Машинное обучение для оптимизации бизнеса на примере реальных кейсов

Константин Коточигов, CleverData, 2017

email: k.kotochigov@cleverdata.ru
skype: konstantin.kotochigov

Обо мне

- Константин Коточигов, CleverData
- Data Scientist
- В компании с 2015 года

О компании

- CleverData, с 2014 года входит в ГК Ланит
- Разработка собственных продуктов в области Digital Advertising
  - 1DMP (data management platform)
  - DMC (data management cloud)
- Консалтинг в области больших данных и машинного обучения

Наши заказчики

Банки	Ритейл	Транспорт
Финансовые организации	Телеком	Сельское хозяйство
Рекламные агенства	Страхование	Электронная коммерция

Основной запрос наших заказчиков

Нужна оптимизация бизнеса, а для этого максимально точный ответ на вопросы:

- Каков портрет моего клиента
- Что на самом деле ему нужно
- Как выстроить оптимальное взаимодействие с ним
- Как повысить его лояльность
- Как увеличить средний чек, LTV, ARPU итд

Как можно этого добиться

- Положиться на чужой опыт
  нанять эксперта

- Развивать свою экспертизу
  выдвинуть ряд гипотез и провести A/B (C/D…) тестирование

- Построить аналитику на исторических данных

Как можно этого добиться

- Положиться на чужой опыт
  нанять эксперта

- Развивать свою экспертизу
  выдвинуть ряд гипотез и провести A/B (C/D…) тестирование

- Построить аналитику на исторических данных
  Это то, что делаем мы

Виды аналитики по “глубине”

1. Дескриптивная
   Что есть в наших данных
   Более или менее все умеют - BI, reporting, visualization

2. Предиктивная
   Экстраполируем знания на новые кейсы

3. Прескриптивная

   Все решения принимает машина
   Огромный спрос на рынке

Виды аналитики по “глубине”

1. Дескриптивная
   Что есть в наших данных
   Более или менее все умеют - BI, reporting, visualization

2. Предиктивная
   Экстраполируем знания на новые кейсы

3. Прескриптивная
   Все решения принимает машина
   Огромный спрос на рынке

Мы делаем 2 и 3

Примеры кейсов

- Модель отклика
  Переход по ссылке, покупка предлагаемого товара

- Модель оттока
  Уход к конкуренту

- Модель удержания
  Отказ от ухода к конкуренту

- Рекомендательная система
  Предложение релевантного продукта

- HR-скоринг
  Прогнозирование показателей сотрудника

Примеры кейсов

- Модель Lookalike
  Найти людей, похожих на покупателей данного товара

- Модель антифрод
  Обнаружить мошенническую сделку/транзакцию

- Модель Collection-скоринг
  Кто из клиентов погасит задолженность

- Модель прогнозирования выручки
  Спрогнозировать финансовый показатель на следующем интервале

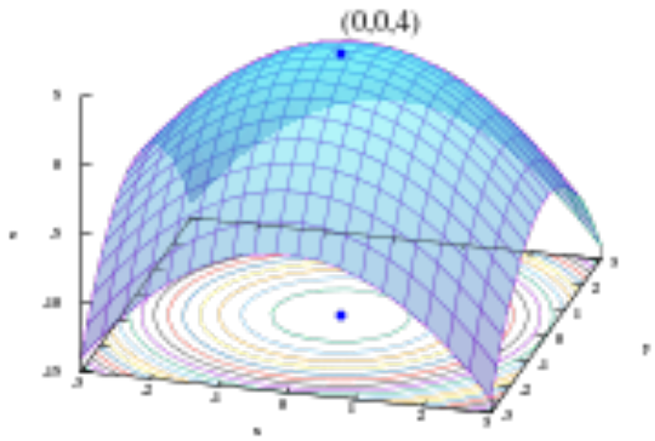
- Модель кластеризации
  Выделить однородные группы посетителей веб-сайта

Постановка задачи

- Есть Y
  - средний чек
  - отклик на рекламное предложение
  - дефолт по кредиту
- Есть X
  - транзакции клиента
  - анкетные данные клиента
  - посещенные страницы на сайте
- Задача: найти такую зависимость f, что Y=f(X)

Постановка задачи

- Модель - приближение реальности => задача аппроксимации f(x)
- Обучение модели = оптимизация критерия качества (количества ошибок)



Если все так просто,
зачем нужны дата саентисты?

