Comparing User Perception of Explanations Developed with XAI Methods

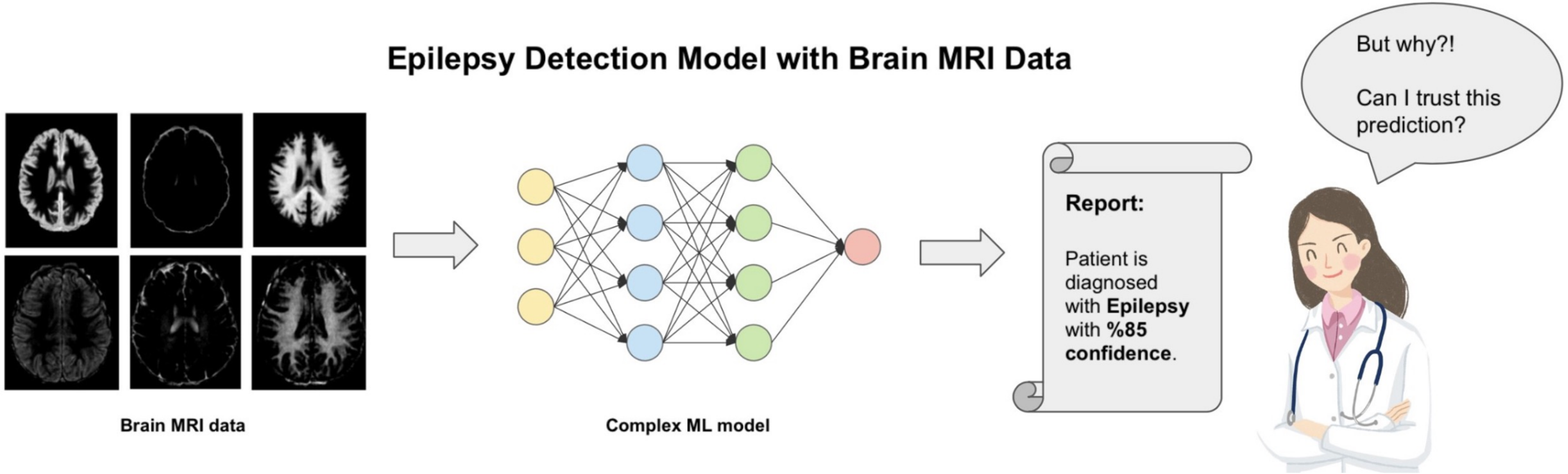
Jonathan Aechtner, Lena Cabrera,
Dennis Katwal, Pierre Onghena,
Diego Penroz Valenzuela and
Anna Wilbik

Machine Learning and Modelling Seminar – April 13

Maastricht University

Introduction

Understand how decisions are made by an AI,
why this decision?



Retrieved from [How XAI can help us trust AI]

Introduction

e**X**plainable **A**rtificial **I**ntelligence (XAI)

→ Methods which aim to be understandable for humans

How to determine the *best* XAI method?

→ User-centric evaluation

Research Questions

1. How do different XAI methods perform on a selection of *evaluation criteria*?
   Which is the *best* performing method?

2. Is there a preference towards local or global explanations for AI experts?

Research Hypotheses

A. AI novices prefer local over global explanations

B. Explanations increase users' trust in a system

Methodology

Design Decisions

Scope of the benchmark study and
the different **building blocks** of the designed questionnaire
were guided by six *design decisions* (D1 - D6)

D1

Target Audience

D2

Use Case

D3

XAI Methods

D4

Black Box Models

D5

Evaluation Criteria

D6

Questionnaire Design

Target Audience & Use Case (D1 & D2)

Target Audience

- Students with different backgrounds:
  - **AI novices**
  - **AI experts**

Use Case

- Admissions process of students for graduate schools

Use Case (D2)

~~University Learning Analytics Dataset~~"

"Graduate Admission" dataset:

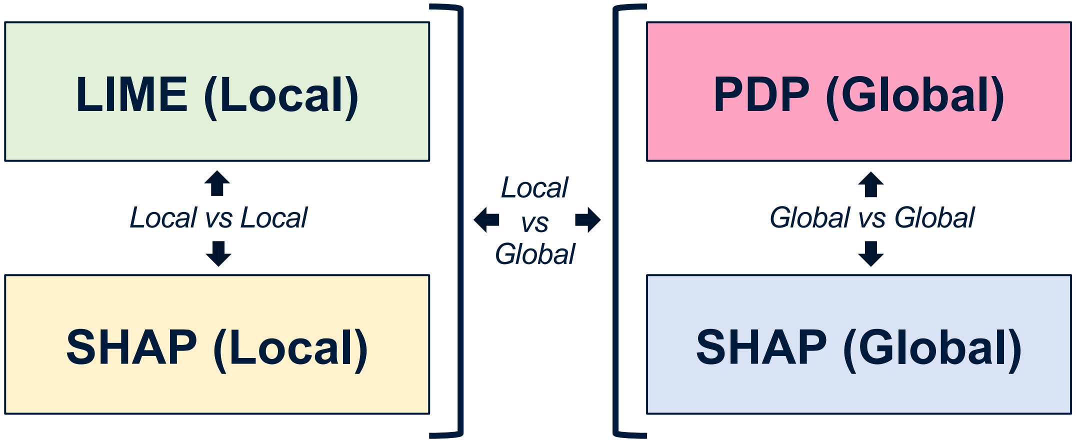
- GRE Scores (out of 340)
- TOEFL Scores (out of 120)
- University Rating (out of 5)
- Statement of Purpose and Letter of Recommendation Strength (out of 5)
- Undergraduate GPA (out of 10)
- Research Experience (either 0 or 1)
- Chance of Admit (ranging from 0 to 1)

- **Chance of Admit : True or False**

Accepted or Rejected?

