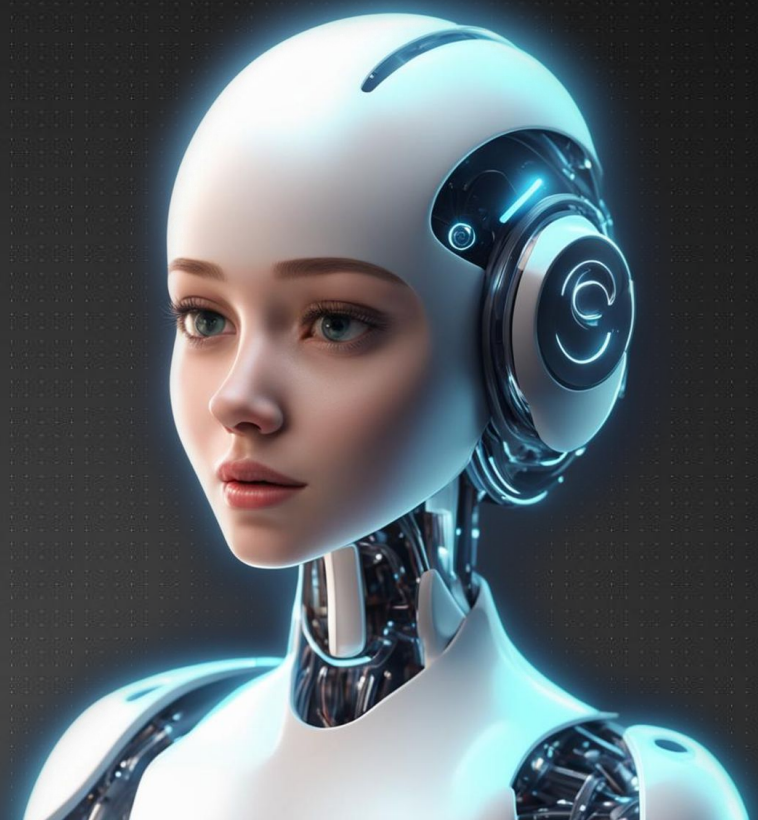
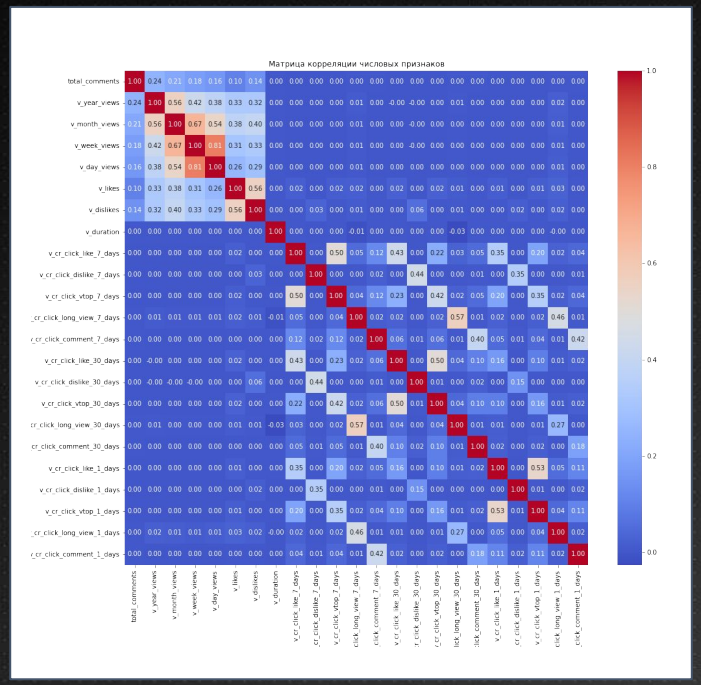
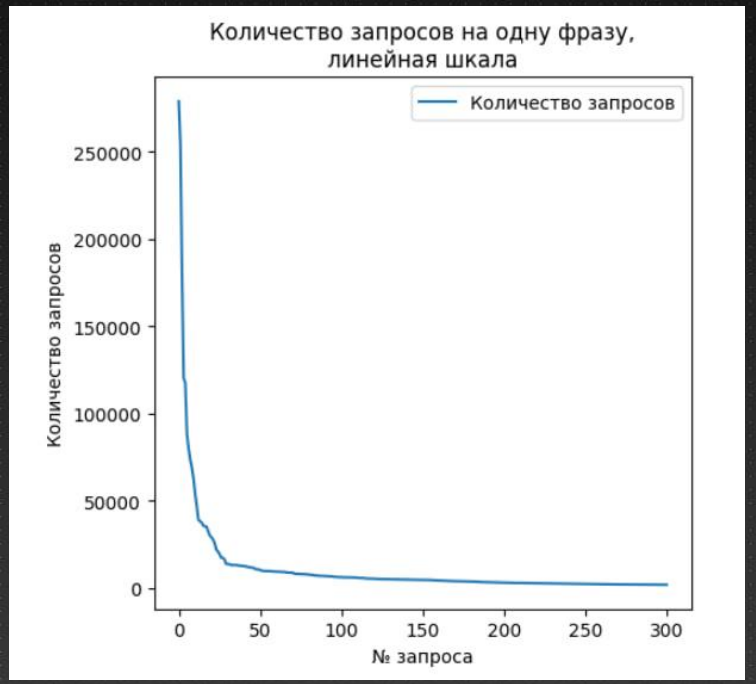
Поисковая система видео