1. Разнообразие подходов
2. Правильная постановка и интерпретация
3. Сложности при реализации

Разнообразие алгоритмов

- Регрессионные модели
- Деревья решений
- Метод опорных векторов
- k-ближайших соседей
- Нейронные сети
- RandomForest
- XGBoost
- Ансамбли классификаторов
- Наивные баесовские классификаторы
- и прочее…

Разнообразие типов данных

- Числовые
- Строковые
- Бинарные
- Порядковые
- Временные ряды
- Текст
- Геоданные
- Изображения
- Аудио/Видео
- И прочее…

Разнообразие методов оптимизации

- Gradient Descent
- Stochastic GD
- MiniBatch
- Evolutionary/Genetic
- Blockwise GD
- итд

Если все так просто,
зачем нужны дата саентисты?

- Разнообразие подходов
- Правильная постановка задачи и интерпретация результатов
- Сложности при реализации

Разнообразие критериев качества

- Accuracy
- Precision
- Recall
- ROC AUC
- PR AUC
- F-measure
- Lift
- RMSE
- R-squared
- LogLoss
- Cross entropy
- и много-много других

Разнообразие способов обучения

- Обучение по размеченным данным
  Классификация
  Регрессия

- Обучение без учителя
  Кластеризация

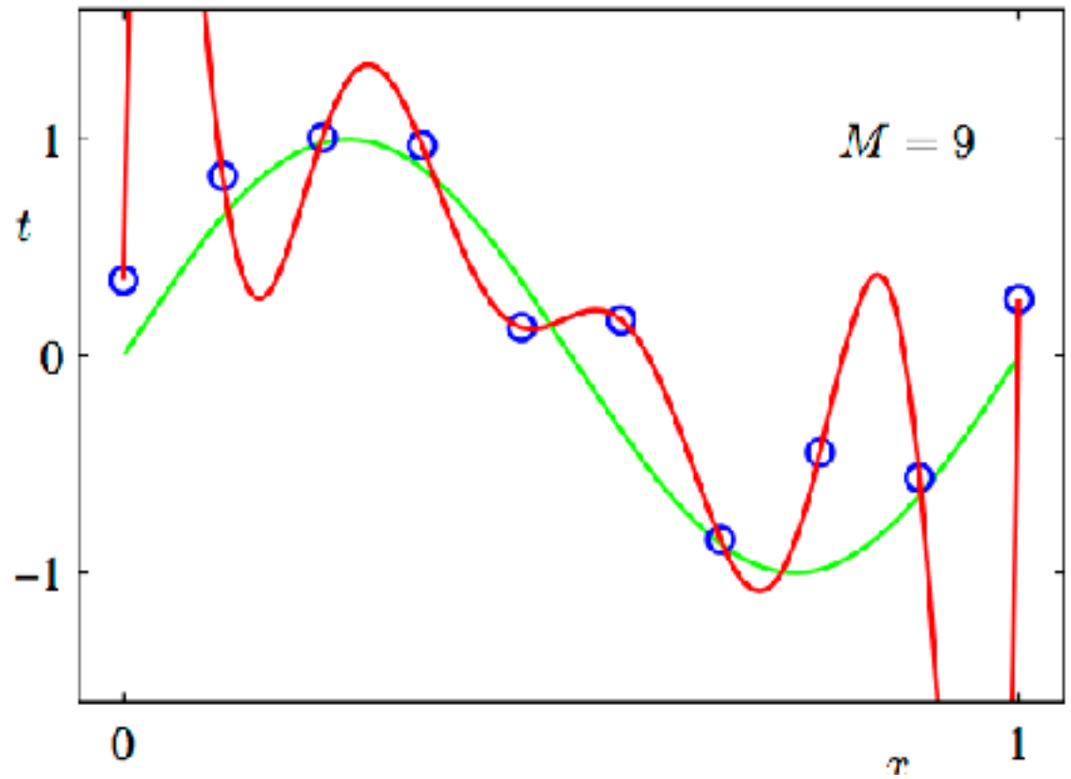
- Обучение с подкреплением
  Искусственный интеллект

