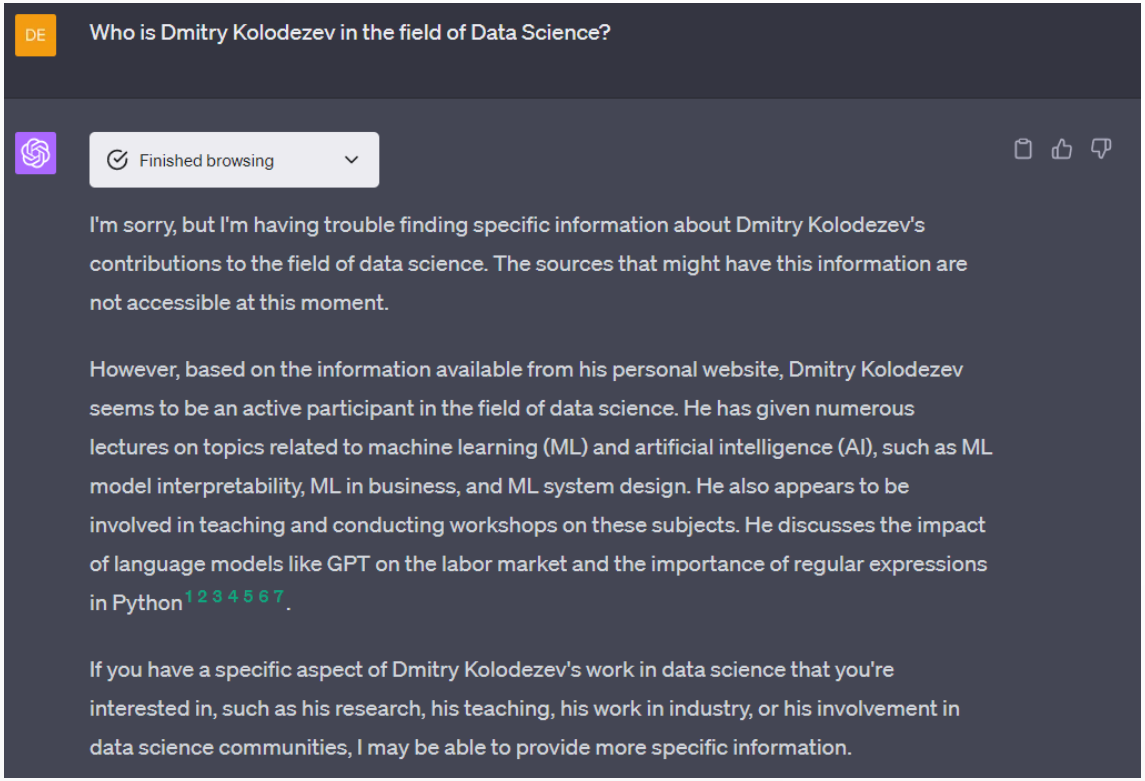
Reliable ML

Устойчивость ML-моделей после выкатки в прод

03.06.2023 ReliableML @ DataFest 2023

Дмитрий Колодезев
ООО Промсофт, Новосибирск

Про меня



KOLODEZEV.RU

О чем мы тут

- Reliable ML
- Устойчивость моделей
- Почему модель может работать плохо
- Как узнать, что она уже
- Что делать
- Когда уже пора
- Что почитать и анонсы

Reliable ML

- Название этой секции на Датафесте
- Канал в телеграмме
- Подход к разработке моделей, максимизирующий практическую полезность ML для бизнеса
  → (мы находимся тут) ←
- Неплохая книга (не наша)
- Будущая книга (в основном про ML Design Doc)

Жизнь после деплоя

- В прошивках оборудования
  - Медицинское оборудование, умные камеры
- В сервисах, построенных вокруг ML
  - Рекомендательные сервисы, поиск
- Как одна из частей бизнес-процесса
  - Скоринг, предсказание продаж, оценка сроков

