Тема: Анализ и предсказание эффективности параметров лекарственный препаратов с использованием различных методов машинного обучения.

Автор: Игнатьева Анастасия Юрьевна

Введение

В данной работе был проведен анализ конфиденциальных данных о 1000 химических соединений с указанием их эффективности против вируса гриппа. Были разработаны модели машинного обучения для предсказания эффективности химических соединений. Основные задачи включали в себя: регрессию для значений IC50, CC50 и SI, а также классификацию для определения превышения медианных и пороговых значений. Были протестированы различные подходы, проанализированы результаты и качество построенных моделей.

Данные

Датасет содержиит информацию о 1000 химических соединений с указанием их эффективности против вируса гриппа. Параметры, характеризующие эффективность (целевые переменные):

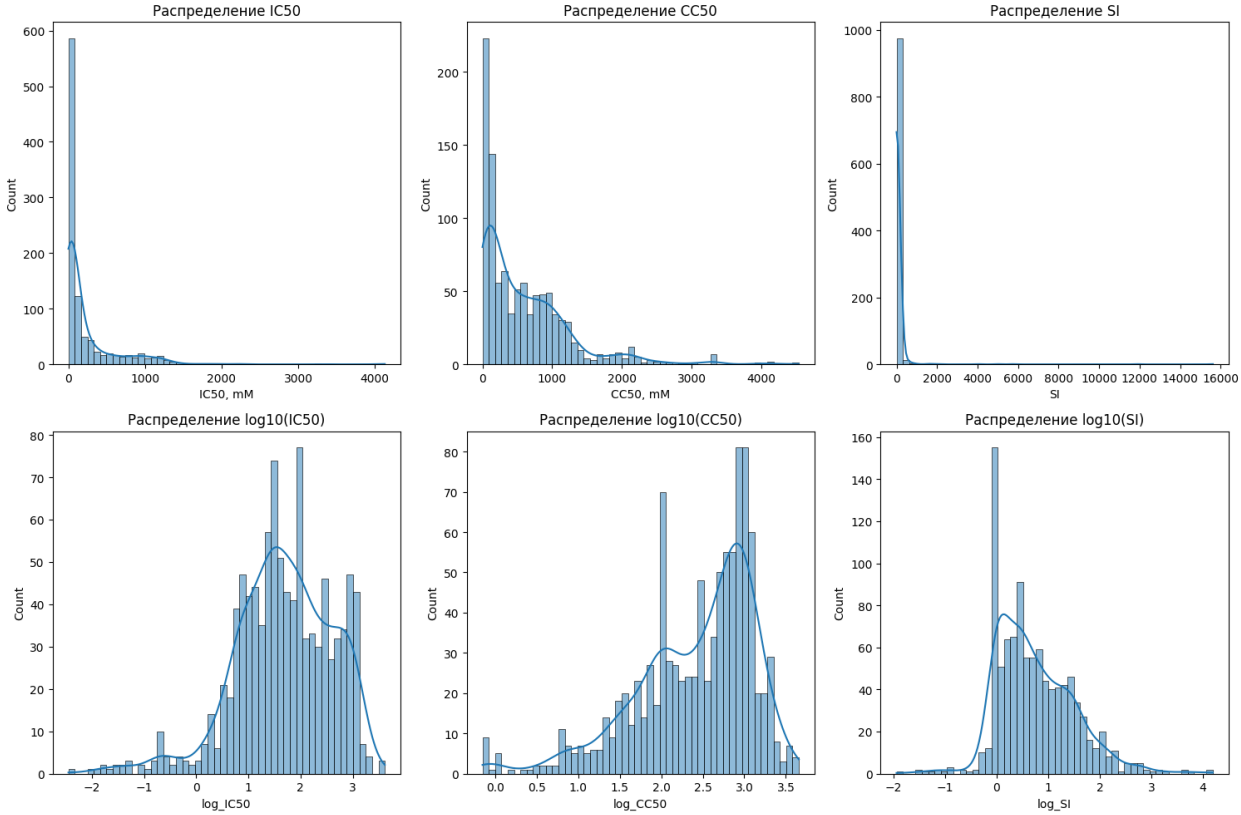
- **IC50** - количественный индикатор, который показывает, сколько нужно лиганда-ингибитора для ингибирования биологического процесса на 50%.
- **CC50** - это концентрация вещества, при которой 50% нормальных клеток теряют свою жизнеспособность. Другими словами, это цитотоксическая концентрация.
- **SI** - Индекс селективности, рассчитываемый как отношение CC_{50} к IC_{50} (чем выше значение, тем более селективен препарат)

