Description of the Process

I. Data Exploration and Cleaning

The dataset consists of 7525 entries, 329 features, and 7 targets.

The targets are independent and equally distributed.

Features were grouped by prefixes and analyzed in batches. While there are no missing values in the dataset, there are many features which have extreme amounts of 0 values (Fig. 1). 17 features has only 0 values, and 2 features have only 1 unique value (other than 0). Another problem determined is an extreme correlation of features to each other (Fig. 2).

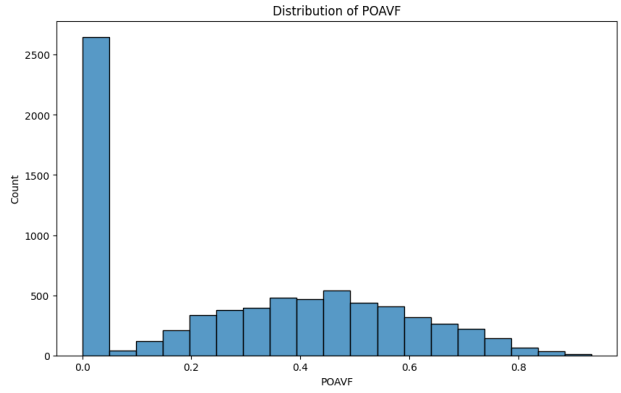

Fig 1. Example of large amount of 0 values POAVF: 35% of 0, while 4526 unique values in total

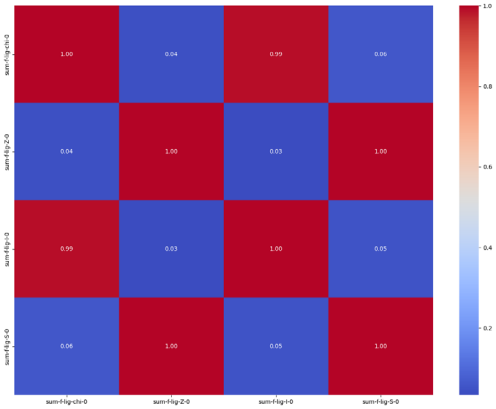

Fig 2. Example of features correlation

Handling highly correlated features and those with cardinality 1 allowed a drastic reduction number of features to 68.

Another observed problem is the vast number of outliers, 1/3 of examples would have to be removed if the threshold is set as much as 10 IQR on the dataset with reduced features. This aspect requires careful further analysis, while was manually handled in the frame of the test task.

II. ML pipeline

A sklearn pipeline was built with an incorporated MinMax scaler, optimization was performed *via* cross-validation with MSE as score. A set of 12 models was tested: ground mean model, variations of linear regression, KNN regressor, tree-based models, including several boosting algorithms, and simple neural network. The process was applied to all features separately; the best estimators for each target were evaluated on the test set (20% of the original dataset).

Additionally, a multiply target regression was tested on the whole set.

Details and results are discussed in the next chapter.

Results & Key Findings

The best estimators for all targets have been selected (Fig. 3, Table 1). The best results were achieved for pure_uptake_CO2_298.00K_16 bar with prediction on test set being better than prediction by mean in 25 times. Good results were also achieved for MPC, logKH_CO2, CO_2 uptake at 0.15 bar and CO2_widomHOA. While MNC can be estimated badly, CO_2 Henry coefficient cannot be reliably estimated at all. It may be related to its' the distribution in the dataset with most values close to 0 (see Fig. 7). Performance of the models on train with cross-validation and test sets demonstrated no or little overfitting (Table 1 & Fig. 8).

Table 1. Results of best models on train and tests sets, with comparison to prediction by mean.

