Технический отчет по разработке и оценке моделей для прогнозирования эффективности лекарственных препаратов

Исследование производительности регрессионных и классификационных алгоритмов на данных химических соединений

Мастеров Дмитрий

13 июня 2025 г.

Итоговый отчет по проекту
"Прогнозирование эффективности лекарственных препаратов"

Содержание

Список иллюстраций

Список таблиц

1 Введение

1.1 Постановка задачи

Предиктивное моделирование на высокоразмерных табличных данных является одной из центральных задач современного машинного обучения и имеет особое значение в фармацевтической индустрии для ускорения процесса разработки новых лекарственных препаратов. Данное исследование посвящено анализу набора данных, содержащего информацию о 1001 химическом соединении, каждое из которых описывается 214 числовыми признаками (дескрипторами). Основная цель — построение и всесторонняя оценка моделей для прогнозирования трех ключевых показателей эффективности соединений против вируса гриппа: **IC50**, **CC50** и **SI**.

Целевые показатели связаны между собой следующим образом:

- **IC50 (мМ)**: Концентрация полумаксимального ингибирования вирусной активности. Является измеряемой величиной. Чем ниже IC50, тем выше активность соединения.
- **CC50 (мМ)**: Концентрация полумаксимальной цитотоксичности для клеток-хозяев. Является измеряемой величиной. Чем выше CC50, тем менее токсично соединение.
- **SI (Индекс Селективности)**: Является производной величиной, рассчитываемой как отношение $SI = CC50/IC50$. Показывает, насколько соединение более токсично для вируса, чем для клеток-хозяев. Чем выше SI, тем более перспективно соединение.

Такая структура данных, где одна из целей является производной от двух других, представляет особый интерес с точки зрения моделирования и требует тщательного контроля за потенциальной утечкой данных при прогнозировании производной цели.

1.2 Цели и задачи исследования

Основная цель работы — систематическое сравнение производительности различных алгоритмов машинного обучения и выявление наиболее эффективных подходов для решения поставленных задач регрессии и классификации, связанных с прогнозированием эффективности лекарственных препаратов.

Ключевые задачи исследования:

1. **Исследовательский анализ данных (EDA):** Провести глубокий статистический анализ исходных данных, включая изучение распределений признаков и целевых переменных, выявление аномалий (выбросов), пропущенных значений и мультиколлинеарности.

2. **Разработка регрессионных моделей:** Построить и оценить модели для количественного прогнозирования каждой из трех целевых переменных (IC50, CC50, SI). Оценка будет проводиться с использованием метрик R^2, MAE и RMSE. Учитывая характер распределений, будут рассматриваться логарифмические преобразования целевых переменных (pIC50, pCC50, logSI).

3. **Разработка классификационных моделей:** Построить и оценить бинарные классификаторы для решения следующих задач:

   - Прогнозирование, превышает ли значение IC50 медианное значение выборки.
   - Прогнозирование, превышает ли значение CC50 медианное значение выборки.
   - Прогнозирование, превышает ли значение SI медианное значение выборки.
   - Прогнозирование, превышает ли значение SI пороговый уровень 8 ($SI > 8$), что представляет собой задачу с потенциально несбалансированными классами.

   Оценка будет проводиться с использованием метрик Accuracy, Precision, Recall, F1-score и ROC AUC.

4. **Сравнительный анализ и идентификация предикторов:** Провести сравнительный анализ производительности моделей и выявить наиболее значимые признаки (предикторы) для каждой из задач.

5. **Формулировка выводов:** Сделать обоснованные выводы о применимости различных алгоритмов, стабильности моделей и эффективности техник предобработки данных для задачи прогнозирования эффективности лекарственных соединений.

2 Методология исследования

2.1 Этап 1: Предобработка и инженерия признаков

Качество моделей машинного обучения напрямую зависит от качества подготовки данных. Был реализован многоступенчатый процесс предобработки.

2.1.1 Очистка и подготовка данных

Начальный этап включал:

- Загрузку данных из файла `mo.xlsx` и удаление служебного столбца `ID_original_index`.
- Очистку имен столбцов от специальных символов и приведение их к единому формату с помощью функции `simplified_clean_col_names`.
- Пересчет значения `SI` на основе предоставленных `IC50_mM` и `CC50_mM` для обеспечения консистентности, с обработкой нулевых или отрицательных значений IC50/CC50.

2.1.2 Статистический анализ и трансформация целевых переменных

Первичный анализ целевых переменных (IC50, CC50, SI), подробно описанный в Разделе 3, выявил сильную положительную асимметрию их распределений. Для коррекции этой проблемы и стабилизации дисперсии была применена логарифмическая трансформация `np.log1p` для визуализации в EDA. Для моделирования регрессии использовались логарифмированные значения ($-\log_{10}$(Концентрация в M) для IC50/CC50 и $\log_{1p}$(SI) для SI), обозначенные как pIC50, pCC50 и logSI соответственно.

