

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

Andrew Yeh^{1,2} and Anhty Ngo^{1,3}

¹ New York University, New York NY 10011, USA {ay1626, an3056}@nyu.edu

² The Wharton School of the University of Pennsylvania, Philadelphia PA 19104, USA
ayeh21@upenn.edu

³ The MITRE Corporation, McLean VA 22102, USA ango@mitre.org

Abstract. As machine learning systems become more ubiquitous, methods for understanding and interpreting these models become increasingly important. In particular, practitioners are often interested both in what features the model relies on and how the model relies on them – the feature’s impact on model predictions. Prior work on feature impact including partial dependence plots (PDPs) and Individual Conditional Expectation (ICE) plots has focused on a visual interpretation of feature impact. We propose a natural extension to ICE plots with ICE feature impact, a model-agnostic, performance-agnostic feature impact metric drawn out from ICE plots that can be interpreted as a close analogy to linear regression coefficients. Additionally, we introduce an in-distribution variant of ICE feature impact to vary the influence of out-of-distribution points as well as heterogeneity and non-linearity measures to characterize feature impact. Lastly, we demonstrate ICE feature impact’s utility in several tasks using real-world data.

1 Introduction

As machine learning (ML) systems become more ubiquitous in human decision making, transparency and interpretability have grown significantly in importance [14]. Some models may not require user trust due to a low-risk nature, e.g. movie recommendation systems. Other problems don’t require top performance and safely rely on highly interpretable models that may not perform as well as black box models. However, when a problem space combines a high risk nature with demands for superior performance, earning the user’s trust in the model is essential.

We distinguish three phases to “trusting” a model: strong performance, model understanding, and prediction understanding (See Figure 1). To distinguish a feature’s contribution to model performance from its contribution to model predictions, we call the former “feature importance” and the latter “feature impact” [11].

* We thank David Rosenberg for his insightful feedback, engagement, and advice in shaping this project from proposal to paper as well for introducing us to the original ICE paper. We thank Anu-Ujin Gerelt-Od without whom this project would not be possible. Last but not least, we thank Lee Kho for her valuable input, ideas, and support.

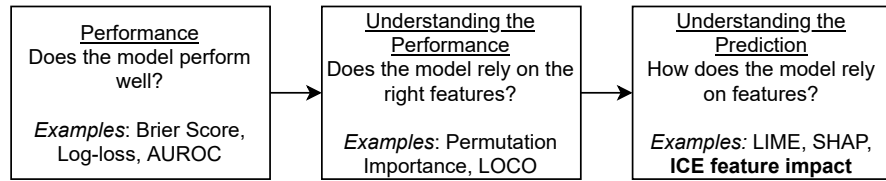


Fig. 1: Three stages in model trust

There exist several visual methods to display feature impact, the relationship between features and predictions, most notably partial dependence plots (PDPs) [6] and individual conditional expectation (ICE) plots [7]. PDPs aggregate the effects of a feature while ICE plots disaggregate divergent effects by plotting individual observations.

Visual tools are highly intuitive and can convey a lot of information in a single plot. However, they have some weaknesses as well. Firstly, visual interpretation is imprecise which makes comparison between features difficult. Secondly, ICE plots in particular can only plot a subset of the observations in the dataset to avoid overcrowding, which can hide outlier observations or overfit extrapolations from view. Thirdly, the cost of visual inspection does not scale well to the number of features—visually inspecting the plots for millions of features, for example, is infeasible.

In this paper, we address these issues and extend ICE plots by extracting feature impact metrics from them (“ICE feature impact”). ICE feature impact is model- and performance-agnostic, meaning it measures the impact of each feature on the prediction only, without regarding the accuracy of that prediction. ICE feature impact also addresses the issues with the visual approach discussed above: it is a precise metric, allowing comparisons between different ICE plots; it takes into account every observation, including outliers, instead of only a subset; and it can be ranked to prioritize inspection of ICE plots to only the most impactful features, allowing the usefulness of ICE plots to scale with the number of features.

We also introduce an in-distribution version of feature impact with a hyperparameter to reduce the influence of out-of-distribution points, and we supplement ICE feature impact with measures of heterogeneity and non-linearity to add depth. Together, these metrics provide a quantitative perspective for understanding feature impact complementary to the qualitative nature of inspecting ICE plots.

2 Related Work

First introduced by Friedman [6], partial dependence plots (PDPs) are a model and performance agnostic method of illustrating the relationships between one or more input variables and the predictions of a black-box model. PDPs estimate the partial dependency by marginalizing over all other features – essentially permuting the at-issue features to specific values across the observed range and then averaging the resulting predictions across training observations.

