Аналитический отчет к курсовой работе по классическому машинному обучению

Цель работы – создание нескольких максимально эффективных моделей для решения задач:

- 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

Исходные данные: данные о 1000 химических соединений с указанием их эффективности против вируса гриппа. Параметры, характеризующие эффективность, обозначаются как IC50, CC50 и SI. Значение SI рассчитывается на основе параметров IC50 и CC50. Все остальные представленные признаки являются числовыми характеристиками химических соединений.

IC50 – это концентрация соединения (в миллимолях), требуемая для подавления вирусной активности на 50%

CC50 – это цитотоксичность (концентрация соединения (в миллимолях), вызывающая гибель 50% клеток)

SI – это индекс селективности, рассчитываемый как отношение CC50 к IC50 (чем выше значение, тем более селективен препарат). Значение SI больше 8 считается хорошим показателем, т.е. препарат эффективен против вирусной инфекции.

Анализ представленных данных (EDA)

Использовано логарифмическое преобразование целевых переменных для приближения распределения к нормальному, поскольку изначально имелась правосторонняя асимметрия.

Пропусков в данных незначительное количество. Их было решено заполнить медианным значением.

Анализ выбросов целевых переменных произведен по методу межквартильного размаха. Для сравнения был использован метод Z-оценки на логарифмированных данных. Методы показали разное количество выбросов. На логарифмированных данных получили маленький процент выбросов по сравнению с результатами по методу межквартильного размаха.

IC50 и CC50 имеют умеренную корреляцию (0.521). При этом по отдельности признаки имеют слабую корреляцию с SI, что логично и объясняется тем, что SI может не иметь однозначного соответствия с IC50 и CC50.

При анализе матрицы корреляций обнаружены признаки, имеющие между собой сильную корреляцию. Такие признаки обозначают связанные между собой параметры химических соединений. Например, молекулярные свойства: MolWt (молекулярный вес), MolMR (мера молекулярного объема и поляризуемости), LabuteASA (площадь доступной растворителю поверхности), NumRotatableBonds (мера гибкости молекулы, ротируемые связи).

Присутствуют признаки с нулевой вариативностью, т.е. имеющие постоянные значения. Такие признаки могут быть удалены.

Решение задачи регрессии IC50

Для задачи регрессии были использованы модели Linear Regression, Ridge Regression, Lasso Regression, Random Forest, XGBoost, LightGBM.

Данные были подготовлены следующим образом: логарифмическое преобразование целевой переменной, удаление признаков с константными значениями, пропущенные значения заполнены медианой.

Наилучшие результаты показала модель случайного леса из 200 деревьев с максимальной глубиной 10 и порогом для разделения узла, равным 5.

Модель объясняет около 45% дисперсии данных ($R^2 = 0.45$).

RMSE = 1.45.

Топ-5 наиболее важных признаков для предсказания IC50: VSA_EState8, VSA_EState4, VSA_EState6, BCUT2D_MRLOW, PEOE_VSA1.

Решение задачи регрессии CC50

Для задачи регрессии были использованы модели Linear Regression, Ridge Regression, Lasso Regression, Random Forest, XGBoost, LightGBM.

Данные были подготовлены следующим образом: логарифмическое преобразование целевой переменной, удаление признаков с константными значениями, пропущенные значения заполнены медианой.

Наилучшие результаты показала модель случайного леса из 100 деревьев с максимальной глубиной 10 и порогом для разделения узла, равным 5.

Модель объясняет около 43% дисперсии данных ($R^2 = 0.43$).

RMSE = 1.14.

Топ-5 наиболее важных признаков для предсказания IC50: BCUT2D_MWLOW, NHOHCount, Kappa1, VSA_EState8, VSA_EState6.

Решение задачи регрессии SI

Для задачи регрессии были использованы модели Linear Regression, Ridge Regression, Lasso Regression, Random Forest, XGBoost, LightGBM.

Данные были подготовлены следующим образом: логарифмическое преобразование целевой переменной, удаление признаков с константными значениями, пропущенные значения заполнены медианой.

Наилучшие результаты показала модель случайного леса из 100 деревьев с максимальной глубиной 10 и порогом для разделения узла, равным 5.

Модель объясняет около 34% дисперсии данных ($R^2 = 0.34$).

RMSE = 1.26.

