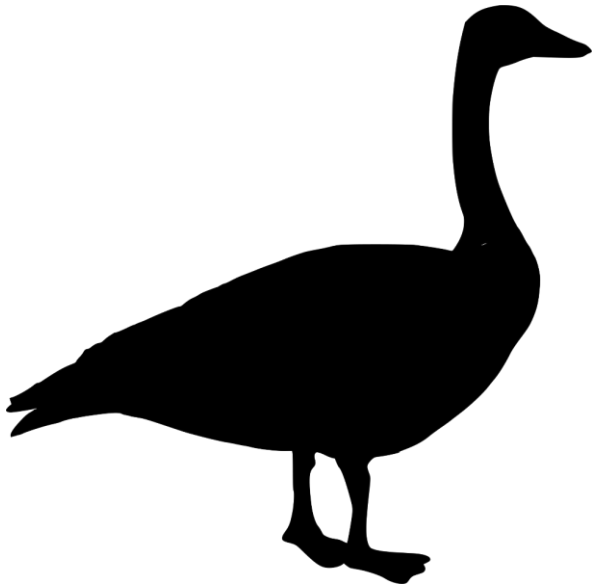
Презентация решения Хакатон Be Coder Направление DataScience



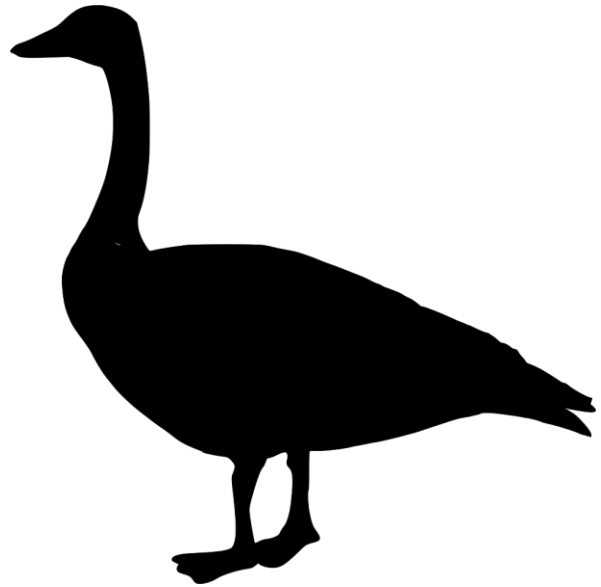
Команда DataGoose:

Елховская Любовь

Иванова Юлия

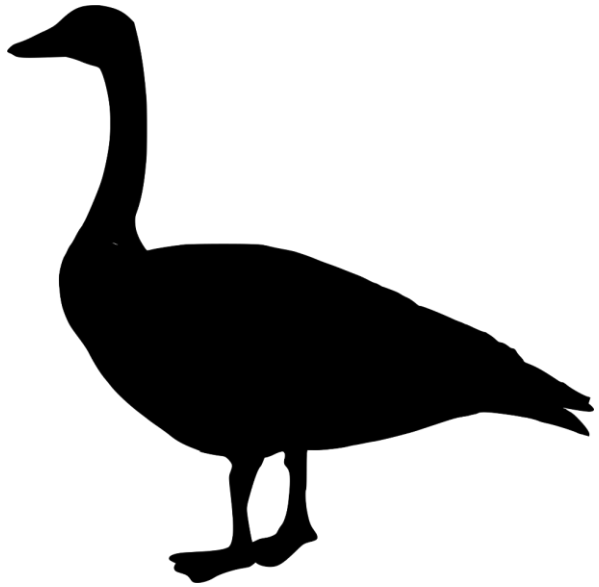
Нагаев Александр

Огородников Владимир

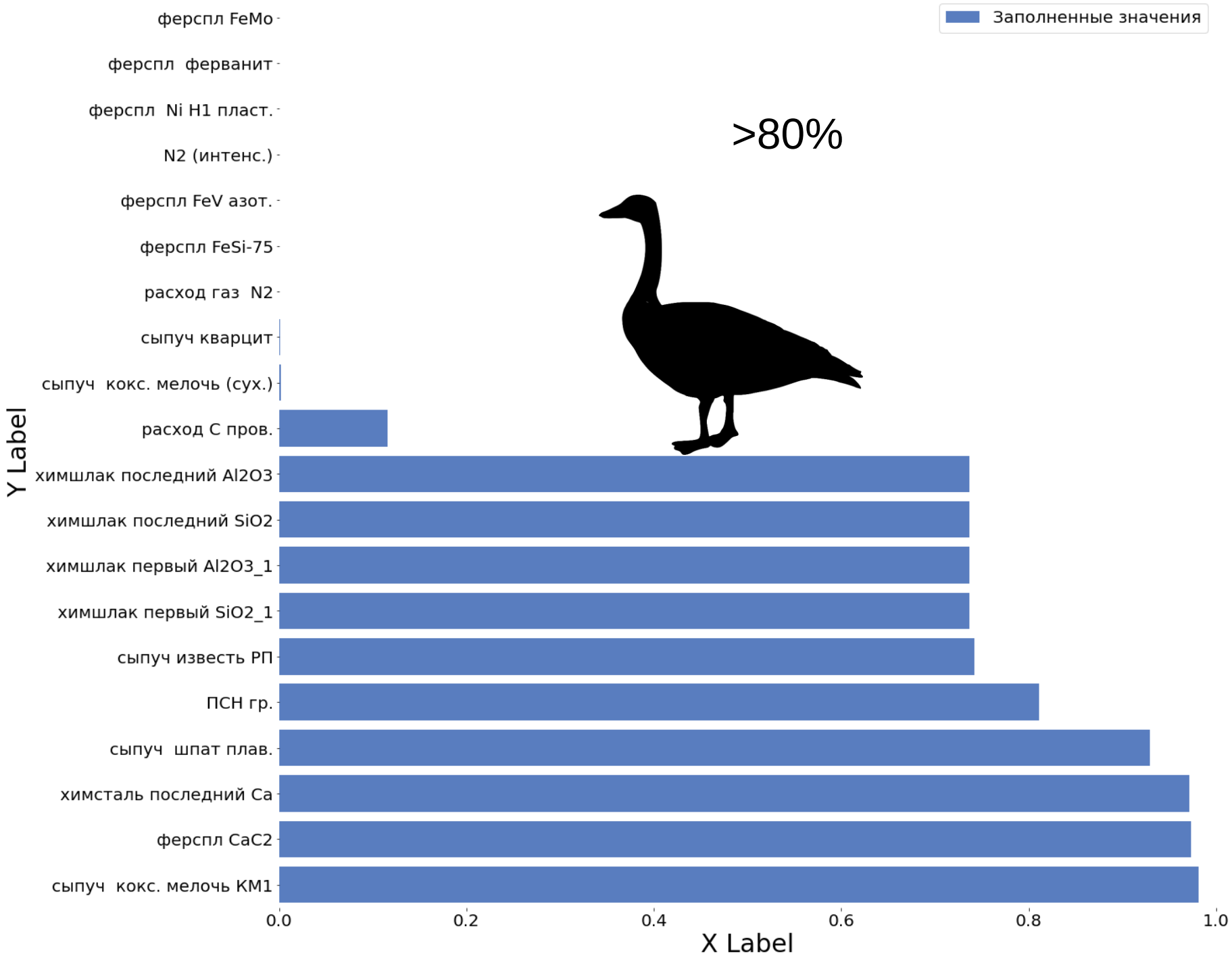


Первый этап: EDA

- Анализ пропусков в данных и стратегии их заполнения
- Исследование параметров (фич):
- Категориальные фичи
- Порядковые фичи
- Численные фичи
- Анализ и чистка выбросов
- Анализ дисперсии в данных
- Анализ корреляции фич
- Создание новых фич



