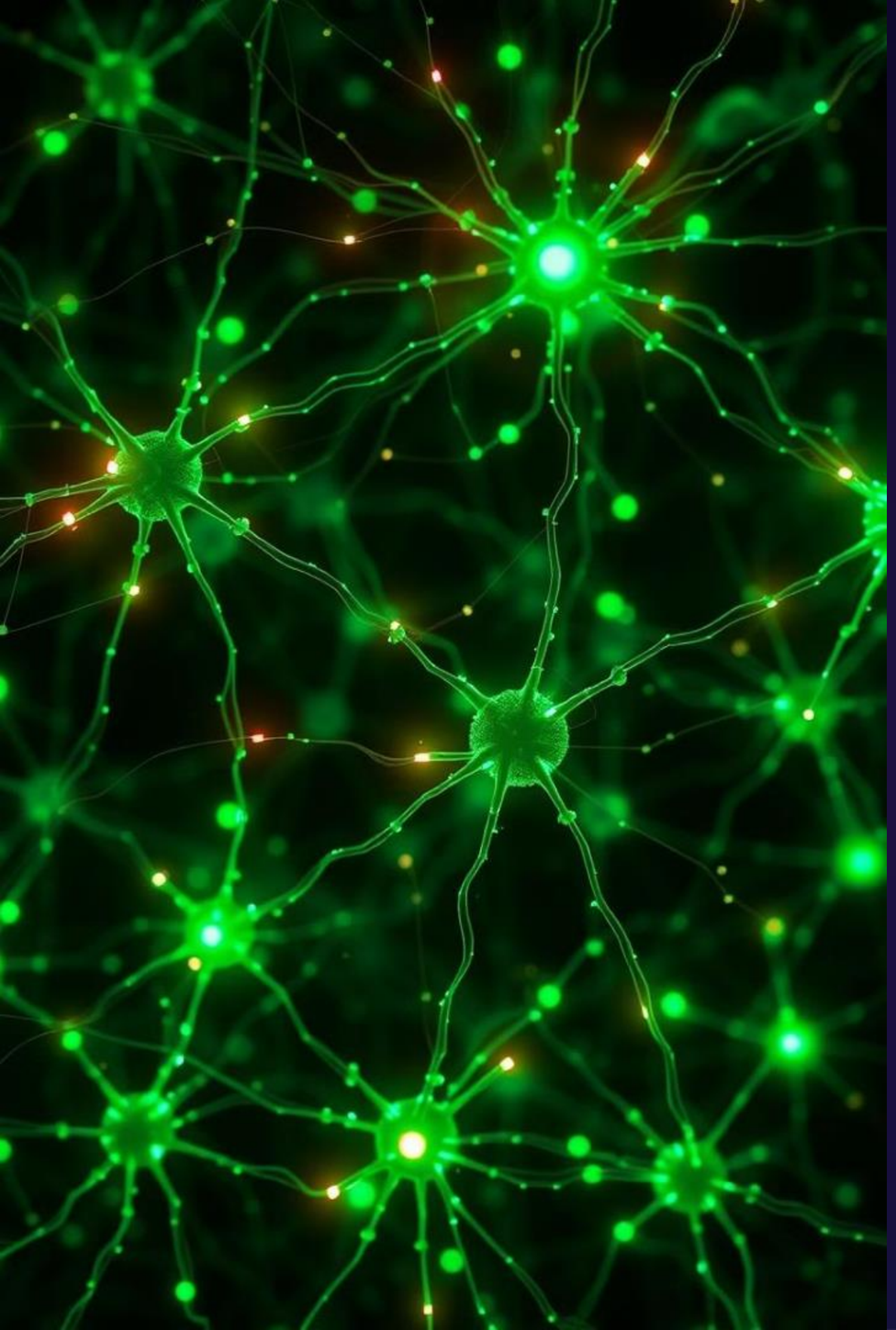


Предсказание социально-демографических характеристик пользователя

Проект Команды «ITимка»

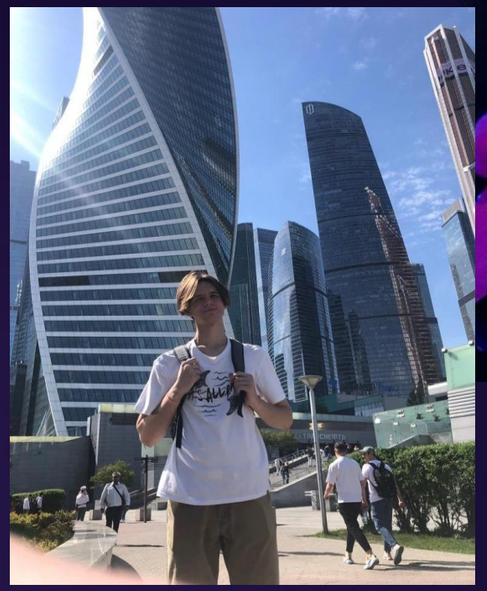
Стек технологий: Python, CatBoost

Стек библиотек: Pandas, NumPy, Pickle, Seaborn, Matplotlib, Scikit-learn, CatBoost, Pytz

Наш корабль идет навстречу Новому!