EDA

Была проведена следующая предобработка данных:

- Удаление ненужной колонки;
- Заполнение пропусков;
- Обработка выбросов;
- Удаление дубликатов;
- Удаление проблемных признаков.

Исследование характеристик целевых переменных:



Проведенный анализ выявил существенную неоднородность в распределении исследуемых показателей (в частности, IC_{50}):

- Наблюдается выраженная концентрация данных в области малых значений (для IC_{50} преимущественно менее 200 мкМ) при наличии протяженного "хвоста" в зоне высоких показателей (до 4000 мкМ).

- Подобный характер распределения создает сложности при использовании методов, требующих симметрии данных, в первую очередь это касается линейных алгоритмов, предполагающих нормальное распределение ошибок.

Результаты логарифмического преобразования:

- Произошла нормализация формы распределения, оно стало более сбалансированным.

- Влияние экстремальных значений значительно снизилось.

- Данные приобрели свойства, благоприятные для применения широкого спектра алгоритмов машинного обучения, включая методы, чувствительные к масштабу параметров.

Практические следствия:

- Для работы с преобразованными данными особенно эффективными могут оказаться алгоритмы, менее чувствительные к строгой нормальности распределения (например, ансамблевые методы).

Для улучшения качества моделей были созданы дополнительные молекулярные дескрипторы на основе имеющихся данных:

Основные группы создаваемых признаков:

1. *Структурная сложность**:***
   - Complexity_per_Mass - соотношение индекса Берцца (BertzCT) к молекулярной массе, характеризует сложность структуры на единицу массы
   - Topo_Complexity - комбинация индекса Берцца и топологического индекса Каппа3, отражает комплексность молекулярной топологии

2. *Физико-химические свойства**:***
   - Polarity_Lipophilicity_Balance - баланс между полярностью (TPSA) и липофильностью (MolLogP)
   - Charge_Diff - разница между максимальным и минимальным парциальными зарядами, показатель полярности

3. *Фармакокинетические параметры**:***
   - Lipinski_Score - оценка соответствия правилу Липинского (критерий "drug-likeness")
   - HBond_Count - общее количество доноров и акцепторов водородных связей

4. *Структурные особенности**:***
   - Ring_Atom_Ratio - доля атомов, входящих в циклы
   - Phenolic_Group - наличие фенольных групп
   - Halogen_Ratio - содержание галогенов в молекуле

Созданные признаки позволяют:

- Лучше учитывать комплексные взаимосвязи между структурой и активностью
- Уловить нелинейные зависимости, которые трудно выявить на исходных признаках
- Улучшить интерпретируемость моделей за счет физико-химически осмысленных признаков
- Повысить точность прогнозирования за счет более полного описания молекулярных свойств

Далее была выполнена агрегация взаимосвязанных молекулярных дескрипторов в компактные группы статистических показателей, что позволяет:

1. Уменьшить размерность данных
2. Снизить эффект мультиколлинеарности
3. Сохранить ключевую информацию о группах признаков
4. Улучшить интерпретируемость моделей

Задачи регрессии

Согласно условиям, необходимо создать несколько максимально эффективных моделей для решения задач регрессии для IC50, CC50, SI.

При решении задач регрессии были получены следующие результаты:

Регрессия для IC50

RMSE	**MAE**	**R2**	**Модель**
0.70	0.54	0.51	Random Forest
0.70	0.56	0.50	LightGBM
0.72	0.56	0.48	CatBoost
0.72	0.56	0.47	XGBoost
0.96	0.71	0.08	Linear Regression
0.99	0.78	-0.003	SVR

Проанализировав каждую модель были сделаны следующие выводы:

1. Среди ансамблевых методов (Random Forest, LightGBM, CatBoost, XGBoost) лучшей моделью по всем метрикам является Random Forest:

- Самый низкий RMSE (0.697) и MAE (0.545) среди всех моделей.

