

REKKO CHALLENGE, 2-е место

Смирнов Евгений

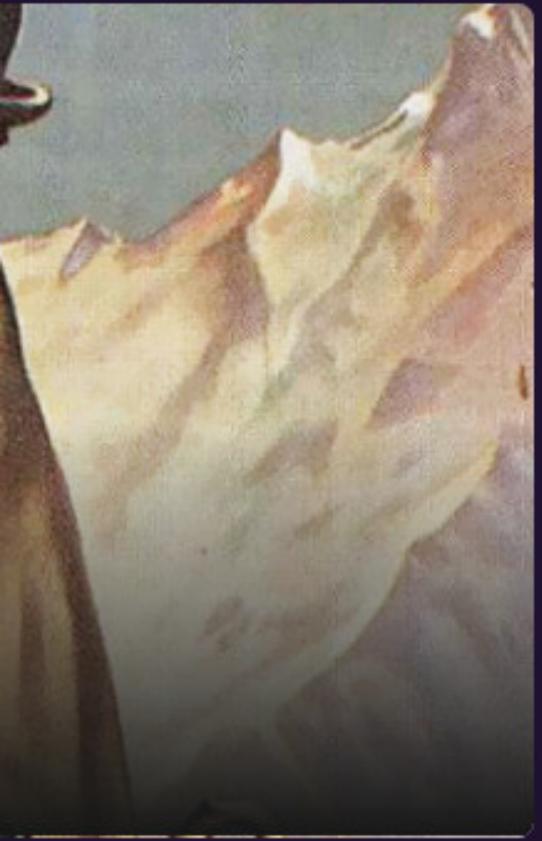


Tinkoff.ru

Обо мне

- Data Science Team Lead
- Магистр МФТИ ФУПМ
- Новичок в соревнованиях по ML

О соревновании



Запомненные
Здесь хранятся
запомненные фильмы
и сериалы

Мои покупки
Здесь хранятся фильмы
и сериалы, которые
вы купили

**История
просмотров**
Здесь хранятся фильмы
и сериалы, которые
вы смотрите

МОИ ФИЛЬМЫ

РЕКОМЕНДАЦИИ

НОВИНКИ

Подписки

КАТАЛОГ



Tinkoff.ru

Описание данных

- Транзакции по покупке/аренде/подписке
- Закладки, рейтинги

element_uid	user_uid	consumption_mode	ts	watched_time	device_type	device_manufacturer
916	503533	S	4.430518e+07	11020	3	99
8494	484023	S	4.430518e+07	5101	0	50
3814	588677	R	4.430517e+07	9014	0	11
1539	252409	S	4.430517e+07	18	5	31
549	264603	S	4.430517e+07	6238	0	50
5951	408050	S	4.430517e+07	20886	0	11
2429	499047	S	4.430517e+07	368576	0	99
4601	23956	S	4.430517e+07	69	5	67
7637	57411	S	4.430517e+07	4971	3	99
3757	574463	S	4.430517e+07	2347	3	99

Описание данных

● Каталог контента

attributes	availability	duration	feature_1	feature_2	feature_3	feature_4	feature_5	type
[31115, 6713, 10906, 31116, 31117, 270, 24431, ...]	[]	80	2.9122e+07	0.57526	0	1.12833	0.654707	movie
[2786, 385, 2799, 3730, 886, 7, 11700, 42, 20, ...]	[purchase, rent]	120	6.61043e+06	0.773224	3	1.11201	0.654707	movie
[31442, 31443, 31444, 31445, 113, 31446, 42, 3...]	[]	80	1.31587e+07	0.699502	0	1.11013	0.68041	movie
[34361, 34362, 23033, 14887, 270, 20089, 43, 25]	[]	20	4.15771e+07	0.702981	0	1.14193	0.654707	series
[26732, 26733, 26734, 9367, 7792, 336, 26735, ...]	[purchase, rent, subscription]	70	3.99958e+07	0.626596	8	1.13008	0.592716	movie
[23108, 23109, 15083, 336, 23110, 123, 42, 43, ...]	[]	100	4.04855e+07	0.693457	0	1.13523	0.68041	movie
[9528, 20806, 20814, 20797, 20815, 270, 20816, ...]	[purchase, rent, subscription]	60	5.89683e+06	0.762507	14	1.12122	0	movie
[20418, 20419, 20420, 20421, 20422, 714, 83, 7...]	[purchase, rent, subscription]	80	2.64084e+07	0.751026	12	1.13523	0.449667	movie
[172, 416, 2839, 5648, 1183, 7, 5683, 32, 11, ...]	[purchase, rent]	110	7.93968e+06	0.699502	7	1.12302	0.592716	movie

Описание данных

- 10м транзакций
- 450к закладок
- 950к рейтингов
- 10к единиц контента
- 500к пользователей

Таргет

Пользователь потребил контент, если он:

- Купил его или взял в аренду
- Посмотрел больше половины фильма
- Посмотрел больше трети сериала

Оценка алгоритма

$$\text{MNAF@20} = \frac{1}{|U|} \sum_{u \in U} \frac{1}{\min(n_u, 20)} \sum_{i=1}^{20} r_u(i) p_u @ i$$

$$p_u @ k = \frac{1}{k} \sum_{i=1}^k r_u(i)$$

$r_u(i)$ – потребил ли пользователь и контент, предсказанный ему на месте i (1 либо 0)

n_u – количество элементов, которые пользователь потребил за тестовый период

U – множество тестовых пользователей

Агрегированный рейтинг

- Доля просмотра фильма * 5
- Доля просмотра сериала * 10
- [Добавление в закладки фильма] * 0.5
- [Добавление в закладки сериала] * 1.5
- [Покупка/аренда контента] * 15
- Рейтинг + 2

