Устойчивость ML-моделей

Как часто дообучать модель и как делать это правильно

2023.04.29 Дата Завтрак @ Красноярск

Дмитрий Колодезев
ООО Промсофт, Новосибирск

Про меня

Консалтинг
MLSystemDesign
Datafest2023
ReliableML
OpenDataScience
Бег
InterpretableML
Промсофт
Датазавтраки
Парус
Управление_ML_командой
Анализ_данных

О чем мы тут

- fit — predict — profit!
- Иногда сразу плохо работает
- Иногда сначала хорошо
- Потом как обычно
- Иногда не понятно, работает ли вообще

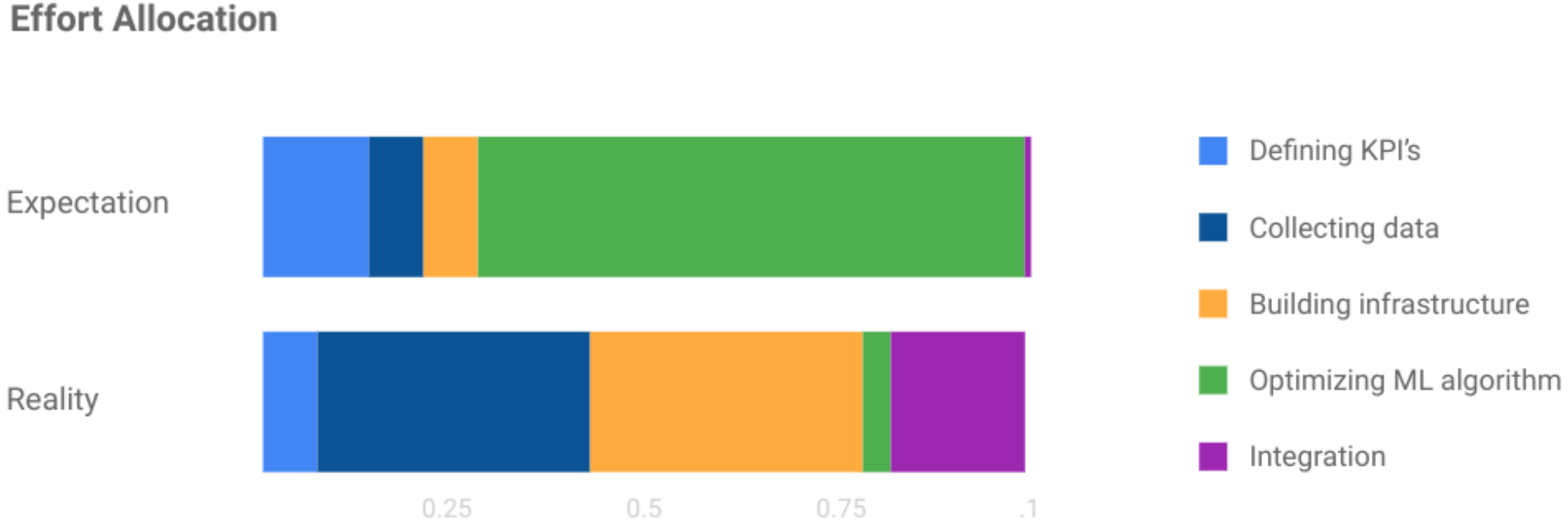
data scientist performing the fit-predict-profit process with joy and ease

Вот об этом и поговорим

- Кто виноват?
- Что делать?
- Кто будет за это отвечать?
- Почему именно я?
- А как правильно?

data scientist in a suit feeling guilty and desperate about not being good enough