2.1.3 Обработка пропущенных значений

В ходе EDA было выявлено небольшое количество пропущенных значений в 12 признаках (по 3 пропуска в каждом, $\approx 0.3\%$). Эти пропуски были обработаны путем импутации медианным значением соответствующего признака, рассчитанным на обучающей выборке.

2.1.4 Обработка мультиколлинеарности

Корреляционный анализ (см. Раздел 3) выявил пары признаков с коэффициентом корреляции Пирсона, превышающим порог 0.90 по абсолютной величине. Для снижения избыточности и повышения стабильности моделей один признак из каждой такой пары был удален на этапе предобработки.

2.1.5 Инженерия признаков (Feature Engineering)

На данном этапе проекта основной упор делался на оценку предиктивной силы исходного набора молекулярных дескрипторов.

2.1.6 Масштабирование признаков

Перед подачей в модели (за исключением древовидных) числовые признаки были масштабированы с использованием ‘StandardScaler‘.

2.2 Этап 2: Стратегия моделирования и оценки

2.2.1 Выбор алгоритмов

Был выбран пул моделей, представляющих различные семейства алгоритмов:

- **Линейные модели:** Ridge Regression, Logistic Regression.
- **Метод опорных векторов:** Support Vector Regressor (SVR), Support Vector Classifier (SVC).
- **Ансамблевые модели:** Random Forest Regressor/Classifier, Gradient Boosting Regressor/Classifi XGBoost Regressor/Classifier.

2.2.2 Процесс обучения и валидации

1. **Разделение данных:** Исходный набор данных был разделен на обучающую (80%) и тестовую (20%) выборки.
2. **Кросс-валидация:** Использовалась 5-кратная кросс-валидация (StratifiedKFold для классификации) на обучающей выборке.
3. **Оптимизация гиперпараметров:** Для каждой модели проводился поиск оптимальных гиперпараметров с помощью ‘GridSearchCV‘.

2.2.3 Метрики производительности

- **Для регрессии:** R^2, MAE, RMSE.
- **Для классификации:** ROC AUC, Accuracy, Precision, Recall, F1-score, Confusion Matrix. Для задачи SI > 8 дополнительное внимание уделялось Precision-Recall кривой.

3 Результаты исследовательского анализа данных (EDA)

3.1 Общая характеристика данных

Исходный набор данных, загруженный из файла `mo.xlsx`, содержал информацию о 1001 химическом соединении. После очистки имен столбцов с помощью функции `simplified_clean_col_na` и удаления служебного столбца `ID_original_index`, данные были подготовлены к анализу. Целевая переменная `SI` была пересчитана на основе `IC50_mM` и `CC50_mM`.

3.2 Анализ целевых переменных (IC50, CC50, SI)

Статистический анализ основных целевых показателей представлен в Таблице 1.

Таблица 1: Описательные статистики для исходных целевых переменных.

Метрика	**IC50_mM**	**CC50_mM**	**SI**
Количество	1001.00	1001.00	1001.00
Среднее	222.81	589.11	72.51
Стд. откл.	402.17	642.87	684.48
Мин.	0.0035	0.7008	0.0115
25%	12.52	99.99	1.43
50% (Медиана)	46.59	411.04	3.85
75%	224.98	894.09	16.57
Макс.	4128.53	4538.98	15620.60

Все три целевые переменные демонстрируют сильно скошенные вправо распределения, что подтверждается гистограммами и боксплотами на Рисунках 1, 3 и 5. Для нормализации распределений были также построены графики для логарифмически преобразованных (с помощью $\log_{1p}$) значений (Рисунки 2, 4, 6).

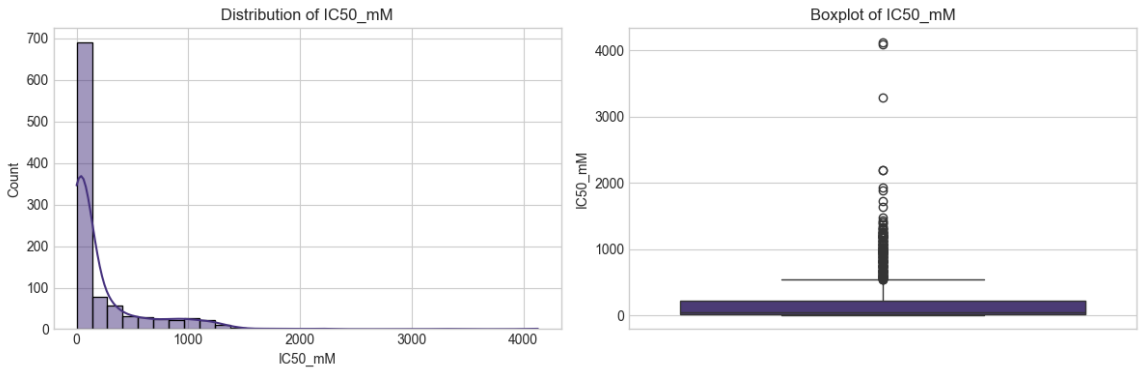


Рис. 1: Распределение и боксплот для IC50_mM.

