Анализ поведения пользователей на сайте «СберАвтоподписка»

Команда хакатона №4

SKILLFACTORY

Участники команды

Екатерина Дорохова

Тимлид, проект решения

Ольга Омельянчук

Анализ данных

Иван Сойко

Предобработка данных

Владимир Гаврилов

ML-инженерия

Андрей Максаков

Frontend-интеграция

Цель проекта

построить модель, которая прогнозирует вероятность совершения пользователем целевого действия (заявка, звонок, клик на кнопку) по данным веб-аналитики.

Функционал модели:

- Оценка эффективности каналов трафика;
- Выявление поведенческих и технических признаков, связанных с конверсией;
- Вывод предсказания в веб-интерфейс для использования специалистами по маркетингу и UX.

Задача и стэк технологий

Задачи

• Предсказать, совершит ли пользователь целевое действие на сайте.
• Использовать реальные данные сессий и событий из Google Analytics*.

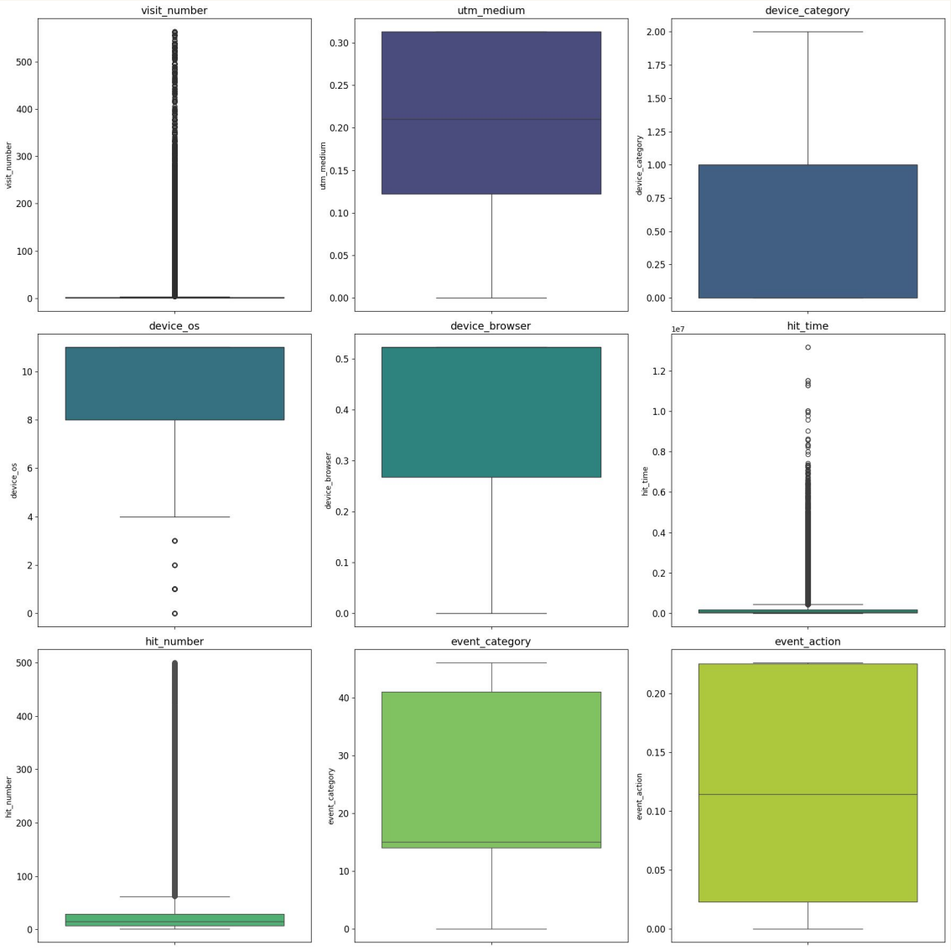
Target - одно из целевых действий (event_category и event_action в ga_hits.pkl) - примеры:

- submit_form
- request_callback
- send_application
- order_button_click
- lead_conversion
- thank_you_page_view

*Источники: 1) ga_sessions.pkl - источник, устройство, география, параметры utm, 2) ga_hits.pkl - действия, поведенческие события.

