DataHow
Coding Challenge

Presented by Adriana Mohap

31 October 2024

Task Description

1. Perform any data analysis you consider relevant.

2. Build any data processing pipeline and model of your choice.

3. Evaluate the performance of your model.

4. Prepare your code so that it can generate predictions and results on a holdout test set (which we will provide before the interview).

Data Description

- X:{name}: Represents measurements of process conditions, which are variables inherent to the process.
- W:{name}: Represents control conditions, which are measurements of the controlled parameters in the process.
- Z:{name}: Represents control setpoints, which are operator-defined control conditions that remain constant throughout the process.
- Y:{name}: Represents process attributes, which are measured at the end of the process.

Task 1: Data Analysis

- Load and preprocess data
- Aggregate descriptive statistics of all variables: mean, standard deviation, min and max values
- Plot the process variable distributions against the target value
- Analyze correlations between all variables

Process Variables (X:):

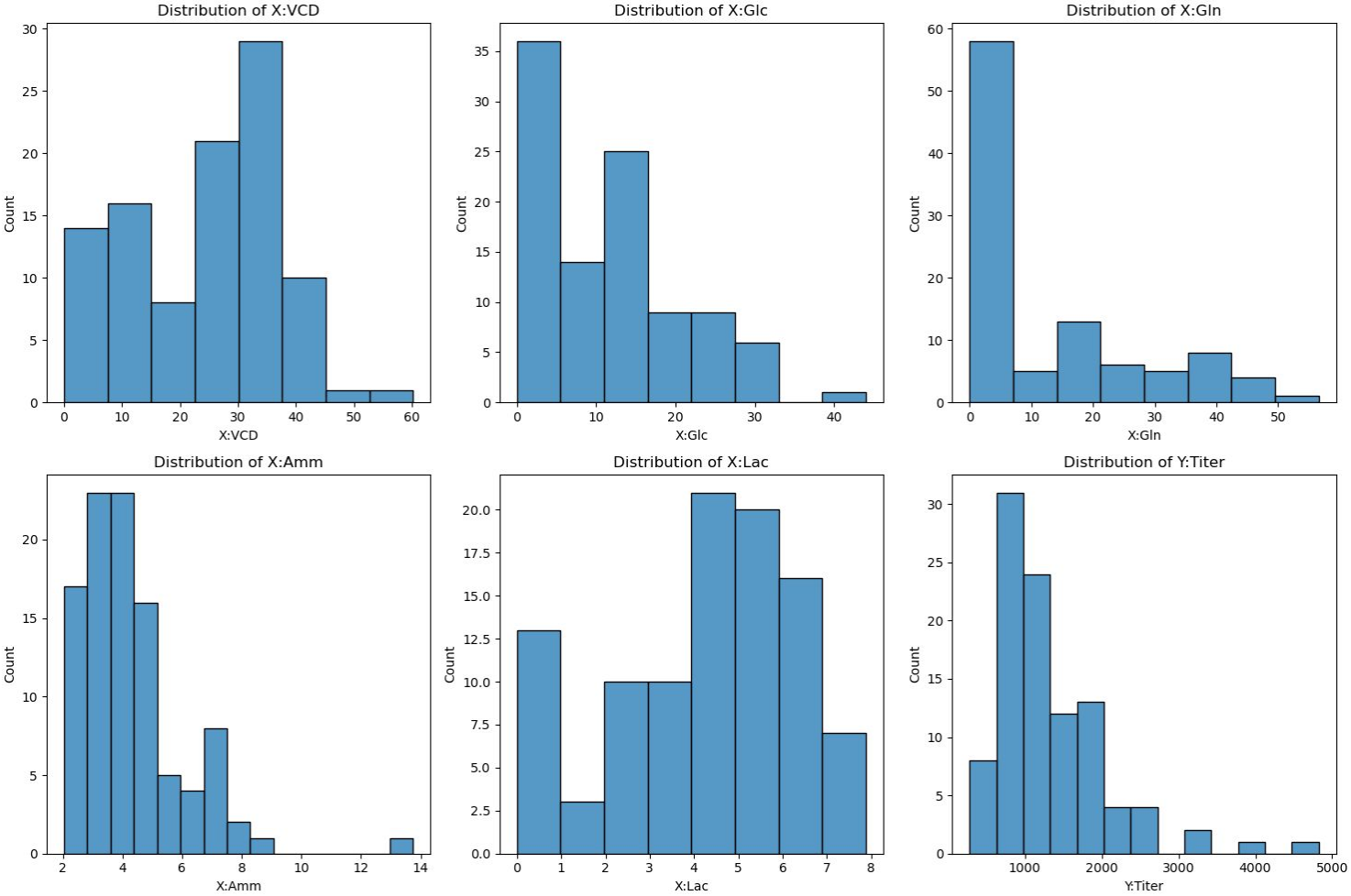
	X:VCD	X:Glc	X:Gln	X:Amm	X:Lac	X:Lysed
count	990.000000	990.000000	990.000000	990.000000	990.000000	990.000000
mean	13.425586	7.210130	7.406499	2.023753	3.271055	0.028822
std	12.377215	6.399089	9.829110	1.809282	2.000817	0.061393
min	0.003421	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.704009	2.910059	0.147580	0.515444	1.580819	0.000000
50%	8.991166	5.133313	4.136096	1.602359	3.488986	0.007234
75%	21.486081	9.738994	9.719193	3.103277	4.843664	0.024610
max	61.242464	44.011310	62.089637	13.744255	8.078131	0.525695

