Characterizing Data Scientists' Mental Models of Local Feature Importance

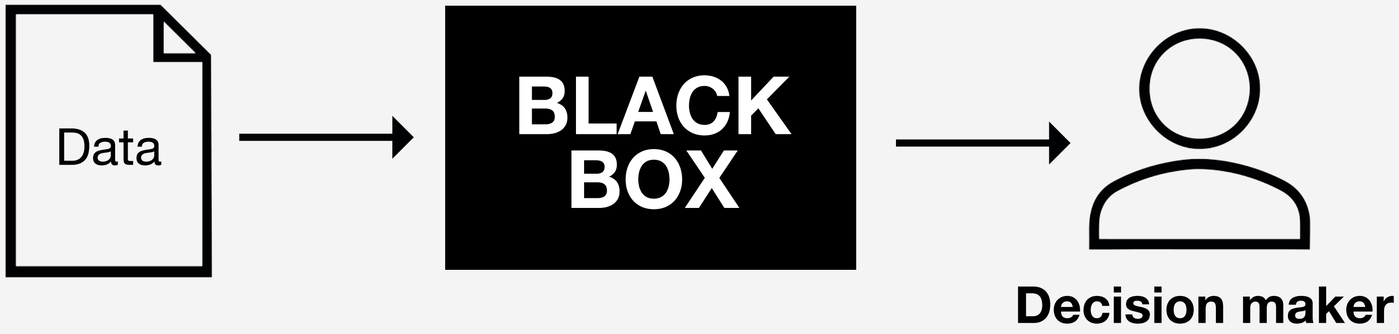
Dennis Collaris, Hilde Weerts, Daphne Miedema,
Jarke J. van Wijk, Mykola Pechenizkiy

Eindhoven University of Technology

Aarhus, Denmark – October 10, 2022

Motivation

Why do we need explanations?



→ **Undertrusting the model**

→ **Overtrusting the model**

→ **Unfair treatment of groups**

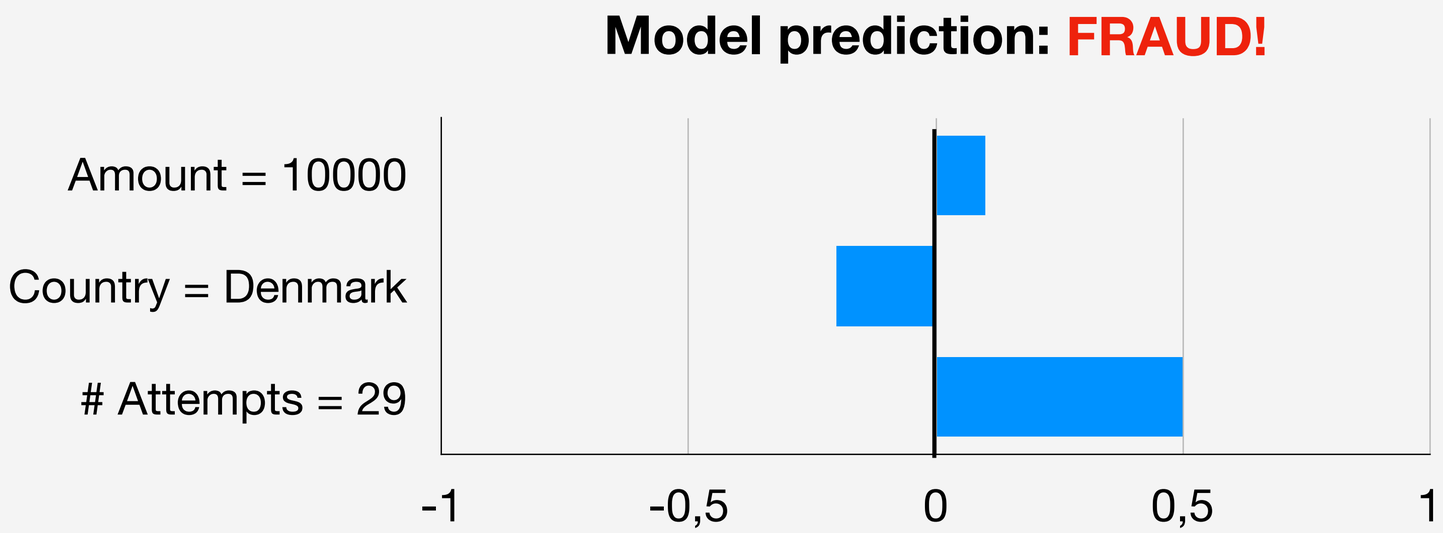
Motivation

How do we explain models?

eXplainable Artificial Intelligence

Motivation

How do we explain models?



