Effects of Uncertainty on the Quality of Feature Importance Explanations

Torgyn Shaikhina*
QuantumBlack
torgyn.shaikhina@quantumblack.com

Umang Bhatt*
University of Cambridge
Partnership on AI

Roxanne Zhang
QuantumBlack

Konstantinos Georgatzis
QuantumBlack

Alice Xiang
Partnership on AI
Sony AI

Adrian Weller
University of Cambridge
The Alan Turing Institute

Abstract

Post-hoc feature importance scores are one method for explaining machine learning model outputs. Such explanations are often used by ML practitioners as part of the model selection process, where they select the best model for their task from a set of candidate models. In this paper, we explore the effects of model uncertainty on the quality of these explanations. Specifically, we develop an approach to quantitatively connect the uncertainty in model predictions with explanation quality in terms of (i) variance, (ii) complexity, (iii) monotonicity, (iv) efficiency, and (v) faithfulness of the explanations. We demonstrate that uncertainty in predictions among a set of candidate models propagates to uncertainty in the feature importance explanations, sometimes resulting in arbitrary explanations for a given sample. We conduct experiments across a range of datasets, model types, and feature importance explanation techniques. Our results show that explanation quality is much poorer for out-of-distribution samples compared to in-distribution (i.e., uncertain vs. certain) samples. We also analyze the effect of the number of candidate models and subsample size on measures of feature importance. Overall, our findings suggest that in the presence of uncertainty, current feature importance explanation techniques are unreliable.

1 Introduction

Widespread research interest in explainable machine learning (ML) has resulted in the development of algorithms that provide local, post-hoc explanations for black-box ML models as feature importance scores (Gilpin et al., 2018). As these techniques gained popularity, ML practitioners both in industry and academia began to raise concerns over the reliability of such explanations (Zhang et al., 2019; Lakkaraju, Arsov, and Bastani, 2020). Aimed at fostering trust in the underlying model, inconsistent explanations can undermine trust in ML altogether (Zhang, Liao, and Bellamy, 2020).

Many have observed a welcome surge in interest and effort towards explainable ML among industry practitioners, who increasingly see explainability as being crucial for model adoption (Bhatt et al., 2020). Indeed, in high-stakes analytics and heavily-regulated domains, explanations can be a defining factor between adopting or discarding an otherwise highly predictive model. Such pressure means that ML practitioners have been keen to bring explainability criteria, often endorsed by human domain experts, into model design considerations. Such criteria can be used to select a model among equally well performing candidates based on the feature attributions the ML practitioners or domain experts find most "sensible."

Existing methods for generating feature importance explanations tend to explain a point estimate, which assumes that only **one** suitable model exists. While Slack et al. (2020) quantify uncertainty of a feature importance explanation for a fixed model and fixed training set, we hypothesize that the fixed model assumption is restrictive. In practice, there can be many models that perform equally well on the same training dataset (Breiman et al., 1984). From a Bayesian perspective, there can be multiple models in a posterior that are equally accurate but have drastically different explanations (Fisher, Rudin, and Dominici, 2019).

Inspired by a real use case (see Section 2.1), we explore the effects of uncertainty, induced by varying the training set used, on the resulting feature importance explanations of the model's output. Our contributions are three-fold:

- We propose uncertainty attributions, a method for obtaining uncertainty estimates for feature importance.
- We examine the effect of uncertainty on explanation quality in terms of the (i) variance, (ii) complexity, (iii) monotonicity, (iv) efficiency, and (v) faithfulness.
- We explore how explanations of an ensemble are related to the explanations of each model in the ensemble.

2 Background

2.1 Motivating Example

Grace is a Data Scientist working in the insurance industry.[1] She is developing a propensity score model to predict the likelihood of a claim event for motor insurance customers. Her current prototype is based on a Random Forest classifier with over 500 features to choose from, covering a broad range of demographic indicators, vehicle properties, past history of driving accidents and claims, as well as features engineered from the telemetry data. She is considering reducing the feature space since she observes that, when ranked by native

[1] Identifying information has been removed from this use case.

impurity-based importance, her top 40 features are responsible for about 80% of the model's predictive power. Now she has multiple candidate Random Forest models, each trained with a different subset of the original 500 features. All of her models are achieving F-score of 0.7-0.75 and AUC of 0.95-0.97 in cross-validation, which is considered excellent for her domain and data.