Target	Train MSE	Test MSE	Mean MSE	$\frac{test\ MSE}{mean\ MSE}$
MNC (minimum negative charge)	0.0389	0.0421	0.0713	2
MPC (maximum positive charge)	0.0374	0.0376	0.189	5
logKH_CO2	0.527	0.476	1.24	3
pure_CO2_Henry_coeff	0.00383	0.00341	0.00414	1
pure_CO2_widomHOA	46.2	38.3	120	3
pure_uptake_CO2_298.00K_0.15bar	0.882	0.911	2.34	3
pure_uptake_CO2_298.00K_16bar	1.15	1.23	30.5	25

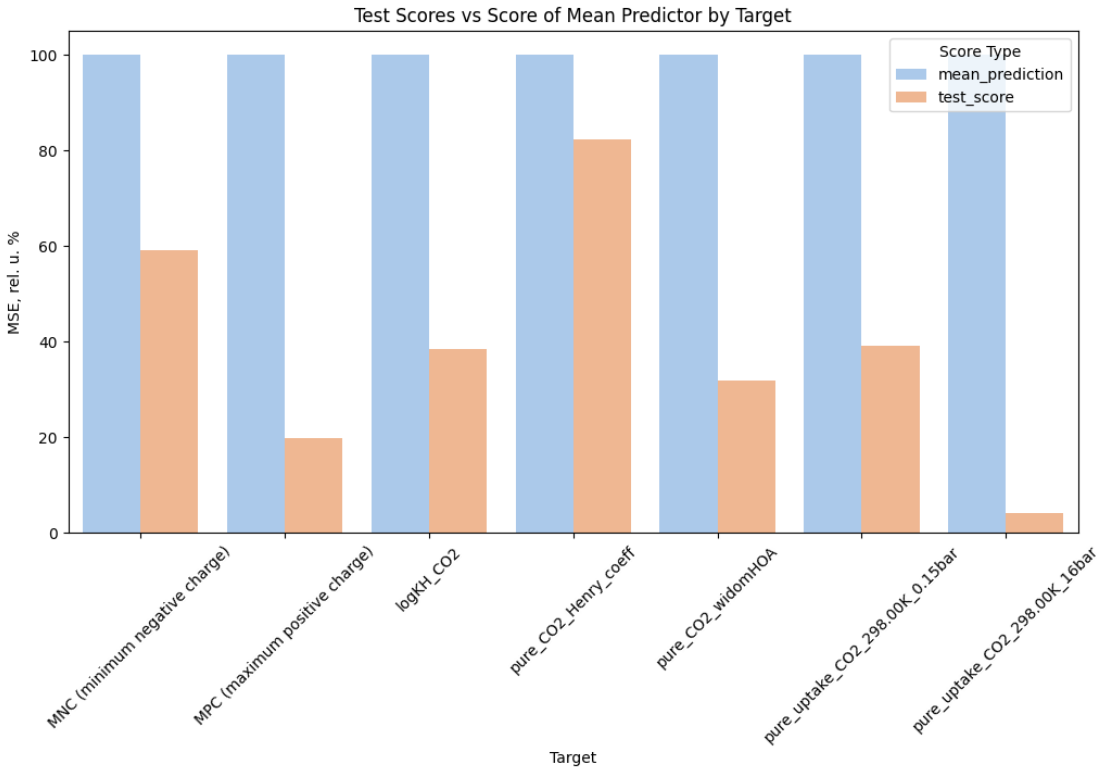

Fig. 3. Performance of the best estimators on test for each target in comparison to prediction by mean (scaled MSE, the less – the better)

In general, linear models handled the task poorly and the best were boosting models, with LGBM being best for 6 out of 7 targets (see Fig. 4 and jupyter notebook). Multilinear regression was also successful for the provided task (for the details, see the jupyter notebook).