Individual Conditional Expectation (ICE) [7] plots disaggregate the average feature impact curve of PDPs into its component, individual observation-curves. This allows

ICE plots to capture heterogeneous relationships that PDPs otherwise miss. We further discuss ICE plots and provide a specific methodology in Section 3.1. Accumulated Local Effects [1] extend PDPs by restricting the permutation of at-issue features within a certain interval as opposed to allowing them to permute from the minimum and maximum possible values as PDPs and ICE plots do. This addresses a weakness in PDPs and ICE plots that permuting the feature value can lead to unrealistic observations when features are correlated and motivates in-distribution ICE feature impact.

Parr et al. [11] distinguishes the idea of “feature impact” from standard feature importance metrics as follows: while feature importance metrics measure how important a feature is to the model’s performance, feature impact metrics measure how variations in feature values impact the prediction, irrespective of performance.

LIME [12] uses an interpretable surrogate model to approximate the feature impact on a local scale around the prediction. Parr et al. [11] proposes a non-parametric feature impact methodology that does not interrogate a fitted model. Instead, they extend the concept of PDPs by calculating the empirical partial dependence of the prediction on the at-issue feature based on the data and then approximating the area under the resulting partial dependence curve with a Riemann’s Sum.

Shapley values [13] detail how to fairly determine the total contribution of each feature to the overall prediction—making it a feature impact metric—by taking into account both a feature’s individual contribution and collaborative contribution together with all possible subsets of features. Shapley values themselves are highly computationally expensive to calculate precisely, though they can be approximated with a Monte Carlo approach [15], Kernel SHAP [10], or Tree SHAP [9]. Tree SHAP differs from other approaches as it relies solely on the training data without interventionist means like permuting the value of features.

3 Methodology

An implementation of ICE feature impact as described below is available in Github.⁴

3.1 ICE Plot Replication

We establish terminology and notation for the remainder of the paper by detailing the ICE replication methodology we use. To replicate ICE plots, we create “phantom observations” from each “real observation” where all not “at-issue features(s)” are constant, but we permute the “at-issue feature(s)”. We then use the phantom observations to interrogate the model.

The exact algorithm is as follows: for at-issue feature(s) \mathbf{x}_S , fitted model \hat{f} , and feature matrix $\mathbf{X} \in \mathbb{R}^{N \times p}$, let there be $n_{\mathbf{x}_S}$ unique values of \mathbf{x}_S found in the data.

1. For each observation $x^{(i)}$, create $n_{\mathbf{x}_S}$ observations with all features the same as in $x^{(i)}$, except for \mathbf{x}_S . Replace \mathbf{x}_S with the $n_{\mathbf{x}_S}$ unique values of feature p found above. This results in $n_{\mathbf{x}_S}$ new observations for each $x^{(i)}$.

⁴ https://github.com/mixerupper/mltools-fi_cate

2. We call the resulting observations “phantom observations”, denoted $x^{(i)}[k]$ which is the k th phantom observation for $x^{(i)}$ with $k = 1, \dots, n_{\mathbf{x}_S}$. For each observation $x^{(i)}$, one of its phantom observations is exactly identical to $x^{(i)}$, and the others are identical except for a permuted \mathbf{x}_S . Combine all $n \cdot n_{\mathbf{x}_S}$ phantom observations into a new feature matrix.
3. Use fitted model \hat{f} to predict \hat{y} for all phantom observations.
4. For each original observation, plot a line composed of the corresponding phantom points with the at-issue feature on the x-axis and \hat{y} on the y-axis. This results in n lines, with each line composed of $n_{\mathbf{x}_S}$ phantom points.

Additionally, if n is large, we sample uniformly from each quantile of \mathbf{x}_S if \mathbf{x}_S is continuous and each value of \mathbf{x}_S if \mathbf{x}_S is categorical to capture the whole distribution.

3.2 ICE Feature Impact

While ICE plots allow visual inspection of feature impact, it does not output any quantitative metrics for comparability. We elicit a numeric feature impact metric from ICE plots in the form of ICE feature impact.

For the sequence of points that make up each observation-curve, we calculate the absolute change in prediction divided by the change in feature ($|\frac{dy}{dx}|$) for each consecutive point. This uses rise over run to quantify the impact of the feature on the prediction value. Then, ICE feature impact is the mean of all the $|\frac{dy}{dx}|$ terms over all phantom points that make up an observation and all observations. To account for features of different scales, we multiply by the standard deviation of that feature. We will see that ICE feature impact has an analogous interpretation to coefficients in a linear model.

