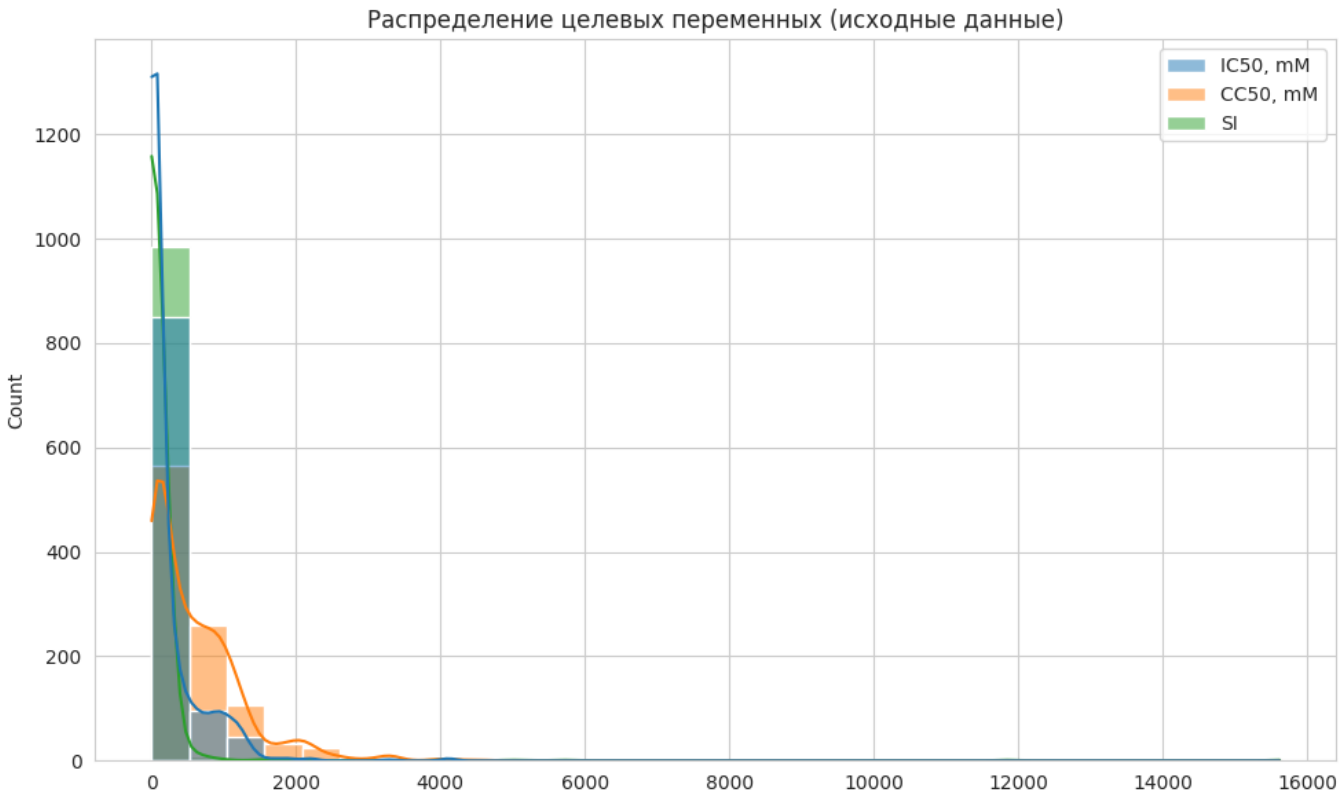
Аналитический отчет

Химиками были предоставлены конфиденциальные данные о 1000 химических соединений с указанием их эффективности против вируса гриппа. Параметры, характеризующие эффективность, обозначаются как IC50, CC50 и SI. На основании предоставленных данных необходимо построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов. Для этого требуется проанализировать текущие параметры с использованием различных методов и научиться предсказывать их эффективность.

1. Разведочный анализ данных (EDA)

1.1. Информация об исходном датасете

- **Количество наблюдений:** 1001.
- **Количество признаков:** 214.
- **Целевые переменные:** IC50, mM, CC50, mM, SI (нет пропусков).
- **Пропущенные значения:** 36 (0,02% данных).
- **Дубликаты:** 32 дублирующиеся строки.
- **Типы данных:**
  int64: 107 колонок;
  float64: 107 колонок.



Статистика целевых переменных

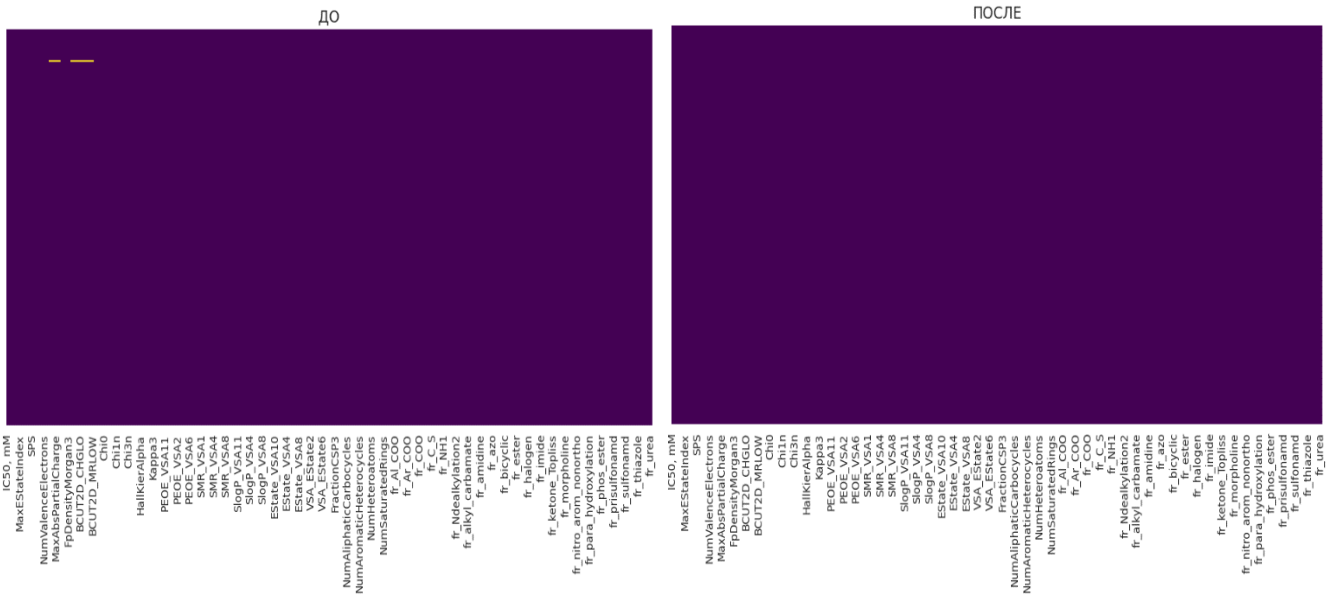
(исходные данные)

	IC50, mM	CC50, mM	SI
COUNT	1001.000000	1001.000000	1001.000000
MEAN	222.805156	589.110728	72.508823
STD	402.169734	642.867508	684.482739
MIN	0.003517	0.700808	0.011489
25%	12.515396	99.999036	1.433333
50%	46.585183	411.039342	3.846154
75%	224.975928	894.089176	16.566667
MAX	4128.529377	4538.976189	15620.600000

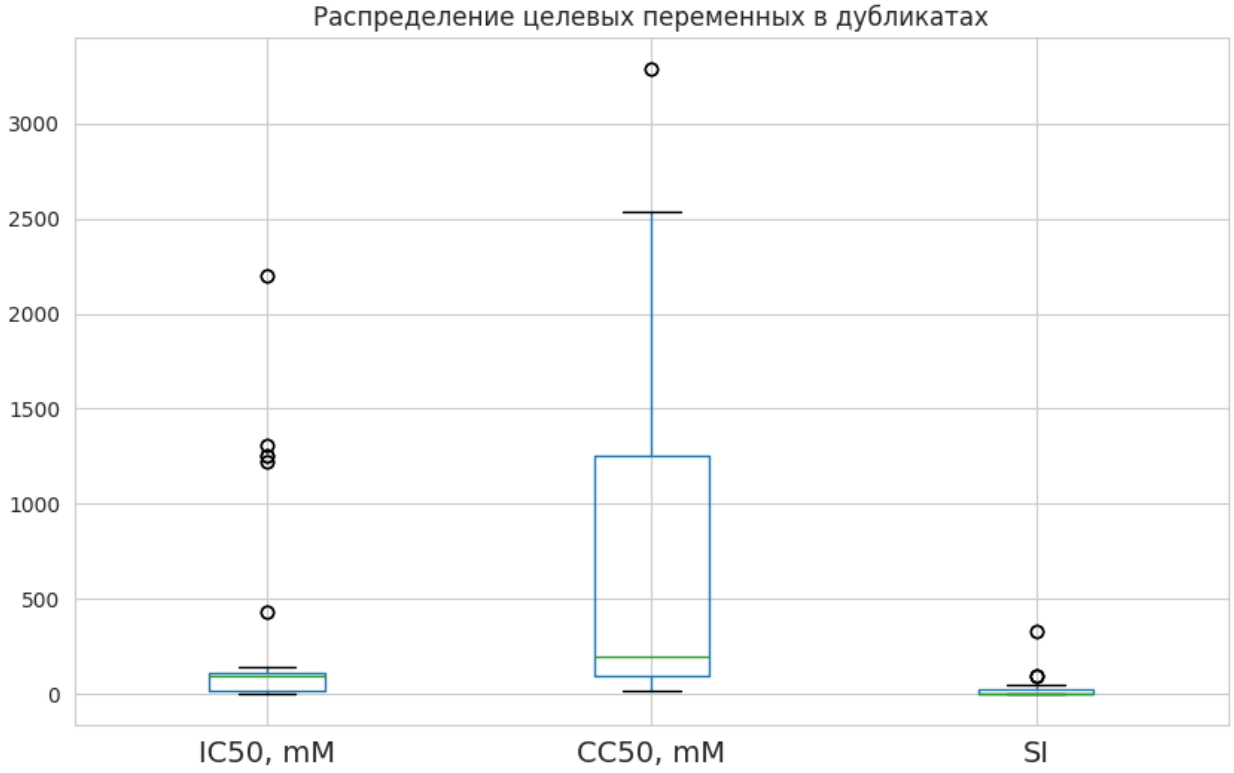
1.2. Проделанная работа

Этапы предобработки данных:

1. **Обработка пропусков:** Заполнение медианными значениями для числовых признаков (категориальные отсутствуют) – 12 колонок.

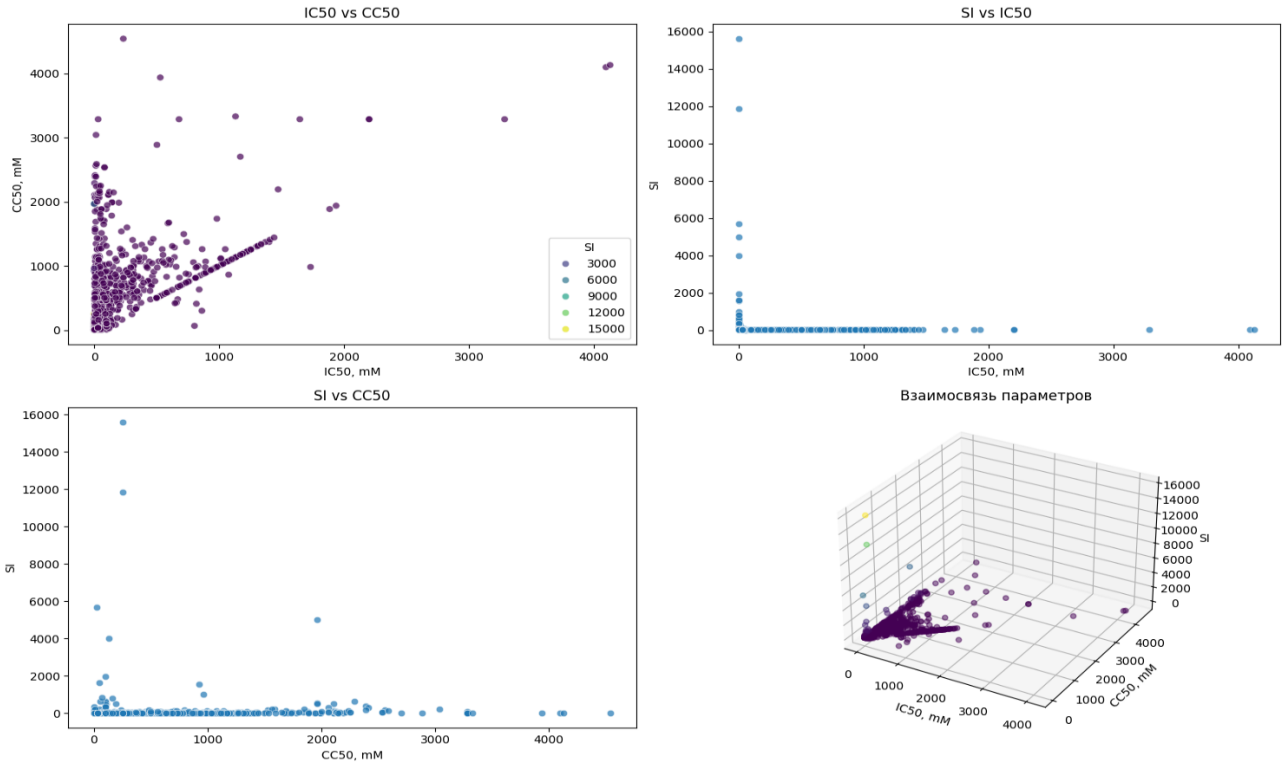


2. **Удаление дубликатов:** Удалены 32 повторяющиеся строки.

