Аналитический отчет

Автор: Алферов Владислав Андреевич

**Исходные данные**

В качестве исходных данных был предоставлен датасет размерностью 1001 строка и 213 столбцов. Краткое описание столбцов предоставлено не было, однако из их названия можно предположить, что они означают:

* IC50, mM – концентрация соединения, необходимая для гибели вирусов на 50%;
* CC50, mM – концентрация соединения, вызывающая гибель 50% клеток;
* SI– индекс селективности, рассчитывается как **CC50 / IC50**;
* MaxAbsEStateIndex**,** MinAbsEStateIndex**,** MaxEStateIndex**,** MinEStateIndex – параметры, описывающие распределение электронной плотности в молекуле;
* Max/MinPartialCharge – максимальный и минимальный парциальные заряды на атомах молекулы;
* Max/MinAbsPartialCharge – абсолютные значения парциальных зарядов;
* MolWt – молекулярная масса соединения;
* ExactMolWt – точная молекулярная масса (с учетом изотопов);
* HeavyAtomMolWt – масса тяжелых атомов (без водородов);
* NumValenceElectrons – количество валентных электронов;
* NumRadicalElectrons – количество неспаренных электронов (радикалов);
* FractionCSP3 – показатель насыщенности;
* MolLogP – мера гидрофобности;
* MolMR – молекулярная рефракция;
* BalabanJ**,** BertzCT – индексы молекулярной сложности;
* Chi**...** – индексы Киера;
* Kappa…– индексы формы молекулы;
* HallKierAlpha – параметр, учитывающий гибридизацию и размер молекулы;
* TPSA– топологическая полярная поверхность;
* LabuteASA – оценка доступной поверхности атомов;
* PEOE\_VSA\***,** SMR\_VSA\***,** SlogP\_VSA\* – дескрипторы, комбинирующие заряд, полярность и гидрофобность;
* fr\_\*– дескрипторы наличия функциональных групп;
* NumHAcceptors, NumHDonors – количество акцепторов и доноров водородных связей;
* NumRotatableBonds –гибкость молекулы;
* RingCount – количество циклов в молекуле;
* BCUT2D\_\* – дескрипторы, основанные на собственных значениях матриц связности;
* Ipc– информационная емкость молекулы;
* EState\_VSA\***,** VSA\_EState\* – комбинация электронных и поверхностных свойств.

У наблюдаемых переменных можно обнаружить следующее:

1. IC50 – среднее значение составляет 222,81, СКО – 402,17 (очень большой разброс значений);
2. CC50 - среднее значение составляет 589,11, СКО – 642,87 (очень большой разброс значений);
3. SI - среднее значение составляет 72.51, СКО – 684.48 (очень большой разброс значений).

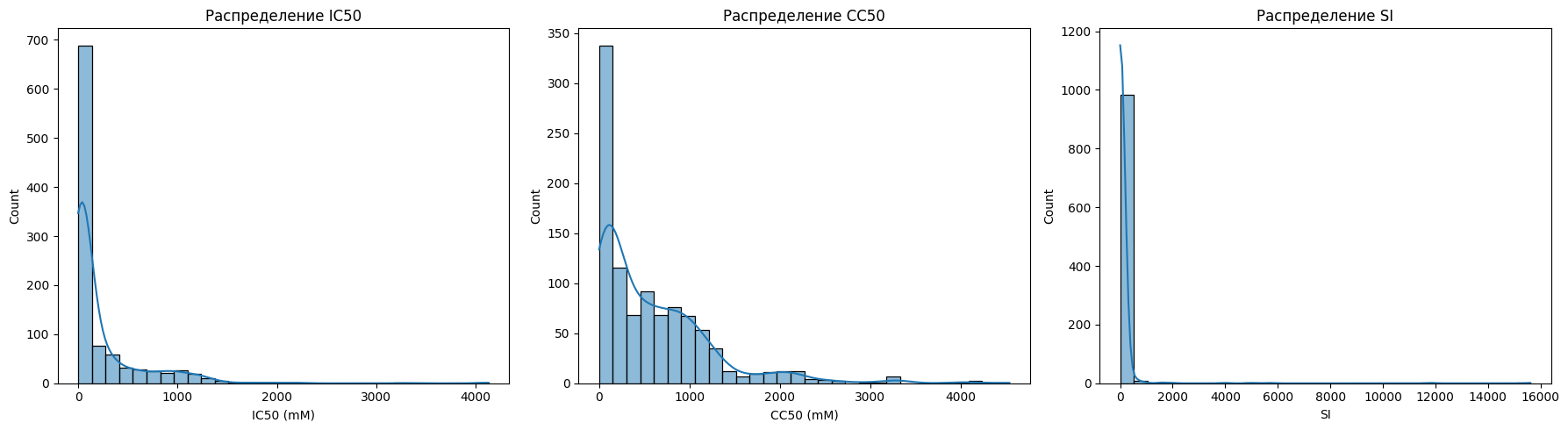
Приведем некоторые дополнительные показатели разбросов данных значений

