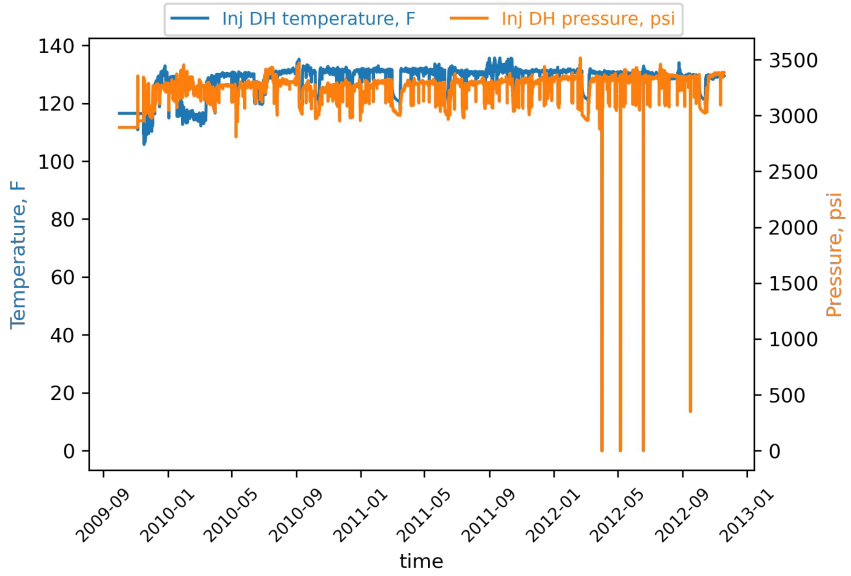
Missing data

2.1. Data analysis

CO2 injection law to predict by ML from PVT properties of CO2 and field data

Single phase approximation of Darcy law holds
(assuming constant skin factor due to Pbh < Shmin)





CO2 is injected as supercritical fluid (single phase). Rate is approximately proportional to delta pressure between BH and Pres. It means that Target variable should strongly depends on pressure changes (dQ/dt ~ dP/dt), and slightly on temperature (=density and viscosity) changes.

2.2. Data pre-processing

Number of missing (NaN) values

SampleTimeUTC	0
Avg_PLT_CO2InjRate_TPH	66
Avg_PLT_CO2VentRate_TPH	66
Avg_CCS1_WHCO2InjPs_psi	194
Avg_CCS1_WHCO2InjTp_F	66
Avg_CCS1_ANPs_psi	160
Avg_CCS1_DH6325Ps_psi	66
Avg_CCS1_DH6325Tp_F	66
Avg_VW1_WBTbgPs_psi	1337
Avg_VW1_WBTbgTp_F	1403
Avg_VW1_ANPs_psi	3977
Avg_VW1_Z11D4917Ps_psi	776
Avg_VW1_Z11D4917Tp_F	755
Avg_VW1_Z10D5001Ps_psi	776
Avg_VW1_Z10D5001Tp_F	755
Avg_VW1_Z09D5653Ps_psi	776
Avg_VW1_Z09D5653Tp_F	755
Avg_VW1_Z08D5840Ps_psi	1275
Avg_VW1_Z08D5840Tp_F	1586
Avg_VW1_Z07D6416Ps_psi	1479
Avg_VW1_Z07D6416Tp_F	1479
Avg_VW1_Z06D6632Ps_psi	1964
Avg_VW1_Z06D6632Tp_F	1964
Avg_VW1_Z05D6720Ps_psi	3509
Avg_VW1_Z05D6720Tp_F	3509
Avg_VW1_Z04D6837Ps_psi	864
Avg_VW1_Z04D6837Tp_F	864
Avg_VW1_Z03D6945Ps_psi	3103
Avg_VW1_Z03D6945Tp_F	1532
Avg_VW1_Z02D6982Ps_psi	1041
Avg_VW1_Z02D6982Tp_F	1041
Avg_VW1_Z01D7061Ps_psi	2157
Avg_VW1_Z01D7061Tp_F	2356
Avg_VW1_Z0910D5482Ps_psi	755
Avg_VW1_Z0910D5482Tp_F	755
Target	67
dtype: int64	

Missing values

Imputation strategy

Missing hours imputation - Interpolation of features

Manual imputation for selected features

K-nearest neighbour imputation

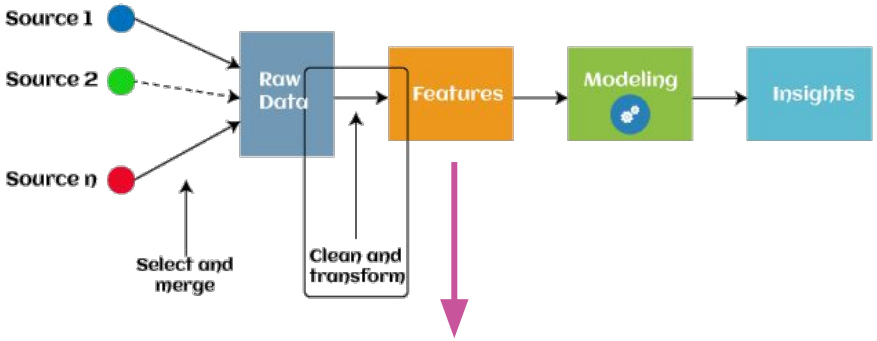
2.2. Data pre-processing - dealing with missing data

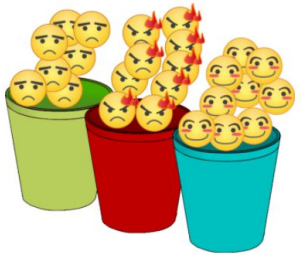
1. We had discovered that ***66 hours*** of data were missing from the train dataset. To resolve this issue, we used ***linear interpolation*** to create new data points.
2. The test dataset has a different challenge: two variables (*Avg_VW1_Z03D6945Ps_psi* and *Avg_VW1_Z03D6945Tp_F*) measured at 6,945 feet had ***66% missing data***. As a result, we decided to ***exclude these variables*** from the training dataset since they couldn't be effectively used for prediction purposes.
3. A ***forward-fill imputation*** method can be utilized to estimate missing values on some features based on preceding observations.
4. Additionally, for the feature *Avg_CCS1_WHCO2InjPs_psi*, we ***manually imputed*** the values by examining nearby values.
5. For all other missing data, we used ***K-Nearest Neighbors (KNN)*** imputation.

