FLOCKTORY

Создание модели, предсказывающей пол пользователя (м/ж) на основе данных о его активности

КОМАНДА “SYNC”

Лосева Надежда (исследование, подготовка данных, логистическая регрессия)
Кабанов Глеб (парсинг данных, градиентный бустинг, автоматизация)
Чернова Татьяна (обработка данных)
Мельчекова Александра (тестирование, отладка)
Арсеньев Александр (обработка данных)

ЦЕЛИ И ЗАДАЧИ ПРОЕКТА

- ПРОБЛЕМА
  - Плохая заполненность признака «пол» (пол точно известен только у 10% пользователей)
- ЗАДАЧА
  - Создать модель машинного обучения, предсказывающую пол пользователя (м/ж) на основе данных о его активности
- ЦЕЛИ
  - Улучшение качества выдачи «подарков» на витрине
  - Определение ЦА продуктов с ограничением по полу
  - Встраивание полученных данных в пайплайн для других моделей

ПРЕДОСТАВЛЕННЫЕ ДАННЫЕ

- Данные в json:
  - Train: 127755 users
  - Val: 27447 users
  - Test: 18784 users

- Распределение target:
  - female 50%
  - male 50%

ДАННЫЕ

- Глубокая иерархия features:
  - **orders (заказы)**
    - **site-id:**
      - orders:
        - created-at
        - items
          - id | count | general-category-path | brand-id
  - **visits (посещения)**
    - **site-id**
    - **first-seen**
    - **last-seen**
    - **visits**
      - visited-at | session-duration | pages-count
  - **site-meta (сайт с регистрацией пользователя)**
    - **site-id** | recency | frequency | monetary
  - **exchange-sessions (показанные сайты)**
    - landed-at | sites | accepted-site-id | accepted-at
  - **last-visits-in-categories**
    - category | last-visit-at

```
Доля пропусков в train['features']:

orders                          0.095855
site-meta                       0.000016
visits                          0.013111
last-visits-in-categories       0.012164
exchange-sessions               0.413949
```

ПРОВЕРЯЕМЫЕ МОДЕЛИ

- Градиентный бустинг (catboost) – первая идея
- Логистическая регрессия – попытка сделать проще
- Метод опорных векторов
- Случайный лес
- Нейронная сеть

ГРАДИЕНТНЫЙ БУСТИНГ

- Градиентный бустинг (catboost):
  - visits [‘site-id’, ‘visits’] 'Accuracy’: 0.64
  - orders [‘site-id’, ‘items.id’, ‘items.brand-id’] 'Accuracy': 0.73

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Features	Accuracy
site-meta ['site-id']	0.80
exchange-sessions ['accepted-site-id']	0.66
exchange-sessions ['clicks']	0.67
exchange-sessions ['clicks'] | ['accepted-site-id']	0.65
last-visits-in-categories ['category']	0.61
orders ['items.id']	0.77
orders ['items.brand-id']	0.63

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ – ФИНАЛЬНАЯ ВЕРСИЯ

- Key features:
  - orders [‘items.id’]
  - site-meta ['site-id’]
- Метод расчета входных данных:
  - Парсинг key features - RegEx
  - Расчет sex_score для key features (доля женщин среди выбравших)
  - Построение модели на sex_scores
- Достоинства:
  - Очень быстрый расчет
  - Высокая точность предсказаний

```
Accuracy: 0.89
Confusion Matrix:
[[10973  1857]
 [  970 11751]]
Classification Report:
              precision    recall  f1-score   support

      female       0.92      0.86      0.89     12830
        male       0.86      0.92      0.89     12721

    accuracy                           0.89     25551
   macro avg       0.89      0.89      0.89     25551
weighted avg       0.89      0.89      0.89     25551
```

НЕЙРОННАЯ СЕТЬ

- Протестировали простые варианты полносвязных сетей с дропаутом
- Есть потенциал для роста метрики

СПАСИБО ЗА ВНИМАНИЕ!

- https://github.com/loseff-n/flocktory