Введение

Создание нового лекарственного препарата — сложный процесс, включающий определение химической формулы, синтез соединений, биологические испытания и клинические тесты. Машинное обучение ускоряет этот процесс, позволяя прогнозировать эффективность химических соединений. В проекте проанализированы данные о 1000 соединениях для предсказания их активности против вируса гриппа. Параметры: IC50 (концентрация, ингибирующая 50% вируса), CC50 (токсичность для 50% клеток), SI (селективный индекс, (SI = \frac{CC50}{IC50})). Соединения с SI > 8 — потенциально эффективные.

Цель — построить модели регрессии для логарифмов IC50, CC50, SI и классификации для определения превышения медианы и SI > 8. Отчет описывает датасет, обработку, моделирование, результаты и QSAR-анализ.

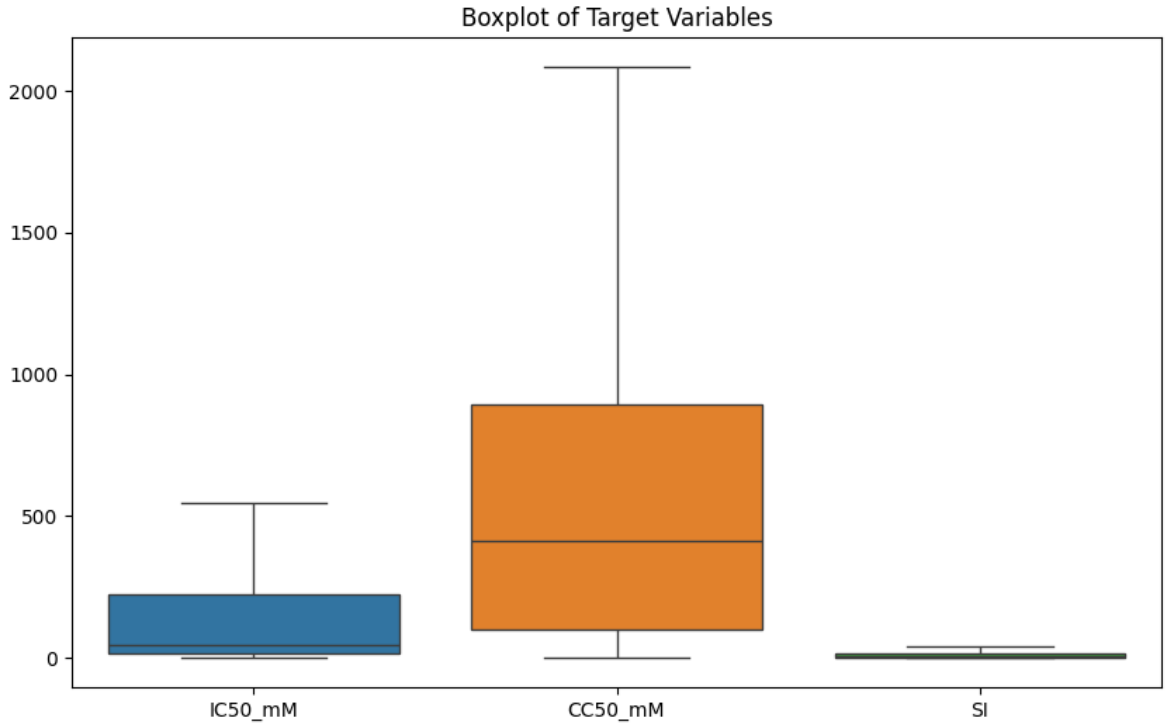
Описание датасета

Датасет содержит 1000 соединений с числовыми признаками, IC50*mM, CC50*mM, SI. Загружен из `data/coursework_data.xlsx`.

Характеристики до обработки

- **Размер**: 1001 строк, 214 столбцов.
- **Пропуски**: 36.
- **Типы данных**: 107 float64, 107 int64.
- **Выбросы**: В IC50*mM, CC50*mM, SI.

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	1001.0	5.000000e+02	2.891100e+02	0.00	250.00	500.00	750.00	1.000000e+03
IC50, mM	1001.0	2.228100e+02	4.021700e+02	0.00	12.52	46.59	224.98	4.128530e+03
CC50, mM	1001.0	5.891100e+02	6.428700e+02	0.70	100.00	411.04	894.09	4.538980e+03
SI	1001.0	7.251000e+01	6.844800e+02	0.01	1.43	3.85	16.57	1.562060e+04
MaxAbsEStateIndex	1001.0	1.083000e+01	3.310000e+00	2.32	9.25	12.18	13.17	1.593000e+01
MaxEStateIndex	1001.0	1.083000e+01	3.310000e+00	2.32	9.25	12.18	13.17	1.593000e+01
MinAbsEStateIndex	1001.0	1.800000e-01	1.700000e-01	0.00	0.05	0.12	0.29	1.370000e+00
MinEStateIndex	1001.0	-9.700000e-01	1.590000e+00	-6.99	-1.33	-0.42	0.06	1.370000e+00
qed	1001.0	5.800000e-01	2.100000e-01	0.06	0.44	0.63	0.74	9.500000e-01
SPS	1001.0	2.949000e+01	1.274000e+01	9.42	18.49	29.29	38.75	6.027000e+01
MolWt	1001.0	3.482600e+02	1.269500e+02	110.16	264.32	315.46	409.28	9.047800e+02
HeavyAtomMolWt	1001.0	3.244700e+02	1.216900e+02	100.08	244.21	293.18	385.26	8.563900e+02
ExactMolWt	1001.0	3.479400e+02	1.268100e+02	110.07	264.14	315.22	408.17	9.042500e+02
NumValenceElectrons	1001.0	1.321200e+02	4.670000e+01	42.00	102.00	120.00	152.00	3.500000e+02
NumRadicalElectrons	1001.0	0.000000e+00	0.000000e+00	0.00	0.00	0.00	0.00	0.000000e+00
MaxPartialCharge	998.0	2.400000e-01	1.300000e-01	-0.02	0.12	0.25	0.34	5.700000e-01
MinPartialCharge	998.0	-4.100000e-01	8.000000e-02	-0.74	-0.48	-0.39	-0.35	-9.000000e-02
MaxAbsPartialCharge	998.0	4.200000e-01	7.000000e-02	0.09	0.36	0.43	0.48	7.400000e-01
MinAbsPartialCharge	998.0	2.300000e-01	1.200000e-01	0.00	0.12	0.25	0.33	5.100000e-01
FpDensityMorgan1	1001.0	1.140000e+00	2.400000e-01	0.22	1.00	1.15	1.33	1.750000e+00
FpDensityMorgan2	1001.0	1.820000e+00	3.200000e-01	0.38	1.63	1.88	2.05	2.620000e+00
FpDensityMorgan3	1001.0	2.420000e+00	4.000000e-01	0.58	2.24	2.50	2.69	3.270000e+00
BCUT2D_MWHI	998.0	2.320000e+01	1.453000e+01	14.51	16.37	16.56	32.09	1.269100e+02
BCUT2D_MWLOW	998.0	9.790000e+00	4.600000e-01	0.93	9.69	9.72	9.98	1.071000e+01
BCUT2D_CHGHI	998.0	2.470000e+00	1.600000e-01	1.83	2.38	2.51	2.59	2.820000e+00
BCUT2D_CHGLO	998.0	-2.350000e+00	1.700000e-01	-2.72	-2.48	-2.36	-2.24	-1.710000e+00
BCUT2D_LOGPHI	998.0	2.470000e+00	1.600000e-01	1.93	2.36	2.49	2.61	2.790000e+00
BCUT2D_LOGPLOW	998.0	-2.400000e+00	1.800000e-01	-2.79	-2.53	-2.40	-2.32	-1.650000e+00
BCUT2D_MRHI	998.0	6.300000e+00	1.090000e+00	4.65	5.77	5.94	6.46	1.411000e+01
BCUT2D_MRLOW	998.0	-7.000000e-02	2.500000e-01	-1.08	-0.16	-0.11	0.06	1.170000e+00
AvgIpc	1001.0	2.820000e+00	4.300000e-01	1.79	2.49	2.80	3.09	3.950000e+00
BalabanJ	1001.0	1.850000e+00	4.500000e-01	0.00	1.53	1.84	2.08	3.790000e+00
BertzCT	1001.0	7.539900e+02	4.359100e+02	113.63	418.37	660.91	963.31	2.263930e+03
Chi0	1001.0	1.767000e+01	6.310000e+00	5.56	13.51	16.07	20.66	4.576000e+01
Chi0n	1001.0	1.452000e+01	5.180000e+00	4.11	11.24	13.30	16.35	3.599000e+01
Chi0v	1001.0	1.485000e+01	5.270000e+00	4.11	11.40	13.63	16.87	3.599000e+01
Chi1	1001.0	1.173000e+01	4.430000e+00	3.86	8.79	10.43	13.88	2.980000e+01
Chi1n	1001.0	8.640000e+00	3.140000e+00	2.13	6.71	7.92	9.82	2.221000e+01
Chi1v	1001.0	8.940000e+00	3.230000e+00	2.13	6.78	8.08	10.30	2.221000e+01
Chi2n	1001.0	7.320000e+00	2.740000e+00	1.30	5.55	6.74	8.61	2.018000e+01
Chi2v	1001.0	7.660000e+00	2.840000e+00	1.30	5.64	7.25	9.04	2.018000e+01
Chi3n	1001.0	5.620000e+00	2.460000e+00	0.75	4.09	5.13	6.66	1.648000e+01
Chi3v	1001.0	5.910000e+00	2.540000e+00	0.75	4.18	5.54	6.96	1.648000e+01
Chi4n	1001.0	4.160000e+00	1.940000e+00	0.42	2.95	3.86	4.87	1.248000e+01
Chi4v	1001.0	4.400000e+00	2.020000e+00	0.42	3.08	4.08	5.19	1.248000e+01
HallKierAlpha	1001.0	-1.890000e+00	1.380000e+00	-6.52	-2.85	-1.56	-0.79	2.700000e-01
Ipc	1001.0	4.831703e+10	1.255969e+12	107.11	18858.35	112366.73	4399307.31	3.951781e+13
Kappa1	1001.0	1.702000e+01	6.370000e+00	4.54	12.77	15.66	19.49	4.691000e+01
Kappa2	1001.0	6.380000e+00	3.000000e+00	1.34	4.37	5.67	7.53	2.037000e+01
Kappa3	1001.0	3.090000e+00	1.720000e+00	0.44	2.00	2.67	3.72	1.267000e+01
LabuteASA	1001.0	1.466800e+02	5.254000e+01	46.23	112.44	132.52	171.54	3.533300e+02
PEOE_VSA1	1001.0	1.263000e+01	1.159000e+01	0.00	5.32	9.84	14.95	1.189100e+02
PEOE_VSA10	1001.0	8.060000e+00	1.261000e+01	0.00	0.00	5.69	11.51	1.079200e+02
PEOE_VSA11	1001.0	4.990000e+00	8.940000e+00	0.00	0.00	0.00	5.78	4.741000e+01
PEOE_VSA12	1001.0	3.410000e+00	5.460000e+00	0.00	0.00	0.00	5.91	2.951000e+01
PEOE_VSA13	1001.0	1.180000e+00	2.620000e+00	0.00	0.00	0.00	0.00	1.772000e+01
PEOE_VSA14	1001.0	3.380000e+00	7.280000e+00	0.00	0.00	0.00	5.97	1.135600e+02
PEOE_VSA2	1001.0	5.970000e+00	5.870000e+00	0.00	0.00	4.79	9.59	3.419000e+01
PEOE_VSA3	1001.0	2.980000e+00	4.960000e+00	0.00	0.00	0.00	4.79	4.346000e+01
PEOE_VSA4	1001.0	2.950000e+00	6.720000e+00	0.00	0.00	0.00	0.00	4.461000e+01
PEOE_VSA5	1001.0	1.760000e+00	4.390000e+00	0.00	0.00	0.00	0.00	2.352000e+01
PEOE_VSA6	1001.0	2.641000e+01	2.063000e+01	0.00	12.13	22.67	37.82	1.289000e+02
PEOE_VSA7	1001.0	4.152000e+01	2.182000e+01	0.00	30.18	38.06	51.37	1.610600e+02
PEOE_VSA8	1001.0	1.649000e+01	1.080000e+01	0.00	6.92	17.02	23.17	5.228000e+01
PEOE_VSA9	1001.0	1.476000e+01	1.253000e+01	0.00	5.69	12.29	19.76	6.545000e+01
SMR_VSA1	1001.0	1.569000e+01	1.348000e+01	0.00	4.92	14.23	23.11	1.189100e+02
SMR_VSA10	1001.0	1.512000e+01	1.323000e+01	0.00	5.71	11.68	22.89	6.841000e+01
SMR_VSA2	1001.0	6.000000e-02	5.600000e-01	0.00	0.00	0.00	0.00	5.530000e+00
SMR_VSA3	1001.0	5.310000e+00	6.670000e+00	0.00	0.00	4.90	9.80	5.906000e+01
SMR_VSA4	1001.0	1.177000e+01	1.223000e+01	0.00	0.00	5.92	17.75	5.736000e+01
SMR_VSA5	1001.0	4.165000e+01	2.959000e+01	0.00	19.28	40.03	57.53	1.743700e+02
SMR_VSA6	1001.0	1.347000e+01	1.623000e+01	0.00	0.00	7.11	18.41	1.047200e+02
SMR_VSA7	1001.0	3.770000e+01	3.172000e+01	0.00	11.65	34.89	64.72	1.329600e+02
SMR_VSA8	1001.0	0.000000e+00	0.000000e+00	0.00	0.00	0.00	0.00	0.000000e+00
SMR_VSA9	1001.0	5.730000e+00	1.097000e+01	0.00	0.00	0.00	5.75	5.750000e+01
SlogP_VSA1	1001.0	6.070000e+00	6.170000e+00	0.00	0.00	5.32	10.17	3.316000e+01
SlogP_VSA10	1001.0	4.590000e+00	7.220000e+00	0.00	0.00	0.00	5.69	3.980000e+01
SlogP_VSA11	1001.0	3.770000e+00	8.450000e+00	0.00	0.00	0.00	5.75	5.750000e+01
SlogP_VSA12	1001.0	4.070000e+00	7.950000e+00	0.00	0.00	0.00	7.60	4.588000e+01
SlogP_VSA2	1001.0	3.307000e+01	2.379000e+01	0.00	16.00	27.95	45.35	2.131400e+02
SlogP_VSA3	1001.0	8.220000e+00	7.720000e+00	0.00	4.74	6.18	12.35	4.355000e+01
SlogP_VSA4	1001.0	1.194000e+01	1.098000e+01	0.00	0.00	11.33	16.75	5.226000e+01
SlogP_VSA5	1001.0	3.959000e+01	2.664000e+01	0.00	21.48	40.03	50.42	1.840800e+02
SlogP_VSA6	1001.0	3.084000e+01	2.518000e+01	0.00	9.98	28.55	48.53	1.267000e+02
SlogP_VSA7	1001.0	3.000000e-01	1.390000e+00	0.00	0.00	0.00	0.00	1.964000e+01
SlogP_VSA8	1001.0	4.030000e+00	7.190000e+00	0.00	0.00	0.00	6.08	4.504000e+01
SlogP_VSA9	1001.0	0.000000e+00	0.000000e+00	0.00	0.00	0.00	0.00	0.000000e+00
TPSA	1001.0	6.140000e+01	4.574000e+01	0.00	29.46	49.74	83.76	4.075000e+02
EState_VSA1	1001.0	1.246000e+01	1.876000e+01	0.00	0.00	5.60	18.12	1.874500e+02
EState_VSA10	1001.0	1.018000e+01	1.062000e+01	0.00	4.39	5.11	15.01	8.139000e+01
EState_VSA11	1001.0	1.400000e-01	9.500000e-01	0.00	0.00	0.00	0.00	1.317000e+01
EState_VSA2	1001.0	1.331000e+01	1.227000e+01	0.00	5.56	11.42	18.12	6.923000e+01
EState_VSA3	1001.0	1.633000e+01	1.388000e+01	0.00	5.92	12.18	22.25	9.326000e+01
EState_VSA4	1001.0	2.024000e+01	1.380000e+01	0.00	11.14	17.75	26.30	9.656000e+01
EState_VSA5	1001.0	1.907000e+01	2.247000e+01	0.00	5.56	12.68	25.68	2.703300e+02
EState_VSA6	1001.0	1.149000e+01	1.364000e+01	0.00	0.00	6.21	19.06	6.224000e+01
EState_VSA7	1001.0	1.395000e+01	1.992000e+01	0.00	0.00	4.90	24.27	1.268500e+02
EState_VSA8	1001.0	2.044000e+01	1.927000e+01	0.00	5.32	17.47	30.99	1.509500e+02
EState_VSA9	1001.0	8.890000e+00	8.480000e+00	0.00	4.42	5.73	13.89	5.661000e+01
VSA_EState1	1001.0	1.465000e+01	1.808000e+01	0.00	0.82	6.48	21.09	1.225900e+02
VSA_EState10	1001.0	1.000000e+00	2.230000e+00	0.00	0.00	0.00	0.00	1.555000e+01
VSA_EState2	1001.0	1.466000e+01	1.353000e+01	-0.59	3.91	12.48	22.74	6.730000e+01
VSA_EState3	1001.0	8.160000e+00	1.510000e+01	0.00	0.00	3.86	12.00	1.563600e+02