Пропуски

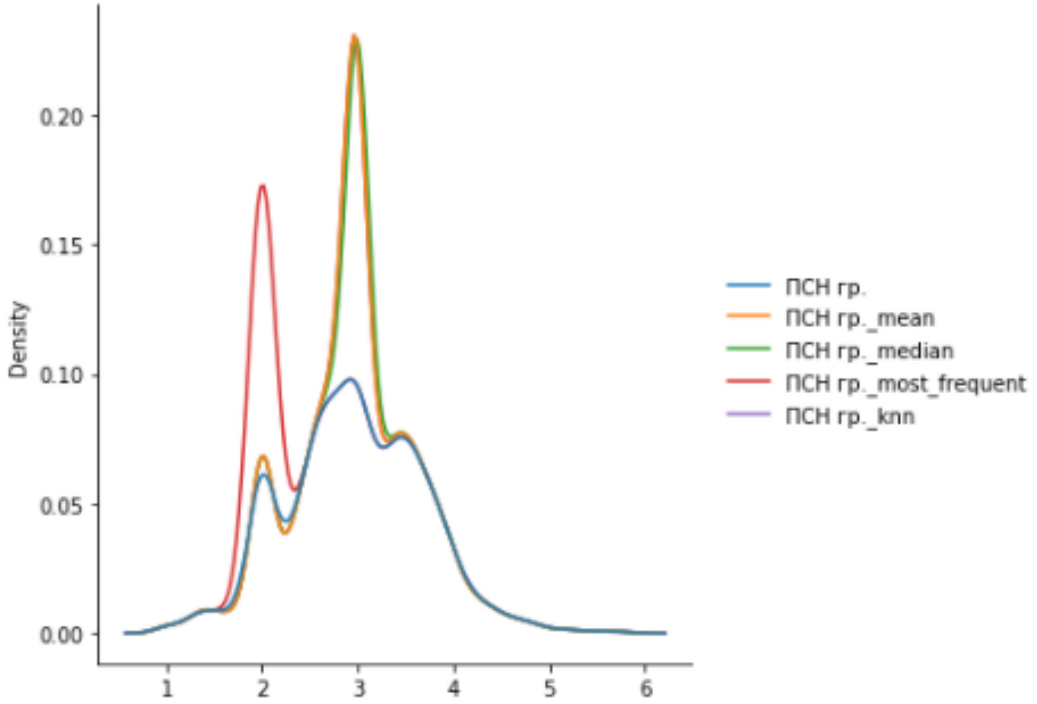


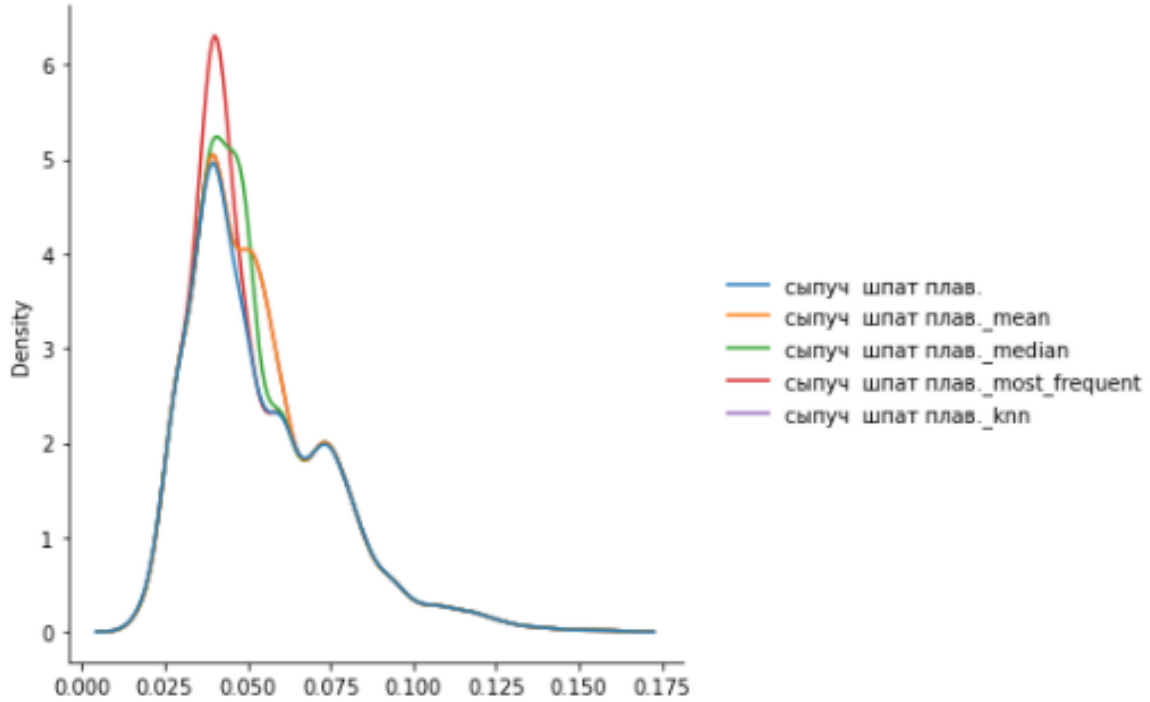
>80%

Самая популярная марка - Э76ХФ

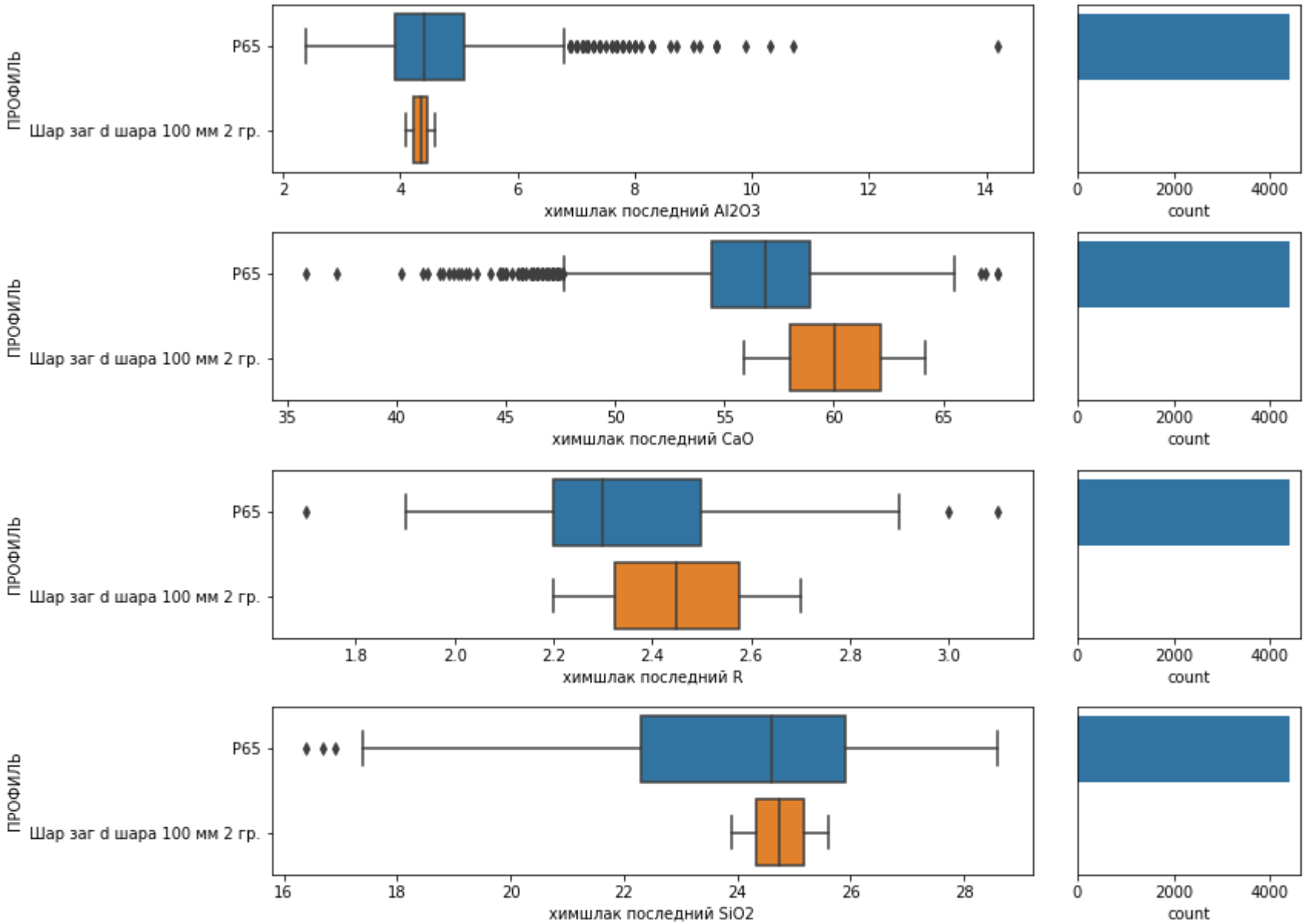
Заполнение пропусков

Я бы выбирал метод, который не меняет распределения

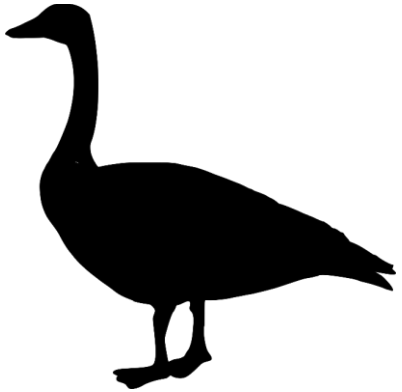




