Solution Quiz:

(a) What is the prediction target in contrast to the prediction?

⇒

(b) What is the model's mechanism in contrast to the data generating process (DGP)?

⇒ The DGP is the true underlying process generating the data. The model is estimated by optimizing the prediction performance, i.e. the model's mechanism tries to be a copy of the DGP but this mostly is not possible.

(c) Yes or No: Shapley values and SHAP values are different names for the same concept.

⇒ No

(d) What are SHAP value functions v in contrast to SHAP values ϕ?

⇒ SHAP value function: a function assigning a value to a coalition

⇒ SHAP values: Contribution of a feature to a prediction

(e) What is the difference between marginal and conditional SHAP?

⇒ The features that are not part of the coalition or the feature of interest are sampled from the marginal or the conditional distribution, respectively.

(f) What is the difference between kernel SHAP and tree SHAP?

⇒

(g) Does the dependence structure in the DGP influence the SHAP result?

⇒

Solution 1:

Your employer, E-Corp, has set up a new special task force, which you are part of. The goal of the task force is to increase the trust of individuals into the AI tools that the company sells to a wide range of businesses, governments and individuals.
The task force was set up as a reaction to criticism of its top-selling AI detection system *Saruman's stone*. The goal of the system is to distinguish innocent civilians from targets in AWS[1] scenarios.
As a first measure to increase trust, you were asked to apply common explanation techniques like SHAP and LIME to the aforementioned AI model. In contrast to the model and the data, which are well kept company secrets,[2] the results of your interpretations will be communicated to a board of independent journalists.
In your mission to increase trust and advance the progress of AI you have the clear task to avoid results that may undermine the progress of E-Corp or the adoption of Saruman's stone.

Now it is up to you to save AI!

(a) Build groups of 2-3 people.

[1] Here, AWS stands for autonomous weapon systems
[2] E-Corp wants to prevent evil terrorists from exploiting the knowledge to harm innocent civilians and especially children.

(b) Fit a random forest classifier with default hyperparameters on the dataset `data.csv`.

(c) Apply SHAP and LIME with default hyperparameters to the dataset. Interpret the results. Which conclusions do you draw (you can use shaper for R and shap for python)?

(d) Since your mission is to increase trust in AI systems, you wonder about ways to improve the explanation results. Try to adjust the interpretation by modifying the hyperparameters of LIME.[3]

Congratulations! E-Corp investor Eter Iehl himself called into your special bonus incentive holiday event[4] to give you the *E-Corp ethical AI award* for tackling unfairness in AI. Unfortunately, just a few days later your holiday is interrupted by an unpleasant report in MIT Tech Review. Tinmit Urbeg publicly accuses E-Corp of cheating by tuning the LIME hyperparameters!
Eter Iehl is super mad and you have to end the holiday. Sad! But then your supervisor has an idea. She once heard that SHAP relies on unrealistic artificial datapoints...

(e) Exploit the extrapolation to adjust the SHAP interpretation such that skin color and beard are considered irrelevant. *Hint: It is sufficient to sketch an approach with pen and paper.*

(f) Can you think of further ways to "improve" the interpretation?

Solution 2:

Discuss with your team:

(a) In what sense are local, post-hoc interpretation techniques helpful in auditing AI systems?

(b) Are local, post-hoc interpretations sufficient to assess the fairness of algorithmic decisions?

(c) How would you design the AI auditing process? What are pitfalls that you should avoid?

[3] If you cannot find hyperparameter configurations for which you get the desired result, you may change the model as well (i.e. reduce the number of estimators or the maximum tree depth).

[4] in the seasteading *Ocean Freedom Nation* in Brazil