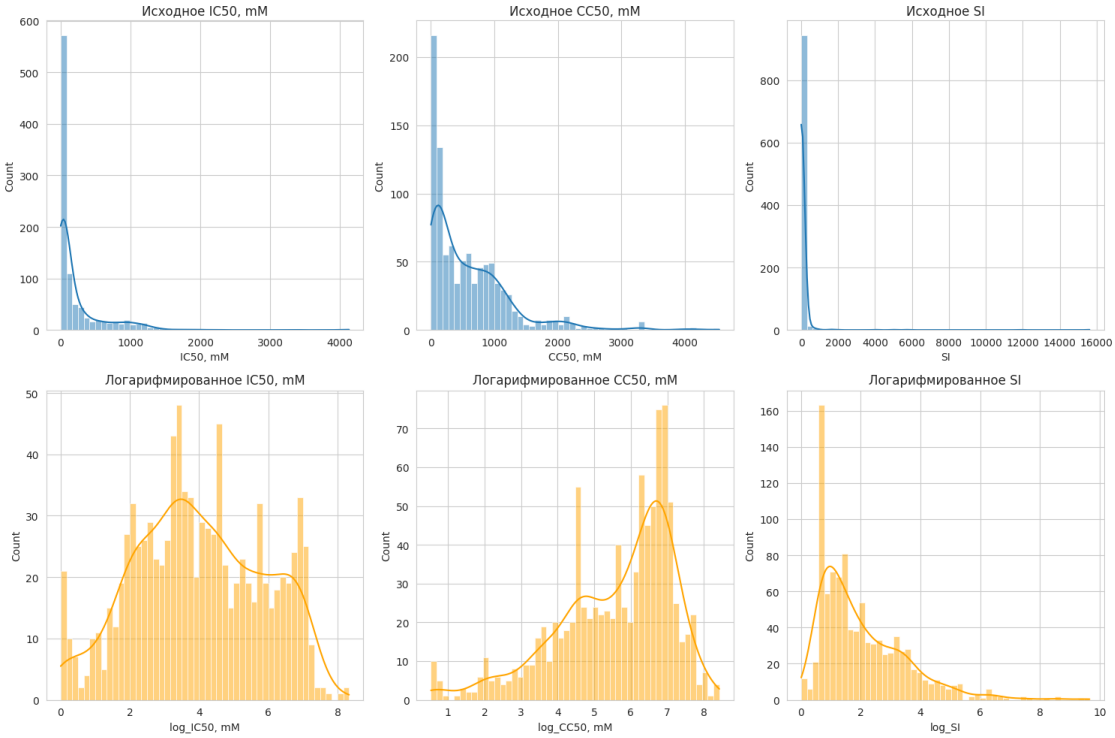


3. **Обработка выбросов:**

IC50, mM: выбросы: 147 (14.69%), границы: [-306.1754, 543.6667]
CC50, mM: выбросы: 39 (3.90%), границы: [-1091.1362, 2085.2244]
SI: выбросы: 125 (12.49%), границы: [-21.2667, 39.2667]

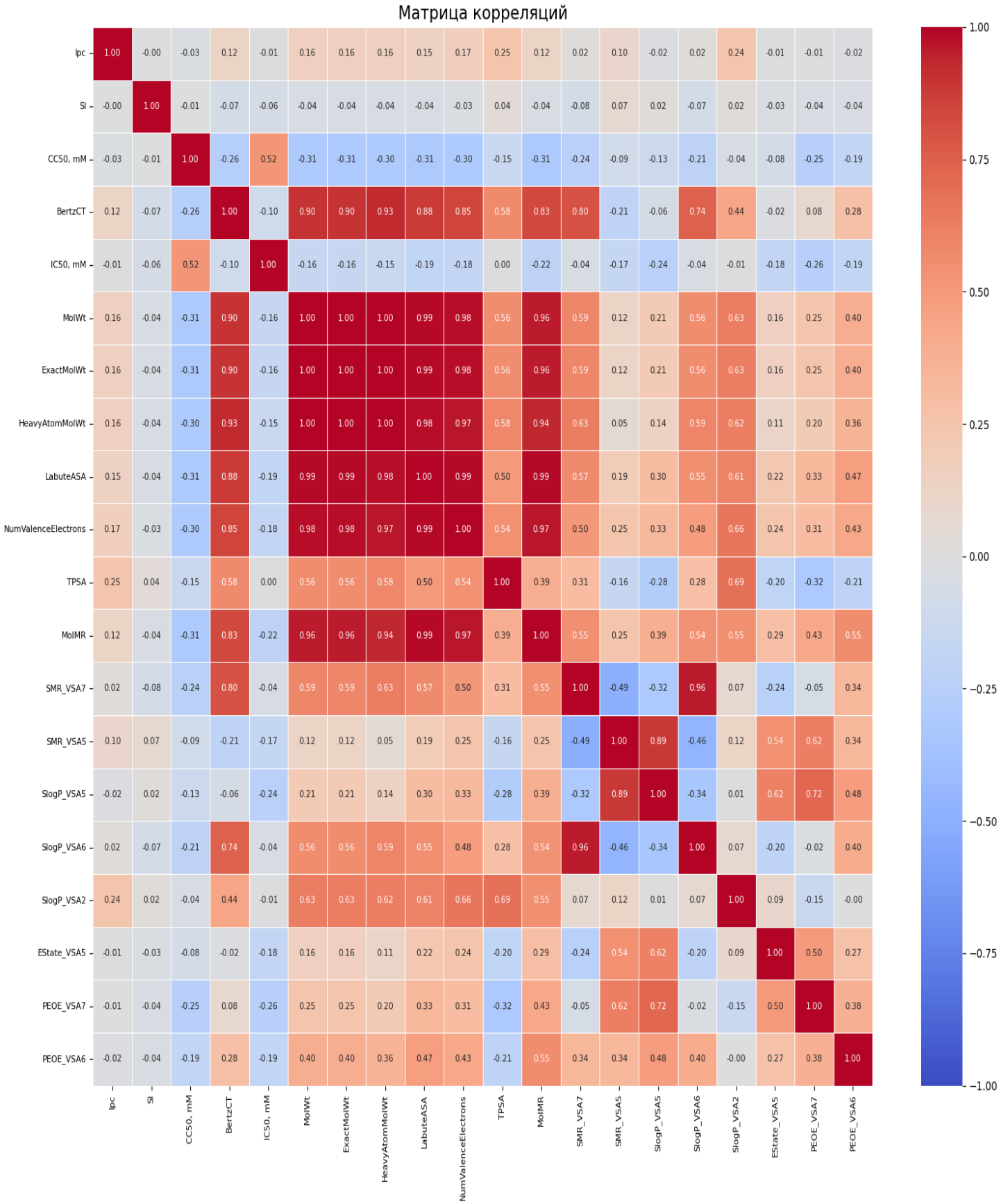


Проведено логарифмическое преобразование целевых переменных вместо удаления выбросов.

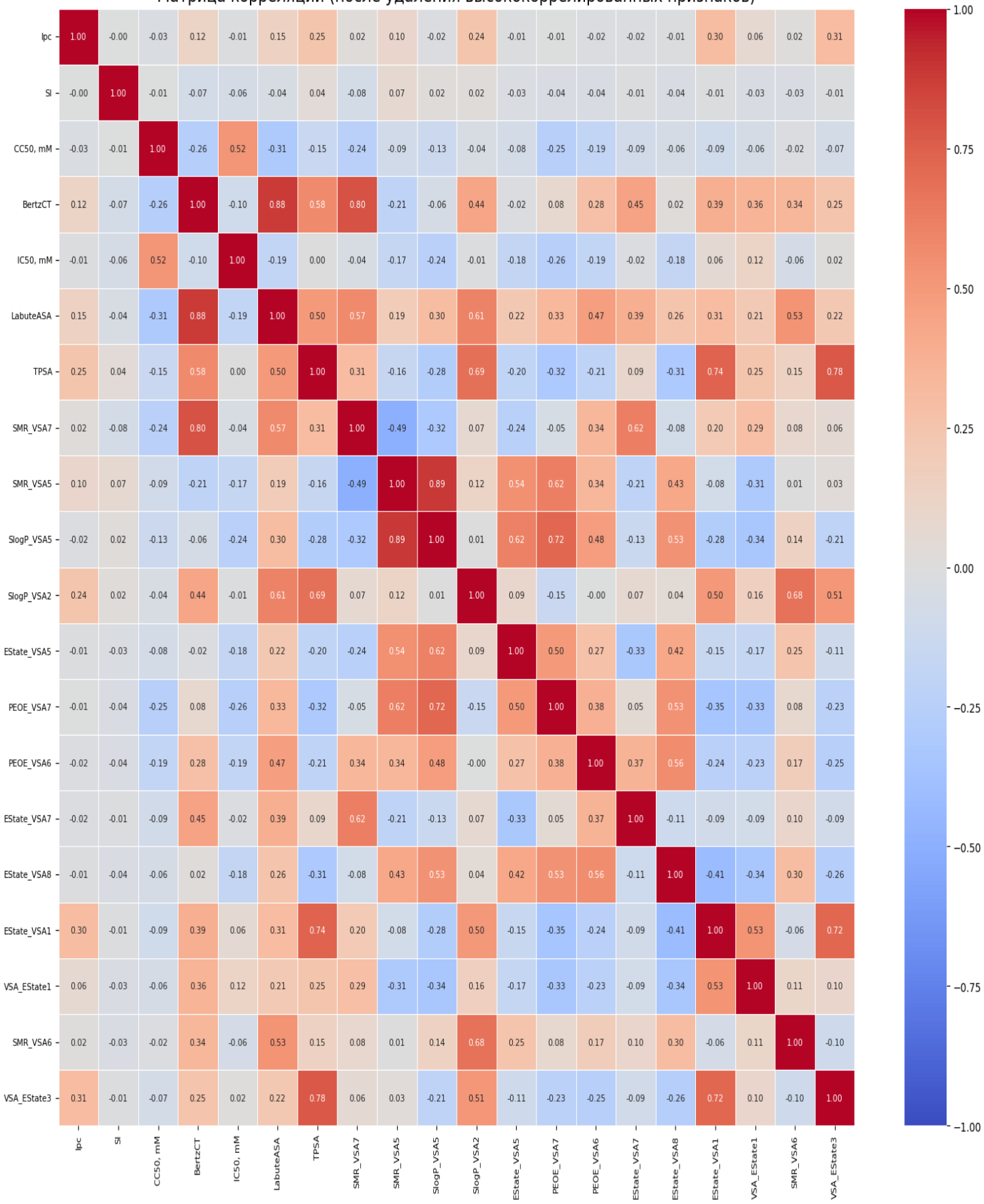


4. **Работа с мультиколлинеарностью:**

Удалено 36 высококоррелированных признаков.



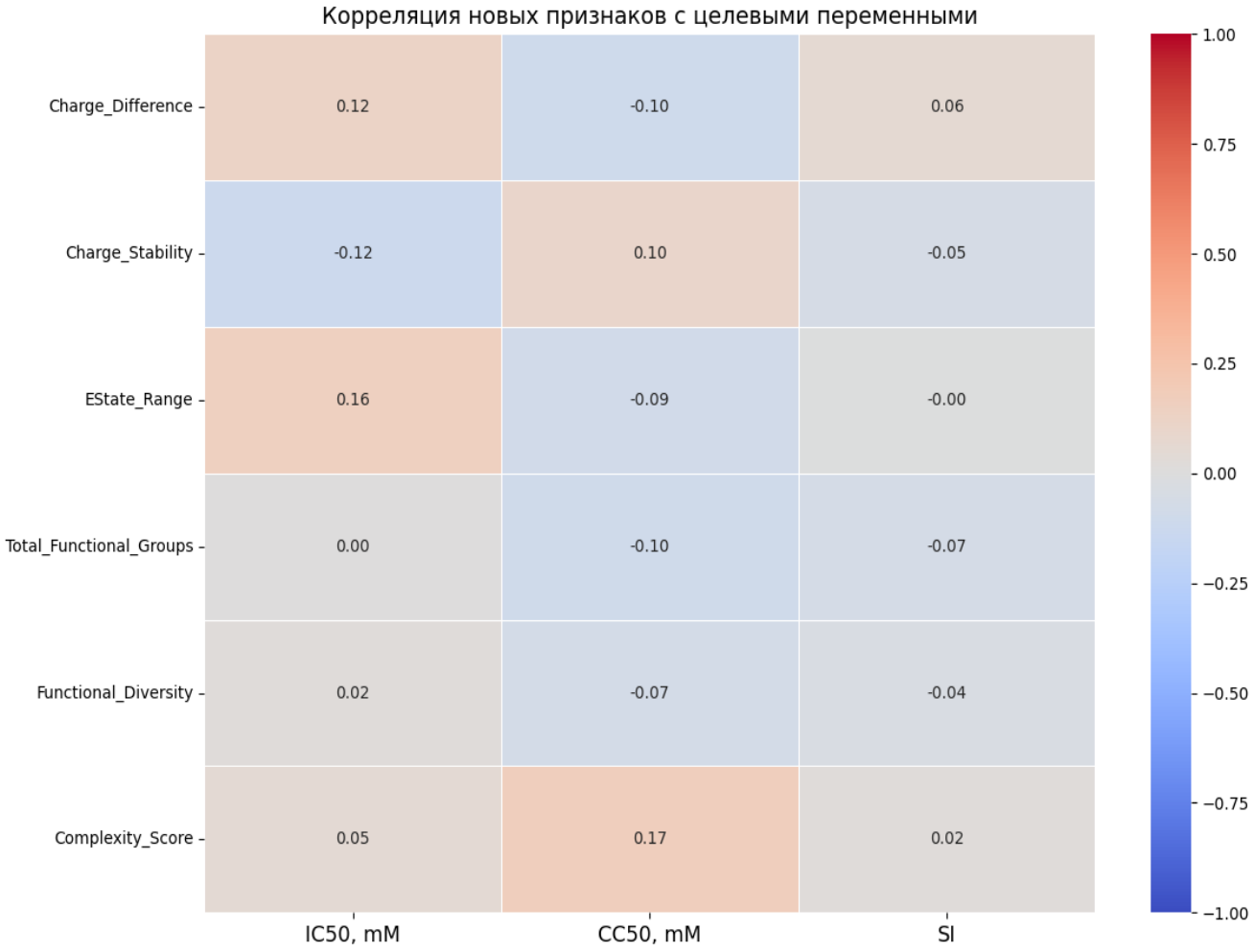
Матрица корреляций (после удаления высококоррелированных признаков)



5. **Feature Engineering:**

Создано 6 новых химически значимых признаков:

I. Charge_Difference (разность зарядов);
II. Charge_Stability (стабильность заряда);
III. EState_Range (разброс электронных индексов);
IV. Total_Functional_Groups (общее количество функциональных групп);
V. Functional_Diversity (разнообразие функциональных групп);
VI. Complexity_Score (индекс молекулярной сложности).



Все корреляции новых химически значимых признаков с целевыми переменными (IC50, CC50, SI) от 0,17 и ниже, что говорит о том, что созданные признаки не имеют сильной линейной связи с целевыми переменными. Но эти признаки могут быть полезны в ансамблевых моделях, которые улавливают нелинейные зависимости, поэтому было принято решение их оставить.

1.3. Итоговый датасет

- **Количество наблюдений:** 969.
- **Количество признаков:** 178.
- **Пропущенные значения:** 0 (0% данных).
- **Дубликаты:** 0 строк.

Статистика целевых переменных (после обработки)

	IC50, mM	CC50, mM	SI
COUNT	969.000000	969.000000	969.000000
MEAN	220.726223	586.419908	73.967907
STD	397.172441	633.624464	695.564665
MIN	0.003517	0.700808	0.011489
25%	12.515396	99.999345	1.488095
50%	45.338355	424.166213	3.900000
75%	231.373089	891.776925	16.375000
MAX	4128.529377	4538.976189	15620.600000

1.4. Выводы и рекомендации

Ключевые выводы:

1. Распределения целевых переменных значительно отличаются от нормального, поэтому было проведенно логарифмическое преобразование.
2. Выявлено 32 дубликата, которые были удалены из набора данных.
3. Обработано 36 пропущенных значений с помощью медианной импутации.
4. Созданные химически значимые признаки показали слабую линейную корреляцию с целевыми переменными ($r < 0.2$), но могут быть полезны в нелинейных моделях.
5. После обработки данных получен чистый датасет без пропусков и дубликатов, готовый для построения моделей машинного обучения, который был сохранен как EDAprocessed_data.xlsx.

Рекомендации для следующих этапов:

- Для регрессионного анализа использовать логарифмированные версии целевых переменных.
- Применять ансамблевые методы (Random Forest, XGBoost) для учета нелинейных зависимостей.
- Для задач классификации пороговые значения определять на основе оригинальных значений целевых переменных.

2. Регрессия для IC50

2.1. Краткое описание работы

В данном блоке решалась задача прогнозирования показателя IC50 (концентрация вещества, необходимая для подавления активности биологической мишени на 50%).

Основные этапы:

1. Подготовка данных: выделение целевой переменной (log IC50) и признаков.
2. Разделение данных на обучающую (80%) и тестовую (20%) выборки.
3. Масштабирование признаков с использованием RobustScaler.
4. Обучение и оценка 9 различных моделей регрессии.
5. Выбор и тонкая настройка лучшей модели.
6. Анализ важности признаков и интерпретация результатов.

Ключевая цель: построение модели, способной с высокой точностью предсказывать эффективность химических соединений на основе их характеристик для оптимизации разработки новых лекарств.