|  |  |  |  |
| --- | --- | --- | --- |
|  | IC50, mM | CC50, mM | SI |
| 25% | 12.52 | 99.999 | 1.43 |
| 50% | 46.59 | 411.04 | 3.85 |
| 75% | 224.98 | 894.09 | 16.57 |
| max | 4128.53 | 4538.98 | 15620.6 |

Из данных показателей можно увидеть тот факт, что данные достаточно сильно разбросаны в датасете, что говорит о сложности в их дальнейшей обработке линейными моделями (что было продемонстрированно на практике).

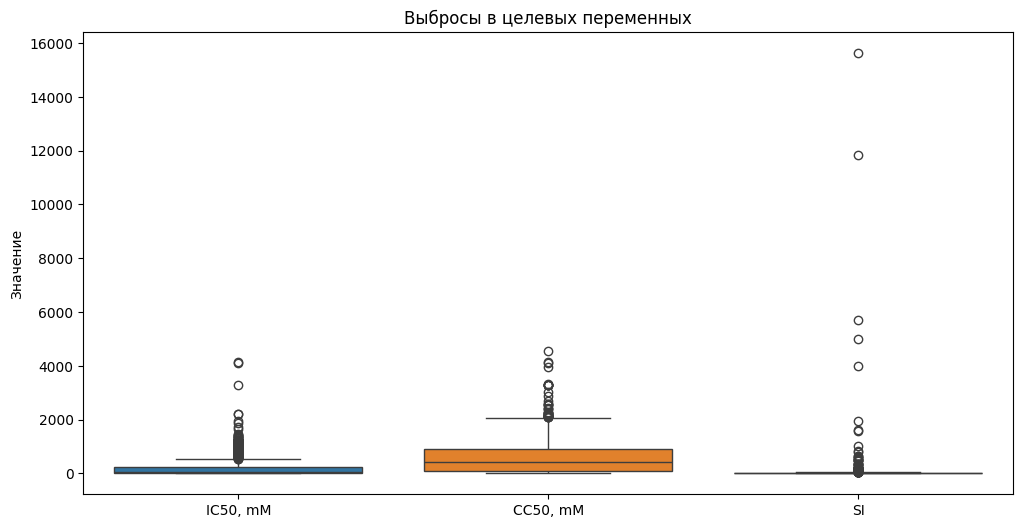
Строки 78, 79, 80 также содержали пропуски в некоторых столбцах (MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, MinAbsPartialCharge, BCUT2D\_MWHI, BCUT2D\_MWLOW, BCUT2D\_CHGHI, BCUT2D\_CHGLO, BCUT2D\_LOGPHI, BCUT2D\_LOGPLOW, BCUT2D\_MRHI, BCUT2D\_MRLOW). В качестве механизма для обработки данных значений было выбрано удалить данные значения из выборки (пропущенных значений меньше 5%, они тяжело заполняются стандартными методами обработки пропусков).

