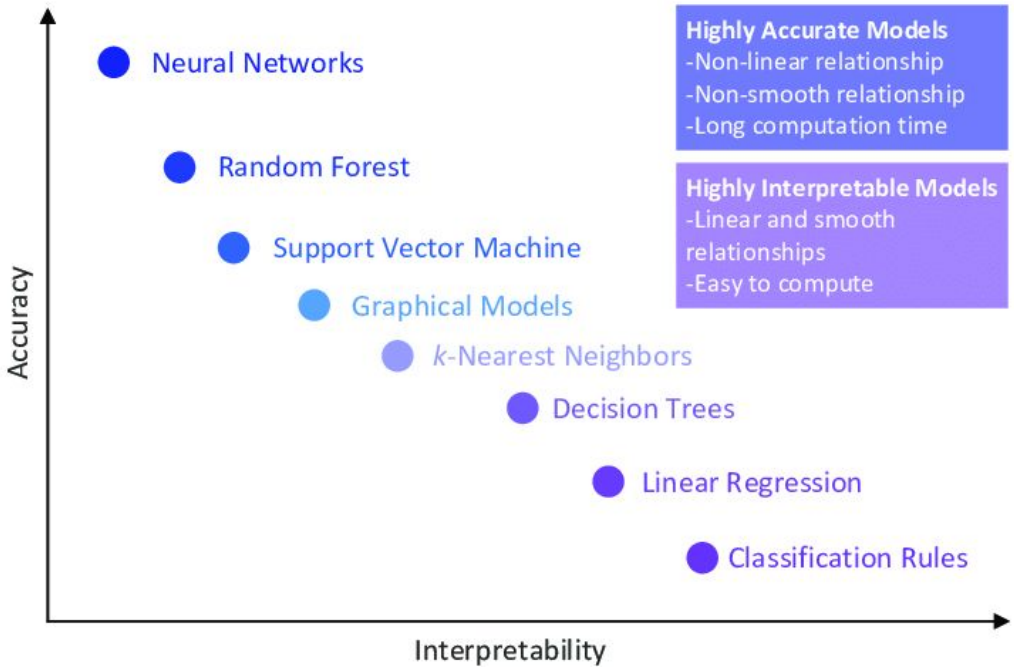
Opening Up/Explaining Black Box Models

LIME - Local Interpretable Model-Agnostic Explanations

@ravibapna

Tension: Interpretability vs Performance

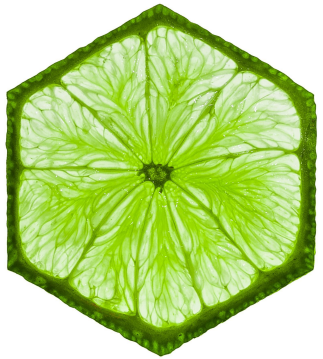


Local Interpretable Model-agnostic Explanations

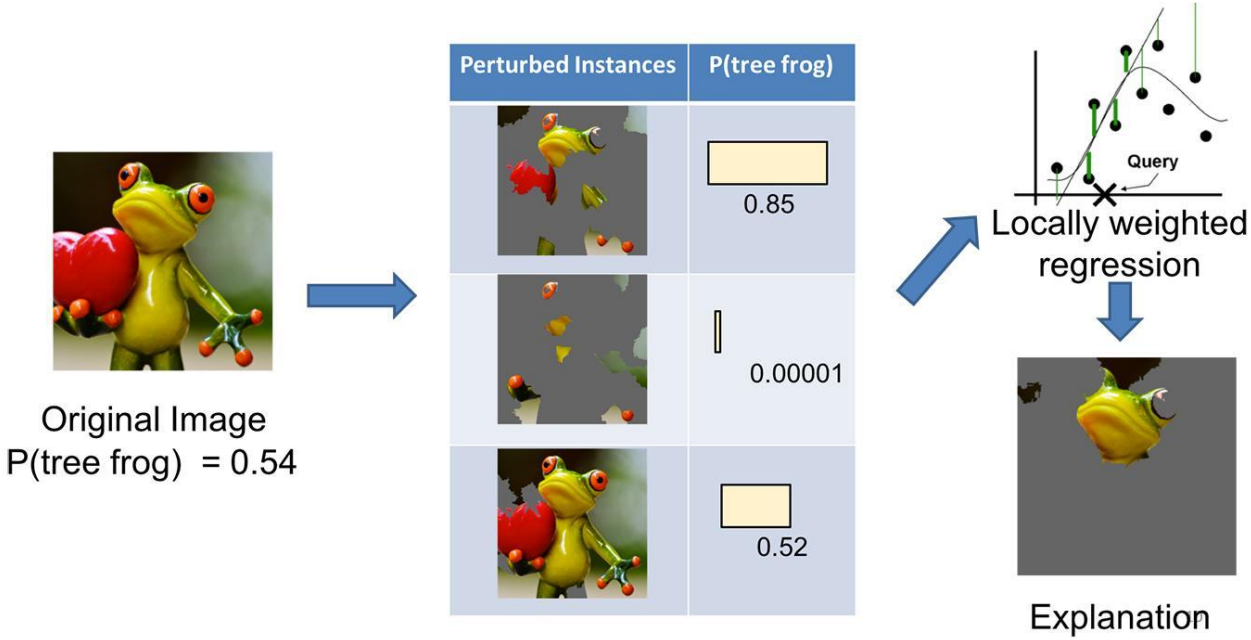
Procedure – Local Perturbations of the Data:

- We assume that a features' contributions to the prediction can be locally approximated by a linear regression. Then:

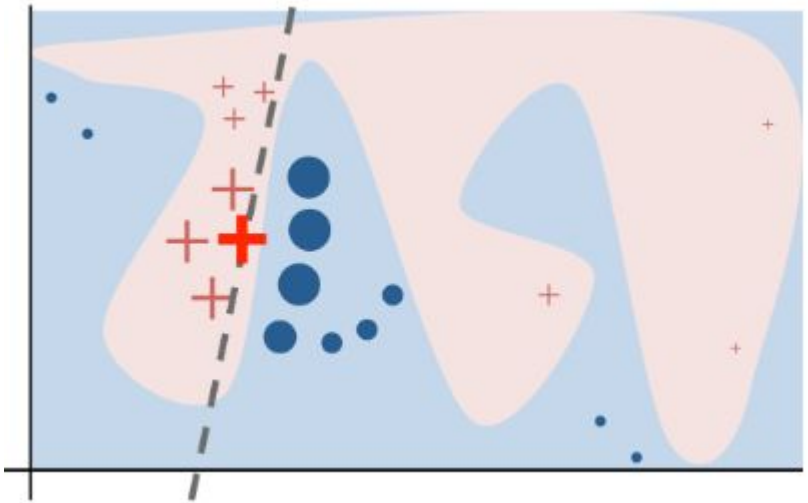
1. For a given prediction, **randomly perturb the observation** (modify its feature values), repeatedly, and **recover the associated predictions** for each synthetic observation.
2. This procedure **yields an observation-specific dataset**. Use the dataset to **estimate a weighted linear regression** on the dataset, weighting observations inversely by their distance / dissimilarity from the original observation.
3. **Beta coefficients** reflect features' localized marginal contributions to the prediction.



Local Interpretable Model-agnostic Explanations



The original model's decision function is represented by the blue/pink background, and is clearly nonlinear



1. The bright red cross is the instance being explained (let's call it X)
2. We sample perturbed instances around X, and weight them according to their proximity to X (weight here is represented by size)
3. We get original model's prediction on these perturbed instances, and then learn a linear model (dashed line) that approximates the model well in the vicinity of X
4. Note that the explanation in this case is not faithful globally, but it is faithful locally around X

Questions?