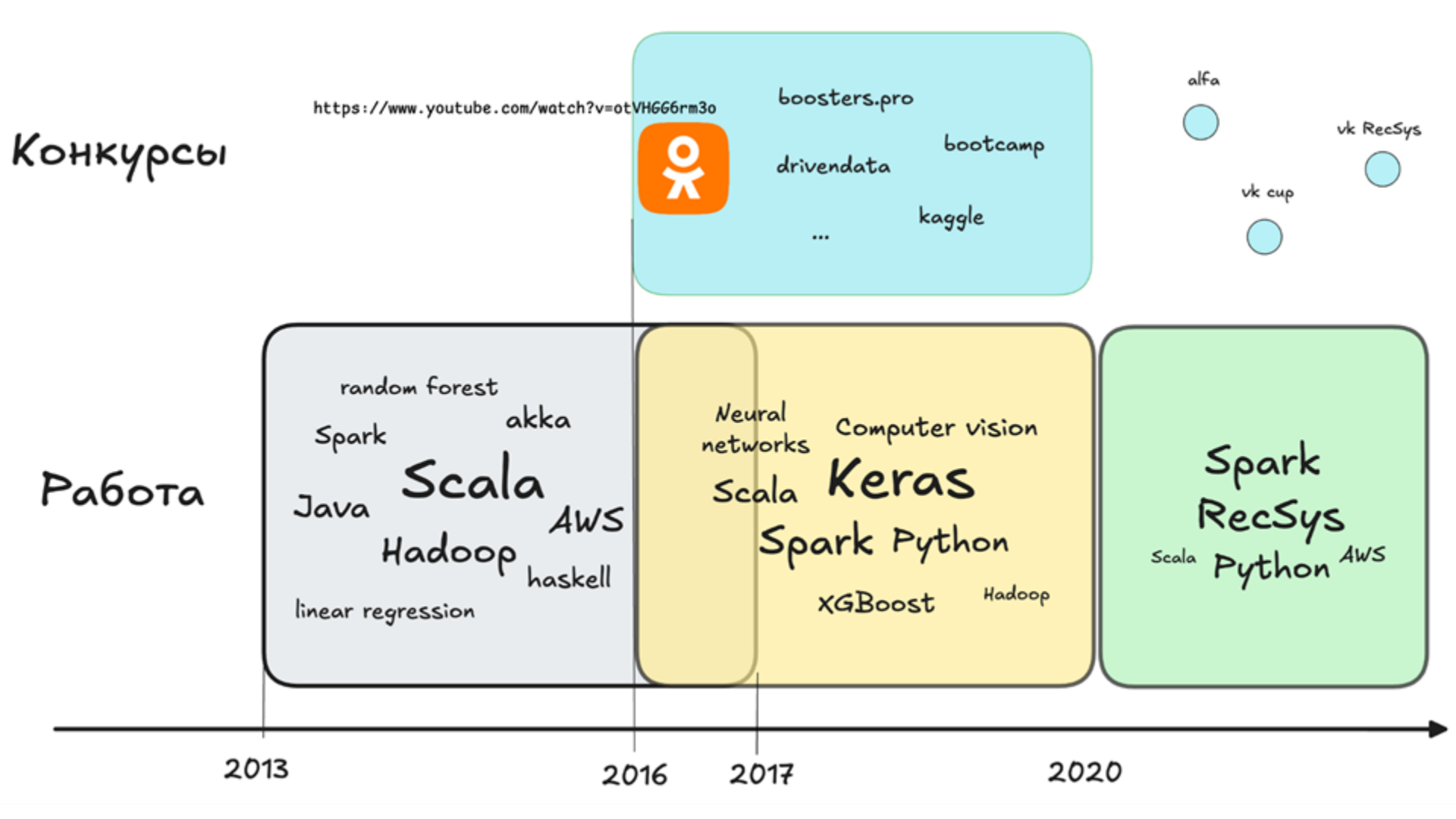
VK RecSys Challenge

catboost is all you need

... для 3го места

Иван Брагин


Конкурсы
https://www.youtube.com/watch?v=otVHGG6rm3o
boosters.pro
bootcamp
drivendata
kaggle
...
alfa
vk RecSys
vk cup
Работа
random forest
akka
Spark
Scala
Java
AWS
Hadoop
haskell
linear regression
Neural networks
Computer vision
Scala
Keras
Spark
Python
XGBoost
Hadoop
Spark
RecSys
Scala
Python
AWS
2013
2016
2017
2020

