VK RecSys Challenge

To like or to dislike?

Александр Пославский

Руководитель группы Core ML, AI VK

Задача

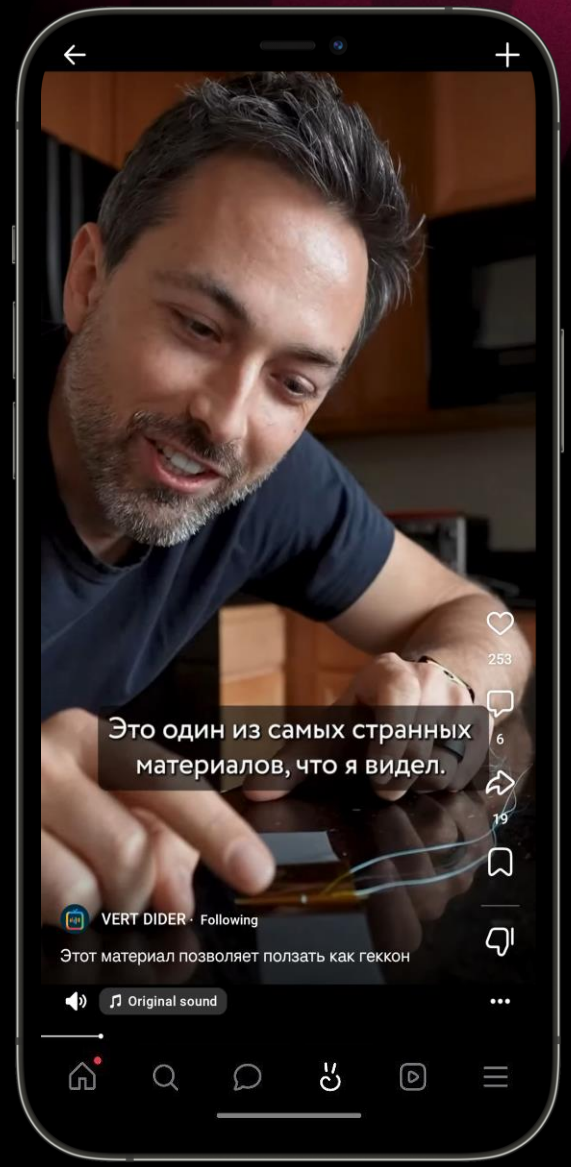
Упорядочить клипы для каждого пользователя так, чтобы клипы с ожидаемо более позитивным фидбэком находились выше.

Статистика

- 10 недель
- 999 участников
- 9255 сабмитов

Призовой фонд

2 000 000 руб

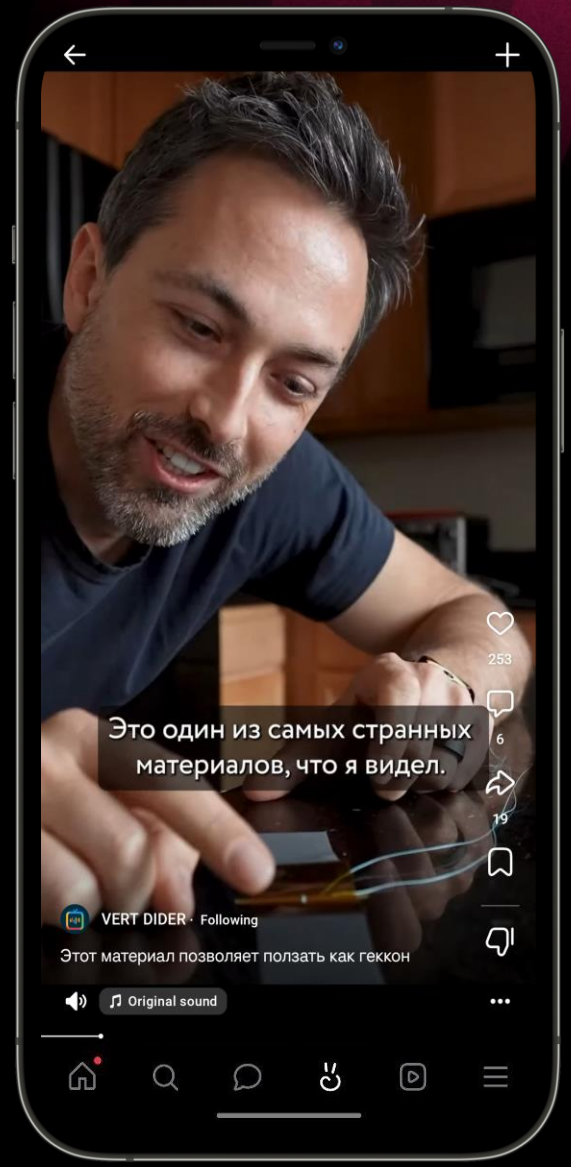


Данные

Пользователи (183k)
- Пол
- Возраст

Клипы (337k)
- Автор
- Длительность
- Эмбеддинг содержимого

Взаимодействия (145m)
- Пользователь
- Клип
- Проведенное время
- Факт лайка
- Факт дизлайка
- Факт шэра
- Факт помещения в закладки