The exact algorithm is as follows: for feature \mathbf{x}_S , let $\sigma_{\mathbf{x}_S}$ denote the standard deviation of \mathbf{x}_S , let n be the number of observations, $n_{\mathbf{x}_S}$ be the number of unique values of \mathbf{x}_S , $x^{(i)}$ be the i th observation, $x^{(i)}[k]$ be the k th phantom observation corresponding to $x^{(i)}$, $\mathbf{x}_S^{(i)}$ be the value of \mathbf{x}_S in observation $x^{(i)}$, $\mathbf{x}_S^{(i)}[k]$ be the value of \mathbf{x}_S in the k th phantom observation corresponding to $x^{(i)}$, and \hat{y} be the fitted model. Thus, the **ICE feature impact** is:

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

The ICE feature impact of \mathbf{x}_S can be interpreted as the absolute change in the predicted value of \hat{y} for each one-unit change in \mathbf{x}_S if \mathbf{x}_S was standardized to a standard deviation of 1 and all other features remained constant. Note that ICE feature impact gives the magnitude of impact, not the direction. Average direction of feature impact can be determined by comparing the ICE feature impact with the value of Equation 1 without an absolute value on the inner summation term.

3.3 In-Distribution ICE Feature Impact

One of the drawbacks of ICE feature impact as introduced in Section 3.2 is that it weights evenly across all phantom points, no matter their likelihood of occurrence in the true feature distribution. This may be concerning if features are highly correlated, and permuting the at-issue feature \mathbf{x}_S takes us out of the feature distribution, e.g., taking the health data from a 9 year old and changing the age to 70 while leaving the other features untouched would give us a phantom observation that has a low likelihood of occurring in reality.

This is a missing data problem with the missing value being the likelihood of the observation. The likelihood is 1 for all true observations and missing for all phantom observations. Let us denote the likelihood of phantom observation $x^{(i)}[k]$ for at-issue feature \mathbf{x}_S with $L_{\mathbf{x}_S}(x^{(i)}[k])$. Then, given this likelihood, the in-distribution ICE feature impact of \mathbf{x}_S is:

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

To estimate $L_{\mathbf{x}_S}(x^{(i)}[k])$, we model likelihood as exponentially decaying with respect to the absolute distance of the at-issue feature’s permutation divided by the feature’s standard deviation:

$$L_{\mathbf{x}_S}(x^{(i)}[k]) = \lambda \frac{|\mathbf{x}_S^{(i)}[k] - \mathbf{x}_S^{(i)}|}{\sigma_{\mathbf{x}_S}} \quad (3)$$

where $0 < \lambda \leq 1$ is a hyperparameter that measures how quickly the weight decays as the phantom feature value differs from the real feature value. Note that $\lambda = 1$ gets us back to ICE feature impact without out-of-distribution considerations.

We can estimate $\sigma_{\mathbf{x}_S}$ as the sample standard deviation of \mathbf{x}_S in the data or as an arbitrarily sophisticated estimate of the standard deviation for the at-issue feature based on the value of all other features for the observation. For example, [5] proposes estimating the conditional distribution of a feature based on all other features using a pseudo-maximum likelihood problem estimated via a single self-attention architecture.

The in-distribution ICE feature impact weights phantom observations closer to the real observation more heavily when measuring feature impact, giving us a perspective on feature impact that is more “true to the data” [4].

4 Real Data

To examine ICE feature impact, we use UC Irvine’s cervical cancer risk factors dataset.⁵ The dataset contains medical information for 858 patients from *Hospital Universitario de Caracas*. There are 32 features including age, number of pregnancies, and use of IUD. The target variable is `Biopsy`, which is binary.

⁵ Cervical Cancer (Risk Factors) Data Set contains a detailed description of the dataset.

4.1 Complementary to Feature Importance

First, we show that ICE feature impact presents an additional dimension to understanding models beyond feature importance.

We train a random forest classifier [2] on the dataset.⁶ We then calculate the following metrics for each feature: ICE feature impact, Tree SHAP [9], Random Forest feature importance [3], and permutation feature importance [2] and normalize them to be positive and sum to 100. We take the correlation between ICE feature impact and alternative metrics and find that the correlation is low (See Table 1). This indicates that ICE feature impact differs substantially from alternatives instead of fulfilling the same function.

Metric	Correlation w/ ICE FI
In-Distribution ICE FI ($\lambda = 0.75$)	0.99
Random Forest Feature Importance	0.36
Permutation Feature Importance	0.35
Tree SHAP Values	0.17

Table 1: Pearson correlation of feature importance and impact metrics with ICE feature impact. All metrics were first normalized to sum to 100. Tree SHAP values were additionally first made positive to remove direction before normalizing to sum to 100.

Table 2 shows the features with the two most positive differences and the features with the two most negative differences between their random forest feature importance and ICE feature impact values.⁷ While `Age` and `Number of Sexual Partners` are highly predictive features and are helpful in reducing impurity of classification, they do not have a strong impact on the model’s predictions itself. On the opposite end of the spectrum, `STDs:molluscum contagiosum` and `STDs:pelvic inflammatory disease` have highly imbalanced feature distributions with the majority of values equal to 0 and therefore are not as helpful for reducing impurity. However, when these factors are present – specifically, when the value is missing and the mean is imputed – they contribute strongly to the model prediction, explaining the higher feature impact.

