Для начала опишем стандартный пайплайн онлайн-рекламы. Наша платформа совершает показ рекламы -> Пользователь может кликнуть на рекламу, в этом случае он переходит на сайт рекламодателя -> На сайте рекламодателя пользователь может совершить конверсию. Такая конверсия называется post-click конверсией. Возможен иной вариант:
Наша платформа совершает показ рекламы -> Пользователь не кликает по рекламе -> Но спустя время совершает конверсию на сайте рекламодателя. Такая конверсия называется post-view конверсией.

Вам предоставлен датасет train_views.parquet, состоящий из показов рекламы различным пользователям. Датасет содержит набор признаков, которые могут повлиять на то, совершит ли пользователь конверсию или нет. Описание признаков находится в таблице features_descr.xlsx. Также имеется датасет train_actions.parquet, в котором перечислены действия, которые пользователи совершили после показа рекламы. Действия следующие: клик, post-view конверсия, либо post-click конверсия. Описание колонок также приложено в features_descr.xlsx. Ключом, по которому можно соединять таблицы train_views.parquet и train_actions.parquet является колонка ssp_event_id. В шаблоне решения baseline.py показано, как можно соединить эти датасеты и получить метки ответов. Файл features_descr.xlsx также содержит описание всевозможных конверсий (см. conversion_name), которые могут совершить пользователи на сайте заказчика. Пользователь может совершить несколько различных конверсий.

Помимо основных датасетов, также есть датасет third_party_conversions.parquet. В нем собраны third-party конверсии. Third-party конверсии - это конверсии, которые были совершены пользователями со сторонних источников (без показа рекламы при помощи нашей платформы). По таким конверсиям известен меньший набор признаков, чем по конверсиям, совершенным после рекламы на нашей платформе. Зато таких конверсий гораздо больше, чем post-click и post-view конверсий. Этот датасет содержит только 3rd-party конверсии, других событий в нем нет. Набор признаков является подмножеством признаков, доступных в датасете train_views.parquet.

Ваша задача: определить вероятность совершения post-click конверсии с conversion_name = ’cart’ (это будет класс 1). Задача бинарной классификации. Предсказывать post-view конверсии и прочие post-click конверсии не нужно, факт их совершения дан вам в качестве дополнительных данных. В тестовом и валидационном доступны все те же признаки, что и в основном датасете train_views.parquet. Все post-view конверсии в тестовом и валидационном датасетах, а также post-click конверсии с conversion_name != ’cart’, относятся к классу 0. Тренировочный, тестовый и валидационный датасеты разделены по времени в хронологическом порядке.

Также вам предоставлен тестовый датасет, на котором можно оценить качество своей модели. Итоговая проверка качества будет производиться на валидационном датасете.

Технические требования и ограничения:
- Допускается использование только open-source библиотек, коммерческие продукты использовать нельзя.
- Ваш проект должен укладываться в 120 Гб SSD, 16 Gb RAM.
- Длительность выполнения инференса ограничена и составляет максимум 1 час.