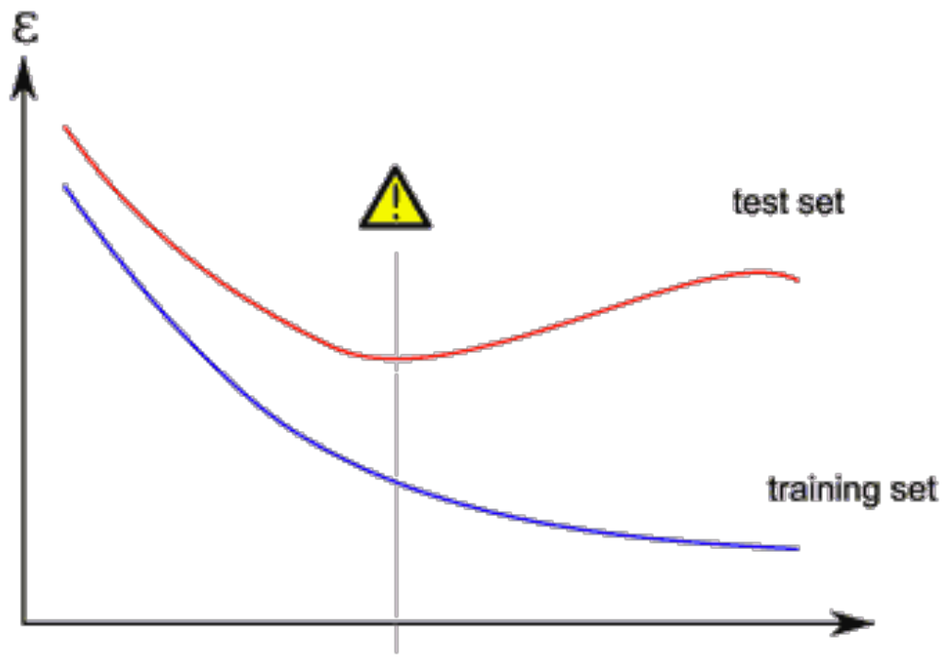
Если все так просто,
зачем нужны дата саентисты?

- Разнообразие подходов
- Правильная постановка и интерпретация
- Сложности при реализации

Проблема переобучения

- Выбирая более сложную модель, всегда легко можем получить 100% точность
- Запоминание vs Обобщение
  - Слишком простая модель будет неточно описывать
  - Слишком сложная - переобучаться
  - Необходим баланс





Проблема переобучения

- Решение
  - Обучение на Train, тестирование на Test
  - Кросс-валидация
  - Упрощение модели путем регуляризации
    - L1, L2 регуляризация
    - Ограничения на параметры
    - Dropout для нейронных сетей

