Interpretable Machine Learning

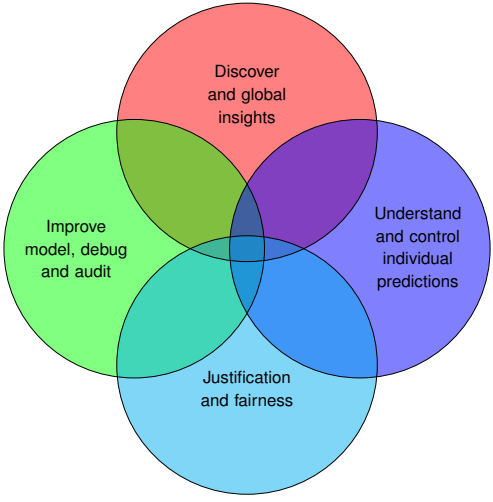
Interpretation Goals

Learning goals

Role of interpretation for...

- ... gaining insides about data
- ... improving model (debug and audit)
- ... understanding and controlling individual predictions
- ... justification and fairness

WHEN DO WE NEED INTERPRETABILITY

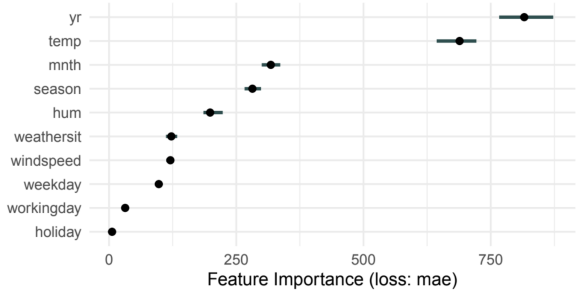


A related presentation can be found in ▸ Adadi and Berrada 2018.

DISCOVER AND GLOBAL INSIGHTS

⇝ 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

⇝ 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.

IMPROVE MODEL, DEBUG AND AUDIT

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

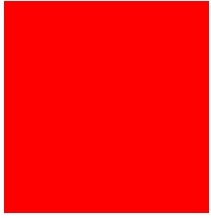
Comment on tank example:

"We made exactly the same mistake in one of my projects on insect recognition. We photographed 54 classes of insects. Specimens had been collected, identified, and placed in vials. Vials were placed in boxes sorted by class. I hired student workers to photograph the specimens. Naturally they did this one box at a time; hence, one class at a time. Photos were taken in alcohol. Bubbles would form in the alcohol. Different bubbles on different days. The learned classifier was surprisingly good. But a saliency map revealed that it was reading the bubble patterns and ignoring the specimens. I was so embarrassed that I had made the oldest mistake in the book (even if it was apocryphal). Unbelievable. Lesson: always randomize even if you don't know what you are controlling for!"

▸ Thomas G. Dietterich

DEBUG AND AUDIT

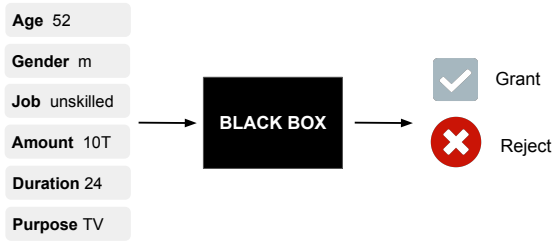
- At first instance nearly all computer programs (CPs) have bugs
  - ⇝ Minimizing bugs in CPs and systems is mandatory
- Process with multiple steps to locate, understand and solve a problem
  - ⇝ Classical debugging
- **In ML** we have a program (CP1) writing another program (CP2)
- Code of CP2 (the ML model) is not readable ⇝ How to debug the model?
  - ⇝ Investigate the data
  - ⇝ Simplify as far as possible
  - ⇝ Verify the mathematics
  - ⇝ Make the code more complex step by step



UNDERSTAND AND CONTROL INDIVIDUAL PREDICTIONS

⇝ 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:

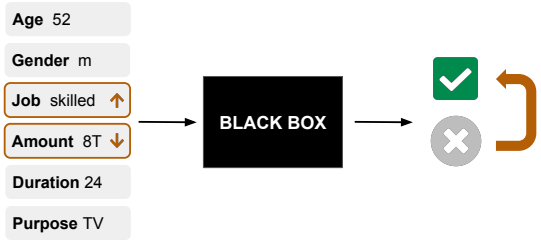
- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**

UNDERSTAND AND CONTROL INDIVIDUAL PREDICTIONS

⇝ 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

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



“If the person was more skilled and the credit amount had been reduced to $8.000, the credit would have been granted."

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

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

JUSTIFICATION AND FAIRNESS

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

Example: COMPAS

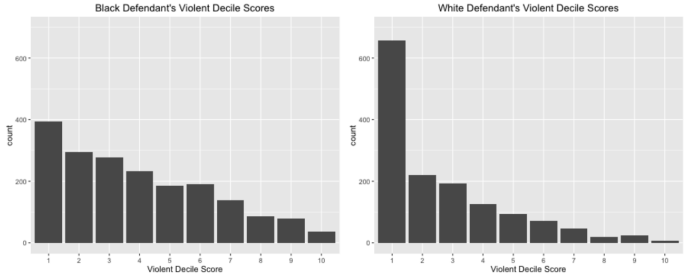
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)
- Commercial algorithm used by judges to assess defendant's likelihood of re-offending
- Predict recidivism risk
  - i.e., criminal re-offense after previous crime, resulting in jail booking
  - different risk levels: high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

⇝ Investigate if and why biased, unexpected or discriminatory predictions were made

Descriptive data analysis:



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

⇝ Model skewed towards low risk for white defendants

⇝ Strong indication that the model is discriminating black defendants

⇝ Use IML to investigate if and how much the model uses the defendants' origin.

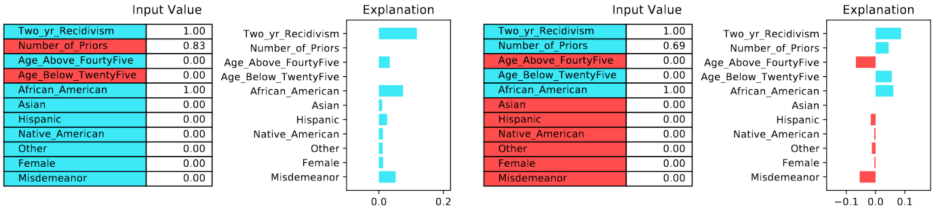
JUSTIFICATION AND FAIRNESS: COMPAS

▸ Alvarez-Melis and Jaakkola 2018

⇝ Investigate if and why biased, unexpected or discriminatory predictions were made

The underlying classifier is a logistic regression. Feature effects analysis for two exemplary defendants, using different interpretation methods (SHAP and LIME):

⇝ The methods give for every feature a number mirroring the impact on violence score.



⇝ In both cases the race (african american) has a noticeable positive impact on violent score