Satisfied with her models' performance, Grace decides to use TreeSHAP to generate local post-hoc explanations of the trained models, and to run those explanations past her team's domain expert. Together they find that some of the candidate models have relied heavily on features related to past history of dangerous driving - which are only available for existing customers - and hence have limited applicability for new customers switching their insurance providers. They further observe that, based on negative Shapley values, the number of past driving accidents or near-misses inversely correlates with the likelihood of the claim event for some drivers – in other words, a few driving accidents in the past are associated with lower risk of future claim than having no driving accidents. Both Grace and the domain expert find this association counter-intuitive and conclude that the features related to past history of driving are not reliable. Instead of investigating the probable predictive uncertainty in play for the customers without the known history of driving, Grace chooses to discard the candidate models that include these features. Subsequently, Grace deploys into production an alternative propensity score model that does not take into account past history of driving, which has achieved equally high cross-validation performance (F-score of 0.7 and AUC of 0.95). However, once in use, Grace and her team notice that the model classifies a higher proportion of new customers as low risk, increasing the amount of manual assessment required by the underwriters. Although Grace suspects that the model fails to detect new customers with high likelihood of future claim events, she is unable to verify this since the true labels are not available for several months. She has to retire the model prematurely out of fear of reputational and commercial damage. Even in low-stakes applications, the practice of using feature importance explanations for model selection risks consequential confirmation bias, especially since it is unclear how uncertainty affects explanations. Raising awareness of such effects is a motivation for this work.

2.2 Feature Importance Explanations

Existing feature importance explanation techniques can be broadly grouped into those that (i) allocate importance to various input features using game theory (e.g., Shapley values (Lundberg and Lee, 2017)), (ii) find (and perhaps manipulate) the partial derivative of the output with respect to an input feature (Baehrens et al., 2010; Smilkov et al., 2017; Sundararajan, Taly, and Yan, 2017), (iii) redistribute relevance in a backwards pass through a deep learning model (Bach et al., 2015), or (iv) fit a linear surrogate model in the neighborhood of a point of interest (Ribeiro, Singh, and Guestrin, 2016). For a comprehensive review of techniques, see Samek et al. (2020).

2.3 Shapley Value Explanations

A prominent class of feature importance explanation methods is based on Shapley values from cooperative game theory (Shapley, 1953). Shapley values are a method for distributing the gains from a cooperative game to its players. In other words, Shapley values denote the marginal contributions of a player to the payoff of a coalitional game. Let T be the number of players, and let $v : 2^T \to \mathbb{R}$ be the characteristic function, where $v(S)$ denotes the contribution of the players in $S \subseteq T$. The Shapley value of player i's contribution (averaging player i's marginal contributions to all possible subsets S) is given by:

$$\phi_i(v) = \frac{1}{|T|} \sum_{S \subseteq T \setminus \{i\}} \binom{T-1}{S}^{-1} (v(S \cup \{i\}) - v(S)).$$

Let $\Phi \in \mathbb{R}^T$ be a Shapley value contribution vector for all players in the game, where $\phi_i(v)$ is the i^{th} element of Φ. The highlight of Shapley values is that they enjoy axiomatic uniqueness guarantees (Shapley, 1953). In the feature importance literature, Lundberg and Lee (2017) formulate a similar problem to where the game's payoff is the predictor's output $y = \boldsymbol{f}(\boldsymbol{x})$, the players are the d features of $\boldsymbol{x}$, and the ϕ_i values represent the contribution of x_i to the game $\boldsymbol{f}(\boldsymbol{x})$. Note that $g(f, x)_i = \phi_i(v)$. Aas, Jullum, and Løland (2019) define a characteristic function v where:

$$v_{\boldsymbol{x}}(S) = \mathbb{E}\big[f(z)|z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s = \boldsymbol{x}_s]}\big]. \tag{1}$$

For a subset of indices $S \subseteq \{1, 2, \dots d\}$, $\boldsymbol{x}_s = \{x_i, i \in S\}$ denotes a sub-vector of input features that partitions the input, $\boldsymbol{x} = \boldsymbol{x}_s \cup \boldsymbol{x}_c$. $\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s = \boldsymbol{x}_s]}$ denotes an input where the features in S are set to the observed values while the rest of the features remain the reference baseline: $\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s = \boldsymbol{x}_s]} = \boldsymbol{x}_s \cup \bar{\boldsymbol{x}}_c$. When $|S| = d$, $\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s = \boldsymbol{x}_s]} = \boldsymbol{x}$. This characteristic function captures the expected model output given when a subset of features take on the value of some reference baseline. This is used by Lundberg and Lee (2017), who show an equivalence between Shapley values and surrogate methods, such as that of Ribeiro, Singh, and Guestrin (2016).