Feature • importance

definition

A *quantitative* assignment of *importance* to the *feature* of a prediction made by a machine learning model.

Motivation

What explanation techniques exist?

Gradient-based (e.g., LIME)

Sensitivity to feature value changes

$$\hat{y} = \alpha + \sum_i \beta_i X_i$$

Ablation-based (e.g., SHAP)

Added value of *inclusion* of a feature

$$\hat{y} = \epsilon + \sum_i \phi_i$$

Motivation

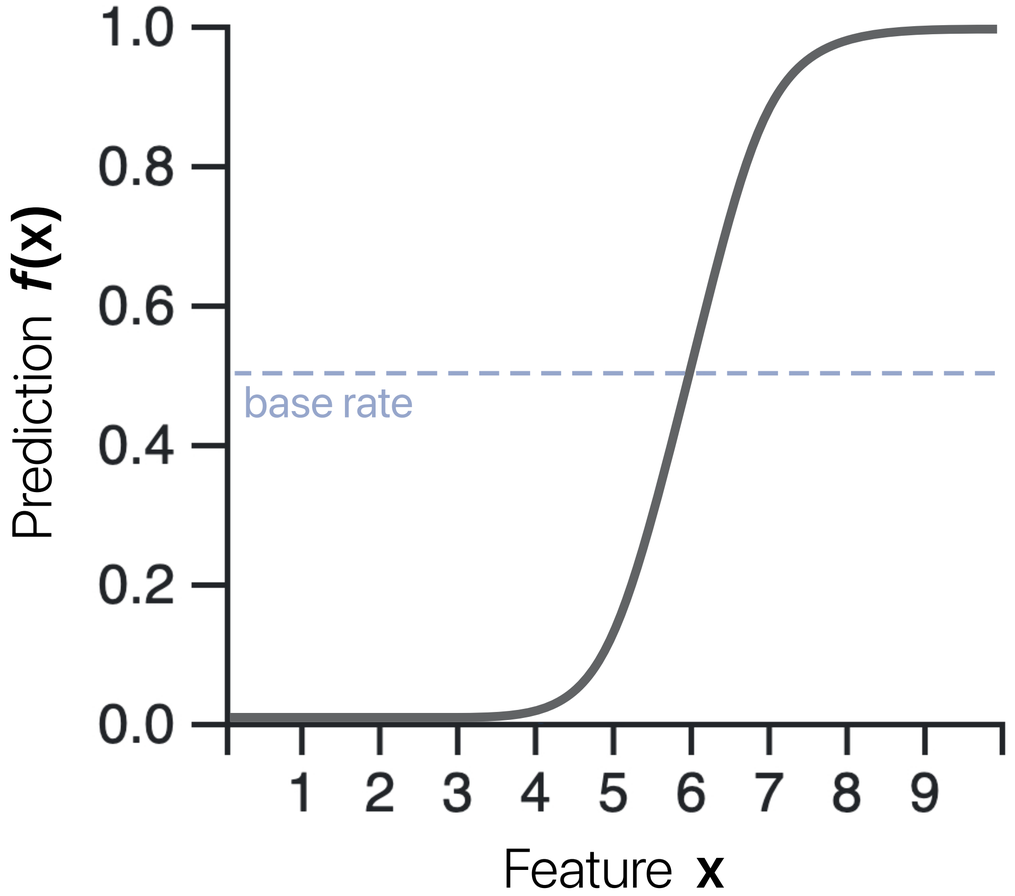
What explanation techniques exist?

Gradient-based (e.g., LIME)

Sensitivity to feature value changes

Ablation-based (e.g., SHAP)