э

Команда с гордым названием «ITимка» соединяет не только университеты, но даже города! И от одного приключения к другому число неравнодушных к миру программирования растет все больше...



Наша команда:

Чудинова Полина – капитан команды, ML разработчик

Цветков Илья – аналитик данных

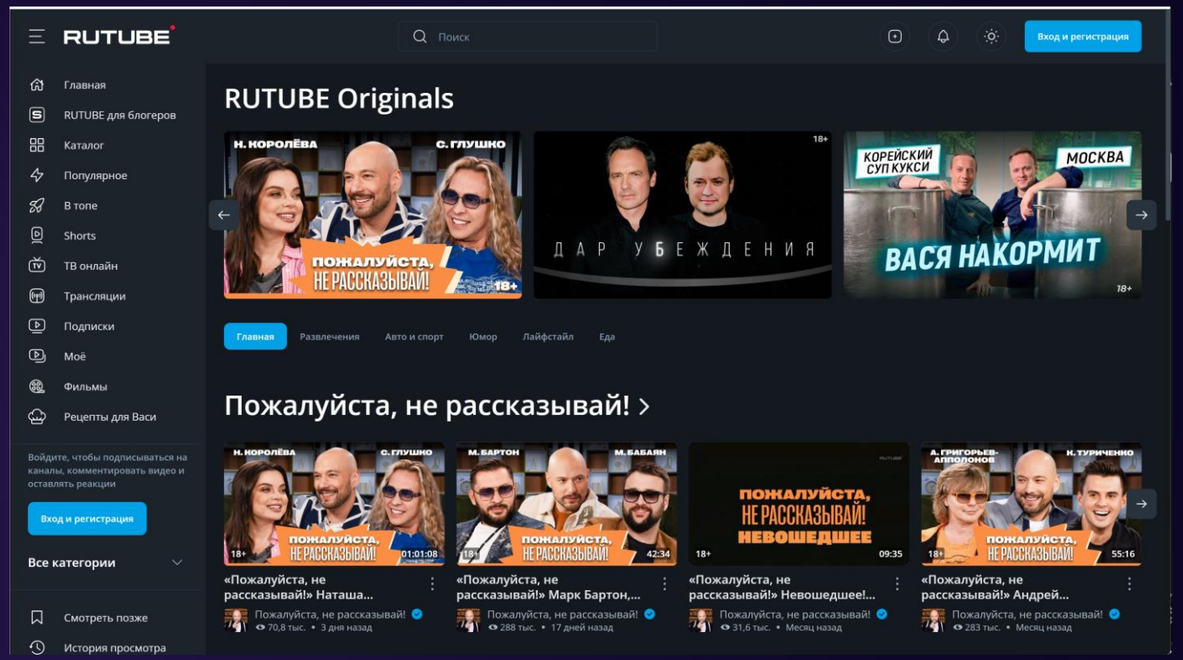
Митрофанов Андрей – исследователь

Кучеровский Макар – ML разработчик

Будкин Лев – аналитик данных

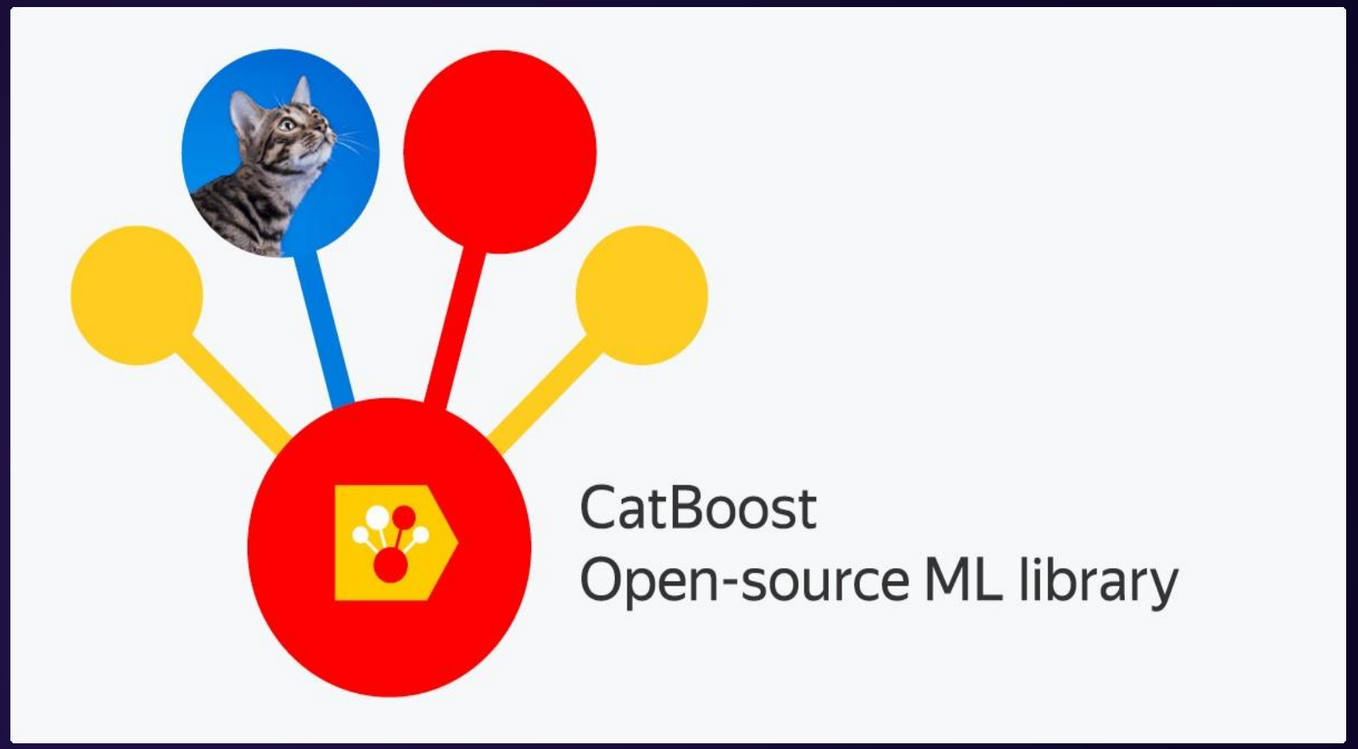
Проблематика

Многие пользователи RUTUBE не указывают возраст и пол, что **усложняет персонализацию рекомендаций и ухудшает пользовательский опыт**. Мы предлагаем разработать **модель**, которая на основе истории просмотров **будет предсказывать возраст и пол пользователей**.



Описание выбранной модели

- **CatBoost**: алгоритм градиентного бустинга.
- **Обработка категориальных признаков**: встроенные методы для работы с категориальными данными.
- **Высокая производительность**: эффективен на малых и больших датасетах.
- **Отсутствие переобучения**: оптимизация для минимизации этого риска.
- **Устойчивость к шуму**: надежность в условиях разнородных данных.
- **Простота использования**: минимальные настройки для достижения хороших результатов.



EDA (Exploratory Data Analysis)

Основная цель EDA — подготовить данные для построения моделей и более глубокого анализа