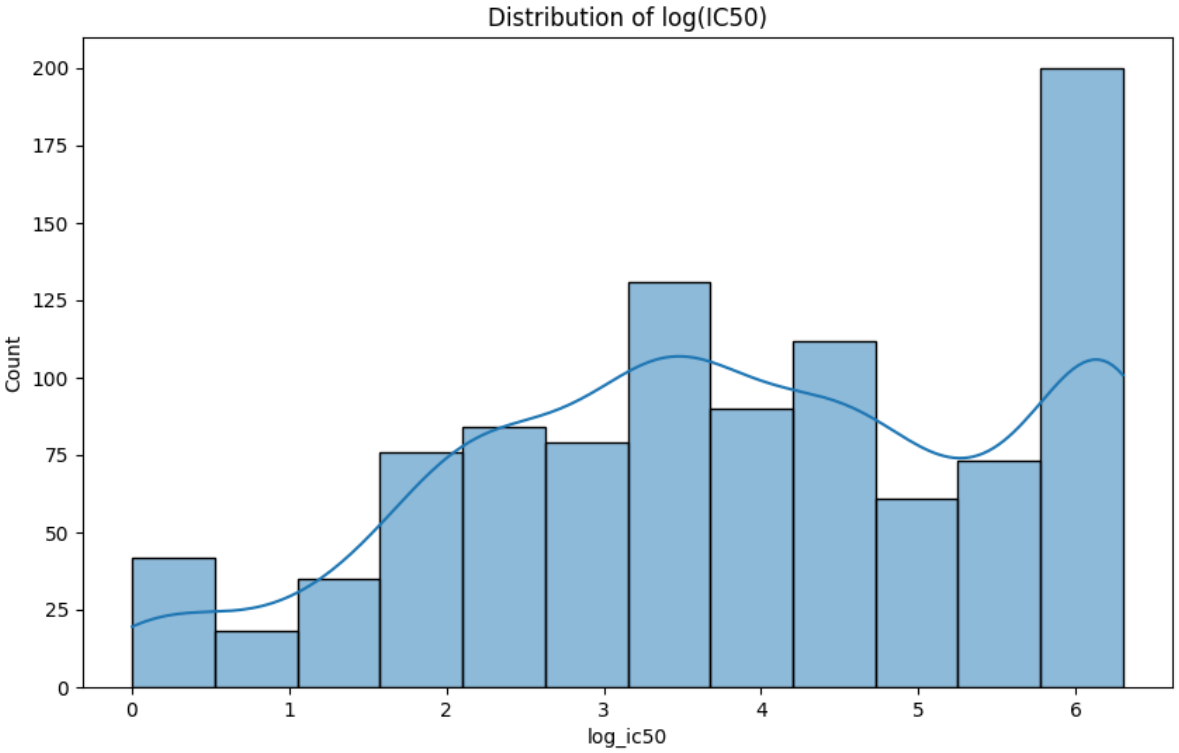


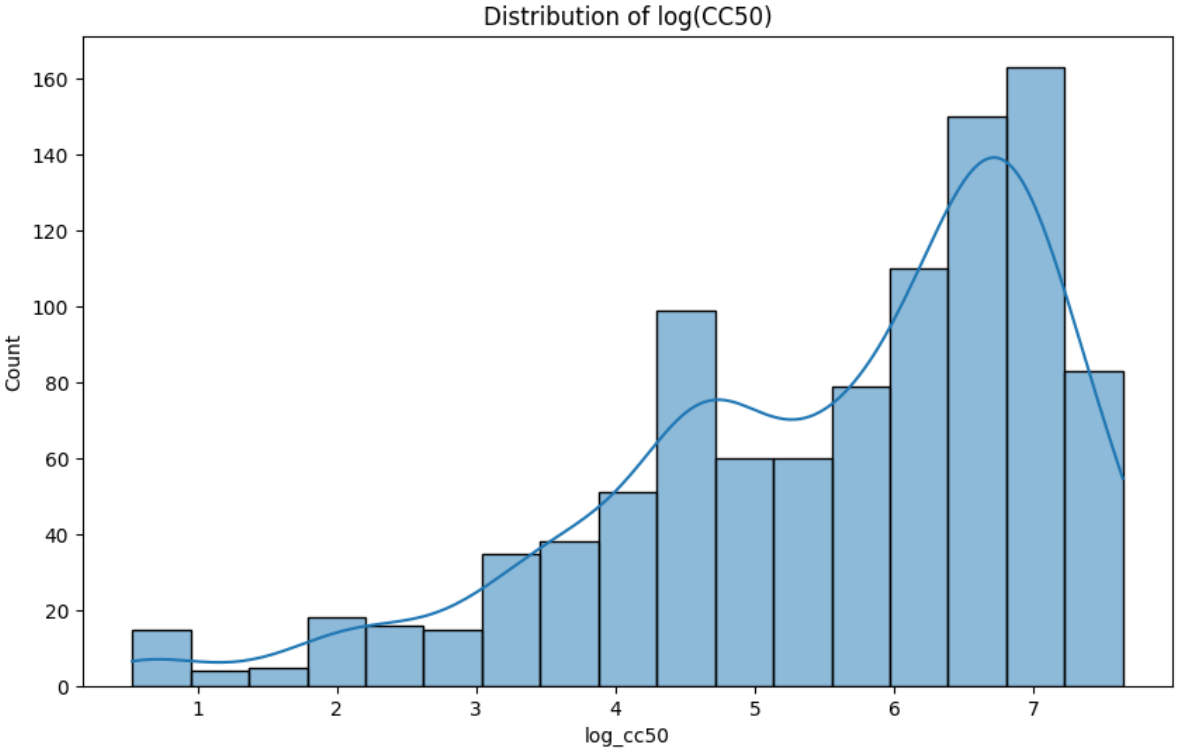
Обработка датасета

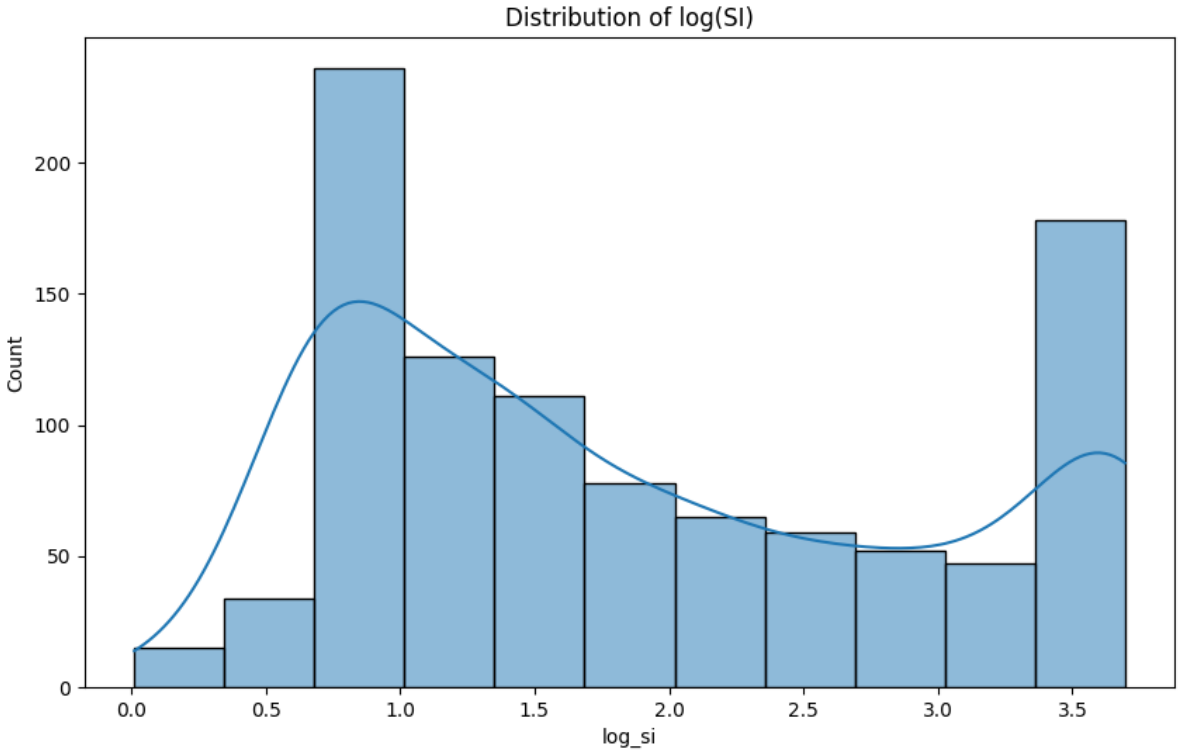
Обработка в `eda.py`:
1. Удалены дубликаты.
2. Переименованы столбцы: `IC50_mM`, `CC50_mM`, признаки — `feature_i`.
3. Выбросы обрезаны по IQR.
4. Пропуски заполнены медианой.
5. Значения ≤ 0 заменены на (10^{-6}).
6. Логарифмированы: `log_ic50`, `log_cc50`, `log_si`.
7. Удалены коррелированные признаки (>0.8).
8. Удалены признаки с низкой дисперсией (<0.01).
9. Исключены NaN/бесконечные значения.

Итог: `data/processed_data.csv`.

	count	mean	std	min	25%	50%	75%	max
feature_0	1001.0	5.000000e+02	2.891100e+02	0.00	250.00	500.00	750.00	1.000000e+03
feature_4	1001.0	1.083000e+01	3.310000e+00	2.32	9.25	12.18	13.17	1.593000e+01
feature_6	1001.0	1.800000e-01	1.700000e-01	0.00	0.05	0.12	0.29	1.370000e+00
feature_7	1001.0	-9.700000e-01	1.590000e+00	-6.99	-1.33	-0.42	0.06	1.370000e+00
feature_8	1001.0	5.800000e-01	2.100000e-01	0.06	0.44	0.63	0.74	9.500000e-01
feature_9	1001.0	2.949000e+01	1.274000e+01	9.42	18.49	29.29	38.75	6.027000e+01
feature_10	1001.0	3.482600e+02	1.269500e+02	110.16	264.32	315.46	409.28	9.047800e+02
feature_15	1001.0	2.400000e-01	1.300000e-01	-0.02	0.12	0.25	0.34	5.700000e-01
feature_19	1001.0	1.140000e+00	2.400000e-01	0.22	1.00	1.15	1.33	1.750000e+00
feature_22	1001.0	2.318000e+01	1.451000e+01	14.51	16.37	16.56	32.09	1.269100e+02
feature_23	1001.0	9.790000e+00	4.600000e-01	0.93	9.69	9.72	9.98	1.071000e+01
feature_24	1001.0	2.470000e+00	1.600000e-01	1.83	2.38	2.51	2.59	2.820000e+00
feature_25	1001.0	-2.350000e+00	1.700000e-01	-2.72	-2.48	-2.36	-2.24	-1.710000e+00
feature_29	1001.0	-7.000000e-02	2.500000e-01	-1.08	-0.16	-0.11	0.06	1.170000e+00
feature_30	1001.0	2.820000e+00	4.300000e-01	1.79	2.49	2.80	3.09	3.950000e+00
feature_31	1001.0	1.850000e+00	4.500000e-01	0.00	1.53	1.84	2.08	3.790000e+00
feature_46	1001.0	4.831703e+10	1.255969e+12	107.11	18858.35	112366.73	4399307.31	3.951781e+13
feature_51	1001.0	1.263000e+01	1.159000e+01	0.00	5.32	9.84	14.95	1.189100e+02
feature_52	1001.0	8.060000e+00	1.261000e+01	0.00	0.00	5.69	11.51	1.079200e+02
feature_53	1001.0	4.990000e+00	8.940000e+00	0.00	0.00	0.00	5.78	4.741000e+01
feature_54	1001.0	3.410000e+00	5.460000e+00	0.00	0.00	0.00	5.91	2.951000e+01
feature_55	1001.0	1.180000e+00	2.620000e+00	0.00	0.00	0.00	0.00	1.772000e+01
feature_56	1001.0	3.380000e+00	7.280000e+00	0.00	0.00	0.00	5.97	1.135600e+02
feature_57	1001.0	5.970000e+00	5.870000e+00	0.00	0.00	4.79	9.59	3.419000e+01
feature_58	1001.0	2.980000e+00	4.960000e+00	0.00	0.00	0.00	4.79	4.346000e+01
feature_59	1001.0	2.950000e+00	6.720000e+00	0.00	0.00	0.00	0.00	4.461000e+01
feature_60	1001.0	1.760000e+00	4.390000e+00	0.00	0.00	0.00	0.00	2.352000e+01
feature_61	1001.0	2.641000e+01	2.063000e+01	0.00	12.13	22.67	37.82	1.289000e+02
feature_62	1001.0	4.152000e+01	2.182000e+01	0.00	30.18	38.06	51.37	1.610600e+02
feature_63	1001.0	1.649000e+01	1.080000e+01	0.00	6.92	17.02	23.17	5.228000e+01
feature_64	1001.0	1.476000e+01	1.253000e+01	0.00	5.69	12.29	19.76	6.545000e+01
feature_66	1001.0	1.512000e+01	1.323000e+01	0.00	5.71	11.68	22.89	6.841000e+01
feature_67	1001.0	6.000000e-02	5.600000e-01	0.00	0.00	0.00	0.00	5.530000e+00
feature_68	1001.0	5.310000e+00	6.670000e+00	0.00	0.00	4.90	9.80	5.906000e+01
feature_69	1001.0	1.177000e+01	1.223000e+01	0.00	0.00	5.92	17.75	5.736000e+01
feature_70	1001.0	4.165000e+01	2.959000e+01	0.00	19.28	40.03	57.53	1.743700e+02
feature_71	1001.0	1.347000e+01	1.623000e+01	0.00	0.00	7.11	18.41	1.047200e+02
feature_74	1001.0	5.730000e+00	1.097000e+01	0.00	0.00	0.00	5.75	5.750000e+01
feature_75	1001.0	6.070000e+00	6.170000e+00	0.00	0.00	5.32	10.17	3.316000e+01
feature_78	1001.0	4.070000e+00	7.950000e+00	0.00	0.00	0.00	7.60	4.588000e+01
feature_79	1001.0	3.307000e+01	2.379000e+01	0.00	16.00	27.95	45.35	2.131400e+02
feature_80	1001.0	8.220000e+00	7.720000e+00	0.00	4.74	6.18	12.35	4.355000e+01
feature_84	1001.0	3.000000e-01	1.390000e+00	0.00	0.00	0.00	0.00	1.964000e+01
feature_85	1001.0	4.030000e+00	7.190000e+00	0.00	0.00	0.00	6.08	4.504000e+01
feature_90	1001.0	1.400000e-01	9.500000e-01	0.00	0.00	0.00	0.00	1.317000e+01
feature_91	1001.0	1.331000e+01	1.227000e+01	0.00	5.56	11.42	18.12	6.923000e+01
feature_92	1001.0	1.633000e+01	1.388000e+01	0.00	5.92	12.18	22.25	9.326000e+01
feature_93	1001.0	2.024000e+01	1.380000e+01	0.00	11.14	17.75	26.30	9.656000e+01
feature_94	1001.0	1.907000e+01	2.247000e+01	0.00	5.56	12.68	25.68	2.703300e+02
feature_95	1001.0	1.149000e+01	1.364000e+01	0.00	0.00	6.21	19.06	6.224000e+01
feature_96	1001.0	1.395000e+01	1.992000e+01	0.00	0.00	4.90	24.27	1.268500e+02
feature_97	1001.0	2.044000e+01	1.927000e+01	0.00	5.32	17.47	30.99	1.509500e+02
feature_98	1001.0	8.890000e+00	8.480000e+00	0.00	4.42	5.73	13.89	5.661000e+01
feature_99	1001.0	1.465000e+01	1.808000e+01	0.00	0.82	6.48	21.09	1.225900e+02
feature_101	1001.0	1.466000e+01	1.353000e+01	-0.59	3.91	12.48	22.74	6.730000e+01
feature_102	1001.0	9.160000e+00	1.519000e+01	-1.26	0.00	3.86	12.00	1.563600e+02
feature_103	1001.0	2.170000e+00	2.910000e+00	-4.58	0.19	1.95	4.15	1.236000e+01
feature_104	1001.0	2.000000e-02	2.560000e+00	-20.30	-0.84	0.63	1.44	6.580000e+00
feature_106	1001.0	3.230000e+00	5.340000e+00	-33.09	0.09	3.61	5.89	3.392000e+01
feature_107	1001.0	5.410000e+00	5.710000e+00	-2.21	1.30	4.24	7.61	3.698000e+01
feature_108	1001.0	5.600000e-01	1.680000e+00	-7.68	0.00	0.00	0.00	1.018000e+01
feature_114	1001.0	8.600000e-01	1.040000e+00	0.00	0.00	1.00	2.00	6.000000e+00
feature_115	1001.0	2.090000e+00	1.360000e+00	0.00	1.00	2.00	3.00	7.000000e+00
feature_117	1001.0	5.100000e-01	7.300000e-01	0.00	0.00	0.00	1.00	5.000000e+00
feature_124	1001.0	5.300000e-01	8.000000e-01	0.00	0.00	0.00	1.00	6.000000e+00
feature_126	1001.0	3.560000e+00	1.570000e+00	0.00	2.00	3.00	4.00	9.000000e+00
feature_127	1001.0	3.440000e+00	2.140000e+00	-5.75	2.45	3.42	4.53	1.282000e+01
feature_129	1001.0	6.000000e-02	2.300000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_132	1001.0	1.000000e-02	1.300000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_135	1001.0	4.000000e-02	1.900000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_136	1001.0	1.900000e-01	7.700000e-01	0.00	0.00	0.00	0.00	8.000000e+00
feature_141	1001.0	4.000000e-02	2.000000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_143	1001.0	1.600000e-01	4.400000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_145	1001.0	3.900000e-01	6.000000e-01	0.00	0.00	0.00	1.00	3.000000e+00
feature_146	1001.0	1.300000e-01	3.400000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_148	1001.0	7.000000e-02	3.000000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_149	1001.0	8.000000e-02	2.900000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_153	1001.0	1.000000e-02	1.100000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_155	1001.0	3.300000e-01	8.000000e-01	0.00	0.00	0.00	0.00	8.000000e+00
feature_156	1001.0	3.400000e-01	7.000000e-01	0.00	0.00	0.00	0.00	4.000000e+00
feature_158	1001.0	3.000000e-01	6.600000e-01	0.00	0.00	0.00	0.00	4.000000e+00
feature_159	1001.0	1.600000e-01	4.200000e-01	0.00	0.00	0.00	0.00	3.000000e+00
feature_165	1001.0	1.450000e+00	1.430000e+00	0.00	0.00	1.00	2.00	7.000000e+00
feature_169	1001.0	2.200000e-01	4.800000e-01	0.00	0.00	0.00	0.00	4.000000e+00
feature_170	1001.0	1.010000e+00	1.410000e+00	0.00	0.00	1.00	1.00	8.000000e+00
feature_171	1001.0	5.000000e-02	2.300000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_175	1001.0	7.000000e-02	2.500000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_176	1001.0	5.000000e-02	2.200000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_177	1001.0	3.000000e-02	1.600000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_180	1001.0	1.900000e-01	5.100000e-01	0.00	0.00	0.00	0.00	3.000000e+00
feature_183	1001.0	4.000000e-02	2.300000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_184	1001.0	2.700000e-01	7.300000e-01	0.00	0.00	0.00	0.00	4.000000e+00
feature_185	1001.0	8.000000e-02	3.700000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_187	1001.0	3.000000e-02	1.600000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_188	1001.0	1.000000e-02	1.100000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_192	1001.0	1.000000e-02	1.000000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_193	1001.0	1.800000e-01	4.800000e-01	0.00	0.00	0.00	0.00	4.000000e+00
feature_198	1001.0	6.000000e-02	2.700000e-01	0.00	0.00	0.00	0.00	2.000000e+00
feature_199	1001.0	2.000000e-02	2.200000e-01	0.00	0.00	0.00	0.00	6.000000e+00
feature_200	1001.0	2.000000e-02	1.500000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_202	1001.0	3.000000e-02	1.700000e-01	0.00	0.00	0.00	0.00	1.000000e+00
feature_203	1001.0	5.000000e-02	3.000000e-01	0.00	0.00	0.00	0.00	4.000000e+00


Distribution of log(IC50)
Count
log_ic50
0
25
50
75
100
125
150
175
200
0
1
2
3
4
5
6


Distribution of log(CC50)
Count
log_cc50
0
20
40
60
80
100
120
140
160
1
2
3
4
5
6
7



Методология

Модели и метрики

Модели: Random Forest, XGBoost, LightGBM, Gradient Boosting, Linear/Logistic Regression, Voting.
- **Регрессия**: MSE, R^2, MAE.
- **Классификация**: Accuracy, F1, Precision, Recall, ROC*AUC, PR*AUC.

Подготовка данных

80/20 разделение, StandardScaler, SMOTE при дисбалансе >10%, Optuna (50 испытаний).

Визуализации

Матрицы ошибок, ROC-кривые, важность признаков, предсказания.