Оказывается, его нужно кормить

- Ожидания от ML-разработки
  - Капитальные затраты
  - Повысят эффективность работы
- Реальность
  - Капитальные затраты
  - Операционные затраты
  - Менее гибкие процессы
  - Эффект внедрения нестабилен

Почему оно сломалось

- Могла сразу работать плохо или не работать вовсе
- Проблемы с данными
  - аномалии, выбросы, редкие значения, пропуски.
  - поменялось распределение данных
- Отказы
  - Программные, аппаратные, организационные
- Проблема с моделью
  - нестабильность, недоопределенность, низкое качество
- Атаки - см Атлас
- Неправильное использование — model card, datasheets

Хрупкость ML-систем

- Модели ищут шаблоны в исторических данных
- Шаблоны фиксированы и скрыты от пользователя
- Меняться будут
  - Бизнес-процессы
  - Классификаторы
  - Схема данных
  - Семантика данных
- Модели не приспосабливаются к изменениям
- И пользователь их приспособить тоже не может

Гибкость ML-систем

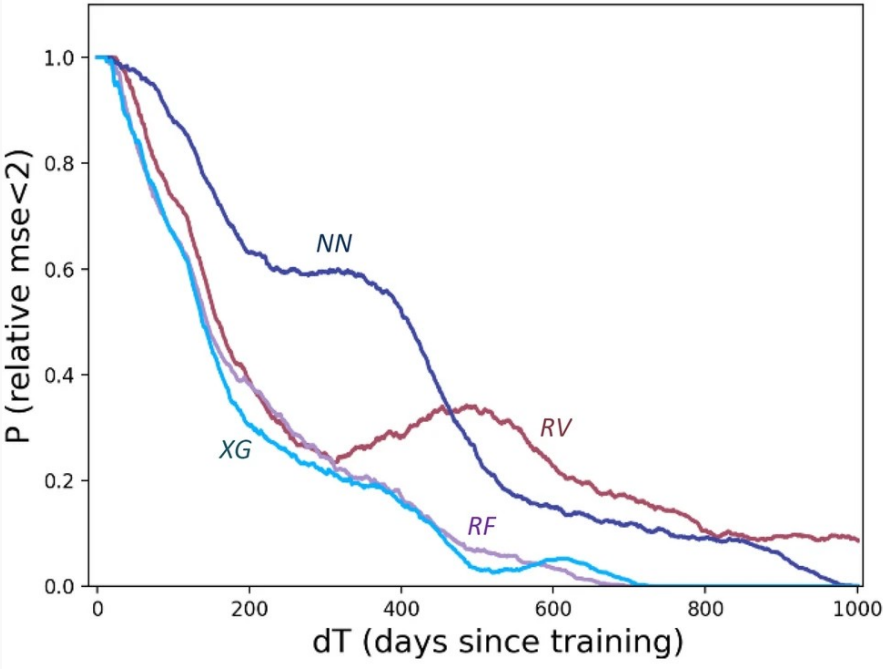
- Настройка порогов и весов
- KNN поверх эмбеддингов
- Регулярные выражения поверх LLM
- Low-Code системы
- Неструктурированные данные
  - Картинки
  - Языковые модели
- Переключение в ручной режим
- Автор в активном поиске антихрупких решений