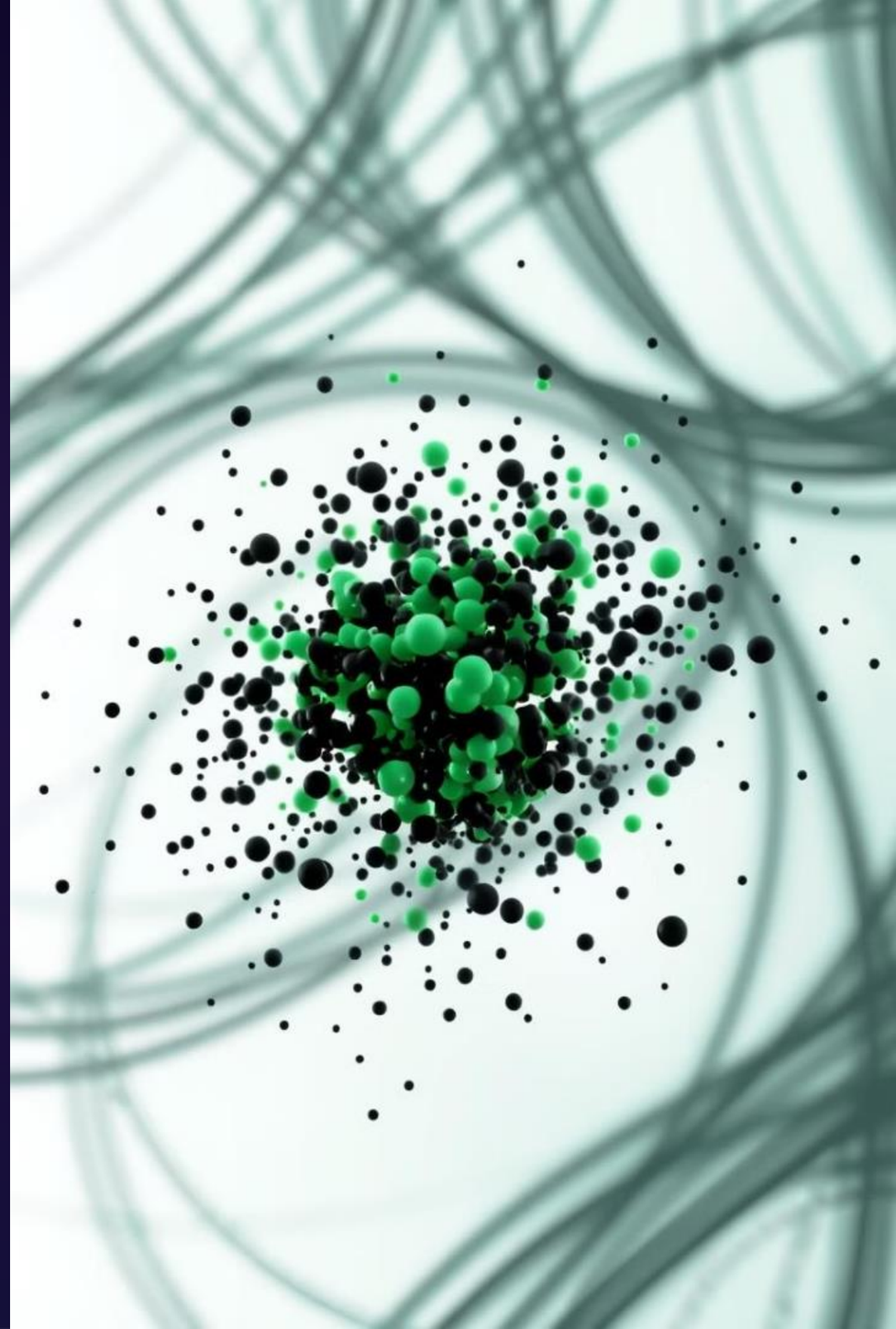
Основные задачи:

1. Описание данных (Исследование распределения переменных. Изучение типов данных и проверка наличия пропущенных значений)