XAI Methods (D3)

- **Model-agnostic**, post hoc XAI methods
- Inclusion of most popular ones: **LIME** and **SHAP**
- ***Local*** vs ***Global***



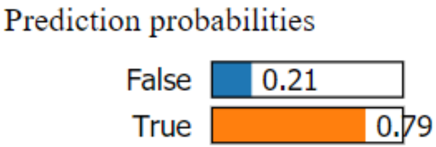
XAI Methods (D3)

General Information	Categorization							
XAI Method	**Specificity**		**Locality**		**Type of explanantion**			
Name	Model-agnostic	Model-specific	Local	Global	Simplification	Feature relevance	Local explanation	Visual Explanation
	can be applied to any machine learning model	*can only be applied to a specific group of models (if model-specific method,*	*explain predictions of a model by*	*explain how a model works*	*Approximate model using a simpler "proxy/surrogat*	*Quantify the influence of each input variable and rank them by*	*explain predictions of a model by investigating its performance on a*	*Generate visualizations to gain insights about e.g. decision boundary or*
LIME	x		x		x		x	
SHAP	x		x	x		x		
PDP	x			x				x

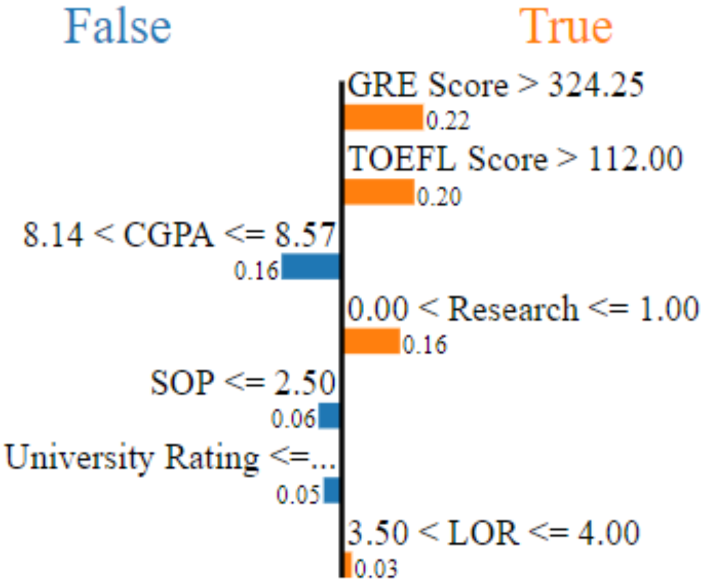
We are able to generate 4 different types of explanations.

XAI Methods (D3)

LIME (Local)



Feature	Value
GRE Score	329.00
TOEFL Score	114.00
CGPA	8.56
Research	1.00
SOP	2.00
University Rating	2.00
LOR	4.00

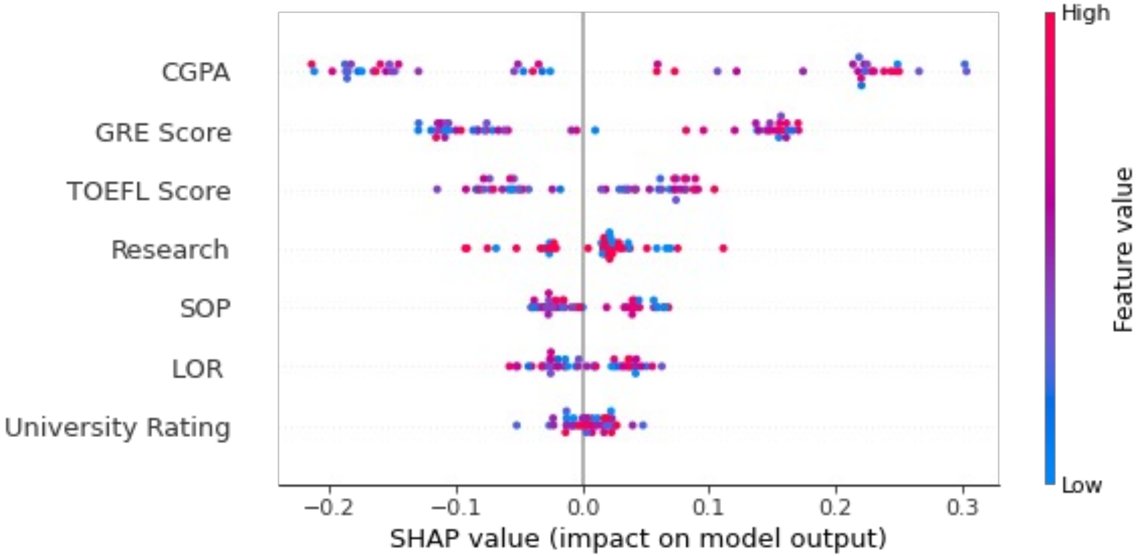


XAI Methods (D3)

SHAP (Local)

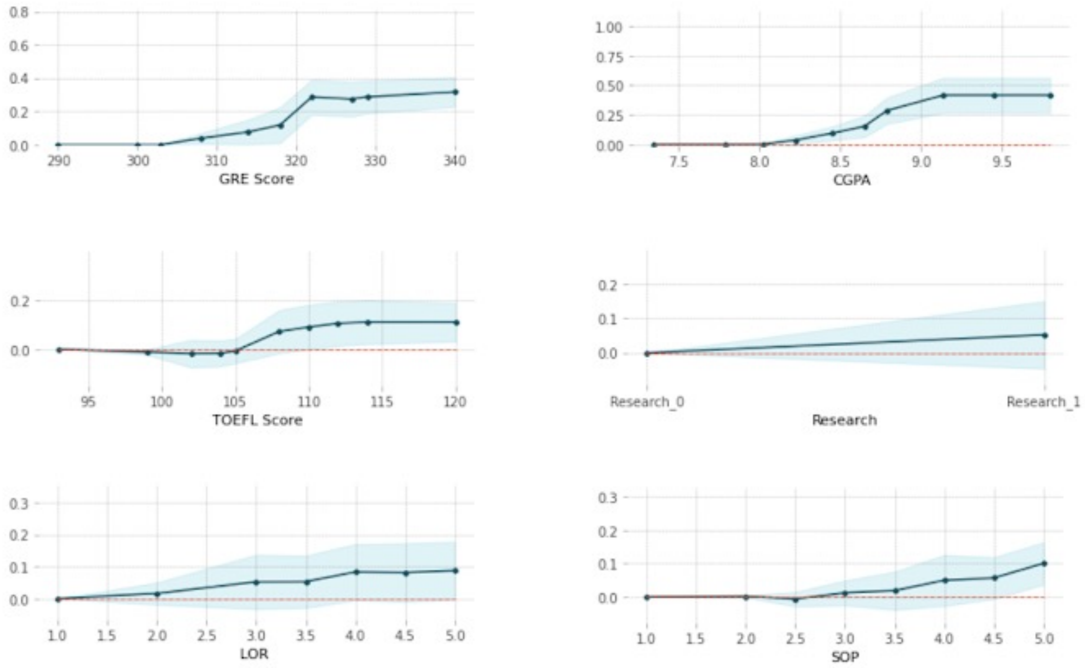
XAI Methods (D3)