Результаты

Регрессия

log_cc50 (`regression_log_cc50.csv`)

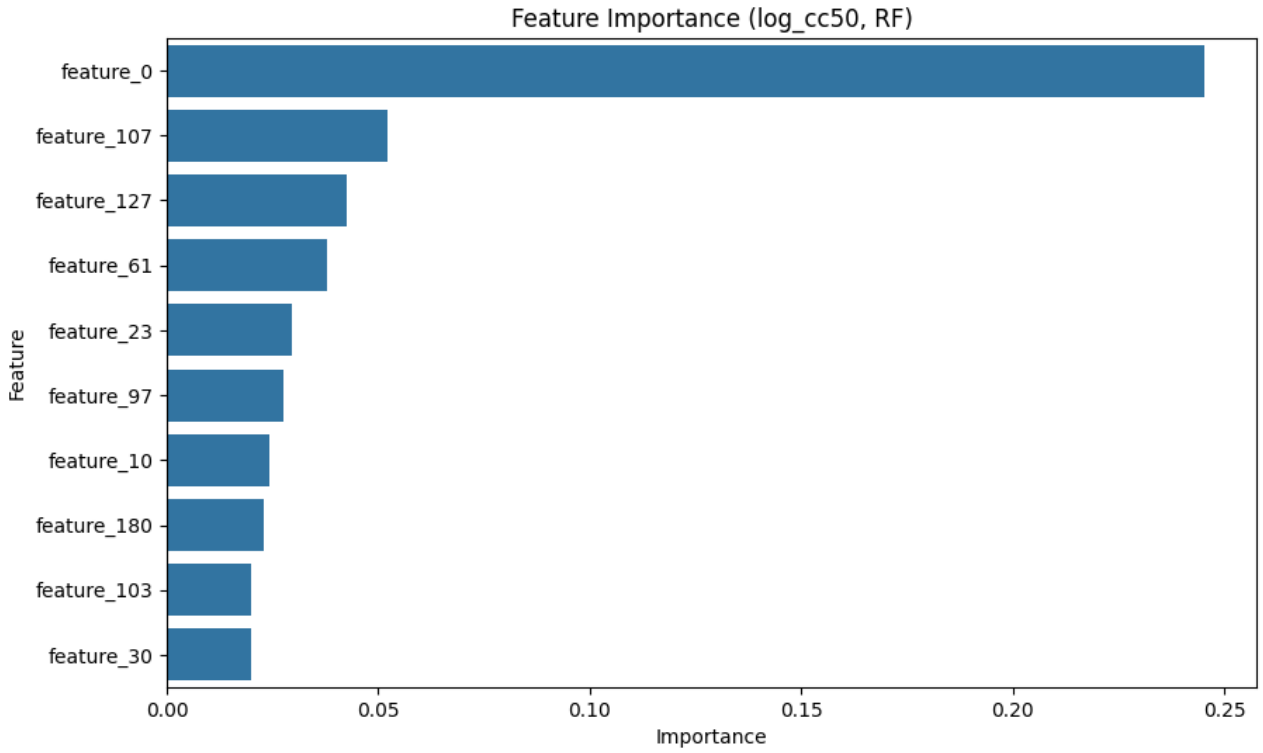
Model	MSE	R2	MAE
RF	1.154	0.476	0.762
XGB	1.201	0.455	0.764

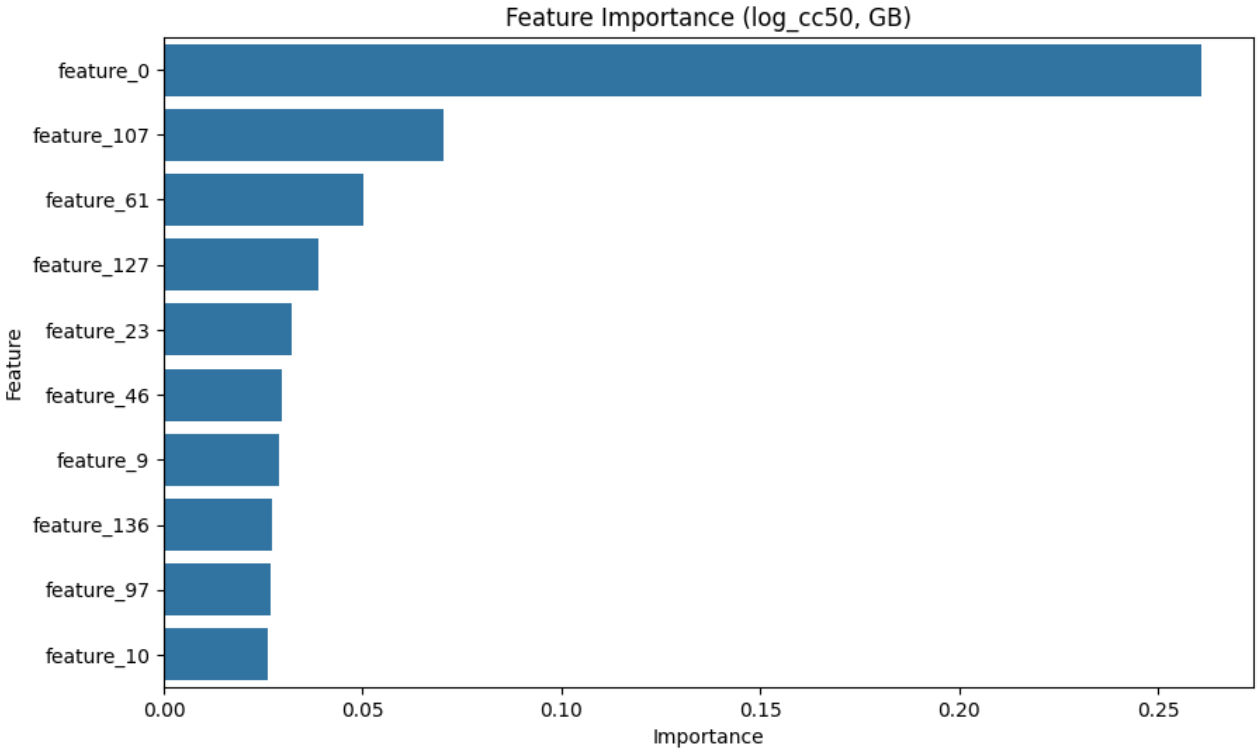
Model	MSE	R2	MAE
LGB	1.24	0.437	0.79
GB	1.142	0.482	0.783
LR	1.493	0.322	0.944
Voting	1.099	0.501	0.749

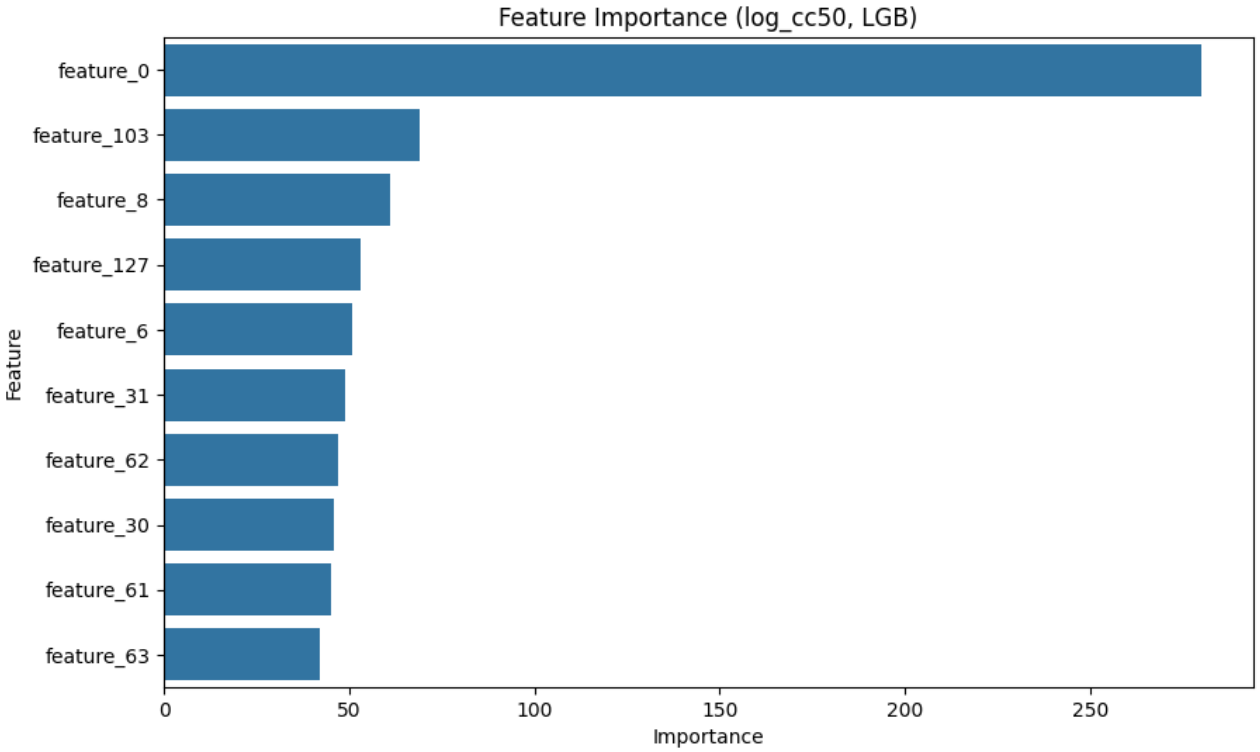
Лучшая модель: Voting (R^2=0.501, MSE=1.099, MAE=0.749).

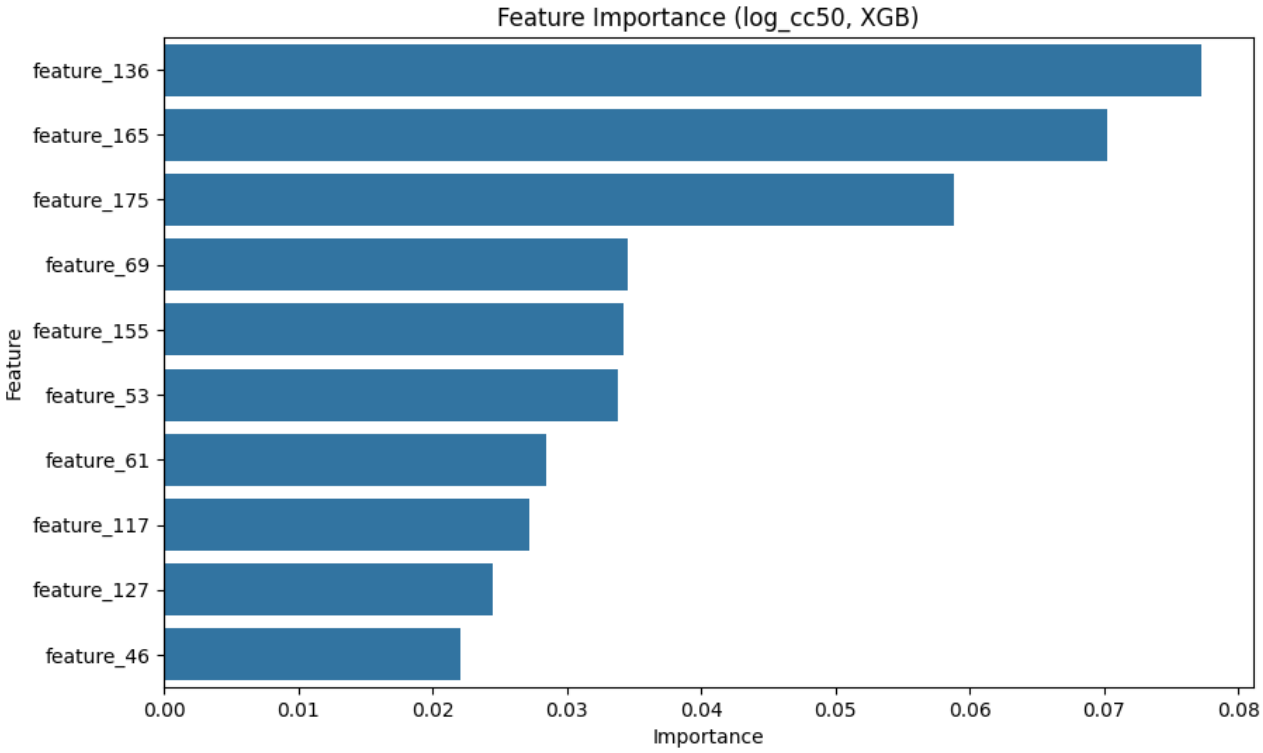
Рекомендации для log_cc50

Для log_cc50 R^2 приемлемый (0.501). Можно улучшить:
- Провести дополнительную настройку гиперпараметров.
- Добавить новые признаки через feature engineering.




Feature Importance (log_cc50, GB)
feature_0
feature_107
feature_61
feature_127
feature_23
feature_46
feature_9
feature_136
feature_97
feature_10
Feature
0.00
0.05
0.10
0.15
0.20
0.25
Importance


Feature Importance (log_cc50, LGB)
feature_0
feature_103
feature_8
feature_127
feature_6
feature_31
feature_62
feature_30
feature_61
feature_63
Feature
0
50
100
150
200
250
Importance



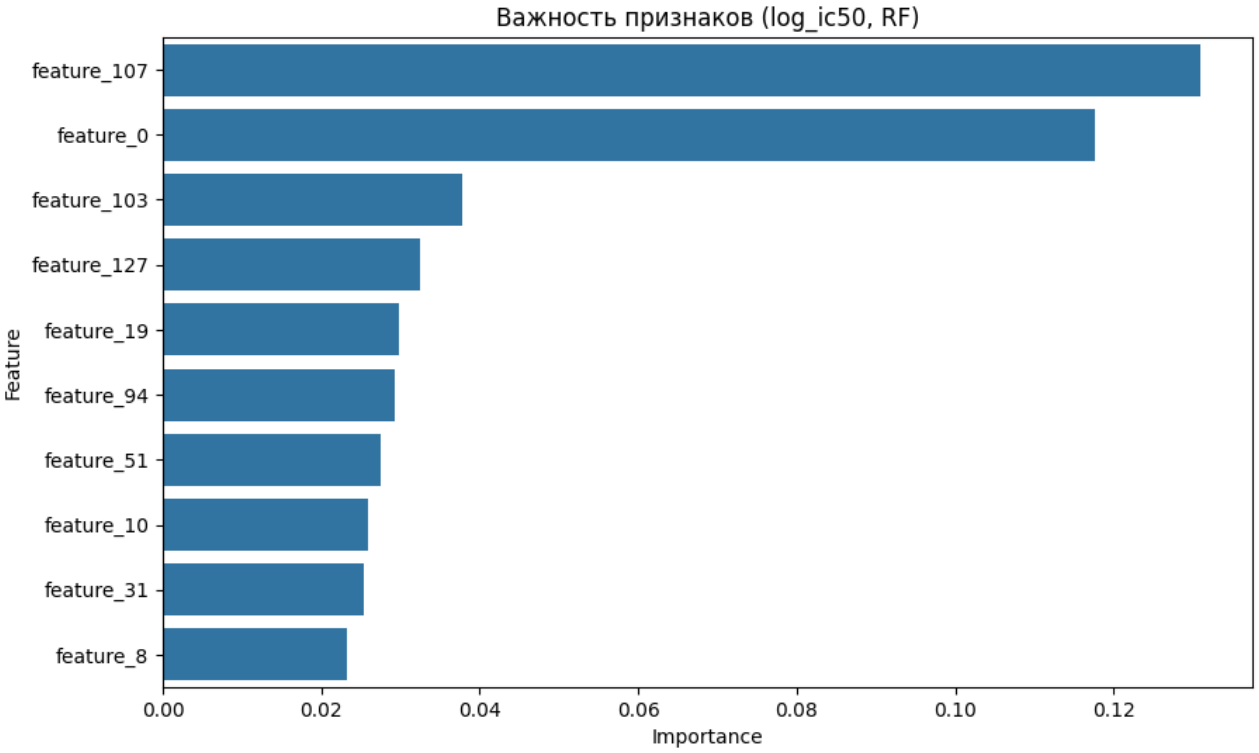
log_ic50 (`regression_log_ic50.csv`)

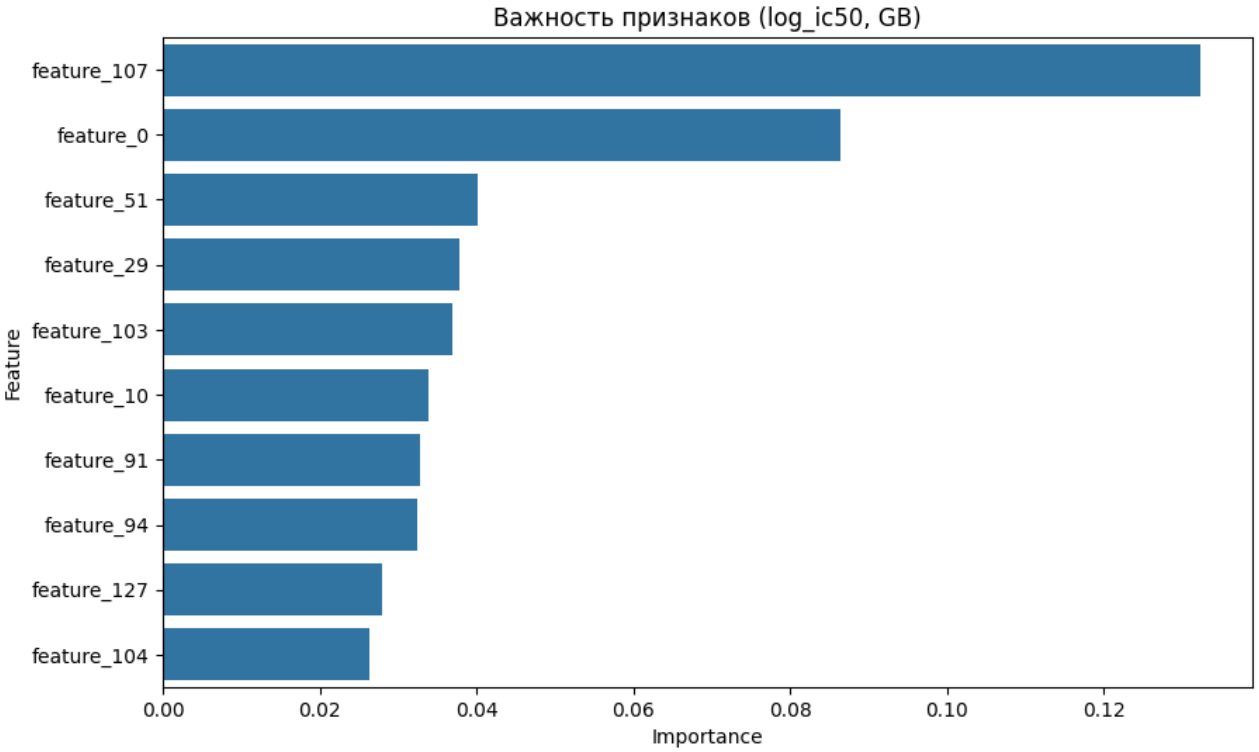
Model	MSE	R2	MAE
RF	1.667	0.476	0.991
XGB	1.747	0.451	1.045
LGB	1.689	0.469	1.022
GB	1.911	0.399	1.095
LR	2.391	0.248	1.256
Voting	1.802	0.433	1.042

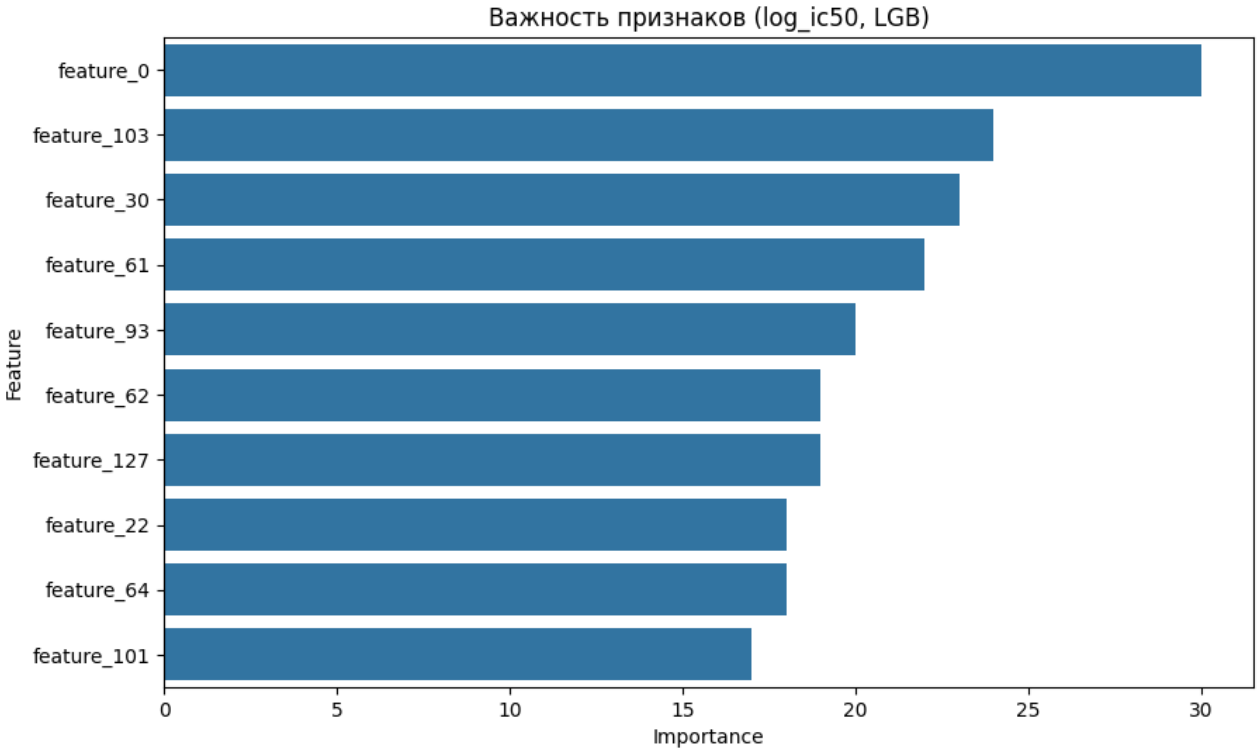
Лучшая модель: RF (R^2=0.476, MSE=1.667, MAE=0.991).

