Interpretable ML

Motivation for Interpretable ML

We want to use ML to make the best predictions possible to make better decisions.

- **Examples**: which houses have the highest risk of lead, which inmates are most likely to return when released.

However, the person making the decision may not be the analyst or there may be a need to explain how the decision was made.

- **Example**: Might need to provide reasoning for intervention for certain inmates for legal reasons.

Motivation for Interpretable ML

Interpretable ML techniques are useful when there is a **need to explain how an ML model reached a prediction**.

This situation is quite common in the social sciences (due to practical/legal reasons)!

Idea behind Interpretable ML Methods

Main goal is still to make the best predictions possible!

But, sometimes, we also want to get **some interpretability** out of the models.

Use these methods when the goal is to make the best predictions … but you need to have **interpretability for applying the predictions**.

Example Use Case

Machine Learning can be used to make decisions on granting bail by determining “flight risk.”

However, using a “black box” model to make these types of decisions can lead to legal issues.

Interpretable ML techniques help to create simplified decision-making instructions for non-experts.

ML Workflow

Collect data on past cases.

Build ML model to predict “flight risk” cases.

Use interpretable ML techniques to **identify variables/relationships**.

(Possible step) Build a simple surrogate model so that non-technical personnel can **make decisions**.

How to Apply Interpretable ML

This should **NOT be used for inference**!

These methods are designed to be **descriptive rather than inferential.** Think of more like using visualizations to show a distribution rather than making definitive statements.

Other Possible Uses

Understanding the models better can also help with identifying areas for **future study using experiments and actual inferential methods**.

Example: If there is a variable that is surprisingly high in variable importance, it might be worth investigating why that might be the case. However, this could simply be a correlation.

Feature Importance

Feature importance values are all relative. Use them to get an idea of what the most important variables might be.

Note: There is no “unit” for feature importance. These are scores used to compare between variables **within a model**. Don’t compare feature importances across models!

Using the PDP/ICE/ALE

These are all **descriptive methods** that you can use to identify different parts of the relationship.

Analogy: Think about describing a distribution. You might use a histogram, a scatterplot, a boxplot, as well as numerical statistics like mean, median, or standard deviation. In the same way, you can **use all or most** of tools like PDP, ICE, ALE, variable importance, etc. **together** to describe the model.

PDP vs. ALE: Global vs. Local

Global means we are looking at the range of all values when trying to interpret the relationship.

Local means that we are only look at a small range of values.

Note: This can make a difference if there are highly correlated values.

- For example, if we have age and income as predictors, we're not likely to have someone who is 10 years old with an income of $200,000.

Surrogate Model

Idea behind surrogate model: We want to build an easily interpretable model that gives predictions that are almost the same as the complicated model.

Notes

- This is for interpretation! Use as much data as you have predictions for.
- The model should be something that is **easily interpretable**. For example, OLS regression.

Using Surrogate Models

Don’t try to interpret coefficients/significance in the same as you would a normal OLS regression model. Remember, **we are not using this to do inference**, just to make machine learning models easier to understand.

Similarly, don’t need to worry about representation and weights as much when using these (particularly true for local compared to global).

Global vs. Local for Surrogate Models

Overall, machine learning models can be quite complex. However, **they might be much simpler on a local level** (over a small range of x values, for example).

Idea behind LIME: Use simple models to explain at local levels.