**Аналитический отчет по курсовой работе (vo\_PJ)**

**Предмет: Классическое машинное обучение**

**Тема: Исследование эффективности лекарственной активности для создания лекарственных препаратов**.

НИЯУ МИФИ, онлайн школа SkillFactory

Онлайн магистратура: группа М24-525

01.04.02 Прикладная математика и информатика

Борщевский Андрей Владимирович

**Цель исследования:**

Создать несколько максимально эффективных моделей машинного обучения для регрессии и классификации, способных предсказывать значения показателей на основе структурных молекулярных дескрипторов.

На основе данных по химическим соединениям построить прогноз эффективности веществ с целью подбора оптимального состава лекарственного препарата, предсказание ключевых показателей активности соединений (IC50, CC50, SI) и их классификации.

Этапы исследования:

1. Провести исследовательский анализ данных (EDA)

2. Построить модели машинного обучения:

Регрессия для IC50

Регрессия для CC50

Регрессия для SI

Классификация: превышает ли значение IC50 медианное значение выборки

Классификация: превышает ли значение CC50 медианное значение выборки

Классификация: превышает ли значение SI медианное значение выборки

Классификация: превышает ли значение SI значение 8

3. Выполнить сравнительный анализ полученных моделей и их результатов, выбрать наиболее качественную модель для решения практической задачи.

Целевые переменные данных, файл: Data\_course\_Classic\_ML.xlsx

Эти параметры используются как целевые при прогнозировании эффективности и безопасности соединений.

Прогноз эффективности сочетания параметров для создания лекарственных препаратов на основе данных датасета с показателями целевых переменных:

1. IC₅₀ (Inhibitory Concentration 50) — концентрация вещества, при которой оно ингибирует 50% активности определённого фермента или клеточного процесса, чем ниже значение IC₅₀, тем сильнее вещество.
2. CC₅₀ (Cytotoxic Concentration 50) — концентрация вещества, при которой 50% клеток погибает или теряет жизнеспособность, чем выше CC₅₀, тем менее токсичным является вещество.
3. SI (Selectivity Index — Индекс селективности) — отношение токсической дозы (CC₅₀) к лечебной дозе (IC₅₀). Показывает, насколько вещество селективно действует — то есть, насколько хорошо оно подавляет целевой эффект, не повреждая клетки.

Описание молекулярных дескрипторов:

1. Физико-химические свойства (Physicochemical Properties)

Дескрипторы, характеризующие общие химические и физические характеристики молекул.

MolWt: молекулярная масса.

HeavyAtomMolWt: масса молекулы без учёта водорода.

ExactMolWt: точная молекулярная масса, рассчитанная на основе точных масс атомов.

NumValenceElectrons: общее количество валентных электронов.

NumRadicalElectrons: количество свободных радикалов (непарных электронов).

MolLogP: логарифм коэффициента распределения между октаном и водой — мера гидрофобности.

MolMR: молекулярная рефрактивность — связана с поляризуемостью молекулы.

TPSA (Topological Polar Surface Area): полярная поверхность молекулы важна для растворимости и проницаемости через биологические мембраны.

LabuteASA: альфа-поверхность молекулы, учитывающая только гидрофобную часть.

qed: Quantitative Estimate of Drug-likeness — численная оценка свойств, характерных для лекарственных препаратов.

SPS: Surface Potential Score — оценка поверхностного потенциала молекулы.

2. Электронные состояния (EState и VSA) характеризуют электронное состояние атомов и их поверхностную доступность.

MaxEStateIndex / MinEStateIndex: максимальный/минимальный индекс электронного состояния атомов в молекуле.

MaxAbsEStateIndex/MinAbsEStateIndex: абсолютные максимальные/минимальные значения индекса электронного состояния.

EState\_VSA1–14: участки поверхностной доступности, основанные на электронном состоянии атомов.

VSA\_EState1–10: другие варианты поверхностной доступности, основанные на электронном состоянии.

3. BCUT-дескрипторы

Способ представить структуру молекулы в виде числового вектора, который учитывает как физико-химические свойства, так и структурные особенности. Считаются с использованием взвешенной матрицы смежности, построенной на основе структуры молекулы, и последующего извлечения собственных значений этой матрицы.