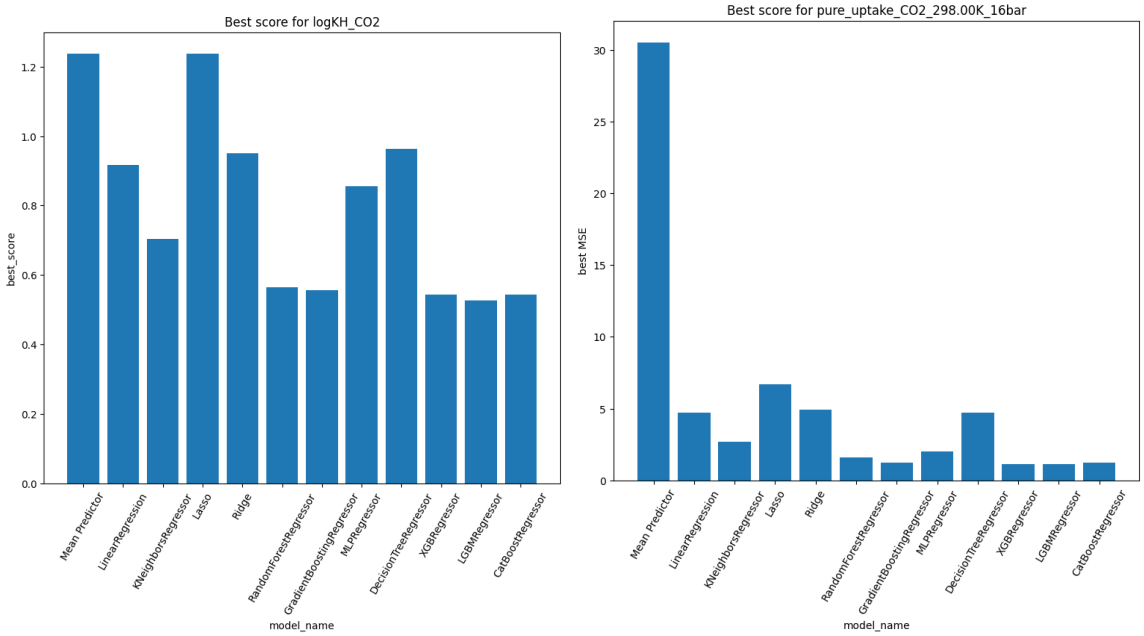

Fig. 4. Comparison of models performing based on cross-validation for logKH_CO2 (left) and pure_uptake_CO2_298.00K (right)

Out of 68 features selected for the models training, the highest importance across the best estimators was for sum-mc_CRY-Z-0-all and D_mc_CRY-S-1-all (Table 2). In most cases, there was a big set of important features with a slight difference across them (Fig. 5 top), however for pure_uptake_CO2_298.00K_16bar an extreme importance was for density (Fig. 5 bottom).

Table 2. Top 3 most important features for best estimator for every target.

Target	1	2	3
MNC (minimum negative charge)	sum-D_mc_CRY-chi-3-all	mc_CRY-I-2-all	CellV [A^3]
MPC (maximum positive charge)	**sum-mc_CRY-Z-0-all**	sum-mc_CRY-chi-0-all	**D_mc_CRY-S-1-all**
pure_CO2_Henry_coeff	Di	**D_mc_CRY-S-1-all**	sum-mc_CRY-chi-0-all
pure_CO2_widomHOA	**D_mc_CRY-S-1-all**	**sum-mc_CRY-Z-0-all**	mc_CRY-S-1-all
pure_uptake_CO2_298.00K_0.15bar	**sum-mc_CRY-Z-0-all**	density [g/cm^3]	total_POV_volumetric
pure_uptake_CO2_298.00K_16bar	density [g/cm^3]	total_POV_volumetric	total_SA_volumetric
logKH_CO2	**sum-mc_CRY-Z-0-all**	**D_mc_CRY-S-1-all**	total_POV_volumetric

