Условие

Есть датасет. Собран он следующим образом:

Из логов были взяты пары ([запрос] - [объект на который кликнул пользователь]). Это положительные примеры (метка 1 в датасете) Для каждого запроса был подобран в пару негативный объект (метка 0) следующим образом: определяем к какой рубрике относится положительный пример; выбираем случайный объект из другой рубрики. Идея в том, что этот пример маловероятно будет релевантным. На этих данных, используя кросс-валидацию, обучались различные модели. Метрики качества были хорошими. При попытки тестирования на реальных данных, качество моделей сильно уступало тестовым метрикам.

Задача

Выявить особенности датасета, которые приводили к данным результатам и объяснить почему так происходило.

Исследование

Первичный анализ и подготовка данных

- Сперва проверил данные на наличие пропусков, распределение целевой переменной.
- Далее привел к более оптимальным типам данных (запрос и рубрика в float32, метки в float8).
- Посмотрел на распределение рубрик. Некоторые рубрики встречаются намного чаще.

Кросс-валидация. Проверка моделей классификации.

- Применил кросс-валидацию для временных рядов.
- Построил модели классификации XGBoost, LightGBM, CatBoost, RandomForest.
- Каждый из них даже без оптимизации гиперпараметров выдаёт неплохие результаты.
- Сначала взял срез датасета в 1 млн строк для первичной оценки. Результаты метрик были на хорошем уровне.
- Далее произвел обучение и тестирование на полном датасете (для бустингов). Оценка почти не изменялась.

Выявление особенностей.

- Первой особенностью датасета является неверное распределение таргета. Для каждого запроса будет 1 положительный пример, но несколько отрицательных. Их распределение не будет 50% на 50%. При поиске чего-то либо всегда выдается список примеров, которые в большей степени будут отрицательной меткой. Поэтому я сгенерировал примерный датасет в котором на каждую положительную метку будет 9 отрицательных. Поэтому модель на реальных данных могла выдавать плохие результаты.

- Второй особенностью датасета является выбор случайной рубрики. Как негативный пример в датасет может попасться как совсем нерелевантная рубрика, так и очень близкой к релевантной. Например, если мы ищем "Ресторан", то случайно выбранной отрицательной меткой в датасет может записаться "Автосервис", а может "Бар". Второй случай, возможно, будет релевантен нашему запросу, но мы все равно поставили ему отрицательную метку.