Высокое качество конкурса

Продолжительность 10 недель

PyTorch

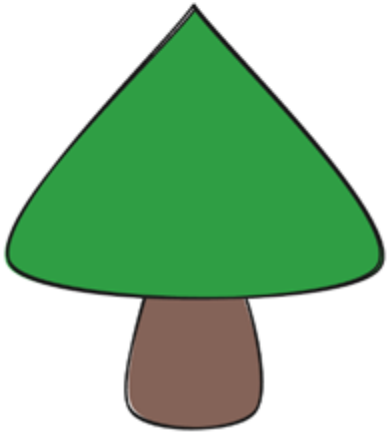
CatBoost

Polars

APACHE Spark™

vast.ai

databricks



Высокое качество конкурса

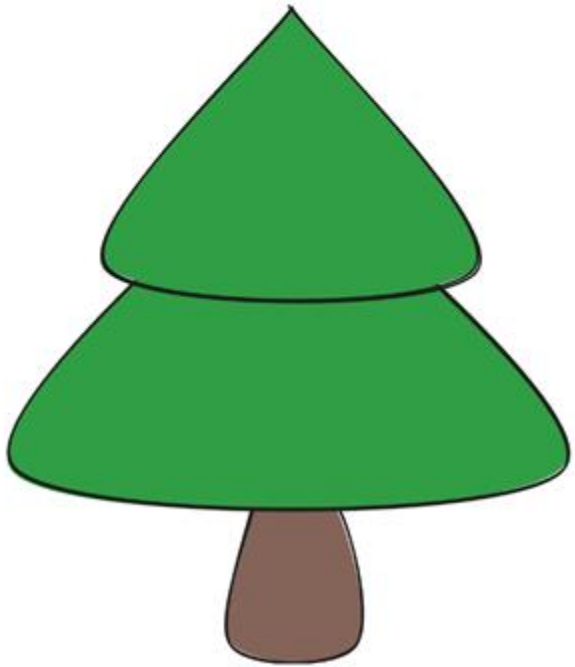
Продолжительность 10 недель

Нестандартная метрика, которую никто не сломал

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

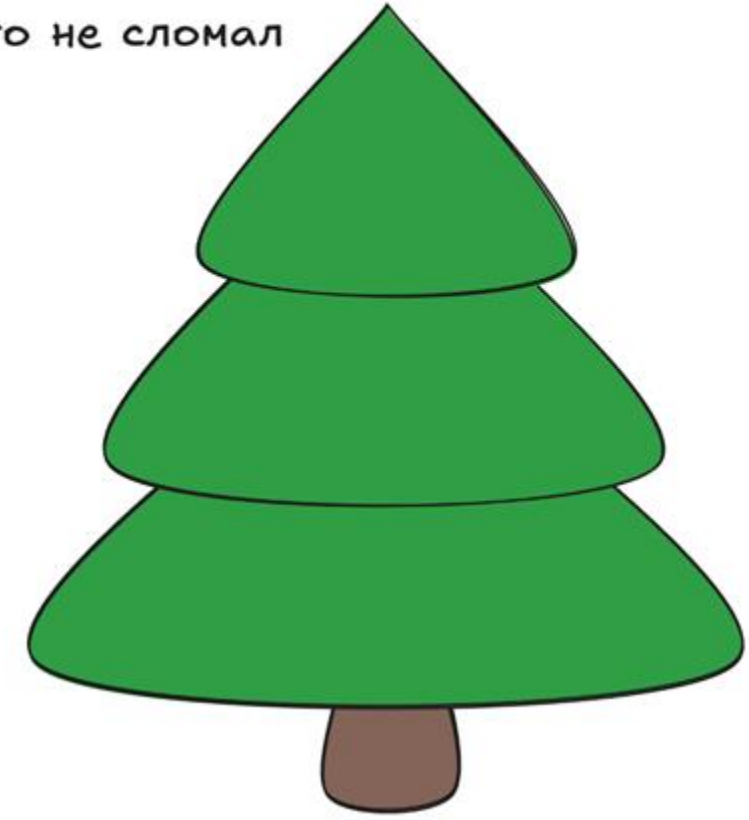
$$[\hat{r}_{ui} \lesssim \hat{r}_{uj}] = \begin{cases} 0 & \text{if } \hat{r}_{ui} > \hat{r}_{uj} \\ 0.5 & \text{if } \hat{r}_{ui} = \hat{r}_{uj} \\ 1 & \text{if } \hat{r}_{ui} < \hat{r}_{uj} \end{cases}$$

$$[r_{ui} < r_{uj}] = \begin{cases} 0 & \text{if } r_{ui} \geq r_{uj} \\ 1 & \text{if } r_{ui} < r_{uj} \end{cases}$$



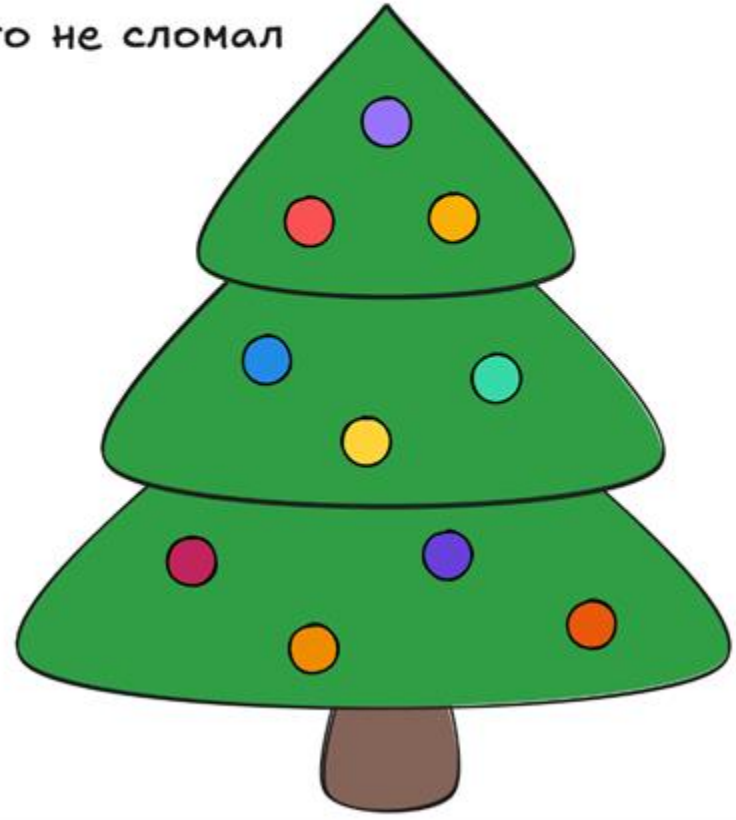
Высокое качество конкурса

Продолжительность 10 недель
Нестандартная метрика, которую никто не сломал
Не было ликов



Высокое качество конкурса

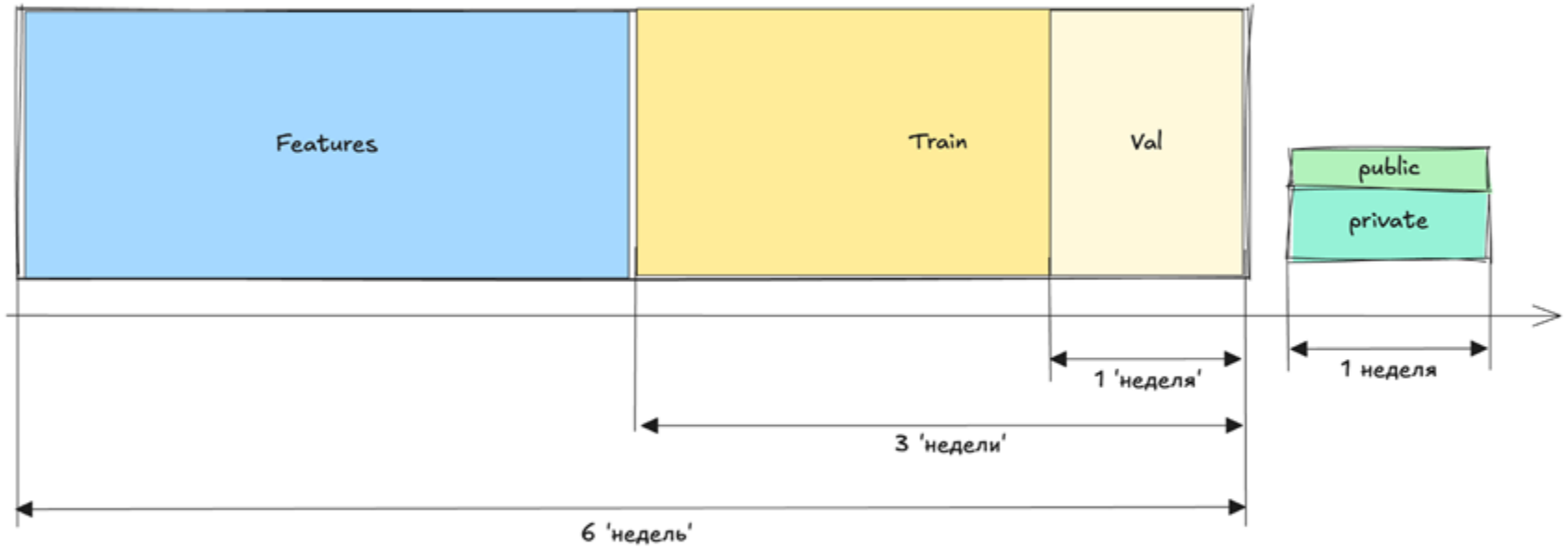
Продолжительность 10 недель
Нестандартная метрика, которую никто не сломал
Не было ликов
Огромное количество данных