**Разведывательный анализ.**

****

В первую очередь было исследовано распределение прогнозируемых в будущем величин, т. к. это прямо влияет на анализ, который будет проводится в дальнейшем. Видно, что у всех параметров есть сильная скошенность распределения вправо, что может говорить о необходимости к логарифмированию данных (что будет использоваться в дальнейшем);

В целом распределение величин похожи друг на друга (это логично исходя из структуры данных, а именно из-за линейной зависимости SI от двух других). Учитывая это, SI впитала в себя наибольшую скошенность из двух других переменных.



Рассматривая выбросы в данных, можно заметить, что они присутствуют во всех наблюдаемых переменных. При этом SI имеет наиболее сильные выбросы. Это стоит также учитывать при прогнозировании в будущем.

Рассмотрим также признаки, которые наиболее сильно коррелируют с SI. Среди них можно выделить: Топ 20 признаков по абсолютной корреляции с SI:

SI 1.000000

BalabanJ 0.164715

fr\_NH2 0.160428

RingCount 0.124835

fr\_Al\_COO 0.102374

fr\_COO2 0.101075

fr\_COO 0.101075

NumAromaticRings 0.088006

FpDensityMorgan1 0.087894

VSA\_EState4 0.087770

VSA\_EState6 0.083298

NHOHCount 0.079056

fr\_benzene 0.078621

NumAromaticCarbocycles 0.078621

SMR\_VSA7 0.076668

MolLogP 0.075862

AvgIpc 0.075706

NumAliphaticHeterocycles 0.075564

BCUT2D\_LOGPHI 0.074508

EState\_VSA2 0.071871

Более подробно корреляция будет рассматриваться при подборе признаков для различных задач, которые будут поставлены перед данным исследованием.

Также в ходе EDA были сформированы диаграммы попарных сравнений основных переменных, однако они не дали ничего интересного для будущего анализа

**Общие данные об анализе через регрессию**

В качестве моделей для регрессии были реализованы следующие модели:

* 1. RandomForestRegressor;
  2. GradientBoostingRegressor;
  3. XGBRegressor.

Выбор моделей происходил с учетом того, что в данные скошены (следовательно, линейные модели будут плохо работать для этих данных). В результате выбор пал на модели по типу деревьев.

В качестве выборов признаков для всех моделей была предпринята попытка выбрать наиболее качественные признаки исходя из модели GBR, которая возвращала feature importance и выбирала признаки выше медианы по важности. Это позволяло отобрать наиболее качественные признаки для модели.

В некоторых случаях использовались дополнительные параметры, добавленные экспертным путем. Об этих параметрах будет сказано в соответствующих разделах.

Также исследовались возможности выбора параметров через RandomForestRegressor, однако он показывал худшие результаты и от него было принято решение отказаться.

**Регрессия по IC50**

По данной переменной не было отобрано других признаков кроме тех, что отобраны GBR. Всего отобрано признаков 105 признаков, отобран 81 признак после удаления мультиколлинеарности. Ввиду их большого количества отображать их неудобно.

Модель случайного леса обученная с использованием GridSearchCV с параметрами {

    'n\_estimators': [100, 200, 300],

    'max\_depth': [5, 10, 15, None],

    'min\_samples\_split': [2, 5, 10],

    'min\_samples\_leaf': [1, 2, 4],

    'max\_features': ['sqrt', 'log2']

}

имела следующие результаты:

Параметры: {'max\_depth': 15, 'max\_features': 'log2', 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 200}

MAE на тесте: 205.232

R² на тесте: 0.451

Модель GBR обученная на GridSearchCV с параметрами {

'n\_estimators': [50, 150, 200],

'learning\_rate': [0.1, 0.15],

'max\_depth': [5, 10],

} имела следующие результаты:

Параметры: {'learning\_rate': 0.1, 'max\_depth': 5, 'n\_estimators': 50}

MAE на тесте: 205.887

R² на тесте: 0.455

В итоге в сравнении видно, что наиболее хорошей оказалась модель градиентного спуска (т. к. МАЕ ниже, R^2 выше)

