Задача VK-1

Команда УрФУ №13

Архипов Данил
Максимова Наталья
Сергеев Владимир
Уткевич Илона

Проект: “Предсказание пола пользователей соцсетей с использованием машинного обучения”.

Цель: Разработать модель для предсказания пола пользователя.

Задача: Помочь рекламодателям эффективно таргетировать праздничные предложения.

Почему это важно: Улучшение качества рекламы, удовлетворение клиентов, увеличение дохода.

Обзор данных

- Датасет с пользовательскими запросами.
- Географическая информация.
- Векторные представления рефереров.

Предварительный анализ данных

Проверка на пропуски

- **Какие столбцы содержат пропуски:**
  - `user_agent`: 1 пропущенное значение.
  - `region_id`: 50,620 пропущенных значений.
- **Обработка пропусков:**
  - Пропуски в `user_agent` были устранены путем удаления строки, так как это минимально влияет на объем данных.
  - Для `region_id` пропуски не заполнялись, так как эта информация может быть использована как отдельная категория (например, "неизвестно").

Распределение целевой переменной

- Целевая переменная `target` (пол):
  - Класс 0 (мужчины): 313,572 записи.
  - Класс 1 (женщины): 287,717 записи.
  - Баланс между классами близок к равному, что снижает необходимость использования методов балансировки данных.

Особенности распределения фичей

- Большинство числовых признаков имеют распределения, близкие к нормальному.
  - Подтверждено визуализацией гистограмм и наложением нормального распределения.

Выявленные закономерности

Распределение ОС пользователей

- Распределение популярных операционных систем (`os`):
  - Android: 389,999 записей.
  - Windows: 136,096 записей.
  - iOS: 16,752 записей.
  - Менее популярные ОС: Mac OS X, Linux, Tizen и другие.

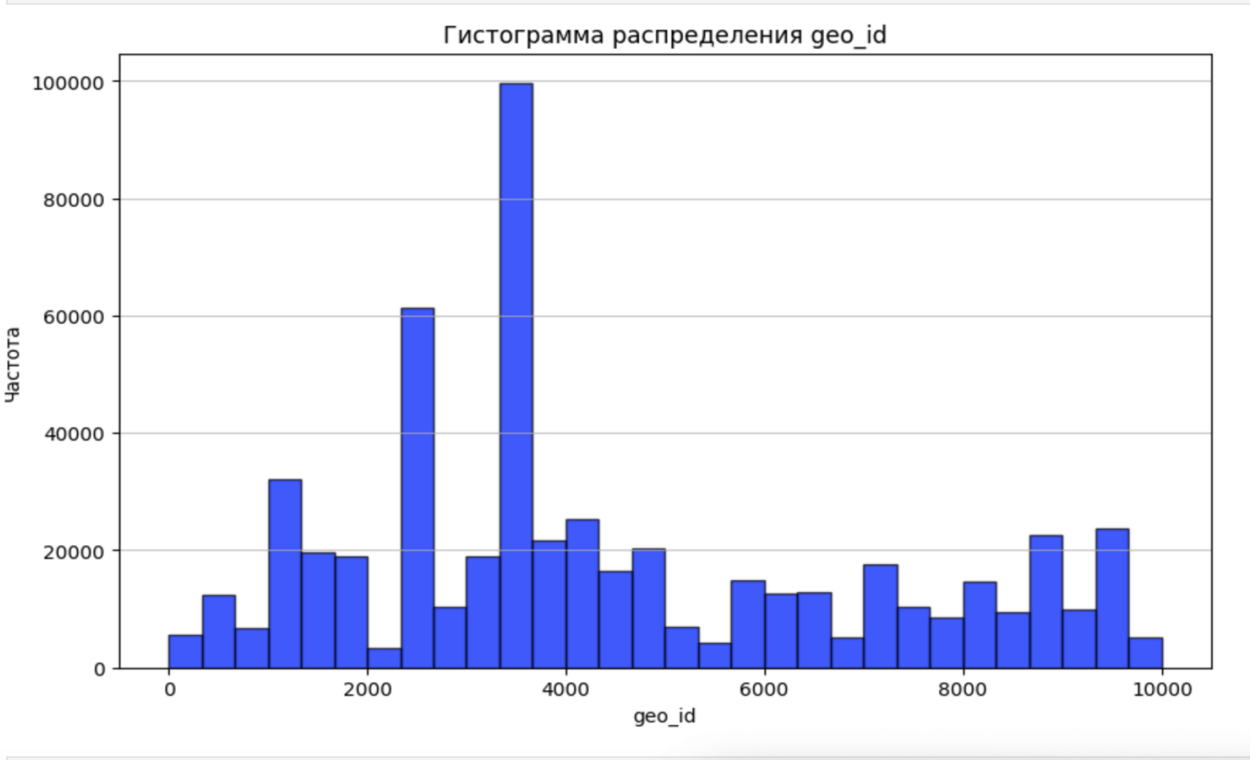
Распределение версий ОС

- Топ-5 популярных версий ОС:
  - `10`: 335,740 записей.
  - `13`: 53,071 записей.
  - `11`: 45,048 записей.
  - `12`: 34,288 записей.
  - `7`: 18,211 записей.

Распределение ОС пользователей

- В выборке присутствуют популярные ОС, такие как Android, Windows, iOS.
- Анализ `os_version` показывает значительное разнообразие версий, что может указывать на технические особенности пользователей.

Построена гистограмма, показывающая частотное распределение



Выбор моделей и гипотезы

Выбор моделей и гипотезы

Используемые алгоритмы

1. **CatBoostClassifier**:
    - Оптимизирован для работы с категориальными фичами.
    - Настроены параметры, включая количество итераций и глубину деревьев.
2. **RandomForestClassifier**:
    - Применяется для числовых данных.
    - Используется как базовый метод для сравнения.

Ожидания и гипотезы

- Существуют прямые зависимости между характеристиками пользователей (локация, ОС) и их полом.
- Точные модели способны учитывать сложные взаимосвязи между параметрами.

Результаты моделей

CatBoostClassifier

- **Лучшие метрики:**
  - Log Loss: `0.4404`
  - Accuracy: `0.4747`
  - Precision: `0.4993`
  - Recall: `0.4854`
  - F1 Score: `0.4922`
  - ROC AUC: `0.4845`
- **Выводы:**
  - Модель показала хорошие результаты для сбалансированного подхода к работе с категориальными фичами.

RandomForestClassifier

- **Лучшие метрики:**
  - Accuracy: `0.3367`
  - Precision: `0.3552`
  - Recall: `0.3244`
  - F1 Score: `0.3391`
  - ROC AUC: `0.3202`
- **Выводы:**
  - Простая модель менее эффективна при работе с текущими данными.

Оптимизация модели (Optuna)

- Применена библиотека **Optuna** для подбора гиперпараметров CatBoost.
- Оптимизированы:
  - `iterations`
  - `learning_rate`
  - `depth`
- Минимизировано значение Log Loss до `0.4404`.

Сравнение моделей

Метрика	CatBoost	Random Forest
Accuracy	0.3367	0.4747
Precision	0.3552	0.4993
Recall	0.3244	0.4854
F1 Score	0.3391	0.4922
ROC AUC	0.3202	0.4845