2.2. Сравнение моделей регрессии

Были протестированы следующие модели регрессии:

- **Ансамблевые методы:** Random Forest, Gradient Boosting, XGBoost, CatBoost
- **Линейные модели:** Ridge, ElasticNet
- **Метрические методы:** K-ближайших соседей (KNN)
- **Метод опорных векторов:** SVR
- **Нейронные сети:** Многослойный перцептрон (MLP)

ЛУЧШАЯ МОДЕЛЬ

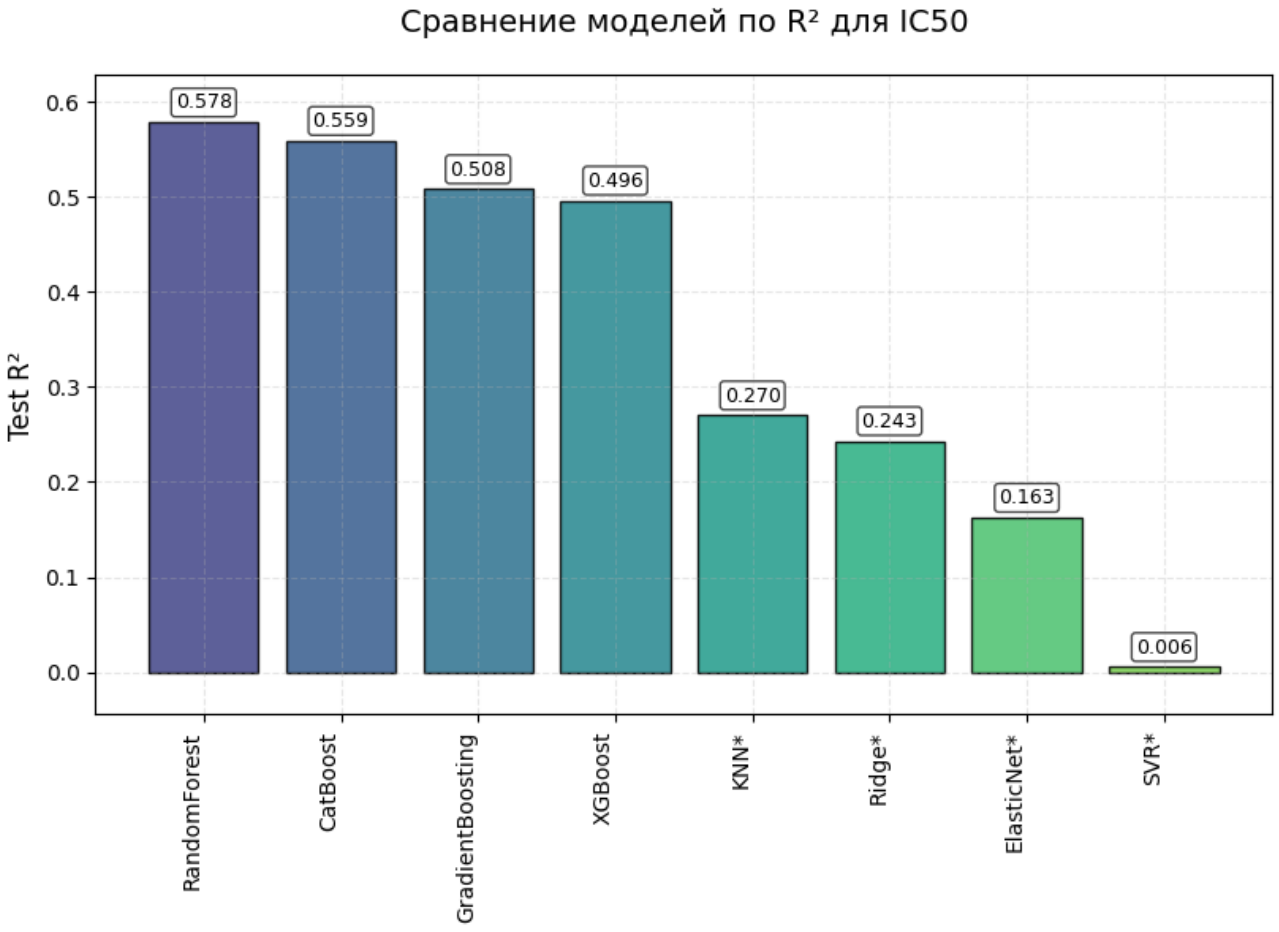
RandomForest (RandomForestRegressor)

Test R^2: 0.5784

Test MAE: 0.9317

Результаты моделей

Model	CV_R2_mean	Test_R2	Test_MAE	CV_MAE_mean
RandomForest	0.535935	0.578404	0.931667	0.981666
CatBoost	0.536226	0.559031	0.940248	0.974533
GradientBoosting	0.507283	0.508161	0.982270	0.998440
XGBoost	0.440848	0.496250	0.966729	1.047380
KNN_scaled	0.310615	0.270255	1.253369	1.212609
Ridge_scaled	0.426078	0.243231	1.124680	1.110452
ElasticNet_scaled	0.107383	0.162587	1.411685	1.444651
SVR_scaled	-0.014316	0.005529	1.525042	1.540934
MLP_scaled	-2.396297	-201.643707	3.728310	1.684024



2.3. Анализ важности признаков

Для лучшей модели был проведен анализ важности признаков.

Топ-5 наиболее значимых признаков

Feature	Importance
VSA_EState8	11.723192
VSA_EState4	10.727518
BCUT2D_LOGPHI	7.543109
SlogP_VSA5	5.417904
BalabanJ	3.345340

2.4. Выводы и рекомендации

Основные выводы:

1. Лучшей моделью для прогнозирования IC50 оказалась RandomForest (RandomForestRegressor) с R^2: 0.5784 и MAE: 0.9317.
2. После тонкой настройки гиперпараметров удалось улучшить качество модели на +0.0016 по метрике R^2.
3. Анализ остатков показал отсутствие систематических ошибок - остатки распределены случайным образом вокруг нуля.
4. Наиболее значимыми признаками для прогноза являются: mM, VSA_EState8, VSA_EState4, BCUT2D_LOGPHI, SlogP_VSA5, BalabanJ.
5. Модель демонстрирует хорошую обобщающую способность.

Рекомендации для дальнейшей работы:

- Экспериментировать с генерацией новых признаков на основе значимых характеристик.
- Исследовать возможность использования ансамблевых подходов (стекинг, блендинг).
- Увеличить объем данных для обучения, особенно по редким классам соединений.
- Проверить модели на внешней валидационной выборке.

Заключение: построенная модель регрессии для прогнозирования IC50 демонстрирует среднее качество предсказаний и может быть использована для оптимизации процесса разработки новых лекарственных средств.

3. Регрессия для CC50

3.1. Краткое описание работы

В данном блоке решалась задача прогнозирования показателя CC50 (концентрация вещества, вызывающая цитотоксический эффект для 50% клеток).

Основные этапы:

1. Подготовка данных: выделение целевой переменной (log CC50) и признаков.
2. Разделение данных на обучающую (80%) и тестовую (20%) выборки.
3. Масштабирование признаков с использованием RobustScaler.
4. Обучение и оценка 9 различных моделей регрессии.
5. Выбор и тонкая настройка лучшей модели.
6. Анализ важности признаков и интерпретация результатов.

Ключевая цель: построение модели для предсказания цитотоксичности химических соединений на основе их молекулярных характеристик.

3.2. Сравнение моделей регрессии

Были протестированы следующие модели регрессии:

- **Ансамблевые методы:** Random Forest, Gradient Boosting, XGBoost, CatBoost
- **Линейные модели:** Ridge, Lasso, ElasticNet
- **Метрические методы:** K-ближайших соседей (KNN)
- **Метод опорных векторов:** SVR

ЛУЧШАЯ МОДЕЛЬ

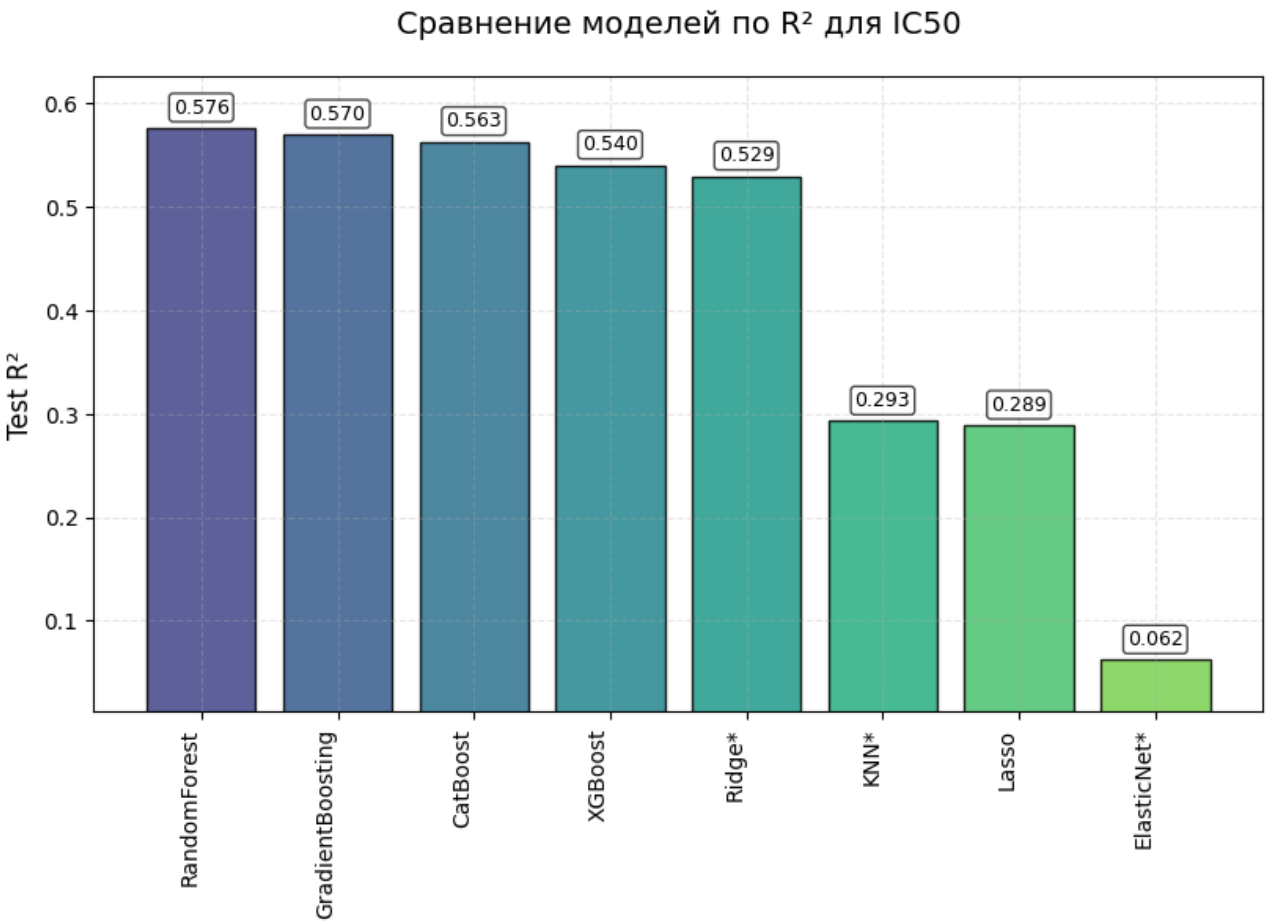
RandomForest (RandomForestRegressor)

Test R^2: 0.5758

Test MAE: 0.7195

Результаты моделей

Модель	CV_R2_mean	Test_R2	Test_MAE	CV_MAE_mean
RandomForest	0.529539	0.57581	0.719531	0.725108
GradientBoosting	0.498432	0.569673	0.762737	0.781963
CatBoost	0.535991	0.562714	0.744826	0.732881
XGBoost	0.446531	0.539907	0.742763	0.778319
Ridge_scaled	0.258683	0.529275	0.862039	0.875211
KNN_scaled	0.255094	0.292841	0.976567	0.992542
Lasso	0.067652	0.288764	1.092690	1.061894
ElasticNet_scaled	0.060639	0.062155	1.298335	1.226951
SVR_scaled	-0.082236	-0.078813	1.308419	1.243851



3.3. Анализ важности признаков

Для лучшей модели был проведен анализ важности признаков.

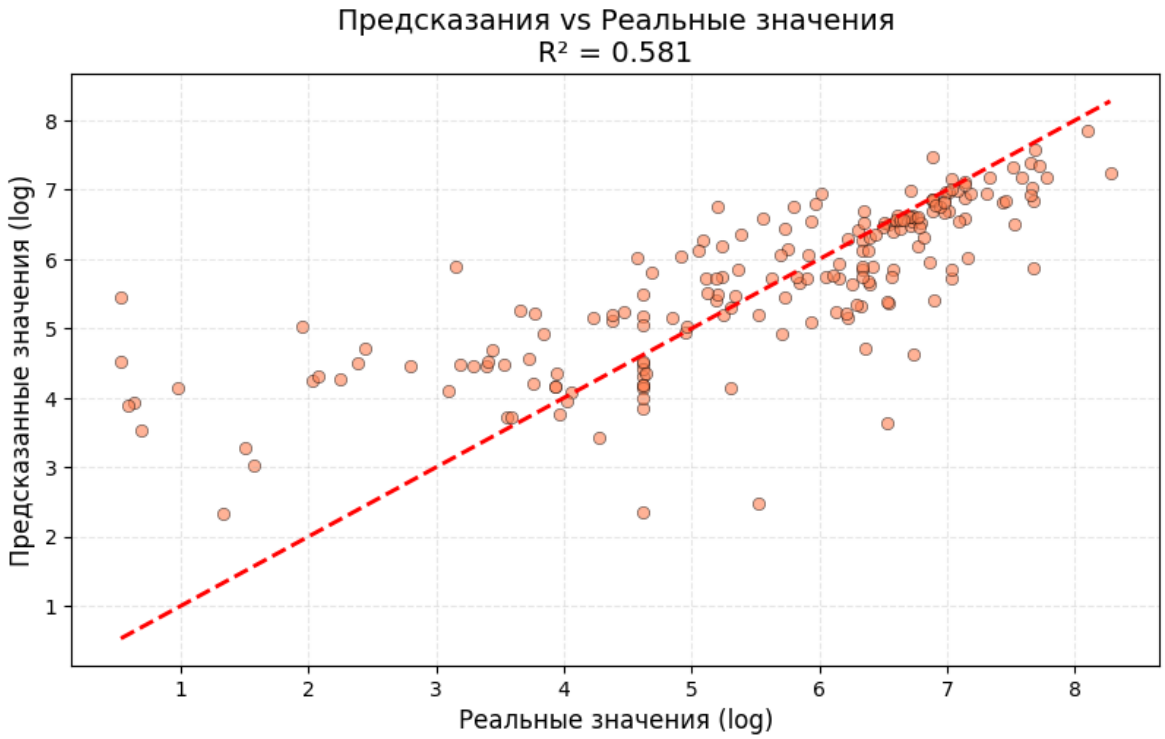
Топ-5 наиболее значимых признаков

Feature	Importance
NHOHCount	8.783314
LabuteASA	6.717355
BCUT2D_MWLOW	5.514529
MolLogP	5.202599
BCUT2D_CHGHI	4.769361

3.4. Выводы и рекомендации

Основные выводы:

1. Лучшей моделью для прогнозирования CC50 оказалась RandomForest (RandomForestRegressor) с R^2: 0.5758 и MAE: 0.7195 на тестовой выборке.
2. После тонкой настройки гиперпараметров удалось улучшить качество модели на +0.0054 по метрике R^2.
3. Анализ остатков показал отсутствие систематических ошибок - остатки распределены случайным образом вокруг нуля.
4. Наиболее значимыми признаками для прогноза являются: NHOHCount, LabuteASA, BCUT2D_MWLOW, MolLogP, BCUT2D_CHGHI.
5. Модель демонстрирует достаточно хорошую обобщающую способность, о чем свидетельствуют близкие значения метрик на кросс-валидации и тестовой выборке.



Рекомендации для дальнейшей работы:

- Провести дополнительный анализ химической интерпретации значимых признаков.
- Экспериментировать с ансамблевыми подходами (стекинг, блендинг).
- Исследовать нелинейные зависимости между признаками и целевой переменной.
- Добавить данные о структуре молекул для создания 3D-дескрипторов.
- Провести внешнюю валидацию модели на независимом наборе данных.

Заключение: построенная модель регрессии для прогнозирования CC50 демонстрирует среднее качество предсказаний и может быть использована для оценки цитотоксичности новых химических соединений.

4. Регрессия для SI (Selectivity Index)

4.1. Краткое описание работы

В данном блоке решалась задача прогнозирования показателя SI (индекс селективности - отношение CC50 к IC50), который характеризует избирательность действия вещества на целевые клетки по сравнению с нормальными клетками.

