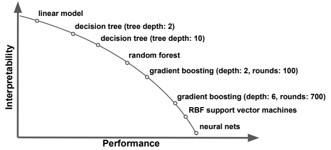
Interpretable Machine Learning

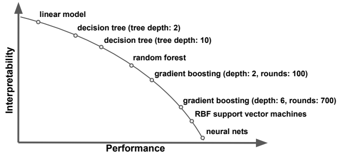
Introduction, Motivation, and History



Learning goals

- Why interpretability?
- Developments until now?
- Use cases for interpretability

Interpretable Machine Learning
Introduction, Motivation, and History

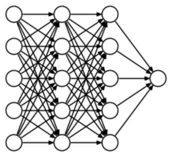


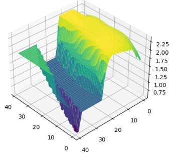
Learning goals

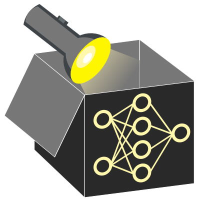
- Why interpretability?
- Developments until now?
- Use cases for interpretability

WHY INTERPRETABILITY?

- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are black boxes, e.g., XGBoost, RBF SVM or DNNs
  ⇝ too complex to be understood by humans
- Some applications are "learn to understand"

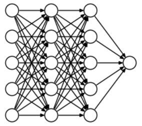


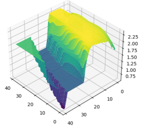


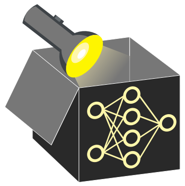


WHY INTERPRETABILITY?

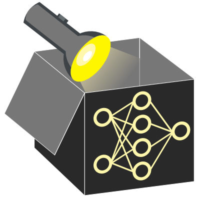
- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are black boxes, e.g., XGBoost, RBF SVM or DNNs
  ⇝ too complex to be understood by humans
- Some applications are "learn to understand"



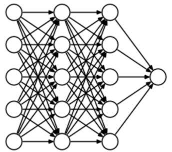


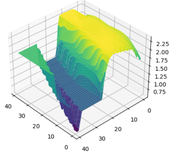


WHY INTERPRETABILITY?

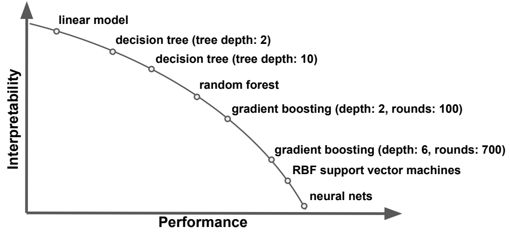


- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are black boxes, e.g., XGBoost, RBF SVM or DNNs
  ⇝ too complex to be understood by humans
- Some applications are "learn to understand"
- When deploying ML models, lack of explanations
  1. hurts trust
  2. creates barriers





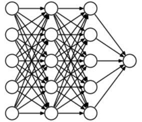
⇝ Many disciplines with required trust rely on traditional models, e.g., linear models, with less predictive performance

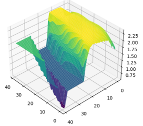


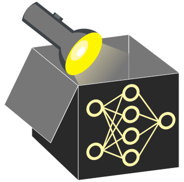
WHY INTERPRETABILITY?

