Data Mining 2

Topic 09 : eXplainable AI

Lecture 01 : eXplainable AI

Dr Kieran Murphy

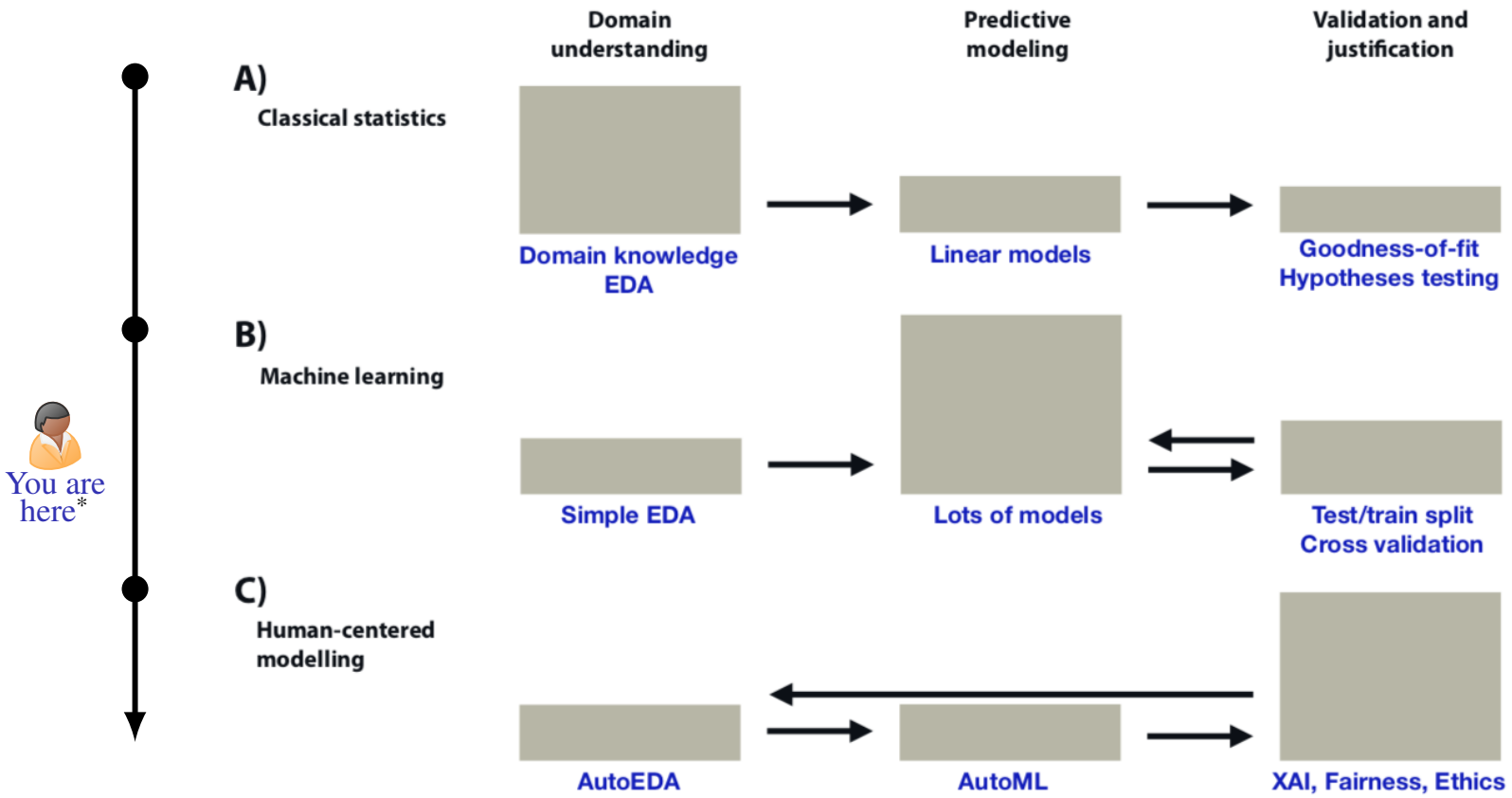
Department of Computing and Mathematics, WIT.
(kmurphy@wit.ie)

Spring Semester, 2021

Outline

- Explainable AI (XAI)

Shift in Relative Importance in Data Science Modelling



(A→B)

- Move from statistical modelling to machine learning (week 4).