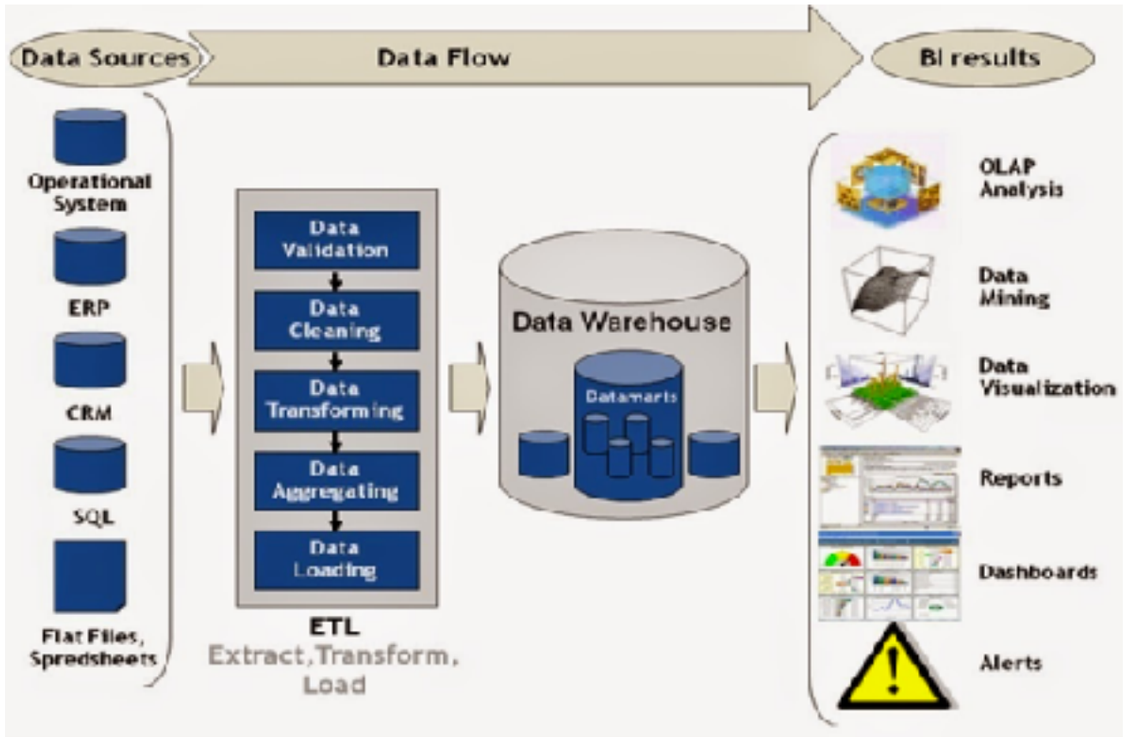
Наличие данных

- В 99% случаев готовых данных нет
  Надо готовить => исследование, экстракция, интеграция, консолидация, агрегация, качество, шедулинг

- Какие данные сработают - неизвестно
  Надо выгружать все по максимуму

- Каких-то данных вообще нет
  Надо организовать сбор

- Если данные большие
  Необходима архитектура хранения и распределенного обучения моделей



до 80% проекта

Качество данных

Что не так с данными

- Ошибки
- Пропуски

Как действовать

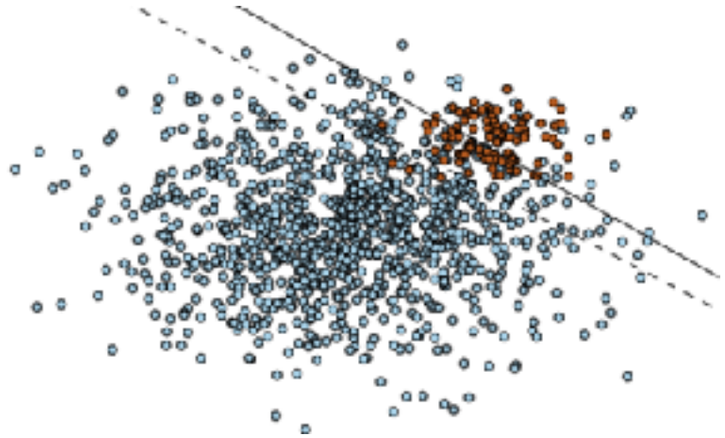
- Исключить некачественный атрибут
- Исключить некачественные кейсы
- Импутация
  - Замена на среднее значение/медиану/новую категорию
  - Предсказание пропущенного значения по контексту

Количество данных

- Сложная модель => требуется больше данных
- Мало данных => модель будет переобучена
  - Rule Of Thumb
    - Минимум 20 кейсов каждого класса на 1 предиктор
    - Переменная с 10 категориями = 10 атрибутов
  - Решение
    - Использовать простую модель
    - Дальше собирать данные

Несбалансированность классов

- Пример. Переход по рекламной ссылке (1% переходов)
- Accuracy Paradox - всё в один класс
- Модель - оптимальна, но абсолютно бесполезна
- Похожий пример из трейдинга
- Решение
  - Задавать цены ошибок 1 и 2 рода в $
  - Использовать ROC AUC вместо Accuracy
  - Oversampling/Undersampling



