

Discoveries and Lessons from the IMPRESS Project

Presented by Illia Rohalskyi

Introducing IMPRESS

- Problem statement
- Why do IMPRESS?
- IMPRESS as a baseline ML project for soft sensors

Approach & Agenda

Load and clean the data

- Get to know the dataset
- How the data was cleaned

Extract statistical features

- What features were extracted and how

Build tree-based models and ensemble them

- Experiments
- Key takeaways

The Dataset

- Online data
- Offline data
- Cleaning process

	aktivsauerstoff	anionischetenside	bsb	carbonathaerte	csb	leitfaehigkeit	nichtionischetenside	oberflaechenspannung	...
0	20.0	27.30	0.0	21.5	2426.0	1105.0	215.0	39.90	...
1	20.0	27.30	0.0	21.5	2426.0	1105.0	215.0	39.90	...
2	20.0	27.30	0.0	21.5	2426.0	1105.0	215.0	39.90	...
3	10.0	1.88	0.0	6.8	29.5	30.0	12.0	61.55	...
4	50.0	260.00	0.0	21.9	3792.0	1100.0	660.0	31.85	...

5 rows × 17 columns

	experimentnummer	timestamp	waschen	spuelen	csbeq	truebung	druck1	druck2	...
4	1.0	1.624966e+09	0	1.0	45.758366	7.649211	0.141451	0.165737	...
5	1.0	1.624966e+09	0	1.0	45.711973	7.676346	0.140628	0.165041	...
6	1.0	1.624966e+09	0	1.0	45.801190	7.839979	0.141379	0.164814	...
7	1.0	1.624966e+09	0	1.0	45.758366	7.852313	0.140800	0.165357	...
8	1.0	1.624966e+09	0	1.0	45.715542	7.893427	0.139941	0.164000	...

5 rows × 21 columns

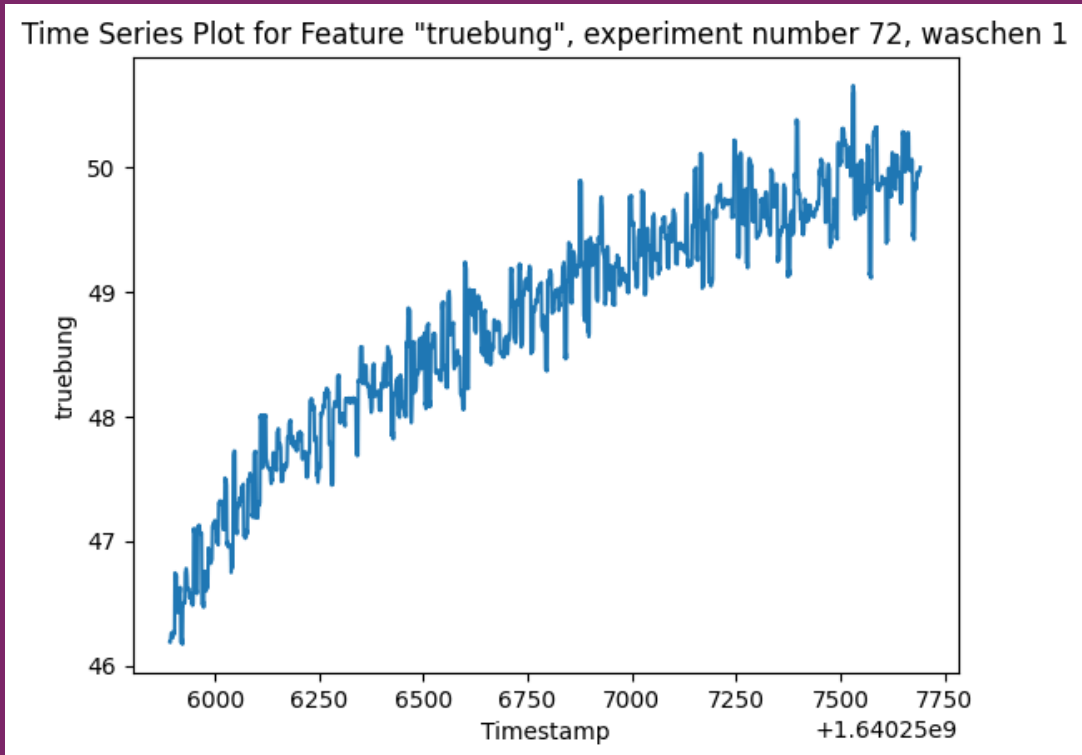
Problem: Data Points with Different Measurements

```
offline_df[(offline_df['experimentnummer'] == 17) & (offline_df['bemerkungen'] == 'W1')]
```

✓ 0.0s Python




it	nichtionischentenside	oberflaechenspannung	...	ph	truebung	wasserstoffperoxid	bemerkungen	timestamp_probeentnahme	timestamp_messung	v
.5	620.0	34.10	...	7.205	38.45	32.70	W1	1618231882	1618231882	
.5	620.0	34.10	...	7.205	38.45	32.70	W1	1618220722	1618220700	
.0	95.0	47.80	...	7.465	4.70	7.25	W1	1618220722	1618220700	
.0	620.0	36.95	...	-1.000	-1.00	-1.00	W1	1627288200	1627300800	

