Дизайн систем машинного обучения

2. Основы проектирования ML-систем

https://ods.ai/tracks/ml-system-design-23

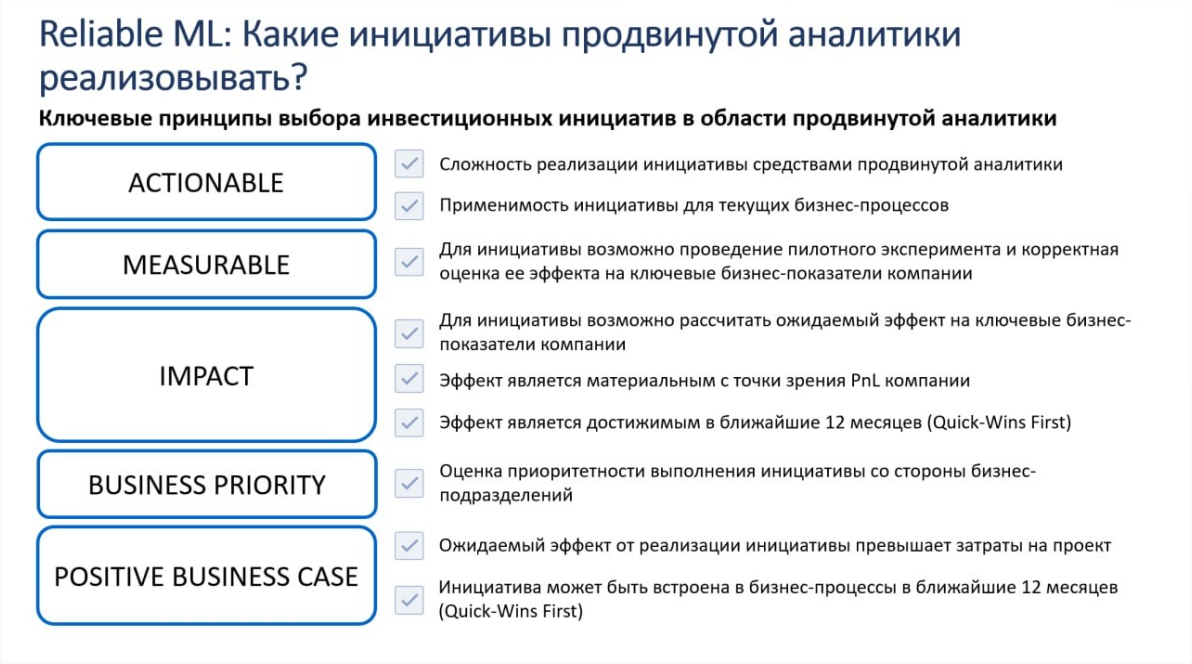
План курса

1) ~~Машинное обучение на практике~~
2) **Основы проектирования ML-систем — Вы находитесь здесь**
3) Обучающие данные
4) Подготовка и отбор признаков
5) Выбор и обучение ML-модели
6) Улучшение модели через данные
7) Оценка качества модели
8) Развертывание
9) Диагностика ошибок и отказов
10) Жизненный цикл модели
11) Потоковые данные
12) Языковые модели в продуктовом окружении
13) Временные ряды и графы
14) Безопасность, этика и восстание машин
15) Интеграция в бизнес-процессы

Инвестиционный цикл

#reliable_ml

Стоит ли запускать проект



#reliable_ml

Как проект может принести пользу

- Явно:
  - увеличить продажи
  - снизить затраты
- Неявно:
  - увеличить удовлетворенность пользователей
  - увеличить удовлетворенность персонала
- Совсем неявно:
  - Увеличить инвестиционную привлекательность
  - Перенести операционные затраты в капитальные

Метрики проекта

- Бизнес-метрики: влияние на ключевые показатели бизнеса
  - Прибыль
  - Доля рынка
  - Привлекательность для инвесторов
  - F1, ROC-AUC, Accuracy, Precision, Recall — интересно, но не нужно
- Инженерные метрики:
  - Задержка
  - Пропускная способность
  - F1, ROC-AUC, Accuracy, Precision, Recall — не всегда можем измерить
- Дерево метрик — от Latency до денег, например