BCUT2D\_MWHI / MWLOW: величины, соответствующие высшей и низшей границе молекулярной массы.

BCUT2D\_CHGHI / CHGLOW: высшая и низшая границы заряда.

BCUT2D\_LOGPHI / LOGPLOW: границы по гидрофобности.

BCUT2D\_MRHI / MRLOW: границы по молекулярной рефрактивности.

4. Топологические индексы (Topological Indices): индексы, которые отражают топологическую структуру молекулы.

Chi0 / Chi0n / Chi0v: нулевой порядок топологического индекса Chi.

Chi1 / Chi1n / Chi1vн: первый порядок индекса Chi.

Chi2n / Chi2v: второй порядок индекса Chi.

Chi3n / Chi3v: третий порядок индекса Chi.

Chi4n / Chi4v: четвёртый порядок индекса Chi.

AvgIpc / Ipc: индекс информации по числу связей (Information Content Index).

BalabanJ: индекс связности, мера структурной сложности молекулы.

BertzCT: индекс сложности молекулы по Берцу.

HallKierAlpha: параметр α из метода Холла-Киер — используется в расчётах поляризуемости.

Kappa1 / Kappa2 / Kappa3: Индексы Каппа — характеризуют размер, форму и разветвлённость молекулы.

5. Дескрипторы поверхностной доступности (Surface Accessibility)

Описывают доступность различных участков молекулы, особенно полезны для моделирования взаимодействия с белками.

PEOE\_VSA1–14: Участки поверхностной доступности, основанные на частичных зарядах PEOE.

SMR\_VSA1–10: Участки поверхностной доступности, основанные на молекулярной рефрактивности.

SlogP\_VSA1–12: Участки поверхностной доступности, связь гидрофобностью (logP) основанные на логП.

6. Молекулярные фрагменты (Fragments): дескрипторы, которые указывают на наличие определённых функциональных групп или структур в молекуле.

fr\_Al\_COO: фрагмент: альфа-карбоксильная группа.

fr\_Al\_OH / fr\_Al\_OH\_noTert: фрагменты: альфа-гидроксильные группы.

fr\_ArN / fr\_Ar\_COO / fr\_Ar\_N / fr\_Ar\_NH / fr\_Ar\_OH: фрагменты: ароматические группы.

fr\_COO / fr\_COO2 / fr\_C\_O / fr\_C\_S / fr\_HOCCN / fr\_Imine: различные функциональные группы (амиды, эфиры, сульфиды и др.).

fr\_NH0 / fr\_NH1 / fr\_NH2 / fr\_N\_O / fr\_Ndealkylation1 / Ndealkylation2 / fr\_Nhpyrrole / fr\_SH: связаны с азотистыми и серосодержащими группами.

fr\_aldehyde / fr\_ester / fr\_ether / fr\_furan / fr\_pyridine / fr\_thiazole / fr\_thiophene: фрагменты: альдегиды, эфиры, эфиры, фураны, пиридин, тиазол, тиофен.

fr\_halogen / fr\_term\_acetylene / fr\_tetrazole / fr\_unbrch\_alkane / fr\_urea: фрагменты: галогены, терминальные ацетилены, тетразолы, алканы, мочевина.

7. Структурные и молекулярные характеристики

HeavyAtomCount: количество тяжёлых атомов (C, O, N и др.).

FractionCSP3: доля sp³-гибридизированных углеродных атомов.

NHOHCount / NOCount: количество NH-OH и N=O групп.

NumAliphaticCarbocycles / NumAliphaticHeterocycles / NumAliphaticRings: количество алифатических циклов.

NumAromaticCarbocycles / NumAromaticHeterocycles / NumAromaticRings: количество ароматических циклов.

NumHAcceptors / NumHDonors: количество атомов, способных принимать/донировать водородные связи.

NumHeteroatoms: количество гетероатомов (атомов, отличных от C и H).

NumRotatableBonds: количество вращающихся связей — влияет на гибкость молекулы.

NumSaturatedCarbocycles / NumSaturatedHeterocycles / NumSaturatedRings: количество насыщенных циклов.

RingCount: общее количество циклов в молекуле.

На основании предоставленных данных от химиков необходимо построить прогноз, позволяющий подобрать наиболее эффективное сочетание параметров для создания лекарственных препаратов.