SHAP (Global)



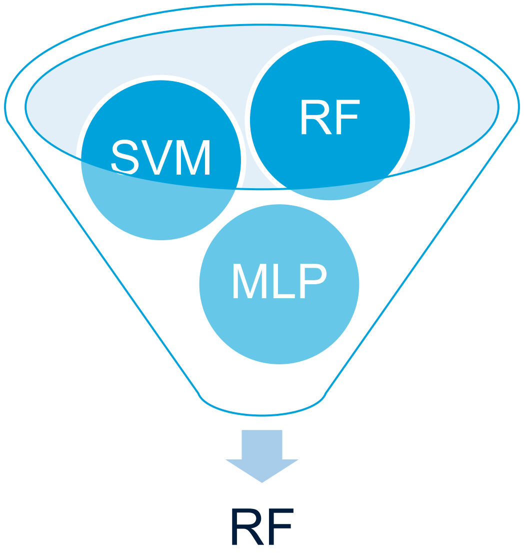
XAI Methods (D3)

PDP (Global)



Black Box Model (D4)

- 3 models implemented: SVM, RF, MLP
  - All performed equally well with similar explanations



Training and validation of Classification models

Random Forest Clf. *vs* Support Vector Machines Clf. *vs* Multi-layer Perceptron Clf.

Very similar performance on test data

```
Accuracy for RandomForestClassifier -> 86.6
and Confusion Matrix is
[[256  26]
 [ 40 178]]
TPR for RandomForestClassifier -> 0.8165137614678899
TNR for RandomForestClassifier -> 0.9078014184397163
```

Evaluation Criteria (D5)

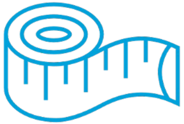
- **Understandability**
  From the explanation, does the user understand how the model makes a decision?

- **Usefulness**
  Is the explanation useful to the user, to make better decisions or to perform an action?

- **Trustworthiness**
  Does the explanation increase the user's trust in the model?

- **Informativeness**
  Does the explanation provide sufficient information to explain how the model makes decisions?

- **Satisfaction**
  Does the explanation of the model satisfy the user?



Questionnaire Design (D6)

Background Questions:
- Language
- Current Education level
- Field of Study
- Courses related to AI taken
- Familiarity with XAI

LIME (Local):
- Relatable Short Story (Context)
- Contextualized Explanation
- Evaluation Criteria (Likert Scale)

Local vs Local

SHAP (Local):
- Relatable Short Story (Context)
- Contextualized Explanation
- Evaluation Criteria (Likert Scale)

Local vs Global

PDP (Global):
- Relatable Short Story (Context)
- Contextualized Explanation
- Evaluation Criteria (Likert Scale)

Global vs Global

SHAP (Global):
- Relatable Short Story (Context)
- Contextualized Explanation
- Evaluation Criteria (Likert Scale)

Questionnaire Design (D6)

Who

We are a group of Master's students at Maastricht University doing research on the topic of Explainable AI.

Explainable AI methods try to explain why an AI arrived at a specific decision. The purpose of this inquiry is to evaluate the quality of these explanations from a user-centric perspective. Based on selected criteria, we aim to assess selected explanation methods and to compare their explanation quality.

What

Duration
+
Privacy
Awareness

The survey will take around 15 minutes, during the survey you can choose not to answer any questions at your own discretion. There are no questions aimed at collecting identifying information such as your name, location, email address and so forth. Additionally, the survey responses will be kept confidential and will only be used for academic/research purposes.

By participating in this survey, you agree that the information gathered through this questionnaire can be used for the aforementioned purposes.

Thank you for taking the time to participate.

Questionnaire Design (D6)

- english-speaking
- program stage
- field of study
- AI experience
- knowledge XAI methods

expert.ai

Questionnaire Design (D6)

Section 3 of 6

Explanation 1

Mary has recently finished her undergraduate program and has begun to think about whether she would like to immediately enrol in a graduate program or look for a job instead. She would be willing to commit the time to apply for a graduate program if the odds of being accepted were favourable. Mary had the feeling that her high GRE scores and glowing letter of recommendation would make up for her poor GPA.