2.3. Feature engineering

The original dataset is expanded using the following:

- Differences in temperature/pressure at different heights
- Lagged features (up to *t*-5)
  - Feature values from previous time steps
  - Percentage changes: The rate of percentage change of every feature from *t* time steps ago
  - Differences: The differences in feature values from *t* times steps ago
- Absolute and log absolute values of all features
- Trend features
  - Simple Moving Average (SMA): 5-time step rolling average
  - Exponential Moving Average (EMA): 5-time step exponentially weighted average





By employing a diverse set of feature engineering techniques, we have created ~2,800 features that capture both temporal and statistical information from the dataset.

2.3. Feature engineering

All new features created contain in their name information to understand how they were created.

Examples of features :

Temperature diff 6632-5001 ft_lag3 : Difference in temperature between altitudes 6632 and 5001 with lag 3 time periods

Avg_CCS1_WHCO2InjTp_F_pct_change3 : Percentage difference between the current WHCO2InjTp_F and same value 3 time periods before

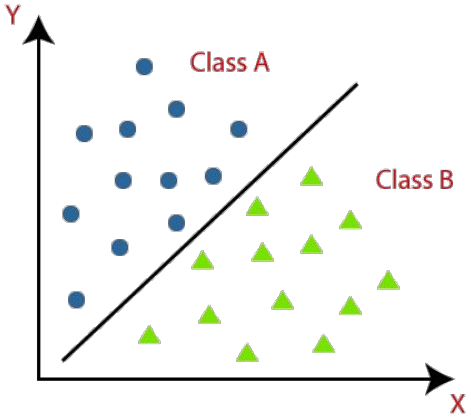
Avg_VW1_Z01D7061Tp_F ema 5 : Exponential moving average of the last 5 values of VW1_Z01D7061Tp_F

3. Machine learning solution

3.1 Proposed approach

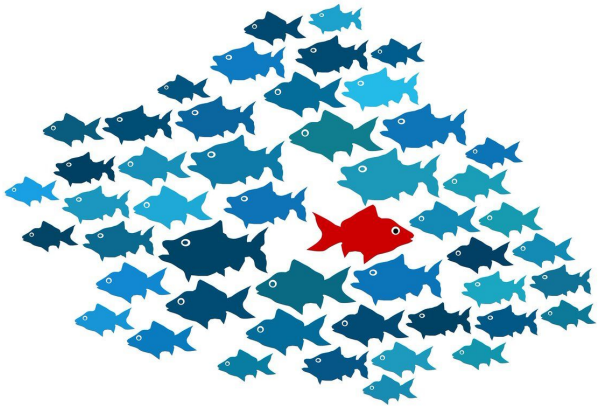
3.2 Classifier

- Feature selection
- Hyperparameter tuning



3.3 Anomalies Regressor

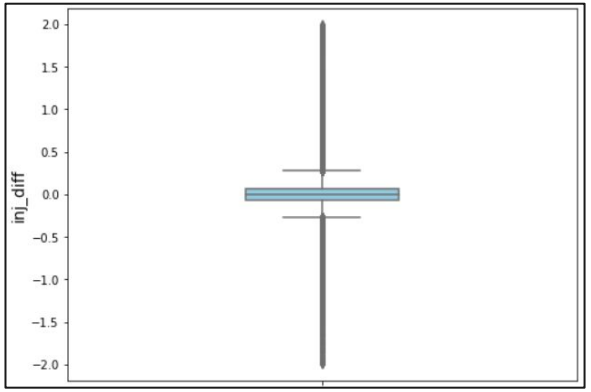
3.3 Low-values Regressor



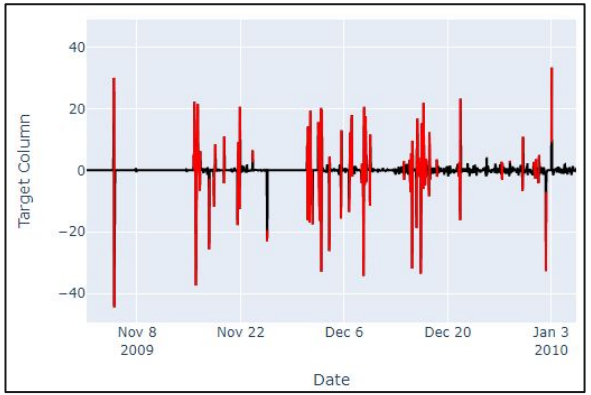
3.1. Proposed approach

Motivation

- Close to 97% of the **Target** has values from **-2** tph/h to **+2** tph/h (noise floor). Given that the metric of interest is **RMSE**, we decided to focus on predicting the 3% of data points outside of the interval.
- The classifier can be thought of as a first screening of the data points, that removes points hard to predict (false negatives) and introduce some noise to make it more robust (false positives).
- If the goal is to predict instances of large fluctuations, the classifier can be used independently.



A) Boxplot of target value within -2 and 2.



B) Target value vs time snippet plot.

3.1. Proposed approach

Methodology

We propose a three-step approach:

1) **Create a classifier to identify anomalies:** The classifier will be trained to separate data points with absolute target value above and below 2 tph/h
   - **Positive class:** Data points with absolute target value greater than 2 tph/h
   - **Negative class:** Data points with absolute target value less than 2 tph/h

2) **Create a regressor to predict the anomalies:**
   - A regressor will be trained on data points the classifier predicts as **positive class**

3) **Create a regressor to predict low-value data points (between -2 and 2):**
   - A regressor will be trained on **true** data points belonging to the **negative class** (classifier's predictions are not used)

3.2. Classifier - Feature selection

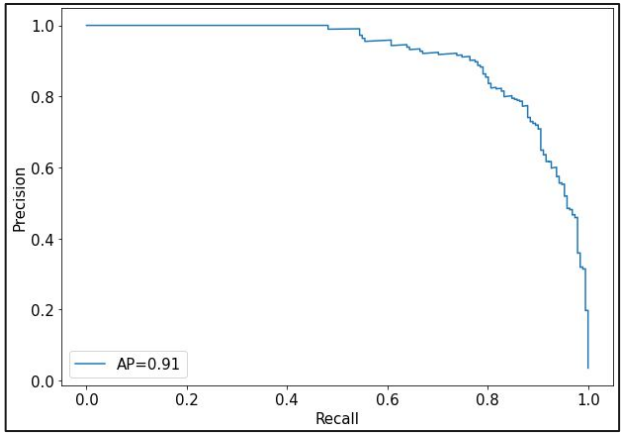
In this step, a first classifier is trained on the first **60%** of the data and validated on the subsequent **20%**.

