Benefits of using multiple (local) post-hoc explanation for Machine Learning

Corentin Boidot, Olivier Augereau, Pierre de Loor, Riwal Lefort

22nd International Conference on Machine Learning and Applications
IEEE December 15-17, 2023 AMLA
Hyatt Regency Jacksonville Riverfront, Florida, USA

Benefits of using multiple post-hoc (local) explanations for Machine Learning

Corentin Boidot, Olivier Augereau, Pierre De Loor, Riwal Lefort
ENIB / Lab-STICC & Crédit Mutuel Arkéa

Context

We pick e-sport winner prediction as a ***binary task on tabular data*** where human and AI could compete. We take a popular game, League of Legends, to recruit student participants with a form of **expertise*** on the task. The analogy with expert task like fraud detection is hindered by the low dimensionality of our setting (d=23 vs. ~80).

*No real form of expertise could be exhibited through this study. Expertise was evaluated only with a survey (discussed with an e-sport professional, but no significant correlation with empirical results could be found.

Explanations do not affect human accuracy on an e-sport prediction task, even with a multi-explanation system.

Human alone : 72±4% AI alone : 74±4%
H-AI team with XAI : 72±8%

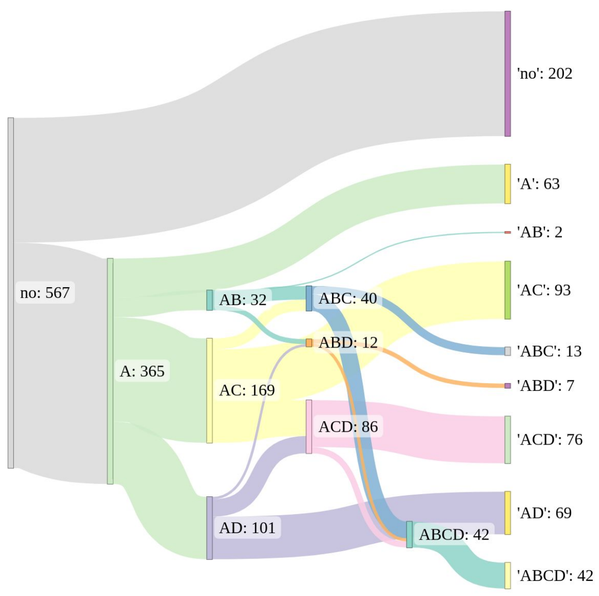
Asking for explanation is also correlated with slower decisions.

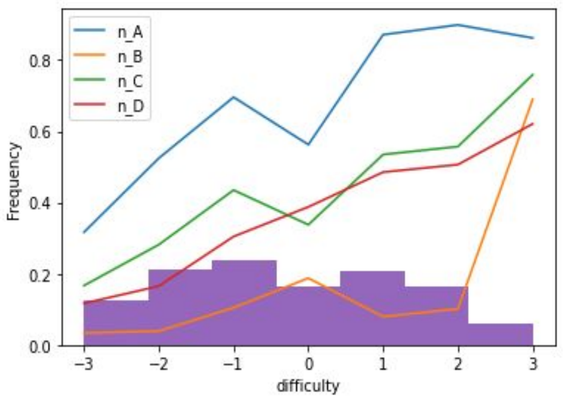
So what?

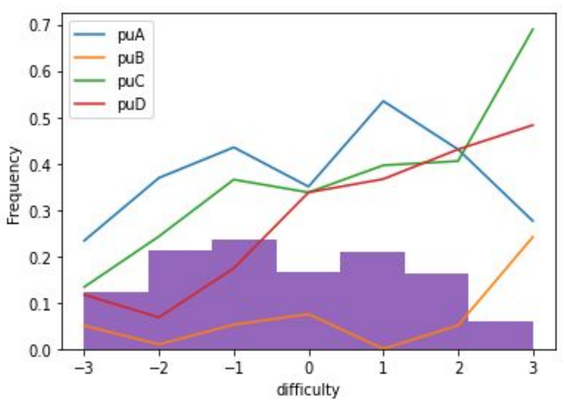
Explanation might still be valuable for users:

- Users mostly appreciated and used the XAI system:
  - Positive results in **perceived ease of use** and **perceived utility**.
- Results show the importance of **perceived case difficulty**:
  - Case difficulty is positively correlated with response time.
  - Case difficulty is positively correlated with the use of the explanations.
  - Case difficulty is positively correlated with perceived utility of explanations.
  - Case difficulty is correlated with the model's confidence toward its prediction
- Results suggest the importance of individual **preferences** towards the explanations:
  - These preferences may not be predictable.
- Decision diversity should be a metric:
  - As accuracy is linked with **human-AI agreement**, this agreement should **always** be measured prior to an XAI application-based evaluation, using human decisions without AI as reference.
  - Participants agreement with the AI was lesser when they had access to the XAI system than in control condition.