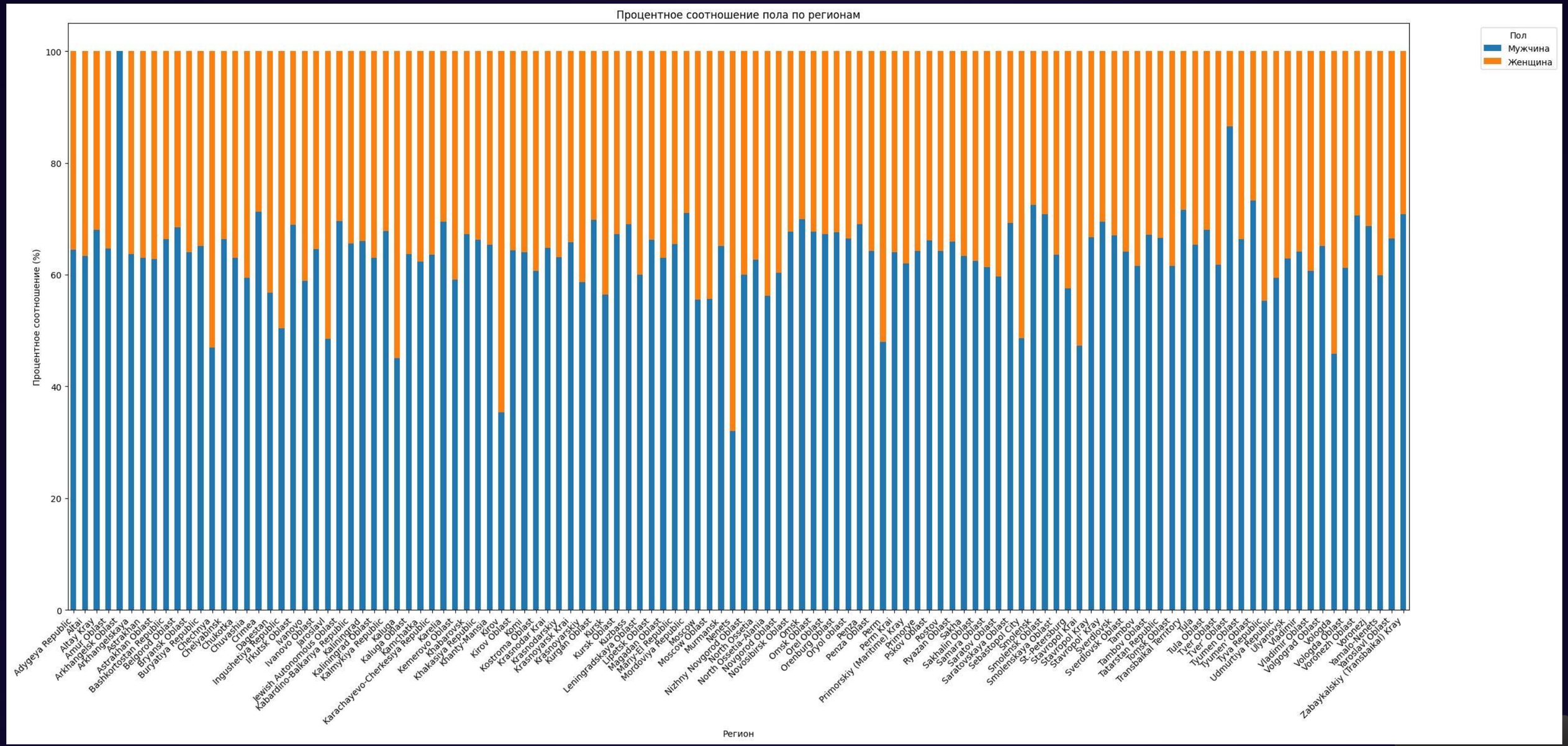
2. Выявление аномалий и выбросов

3. Взаимосвязи между переменными

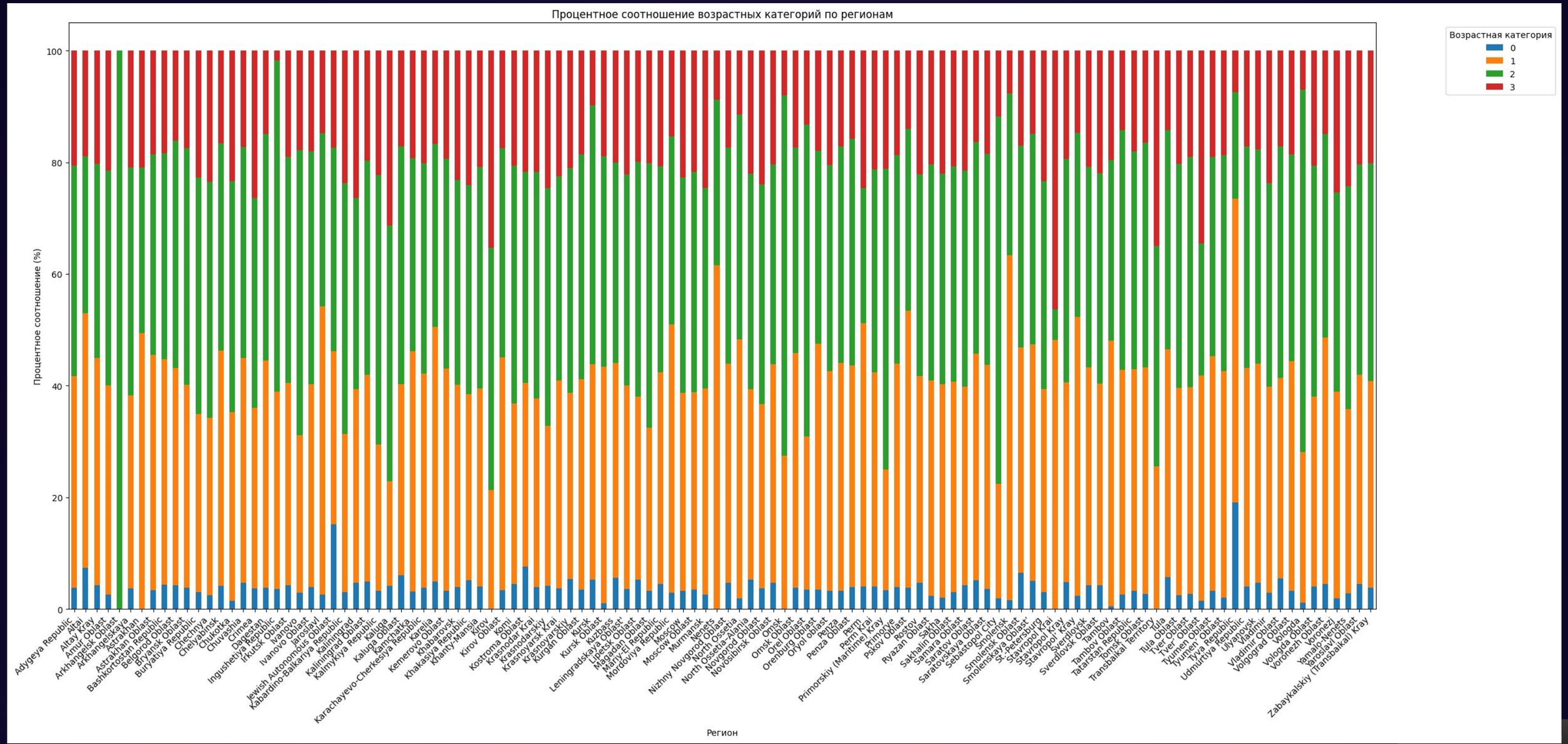
4. Визуализация данных



Процентное соотношение пола по регионам



Процентное соотношение возрастных категорий по регионам



Предобработка и обогащение датасета

1. **Загрузка и объединение данных:** Объединены данные, удалены строки с пропусками.
2. **Преобразование временной метки**: Время конвертировано в местное с учетом региона пользователя; извлечены новые признаки: hour, day_of_week, month, season и создан признак watch_time_category.
3. **Анализ активности пользователей**: Проведена группировка по viewer_uid для подсчета общего времени просмотра и количества видео; обогащение данных вероятностями пола и возрастных категорий по регионам и видео.

Создание синтетических данных для 0 и 3 категорий возраста

Был проведен анализ зависимости категории возраста от различных показателей. Далее было получены вероятности всех возможных вариантов пользователей для различных возрастных категорий

После этого мы случайным образом выбирали значения для различных атрибутов, таких как регион, тип устройства и возраст, а также генерировали временные метки и уникальные идентификаторы.

Вот Оно! Эврика!