XGBoost model with semi-optimized hyperparameter is trained using all ~3000 generated features. Given the robust capabilities of XGBoost, it is not necessary to eliminate correlated features or an exhausting imputation of missing values.

The top K features for predicting anomalies are:

- Top 10 features by ***gain*** using XGBoost's importance list
- Top 10 features by ***weight*** using XGBoost's importance list
- Top 10 features by ***SHAP value***

Plots show the results in the first 20% validation set.



A) Validation precision-Recall curve

		Actual	
		Negative (0)	Positive (1)
Predicted	Negative (0)	5279	42
	Positive (1)	17	149

B) Validation confusion matrix

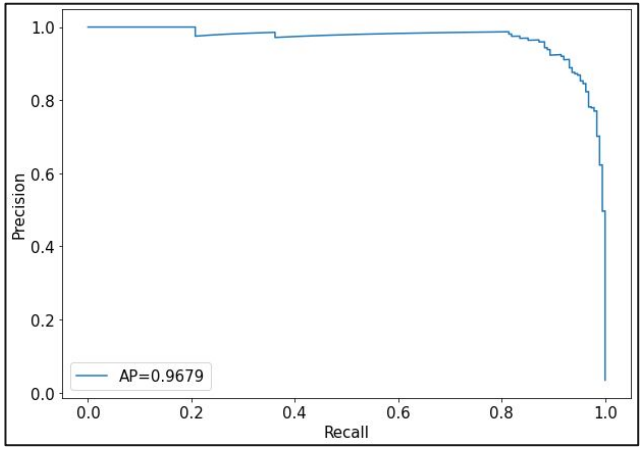
3.2. Classifier - Hyperparameter tuning

A final classifier is trained on **80%** of the data and validated with the last **20%**. The **top K features** selected from the previous step were used. We performed a **grid search of hyperparameters** to maximize validation AUC-PR.

Plots show the results on the second 20% validation set.

The classifier trained with 100% of the data will give us predicted labels, later used by the *anomalies regressor*:

- **1052** true positives
- **73** false positives



A) Validation Precision-Recall curve

		Actual	
		Negative (0)	Positive (1)
Predicted	Negative (0)	5282	17
	Positive (1)	13	175

B) Validation confusion matrix

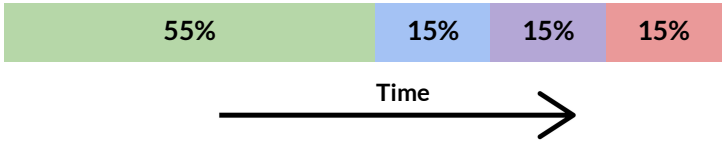
3.3. Anomalies Regressor

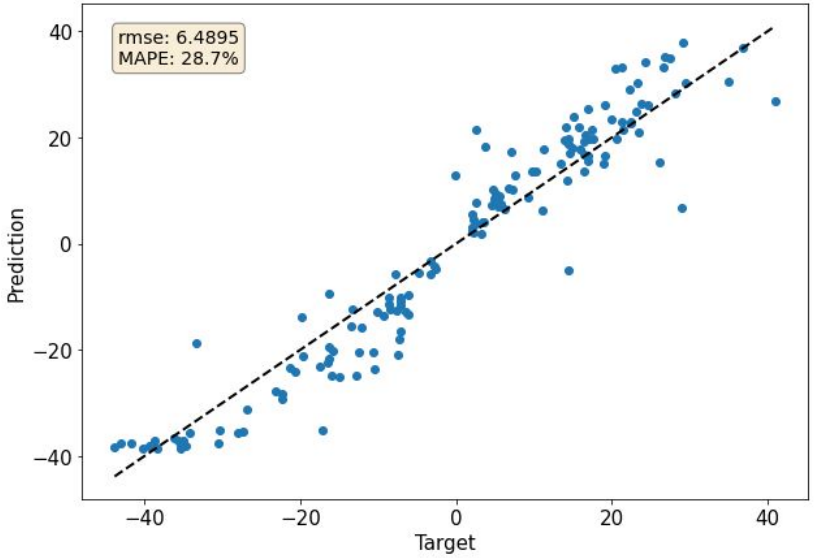
The regressor is trained on the data points the **classifier predicts as positive** (absolute target value greater than 2). The classifier removes data points hard to classify - hence, hard to predict.

The data splits used are **55%-15%-15%-15%**.

XGBoost model is also used here and the same framework used for the classifier is adopted:

- **Feature selection:** Using the first 55-15 split, we select the top K features.
- **Hyperparameter tuning:** Using the first 70-15 split, we select the best hyperparameters.
- **Final validation:** The regressor is trained on the first 85% of the data and the last 15% is used to select the best model.