2.4 Uncertainty Estimates

In the context of this manuscript, we define uncertainty as the variance in predicted probabilities over an ensemble of models from the same model class and with the same model specification. Confidence is defined as the absolute difference between the mean in predicted probabilities over that same ensemble of models and $\frac{1}{k}$, where k is the number of classes.

We posit that we can obtain uncertainty estimates for feature importance explanations by quantifying the variance in importance scores across the each of the candidate models that Grace found in Section 2.1. While Slack et al. (2020) vary the perturbation region of a single Shapley value calculation to obtain uncertainty estimates with respect to a fixed model, we consider feature importance across multiple models (effectively sampling from the posterior of models, given the training samples) to get a mean and variance for a feature importance score. Though we can apply this to any explanation method, we propose our framework in the context of Shapley value explanations.

3 Method: Uncertainty Attributions

In this section, we propose a method for obtaining uncertainty estimates for (Shapley value) feature importance explanations by examining feature importance scores averaged over multiple models in the posterior. Note Equation 1 depends on f, since we fix the f to be the characteristic function. We can explicitly write this dependence as $v(S, f)$. Let the following quantity be the model class feature importance:

$$\phi_i(v)^* = \int_{f \in \mathcal{F}} P(f|\mathcal{D})\phi(v, f)df, \tag{2}$$

where $\phi(v, f) = \frac{1}{|T|}\sum_{S \subseteq T \setminus \{i\}} \binom{T-1}{S}^{-1}(v(S \cup \{i\}, f) - v(S, f))$ and $P(f|\mathcal{D})$ denotes the prior over our model f given dataset $\mathcal{D}$. This quantity captures what the average Shapley value of a data point is across all possible models in some family of functions $\mathcal{F}$. We can interpret these values as the Shapley value of a feature under a selected model class and a given dataset. We deem this Model Class Shapley. Marginalizing over all possible models yields $\phi_i(v)^*$. We weight the Shapley value of each individual model by the likelihood of the model itself. The above is intractable, but we can approximate it via MC Sampling, as follows:

$$\tilde{\phi}_i(v) = \sum_{f \in \mathcal{B}} w_f \phi(v, f) = \mathbb{E}_{f \in \mathcal{B}}[\phi(v, f)], \tag{3}$$

where $\mathcal{B}$ is a finite set of models from $\mathcal{F}$ and w_f captures the likelihood of f. For example, if we assume each model in $\mathcal{B}$ is equally likely, then $w_f = \frac{1}{|\mathcal{B}|}$. We can also weight each model by its accuracy on a validation set. If $\mathcal{B} = \mathcal{F}$, then $\tilde{\phi}_i(v) = \phi_i(v)^*$. In addition to the Model Class Shapley value, we can obtain uncertainty estimates by obtaining the variance over the $|\mathcal{B}|$ Shapley values for feature x_i:

$$s_i(\mathcal{B}, x) = \mathbb{V}_{f \in \mathcal{B}}[\phi_i(v, f)]. \tag{4}$$

Note ϕ can be any feature importance method. We may also want to know how much of the feature importance's variance stems from the ensembling procedure or the explainer itself. We now relate the Shapley values of the ensemble to the Shapley value of each model itself.

Theorem 1. Let f_e be an ensemble predictor defined as $f_e(x) = \frac{1}{|\mathcal{B}|}\sum_{f \in \mathcal{B}} f(x)$. When ϕ is the Shapley value and v is defined per Equation 1, $\phi(v, f_e) = \mathbb{E}_f\left[\phi(v, f)\right]$.

A detailed proof can be found in the supplementary material. Theorem 1 suggests we expect that the Shapley value of an ensemble is equal to the average Shapley value of the constituent models. This corroborates with our empirical findings: based on experiments involving ensembles of 10 to 100 Multilayer Perceptrons, the two Shapley values are equal to 5 decimal places.

4 Experimental Setup

To study the effect of uncertainty on feature importance explanations, we compare explanation quality between (i) out-of-distribution (OOD) samples – those that the model is uncertain about, and (ii) in-distribution samples – those that the model is certain about (i.e., non-OOD). We deem a point as being OOD if it is dispersed further than 1.5 times the interquartile range (IQR) from lower (for confidence) or upper (for uncertainty) quartile.

In total, we conducted over 168 hours of experiments have on a 24-core Intel(R) Xeon(R) E5-2620 v3 @ 2.40GHz processing unit with access to 60 GB RAM.