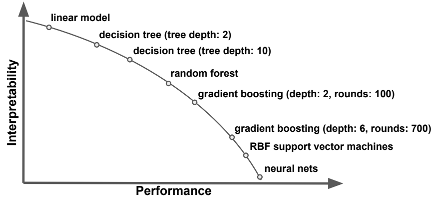
- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are black boxes, e.g., XGBoost, RBF SVM or DNNs
  ⇝ too complex to be understood by humans
- Some applications are "learn to understand"
- When deploying ML models, lack of explanations
  1. hurts trust
  2. creates barriers

⇝ Many disciplines with required trust rely on traditional models, e.g., linear models, with less predictive performance

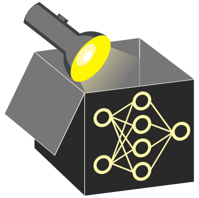






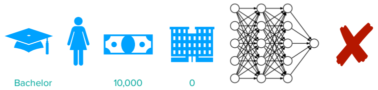


INTERPRETABILITY IN HIGH-STAKES DECISIONS

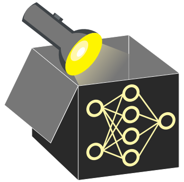


Examples of critical areas where decisions based on ML models can affect human life

- Credit scoring and insurance applications
  ▸ Society of Actuaries
  - Reasons for not granting a loan
  - Fraud detection in insurance claims

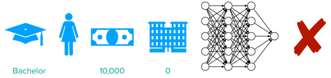


INTERPRETABILITY IN HIGH-STAKES DECISIONS

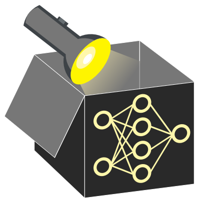


Examples of critical areas where decisions based on ML models can affect human life

- Credit scoring and insurance applications
  ▸ Click for source
  - Reasons for not granting a loan
  - Fraud detection in insurance claims

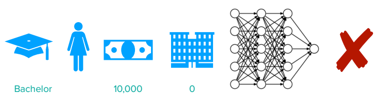


INTERPRETABILITY IN HIGH-STAKES DECISIONS

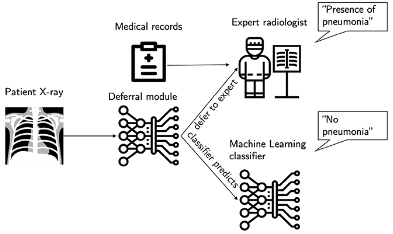


Examples of critical areas where decisions based on ML models can affect human life

- Credit scoring and insurance applications
  ▸ Society of Actuaries
  - Reasons for not granting a loan
  - Fraud detection in insurance claims

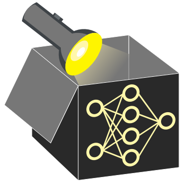


- Medical applications
  - Identification of diseases
  - Recommendations of treatments
- . . .



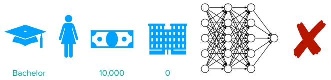
▸ Miliard (2020)

INTERPRETABILITY IN HIGH-STAKES DECISIONS

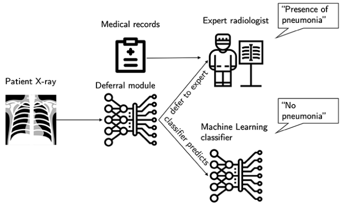


Examples of critical areas where decisions based on ML models can affect human life

- Credit scoring and insurance applications ▸ Click for source
  - Reasons for not granting a loan
  - Fraud detection in insurance claims

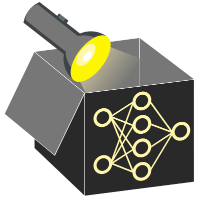


- Medical applications
  - Identification of diseases
  - Recommendations of treatments
- . . .



