NATIONAL RESEARCH UNIVERSITY

DA in Python

Cell viability prediction

Dokin
Zaripov
Yartsev
Epifanov

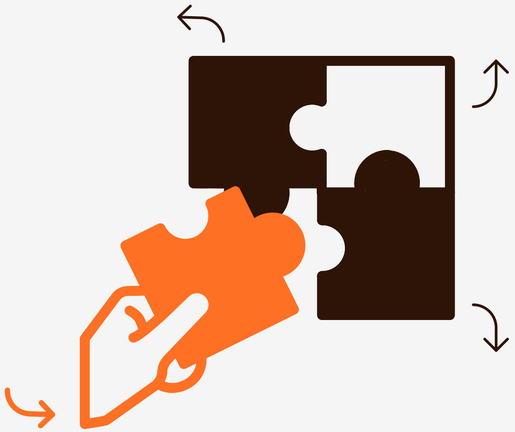
Team

24 ZIOC RAS

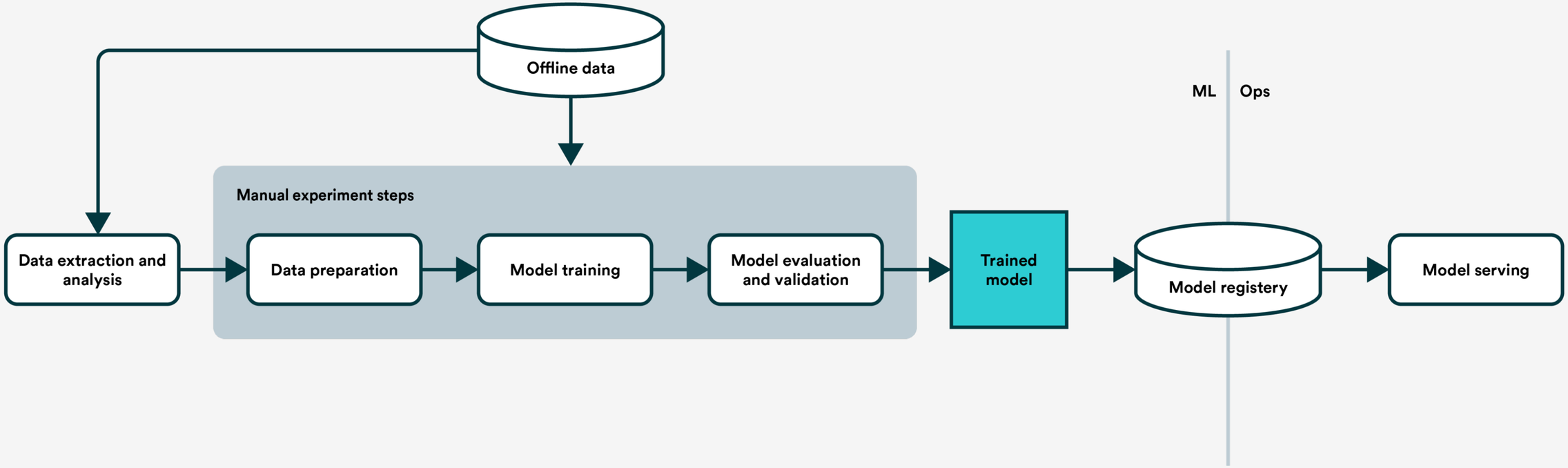
Alpha male

Tatarchad

Pipeline




NORD STREAM 2


Offline data
ML
Ops
Manual experiment steps
Data extraction and analysis
Data preparation
Model training
Model evaluation and validation
Trained model
Model registery
Model serving

