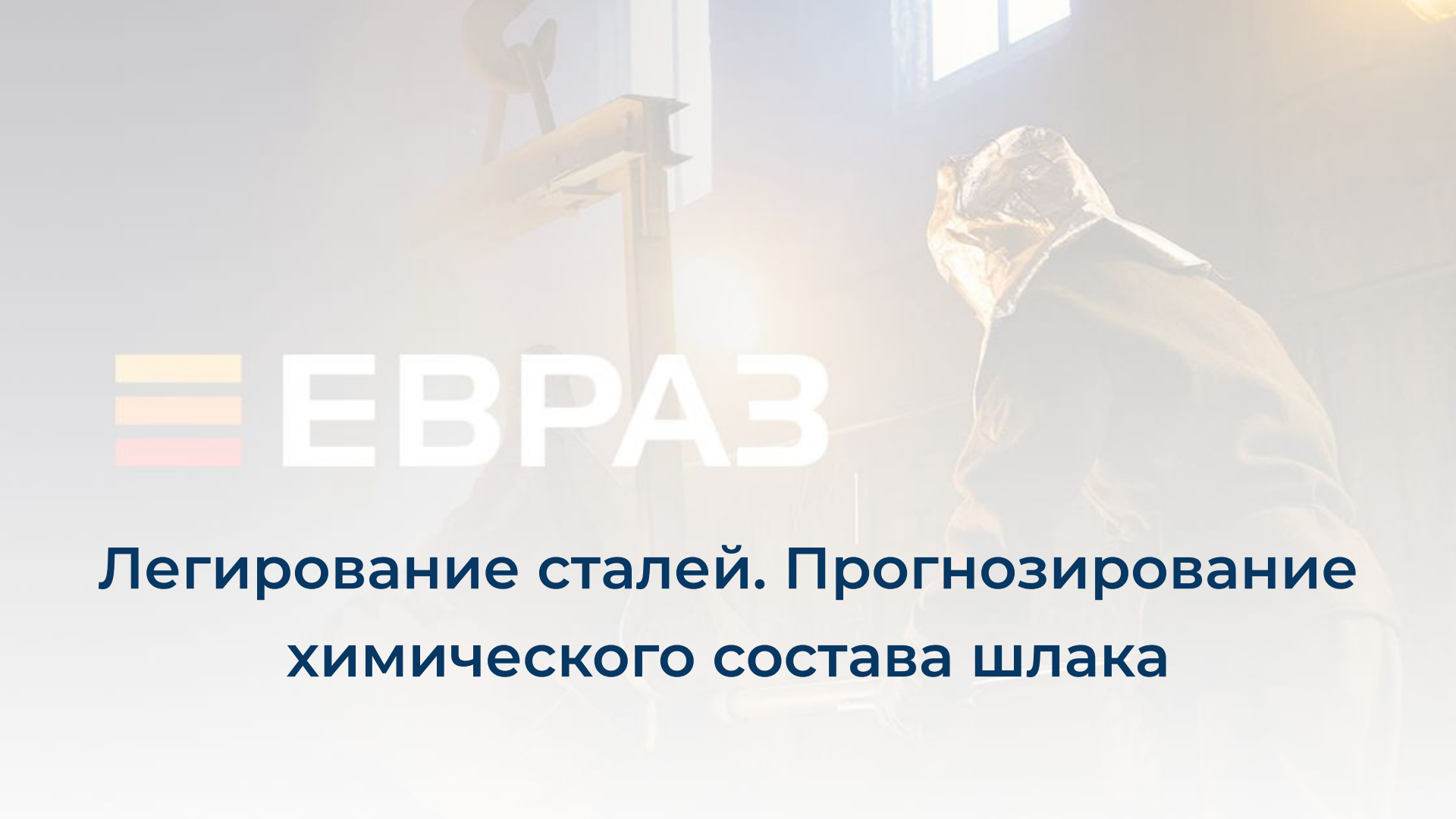

Легирование сталей. Прогнозирование химического состава шлака

Цель соревнования

Изучить данные физико-химического процесса легирования сталей и создать алгоритм определения химического состава шлака по исходным данным

Метрики качества

R2	отношение дисперсий: предсказанного и настоящего , чем больше тем лучше
MAE	в отличие от MSE слабее штрафует за бОльшие ошибки, чем меньше тем лучше
MSE	квадрат отклонения классическая метрика для задач регрессии, чем меньше тем лучше
MAPe	плохо подходит если в данных много нулей, чем меньше тем лучше

Data

Для кроссвалидации использовалось ~4200 элементов, с разделением на 10 частей

Все пропуски заменены на 0, по условию.