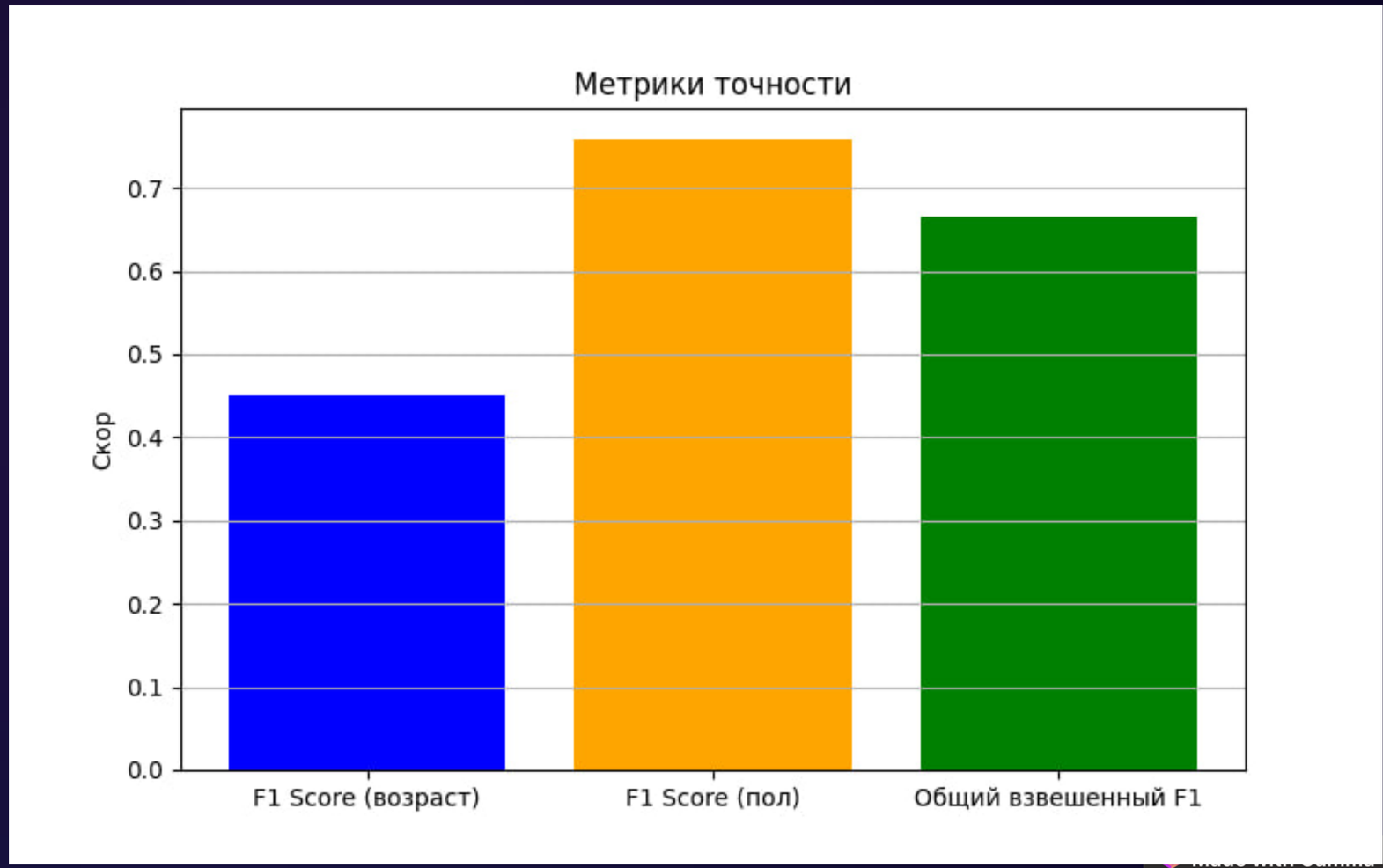
Наше решение

Наши модели продемонстрировали общую взвешенную точность **F1 = 68%**

F1 по возрасту = 44,9%

F1 по полу = 77,8%

На валидационных данных



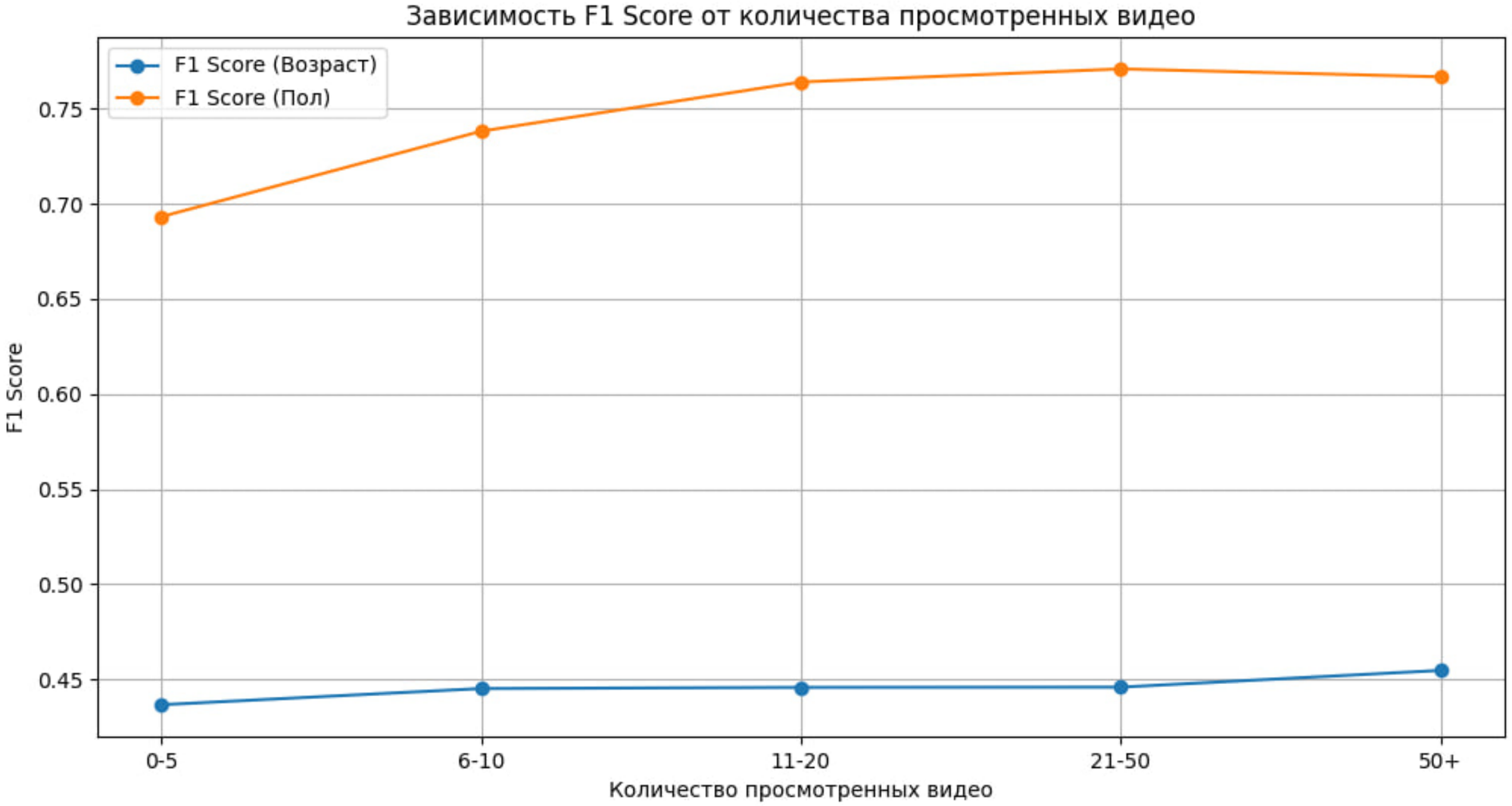
Анализ и оценка производительности моделей

Оценка моделей: Мы смогли оценить, как хорошо модели справляются с задачей классификации в зависимости от количества просмотренных видео. Это важно для понимания их эффективности в разных условиях.