Feature Extraction

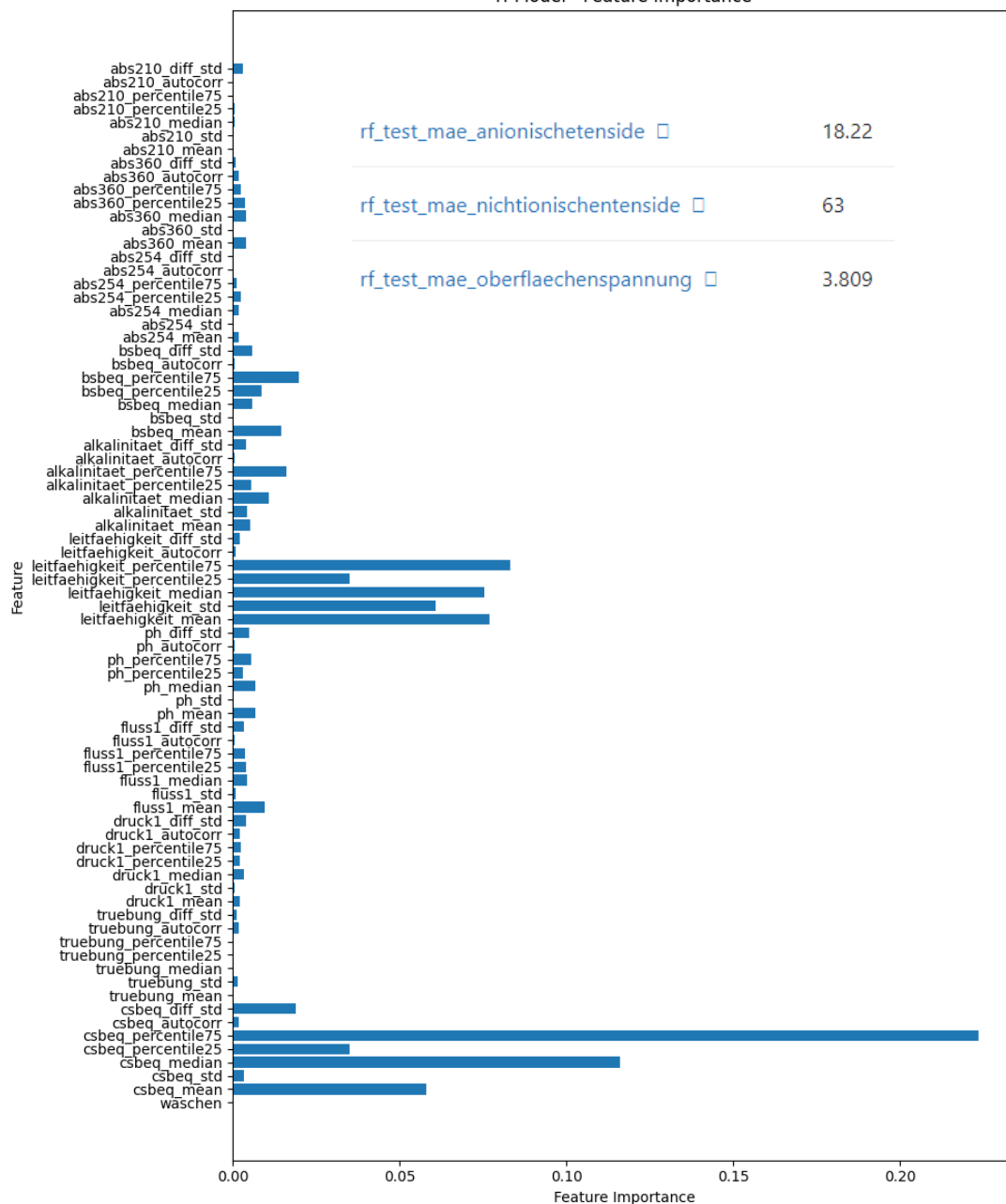


- **Mean, Median, 25%, 75%, Standard Deviation**
- **Autocorrelation, Difference in STD tried but did not help**

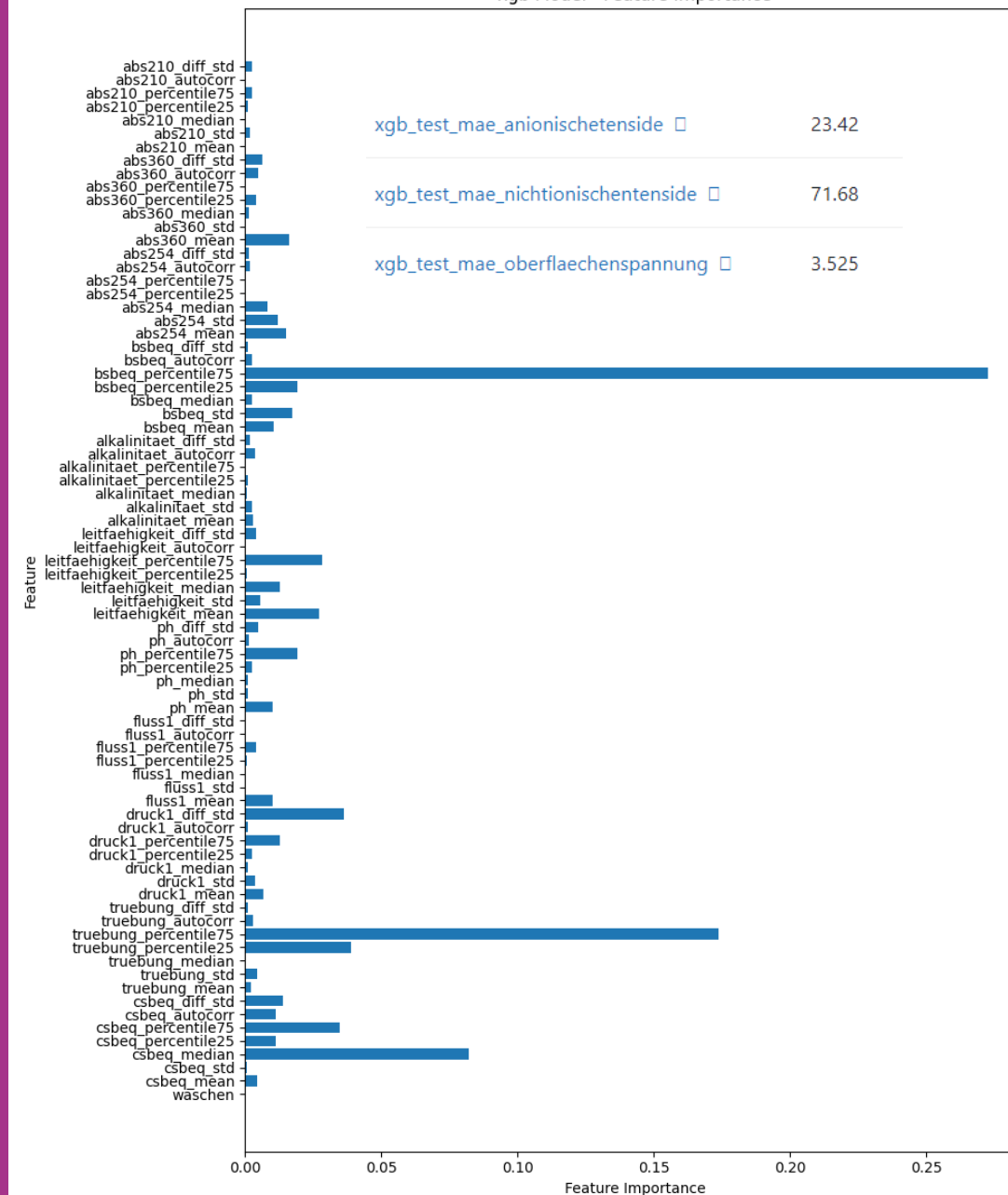
Experiment 1: initial model training results