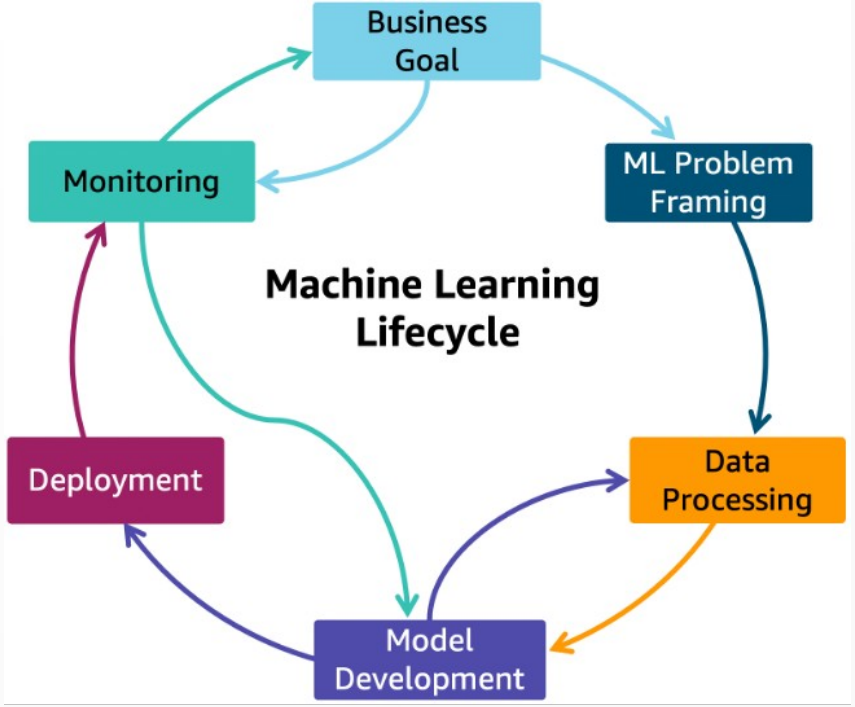
Как мы делаем ML



https://medium.com/thelaunchpad/the-ml-surprise-f54706361a6c

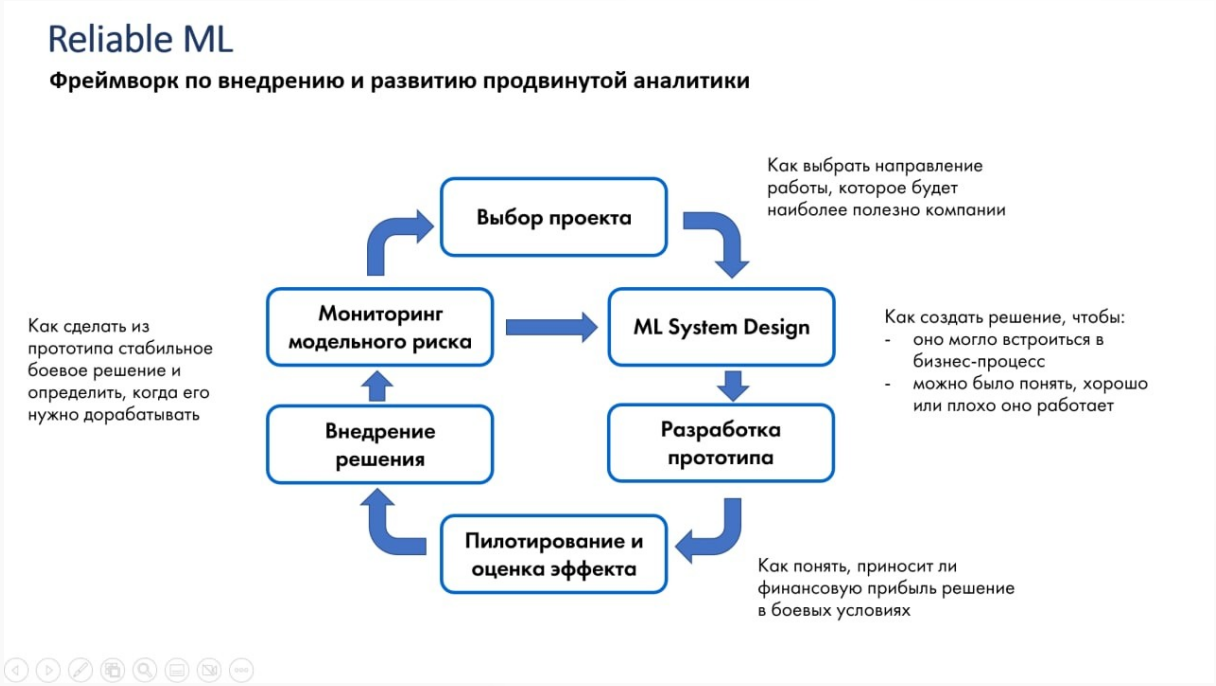
Круговорот моделей в природе

CRISP-DM



AWS ML LifeCycle

Рекламная пауза



#ReliableML

Оказывается, его надо кормить

- Ожидания:
  тратим деньги
  получаем систему
  система приносит деньги
- Реальность:
  тратим деньги
  получаем систему
  система требует денег

data scientist in a suit looking inside his empty leather wallet, white bills with red numbers flying around him in a chaotic manner, blue background

Хрупкость ML систем

- Нассим Талеб «Антихрупкость»
- Александр Бындю «Антихрупкость в IT»
- Мы заменяем гибких людей на жесткие системы
