Interpretacja predykcji metod ML i DM
Uwagi wstępne

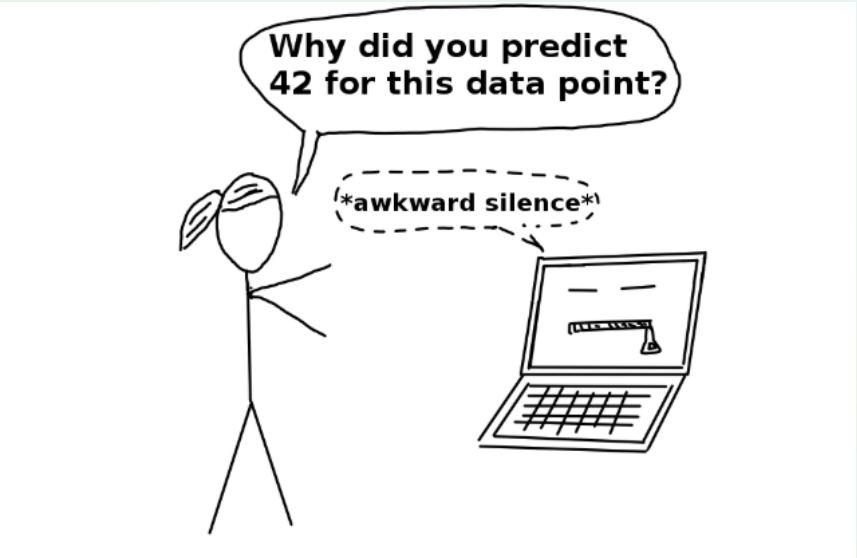
Jerzy Stefanowski

Institute of Computing Sciences,
Poznań University of Technology
Poland

Wersja 2020

Motywacje

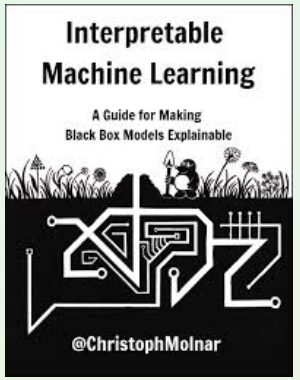
- Cytat [Molnar] – If a machine learning model performs (predicts) well, why do not we just trust the model?
- Why do we express doubts to prediction accuracy and cross validation estimation?
- Spróbuj odpowiedzieć



Simple ML or Data Scientists

Motywacje – ograniczenia automatycznej predykcji

- Predykcja - to to jedna ze możliwych perspektyw ML, w rzeczywistych zastosowaniach inteligentnych systemów oczekujemy także innych perspektyw
- Czy ludzie ufają tzw. black box models?
- Cytaty:
  - If a machine learning model performs well, why do not we just trust the model and ignore why it made a certain decision? “The problem is that a single metric, such as classification accuracy, is an incomplete description of most real-world tasks.” (Doshi-Velez and Kim 2017)
  - Do you just want to know what is predicted? For example, the probability that a customer will churn or how effective some drug will be for a patient. Or do you want to know why the prediction was made and possibly pay for the interpretability with a drop in predictive performance? (Molnar book)