Name	Value
ensemble_test_mae_anionischetenside 	19.8
ensemble_test_mae_nichtionischetenside 	58.99
ensemble_test_mae_oberflaechenspannung 	3.583




rf Model - Feature Importance

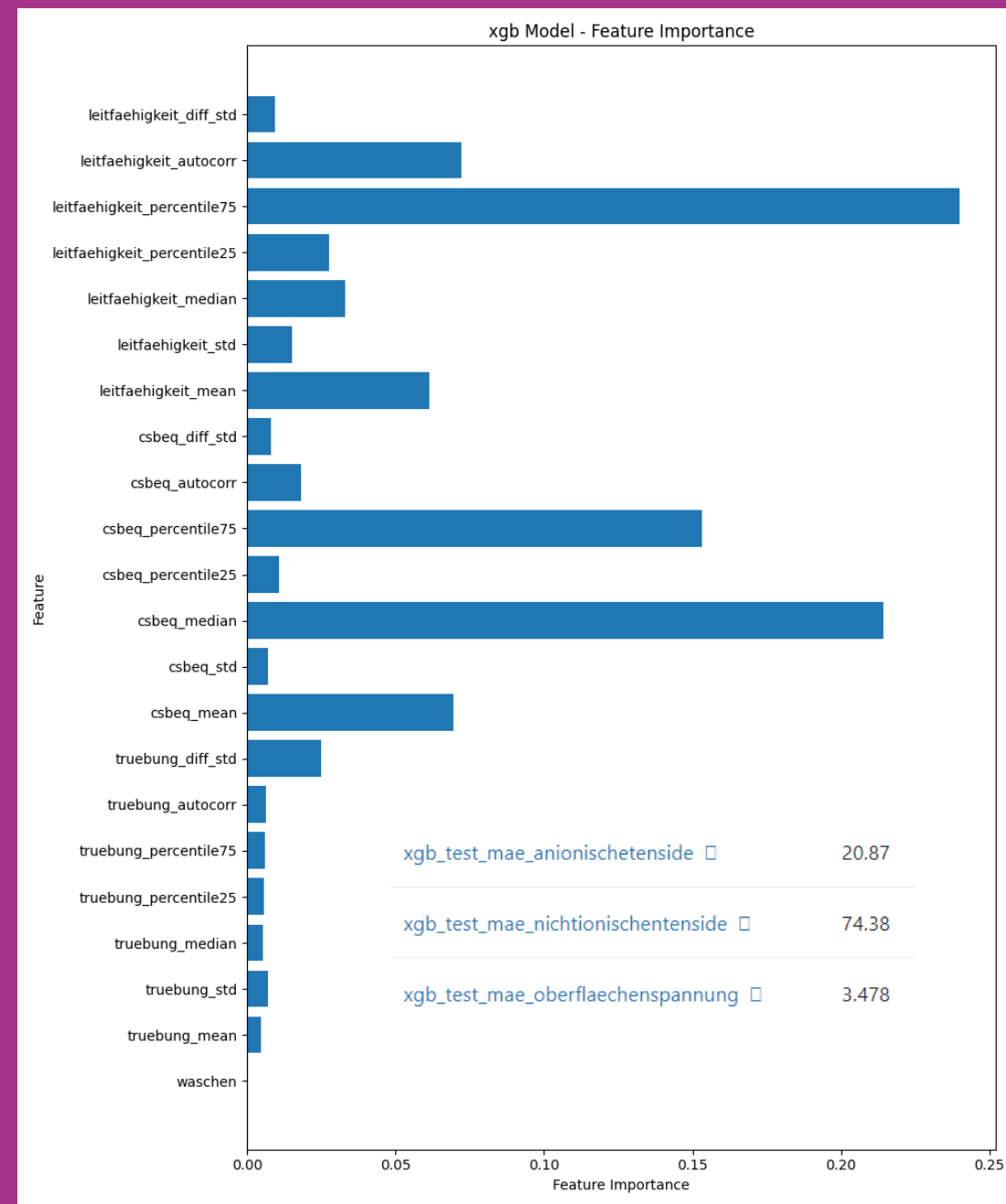
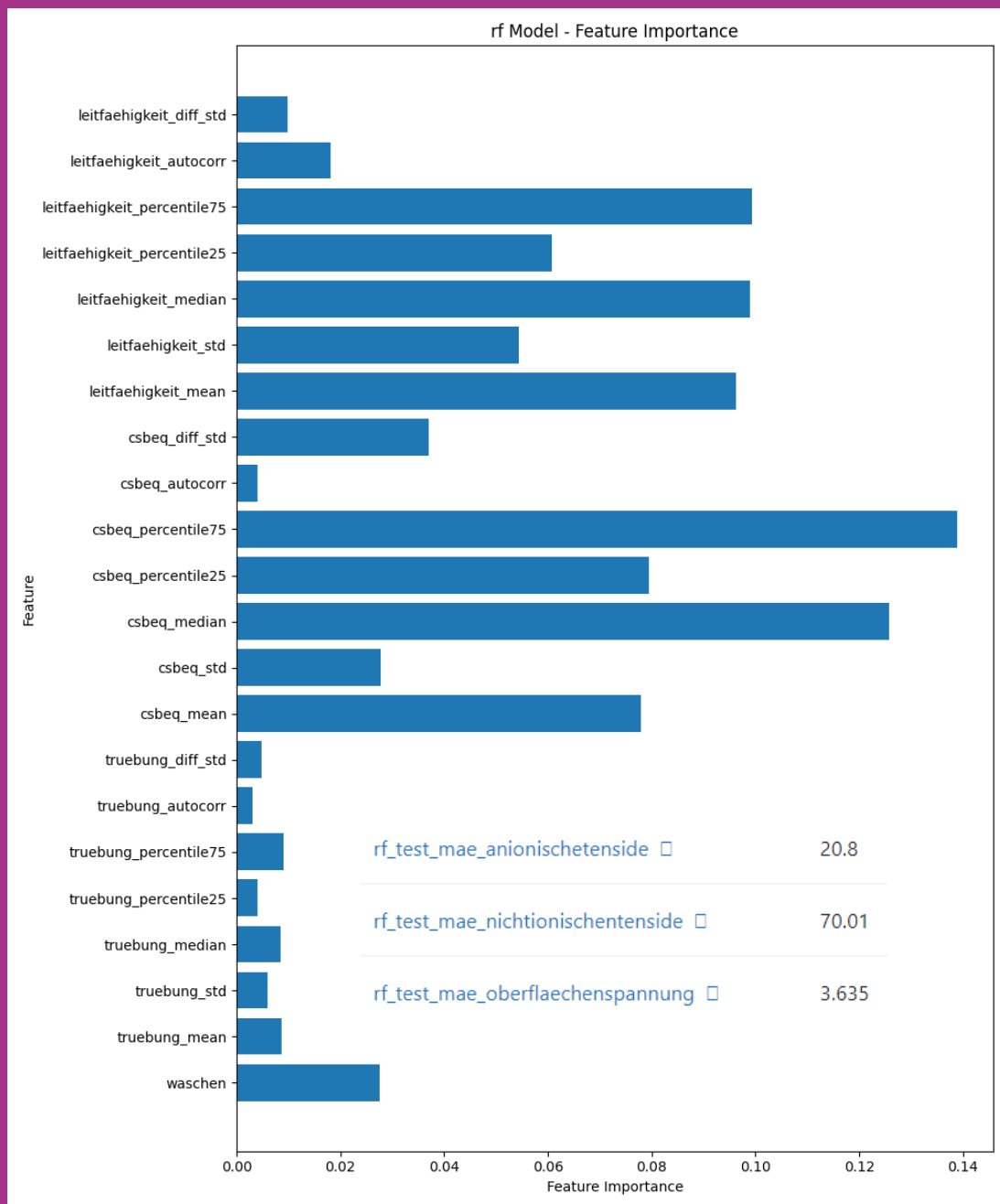


