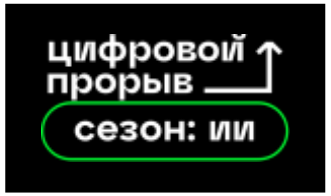
Предсказание социально-демографических характеристик пользователя

Цифровой прорыв 2024
Всероссийский хакатон, Москва



ЛИФТ

состав команды

Руслан Латипов

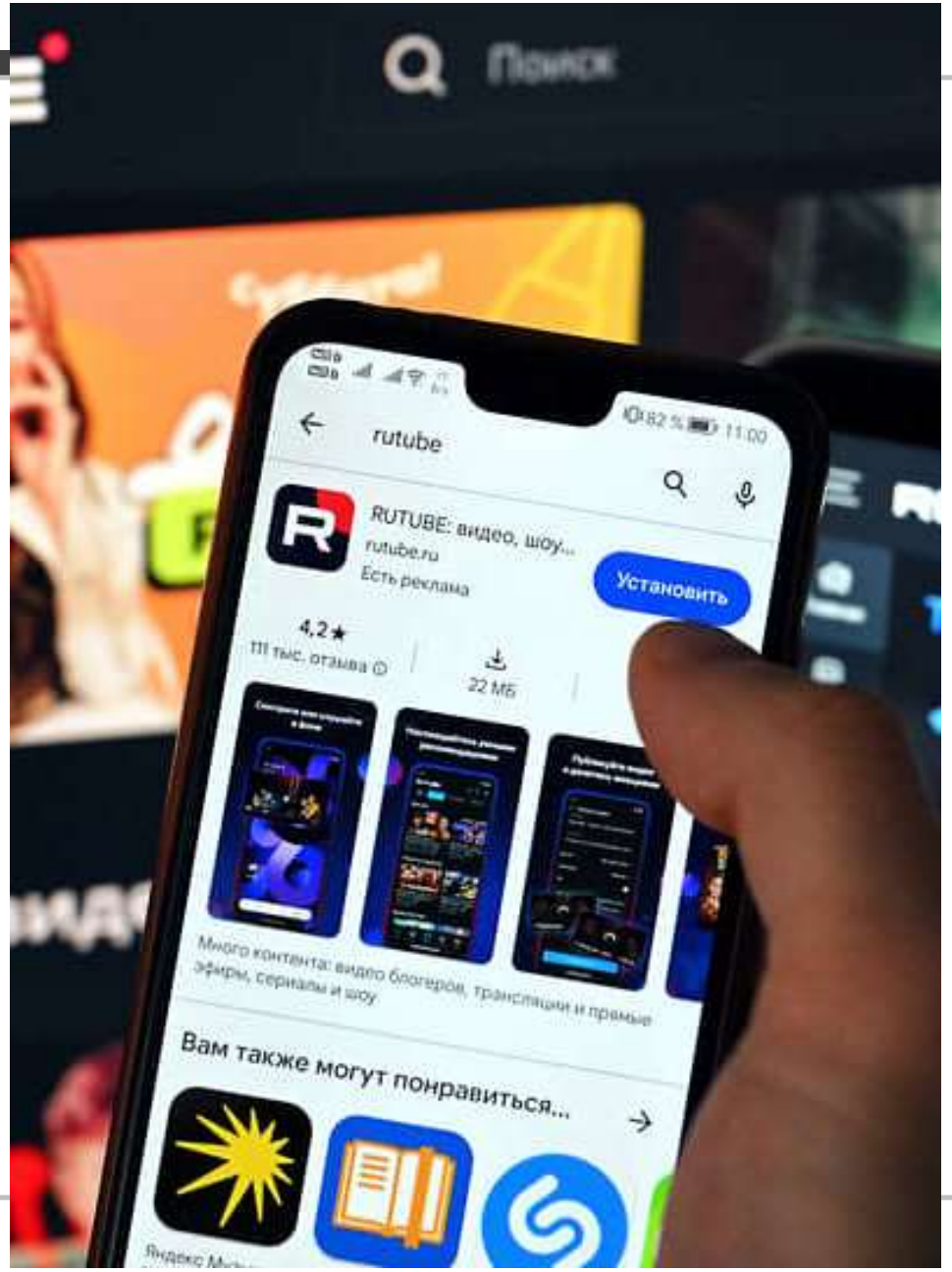
Full Stack Developer, Зеленодольск, @rus_lat116