Data engineering

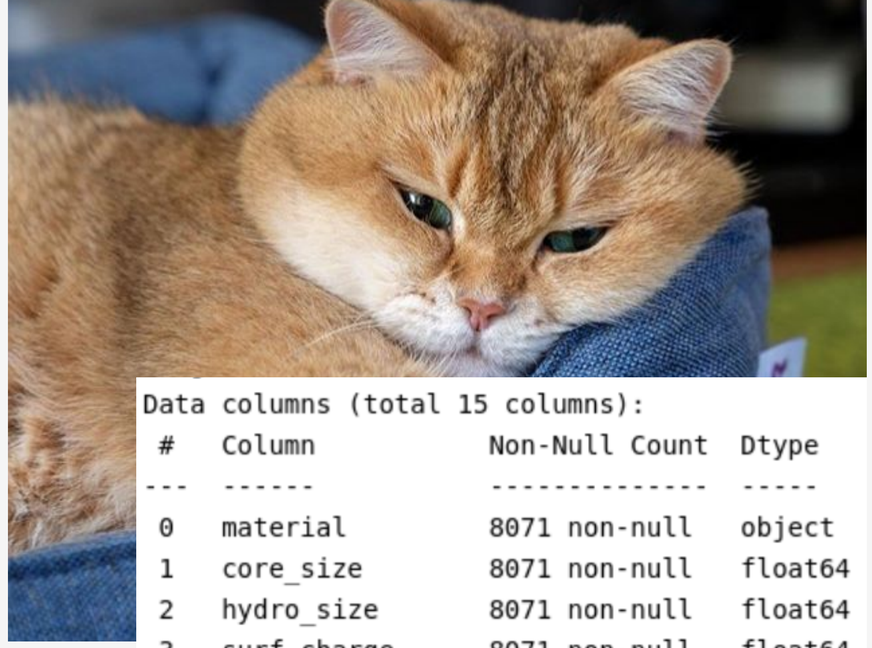
Problems:

- Many NaN values
- Outliers
- Not a single dataset
- Feature naming troubles

Results:

- No time to sleep
- Descriptors naming issues solved
- A single dataset creates

Preprocessing



	material	core_size	hydro_size	surf_charge	is_cancer_cell	dose	viability	material_type	i
0	CuO	12.8	313.8	0.0	1	200.0	9.1000	I	
1	ZnO	22.6	114.7	0.0	0	200.0	9.5000	I	
2	ZnO	22.6	114.7	0.0	0	100.0	10.2000	I	
3	ZnO	22.6	69.4	0.0	1	100.0	11.0000	I	
4	Mn2O3	51.5	291.7	0.0	1	200.0	11.3000	I	
...	...	...	...	...	...	...	...	...	
8066	ZnO	35.6	236.0	-41.6	1	1.0	127.4363	I	
8067	ZnO	35.6	236.0	-41.6	1	10.0	116.3751	I	
8068	ZnO	35.6	236.0	-41.6	1	100.0	40.8796	I	
8069	ZnO	35.6	236.0	-41.6	1	0.1	86.8566	I	
8070	ZnO	35.6	236.0	-41.6	1	1.0	84.4578	I	

