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

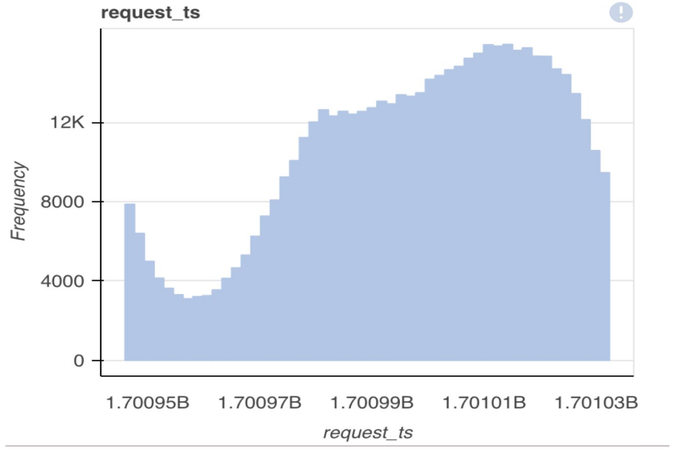


Рис 1. Промежуток с максимальным количеством запросов

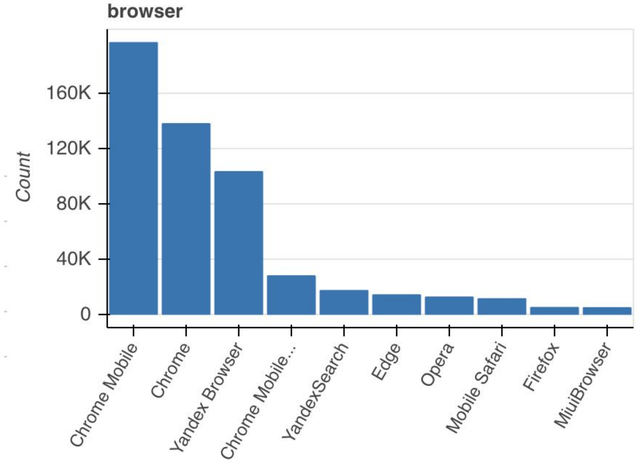


Рис 3. Использование пользователями браузеров

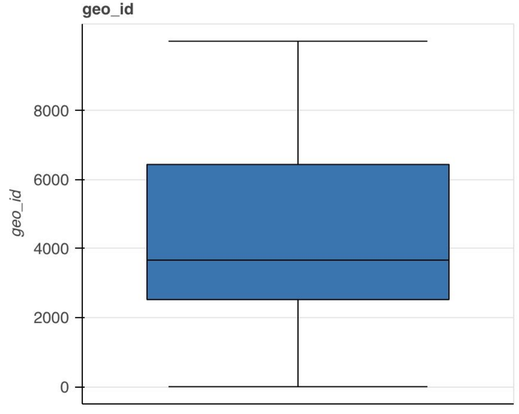