(B→C)

- Increased automation of the hyper-parameter tuning, model fitting, model selection, feature engineering, EDA stages.
- Resulting models are better (score higher) but are more opaque. œ
- Increased need for model generic tools to explain/justify predictions

*Explanatory Model Analysis, `pbiecek.github.io/ema/`

Why Should We Care About XAI?

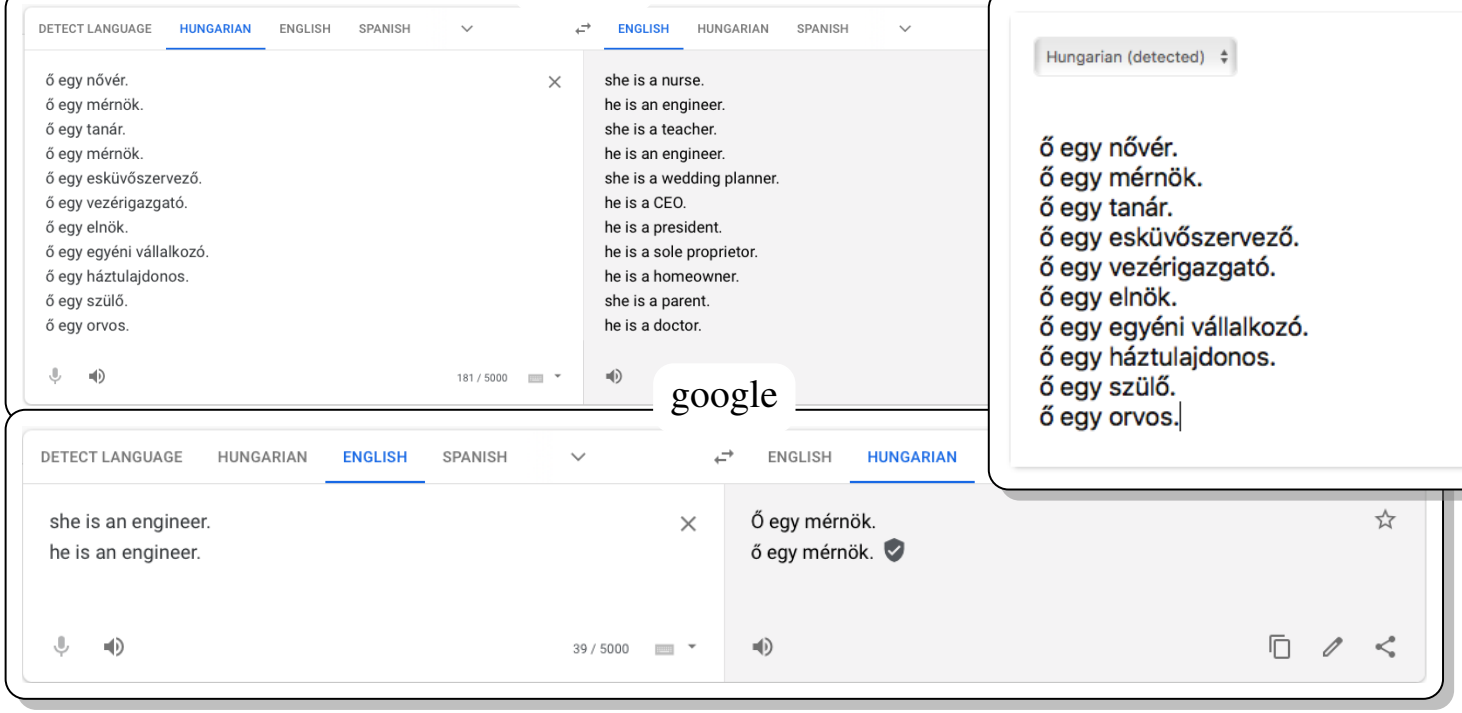
> "With great power there must also come great responsibility"
>
> — Peter Parker principle, Marvel Comics