Стэк проекта: pandas, matplotlib, numpy, seaborn, scikit-learn, CatBoost, streamlit (web-интерфейс), category_encoders (для кодирования признаков)

Этап 1. Предобработка и объединение данных



Чтение и изучение данных

Файлы ga_sessions.pkl и ga_hits.pkl - анализ структуры и содержания.

Фильтрация уникальных сессий

Обнаружение и обработка пропусков, дубликатов, аномалий.

Объединение по session_id

Данные из ga_sessions и ga_hits объединены по идентификатору сессии (session_id).

Отбор значимых связей, построение гипотез

Исследование признаков в разрезе связи с целевой переменной.

Результаты предобработки: структура и полнота данных в целом удовлетворительные, пропуски и некорректные значения локальны и некритичны, числовые и категориальные типы данных определены корректно, дубликатов не обнаружено. Выделена переменная target на основе целевых event_action.

Этап 2. Разведочный анализ данных (EDA) и визуализация

Анализ распределений

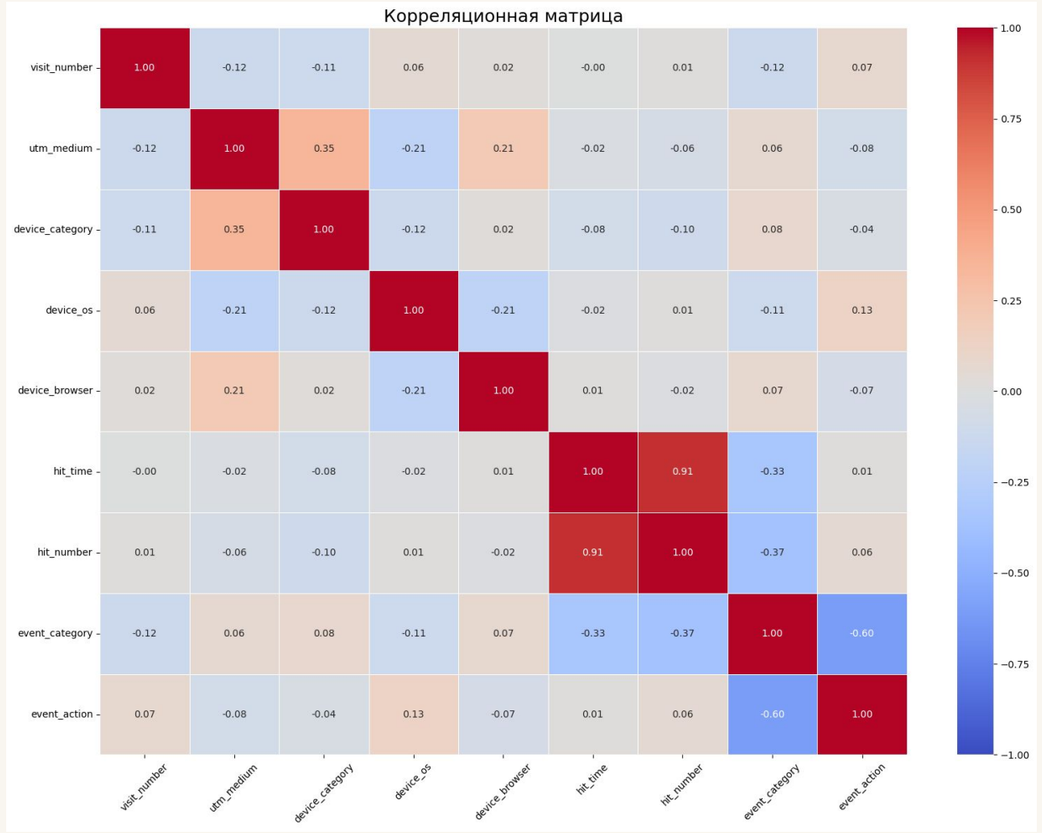
Построены гистограммы для каждого признака для визуальной оценки их распределения и выявления аномалий, групповые барплоты по target, Boxplot'ы

Выявление выбросов и редких категорий

Например, в признаке visit_number были обнаружены выбросы и идентифицированы редкие категории.