Прокси-метрика

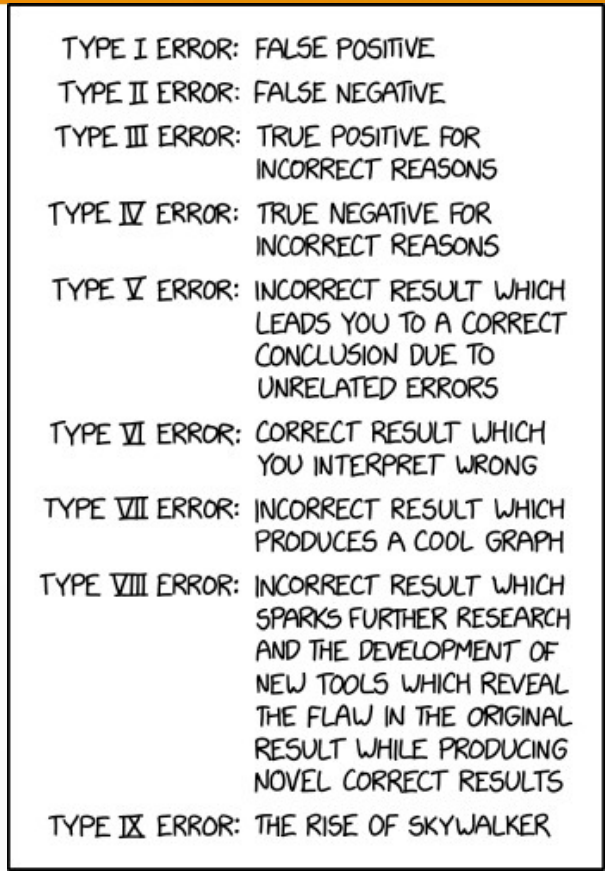
- Если что-то не можем измерить
- Если не можем измерить в нужное время
- Бывает: модель сломалась, выручка выросла
- Прокси-метрика из знаний о бизнесе
- Прокси-метрика из корреляций в данных
- Комбинировать ML и знания о бизнесе

https://habr.com/ru/company/retailrocket/blog/591205/

ML-метрики

- Что будем измерять
- Как будем измерять
- Baseline — с чем сравнивать
- Цена ошибки 1 и 2 рода
- Требования к качеству
- Оценка достоверности
- Надежность, масштабируемость

https://www.explainxkcd.com/wiki/index.php/2303:_Error_Types



Baseline

- Прежде чем улучшать, найдите, с чем сравнить
  - Существующее на рынке решение
  - Простое решение на правилах
  - Прототип с использованием чужого API
  - Качество решений человека
  - Решения конкурентов

Разная цена ошибки

- False positive
  - Нашли то, чего нет
- False negative
  - Не нашли то, что есть
- Как мы измерим цену ошибки?
- Сколько стоит убитый пациент?
- Как измерить удобство?
- Unknown Unknowns — не знаем, что это нужно измерять

Требования к качеству модели

- Самоуправляемый автомобиль:
  - Высокие требования к качеству
- Подсказки при вводе текста
  - Не такие высокие требования
- Рекомендации в интернет-магазине
  - Даже если совсем не то, все равно найдут поиском и купят
- Как формализовать?
  - Например, попробовать выразить в деньгах
  - Насколько вырастет прибыль, если увеличить accuracy на 1%

Дизайн систем машинного обучения

- Процесс принятия решений про
  - Интерфейс
  - Алгоритмы, данные
  - Программную инфраструктуру
  - Оборудование
- Чтобы соответствовать требованиям к
  - Надежности (reliable)
  - Масштабируемости (scalable)
  - Обслуживаемости (maintainable)
  - Адаптируемости (adaptable)

