Курсовой отчёт

ПРОГНОЗИРОВАНИЕ ЭФФЕКТИВНОСТИ ЛЕКАРСТВЕННЫХ СОЕДИНЕНИЙ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Натан Кловер | июнь 2025

Введение

На основании предоставленных химических данных стояла задача построить прогноз, позволяющий определить эффективность веществ для создания новых лекарственных препаратов. Ключевая цель — выявить связи между химическими дескрипторами и показателями активности веществ, такими как IC50, CC50 и SI, а также уметь классифицировать соединения как потенциально эффективные или неэффективные.

Мы последовательно решили 7 задач:

- Регрессия: предсказание IC50
- Регрессия: предсказание CC50
- Регрессия: предсказание SI
- Классификация: IC50 > медианы
- Классификация: CC50 > медианы
- Классификация: SI > медианы
- Классификация: SI > 8 (важный медицинский порог)

Для каждой задачи были обучены и сравнены несколько моделей. Это позволило выявить закономерности, сильные и слабые стороны разных подходов, а также предложить направления для улучшения.

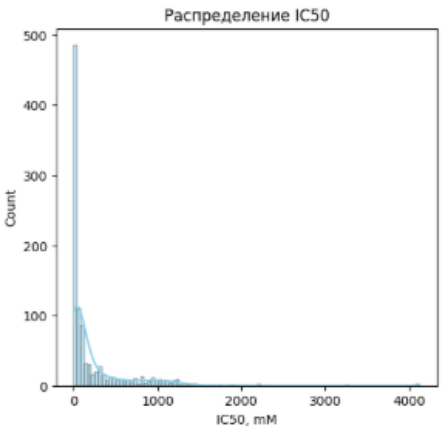
Разведочный анализ данных (EDA)

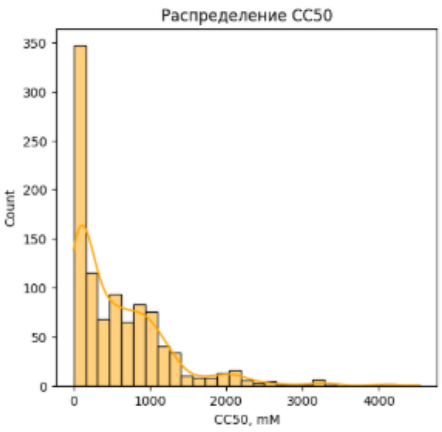
Перед построением моделей был проведён EDA. Основные шаги:

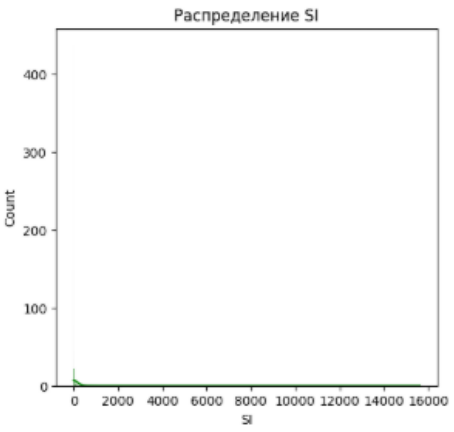
- Удалены пропуски и проверена чистота данных
- Построены распределения IC50, CC50 и SI — они оказались смещены вправо
- Проведён корреляционный анализ — высокая корреляция между IC50 и CC50
- Обнаружены выбросы — особенно среди SI > 10000
- Некоторые признаки почти не варьируются — удалены