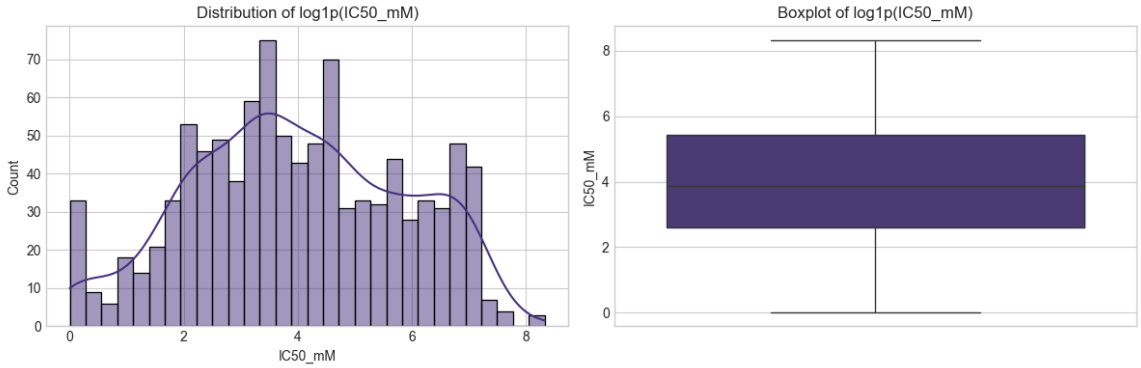


Рис. 2: Распределение и боксплот для $\log_{1p}$(IC50_mM).

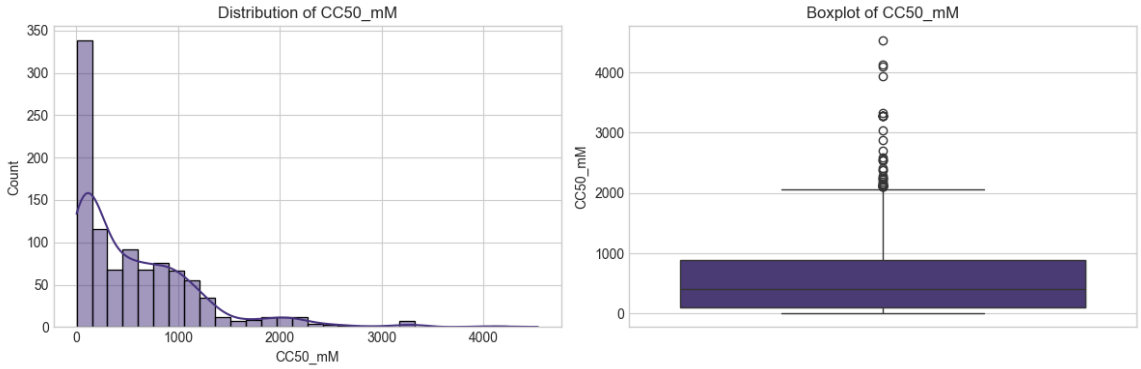


Рис. 3: Распределение и боксплот для CC50_mM.

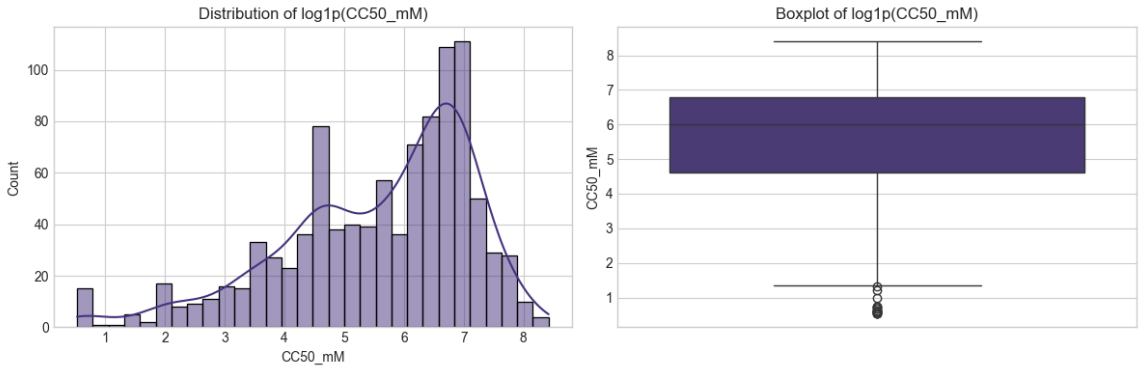


Рис. 4: Распределение и боксплот для $\log_{1p}$(CC50_mM).