- Predictive models are only as good as the data they are trained on:
  - Google's image object identification was trained on a dataset of predominantly white people — resulting in classifying some black people as gorillias (WSJ, July 2015)
  - Nikon S630, Hewlett-Packard's MediaSmart web camera, ...
- Mass adoption of (the more "effective") models means that pre-existing systematic biases are propagated and become more baked in.
  - Amazon's same-day delivery service was unavailable for ZIP codes in predominantly black neighbourhoods — correlated to areas affected by mortgage redlining in the 1960s. (Bloomberg, 2016)
  - Amazon developed a internal recruiting recommendation engine but was initially trained on male dominated data and even after subsequence attempts to remove gender the model learnt to use proxies for gender (membership of women's tennis etc, or 'male oriented' verbs). (Reuters, 2019)
- Decisions made by models are harder to refute (Think Little Britain's "Computer says 'NO' ")
  - Software that assessed the risk of recidivism in criminals was twice as likely to mistakenly flag black defendants as being at a higher risk of committing future crimes. (ProPublica, 2016)

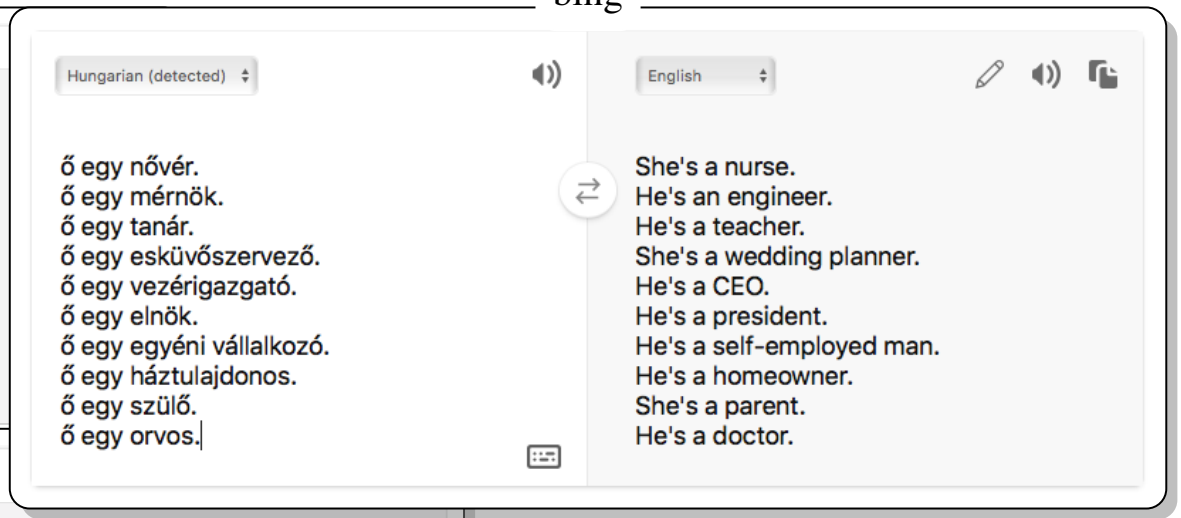
Example: Assuming Gender Based on Stereotypes

What do† nurses, teachers, wedding planners, or parents have, that engineers, CEOs, presidents, sole proprietors, home owners, and doctors don't? . . . and it is not just Google