True vs predicted plot of the 15% hold-out set

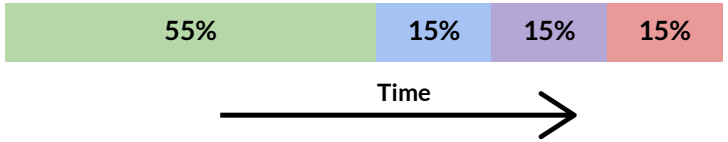
3.3. Low-values Regressor

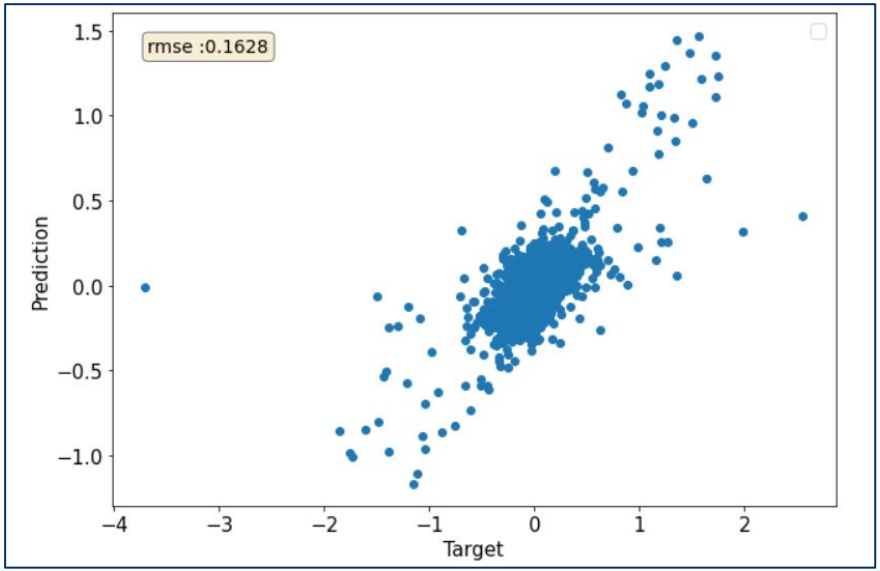
The regressor is trained on the **true** data points the **belonging to the negative class** (absolute target value less than 2). We believe using the true points versus the predictions of the classifier helps remove outlier points and makes it easier to train.

The data splits used are **55%-15%-15%-15%**.

XGBoost model is also used here and the same framework used for the classifier is adopted:

- **Feature selection:** Using the first 55-15 split, we select the top K features.
- **Hyperparameter tuning:** Using the first 70-15 split, we select the best hyperparameters.
- **Final validation:** The regressor is trained on the first 85% of the data and the last 15% is used to select the best model.





True vs predicted plot of the 15% hold-out set

4. Results

4.1 Putting it all together

4.2 Classifier SHAP values

4.3 Positive class regressor SHAP values

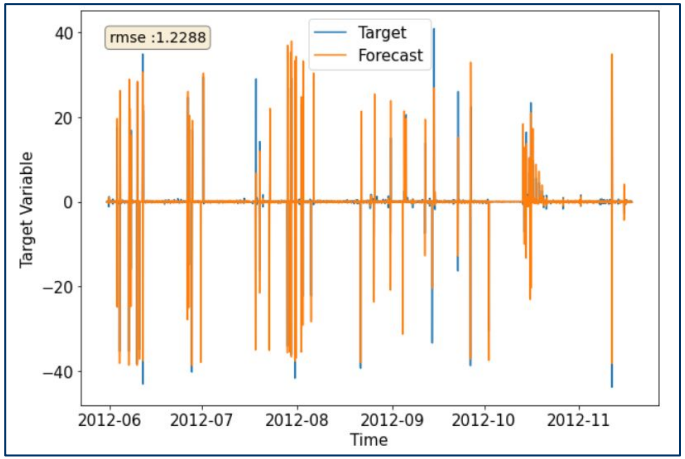
4.4 Negative class regressor SHAP values

4.1. Putting it all together

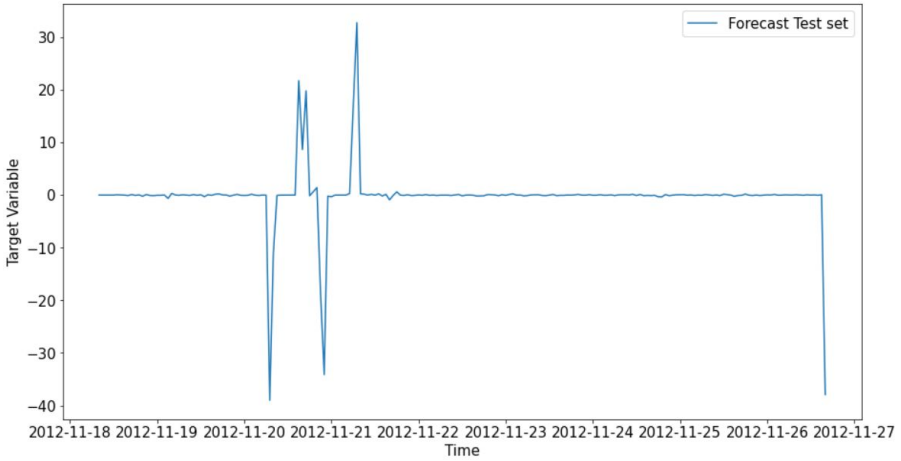
Nice job! Your prediction scored

0.9542

The **classifier** is initially employed to determine which regressor is needed. For data points **predicted** as belonging to the **positive class**, the **anomalies regressor** is used to estimate the target value. For data points **predicted** as belonging to the **negative class**, the **second regressor** is used. Considering that lagged features rely on prior time steps, the first 5 data points in the test set cannot be reliably predicted (therefore, imputed to 0s).



