ВК

ИИ по следам пользователей

Банда пяти 2.0

the team

Даня
ML Dev

Саша
Analytic

Артём
Social
Dev

Мила
Algorithmist

Антоша
Python Dev

the task

clickstream -> url

the problems

01

Низкая доля положительного класса

02

Необходимость экстракции признаков

03

Пропуски и шумы

Падающ
VK Cl

data preprocessing

- Экстракция признаков: статистика по каждому юзеру и временным интервалам
- Детекция пропусков и шумов
- Приведение токенов к однородному виду (транслит + приведение к норм. форме)

the facts

24.3%

записей имеют положительный класс

0.23%

невалидных записей

Самые популярные токены:
новости, займы, кредиты, смотреть, купить

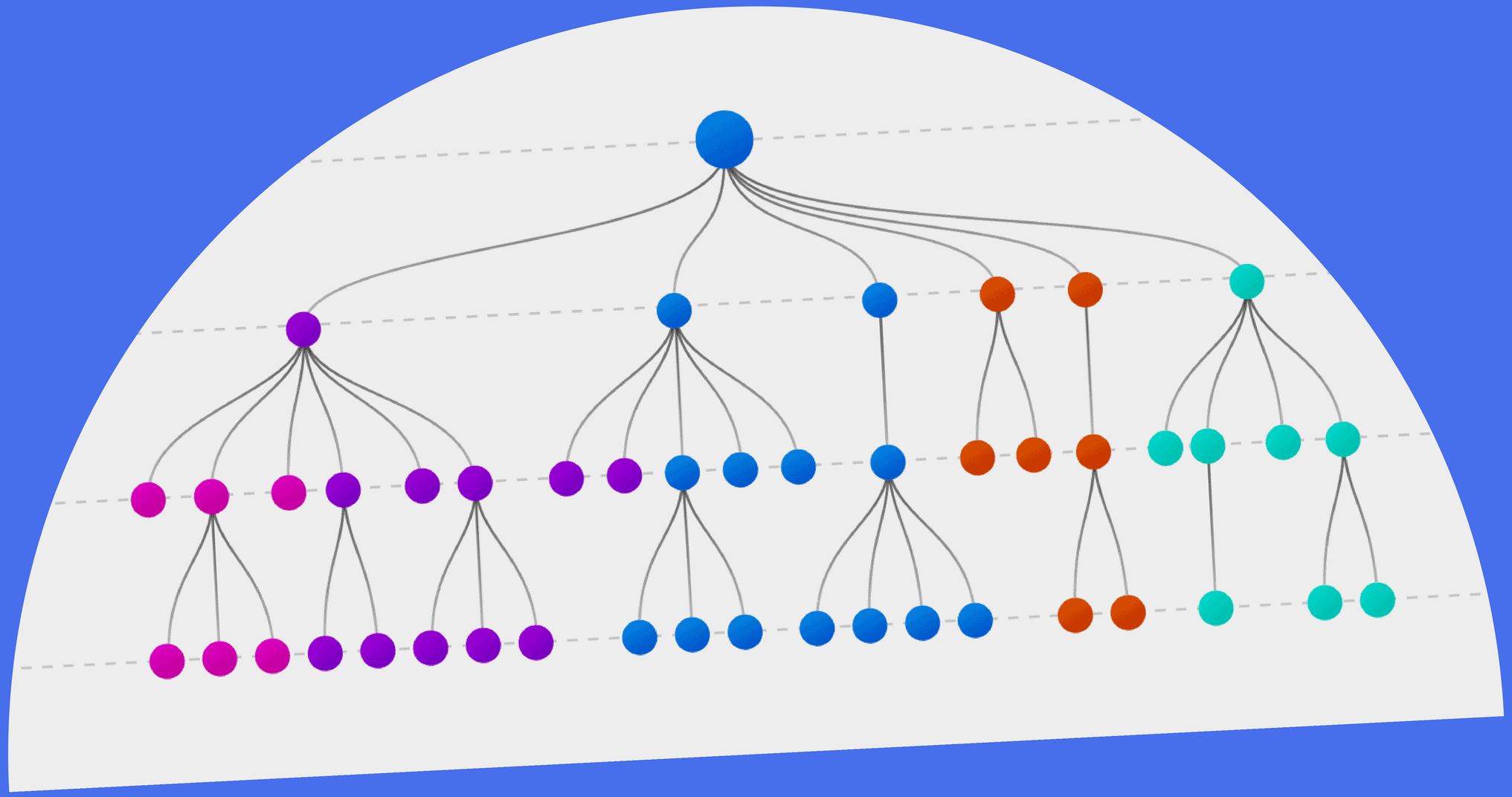
Самые непопулярные токены:
фамилии,
токены с грамматическими ошибками,
жаргонизмы

