Анализ сайта «СберАвтоподписка»

анализ и предсказание целевых действий пользователей

Николай Борзяк,
kolkingen@gmail.com

Цели и задачи

Разведочный анализ данных

Ознакомление с данными

Оценка полноты и чистоты

Базовая чистка дубликатов, пропусков

Оценка распределений и отношений

Создать и обучить модель для

Предсказания **целевых действий**

«Оставить заявку» или «Заказать звонок»

Целевая метрика **ROC-AUC > 0.65**

Упаковать модель в сервис

Принимающий на вход признаки

visit_*, utm_*, device_*, geo_*

И возвращающий **0 или 1**, где

1 - пользователь совершил целевое действие

Реализация проекта

На входе файлы с сессиями и событиями на сайте

ga_sessions.csv - сессии:

- 1.8 млн объектов
- 18 признаков
- 4 колонки дают численные признаки
- 3 переменные бесполезны
- 11 категориальных признаков

ga_hits.csv - события:

- 15.7 млн объектов
- 11 признаков
- **event_action** -> целевая переменная
- 2.7% сессий с целевыми действиями

Когда целевых действий больше?

- Днём
- В начале недели
- При повторных посещениях
- Не из социальных сетей
- С органического трафика
- С компьютера
- Из Москвы и области

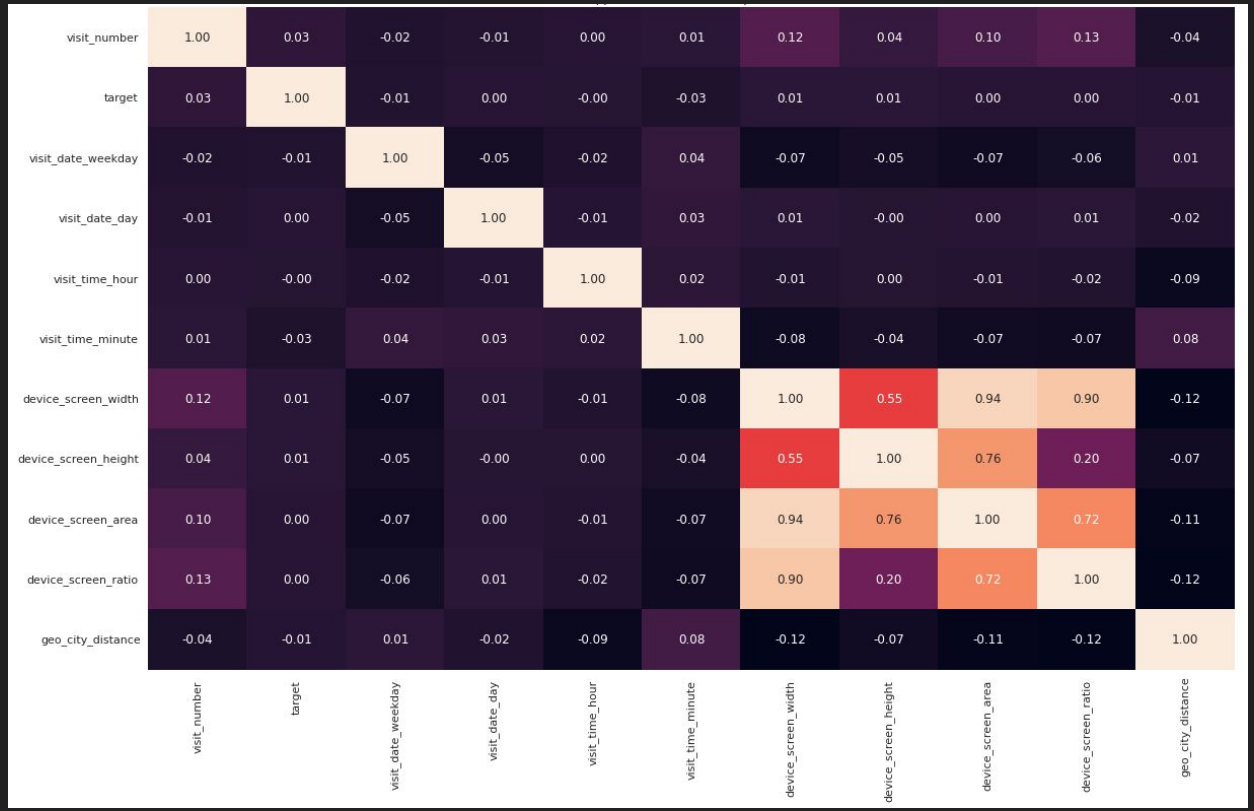
Целевые действия по посещениям:

4+	4.4%
3	3.8%
2	3.2%
1	2.4%

Дополнительные признаки

Из даты и времени	день недели, час
Органический трафик	
Трафик из соцсетей	
Из размера экрана	ширина, площадь экрана
Для городов	московская область
Расстояние до Москвы	

Корреляция численных признаков



Подготовка данных

- Создание признаков
- Численные преобразования
- Категориальные преобразования
- Выбор признаков

```
('indexer', FunctionTransformer(set_index)),
('imputer', FunctionTransformer(fill_missings)),
('engineer', FunctionTransformer(create_features)),
('dropper', DropFeatures([...])),

('normalization', YeoJohnsonTransformer()),
('outlier_remover', Winsorizer()),
('scaler', StandardScaler()),

('rare_encoder', RareLabelEncoder(tol=0.05, replace_with='rare')),
('onehot_encoder', OneHotEncoder(drop_last_binary=True)),
('bool_converter', FunctionTransformer(converse_types)),

('constant_dropper', DropConstantFeatures(tol=0.99)),
('duplicated_dropper', DropDuplicateFeatures()),
('correlated_dropper', DropCorrelatedFeatures(threshold=0.8)),
```

LightGBM - лучше

- Высокий ROC-AUC
- Быстрое обучение
- Интерпретируемая
- Предсказывает вероятность

Модель	ROC-AUC
Гистограммный бустинг	0.7070
LightGBM	**0.7066**
CatBoost	0.7062
Нейронная сеть	0.70
XGBoost	0.69
Логистическая регрессия	0.67
Классификатор Байеса	0.65
Случайный лес	0.63
Метод опорных векторов	0.62
Дерево решений	0.52

Оптимизация модели

Производительность:

800 деревьев

0.07 скорость обучения

‘GOSS’ тип бустинга

0.95 порог для коррелируемых признаков

Регуляризация:

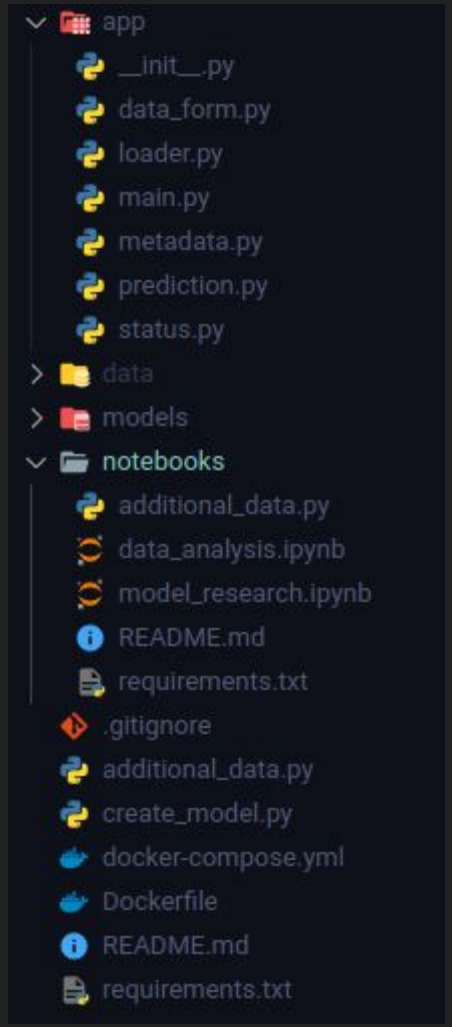
L1 регуляризация - 10

L2 регуляризация - 10

26 листьев в дереве

Создание сервиса

- Сервис в отдельном модуле
- Создание модели - отдельно
- FastAPI + Uvicorn + Pydantic
- Упакован в Docker