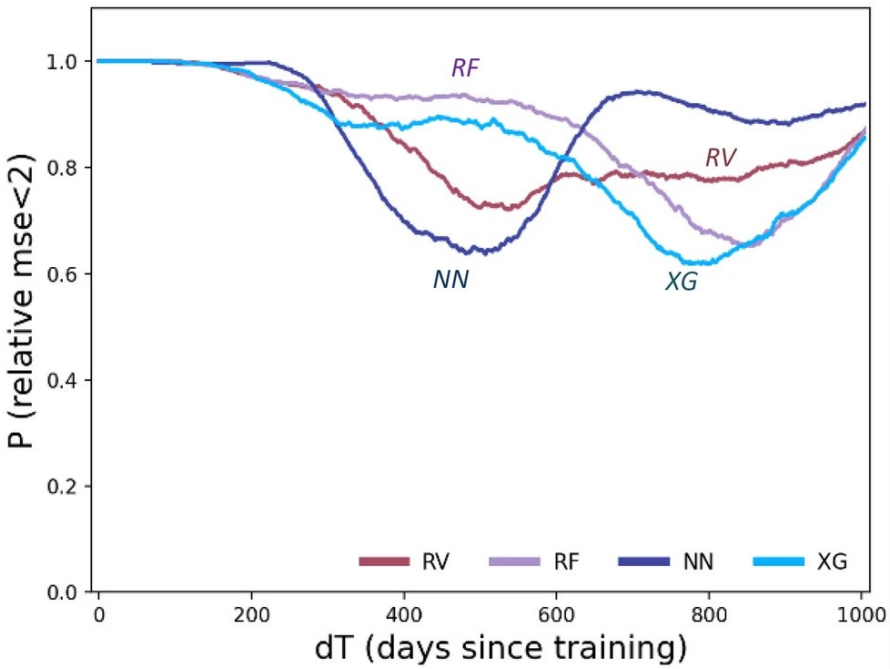
Как узнаем, что уже?

- Селфчеки т.е. health tests
- Прогноз скорости деградации на исторических данных
- Скорость вызревания разметки
  - Когда мы узнаем, что предсказание было правильным
  - Быстрая: можем позволить деградировать
  - Медленная: не можем позволить деградировать
- Если метки вызревают быстро
  - Отслеживаем качество: SLO, SLI, SLA