Added value of *inclusion* of a feature



Motivation

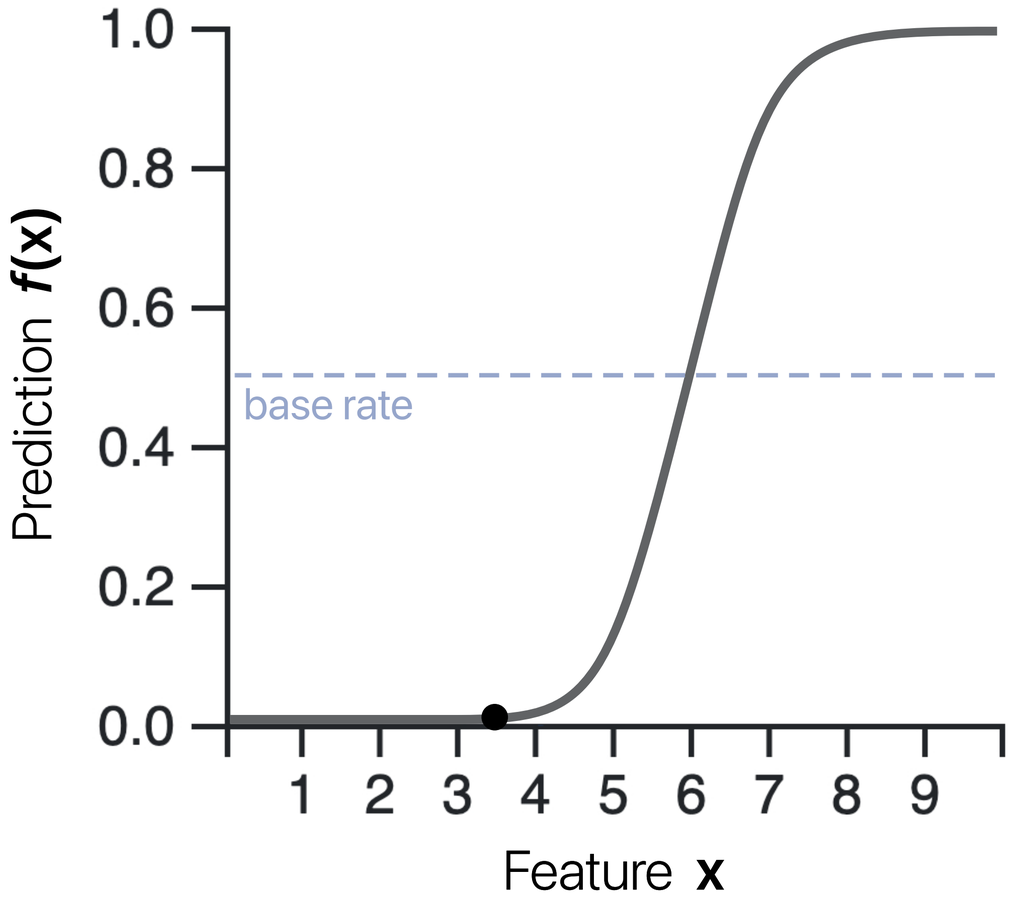
What explanation techniques exist?

Gradient-based (e.g., LIME)

Sensitivity to feature value changes

Ablation-based (e.g., SHAP)

Added value of *inclusion* of a feature



Motivation

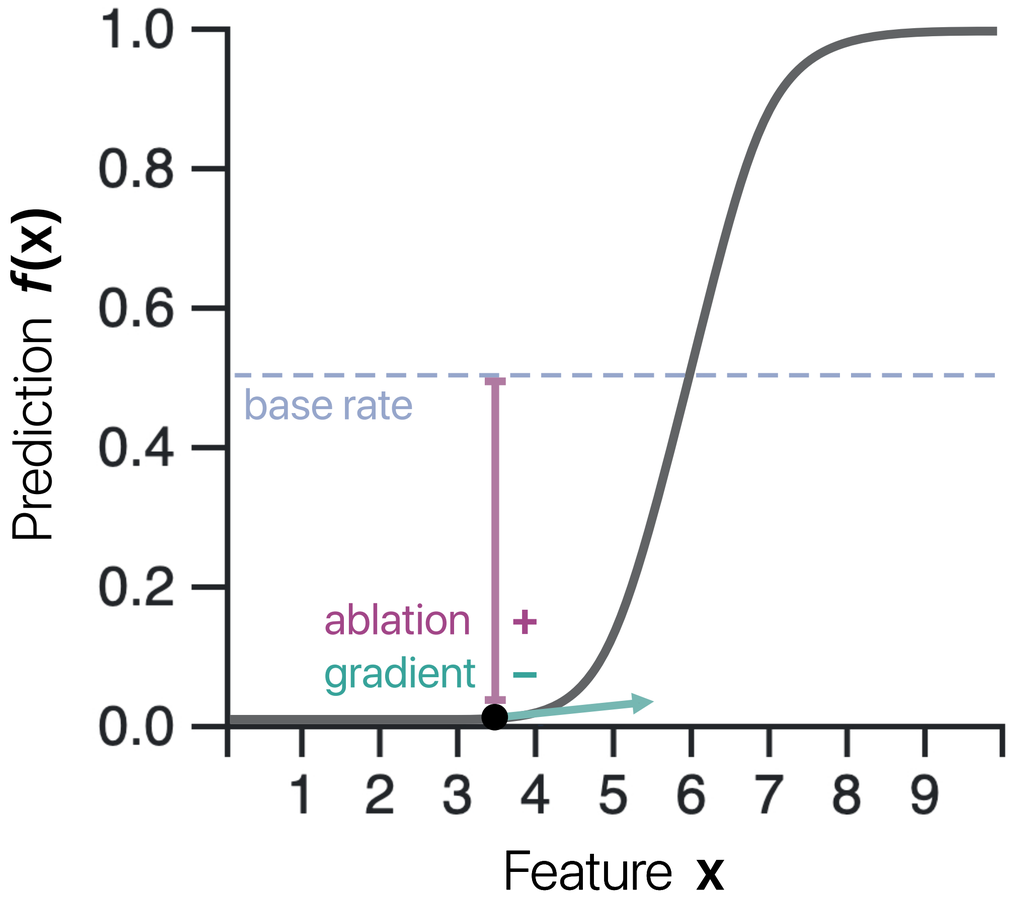
What explanation techniques exist?