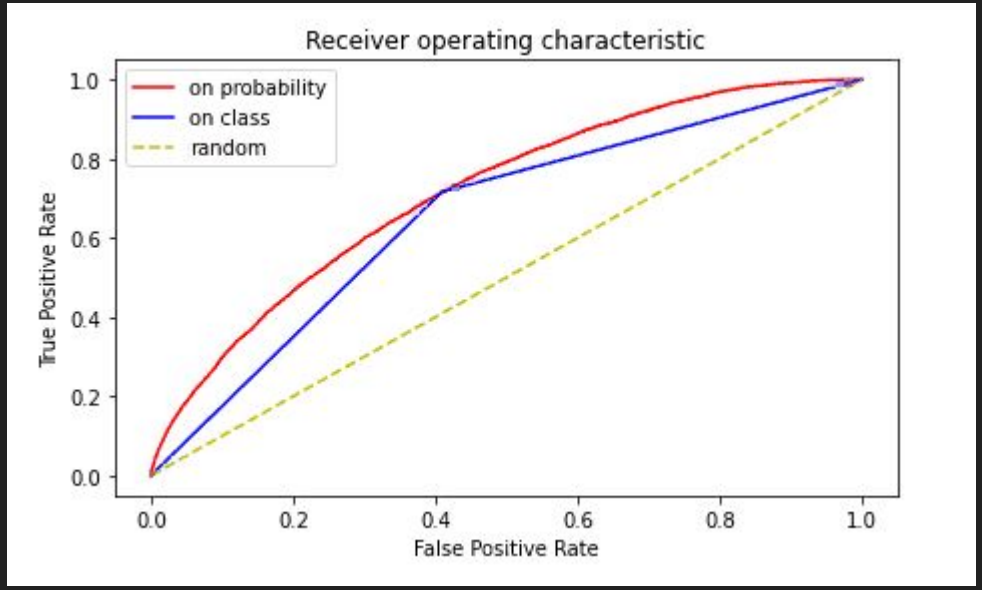
Результаты проекта

Метрики модели

ROC-AUC	**0.715**
ROC-AUC (classes)	0.654
ACCURACY	0.594
PRECISION	0.046
RECALL	0.717
F1	0.087
Порог вероятностей	0.0257
Переобучение	Нет

Матрица ошибок:

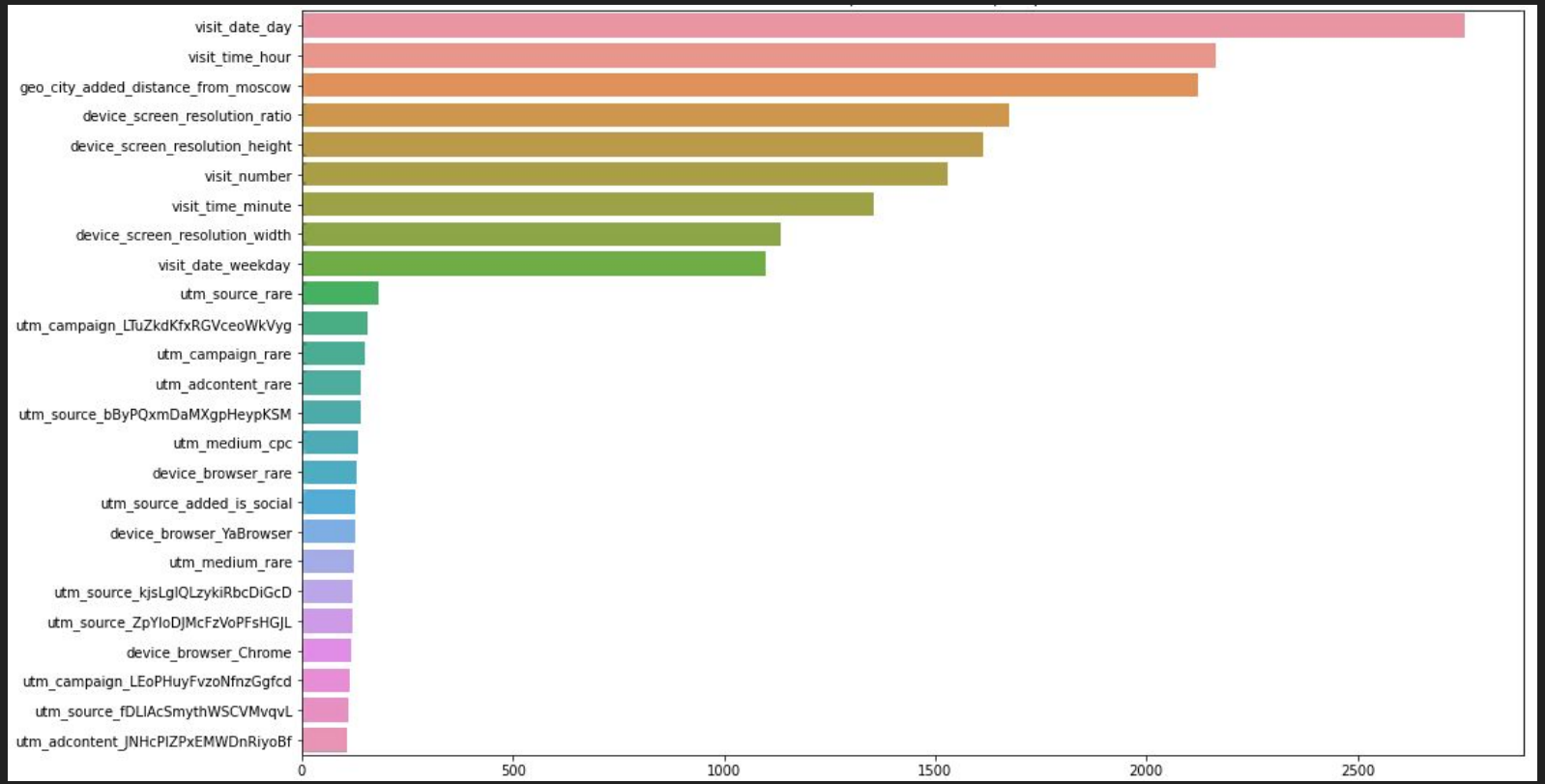
prediction|	0.0	1.0
true label|		
0.0	114853	79737
1.0	1532	3878