google



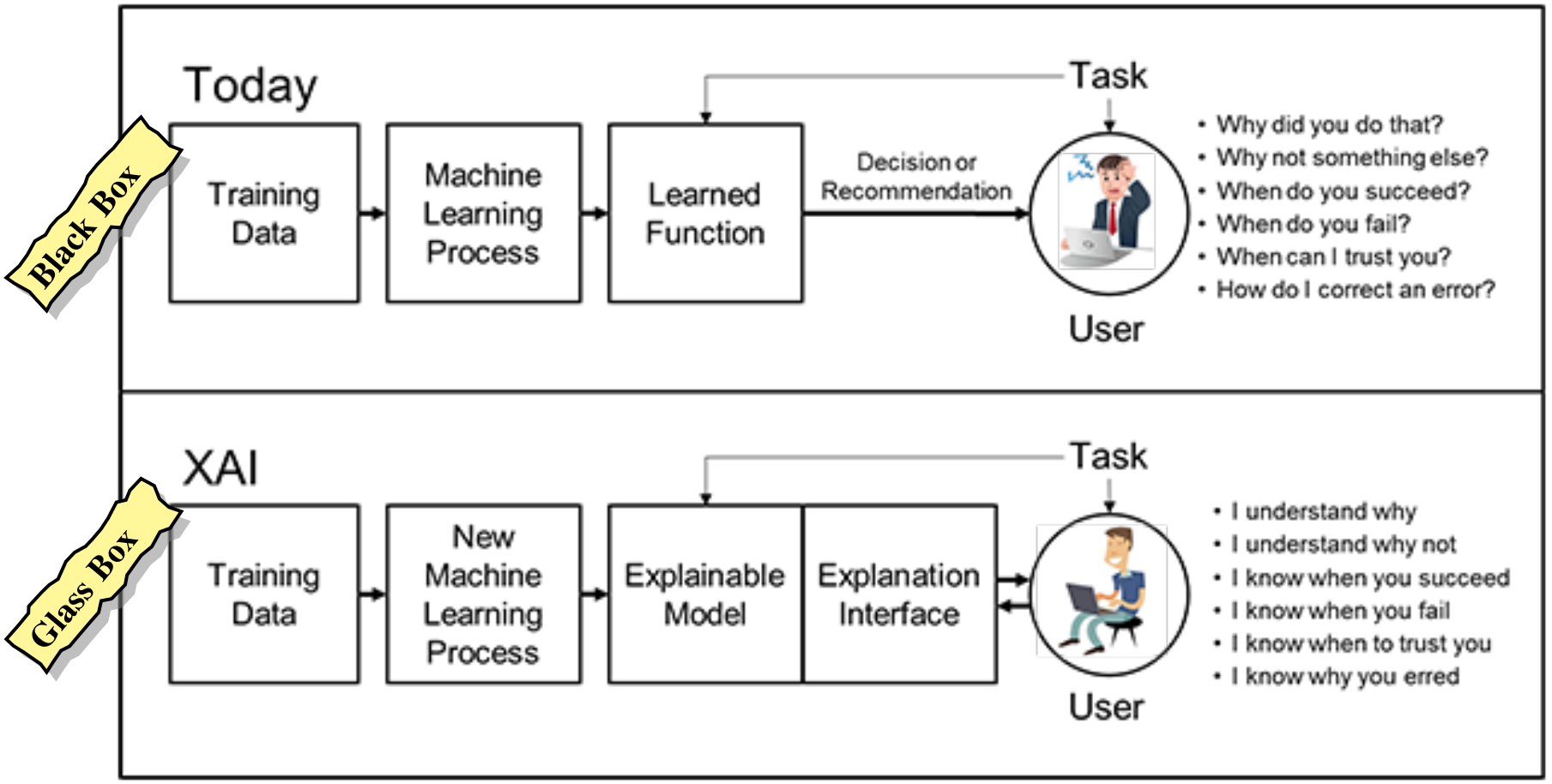
bing



google

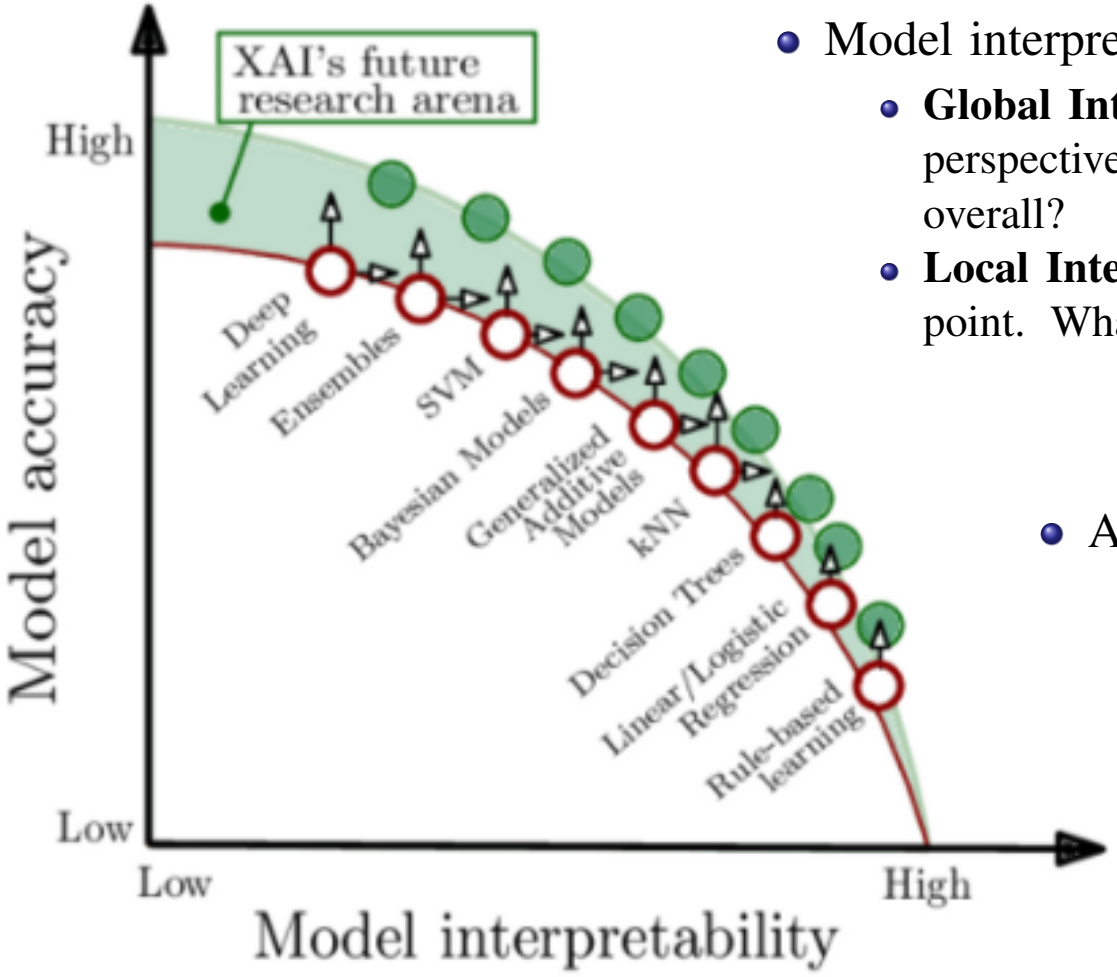
†Full disclosure — I knew about nurse, engineer, wedding planner and CEO before. The other professions were my first picks. The only surprise is shop assistant. (March 8, 2021).

