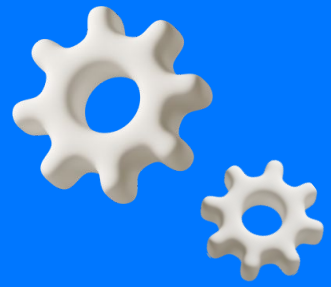


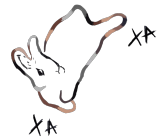
КЕЙС ОТ ВК:

ПРОГНОЗ, ОСУЩЕСТВЛЕНИЯ ЦЕЛЕВОГО ДЕЙСТВИЯ ПОЛЬЗОВАТЕЛЕМ В ЗАВИСИМОСТИ ОТ ЕГО КЛИКСТРИМА, СЧИТАЯ МЕТРИКОЙ КАЧЕСТВА ПОСТРОЕННОЙ МОДЕЛИ ROC-AUC SCORE

Решение команды “Московские зайцы”

РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

цифровой прорыв сезон: ИИ



ДАННЫЕ

- CLIENT_ID
- RETRO_DT
- TOKENS
- URLS_HASHED
- DEF

```
'code 1 историй 1 scf 1 шаг 1 деньги 3 серый 1 авиабилеты 8 tova
tivation 2 faq 1 faw 1 оттенок 1 оформить 2 cool 1 надежные 2 те
completion 1 agreement 1 выгодной 1 address 1 телефона 1 фильм 1
1 новости 2 средство 1 cveta 1 оплата 1 msk 1 passport 1 src 1 c
секс 1 дешево 8 отправлена 1 друзьями 1 slova 2 нальчик 3 recove
чшие 1 mae 3 mag 2 law 1 maska 1 short 2 natural 1 helsinki 1 чи
ения 1 scoring 4 vivus 1 tsi 1 предложения 1 offers 1 дешевых 8
ора 1 landings 1 pam 1 оформление 1 kfc 1 займы 18 икеа 2 робот
```

СЛОЖНОСТИ

- (Very) Big Data
- Неоднородность

```
'1e833434273e04ba76cfcfb4b48ad21b 3 aee71c8d18
8cfcf7 1 39b0d1a68355c8a34807b35c43c507d5 1 42
2eca6e29bdda56 2 67ce52bfdc907a586499417065490
1dafd37986235c6c61158e0 2 c759ea7aaf7a4c4606f7
7b7c8647c131f56fe147a9612e1a766 1 e3fbe7c0fa70
7029 1 e55c08a3538daa74ee8e3c581a9feafe 4 bb55
1b047a709eb5 1 41e46149fec1622332d023d2d6cb5a0
7b471d3ff84d02d2e652 2 1372a402b2bce58430fb6d8
20dc2fb4615d172f1b3d105ae912 2 64a7275a3ae205e
2 10 1a949a82f9f6be9096b327557b848d16 1 7405aa
0ac4c9b476f 1 2ea04a31bf4dda27ccd3f7780e870db9
7de7b89f95106425b0f 8 cee8121c8e56150bd0719196
79c10dfe11edad7105c77f7bdad 1 f6e856d1ae5b93db
3 1192d043fa15f11a5f0412c8b978a4c2 2 c55309e36
1da44d93 1 795c2e25b616f4dcf949b7593cdbf6b1 2
92e1d2a3a99afaa1 4 9a5add12fb2126e6d11ecefc94c
```

Наш подход

- Протестировать различные базовые пайпланы для быстрого получения метрики и понимания, в каком направлении двигаться.
- Избавиться от сильного зашумления и высокой размерности данных.
- Сформировать фичи - векторные представления при помощи различных алгоритмов.
- Обучить модели на этих фичах и получить метрики.

ЭВРИСТИЧЕСКОЕ ПОНИЖЕНИЕ РАЗМЕРНОСТИ

- Убрали самые редкие url-ы и токены
- Посчитали для токенов и url-ов коэффициенты как отношение количества их встречаемости в положительном и нулевом классах
- Исключили те, которые встречаются сопоставимое количество раз в обоих классах

ВЕКТОРНЫЕ ПРЕОБРАЗОВАНИЯ ДАННЫХ

1. Получили эмбеддинги токенов при помощи sentence-transformers.
Тестировали distilrubert и LaBSE.

2. Обучили SVD и на токенах, и на url. Получили средний эмбеддинг для tokens и urls_hashed.

3. Обучили Word2vec на url-sequences. Собрали взвешенный эмбеддинг для urls_hashed.

МОДЕЛИ

Базовые подходы:

- BERT
- CatBoostClassifier на текстовых фичах
- CatBoostClassifier на TF-IDF

Метрика ~ 0.6

Наш подход:

- Объединение фичей
- (LaBSE Embeddings + SVD + Word2Vec + Latent Dirichlet Allocation)
- Репрезентативные текстовые фичи
- CatBoostClassifier + отбор фичей + StratifiedKFold

Метрика ~ **0.72**

ДАЛЬНЕЙШЕЕ РАЗВИТИЕ

- Протестировать большее количество векторных представлений: SVD на последовательностях, эмбеддинги трансформеров по словам и другие
- Использовать вместо CatBoost другие алгоритмы, такие как LightGBM, TabNET, LAMA
- Проверить иные варианты понижения размерности: другие трешхолды для нашего метода, либо иные методы

Команда “Московские зайцы”

Анна Беляева

Дизайнер

РЭУ им. Г.В. Плеханова,
Аналитик и дизайнер в
“Территории.РФ”

Даниил Степанов

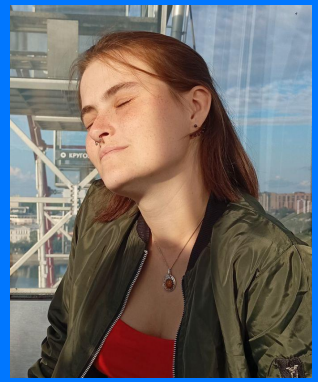
ML engineer

МИСиС,
Data scientist
в “РСХБ-Интех”

Юрий Баландин

Data scientist

ИТМО,
Продуктовый аналитик
в АО “Тинькофф банк”



Анастасия Алимова

Аналитик

СПбГУ