Машинное обучение в бизнесе

Постановка задачи. Подходы к решению.

Лекция 9, часть 1

Дмитрий Колодезев
НГУ 2019 весна

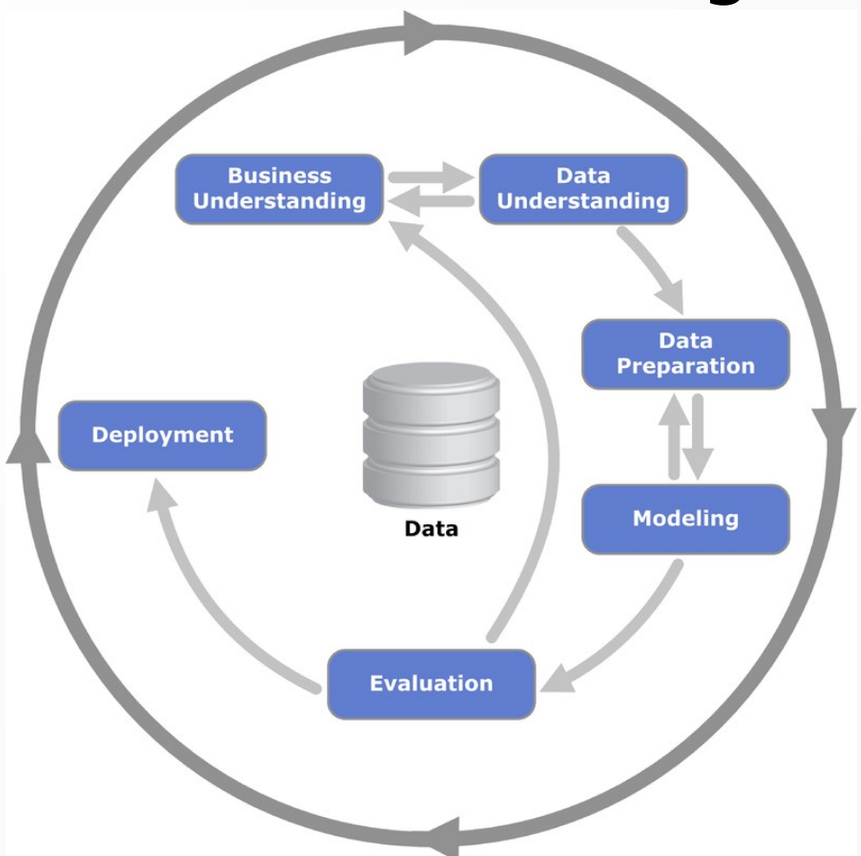
План лекции

- CRISP-DM
- ДНК
- Правильно поставленные задачи
- Оценка ROI на пальцах
- Проблемы
- Решения
- Kaggle & Co

CRISP-DM (1999)

CRoss Industry Standard Process for Data Mining

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

Дано-Найти-Критерий

- Дано
  - Что есть? Что может быть? Что будет?
- Найти
  - Что, собственно, ищем?
- Критерий
  - Метрика и способ оценивания модели
- Смысл системы - вне ее
  - Ту ли задачу решаем?

"data scientists ... incorporate data science into the operation of a product or service, using data in smart ways to provide value."

Edo Wilder-James https://www.svds.com/how-do-you-build-a-data-product/

ДНК - погода

Как продажи товара зависят от погоды?

- Дано:
  - Данные о продажах
- Найти
  - Как зависит спрос на товар от погоды.
- Модель и критерий
  - Линейная регрессия, p-value? R^2? RMSE на отложенной выборке?

https://www.shopolog.ru/metodichka/attracting-clients/dozhd-sneg-ili-solntse-kak-riteylery-ispolzuyut-pogodu-dlya-povysheniya-prodazh/

ДНК - погода

Зачем это нам? Чтобы больше продать

- Давайте изменим погоду, чтобы увеличить продажи!

- Исходная задача — корректировка рекламы
  - Больше/меньше отклик на рекламу? Его и измерять
  - Учесть другие факторы — сезон, дни недели
  - Есть ли у нас прогноз погоды нужной точности?
- Подход
  - Прерванные временные ряды, например https://en.wikipedia.org/wiki/Interrupted_time_series

ДНК - продажи

Приоритизация продаж

- Дано:
  - Данные о продажах
- Найти
  - Какие товары скорее всего согласится купить клиент?
- Критерий
  - Точность (DCG, NDCG) > X?

https://www.coursera.org/lecture/data-analysis-applications/mietriki-kachiestva-ranzhirovaniia-OkLNB
https://en.wikipedia.org/wiki/Learning_to_rank

http://www.action-mcfr.ru/blog/Kapaev/Pro-analitiku-v-novyh-prodazhah

ДНК - продажи

- Зачем это нам? Чтобы больше продать
- Задача:
  - Получить больше отдачи на один звонок продавца
- Подход
  - Uplift

https://en.wikipedia.org/wiki/Uplift_modelling
https://robo-hunter.com/news/model-uplift-kak-magazini-reshayt-davat-klientam-skidku-ili-net14668

ДНК — квалификация лида

Оценка привлекательности клиента

- Дано:
  - Данные о продажах
- Найти
  - Оценить вероятность продажи
- Критерий
  - BCE / Точность на отложенной выборке?

https://rdipietro.github.io/friendly-intro-to-cross-entropy-loss/

ДНК - квалификация лида

- Зачем это нам? Чтобы больше продать
- Задача:
  - Получить больше отдачи на один входящий лид
- Подходы
  - Оценка вероятности/срока первой покупки
  - Оценка суммы первой покупки, RMSE, MAE
  - LTV RMSE, MAE на отложенной выборке

Оценка ROI — правило 1%

- Сколько денег принесет улучшение метрики на 1%?

  Например, если мы будем предсказывать спрос на 1% точнее, сколько денег удастся заработать? За счет чего?

  - оптимизация складских запасов, меньше денег заморожено в товарных запасах
  - увеличение продаж, товар будет в наличии
- Главный вопрос к модели — И что?

~~Все плохо~~ Проблемы

- Регрессия к среднему — модель в жизни работает хуже, чем у аналитика
- Распределение признаков меняется со временем (версия iphone, год выпуска машины)
- Поведение потребителей меняется
- Собираемые данные меняются (формат и состав данных)
- Модель — скрытые знания (нет синергии)

~~Выкрутимся~~ Решения

- Мониторинг качества работы модели. Закладывать при внедрении
- Дообучение на новых данных
- Проверка распределений признаков
- Интерпретируемость модели (про это будет митап)

Kaggle как ML-витрина

- На Kaggle задачи не как в жизни.
  ДНК есть, проблем нет
- Вашу задачу решали на Kaggle
- Или решали похожую на нее
- Kaggle как чеклист — что из того, что там решают, можно применить в вашей отрасли?
- Kaggle как реестр идей для стартапов