Метрика

$$ROC\ AUC = \frac{1}{|U|}\sum_{u \in U} \frac{\sum_{i \in I_u}\sum_{j \in I_u}[r_{ui} < r_{uj}][\hat{r}_{ui} \lesssim \hat{r}_{uj}]}{\sum_{i \in I_u}\sum_{j \in I_u}[r_{ui} < r_{uj}]}$$

$$[\hat{r}_{ui} \lesssim \hat{r}_{uj}] = \begin{cases} 0, \hat{r}_{ui} > \hat{r}_{uj} \\ 0.5, \hat{r}_{ui} = \hat{r}_{uj} \\ 1, \hat{r}_{ui} < \hat{r}_{uj} \end{cases} \qquad [r_{ui} < r_{uj}] = \begin{cases} 0, r_{ui} \geq r_{uj} \\ 1, r_{ui} < r_{uj} \end{cases}$$

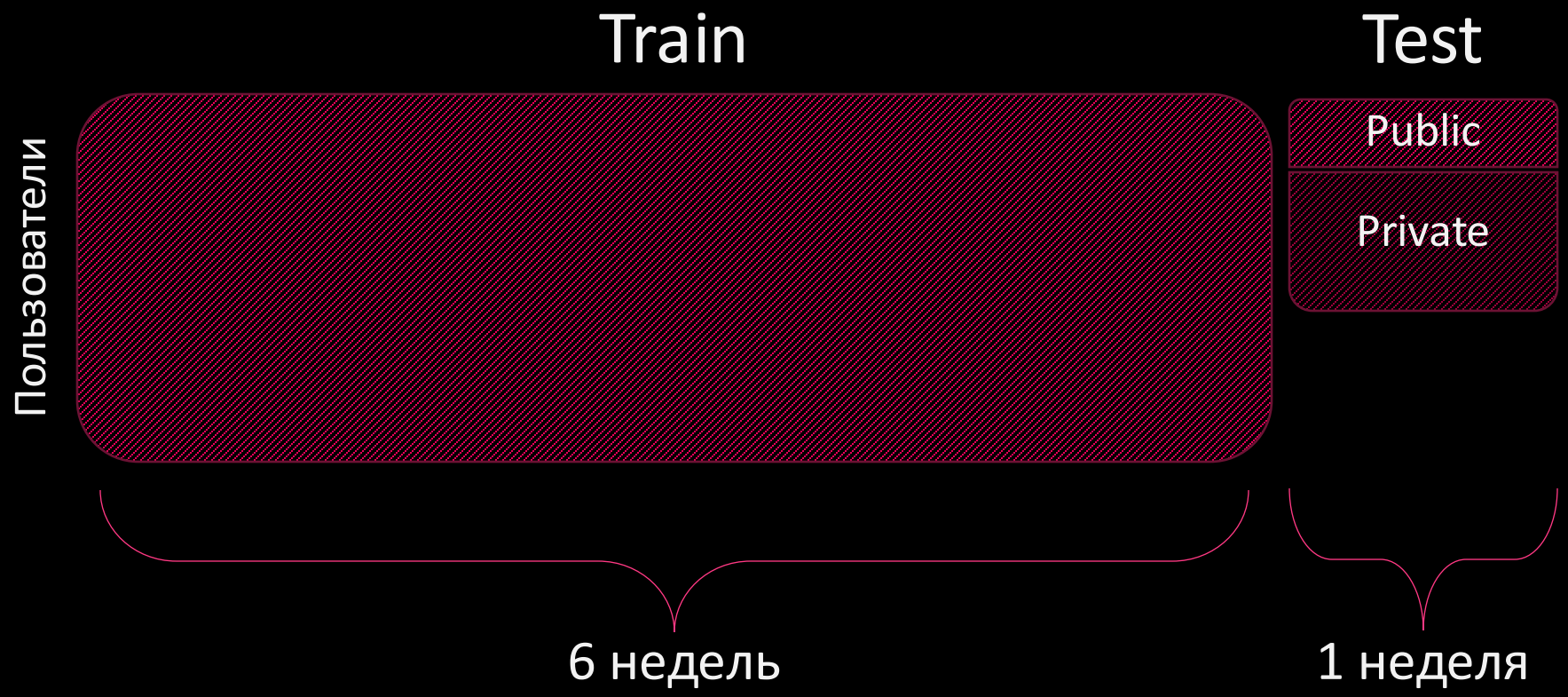
r_{ui} – Истинная реакция пользователя u на айтем i