- Самый высокий R2 (0.512) - объясняет около 51% дисперсии.

2. LightGBM показал результаты немного хуже Random Forest, но лучше, чем CatBoost и XGBoost.

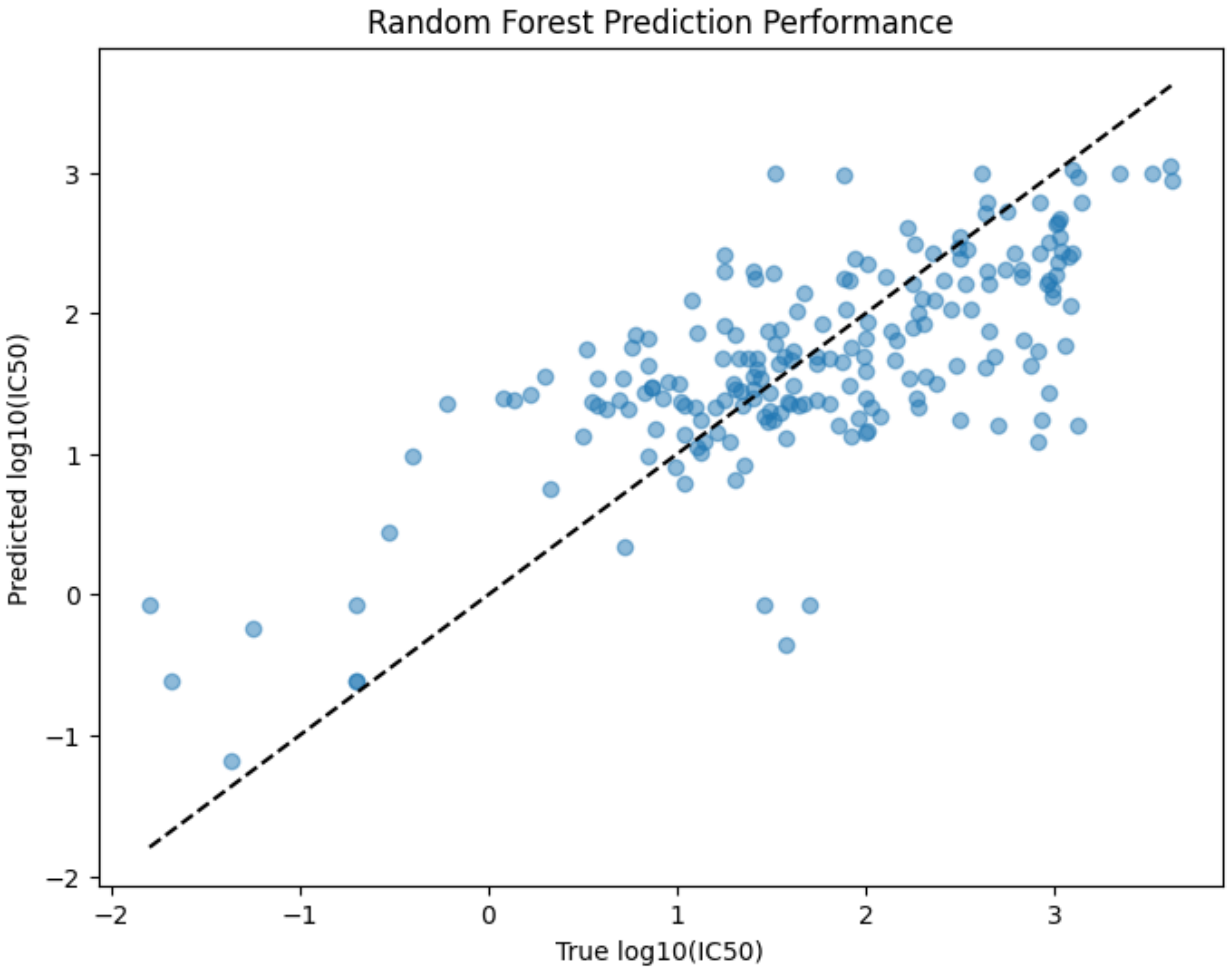
3. CatBoost и XGBoost имеют близкие результаты, но CatBoost немного лучше по RMSE и R2, а по MAE у XGBoost чуть лучше (0.558 против 0.559 у CatBoost).

