МФТИ

Отчет команды 30

ВК Задача 1

Название и состав команды

Состав команды и роли участников

- Екатерина Чуйко – Анализ существующих решений, выбор моделей
- Бушуева Татьяна – Предобработка данных, анализ и поиск инсайтов
- Семен Шульга – Обучение моделей
- Егор Лапенков – Обучение моделей
- Арсалан Алексеев – Обучение моделей

Контекст задачи

Задача: Предсказание пола пользователей на основе предоставленных данных

Ссылка на задачу: cups.online/ru/tasks/1923

Ключевые особенности данных:

- Табличные данные: geo_id, user_agent, referer_vectors и другие.
- Целевой признак: пол пользователя (0 или 1).
- Равномерное распределение таргета.

Проведенные исследования

1. Анализ существующих решений:

Использованы источники научных публикаций, проектов и открытых библиотек.

Рассмотрены три ключевых подхода:

- *Логистическая регрессия*: простая, быстро обучаемая модель.
- *Градиентный бустинг*: точность, работа с нелинейностями.
- *CatBoost*: эффективен для категориальных признаков.
- *TabNet*: точность, работа с нелинейными признаками за счет глубокого обучения, интерпретируемость

2. Обзор литературы:

- *Mortrey/gender_predict* – GitHub
- *Osmanov-Bairam/Gender-detection – GitHub*
- *CatBoost documentation*
- *Habr* – Нейросети для предсказания пола и возраста
- *https://habr.com/ru/articles/534186/, https://pypi.org/,*

и другие источники

Предобработка данных

1. Основные этапы обработки:
- Удаление пропусков и дубликатов.
- Добавление временных признаков (день недели, время суток).
- Кодирование категориальных признаков.

2. Исследование данных:
- Анализ временных пиков активности.
- Географическое распределение пользователей.

3. Выводы:
- Целевая переменная сбалансирована.
- Большинство пользователей из одной страны.
- Поведение пользователей различается в зависимости от пола.