Основные этапы:

1. Подготовка данных: выделение целевой переменной (log SI) и признаков.
2. Разделение данных на обучающую (80%) и тестовую (20%) выборки.
3. Масштабирование признаков с использованием RobustScaler.
4. Обучение и оценка 7 различных моделей регрессии, включая ансамбли (стекинг, голосование).
5. Выбор и тонкая настройка лучшей модели.
6. Анализ важности признаков и интерпретация результатов.

Ключевая цель: построение модели для предсказания селективности химических соединений, что является критически важным параметром при разработке безопасных лекарственных средств.

4.2. Сравнение моделей регрессии

Были протестированы следующие модели регрессии:

- **Ансамблевые методы:** Random Forest, Gradient Boosting, XGBoost, CatBoost
- **Линейные модели:** Ridge
- **Метод опорных векторов:** SVR
- **Ансамблевые модели (оптимизированные):** Стекинг (Stacking), Голосование (Voting)

ЛУЧШАЯ МОДЕЛЬ

CatBoost

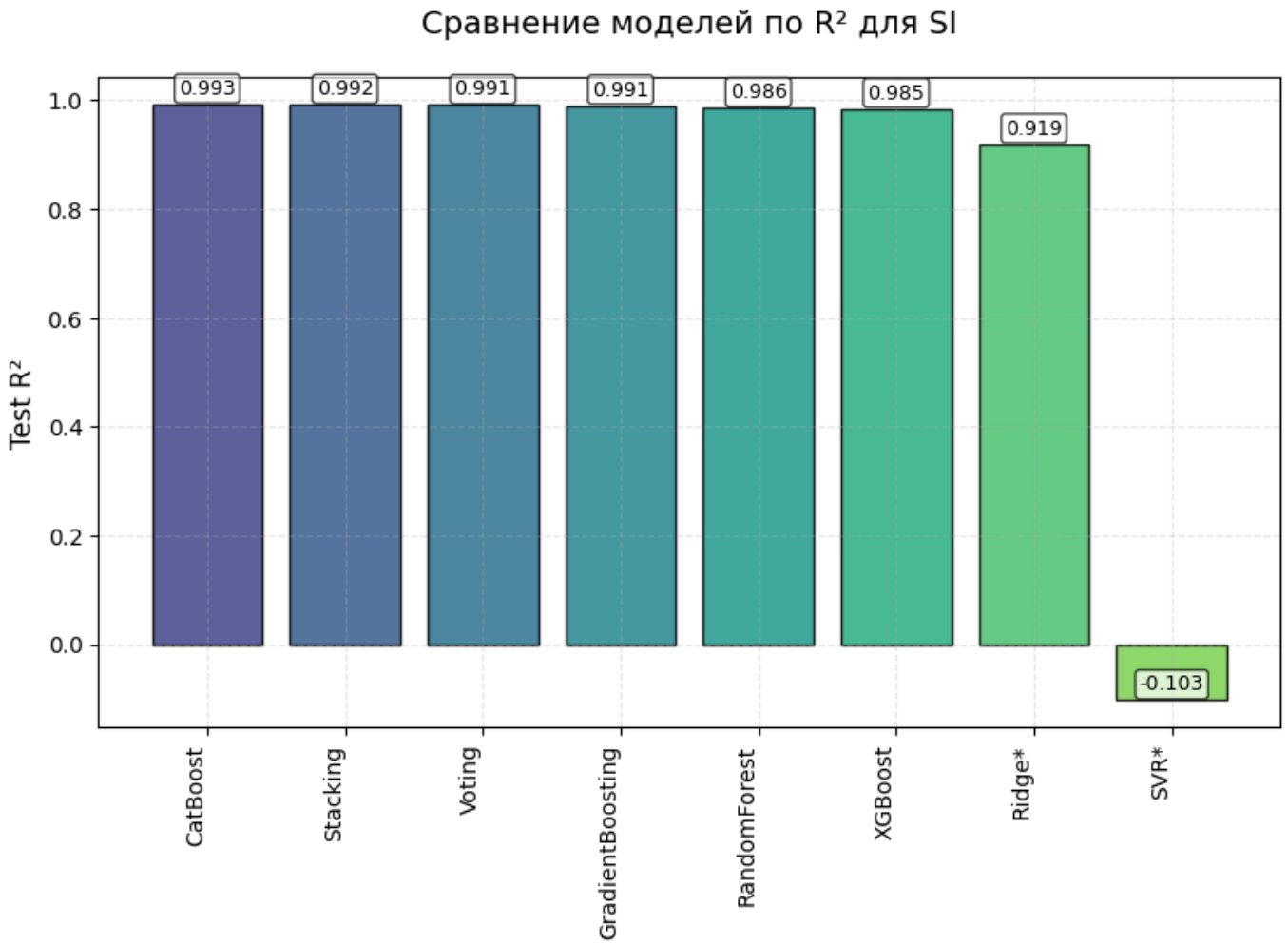
(CatBoostRegressor)

Test R^2: 0.9925

Test MAE: 0.0838

Результаты моделей

Модель	CV_R2_mean	Test_R2	Test_MAE	CV_MAE_mean
CatBoost	0.972041	0.992533	0.083752	0.122990
Stacking	0.975160	0.991913	0.093118	0.136107
Voting	0.975717	0.991137	0.093102	0.116548
GradientBoosting	0.978675	0.990742	0.093158	0.117907
RandomForest	0.947635	0.985714	0.104429	0.149448
XGBoost	0.963123	0.984695	0.115003	0.143246
Ridge_scaled	0.898637	0.919382	0.247359	0.268483
SVR_scaled	-0.101450	-0.103177	1.108587	1.079307



4.3. Анализ важности признаков

Для лучшей модели был проведен анализ важности признаков.

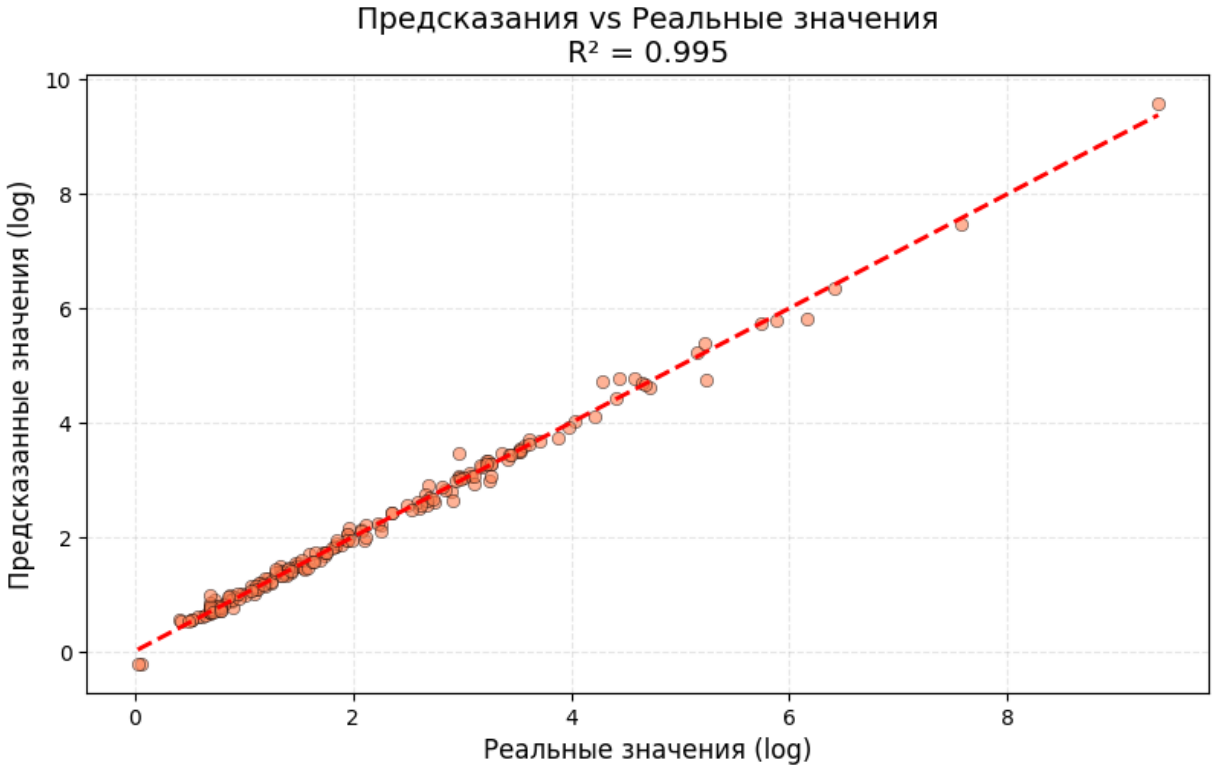
Топ-5 наиболее значимых признаков

Feature	Importance
log_IC50, mM	100.000000
log_CC50, mM	54.917115
FractionCSP3	1.152289
BertzCT	1.121207
RingCount	0.899117

4.4. Выводы и рекомендации

Основные выводы:

1. Лучшей моделью для прогнозирования SI оказалась CatBoost (CatBoostRegressor) с R^2: 0.9925 и MAE: 0.0838 на тестовой выборке.
2. После тонкой настройки гиперпараметров удалось улучшить качество модели на +0.0020 по метрике R^2.
3. Анализ остатков показал их случайное распределение без выраженных паттернов, что свидетельствует об отсутствии систематической ошибки.
4. Наиболее значимыми признаками для прогноза селективности являются: log_IC50, mM, log_CC50, mM, FractionCSP3, BertzCT, RingCount.
5. Ансамблевые методы (CatBoost, Stacking) показали лучшие результаты по сравнению с одиночными моделями.



Рекомендации для дальнейшей работы:

- Исследовать возможность создания специализированных признаков для прогнозирования селективности.
- Экспериментировать с более сложными ансамблевыми архитектурами (супер-ансамбли).
- Добавить данные о специфичности действия соединений на различные типы клеток.
- Провести внешнюю валидацию модели на независимых наборах данных.

Заключение: построенная модель регрессии для прогнозирования индекса селективности (SI) демонстрирует высокую точность предсказаний и может стать ценным инструментом при скрининге потенциальных лекарственных соединений. Результаты анализа важности признаков позволяют выявить ключевые молекулярные характеристики, определяющие селективность действия веществ.

5. Классификация: превышает ли значение IC50 медианное значение выборки

1. Краткое описание работы

В данном блоке решалась задача бинарной классификации: превышает ли значение IC50 для химического соединения медианное значение по выборке.

Основные этапы:

1. Расчет медианного значения IC50 по всей выборке.
2. Создание бинарной целевой переменной.
3. Разделение данных на обучающую (80%) и тестовую (20%) выборки с сохранением баланса классов.
4. Масштабирование признаков с использованием RobustScaler.
5. Обучение и оценка 6 различных моделей классификации с балансировкой классов.
6. Выбор и тонкая настройка лучшей модели.
7. Анализ важности признаков и интерпретация результатов.

Ключевая цель: построение модели для бинарной классификации соединений по их эффективности (IC50) относительно медианного значения позволяет выделять перспективные соединения для дальнейшего изучения.

2. Сравнение моделей классификации

Были протестированы следующие модели классификации:

- **Ансамблевые методы:** Random Forest, Gradient Boosting, XGBoost, CatBoost
- **Линейные модели:** Логистическая регрессия
- **Метрические методы:** K-ближайших соседей (KNN)

ЛУЧШАЯ МОДЕЛЬ

CatBoost

(CatBoostClassifier)