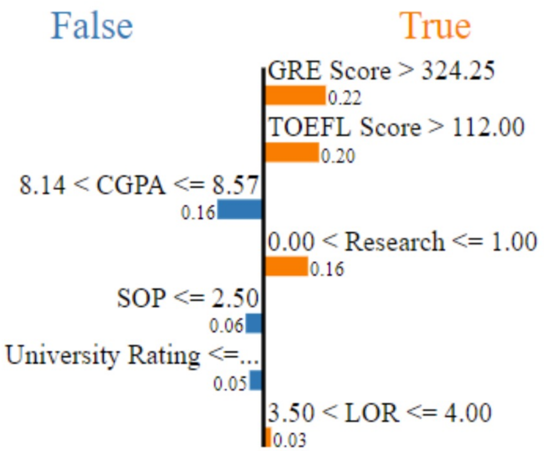
To help her decision making process, she decided to reach out to an education consultancy who could help her identify her prospects of being accepted for a graduate program. The education consultancy used an AI system based on historic data to evaluate the chance of students being accepted. She was asked to provide the following information in order to receive an evaluation:

Story to give context information

Questionnaire Design (D6)

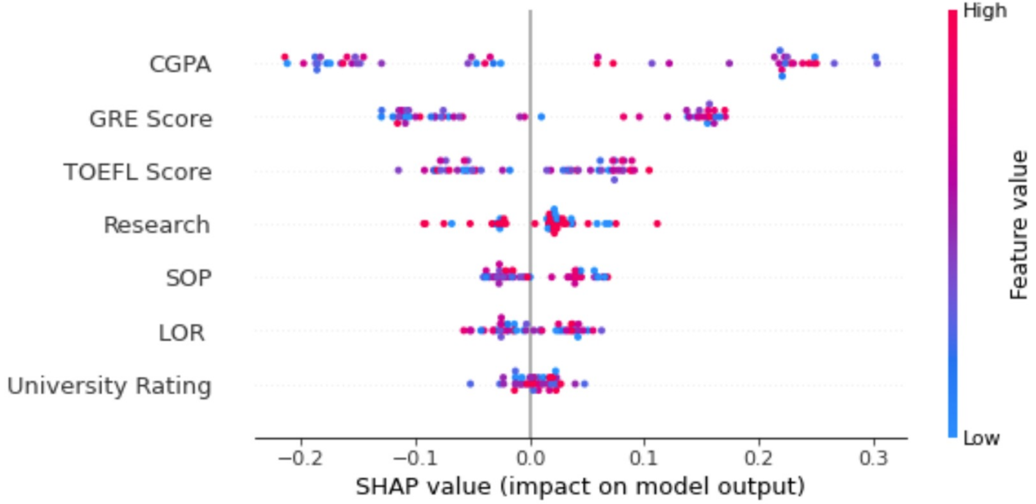
“Mary had the feeling that her high GRE scores and glowing letter of recommendation would make up for her poor GPA.”

	Scale	Mary’s Score
GRE Score	0-340	329
TOEFL Score	0-120	114
University Rating	0-5	2
Statement of Purpose Strength (SOP)	0-5	2
Recommendation Letter Strength (LOR)	0-5	4
CGPA	1-10	8.56
Research Experience	0 or 1	1



Questionnaire Design (D6)

"It was generated by Dream University's AI system which was based on past applicants at the university."



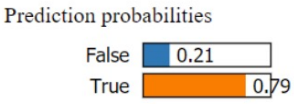
Feature values are ambiguously spreaded:

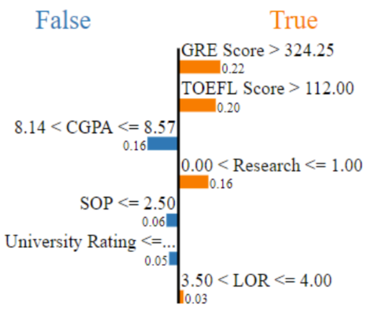
"Since the university is highly ranked with competitive students, their dataset also only contains students with high grades."

Questionnaire Design (D6)

	Scale	Mary's Score
GRE Score	0-340	329
TOEFL Score	0-120	114
University Rating	0-5	2
Statement of Purpose Strength (SOP)	0-5	2
Recommendation Letter Strength (LOR)	0-5	4
CGPA	1-10	8.56
Research Experience	0 or 1	1

Based on her information, she received the following explanation:





Feature	Value
GRE Score	329.00
TOEFL Score	114.00
CGPA	8.56
Research	1.00
SOP	2.00
University Rating	2.00
LOR	4.00

Raw explanation
generated by
XAI method

Questionnaire Design (D6)

Accompanying the explanation were the following descriptions:

The explanation shows the importance of each feature for Mary's evaluation. Additionally, she was told that true referred to the probability of being accepted and false referred to the probability of being rejected.

Additional description of explanation to disambiguate

From the explanation I understand how the system makes a decision.

	1	2	3	4	5	6	7	
Disagree	○	○	○	○	○	○	○	Agree

The explanation is useful to me, for making better decisions or to perform an action.

	1	2	3	4	5	6	7	
Disagree	○	○	○	○	○	○	○	Agree

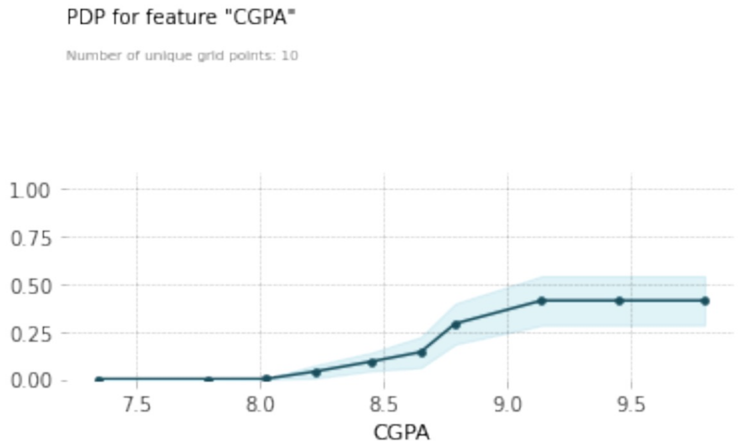
Evaluation / Rating on likert scale

Questionnaire Design (D6)

To assist non-technical people in interpreting the XAI graph

Partial dependency {GPA} → {acceptance}:

"In the y-axis, positive values mean that there is a higher **likeness** or chance of being accepted, while zero implies no average impact on being accepted according to the model."