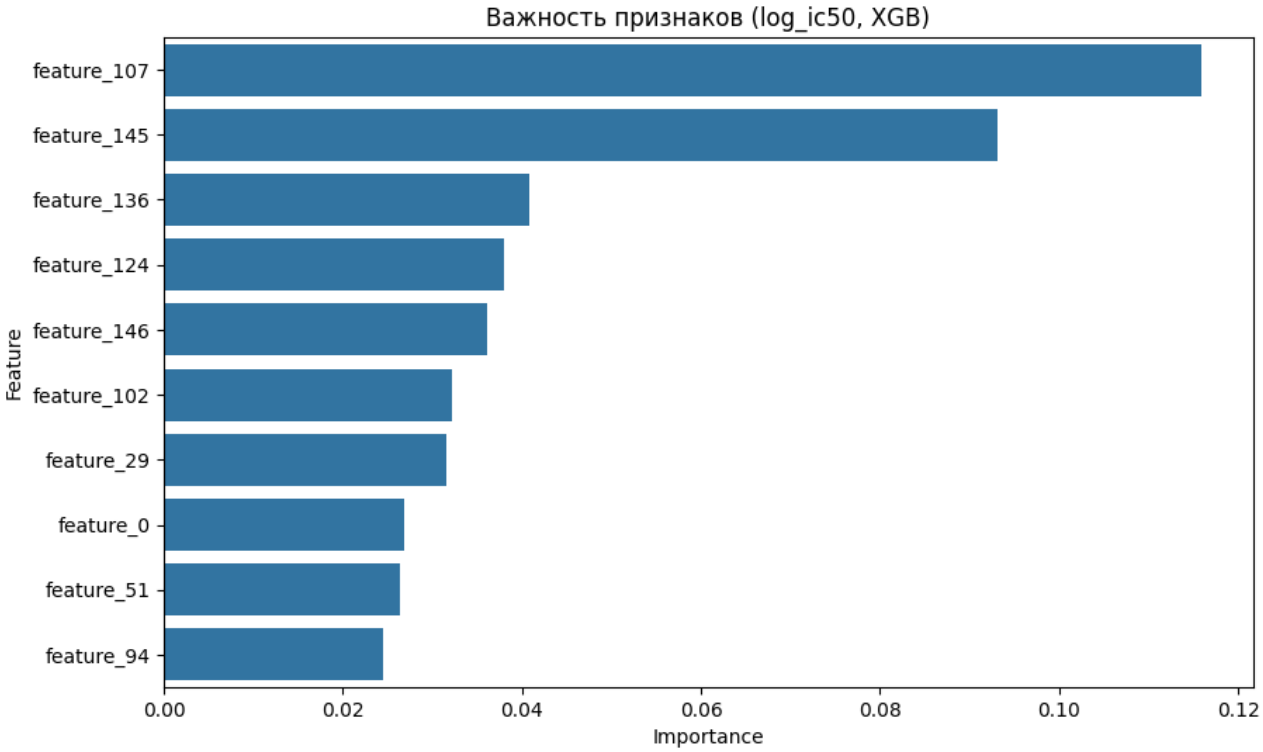
Рекомендации для log_ic50

Для log_ic50 R^2 низкий (0.476). Рекомендуется:
- Проверить данные на выбросы с помощью Isolation Forest.
- Применить SMOTE для балансировки данных, если наблюдается дисбаланс.
- Рассмотреть PCA для снижения размерности признаков.
- Использовать более сложные модели, например, Stacking Regressor.


Важность признаков (log_ic50, RF)
feature_107
feature_0
feature_103
feature_127
feature_19
feature_94
feature_51
feature_10
feature_31
feature_8
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance


Важность признаков (log_ic50, GB)
feature_107
feature_0
feature_51
feature_29
feature_103
feature_10
feature_91
feature_94
feature_127
feature_104
Feature
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Importance





log_si (`regression_log_si.csv`)

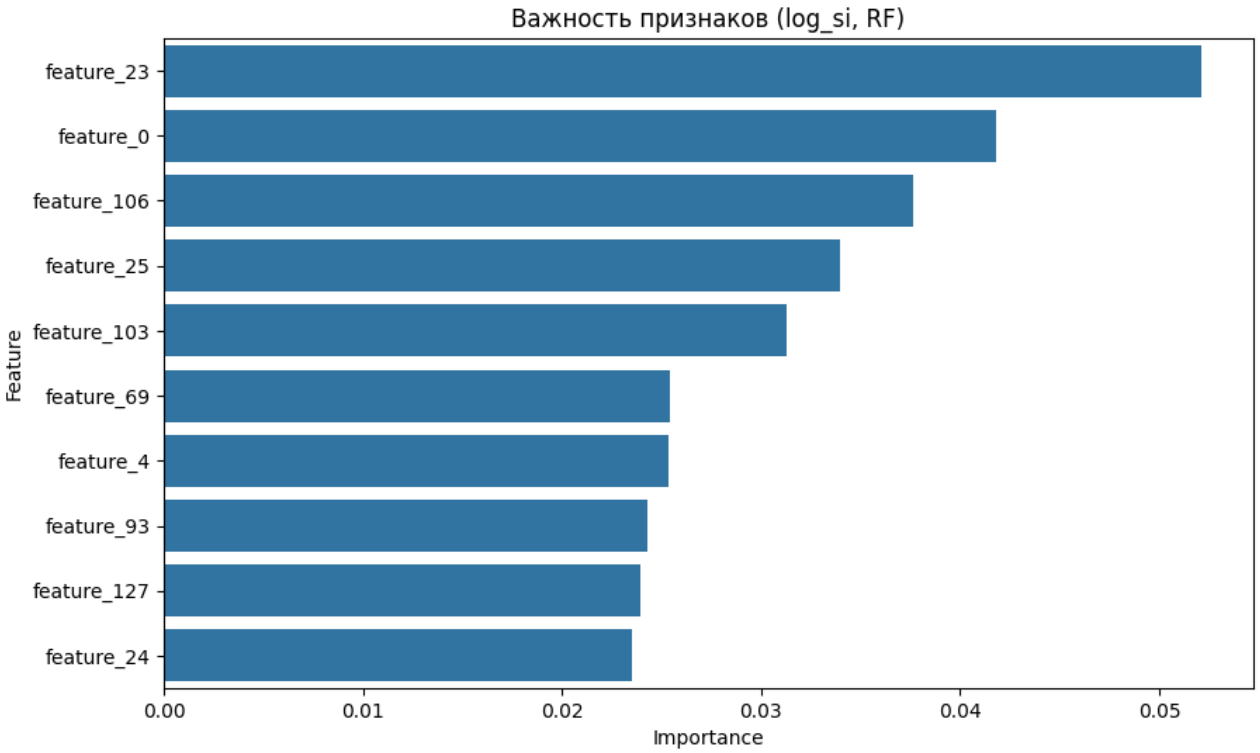
Model	MSE	R2	MAE
RF	0.883	0.261	0.757

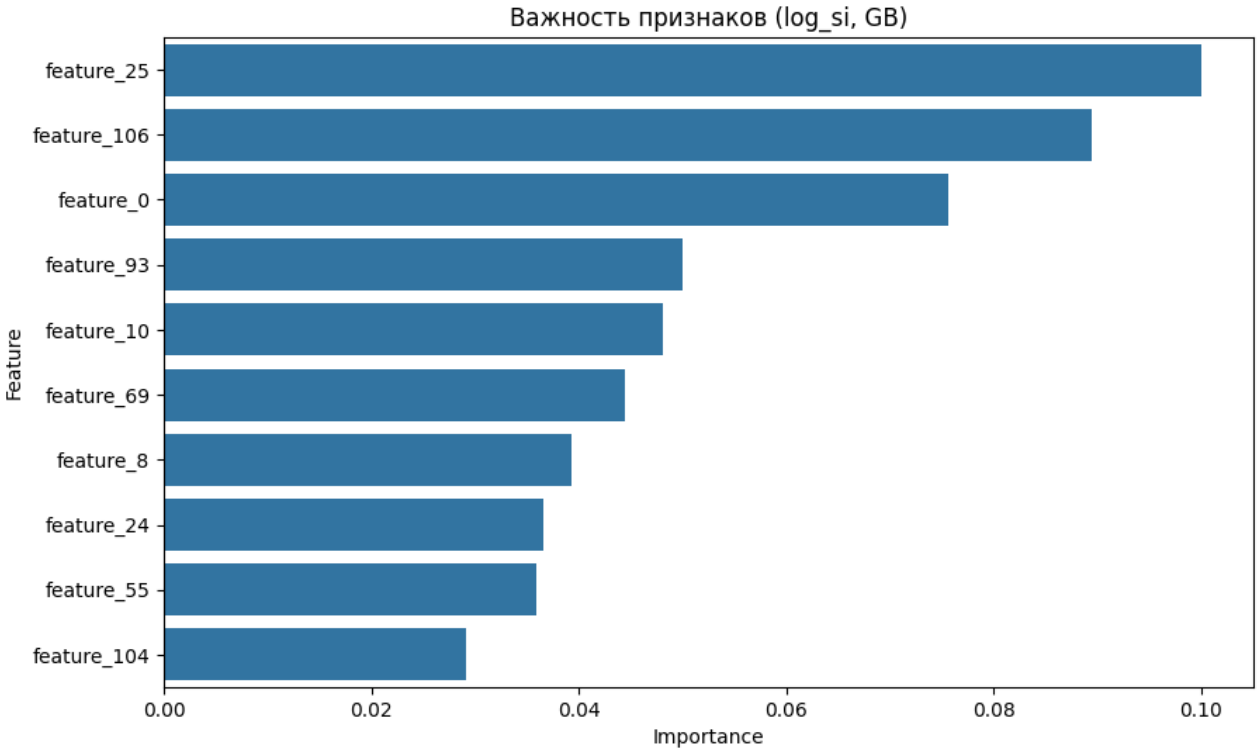
Model	MSE	R2	MAE
XGB	0.962	0.195	0.789
LGB	0.911	0.238	0.762
GB	0.932	0.22	0.79
LR	1.13	0.055	0.869
Voting	0.926	0.226	0.749

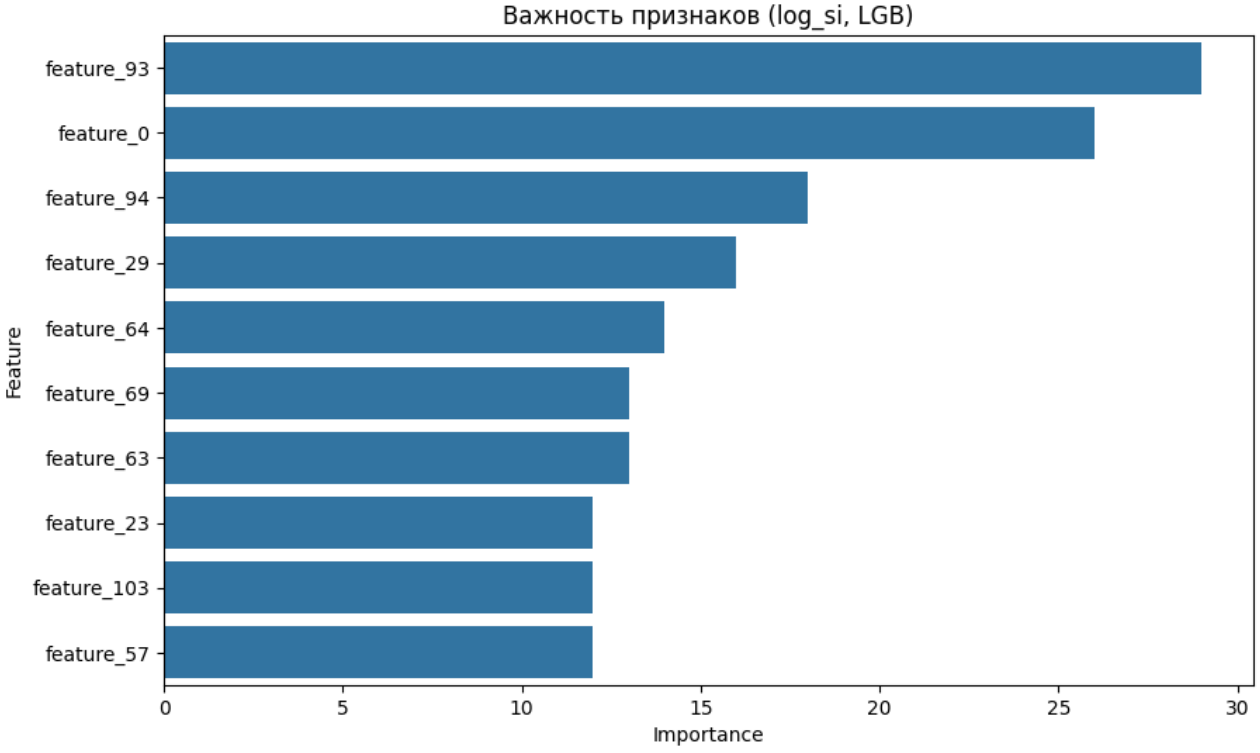
Лучшая модель: RF (R^2=0.261, MSE=0.883, MAE=0.757).

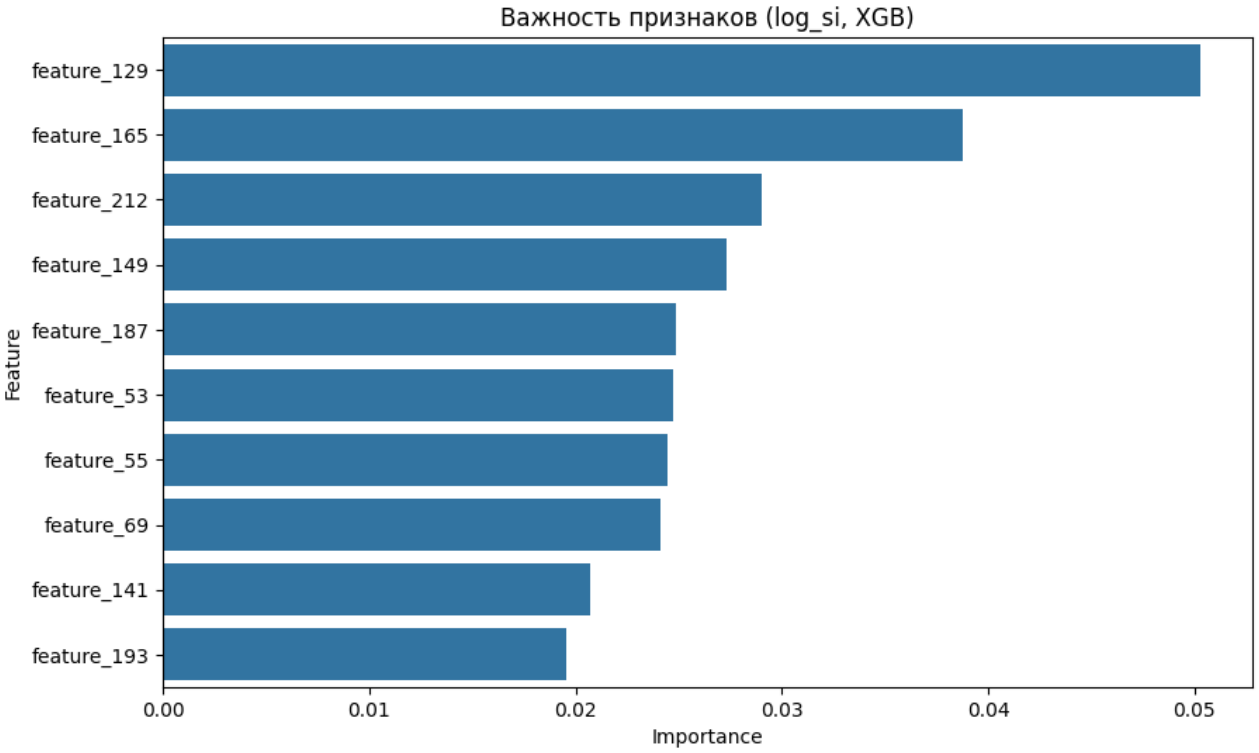
Рекомендации для log_si

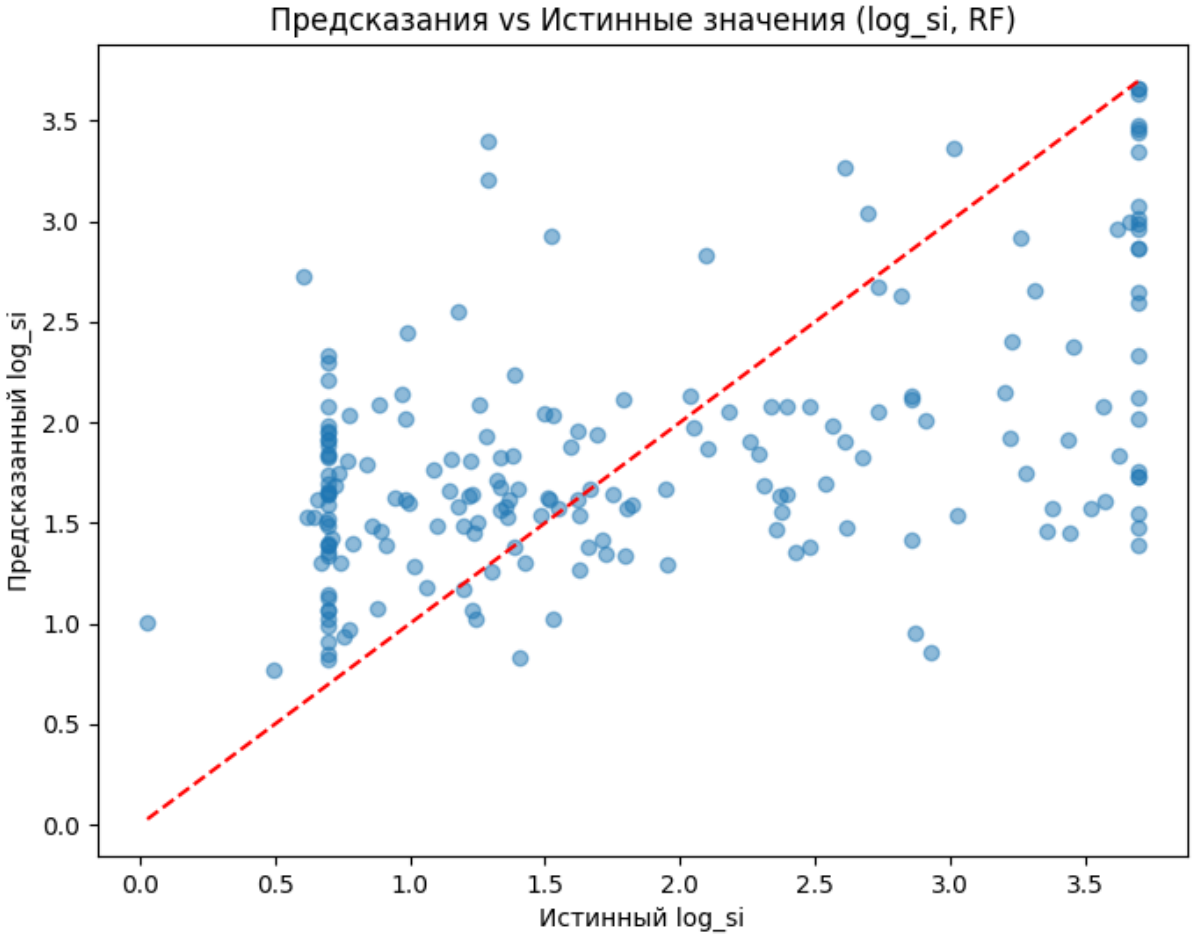
Для log_si R^2 низкий (0.261). Рекомендуется:
- Проверить данные на выбросы с помощью Isolation Forest.
- Применить SMOTE для балансировки данных, если наблюдается дисбаланс.
- Рассмотреть PCA для снижения размерности признаков.
- Использовать более сложные модели, например, Stacking Regressor.