4.2 Interpretability: Analogous to Linear Regression Coefficients

In the base case of analyzing a linear regression model, ICE feature impact values are exactly the absolute value of the linear regression coefficients. We also calculated ICE feature impact for the pseudo-linear models of Logistic Regression and linear SVMs. Table 3 shows that the resulting model coefficients are strongly correlated with the corresponding ICE feature impact values.

⁶ We use the `sklearn` package with parameters of 500 trees, a random state seed of 20, and the default values for the remaining parameters. As this exercise is about model interpretability, we did not tune the model to improve performance.

⁷ See Appendix C for the full feature impact table, Appendix A for the ICE plots for all features, and Appendix B for the centered ICE plots (c-ICE) [7] for all features.

Feature	ICE FI	Native Feature Importance	Difference
STDs:molluscum contagiosum	9.8	0.1	9.6
STDs:pelvic inflammatory disease	8.9	0.1	8.8
Number of sexual partners	1.0	9.9	-8.9
Age	3.4	17.8	-14.3

Table 2: Feature impact table for features in cervical cancer dataset with two largest and most negative difference between Random Forest feature importance and ICE Feature Impact.

Model	ICE Feature Impact	
	Base	In-Dist
Linear Regression	1	1
Logistic	0.73	0.8
SVM	0.9	0.98

Table 3: Pearson correlation of ICE feature impact values with absolute value of coefficients of linear and pseudo-linear models.

These results show that ICE feature impact can be interpreted analogously to linear regression coefficients with features standardized to a unit standard deviation.

4.3 Quantifying Heterogeneity and Non-Linearity

In linear models, knowing feature impact means knowing exactly where predictions come from. In non-linear models, however, the relationship between features and the model prediction can be more complex: in particular, the relationship can be heterogeneous – different across observations – or non-linear – different across the feature’s support. We propose measures of heterogeneity and non-linearity to allow the practitioner a more nuanced understanding of ICE feature impact.

Let heterogeneity be 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. Then, following the notation described in Section 3.1, the heterogeneity of feature \mathbf{x}_S is:

$$\mathbf{HE}(\mathbf{x}_S) = \frac{\sigma_{\mathbf{x}_S}}{n_{\mathbf{x}_S}} \sum_{k=1}^{n_{\mathbf{x}_S}} SD_{i \in \{1, \dots, n\}} \left(\frac{\hat{y}(x^{(i)}[k]) - \hat{y}(x^{(i)}[k-1])}{x_S^{(i)}[k] - x_S^{(i)}[k-1]} \right) \quad (4)$$

where the standard deviation is taken for fixed k across all real observations. The lower the heterogeneity metric, the more similar the shape of observation-curves are at each point. For linear regressions and additive models like GAM [8], the heterogeneity metric is zero since the effect of a feature on the prediction is the same across all observations.

Let non-linearity be the degree to which features have a non-linear relationship with the model’s predictions, i.e. how much the effect of a feature varies across the support

for a given observation. For features with low non-linearity, the corresponding ICE feature impact can be interpreted as close to a linear regression coefficient, even if the underlying model is non-linear. We quantify non-linearity as follows:

$$\mathbf{NL}(\mathbf{x}_S) = \frac{\sigma_{\mathbf{x}_S}}{n} \sum_{i=1}^n SD_{k \in \{1, \dots, n_{\mathbf{x}_S}\}} \left(\frac{\hat{y}(x^{(i)}[k]) - \hat{y}(x^{(i)}[k-1])}{x_S^{(i)}[k] - x_S^{(i)}[k-1]} \right) \quad (5)$$

where the standard deviation is taken for fixed i across all corresponding phantom observations. For linear regressions, the non-linearity is equal to 0 as desired since the effect of a feature is constant across the feature’s support.

Table 4 shows the heterogeneity and non-linearity of the features listed in Table 2.⁸ Note that the features with the largest positive differences between feature impact and feature importance have higher heterogeneity but similar non-linearity compared to the features with the largest negative differences. This is because ICE feature impact captures heterogeneity through taking the absolute value of the feature impact $\frac{dy}{dx}$ units but does not discriminate between non-linear or linear relationships.

Feature	Feature Impact	Heterogeneity	Non-Linearity
STDs:molluscum contagiosum	9.8	0.27	0.19
STDs:pelvic inflammatory disease	8.9	0.23	0.17
Number of sexual partners	1.0	0.05	0.04
Age	3.4	0.11	0.18

Table 4: ICE feature impact, heterogeneity, and non-linearity for features in cervical cancer dataset with the two most positive and most negative differences between Random Forest feature importance and ICE Feature Impact.

5 Discussion

Building upon efforts to interpret machine learning models, we extend ICE plots by drawing out ICE feature impact, a measure of the relationship between features and model predictions. ICE feature impact is uncorrelated with alternative feature importance metrics, highlighting features that are impactful to predictions but do not contribute as strongly to model performance. It has a highly interpretable form and is analogous to linear regression coefficients.