Survey Process

Trial Run

- Send questionnaire to a few “test participants“
- Receive **feedback** & estimate time
- Incorporate feedback

Final Roll-Out

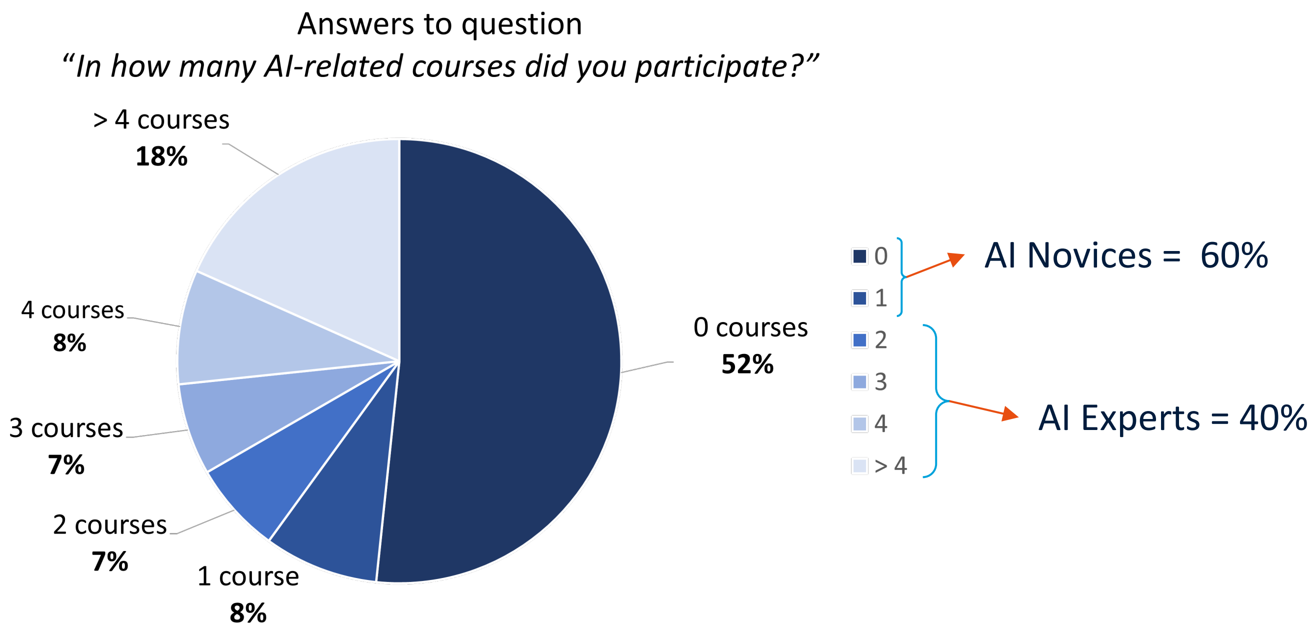
- Distribution through multiple channels
- Survey was open for 3-4 weeks
- Ultimately received **60 responses**

Google Forms: https://forms.gle/R6SfrwX13EVwgyfJA

Results and Discussion

Maastricht University

Background of Respondents



Research Questions

Respondents' Evaluation of XAI Methods

Overview of Results

	Understandability	Usefulness	Trust	Informativeness	Satisfaction
LIME	4.77 ±1.61	4.79 ±1.49	4.74 ±1.66	4.33 ±1.74	4.08 ±1.68
SHAP (local)	4.03 ±1.61	3.90 ±1.53	3.83 ±1.55	3.37 ±1.59	3.50 ±1.47
SHAP (global)	4.00 ±1.85	3.77 ±1.93	3.85 ±2.02	3.54 ±1.78	3.50 ±1.89
PDP	**5.28** ±1.59	**5.25** ±1.64	**4.84** ±1.79	**5.10** ±1.60	**5.08** ±1.64

Mean score on 7-point likert scale with standard deviation for all evaluation criteria

Comparison of Criteria

Usefulness

	Mean (SD)
LIME	4.79 (1.49)
SHAP (local)	3.90 (1.53)
SHAP (global)	3.77 (**1.93**)
PDP	**5.25** (1.64)

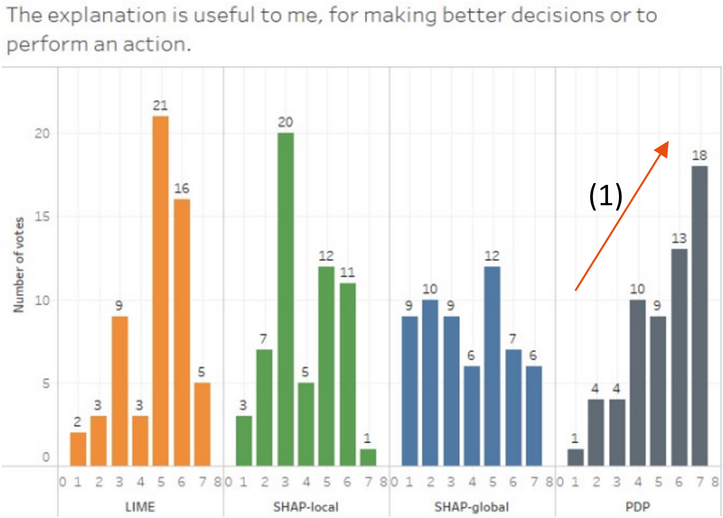


Table 3.2: Evaluation of *usefulness* for all XAI methods

Comparison of Criteria

	Mean (SD)
LIME	4.33 (1.74)
SHAP (local)	3.37 (1.59)
SHAP (global)	3.54 (**1.78**)
PDP	**5.10** (1.60)