Топ-5 наиболее важных признаков для предсказания IC50: VSA_Estate6, VSA_Estate8, SMR_VSA7, BCUT2D_CHGLO, BCUT2D_MRLOW.

Решение задачи классификации IC50 > медианы

Для задачи регрессии были использованы модели Logistic Regression, Random Forest, XGBoost, LightGBM.

Данные были подготовлены следующим образом: вычислено значение медианы (Медиана IC50: 46.59), создана целевая бинарной переменной (0 – если IC50 ≤ медианы, 1 – если IC50 > медианы), удалены признаки с константными значениями, пропущенные значения заполнены медианой, проверен баланс классов.

Наилучшие результаты показала модель случайного леса из 100 деревьев с максимальной глубиной 10 и порогом для разделения узла, равным 5.

Модель имеет точность (accuracy) 73% (accuracy=0.73). F1-score=0.73. ROC-AUC представлен на рисунке ниже.

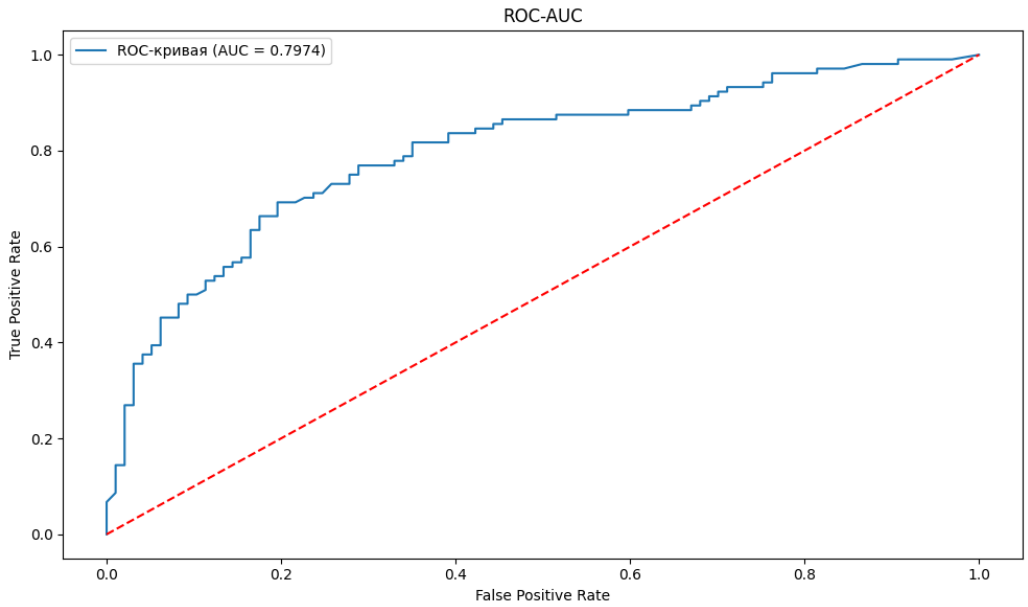

Рисунок 1 - ROC-AUC для классификации IC50 > медианы

Топ-5 наиболее важных признаков для предсказания IC50: VSA_Estate8, BCUT2D_MRLOW, SlogP_VSA5, Kappa3, PEOE_VSA7.

Решение задачи классификации CC50 > медианы

Для задачи регрессии были использованы модели Logistic Regression, Random Forest, XGBoost, LightGBM.

Данные были подготовлены следующим образом: вычислено значение медианы (Медиана CC50: 411.04), создана целевая бинарной переменной (0 – если CC50 ≤ медианы, 1 – если CC50 > медианы), удалены признаки с константными значениями, пропущенные значения заполнены медианой, проверен баланс классов.

Наилучшие результаты показала модель случайного леса из 100 деревьев с максимальной глубиной 10 и порогом для разделения узла, равным 5.

Модель имеет точность (accuracy) 79% (accuracy=0.79). F1-score=0.79. ROC-AUC представлен на рисунке ниже.