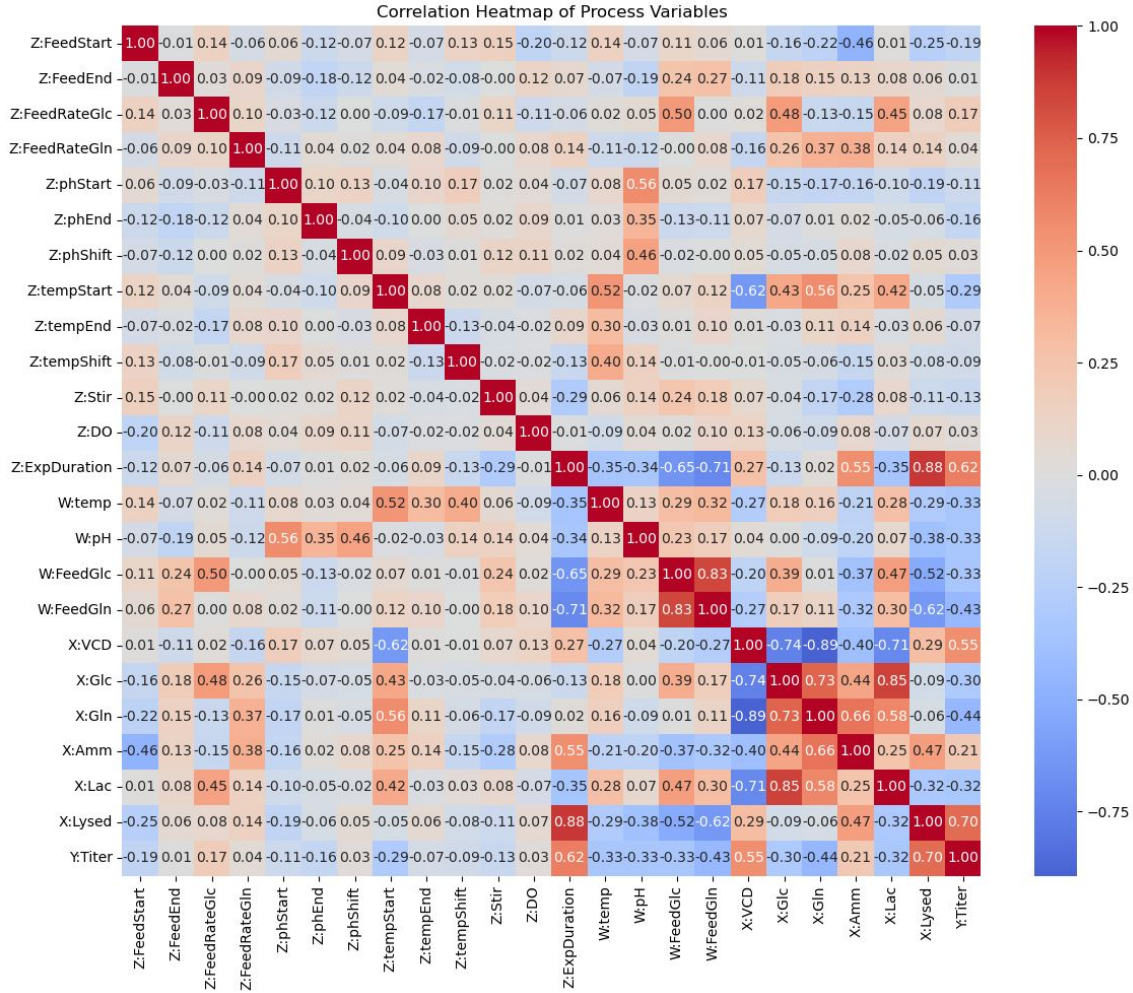
Control Variables (W:):