Информация о DataFrame:

Количество записей: 1001

Индекс: от 0 до 1000

Количество колонок: 213 Статистика по колонкам: Название первой колонки: IC50

Тип данных первой колонки: float64

Последняя колонка: fr\_urea

Тип данных последней колонки: int64

Динамика типов данных:

float64: 107 колонок

int64: 106 колонок

Использование памяти: 1.6 MB

**1. Качество проведённого EDA**

Проверка типов данных

Поиск нечисловых столбцов

Получение описания, статистики, включая все типы данных

Замена имен столбцов 'IC50, mM': 'IC50', 'CC50, mM': 'CC50'

Вывод колонок с категориальным типом

Вывод колонок с числовым типом

Cтатистики для числовых столбцов

информацию о столбцах, состоящих только из нулей

Удаление нулевых столбцов из исходного датафрейма

Построение гистограмм с KDE и линиями медианы и среднего

Построения боксплотов для визуализации выбросов

Вывод:

IC50, CC50 и SI характеризуются правосторонней асимметрией распределений с наличием экстремальных значений, особенно выраженных в случае SI. Это подтверждается тем, что медиана меньше среднего значения во всех случаях, что указывает на влияние длинных правых хвостов.

Таким образом, данные особенности будем интерпретировать как важные характеристики целевых переменных , которые могут оказывать существенное влияние на моделирование и прогнозирование.

Анализ свойств молекул:

1. MaxAbsEStateIndex (максимальное абсолютное значение электрохимического состояния)

Среднее = 10.89

Медиана = 12.20

Умеренная левосторонняя асимметрия, низкая вариативность. Показатель стабильный, данные однородны, с хорошей концентрацией значений около медианы. Свойства молекул по этому признаку относительно однородны, без выраженных выбросов.

2. MolWt (молекулярная масса)

Среднее = 351.49

Медиана = 318.37

Сильная правосторонняя асимметрия, высокая вариативность. Присутствуют молекулы с чрезвычайно высокой массой.

Все корреляции:

IC50 & CC50: 0.52

IC50 & SI: -0.06

IC50 & MolWt: -0.16

IC50 & MolLogP: -0.23

IC50 & MaxAbsEStateIndex: 0.11

CC50 & SI: -0.01

CC50 & MolWt: -0.31

CC50 & MolLogP: -0.22

CC50 & MaxAbsEStateIndex: -0.11

SI & MolWt: -0.04

SI & MolLogP: -0.08

SI & MaxAbsEStateIndex: 0.00

MolWt & MolLogP: 0.42

MolWt & MaxAbsEStateIndex: 0.39

MolLogP & MaxAbsEStateIndex: -0.15

Сильные корреляции (>= 0.7 или <= -0.7):

Нет сильных корреляций.

Выводы:

Анализ мультиколлинеарности показывает наличие сильных линейных зависимостей между некоторыми признаками, что может повлиять на качество и интерпретируемость моделей.

Сильная корреляция (≥ 0.99):

ExactMolWt & MolWt : 0.999999

Chi1 & HeavyAtomCount : 0.998742

Эти пары признаков демонстрируют очень высокую степень взаимосвязи, почти полную линейную зависимость.

Рекомендации: удалить некоторые сильно коррелирующие пары.

Численный анализ целевых переменных:

Столбец: IC50

Среднее: 220.73

Медиана: 45.34

IQR (размах межквартильного диапазона): 218.86

Количество выбросов (ниже Q1 - 1.5IQR): 0

Количество выбросов (выше Q3 + 1.5IQR): 138

Столбец: CC50

Среднее: 586.42

Медиана: 424.17

IQR (размах межквартильного диапазона): 791.78

Количество выбросов (ниже Q1 - 1.5IQR): 0

Количество выбросов (выше Q3 + 1.5IQR): 35

Столбец: SI

Среднее: 73.97

Медиана: 3.90

IQR (размах межквартильного диапазона): 14.89

Количество выбросов (ниже Q1 - 1.5IQR): 0

Количество выбросов (выше Q3 + 1.5IQR): 120

Вывод:

IC50, CC50 и SI характеризуются правосторонней асимметрией распределений с наличием экстремальных значений, особенно выраженных в случае SI. Это подтверждается тем, что медиана меньше среднего значения во всех случаях, что указывает на влияние длинных правых хвостов.