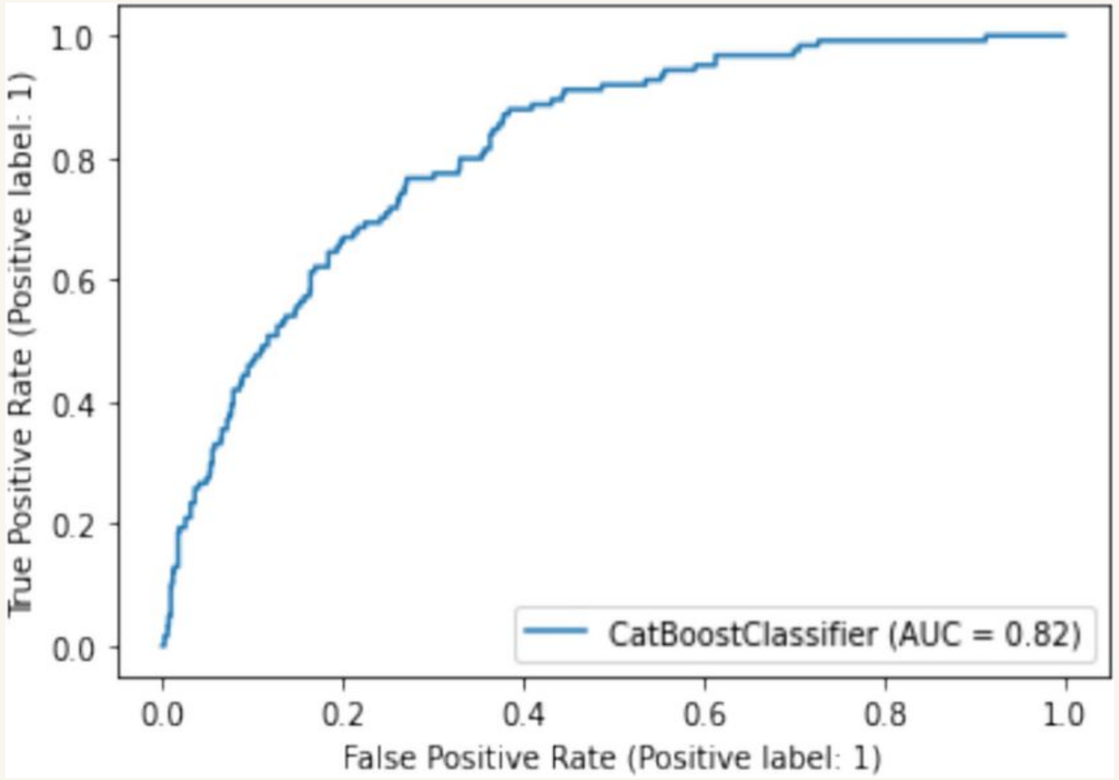
Связь с конверсией

Построена тепловая карта корреляций. Выявлена значимая связь между показателем конверсии и такими признаками, как источник трафика (utm_source), город (geo_city) и категория устройства (device_category).



Наблюдения: почти все сессии из России, в действиях пользователей (event_action, event_category) есть сильные лидеры и длинный "хвост" редких значений, кампании (utm_campaign) и источники (utm_source) демонстрируют сильную неравномерность (преобладают 3 значения, остальное — шум), признаки visit_number, hit_number, hit_time скошены вправо + много выбросов - вероятно, связано с поведением "лояльных" пользователей.

Этап 3. Моделирование с CatBoostClassifier



Выбор модели

В качестве основной модели был выбран **CatBoostClassifier** (градиентный бустинг) благодаря его способности напрямую работать с категориальными признаками без необходимости ручного кодирования и тем самым учесть большое количество категорий. Другое важное для задачи преимущество - устойчивость модели к дисбалансу классов.

Процесс моделирования

Были отобраны 15 ключевых признаков, данные разделены на обучающую и тестовую выборки.

Подбор оптимальных гиперпараметров для обучения модели проводился с помощью `RandomizedSearchCV`.

Оценка качества модели выполнялась по метрике ROC-AUC.

Достигнут показатель **ROC-AUC = 0.82** - то есть хорошая прогностическая способности модели - она различает пользователей, которые совершат целевое действие, от тех, кто этого не сделает.

Этап 4. Реализация web-приложения на Streamlit

Ввод признаков

Пользователь вводит значения 15 признаков сессии через интерактивные формы.

Обработка данных

Введенные данные передаются в модель.

Отображение результата

Вероятность конверсии выводится пользователю в реальном времени.

Прогнозирование

Модель рассчитывает вероятность совершения целевого действия.