xgb Model - Feature Importance



Experiment 2: Most important features




Name	Value
ensemble_test_mae_anionischetenside 	20.82
ensemble_test_mae_nichtionischetenside 	67.2
ensemble_test_mae_oberflaechenspannung 	3.558

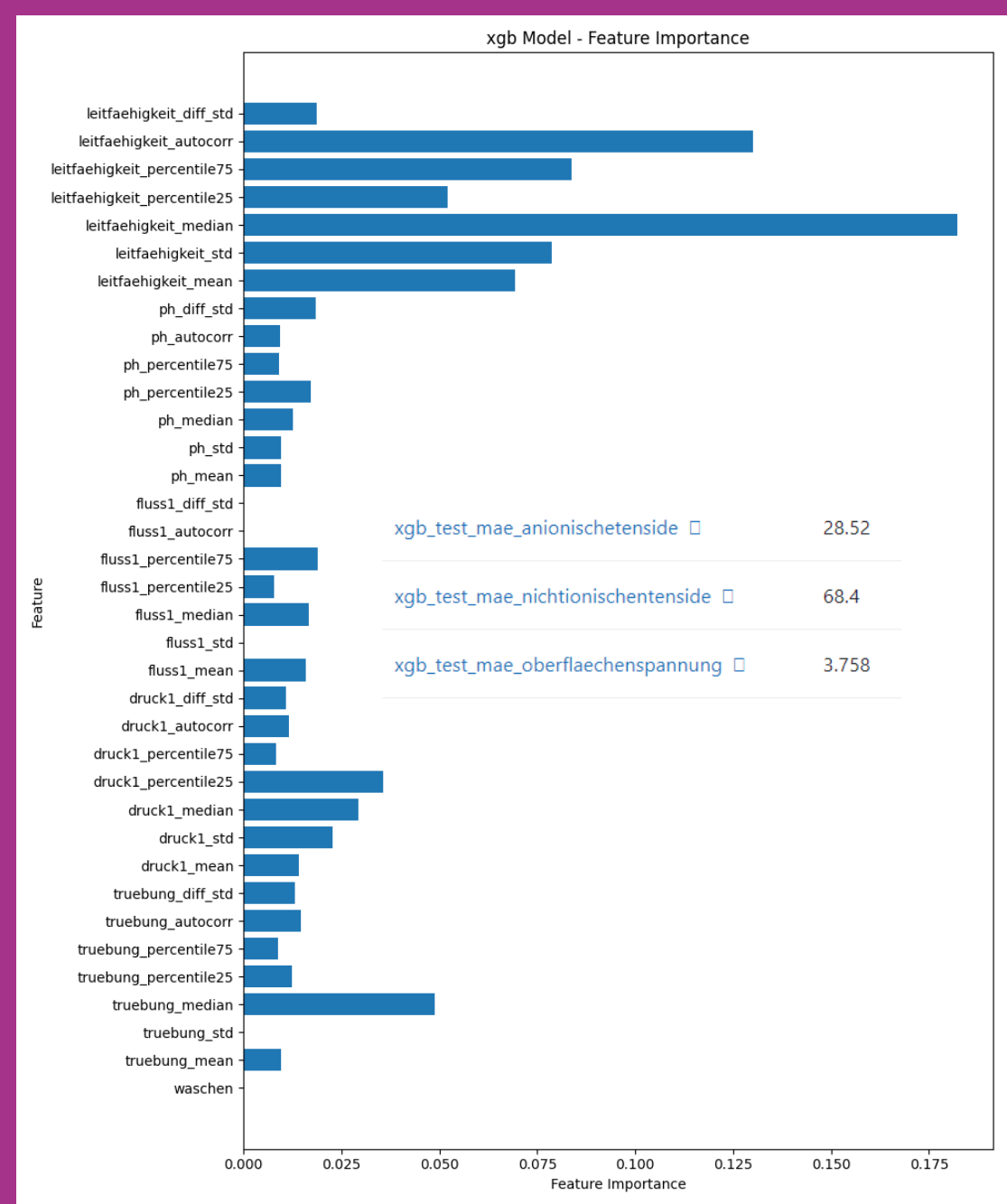
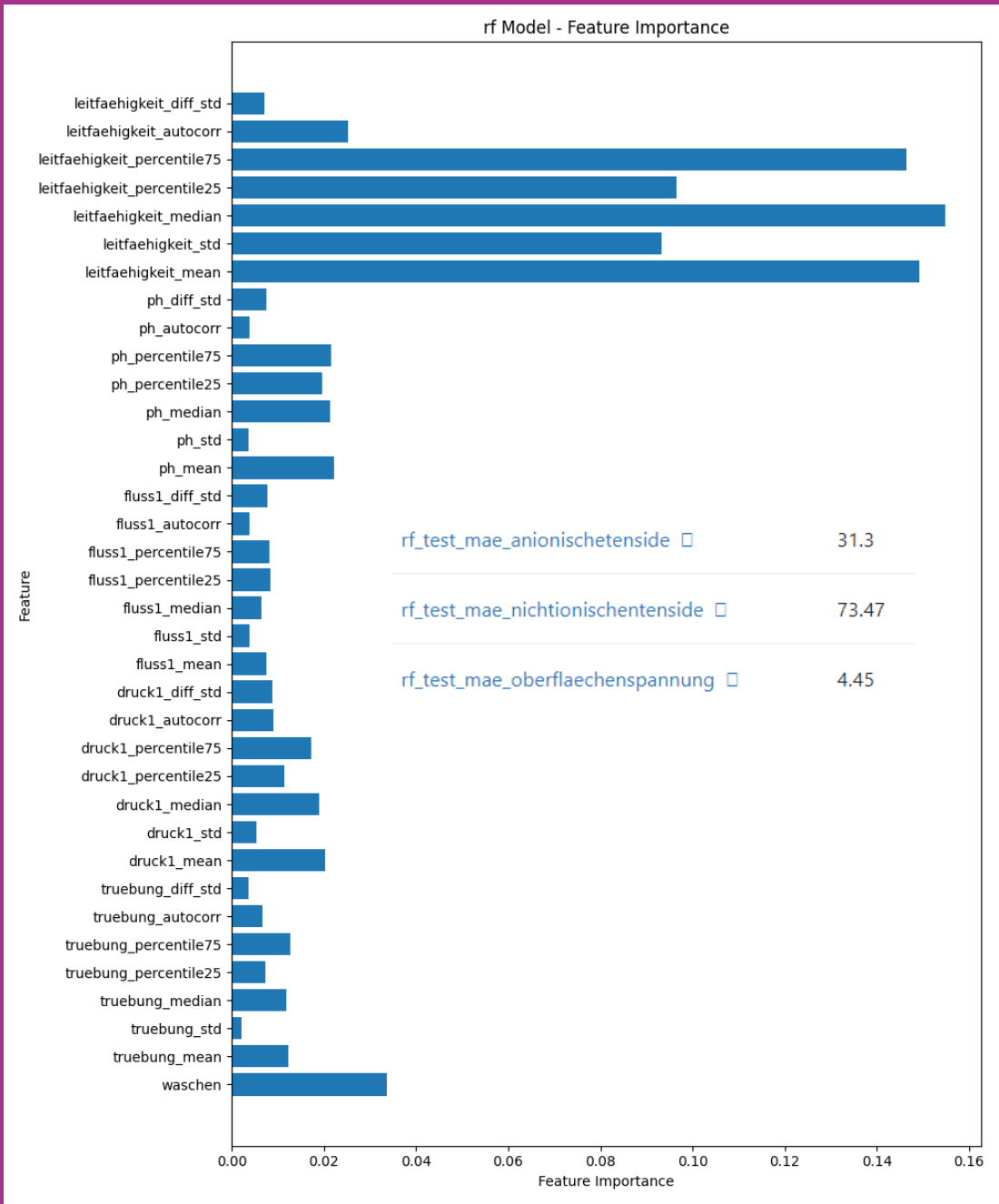