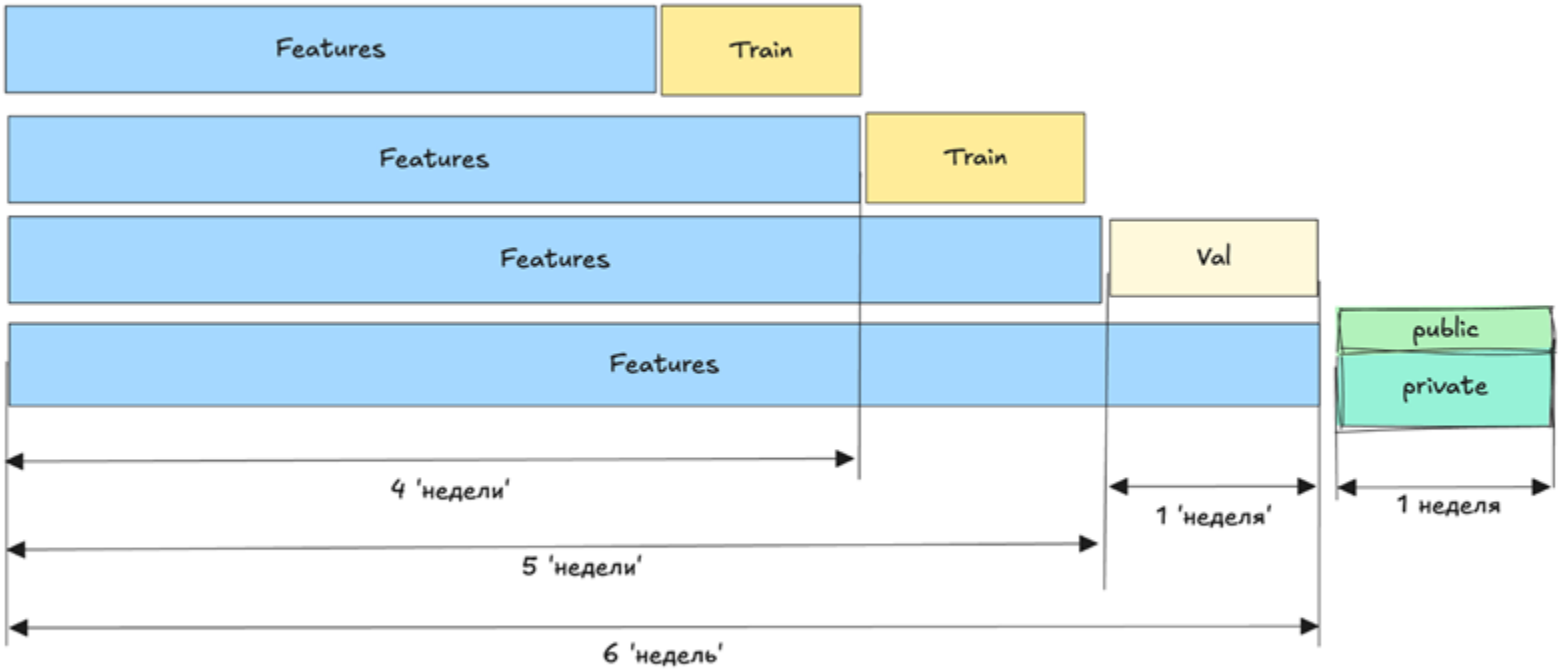
Высокое качество конкурса

Продолжительность 10 недель
Нестандартная метрика, которую никто не сломал
Не было ликов
Огромное количество данных
Разбор решений на дата ёлке