Motywacje – ograniczenia automatycznej predykcji

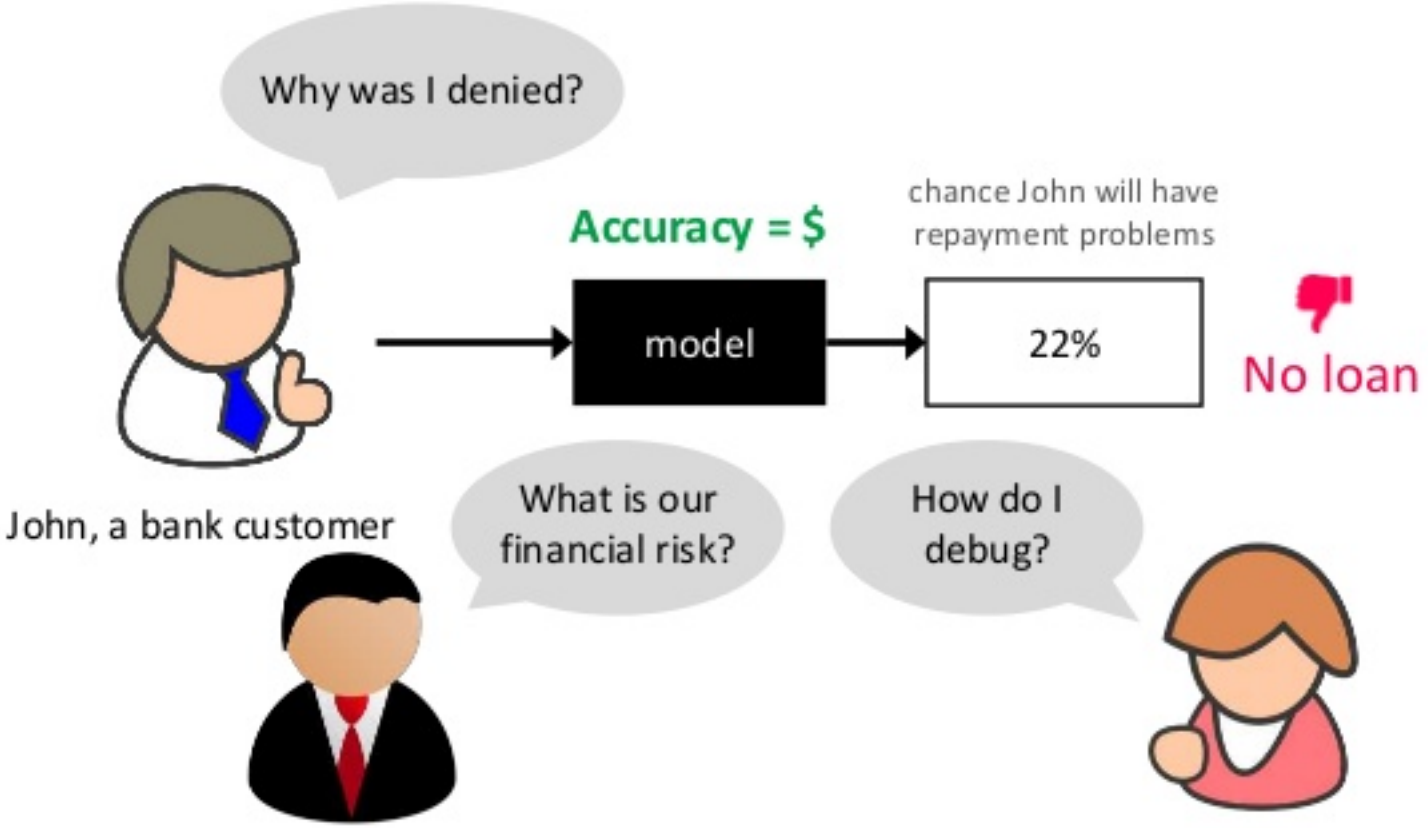
Ponadto

- Inne miary określają inne perspektyw
  - “a correct prediction only partially solves your original problem”.
- Zamknięty vs. otwarty świat – czy dysponujemy wystarczającymi danymi oraz przygotowaliśmy dogłębne testy?
- Jak radzić sobie z krytycznymi błędami systemu oraz zaskakującymi nietypowymi sytuacjami.
- Uczenie nadzorowane jednoetapowej klasyfikacji jest ograniczone w stosunku do rzeczywistych złożonych procesów decyzyjnych.
- Mała odporność na tzw. ataki zewn. „hacking and adversarial attacks”

Potrzeby rzeczywistych zastosowań

- Interpretowalność modeli uczenia maszynowego jest niezbędna dla pozyskania zaufania ludzi wobec takich systemów, zwłaszcza jeśli automatyczne podjęte decyzje są zaskakujące, nawet dla ekspertów [Stefanowski, Woźniak]