	МАРКА	t вып-обр	t обработка	t под током	t продувка	ПСН гр.	чист расход C	чист расход Cr	чист расход Mn	чист расход Si	чист расход V	температура первая	температура последняя	Ar (интенс.)	N2 (интенс.)	эл. энергия (интенс.)	произв жидкая сталь
0	Э76ХФ	29,0	45,3666667	24,4	41,0333333	NaN	0,45646	0,059572	0,117446	0,104762	0,0409383	1557,0	1580,0	13,6067425	NaN	12809,0163934	115,5
1	Э76ХФ	26,0	44,0666667	13,8666667	44,0666667	NaN	0,359285	0,083738	0,160923	0,110327	0,0400831	1601,0	1591,0	8,074721	NaN	12816,3461538	111,6
2	Э76ХФ	24,0	43,35	17,95	43,35	NaN	0,331665	0,08149	0,132332	0,13986	0,0416225	1593,0	1586,0	13,801968	NaN	12511,4206128	115,8
3	Э76ХФ	17,0	46,1833333	19,8166667	46,1833333	NaN	0,377945	0,133194	0,221605	0,165186	0,0420497	1589,0	1589,0	12,6649585	NaN	12998,1497056	116,3
4	Э76ХФ	20,0	48,5	17,0333333	48,5	NaN	0,389875	0,105094	0,169459	0,143024	0,0409667	1597,0	1592,0	10,2983505	NaN	12987,4755382	115,0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7036	Э90ХАФ	41,0	42,3333333	23,2	42,1666667	3,3806283	0,53708	0,10564	0,128626	0,101552	0,0921946	1552,0	1582,0	10,0156347	11,1591276	13388,7931034	114,4
7037	Э90ХАФ	36,0	46,5333333	16,1833333	46,3833333	NaN	0,555875	0,122876	0,156558	0,155113	0,0920241	1576,0	1580,0	17,7615538	21,443787	12053,1410917	112,0
7038	Э90ХАФ	42,0	47,5666667	23,0166667	47,1	2,2630044	0,548385	0,111756	0,124018	0,120513	0,0935034	1600,0	1600,0	8,9295723	8,2936906	12098,1897176	115,6
7039	Э90ХАФ	45,0	46,0333333	17,5333333	45,6833333	3,0	0,60135	0,147896	0,19077	0,15436	0,0951099	1612,0	1587,0	14,727117	15,484507	12726,6159696	116,9
7040	Э90ХАФ	48,0	52,0333333	21,7	50,2333333	3,0	0,5494	0,140112	0,163476	0,100967	0,0936424	1580,0	1576,0	20,9973742	19,8473684	12229,4930876	115,5

7041 rows × 81 columns

Features

В отборочном задании нельзя было использовать данные из будущего с пометкой “последний”

В финальном задании можно было использовать все имеющиеся

Решение

1. я загрузил данные в разные модели и выбрал лучшую согласно метрикам

Метрики: отборочный этап

химшлак последний Al2O3

	Column	Model Name	R2	MAE	MSE	MAPe
0	химшлак последний Al2O3	XGBRegressor	0.739275	0.338354	0.307893	0.098108
1	химшлак последний Al2O3	LGBMRegressor	0.747387	0.341773	0.314094	0.102445
2	химшлак последний Al2O3	TabNet	0.943267	0.327076	0.292709	0.105351

Параметры дефолтные

Метрики: отборочный этап

химшлак последний CaO

	Column	Model Name	R2	MAE	MSE	MAPe
0	химшлак последний CaO	XGBRegressor	0.895744	2.413937	10.676451	0.049618
1	химшлак последний CaO	LGBMRegressor	0.895750	2.400903	10.704460	0.049251
2	химшлак последний CaO	TabNet	0.971344	2.282357	9.952410	0.047143

Параметры дефолтные

Метрики: отборочный этап

химшлак последний R

	Column	Model Name	R2	MAE	MSE	MAPe
0	химшлак последний R	XGBRegressor	0.855428	0.127898	0.028758	0.060853
1	химшлак последний R	LGBMRegressor	0.855177	0.128024	0.029141	0.060880
2	химшлак последний R	TabNet	0.956060	0.125069	0.027699	0.058753

Параметры дефолтные

Метрики: отборочный этап

химшлак последний SiO2

	Column	Model Name	R2	MAE	MSE	MAPe
0	химшлак последний SiO2	XGBRegressor	0.854957	1.133493	2.752073	0.068402
1	химшлак последний SiO2	LGBMRegressor	0.854114	1.127226	2.784321	0.066609
2	химшлак последний SiO2	TabNet	0.980732	1.061235	2.555623	0.068341

Параметры дефолтные

Метрики: отборочный этап

сыпуч известь РП

	Column	Model Name	R2	MAE	MSE	MAPe
0	сыпуч известь РП	XGBRegressor	0.347109	0.040141	0.002848	0.463609
1	сыпуч известь РП	LGBMRegressor	0.359543	0.039586	0.002764	0.458894
2	сыпуч известь РП	TabNet	0.532624	0.038075	0.002707	0.397542

Параметры дефолтные

Выводы

1. Лучше всего использовать метрику R2 или MSE, так как первая сравнивает похожесть распределения данных, а вторая учитывает среднюю ошибку
2. Модель TabNet превосходит с дефолтными параметрами все остальные модели, но тренируется в 10 раз дольше (~1час)
3. -