We also propose in-distribution ICE feature impact to downweight out-of-distribution observations and the heterogeneity and non-linearity measures that add dimensionality to our characterization of ICE feature impact.

Altogether, ICE feature impact provides a different perspective from traditional feature importance methods, complements ICE plots, and serves as an alternative to SHAP values in understanding where a model’s predictions come from.

⁸ See Appendix D for heterogeneity and non-linearity for all features.

Bibliography

- [1] Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(4), 1059–1086 (2020)
- [2] Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001)
- [3] Breiman, L.: Manual on setting up, using, and understanding random forests v3. Statistics Department University of California Berkeley 1, 1–58 (2002)
- [4] Chen, H., Janizek, J.D., Lundberg, S., Lee, S.: True to the model or true to the data? *arXiv preprint arXiv:2006.16234* (2020)
- [5] Fakoor, R., Mueller, J., Erickson, N., Chaudhari, P., Smola, A.J.: Fast, accurate, and simple models for tabular data via augmented distillation. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) *Advances in Neural Information Processing Systems 33. NeurIPS* (2020)
- [6] Friedman, J.H.: Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29(5), 1189 – 1232 (2001)
- [7] 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)
- [8] Hastie, T., Tibshirani, R.: *Generalized Additive Models*. *Statistical Science* 1(3), 297 – 310 (1986)
- [9] Lundberg, S.M., Erion, G.G., Lee, S.: Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888* (2018)
- [10] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017)
- [11] Parr, T., Wilson, J.D., Hamrick, J.: Nonparametric feature impact and importance. *arXiv preprint arXiv:2006.04750* (2020)
- [12] Ribeiro, M.T., Singh, S., Guestrin, C.: “Why should I trust you?”: Explaining the predictions of any classifier. In: *KDD*. pp. 1135–1144. ACM (2016)
- [13] Shapley, L.: A value for n-person games. In: Kuhn, H., Tucker, A. (eds.) *Contributions to the Theory of Games*, vol. 2. Princeton Press, Princeton, NJ (1953)
- [14] Varshney, K.R.: Engineering safety in machine learning. In: *ITA*. pp. 1–5. IEEE (2016)
- [15] Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41, 647–665 (12 2013)