our solution

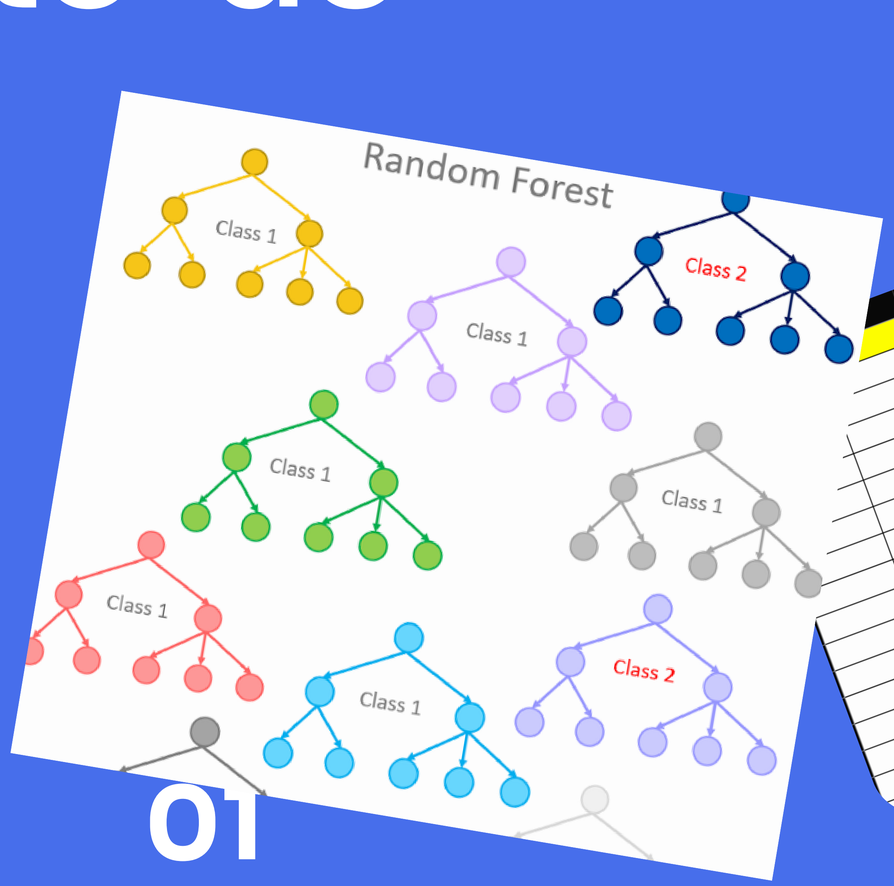
Стек

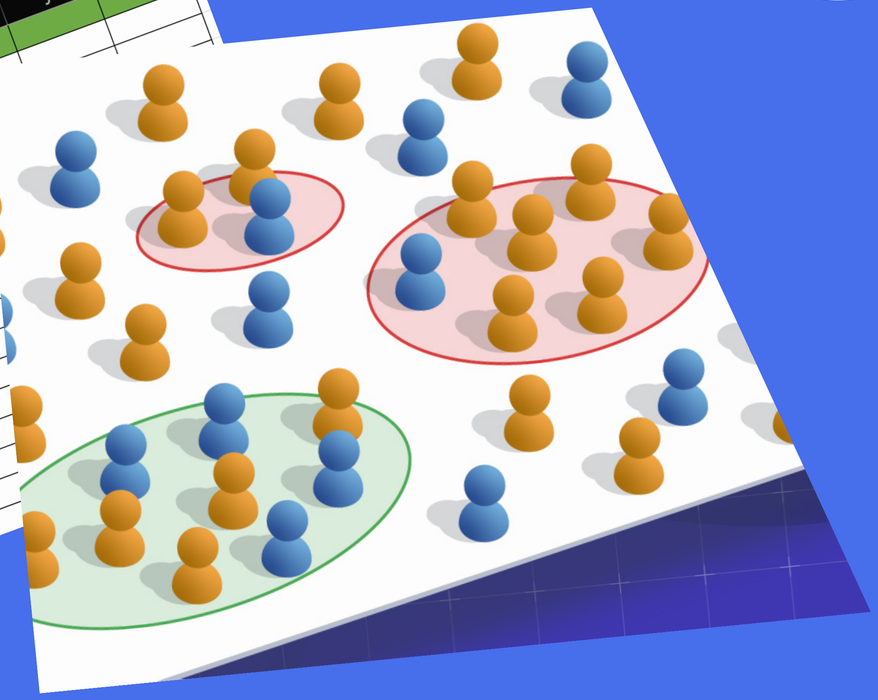
Pandas — для обработки
Numpy, Matplotlib — для визуалиции
Sklearn — для построения моделей
Pymorphy2 — для анализа тегов

decision tree



to-do





01

Расширить модель до леса (Random Forest)

02

Взять больше признаков

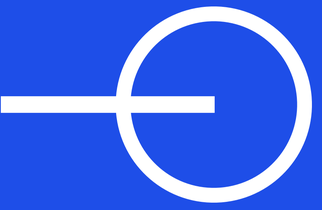
03

Увеличить количество выборок // Обучить на другом сете

thank you

for your consideration

band of five 2.0



Если хочешь быть королем джунглей, мало просто вести себя как король, нужно быть королем. И никогда нельзя сомневаться, ведь сомнения порождают хаос и ведут к краху. Этому научила меня моя королева.

— **Mickey Pearson**

ВК