Таким образом, данные особенности будем интерпретировать как важные характеристики целевых переменных, которые могут оказывать существенное влияние на моделирование и прогнозирование.

Численный анализ свойств молекул:

Свойство: MaxAbsEStateIndex

Среднее значение: 10.89

Медиана: 12.20

Стандартное отклонение: 3.28

Минимальное значение: 2.32

Максимальное значение: 15.93

Свойство: MolWt

Среднее значение: 351.49

Медиана: 318.37

Стандартное отклонение: 127.39

Минимальное значение: 110.16

Максимальное значение: 904.78

Анализ свойств молекул:

1. MaxAbsEStateIndex (максимальное абсолютное значение электрохимического состояния)

Среднее = 10.89

Медиана = 12.20

Умеренная левосторонняя асимметрия, низкая вариативность. Показатель стабильный, данные однородны, с хорошей концентрацией значений около медианы. Свойства молекул по этому признаку относительно однородны, без выраженных выбросов.

1. MolWt (молекулярная масса)

Среднее = 351.49

Медиана = 318.37

Сильная правосторонняя асимметрия, высокая вариативность. Присутствуют молекулы с чрезвычайно высокой массой.

Все корреляции:

IC50 & CC50: 0.52

IC50 & SI: -0.06

IC50 & MolWt: -0.16

IC50 & MolLogP: -0.23

IC50 & MaxAbsEStateIndex: 0.11

CC50 & SI: -0.01

CC50 & MolWt: -0.31

CC50 & MolLogP: -0.22

CC50 & MaxAbsEStateIndex: -0.11

SI & MolWt: -0.04

SI & MolLogP: -0.08

SI & MaxAbsEStateIndex: 0.00

MolWt & MolLogP: 0.42

MolWt & MaxAbsEStateIndex: 0.39

MolLogP & MaxAbsEStateIndex: -0.15

Сильные корреляции (>= 0.7 или <= -0.7):

Нет сильных корреляций.

Выводы:

Анализ мультиколлинеарности показывает наличие сильных линейных зависимостей между некоторыми признаками, что может повлиять на качество и интерпретируемость моделей.

Сильная корреляция (≥ 0.99):

ExactMolWt & MolWt : 0.999999

Chi1 & HeavyAtomCount : 0.998742

Эти пары признаков демонстрируют очень высокую степень взаимосвязи, почти полную линейную зависимость.

Рекомендации: удалить некоторые сильно коррелирующие пары.

Вывод:

1. Тест Шапиро-Уилка Все три признака (CC50, IC50, SI) не соответствуют нормальному распределению , так как: Значения статистики значительно меньше 1. Все p-значения < 0.05, что позволяет отвергнуть нулевую гипотезу о нормальности.
2. Тест Андерсона-Дарлина CC50

Статистика: 42.665

Критические значения:

15% — 0.574

30% — 0.653

45% — 0.784

60% — 0.914

75% — 1.088

* Статистика больше всех критических значений, что подтверждает отсутствие нормальности.

IC50 Статистика: 134.582 Критические значения те же, что и выше.

* Статистика значительно превышает все критические значения, что также указывает на отсутствие нормальности.

SI Статистика: 206.889

* Аналогично предыдущим признакам, статистика намного больше всех критических значений , что подтверждает сильное отклонение от нормального распределения.
* Общий вывод Ни один из исследуемых признаков (CC50, IC50, SI) не следует нормальному распределению ни по тесту Шапиро-Уилка, ни по тесту Андерсона-Дарлина.

Создан очищенный̆ датафрейм df\_standardscaler.csv, подготовленный для дальнейших этапов построения моделей.

Применяемые модели:

* 1. DecisionTreeRegressor

Модель на основе дерева решений, которое строит правила для разделения данных и делает предсказания на основе значений признаков.

Преимущества:

простая в интерпретации, не требует сложной предобработки данных, быстро обучается на небольших наборах данных.

Недостатки:

склонна к переобучению, особенно если дерево глубокое, менее точна по сравнению с ансамблевыми методами, не устойчива к шуму и выбросам.

Гиперпараметры:

max\_depth — максимальная глубина дерева.

min\_samples\_split — минимальное количество образцов для разделения узла.

min\_samples\_leaf — минимальное количество образцов в листе.