A) Prediction of the last 15% data points in the train set



B) Test set predictions

4.2. Classifier SHAP values

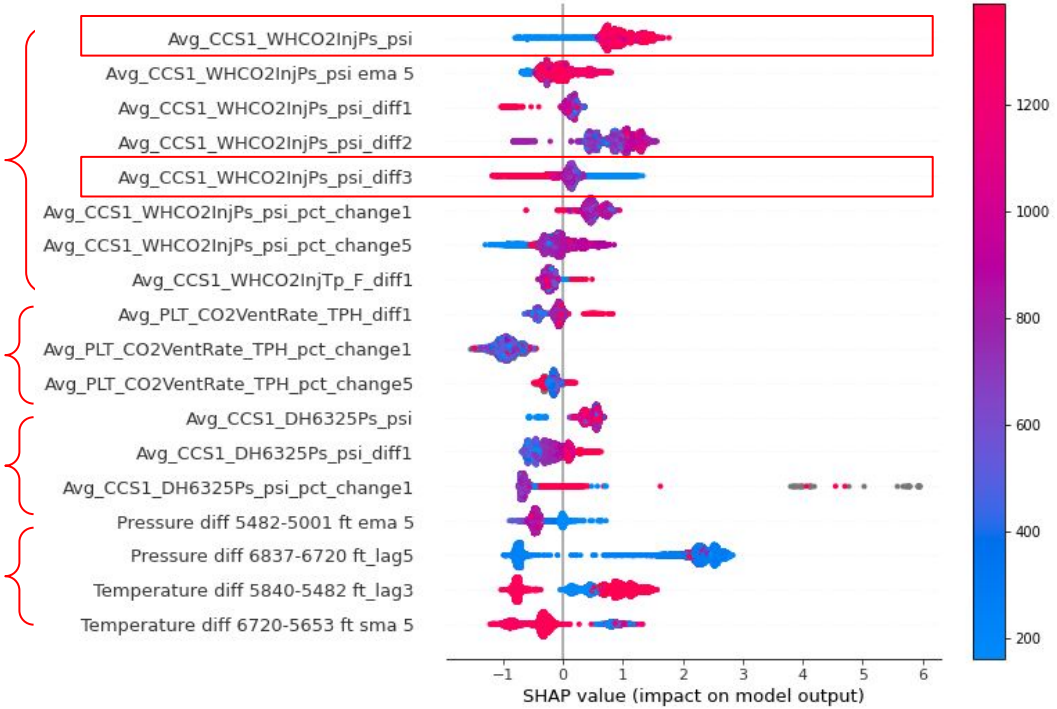
Highlighted features exhibit a monotonic relationship between their values and contributions to the classifier's probability output

CCS1_WHOCO2Inj derived features
(CCS1 wellhead pressure/temperature)

PLT_CO2VentRate derived features
(CO2 vent rate)

CCS1_DH6325Ps derived features
(CCS1 downhole pressure)

Pressure and temperature differences



4.3. Positive class regressor SHAP values

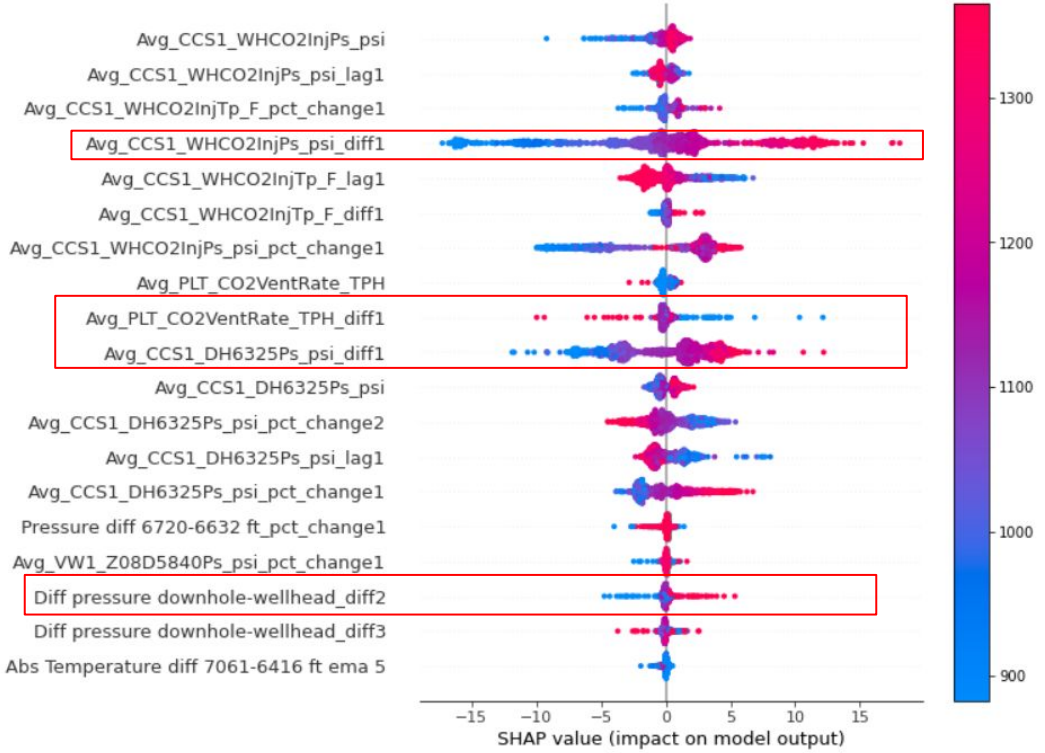
Highlighted features exhibit a monotonic relationship between their values and contributions to the regressor's prediction

CCS1_WHOCO2Inj derived features
(CCS1 wellhead pressure/temperature)

PLT_CO2VentRate derived features
(CO2 vent rate)

CCS1_DH6325Ps derived features
(CCS1 downhole pressure)

Pressure and temperature differences



4.3. Negative class regressor SHAP values

Highlighted feature exhibits a monotonic relationship between its value and contribution to the regressor's prediction