Модели первого уровня

- LightFM (0.038 VAL, 0.033 LB)

$$\text{score}(u, c) = \langle q_u, q_c \rangle + b_c + b_u$$

- BM25Recommender (0.035 VAL, 0.03 LB)

$$w_j(\bar{d}, C) := \frac{(k_1 + 1)d_j}{k_1((1 - b) + b\frac{dl}{avdl}) + d_j} \log \frac{N - df_j + 0.5}{df_j + 0.5}$$

Блендинг

LightFM + BM25Recommender (LB 0.0347)

rank	element
0	3567
1	2639
2	4141
3	3947
4	71
5	427
6	4070
7	6377
8	1636
9	9472
10	9491

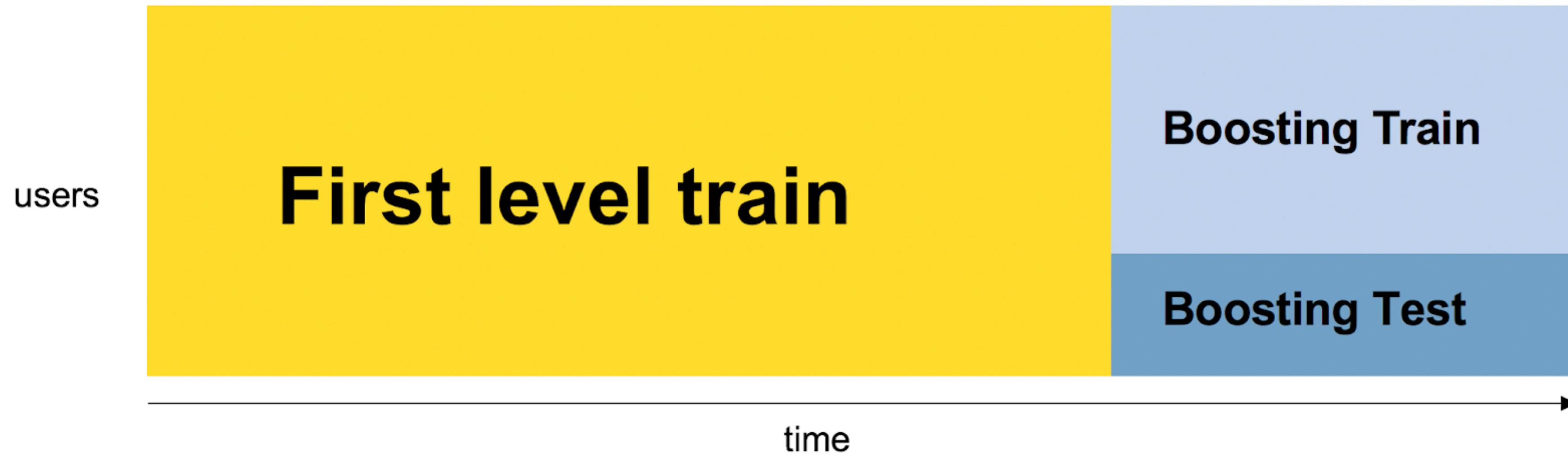
+

rank	element
0	3567
1	4141
2	3947
3	2639
4	71
5	4548
6	9179
7	6377
8	9491
9	427
10	9472

=

rank	element
0	3567
1	4141
2	2639
3	3947
4	71
5	427
6	6377
7	9491
8	9472
9	4548
10	1636

Модель второго уровня



Признаки

- score + rank (bm25+lightfm) + catalog features = 0.0359 LB
- - feature1 = 0.0367 LB
- + (item + user) bias = 0.0388 LB
- + change rankings = 0.0395 LB
- + rare element filter + user embeddings = 0.0429 LB
- + new bookmark = 0.0447 LB
- + start, end, diff time = 0.0457 LB
- + bookmark_cnt, bookmark_percentage = 0.0453 LB
- + usr_buy_cnt = 0.04515 LB

Не зашло

- Атрибуты контента
- Атрибуты устройств
- Эмбеддинги контента
- Доля фильмов/сериалов у пользователя
- Понижение веса популярного контента
- CatBoost

Финальный блендинг

- Квантиль для разбиения на train/test
- Метод агрегации рейтинга
- Обучение с непопулярным контентом

Public Score

Место	Команда	Решений	Результат	Последнее решение
1	Yareg	31	0.0485883	08.04.2019 00:07
2	gdanschin	47	0.0476853	18.04.2019 23:53
3	alexey_g	97	0.0473816	18.04.2019 21:44
4	stason	144	0.0471327	18.04.2019 23:36
5	hype	191	0.0469678	18.04.2019 23:11

Private Score

Место	Команда	Решений	Награда	Скор
1	alexey_g	97	 Золото	0.0481331
2	hype	191	 Золото	0.0478976
3	Yareg	31	 Золото	0.0478931
4	gdanschin	47	 Золото	0.0478442
5	stason	144	 Золото	0.0469455

Интересные факты

- Топ1 решение оказалось хуже продуктовой модели Okko 0.048 vs 0.062
- В соревнования изменили датасет, для старых участников добавили 30 сабмитов
- Валидация не всегда коррелировала с LB

Спасибо!

Смирнов Евгений

e.smirnov6@tinkoff.ru