Рис 2. Медиальное значение geo_id

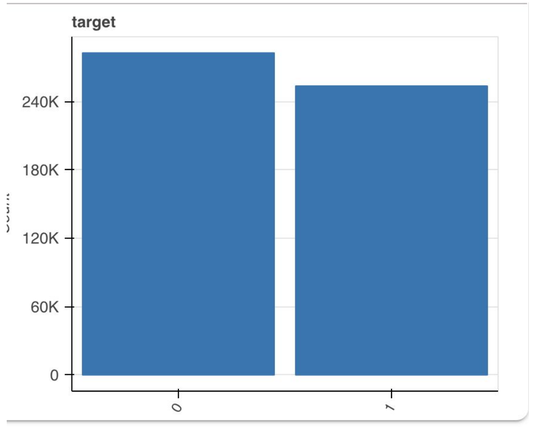


Рис 4. Распределение таргета 1 и 0

Предобработка данных

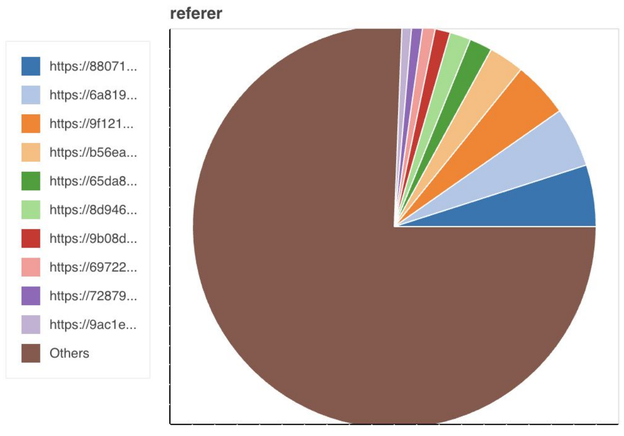


Рис 5. Разнообразие ссылок в данных

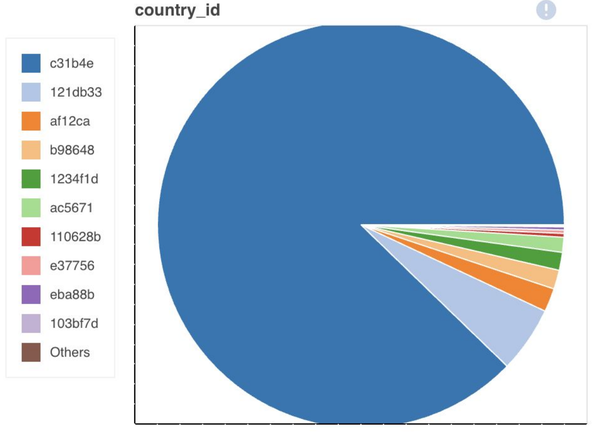


Рис 6. Распределение стран пользователей