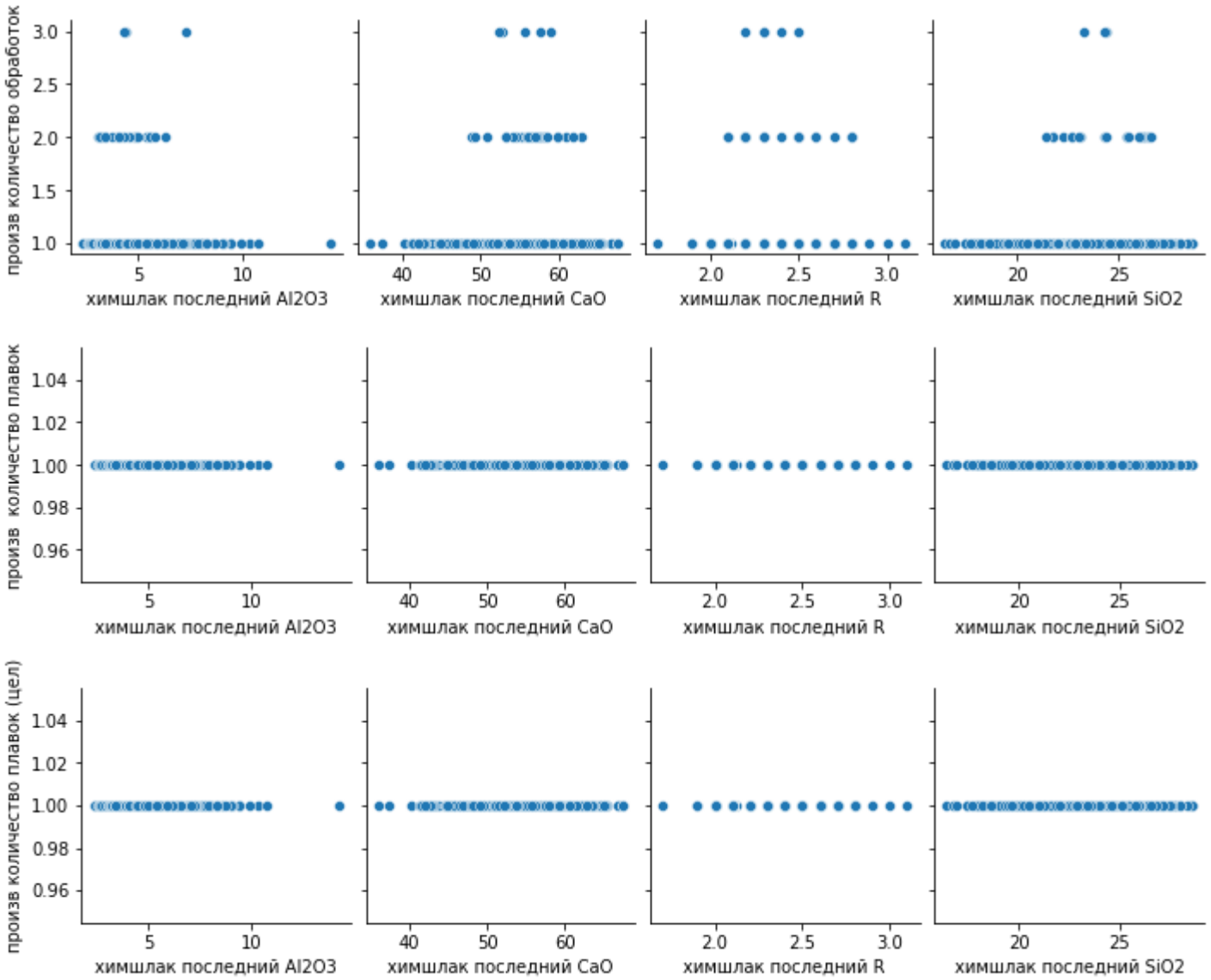
Категориальные фичи



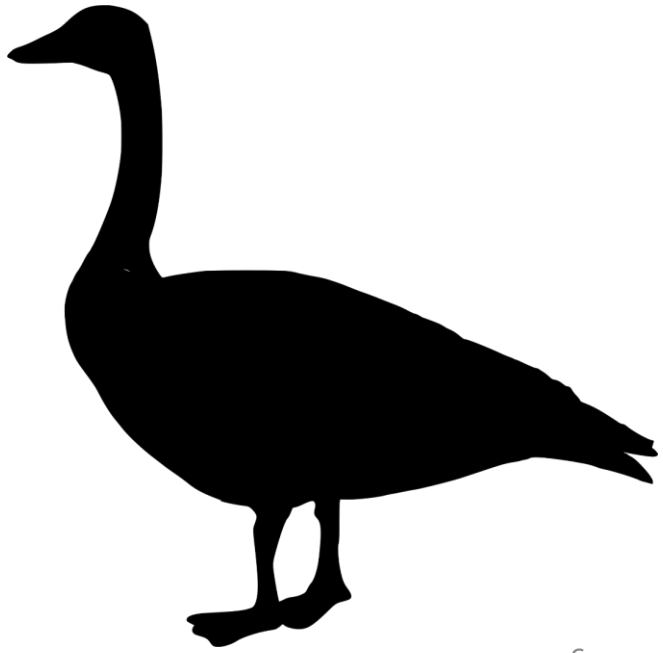
Перекос в сторону P65



Порядковые фичи

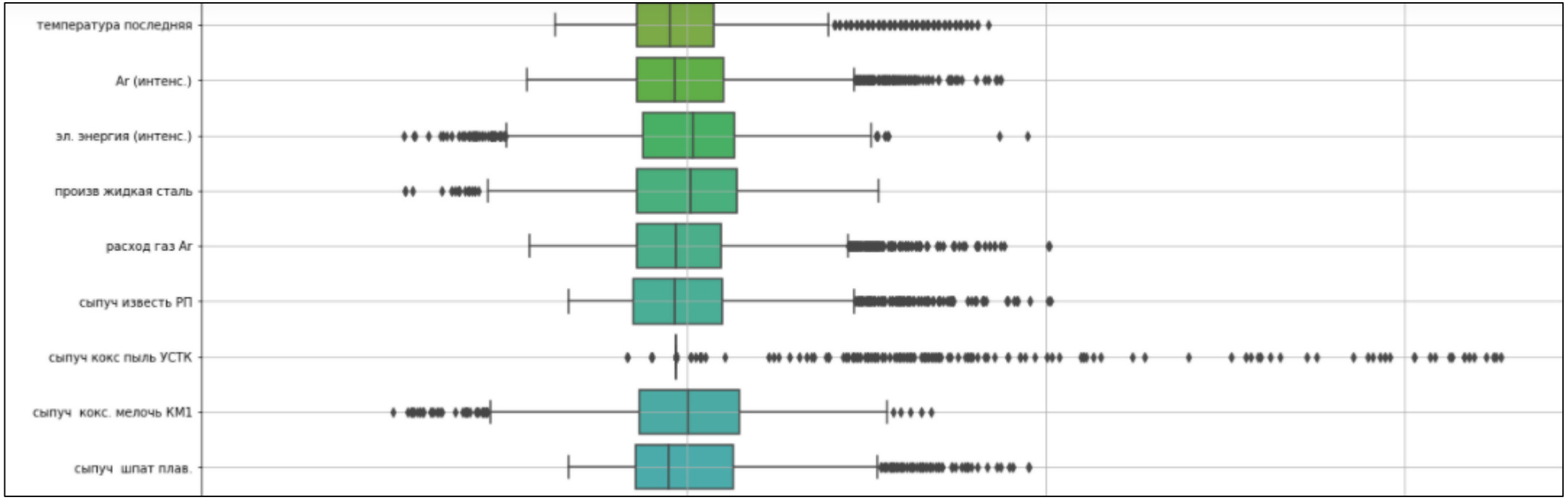


'произв количество плавок' и 'произв количество плавок (цел)' имеют одно фиксированное значение для всех таргетов.