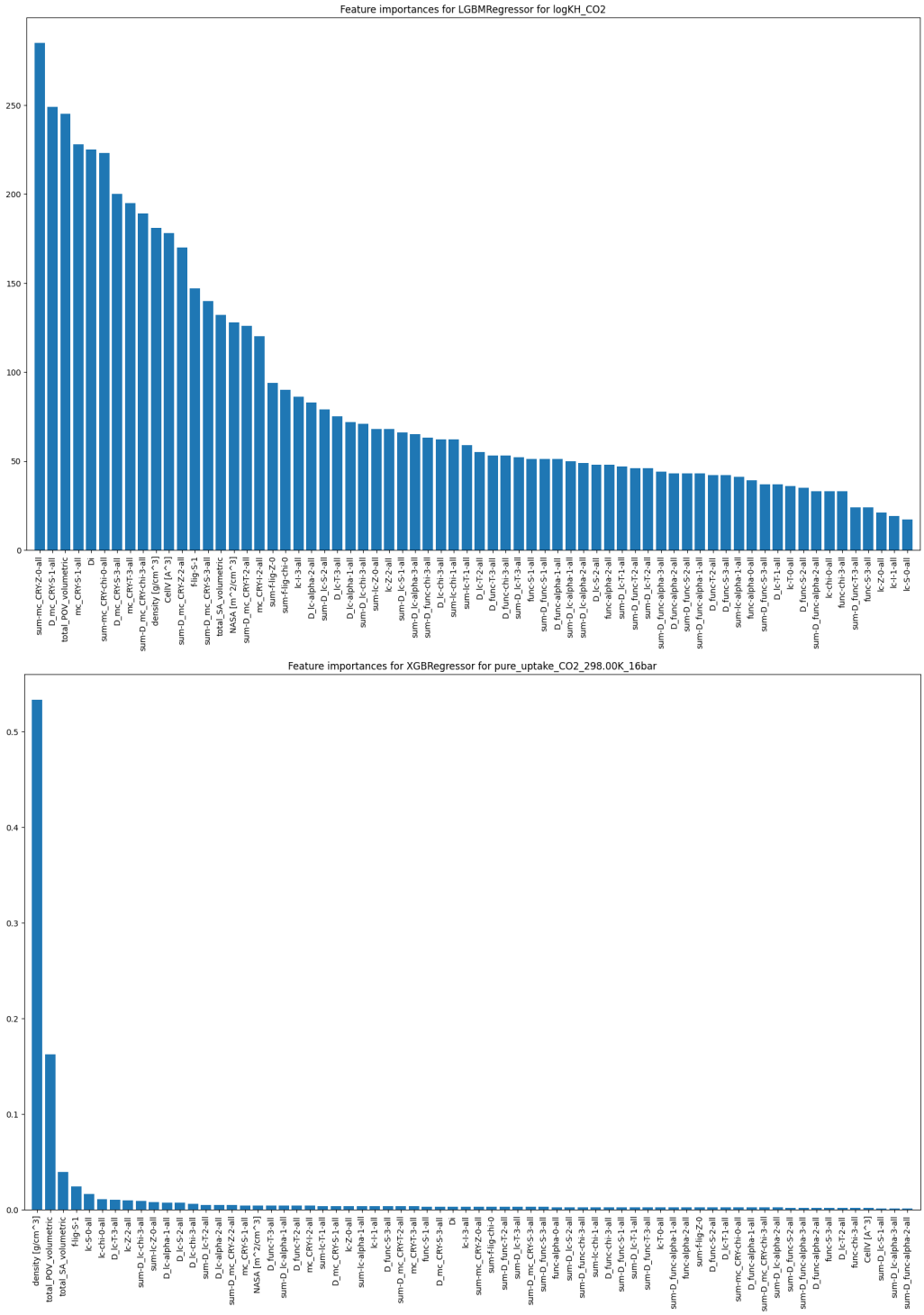

Fig. 5. Feature importance for logKH_CO2 – typical example (top), and pure_uptake_CO2_298.00K_16bar bottom

Possible Following Steps

The first goal would be to **work on features**, namely:

1) Evaluating data sources quality and integrity
2) Select of main features from highly correlated ones based on domain knowledge and their experimental availability
3) Work with outliers with the help of data sources and chemical knowledge
4) Choose the best normalization/scaling method for features based on their distribution
5) It may be valuable to deep into clusters in target distribution, especially for MNC & MPC (Fig. 6)

Then, **ML-part** should be expanded:

1) Evaluate various metrics for the regression problem, based on the target's distribution, it may be beneficial to use an unbalanced one
2) More detailed hyperparameter optimization, and additional models should be evaluated as well (kernel-powered methods, passive-aggressive algorithms, designed small neural networks)
3) Additional round of features selection, and engineering after the initial prediction results
4) Feature importance for various targets is of great importance to understand their relationships and should be studied in more detail. Residual analysis should also be performed
5) There is potential to explore multi-target regression further

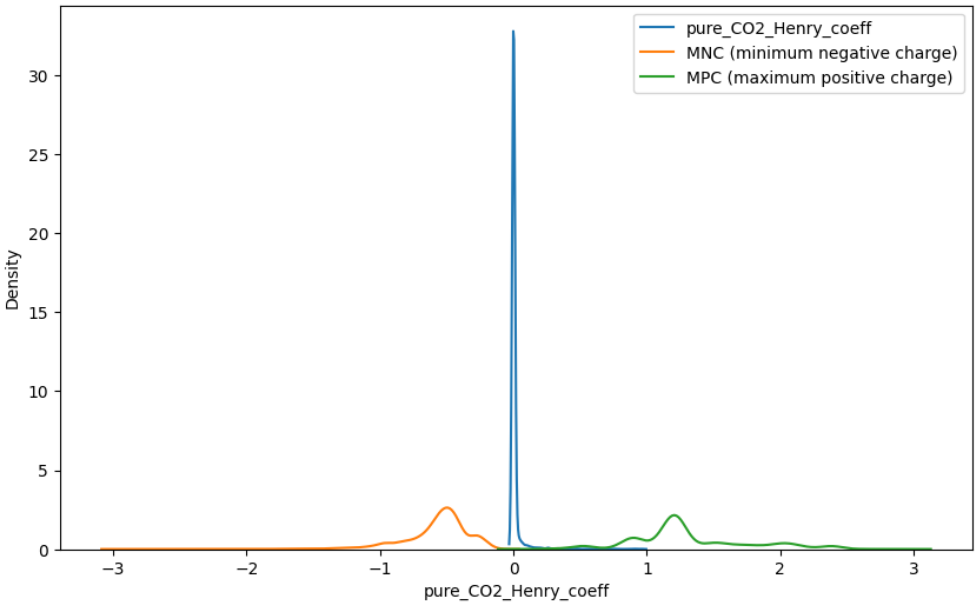

Fig. 6. MNC & MPC distribution

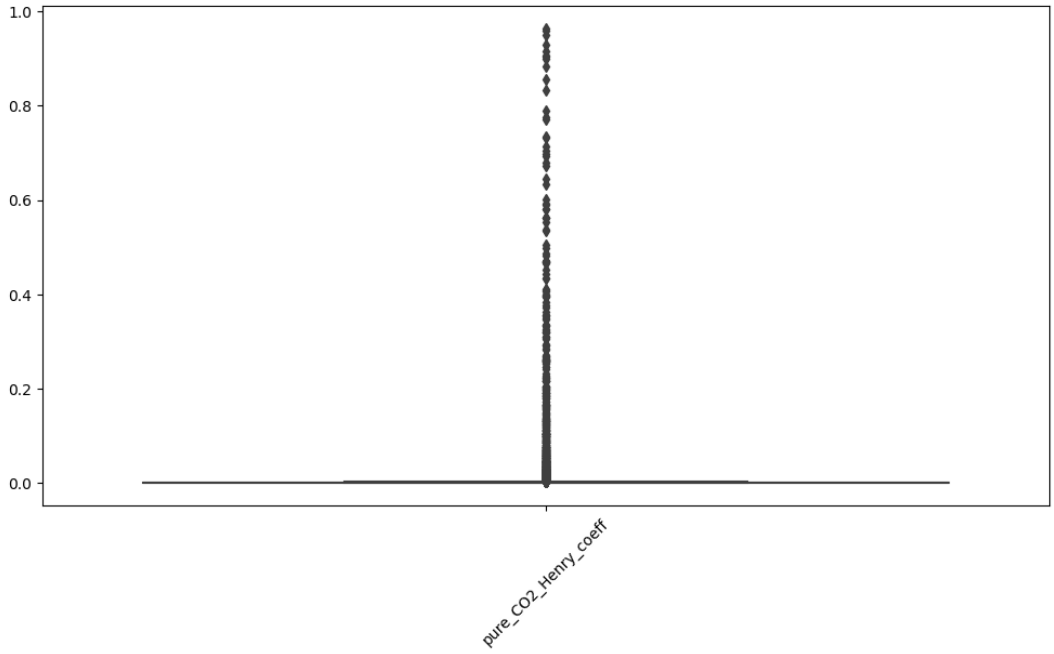

Fig. 7. Boxplot for CO_2 Henry coefficient

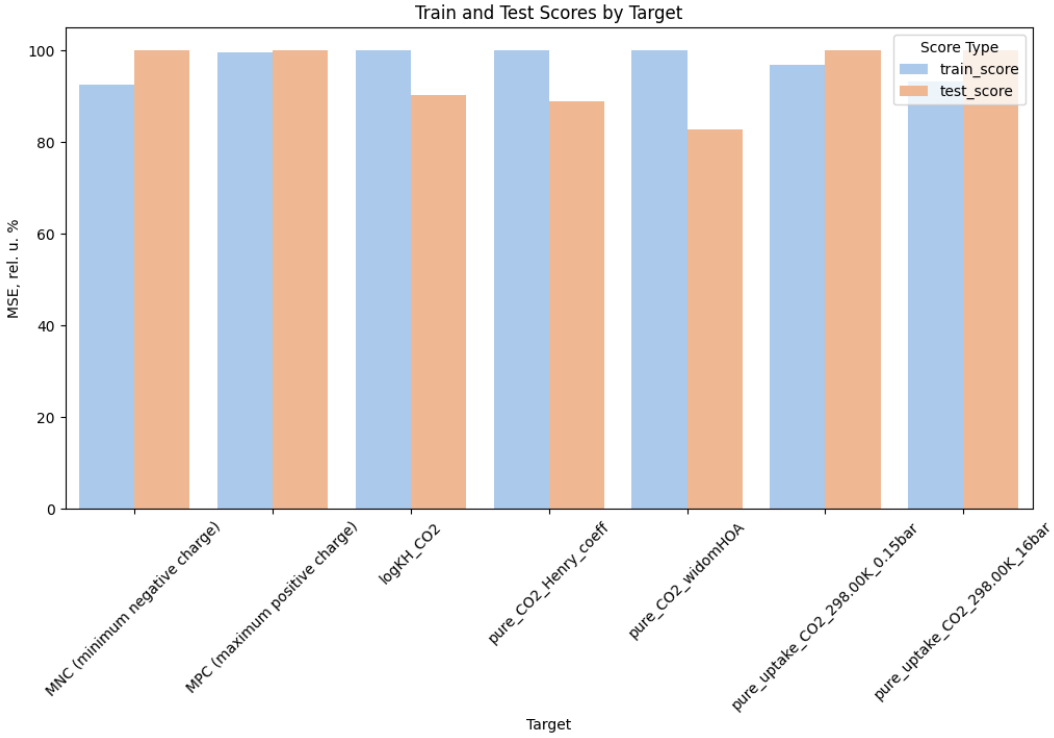

Fig. 8. Performance of the best models on train and test sets (scaled)