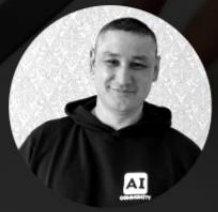
Анализ сайта «СберАвтоподписка»

Итоговая работа по курсу Data Science



Автор работы:
Фатихов Салих

Содержание

Фатихов Салих Загирович

г. Уфа

Цель обучения:
Расширение области знаний для профессионального и карьерного роста
Интерес к области искуственного интеллекта

Деятельность:
Главный специалист в сфере инжиниринга нефтегазовых месторождений

Образование:
Кандидат физико-математических наук

Научные интересы:
Развитие методов исследований, автоматизации и обработки различных данных применительно к нефтегазовым месторождениям

Решаемая задача

«СберАвтоподписка» — это сервис долгосрочной аренды автомобилей для физлиц.

Задачей для финальной работы ML инженера поставлена разработка сервиса, подсказывающего, совершит пользователь сайта целевое действие или нет. Для этого необходимо:

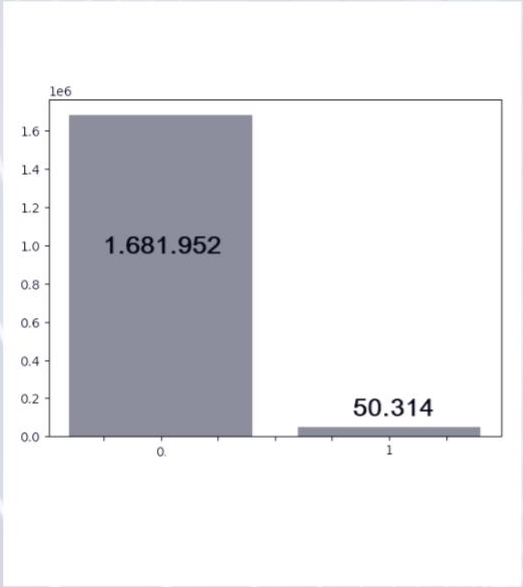
- Подготовить датасет
- Обучить модель
- Создать localhost web app

Требование к точности модели: не менее 0,65 по метрике ROC-AUC.

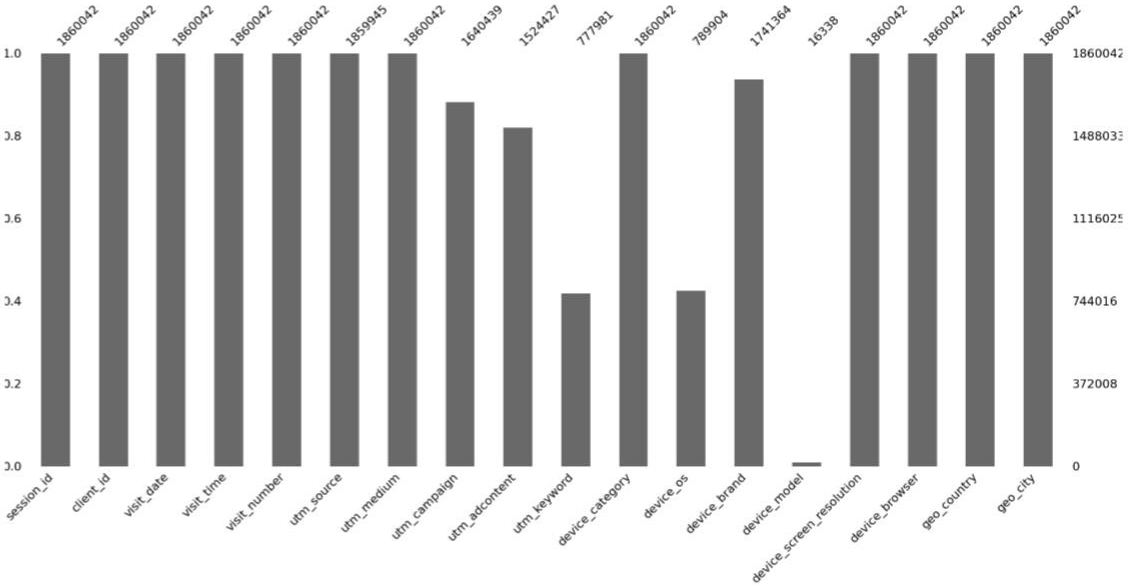
Содержание — модель, берущая на вход строку с данными по визиту (согласно схеме данных) и отдающая на выход результат предсказания по отдельному событию в числовом формате 0|1.