* 1. RandomForestRegressor

Ансамблевый алгоритм, который использует множество деревьев решений , каждое из которых обучается на случайной подвыборке данных и признаков. Результат — среднее значение предсказаний всех деревьев.

Преимущества: Устойчива к переобучению, подходит для работы с большими наборами данных и множеством признаков, обладает хорошей точностью и стабильностью.

Недостатки: требует больше времени и памяти по сравнению с одним деревом, меньше интерпретируема по сравнению с Decision Tree.

Гиперпараметры:

n\_estimators — количество деревьев в лесу.

max\_depth — максимальная глубина каждого дерева.

min\_samples\_split, min\_samples\_leaf — контролируют размер узлов и листьев.

* 1. XGBRegressor (XGBoost Regressor) - реализация алгоритма градиентного бустинга с использованием деревьев решений, известен своей скоростью, эффективностью и возможностью оптимизации.

Преимущества: высокая точность прогнозирования, хорошо работает с большими объёмами данных, поддерживает регуляризацию, что помогает бороться с переобучением.

Недостатки: требует тщательной настройки гиперпараметров, может быть менее устойчивой к пропускам и категориальным данным без предварительной обработки.

Гиперпараметры:

learning\_rate — скорость обучения.

max\_depth — максимальная глубина дерева.

n\_estimators — количество деревьев.

subsample — доля данных, используемых для обучения одного дерева.

colsample\_bytree — доля признаков, используемых для построения дерева.

* 1. CatBoostRegressor

Модель на основе градиентного бустинга деревьев, разработанная компанией CatBoost. Отличается тем, что автоматически обрабатывает категориальные признаки, не требуя их предварительного кодирования.

Преимущества: лучшая обработка категориальных признаков, автоматическое управление отсутствующими данными, демонстрирует лучшую производительность благодаря внутренним оптимизациям.

Недостатки: может потреблять больше памяти, меньше контроля над процессом обучения по сравнению с XGBoost.

Гиперпараметры:

learning\_rate — скорость обучения.

depth — глубина деревьев.

iterations — количество итераций (деревьев).

l2\_leaf\_reg — параметр L2-регуляризации.

verbose — уровень вывода информации при обучении.

1. Регрессия для IC50



Основные выводы

1. CatBoostRegressor показывает наилучшие результаты Имеет минимальный MSE, что указывает на наименьшую среднеквадратичную ошибку предсказаний. Также имеет лучший R², что говорит о немного лучшей способности модели объяснять вариацию целевой переменной IC50. MAE также близок к минимальному, что подтверждает стабильность модели.

Вывод: CatBoostRegressor — самая точная модель из всех тестированных.

1. XGBoost и Random Forest имеют схожее качество по всем метрикам они находятся в очень тесной конкуренции. Разница между ними незначительна, особенно по R² и MAE. Это может говорить о том, что обе модели одинаково хорошо подходят для задачи.

Вывод: XGBoost и Random Forest демонстрируют сопоставимое качество.

1. DecisionTreeRegressor хуже остальных моделей, у неё наивысший MSE и MAE, а R² самый низкий. Это может быть связано с переобучением или недостаточной сложностью модели. Простые деревья решений могут не улавливать сложные зависимости в данных.

Вывод: DecisionTreeRegressor не рекомендуется использовать как основную модель в данной задаче.

1. Гиперпараметры CatBoostRegressor оказались оптимальными

Лучшие параметры:

learning\_rate = 0.021

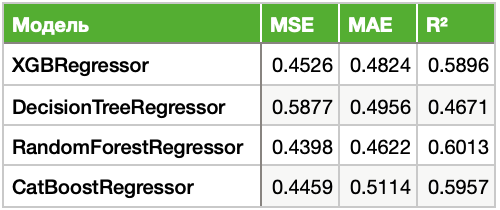
depth = 5

iterations = 269

l2\_leaf\_reg = 0.19

Эти значения обеспечивают хороший баланс между скоростью обучения и качеством предсказаний, что делает модель надежной и эффективной.

1. Все модели показывают низкие значения R² Все модели имеют R² < 0.2, что указывает на то, что модели не очень хорошо объясняют вариацию целевой переменной.
2. Регрессия для CC50



1. RandomForestRegressor показывает наилучшие результаты, меет минимальное значение MSE, наименьшую ошибку предсказания.

Также имеет лучший R², эта модель лучше всего объясняет вариацию целевой переменной CC50.