Informativeness

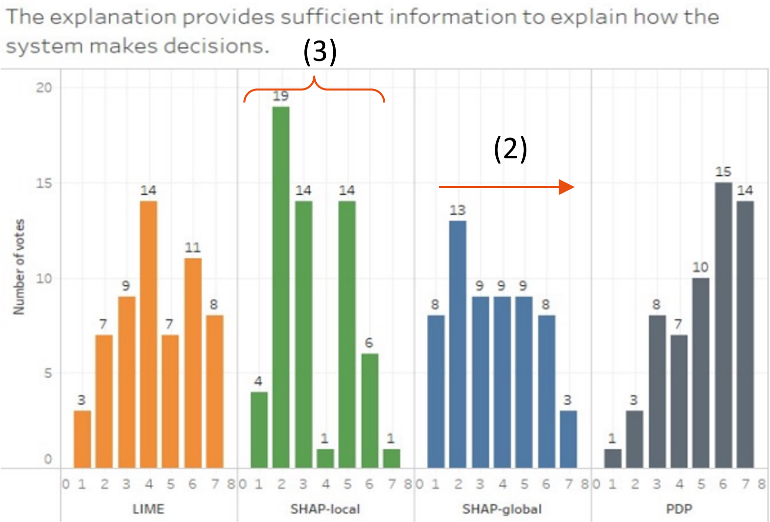


Table 3.3: Evaluation of *informativeness* for all XAI methods

Comparison of Criteria

Understandability

	Mean (SD)
LIME	4.77 (1.61)
SHAP (local)	4.03 (1.61)
SHAP (global)	4.00 (**1.85**)
PDP	**5.28** (1.59)

Mean > 4

Satisfaction

	Mean (SD)
LIME	4.08 (1.68)
SHAP (local)	3.50 (1.47)
SHAP (global)	3.50 (**1.89**)
PDP	**5.08** (1.64)

Research Question 2

Is there a preference towards local or global explanations for AI experts?

	local	global	p-value (Welch)
Understandability	4.38	**4.77**	0.21
Usefulness	4.42	**4.69**	0.41
Trustworthiness	4.04	**4.21**	0.65
Informativeness	3.52	**4.38**	**0.02**
Satisfaction	3.63	**4.42**	**0.02**

AI experts' evaluation of local and global methods (mean)

- First and foremost, unbiased evaluation, as scope was not mentioned
- AI experts have the knowledge to successfully derive additional information from global methods

Research Hypotheses

Hypothesis A: AI novices prefer local over global explanations.

Why might this be true?

Local explanations aim to explain the reasoning of a model for the results for an individual user query.

→

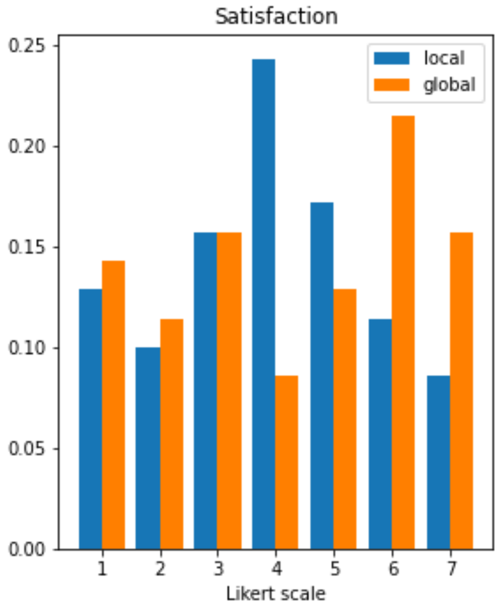
Less overwhelming for novices

Hypothesis A

AI novices prefer local over global explanations

	Local	Global
Understandability	4.43	**4.56**
Usefulness	4.31	**4.40**
Trustworthiness	**4.46**	4.45
Informativeness	4.09	**4.29**
Satisfaction	3.91	**4.21**

AI novices' evaluation of local and global methods



Hypothesis B: Explanations increase users' trust in a system.

Why might this be true?

The intuition behind the second hypothesis is that an ML model is expected to be trusted more by students when its prediction is complemented with an explanation.

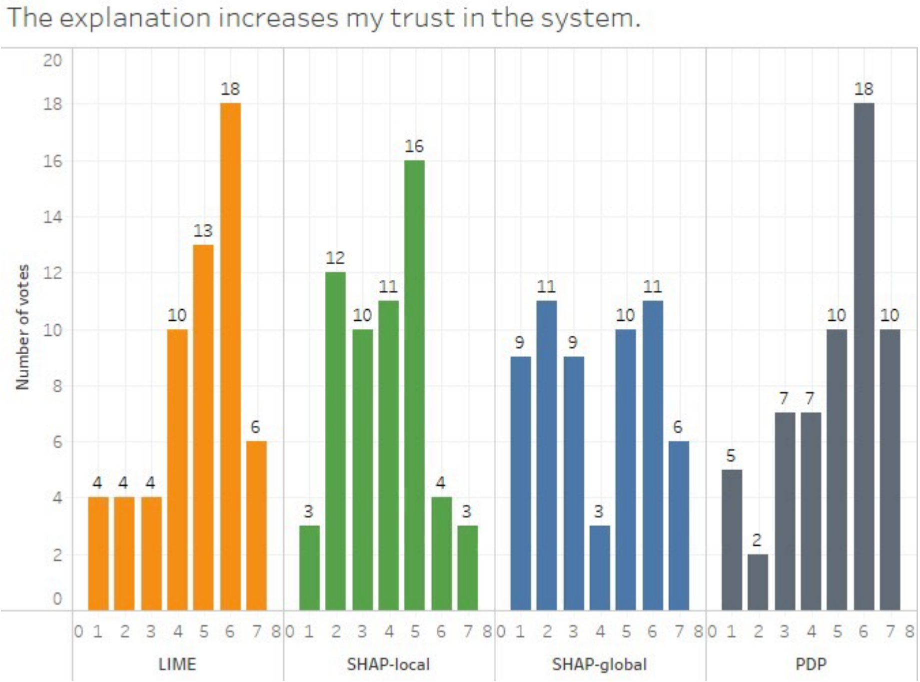
Trust is crucial for effective human interaction with AI systems

Hypothesis B

Explanations increase users' trust in a system

	Mean
LIME	**4.74** ±1.66
SHAP (local)	3.83 ±1.55
SHAP (global)	3.85 ±2.02
PDP	**4.84** ±1.79

Mean score on 7-point likert scale with standard deviation for *trust* evaluation criteria



Additional Findings

AI Novices Prefer PDP

Is the scoring for AI experts' greater than the one of AI novices for all XAI methods ?