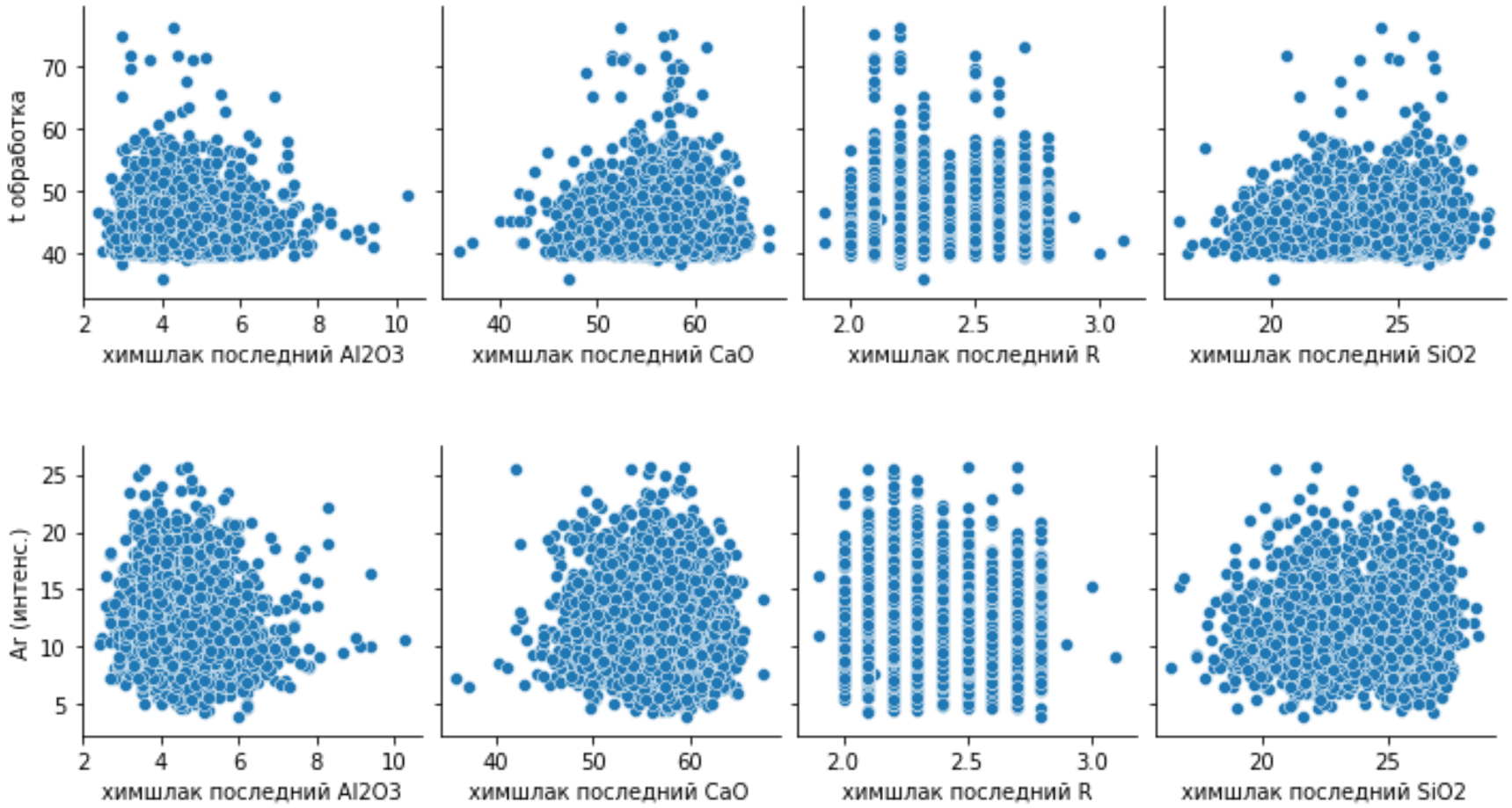


Численные фичи и выбросы

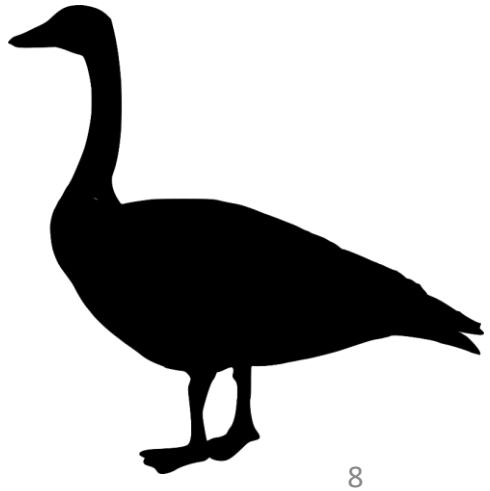
Было бы неплохо почистить выбросы



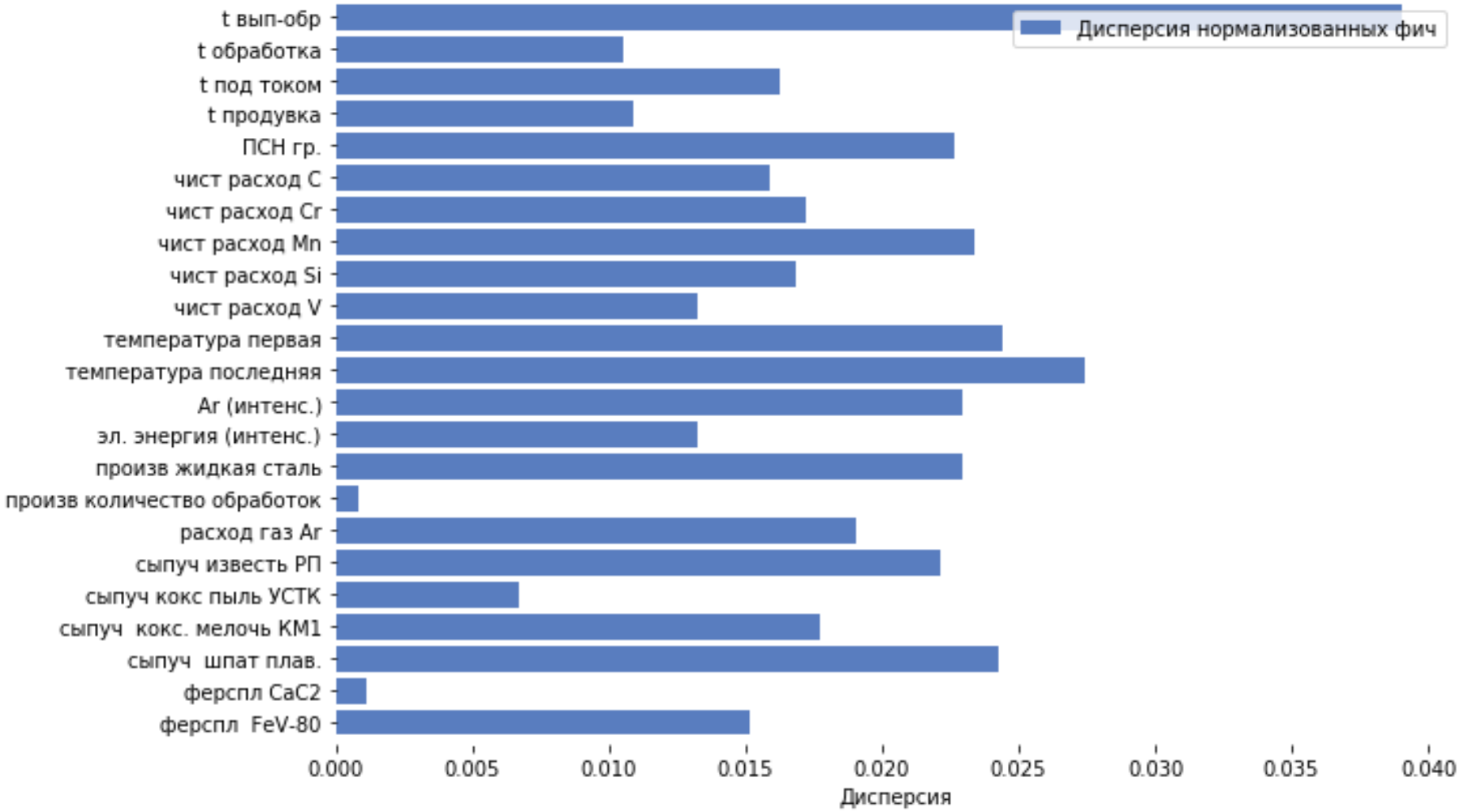
Численные фичи и зависимости



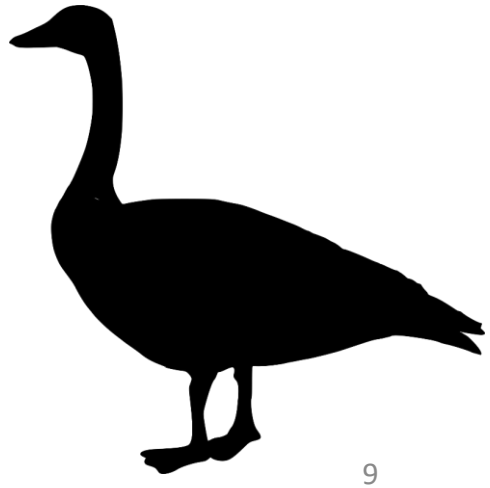