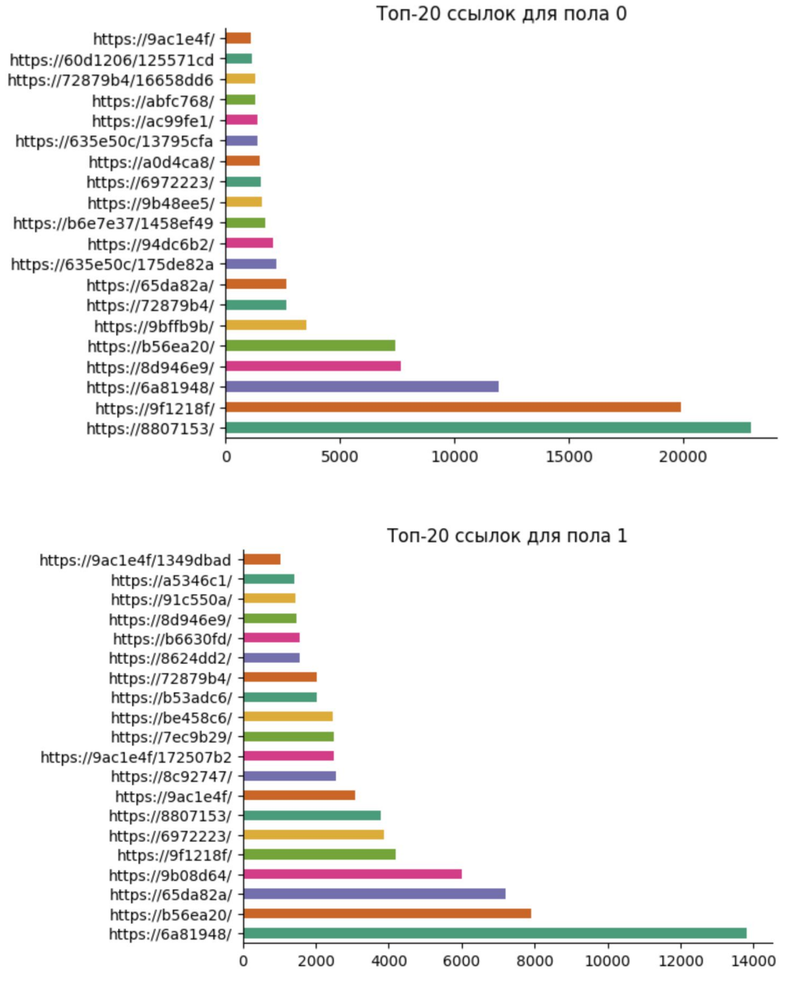


Рис 7. Распределение ссылок для таргета 1 и 0

Результаты обучения моделей: градиентный бустинг

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

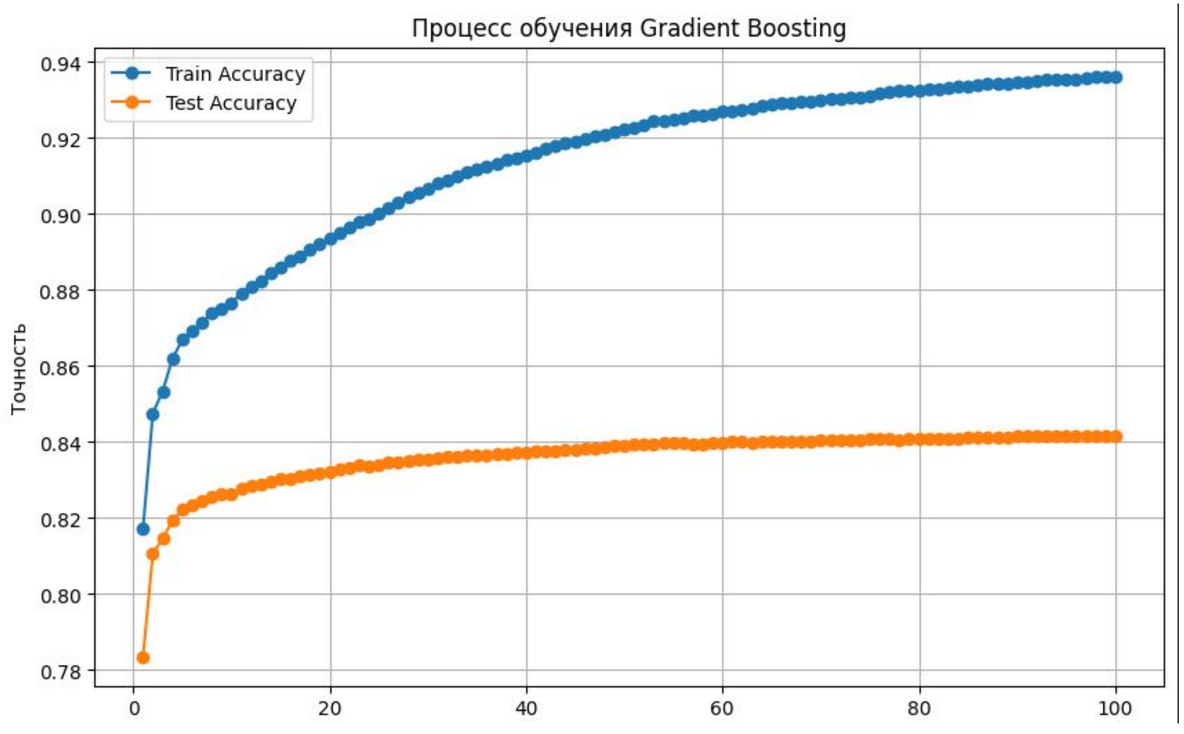
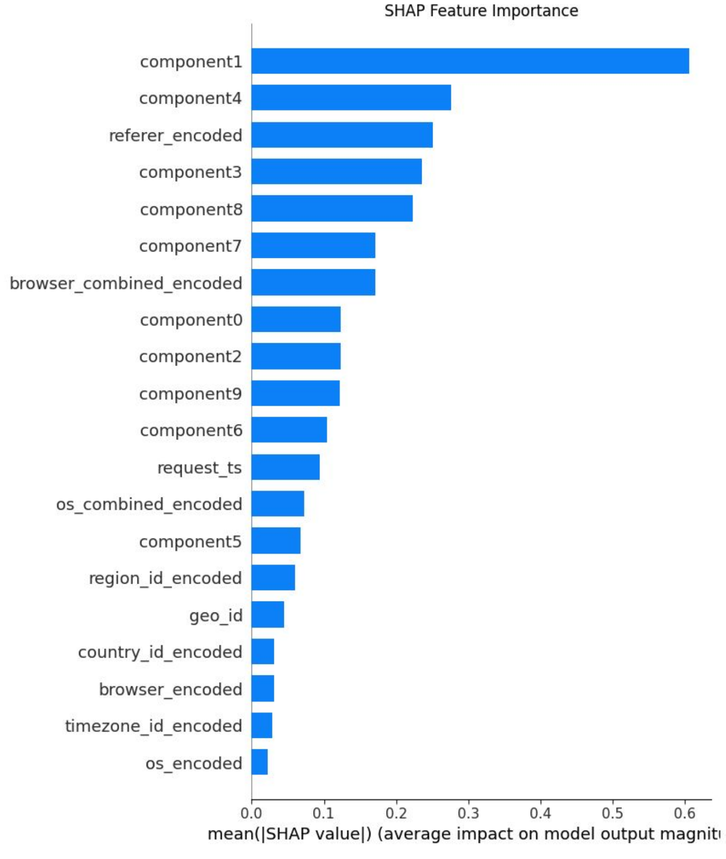
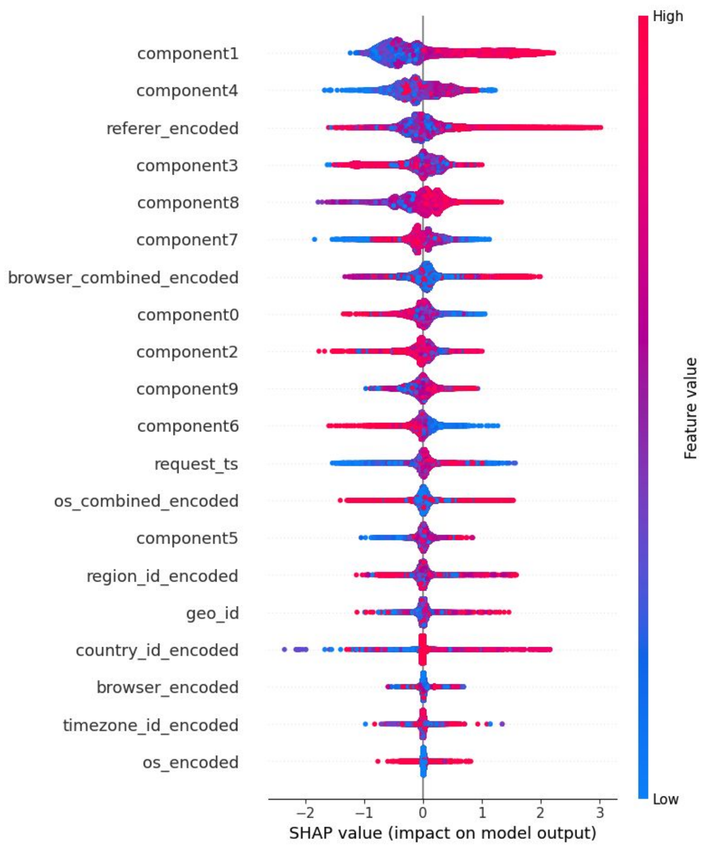


Рис 8. Обучение модели с заданными параметрами

Результаты обучения моделей: градиентный бустинг

После обучения также построили графики важности признаков





Результаты обучения моделей: TabNet

Классификатор на основе нейросетевой модели TabNet показал на тестовой выборке 81% точности

