Interpretable Machine Learning

Introduction to Interpretable Machine Learning

Learning goals

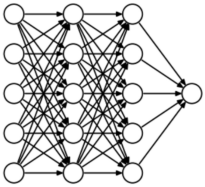
- Introduction and Motivation
- Dimensions of Interpretability
- Bike Sharing Dataset

WHY INTERPRETABILITY?

- ML: huge potential to aid decision-making process due to its predictive performance
- ML models are often black boxes, e.g., XGBoost, RBF SVM or DNNs
  - ⇝ too complex to be understood by humans
- Lack of explanation
  1. hurts trust
  2. creates barriers

⇝ Harder to adapt for critical areas with decisions affecting human life

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

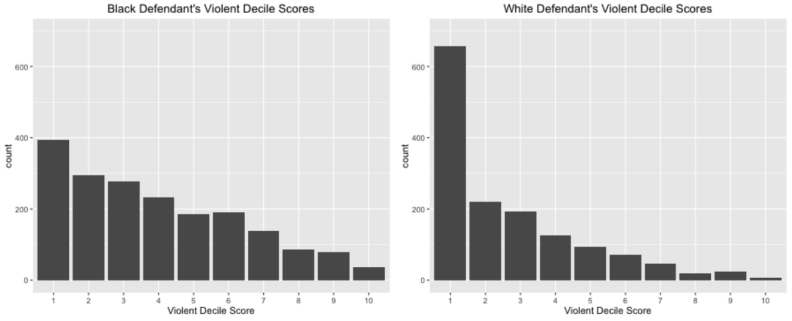


▸ Liu 2021

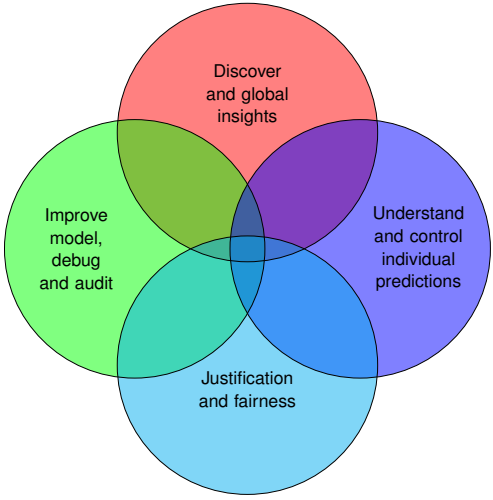
NEED FOR INTERPRETABILITY IN HIGH-STAKES DECISIONS

Need for interpretability also 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)
  $\rightsquigarrow$ *EU Regulations on Algorithmic Decision-Making and a "Right to Explanation"*
- *Ethics guidelines for trustworthy AI* ▸ European Commission (2019)



WHEN DO WE NEED INTERPRETABILITY

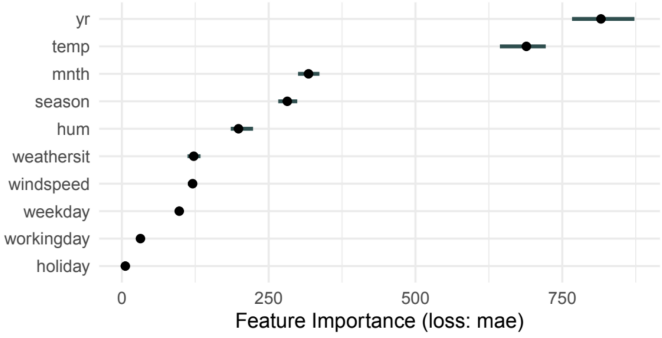


A related presentation can be found in ▸ Adadi and Berrada 2018.

DISCOVER AND GLOBAL INSIGHTS

$\rightsquigarrow$ Gain insights about data, distribution and model

Example: Bike Sharing Dataset (predict number of bike rentals per day)
Exemplary question: Which feature influences the model performance and to what extent?



- Year (`yr`) and Temperature (temp) most important features
- Holiday (`holiday`) less important (Can we drop it?)