	W:temp	W:pH	W:FeedGlc	W:FeedGln
count	990.000000	990.000000	990.000000	990.000000
mean	36.883461	6.941144	2.963942	5.250168
std	0.649905	0.326060	1.987881	3.084908
min	35.090909	6.025253	0.000000	0.000000
25%	36.414141	6.696970	0.000000	0.000000
50%	36.878788	6.929293	3.272727	6.696970
75%	37.404040	7.212121	4.606061	7.343434
max	37.989899	7.494949	5.979798	7.989899

Setpoint Variables (Z:):

	Z:FeedStart	Z:FeedEnd	Z:FeedRateGlc	Z:FeedRateGln	Z:phStart	Z:phEnd	Z:phShift	Z:tempStart	Z:tempEnd	Z:tempShift	Z:Stir	Z:DO	Z:ExpDuration
count	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	2.000000	11.000000	4.000000	7.000000	7.000000	6.500000	10.000000	37.000000	36.000000	10.000000	200.000000	55.000000	8.900000
std	0.710669	1.214392	1.154642	0.577321	0.288660	0.288660	2.348436	0.577321	0.577321	2.348436	28.866041	14.433020	2.032563
min	1.000000	9.000000	2.020202	6.010101	6.505051	6.005051	6.000000	36.010101	35.010101	6.000000	150.505051	30.252525	7.000000
25%	1.750000	10.000000	3.020202	6.510101	6.755051	6.255051	8.000000	36.510101	35.510101	8.000000	175.505051	42.752525	7.000000
50%	2.000000	11.000000	4.000000	7.000000	7.000000	6.500000	10.000000	37.000000	36.000000	10.000000	200.000000	55.000000	8.500000
75%	2.250000	12.000000	4.979798	7.489899	7.244949	6.744949	12.000000	37.489899	36.489899	12.000000	224.494949	67.247475	10.000000
max	3.000000	13.000000	5.979798	7.989899	7.494949	6.994949	14.000000	37.989899	36.989899	14.000000	249.494949	79.747475	14.000000


Distribution of X:VCD
Count
X:VCD
Distribution of X:Glc
Count
X:Glc
Distribution of X:Gln
Count
X:Gln
Distribution of X:Amm
Count
X:Amm
Distribution of X:Lac
Count
X:Lac
Distribution of Y:Titer
Count
Y:Titer