Этапы решения задачи

Знакомство с датасетом



- Целевая переменная сильно несбалансирована



- Фичи для модели категориальные
- Есть пустые поля

Этапы решения задачи

Стратегия работы с фичами

- Пустые поля заполнял значением "other"
- По полю "канал привлечения" использовал имеющуюся информацию и сделал бинарным как реклама в соцсетях или нет
- Аналогично по полю тип привлечения - органический трафик или нет
- По полю браузер объединил категории содержащие instagram
- Во всех полях использовал 2 стратегии сокращения количества категорий: объединял категории по которым все целевые значения равны нулю; объединял поля, количество строк по которым меньше граничного значения (30)
- Все категориальные переменные преобразовывал с помощью One Hot Encoder

Модели машинного обучения

Random Forest

ROC-AUC: 0.579

Neural Network

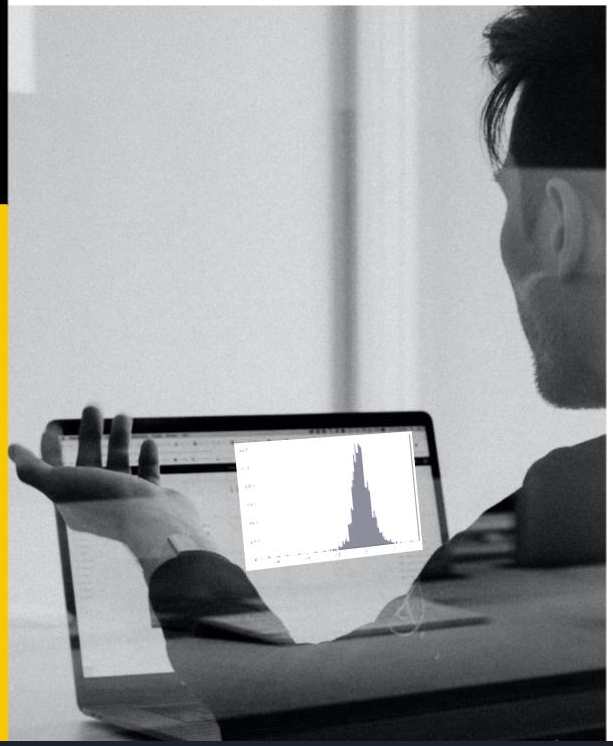
ROC-AUC: 0.662

Linear Regresion
ROC-AUC: 0.705

C-Support Vector Classification
ROC-AUC: 0.71

jupyter

Анализ влияния крайних фич в логистической регрессии



- Выделены фичи сильно влияющие на предсказание 0 или 1
- Среди предсказывающих 1 выделены id ключевых слов и рекламных компаний
- Как и ожидалось 0 предсказывают фичи выделенные в категорию как не содержащую положительные таргеты
- Дополнительно выделены предсказывающие ноль категории среди полей utm_campaign, utm_adcontent, utm_keyword

Web App Сервис

Разработан пайплайн обучения модели машинного обучения, модель сохранен с помощью пакета dill

Модель обернута в web сервис с помощью библиотеки fast api которая передает результат предсказания по post запросу

Демонстрация:

https://disk.yandex.ru/i/ZHExhP9rHPXjBQ

Использовано при работе

- **ПО:** Jupyter Notebook, PyCharm
- **Библиотеки:** pandas, numpy, scipy, scikt-learn, fastapi, matplotlib
- **Репозиторий в:** GitHub
  https://github.com/fsalih/DS-intro-final

Благодарю за внимание!

Machine Learning