IMPROVE MODEL, DEBUG AND AUDIT

$\rightsquigarrow$ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

A cautionary tale (which never actually happened):

- Creation of neural network to detect tanks
- Model shows good predictive performance in training data set
- Application outside training data set: failure

IMPROVE MODEL, DEBUG AND AUDIT

$\rightsquigarrow$ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

A cautionary tale (which never actually happened):

- Creation of neural network to detect tanks
- Model shows good predictive performance in training data set
- Application outside training data set: failure
- Reasons vary depending on the source, in general: NN based its decision on irrelevant points.

IMPROVE MODEL, DEBUG AND AUDIT

$\rightsquigarrow$ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

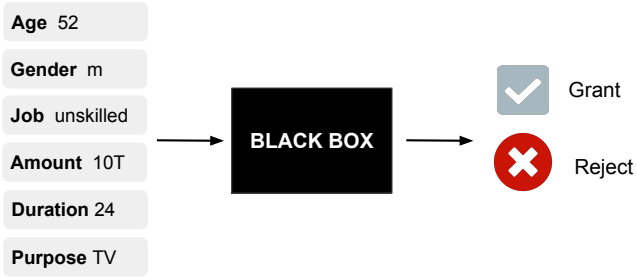
A cautionary tale (which never actually happened):

- Creation of neural network to detect tanks
- Model shows good predictive performance in training data set
- Application outside training data set: failure
- Reasons vary depending on the source, in general: NN based its decision on irrelevant points.
- E.g. model detecting weather situations: Tanks always photographed under cloudy skies; photos without tanks always taken in sunny weather.

UNDERSTAND AND CONTROL INDIVIDUAL PREDICTIONS

$\rightsquigarrow$ Explaining individual decisions can prevent unwanted actions based on the model

Example: Credit Risk Application. **x**: customer and credit information; y: grant or reject credit



Questions:

- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**

JUSTIFICATION AND FAIRNESS

$\rightsquigarrow$ Investigate if and why biased, unexpected or discriminatory predictions were made

Example: COMPAS

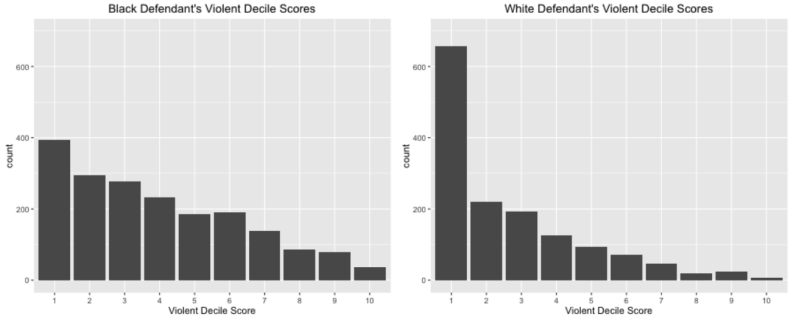
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
- Commercial algorithm used by judges to assess defendant's likelihood of re-offending
- Predict recidivism risk
  - i.e., criminal re-offense after previous crime, resulting in jail booking
  - different risk levels: high risk, medium risk or low risk
- Evaluation of recidivism risk based on a questionnaire the defendant has to answer

JUSTIFICATION AND FAIRNESS: COMPAS

▸ Larson et al. 2016

$\rightsquigarrow$ Investigate if and why biased, unexpected or discriminatory predictions were made

Descriptive data analysis:



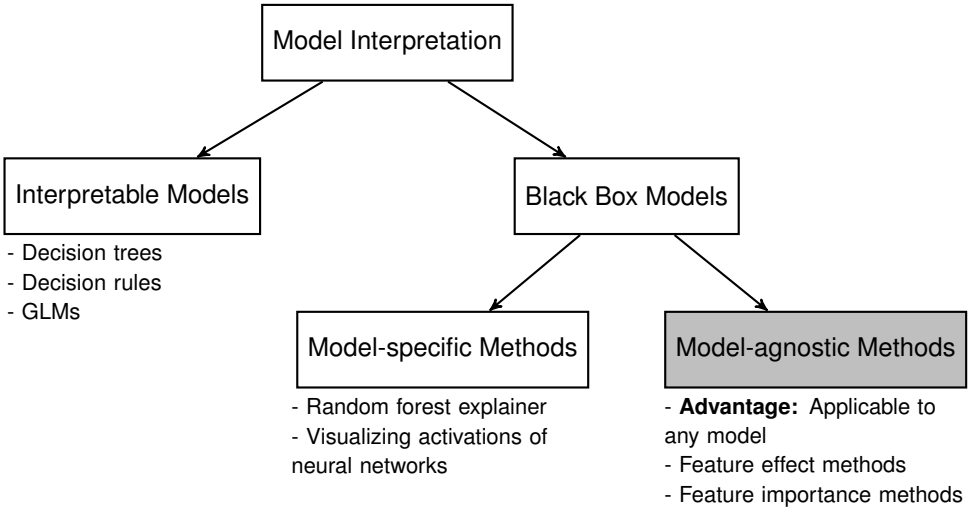
Decile score: 1 (low risk) to 10 (high risk)