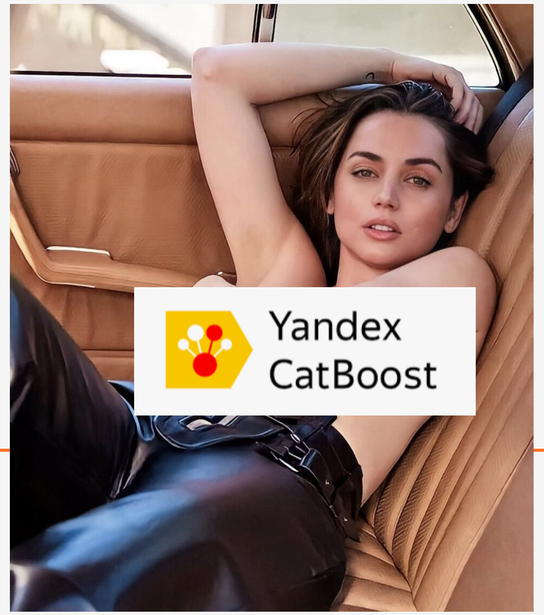
8071 rows × 15 columns

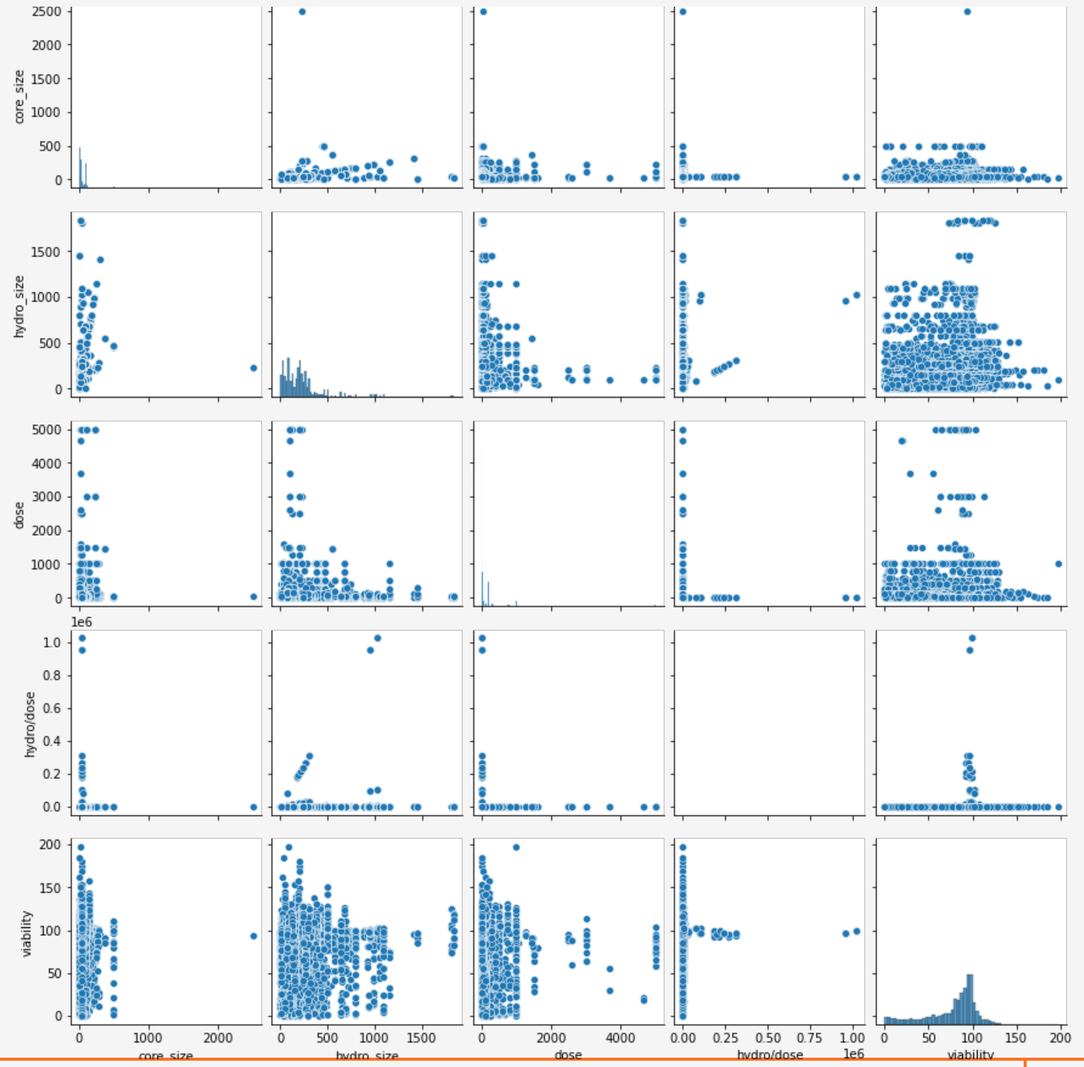
```
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   material         8071 non-null   object
 1   core_size        8071 non-null   float64
 2   hydro_size       8071 non-null   float64
 3   surf_charge      8071 non-null   float64
 4   is_cancer_cell   8071 non-null   int64
 5   dose             8071 non-null   float64
 6   viability        8071 non-null   float64
 7   material_type    8071 non-null   object
 8   is_human_cell    8071 non-null   object
 9   surf_charge_cat  8071 non-null   int64
 10  cell_age         8071 non-null   object
 11  cell_origin      8071 non-null   object
 12  cell_type        8071 non-null   object
 13  cell_line        8071 non-null   object
 14  hydro/dose       8071 non-null   float64
dtypes: float64(6), int64(2), object(7)
memory usage: 945.9+ KB
```