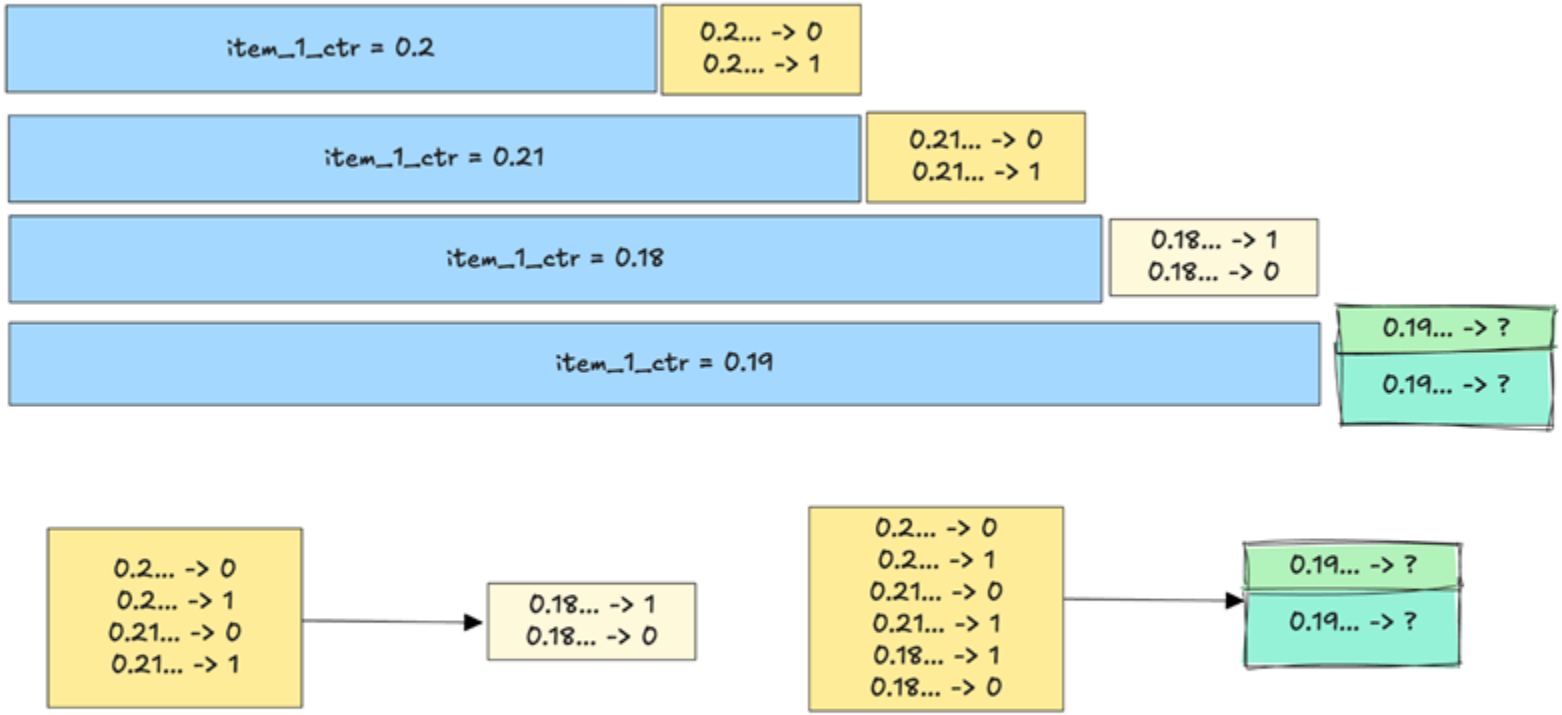
Разбиение данных



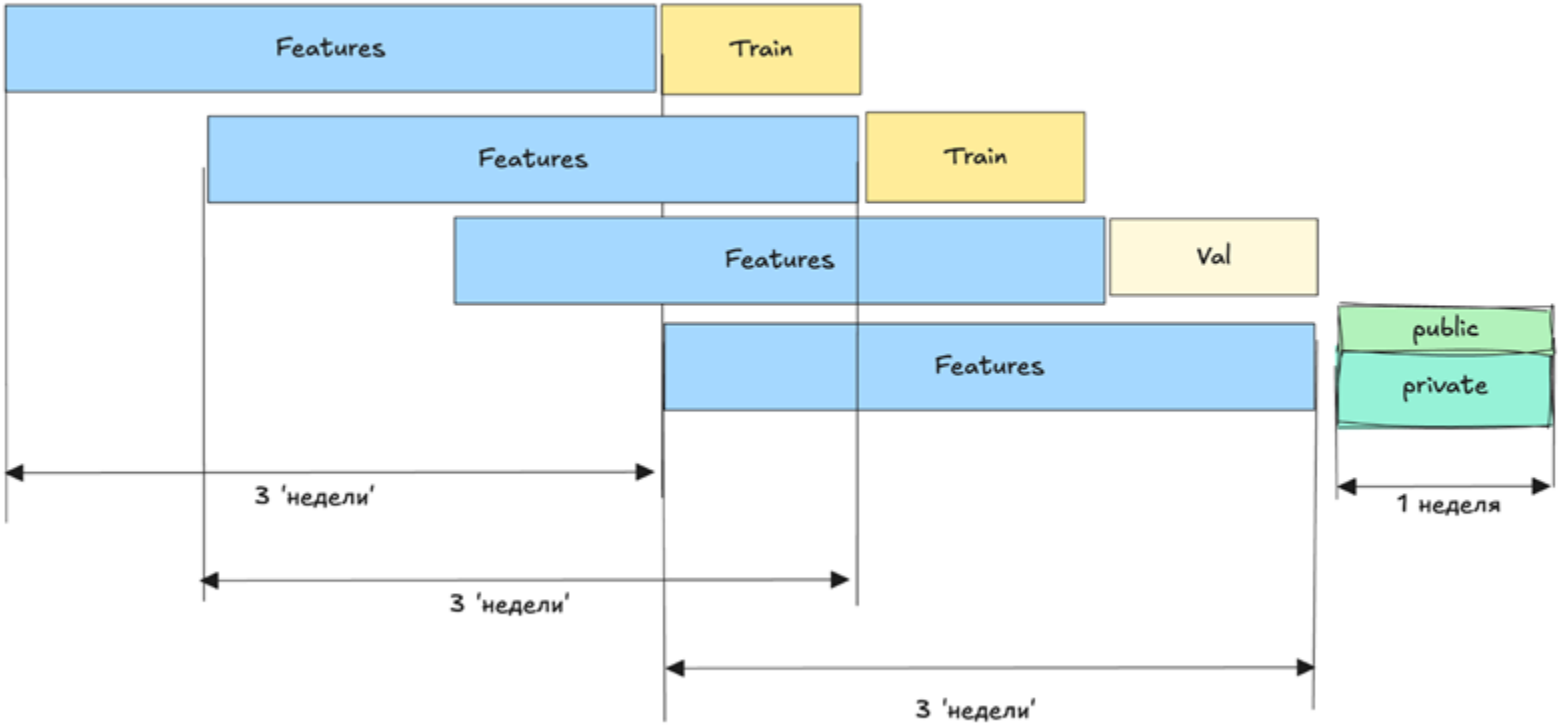
Относительные фичи



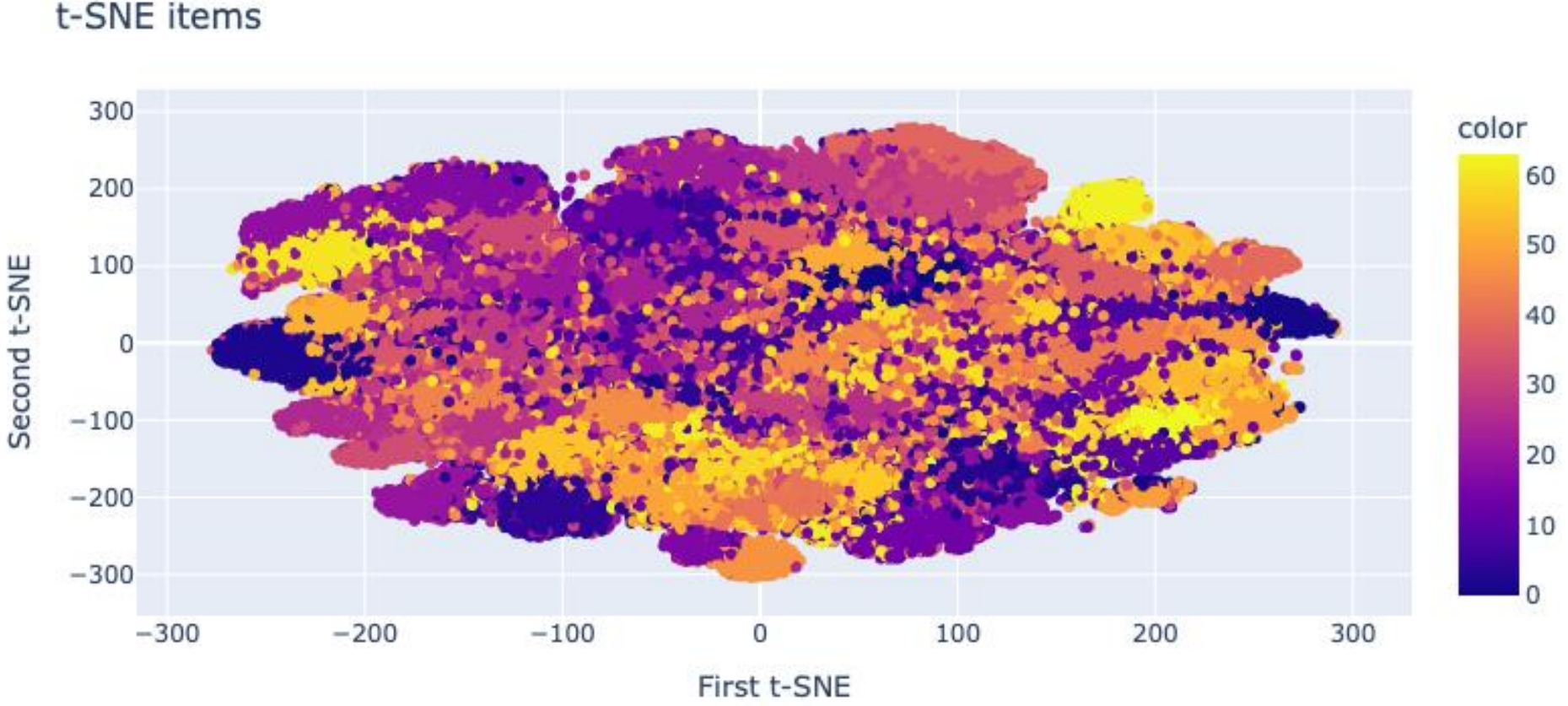
Пример фичи CTR

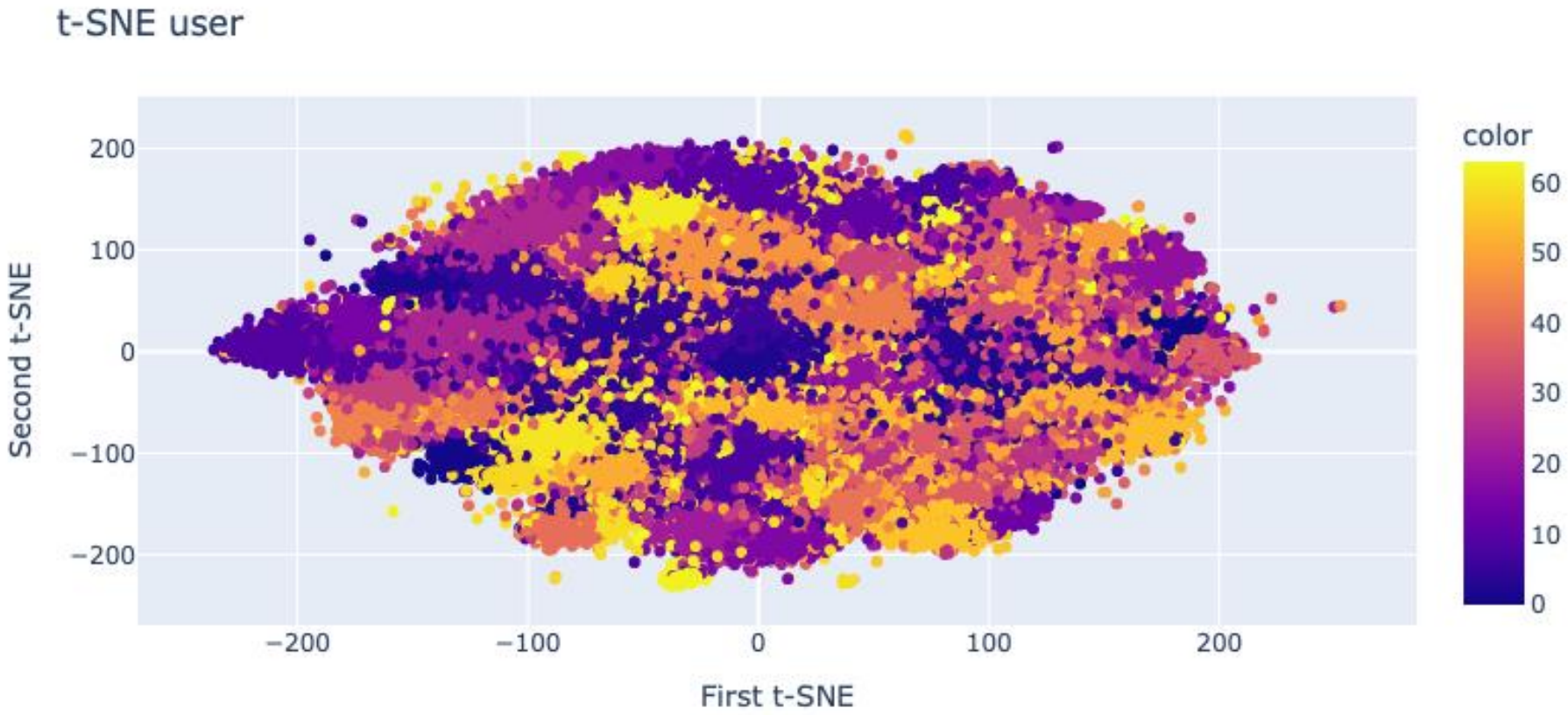


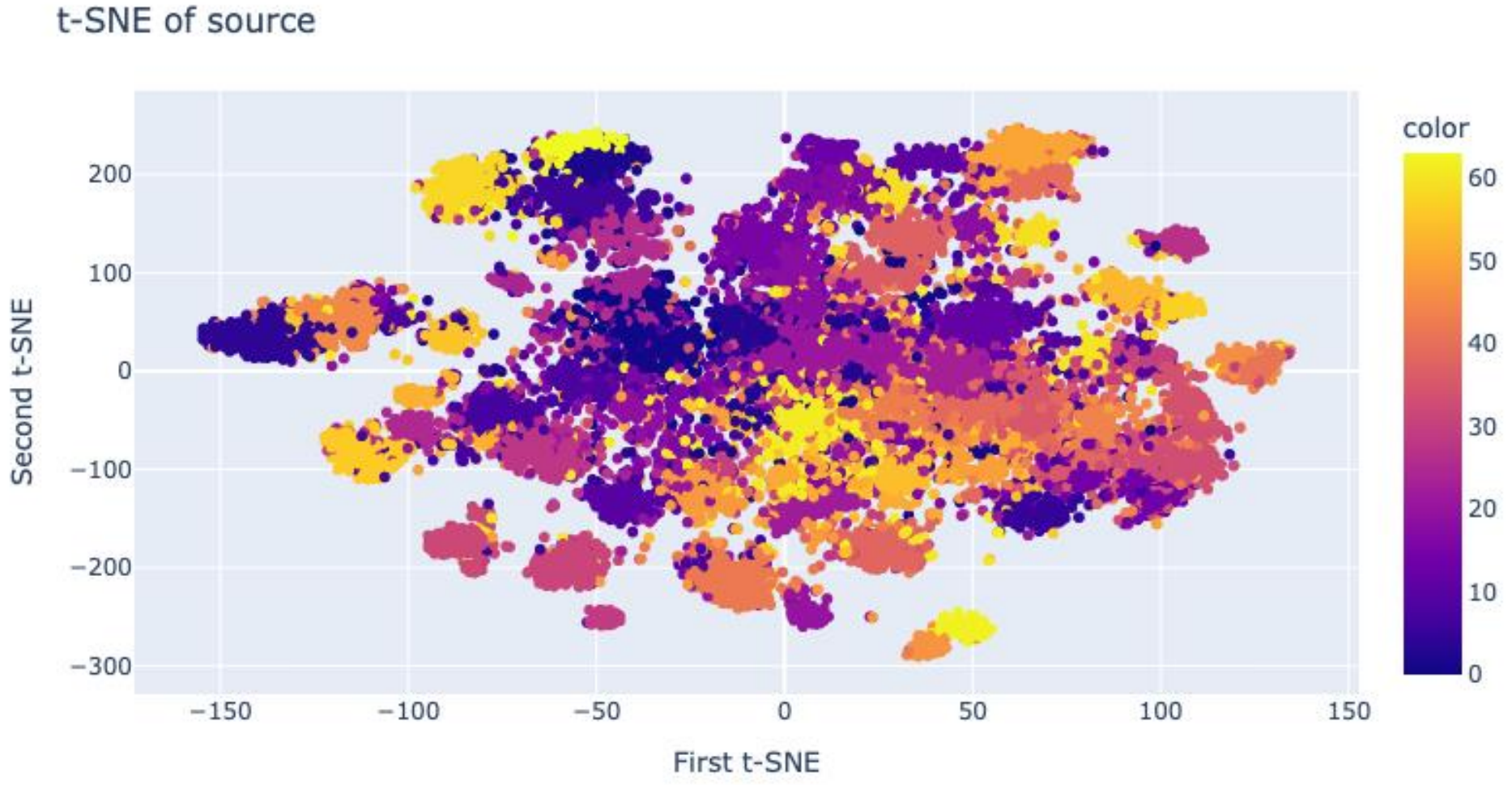
Абсолютные значения фичей



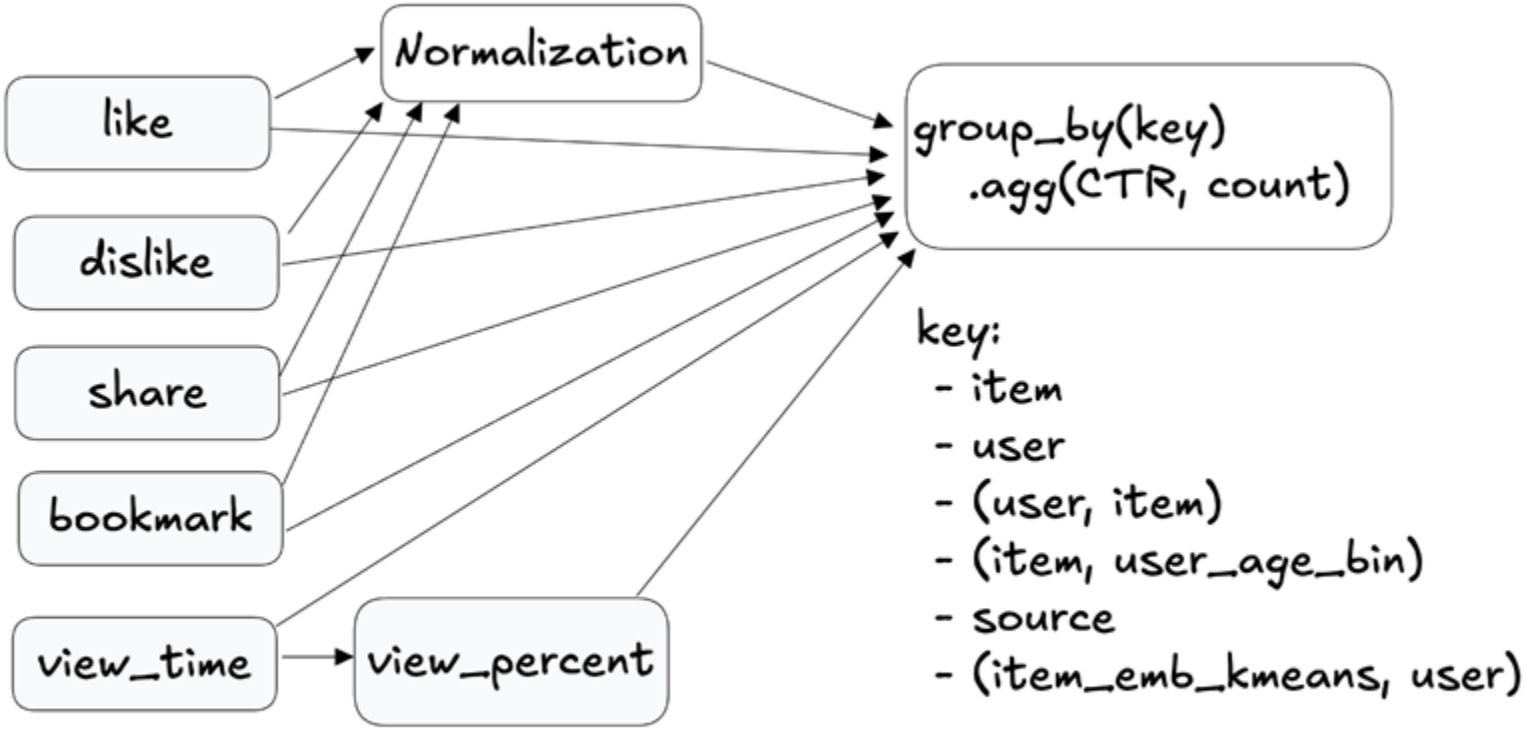
Контентные признаки


t-SNE items
300
200
100
0
−100
−200
−300
Second t-SNE
−300
−200
−100
0
100
200
300
First t-SNE
color
60
50
40
30
20
10
0


t-SNE user
200
100
0
−100
−200
Second t-SNE
−200
−100
0
100
200