$\rightsquigarrow$ Model skewed towards low risk for white defendants

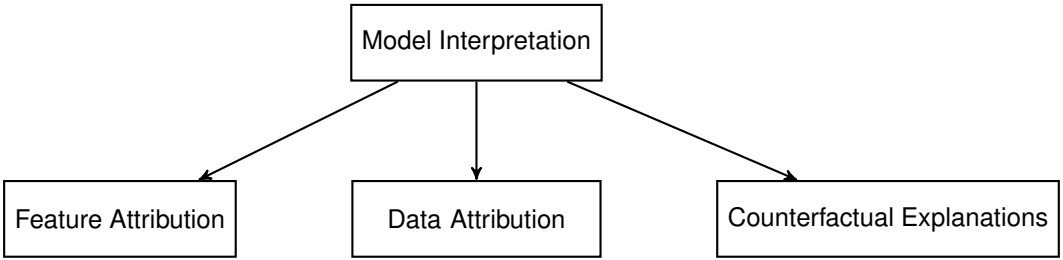
$\rightsquigarrow$ Strong indication that the model is discriminating black defendants

$\rightsquigarrow$ Use IML to investigate if and how much the model uses the defendants' origin.

INTRINSIC VS. MODEL-AGNOSTIC



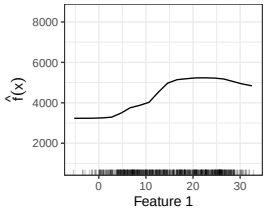
TYPES OF EXPLANATIONS



GLOBAL VS. LOCAL

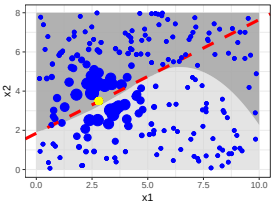
Global interpretation methods explain the model behavior for the entire input space by considering all available observations:

- Permutation Feature Importance (PFI)
- Partial Dependence (PD) plots
- Accumulated Local Effect (ALE) plots
- ...



Local interpretation methods explain the model behavior for single data instances:

- Individual Conditional Expectation (ICE) curves
- Local Interpretable Model-Agnostic Explanations (LIME)
- Shapley values, SHAP
- ...



BIKE SHARING DATASET ▸ Fanaee-T and Gama (2014)

- Daily counts of rented bikes in Washington D.C. from ▸ Capital-Bikeshare
- Feature description:
  - `cnt`: count of total rental bikes (prediction target for regression)
  - `season`: season (1: WINTER, 2: SPRING, 3: SUMMER, 4: FALL)
  - `yr`: year (0: 2011, 1: 2012)
  - `mnth`: month of year (1: JAN, ..., 12: DEC)
  - `holiday`: day is holiday (0: NO HOLIDAY, 1: HOLIDAY)
  - `weekday`: day of the week (1: SUN, 2: MON, ..., 7: SAT)
  - `workingday`: day is not a weekend or holiday (0: NO WORKING DAY, 1: WORKING DAY)
  - `weathersit`: weather situation (1: GOOD, 2: MISTY, 3: RAIN/SNOW/STORM)
  - `temp`: temperature in Celsius
  - `hum`: humidity in percent
  - `windspeed`: wind speed in km/h
  - `days_since_2011`: Number of days since January 1st, 2011 (start of historical log)
    $\rightsquigarrow$ accounts for the trend over time

BIKE SHARING DATASET - EDA