The need for XAI‡



‡DARPA, Our Research, XAI, Dr. Matt Turek

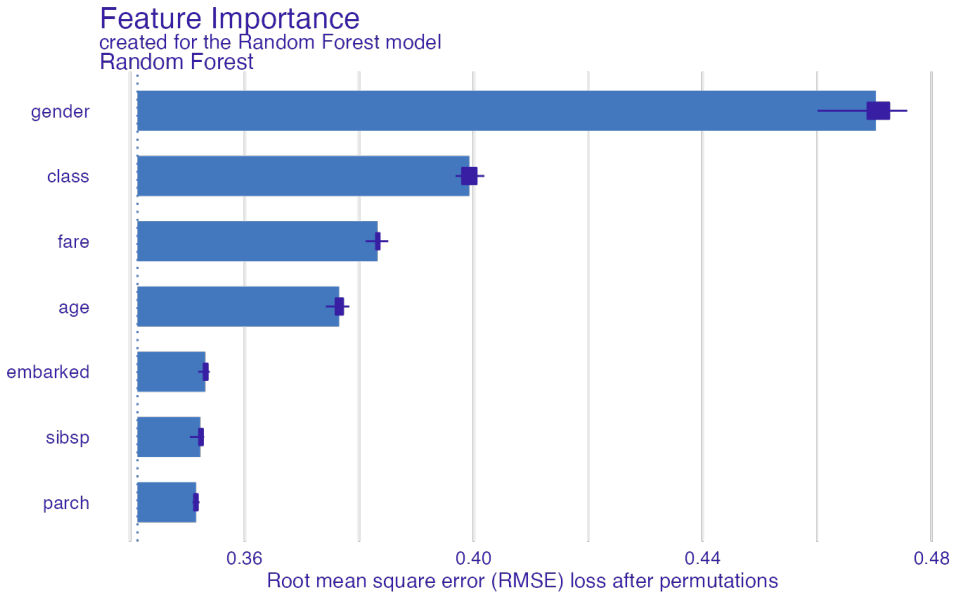
Accuracy vs Interpretability Trade Off

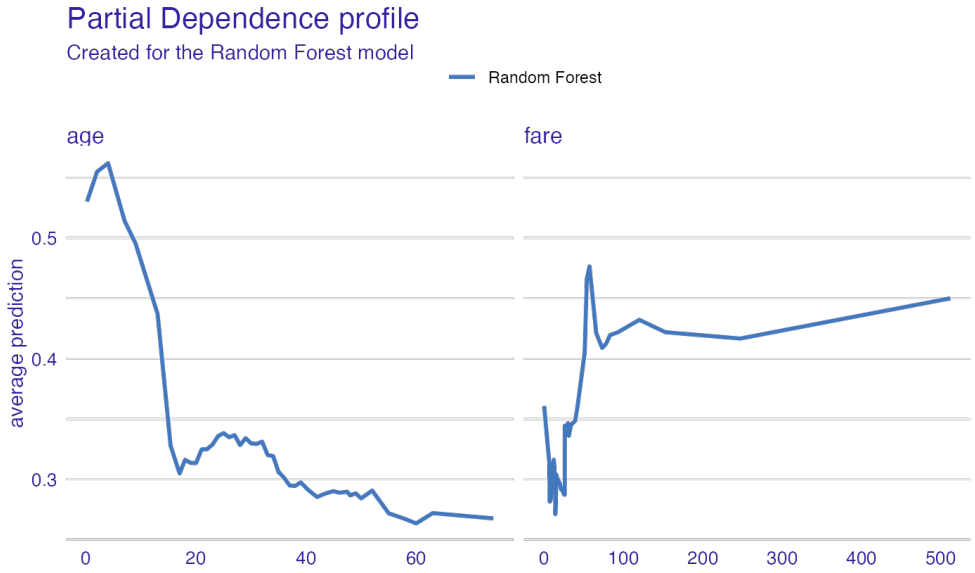


- Model interpretability can be examined in two levels:
  - **Global Interpretation** examines the model from an over all perspective. Which features are important and how important overall?
  - **Local Interpretation** is focused on a individual observation/data point. What features contributed to this prediction?

- Active area of research and development:
  - **LIME**
    Local Interpretable Model-Agnostic Explanations
  - **SHAP**
    Shapley Additive Explanations
  - **DALEX**
    moDel Agnostic Language for Exploration and eXplanation
  - ...