RandomForestRegressor — лучшая модель по метрикам качества среди всех рассмотренных.

1. XGBoost и RandomForestRegressor конкурируют в тестах. Обе модели имеют очень близкие значения MSE и R². Разница между ними незначительна, особенно учитывая стабильность Random Forest.

XGBoost и RandomForestRegressor являются сопоставимыми по качеству.

1. CatBoostRegressor демонстрирует средние результаты. У CatBoostRegressor среднее значение MSE и немного худший R² по сравнению с RandomForestRegressor.

MAE у CatBoostRegressor даже ниже, чем у остальных моделей.

CatBoostRegressor имеет хорошие, результаты.

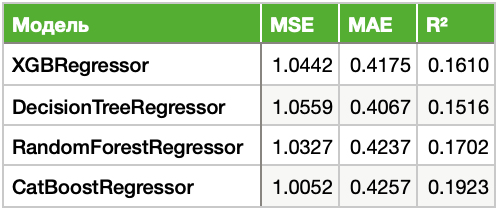
1. DecisionTreeRegressor хуже остальных моделей. Имеет самое высокое значение MSE и MAE, и самый низкий R².

DecisionTreeRegressor не рекомендуется использовать как основную модель в данной задаче.

Заключение:

Модель RandomForestRegressor показала себя лучше всего по всем ключевым метрикам, поэтому используем RandomForestRegressor как основную модель для дальнейших предсказаний.

1. Регрессия для SI



Выводы:

1. CatBoostRegressor показала наилучшее качество среди всех рассмотренных моделей по всем ключевым метрикам:

MSE = 1.0052

R² = 0.1923

Выбранны гиперпараметры:

learning\_rate: 0.2338

depth: 8

iterations: 547

l2\_leaf\_reg: 0.0015

имеет наименьшее значение MSE, что говорит о лучшей точности предсказаний, демонстрирует наивысший R², что указывает на лучшую объясняющую способность модели.

1. RandomForestRegressor — показывает очень близкие значения MSE и R² к CatBoostRegressor.
2. XGBoostRegressor и DecisionTreeRegressor — средние результаты. У XGBoostRegressor худший R², чем у Random Forest и CatBoost. DecisionTreeRegressor имеет незначительно худшие метрики, особенно по сравнению с CatBoostRegressor.

Модели для классификации:

* + 1. DecisionTreeClassifier - строит дерево решений на основе значений признаков для разделения данных на классы.

Основные гиперпараметры:

max\_depth: максимальная глубина дерева (ограничивает сложность модели).

min\_samples\_split: минимальное количество образцов для разделения узла.

min\_samples\_leaf: минимальное количество образцов в листовом узле.

criterion: критерий оценки качества разделения (по умолчанию 'gini').

1. RandomForestClassifier — ансамблевая модель, состоящая из множества деревьев решений. Каждое дерево обучается на случайной подвыборке данных и признаков, а окончательный прогноз делается голосованием или усреднением результатов.

Основные гиперпараметры:

n\_estimators: количество деревьев в ансамбле.

max\_depth: максимальная глубина каждого дерева.

min\_samples\_split, min\_samples\_leaf: контролируют размер узлов и листьев.

criterion: критерий разделения ('gini' или 'entropy').

1. XGBClassifier (Extreme Gradient Boosting Classifier) — это реализация алгоритма градиентного бустинга, которая использует последовательное обучение деревьев с корректировкой ошибок.

Основные гиперпараметры:

learning\_rate: скорость обучения (малое значение улучшает качество, но увеличивает время обучения).

max\_depth: максимальная глубина деревьев.

n\_estimators: количество деревьев в ансамбле.

subsample: доля данных, используемых для обучения одного дерева.

colsample\_bytree: доля признаков, используемых для построения дерева.

4. CatBoostClassifier — это реализация градиентного бустинга, специально адаптированная для работы с категориальными признаками, не требующая их предварительного кодирования. Также поддерживает автоматическое управление пропущенными значениями.

Основные гиперпараметры:

learning\_rate: скорость обучения.