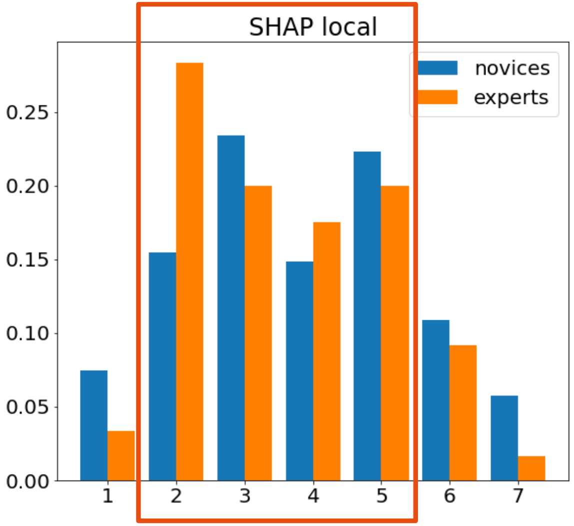
NO!!!

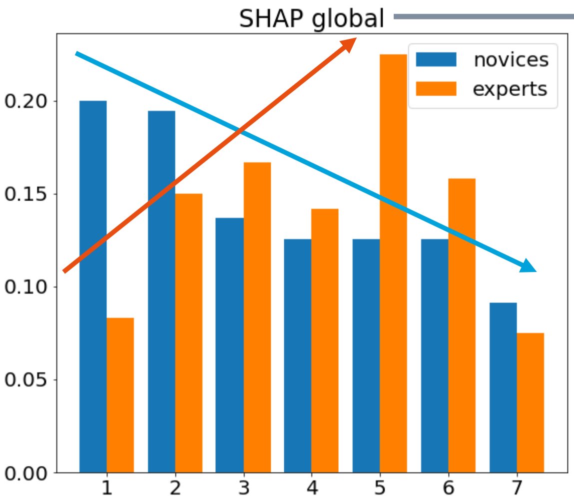
	AI experts	AI novices	p-values (Welch)
Understandability	5.08 (1.47)	**5.43 (1.66)**	0.41
Usefulness	5.13 (1.51)	**5.34 (1.72)**	0.62
Trustworthiness	4.50 (1.66)	**5.09 (1.84)**	0.22
Informativeness	5.04 (1.51)	**5.14 (1.66)**	0.81
Satisfaction	4.92 (1.66)	**5.20 (1.64)**	0.53

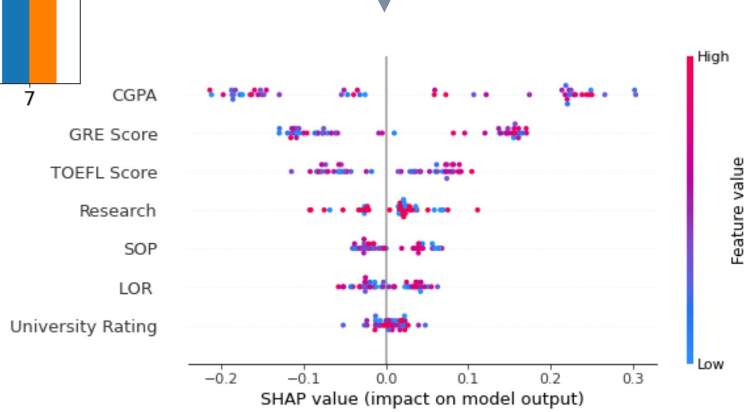
Mean and standard deviations for all evaluations regarding **PDP**

Discrepancy of SHAP

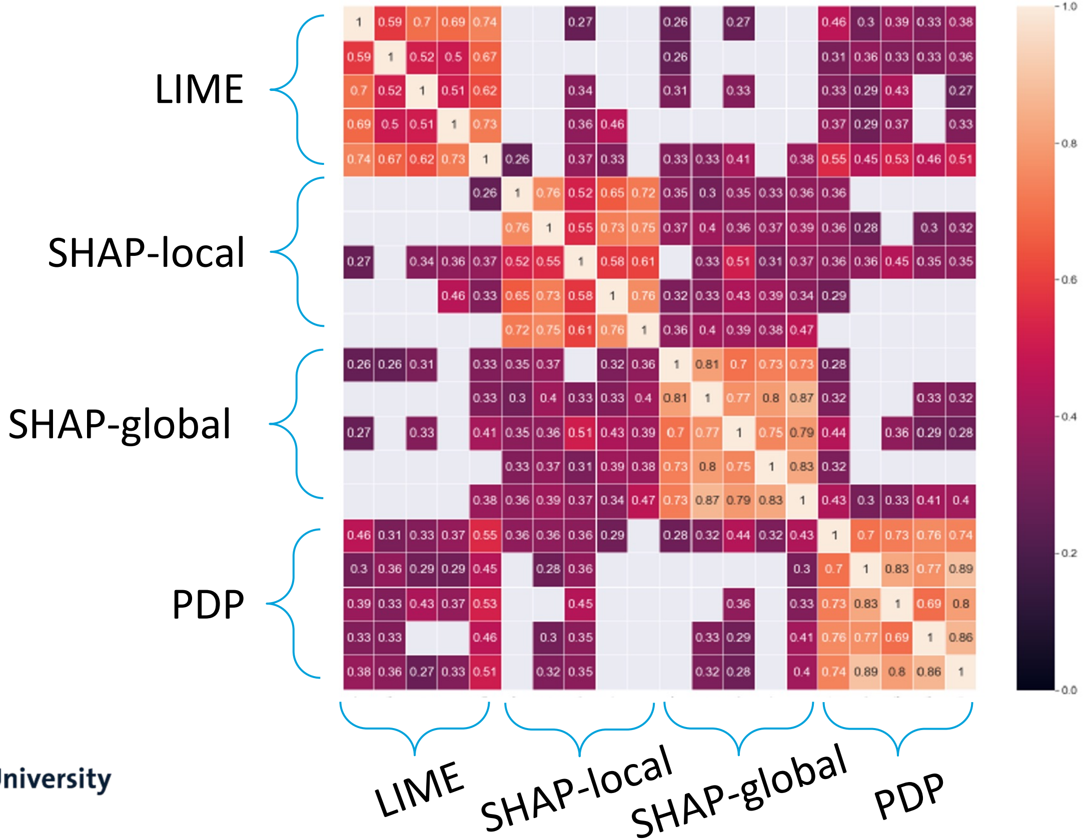
Discrepancy between SHAP – AI novices and experts