EDA

- No linear correlation between features and target variable

- Linear models will be unefficient, Gradient boosting will be used



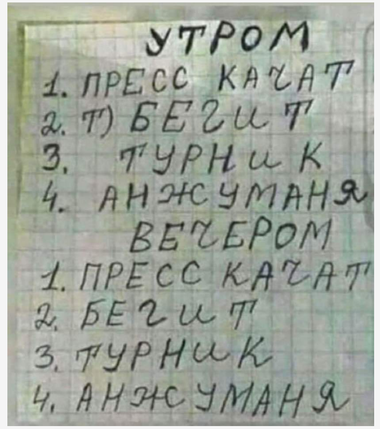


Training

EDA

Make baseline

Optimize

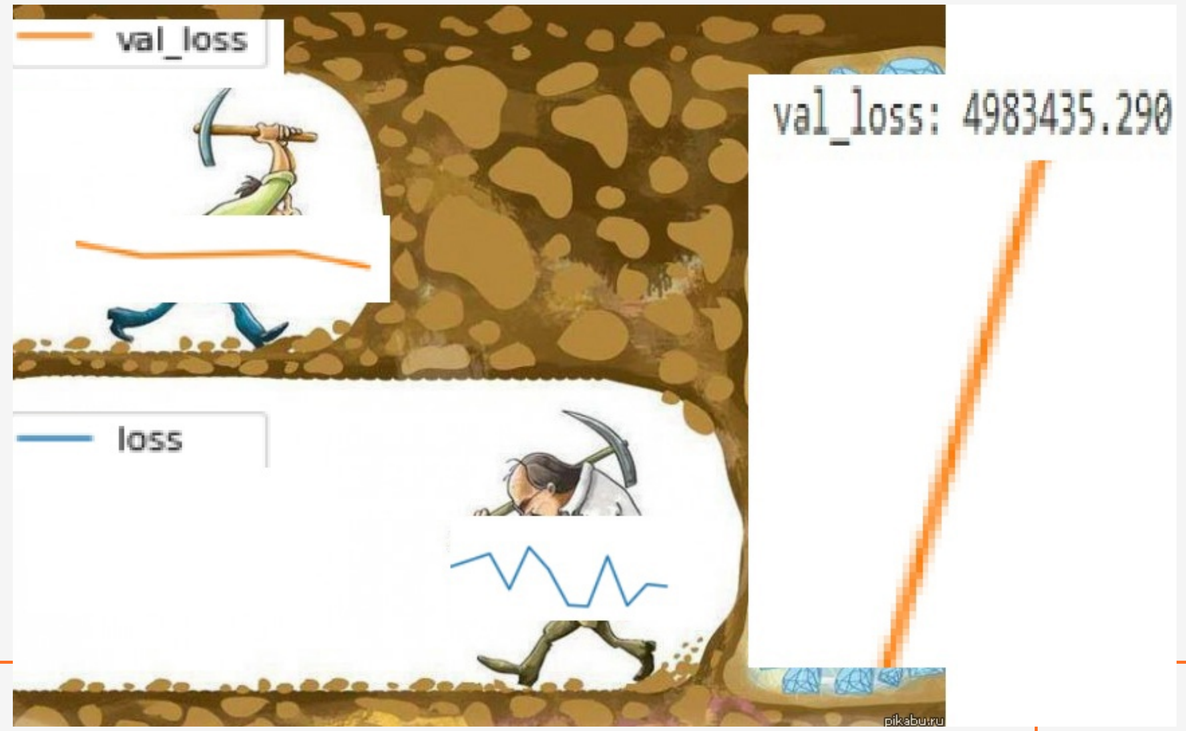


Fix bugs
Repeat

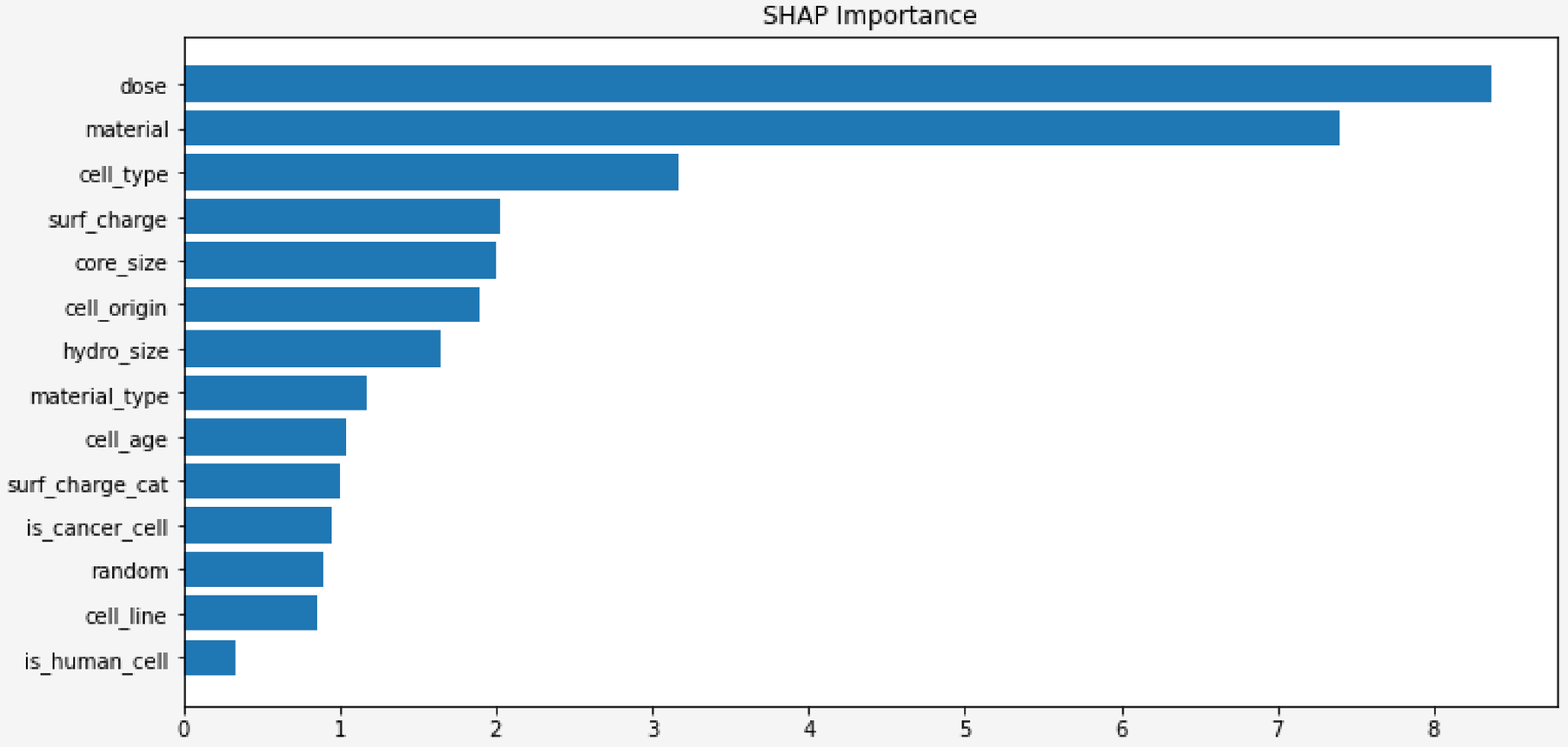
Training: details

Yandex CatBoost

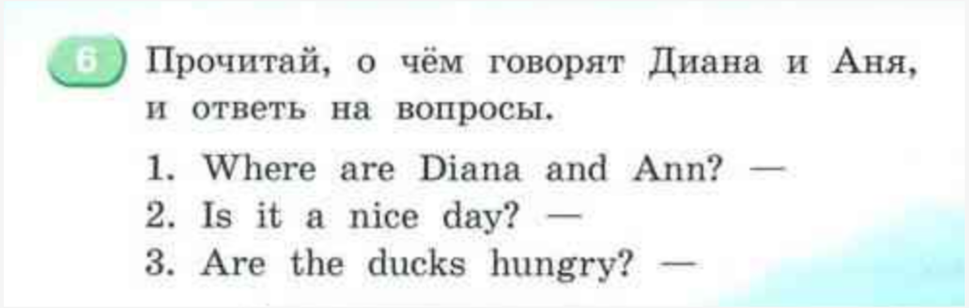
- CatBoost Regressor chosen because of convinience
- Loss function - RMSE
- Approx. 25k iterations



Baseline: interpretation



Baseline: interpretation