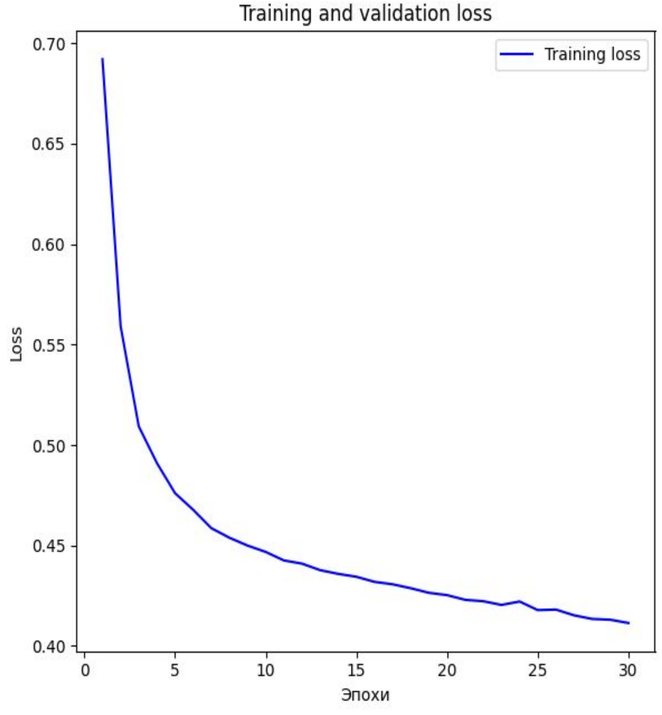
Оптимальные параметры по итогам сравнения результатов:

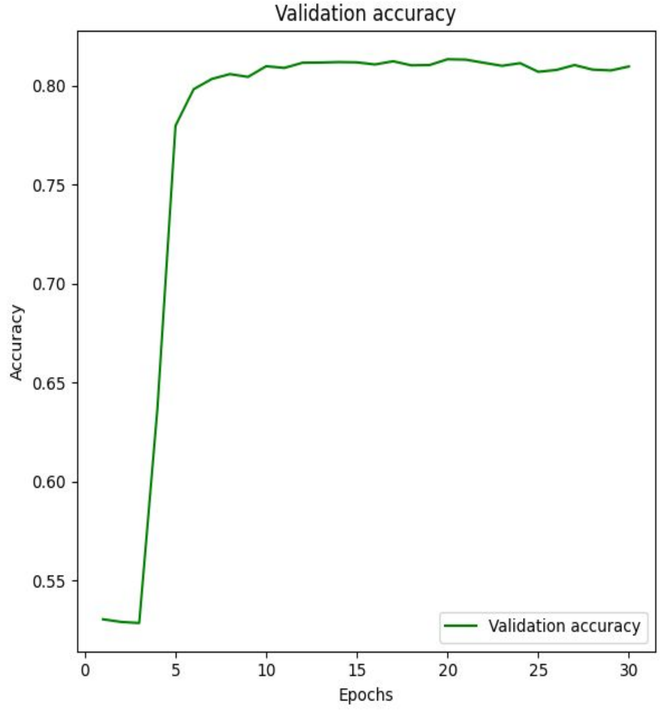
- batch_size=512
- остальные параметры - by default

Наилучший результат на 29-ой эпохе

Встренная в модель важность признаков показала, что наибольший вклад в предсказание вносят:

- domain
- path
- browser
- browser_version
- os_version





Итог

Мы провели полный цикл работы:

1. Анализ данных и их предобработка.
2. Обучение и настройка моделей.
3. Анализ и интерпретация результатов.

Лучшие результаты:

- Градиентный бустинг — точность 84%.
- TabNet — точность 81%, но высокая интерпретируемость.

Планы:

- Улучшение гиперпараметров.
- Тестирование дополнительных моделей.
- Повышение интерпретируемости с помощью SHAP.