Lecture-11: XAI continued

Techniques of explanations

- Global → overall model
- Local → specific predictions
- Some models are more interpretable / explainable than others
- Need for explaining “black-box” models for making decisions
  - LIME → Learn a more, explainable surrogate model for a point’s neighborhood by creating more points in the neighbourhood (by perturbing the inputs)
  - Model-agnostic, but hard for building global mental models

SHAP

- Shapley values from Game theory
- Intuition: See how much each feature contributes to the model's predictions (at a point as well as overall)
  - Great globally, and locally!

Shapley values

- Suppose there is a team T= {1,2…p}, playing a game G.
- All could play, or a subset forming a team T.
- The team overall won a prize v.
- Shapley value, $\phi_m\,(v)$ = fair share of prize given to member m.

$$\varphi_m(v) = \frac{1}{p}\sum_{S}\frac{[v(S \cup \{m\}) - v(S)]}{\binom{p-1}{k(S)}}, \qquad m = 1,2,3,\ldots p$$

How is this arrived?

$$\varphi_m(v) = \frac{1}{p}\sum_{S} \frac{[v(S \cup \{m\}) - v(S)]}{\binom{p-1}{k(S)}}, \qquad m = 1,2,3,\ldots p$$

- Take all possible playing subsets S from T without m.
- Compute the prize to be won by that team without S, and the same team with S.
- The difference in prize shows the marginal contribution of player m.

Has nice properties

- Efficiency: The total of individual contributions is equal to the team's value (total prize = sum of individual payoffs)

$$\sum_{m=1}^{m=p} \varphi_m(v) = v(T)$$

- Symmetry: Same contribution results in same share of prize

$$\text{if } v(S \cup \{m\}) = v(S \cup \{n\}), \text{ then } \varphi_m(v) = \varphi_n(v)$$

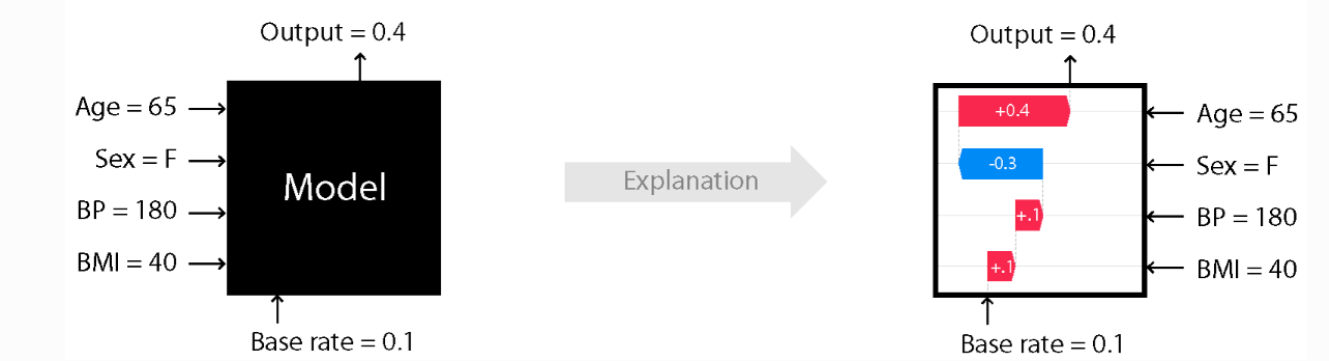
- Linearity: If the team participates in multiple projects/rounds/games with values v and u, the sum of values and contributions can just be added up.

$$\varphi_m(v + u) = \varphi_m(v) + \varphi_m(u)$$

$$\varphi_m(av) = a\,\varphi_m(v)$$

Applied to machine learning

- Think of ensemble models / DNNs
- We need ways of explanation of each layer / model
- Attributes the overall model predictiveness to a feature
- Summative → so some linear summation of features



Intuition

- Suppose we have a linear model, learning f:

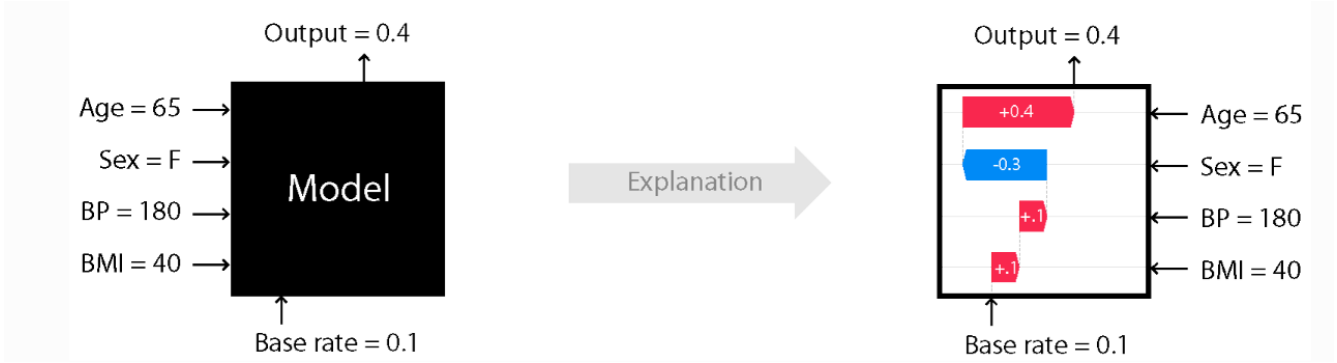
$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n.$$

- Contribution of i-th feature x_i (partial dependence on x_i) for an instance:
  - | model's prediction with x_i - model's prediction with x_i unknown|
  - Unknown = expected value of x_i, $E(x_i)$ / pre-determined.

$$\begin{aligned}\varphi_i^{additive}(x) &= \beta_0 + \cdots + \beta_i x_i + \cdots + \beta_n x_n - (\beta_0 + \cdots + \beta_i \mathbb{E}[x_{,i}] + \cdots + \beta_n x_{,n}) \\ &= \beta_i (x_i - E[X_i]).\end{aligned}$$

- Account for every subset of features, since features interact

Remember good explainability



NEEDS
INPUT
OUTPUT
PERFORMANCE
WORKING
WHY
WHY NOT
WHAT IF
HOW TO BE THAT
HOW TO REMAIN HERE

Pros and Cons

- Model agnostic
- Works for complex, diverse models
  - Ensemble, layers of a neural network, text, …
- Easy to visualize and interpret
- Cons → compute heavy, especially as we try to understand intermediate layers for images, etc.
- Misread
  - Contribution is NOT causation.

Other techniques: Counterfactuals

- Counterfactual → if X hadn't happened, Y wouldn't have happened.
- Can we provide such counterfactual explanations of features, leading to a local prediction?
- Change features, see what changes the prediction, and then use that as a counterfactual
  - Great for answering "why not" / "why"
  - Human-friendly explanations
  - Not be read as causation

Summary: Model-agnostic explanations

- The SIPA principle
  - Sample from the data
  - Intervention on the sampled data (e.g., perturbation)
  - Predict on the manipulated data
  - Aggregate the results
- LIME, SHAP, counterfactuals, …

Need for careful presentation / design

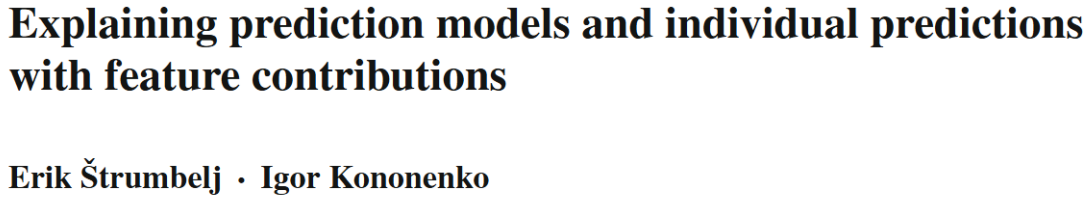
- Overloaded, information heavy
- Present on-demand
- Allow for hailing, preserving and dismissing explanations of interest
- Present a variety (visuals, text, even types)
  - Emerging consensus: No one size fits all

Evaluating explanations

- Use of user studies & usability studies
- Ask people (users) be the judge!
- One way:
  - Build explanations / sketch
  - Show people
  - Ask questions about the model
  - Ask to answer with this tool
  - Evaluate confusion & correctness

Read the official docs:

- Required reading:



- https://shap.readthedocs.io/en/latest/index.html
- At some point, you will have to use it for your project.
- Great general reading on interpretable / explainable ML [SKIM]
  - https://christophm.github.io/interpretable-ml-book/