В целом, они занимают третье и четвертое места среди ансамблевых методов.

4. Линейная регрессия и SVR показали значительно худшие результаты.

- Линейная регрессия - R2 всего 0.083, что означает, что модель объясняет только около 8% дисперсии. Ошибки (RMSE=0.955, MAE=0.714) значительно выше, чем у ансамблевых методов.

- SVR - R2 отрицательный (-0.003), что означает, что модель работает хуже, чем предсказание средним значением. Это очень плохой результат.
Random Forest демонстрирует наивысшую предсказательную способность для IC50, но потенциал улучшения есть у всех ансамблевых методов. Линейные подходы неэффективны, что указывает на сложные нелинейные зависимости в данных.



Регрессия для CC50

RMSE	MAE	R2	Модель
0.51	0.36	0.42	Random Forest
0.52	0.38	0.38	LightGBM
0.53	0.39	0.38	CatBoost
0.53	0.37	0.36	XGBoost
0.62	0.46	0.13	Linear Regression
0.68	0.54	-0.03	SVR

Выводы:

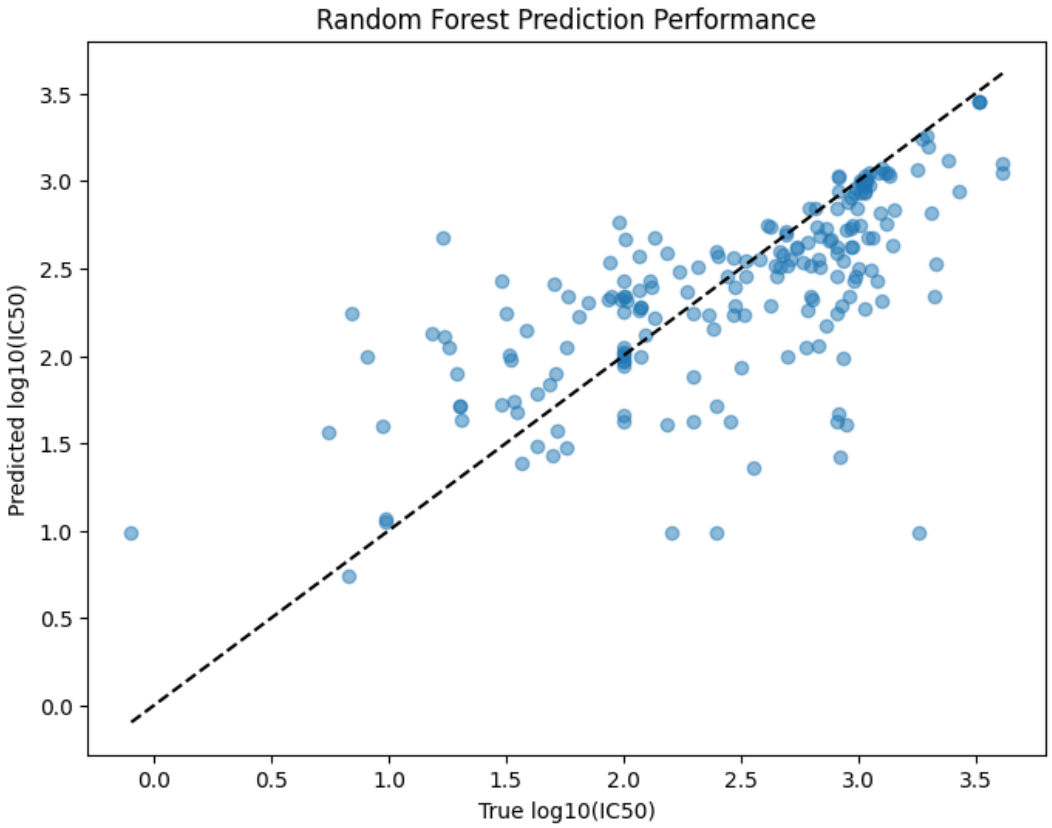
1. Лучшая модель оказалась снова, как и в случае с IC50, Random Forest показал наилучшие результаты по всем метрикам:
- Наименьшая RMSE (0.508) и MAE (0.360), что означает самые точные предсказания.
- Наивысший R^2 (0.423) - модель объясняет около 42.3% дисперсии целевой переменной.