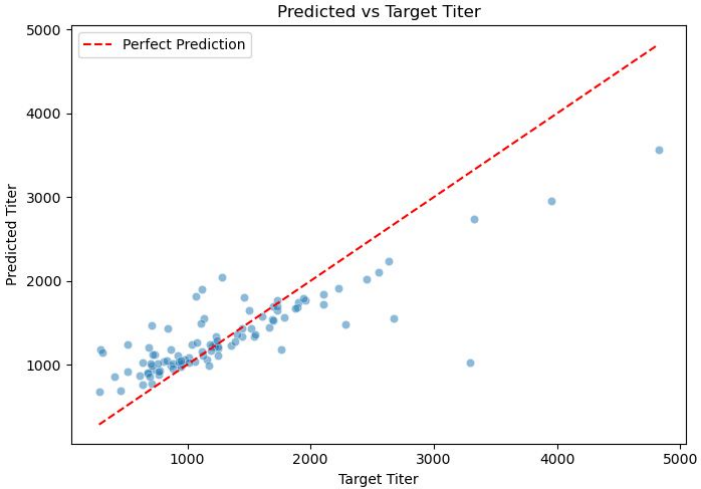

Correlation Heatmap of Process Variables
Z:FeedStart 1.00 -0.01 0.14 -0.06 0.06 -0.12 -0.07 0.12 -0.07 0.13 0.15 -0.20 -0.12 0.14 -0.07 0.11 0.06 0.01 -0.16 -0.22 -0.46 0.01 -0.25 -0.19
Z:FeedEnd -0.01 1.00 0.03 0.09 -0.09 -0.18 -0.12 0.04 -0.02 -0.08 -0.00 0.12 0.07 -0.07 -0.19 0.24 0.27 -0.11 0.18 0.15 0.13 0.08 0.06 0.01
Z:FeedRateGlc 0.14 0.03 1.00 0.10 -0.03 -0.12 0.00 -0.09 -0.17 -0.01 0.11 -0.11 -0.06 0.02 0.05 0.50 0.00 0.02 0.48 -0.13 -0.15 0.45 0.08 0.17
Z:FeedRateGln -0.06 0.09 0.10 1.00 -0.11 0.04 0.02 0.04 0.08 -0.09 -0.00 0.08 0.14 -0.11 -0.12 -0.00 0.08 -0.16 0.26 0.37 0.38 0.14 0.14 0.04
Z:phStart 0.06 -0.09 -0.03 -0.11 1.00 0.10 0.13 -0.04 0.10 0.17 0.02 0.04 -0.07 0.08 0.56 0.05 0.02 0.17 -0.15 -0.17 -0.16 -0.10 -0.19 -0.11
Z:phEnd -0.12 -0.18 -0.12 0.04 0.10 1.00 -0.04 -0.10 0.00 0.05 0.02 0.09 0.01 0.03 0.35 -0.13 -0.11 0.07 -0.07 0.01 0.02 -0.05 -0.06 -0.16
Z:phShift -0.07 -0.12 0.00 0.02 0.13 -0.04 1.00 0.09 -0.03 0.01 0.12 0.11 0.02 0.04 0.46 -0.02 -0.00 0.05 -0.05 -0.05 0.08 -0.02 0.05 0.03
Z:tempStart 0.12 0.04 -0.09 0.04 -0.04 -0.10 0.09 1.00 0.08 0.02 0.02 -0.07 -0.06 0.52 -0.02 0.07 0.12 -0.62 0.43 0.56 0.25 0.42 -0.05 -0.29
Z:tempEnd -0.07 -0.02 -0.17 0.08 0.10 0.00 -0.03 0.08 1.00 -0.13 -0.04 -0.02 0.09 0.30 -0.03 0.01 0.10 0.01 -0.03 0.11 0.14 -0.03 0.06 -0.07
Z:tempShift 0.13 -0.08 -0.01 -0.09 0.17 0.05 0.01 0.02 -0.13 1.00 -0.02 -0.02 -0.13 0.40 0.14 -0.01 -0.00 -0.01 -0.05 -0.06 -0.15 0.03 -0.08 -0.09
Z:Stir 0.15 -0.00 0.11 -0.00 0.02 0.02 0.12 0.02 -0.04 -0.02 1.00 0.04 -0.29 0.06 0.14 0.24 0.18 0.07 -0.04 -0.17 -0.28 0.08 -0.11 -0.13
Z:DO -0.20 0.12 -0.11 0.08 0.04 0.09 0.11 -0.07 -0.02 -0.02 0.04 1.00 -0.01 -0.09 0.04 0.02 0.10 0.13 -0.06 -0.09 0.08 -0.07 0.07 0.03
Z:ExpDuration -0.12 0.07 -0.06 0.14 -0.07 0.01 0.02 -0.06 0.09 -0.13 -0.29 -0.01 1.00 -0.35 -0.34 -0.65 -0.71 0.27 -0.13 0.02 0.55 -0.35 0.88 0.62