6 Прочитай, о чём говорят Диана и Аня, и ответь на вопросы.

1. Where are Diana and Ann? —
2. Is it a nice day? —
3. Are the ducks hungry? —

- dose, material -> the most useful features
- core_size, hydro_size -> the most important real features

- cell_age, is_cancer_cell, is_human_cell -> less important than the noise feature, should be eliminated

Feature engineering

Updates:

- New feature:
  hydro_size / dose ratio

- ~ particle concentration

$$C = \frac{n}{V}$$

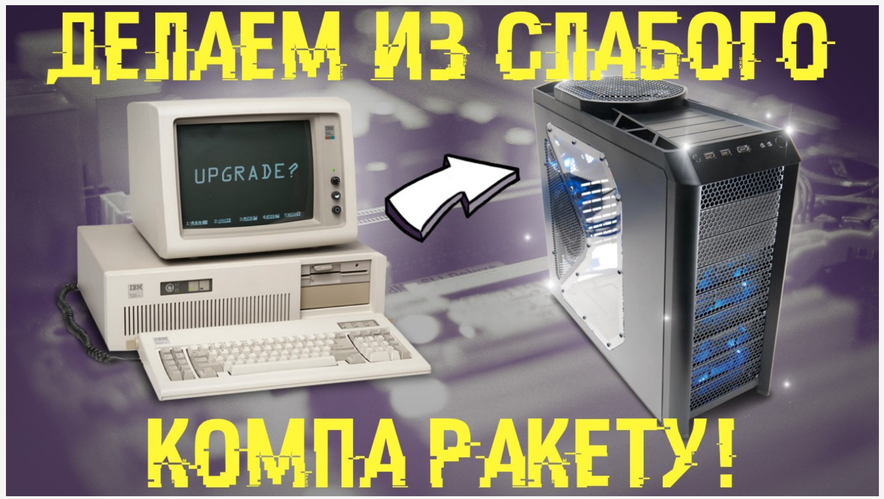
Концентрация растворов. Способ...
calc.ru