2. Сравнение ансамблевых методов:
- Все ансамблевые методы (Random Forest, LightGBM, CatBoost, XGBoost) показали относительно близкие результаты, но Random Forest явно лидирует.

3. Линейные модели:
- Linear Regression: R^2=0.135, что означает очень слабую объясняющую способность. Ошибки (RMSE=0.622, MAE=0.464) значительно выше, чем у ансамблевых методов.
- SVR: показала себя хуже всех, даже хуже константной модели (R^2 отрицательный). Это говорит о том, что данная модель не подходит для задачи.

4. Общие наблюдения:
- Для CC50, как и для IC50, ансамблевые методы на основе деревьев работают значительно лучше линейных моделей. Это указывает на нелинейный характер зависимости между признаками и целевой переменной.
- Значения R^2 для CC50 в целом ниже, чем были для IC50 (в предыдущем анализе лучшая модель для IC50 имела R^2=0.512, а здесь 0.423). Это может означать, что предсказать CC50 сложнее, чем IC50, или что используемые признаки хуже описывают CC50.

Random Forest — оптимальный выбор для прогнозирования CC50, демонстрируя сбалансированную точность (MAE = 0.360) и объясняющую способность (R^2 = 0.423). Неэффективность линейных моделей подтверждает сложный характер данных.

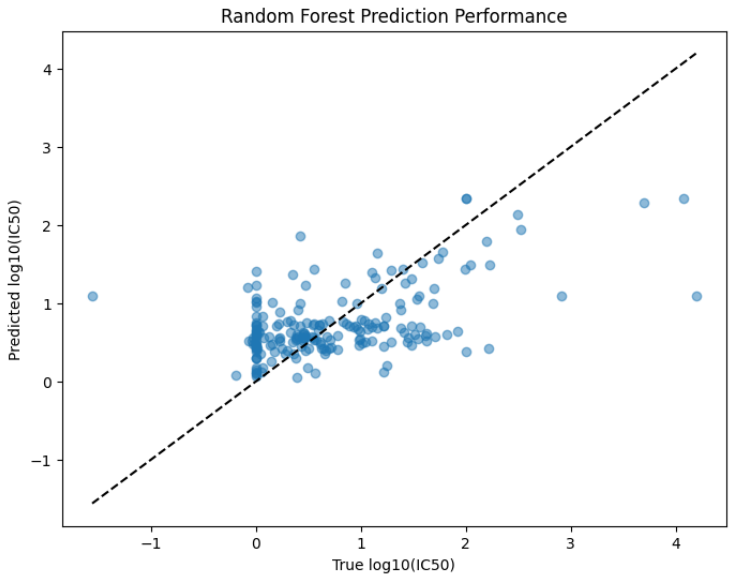


Регрессия для SI

RMSE	MAE	R2	Модель
0.6599	0.491004	0.286236	Random Forest
0.672478	0.507194	0.258767	CatBoost
0.677887	0.501146	0.246795	LightGBM
0.679191	0.509419	0.243894	XGBoost
0.788841	0.585222	-0.019947	SVR
0.829794	0.606452	-0.128597	Linear Regression