https://kolodezev.ru/mlsysd8.html

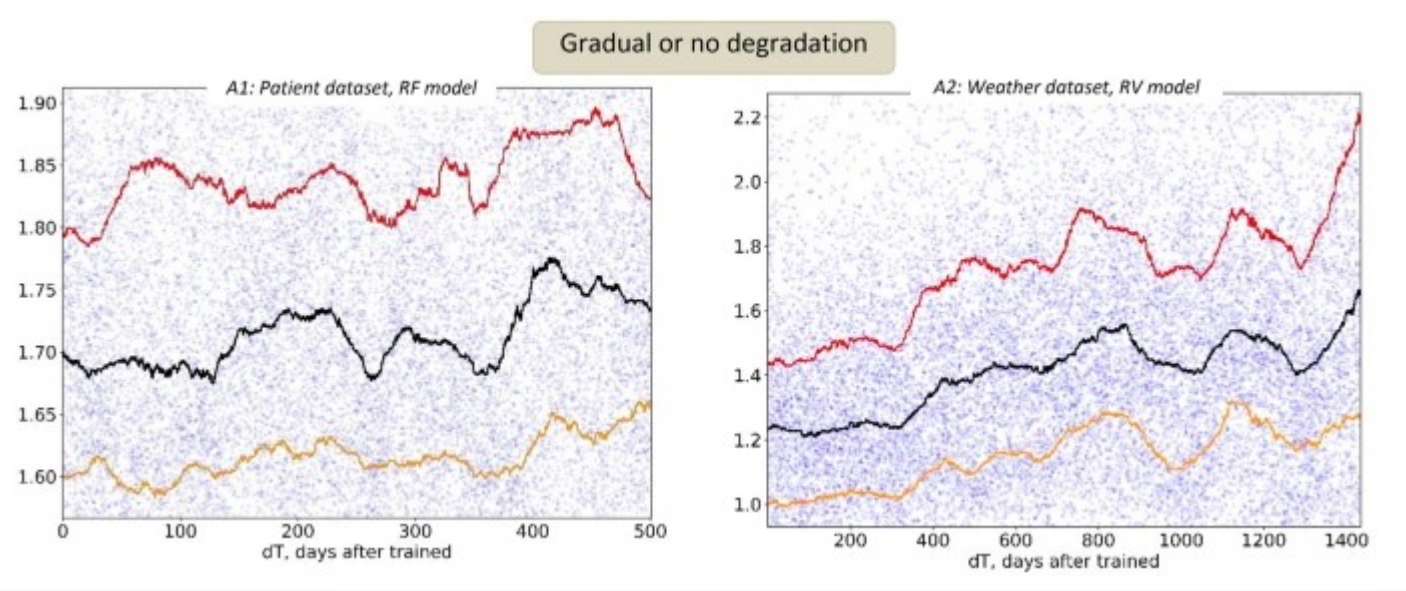
Деградируем





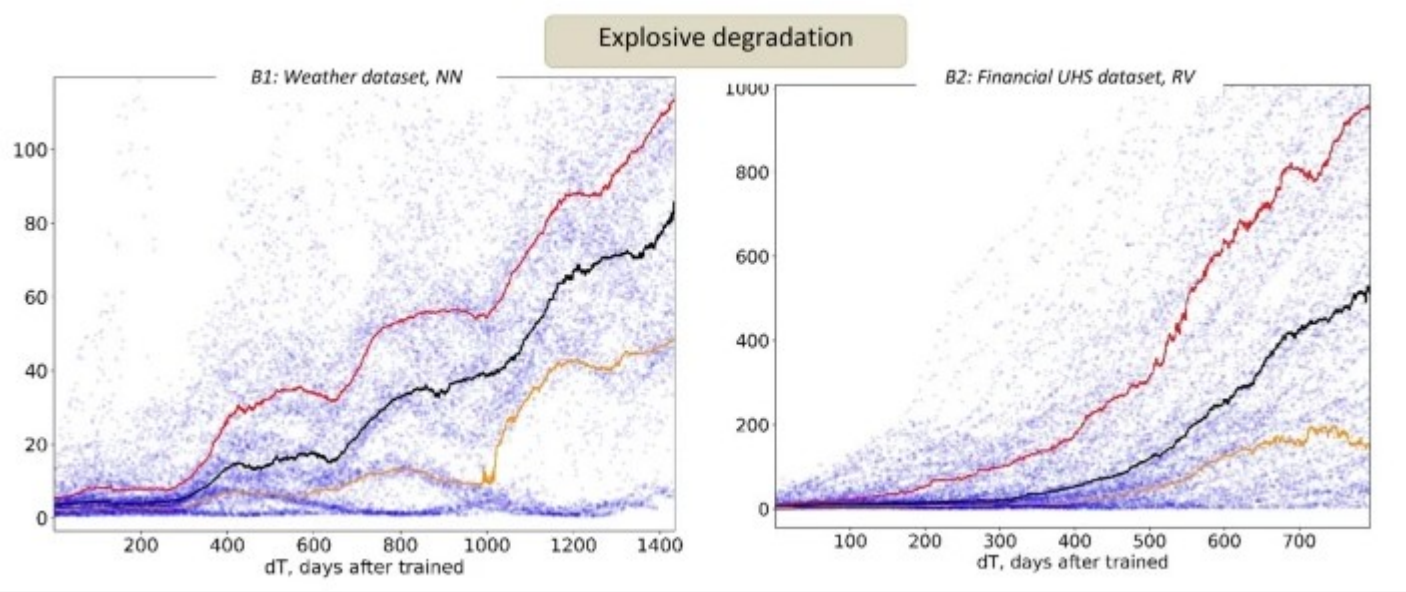
https://www.nature.com/articles/s41598-022-15245-z