- Когда окружение меняется, системы ломаются

data scientist in a suit holding complex diagram elements made of glass, several are falling, several are broken on the floor

Давайте сделаем устойчиво

- Устойчивость
- Меняются данные
- Меняется мир,
  стоящий за данными
- Нас атакуют!

data scientist in a suit holding in his hands a complex but stable construction made out of different glass elements

Проблемы с данными

- Выбросы / аномалии
  - Обрезать + флаг обрезки
- Пропущенные значения
  - Учить с пропусками
- Редкие значения
  - Добавлять эвристики
- Другие распределения
  - Rejection class

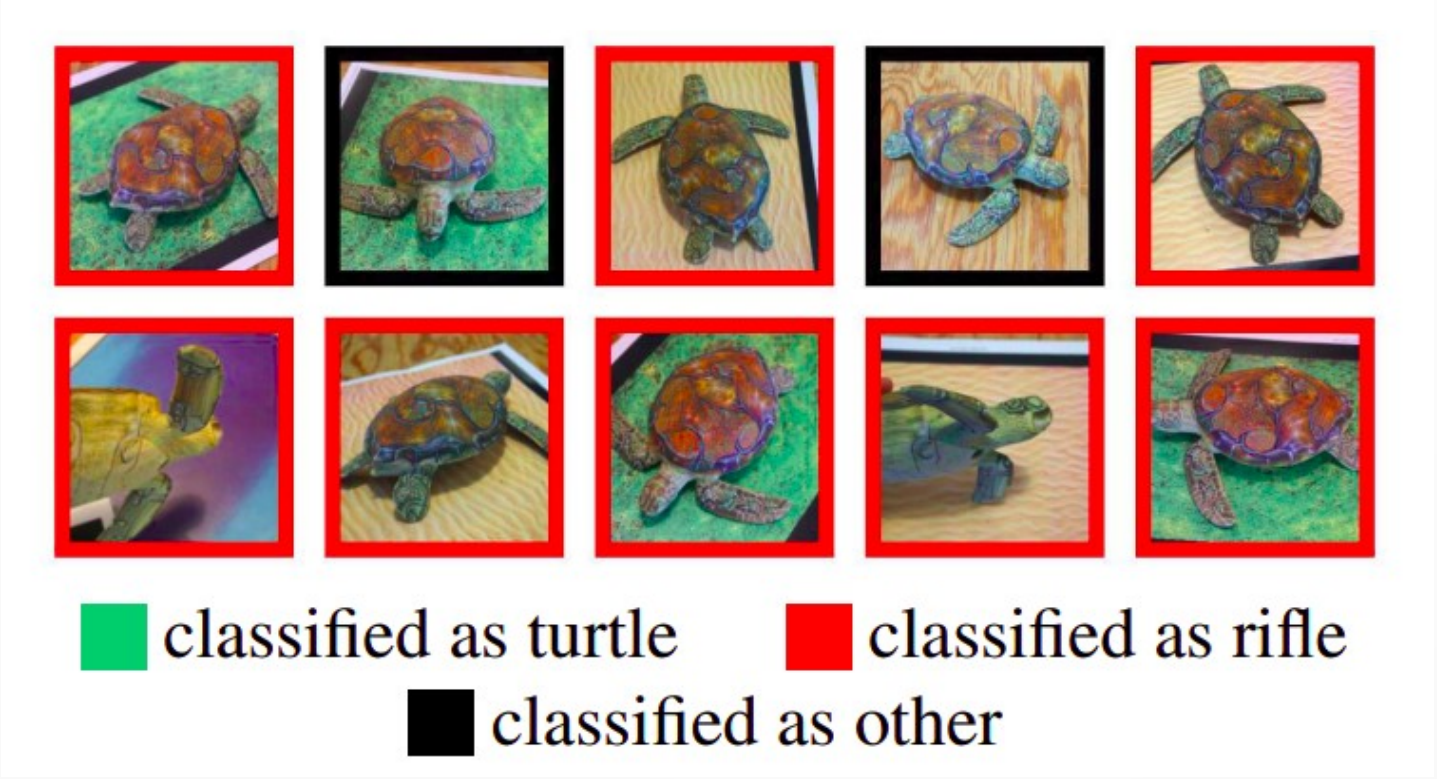
data scientist thinking about chaotic diagrams in thought bubble