4.1 Datasets and Models

The experiments described in this work involve two commonly-used ML classification models: Gradient Boosted Trees (GBTs) and Multilayer Perceptrons (MLPs), trained on three publicly-available tabular datasets: Adult census income (Kohavi and Becker, 1996), COMPAS recidivism score (ProPublica, 2017), and Wisconsin Breast Cancer diagnostic (Kohavi and Becker, 1995). In total we have 6 model-dataset pairs. We split each dataset into separate training and test sets, train the classification models on the training set, and compute feature attributions for each point in the test set.

For each model-dataset pair, we iteratively sample from the posterior across initial states with a uniform prior. In each iteration, only the model initialisations and the subset of training points used are permuted, while the model architecture, training procedure, and evaluation parameters remain fixed. This approach is equivalent to the simple averaging ensemble from Lakshminarayanan, Pritzel, and Blundell (2017). The resulting point estimates make up the distribution of model's predicted probabilities, where the expected value defines the model's confidence, and its variance defines the epistemic uncertainty. The corresponding predictions on a test set (20% of all samples) are subsequently explained using one of the feature importance explanation techniques described in the following subsection.

4.2 Explainers

For MLPs and their ensembles, the feature importance methods we use are (i) Integrated Gradients (Sundararajan, Taly, and Yan, 2017), and (ii) SmoothGrad (Smilkov et al., 2017), which is a noise-tunnel modification of Integrated Gradients. We use the Captum implementation (CaptumAI, 2020) of both of the above methods. For GBTs, we utilize (i) exact Shapley importance scores, (ii) TreeSHAP (Lundberg et al., 2020), and (iii) KernelSHAP (Lundberg and Lee, 2017) implementations. In total, we examine 5 explainers.

4.3 Evaluation Criteria

We evaluate the effect of uncertainty on explanation quality with respect to the following metrics:

- *Variance* (in feature importance scores): as defined by Equation 4, averaged across the dataset. This may be reported per feature or averaged over all features. A low variance is desired.
- *Complexity*: the entropy of the probability distribution made by the fractional contributions of each feature to the magnitude of the attribution. A low complexity is desired, with the simplest explanation being comprised of a single feature attribution (Bhatt, Weller, and Moura, 2020).

		Variance	Complexity	Monotonicity	Efficiency	Faithfulness
Exact Shapley	non-OOD	0.07 ± 0.00	1.88 ± 0.22	0.88 ± 0.04	1.00 ± 0.00	0.00 ± 0.24
	OOD	$\mathbf{0.39 \pm 0.00}$	$\mathbf{1.96 \pm 0.14}$	$\mathbf{0.85 \pm 0.05}$	1.00 ± 0.00	$\mathbf{-0.01 \pm 0.23}$
TreeSHAP	non-OOD	0.08 ± 0.00	1.90 ± 0.20	0.88 ± 0.04	0.99 ± 0.09	0.00 ± 0.24
	OOD	$\mathbf{0.44 \pm 0.00}$	$\mathbf{1.96 \pm 0.13}$	$\mathbf{0.85 \pm 0.04}$	$\mathbf{0.97 \pm 0.18}$	$\mathbf{-0.01 \pm 0.23}$
KernelSHAP	non-OOD	0.19 ± 0.00	1.81 ± 0.23	0.89 ± 0.05	1.00 ± 0.00	0.00 ± 0.24
	OOD	$\mathbf{0.809 \pm 0.00}$	$\mathbf{1.85 \pm 0.16}$	$\mathbf{0.82 \pm 0.05}$	1.00 ± 0.00	0.00 ± 0.23
Integrated Gradients	non-OOD	20.00 ± 5.34	1.53 ± 0.22	0.96 ± 0.03	$\mathbf{0.99 \pm 0.10}$	0.00 ± 0.16
	OOD	$\mathbf{46.48 \pm 18.61}$	$\mathbf{1.55 \pm 0.18}$	0.96 ± 0.03	1.00 ± 0.03	0.00 ± 0.13
Smoothgrad	non-OOD	20.35 ± 0.52	1.92 ± 0.11	0.92 ± 0.04	1.00 ± 0.10	0.00 ± 0.13
	OOD	19.97 ± 0.43	$\mathbf{1.98 \pm 0.06}$	0.92 ± 0.03	1.00 ± 0.09	$\mathbf{-0.02 \pm 0.10}$

Table 1: Comparison of the explanation quality on OOD and in-distribution samples of the Adult dataset in terms of the mean and standard deviation of quality scores averaged across the ensemble of models. Values in **bold** indicate where the degradation in quality was statistically significant ($p<0.05$).