Данные сильные и независимые



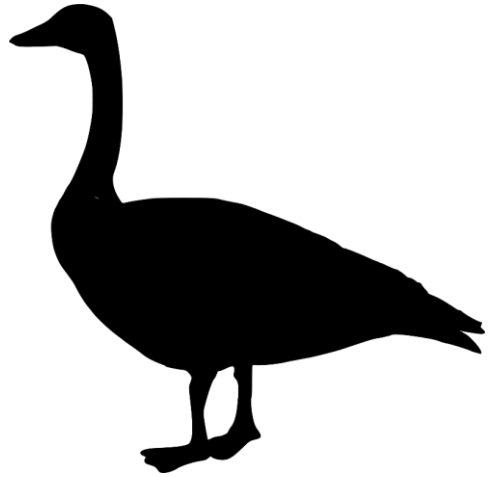
Дисперсия в данных (нормализованная)



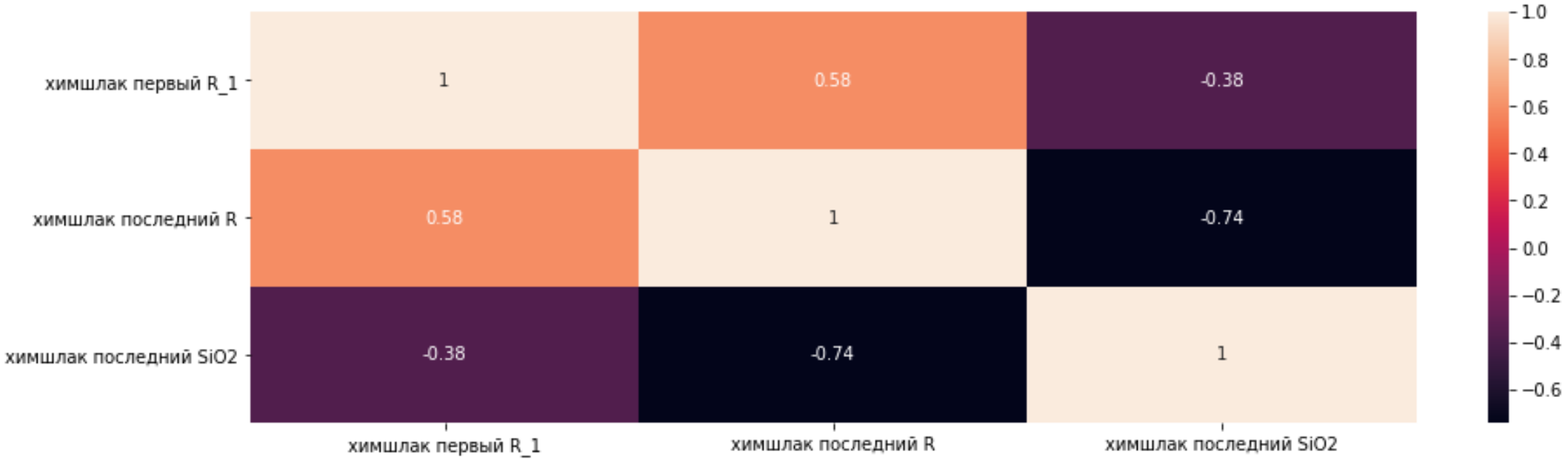
Околонулевая дисперсия фич плохо влияет на предсказательную модель