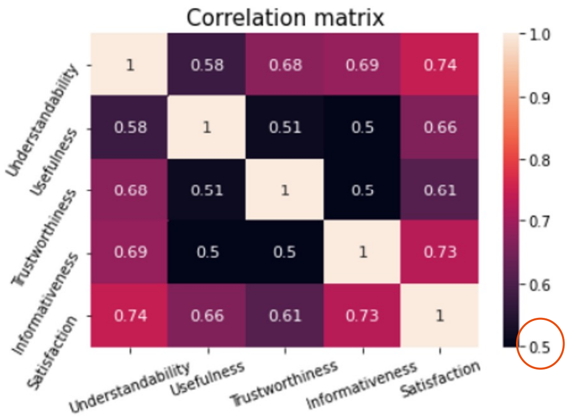
Correlations Between Criteria and Across Methods



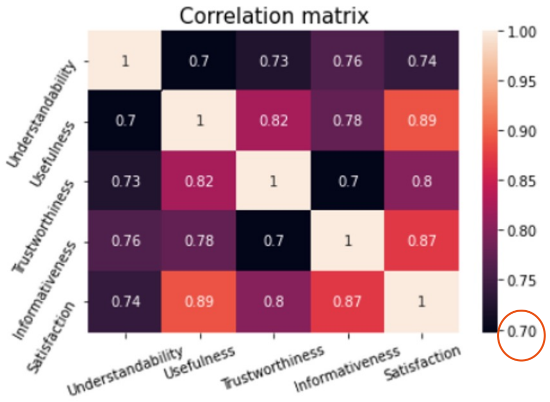
Correlation Analysis

Spearman rank correlation

- Correlation between criteria
- Increasing correlation within a method, from first to last method in questionnaire



(a) LIME



(d) PDP

Conclusion

Conclusion

Ranking of XAI Methods:

1.PDP
2. LIME
3. SHAP Local and Global

High Correlation within a XAI method
Low Correlation over all methods

Conclusion

AI Experts

PDP (+)
global SHAP (+)

Preference for Global:

Significant for ***Satisfaction*** and ***Informativeness***

AI Novices

PDP (+)
global SHAP (-)

Conclusion

Do explanations increase trust in a system?

1.PDP and LIME (> Neutral)

2. SHAP Local and Global (< Neutral)

Ranking

1. PDP
2. LIME
3. SHAP Local and Global

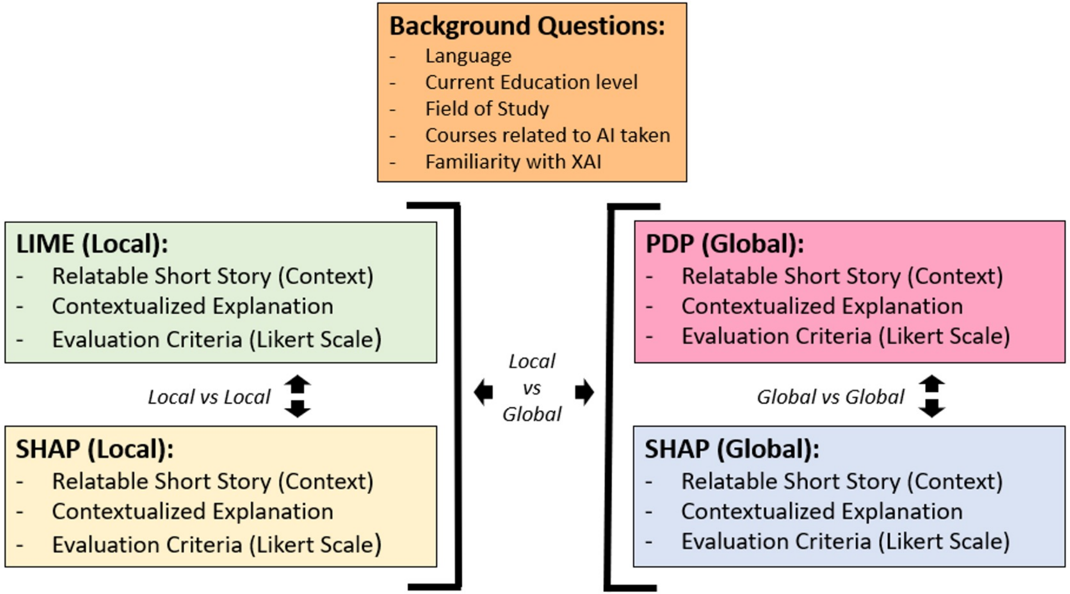
Trust

1. PDP and LIME
2. SHAP Local and Global

AI Novices

Need for Tailored Explanation

Q & A Time



References

XAI Methods

1 https://github.com/marcotcr/lime
2 https://github.com/slundberg/shap
3 https://github.com/SauceCat/PDPbox

Evaluation Criteria

4 Hoffman, Robert R, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. "Metrics for explainable AI: Challenges and prospects."

5 Dieber, Jürgen, and Sabrina Kirrane. 2020. "Why model why? Assessing the strengths and limitations of LIME."

6 Chromik, Michael, and Martin Schuessler. 2020. "A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI."

Google Forms: https://forms.gle/eXsFBCRzGhS8NQjV7