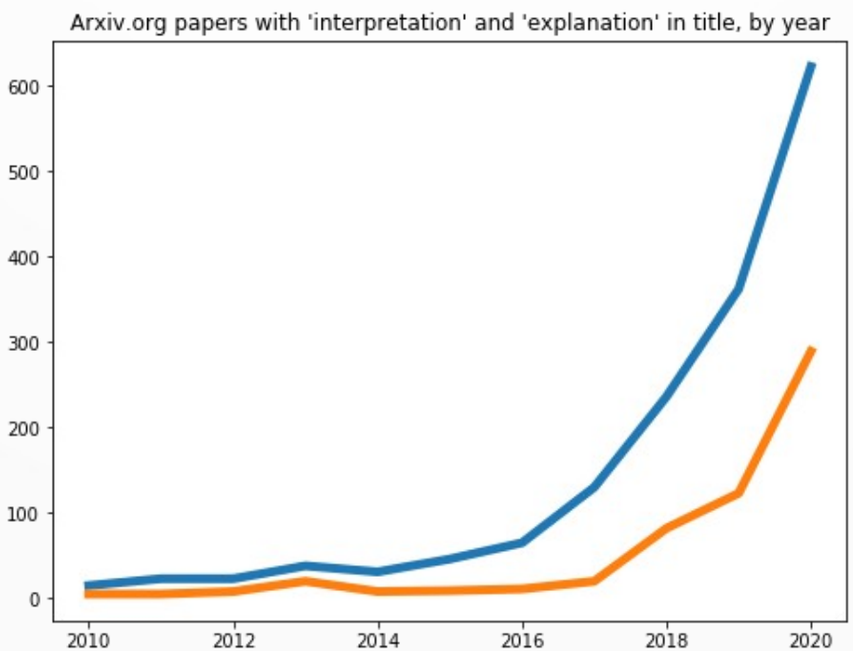
ДатаЕлка-2020
Интерпретируем модели машинного обучения

Дмитрий Колодезев
ООО Промсофт, Новосибирск

Итоги года

Это я, проинтерпретировал очередную модель

Интерпретируемость ML-моделей повзрослела:
- Shapley Values в каждом утюге
- на датафесте 7 докладов
- канал #interpretable_ml в слаке ODS



А кстати, про что это?

- Как модель принимает решения?
- Можем ли мы полагаться на эти решения?
- Есть инструменты для текста, картинок, табличек
- Обзор инструментов на последнем Датафесте
- Для первого знакомства с темой гуглить
  - SHAP https://github.com/slundberg/shap
  - Interpretml https://github.com/interpretml/interpret
  - Grad-CAM http://gradcam.cloudcv.org/
  - Alibi https://github.com/SeldonIO/alibi
  - Christoph Molnar https://christophm.github.io/interpretable-ml-book/