W:temp 0.14 -0.07 0.02 -0.11 0.08 0.03 0.04 0.52 0.30 0.40 0.06 -0.09 -0.35 1.00 0.13 0.29 0.32 -0.27 0.18 0.16 -0.21 0.28 -0.29 -0.33
W:pH -0.07 -0.19 0.05 -0.12 0.56 0.35 0.46 -0.02 -0.03 0.14 0.14 0.04 -0.34 0.13 1.00 0.23 0.17 0.04 0.00 -0.09 -0.20 0.07 -0.38 -0.33
W:FeedGlc 0.11 0.24 0.50 -0.00 0.05 -0.13 -0.02 0.07 0.01 -0.01 0.24 0.02 -0.65 0.29 0.23 1.00 0.83 -0.20 0.39 0.01 -0.37 0.47 -0.52 -0.33
W:FeedGln 0.06 0.27 0.00 0.08 0.02 -0.11 -0.00 0.12 0.10 -0.00 0.18 0.10 -0.71 0.32 0.17 0.83 1.00 -0.27 0.17 0.11 -0.32 0.30 -0.62 -0.43
X:VCD 0.01 -0.11 0.02 -0.16 0.17 0.07 0.05 -0.62 0.01 -0.01 0.07 0.13 0.27 -0.27 0.04 -0.20 -0.27 1.00 -0.74 -0.89 -0.40 -0.71 0.29 0.55
X:Glc -0.16 0.18 0.48 0.26 -0.15 -0.07 -0.05 0.43 -0.03 -0.05 -0.04 -0.06 -0.13 0.18 0.00 0.39 0.17 -0.74 1.00 0.73 0.44 0.85 -0.09 -0.30
X:Gln -0.22 0.15 -0.13 0.37 -0.17 0.01 -0.05 0.56 0.11 -0.06 -0.17 -0.09 0.02 0.16 -0.09 0.01 0.11 -0.89 0.73 1.00 0.66 0.58 -0.06 -0.44
X:Amm -0.46 0.13 -0.15 0.38 -0.16 0.02 0.08 0.25 0.14 -0.15 -0.28 0.08 0.55 -0.21 -0.20 -0.37 -0.32 -0.40 0.44 0.66 1.00 0.25 0.47 0.21
X:Lac 0.01 0.08 0.45 0.14 -0.10 -0.05 -0.02 0.42 -0.03 0.03 0.08 -0.07 -0.35 0.28 0.07 0.47 0.30 -0.71 0.85 0.58 0.25 1.00 -0.32 -0.32
X:Lysed -0.25 0.06 0.08 0.14 -0.19 -0.06 0.05 -0.05 0.06 -0.08 -0.11 0.07 0.88 -0.29 -0.38 -0.52 -0.62 0.29 -0.09 -0.06 0.47 -0.32 1.00 0.70
Y:Titer -0.19 0.01 0.17 0.04 -0.11 -0.16 0.03 -0.29 -0.07 -0.09 -0.13 0.03 0.62 -0.33 -0.33 -0.33 -0.43 0.55 -0.30 -0.44 0.21 -0.32 0.70 1.00
Z:FeedStart
Z:FeedEnd
Z:FeedRateGlc
Z:FeedRateGln
Z:phStart
Z:phEnd
Z:phShift
Z:tempStart
Z:tempEnd
Z:tempShift
Z:Stir
Z:DO
Z:ExpDuration
W:temp
W:pH
W:FeedGlc
W:FeedGln
X:VCD
X:Glc
X:Gln
X:Amm
X:Lac
X:Lysed
Y:Titer
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

