Из чего состоит хороший ML-проект

2023.04.27 @ ИТМО

Дмитрий Колодезев, ООО Промсофт

Про меня и эту презентацию

- На фото — я еще не умел руководить проектами
- Но рулить уже пытался
- Успешно завершил и провалил много проектов
- Рассказ — про то, как начать проект правильно

ML-проект

- Проект, значительную часть которого составляет разработка и/или внедрение модели машинного обучения
- Сам проект обычно не про машинное обучение, машинное обучение в нем — волшебная смазка, заставляющая его работать быстрее
- Улучшаем процесс
  - Быстрее, лучше, дешевле, проще, повторяемее
- Улучшаем надпроцесс
  - Быстрый старт, масштабирование, свертывание

Гипотеза

- Предположение, которое не имеет достаточных фактических подтверждений, но представляется вероятным и не опровергнуто →
- Принцип верифицируемости →
  - Существует возможность подтвердить
- Принцип фальсифицируемости →
  - Существует возможность опровергнуть
- Большинство «фактов» и «ограничений» в бизнесе — гипотезы.

Составляющие хорошего ML-проекта

- Насущная проблема
- Доступные в нужное время данные
- Доступ к эксперту предметной области
- Гипотезы о том, как можно решить проблему
- Возможность проверять гипотезы
- Возможность внедрить решение
- Возможность измерить результат

Насущная проблема

- Отличный вариант — проблема есть у ваших знакомых, и они готовы платить за ее решение
- Если люди не готовы платить за решение проблемы, скорее всего:
  - или проблемы нет (и тогда не надо ее решать)
  - или в вас не верят (и надо спросить — почему?)
  - чаще — и то и другое
- Просите больше денег. Просите денег вперед.
  Так вы проверите гипотезу, что проблема реальна
- Скепсис экспертов обычно обоснован. Но это тоже гипотезы

Гипотеза о проблеме

- Если проблема была, ее кто-то как-то решил
- Решение было оптимально тогда и там
- Проблема могла возникнуть у новых людей
- Могли измениться условия задачи
  - каждая новая проблема это возможность
  - каждая новая технология это возможность
- Какое-то плохое решение стало хорошим
- Главное в гипотезе — как ее проверить

Если проблемы нет

- Возьмите отрасль, которая вам любопытна
- Поищите роли
- Посмотрите, какие задачи решаются в этой роли
- Посмотрите, куда уходит много денег, времени, трудозатрат, эмоциональных ресурсов.
- Там может прятаться проблема

Данные, доступные вовремя

- Машинное обучение / искуственный интеллект построен на поиске и использовании шаблонов в данных.
- Нет данных — нет машинного обучения.
- Все заказчики уверены, что данные у них есть
- Но нет
- Гипотеза о данных — какие есть, как проверить?
- Просите примеры данных вперед

Если данных нет

- Можно собрать самим
- Можно собрать «из мусора» - логов, журналов звонков, которые не принято анализировать
- Неструктурированные данные часто - возможность
- Можно сгенерировать самим (синтетические данные)
- Гипотезу о том, что синтетические данные похожи на те данные, которые будут у заказчика, надо проверять

Данных нет вовремя

- Надо учить модель — данных нет
- Надо делать предсказание — данных нет
- Разметка «вызревает» долго
- Событий важного класса мало
- Мир изменился, данные устарели
- Мир остался тем же, данные более недоступны

Эксперт предметной области

- Кто-то должен понимать, как работает процесс, который вы хотите улучшить
- Показывайте свое решение специалистам. Чаще.
- Пусть расскажут, почему не сработает
- В отрасли есть вещи, «о которых все знают», и вам про них не скажут
- Идите в Гэмба — к прилавку, на склад, на завод, подслушивайте, подсматривайте и переспрашивайте

Гипотезы о решении

- Если мы что-то добавили, что-то нужно выкинуть
- Гипотезы «на минус миллион»
- Ищите самую дешевую в проверке гипотезу
- Как вы поймете, что решение работает?
- Смотрим назад из будущего, и рассказываем, как мы туда попали

Эксперименты

- Гипотезы, которые мы не можем проверить — бесполезны
- Гипотезы, которые мы проверили неправильно — опасны, как ненадежные ступеньки
- Планирование, проведение и анализ эксперимента — учиться у рекламистов и эпидемиологов
- Скорость проверки гипотез важнее их качества

Возможность внедрения

- Гипотеза о возможности внедрения самая важная
- Размер эффекта
- Доступность данных
- Дополнительный персонал и оборудование
- Как изменится поддержка и обучение
- Станет ли бизнес-процесс менее гибким?
- Какие процессы и потоки информации сломаются?
- Кому невыгодно внедрение и почему?

Параллельные реальности

- Все вышеописанное — решаемо, если договориться
  - У проекта много заинтересованных лиц
  - У каждого своя «параллельная реальность»
- Большинство людей участвуют впервые
- Важные вопросы остаются незаданными
- Решение — ML Design Doc — фреймворк, позволяющий избежать грубых ошибок при планировании ML-проекта

Вопросы лучше потом

Слайды тут

dkolodezev

promsoft

dmitry_kolodezev

https://kolodezev.ru/download/ml_project_start.pdf