Наличие только одного класса

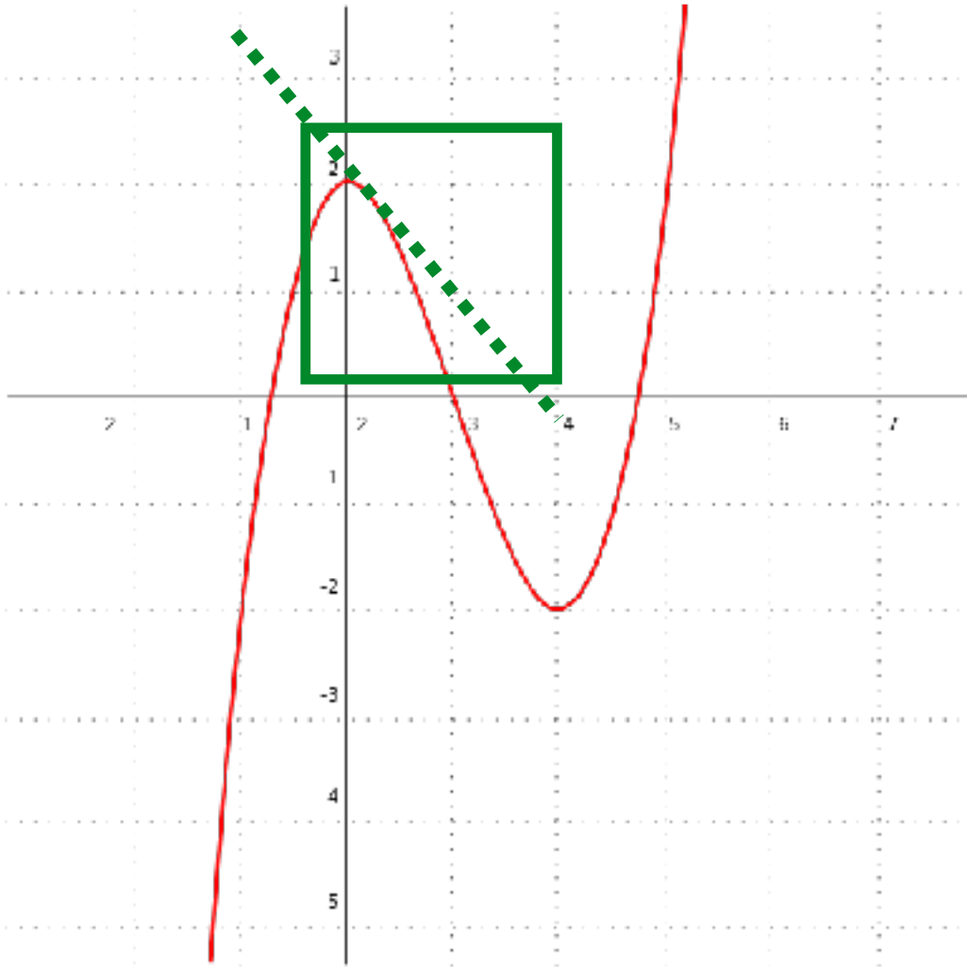
- Пример. Аудитория интернет магазина.

- Есть кнопка Like, но нет кнопки Dislike
  Знаем кому нравится, не знаем кому не нравится