10^-3	Cohort	Age	Workclass	Years in Education	Marital Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country
Exact Shapley	non-OOD	0.142	0.052	0.082	0.115	0.095	0.161	0.015	0.024	0.042	0.023	0.086	0.021
	OOD	**0.851**	**0.390**	**0.538**	**0.319**	**0.509**	**0.464**	**0.165**	**0.103**	**0.177**	**0.267**	**0.641**	**0.293**
TreeSHAP	non-OOD	0.184	0.053	0.092	0.116	0.104	0.147	0.018	0.024	0.079	0.023	0.109	0.022
	OOD	**0.932**	**0.332**	**0.633**	**0.341**	**0.496**	**0.543**	**0.145**	**0.095**	**0.481**	**0.351**	**0.696**	**0.233**
KernelSHAP	non-OOD	0.241	0.068	0.178	0.299	0.251	0.604	0.017	0.026	0.099	0.062	0.148	0.248
	OOD	**1.303**	**0.579**	**1.014**	**0.865**	**1.165**	**1.620**	**0.202**	**0.163**	**0.333**	**0.485**	**1.011**	**0.868**
Integrated Gradients	non-OOD	16.554	8.758	26.606	**37.233**	25.819	55.443	5.467	15.611	7.590	12.597	14.390	13.965
	OOD	**47.485**	**37.694**	**100.807**	34.555	**81.930**	**96.448**	**7.780**	**27.302**	**28.773**	**15.651**	**56.416**	**22.881**
SmoothGrad	non-OOD	17.842	5.159	25.253	**24.581**	13.718	**45.294**	**3.721**	**8.821**	**67.266**	**7.871**	14.739	9.936
	OOD	**19.513**	**7.399**	**30.758**	17.057	**17.091**	38.126	3.639	7.092	61.986	6.952	**18.297**	**11.750**

Table 2: Variance in feature attribution compared between OOD and in-distribution samples of the Adult dataset. Presented is the mean of the Variance metric across the ensemble. Values are multiplied by 10^3 for clarity of presentation. Numbers in **bold** indicate where the degradation in quality is statistically significant ($p<0.05$).

- *Monotonicity*: measures the changes in model performance when incrementally adding each attribute in order of increasing importance. As each feature is added, the performance of the model should correspondingly increase (or decrease, if that feature's attribution is negative), thereby resulting in monotonically increasing model performance (Luss et al., 2019). Per Young (1985), monotonicity ensures that feature importance scores satisfy the Shapley axioms of additivity and of null value ("dummy").
- *Efficiency*: a property ensuring that the feature attributions add up to the difference between model prediction and a baseline (baseline is typically set to the global mean of the training data). For exact Shapley values via the KernelSHAP implementation and for Integrated Gradients, the efficiency is expected to be equal to 1 by design. Also known as "local accuracy" or "completeness".
- *Faithfulness*: measures the correlation of the feature importance with a "ground truth," commonly defined as difference in predicted value when a given feature is occluded (Melis and Jaakkola, 2018; Yeh et al., 2019).

5 Experimental Results

Our results are consistent across all three datasets. In this section, we describe our results for the Adult dataset. Additional results can be found in the Supplementary Material.

5.1 Explanation Quality: Overall

Our experiments quantitatively compare and contrast the quality of feature importance explanations under uncertainty based on the 5 metrics in Section 4.3. The results averaged over all features are shown in Table 1. We observed a statistically significant decrease in explanation quality across all 5 evaluation criteria (see Section 4.3) and 5 types of explainers (see Section 4.2). As summarized in Table 1, *Complexity* and *Monotonicity* degrade significantly for all explainers considered. For explainers which are not 100% efficient by design (TreeSHAP) efficiency was also lower in OOD samples.

Our experimental results also confirm that exact Shapley values and their KernelSHAP approximation are locally accurate and fully meet the *Efficiency* axiom (i.e. scores equal to 1 in both OOD and non-OOD samples), however, the Integrated Gradients and SmoothGrad implementations show surprising deviation from the *Efficiency* axiom. This discrepancy could be due to the choice of the baseline, which is set to the global mean of the training set, whilst the explanations are computed for samples in the test set. With respect to *Faithfulness*, the differences between OOD and non-OOD samples are less pronounced, since statistically significant ($p<0.05$) degradation in explanation quality for points with low uncertainty was only observed in exact Shapley, TreeSHAP, and SmoothGrad explanations.