First t-SNE
color
60
50
40
30
20
10
0


t-SNE of source
color
60
50
40
30
20
10
0
200
100
0
−100
−200
−300
Second t-SNE
−150
−100
−50
0
50
100
150
First t-SNE

Основанные на действиях



Калибровка действий пользователя

Ваш рейтинг 4.9/5

Вам поставили

Ваш рейтинг 3.8/5

Фреймворк

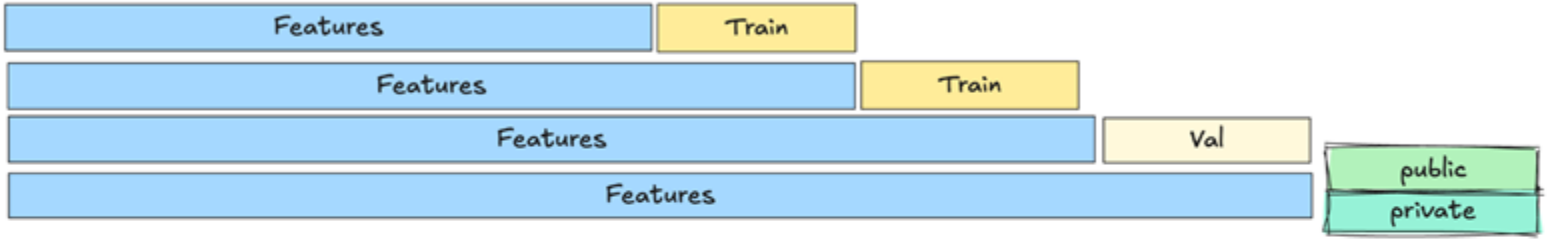
Подготовка фичей

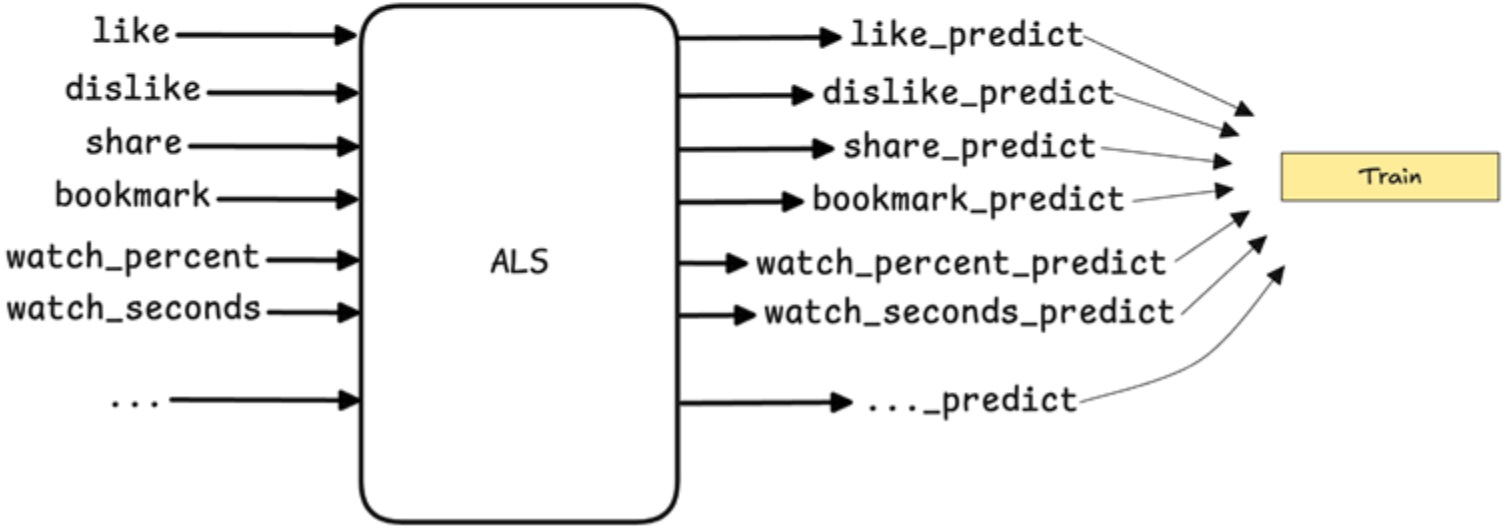
```
class FeaturesLoop():  1 usage  new *
    train_dataset: DataFrame = ...
    source_weeks: float = 3
    relative_features: bool = False
    train_step_weeks: float = 0.5
    def apply_transform(self, func: callable):  new *
        for step in range(...):
            features_df = train_dataset.filter(...)
            train_df = train_dataset.filter(...)
            features = func(features_df)
            train_df.join(features).write(...)
def user_features(df):  1 usage  new *
    return df.groupBy('user_id').agg(...)
def item_features(df):  1 usage  new *
    return df.groupBy('item_id').agg(...)
loop = FeaturesLoop(source_weeks=2)
loop.apply(user_features)
loop.apply(item_features)
```