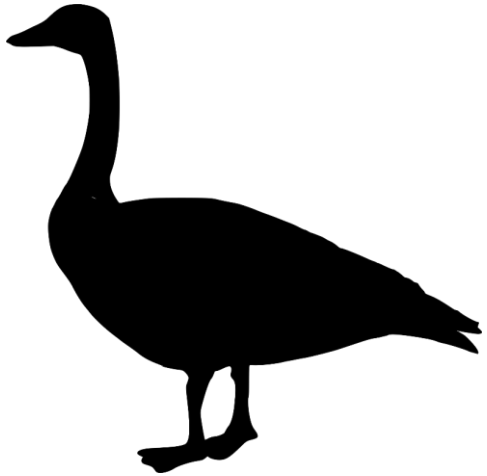


Корреляция таргетов



Химшлак последний R отрицательно коррелирует с Химшлак последний SiO2



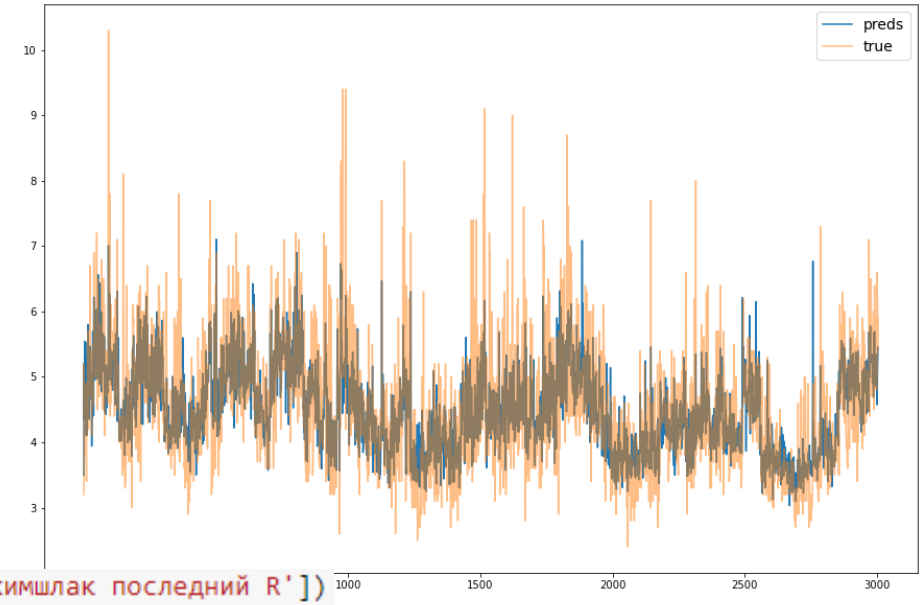


Предсказание таргетов

Пайплайн предсказания целевого состава химшлака

1. Предсказываем «химшлак последний Al2O3»
   - 65 фич (отбор+engineering)
   - RidgeCV (регуляризация)

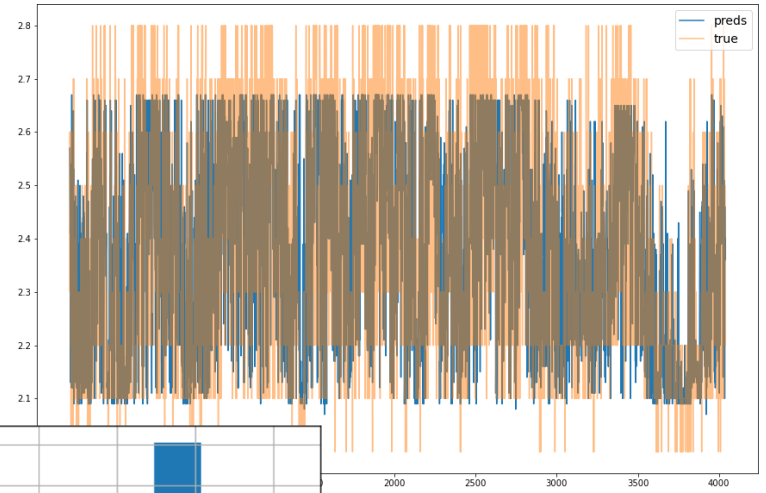
	train_R2	train_MAE	train_MSE	train_MAPe	test_R2	test_MAE	test_MSE	test_MAPe
0	0.520956	0.101672	0.018662	6.979809	0.50475	0.103632	0.018719	7.121529



2. Предсказываем «химшлак последний R»
   - + фича-предсказания «химшлак последний Al2O3»
   - «категоризация» таргета + SMOTE
   - RandomForest

```
Counter(data['химшлак последний R'])
```

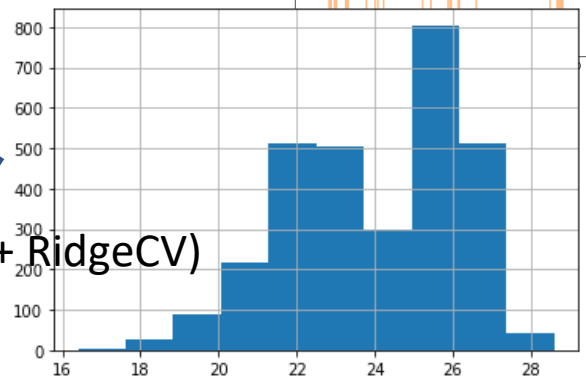
```
Counter({2.6: 373,
         2.3: 690,
         2.5: 399,
         2.2: 907,
         2.0: 111,
         2.1: 601,
         2.4: 355,
         2.7: 342,
         2.8: 262})
```



	train_R2	train_MAE	train_MSE	train_MAPe	test_R2	test_MAE	test_MSE	test_MAPe
0	0.665211	0.116723	0.022321	4.90639	0.625839	0.122838	0.024941	5.183411

3. Предсказываем «химшлак последний SiO2»
   - + фича-предсказания «химшлак последний R»
   - + фича-произведение предыдущих предсказаний таргетов
   - RidgeCV (регуляризация)

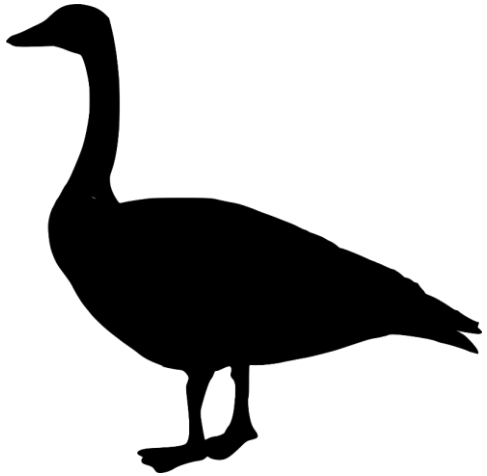
	train_R2	train_MAE	train_MSE	train_MAPe	test_R2	test_MAE	test_MSE	test_MAPe
0	0.236431	1.543365	3.620202	6.619005	0.21484	1.571822	3.786306	6.726546