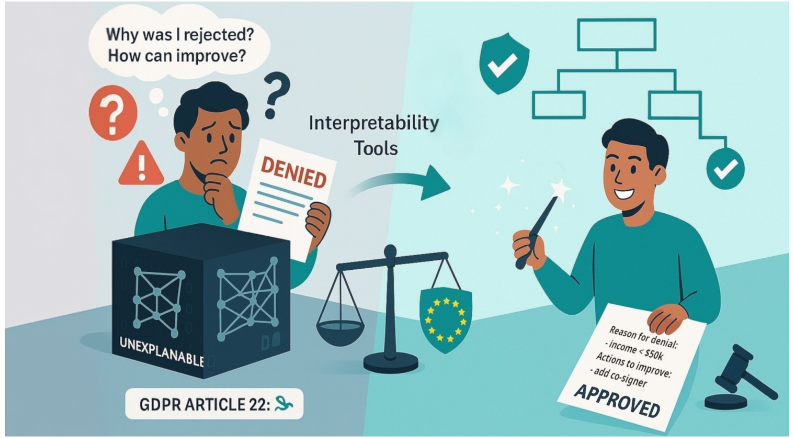
Miliard (2020) ▸ Click for source

NEED FOR INTERPRETABILITY

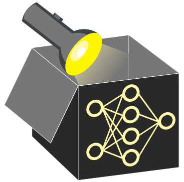


Need for interpretability becoming increasingly important from a legal perspective

- General Data Protection Regulation (GDPR) requires for some applications that models have to be explainable ▸ Goodman & Flaxman (2017)
  ⇝ *EU Regulations on Algorithmic Decision-Making and a "Right to Explanation"*
- *Ethics guidelines for trustworthy AI* ▸ European Commission (2019)

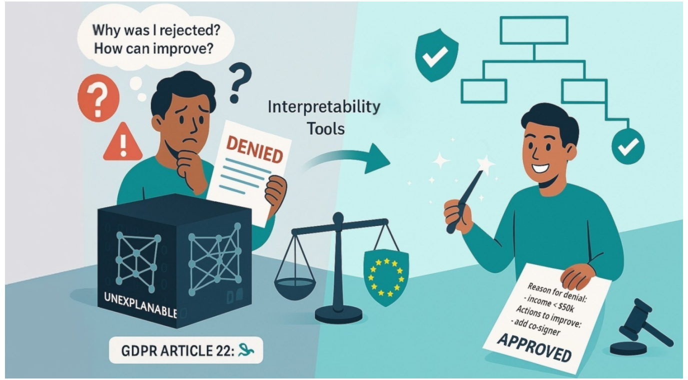


NEED FOR INTERPRETABILITY

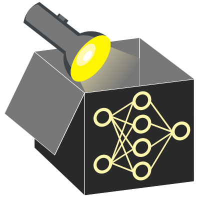


Need for interpretability becoming increasingly important from a legal perspective

- General Data Protection Regulation (GDPR) requires for some applications that models have to be explainable ▸ Flaxman 2017
  ⇝ *EU Regulations on Algorithmic Decision-Making and a "Right to Explanation"*
- *Ethics guidelines for trustworthy AI* ▸ "European Commission" 2019



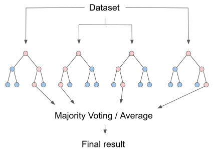
BRIEF HISTORY OF INTERPRETABILITY



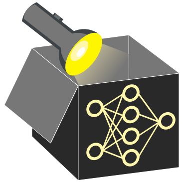
- 18th and 19th century:
  Linear regression models (Gauss, Legendre, Quetelet)
- 1940s:
  Emergence of sensitivity analysis (SA)
- Middle of 20th century:
  Rule-based ML, incl. decision rules and decision trees
- 2001:
  Built-in feature importance measure of random forests
- >2010:
  Explainable AI (XAI) for deep learning
- >2015:
  IML as an independent field of research

▸ Carl Friedrich Gauss

▸ Wikipedia



BRIEF HISTORY OF INTERPRETABILITY



- 18th and 19th century:
  Lin. regression models (Gauss, Legendre, Quetelet)
- 1940s:
  Emergence of sensitivity analysis (SA)
- Middle of 20th century:
  Rule-based ML, incl. decision rules and trees
- 2001:
  Built-in feature imp. measure of random forests
- >2010:
  Explainable AI (XAI) for deep learning
- >2015:
  IML as an independent field of research

Carl Friedrich Gauss

▸ Click for source

Wikipedia

▸ Click for source