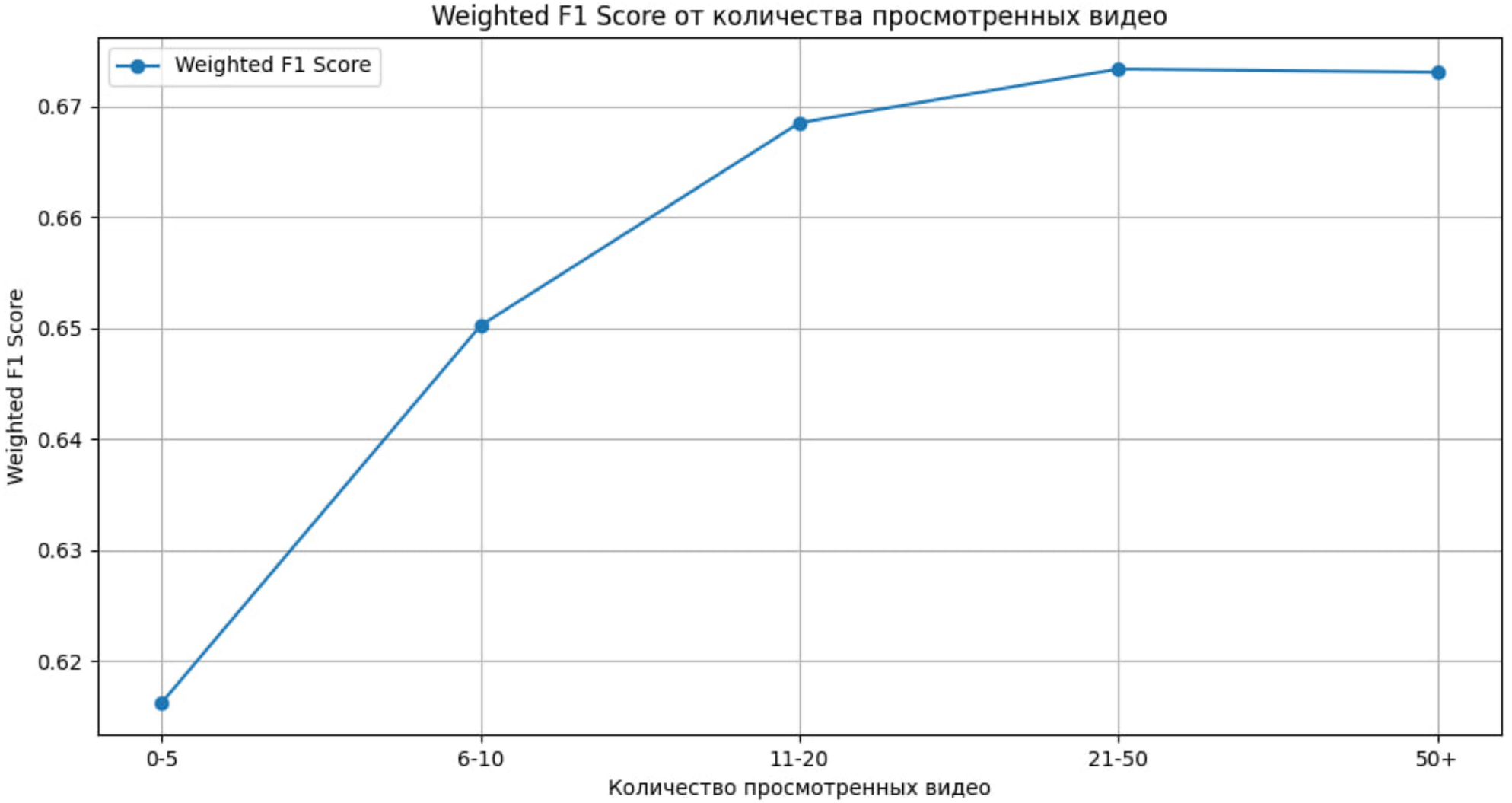
Анализ метрик: Мы смогли вычислить различные метрики (F1, precision, recall) для каждой группы пользователей. Это помогает выявить, в каких случаях модели работают лучше или хуже.

Определение порогов: Мы смогли определить минимальное количество просмотренных видео, необходимое для достижения хороших показателей моделей, что может быть полезно для оптимизации пользовательского опыта и бизнес-процессов.

Наше решение



Наше решение



Наше решение

Вывод: более выгодно и с точки зрения бизнеса, и с точки зрения точности моделей начинать предсказывать пол и категорию возраста незарегистрированного пользователя с **21-го просмотренного им** видео.

Спасибо за Внимание!

Капитан корабля – Чудинова Полина

Телефон: +7 (969)075-47-18

Телеграмм: @Lin_Lin2021