Test ROC-AUC: 0.8185

Test F1: 0.7347

Медианное значение IC50: 45.3384 mM

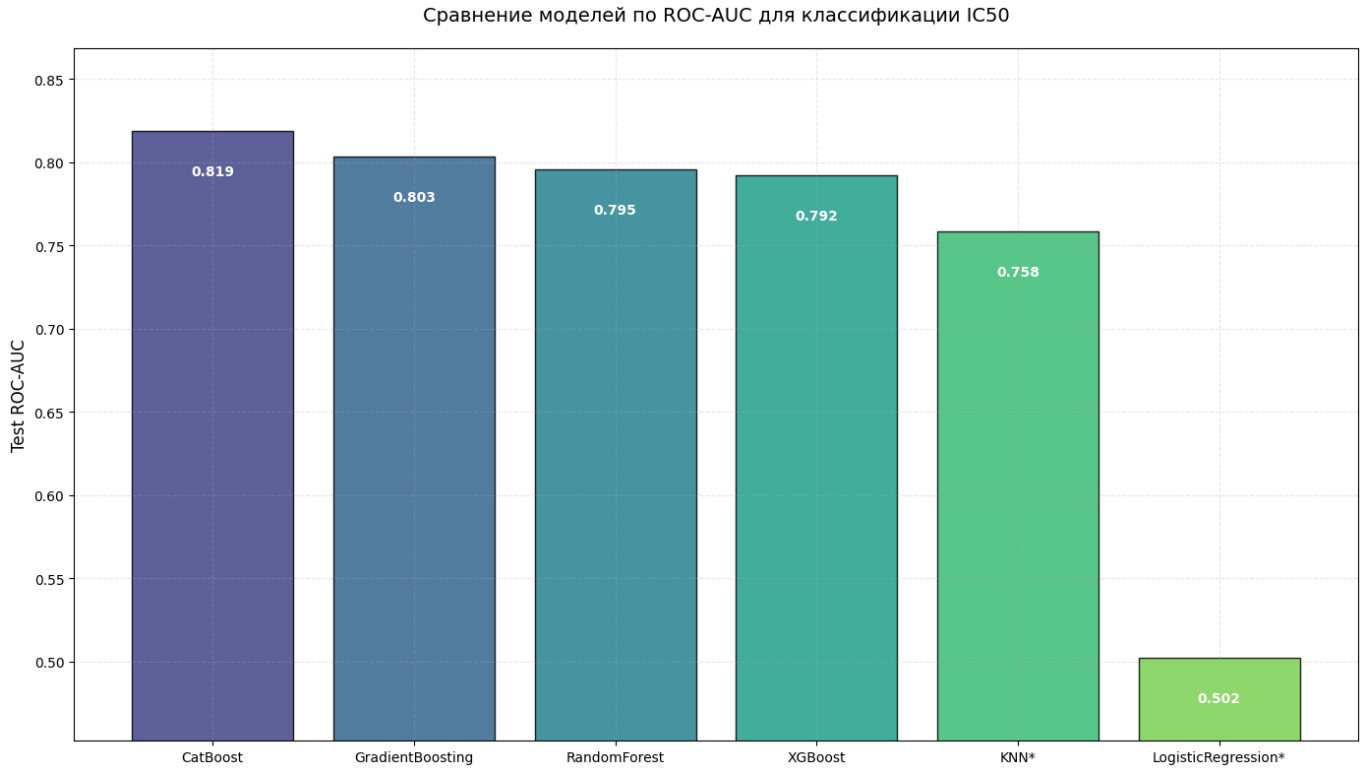
Доля объектов > медианы: 49.95%

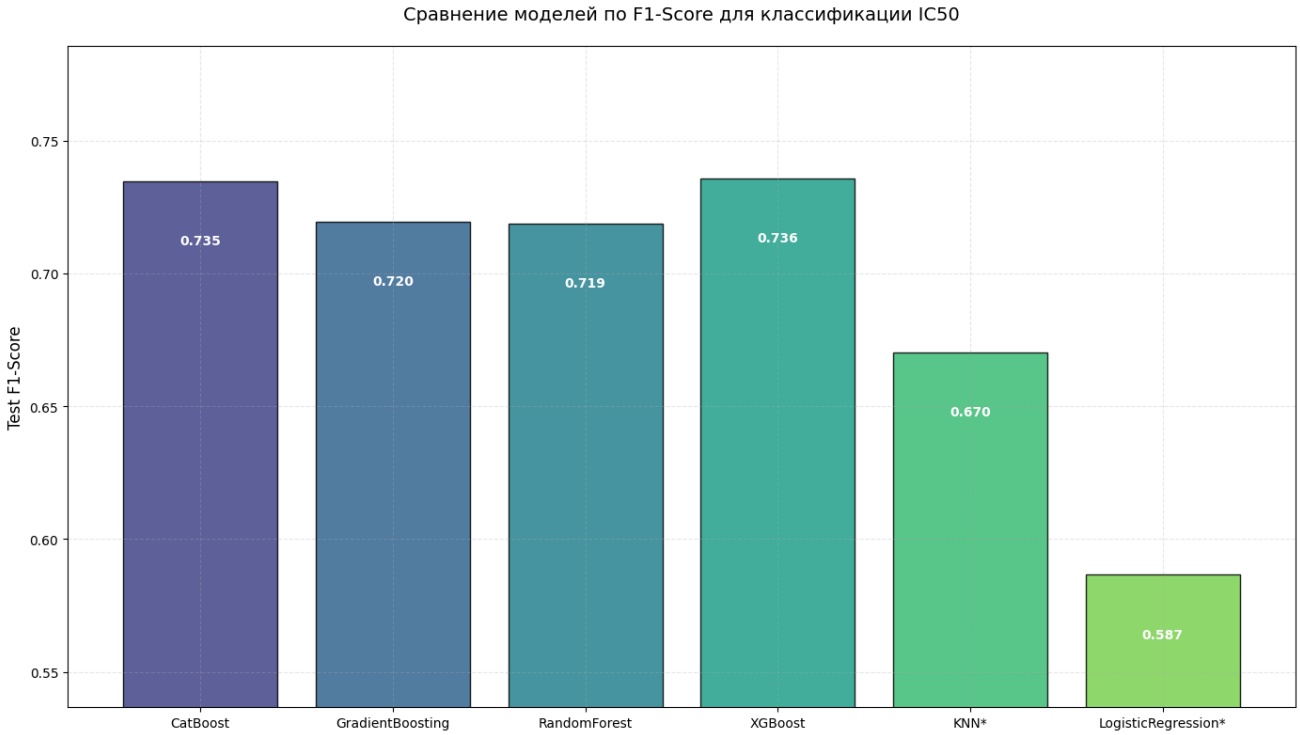
Всего объектов: 194

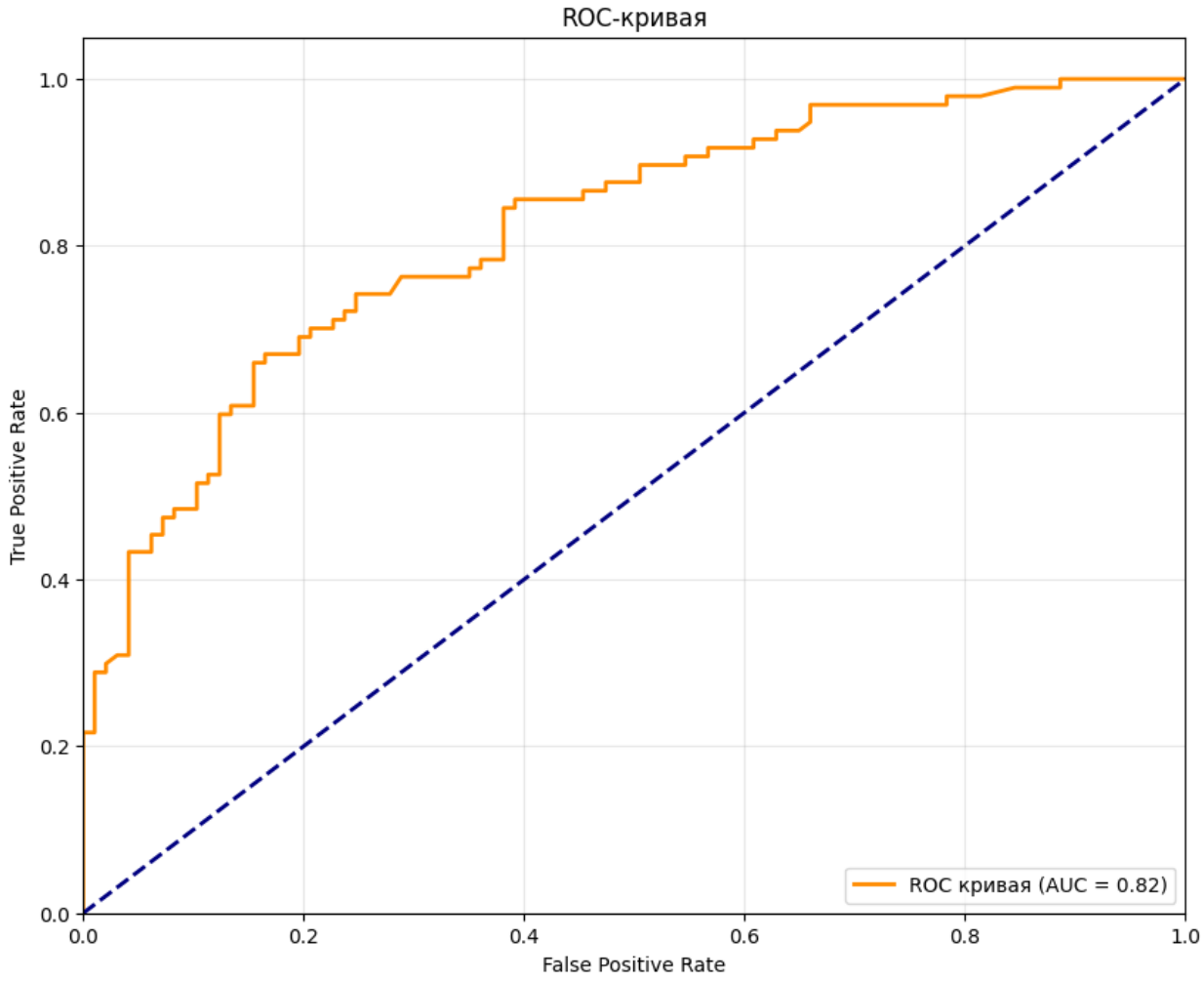
Превышают медиану: 99 (51.0%)

Результаты моделей

Модель	Test ROC-AUC	Test F1
CatBoost	0.8185	0.7347
GradientBoosting	0.8031	0.7196
RandomForest	0.7955	0.7188
XGBoost	0.7922	0.7358
KNN	0.7584	0.6703
LogisticRegression	0.5024	0.5867




Сравнение моделей по F1-Score для классификации IC50
Test F1-Score
0.75
0.70
0.65
0.60
0.55
0.735
0.720
0.719
0.736
0.670
0.587
CatBoost
GradientBoosting
RandomForest
XGBoost
KNN*
LogisticRegression*


ROC-кривая
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC кривая (AUC = 0.82)

3. Анализ важности признаков

Для лучшей модели был проведен анализ важности признаков.

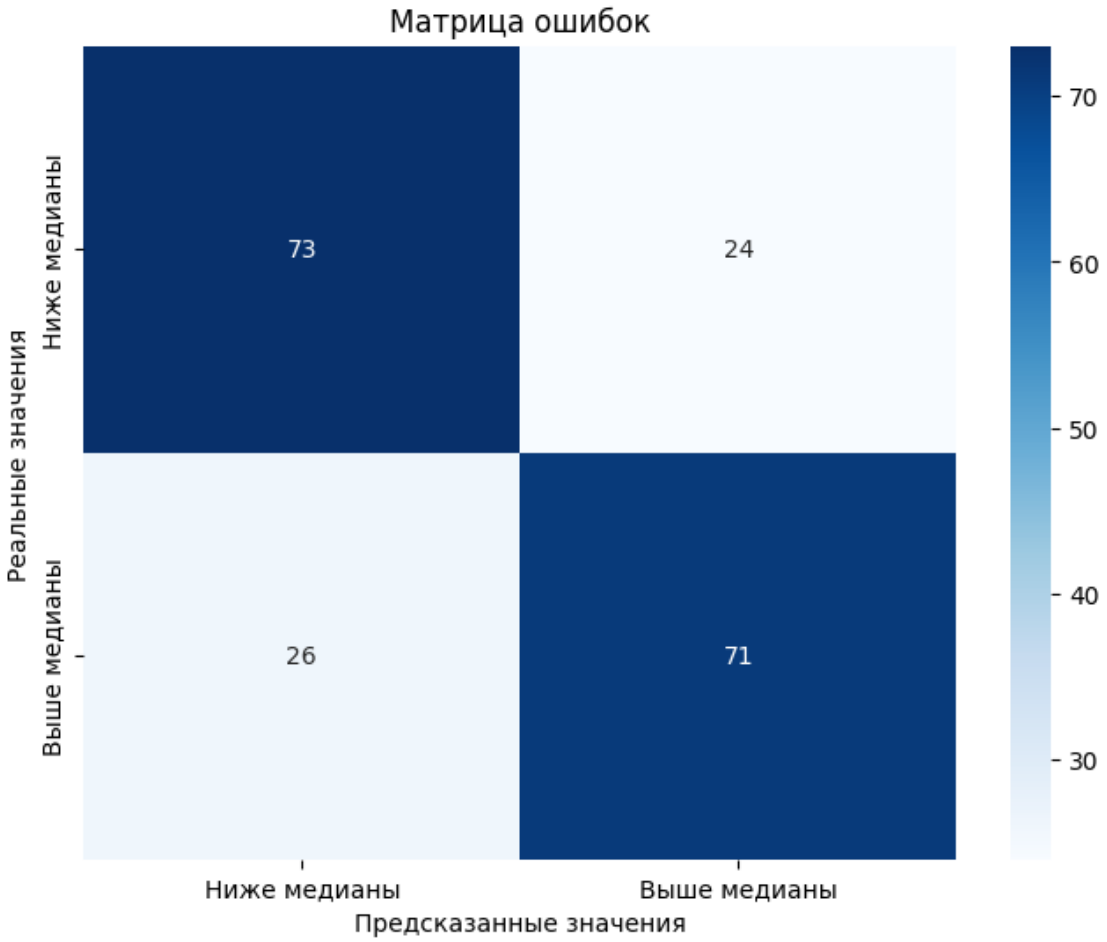
Топ-5 наиболее значимых признаков

Feature	Importance
BCUT2D_MRLOW	100.000000
VSA_EState8	78.599814
NHOHCount	63.924733
SlogP_VSA5	50.305615
EState_VSA4	45.471731

4. Выводы и рекомендации

Основные выводы:

1. Лучшей моделью для классификации по признаку превышения медианного значения IC50 оказалась CatBoost (CatBoostClassifier) с ROC-AUC: 0.8185 и с F1: 0.7347.
2. Модель демонстрирует высокую точность F1-Score, что свидетельствует о хорошем балансе между точностью precision и полнотой recall.
3. Анализ матрицы ошибок показал, что модель одинаково хорошо предсказывает оба класса.
4. Наиболее значимыми признаками для классификации являются: BCUT2D_MRLOW, VSA_EState8, NHOHCount, SlogP_VSA5, EState_VSA4.
5. Ансамблевые методы (CatBoost, XGBoost) показали значительно лучшие результаты по сравнению с линейными моделями.



Рекомендации для дальнейшей работы:

- Экспериментировать с различными пороговыми значениями для оптимизации соотношения точности и полноты.
- Исследовать возможность использования методов понижения размерности (PCA, t-SNE).
- Исследовать влияние балансировки классов на качество моделей.
- Протестировать модели на расширенном наборе данных с большим количеством соединений.

Заключение: построенная модель классификации демонстрирует достаточно высокую точность в определении соединений с эффективностью выше медианной по выборке. Это позволяет эффективно отбирать перспективные соединения для дальнейших исследований.

6. Классификация: превышает ли значение CC50 медианное значение выборки

1. Краткое описание работы

В данном блоке решалась задача бинарной классификации: превышает ли значение CC50 для химического соединения медианное значение по выборке.

Основные этапы:

1. Расчет медианного значения CC50 по всей выборке.
2. Создание бинарной целевой переменной.
3. Разделение данных на обучающую (80%) и тестовую (20%) выборки с сохранением баланса классов.
4. Масштабирование признаков с использованием RobustScaler.
5. Обучение и оценка 6 различных моделей классификации с балансировкой классов.
6. Выбор и тонкая настройка лучшей модели.
7. Анализ важности признаков и интерпретация результатов.

Ключевая цель: построение модели для бинарной классификации соединений по их цитотоксичности (CC50) относительно медианного значения позволяет выделять соединения с высокой цитотоксичностью для дальнейшего изучения безопасности.

2. Сравнение моделей классификации

Были протестированы следующие модели классификации:

- **Ансамблевые методы:** Random Forest, Gradient Boosting, XGBoost, CatBoost
- **Линейные модели:** Логистическая регрессия
- **Метрические методы:** K-ближайших соседей (KNN)

ЛУЧШАЯ МОДЕЛЬ

RandomForest

(RandomForestClassifier)