Additional figures



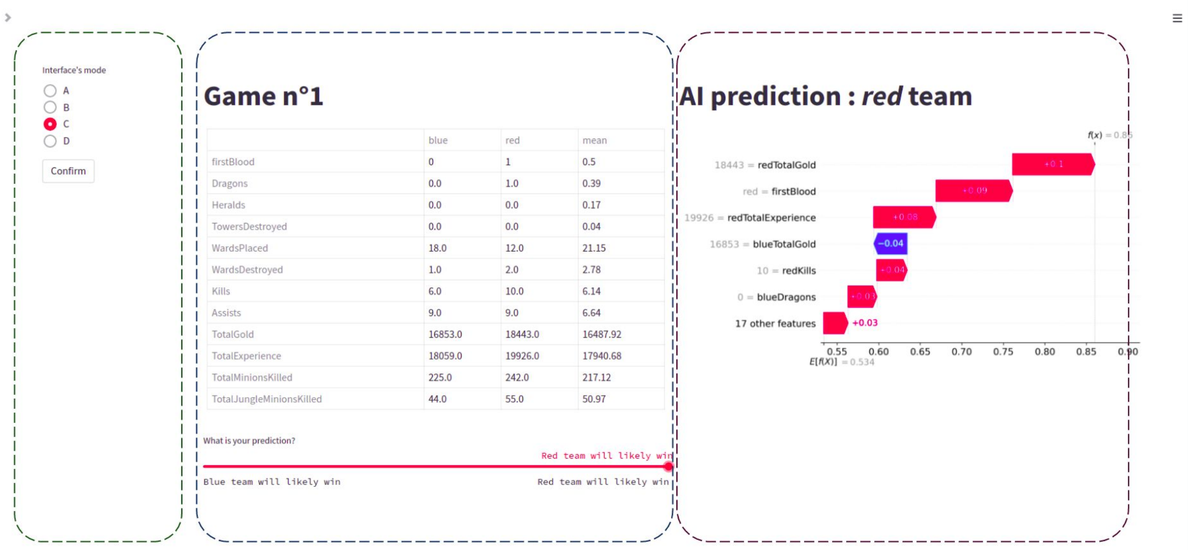




Our XAI system :

The ML model is a random forest. The decision making interface displays data first (II.). The user must click (I.) to gain access to AI prediction and score (III.). The 3 explanations (III.) are available through a radio button (I.).

Decision interface example (with explanation B: LIME)



I. II. III.

The explanation methods used:

A/ **Confidence score**, without further calibration.

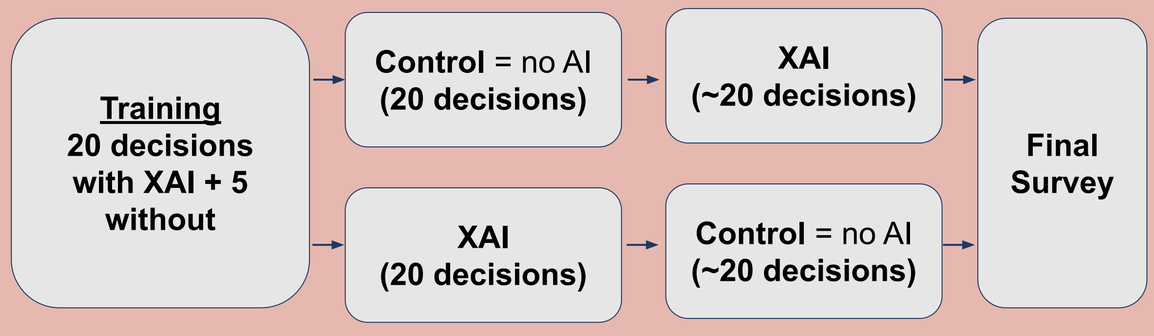
B/ **Surrogate Model**, using a simple **rule** model: skope-rules.

C/ **LIME**, computed with 7k data samples, displayed through ***shap*** library.

D/ **Nearest Neighbor**, with additional information (class, AI score, search for contrast).

Within-subject study
N=27 participants

~1h



Comparisons between explanations are done through an observational study inside XAI condition

Additional results

Objective and subjective metrics evaluation explanations

	Users	Use rate	Usefulness	Time spent	Interpretability
Confidence	24	76%	66%	3.23s	2.99
Surrogate rule	10	19%	44%	4.56s	-0.99
LIME	21	49%	88%	6.50s	1.38
Neighbor	19	42%	75%	7.07s	0.27