Overall, we find that feature importance explanations fail

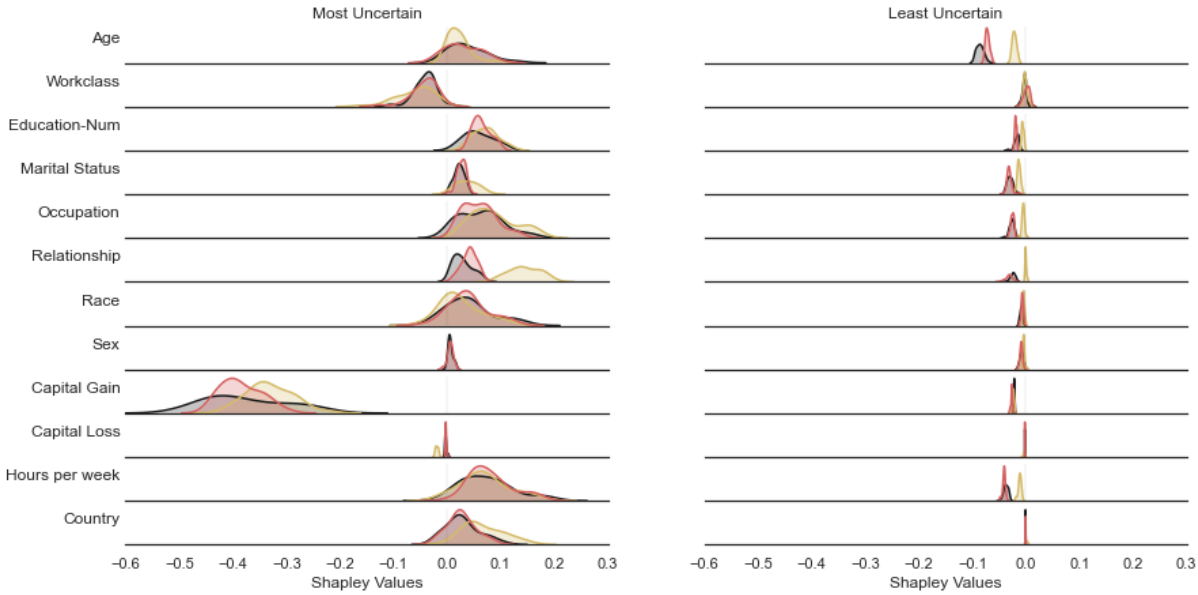


Figure 1: Feature importance explanations of samples of the Adult dataset computed using Exact (red), Tree (black) and Kernel (yellow) SHAP. Left: Most uncertain sample (OOD). Right: Least uncertain sample (non-OOD). Feature names are listed on the left. Each feature is presented as a distribution of its Shapley values associated with the ensemble of 30 Gradient Boosted Trees.