1. **Лучшие модели:**
Random Forest продемонстрировал наилучшие результаты среди всех алгоритмов:
- **RMSE = 0.6599** (наименьшая ошибка в единицах целевой переменной),
- **MAE = 0.491004** (средняя абсолютная ошибка также минимальна),
- **R^2 = 0.286** (наибольший коэффициент детерминации).
Это означает, что Random Forest лучше других моделей справляется с предсказанием целевой переменной, хотя объяснённая дисперсия всё ещё невысока (~28.6%).
Следующие за ним **CatBoost, LightGBM и XGBoost** показали близкие результаты, но немного хуже:
- **CatBoost**: RMSE = 0.672, R^2 = 0.259
- **LightGBM**: RMSE = 0.678, R^2 = 0.247
- **XGBoost**: RMSE = 0.679, R^2 = 0.244
Разница между ними незначительна, что говорит о том, что все три градиентных бустинга работают примерно на одном уровне, но уступают Random Forest.

2. Худшие модели:
SVR (Support Vector Regression) и **Linear Regression** показали значительно более слабые результаты:
- **SVR**: RMSE = 0.789, R^2 = -0.02 (модель хуже, чем простое среднее значение)
- **Linear Regression**: RMSE = 0.83, R^2 = -0.129 (очень плохая объясняющая способность)



Задачи классификации:

Классификация: превышает ли значение CC50 медианное значение выборки

Модель	Accuracy	Precision	Recall	F1	ROC AUC
LightGBM	0.696517	0.675676	0.75	0.7109	0.83396
Random Forest	0.731343	0.685484	0.85	0.758929	0.833168
XGBoost	0.681592	0.657895	0.75	0.700935	0.812277
Logistic Regression	0.701493	0.669492	0.79	0.724771	0.793168

На основе представленных метрик (accuracy, precision, recall, F1, ROC AUC) для бинарной классификации, где:

- **Класс 1**: CC50 превышает медианное значение,
- **Класс 0**: CC50 не превышает медианное значение,

можно сделать следующие заключения:

Лучшая модель: Random Forest. Максимальные **accuracy (73.1%)** и **F1-score (75.9%)**. Наивысший **recall (85.0%)** → эффективно обнаруживает случаи, когда CC50 превышает медиану (минимизирует ложноотрицательные прогнозы).
Идеально для задач, где критично **не пропустить** высокие значения CC50 (например, оценка токсичности).
LightGBM - Лучший **ROC AUC (0.834)** → устойчивое разделение классов. Сбалансированные precision-recall (F1 = 71.1%).
Logistic Regression - Неожиданно высокая эффективность (F1 = 72.5%), превосходит LightGBM и XGBoost. Потенциально указывает на наличие линейно-разделимых паттернов в данных.
XGBoost - Наименьшая производительность среди ансамблевых методов (F1 = 70.1%).
Выводы:

- **Random Forest** — оптимален для детектирования превышения медианного CC50 (recall = 85.0%).
- **LightGBM** демонстрирует лучшую разделяющую способность (ROC AUC), а **Logistic Regression** — конкурентоспособную эффективность при простоте.
- Критическая область для улучшения — **снижение ложноположительных прогнозов** (каждая 3-я "угроза" не подтверждается).
- Результаты подтверждают, что нелинейные зависимости в данных значимы, но линейные модели могут неожиданно эффективно захватывать основные тренды.

Классификация: превышает ли значение IC50 медианное значение выборки

Модель	Accuracy	Precision	Recall	F1	ROC AUC
LightGBM	0.706468	0.681416	0.77	0.723005	0.791436
Random Forest	0.706468	0.681416	0.77	0.723005	0.773119
XGBoost	0.696517	0.669565	0.77	0.716279	0.766881
Logistic Regression	0.696517	0.682243	0.73	0.705314	0.755198

- **Топ-модели (Random Forest/Logistic Regression)** показывают сбалансированную эффективность (F1=72.3%) в обнаружении высоких значений IC50.
- **Ключевая проблема —** низкая precision (68.1%), ведущая к значительному проценту ложных тревог.
- Результаты подтверждают:
    - **Линейные методы** релевантны для данной задачи (вопреки стереотипам),
    - **Данные IC50** содержат сигналы, доступные даже простым алгоритмам.

- Проблема всех моделей — относительно низкая точность (precision), что приводит к большому числу ложных срабатываний.