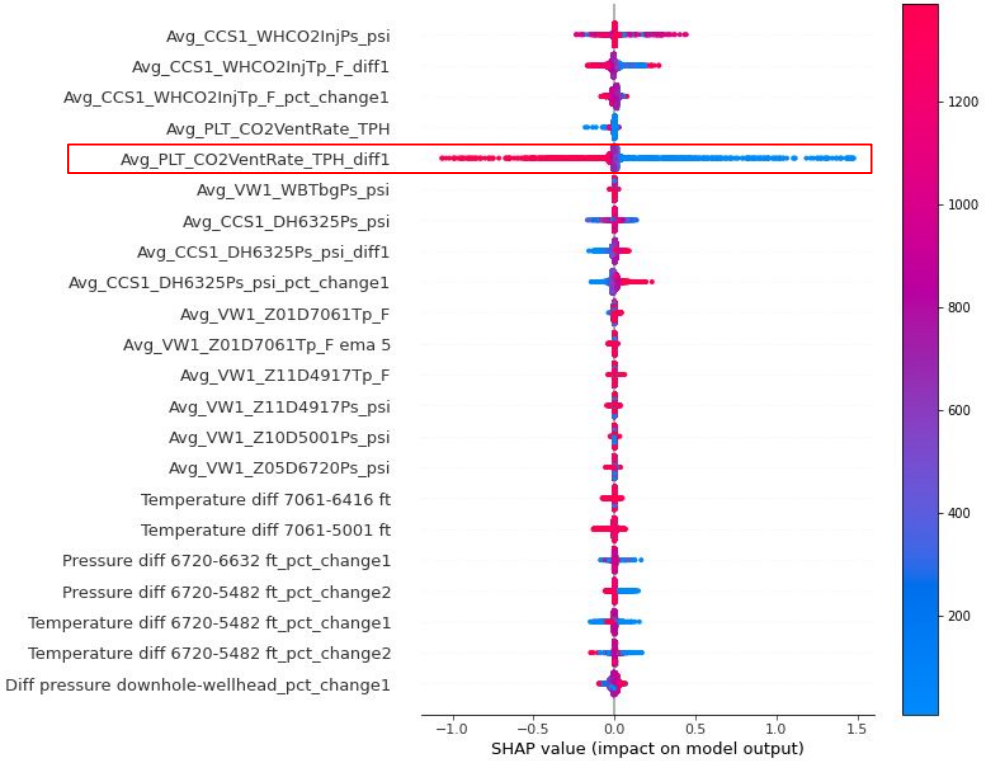
CCS1_WHOCO2Inj derived features
(CCS1 wellhead pressure/temperature)

PLT_CO2VentRate derived features
(CO2 vent rate)

CCS1_DH6325Ps derived features
(CCS1 downhole pressure)

Pressure and temperature non-transformed values

Pressure and temperature differences



5. Discussion

Addressing this challenge involved the following steps:

- Analyzing and cleaning the data
- Conducting feature engineering to enhance the feature set
- Implementing a method that initially screens data with a classifier, followed by the development of a separate regression models for each class
- Carrying out feature selection and hyperparameter tuning for each model to minimize the metric of interest

In addition, we took measures to minimize target leakage risks:

- Partitioning the data into multiple validation sets, taking the time component into account
- Removing data points with shared information between training and validation sets due to lagged features

The final model selection was based on results from the last 15% hold-out validation set. We explored various alternatives, such as excluding the classifier and using a single regression model, or training the regressors with data points from the true positive class only. For further information, please refer to the backup slides.

5. Discussion

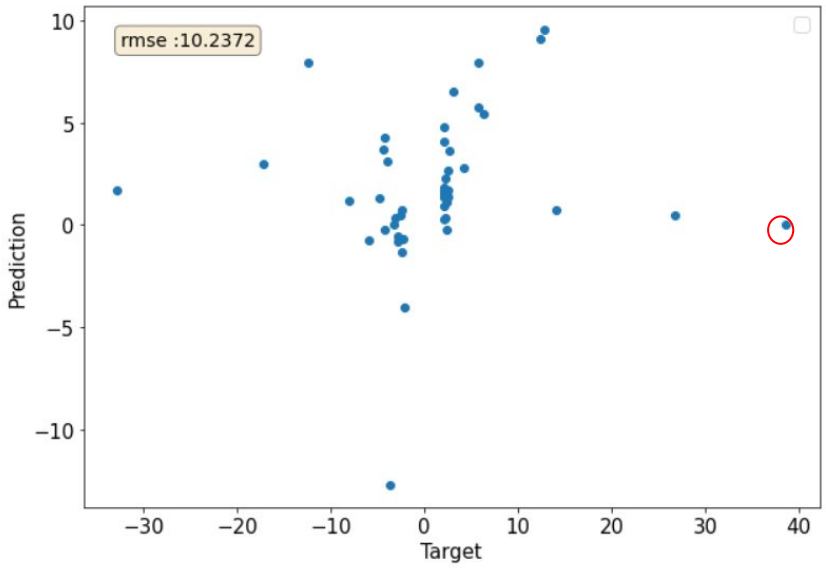
After analysis of the most important features across models, we can conclude that only a few measurements are needed to predict the injection rate delta with high degree of accuracy:

- **CCS1 wellhead pressure/temperature**
- **CO2 vent rate**
- **CCS1 downhole pressure**
- **Extra measurements from the fiber optic are useful but not necessary**

SHAP value interpretation:

The SHAP values represent the average contribution of each feature to the model's prediction (probability for the classifier and estimation for the regressor). The colors indicate the feature values, red for higher values and blue for low values. For this given example, the feature Avg_CCS1_WHCO2InjPs_psi contributes positively to the prediction when it has high values and negatively when having low values. Similarly, other features can be interpreted.

5. Discussion : Why do we use predicted labels to train the regressor ?



Regressor model results on points mislabeled by classifier

After doing an analysis of the points that were **misclassified** we realized that there injection difference value was **not necessarily close** to the [-2,2] border but ranged from [-30,40].

After an analysis of the **values** of the **different features** we noticed that most points don't show values that would led to being predicted as an **abnormal point** (absolute value higher than 2).

We believe that those points might be either **errors in the data** or some times where the **injection stopped** for maintenance (or any other reason)

As we can see in the graph on the left, the regressor has a **hard time predicting** for those points and since some of them have a high value, the regressor will try to train to predict them which can lead to **errors predicting** the points that do belong to class 1