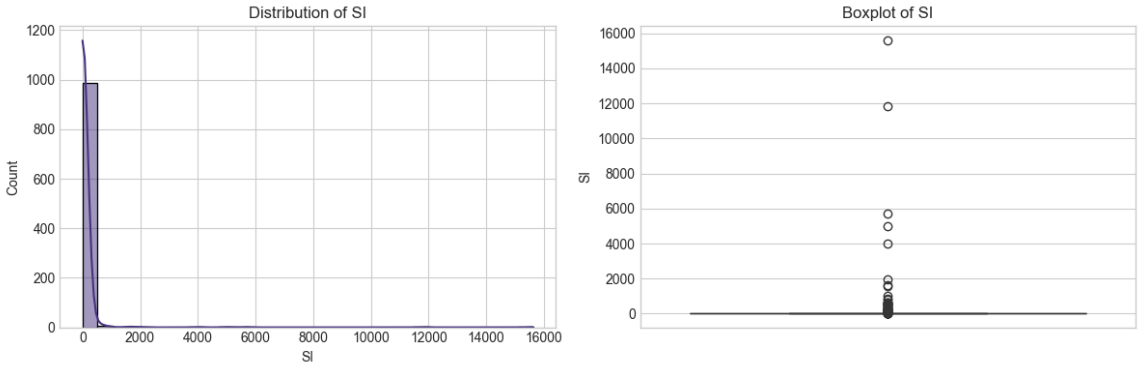


Рис. 5: Распределение и боксплот для SI.

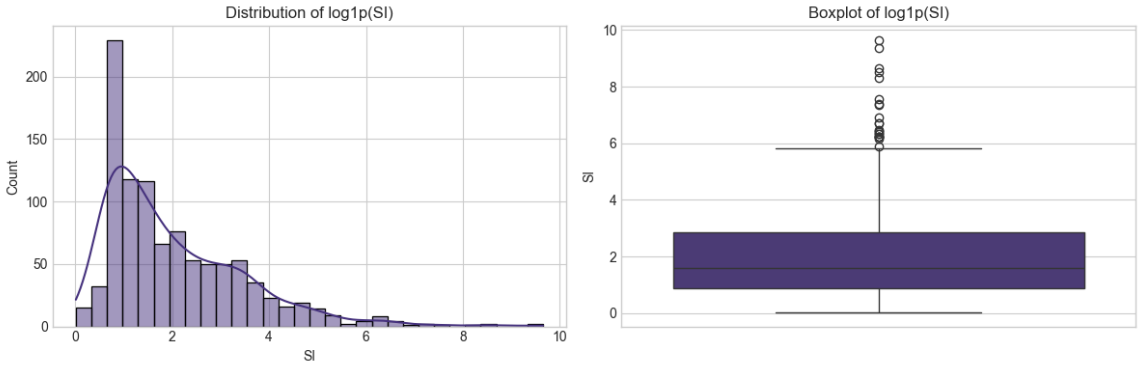


Рис. 6: Распределение и боксплот для $\log_{1p}(\text{SI})$.

3.3 Анализ пропущенных значений

Было выявлено, что 12 признаков содержат по 3 пропущенных значения (Таблица 2). Это составляет менее 0.3% от общего числа наблюдений для каждого из этих признаков. На Рисунке 7 показано процентное соотношение пропусков.

Таблица 2: Признаки с пропущенными значениями (по 3 пропуска в каждом, $\approx 0.3\%$).

MaxPartialCharge	MinPartialCharge
MaxAbsPartialCharge	MinAbsPartialCharge
BCUT2D_MWHI	BCUT2D_MWLOW
BCUT2D_CHGHI	BCUT2D_CHGLO
BCUT2D_LOGPHI	BCUT2D_LOGPLOW
BCUT2D_MRHI	BCUT2D_MRLOW