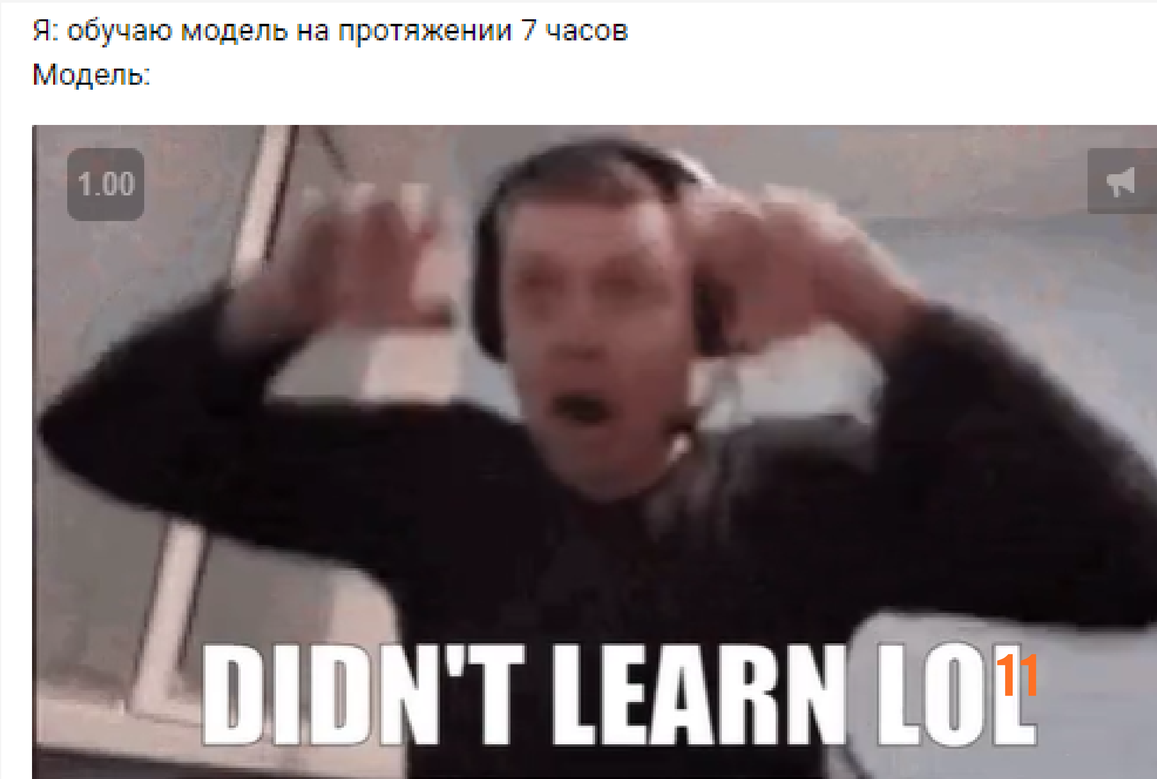
Чем полезна концентрация? | П...
shkolazhizni.ru