Важность признаков (log_si, GB)
feature_25
feature_106
feature_0
feature_93
feature_10
feature_69
feature_8
feature_24
feature_55
feature_104
Feature
0.00
0.02
0.04
0.06
0.08
0.10
Importance


Важность признаков (log_si, LGB)
feature_93
feature_0
feature_94
feature_29
feature_64
feature_69
feature_63
feature_23
feature_103
feature_57
Feature
0
5
10
15
20
25
30
Importance


Важность признаков (log_si, XGB)
feature_129
feature_165
feature_212
feature_149
feature_187
feature_53
feature_55
feature_69
feature_141
feature_193
Feature
0.00
0.01
0.02
0.03
0.04
0.05
Importance


Предсказания vs Истинные значения (log_si, RF)
Предсказанный log_si
Истинный log_si
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Классификация

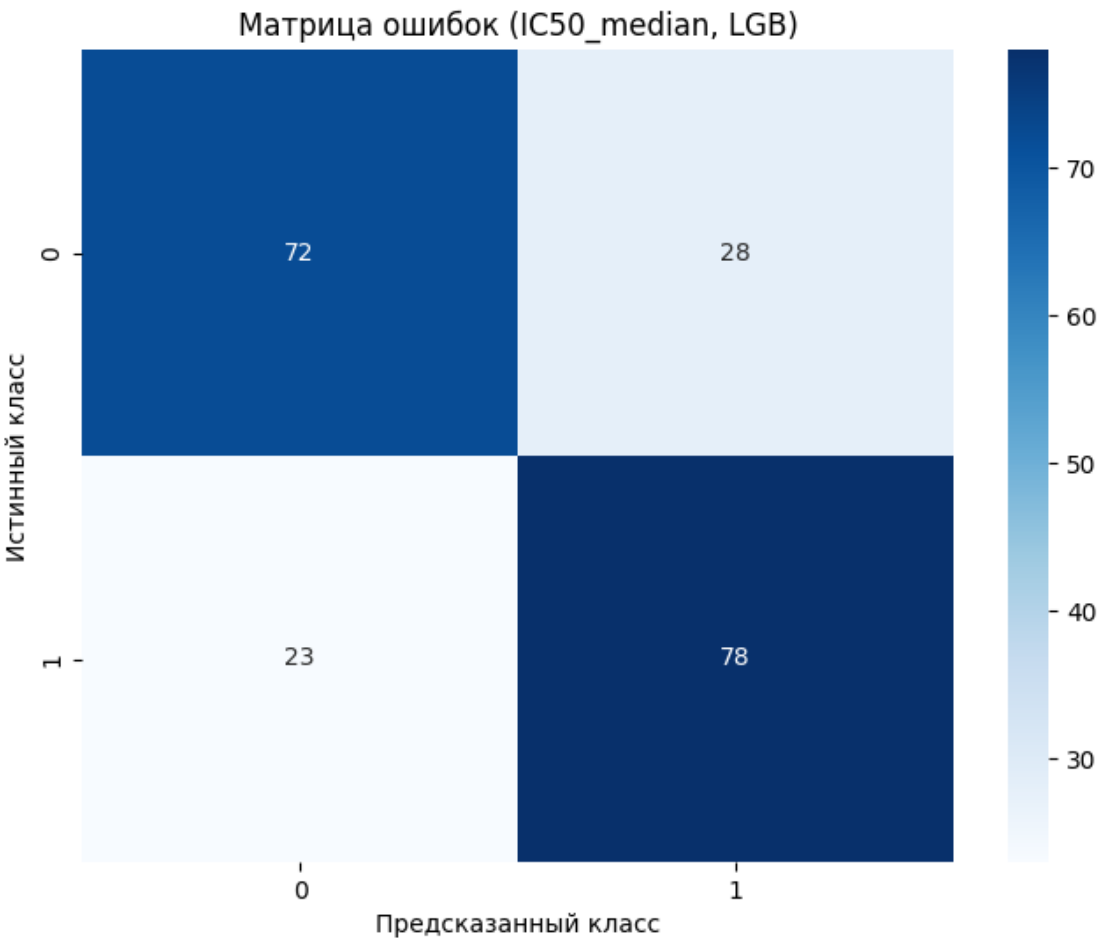
IC50_median (`classification_ic50_median.csv`)

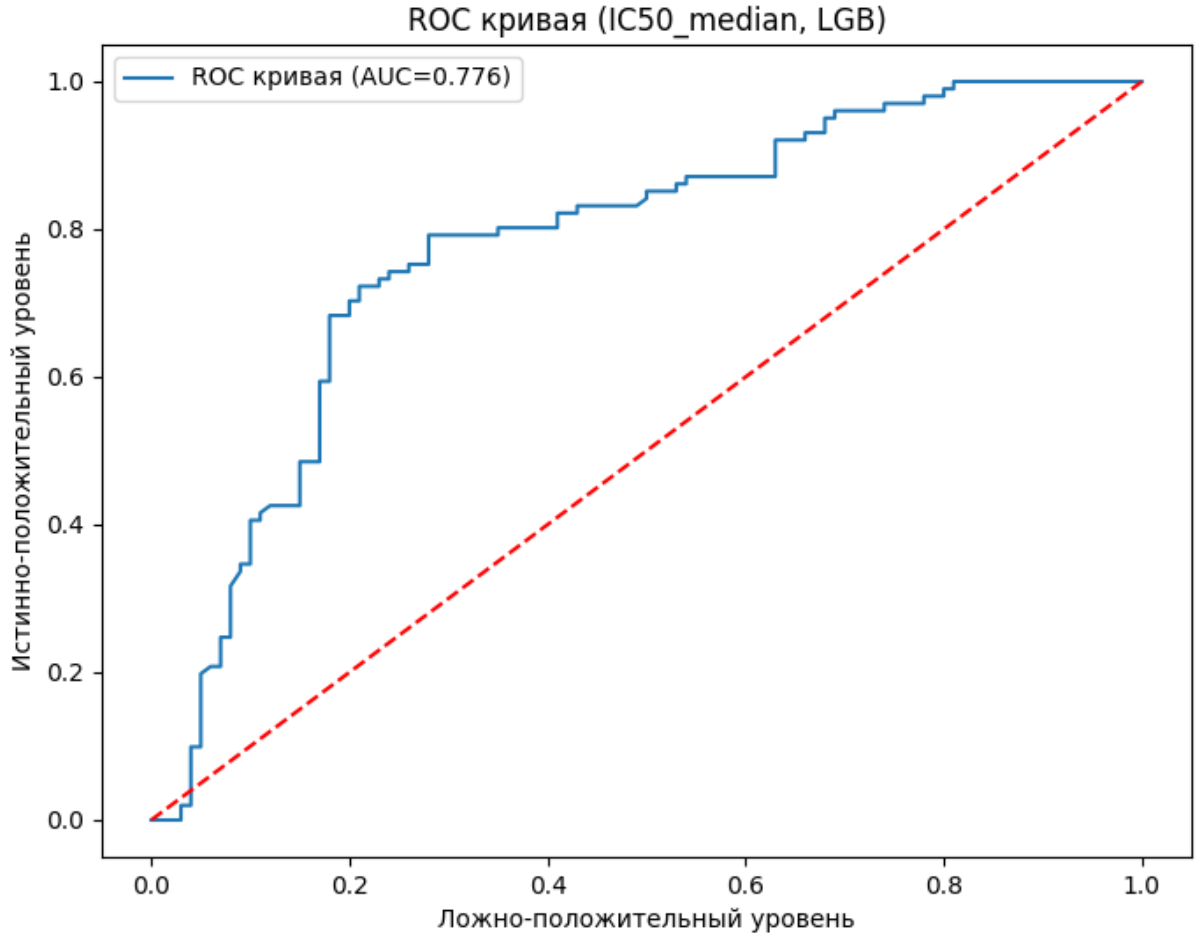
Model	Accuracy	F1	Precision	Recall	ROC_AUC	PR_AUC
RF	0.731	0.74	0.72	0.762	0.777	0.732
XGB	0.706	0.715	0.698	0.733	0.76	0.716
LGB	0.746	0.754	0.736	0.772	0.776	0.713
GB	0.726	0.732	0.721	0.743	0.762	0.737
LR	0.667	0.676	0.66	0.693	0.736	0.727
Voting	0.726	0.739	0.709	0.772	0.783	0.741

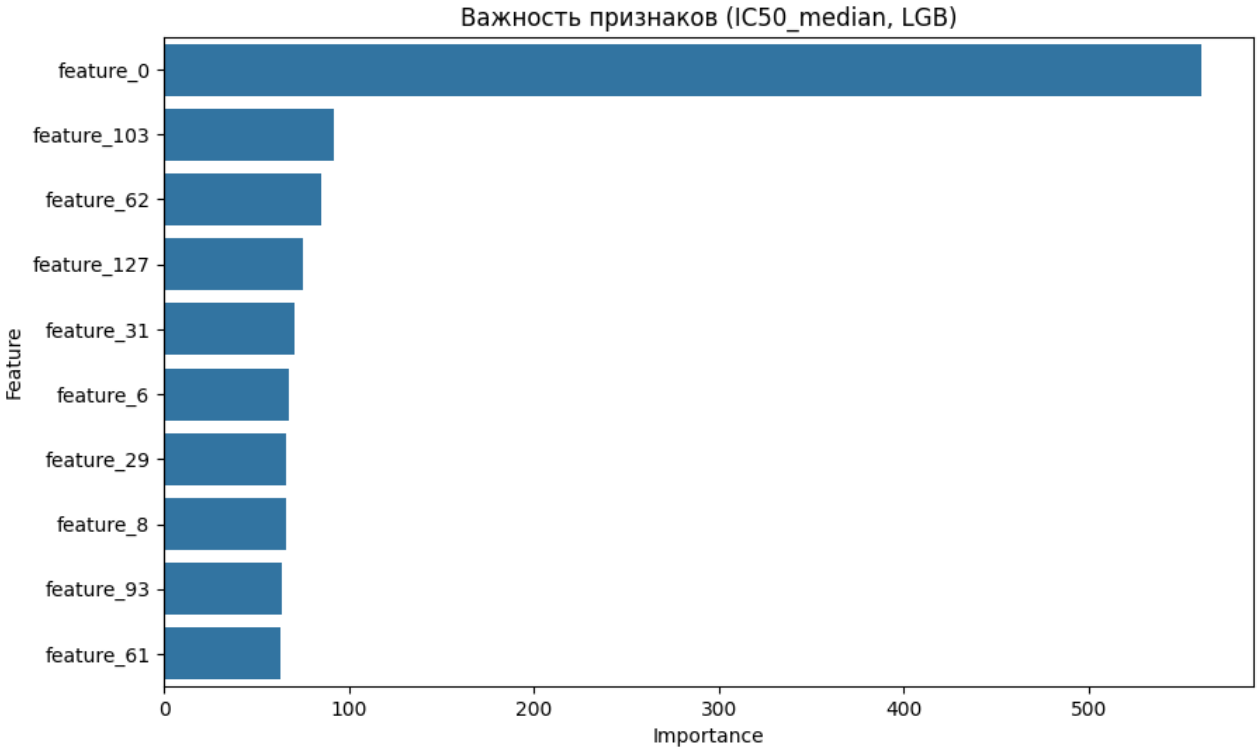
Лучшая модель: LGB (F1=0.754, Accuracy=0.746, Precision=0.736, Recall=0.772, ROC_AUC=0.776).

Рекомендации для IC50_median

Для IC50_median метрики приемлемые (F1=0.754, Precision=0.736, Recall=0.772). Можно улучшить:
- Провести дополнительную настройку гиперпараметров.
- Проверить важность признаков для исключения лишних.


Матрица ошибок (IC50_median, LGB)
72
28
23
78
0
1
Истинный класс
0
1
Предсказанный класс
70
60
50
40
30


ROC кривая (IC50_median, LGB)
ROC кривая (AUC=0.776)
Истинно-положительный уровень
Ложно-положительный уровень
0.0
0.2
0.4
0.6
0.8
1.0


Важность признаков (IC50_median, LGB)
feature_0
feature_103
feature_62
feature_127
feature_31
feature_6
feature_29
feature_8
feature_93
feature_61
Feature
0
100
200
300
400
500
Importance

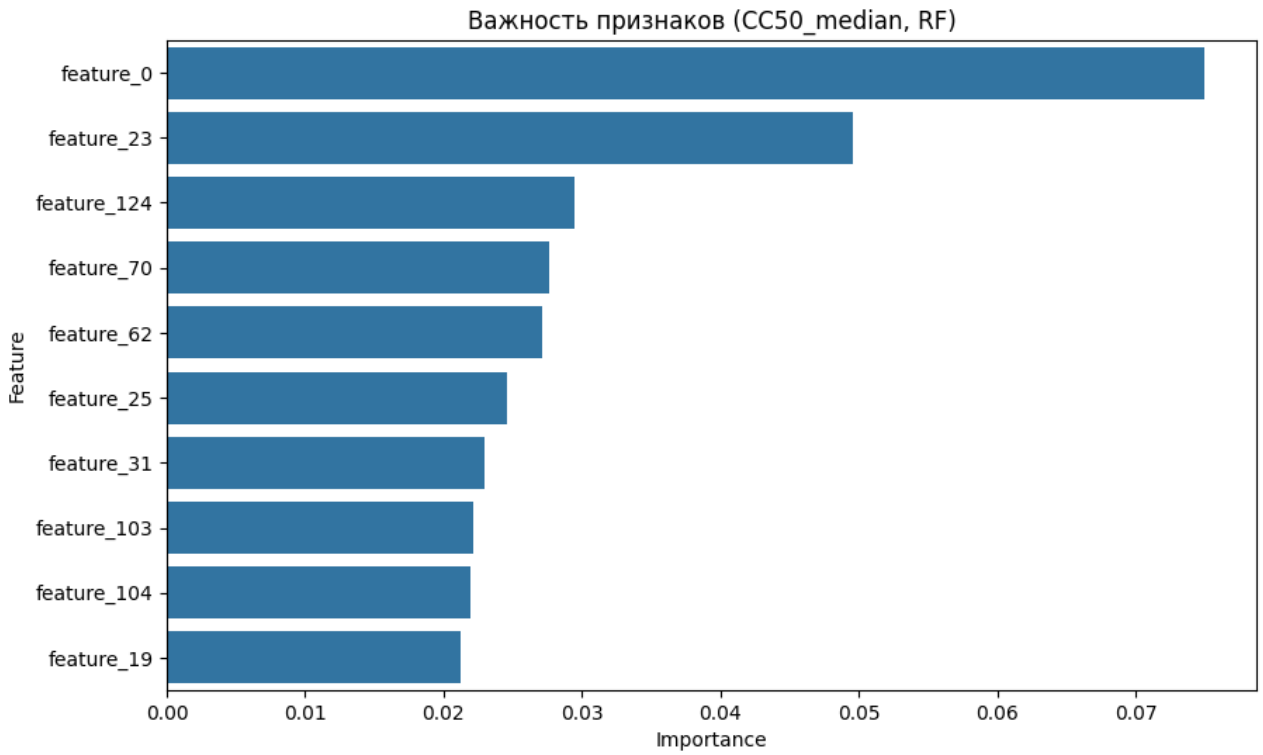
CC50_median (`classification_cc50_median.csv`)

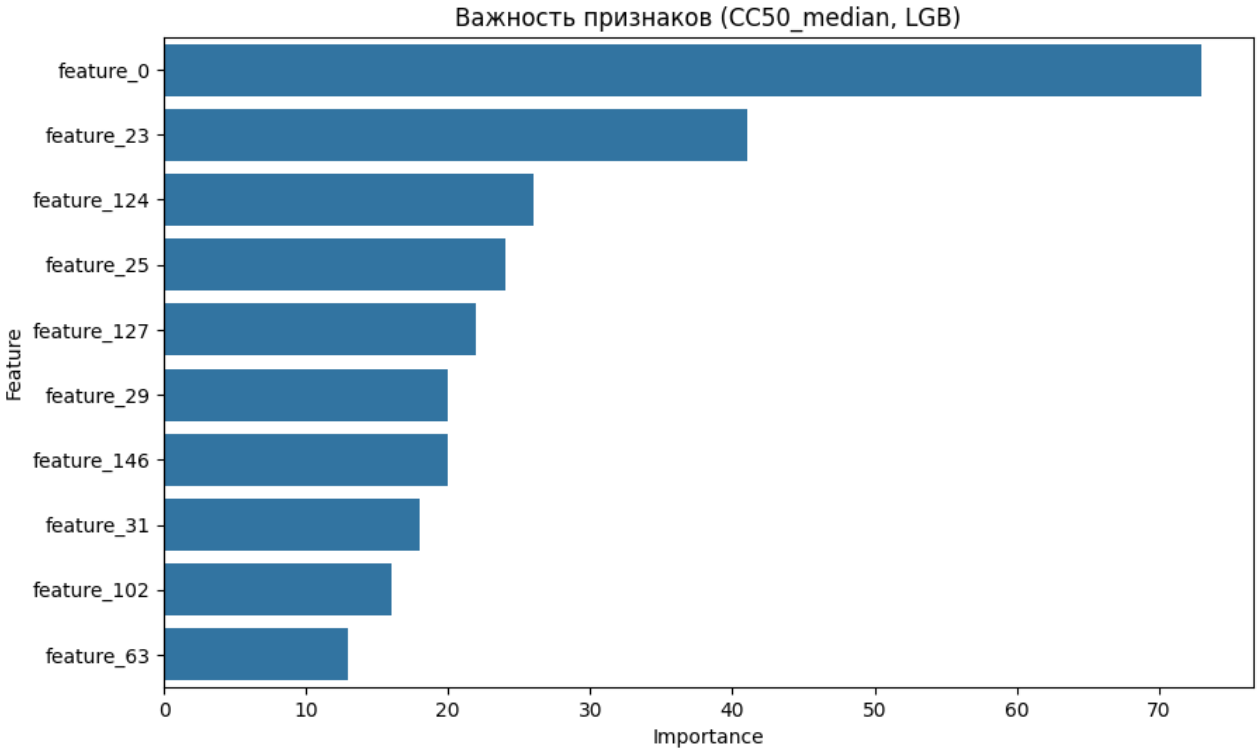
Model	Accuracy	F1	Precision	Recall	ROC_AUC	PR_AUC
RF	0.761	0.767	0.752	0.782	0.852	0.858
XGB	0.741	0.764	0.706	0.832	0.826	0.812
LGB	0.741	0.75	0.729	0.772	0.836	0.836
GB	0.766	0.771	0.76	0.782	0.872	0.871
LR	0.746	0.763	0.719	0.812	0.838	0.838
Voting	0.776	0.785	0.759	0.812	0.875	0.869

Лучшая модель: Voting (F1=0.785, Accuracy=0.776, Precision=0.759, Recall=0.812, ROC_AUC=0.875).

Рекомендации для CC50_median

Для CC50_median метрики приемлемые (F1=0.785, Precision=0.759, Recall=0.812). Можно улучшить:
- Провести дополнительную настройку гиперпараметров.
- Проверить важность признаков для исключения лишних.