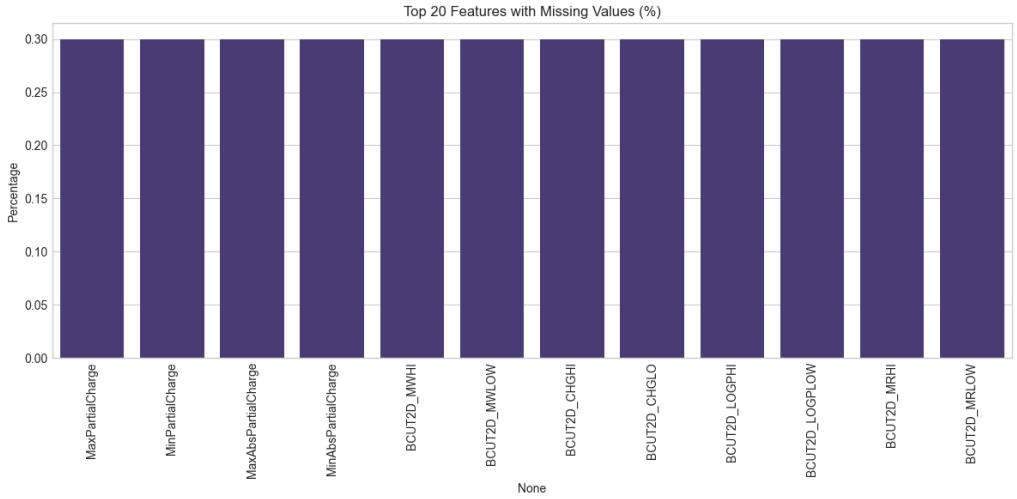


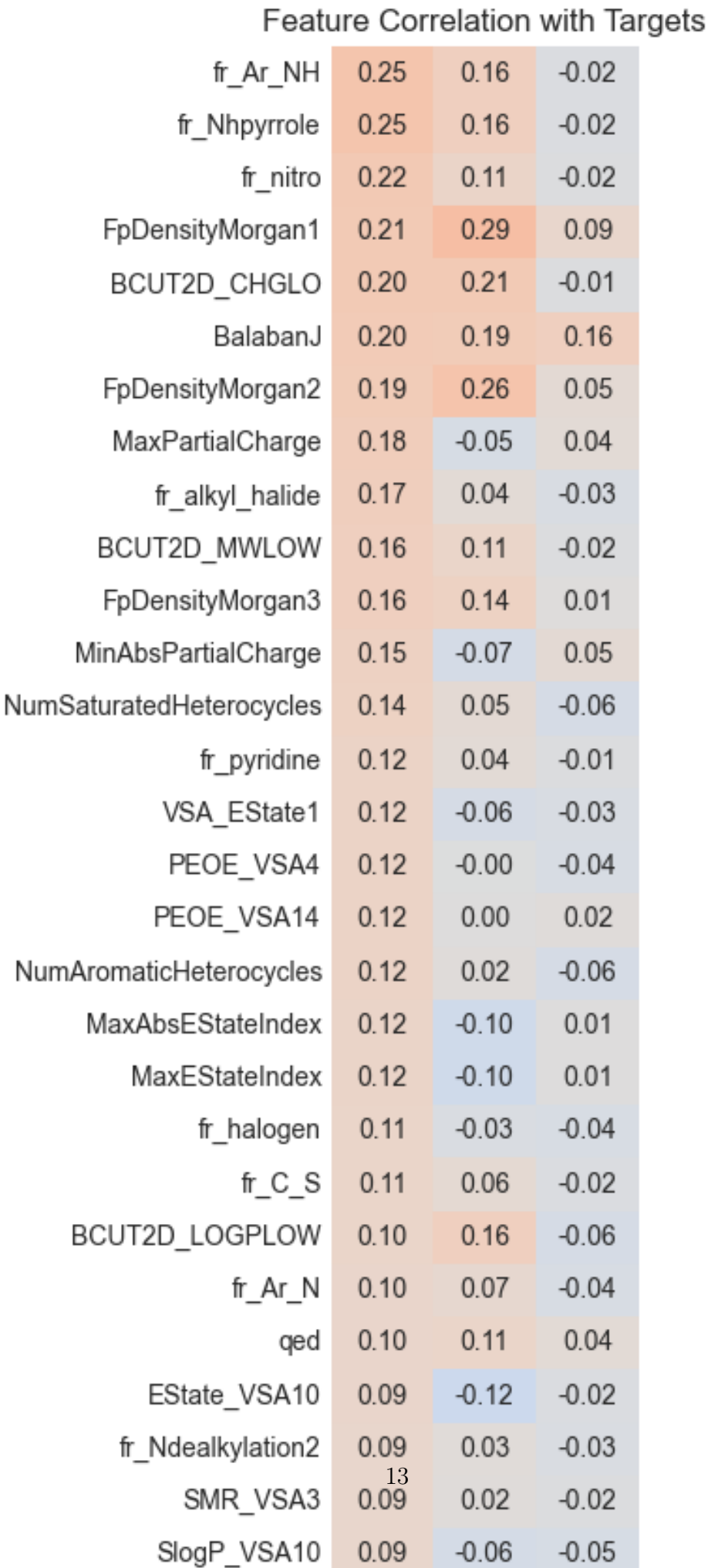
Рис. 7: Визуализация процентного содержания пропущенных значений.