Optimization

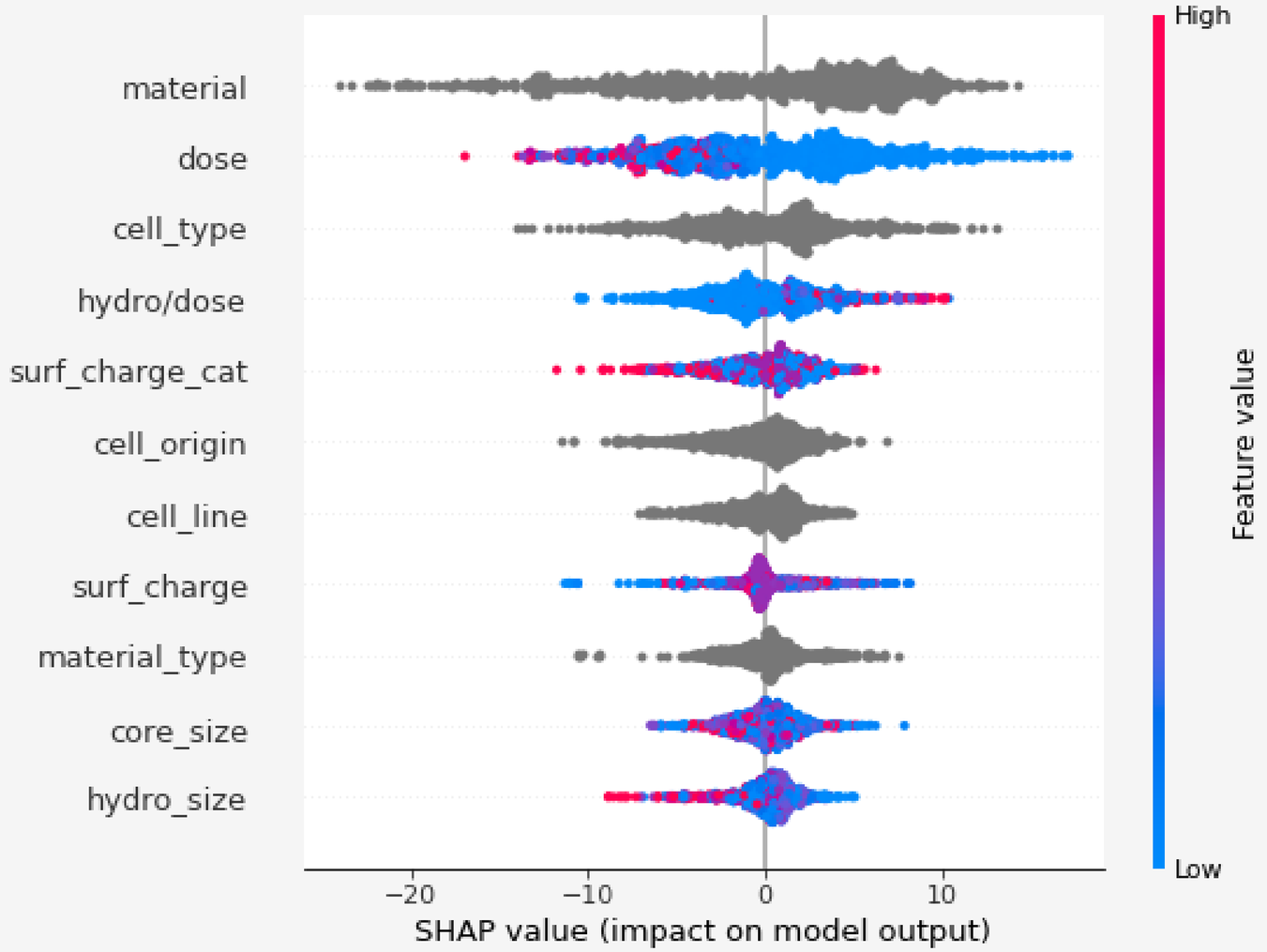
OPTUNA

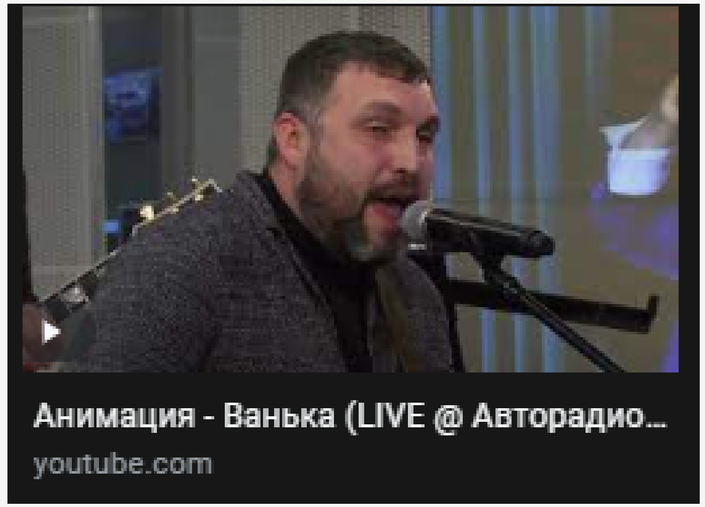
- Optuna framework used for hyperparameter correction
- Cross validation used to control the quality
- Iteration number increased to 50k
- Early stopping added