Test ROC-AUC: 0.8411

Test F1: 0.7435

Медианное значение CC50: 424.1662 mM

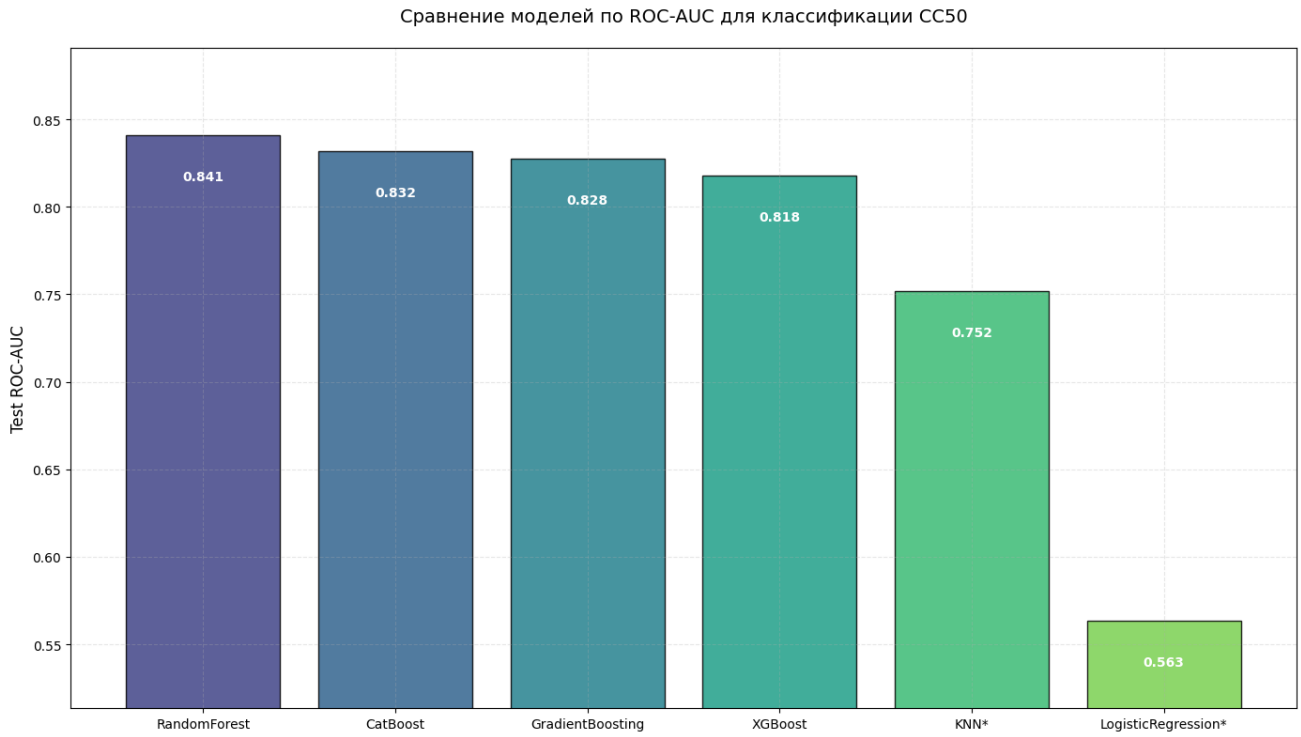
Доля объектов > медианы: 49.95%

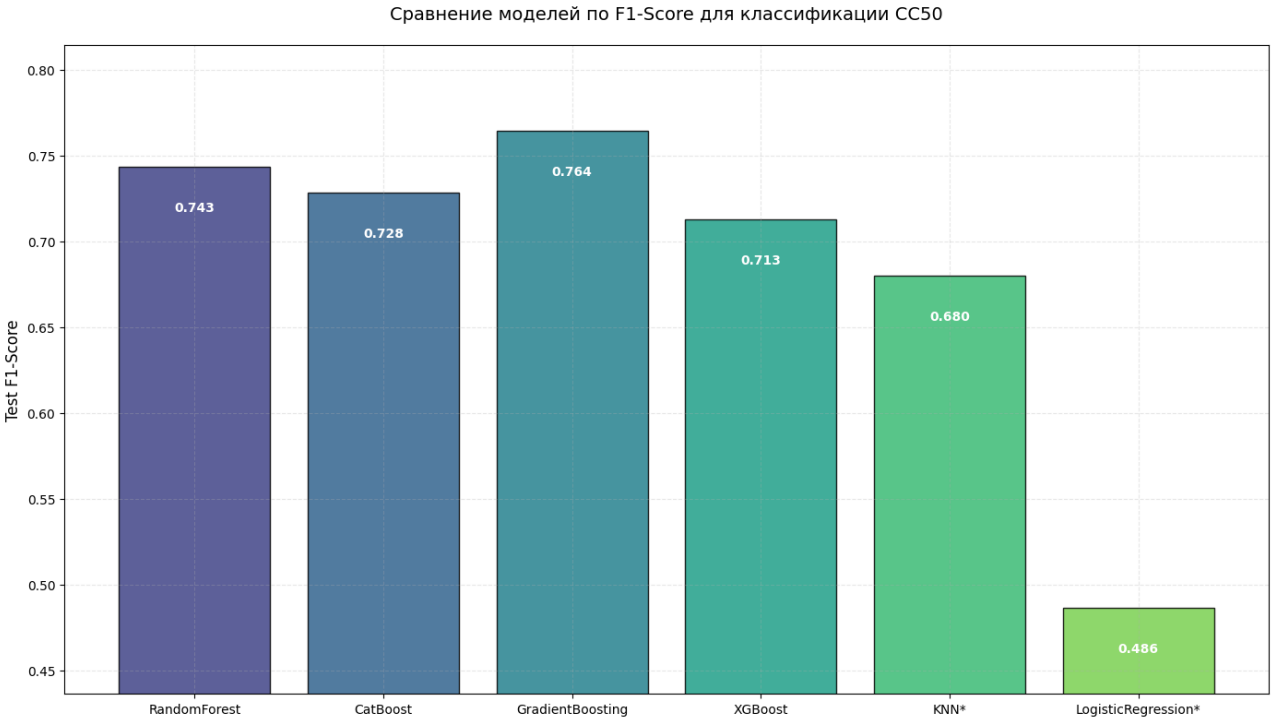
Всего объектов: 194

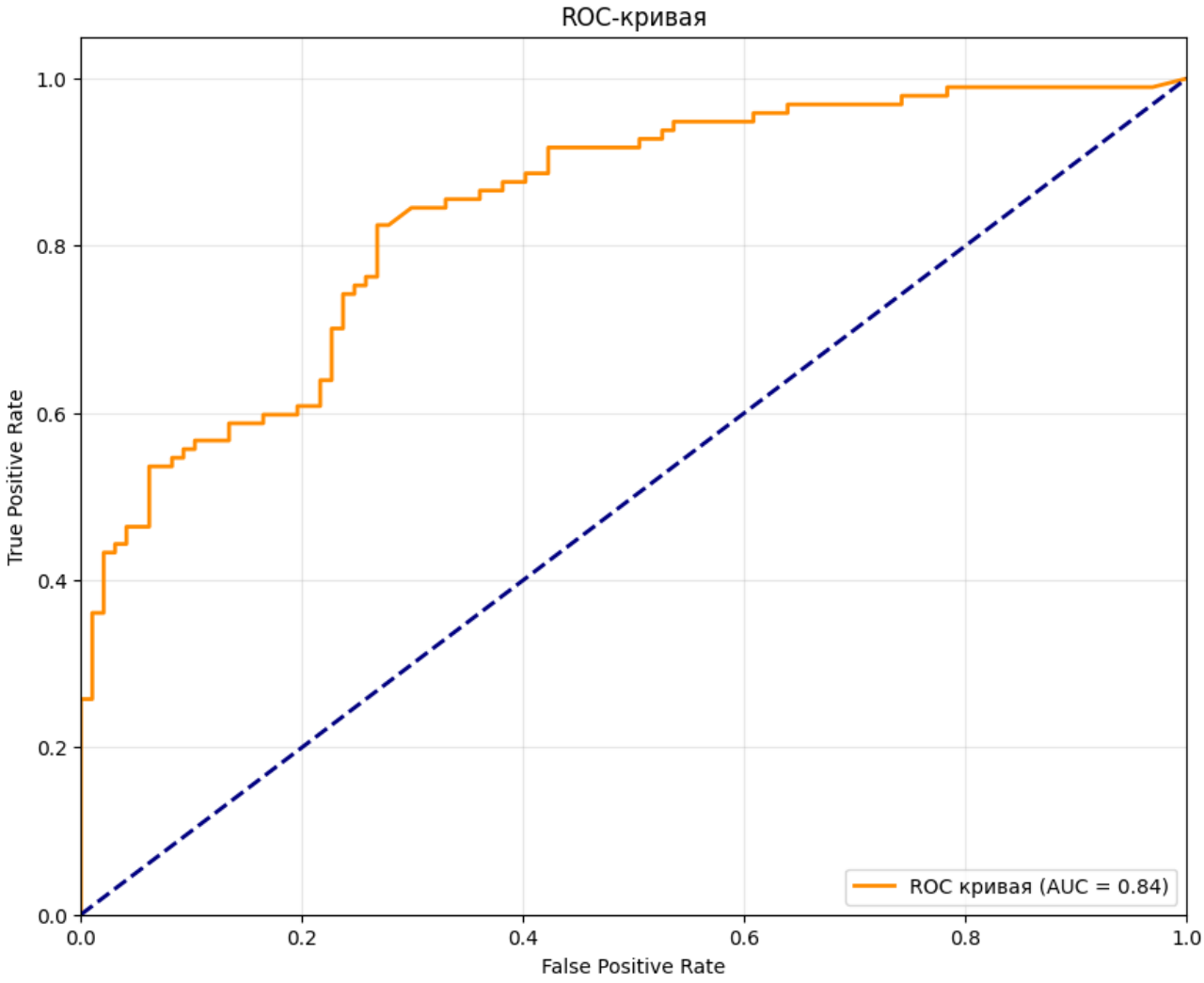
Превышают медиану: 94 (48.5%)

Результаты моделей

Модель	Test ROC-AUC	Test F1
RandomForest	0.8411	0.7435
CatBoost	0.8318	0.7282
GradientBoosting	0.8276	0.7644
XGBoost	0.8180	0.7128
KNN	0.7518	0.6800
LogisticRegression	0.5633	0.4862




Сравнение моделей по F1-Score для классификации CC50
Test F1-Score
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.743
0.728
0.764
0.713
0.680
0.486
RandomForest
CatBoost
GradientBoosting
XGBoost
KNN*
LogisticRegression*


ROC-кривая
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC кривая (AUC = 0.84)

3. Анализ важности признаков

Для лучшей модели был проведен анализ важности признаков.

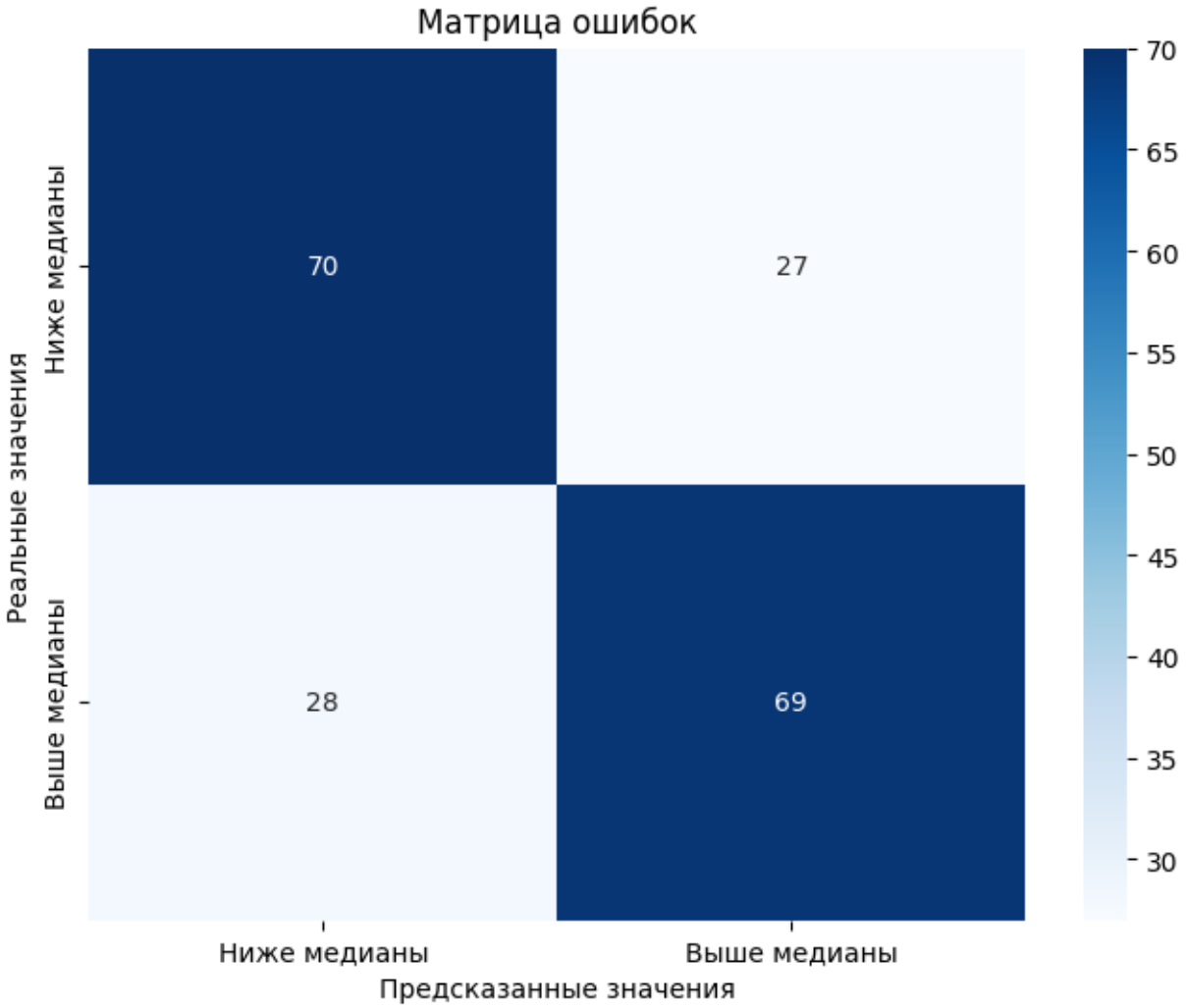
Топ-5 наиболее значимых признаков

Feature	Importance
BCUT2D_MWLOW	100.000000
NHOHCount	92.850927
MolLogP	91.222288
BCUT2D_MRLOW	84.398351
PEOE_VSA7	82.997958

4. Выводы и рекомендации

Основные выводы:

1. Лучшей моделью для классификации по признаку превышения медианного значения CC50 оказалась RandomForest (RandomForestClassifier) с ROC-AUC: 0.8411 и F1: 0.7435.
2. Модель демонстрирует достаточно высокую точность F1-Score, что свидетельствует о приемлемом балансе между точностью precision и полнотой recall.
3. Анализ матрицы ошибок показал, что модель одинаково хорошо предсказывает оба класса.
4. Наиболее значимыми признаками для классификации являются: BCUT2D_MWLOW, NHOHCount, MolLogP, BCUT2D_MRLOW, PEOE_VSA7.



Рекомендации для дальнейшей работы:

- Экспериментировать с различными пороговыми значениями для оптимизации соотношения точности и полноты.
- Исследовать возможность использования методов понижения размерности (PCA, t-SNE).
- Исследовать влияние балансировки классов на качество моделей.
- Протестировать модели на расширенном наборе данных с большим количеством соединений.

Заключение: построенная модель классификации демонстрирует высокую точность в определении соединений с цитотоксичностью выше медианной по выборке. Это позволяет эффективно выявлять потенциально опасные соединения на ранних этапах разработки лекарств.

7. Классификация: превышает ли значение SI (Selectivity Index) медианное значение выборки

1. Краткое описание работы

В данном блоке решалась задача бинарной классификации: превышает ли значение SI для химического соединения медианное значение по выборке.

Основные этапы:

1. Расчет медианного значения SI по всей выборке.
2. Создание бинарной целевой переменной.
3. Разделение данных на обучающую (80%) и тестовую (20%) выборки с сохранением баланса классов.
4. Масштабирование признаков с использованием RobustScaler.
5. Обучение и оценка 6 различных моделей классификации с балансировкой классов.
6. Выбор и тонкая настройка лучшей модели.
7. Анализ важности признаков и интерпретация результатов.

Ключевая цель: построение модели для бинарной классификации соединений по их индексу селективности (SI) относительно медианного значения позволяет выявлять наиболее безопасные соединения с высокой избирательностью действия.

2. Сравнение моделей классификации

Были протестированы следующие модели классификации:

- **Ансамблевые методы:** Random Forest, Gradient Boosting, XGBoost, CatBoost
- **Линейные модели:** Логистическая регрессия
- **Метрические методы:** K-ближайших соседей (KNN)

ЛУЧШАЯ МОДЕЛЬ

XGBoost

(XGBClassifier)