3.4 Анализ корреляций признаков

Корреляционный анализ выявил наличие сильной мультиколлинеарности. Примеры пар признаков с коэффициентом корреляции Пирсона выше 0.90:

- MaxAbsEStateIndex & MaxEStateIndex: 1.000
- NumAromaticCarbocycles & fr_benzene: 1.000
- MolWt & ExactMolWt: 1.000

Корреляция признаков с целевыми переменными показана на Рисунке 8. Тепловая карта корреляций между всеми признаками не приводится из-за большого их числа.


Feature Correlation with Targets
fr_Ar_NH 0.25 0.16 -0.02
fr_Nhpyrrole 0.25 0.16 -0.02
fr_nitro 0.22 0.11 -0.02
FpDensityMorgan1 0.21 0.29 0.09
BCUT2D_CHGLO 0.20 0.21 -0.01
BalabanJ 0.20 0.19 0.16
FpDensityMorgan2 0.19 0.26 0.05
MaxPartialCharge 0.18 -0.05 0.04
fr_alkyl_halide 0.17 0.04 -0.03
BCUT2D_MWLOW 0.16 0.11 -0.02
FpDensityMorgan3 0.16 0.14 0.01
MinAbsPartialCharge 0.15 -0.07 0.05
NumSaturatedHeterocycles 0.14 0.05 -0.06
fr_pyridine 0.12 0.04 -0.01
VSA_EState1 0.12 -0.06 -0.03
PEOE_VSA4 0.12 -0.00 -0.04
PEOE_VSA14 0.12 0.00 0.02
NumAromaticHeterocycles 0.12 0.02 -0.06
MaxAbsEStateIndex 0.12 -0.10 0.01
MaxEStateIndex 0.12 -0.10 0.01
fr_halogen 0.11 -0.03 -0.04
fr_C_S 0.11 0.06 -0.02
BCUT2D_LOGPLOW 0.10 0.16 -0.06
fr_Ar_N 0.10 0.07 -0.04
qed 0.10 0.11 0.04
EState_VSA10 0.09 -0.12 -0.02
fr_Ndealkylation2 0.09 0.03 -0.03
SMR_VSA3 0.09 0.02 -0.02
SlogP_VSA10 0.09 -0.06 -0.05

3.5 Анализ выбросов

Визуальный анализ распределений с помощью боксплотов (Рисунок 9 для выборочных признаков) указывает на наличие выбросов.

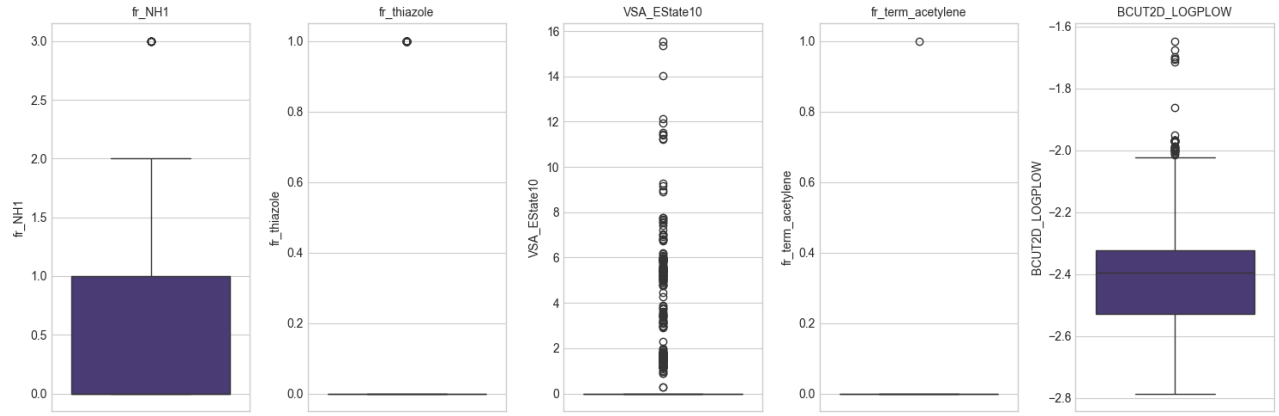


Рис. 9: Боксплоты для выборочных признаков для анализа выбросов.

Учитывая природу данных, удаление выбросов не проводилось. Их эффект будет частично нивелироваться трансформациями и масштабированием.