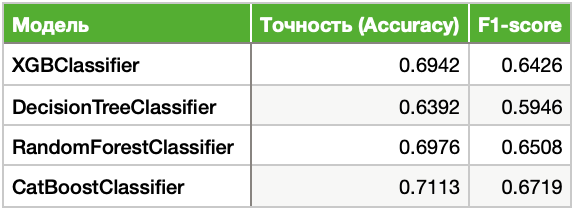
depth: глубина деревьев.

iterations: количество итераций (деревьев).

l2\_leaf\_reg: коэффициент L2-регуляризации.

verbose: уровень детализации вывода информации во время обучения.

1. Классификация: превышает ли значение IC50 медианное значение выборки

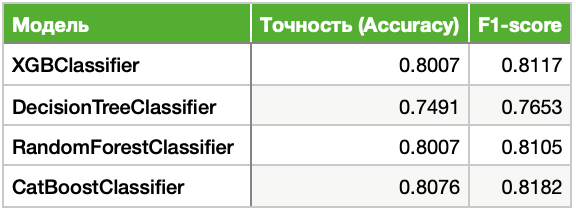


1. CatBoostClassifier показывает наилучшие результаты, имеет наивысшую точность (Accuracy) и F1-score, что говорит о том, что модель лучше всего справляется с задачей.
2. RandomForestClassifier демонстрирует стабильное качество Следом за CatBoost следует Random Forest с немного худшими, но всё ещё хорошими значениями метрик.
3. XGBoostRegressor имеет средние результаты, показатели Accuracy и F1-score у XGBoost близки к Random Forest. Разница между ними незначительна, но она всё же есть в пользу CatBoost и Random Forest.
4. DecisionTreeClassifier хуже остальных моделей, имеет самые низкие значения Accuracy и F1-score, что указывает на низкую способность обобщать данные.

Вывод:

* CatBoostClassifier — лучшая модель для текущей задачи классификации. Оптимальные гиперпараметры для CatBoostClassifier обеспечивают хороший баланс между скоростью обучения и качеством предсказаний, что делает модель надежной и эффективной.
* RandomForestClassifier может быть альтернативой, которая также эффективно решает задачу.
* XGBoostRegressor — подходящая модель, но не лидирует по качеству.
* DecisionTreeClassifier не рекомендуется использовать как основную модель в данной задаче.

1. Классификация: превышает ли значение CC50 медианное значение выборки



1. CatBoostClassifier показывает наилучшие результаты, меет наивысшую точность (Accuracy) и F1-score, что говорит о его лучшей способности корректно классифицировать объекты.
2. XGBoost и RandomForest имеют сопоставимые результаты, обе модели демонстрируют очень близкие значения метрик: Accuracy: 0.8007 F1-score: ~0.811
3. DecisionTreeClassifier уступает остальным моделям, имеет значительно худшую точность и F1-score, чем остальные модели, не смогло полностью выявить сложные закономерности в данных.
4. Высокие значения метрик указывают на хорошее качество моделирования Все модели показали выше среднего Accuracy и F1-score, что говорит о хорошем соответствии модели реальным данным. Особенно высокий F1-score у CatBoost (~0.818), что подтверждает баланс между precision и recall.

Вывод:

* CatBoostClassifier — самая эффективная модель для данной задачи, является лидером по всем ключевым метрикам.
* XGBoost и RandomForestClassifier являются альтернативами, которые могут использоваться в зависимости от требований к интерпретируемости и производительности.
* DecisionTreeClassifier не рекомендуется использовать как основную модель в этой задаче.

1. Классификация: превышает ли значение SI медианное значение выборки

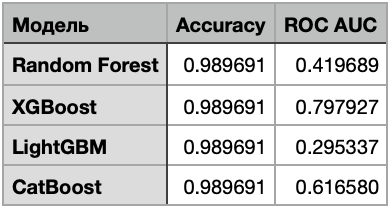


1. CatBoostClassifier показывает наилучшие результаты, имеет наивысший F1-score, что говорит о хорошем балансе между precision и recall. Уровень точности (Accuracy) у CatBoost совпадает с XGBoost, но F1 немного выше, что делает его предпочтительным.
2. XGBoostClassifier демонстрирует высокую стабильность, имеет точность 0.6942, равную CatBoost. F1-score чуть ниже, но всё же достаточно хорош. Это указывает на то, что модель хорошо справляется с задачей.
3. RandomForestClassifier имеет средние результаты, показывает ниже Accuracy и F1, чем CatBoost и XGBoost. Однако он остаётся более устойчивым к переобучению и может быть полезен в случае, если важна интерпретируемость модели или её работоспособность при наличии шума.
4. DecisionTreeClassifier значительно хуже остальных моделей, имеет самый низкий Accuracy и F1-score, что указывает на недообучение или переобучение.

