Digital Reputation Challenge

Письменный Алексей
МГУ ВМК ММП

22 октября 2019 г.

Первая идея

- По предположению, числа в X2 соответствуют адресам, которые посещал пользователь
- Для каждого id из тестовой выборки посмотрим, сколько «общих сайтов» у него с каждым id из обучающей выборки
- Для каждого id из тестовой выборки усредним Y с соответствующими весами и получим ответ
- Результат: pub: 0.595662 priv: 0.59579

Дальнейшее развитие

- Вес был равен числу «общих сайтов», то есть «общий сайт» соответствовал числу 1
- Заменим для i-го сайта 1 на $\log(D/d_i)$, где D - общее число посещенных сайтов в train (повторы учитываются), d_i - число посещений конкретного сайта
- pub: 0.597172 (+0.0015) priv: 0.599857 (+0.0041)

Дальнейшее развитие

- Усилим влияние редких сайтов
- Веса распределены примерно с 6.5 до 13, вычтем из каждого 2

  pub: 0.59737 (+0.0002) priv: 0.600342 (+0.0005)

Дальнейшее развитие

- Вычтем 4:

  pub: 0.597546 (+0.00018) priv: 0.601108 (+0.00077)

- Вычтем 6:

  pub: 0.597693 (+0.00015) priv: **0.602368** (+0.00126)

- Приступим к обучению

XGBoost

- Есть XGBoost, обученный на X1, и первый метод
- Каждый столбец (из 5) - результат алгоритма с лучшим результатом на public

  pub: 0.61306 (+0.0153) priv: 0.603183 (+0.0008)
- XGBoost показывает лучший результат на public для 1-й и 3-й целевой переменной
- Теперь видна разница

Ансамблирование

- Переводим значения в ранги и считаем усредненный ранг для каждого id
- Вес: 0.7 для лучшей модели на public и 0.3 для второй (возможно неверная стратегия)
- По 5-му столбцу ансамблирования нет в итоговых посылках (а зря)
- Итоговый результат: pub: 0.617597 (+0.0045) priv: **0.611575** (+0.0084), 22-е место