

#### Первоначальная очистка данных:

1. Удаляем данные за 15 число, а также те, что выходят за рамки месяца.
2. Столбец `fulldate` содержит даты, которые не совпадают с `ts`, причем они соответствуют дубликатам записей, судя по одинаковым значениям `lsc`, `cid`, `ts` – удаляем.
3. ID из разных подгрупп по условию не должны пересекаться, так как относятся к разным периодам времени и способу идентификации, поэтому найдем пересечение групп ID и посмотрим в какой группе они в основном проявляли активность – как оказалось во второй, поэтому удаляем строки с ними из первой половины месяца.
4. По итогу количество ID оказалось не равным, что говорит о том, что не все пользователи проявляли активность в течении всего месяца, а значит не имеют пары.
5. Также удалим ошибочные записи с эталонными ID попавшие не в свою группу.

#### Подготовка обучающего и предсказательного датасета:

1. Считаем для каждого юзера количество действий по каждой категории фичей `lsc` и `cid`, а также общее количество действий, и долю действий по каждой категории фичей `lsc` и `cid`.
2. Далее у каждого пользователя определяем топ-20 категорий фичей `lsc` и `cid`, проанализировав самых активные эталонные пары можно отметить, что большая часть действий и пересечений находится в этом диапазоне рангов.
3. После анализа взаимосвязи распределения времени активности внутри дня и по дням недели, можно сделать следующие выводы:
  - Дисперсия распределения времени активности внутри дня юзера слабо меняется для эталонных пользователей - но сильно отличается между разными юзерами;
  - Мода активности по дням недели особенно сильно разнится у разных юзеров - можно усилить эффект экспоненцированием и в дальнейшем вычислить дельту.
4. Генерируем все возможные пары ID, которые гарантированно относятся к False классу, путем кросс-джойна эталонный ID одной группы с неизвестными другой и наоборот, конкатенируем, True и False пары, джойним к этим парам метрики из пункта 1,2,3.
5. Для всех этих пар рассчитываем следующие метрики, говорящие о различиях в паттерне поведения юзеров:
  - долю совпадений в топ-20 по использованию категориальных фичей `cid` и `lsc` (в случае если ID соответствуют разным пользователям, метрика будет уменьшаться до нуля, в обратном случае стремиться к единице)
  - дисперсию распределений разности долей использования топ-20 категориальных фичей `cid` и `lsc` (в случае если ID соответствуют разным пользователям, метрика будет расти, в обратном случае стремиться к нулю)
  - модуль доли количества общей активности относительно максимума в паре (в случае если ID соответствуют разным пользователям, метрика будет расти до 1, в обратном случае стремиться к нулю)
  - модуль доли экспоненты моды распределения дней недели активности относительно максимума в паре (в случае если ID соответствуют разным пользователям, метрика будет расти до 1, в обратном случае стремиться к нулю)
  - модуль доли дисперсии распределения часов активности относительно максимума в паре (в случае если ID соответствуют разным пользователям, метрика будет расти до 1, в обратном случае стремиться к нулю).
6. Все старые метрики, относящиеся к одному пользователю удаляем.
7. Создаем аналогично датасет в котором будем предсказывать неизвестные пары, кросс-джойним все одинокие ID и далее по пунктам 1,2,3,5,6.
8. Методом SMOTE балансируем классы в обучающем датасете, путем генерации строк со схожими значениями метрик.

#### Обучение, валидация и предсказание:

1. Обучаем 2 варианта обучающих датасетов – несбалансированный и сбалансированный методом SMOTE.
2. Разбиваем на тренировочный и тестовый датасеты наши обучающие датасеты, в случае несбалансированного датасета используем стратификацию.
3. Используя кросс-валидацию с 4 фолдами рассчитываем метрики F1, точность и полноту, в случае несбалансированного датасета увеличиваем веса True класса в (кол-во False/ кол-во True) раз.
4. Далее настраиваем гипер-параметры и обучаем модель catboost-classifier, которую затем проверяем на тестовом датасете.
5. Проводим обучение на предсказательном датасете – появляются коллизии (один ID состоит сразу в нескольких группах), избавляемся от которых с помощью удаления пар с меньшей вероятностью.