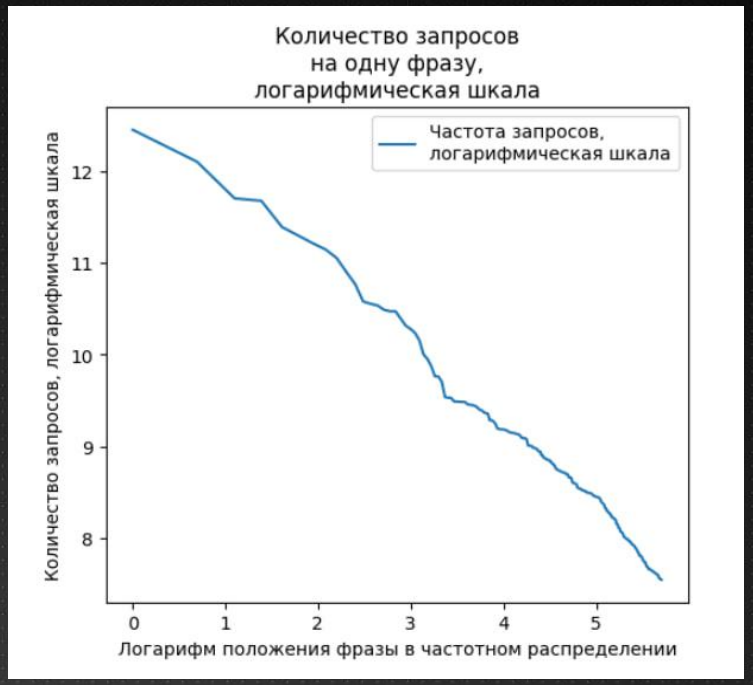
Стек технологий:
Python, Pandas, NumPy. CatBoostRanker, Faiss, SentenceTransformer, SHAP



Анализ данных







Распределение частотности слов по закону Ципфа

Исследование моделей для эмбедингов

Лучшая точность

cointegrated/rubert-tiny2

Это обновленная версия cointegrated/rubert-tiny : небольшого российского кодировщика на основе BERT с высококачественным встраиванием предложений.

Отличия от предыдущей версии заключаются в следующем:

- увеличен словарный запас: 83828 токенов вместо 29564;
- поддерживаемые последовательности большего размера: 2048 вместо 512;
- модель ориентирована только на русский язык.

LaBSE

distiluse-base-multilingual-cased-v2

paraphrase-multilingual-mpnet-base-v2

rubert-tiny2-retriever

paraphrase-multilingual-MiniLM-L12-v2

Проверка орфографии

1. Проверка правописания

Первая часть проверяет правописание и предлагает возможные варианты исправления. Базируется на свободном ПО Hunspell, одной из лучших программ для проверки орфографии.

→ превед -> привет

2. Разделение слипшихся слов.

Вторая часть ищет слипшиеся слова и разделяет их. Основана на библиотеке wordninja, исправленной для работы с кириллицей.

→ приветкакдела -> привет как дела

3. Исправление раскладки

Третья часть - простой алгоритм подстановки для исправления слов, написанных в неправильной раскладке.

→ ghbdtn -> привет

Технологии:

- Hunspell — свободное ПО для проверки орфографии языков со сложной системой словообразования и морфологией. Выбран за свою скорость и качество обработки грамматики.
- wordninja — свободное ПО для разделения слипшихся слов по морфологическим признакам. Изначально библиотека предназначена только для английского языка. Был доработан нами для работы с русским языком.

Ход решения

Исправление ошибок ввода

Проверка на неверную раскладку на клавиатуре с использованием spellchecke.