XAI Example: Titanic (via DALEX) — Global Interpretation





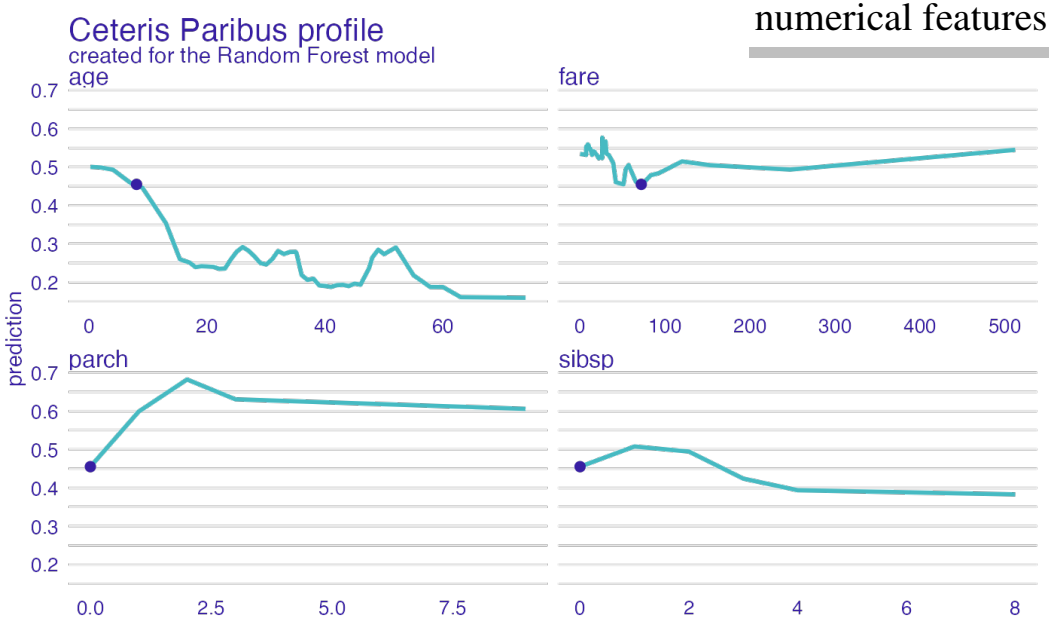
- The important feature is gender.
- Next three important features are class (1,2,3), age and fare.

- Kids under 5 have much higher probability of survival, drops significantly near 20.
- Fare reflects passenger class, why the spike?

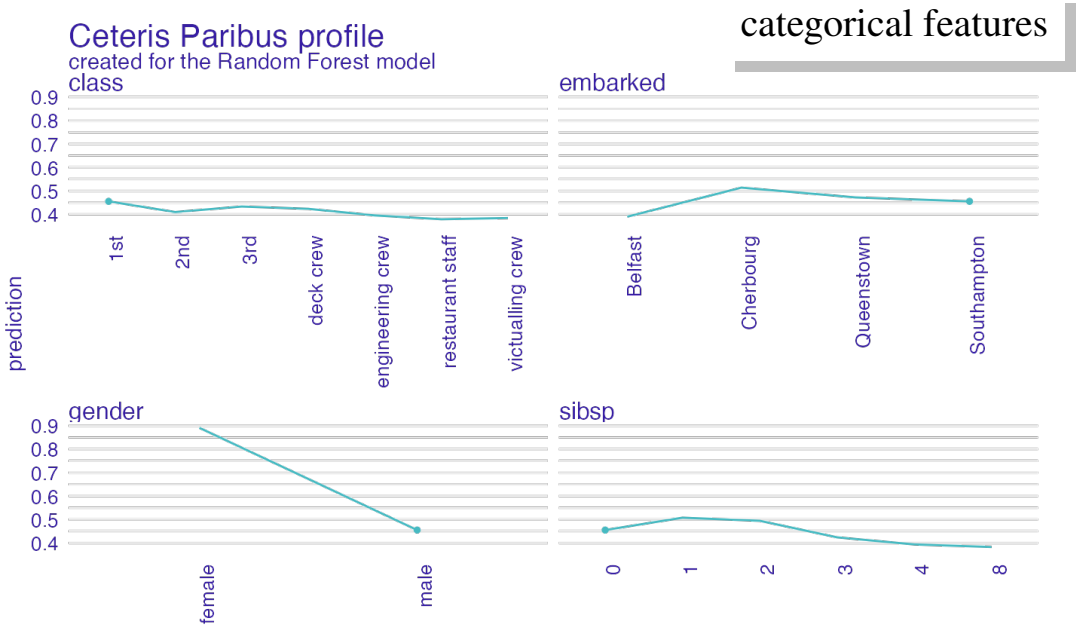
XAI Example: Titanic (via DALEX) — Local Interpretation

Lets break down explanation for model predictions for a fictitious, Ceteris Paribus, an 8 years old male, in 1st class that embarked from port C . . .

numerical features



categorical features



- It looks like the most important feature for this passenger is age and sex.
- His odds for survival are higher than for the average passenger. Mainly because of the young age and class, despite being a male.