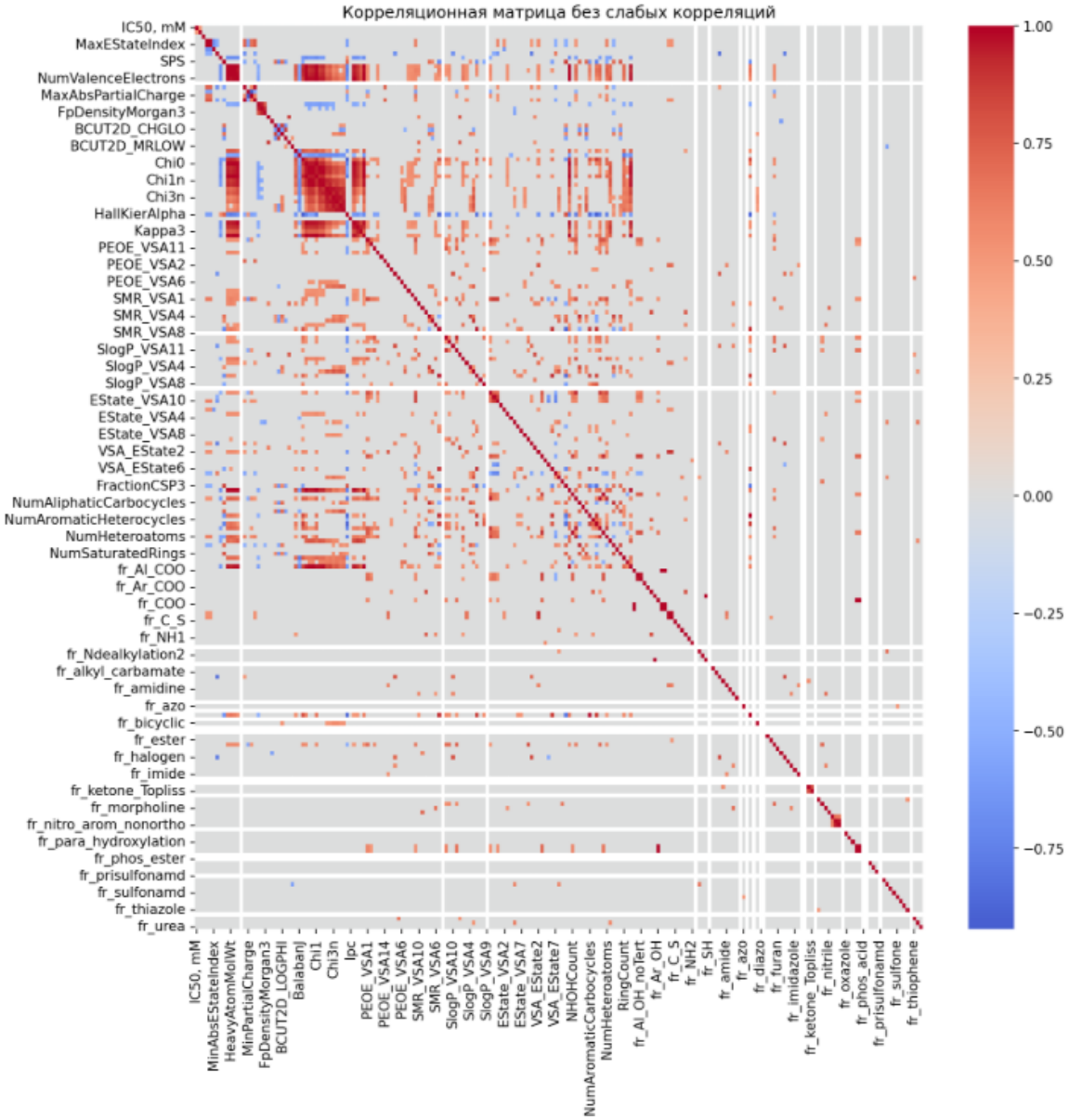
Интересные наблюдения:

- У SI много экстремальных значений, часто из-за деления на очень малые CC50
- IC50 и CC50 распределены логнормально, логарифмирование улучшает стабильность моделей
- Некоторые дескрипторы имеют высокую дисперсию, особенно связанные с массой и числом атомов









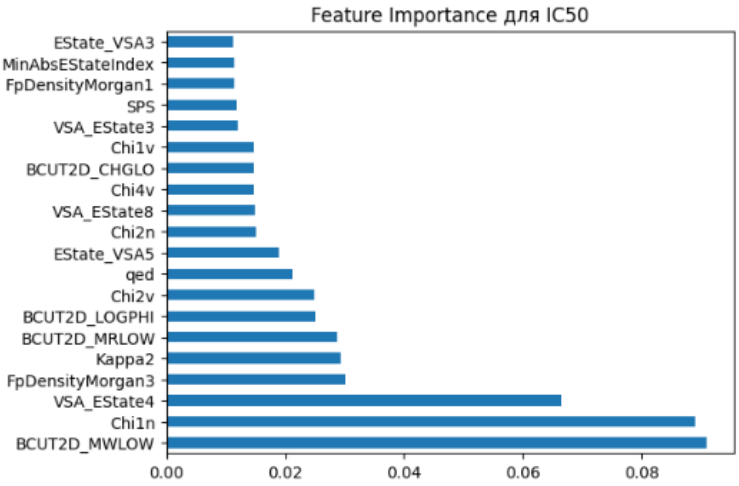
Данные достаточно разнообразные: нет тотальной мультиколлинеарности