5. Discussion : Why do we use predicted labels to train the regressor ?

Features	False high Value	Shap impact False high	True high Value	Shap impact True high
Avg_CCS1_WHCO2InjPs_psi_diff1	-2.793625	0.256342	1302.711877	11.655196
Avg_CCS1_WHCO2InjPs_psi_pct_change1	-0.002534	2.153897	NaN	5.708593
Avg_CCS1_WHCO2InjTp_F_lag1	80.732462	0.220311	0.000000	4.586085
Avg_CCS1_WHCO2InjTp_F_diff1	-0.161511	-0.220111	95.553879	2.643117
Avg_CCS1_DH6325Ps_psi_diff1	-9.951215	-0.293291	-0.294719	2.632650
Avg_CCS1_WHCO2InjTp_F_pct_change1	-0.002001	-0.065340	NaN	1.711510
Avg_CCS1_DH6325Ps_psi_pct_change1	-0.003025	-0.921192	-0.000091	1.370582
Avg_CCS1_DH6325Ps_psi_pct_change2	-0.002924	-0.024919	0.000849	1.195470
Avg_CCS1_WHCO2InjPs_psi	1099.778038	-0.243055	1302.711877	1.116126
Avg_CCS1_DH6325Ps_psi	3280.075168	0.263499	3252.473927	1.095933
Avg_CCS1_WHCO2InjPs_psi_lag1	1102.571663	0.274964	0.000000	0.390300
Abs Temperature diff 7061-6416 ft ema 5	6.593172	0.095052	4.243114	0.228118
Avg_VW1_Z08D5840Ps_psi_pct_change1	0.000000	-0.060719	0.000002	0.183352
Diff pressure downhole-wellhead_diff3	-32.856002	-0.023852	-591.238701	0.152396
Pressure diff 6720-6632 ft_pct_change1	-0.000200	0.030879	-0.000179	0.049242
Avg_PLT_CO2VentRate_TPH	0.000000	-0.422363	0.000000	0.011324
Diff pressure downhole-wellhead_diff2	50.311192	0.021769	-589.152898	-0.146839
Avg_PLT_CO2VentRate_TPH_diff1	0.000000	-0.181818	0.000000	-0.186481
Avg_CCS1_DH6325Ps_psi_lag1	3290.026383	-0.824091	3252.768646	-0.418108

This table shows the **results of the regressor** on two points. The first one is a point that has a **high value** (close to 40, red point circled on the previous graph) but was **misclassified**, the second one also has a **high value** but was **correctly classified.**

As seen in the previous slides, the **first feature** is one of the **most important** to predict the output and is **positively correlated** with the output. The **higher** the difference in CO2 pressure with previous time period the **higher** the injection rate difference. When we look at the false high value (the point that was not classified as 1) we can see that the value for the first feature is even **negative**, so there **would be no reason** for this point to have a high value, it's most probably a **shut down or a maintenance.**

6. Conclusion

Nice job! Your prediction scored

0.9542

- The case study on ML Challenge presented an analysis of CO2 Containment prediction.
- The study used a combination of machine learning algorithms: **3 XGBoost models (1 classifier and 2 regressors)** to achieved an **RMSE** of **0.95** in the test set.
- **Feature engineering** part involved selecting and transforming relevant variables to create new features that better capture the underlying patterns in the data.
- After fine-tuning, the **XGBoost classifier** achieves up to 0.96 area under the Precision-Recall curve in predicting if a data point has an absolute target value higher or lower than 2.
- Dividing the data into two distinct regimes and developing two individual regression models simplifies the problem.
- Overall, the study demonstrates the effectiveness of machine learning in solving the problem stated which we believe to be predicting changes in injection rate and understanding what makes the injection rate go down from its original value.

References

1. A.T. Akono, G. Davila, J. Druhan, Z. Shi, K. Jessen, T. Tsotsis, S. Fuchs, D. Crandall, L. Dalton, M.K. Tkach, A.L. Goodman, S. Frailey, C.J. Werth. A review of geochemical–mechanical impacts in geological carbon storage reservoirs. Greenhouse Gas Sci Technol, 9 (2019), pp. 474-504, 10.1002/ghg.1870

2. Bauer RA, Carney M and Finley RJ, Overview of microseismic response to CO2 injection into the Mt. Simon saline reservoir at the Illinois Basin-Decatur Project. Int J Greenh Gas Control 54:378–388 (2016).

3. Bauer RA, Will R, Jaques P, Smith V and Payne WG, Pre-thru post-injection monitoring of microseismicity at Illinois Basin Decatur Project and static and dynamic modelling efforts for monitoring and event prediction (2016). Available: http://sequestration.org/resources/PAGMay2016Presentations/09aBauer_Will_2016-May-16_Microseismicity_IBDP.pdf [29 April 2019].

4. Will R, El-Kaseeh G, Jaques P, Greenberg S and Finley R, Microseismic data acquisition, processing, and event characterization at the Illinois Basin-Decatur Project. Int J Greenh Gas Control 54:404–420 (2016).

5. Frailey SM and Finley RJ, Overview of the midwest geologic storage consortium pilot projects, in SPE International Conference on CO2 Capture, Storage, and Utilization, SPE 139746. Society of Petroleum Engineers, Richardson, TX (2010).

6. Frailey SM, Damico J and Leetaru H, Reservoir characterization of the Mt. Simon Sandstone, Illinois Basin, USA. Energy Procedia 4:5487–5494 (2011)