Юрий Дон

Data Science, Краснодар,
@Yuriy_Nikitich

Верт-Миллер Алексей

Data Science, Архангельск, @alexwert3



Задача

Пользователи RUTUBE не всегда указывают свои данные, такие как возраст и пол, что затрудняет формирование портрета пользователя и создание персонализированных рекомендаций. Это ограничивает возможности платформы в предоставлении контента, который наиболее подходит интересам и потребностям пользователей, тем самым ухудшая пользовательский опыт.

Необходимо разработать модель, которая на основе истории просмотров сможет предсказывать пол и возраст пользователя. Подобные данные смогут служить сильными признаками в моделях рекомендательной системы, поиска и модерации контента платформы.

Данные

обзор данных

Пользователи

180 012 пользователей
4 возрастные категории
половой признак

Временные зоны

дополнительно получены данные для приведения времени к региональному

История просмотров

история просмотров видео
за июнь 2024 года

Видео

354 367 уникальных видео
14 000 категорий
54 автора

EDA

анализ данных

1 ## Обнаружены «мульти» пользователи

пользователи смотрят видео с одного устройства (дети/родители, трансляция видео с устройства на TV - семейный просмотр, регистрация аккаунта на ребенка)

Рекомендация

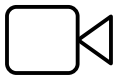
рекомендовать к просмотру видео для семейного просмотра (новая категория видео)

2 ## Дисбаланс классов по возрасту

доля возрастной категории до 20 лет составляет менее 10%

3 ## Наиболее популярное видео для всех категорий

телепередача - «Битва экстрасенсов», разных сезонов



Рекомендация

рекомендовать данное видео пользователям у которых нет истории просмотров

FE

генерация признаков

1

Приведение времени к региональному

данный признак позволяет формировать дополнительные признаки с учетом дня недели, времени суток, выходных дней

2

Текстовое описание видео

данный признак позволяет выявить уникальные слова в описание видео характерные для отдельно взятой категории по полу или возрасту

3

Определение категориальных признаков для модели

операционная система, тип устройства, тип браузера, название клиента браузера

4

Разряженная матрица признаков

применен подход используемый в рекомендательных системах, сформирована матрица взаимодействия пользователя и видео значениями в которой являются отношение длительности просмотра видео пользователем к общей длительности видео (рейтинг видео у пользователя), в результате факторизации данной матрицы получены эмбединги характеризующие пользователя

5

Частотные признаки

выявление популярных видео у пользователей по авторам и категориям видео на истории всех просмотров

ML

модель

Градиентный бустинг

- CatBoostClassifier
- LightGBM

1. библиотеки с открытым исходным кодом, которые предоставляют эффективную и действенную реализацию алгоритма градиентного бустинга

2. LightGBM расширяет алгоритм градиентного бустинга, добавляя тип автоматического выбора объектов, а также фокусируясь на примерах бустинга с большими градиентами. Это может привести к резкому ускорению обучения и улучшению прогнозных показателей

3. CatBoost – алгоритм, разработанный Yandex это гармоничное сочетание инноваций и эффективности, особенно когда дело доходит до работы с категориальными данными.

Исследование

решение проблемы «холодного старта»

1

Сформированы 2 группы

группы сформированы на основании количества просмотров видео пользователем и общего времени просмотра видео на видеохостинге Rutube

2

Валидация групп

оценка метрик на валидации каждой группы

3

Подбор пороговых значений

определение оптимального количества и времени просмотров для идентификации пользователя

Вывод

пользователи, которые просмотрели менее 10 видео, а также общая длительность просмотра видео у которых менее 10 минут плохо поддаются определению поло-возрастных характеристик

ЗНАЕМ ВСЕ

о пользователях Rutube