- Нужны обходные пути
  Например, трактовать отсутствие покупки как Dislike

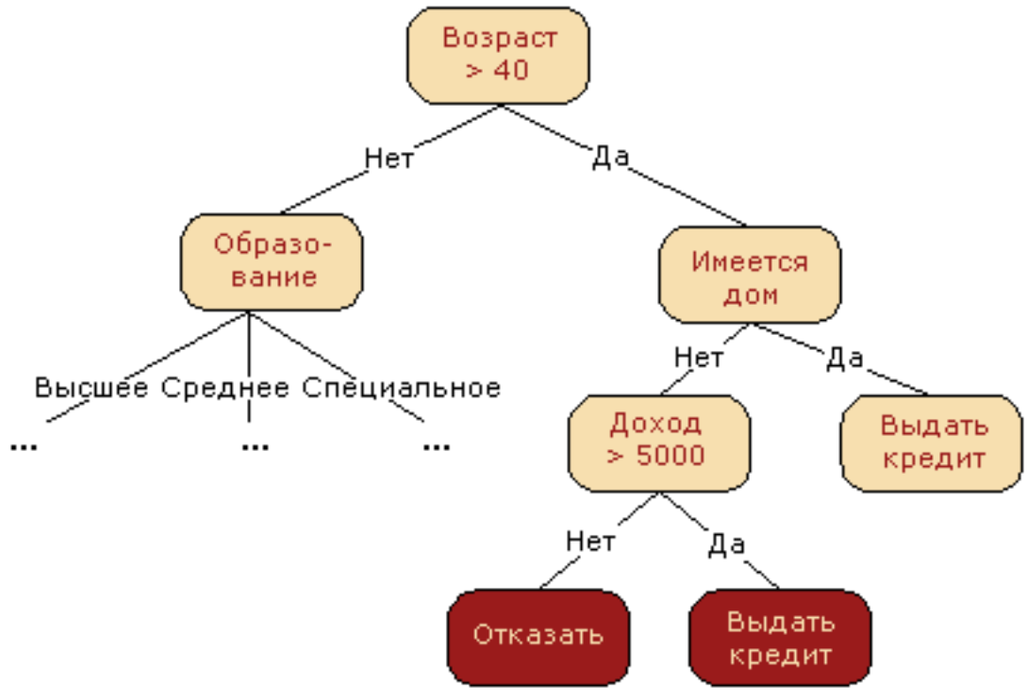
Смещенность выборки

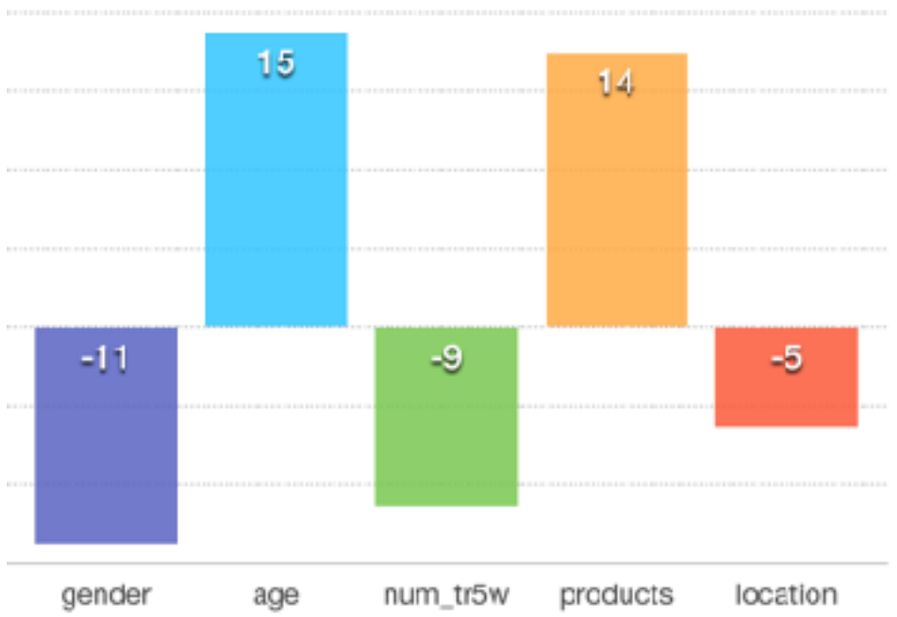
- Пример. Маркетинговый отдел банка.
- Запущенная ранее стратегия отбора перспективных клиентов
- История коммуникаций есть только по этим клиентам
- Решение
  - Запускать доп. тестовую кампанию по всем клиентам (холодный старт)
  - Оставить предварительно отобранный сегмент
  - Экстраполировать модель на всю базу 👎



Интепретируемость модели

- Модель должна быть понятна
- Сложная модель => менее интерпретируемая
- Для банков это требование регулятора
  Почему модель отнесла заявителя в класс “отказать”
- Стандарт для банков
  логистическая регрессия
  деревья решений





Параметры типичного ML-проекта

- Проект ~ создание рабочей предиктивной модели

- Длительность: 1-12 месяцев:

- Подготовка данных
  плохо автоматизируема
  прогнозируема

- Моделирование
  автоматизируема
  непрогнозируема

Роли на проекте

- Менеджер проекта
- Data Scientist
- Бизнес-аналитик
- Разработчики
- DevOps
- Тестировщик

Параметры типичного ML-проекта

Данные

- CSV-файл
- Выгрузка таблиц
- Доступ в БД

Требования к модели

- Классификация
- Скоринг

Формат результата

- Формула
- Скрипт
- Workflow
- Полноценное решение (сбор, хранение, аналитика, отчетность)

Среды моделирования

- Языки прототипирования
  Python, R, Matlab

- UI решения
  SPSS, SAS, Rapidminer, Knime

- Облачные решения
  Microsoft Azure, Amazon AWS, IBM Bluemix

- BigData решения

  Spark ML

Если точность модели не устраивает

- Собрать/рассчитать новые данные
- Попробовать новые алгоритмы
- Решать другую задачу

Усложнение модели редко дает большой эффект

90% результата достигается на первых 10% усилий

Ключевые особенности любого ML-проекта

- Неопределенность цели
- Неопределенность результата
- Неопределенность алгоритма решения

…И за это еще платят!

О методологии CRISP-DM

- Бизнес-анализ
  Business Understanding
- Анализ данных
  Data Understanding
- Подготовка данных
  Data Preparation
- Моделирование
  Modeling
- Оценка решения
  Evaluation
- Внедрение
  Deployment

Мифы о машинном обучении

- Машинное обучение = Big Data
- Вся математика придумана в 70е годы, сейчас ничего нового
- Машинное обучение = нейронные сети
- Через пару лет построят искусственный интеллект, который заменит большинство профессий
- Чем сложнее модель, тем она лучше работает

Вопросы?

email: k.kotochigov@cleverdata.ru
skype: konstantin.kotochigov