- Разработано веб-приложение на **Streamlit**, позволяющее пользователю взаимодействовать с построенной моделью.
- Оно запускается локально (`localhost`) и обеспечивает быструю интеграцию с сериализованной моделью (`model.pkl`).
- Время отклика не превышает 3 секунд, что делает приложение быстрым и отзывчивым.

Результаты проекта: от модели к практическому применению

Определены факторы конверсии

Выявлены основные факторы, влияющие на вероятность целевого действия: канал привлечения, устройство и город пользователя.

Модель бинарной классификации

Разработана надёжная производительная модель (ROC-AUC = 0.82), с высокой эффективностью прогнозирования конверсии.

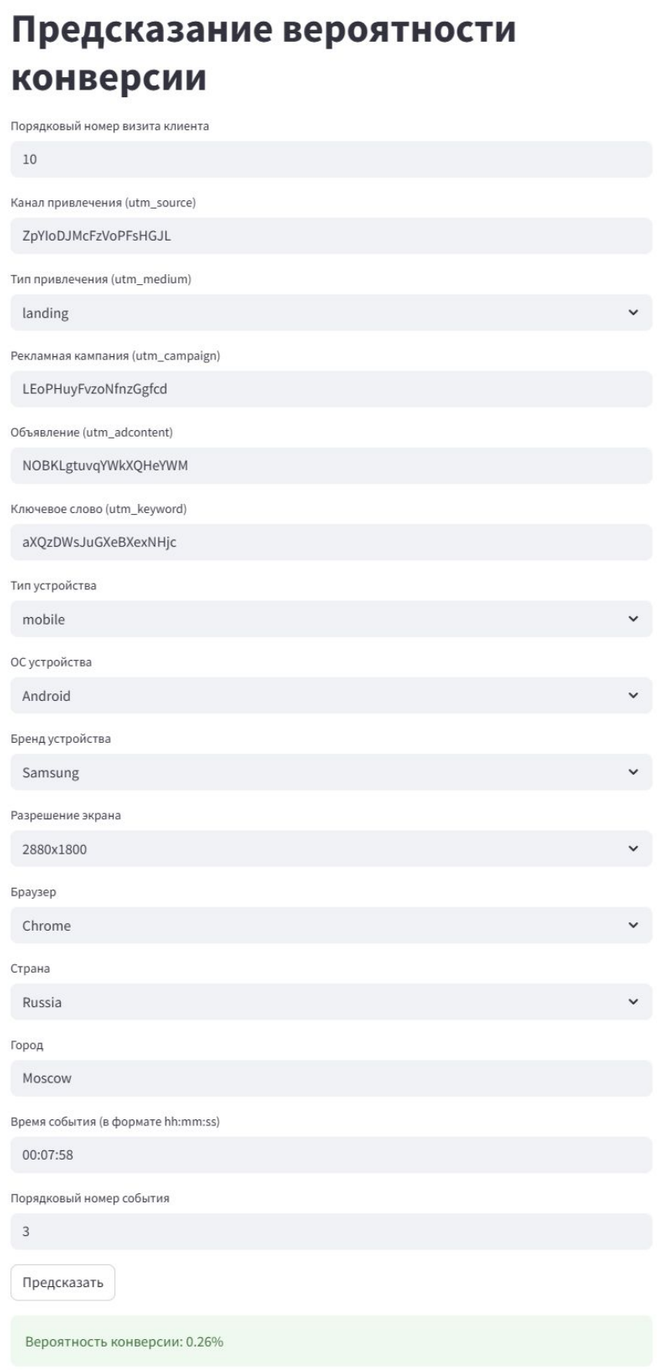
Веб-калькулятор конверсии

Реализован интерактивный веб-интерфейс для быстрого расчёта вероятности конверсии.

Проект демонстрирует, как поведенческие данные могут быть использованы для:

- улучшения пользовательского опыта (UX),
- адаптации рекламных кампаний под целевую аудиторию,
- повышения общей эффективности веб-сайта.