Разъединение слов, где забыли поставить пробел.

Исправление орфографических ошибок

Изменён алгоритм формирования данных

Изменение основано на свойстве распределении частотности фраз по закону Ципфа.

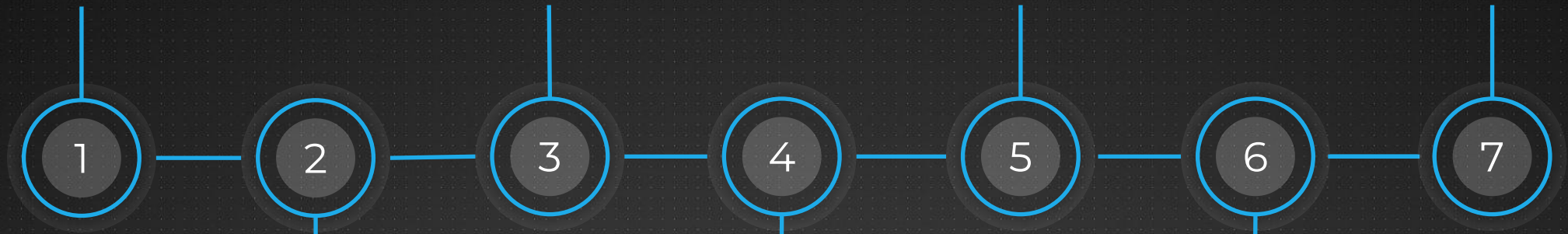
Оценка и результаты:

Вычислены метрики NDCG для тренировочного, валидационного и тестового наборов данных.

Проведена оценка результатов с учетом различных top-k значений.

Код решения

Выложен на github, готова презентация, тизер



Features Engineering

Произведено обогащение данных.

Удалены выбросы по квантилям просмотров и лайков.

Созданы новые признаки, такие как соотношение лайков и дизлайков, отношение просмотров к лайкам и другие.

Добавлена информация о возрасте канала, возрасте видео, времени публикации и другие характеристики.

Обучение модели ранжирования:

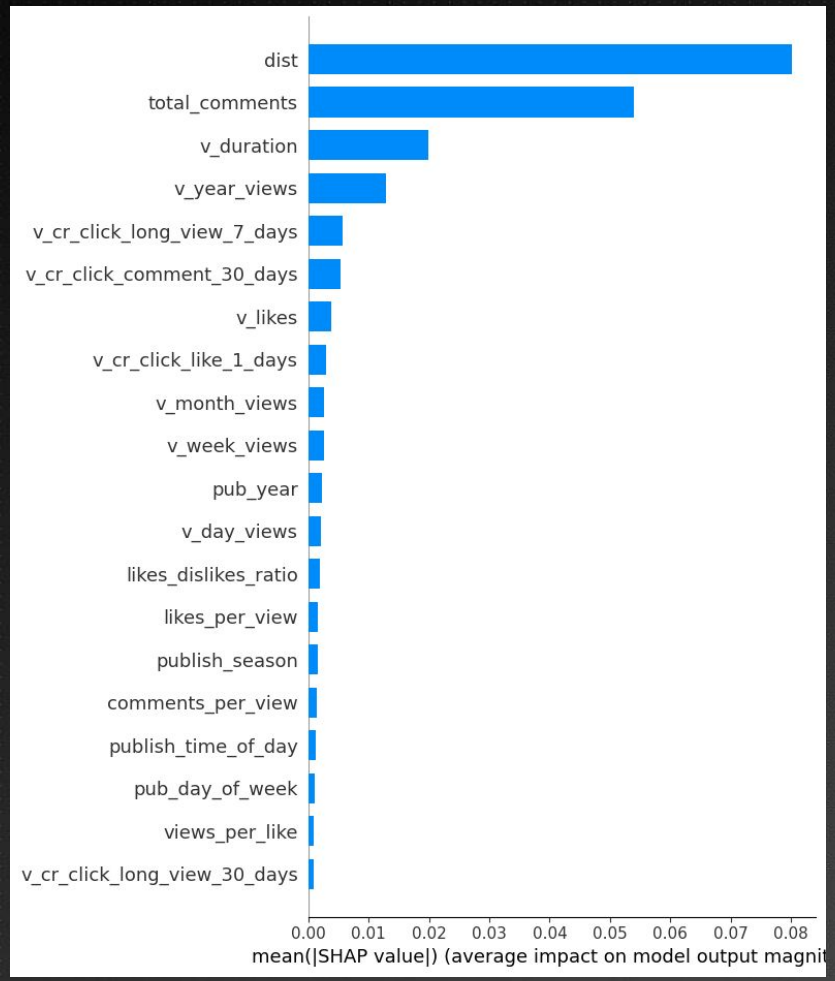
Использован CatBoostRanker для обучения модели с учетом подготовленных данных.