- Samek, Muller zauważają potrzeby w stosowaniu systemów predykcyjnych:
  - Weryfikowalność poprawności działania systemu predykcyjnego [przykład wykrycia niedoskonałości danych w problemie diagnozy zapalenia płuc].
  - Zrozumienie działania i poprawa modelu
  - Uczenie się od systemów sztucznej inteligencji
  - Zgodność z prawodawstwem i postulatami regulacyjnymi



Za wykład Scott Lundberg'a h20 world nyc 2019

Potrzeby interpretacji modeli predykcyjnych

Inni:

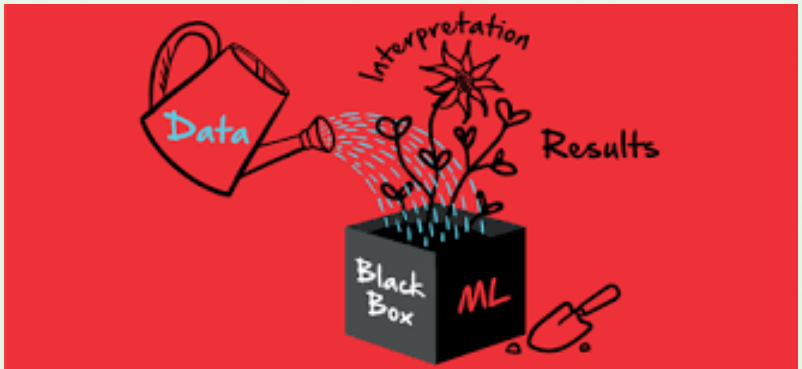
- W systemach wysokiego ryzyka, pomyłki mają krytyczne znaczenie
- Naturalne dla ciekawości człowieka
- Niezbędne w naukowym odkrywaniu oraz zrozumieniu świata (poszukuje się praw)
  - Badacze powinni zrozumieć ...
- Konieczne w niektórych obszarach zastosowań – prawo, administracja, decyzje finansowe, medycyna, itp.
- -> lecz nie w każdej dziedzinie!
- Ważne dla pozyskania tzw. społecznego zaufania wobec zbyt złożonych systemów AI i Robotyki
- Mocno wspomaga “debugging tool for detecting bias in machine learning models. ”

Specyfika b. trafnych modeli “black box”

Za “Why Should I Trust You?” Explaining the Predictions of Any Classifier (KDD2016) by Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, the interpretability for a black-box model is required.

Ponadto postulują:

- **Trusting a prediction**: a user will trust an individual prediction to act upon. No user wants to accept a model prediction on blind faith, especially if the consequences can be catastrophic.
- **Trusting a model:** the user gains enough trust that the model will behave in reasonable ways when deployed. Although in the modeling stage accuracy metrics are used on multiple validation datasets to mimic the real-world data, there often exist significant differences in the real-world data. Besides using the accuracy metrics, we need to test the individual prediction explanations

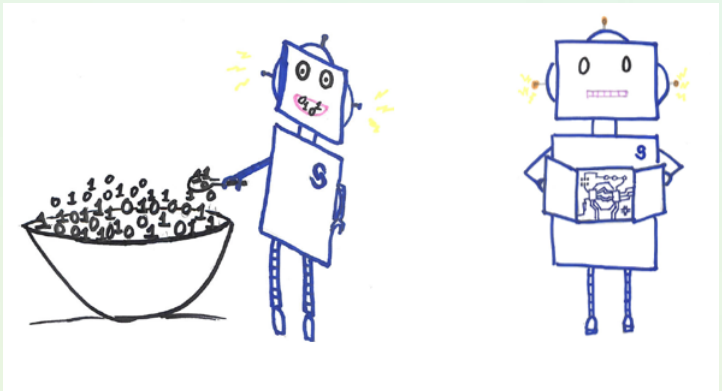


Explainable AI - XAI

- **Fairness**: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups.
- **Privacy**: Ensuring that sensitive information in the data is protected.
- **Reliability or Robustness**: Ensuring that small changes in the input do not lead to large changes in the prediction.
- **Causality**: Check that only causal relationships are picked up.
- **Trust**: It is easier for humans to trust a system that explains its decisions compared to a black box.