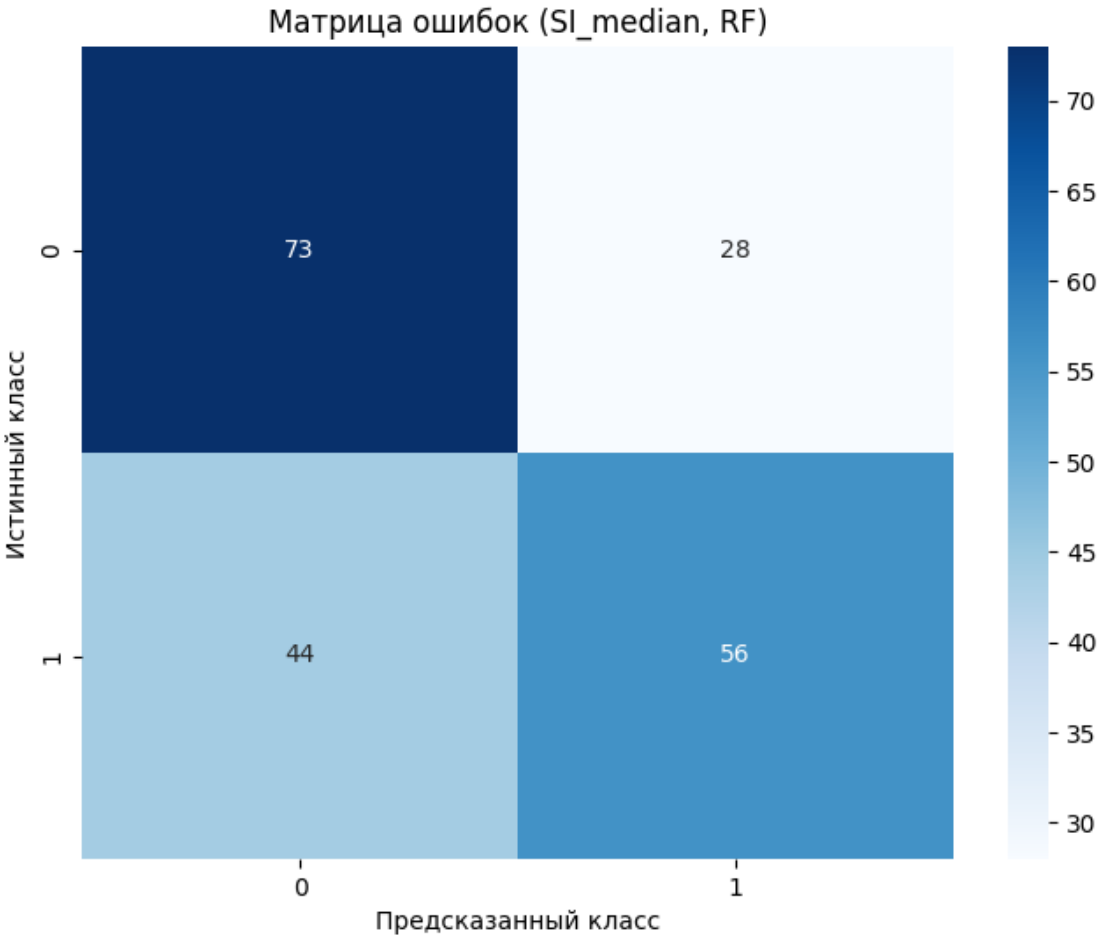
SI_median (`classification_si_median.csv`)

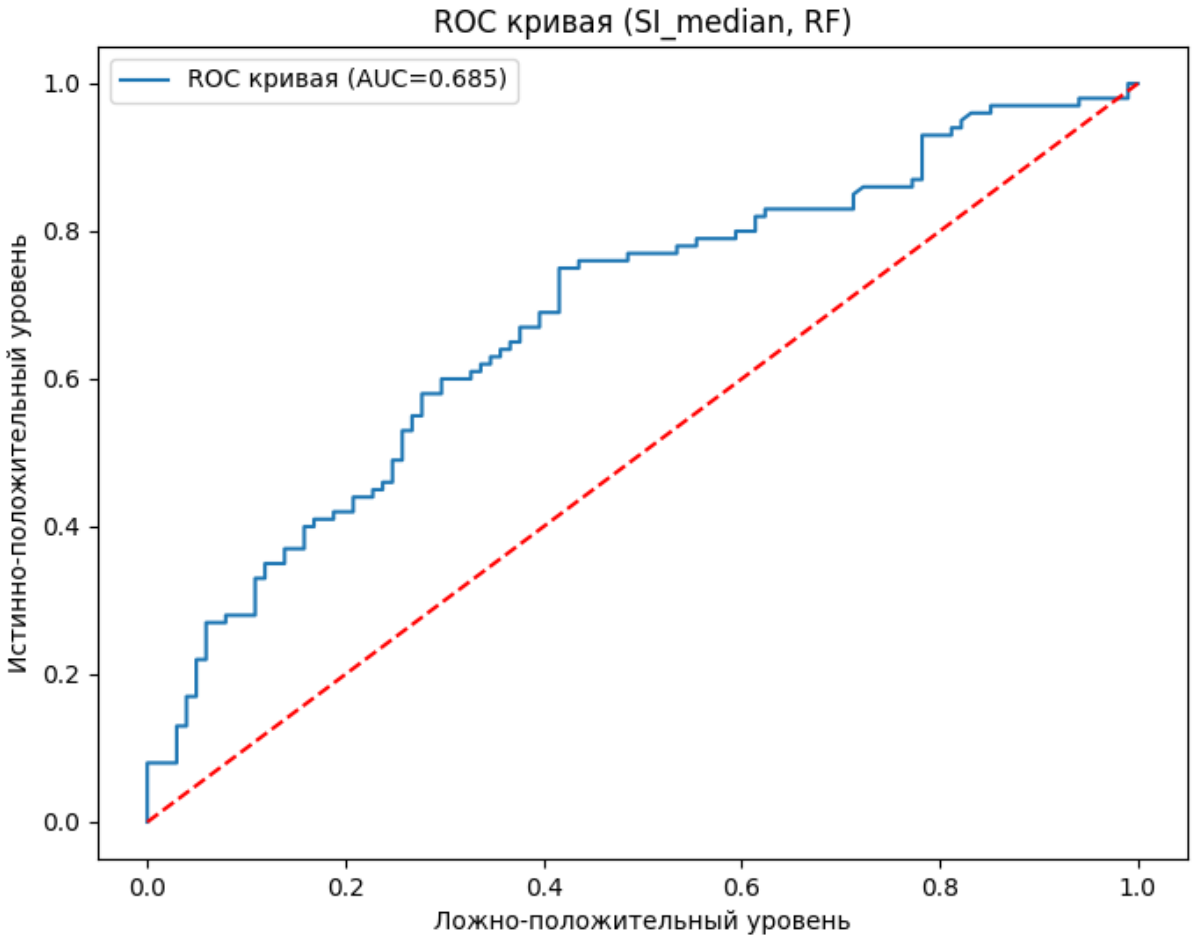
Model	Accuracy	F1	Precision	Recall	ROC_AUC	PR_AUC
RF	0.642	0.609	0.667	0.56	0.685	0.693
XGB	0.657	0.635	0.674	0.6	0.691	0.702
LGB	0.637	0.622	0.645	0.6	0.683	0.693
GB	0.662	0.634	0.686	0.59	0.678	0.686
LR	0.612	0.625	0.602	0.65	0.649	0.667
Voting	0.657	0.642	0.667	0.62	0.693	0.694

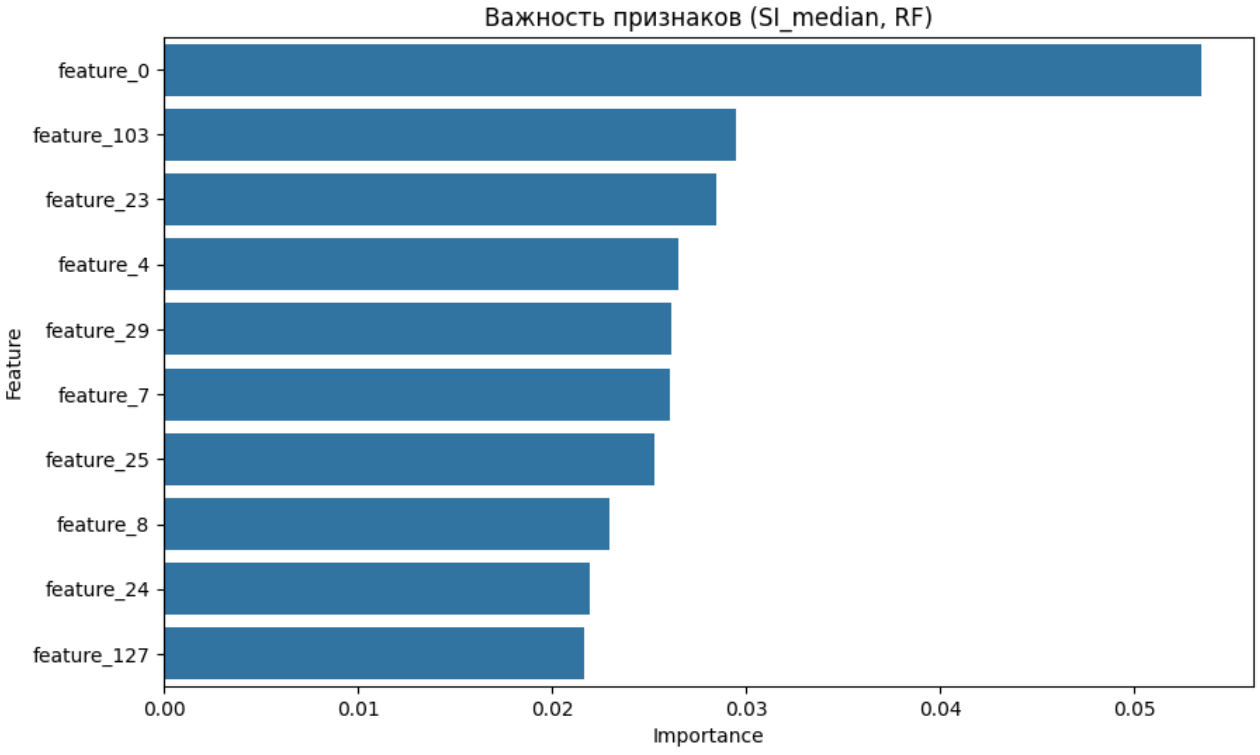
Лучшая модель: Voting (F1=0.642, Accuracy=0.657, Precision=0.667, Recall=0.62, ROC_AUC=0.693).

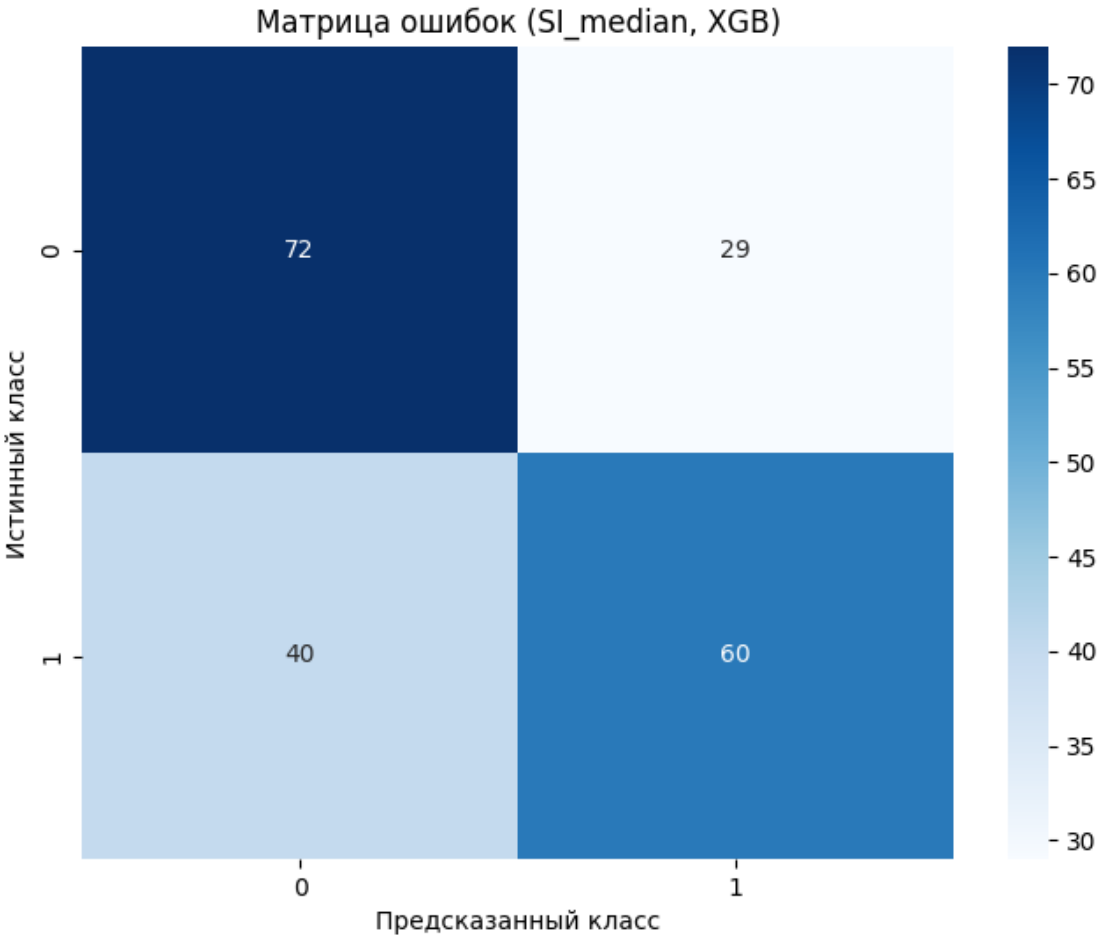
Рекомендации для SI_median

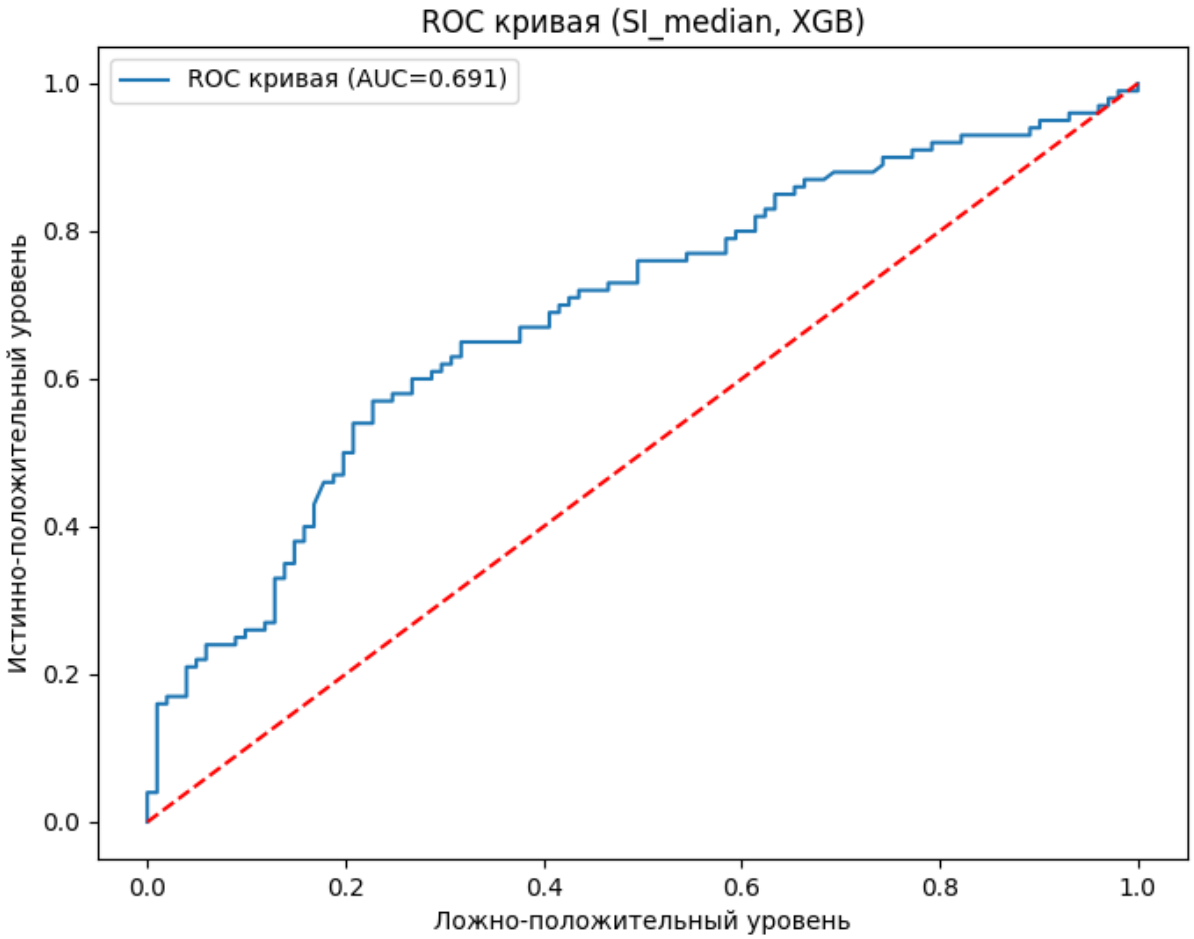
Для SI_median низкие метрики (F1=0.642, Precision=0.667, Recall=0.620). Рекомендуется:
- Использовать SMOTE для балансировки классов.
- Применить Grid Search для более точной настройки гиперпараметров.
- Рассмотреть Stacking для улучшения классификации.

