Проект: Предсказание знака изменения свободной энергии связывания (ddG) на основе данных SKEMPI 2.0

Цель проекта: Реализовать модель машинного обучения, предсказывающую знак изменения свободной энергии связывания (ddG) на основе данных из базы SKEMPI 2.0.

* Задачи проекта:
  1. Анализ датасета, в частности:
     + Ресерч по теме с целью выделения ключевых признаков, влияющих на знак изменения свободной энергии связывания
     + Разведочный анализ датасета – оценка внутренней структуры датасета, очищение от nan значений, удаление незначащих категорий
     + Приведение датасета в вид, который впоследствие можно передать модели
  2. Выбор подходящей модели исходя из датасета
  3. Оценка работы модели

Описание решения:

1. Прежде всего для работы с датасетом необходимо разобраться в том, какие из данных являются ключевыми для определения знака свободной энергии связывания. Исходя из проведенного ресерча можно заключить, что такими являются данные из следующих столбцов:
   1. Mutation(s)\_cleaned – мутации в фомате TP11N, где T – исходная аминокислота, P – цепочка аминокислот, в которой произошла мутрация, 11 – условная координата мутации, N – результирующая аминокислота
   2. iMutation\_Location(s) -- Расположение мутаций (мутаций) в месте связывания или вдали от него
   3. Hold\_out\_type -- Удерживающий тип. Некоторые комплексы классифицируются как протеаза-ингибитор (Pr/PI), антитело-антиген (AB/AG) или pMHC-TCR (TCR/pMHC).
   4. Affinity\_mut\_parsed – необходим для подсчета ddG
   5. Affinity\_wt\_parsed– необходим для подсчета ddG
   6. Protein 1 – участвующий в мутации белок
   7. Protein 2 – участвующий в мутации белок
   8. Temperature – необходим для подсчета ddG
2. После был проведен разведочный анализ – удалены строки, в которых отсутствовали значения Affinity\_wt\_parsed, Affinity\_mut\_parsed, Temperature; Так как в ходе соединения цепочек амонокислот может происходить несколько мутаций, то было подсчитано количество всех возможных мутаций и Hold\_out\_type в ходе каждого соединения. В случает, если определенное количество было представлено незначительно в датасете, то записи с ними удалялись. Далее было проведено разделение мутаций в формат T – исходная аминокислота, P – цепочка аминокислот, в которой произошла мутрация, 11 – условная координата мутации, N – результирующая аминокислота для каждой мутации внутри связи для дальнейшей передачи в модель. После был произведено вычисление знака ddg и оценка сбалансированности таргета.

Так как большая часть признаков категориальная, размер датасета небольшой (7к значений) и решаемая задача – задача бинарной классификации, то была выбрана модель на основе градиентного бустинга, так как она лучше всего в таких условиях справляется с поставленной задачей. В частности была выбрана модель catboost от яндекса, так как она является лучшим представителем таких моделей, особенно в случае категориальных данных. Его быстрее обучать и проще настроить, поэтому как бейзлайн он очень хорошо подходит. Для подготовки тестовой выборки было решено использовать кросс-валидацию, так как она хорошо работает в условиях ограниченной выборки. В ходе обучения модель (с перебором гиперпараметров depth и learning\_rate) отработала следующим образом:

Mean CV F1: 0.8020 ± 0.0114

=== Grid Search Results ===

lr=0.03, depth=4: F1 = 0.7903 ± 0.0035

lr=0.03, depth=6: F1 = 0.7948 ± 0.0060

lr=0.03, depth=8: F1 = 0.7948 ± 0.0049

lr=0.05, depth=4: F1 = 0.7949 ± 0.0075

lr=0.05, depth=6: F1 = 0.7960 ± 0.0076

lr=0.05, depth=8: F1 = 0.7998 ± 0.0085

lr=0.1, depth=4: F1 = 0.7996 ± 0.0027

lr=0.1, depth=6: F1 = 0.7980 ± 0.0127

lr=0.1, depth=8: F1 = 0.8014 ± 0.0123

Best parameters: {'learning\_rate': 0.1, 'depth': 8} (F1 = 0.8014)

Для определения качества работы была использована F1-score, так как он подходит для работы в условиях несбалансированной выборки. По нему видно, что предсказание модели удовлетворительно и что её точности как минимум выше случайной.

В ходе работы была определена важность признаков:  
Hold\_out\_type: 5.88

Protein 1: 13.74

Protein 2: 8.70

iMutation\_Location(s)\_1: 13.86

iMutation\_Location(s)\_2: 2.04

orig\_aa\_1: 17.18

chain\_1: 7.52

residue\_num\_1: 15.26

mut\_aa\_1: 11.22

orig\_aa\_2: 1.00

chain\_2: 1.11

residue\_num\_2: 0.64

mut\_aa\_2: 1.87

Из чего можно заключить, что основной вклад в предсказание вносят следующие признаки -- iMutation\_Location(s), Protein 1, orig\_aa, mut\_aa, то есть область мутации, мутирующие аминокислоты и участвующие белки. Это может быть использовано для переработки датасета.