Gradient-based (e.g., LIME)

Sensitivity to feature value changes

Ablation-based (e.g., SHAP)

Added value of *inclusion* of a feature



Motivation

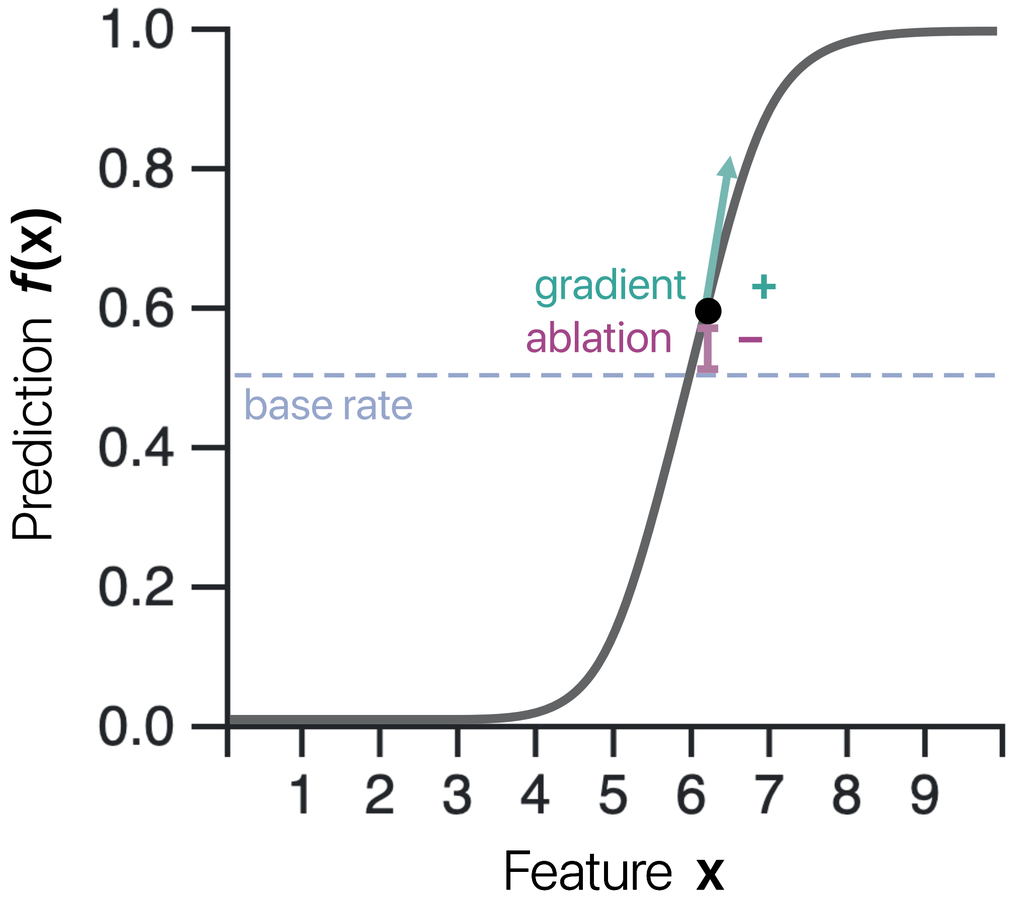
What explanation techniques exist?

Gradient-based (e.g., LIME)

Sensitivity to feature value changes

Ablation-based (e.g., SHAP)

Added value of *inclusion* of a feature



Motivation

Which technique do I use?

Motivation

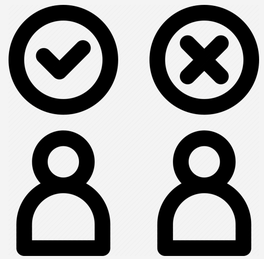
Which technique do I use?