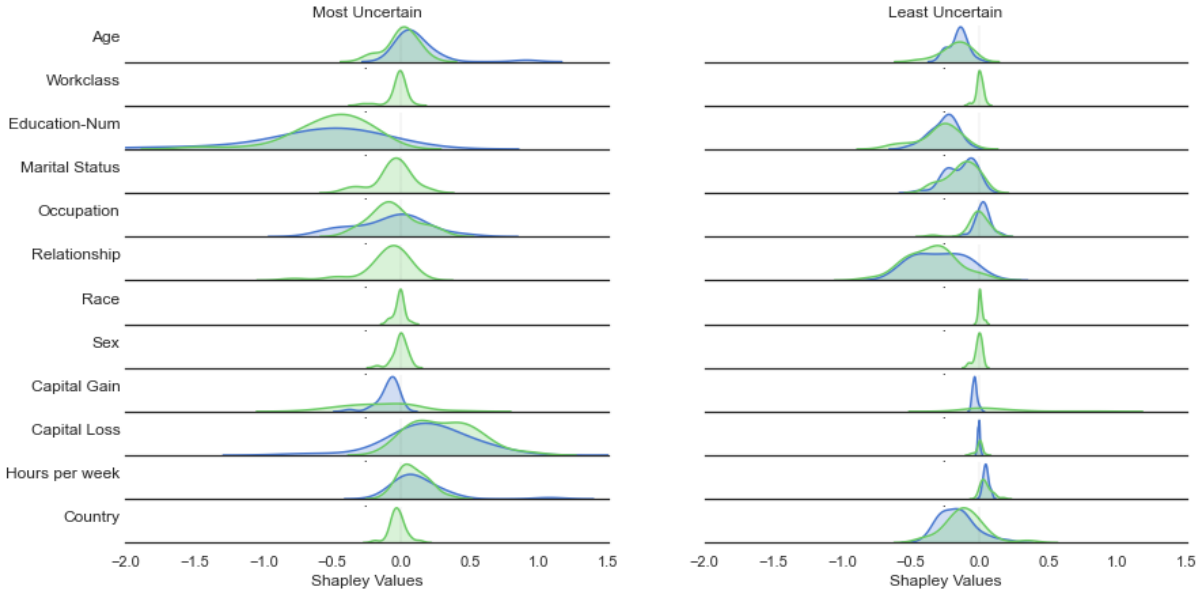


Figure 2: Feature importance explanations of samples of the Adult dataset computed using Integrated Gradients (blue) and SmoothGrad (green). Left: Most uncertain sample (OOD). Right: Least uncertain sample (non-OOD). Feature names are listed on the left. Each feature is presented as a distribution of its explanations associated with the ensemble of 30 Multilayer Perceptrons.

on uncertain (OOD) points; specifically, evaluating the quality of explanations shows that both Shapley value explanations and Integrated Gradient explanations for OOD data are more complex, less monotone, and less faithful than explanations for non-OOD (i.e., in-distribution) data.

5.2 Explanation Quality: Variance

We now focus on one facet of explanation quality, *Variance*. In Table 2, we break down the explanation *Variance* for OOD and non-OOD samples by feature in the Adult dataset. Our experiments across a range of explainers support the hypothesis that uncertainty in prediction (i.e., disagreement in ensemble predictions) propagates to uncertainty in individual feature importance scores, resulting in higher variance in

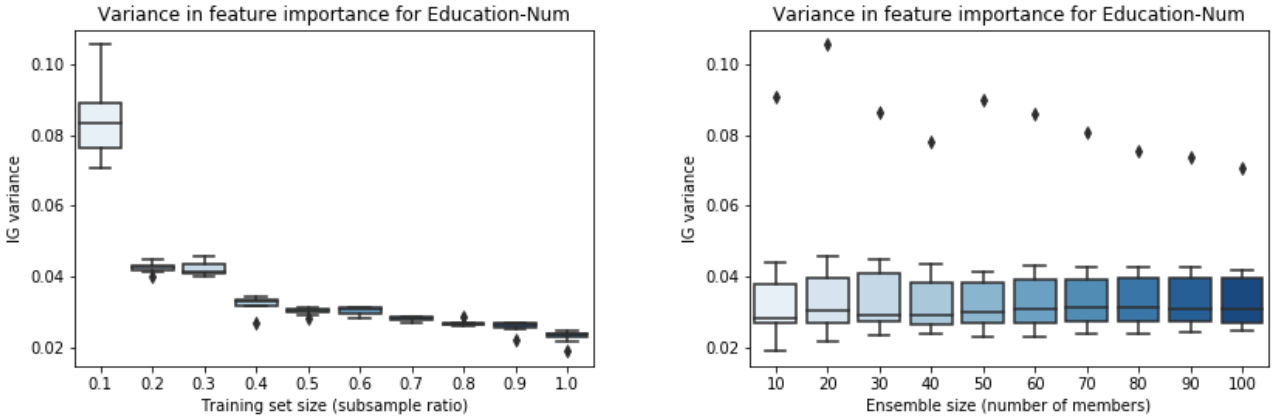


Figure 3: Left: As the subsample ratio, $\frac{k}{n}$, increases, our variance estimates converge. Right: Ensemble size, $|\mathcal{B}|$, leads to negligible improvement in the quality of our uncertainty estimates after 60 model ensembles.

feature importance scores across the ensemble.

As an illustrative example, Figure 1 shows two samples in the Adult datset using a GBT model: the most uncertain sample in the dataset (OOD), and the least uncertain sample in the dataset (non-OOD). We find that there is a larger spread (i.e., variance) in feature importance scores attributed to each of the 12 features for the OOD sample. In contrast, an in-distribution sample – i.e. a point on which the ensemble of models had agreement – yields low variance in the explanations. These observations also hold for a different model class (MLPs), and different explainers (Integrated Gradients and SmoothGrad) as shown in Figure 2. In the supplementary material, we cover an example of uncertainty attribution for MNIST digits with feature importance (LeCun, 1998).