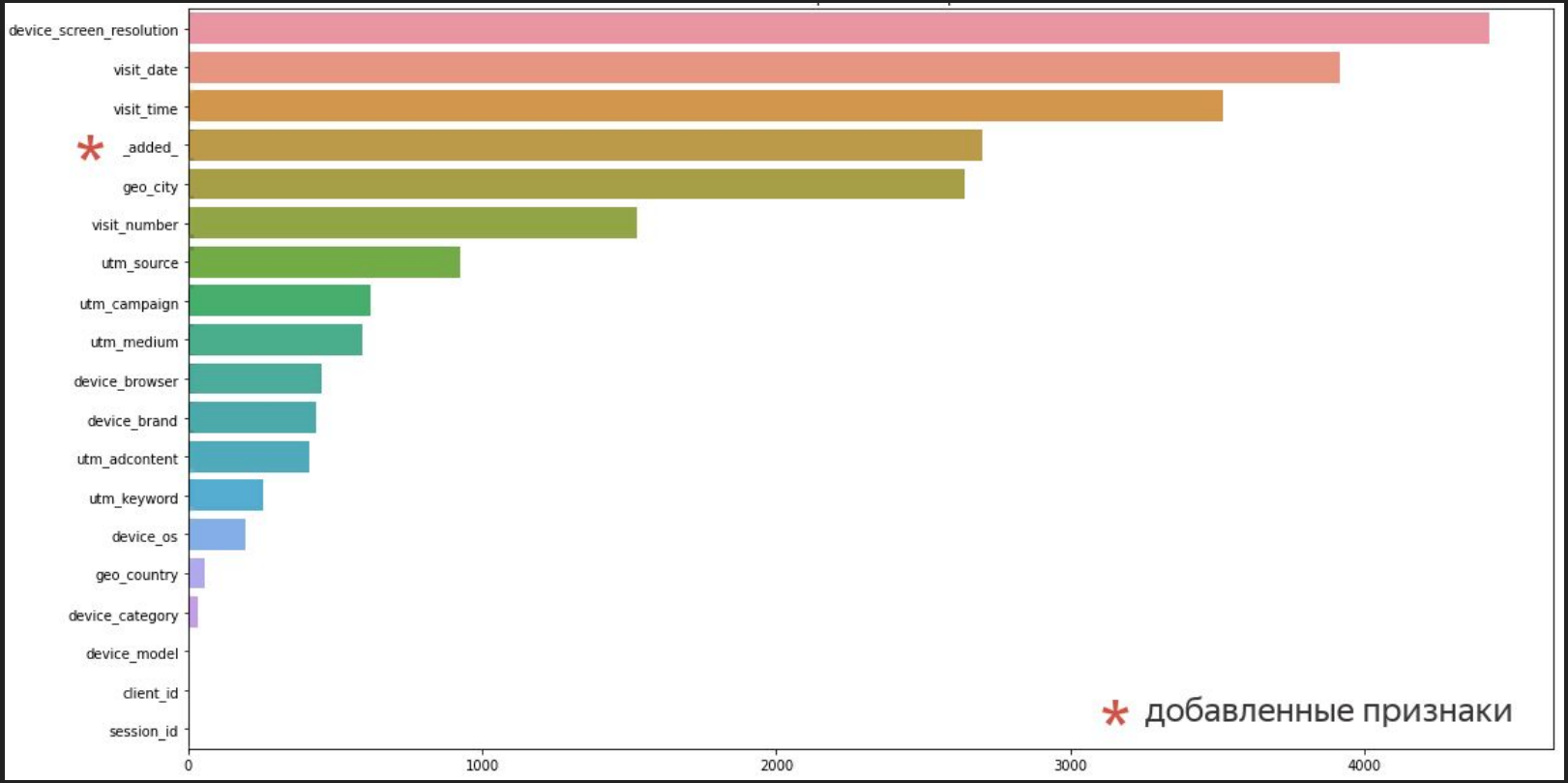
Результаты обработки данных

- 67 признаков (ещё 8 удалены)
- 9 из них - численные
- 180 тыс. дубликатов
- Нет корреляций с целевой переменной

Самые важные созданные признаки



Важность оригинальных признаков



Работа сервиса

- Работает через **uvicorn** или **docker**
- Предсказание за 2 секунды
- Возвращает статус и метаданные
- Предсказание класса или вероятности
- Для одного объекта и для множества

http://127.0.0.1:8000/predict_many


```
[
  {
    "session_id": "9055434745589932991.1637753792.1637753792",
    "prediction": 0
  },
  {
    "session_id": "905544597018549464.1636867290.1636867290",
    "prediction": 0
  }
]
```


http://127.0.0.1:8000/status


```
"Сервис работает."
```


http://127.0.0.1:8000/version


```
{
  "name": "SberAutopodpiska: target event prediction",
  "description": "Модель по предсказанию совершения пользователем одного из целевых действий \"Заказать звонок\" или \"Оставить заявку\" на сайте сервиса СберАвтоподписка.",
  "version": 1,
  "author": "Nikolai Borziak",
  "model_type": "LGBMClassifier",
  "training_datetime": "2022-11-25 07:54:19.080034",
  "threshold": 0.026015066085760746,
  "metrics": {
    "roc_auc": 0.7160244331913947,
    "roc_auc_by_class": 0.6541750943426337,
    "accuracy": 0.59894,
    "precision": 0.04672047702152414,
    "recall": 0.7125693160813309,
    "f1": 0.08769136279884442
  }
}
```

Как улучшить модель?

- Больше данных
- Ребалансировка классов
- Поиск лучших гиперпараметров
- Больше новых признаков
- Более сложные модели

Выводы

- ROC-AUC = 0.715
- Время предсказания < 3 секунд
- Анализ и чистка данных проведены
- Проведена генерация новых признаков
- Важность признаков проанализирована