Bringing a Ruler Into the Black Box: Uncovering Feature Impact from Individual Conditional Expectation (ICE) Plots

NYU

Andrew Yeh
Wharton/NYU

Anhthy Ngo
MITRE/NYU

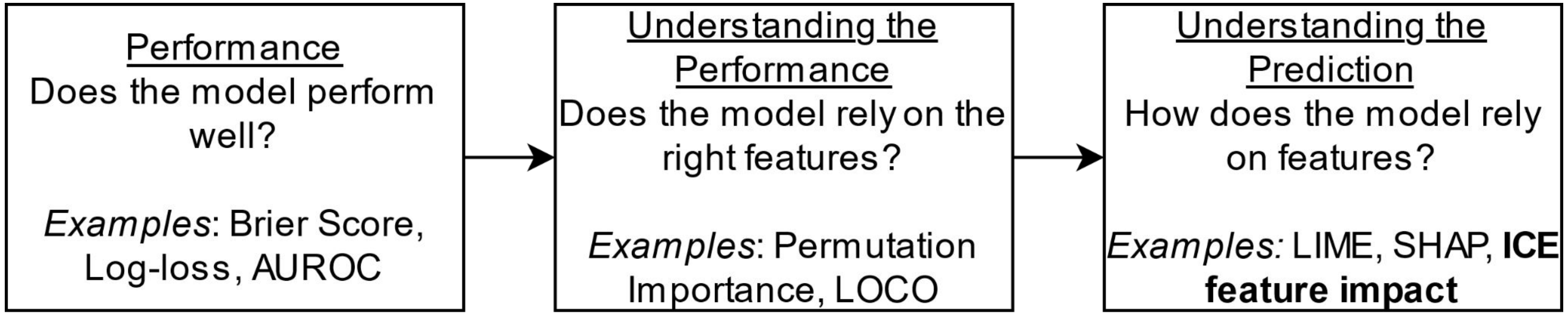
Wharton UNIVERSITY of PENNSYLVANIA

MITRE

aimlai@ecmlpkdd'21

Introduction

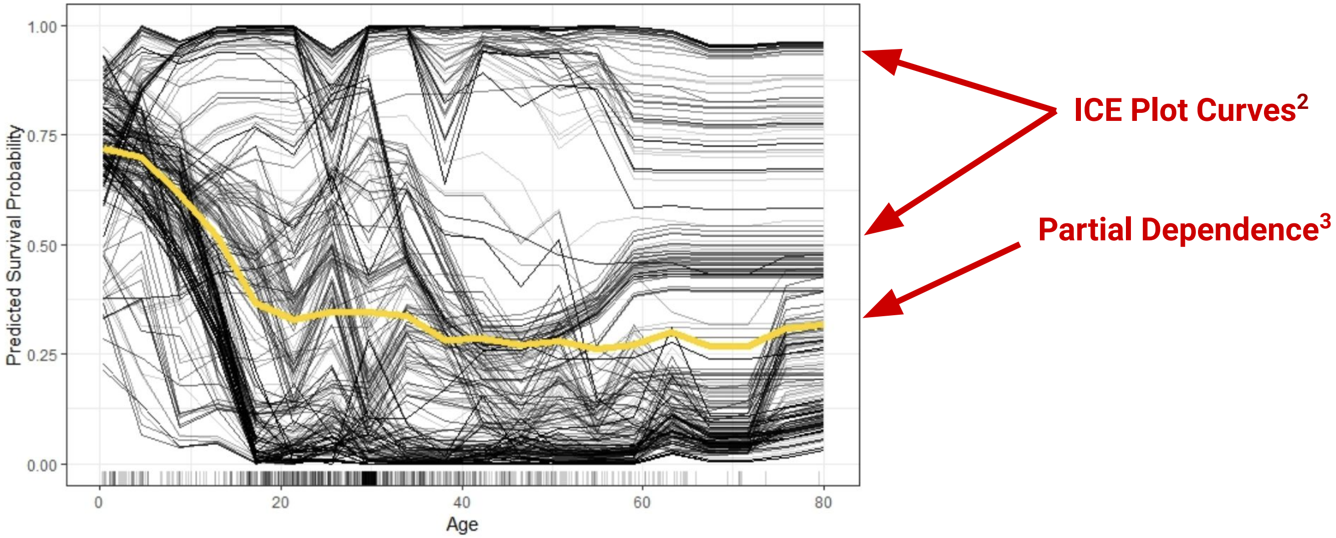
- ICE Feature Impact is an extension of ICE plots that provides a quantitative measure of how features impact predictions.
- A model-agnostic and performance agnostic “feature impact” metric in contrast to feature importance metrics.[1]
- Highly intuitive “linear regression”-like coefficients, complementary to existing metrics, and with additional depth beyond a point estimate



[1] Parr, T., Wilson, J.D., Hamrick, J.: Nonparametric feature impact and importance. arXiv preprint arXiv:2006.04750 (2020)

Background: Visual Feature Impact Tools

ICE plots provide visual tools to understand feature impact from models.[1]



Advantages: Intuitive, efficient at conveying information

Disadvantages: Imprecise, overcrowded, does not scale to large number of features