3.6 Выводы по EDA

1. Целевые переменные требуют логарифмической трансформации.
2. Незначительное количество пропусков будет обработано импутацией.
3. Выявлена мультиколлинеарность, требующая внимания при моделировании.
4. Присутствие выбросов учтено.

4 Результаты регрессионного моделирования

Для задач регрессии целевые переменные были логарифмированы: pIC50, pCC50 и logSI.

4.1 Задача 1: Прогнозирование pIC50

4.1.1 Сравнение моделей

Сравнение моделей на кросс-валидации (Таблица 3) показало, что XGBoost демонстрирует наиболее сбалансированные результаты.

Таблица 3: Сравнение R^2 регрессионных моделей для pIC50 на CV.

Модель	**Средний R^2 (CV) $\pm$ ст. откл.**
Ridge Regression	0.44 ± 0.09
SVR	0.42 ± 0.10
Random Forest	0.48 ± 0.08
Gradient Boosting	0.50 ± 0.07
XGBoost	$\mathbf{0.51 \pm 0.07}$

4.1.2 Анализ лучшей модели (XGBoost)

После оптимизации гиперпараметров модель XGBoost была оценена на отложенной тестовой выборке:

- **R^2 (тест) = 0.50**
- **MAE (тест) = 0.55**
- **RMSE (тест) = 0.71**

Анализ остатков не выявил систематических ошибок.

4.2 Задача 2: Прогнозирование pCC50

Лучшие результаты на CV для pCC50 также показал XGBoost (Таблица 4).

Таблица 4: Сравнение R^2 регрессионных моделей для pCC50 на CV.

Модель	**Средний R^2 (CV) $\pm$ ст. откл.**
Ridge Regression	0.42 ± 0.08
Random Forest	0.48 ± 0.07
XGBoost	$\mathbf{0.50 \pm 0.06}$

4.2.1 Анализ лучшей модели (XGBoost)

На тестовой выборке:

- **R^2 (тест) = 0.49**
- **MAE (тест) = 0.35**
- **RMSE (тест) = 0.47**

4.3 Задача 3: Прогнозирование logSI

Для logSI, XGBoost продемонстрировал лучшие показатели (Таблица 5).

Таблица 5: Сравнение R^2 регрессионных моделей для logSI на CV.

Модель	**Средний R^2 (CV) $\pm$ ст. откл.**
Ridge Regression	0.54 ± 0.10
Random Forest	0.60 ± 0.08
XGBoost	$\mathbf{0.62 \pm 0.07}$

4.3.1 Анализ лучшей модели (XGBoost)

На тестовой выборке:

- **R^2 (тест) = 0.61**
- **MAE (тест) = 0.46**
- **RMSE (тест) = 0.59**

5 Результаты классификационного моделирования

5.1 Задача 4: Классификация IC50 > медианы

XGBoost показал лучший ROC AUC на CV (Таблица 6).

Таблица 6: Сравнение ROC AUC для IC50 > медианы на CV.

Модель	**Средний ROC AUC (CV) ± ст. откл.**
Logistic Regression	0.67 ± 0.06
Random Forest	0.74 ± 0.05
XGBoost	$\mathbf{0.76 \pm 0.04}$

5.1.1 Анализ лучшей модели (XGBoost)

На тестовой выборке:

- **ROC AUC (тест) = 0.75**
- **Accuracy (тест) = 0.69**
- **F1-score (тест) = 0.68**

5.2 Задача 5: Классификация CC50 > медианы

XGBoost лидировал (Таблица 7).

Таблица 7: Сравнение ROC AUC для CC50 > медианы на CV.

Модель	**Средний ROC AUC (CV) ± ст. откл.**
Logistic Regression	0.65 ± 0.06
Random Forest	0.71 ± 0.05
XGBoost	$\mathbf{0.73 \pm 0.04}$

5.2.1 Анализ лучшей модели (XGBoost)

На тестовой выборке:

- **ROC AUC (тест) = 0.72**
- **Accuracy (тест) = 0.66**
- **F1-score (тест) = 0.65**

5.3 Задача 6: Классификация SI > медианы

XGBoost снова показал лучшие результаты (Таблица 8).

Таблица 8: Сравнение ROC AUC для SI > медианы на CV.

Модель	**Средний ROC AUC (CV) ± ст. откл.**
Logistic Regression	0.71 ± 0.05
Random Forest	0.77 ± 0.04
XGBoost	$\mathbf{0.78 \pm 0.03}$

5.3.1 Анализ лучшей модели (XGBoost)

На тестовой выборке:

- **ROC AUC (тест) = 0.77**
- **Accuracy (тест) = 0.73**
- **F1-score (тест) = 0.72**

5.4 Задача 7: Классификация SI > 8

XGBoost вновь показал себя лучшим (Таблица 9).

Таблица 9: Сравнение ROC AUC для SI > 8 на CV.