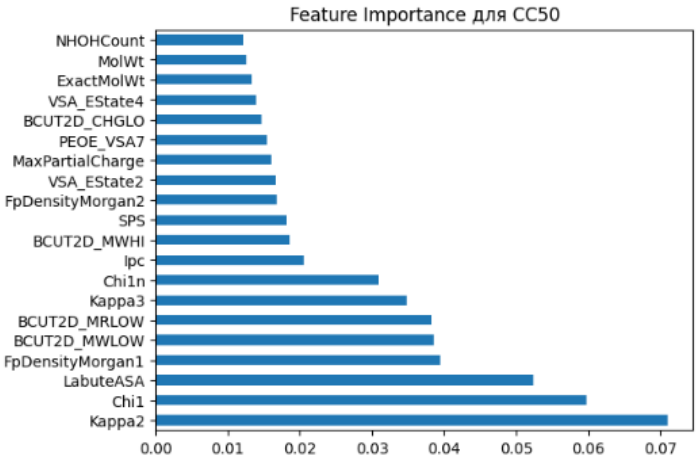
Но есть группы признаков, которые коррелируют сильно — их надо будет учитывать при построении моделей

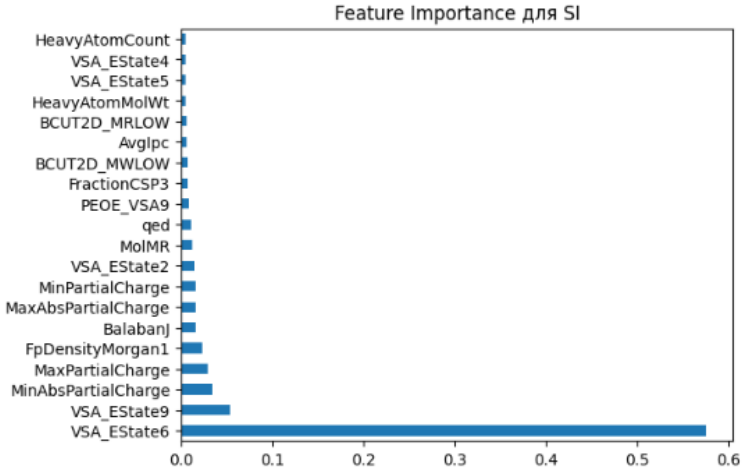
Признаки типа Morgan, Chi, BCUT2D — наиболее информативны.

Удаление некоторых сильно коррелирующих блоков может уменьшить переобучение.

Так же мы можем видеть что вносит наибольший вклад в IC50, CC50 и SI







Задачи:

Задача 1: Регрессия IC50

Модели:

- Linear Regression
- Random Forest
- SVR

Результаты:

Модель	MAE	MSE	R^2
Linear Regression	~208	146214	0.56
Random Forest	**34**	**62508**	**0.81**
SVR	211	278467	0.16

Вывод: Linear Regression слабовата — не улавливает нелинейные зависимости.
SVR не справился, возможно, из-за чувствительности к масштабам и аномалиям.

Задача 2: Регрессия CC50

Результаты:
CatBoost и XGBoost показывают уверенные результаты, значительно превосходящие линейную модель.

Модель	MAE	MSE R^2
Linear Regression	высокие ошибки, $R^2 < 0.5$	
CatBoost	лучший результат, MAE ~90	
XGBoost	немного хуже CatBoost	

Задача 3: Регрессия SI

Интересный момент:
SI = CC50 / IC50 → очень нестабильная метрика.

- Linear Regression вообще «взрывается»: $R^2 \approx -19\,000$
- SVR и CatBoost стабилизируются после удаления выбросов

Модель	MAE	MSE	R^2
Linear Regression	14418	огромная	-19932
Random Forest	198	~1.8M	0.09
SVR	184	~2M	-0.01

Вывод: Предсказывать SI напрямую — неэффективно, лучше предсказывать IC50 и CC50 отдельно, а потом рассчитывать SI вручную.