Изменился мир

- Дрейф данных
- Пропали признаки
- Новые признаки
- Новые токены
- Новые классы
  в тяжелых моделях

steampunk laptop

Нас атакуют



https://arxiv.org/abs/1707.07397

Вояки пока держатся

KEY FINDINGS

- Adversarial attacks designed to hide objects pose less risk to U.S. Department of Defense applications than academic research implies.
- In the real world, such adversarial attacks are difficult to design and deploy because of high knowledge requirements and infeasible attack vectors—there are often less expensive, more practical, and more effective nonadversarial techniques.
- Fusing data and predictions across sensor modalities, signal-sampling rates, and image resolution can further mitigate the risk of adversarial attacks.

CM RAND_RRA866-1

data scientist in a modern suit holding a sword piercing through complex structure made out of different diagram elements

Банки пока держатся

Data Fusion Contest 2023

Ежегодное соревнование по машинному обучению Data Fusion Contest. В 2023 году это турнир по Adversarial ML между командами атакующих и защищающих ML модели на транзакционных данных.

21 18 17 14 11 9 4 4 3 2 2 1 +

Ежегодное соревнование Data Fusion Contest 2023 продолжается! Регистрация открыта для участников до 2 апреля!

Вас ждёт уникальное соревнование по атакам и защите моделей машинного обучения в турнирном формате:

🗡️ В задаче **Атака** участники будут создавать атаки на нейросеть, обученную на данных транзакций.

🛡️ В задаче **Защита** — наоборот, учиться защищать свои модели от заранее оговоренного вида атак.

🏆 Призеров определят **Турниры** — лучшие команды обеих задач столкнутся друг с другом за призовой фонд в **2 000 000 рублей**!

https://ods.ai/tracks/data-fusion-2023-competitions

Поисковые машины - нет

накрутка поисковых подсказок

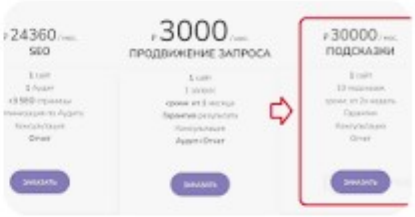
Все | Видео | Картинки | Новости | Книги | Ещё | Инструменты

Результатов: примерно 2 430 (0,38 сек.)

Накрутка поисковых подсказок относится к «серым» способам продвижения.

...

