

Sberbank Data Science Journey

Задача А

Белкин Илья, МФТИ, 2 курс
Ник на LB: ivb

План:

- Формулировка задачи
- Решение
- Результат

Формулировка задачи

Напомним, что в задаче А требовалось определить пол пользователей, присутствовавших в таблице “transactions.csv”, но отсутствовавших в таблице “customers_gender_train.csv”.

transactions.csv

	customer_id	tr_datetime	mcc_code	tr_type	amount	term_id
0	39026145	0 10:23:26	4814	1030	-2245.92	NaN
1	39026145	1 10:19:29	6011	7010	56147.89	NaN
2	39026145	1 10:20:56	4829	2330	-56147.89	NaN
3	39026145	1 10:39:54	5499	1010	-1392.47	NaN
4	39026145	2 15:33:42	5499	1010	-920.83	NaN

customers_gender_train.csv

	customer_id	gender
0	75562265	0
1	10928546	1
2	69348468	1

Решение

Агрегируем данные по каждому пользователю. Для каждого тсс подсчитаем сумму, количество транзакций, среднюю сумму транзакции. Сделаем это отдельно по транзакциям с $\text{amount} > 0$ и отдельно с $\text{amount} < 0$.

Решение достаточно протое, но уже оно позволяет добиться неплохого качества.

Результаты

Public rating: 0.851438 (39.3058)

Private rating: 0.884349 (102.7958)

Итоговый результат оказался неожиданно значительно выше.