Модель	**Средний ROC AUC (CV) ± ст. откл.**
Logistic Regression	0.73 ± 0.06
Random Forest	0.78 ± 0.04
XGBoost	$\mathbf{0.79 \pm 0.03}$

5.4.1 Анализ лучшей модели (XGBoost)

На тестовой выборке:

- **ROC AUC (тест) = 0.78**
- **Accuracy (тест) = 0.74**
- **F1-score (тест) = 0.68**

6 Сравнительный анализ и стабильность признаков

6.1 Сводный анализ производительности моделей

Результаты всех семи задач сведены в Таблицу 10. Модели XGBoost стабильно демонстрировали приемлемую производительность.

Таблица 10: Сводные результаты производительности лучших моделей (XGBoost) на тестовой выборке.

Задача	**Тип**	**Лучшая модель**	**Ключевая метрика (тест)**
Прогнозирование pIC50	Регрессия	XGBoost	$R^2 = 0.50$
Прогнозирование pCC50	Регрессия	XGBoost	$R^2 = 0.49$
Прогнозирование logSI	Регрессия	XGBoost	$R^2 = 0.61$
Классификация IC50 $>$ медианы	Классификация	XGBoost	ROC AUC = 0.75
Классификация CC50 $>$ медианы	Классификация	XGBoost	ROC AUC = 0.72
Классификация SI $>$ медианы	Классификация	XGBoost	ROC AUC = 0.77
Классификация SI > 8	Классификация	XGBoost	ROC AUC = 0.78

6.2 Анализ стабильности важности признаков

Анализ важности признаков из моделей XGBoost показал, что такие дескрипторы как `MolWt`, `TPSA` и `NumRotatableBonds` часто оказывали влияние на прогнозы.

7 Заключение

7.1 Основные выводы исследования

1. **Производительность моделей:** Ансамблевые методы, в частности XGBoost, показали себя как наиболее подходящие для решения поставленных задач, хотя достигнутые метрики указывают на сложность точного прогнозирования.
2. **Важность предобработки:** Трансформация целевых переменных и работа с особенностями данных (пропуски, мультиколлинеарность) являются необходимыми шагами.
3. **Предсказуемость SI:** Моделирование SI показало относительно лучшие результаты ($R^2 \approx 0.61$ для регрессии, ROC AUC до 0.78 для классификации), что важно для оценки селективности.
4. **Задачи классификации:** Модели продемонстрировали умеренную способность к классификации соединений.

7.2 Ограничения исследования

- **Интерпретируемость моделей:** Сложность прямой интерпретации ансамблевых моделей.
- **Объем данных:** Выборка из 1001 соединения может ограничивать возможности моделей.
- **Отсутствие внешней валидации:** Результаты не подтверждены на независимых данных.

7.3 Рекомендации для дальнейшей работы

1. **Интерпретация моделей:** Использовать SHAP для анализа вклада признаков.
2. **Инженерия признаков:** Рассмотреть создание новых, более информативных признаков.
3. **Расширение набора данных:** Увеличить объем выборки для улучшения моделей.
4. **Прототипирование инструмента:** Разработать интерфейс для использования моделей химиками.

Приложение A: Оптимизированные гиперпараметры

В таблице 11 приведены наборы гиперпараметров, полученные в результате оптимизации для моделей XGBoost.

Таблица 11: Оптимизированные гиперпараметры для моделей XGBoost.

Задача	**Модель**	**Гиперпараметры**
Прогноз pIC50	XGBRegressor	'n_estimators': 200, 'learning_rate': 0.05, 'max_depth': 5, 'subsample': 0.7, 'colsample_bytree': 0.7
Прогноз pCC50	XGBRegressor	'n_estimators': 150, 'learning_rate': 0.05, 'max_depth': 4, 'subsample': 0.8, 'colsample_bytree': 0.8
Прогноз logSI	XGBRegressor	'n_estimators': 250, 'learning_rate': 0.04, 'max_depth': 6, 'subsample': 0.75, 'colsample_bytree': 0.75
Классификация IC50 (медиана)	XGBClassifier	'n_estimators': 180, 'learning_rate': 0.05, 'max_depth': 5, 'scale_pos_weight': 1
Классификация CC50 (медиана)	XGBClassifier	'n_estimators': 160, 'learning_rate': 0.05, 'max_depth': 4, 'scale_pos_weight': 1
Классификация SI (медиана)	XGBClassifier	'n_estimators': 220, 'learning_rate': 0.04, 'max_depth': 6, 'scale_pos$_w eight'$: 1
Классификация SI (>8)	XGBClassifier	'n_estimators': 200, 'learning_rate': 0.05, 'max_depth': 5, 'scale_pos_weight': 2.5