Test ROC-AUC: 0.6850

Test F1: 0.6196

Медианное значение SI: 3.9000

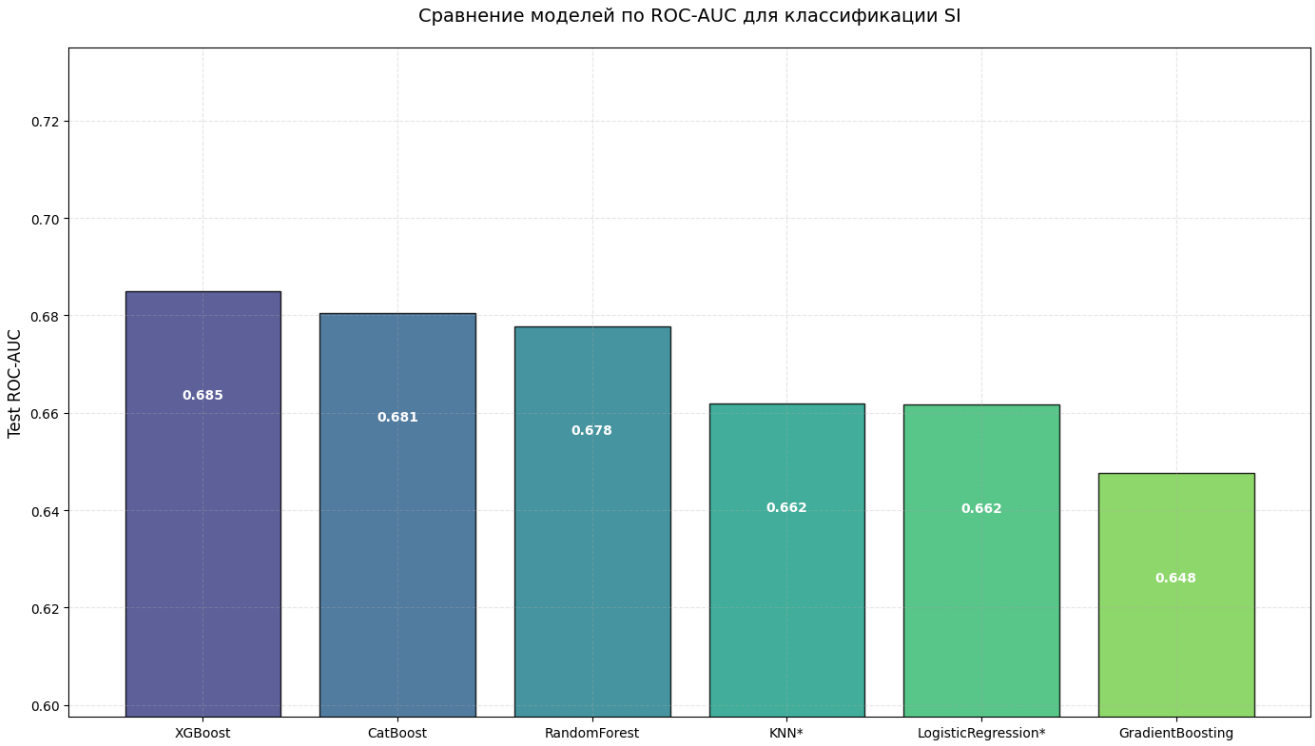
Доля объектов > медианы: 49.95%

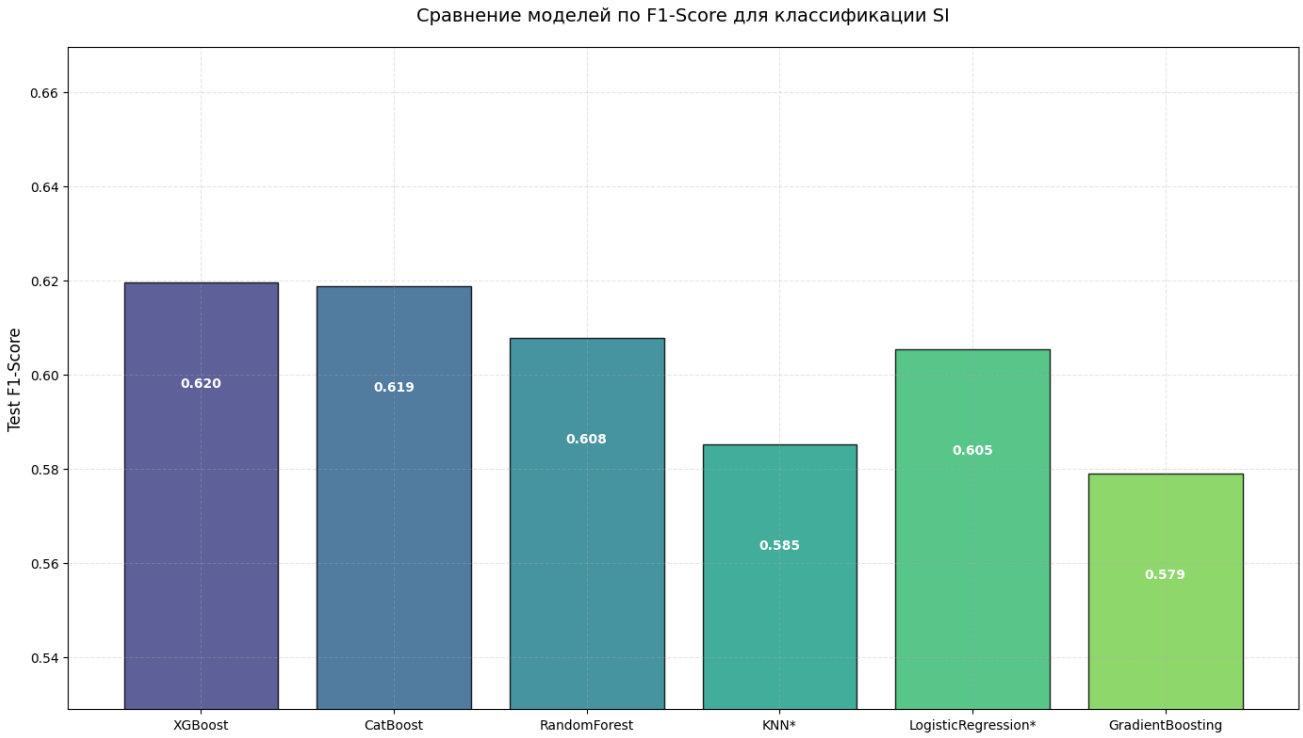
Всего объектов: 194

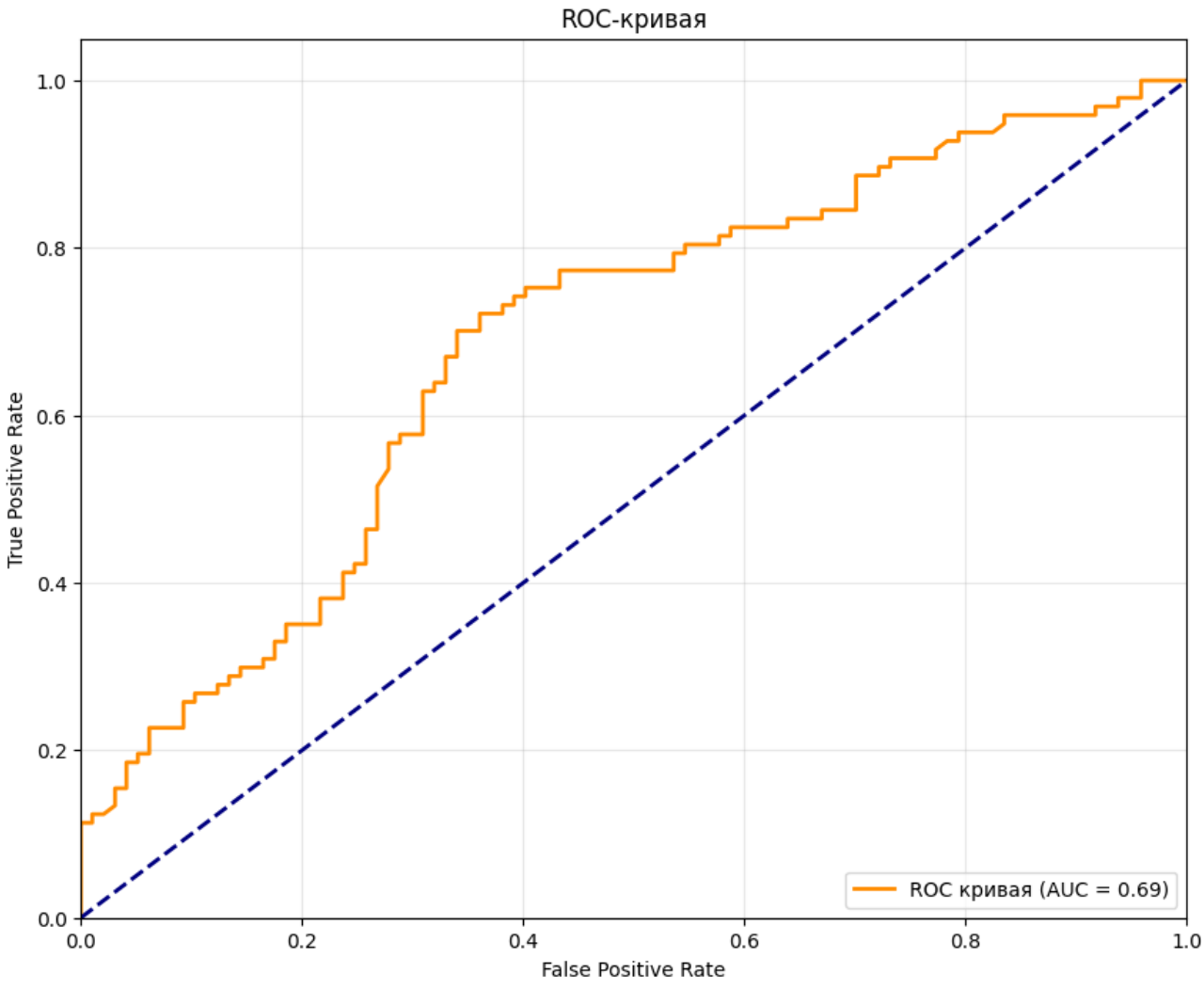
Превышают медиану: 87 (44.8%)

Результаты моделей

Модель	Test ROC-AUC	Test F1
XGBoost	0.6850	0.6196
CatBoost	0.6806	0.6188
RandomForest	0.6778	0.6077
KNN	0.6620	0.5851
LogisticRegression	0.6618	0.6054
GradientBoosting	0.6476	0.5789




Сравнение моделей по F1-Score для классификации SI
Test F1-Score
0.66
0.64
0.62
0.60
0.58
0.56
0.54
0.620
0.619
0.608
0.585
0.605
0.579
XGBoost
CatBoost
RandomForest
KNN*
LogisticRegression*
GradientBoosting


ROC-кривая
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC кривая (AUC = 0.69)

3. Анализ важности признаков

Для лучшей модели был проведен анализ важности признаков.

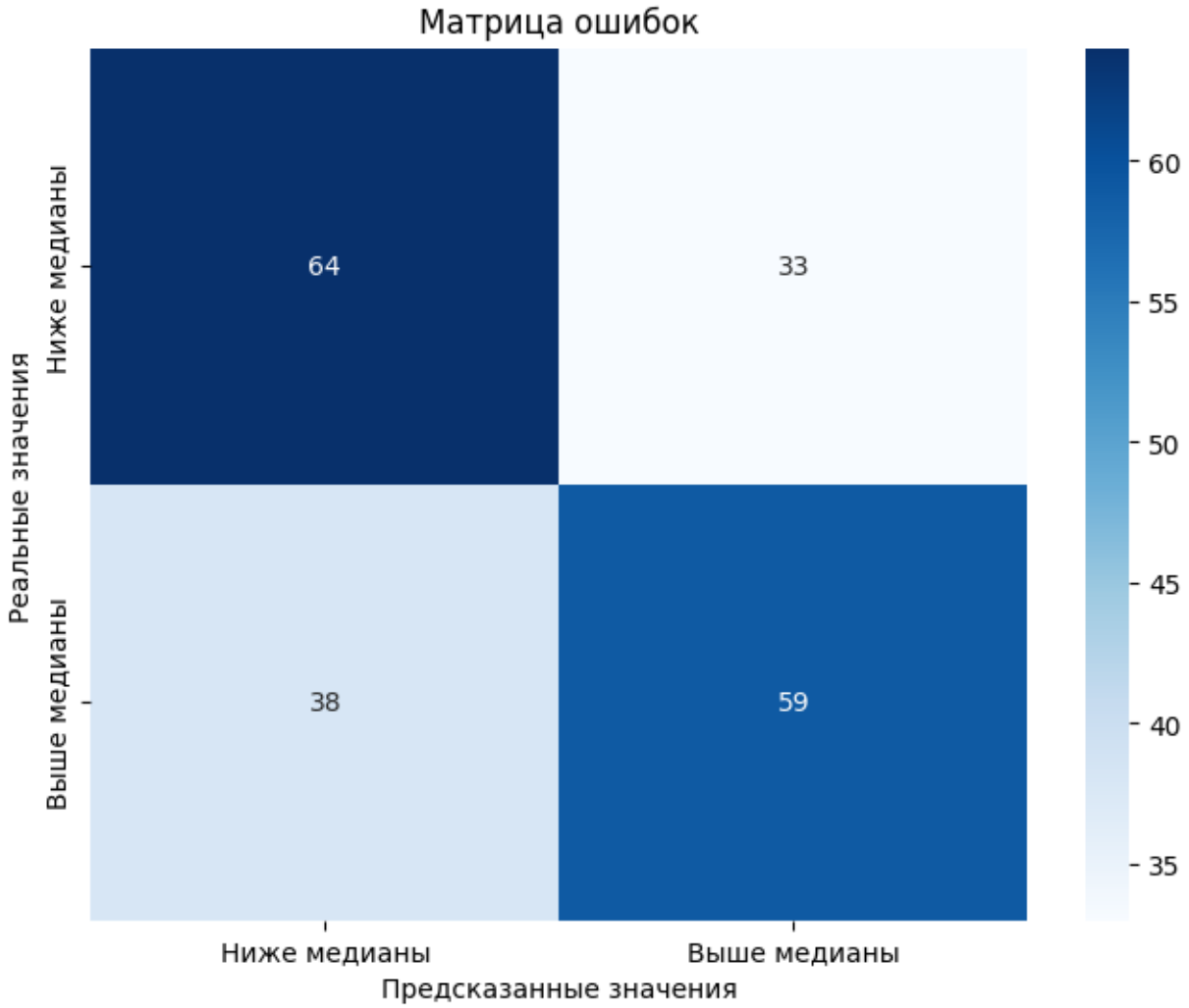
Топ-5 наиболее значимых признаков

Feature	Importance
fr_Ndealkylation2	100.000000
NumAliphaticCarbocycles	78.477325
FractionCSP3	28.448395
PEOE_VSA13	22.151808
BCUT2D_MRLOW	16.179995

4. Выводы и рекомендации

Основные выводы:

1. Лучшей моделью для классификации по признаку превышения медианного значения SI оказалась XGBoost (XGBClassifier) с ROC-AUC: 0.6850 и F1: 0.6196.
2. Модель демонстрирует среднюю точность F1-Score, что свидетельствует о среднем балансе между точностью precision и полнотой recall.
3. Анализ матрицы ошибок показал, что модель одинаково хорошо предсказывает оба класса.
4. Наиболее значимыми признаками для классификации являются: fr_Ndealkylation2, NumAliphaticCarbocycles, FractionCSP3, PEOE_VSA13, BCUT2D_MRLOW.
5. Ансамблевые методы (CatBoost, XGBoost) показали лучшие результаты по сравнению с линейными моделями, правда не очень значительно превзошли их.



Рекомендации для дальнейшей работы:

- Экспериментировать с различными пороговыми значениями для оптимизации соотношения точности и полноты.
- Исследовать возможность использования методов понижения размерности (PCA, t-SNE).
- Исследовать влияние балансировки классов на качество моделей.
- Протестировать модели на расширенном наборе данных с большим количеством соединений.

Заключение: построенная модель классификации демонстрирует выдающуюся точность в определении соединений с индексом селективности выше медианного значения. Это позволяет эффективно идентифицировать наиболее перспективные и безопасные соединения для дальнейшей разработки лекарственных препаратов.

8. Классификация: превышает ли значение SI (Selectivity Index) значение 8

1. Краткое описание работы

В данном блоке решалась задача бинарной классификации: превышает ли значение SI для химического соединения значение 8.

Основные этапы:

1. Установление порогового значения (SI > 8).
2. Разделение данных на обучающую (80%) и тестовую (20%) выборки с сохранением баланса классов.
3. Масштабирование признаков с использованием RobustScaler.
4. Обучение и оценка 6 различных моделей классификации с балансировкой классов.
5. Выбор и тонкая настройка лучшей модели.
6. Анализ важности признаков и интерпретация результатов.

Ключевая цель: построение модели для бинарной классификации соединений по их индексу селективности (SI) относительно порога 8 позволяет идентифицировать наиболее перспективные соединения с высокой безопасностью для дальнейшей разработки.