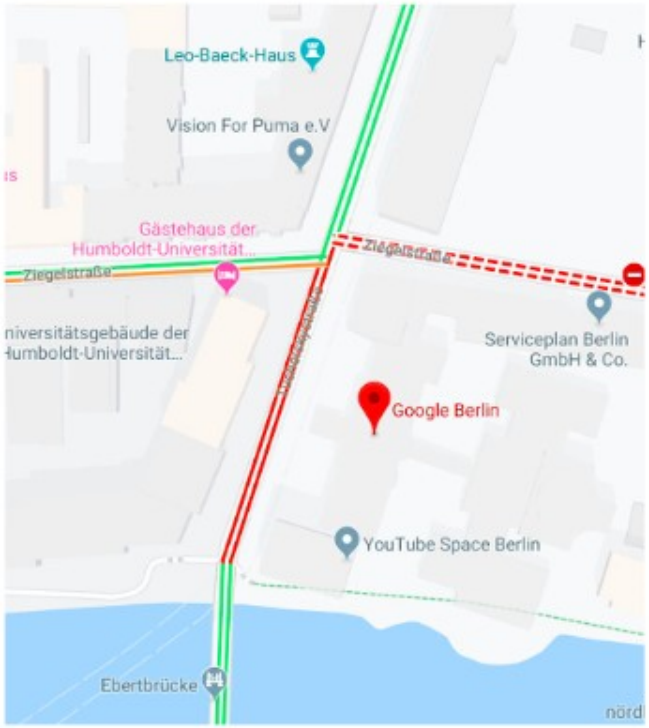
Эксперты, которые плотно работают с накруткой поисковых подсказок, отмечают следующие санкции за некачественную накрутку:



1. Чистка списка **подсказок**. ...
2. Временный или постоянный фильтр на вывод **подсказки** или бренда на определенном ключевом слове.

Ещё • 24 сент. 2022 г.

Против тачки нет приема



https://www.simonweckert.com/googlemapshacks.html

Что делать-то?

- Метки вызревают быстро:
  - Мониторить качество модели
- Метки вызревают медленно:
  - Мониторить прокси-метрики
  - Мониторить распределение предсказаний
  - Мониторить распределение признаков
- Валидировать входные данные
- Моделировать невязку модели
- Оценивать неуверенность модели

И вот мы знаем, что пора

- Собрать новые данные
- Дообучить модель
- Сравнить с текущей моделью
- Оставить ту, что лучше
- Повторить