Нет ограничений — нет дизайна

- Ограничения — частный случай требований
- Источники требований (SWEBOK v 3.0):
  - Цели (бизнес-задачи)
  - Знания предметной области
  - Заинтересованные лица проекта
  - Бизнес-правила и регулирование
  - Окружение, в котором будет работать система
  - Организации, создающие и эксплуатирующие систему

Железные ограничения

- Пропускная способность (сети, интерфейса, шины)
- Место для хранения данных
- Скорость поиска / произвольного доступа
- Размер оперативной памяти/ кэша / VRAM / ...
- Структура хранилища и условия доступа к нему
- Задержки между обращениями
- Систематические отказы оборудования
- Ограничения по энергоснабжению и охлаждению

Пример — подсчет людей на видео

- Пусть наш сервис обрабатывает видеопоток с камер. На старте у нас будет 10 клиентов, через год мы хотели бы иметь 100.
- У клиента будет от 1 до 100 камер, в среднем 20
- Видеопоток 6,5 Мбит/сек.
- Клиент не сможет отдавать в интернет 650 Мбит.
- Мы не сможем принять 13Гбит
- Расходимся

https://tfortis.ru/support/calc-ip/

Работаем с ограничениями

- Меньше видеопоток — пусть 1Мбит
- Для клиентов с 30 камерами и более ставим сервер на их площадке (их будет 10%)
- Пусть клиент запрашивает сервер из пула серверов
- Итого средний трафик с клиента — 5Мбит, 450Мбит, вполне подьемно
- Балансировку можно доверить облаку (но дорого)
- Лучше прямо в камере считать, конечно

Бизнес-ограничения

- Время ответа системы
- Требуемая пиковая нагрузка
- Требуемая пропускная способность
- Требования к стоимости обработки
- Требования к надежности
- Требования к поддерживаемости
- Требования к квалификации персонала

Организационные ограничения

- Наличие и квалификация персонала
- Уже имеющаяся у организации инфраструктура
- Деньги
- Процессы
- Время(можно купить), 80% на доводку решения
  - Мощнее сервера
  - Больше людей на разметке
  - Больше людей в разработке
  - Купить готовое решение

Существующие системы

- Проблема интеграции в существующие системы
  - Получить данные
  - Загрузить данные обратно
  - Превзойти качество существующей системы
- Использование имеющейся инфраструктуры
- Обучение пользователей и операторов системы
- Изменение бизнес-процессов

Compliance & Privacy

- Можем ли мы отдать данные на разметку?
- Можем ли мы хранить данные в облаке?
- В какой стране мы должны хранить данные?
- Можем ли мы собирать данные пользователей?
- Обязаны ли мы шифровать данные?
- Можем ли мы использовать сторонние сервисы?
- Какие законы мы должны соблюдать?

Надежность Reliability

- Свойство объекта сохранять во времени в установленных пределах значения всех параметров, характеризующих способность выполнять требуемые функции в заданных условиях применения, технического обслуживания, хранения и транспортирования →
- Как мы поймем, что все работает правильно?
- Как мы отличим ошибочный предикт от хорошего?
- Подробнее — в лекции про мониторинг.

Масштабируемость Scalability

- Способность системы, сети или процесса справляться с увеличением рабочей нагрузки (увеличивать свою производительность) при добавлении ресурсов (обычно аппаратных) →
- Если повезет, вам потребуется масштабируемость
- Растет трафик и объем данных
- Растет количество моделей и сценариев использования
- Обсудим в дальнейших лекциях

Обслуживаемость Maintainability

- Приспособленность к восстановлению работоспособного состояния после отказа или повреждения →
- Аварии серверов и сетей связи
- Проблемы с данными
- Проблемы с персоналом
- Обновление оборудования и схемы данных

Адаптируемость Adaptability

- Способность адаптироваться к меняющимся обстоятельствам →
- Меняются:
  - Бизнес-требования
  - Структура данных
  - Доступность данных
  - Оборудование
  - Инфраструктурные сервисы
  - Распределение данных (data shift, target drift)

Дополнительные материалы

- Building Secure and Reliable Systems, глава 4
- ML system design: 200 case studies to learn from