Чтение фичей и обучение

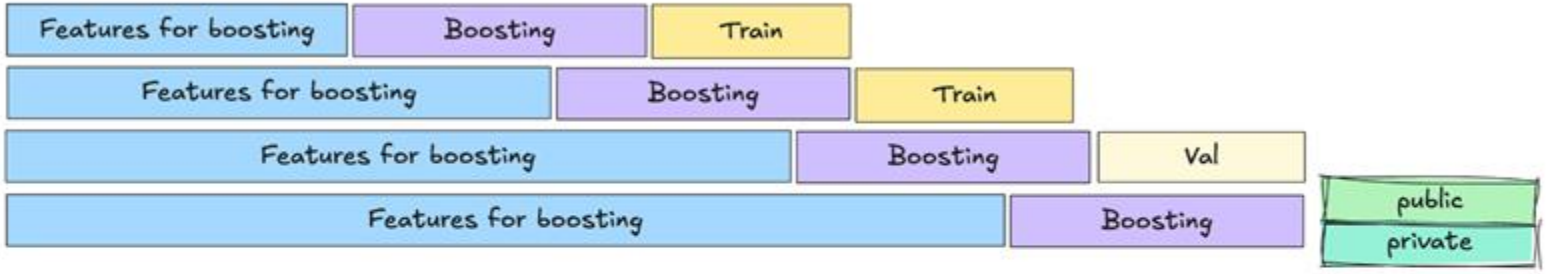
```
user_features = Feature('user_features')
item_features = Feature('item_features')
train_dataset = (train_dataset
                 .join(user_features)
                 .join(item_features))
Catboost.train(train_dataset)
```

Любой фидбэк в ALS





Любой фидбэк в Boosting



Фичи за 2 недели

Features for boosting

Фичи для первого бустинга считаются, как и для конечного обучения
- количество лайков
- ctr
- als likes_predict
- mean watch percent
- ...

Boosting

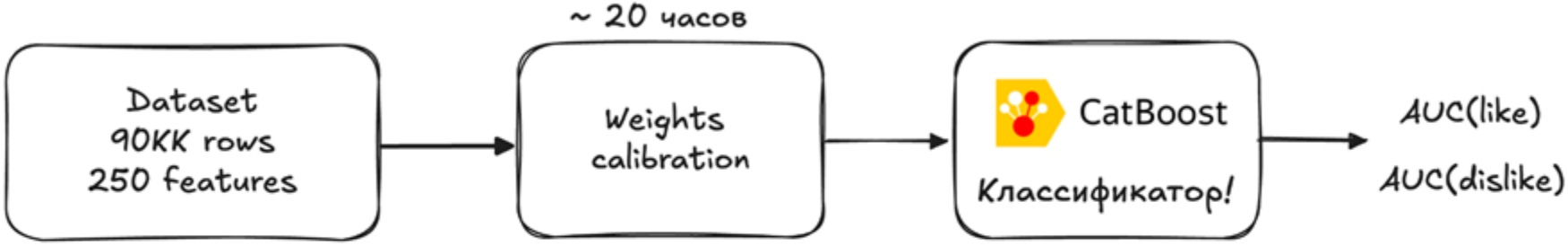
2_weeks_features -> target

где target может быть like, dislike, watch_percent ...

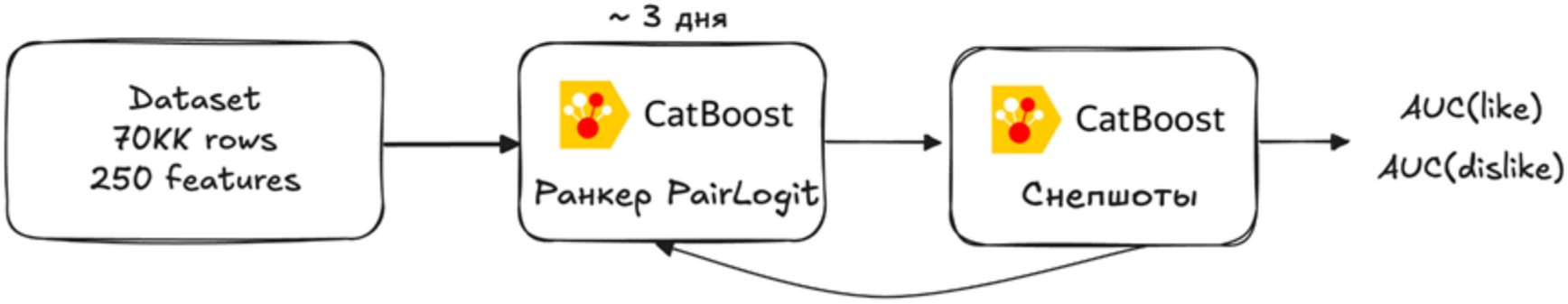
Train

boosting_like_predict,
boosting_dislike_predict...
als_like_predict, als_wp_predict...
item_ctr, item_norm_ctr, item_likes_couint...
user_ctr, user_item_cos...
source_ctr...