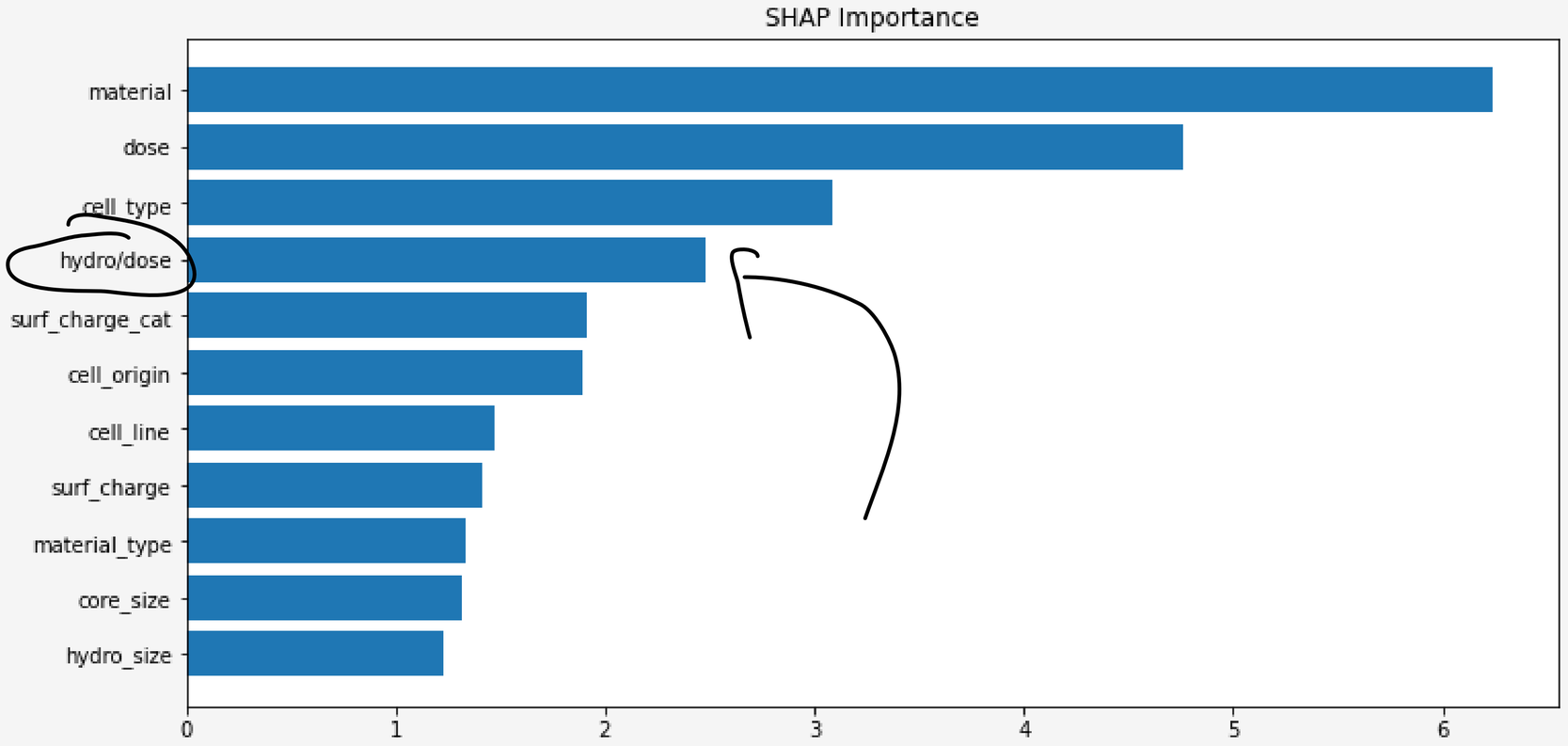


Interpretation





Interpretation



New feature fate:

- 4th most important from 11 + better than parent hydro_size

Timeline

в первый раз он выдал мне 56 MAE, что на 0-115 рендже с несколькими выбросами почти похоже на случайность 03:59

r2_score:-3.8973293515542906e-05 14:51

r2_score:-0.9000419442647092
MAE:23.186936156271067 15:18

Out of folds MAE = 13.200534747446937 22:31

The end

13.2005
MEAN MAE

- ML model for predicting cell viability trained and evaluated
- The results uploaded to GitHub


```
Fold 1 | MAE: 13.113173720472396
Fold 2 | MAE: 13.388517347260217
Fold 3 | MAE: 13.220920966244766
Fold 4 | MAE: 12.736958657051733
Fold 5 | MAE: 13.543170715475108

Mean MAE = 13.200548281300843
Out of folds MAE = 13.200534747446937
```