**Регрессия по СС50**

Регрессия по CC50 имела в себе некоторые теоретические признаки, которые могут быть важны для дальнейшего анализа. В ходе добавления этих признаков удалось увеличить R^2 на 2%. Среди признаков есть признаки, отвечающие за токсичность соединения ('MolLogP', 'TPSA', 'NumHDonors', 'NumHAcceptors', 'fr\_halogen'), а также молекулярные дескрипторы ('BCUT2D\_CHGLO', 'BCUT2D\_LOGPLOW', 'VSA\_EState4'). В данной модели были обучены обычный градиентный бустинг, а также XGB. Как и ожидалось, лучше оказался XGB:

Лучшие параметры: {'colsample\_bytree': 0.9, 'gamma': 0, 'learning\_rate': 0.1, 'max\_depth': 5, 'n\_estimators': 50, 'reg\_alpha': 0, 'subsample': 0.9}

MAE на тестовых данных: 309.4767383383203

R2 на тестовых данных: 0.4250237642392869

**Регрессия по SI**

Ввиду того, что в SI наибольшее количество выбросов, а распределение сильнее всего скошено для него потребовалось больше предобработки. Во-первых были исключены 1 процентиль с каждой стороны. Для борьбы со скошенность распределения переменная была прологарифмирована перед поиском необходимых показателей и обучением модели. Также были подобраны экспертные параметры: 'MolLogP', 'TPSA', 'NumHDonors', 'NumHAcceptors', 'fr\_halogen', 'qed', 'FractionCSP3', 'SPS'.

Наилучшей оказалась модель XGB с параметрами:

Лучшие параметры: {'colsample\_bytree': 0.9, 'gamma': 0.1, 'learning\_rate': 0.05, 'max\_depth': 5, 'n\_estimators': 50, 'reg\_alpha': 0.1, 'subsample': 0.9}

MAE на тестовых данных: 0.8203366997961472

R2 на тестовых данных: 0.34268901586239153

Также была предпринята попытка расчитать ансамль из 2 прошлых моделей, но это не привело к результатам.

**Классификация: превышает ли значение IC50 медианное значение выборки**

В задачах классификации использовался RandomForestClassifier для поиска необходимых данных. Также не было исключение признаков по мультиколлинеарности. Для RandomForest в Pipeline были добавлены StandatScaler для нормализации данных и обучение проводилось с использованием GridSearchCV. В качестве метрики оптимизации использовалась f1-score (ввиду неизвестности, что необходимо больше: безопасность лекарства или его эффективность). Для первой классификации получились следующие значения:

Best parameters: {'colsample\_bytree': 0.8, 'gamma': 0, 'learning\_rate': 0.05, 'max\_depth': 5, 'n\_estimators': 50, 'reg\_alpha': 0, 'subsample': 0.8}

Accuracy 0.73, f1-score (0.72, 0.74)

**Классификация: превышает ли значение CC50 медианное значение выборки**

В данной классификации, как и в регрессии были использованы специальные переменные (['MolLogP', 'TPSA', 'fr\_halogen', 'NumHDonors', 'BCUT2D\_CHGLO’). Лучшим оказался XGB с параметрами {'colsample\_bytree': 0.9, 'gamma': 0, 'learning\_rate': 0.1, 'max\_depth': 5, 'n\_estimators': 100, 'reg\_alpha': 0, 'subsample': 0.8}, accuracy 0.79, f1 (0.77, 0.8)

**Классификация: превышает ли значение SI медианное значение выборки**

Как и в регрессии здесь использовалось ограничение выборки и логарифмирование, однако хорошего качества модели не удалось добиться. XGB показал следующие результаты: f1-score (0.67, 0.67), accuracy (0.67)

**Классификация: превышает ли значение SI значение 8**

Этот классификатор не отличается от предыдущего. Однако в данной модели лучшей оказалась модель леса. Результаты: accuracy – 0.68, f1 (0.75, 0.56). Очевидно, что второй класс прогнозируется много хуже, чем первый, что связанно с дисбалансом классов. Для борьбы с дисбалансом была предпринята стратификация и другие алгоритмы внутри моделей.