Предсказуемая деградация



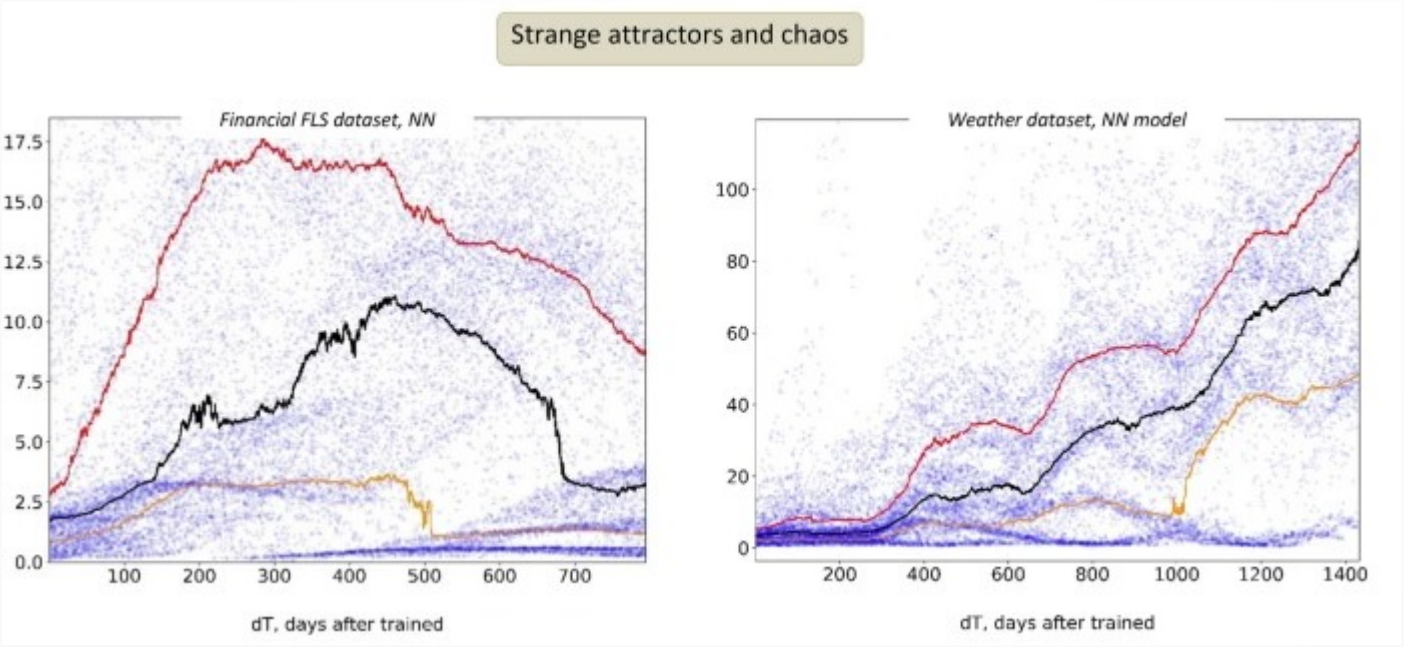
https://www.nature.com/articles/s41598-022-15245-z

Взрывная деградация



https://www.nature.com/articles/s41598-022-15245-z

Черт-те что



https://www.nature.com/articles/s41598-022-15245-z

Медленно вызревающие метки

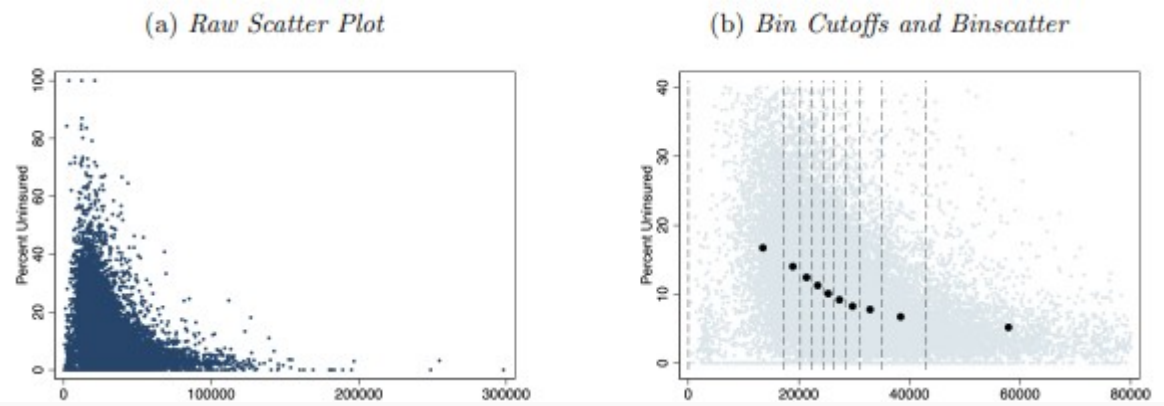
- Прокси-метрики
- Мониторинг распределения предсказаний
- Мониторинг сдвига данных
- Валидация входных данных
- Оценка уверенности модели
- Моделирование невязки
- Слушать пользователей
- Повышать прозрачность решения
- Запасные простые модели