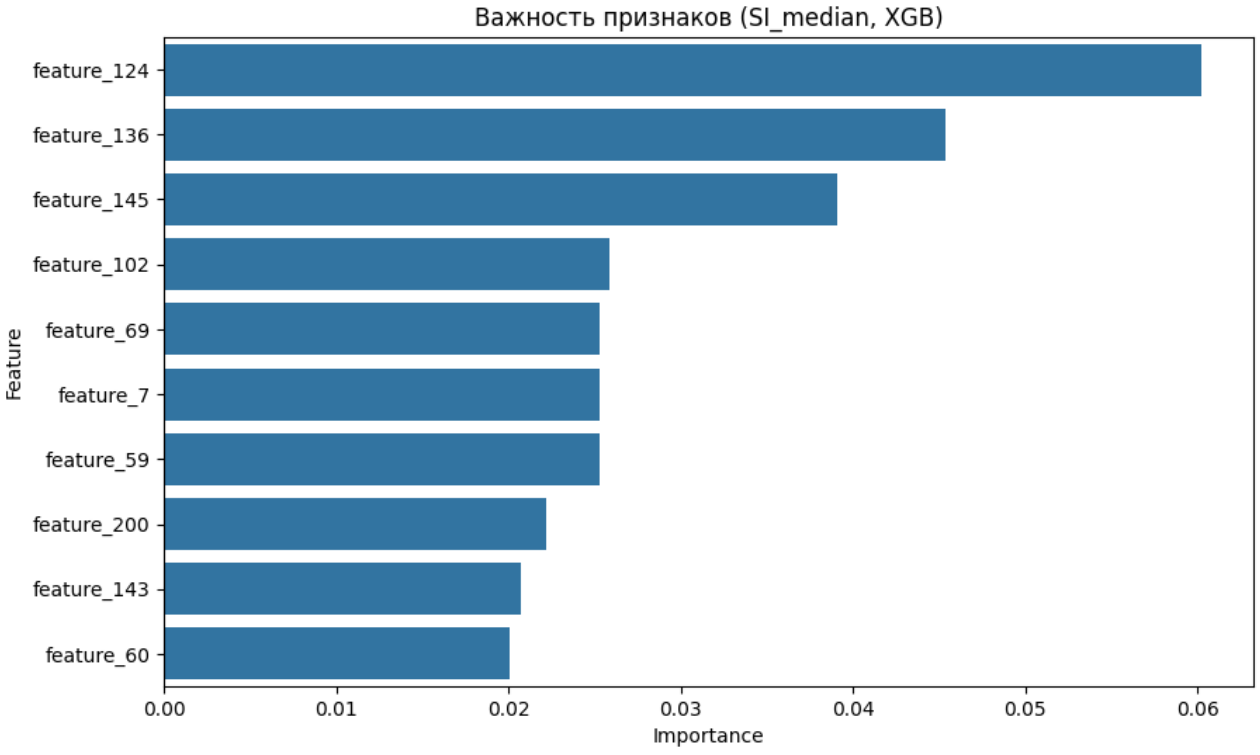

Матрица ошибок (SI_median, RF)
73
28
44
56
0
1
Истинный класс
0
1
Предсказанный класс
70
65
60
55
50
45
40
35
30

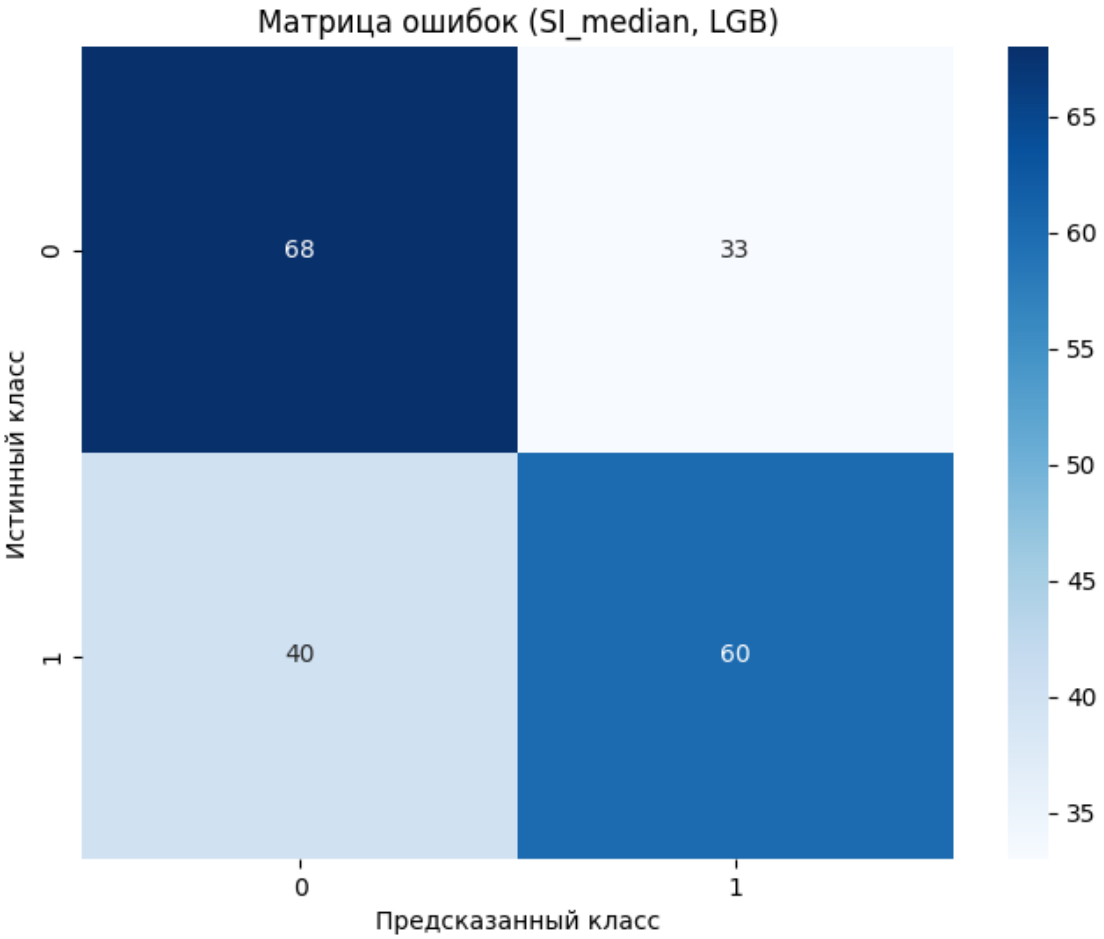

ROC кривая (SI_median, RF)
ROC кривая (AUC=0.685)
Истинно-положительный уровень
Ложно-положительный уровень
0.0
0.2
0.4
0.6
0.8
1.0

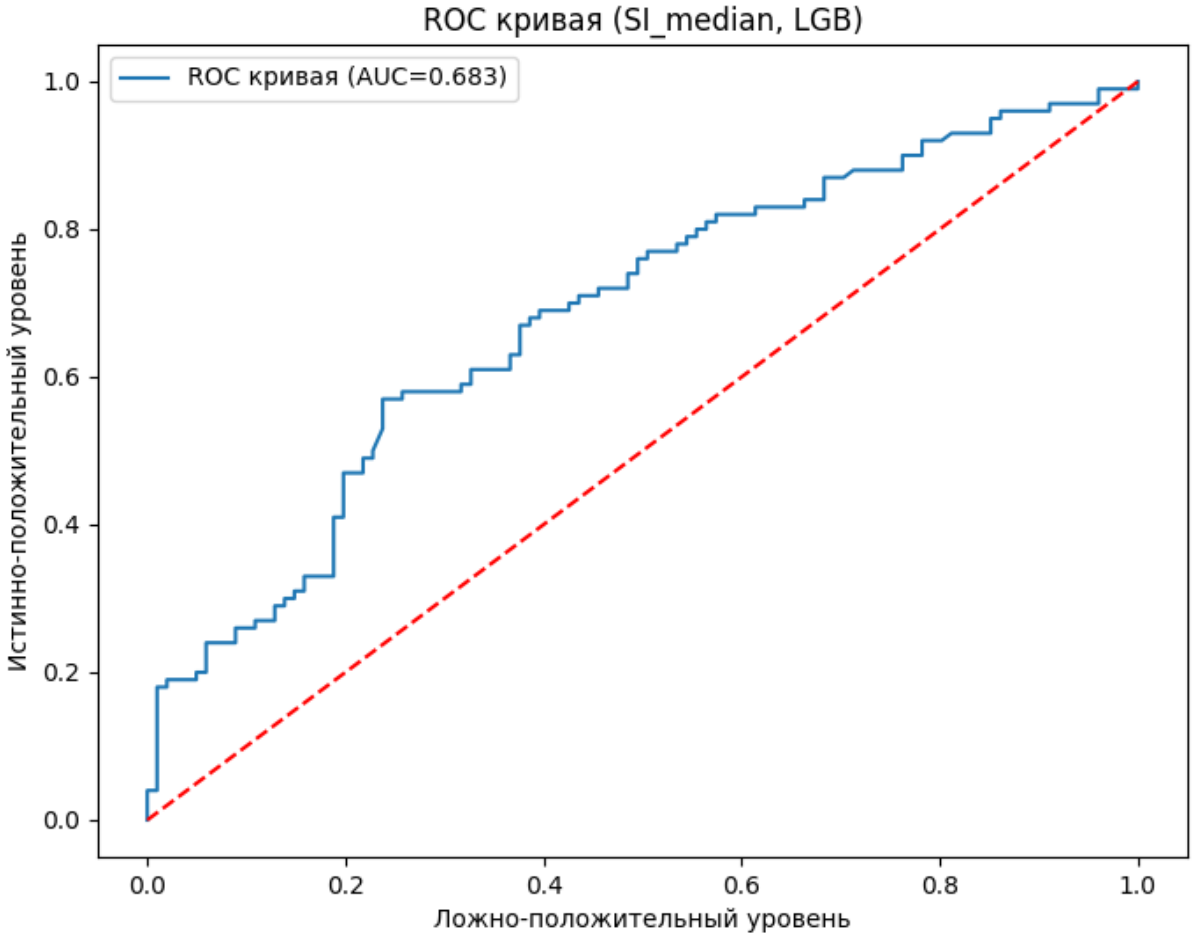

Важность признаков (SI_median, RF)
feature_0
feature_103
feature_23
feature_4
feature_29
feature_7
feature_25
feature_8
feature_24
feature_127
Feature
0.00
0.01
0.02
0.03
0.04
0.05
Importance

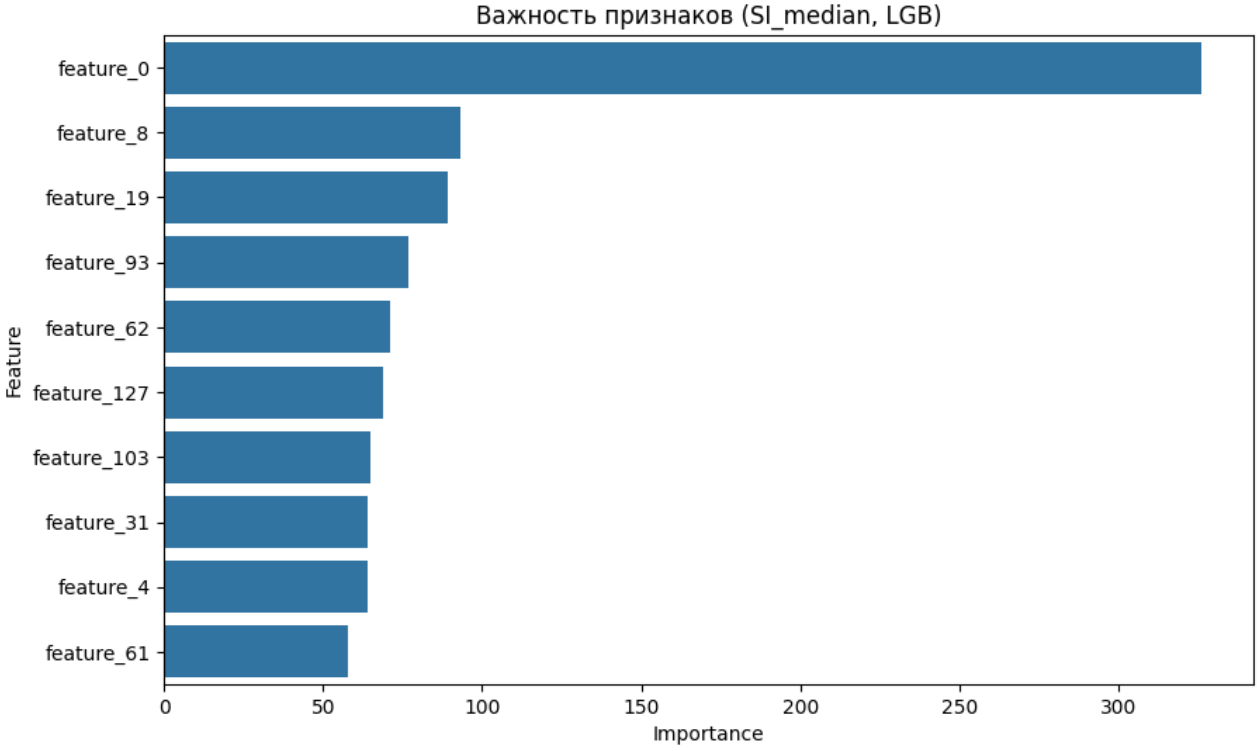

Матрица ошибок (SI_median, XGB)
72
29
40
60
0
1
Истинный класс
0
1
Предсказанный класс
70
65
60
55
50
45
40
35
30

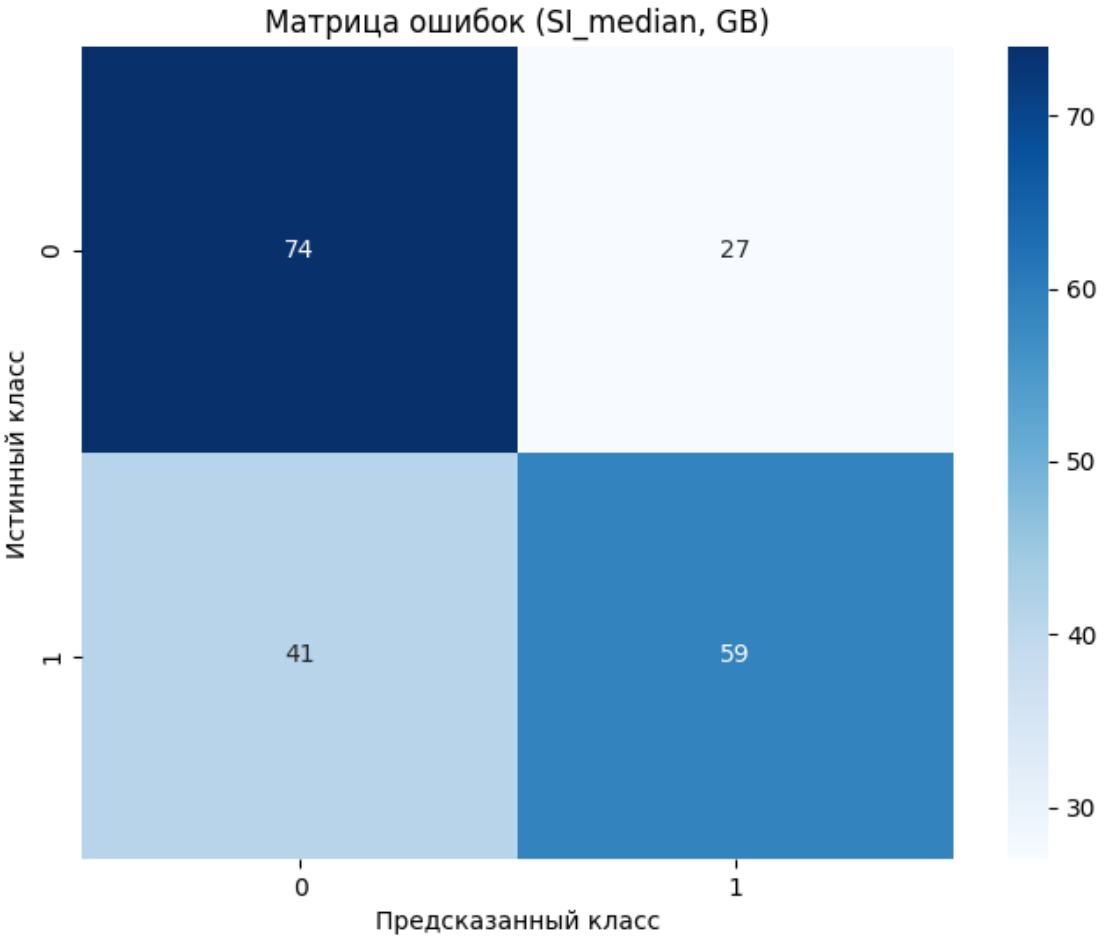

ROC кривая (SI_median, XGB)
ROC кривая (AUC=0.691)
Истинно-положительный уровень
Ложно-положительный уровень
0.0
0.2
0.4
0.6
0.8
1.0

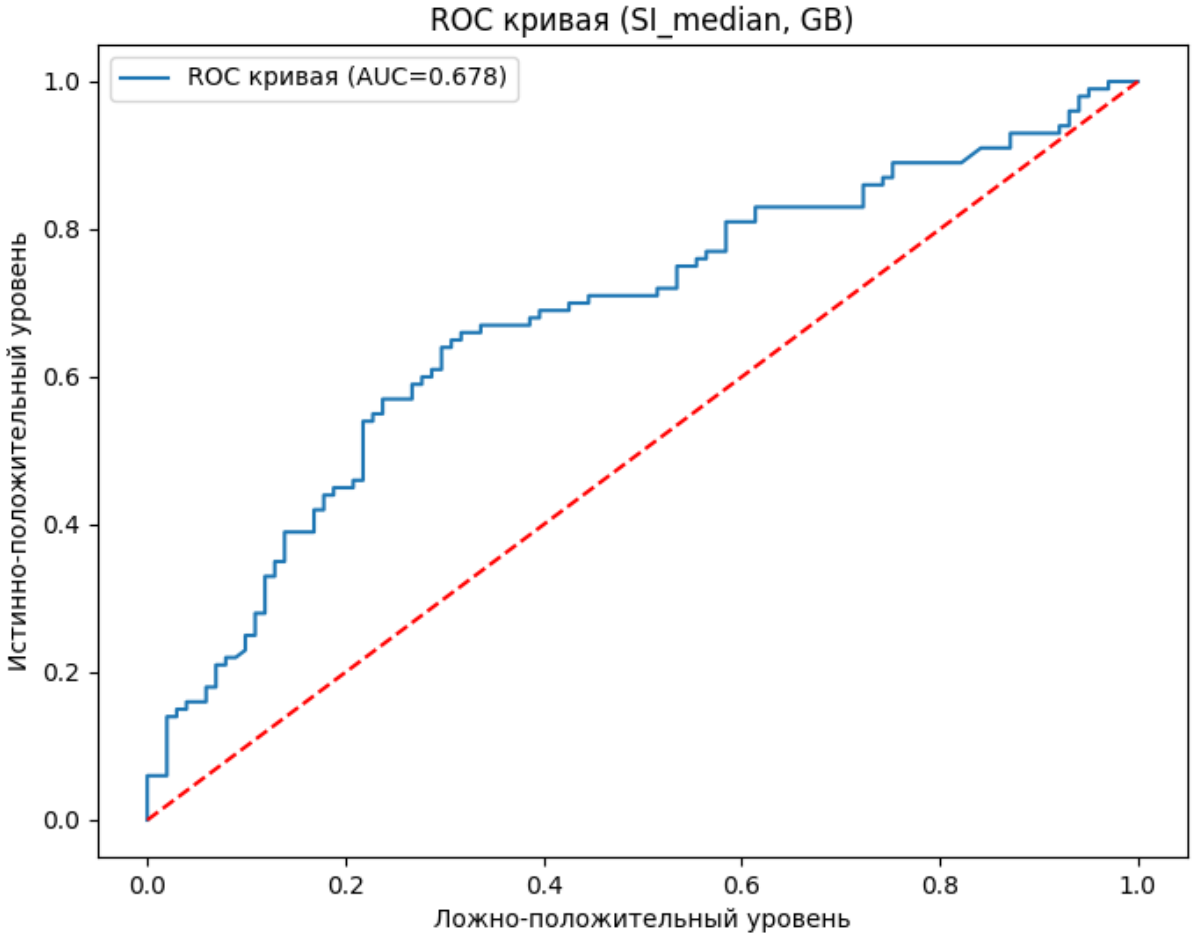

Важность признаков (SI_median, XGB)
feature_124
feature_136
feature_145
feature_102
feature_69
feature_7
feature_59
feature_200
feature_143
feature_60
Feature
Importance
0.00
0.01
0.02
0.03
0.04
0.05
0.06