Классификации: медианы

IC50 > медианы

Модель	**Accuracy**	**F1**	**ROC-AUC**
Logistic Regression	0.70	0.71	0.70
Random Forest	**0.745**	**0.76**	**0.75**
CatBoost	0.72	0.73	0.72
XGBoost	0.705	0.71	0.71

CC50 > медианы

CatBoost и XGBoost победили с результатом до **0.77 ROC-AUC**
Logistic Regression стабильно около **0.72**

SI > медианы

Модель	**Accuracy**	**F1**	**ROC-AUC**
Logistic Regression	0.62	0.61	0.62
Random Forest	**0.66**	**0.63**	**0.66**
CatBoost и XGBoost — около 0.62			

Классификация SI > 8

Почему важно: порог 8 — ориентир фармкомпаний на "терапевтически значимые" соединения.

Модель	**Accuracy**	**F1**	**ROC-AUC**
Logistic Regression	0.695	0.573	0.669
CatBoost	0.670	0.535	0.639
XGBoost	**0.720**	**0.605**	**0.694**

Вывод: задача сложнее, чем медианная классификация, классы несбалансированные. XGBoost здесь показывает лучшую устойчивость.

Идеи для улучшения

Устранить дисбаланс классов с помощью SMOTE / class_weight='balanced'

Feature Engineering: создать ratio-дескрипторы (масса/атомы, H-бонды/масса и т.д.)

Логарифмировать IC50, CC50 перед SI

Убрать выбросы SI > 10000 или ограничить логарифмом

Использовать ансамбли моделей (Stacking)

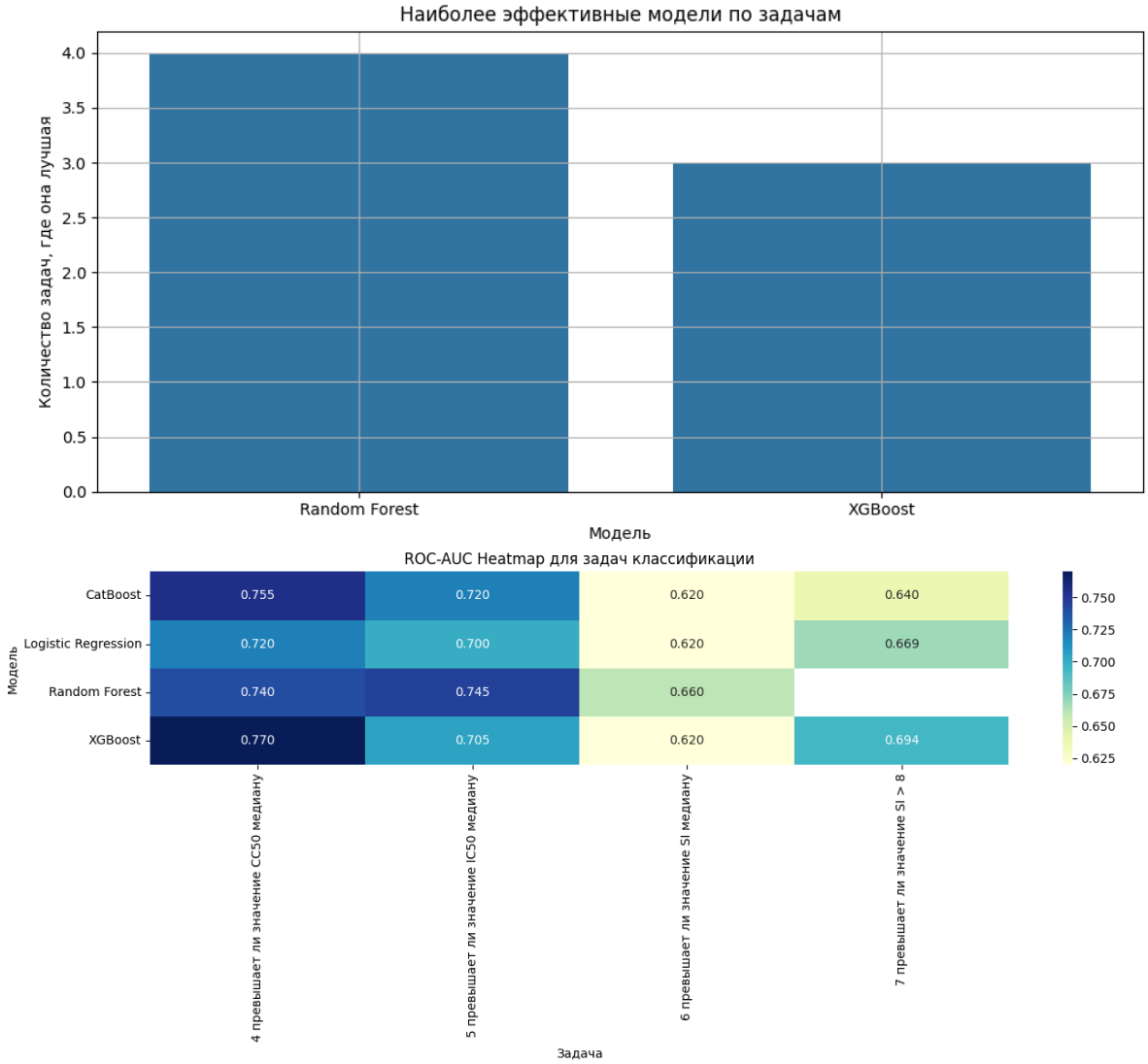
Добавить SHAP-анализ — почему модель делает именно такое предсказание