Appendix A ICE Plots for Cervical Cancer Data

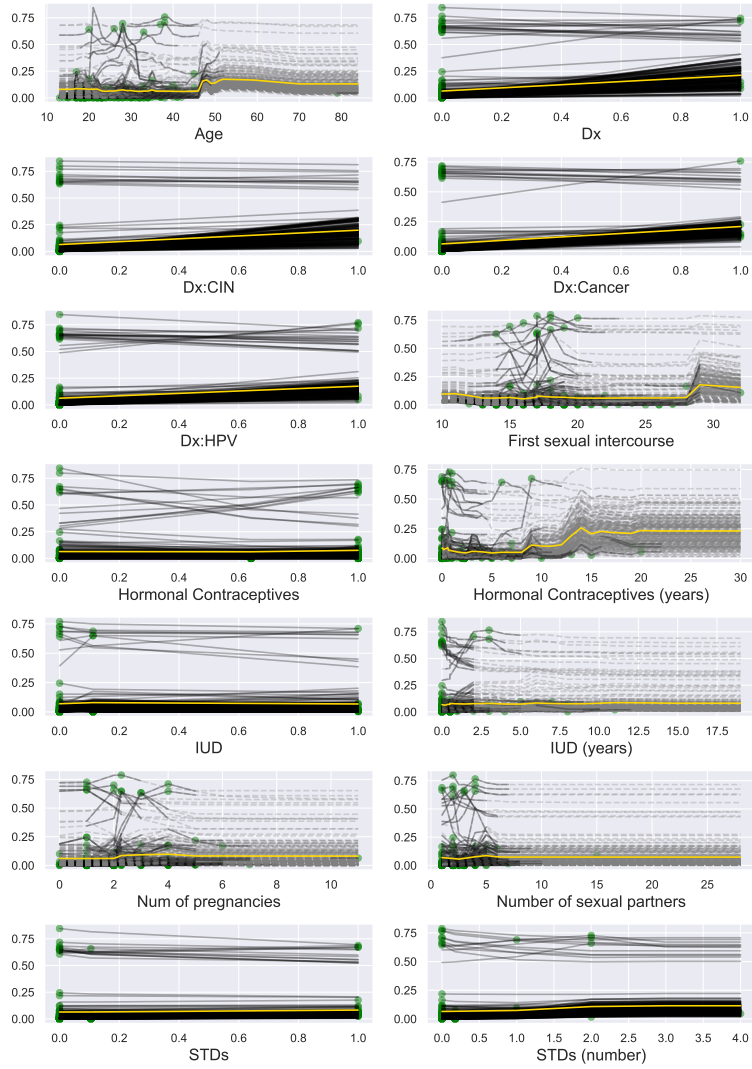
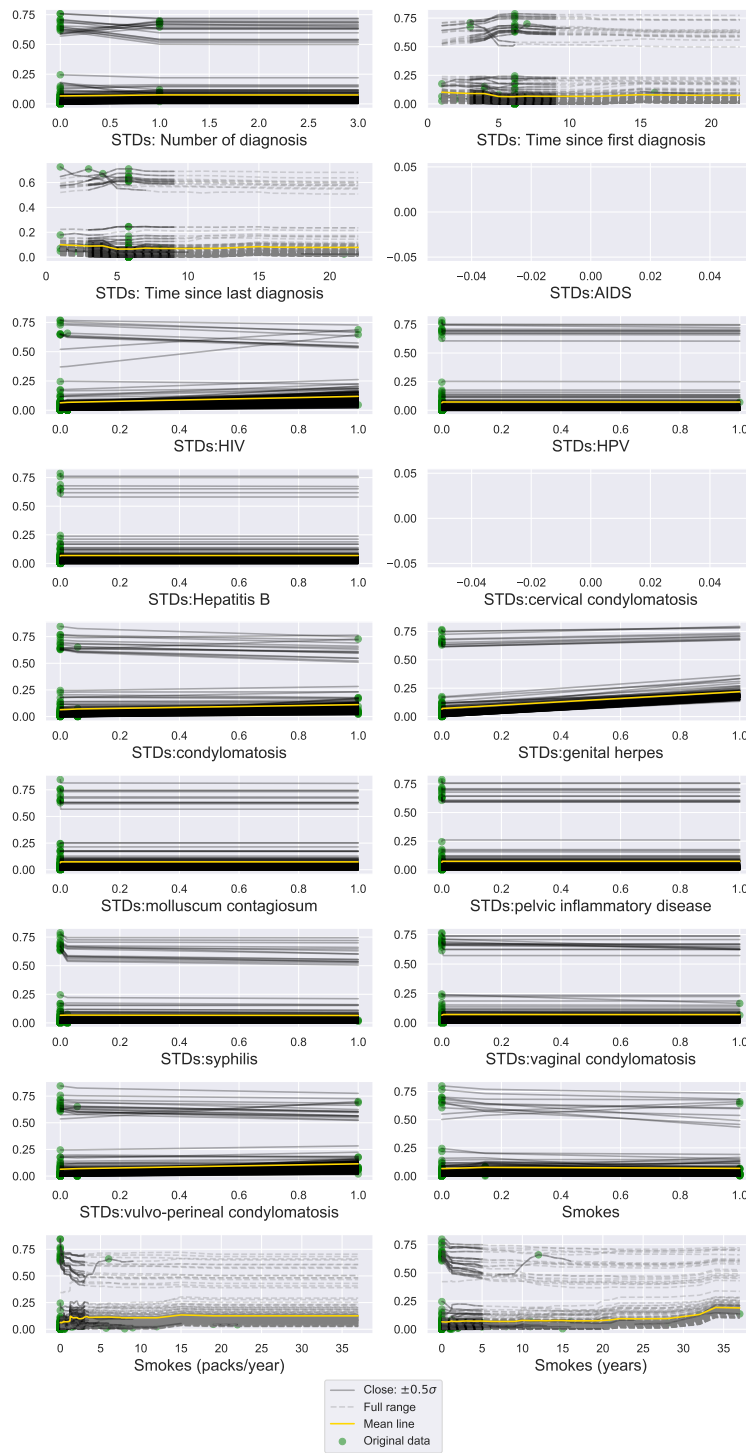


Fig. 2: ICE plots for all features in cervical cancer dataset, following the methodology described in Section 3.1. Each green dot represents a different observation, and the corresponding line shows how varying the observation’s at-issue feature value affects the model’s prediction. Observation-lines are solid within $\frac{1}{2}$ a standard deviation (of the at-issue feature) and dotted outside that range.