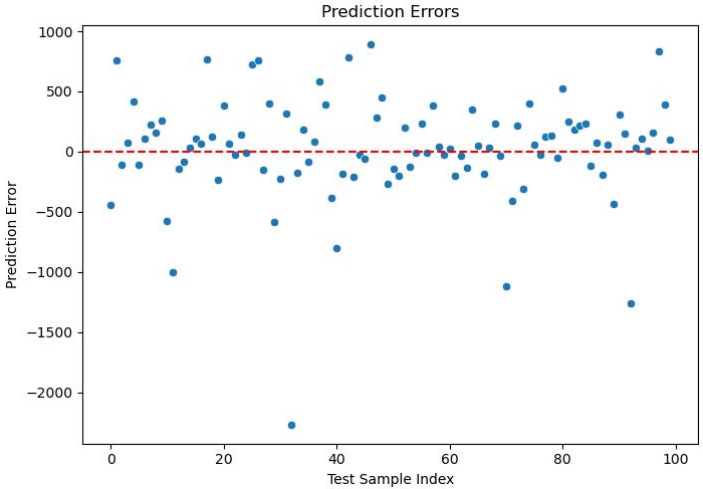
Task 2: Build Data Processing Pipeline and Model

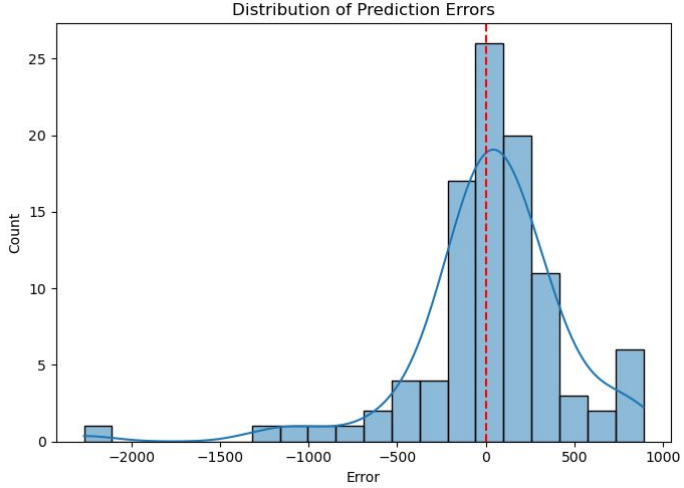
- Build two functions for feature engineering:
  - engineer_features(process_data): **X variables**: mean, std, min, max; **W variables**: mean only; **Z variables**: initial values only
  - engineer_features_baseline(process_data): Baseline feature engineering using **only mean values of X** variables
- Random forest regression as model
  - no assumptions about the underlying distribution of the data
  - works well with high dimensional data and collinearity in features

