Национальный исследовательский ядерный университет
«МИФИ»

Классическое машинное обучение

Курсовая работа (vo_PJ)

Исследование лекарственной активности

Студент: Лобас Фанис Станиславович

Введение

Цель работы

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

Задачи исследования

1. Провести исследовательский анализ данных (EDA) и оценить информативность признаков.
2. Построить модели машинного обучения:

- Регрессия:
  - Прогноз значения IC50
  - Прогноз значения CC50
  - Прогноз значения SI
- Классификация:
  - Бинарный прогноз: превышает ли IC50 медианное значение
  - Бинарный прогноз: превышает ли CC50 медианное значение
  - Бинарный прогноз: превышает ли SI медианное значение
  - Бинарный прогноз: превышает ли SI значение 8

3. Выполнить сравнительный анализ качества моделей по метрикам:
4. Выбрать наиболее эффективные модели и обосновать выбор.
5. (Предложить рекомендации по использованию финальной модели в практической работе.)

Глава 1

Предобработка данных

Описание датасета

Датасет представляет собой таблицу, содержащую данные по 1001 химическому соединению. Каждая строка соответствует одному веществу, столбцы — его физикохимическим признакам и биологической активности. Число колонок – 214: 107 колонок с типом float64 и 107 колонок с типом int64.

Выделим целевые и контрольные переменные:

- Колонка IC50, mM - Концентрация ингибитора, при которой подавляется 50% активности (мера активности соединения).
- Колонка CC50, mM - Концентрация, вызывающая 50% токсичности (мера токсичности).
- Колонка SI - Selectivity Index = CC50 / IC50 (чем выше, тем лучше: высокая активность и низкая токсичность).

Остальные колонки - это признаки, которые описывают структурные, физико-химические и молекулярные свойства соединений.

Разведочный анализ

1. Удалим колонку Unnamed: 0, т.к. значения – порядковые номера, не влияющие на целевые переменные;
2. Имеются пропуски:

MinAbsPartialCharge	3
MaxPartialCharge	3
MinPartialCharge	3
MaxAbsPartialCharge	3
BCUT2D_MWHI	3
BCUT2D_MRLOW	3
BCUT2D_MRHI	3
BCUT2D_LOGPLOW	3
BCUT2D_LOGPHI	3
BCUT2D_CHGLO	3
BCUT2D_CHGHI	3
BCUT2D_MWLOW	3

заполняем пропуски средним для него значением;

3. Удалим колонки с нулевыми значениями. Таких колонок – 18.

	NumRadicalElectrons	SMR_VSA8	SlogP_VSA9	fr_N_O	fr_SH	fr_azide	fr_barbitur	fr_benzodiazepine	fr_diazo	fr_dihydropyridine	fr_isocyan	fr_isothiocyan	fr_lactam	fr_nitroso	fr_phos_acid	fr_phos_ester	fr_prisulfonamd	fr_thiocyan	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4. Удалим дубликаты. Осталось 969 rows × 213 columns
5. Удаляем строки, где значение "IC50, mM", "CC50, mM" и "SI" выше 98 перцентиля
6. Сохраним результат в файл Обработанный.csv

Глава 2

Решение задачи регрессии IC50

Цель данного этапа – найти максимально эффективную модель, способную предсказывать значения **IC50.**

Выбраны пять моделей, обычную линейную регрессию, регуляризацию Ridge и Lasso, RandomForest, XGBoost и их разные гиперпараметры для подбора. Для каждой модели были подобраны лучшие гиперпараметры с помощью GridSearch.
Модели линейной регрессии, регуляризации Ridge и Lasso показывают низкое качество прогноза, особенно по метрике R^2. Эти модели недостаточно подходят для данной задачи без дополнительной работы над признаками или преобразования данных, т.к. они плохо работают с нелинейными признаками.
Исследование показало, что наилучшее качество прогнозирования IC50 достигается при использовании модели RandomForest, XGBoost с гиперпараметрами. Полученная модели дают $R^2 > 0.56$, что является наилучшим результатом среди всех исследованных моделей и методов настройки.
Решение задачи регрессии CC50 выполняется аналогично решению задачи регрессии IC50, где результаты почти не отличаются.
Решение задачи регрессии SI выполняется аналогично решению задачи регрессии IC50, но результаты отличаются, где модели RandomForest, XGBoost дают несколько худшие показатели R^2.

Глава 3
Решение задачи классификации IC50
Цель данного этапа – превышает ли значение **IC50** медианное значение выборки.
Выбраны четыре модели, логистическую регрессию, RandomForest, XGBoost, KNN и их разные гиперпараметры для подбора. Провёл перебор гиперпараметров для нескольких моделей, сохранил лучшие параметры, а также метрики accuracy и ROC AUC для каждой модели. В итоге сформировал таблицу с результатами для удобного сравнения.
Решение задачи классификации CC50 выполняется аналогично решению задачи классификации IC50, но результаты отличаются, где модели RandomForest, XGBoost дают несколько лучшиие показатели по ROC AUC.
Решение задачи классификации SI выполняется аналогично решению задачи классификации IC50, удалив все целевые признаки. Результаты по ROC AUC хуже, но уже модель RandomForest здесь является лучшей из выбранных.
Решение задачи классификации SI>8 выполняется аналогично решению задачи классификации IC50, удалив все целевые признаки. Результаты по ROC AUC хуже, но уже модель XGBoost здесь является лучшей из выбранных.

Вывод

XGBoost — лучшая и стабильная модель с хорошим балансом между точностью и полнотой.

Random Forest показывает худший результат, хотя и остаётся приемлемыми для начального анализа.

Линейные модели такие как LinearRegression, Ridge, Lasso не работают с нелинейными признаками.

В задачах классификации можно использовать ансамбль моделей.

Рекомендуется логарифмировать все целевые значения перед обучения, чтобы стабилизировать обучение и повысить точность.