Appendix B c-ICE Plots for Cervical Cancer Data

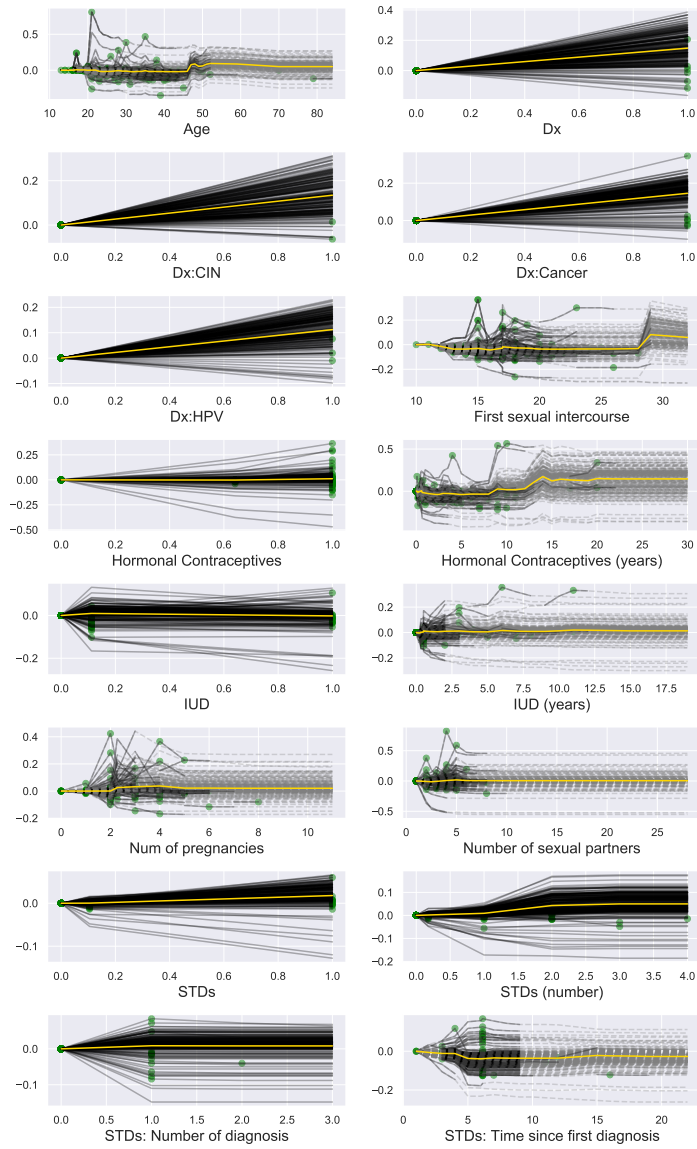
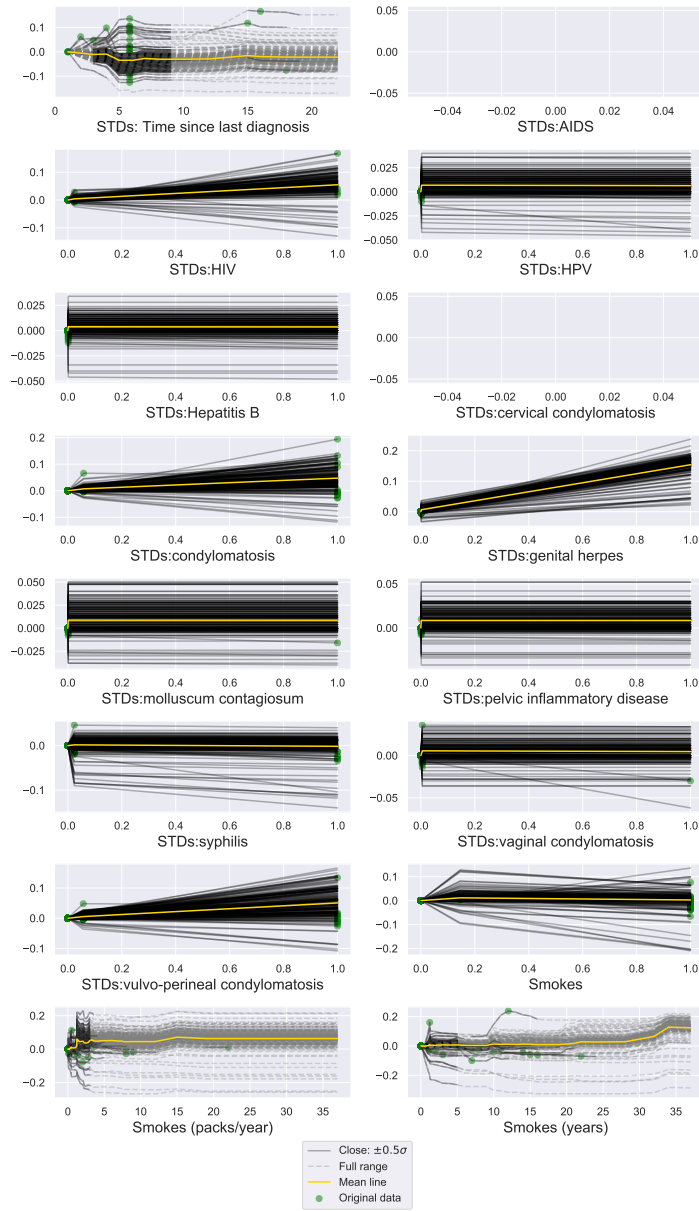


Fig. 3: Centered ICE (c-ICE) plots [7] for all features in cervical cancer dataset. c-ICE plots are equivalent to ICE plots but with the starting \hat{y} value centered to zero such that the lines represent the change in \hat{y} instead of its value.