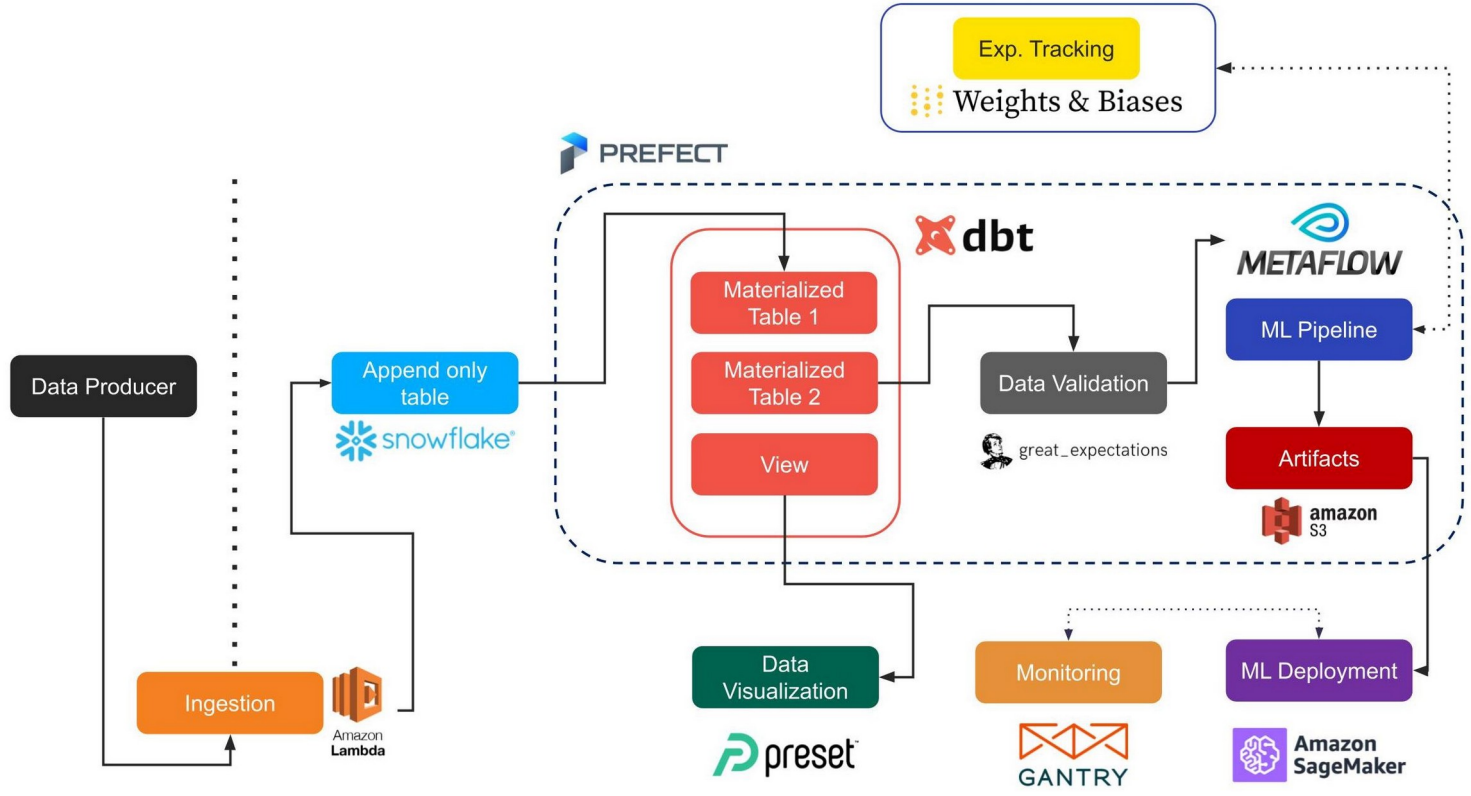
Как часто?

- Замена модели стоит денег
- Снижение качества модели измерить в деньгах
- Моделируем на исторических данных
  - Учим на данных запрошлый год и смотрим на ошибку
  - Полагаем скорость устаревания постоянной
  - Оцениваем, сколько мы теряем
- Считаем, как выгоднее
- Чаще чем половина срока службы датчиков

Есть еще онлайн-модели

- Например, River
- Переобучается на ходу
- Много узлов — везде по-разному
- Колебания сезонные — все время не так
- Надо пробовать, редко работает

Практический пример



https://github.com/jacopotagliabue/you-dont-need-a-bigger-boat

Все ходы записываем

- Хорошо бы сохранять:
  - Вектор признаков
  - Рассчитанный предикт
  - Версия модели
- Когда вызревают метки, добавлять
  - Дообучаем
  - Мониторим
  - Тонко настраиваем
- Дорого хранить, но очень полезно

Как решать, кто лучше?

- A/B тесты для терпеливых
- Interleaving для торопливых
- Канареечный деплой для уверенных
- Теневой деплой для осторожных
- Не глядя — для тех, кто любит приключения

Итого

- Моделируйте устаревание модели
- Проектируйте с учетом стоимости владения
  - Мониторинг
  - Переобучение / дообучение
  - Замена модели

Почитать

- Блог Evidently.AI
- 10 глава курса ML System Design
- Introduction to streaming for data scientists
- Переобучению быть или не быть
- Пару слов о дрейфе данных

Вопросы

Слайды тут

dkolodezev

promsoft

dmitry_kolodezev

https://kolodezev.ru/download/2023_ekaterinburg.pdf