$\hat{r}_{ui}$ – Предсказанный скор пользователя u на айтем i

U – Множество тестовых пользователей

I_u – Множество тестовых айтемов пользователя u

Train/test split

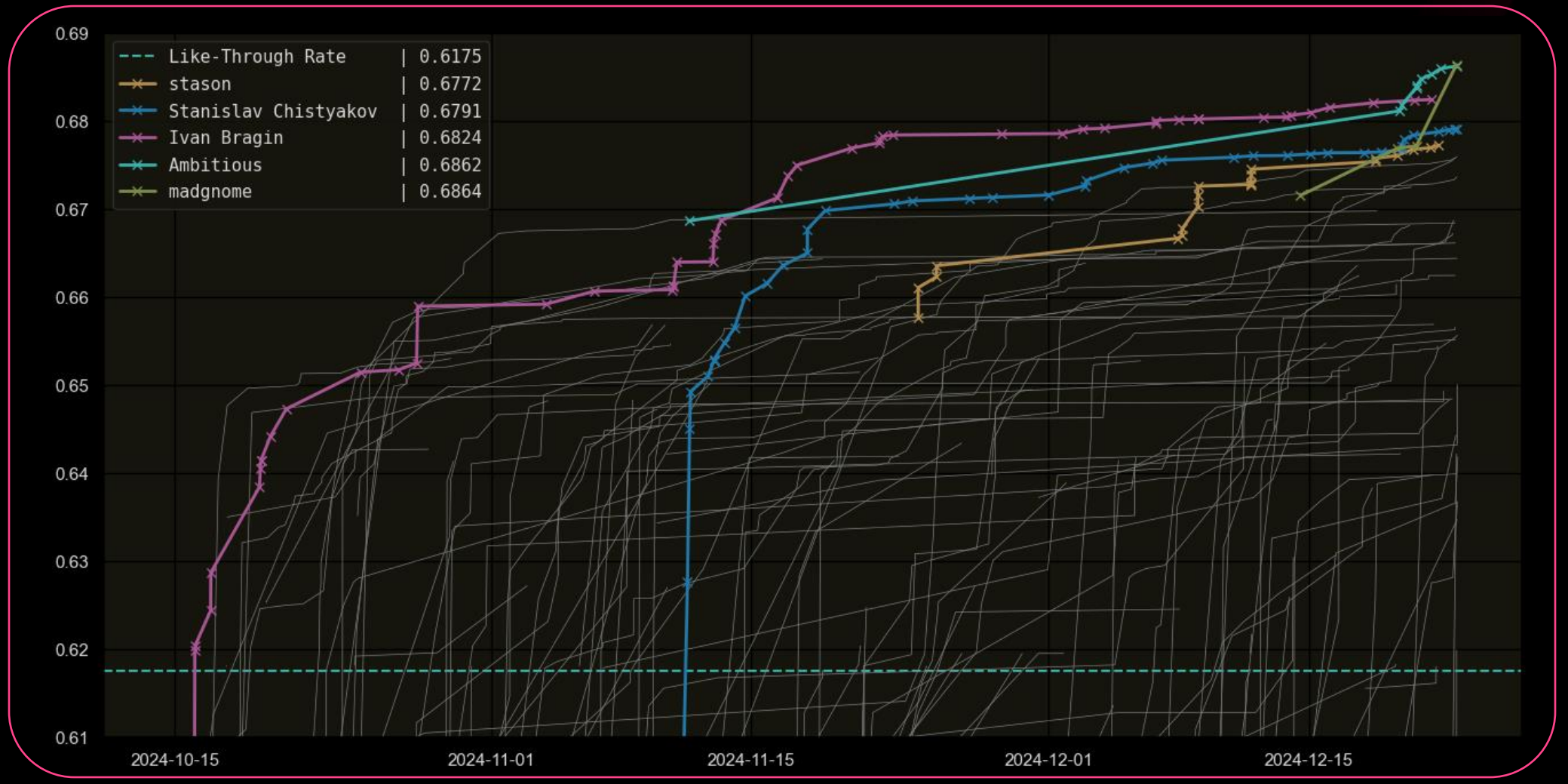


Like-Through Rate

```
train = pl.scan_parquet('train_interactions.parquet')
items_stats = train.group_by('item_id').agg(pl.len().alias('shows'),
                                pl.col('like').sum().alias('likes'))
test_pairs = pl.scan_csv('test_pairs.csv')
test_pairs = test_pairs.join(items_stats, how='left',
                        on=pl.col('item_id').cast(pl.Int64), coalesce=True)
test_pairs = test_pairs.with_columns(predict=pl.col('likes') / pl.col('shows'))
validate_pauc(closed_test, test_pairs.collect())
```

ROC AUC 0.6175 – топ 33% команд

История лидеров



Звездный состав

Михаил Трапезников

Senior MLE, AI VK

Валентина Стецюк

DevRel, AI VK

@DS_LEAGUE