Appendix C Feature Impact Table for Cervical Cancer Data

Feature	ICE	ICE ID ($\lambda = 0.75$)	Random Forest	Tree SHAP	Permutation
Age	3.4	2.6	17.8	11.3	13.3
Number of sexual partners	1.0	1.1	9.9	6.3	9.9
First sexual intercourse	16.9	16.9	12.3	7.2	12.1
Num of pregnancies	1.3	1.4	10.0	0.0	13.1
Smokes	1.3	1.1	1.4	0.2	1.9
Smokes (years)	1.8	1.8	3.9	6.8	4.6
Smokes (packs/year)	7.0	6.6	3.7	1.9	4.4
Hormonal Contraceptives	0.8	0.6	2.9	10.8	5.1
Hormonal Contraceptives (ye...	9.0	7.8	15.6	7.8	15.9
IUD	2.2	2.0	2.2	3.4	3.0
IUD (years)	3.7	4.0	3.6	4.6	4.6
STDs	0.6	0.4	0.5	0.7	0.1
STDs (number)	0.6	0.6	1.1	1.2	1.0
STDs:condylomatosis	1.2	1.0	0.6	1.8	0.5
STDs:cervical condylomatosis	0.0	0.0	0.0	0.0	0.0
STDs:vaginal condylomatosis	3.8	4.1	0.3	0.8	0.0
STDs:vulvo-perineal condylo...	1.1	0.9	0.6	1.5	0.5
STDs:syphilis	2.1	2.2	0.4	1.1	0.0
STDs:pelvic inflammatory di...	8.9	9.9	0.1	1.3	0.0
STDs:genital herpes	6.8	7.4	0.9	2.4	0.5
STDs:molluscum contagiosum	9.8	10.8	0.1	1.0	0.0
STDs:AIDS	0.0	0.0	0.0	0.0	0.0
STDs:HIV	1.2	1.2	1.1	1.8	1.9
STDs:Hepatitis B	5.7	6.3	0.2	1.5	0.0
STDs:HPV	5.4	5.9	0.1	1.0	0.0
STDs: Number of diagnosis	0.2	0.2	0.7	4.7	0.0
STDs: Time since first diag...	0.3	0.3	2.0	1.2	1.5
STDs: Time since last diagn...	0.3	0.3	1.7	2.7	0.0
Dx:Cancer	1.0	0.7	1.7	3.5	2.1
Dx:CIN	0.6	0.5	1.2	3.9	1.1
Dx:HPV	0.8	0.6	1.6	4.0	1.6
Dx	1.2	0.9	1.7	3.7	1.1

Table 5: Feature impact table for all features in cervical cancer dataset. All feature impact/importance metrics have been made positive and normalized to sum to 100. The ordering of features is as ordered in the original dataset.

Appendix D Heterogeneity and Non-Linearity of ICE Feature Impact for Cervical Cancer Data

Feature	ICE Feature Impact	ICE Heterogeneity	ICE Non-Linearity
Age	0.08	0.11	0.18
Number of sexual partners	0.02	0.05	0.04
First sexual intercourse	0.38	0.59	1.63
Num of pregnancies	0.03	0.05	0.07
Smokes	0.03	0.04	0.02
Smokes (years)	0.04	0.06	0.12
Smokes (packs/year)	0.16	0.32	0.60
Hormonal Contraceptives	0.02	0.04	0.01
Hormonal Contraceptives (ye...	0.20	0.34	0.47
IUD	0.05	0.07	0.04
IUD (years)	0.08	0.12	0.28
STDs	0.01	0.02	0.01
STDs (number)	0.01	0.02	0.01
STDs:condylomatosi	0.03	0.03	0.01
STDs:cervical condylomatosi	0.00	0.00	0.00
STDs:vaginal condylomatosi	0.08	0.11	0.07
STDs:vulvo-perineal condylo...	0.02	0.03	0.01
STDs:syphilis	0.05	0.07	0.04
STDs:pelvic inflammatory di...	0.20	0.23	0.17
STDs:genital herpes	0.15	0.18	0.13
STDs:molluscum contagiosum	0.22	0.27	0.19
STDs:AIDS	0.00	0.00	0.00
STDs:HIV	0.03	0.03	0.02
STDs:Hepatitis B	0.13	0.17	0.11
STDs:HPV	0.12	0.14	0.10
STDs: Number of diagnosis	0.00	0.00	0.00
STDs: Time since first diag...	0.01	0.01	0.02
STDs: Time since last diagn...	0.01	0.01	0.02
Dx:Cancer	0.02	0.01	0.00
Dx:CIN	0.01	0.01	0.00
Dx:HPV	0.02	0.01	0.00
Dx	0.03	0.02	0.00

Table 6: Heterogeneity and non-linearity dimensions for all features in cervical cancer dataset. The raw ICE feature impact is presented without normalization to sum to 100.