Ponadto – postulaty dokumentów EU tzw. Trustworthy Intelligence

Potrzeby przemysłowe -> Interpretability is required for regulated industry to adopt machine learning




Human agency and Oversight
Technical robustness and Safety
Privacy and Data Governance
Transparency
Diversity, Non-Discrimination and Fairness
Societal and Environmental wellbeing
Accountability
To be continuously evaluated and addressed throughout the AI system's life cycle


The European Commission's
HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE
AI
DRAFT
ETHICS GUIDELINES FOR TRUSTWORTHY AI

Interpretowalność ML

Próby definicji:

- What is Machine Learning Interpretability? “The ability to explain or to present in understandable terms to a human.”
  - Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning.” arXiv preprint. 2017. Complex models – although accurate are often not explainable black boxes

Interpretowalność predykcji, tzn. odpowiedzi na pytanie dlaczego pojęta taką, a nie inna decyzję, **interpretowalności modelu**, tzn. jak wygląda typowy reprezentant danej kategorii oraz omawianej przed chwilą **interpretowalności danych**, tzn. jakie warunki muszą spełniać dane wejściowe aby była możliwość interpretacji [stefanowski, woźniak 2019]

Elementy terminologiczne ad. interpretowalność

Pojęcie zrozumienia (ang. model comprehension) przez człowieka jest najważniejszych aspektów dobrej interpretowalności, lecz może zarówno dotyczyć zrozumienia procesu zbudowania modelu, jego ewentualnej reprezentacji, jak również zasad jego działania.

Wymienia się także inne dodatkowe pojęcia

- przejrzystość i czytelność,
- wierność odwzorowania zależności,
- dopasowanie do zdolności poznawczych odbiorcy,
- zaufanie do modelu,
- zdolność udzielania wyjaśnień

Zdolność udzielania wyjaśnień (ang. explanability)

System powinien móc udzielić wyjaśnień, dlaczego rekomenduje określoną decyzję.

- Rozróżnia się poziomy:
  - Funkcjonalnej interpretacji (jak i dlaczego model działa) i
  - Nisko-poziomowe algorytmiczne zrozumienie detali algorytmu, wpływu parametrów itd.
- Należy „powiązać atrybuty opisujące przykłady z wyjściem systemu w prosty, informatywny oraz posiadający znaczenie sposób”.
- Użytkownicy złożonych systemów muszą mieć możliwość odtworzenia całego procesu podejmowania decyzji

Najpopularniejsze podejścia

- Feature summary statistics, evaluating their role, impact
- Visualization (partial dependency plots, heat maps for ANN, …)
- Model internal parameters (easier for typical interpretable models)
- Focus on some data elements (interpreting some data points / example based explanations)
- Use (or transform into) Intrinsically interpretable model: Needs to explain models / their predictions and relation to data

Inne podziały:

Model-specific vs. model agnostic solutions

Globalne vs. lokalne

Uwagi nt podziału metod

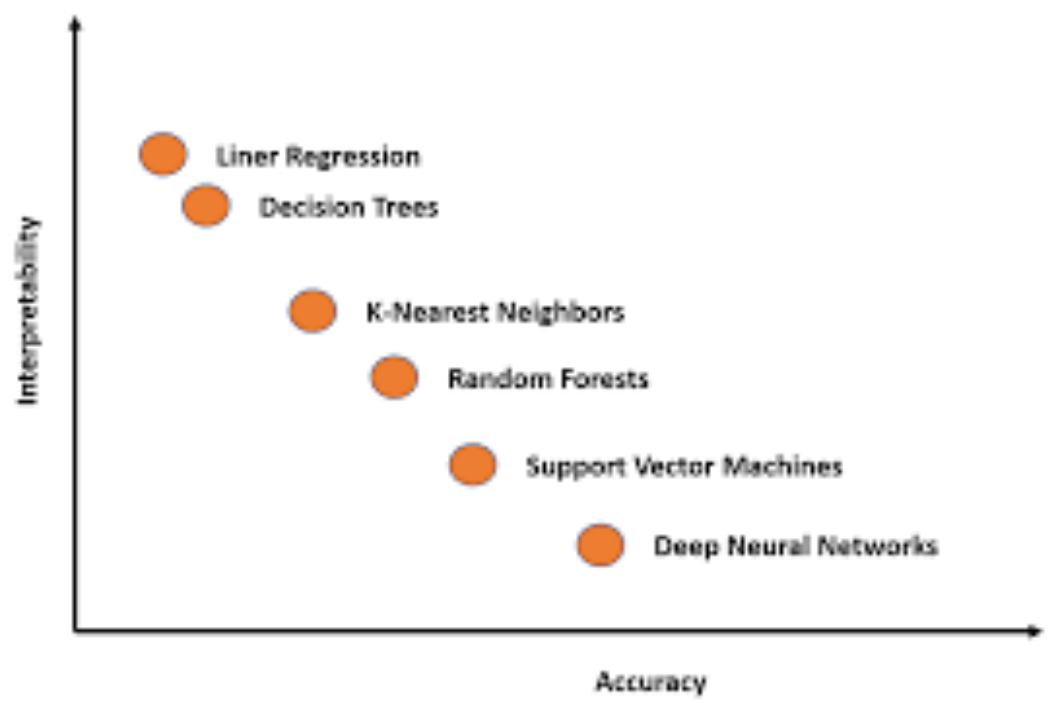
Holzinger rozróżnia dwa rodzaje wyjaśnialności:

- post-hoc
- oraz ante-hoc systemy

Post-hoc /agnostic models/ poszukuje się zastępczych modeli (surrogate) wspierających lokalne wyjaśnienia działania głównego modelu.

Ante-hoc - odnoszą się do przejrzystych reprezentacji wiedzy i starają się wprowadzić do procesu ich uczenia kryteria uwzględniające interakcje z człowiekiem i jego oczekiwania do łatwiejszego zrozumienia.

Typical interpretable models



Rys. za MC.AI

Typowe modele interpretowalne:

Linear models,

Regression, logistic regression, Generalized linear models

Decision trees

Decision rules,

Naive Bayes classifiers

K-nearest neighbors

Poczytaj więcej

Książki:

- Christoph Molnar, “Interpretable Machine Learning: A Guide for making black box models explainable”
- Przemysław Biecek, Explanatory Model Analysis (book under preparation)

Artykuły:

- Bibal A., Frenay B.: Interpretability of machine learning models and representations: an introduction. W: Proceedings of ESANN 2016
- Bratko I.: Machine learning: between accuracy and interpretability.
- Freitas A.: Comprehensible classification models: a position paper. ACM SIGKDD Exploration Newsletter, Vol. 15, Nr 1, 2014
- Guidotti R, Monreale A., Ruggieri S, Turini F., Giannotti F, Pedreschi D.: A Survey of Methods for Explaining Black Box Models, ACM Comput. Surv., 2018
- Ribeiro M. T., Singh S., Guestrin C.: Why should i trust you? Explaining the predictions of any classifier, W: Proc. of the 22nd ACM SIGKDD 2015
- Samek, W., Wiegand, T., Müller, K. R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296, 2017.
- Stefanowski J. Woźniak M.: Interpretacja modeli uczonych ze złożonych danych medycznych. W Informatyka w medycynie 2019.

Pytania lub komentarze?

Chwilowa przerwa techn.

Kontakt via:
Jerzy.Stefanowski@cs.put.poznan.pl

or www.cs.put.poznan.pl/jstefanowski