Experiment 3:




Cheapest features to

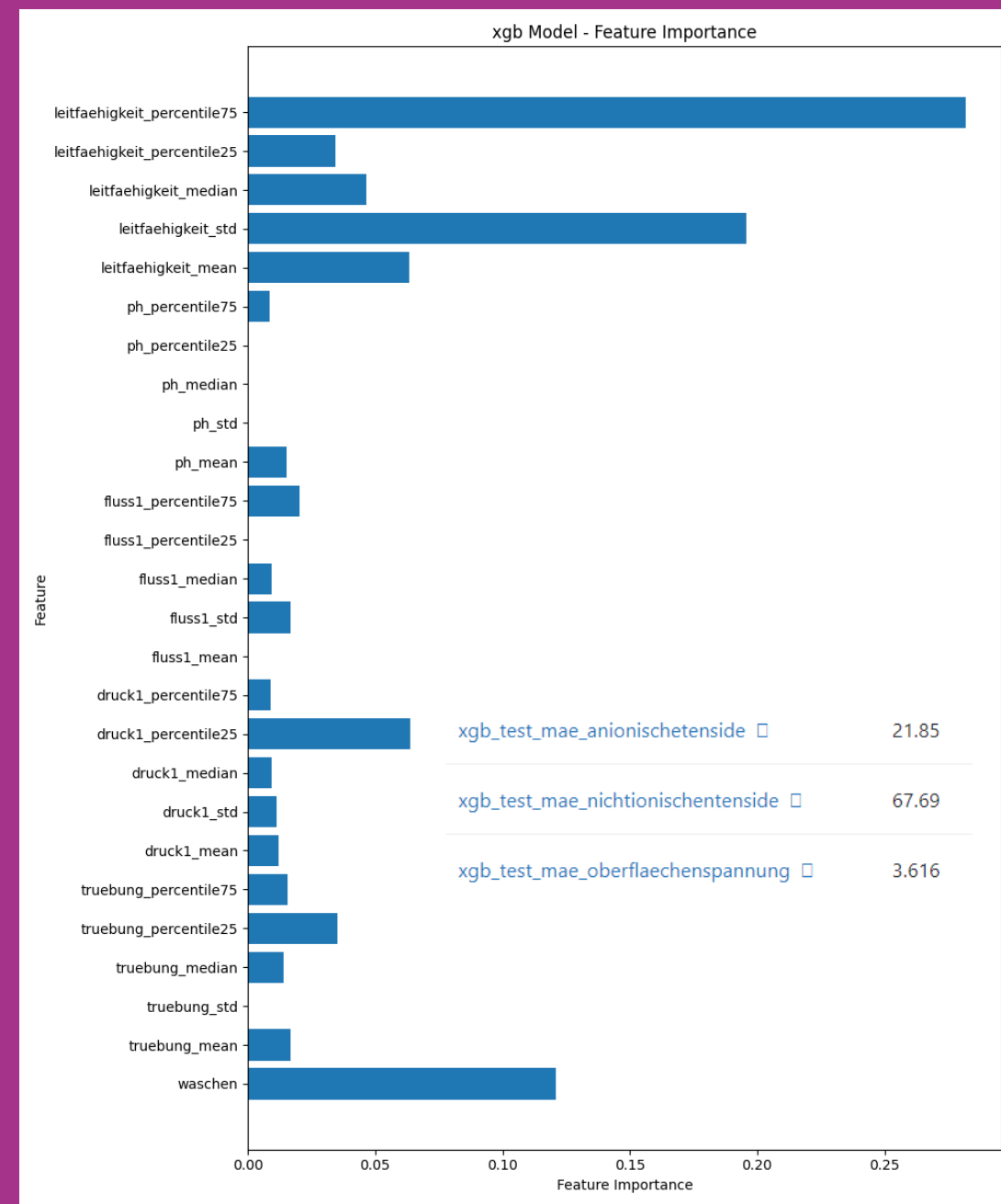
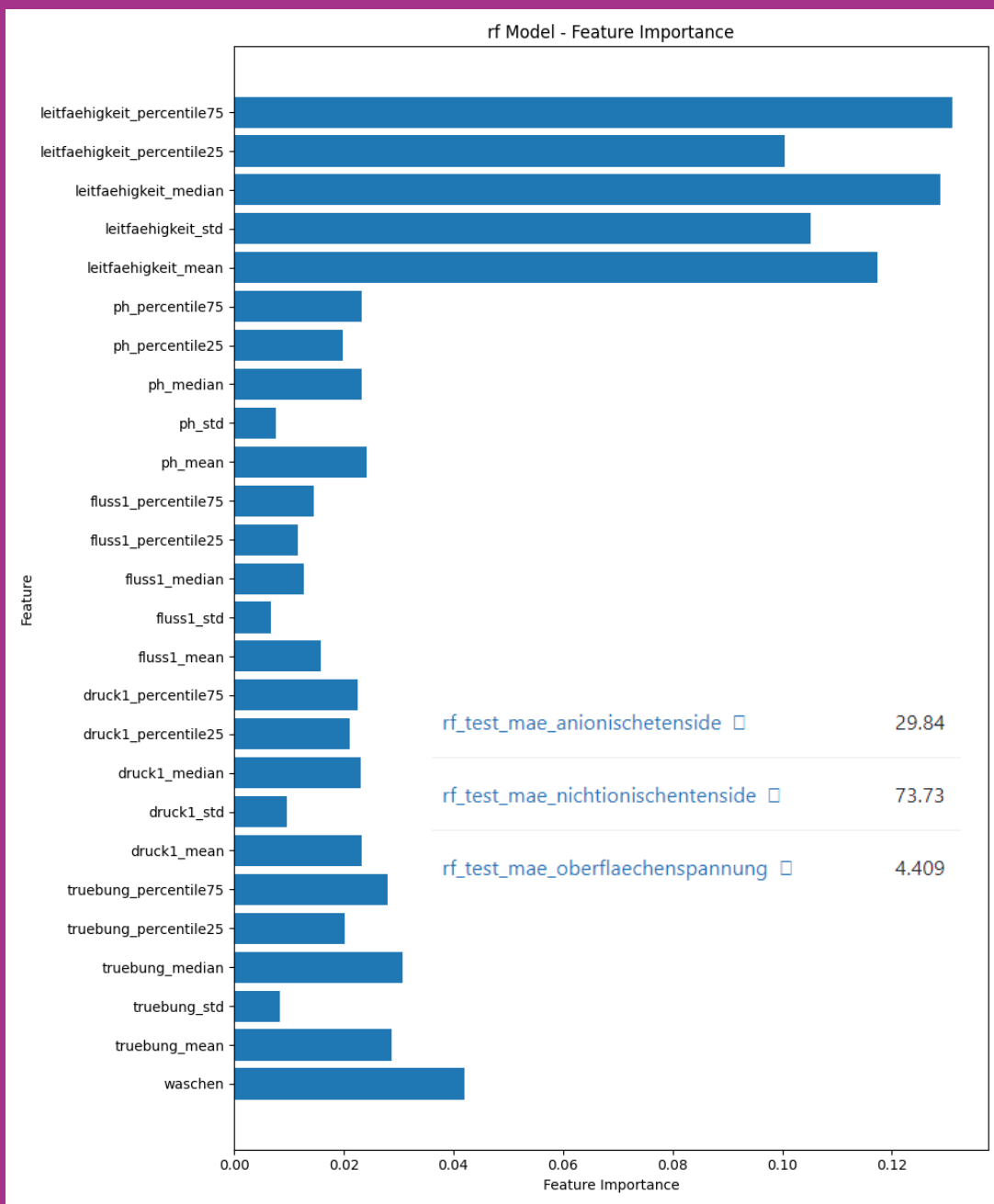
measure

Name	Value
ensemble_test_mae_anionischetenside 	24.73
ensemble_test_mae_nichtionischetenside 	65.68
ensemble_test_mae_oberflaechenspannung 	3.98