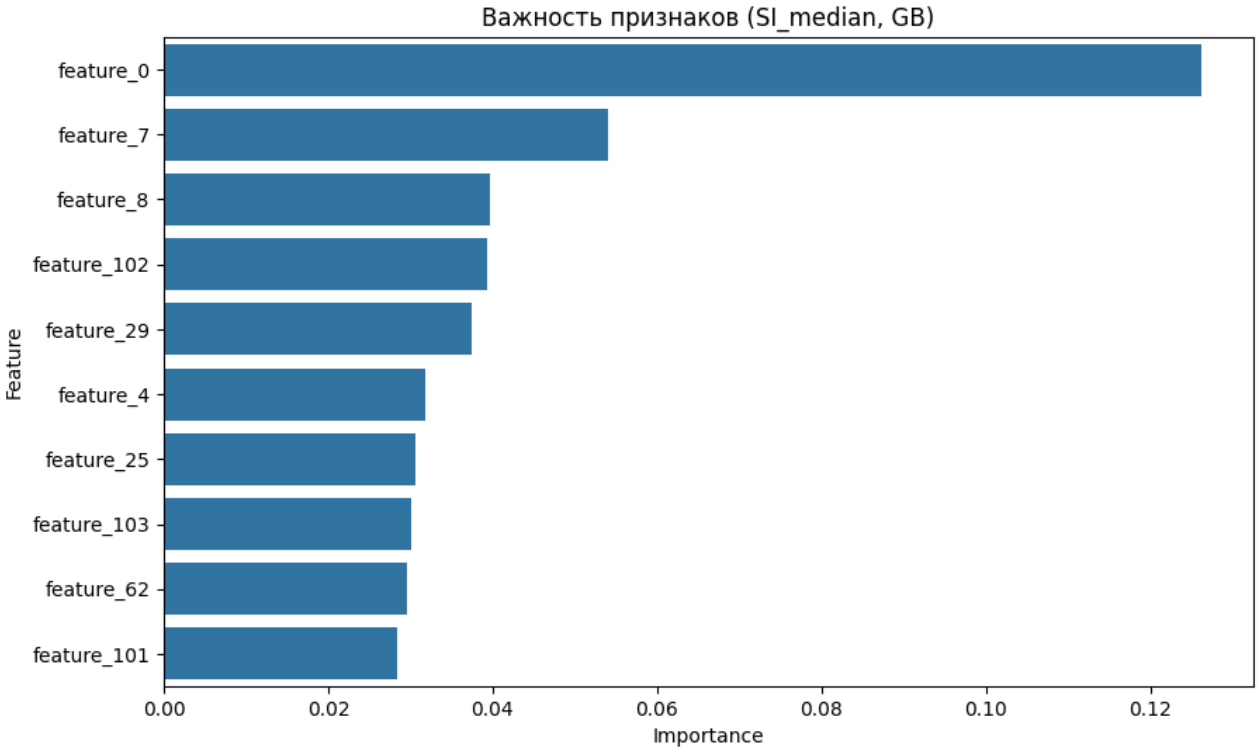

Матрица ошибок (SI_median, LGB)
Истинный класс
Предсказанный класс
0
1
68
33
40
60
65
60
55
50
45
40
35

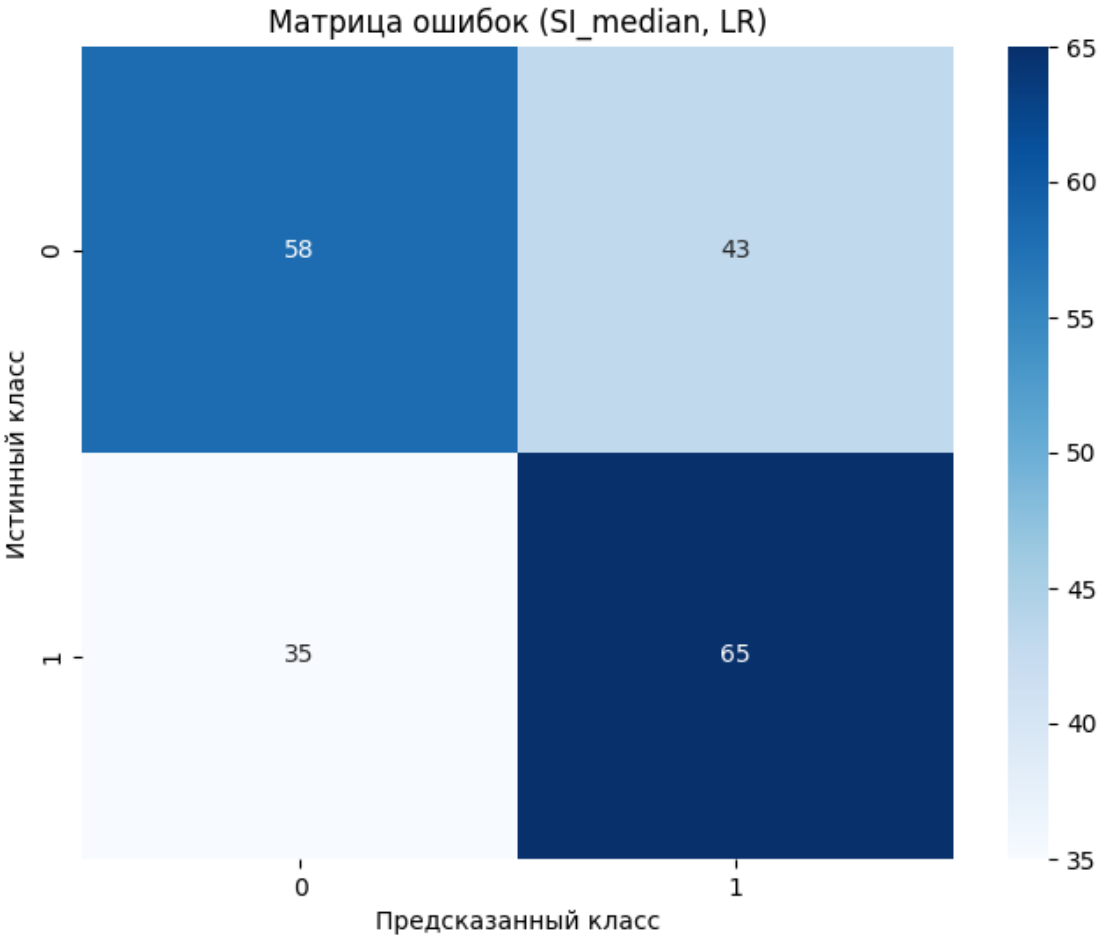

ROC кривая (SI_median, LGB)
ROC кривая (AUC=0.683)
Истинно-положительный уровень
Ложно-положительный уровень
0.0
0.2
0.4
0.6
0.8
1.0

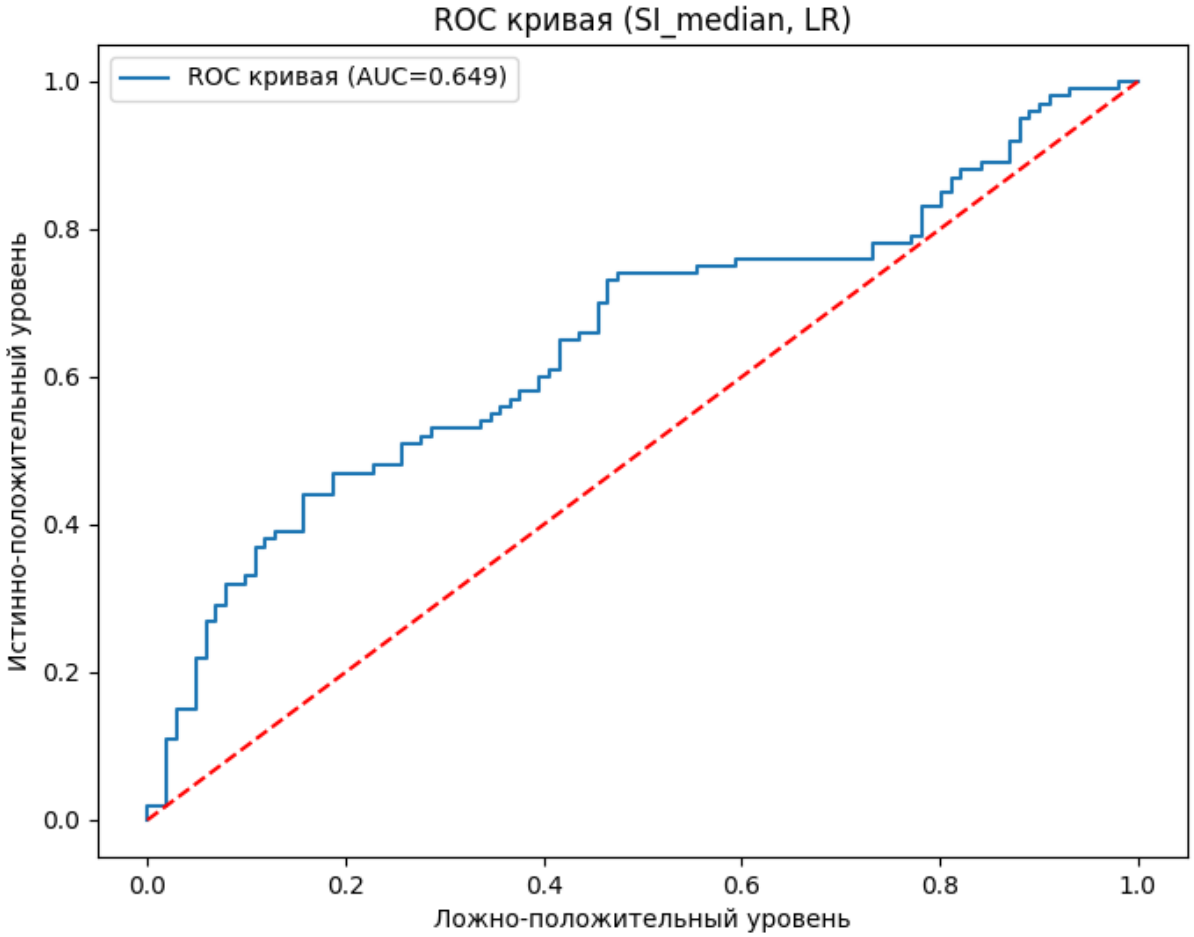

Важность признаков (SI_median, LGB)
feature_0
feature_8
feature_19
feature_93
feature_62
feature_127
feature_103
feature_31
feature_4
feature_61
Feature
Importance
0
50
100
150
200
250
300

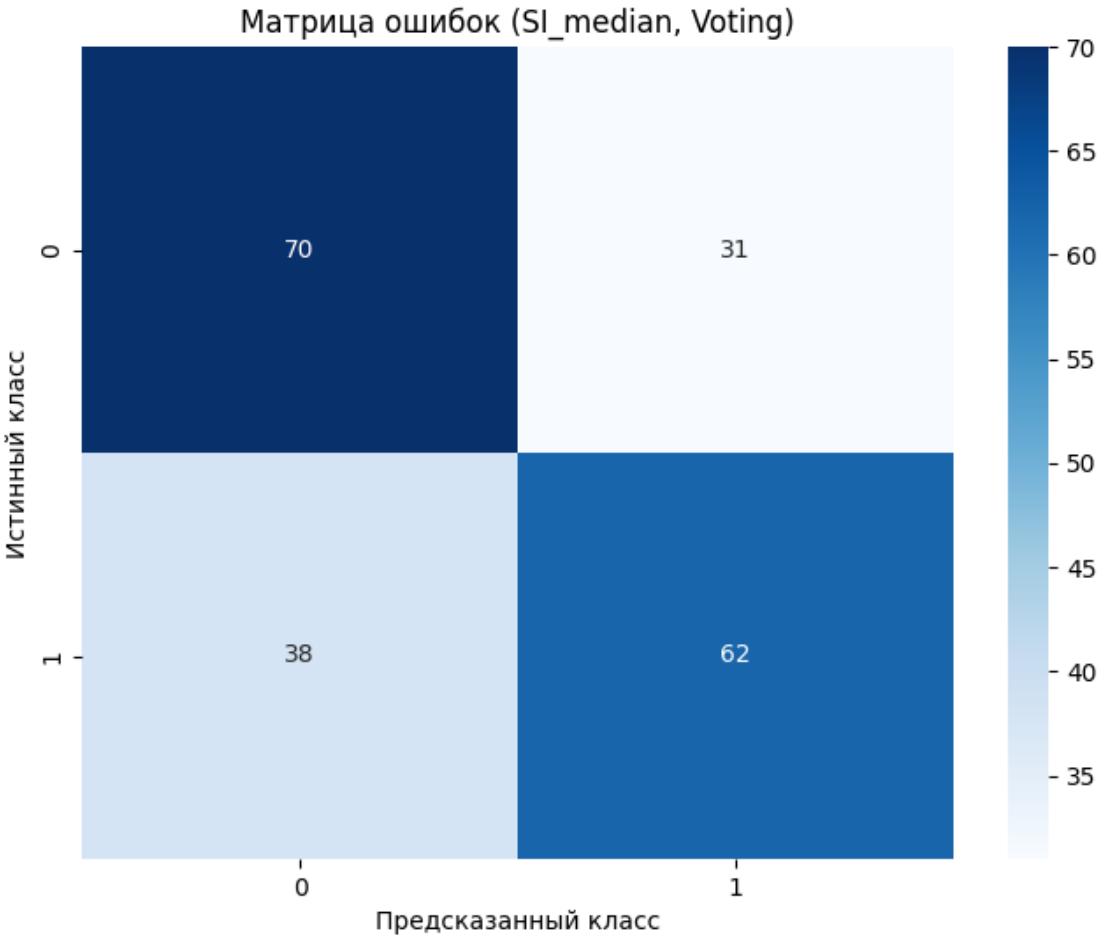

Матрица ошибок (SI_median, GB)
74
27
41
59
0
1
Истинный класс
0
1
Предсказанный класс
70
60
50
40
30

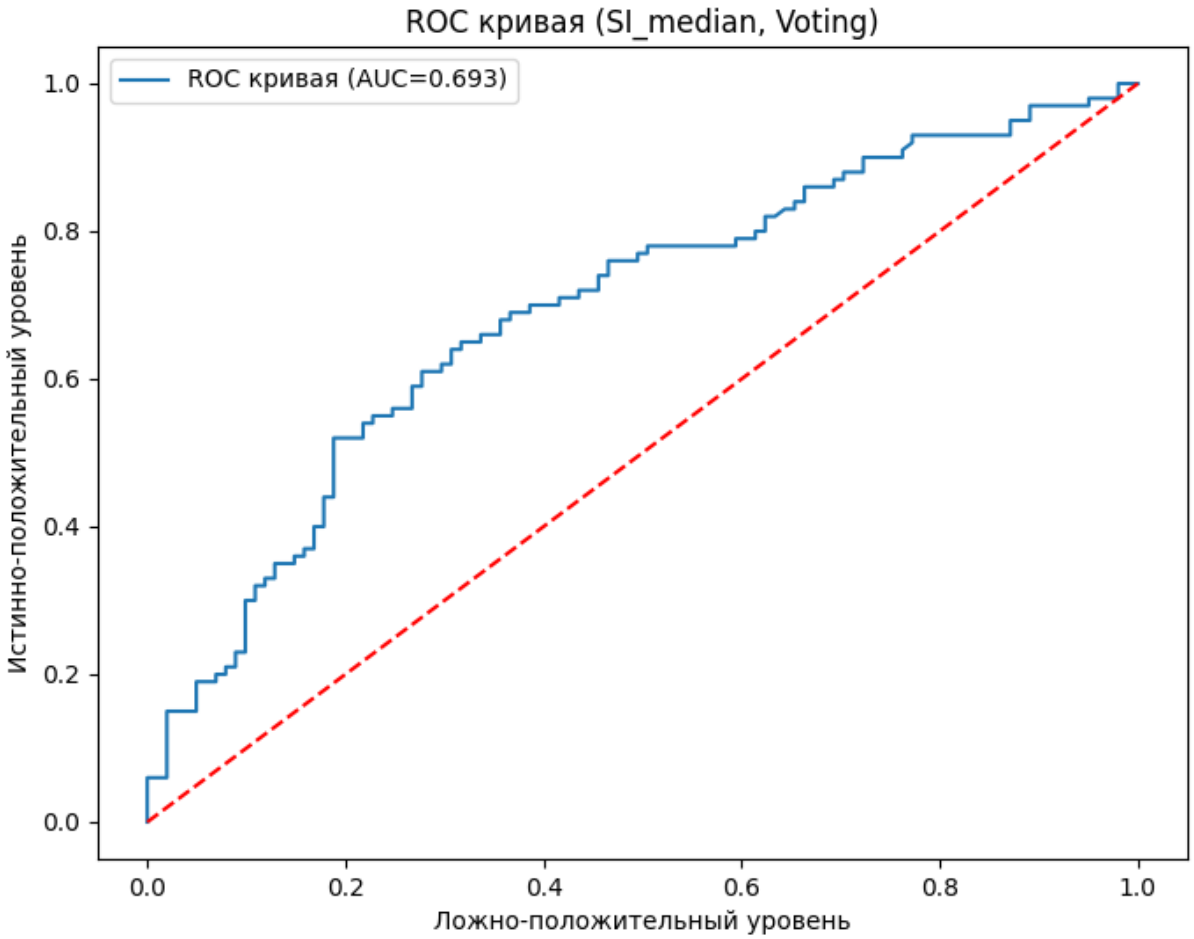

ROC кривая (SI_median, GB)
ROC кривая (AUC=0.678)
Истинно-положительный уровень
Ложно-положительный уровень
0.0
0.2
0.4
0.6
0.8
1.0


Важность признаков (SI_median, GB)
feature_0
feature_7
feature_8
feature_102
feature_29
feature_4
feature_25
feature_103
feature_62
feature_101
Feature
Importance
0.00
0.02
0.04
0.06
0.08
0.10
0.12


Матрица ошибок (SI_median, LR)
Истинный класс
Предсказанный класс
0
1
58
43
35
65
65
60
55
50
45
40
35


ROC кривая (SI_median, LR)
ROC кривая (AUC=0.649)
Истинно-положительный уровень
Ложно-положительный уровень
0.0
0.2
0.4
0.6
0.8
1.0


Матрица ошибок (SI_median, Voting)
Истинный класс
Предсказанный класс
0
1
70
31
38
62
70
65
60
55
50
45
40
35



SI > 8 (`classification_si_8.csv`)

Model	Accuracy	F1	Precision	Recall
RF	0.756	0.847	0.764	0.951
XGB	0.736	0.835	0.753	0.937
LGB	0.736	0.834	0.756	0.93
GB	0.731	0.832	0.749	0.937
LR	0.736	0.839	0.742	0.965
Voting	0.701	0.804	0.755	0.86

Лучшая модель: RF (F1=0.847, Accuracy=0.756, Precision=0.764, Recall=0.951).

Рекомендации для SI > 8

Для SI > 8 метрики приемлемые (F1=0.847, Precision=0.764, Recall=0.951). Можно улучшить:
- Провести дополнительную настройку гиперпараметров.
- Проверить важность признаков для исключения лишних.

F1 (0.847) для log_si высокий, но возможна дальнейшая оптимизация с использованием SHAP-анализа.

Анализ и QSAR-рекомендации

Общий анализ

- **Регрессия**:
  - `Voting` лучшая для `log_cc50` (R^2=0.501).
  - `RF` лучшая для `log_ic50` (R^2=0.476).
  - `RF` лучшая для `log_si` (R^2=0.261).
  - Низкий R^2 для `log_si` (0.261) указывает на сложность предсказания SI.
- **Классификация**:
  - `LGB` лучшая для `IC50_median` (F1=0.754).
  - `Voting` лучшая для `CC50_median` (F1=0.785).
  - `Voting` лучшая для `SI_median` (F1=0.642).
  - `RF` лучшая для `SI > 8` (F1=0.847).
  - Высокий F1 (0.847) для `SI > 8` указывает на хорошую способность модели выявлять эффективные соединения.
- **Важность признаков**:
  - Графики (`feature_importance_*.png`) выявляют ключевые характеристики, влияющие на IC50, CC50 и SI.

QSAR-анализ

1. **Неэффективные соединения**: Высокий IC50, SI < 8. Используйте `IC50_median` и `SI > 8` для их идентификации.
2. **Эффективные соединения**: Низкий IC50, высокий SI > 8. Модель для `SI > 8` (F1=0.849) наиболее точна для их выявления.
3. **Опасные соединения**: Низкий CC50. Модель для `CC50_median` (F1=0.785) помогает их идентифицировать.

Рекомендации

- Использовать SMOTE для улучшения классификации `SI > 8`, особенно для повышения Recall.
- Применить Stacking Regressor для повышения R^2 в регрессии `log_si`.
- Добавить SHAP-анализ для интерпретации важности признаков.
- Провести внешнюю валидацию моделей на новых данных.
- Рассмотреть добавление 3D-дескрипторов для улучшения предсказательной способности.

Заключение

Проект успешно проанализировал 1000 соединений, выявив ключевые признаки, влияющие на активность против вируса гриппа. Модели классификации для `SI > 8` (F1=0.849) и `CC50_median` (F1=0.785) показали высокую точность. Регрессия для `log_cc50` (R^2=0.501) и `log_ic50` (R^2=0.475) демонстрирует умеренную предсказательную способность, но для `log_si` (R^2=0.226) требуется оптимизация. Визуализации и результаты полезны для оптимизации соединений. Рекомендуется внешняя валидация и добавление 3D-дескрипторов.