What Exploratory mixed-methods survey

Survey

What Exploratory mixed-methods survey

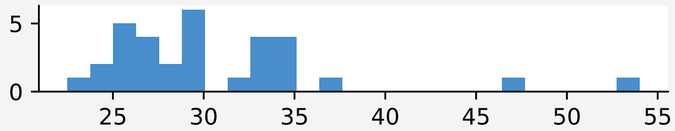
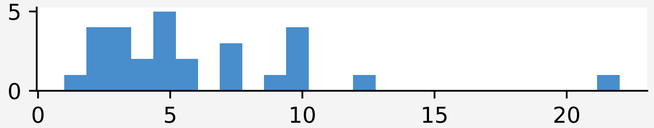
Survey



Qualitative + Quantitative

Method and results

Who Demographics

Type	**Answers** (count)
Gender	Male **(20)**, Female **(12)**, Prefer not to disclose **(2)**, Prefer to self-describe **(0)**
Age	
Location	Netherlands **(24)**, United States **(2)**, Prefer not to disclose **(2)**, Colombia **(1)**, India **(1)**, Singapore **(1)**, Spain **(1)**, Switzerland **(1)**, United Kingdom **(1)**
Role	Data scientist **(18)**, (Data science) researcher **(6)**, Software/data/AI engineer **(3)**, PhD candidate **(3)**, (Data science) consultancy **(2)**, AVP (Assistant Vice President) **(1)**, Prefer not to disclose **(1)**
Data science experience (years)	
Familiarity	Linear regression coefficients **(30)**, Random Forest feature importance **(29)**, Shapley values **(23)**, LIME **(23)**, Permutation importance **(18)**, Saliency maps (e.g., GradCAM) **(11)**, treeinterpreter **(9)**, Anchors **(2)**, DeepLift **(1)**

RQ1 How do data scientists define feature importance?

Open questions:

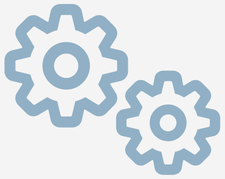
- Explain in your own words what feature importance is, in **the context of machine learning.**
- What does it mean to you when a feature is important in **a (trained) machine learning model**? *(global)*
- What does it mean to you when a feature is important for **an individual prediction**? *(local)*

RQ1 How do data scientists define feature importance?

Three main themes were identified:

Locality

Explanandum



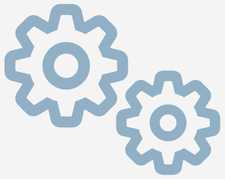
Underlying mechanism

RQ1 How do data scientists define feature importance?

Three main themes were identified:

Locality

Explanandum



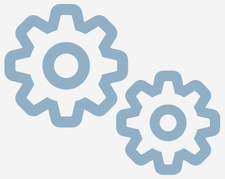
Underlying mechanism

RQ1 How do data scientists define feature importance?

Three main themes were identified:

Locality

Explanandum



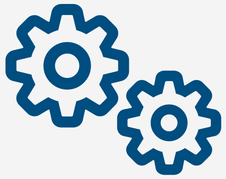
Underlying mechanism

RQ1 How do data scientists define feature importance?

Three main themes were identified:

Locality

Explanandum



Underlying mechanism

RQ2 What are the data scientists' expectations of properties of local feature importance?

I expect that if the model is relatively **uncertain** about a risk prediction, the sum of feature importances is relatively **high**.

◯ Strongly disagree	◯ Disagree	◯ Neutral	◯ Agree	◯ Strongly agree

I expect that if the model is relatively **certain** about a risk prediction, the sum of feature importances is relatively **high**.

◯ Strongly disagree	◯ Disagree	◯ Neutral	◯ Agree	◯ Strongly agree

RQ2 What are the data scientists' expectations of properties of local feature importance?

Expected properties:

- Robustness and stability
- Causality and sampling distribution
- Additivity

Varying expectations for:

- Selectivity
- Actionability and proportionality

RQ2 What are the data scientists' expectations of properties of local feature importance?