Experiment 4: Experiment 3 Without Autocorr and Diff STD

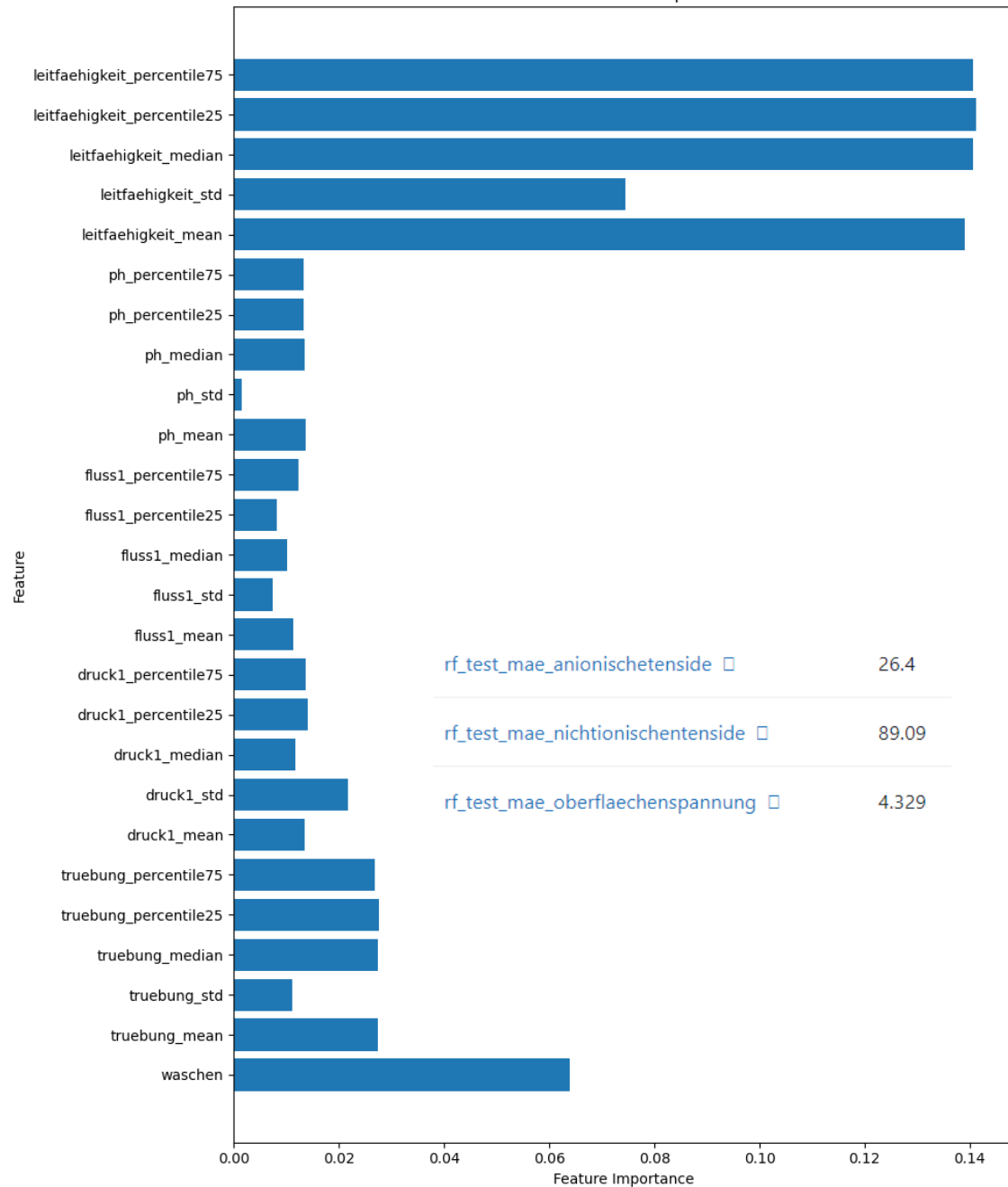
Name	Value
ensemble_test_mae_anionischetenside 	20.82