Модель обучена на тренировочном наборе с валидацией, используя метрику QueryRMSE.

Интерпретация features:

Использована библиотека SHAP для оценки важности признаков

Интерпретация влияния features на точность модели



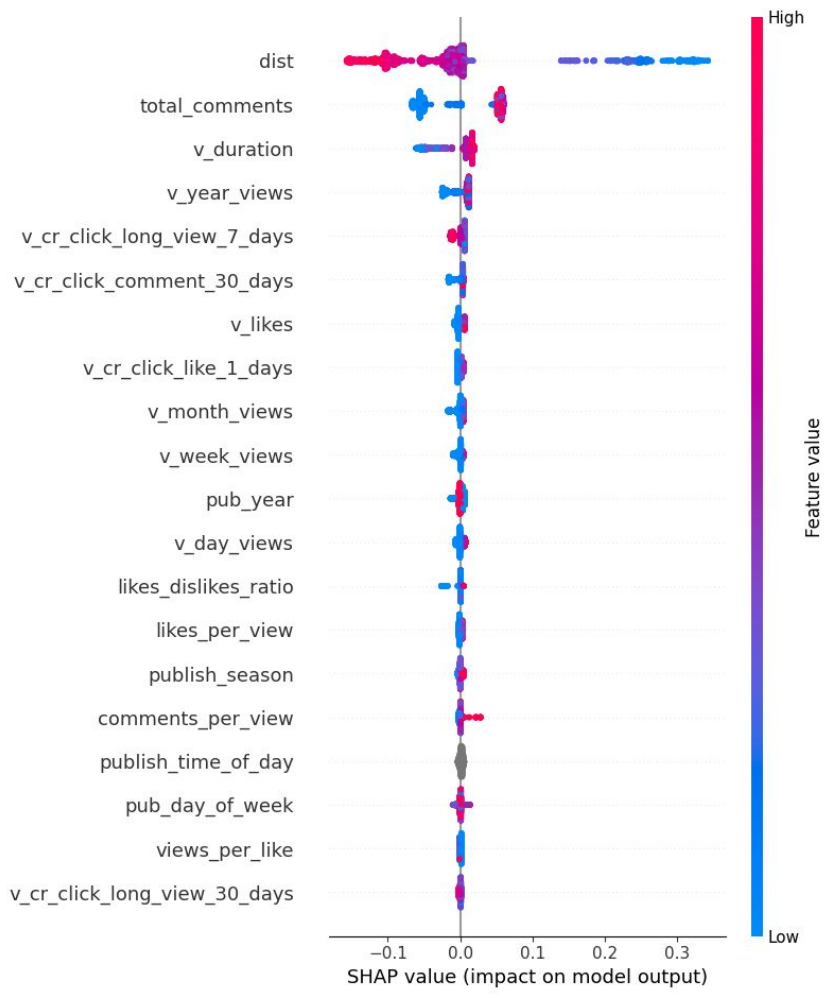
Здесь мы видим топ 20 признаков, имеющих наибольшее значение на предсказание модели.

Важным признаком является дистанция между векторами, их близость.

Данный признак отсутствовал в baseline и был добавлен нами.

Распределение признаков заметно меняется до и после фильтрации по квантилям

Интерпретация влияния features на точность модели



Здесь представлены показатели SHAP (SHapley Additive exPlanations):

Значения слева от центральной вертикальной линии — это негативные примеры, справа — позитивные по матрице ошибок предиктивной модели;

Толщина линии прямо пропорциональна количеству точек наблюдения;

Чем краснее точки, тем большее значение имеет признак в этой точке. Серые признаки - неизмеримы.

Дальнейшее развитие

1. Персонализация
   Добавляем вектор пользователя в модель для персонализации выдачи.

2. Учёт “последнего”, “долгого”, “быстрого” клика
   Механизм, аналогичный как в поисковых системах

Наша команда

Леонид Чесников

- DS, AI-engineer
- @FatherKomm

Иван Василевский

- Data scientist
- @Mrghoste0

Владимир Губин

- Data scientist
- @FiveCharactersName

Артём Качалкин

- Data scientist
- @anarakinson

Алексей Домненко

- DS, fullstack
- @domnenko_a_n