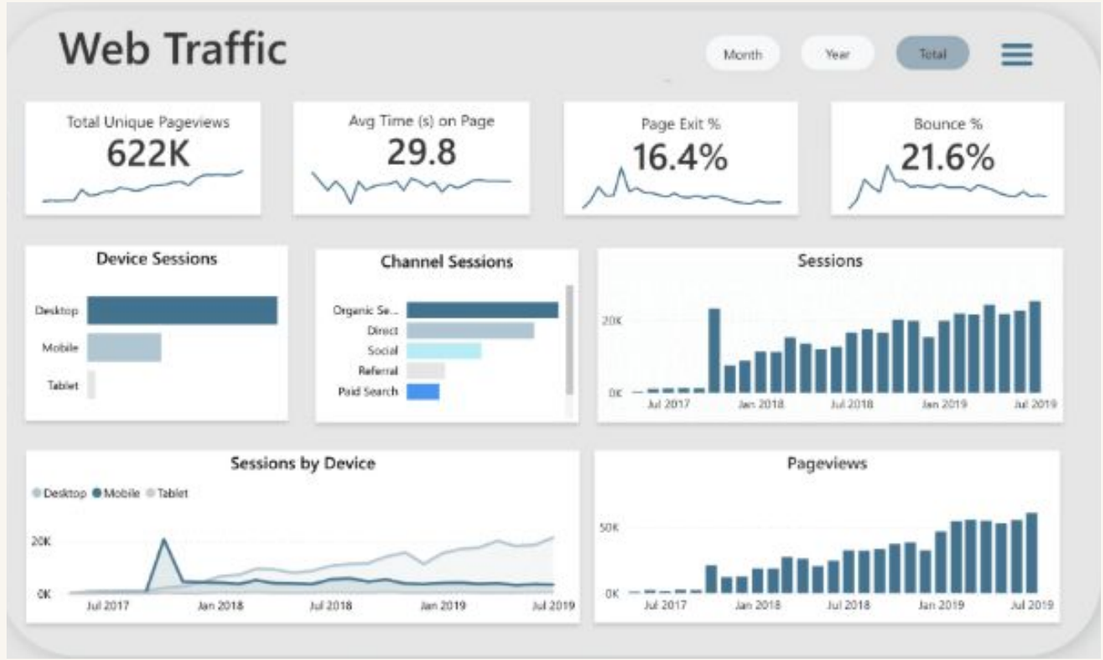
Инструмент может быть основой для принятия управленческих решений, направленных на удержание и возврат клиентов, выявление точек роста, снижение затрат и увеличение доходов.



Выводы и перспективы развития проекта

Что реализовано:

- Построена и протестирована интерпретируемая модель.
- Создан визуальный EDA с понятными выводами.
- Реализован работающий web-интерфейс.



Что можно улучшить:

1. Объединение категориальных признаков по смыслу:

Использование семантической группировки значений (например, через SentenceTransformer) позволит сократить количество уникальных значений и устранить дублирующиеся по смыслу категориальные значения. Это откроет возможность использовать классические модели (Logistic Regression, Decision Tree) без потери качества и обеспечит контроль стабильности модели при изменении входных данных.

2. Дополнительные признаки:

Добавление поведенческих фичей (глубина просмотра, длительность сессии, порядок событий) для повышения точности модели.

3. Расширение интерфейса:

Добавление функционала пакетной обработки, генерации отчётов и экспорта данных в Excel/BI, создание дашбордов.

4. Интеграция с CRM:

Связывание модели с CRM для онлайн-оценки конверсий и автоматизации процессов.

5. Настройка аллертов:

Добавление отправки уведомлений пользователю при критическом изменении показателей и тенденций.

Интересные наблюдения и выводы

Зависимость конверсии от географии и utm-источников

Модель наглядно продемонстрировала, что местоположение пользователя и канал, через который он пришёл на сайт, являются сильными предикторами целевого действия. Это подчеркивает важность геотаргетинга и оптимизации источников трафика.

Эффективность CatBoost

Использование CatBoost позволило значительно упростить процесс предобработки данных, так как эта модель отлично работает напрямую с категориальными признаками, избавляя от необходимости ручного кодирования и сложной инженерии признаков.

Скорость разработки с Streamlit

Streamlit оказался незаменимым инструментом для быстрой разработки и демонстрации интерактивного веб-приложения. Он позволил в короткие сроки создать функциональный интерфейс для показа работы ML-модели в реальном времени.

Интересно наблюдать, как даже короткие сессии с правильно подобранными признаками могут давать высокую вероятность конверсии. Это подтверждает, что качество данных и релевантность признаков играют ключевую роль в точности прогнозирования.

Спасибо за внимание!

Команда хакатона №4

SKILLFACTORY