Предобработка данных

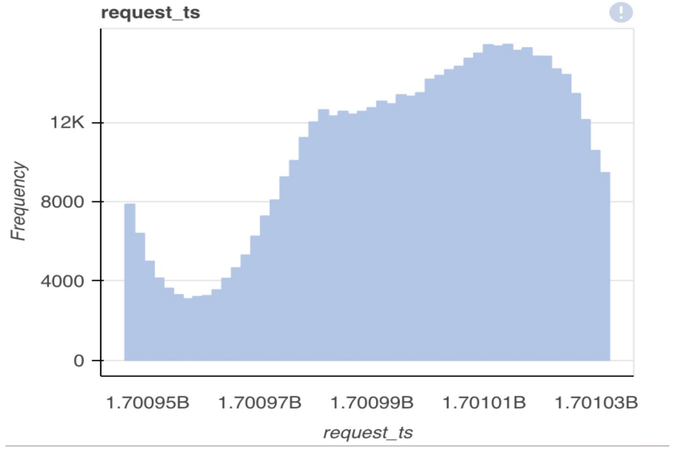

Рис 1. Промежуток с максимальным количеством запросов

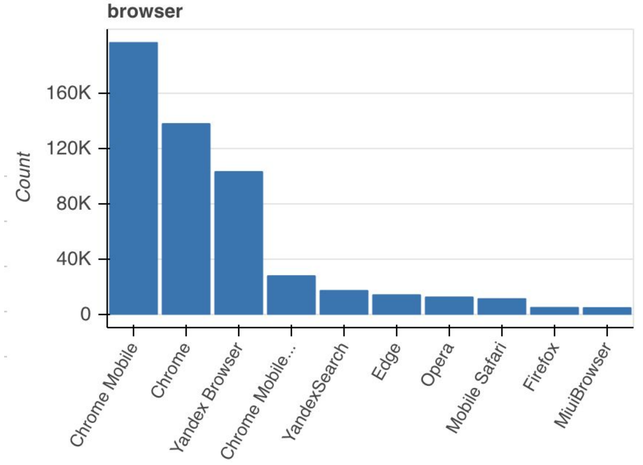

Рис 3. Использование пользователями браузеров

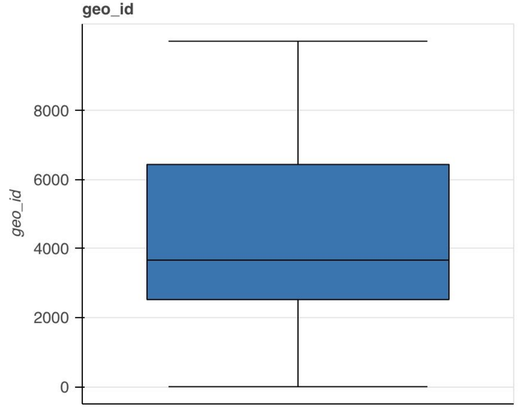

Рис 2. Медиальное значение geo_id

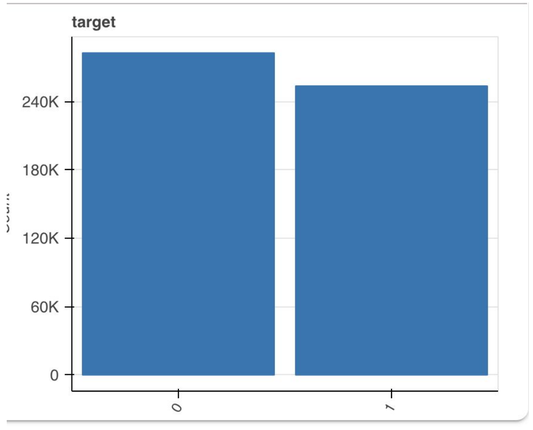

Рис 4. Распределение таргета 1 и 0

Предобработка данных

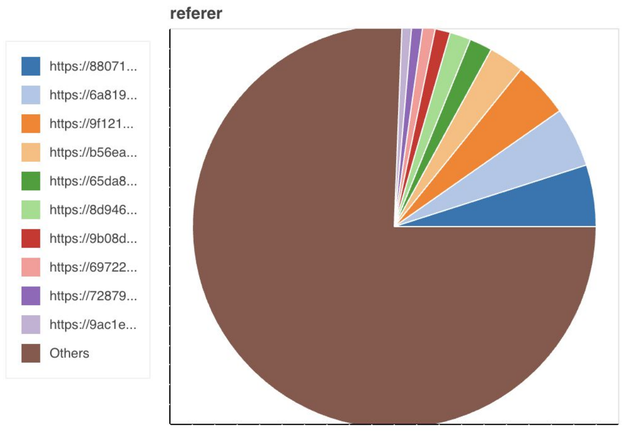

Рис 5. Разнообразие ссылок в данных

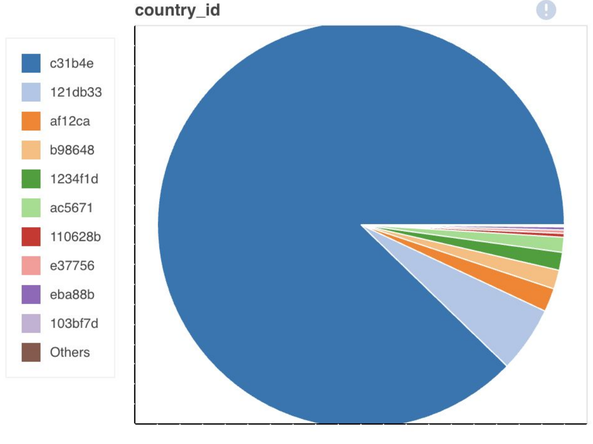

Рис 6. Распределение стран пользователей

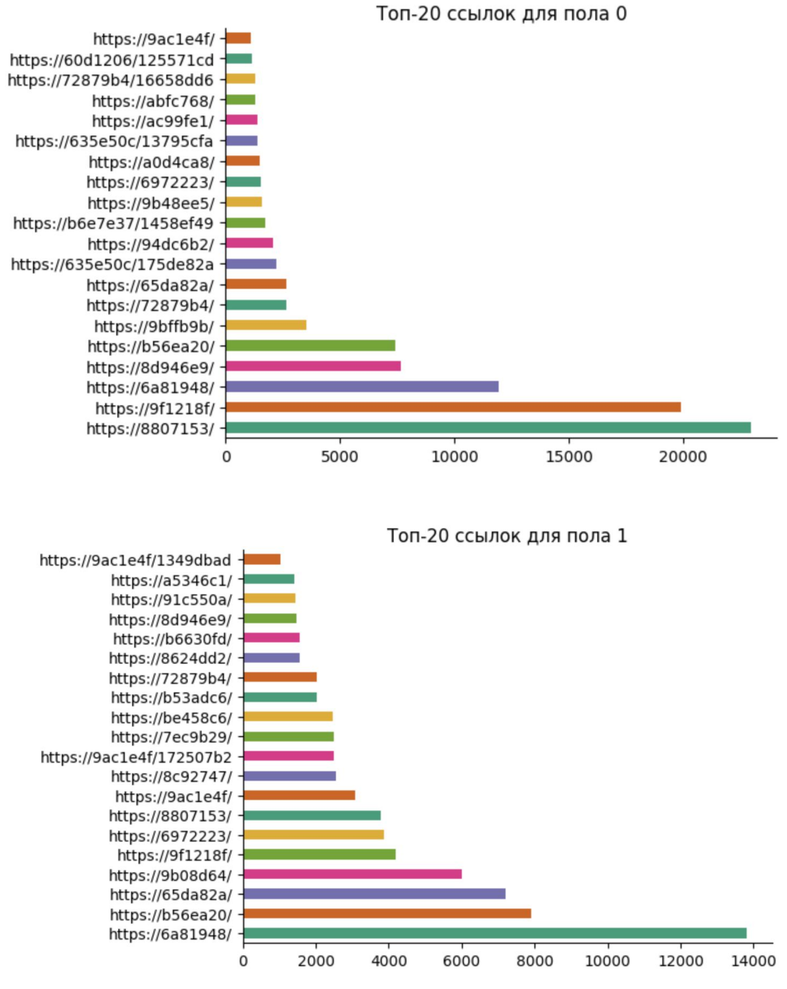

Рис 7. Распределение ссылок для таргета 1 и 0

Результаты обучения моделей

1. **Обученные модели:**
- Логистическая регрессия: базовый результат.
- Градиентный бустинг: точность 84%.
2. **Лучшие гиперпараметры:**
- Глубина деревьев: 15.
- Количество деревьев: 100.

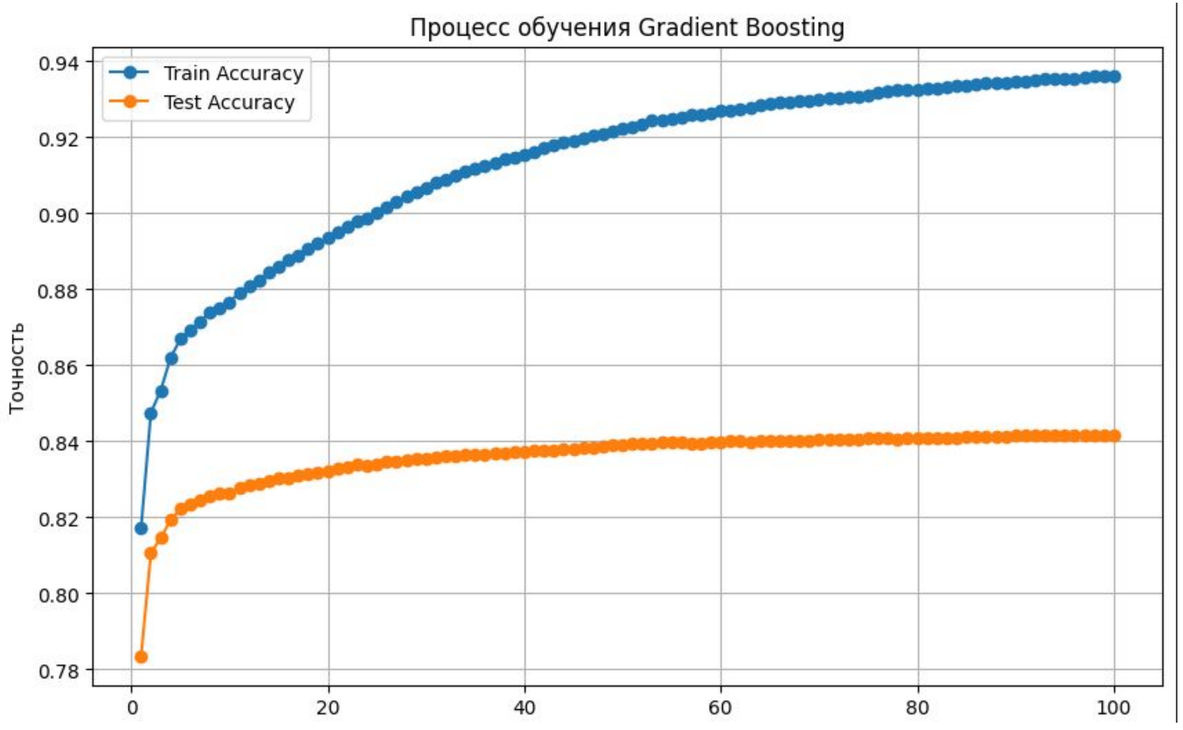

Рис 8. Обучение модели с заданными параметрами

Планы по обучению моделей

В качестве одного из возможных решений предложено обучить модель с архитектурой TabNet, что позволит проверить наличие/отсутствие преимуществ от использования нейросетевых моделей при обработке табличных данных.

В настоящее время:

- изучена документация;
- подготовлены данные для передачи в модель;
- написан код для обучения и тестирования модели.

Table 2: Performance for Forest Cover Type dataset.

Model	*Test accuracy (%)*
XGBoost	89.34
LightGBM	89.28
CatBoost	85.14
AutoML Tables	94.95
TabNet	**96.99**

Результаты авторов оригинальной статьи впечатляют . Интересно проверить

В связи с длительностью процесса обучения результаты в данной части будут представлены к следующему дедлайну.

Планы

1. **Улучшение моделей:**
   - Тонкая настройка гиперпараметров.
   - Тестирование дополнительных моделей (например, XGBoost, TabNet).

2. **Работа с признаками:**
   - Создание новых признаков (взаимодействие geo и referer).

3. **Повышение интерпретируемости:**
   - Использование SHAP для анализа влияния признаков.

4. **Визуализация прогресса:**
   - Добавление графиков с улучшением метрик.