BACK-UP slides

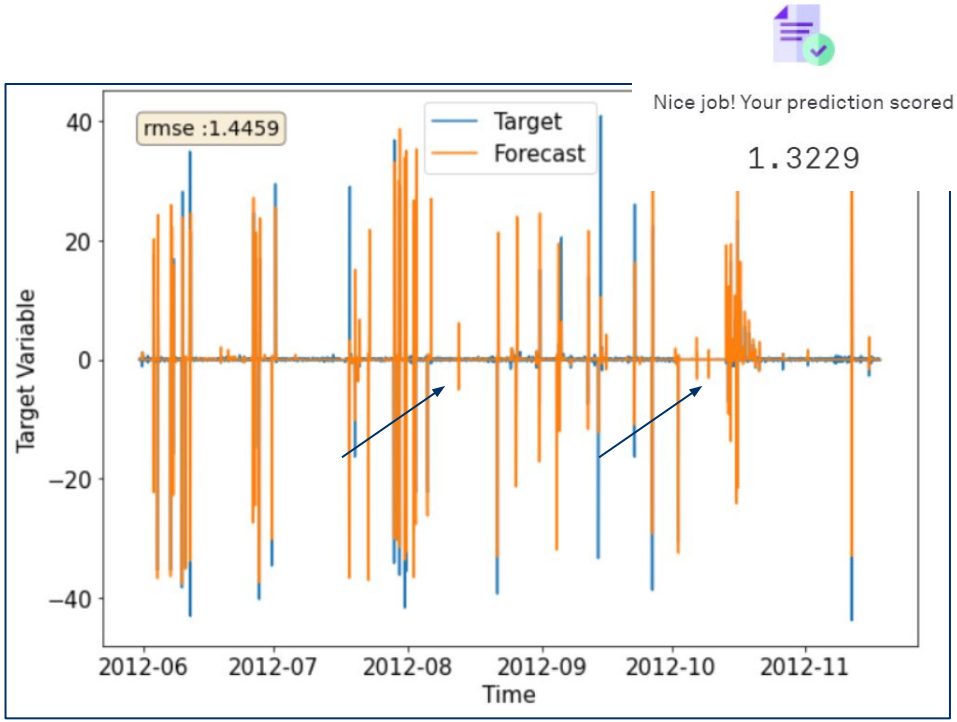
Other approaches

Best model
Validation RMSE: 1.228
Leaderboard RMSE: 0.95

We trained a regressor model (XGBoost) **using all data (no classifier)**. We followed the same procedure as with our best model:

- Split data 55-15-15-15.
- Feature selection (55-15)
- Hyperparameter tuning (70-15)

The results of this regressor show a lower performance in the 15% hold-out validation set **(1.44 rmse vs 1.22)**.



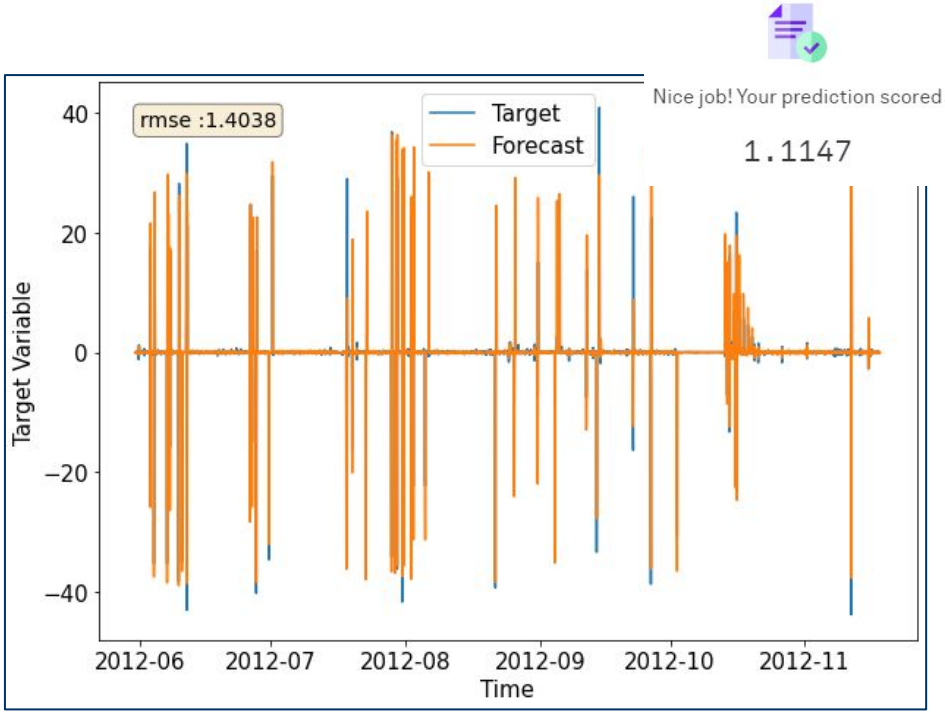
Other approaches

Best model
Validation RMSE: 1.228
Leaderboard RMSE: 0.95

We trained the *anomalies regressor* model (XGBoost) using data points with absolute target values greater than 2 **(regardless of the classifier's prediction)**. The low-values regressor is the same as our best model. We followed the same procedure as with our best model:

- Split data 55-15-15-15.
- Feature selection (55-15)
- Hyperparameter tuning (70-15)

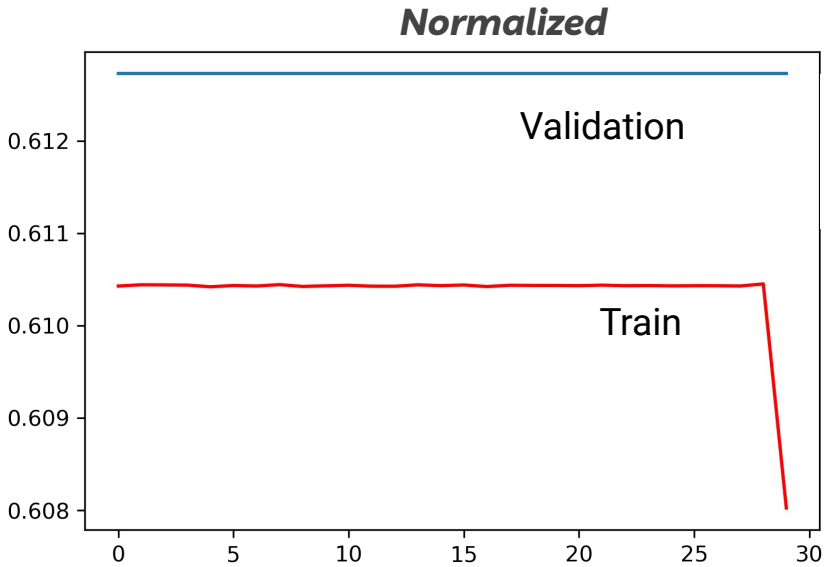
The results of this regressor show a lower performance in the 15% hold-out validation set **(1.40 rmse vs 1.22)**.



Other approaches

We have tried RNN LSTM model with time component only. The results are shown below:

Predicted: last 15%
Epochs: 3
Batches: 32



RMSE

```
Epoch 1/3
851/851 [==============================] - 82s 94ms/step - loss: 0.0266 - accuracy: 3.6728e-05
Epoch 2/3
851/851 [==============================] - 81s 95ms/step - loss: 0.0088 - accuracy: 3.6728e-05
Epoch 3/3
851/851 [==============================] - 81s 95ms/step - loss: 0.0067 - accuracy: 3.6728e-05
```