Этапы проекта по анализу данных

Элен Теванян

16 мая 2023

План на курс

- Разбираем 1 сквозную задачу на проектах
- Предлагаем 2-3 задачи от нас, либо инициативная задача
- Каждую неделю делаем домашку: повторяем на данных то, что изучили
- Хорошая новость – делаем домашку в командах
- Дедлайн – следующее занятие
- Последнее занятие – презентация результатов

Задание в конце пары

1. Попасть в чат
   https://t.me/+IkY1XnU6PQRkNTQy

Домашка после сегодня:

1. Найти себе команду – или работать одному
2. Выбрать датасет
3. (опц.) Уметь открывать его в коллабе

Сквозная задача на курс

Сквозная задача на курс

- Отток пользователей из Телеком-компании

Сквозная задача на курс

- Отток пользователей из Телеком-компании
- Датасет публичный, не является NDA
- Детали датасета обсудим сегодня в процессе

Сквозная задача на курс

- Отток пользователей из Телеком-компании
- Датасет публичный, не является NDA
- Детали датасета обсудим сегодня в процессе

Этапы проекта по анализу данных

1. **Постановка задачи**

Отток пользователей

Отток пользователей

- ML – это только инструмент

Отток пользователей

- ML – это только инструмент

- Зачем мы делаем эту задачу?

Отток пользователей

- ML – это только инструмент

- Зачем мы делаем эту задачу?
- Какие есть способы её решить?

Отток пользователей

- ML – это только инструмент

- Зачем мы делаем эту задачу?
- Какие есть способы её решить?
- Сколько нам будет стоить каждое из решений?

Отток пользователей

- ML – это только инструмент

- Зачем мы делаем эту задачу?
- Какие есть способы её решить?
- Сколько нам будет стоить каждое из решений?
- Есть ли уже рабочие, почти бесплатные решения?

Отток пользователей

- ML – это только инструмент

- Зачем мы делаем эту задачу?
- Какие есть способы её решить?
- Сколько нам будет стоить каждое из решений?
- Есть ли уже рабочие, почти бесплатные решения?
- Какие данные помогут решить задачу?

Этапы проекта по анализу данных

1. Постановка задачи
2. **Сбор данных**

Сбор данных: какие есть варианты

Сбор данных: какие есть варианты

- Что-то есть собранное:

Сбор данных: какие есть варианты

- Что-то есть собранное:
  - В виде отдельных файлов (.csv, xlsx, …)
  - В облаке
  - В базе данных

Сбор данных: какие есть варианты

- Что-то есть собранное:
  - В виде отдельных файлов (.csv, xlsx, …)
  - В облаке
  - В базе данных
- Есть структурированные источники, из которых можно собрать датасет:
  - Витрины в БД

Сбор данных: какие есть варианты

- Что-то есть собранное:
  - В виде отдельных файлов (.csv, xlsx, …)
  - В облаке
  - В базе данных
- Есть структурированные источники, из которых можно собрать датасет:
  - Витрины в БД
- Спарсить данные

Сбор данных: какие есть варианты

- Что-то есть собранное:
  - В виде отдельных файлов (.csv, xlsx, …)
  - В облаке
  - В базе данных
- Есть структурированные источники, из которых можно собрать датасет:
  - Витрины в БД
- Спарсить данные
- Купить данные

Этапы проекта по анализу данных

1. Постановка задачи
2. Сбор данных
3. **Очистка и обработка данных**

Мусор на входе – мусор на выходе

Минимальные проверки

Минимальные проверки

- Дубликаты

Минимальные проверки

- Дубликаты
- Пропуски

Минимальные проверки

- Дубликаты
- Пропуски
- Неадекватные значения

Минимальные проверки

- Дубликаты
- Пропуски
- Неадекватные значения
- Выбросы

Минимальные проверки

- Дубликаты
- Пропуски
- Неадекватные значения
- Выбросы
- ….

Этапы проекта по анализу данных

1. Постановка задачи
2. Сбор данных
3. Очистка и обработка данных
4. **Разведочный анализ данных**

Что скрыто в данных?

Что скрыто в данных?

- Считаем описательные статистики:

Что скрыто в данных?

- Считаем описательные статистики:
  - Количество данных, мин, макс
  - Среднее, дисперсию
  - Процентили
  - …..

  Кандидат в выбросы = любое значение, которые вне диапазона (LQ – 1.5IQR; UQ + 1.5IQR)
- IQR = UQ – LQ

Что скрыто в данных?

- Считаем описательные статистики:
  - Количество данных, мин, макс
  - Среднее, дисперсию
  - Процентили
  - …..
- Рисуем графики:
  - Гистограммы
  - Точечные диаграммы
  - Боксплоты
  - …

Этапы проекта по анализу данных

1. Постановка задачи
2. Сбор данных
3. Очистка и обработка данных
4. Разведочный анализ данных
5. **Моделирование**

Моделируем

- Определяем класс задачи:
  - Регрессия/классификация/что-то еще

Моделируем

- Определяем класс задачи:
  - Регрессия/классификация/что-то еще
- Выбираем пул моделей для обучения

Моделируем

- Определяем класс задачи:
  - Регрессия/классификация/что-то еще
- Выбираем пул моделей для обучения
- Обучаем и настраиваем их

Моделируем

- Определяем класс задачи:
  - Регрессия/классификация/что-то еще
- Выбираем пул моделей для обучения
- Обучаем и настраиваем их
- Валидируем и выбираем лучшую

Этапы проекта по анализу данных

1. Постановка задачи
2. Сбор данных
3. Очистка и обработка данных
4. Разведочный анализ данных
5. Моделирование
6. **AB-тесты**

Зачем этот этап?

Зачем этот этап?

- Есть экономический смысл продуктивизировать разработку

Зачем этот этап?

- Есть экономический смысл продуктивизировать разработку
- Работа над проектом не заканчивается на оттюнингованной модели

Зачем этот этап?

- Есть экономический смысл продуктивизировать разработку
- Работа над проектом не заканчивается на оттюнингованной модели

Что делаем

- Глобально отвечаем на вопрос, стало ли от нашей разработки лучше

Что делаем

- Допустим, мерим успешность всех активностей по оттоку через ретеншн.

Что делаем

- Допустим, мерим успешность всех активностей по оттоку через ретеншн.

- Без алгоритмов естественным образом возвращались 2% из оттока

Что делаем

- Допустим, мерим успешность всех активностей по оттоку через ретеншн.

- Без алгоритмов естественным образом возвращались 2% из оттока
- Был один алгоритм определения оттока, помогал вернуть 30%

Что делаем

- Допустим, мерим успешность всех активностей по оттоку через ретеншн.

- Без алгоритмов естественным образом возвращались 2% из оттока
- Был один алгоритм определения оттока, помогал вернуть 30%
- Мы сделали другой, с ML – какой прирост даст она?

Что делаем

- Разбиваем на 2 группы
- Для одной группы работает старый вариант
- Для второй группы работает новый вариант

- Через некоторое время оцениваем результаты

Подводные камни

- Уйма
- Плохо задизайненный тест и мало доверия к разработке – можно прощаться с идеей
- Плохой тест может не показать хорошие результаты, которые могли бы быть
- Хороший тест может не показать хорошие результаты – прощаемся с идеей

Этапы проекта по анализу данных

1. Постановка задачи
2. Сбор данных
3. Очистка и обработка данных
4. Разведочный анализ данных
5. Моделирование
6. AB-тесты
7. **Продакшн**

Продакшн?

- Упаковать в докер
- Обернуть в микросервис
- Привлечь фронт/бэк, чтобы передавать в приложения
- …