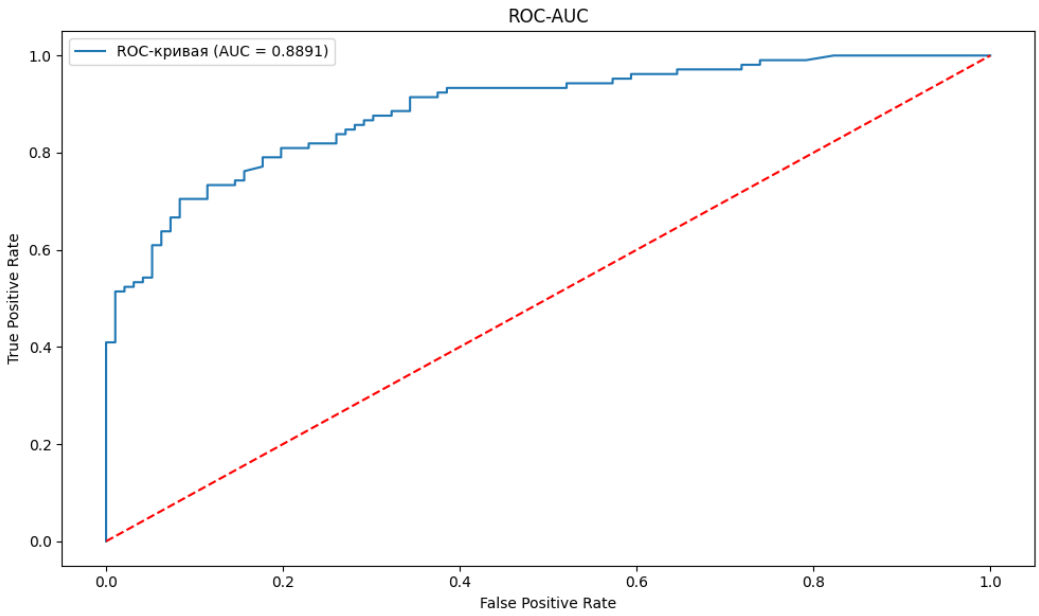

Рисунок 2 - ROC-AUC для классификации CC50 > медианы

Топ-5 наиболее важных признаков для предсказания CC50: NHOHCount, PEOE_VSA7, VSA_EState4, SMR_VSA5, BCUT2D_MWLOW.

Решение задачи классификации SI > медианы

Для задачи регрессии были использованы модели Logistic Regression, Random Forest, XGBoost, LightGBM.

Данные были подготовлены следующим образом: вычислено значение медианы (Медиана SI: 3.85), создана целевая бинарной переменной (0 – если SI ≤ медианы, 1 – если SI > медианы), удалены признаки с константными значениями, пропущенные значения заполнены медианой, проверен баланс классов.

Наилучшие результаты показала модель случайного леса из 50 деревьев с максимальной глубиной 10 и порогом для разделения узла, равным 5.

Модель имеет точность (accuracy) 67% (accuracy=0.67). F1-score=0.64. ROC-AUC представлен на рисунке ниже.