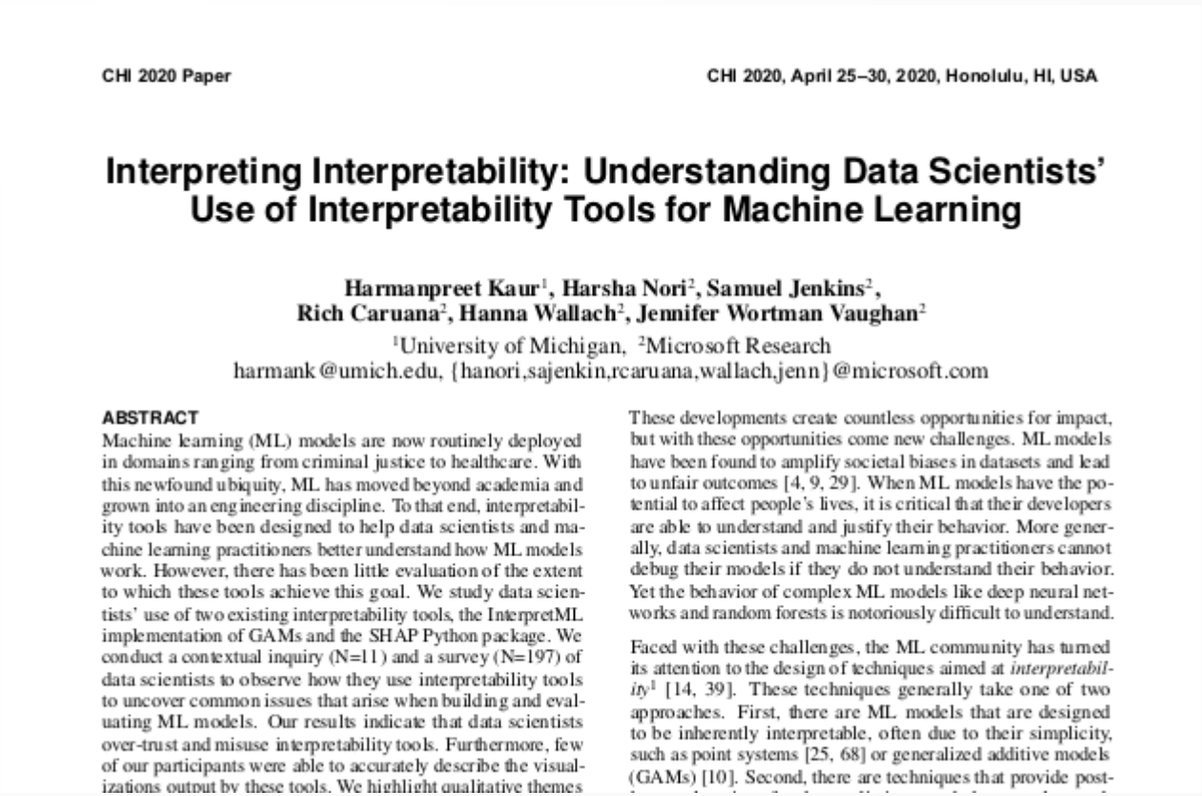
Объясните мне это

- Заказчики требуют
- Тимлиды требуют
- Регуляторы требуют
- Самим интересно:

  Как же она все-таки работает?

Лишь бы график красивый

- У нас был датасет
- Пачка графиков
- 11 датасатанистов
- SHAP
- GAM
- …



Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning

Harmanpreet Kaur[1], **Harsha Nori**[2], **Samuel Jenkins**[2], **Rich Caruana**[2], **Hanna Wallach**[2], **Jennifer Wortman Vaughan**[2]
[1]University of Michigan, [2]Microsoft Research
harmank@umich.edu, {hanori,sajenkin,rcaruana,wallach,jenn}@microsoft.com

ABSTRACT
Machine learning (ML) models are now routinely deployed in domains ranging from criminal justice to healthcare. With this newfound ubiquity, ML has moved beyond academia and grown into an engineering discipline. To that end, interpretability tools have been designed to help data scientists and machine learning practitioners better understand how ML models work. However, there has been little evaluation of the extent to which these tools achieve this goal. We study data scientists' use of two existing interpretability tools, the InterpretML implementation of GAMs and the SHAP Python package. We conduct a contextual inquiry (N=11) and a survey (N=197) of data scientists to observe how they use interpretability tools to uncover common issues that arise when building and evaluating ML models. Our results indicate that data scientists over-trust and misuse interpretability tools. Furthermore, few of our participants were able to accurately describe the visualizations output by these tools. We highlight qualitative themes

These developments create countless opportunities for impact, but with these opportunities come new challenges. ML models have been found to amplify societal biases in datasets and lead to unfair outcomes [4, 9, 29]. When ML models have the potential to affect people's lives, it is critical that their developers are able to understand and justify their behavior. More generally, data scientists and machine learning practitioners cannot debug their models if they do not understand their behavior. Yet the behavior of complex ML models like deep neural networks and random forests is notoriously difficult to understand.

Faced with these challenges, the ML community has turned its attention to the design of techniques aimed at *interpretability*[1] [14, 39]. These techniques generally take one of two approaches. First, there are ML models that are designed to be inherently interpretable, often due to their simplicity, such as point systems [25, 68] or generalized additive models (GAMs) [10]. Second, there are techniques that provide post-

http://www.jennwv.com/papers/interp-ds.pdf

Все очень быстро

https://arxiv.org/pdf/2002.11097.pdf

Problems with Shapley-value-based explanations as feature importance measures

https://papers.nips.cc/paper/2020/file/0d770c496aa3da6d2c3f2bd19e7b9d6b-Paper.pdf

Asymmetric Shapley values

https://papers.nips.cc/paper/2020/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf

Causal Shapley Values

Кстати, огонь

Кирилл Быков
обзор со свежего
NeurIPS 2020
горячо рекомендую.

@TUBerlin_UMI

https://umilab.medium.com/neurips-2020-review-on-explainable-ai-xai-5de76381bf43

Глаза разбегаются

- Causal Interpretability
- Trust scores
- Uncertainty estimation
- XAI for AutoML
- ...

Дед Мороз, давай уже

- Соревнование по интерпретации ML-моделей
- Онлайн-курс по XAI
- Удачи
- Здоровья
- Всего хорошего!

С наступающим Новым Годом!

Слайды тут

dkolodezev

promsoft

dkolodezev

d_key

dmitry_kolodezev

https://kolodezev.ru/download/dataelka2020.pdf