Expected properties:

- Robustness and stability
- Causality and sampling distribution
- Additivity

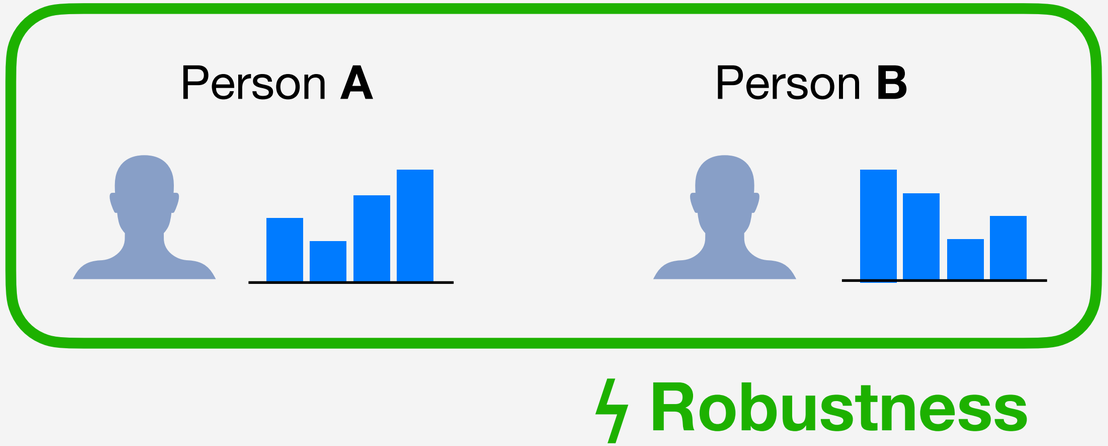
Varying expectations for:

- Selectivity
- Actionability and proportionality

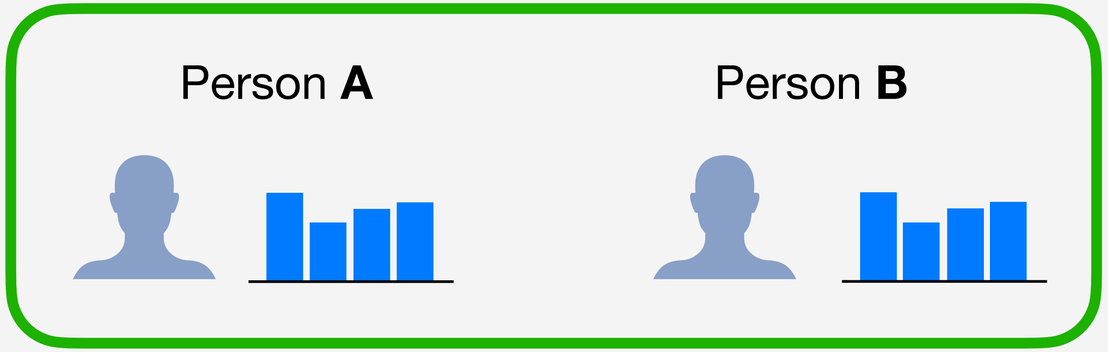
Method and results

RQ2 ❶ Robustness & ❷ Stability

RQ2 ❶ Robustness & ❷ Stability

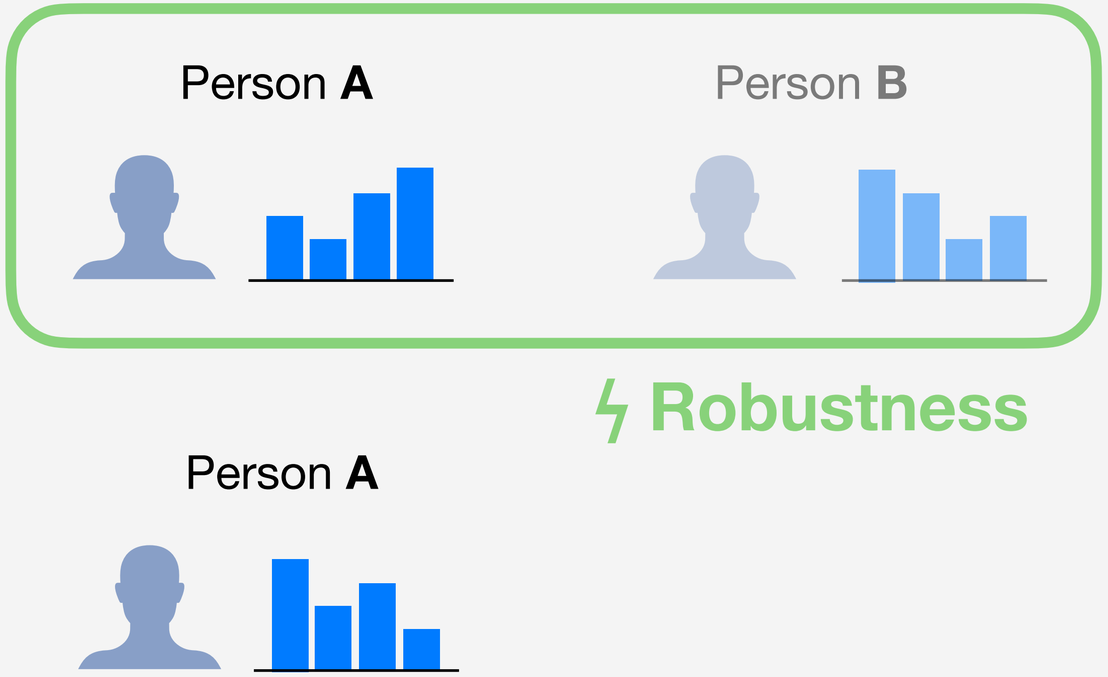


RQ2 ❶ Robustness & ❷ Stability



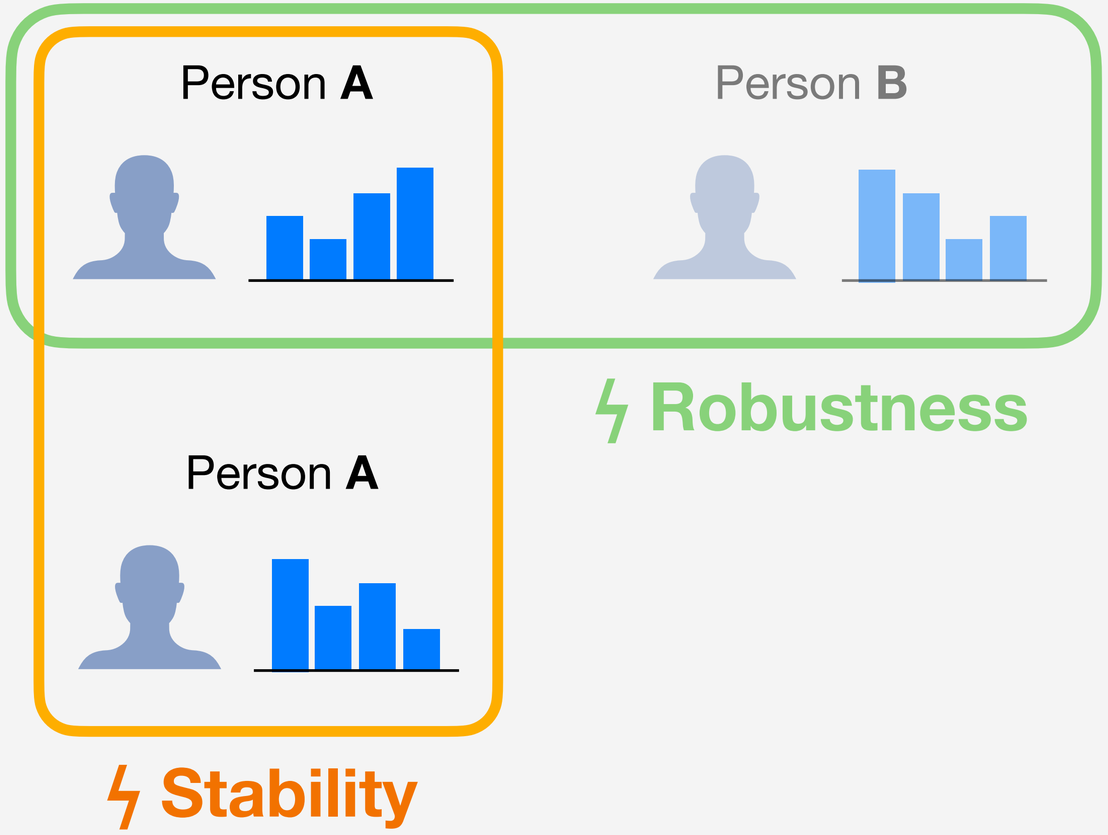
✓ **Robustness**

RQ2 ❶ Robustness & ❷ Stability

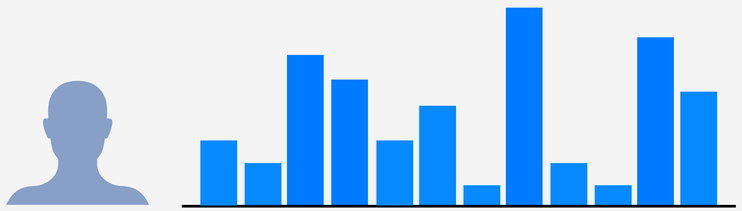


Method and results

RQ2 ❶ Robustness & ❷ Stability



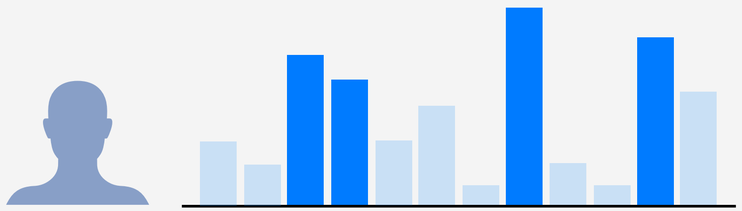
RQ2 ❸ Selectivity



Selectivity

Method and results

RQ2 ❸ Selectivity



✓ **Selectivity**

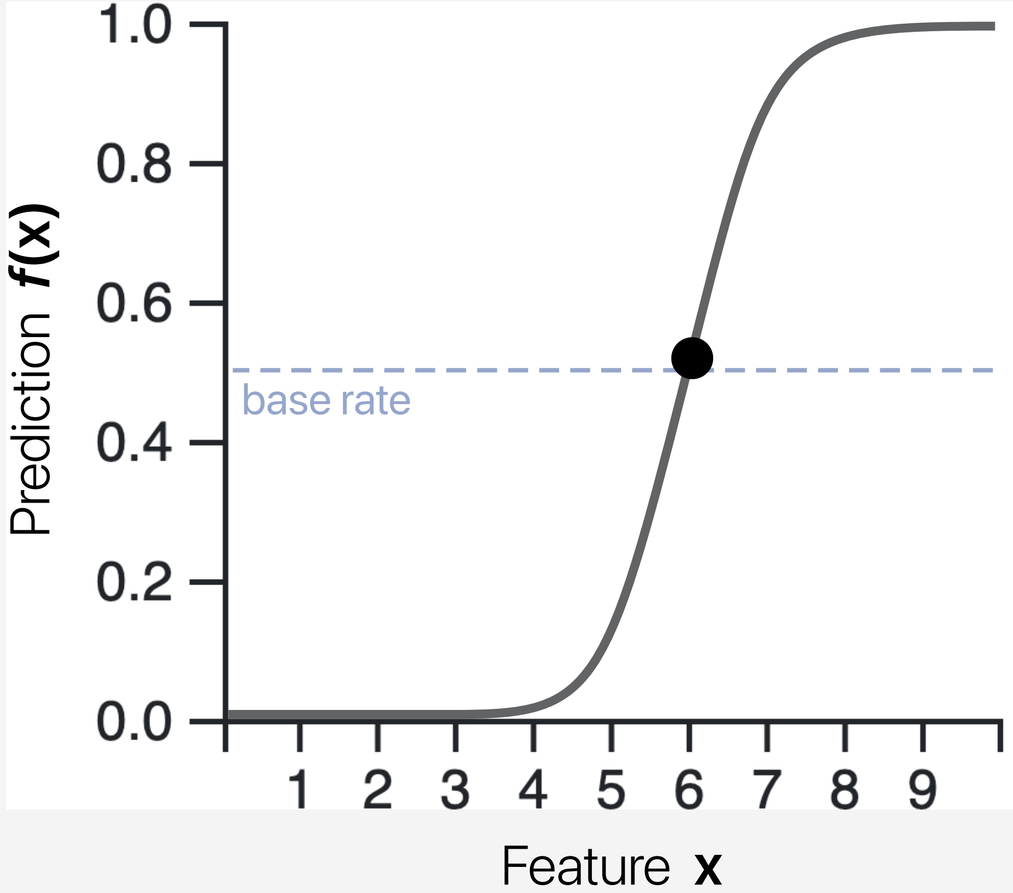
RQ2 ❹ Actionability & ❺ Proportionality

	Actionable	Proportional
Gradient-Based (LIME)	✓	
Ablation-Based (SHAP)		✓

Method and results

RQ2 ❹ Actionability & ❺ Proportionality

	Actionable	Proportional
Gradient-Based (LIME)	✓	
Ablation-Based (SHAP)		✓



What do we conclude

- The definition and computation of “feature importance” can vary between methods resulting in different **properties** of feature importance values
- Data scientists expect feature importance values to be *robust* and *stable*
- Data scientists have **varying** expectations for *selectivity*, *actionability*, and *proportionality*.
- Data scientists have expectations of properties that are **incompatible** with existing techniques

What technique should you pick?

None of the existing techniques matched all participants’ expectations!

What technique should you pick?

None of the existing techniques matched all participants' expectations!

Feature importance →

Feature sensitivity
for gradient-based techniques.

Feature attribution
for ablation-based techniques.

Characterizing Data Scientists' Mental Models of Local Feature Importance

Dennis Collaris, Hilde Weerts, Daphne Miedema,
Jarke J. van Wijk, Mykola Pechenizkiy

Eindhoven University of Technology

Aarhus, Denmark – October 10, 2022