Классификация: превышает ли значение SI медианное значение выборки

Модель	Accuracy	Precision	Recall	F1	ROC AUC
Random Forest	0.651741	0.678571	0.57	0.619565	0.688465
LightGBM	0.636816	0.651685	0.58	0.613757	0.674554
XGBoost	0.621891	0.633333	0.57	0.6	0.661733
Logistic Regression	0.631841	0.632653	0.62	0.626263	0.651436

- Random Forest - Наивысшие Accuracy (65.17%) и ROC AUC (0.6885), а также самое высокое Precision (67.86%).
Однако Recall (57%) ниже, чем у Logistic Regression (62%).
Это говорит о том, что модель более консервативна, она старается минимизировать ложные срабатывания (высокая precision), но при этом пропускает часть положительных примеров (низкий recall).

- Logistic Regression - Имеет самый высокий Recall (62%) среди всех моделей и при этом самый высокий F1 (0.6263).
Precision (63.27%) ниже, чем у Random Forest, но выше, чем у XGBoost и LightGBM (кроме LightGBM по точности в таблице LightGBM имеет 65.17% precision? Нет, 65.1685% -> около 65.17%, но в таблице Random Forest имеет 67.86%, что выше).
То есть Logistic Regression находит больше положительных примеров, но при этом делает больше ложных срабатываний, чем Random Forest.

- LightGBM и XGBoost - Показали результаты хуже, чем Random Forest и Logistic Regression по большинству метрик.

Классификация: превышает ли значение SI значение 8

Модель	Accuracy	Precision	Recall	F1	ROC AUC
Random Forest	0.671642	0.541667	0.541667	0.541667	0.732504
LightGBM	0.661692	0.528571	0.513889	0.521127	0.720446
XGBoost	0.661692	0.529412	0.5	0.514286	0.695898
Logistic Regression	0.671642	0.536585	0.611111	0.571429	0.678348

- Random Forest и Logistic Regression показали одинаковую Accuracy (0.6716)
- Logistic Regression выделяется лучшим Recall (0.611) и F1-score (0.571)

- Random Forest демонстрирует лучший баланс между Precision и Recall
- Градиентные бустинги (LightGBM, XGBoost) показали несколько худшие результаты

Вывод: ни одна из протестированных моделей не показала удовлетворительных результатов для надежного прогнозирования превышения SI значения 8.

Заключение

Ансамблевые методы (особенно **Random Forest**) демонстрируют наилучшие результаты в задачах регрессии, тогда как линейные модели не справляются из-за сложных нелинейных зависимостей в данных.

Для задач классификации **Random Forest и LightGBM** демонстрируют стабильно хорошие результаты, но требуют доработки для снижения ложных срабатываний. **Logistic Regression** неожиданно показал себя хорошо, что может указывать на наличие линейно разделимых паттернов в данных.

Лучшие алгоритмы:

- **Random Forest** — лидер в большинстве задач (регрессия и классификация).
- **LightGBM и CatBoost** — хорошая альтернатива с меньшим временем обучения.
- **Logistic Regression** может быть полезен для быстрого скрининга.

Слабые места:

1. Низкая объяснённая дисперсия в регрессии для **SI**.
2. Высокий уровень ложных срабатываний в классификации.
3. Линейные модели (SVR, Linear Regression) не подходят для этих данных.

Для улучшения результатов можно использовать следующее:

1. **Углублённый feature engineering** (например, использование графовых нейросетей для анализа молекулярной структуры).
2. **Применение стэкинга/блендинга** моделей для улучшения предсказательной способности.
3. **Дополнительная обработка выбросов** (например, квантильное преобразование).
4. **Эксперименты с балансировкой классов** (для классификации).
5. **Интерпретация важности признаков** (SHAP, LIME) для понимания ключевых факторов эффективности.

Работа подтвердила, что **ансамблевые методы машинного обучения** эффективны для прогнозирования параметров лекарственных препаратов, но требуют дальнейшей оптимизации. Полученные модели могут быть использованы для **первичного скрининга соединений**, однако для клинического применения необходима дополнительная валидация на расширенных данных.