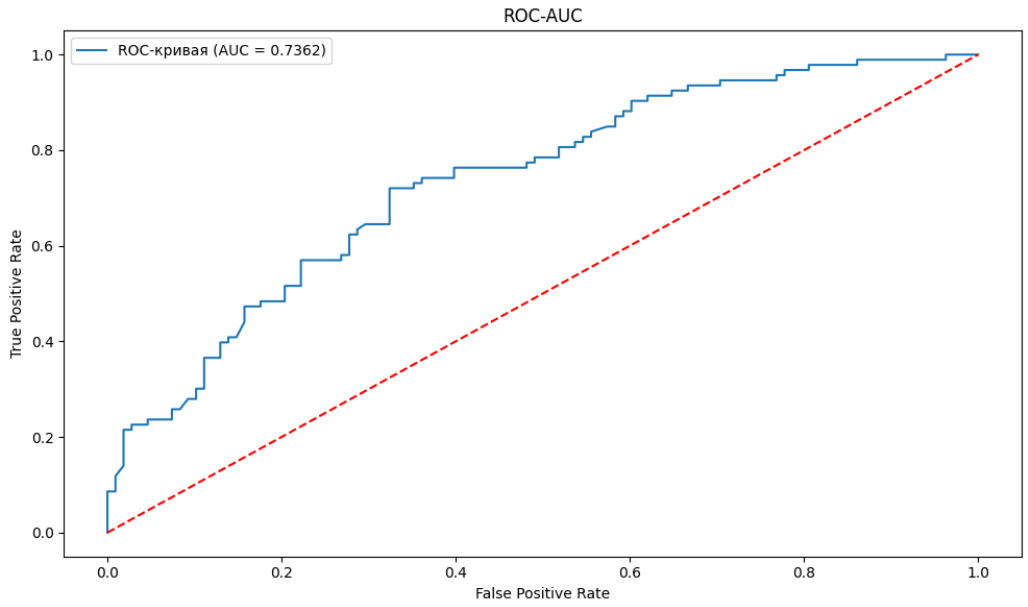

Рисунок 3 - ROC-AUC для классификации SI > медианы

Топ-5 наиболее важных признаков для предсказания SI: BCUT2D_MRLOW, BCUT2D_MWLOW, VSA_EState4, BCUT2D_LOGPHI, MinEStateIndex.

Решение задачи классификации SI > 8

Для задачи регрессии были использованы модели Logistic Regression, Random Forest, XGBoost, LightGBM.

Данные были подготовлены следующим образом: создана целевая бинарной переменной (0 – если SI ≤ 8, 1 – если SI > 8), удалены признаки с константными значениями, пропущенные значения заполнены медианой, проверен баланс классов.

Наилучшие результаты показала модель XGBoost из 50 деревьев с максимальной глубиной 5 и скоростью обучения 0.1.

Модель имеет точность (accuracy) 72% (accuracy=0.72). F1-score=0.53. ROC-AUC представлен на рисунке ниже.

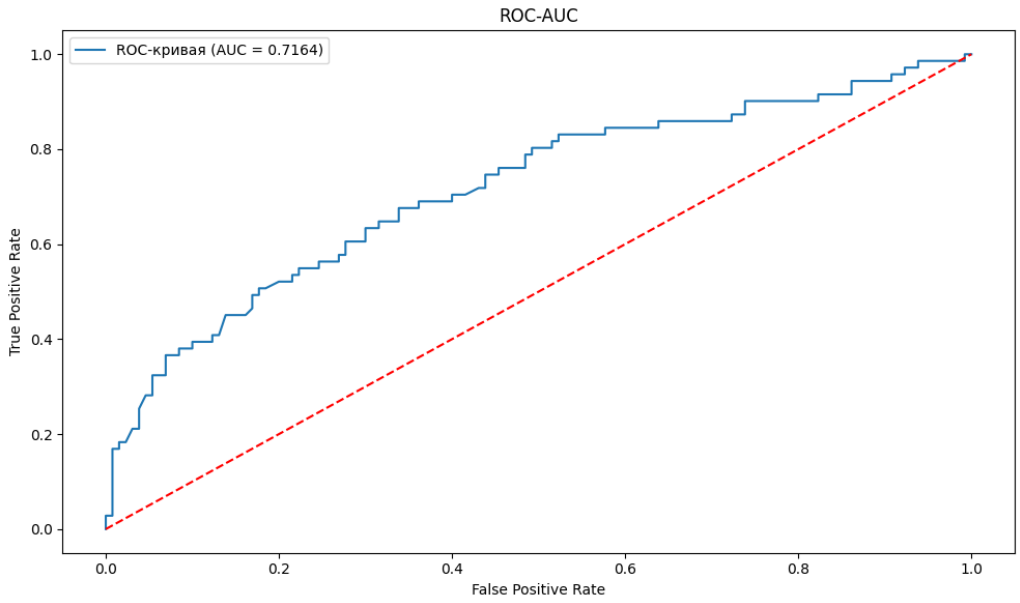

Рисунок 3 - ROC-AUC для классификации SI > 8

Топ-5 наиболее важных признаков для предсказания SI: SMR_VSA7, fr_allylic_oxid, SMR_VSA4, BCUT2D_CHGLO, fr_ether.

Выводы

В ходе работы были изучены предоставленные данные о химических соединениях, были построены модели для регрессии и классификации параметров эффективности: IC50, CC50 и SI.

Наилучшие результаты показали модели случайного леса и XGBoost. Оптимальные гиперпараметры подбирались для каждой задачи отдельно.

Для дальнейшего продолжения исследования и улучшения результатов предлагаются следующие рекомендации:

- Создание новых признаков, объединяющих связанные между собой признаки, чтобы подобрать наиболее подходящие сочетания параметров для разработки лекарственных средств;
- Снижение размерности с помощью метода главных компонент;
- Проведение более тщательного анализа выбросов и аномальных значений;
- Тестирование других моделей с подбором гиперпараметров.