Интерпретируемость и тд

- Интерпретируем предсказания
- Не должны удивлять экспертов
- Необходимо, но недостаточно — можно накрутить
- Парадокс прозрачности
  - Прозрачные модели надежнее
  - Прозрачные модели проще обманывать
- SHAP и LIME не моделируют причинность
- Посмотрите на контрфактические примеры

Кстати, EDA

- Идеальная модель выучит шаблоны в данных
- Шаблоны в данных можно посмотреть без модели
- Mean target plot, например
- Мое новое любимое binsreg - см On Binscatter

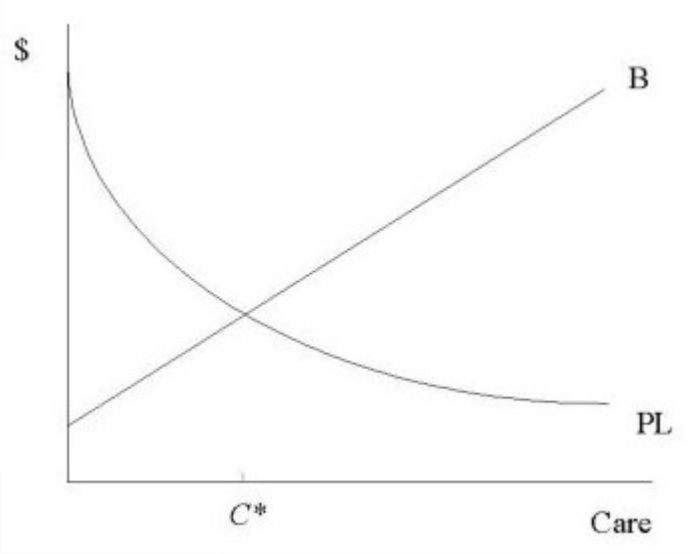


Предотвратим

- Аудит процесса подготовки данных
  - df.fillna() df.dropna()
- Детектирование аномалий
  - Клиппирование + флаг
  - Добавлять аномалии специально
- Заполнение пропущенных значений + флаг
  - Выбрасывать признаки специально
- Контролируемая деградация
- Проработанный ML Design Doc — видео, шаблон

Когда уже пора

- Сколько стоит некачественный предикт?
- Часто не знают
- Сколько стоит мониторить?
- Сколько стоит дообучить?
- Неопределенность
  - Ущерб случаен
  - Успех не гарантирован
  - Например, двойной запас



https://cyber.harvard.edu/bridge/LawEconomics/neg-liab.htm

Что мешает дообучать модели

- Малый поток данных
  - Синтетика и аугментация
- Медленное вызревание меток
  - Конструируем обратную связь, слабые метки
- Громоздкая подготовка данных
  - Сохранять предикты и признаки
- Трудно сравнивать модели
  - Интерливинг вместо AB тестов, откат модели

Анонс

- Курс по ML System Design, второй запуск осенью
- Переработанный материал
- С ~~блекджеком~~ лабораторными работами
- С промежуточными тестами
- С разделом про устойчивость моделей
- Только онлайн

Что почитать

- Supervisory Guidance On Model Risk Management
- Machine Learning for High-Risk Applications
- Graceful Degradation and Related Fields
- На DataFest 2023 много докладов про это
- Блог nannyML
- Блог Evidently AI
- Про устойчивость моделей
- Курс ML System Design

Вопросы

Слайды тут

dkolodezev

promsoft

dmitry_kolodezev

https://kolodezev.ru/download/model_sustainability_2023.pdf