Вывод:

* CatBoostClassifier — лучшая модель по метрикам качества среди всех рассмотренных, использовать как основную модель для дальнейших предсказаний.
* XGBoostClassifier — надёжная альтернатива CatBoostClassifier.
* RandomForestClassifier — подходящий вариант, если требуется компромисс между качеством и интерпретируемостью.
* DecisionTreeClassifier не рекомендуется использовать как основную модель в этой задаче.

08. Классификация: превышает ли значение SI значение 8



1. Все модели имеют одинаковое значение Accuracy, все модели показывают очень высокую точность (Accuracy ≈ 0.99). Возможно указывать на высокий процент объектов одного класса и модель часто предсказывает доминирующий класс. Accuracy не является надёжным критерием в данной задаче, так как не отражает баланс между precision и recall.
2. XGBoost превосходит остальные по ROC AUC, у XGBoost наивысшее значение ROC AUC (≈ 0.80), что говорит о том, что он лучше всего справляется с различением положительного и отрицательного классов. Это делает его наиболее подходящим для задачи, где важны оба класса и нужно учитывать полноту и точность предсказаний.
3. LightGBM имеет наихудший ROC AUC, у LightGBM ROC AUC = 0.295, что почти эквивалентно случайному выбору класса, такой низкий ROC AUC говорит о том, что модель не способна эффективно разделять классы.
4. CatBoost занимает среднюю позицию, имеет ROC AUC = 0.617, что намного выше LightGBM, но ниже XGBoost.

Вывод: CatBoost может быть использован как альтернатива, но не является оптимальной.

XGBoost — лучшая модель по метрике ROC AUC.

LightGBM показывает худшие результаты и требует пересмотра признаков или параметров.

Рекомендации:

Использовать XGBoost как основную модель, если важно качество классификации.

Заключение:

Применяемые модели обеспечивают разнообразные подходы к решению задач классификации, используются для сравнения их производительности на одной и той же задаче, что позволяет выбрать наиболее подходящую для приведенных данных и целей

Общие выводы по моделированию задач классификации:

- Высокая точность (Accuracy) у всех моделей, все рассмотренные модели показали очень близкие значения Accuracy, что говорит о хорошем уровне общего соответствия предсказаний реальным данным.

Это может быть связано с тем, что датасет не сбалансирован, и большинство образцов принадлежит к одному из классов.

- Значительные различия в ROC AUC

- XGBoostClassifier - 0.81 наилучший результат среди всех моделей.

- CatBoostClassifier ~ 0.80 близок к XGBoost, демонстрирует высокую способность к разделению классов.

- RandomForestClassifier ~ 0.64 значительно ниже, что указывает на

меньшую эффективность в определении положительных случаев.

- LightGBMClassifier ~ 0.30 — наименьший показатель, свидетельствует о

плохой способности различать классы.

- Контрарное сочетание метрик

Несмотря на одинаковые или близкие значения Accuracy, разница в ROC AUC подчеркивает, что: XGBoost и CatBoost лучше справляются с задачей, особенно при не сбалансированных данных.

Random Forest и LightGBM имеют меньшую способность к различению классов, что делает их менее надежными для задач, где важны оба класса.

Рекомендации по выбору модели:

CatBoostClassifier - наиболее подходящая модель, лучшая по всем ключевым метрикам. Подходит для задач с категориальными признаками.

XGBoostClassifier - хорошая альтернатива, показывает чуть худшие результаты, чем CatBoost, но всё ещё высокий уровень производительности.

RandomForestClassifier - приемлемый результат, но уступает первым двум моделям, может использоваться при ограниченных ресурсах.

DecisionTreeClassifier - не рекомендуется, слабые результаты по всем метрикам, склонна к переобучению.

LightGBMClassifier - слабый результат.