ensemble_test_mae_nichtionischetenside 	67.2
ensemble_test_mae_oberflaechenspannung 	3.558



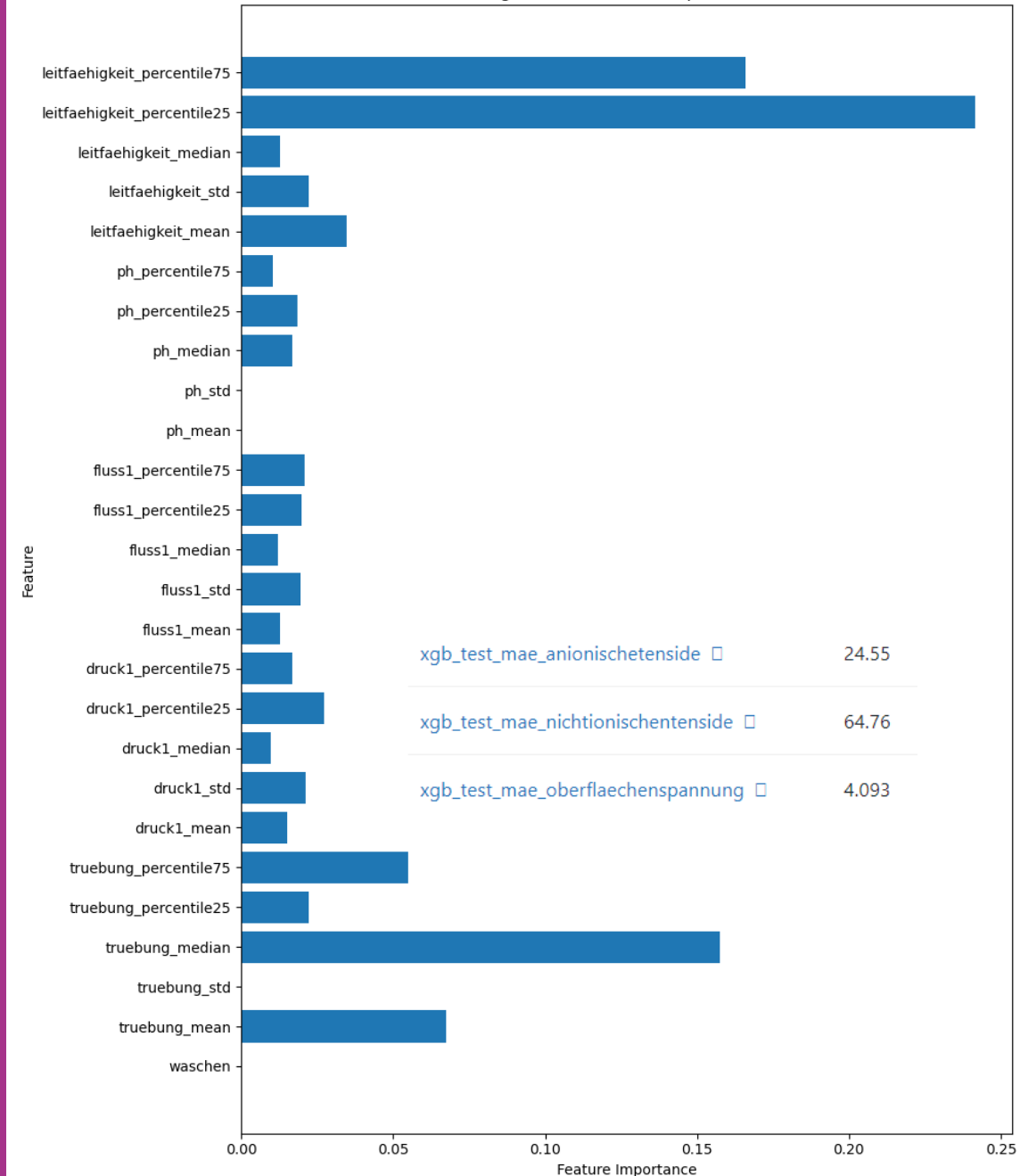
Experiment 5: Experiment 4 last 700 data points