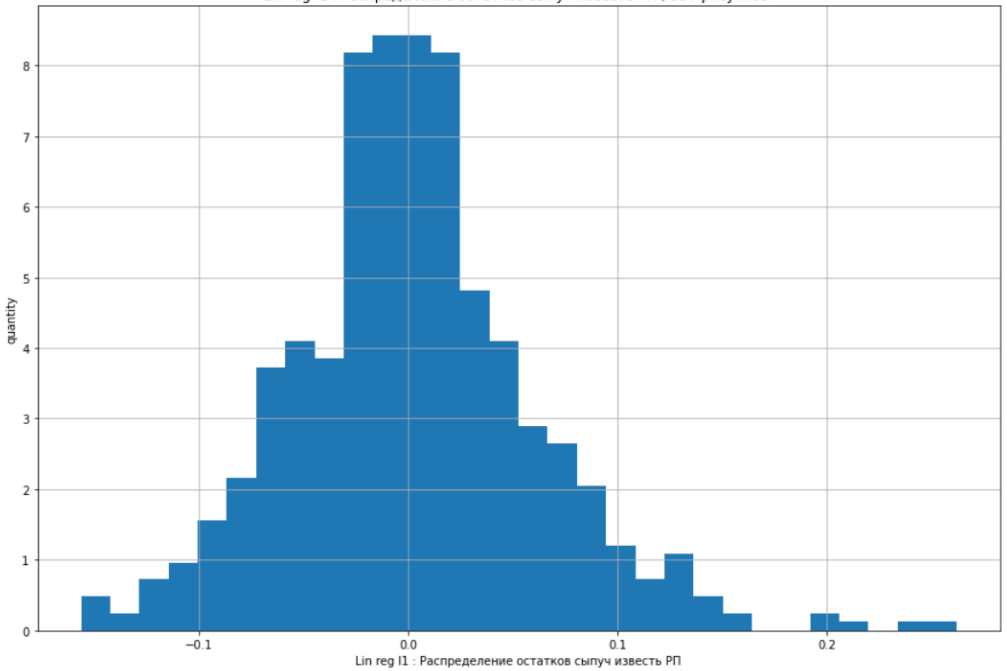
4. Предсказываем «химшлак последний CaO» (пред. Предсказания + RidgeCV)

	train_R2	train_MAE	train_MSE	train_MAPe	test_R2	test_MAE	test_MSE	test_MAPe
0	0.227888	2.570239	10.940711	4.674918	0.183668	2.655558	11.42191	4.802292

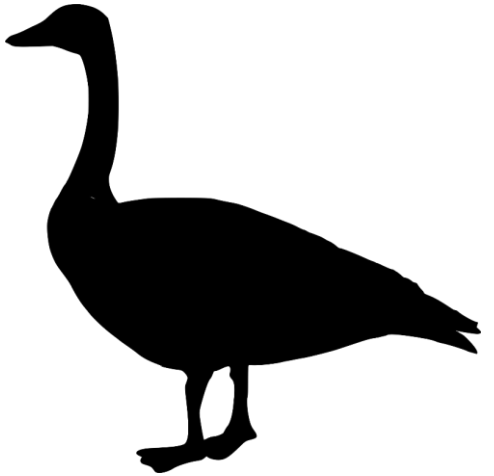
бимодальщина…

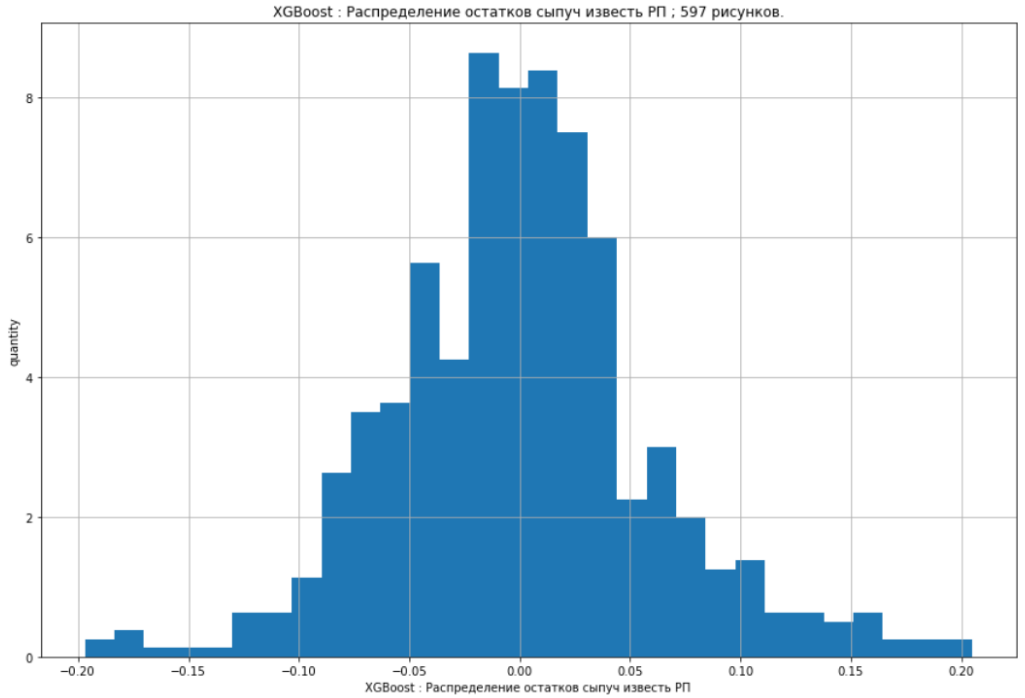


Финальное предсказание



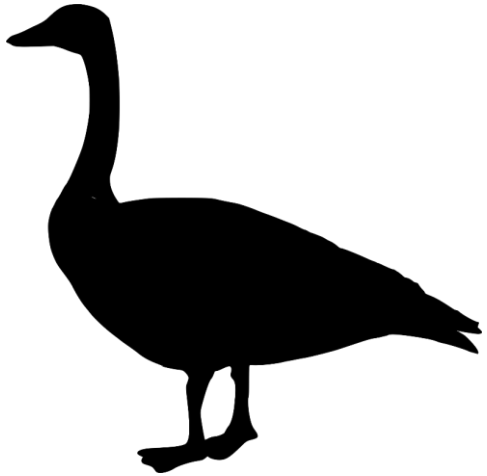
- Mean : 0.0027000228480781015
- Standard deviation : 0.058019236451975875
- Minimum : -0.15598137699593867
- Maximum : 0.26154918013184675
- Quantile 0.1 : -0.0663764859771178
- Quantile 0.9 : 0.0726327856245181





- Mean : 0.0006479496963858799
- Standard deviation : 0.05890657620136332
- Minimum : -0.19690929460525514
- Maximum : 0.20464485883712769
- Quantile 0.1 : -0.07205058488845825
- Quantile 0.9 : 0.0716210594177246

	R2	MAE	MSE	MAPE
XGBoost	0.434344	0.038558	0.002615	41.976033
Lin reg l2	-8.112844	0.041457	0.040797	44.884806
knn	-0.194701	0.056514	0.005540	61.543900
Random forest	0.429636	0.039284	0.002621	42.694805



Сервис

Спасибо за внимание☺