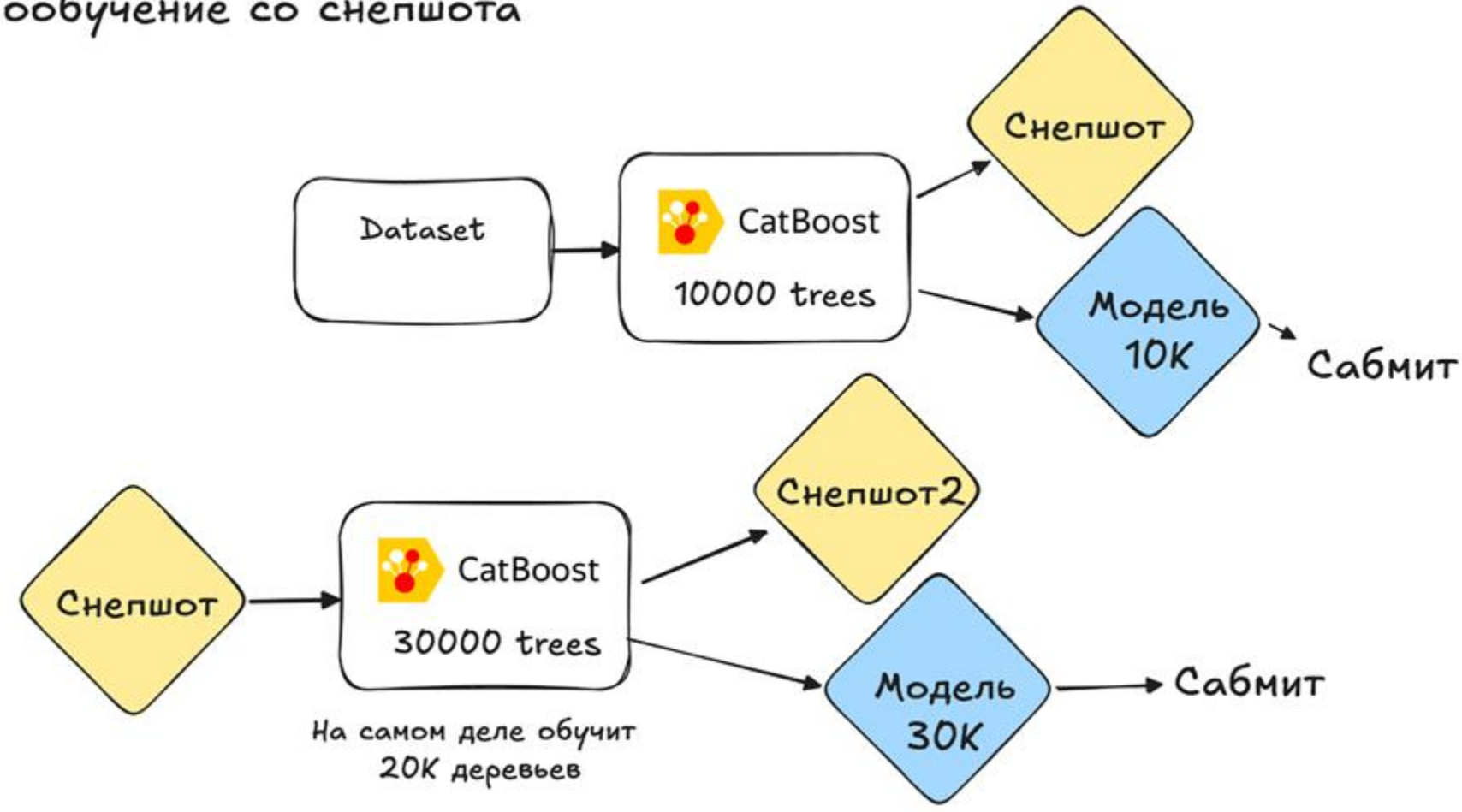
Обучение



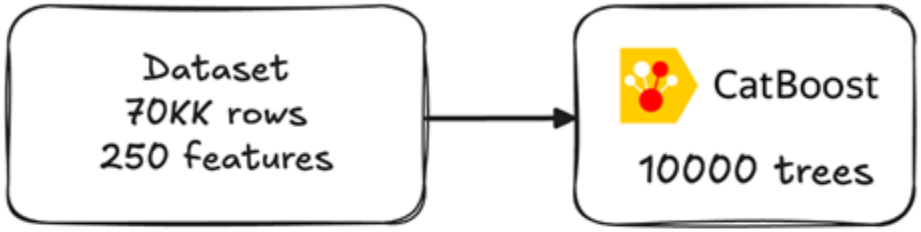
На последней неделе перешел на ранкер

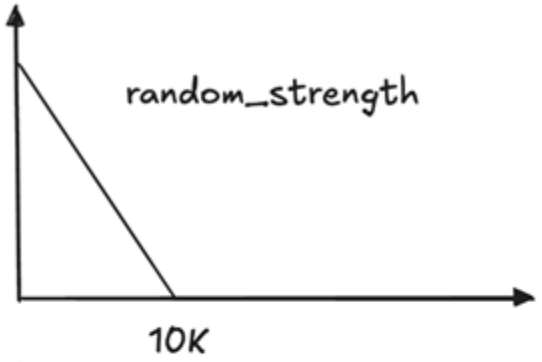


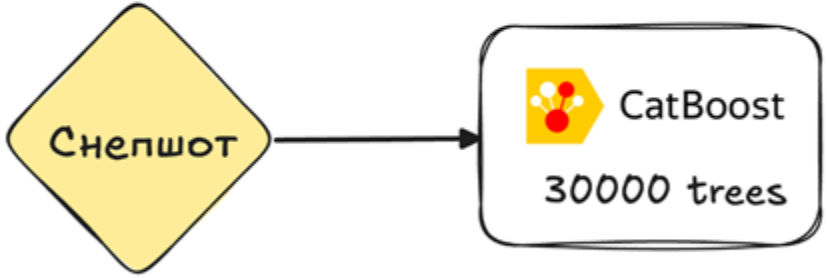
Дообучение со снепшота

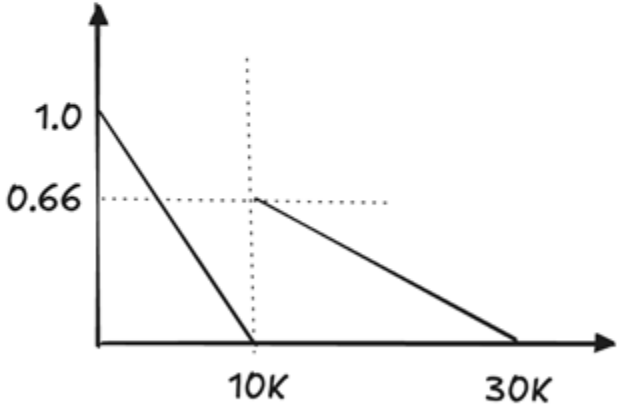


Проблема дообучения со снепшота









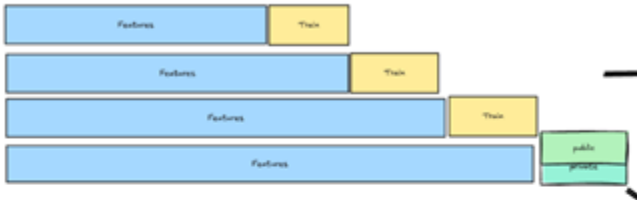
The scores have no randomness. A normally distributed random variable is added to the score of the feature. It has a zero mean and a variance that decreases during the training. The value of this parameter is the multiplier of the variance.

*https://catboost.ai/docs/en/references/training-parameters/common

Железо

m:11184	host:43425	Saitama, JP	H12SSL-i	↑2969 Mbps		verified	$0.551/hr
vast.ai	1x RTX 4090		PCIE 4.0,16x 23.8 GB/s	↓6815 Mbps 199 ports		Max Duration 2 mon, 14d	
Type #15876758	82.6 TFLOPS	24 GB	AMD EPYC 7K62 ...	WD Blue SN570 2...	97.3 DLPerf	Reliability	RENT
	Max CUDA: 12.2	875.9 GB/s	48.0/96 cpu 258/516 GB	3409 MB/s 705.0 GB	176.6 DLP/$/hr	99.927%	

Вопросы




```
train_data_with_features = Train
.join(features.groupBy(key1).agg(ctr, count ...))
.join(features.groupBy(key2).agg(ctr, count ...))
...
.join(features.groupBy(keyN).agg(ctr, count ...))
```

```
test_data_ Test
.join(features.groupBy(key1).agg(ctr, count ...))
.join(features.groupBy(key2).agg(ctr, count ...))
...
.join(features.groupBy(keyN).agg(ctr, count ...))
```

Polars

4090 TI
256 ram

vast.ai

CatBoost

PairLogit

Различные близости видео с
- другими видео
  - например самым популярным
- юзерами, где юзер это
  - mean(user_likes)
  - mean(user_dislikes)
  - mean(user_views)
- источниками, как с юзерами

ctr и count по фидбэку: like, dislike, share, view_percent...

по ключам: item, source, user, age_bin, gender, kmeans и их комбинациям

boosting и als как фичи