Name	Value
ensemble_test_mae_anionischetenside 	24.66
ensemble_test_mae_nichtionischetenside 	73.39
ensemble_test_mae_oberflaechenspannung 	4.191

rf Model - Feature Importance



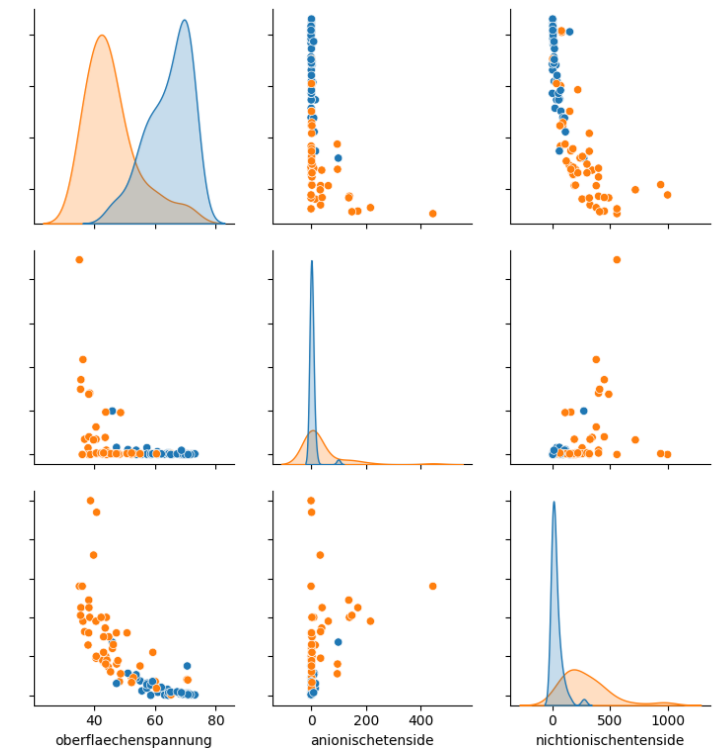
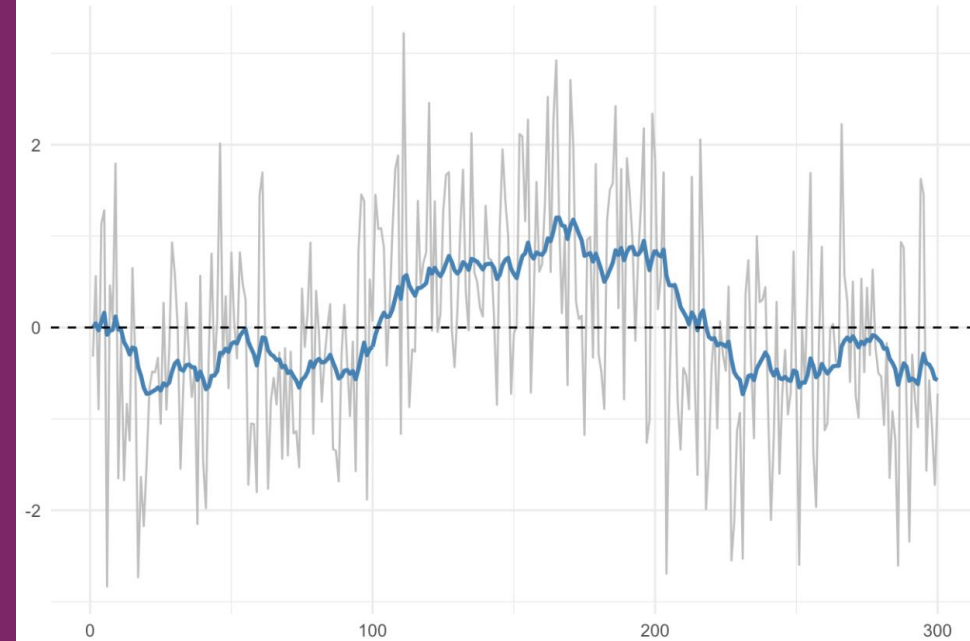
xgb Model - Feature Importance



OUTLOOK

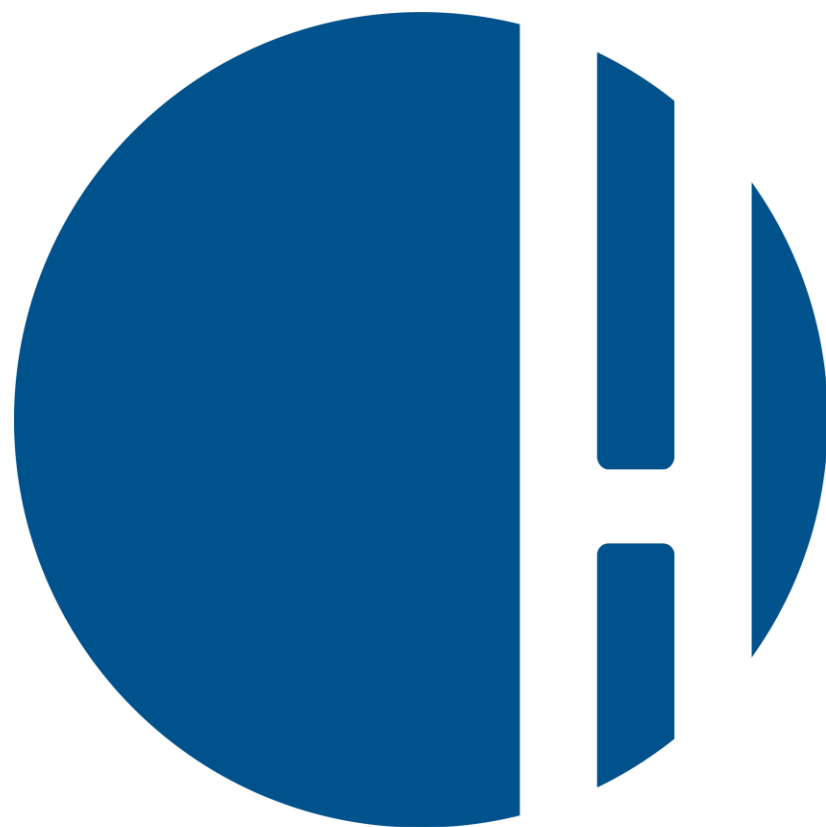
What will be done next

- Smoothing filter for time series
- Rinsing (blue) vs washing (orange) water quality classification problem



Q&A

**Thank you for
your attention!**



HOHENSTEIN