Вывод

Не могу сказать что это лучшее что я мог сделать, но что-то получилось.

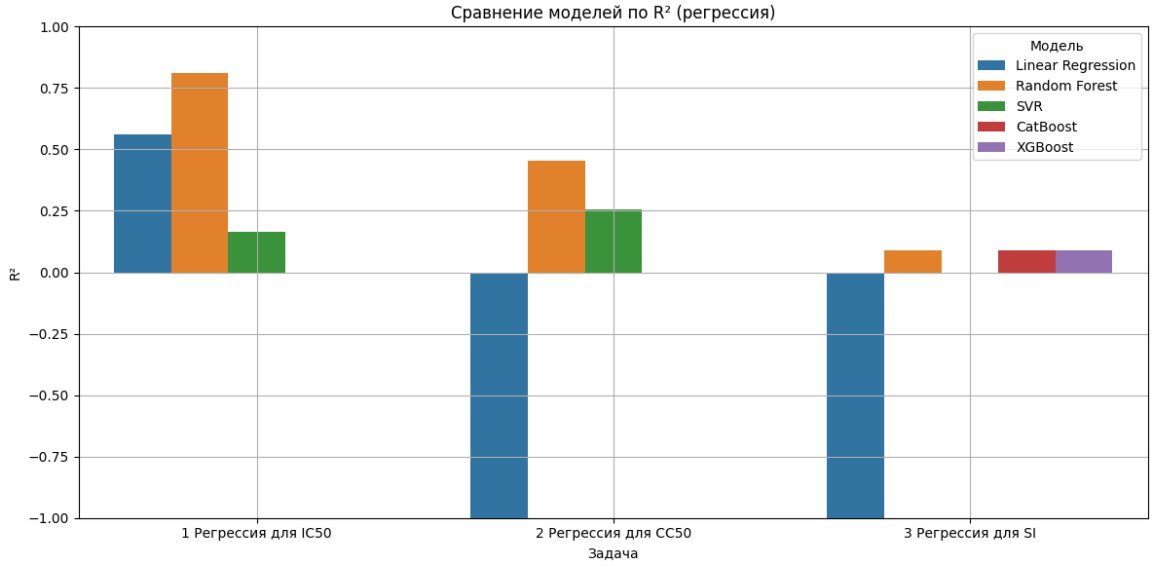
CatBoost и XGBoost уверенно справляются почти во всех задачах, особенно при классификации.
Random Forest отлично показывает себя в регрессии IC50.
SI — нестабильная метрика, её лучше считать вручную.

Машинное обучение позволяет упростить отбор кандидатов на лекарство, но требует аккуратного подхода к данным. В дальнейшем важно улучшать качество и баланс датасета, использовать интерпретируемые модели, и проверять результаты медицинскими экспертами.



Рекомендую:

- Для IC50 — Random Forest
- Для CC50 — CatBoost
- Для SI > 8 — XGBoost
- Избегать прямого регрессирования SI



Спасибо за внимание, надеюсь не сильно плохо!