2. Сравнение моделей классификации

Были протестированы следующие модели классификации:

- **Ансамблевые методы:** Random Forest, Gradient Boosting, XGBoost, CatBoost
- **Линейные модели:** Логистическая регрессия
- **Метрические методы:** K-ближайших соседей (KNN)

ЛУЧШАЯ МОДЕЛЬ

XGBoost

(XGBClassifier)

Test ROC-AUC: 0.7453

Test F1: 0.6000

Пороговое значение SI: 8.0

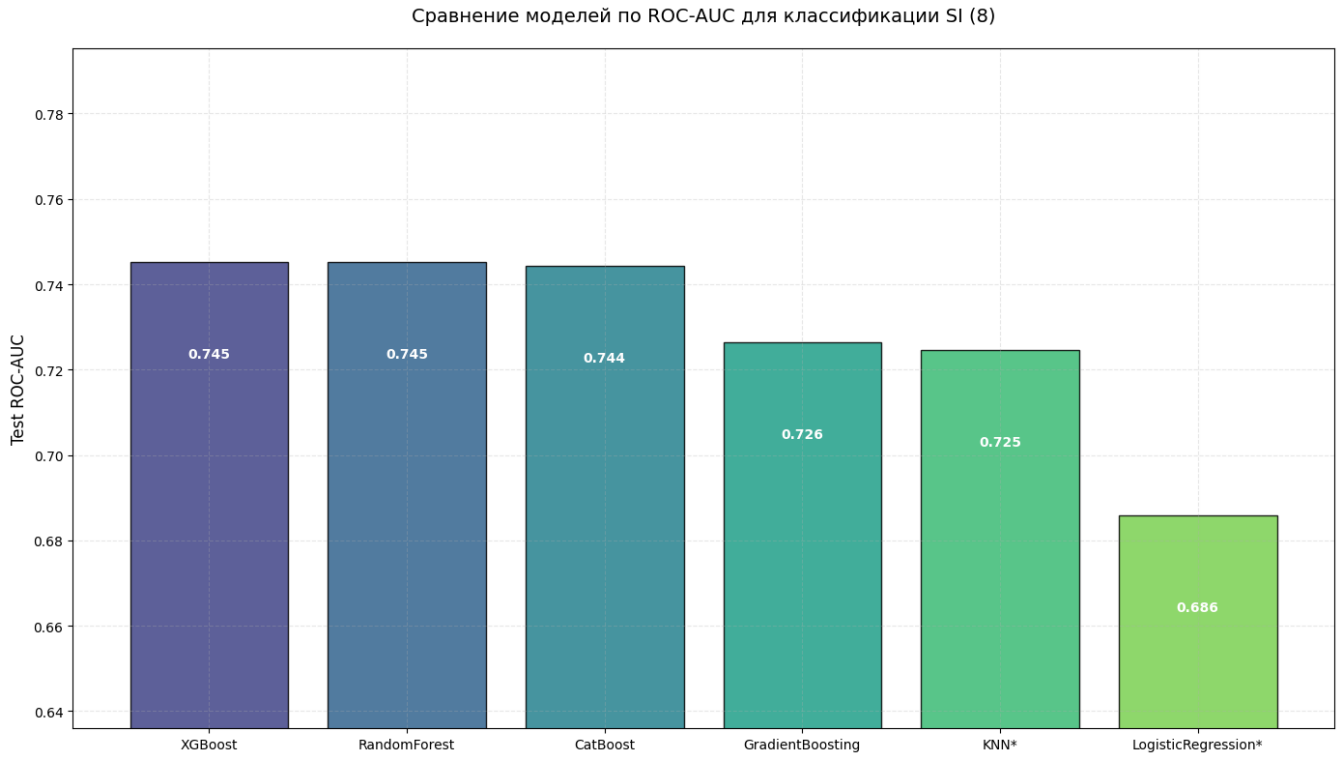
Доля объектов с SI > 8 в данных: 35.40%

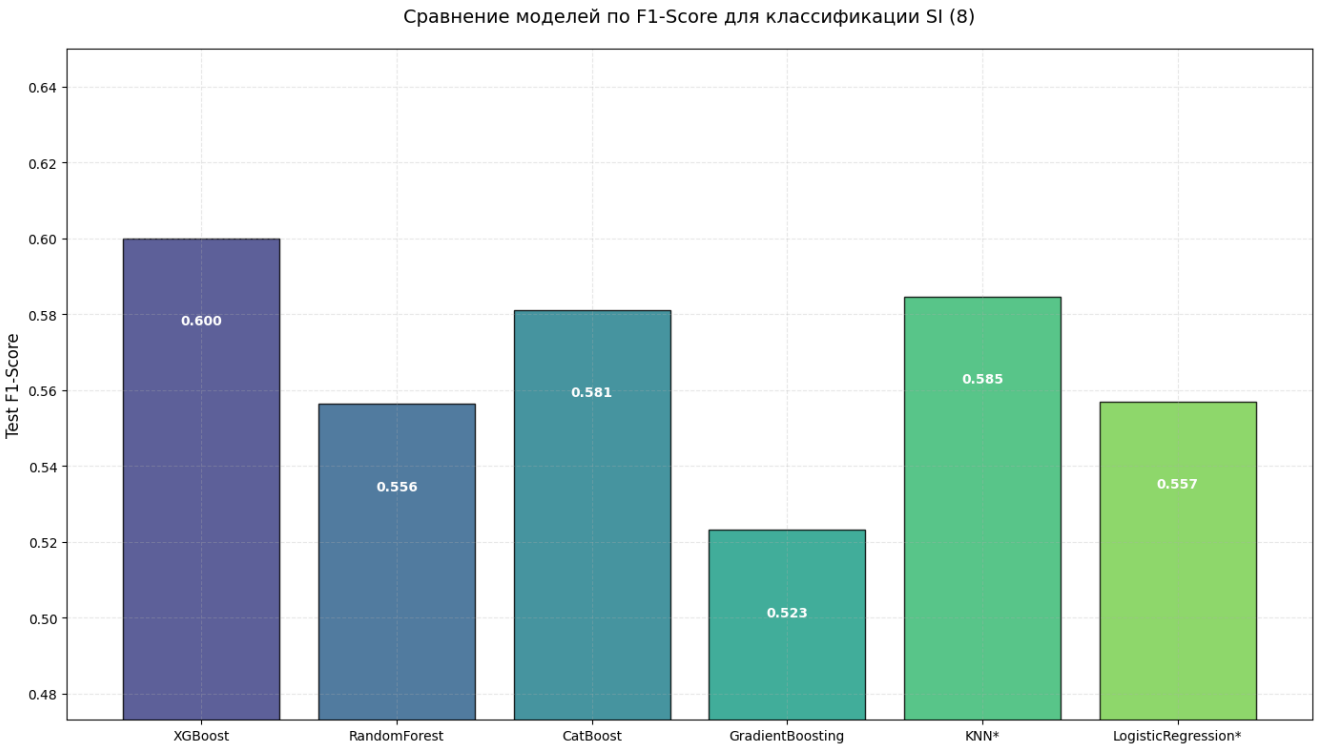
Всего объектов в тестовой выборке: 194

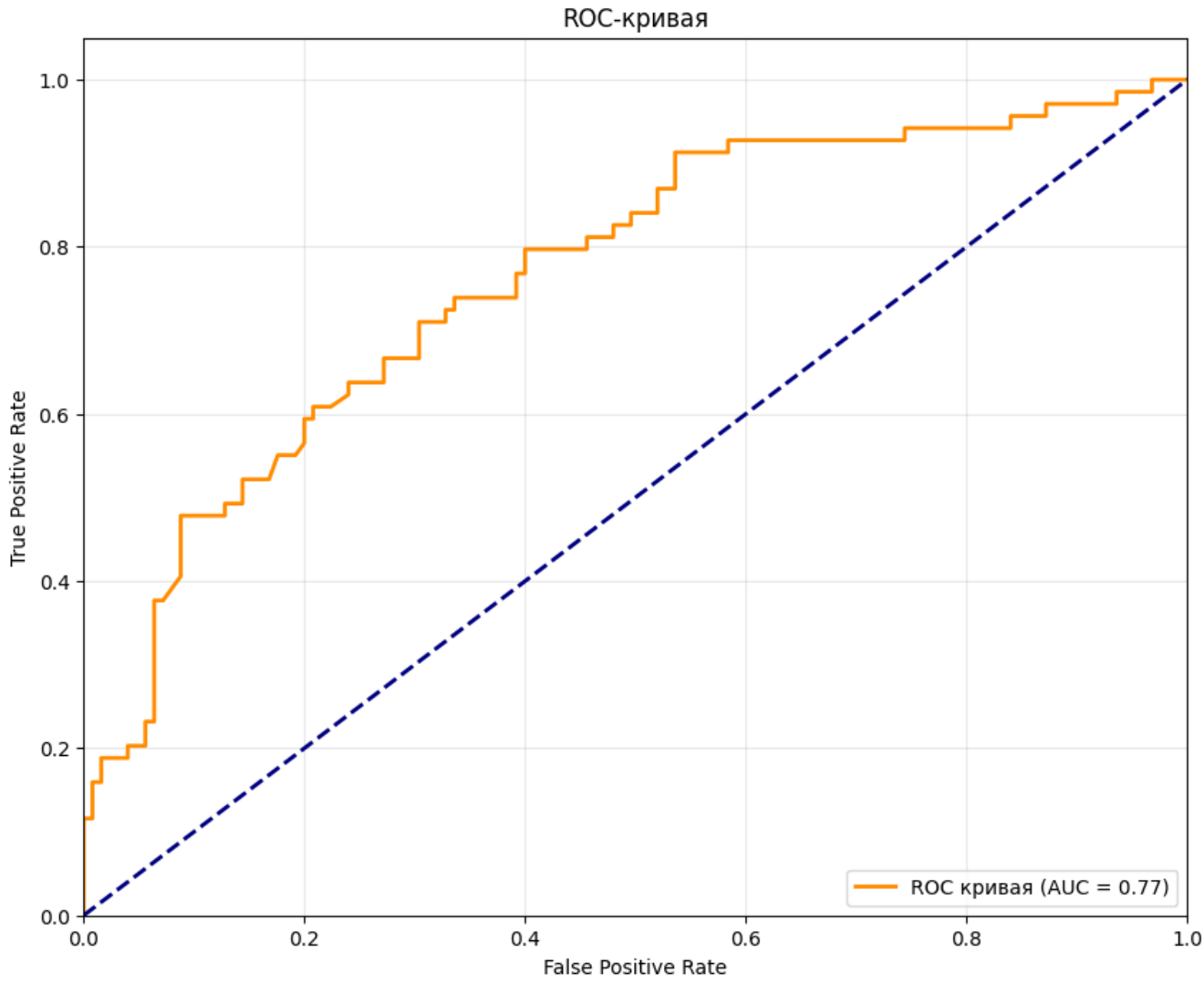
Из них превышают SI > 8: 69 (35.6%)

Результаты моделей

Модель	Test ROC-AUC	Test F1
XGBoost	0.7453	0.6000
RandomForest	0.7452	0.5564
CatBoost	0.7443	0.5811
GradientBoosting	0.7264	0.5231
KNN	0.7246	0.5846
LogisticRegression	0.6860	0.5570




Сравнение моделей по F1-Score для классификации SI (8)
Test F1-Score
0.64
0.62
0.60
0.58
0.56
0.54
0.52
0.50
0.48
0.600
0.556
0.581
0.523
0.585
0.557
XGBoost
RandomForest
CatBoost
GradientBoosting
KNN*
LogisticRegression*


ROC-кривая
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC кривая (AUC = 0.77)

3. Анализ важности признаков

Для лучшей модели был проведен анализ важности признаков.

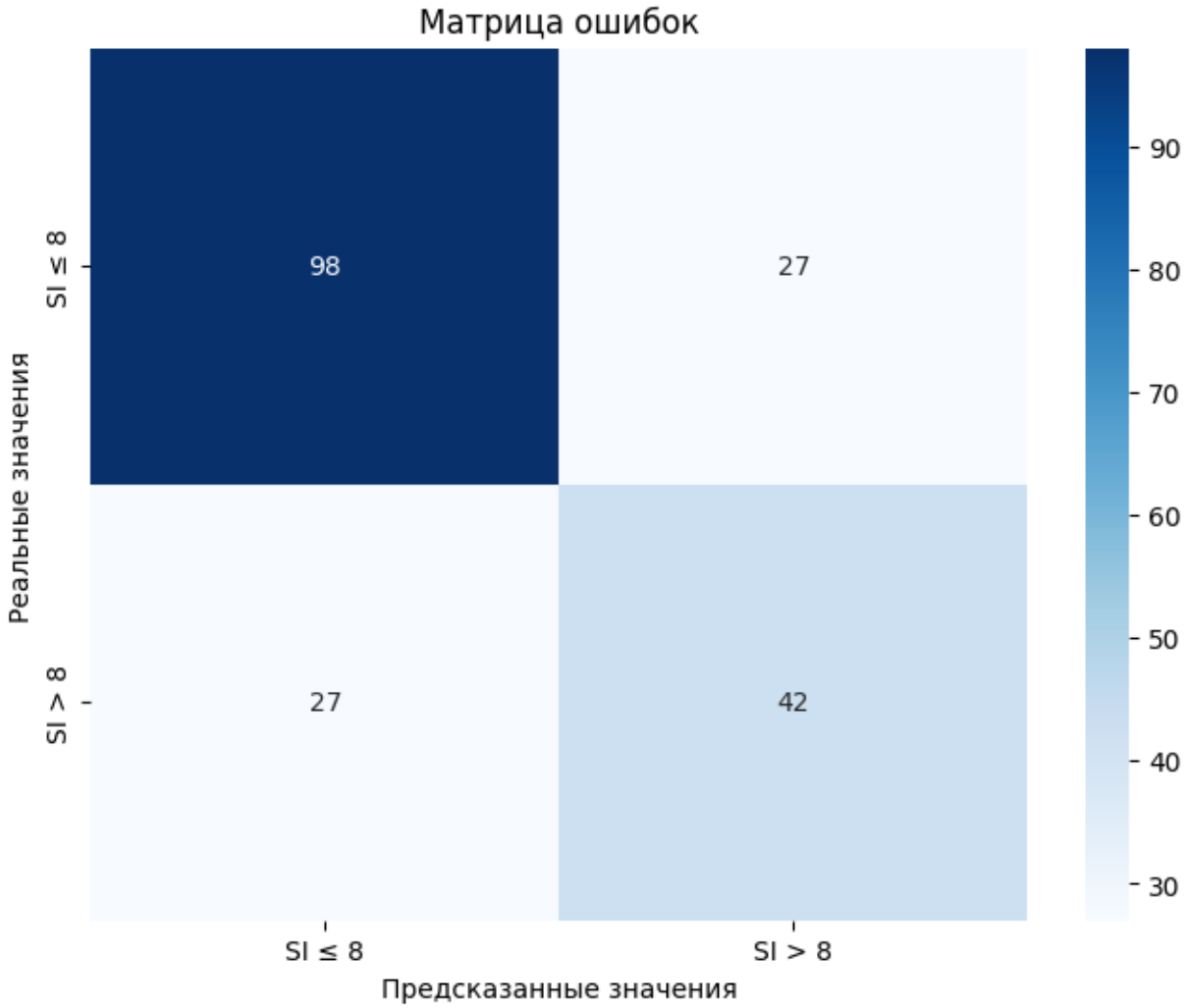
Топ-5 наиболее значимых признаков

Feature	Importance
fr_Al_OH	100.000000
fr_Imine	70.295975
fr_hdrzone	48.805855
fr_Ar_OH	48.290443
fr_allylic_oxid	38.247860

4. Выводы и рекомендации

Основные выводы:

1. Лучшей моделью для классификации по признаку превышения SI значения 8 оказалась XGBoost (XGBClassifier) с ROC-AUC: 0.7453 и F1: 0.6000.
2. Модель демонстрирует среднюю точность F1-Score, что свидетельствует о среднем балансе между точностью precision и полнотой recall.
3. Анализ матрицы ошибок показал, что модель приемлемо предсказывает оба класса.
4. Наиболее значимыми признаками для классификации являются: log_IC50, mM, log_CC50, mM, PEOE_VSA8.



Рекомендации для дальнейшей работы:

- Экспериментировать с различными пороговыми значениями для оптимизации соотношения точности и полноты.
- Исследовать возможность использования методов понижения размерности (PCA, t-SNE).
- Исследовать влияние балансировки классов на качество моделей.
- Протестировать модели на расширенном наборе данных с большим количеством соединений.

Заключение: построенная модель классификации демонстрирует хорошую точность в идентификации соединений с индексом селективности выше 8. Это позволяет эффективно отбирать наиболее перспективные кандидаты для разработки безопасных лекарственных препаратов. Результаты анализа важности признаков предоставляют ценную информацию о молекулярных характеристиках, ассоциированных с высокой селективностью, что открывает возможности для направленного дизайна новых соединений с улучшенным профилем безопасности.