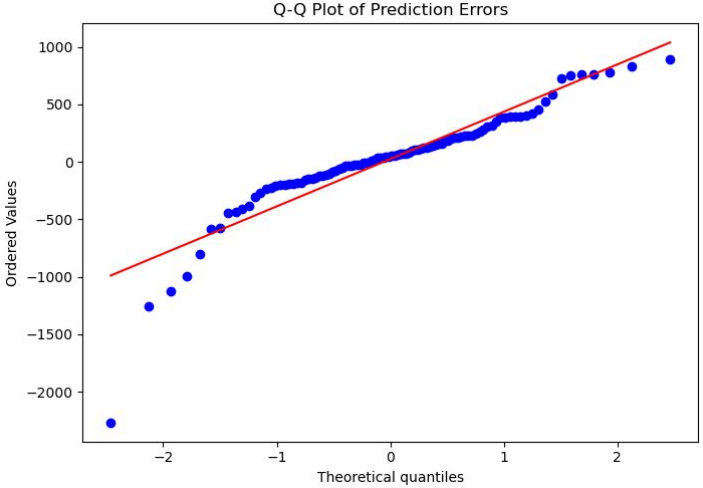
Task 3: Evaluate Performance

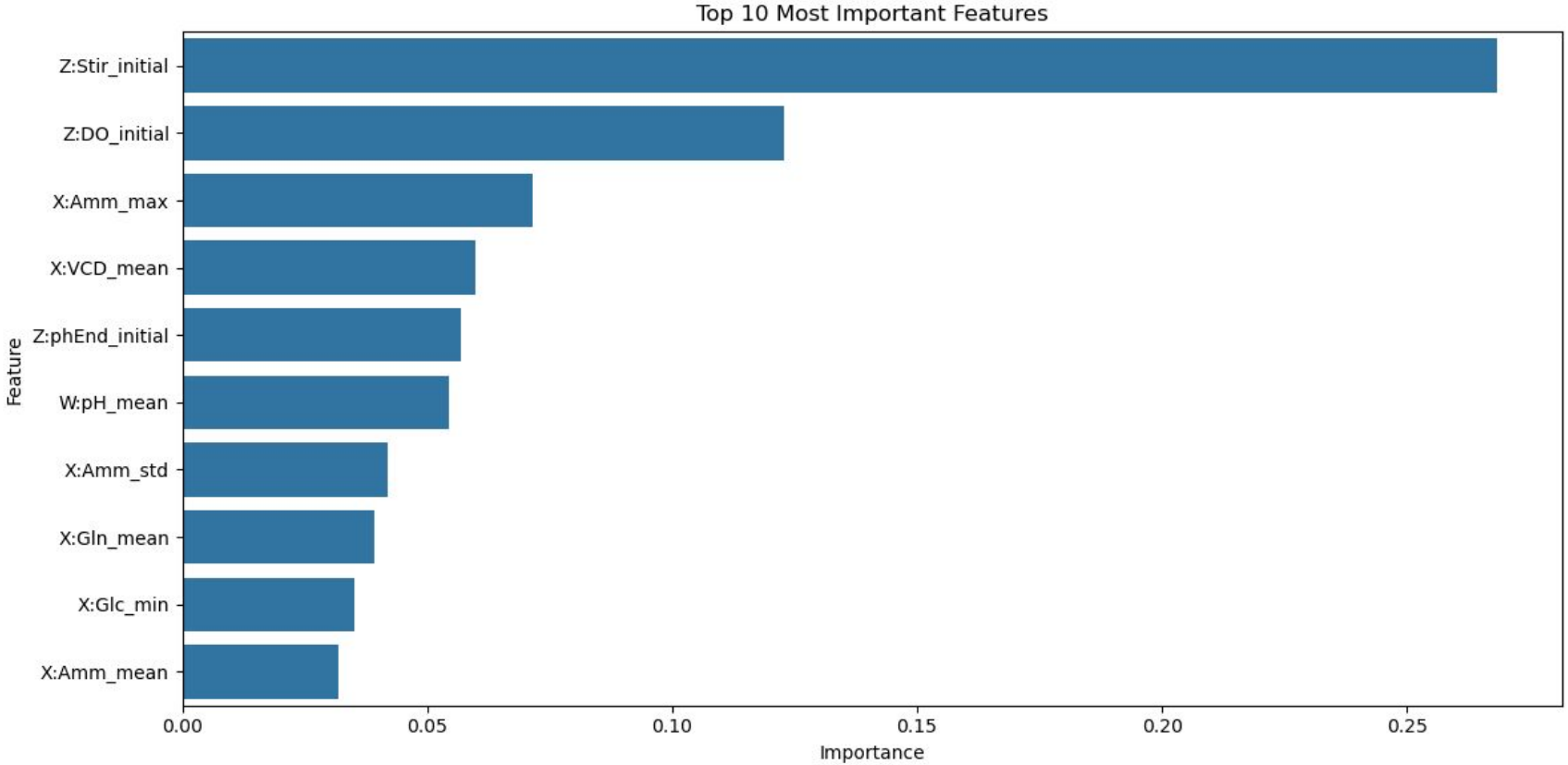
- Metrics over the Training and Test set:
  - R Squared
  - RMSE
  - MAE
  - MAPE
- Visual Inspection of the predictions and the residuals


Predicted vs Target Titer
Perfect Prediction
Predicted Titer
Target Titer
5000
4000
3000
2000
1000
1000
2000
3000
4000
5000


Prediction Errors
Prediction Error
Test Sample Index
1000
500
0
−500
−1000
−1500
−2000
0
20
40
60
80
100


Distribution of Prediction Errors
Count
Error
25
20
15
10
5
0
−2000
−1500
−1000
−500
0
500
1000


Q-Q Plot of Prediction Errors
Ordered Values
Theoretical quantiles
1000
500
0
−500
−1000
−1500
−2000
−2
−1
0
1
2


Top 10 Most Important Features
Z:Stir_initial
Z:DO_initial
X:Amm_max
X:VCD_mean
Z:phEnd_initial
W:pH_mean
X:Amm_std
X:Gln_mean
X:Glc_min
X:Amm_mean
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Importance

Future Work

- Gain domain knowledge about the process variables in order to adjust the selected features
- Create features that accurately reflect their relationships (interaction terms, growth rates etc.)
- Drop features that show high correlations
- Perform hyperparameter tuning; Cross-validation
- Use time variable as separate variable
- Select a model, which is able to capture time series data more accurately (for example LSTM or Recurrent networks) or use ensemble techniques