[1] Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24 (09 2013)
[2] Plot taken from Limitations of Interpretable Machine Learning Methods by Altmann et. al (2020).
[3] Friedman, J.H.: Greedy function approximation: A gradient boosting machine.The Annals of Statistics 29(5), 1189 – 1232 (2001)

ICE Feature Impact

Feature impact: Feature impact is how strongly a change in the feature impacts the prediction. We quantify it as rise-over-run: change in model prediction divided by the corresponding change in the feature.

$$\mathbf{FI}(\mathbf{x}_S) = \frac{\sigma_{\mathbf{x}_S}}{n \cdot (n_{\mathbf{x}_S} - 1)} \sum_{i=1}^{n} \sum_{k=2}^{n_{\mathbf{x}_S}} \left| \frac{d\hat{y}(x^{(i)}[k])}{dx_S^{(i)}[k]} \right|$$

$$\approx \frac{\sigma_{\mathbf{x}_S}}{n \cdot (n_{\mathbf{x}_S} - 1)} \sum_{i=1}^{n} \sum_{k=2}^{n_{\mathbf{x}_S}} \left| \frac{\hat{y}(x^{(i)}[k]) - \hat{y}(x^{(i)}[k-1])}{x_S^{(i)}[k] - x_S^{(i)}[k-1]} \right|$$

Phantom observation: A phantom observation corresponds to a real observation when all not at-issue features are equal to their values in the real observation, but the at-issue feature is permuted. Each phantom observation is a point in the ICE plot, and all the phantom observations for one real observation are one line.

ICE Feature Impact: The average feature impact over all phantom observations that correspond to an observation and all observations.

In-Distribution ICE Feature Impact

A weakness of ICE plots and permuting features to interrogate the model is that it ignores likelihood and feature correlations.

$$\mathbf{IDFI}(\mathbf{x}_S) \approx \frac{\sigma_{\mathbf{x}_S}}{\sum_{i=1}^{n} \sum_{k=2}^{n_{\mathbf{x}_S}} L_{\mathbf{x}_S}} \sum_{i=1}^{n} \sum_{k=2}^{n_{\mathbf{x}_S}} L_{\mathbf{x}_S}(x^{(i)}[k]) \left| \frac{\hat{y}(x^{(i)}[k]) - \hat{y}(x^{(i)}[k-1])}{\mathbf{x}_S^{(i)}[k] - x_S^{(i)}[k-1]} \right|$$

$$L_{\mathbf{x}_S}(x^{(i)}[k]) = \lambda^{\frac{|\mathbf{x}_S^{(i)}[k] - \mathbf{x}_S^{(i)}|}{\sigma_{\mathbf{x}_S}}} \qquad \lambda \in (0, 1]$$

In-distribution ICE Feature Impact addresses this issue by weighting the phantom observations and self-normalizing by likelihood. Accepts arbitrarily complicated likelihood functions.

Heterogeneity and Non-Linearity

Heterogeneity: The degree to which the pattern of ICE curves varies across observations, i.e. the feature impact is heterogeneous when its impact is higher on some observations and lower on others.

$$\mathbf{HE}(\mathbf{x}_S) = \frac{\sigma_{\mathbf{x}_S}}{n_{\mathbf{x}_S}} \sum_{k=1}^{n_{\mathbf{x}_S}} SD_{i \in \{1,\ldots,n\}} \left(\frac{\hat{y}(x^{(i)}[k]) - \hat{y}(x^{(i)}[k-1])}{x_S^{(i)}[k] - x_S^{(i)}[k-1]} \right)$$

Non-linearity: The degree to which features have a non-linear relationship with the model's predictions, i.e. how much the impact of a feature varies across the feature's support.

$$\mathbf{NL}(\mathbf{x}_S) = \frac{\sigma_{\mathbf{x}_S}}{n} \sum_{i=1}^{n} SD_{k \in \{1,\ldots,n_{\mathbf{x}_S}\}} \left(\frac{\hat{y}(x^{(i)}[k]) - \hat{y}(x^{(i)}[k-1])}{x_S^{(i)}[k] - x_S^{(i)}[k-1]} \right)$$

Experiment with Cancer Dataset

- Data: 32 features with “Biopsy” as binary target feature
- Model: Random Forest to take advantage of native feature importance metrics for comparison

Selected Findings

- ICE Feature Impact exhibits low correlation with alternative metrics, e.g. permutation feature importance and Tree SHAP
- ICE Feature Impact has perfect correlation with linear regression coefficients (magnitude only) and is strongly correlated with the underlying coefficients for pseudo-linear models, e.g. Logistic Regression and SVM
- For non-linear models, we have measures for heterogeneity and non-linearity

Conclusion

ICE Feature Impact is a highly interpretable measure of feature impact drawn out from ICE plots.

- Model and performance agnostic: uncovers the “linear regression coefficients” analogy for any black-box model
- Measures feature impact complementary to feature importance
- Extended with in-distribution version, heterogeneity, and non-linearity

Use ICE Feature Impact to build model trust in how the model arrives at its predictions from the features.

Questions?