5.3 Hyperparameter Sensitivity

To obtain our uncertainty estimates, there are two hyperparameters that affect the quality of our estimate: $|\mathcal{B}|$ the number of models we train and k points to sample from n datapoints in $\mathcal{D}$. We call $\frac{k}{n}$ our subsample ratio that is the training set size for a given model in our ensemble. In Figure 3 (Left), we see that ensemble size has negligible effect on variance quality, beyond 60 ensembles for a particular feature (Education-Num from the Adult dataset): similar results are reported in the supplementary material for other features. In Figure 3 (Right), we see that as the subsample ratio increases, variance estimates for a particular feature (again Education-Num) converge: similar results are reported in the supplementary material for other features. In Figure 4, we relate subsample ratio and ensemble size in terms of the average variance in feature importance (per Equation 4) across the Adult test set. Note that the F1 score for each ensemble, where the subsample ratio was greater than 0.2, was around 0.68. In future work, we hope to formally relate ensemble size, subsample ratio, and their effect on the converge of the variance in feature importance, using the machinery established by Politis, Romano, and Wolf (2001).

6 Discussion

We strongly suggest that existing feature importance techniques should not be used on samples for which a model is uncertain. We notice that feature importance scores for OOD data perform poorly on multiple quantitative evaluation criteria. We suggest that data practitioners convey the uncertainty in explanations themselves as opposed to simply not using post-hoc explanation techniques for points with high predictive uncertainty. As such, we develop a scheme to calculate uncertainty estimates for feature importance scores. These uncertainty estimates can be used to detect OOD data and to develop an uncertainty attribution, an alternative to feature importance that captures where in input space an uncertainty estimate lies. We report the variance of the feature importance explanations across the ensemble as the uncertainty attribution. Instead of attributing importance to each feature, uncertainty attributions denote which inputs are contributing to a model's uncertainty. Future work can leverage generative modelling to develop uncertainty explanations that not only denote how much predictive uncertainty is associated with each input feature but also provide an actionable change for an input to achieve a specified model confidence, similar to Booth et al. (2020) and Antorán et al. (2021). Moreover, we hope future work can establish theory for uncertainty attributions as an alternative for feature importance.

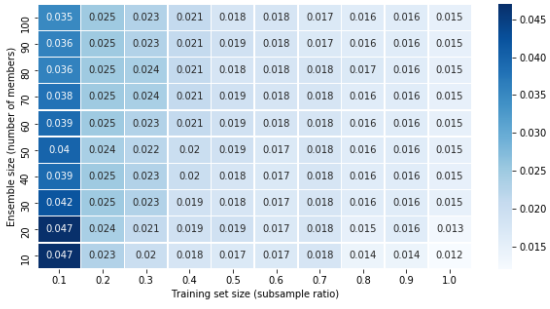


Figure 4: The interplay between ensemble size and subsample ratio: subsample ratio has a stronger effect on the convergence of our variance estimates.

References

Aas, K.; Jullum, M.; and Løland, A. 2019. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv preprint arXiv:1903.10464*.

Antorán, J.; Bhatt, U.; Adel, T.; Weller, A.; and Hernández-Lobato, J. M. 2021. Getting a CLUE: A method for explaining uncertainty estimates. In *International Conference on Learning Representations*.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7).

Baehrens, D.; Schroeter, T.; Harmeling, S.; Kawanabe, M.; Hansen, K.; and MÃžller, K.-R. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11(Jun):1803–1831.

Bhatt, U.; Xiang, A.; Sharma, S.; Weller, A.; Taly, A.; Jia, Y.; Ghosh, J.; Puri, R.; Moura, J. M.; and Eckersley, P. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657.

Bhatt, U.; Weller, A.; and Moura, J. M. 2020. Evaluating and aggregating feature-based model explanations. In *IJCAI*.

Booth, S.; Zhou, Y.; Shah, A.; and Shah, J. 2020. Bayes-probe: Distribution-guided sampling for prediction level sets. *arXiv preprint arXiv:2002.10248*.

Breiman, L.; Friedman, J.; Stone, C. J.; and Olshen, R. A. 1984. *Classification and regression trees*. CRC press.

CaptumAI. 2020. Captum v0.2.0 release.

Fisher, A.; Rudin, C.; and Dominici, F. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20(177):1–81.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89. IEEE.

Kohavi, R., and Becker, B. 1995. Breast cancer wisconsin (diagnostic) data set.

Kohavi, R., and Becker, B. 1996. Adult data set.

Lakkaraju, H.; Arsov, N.; and Bastani, O. 2020. Robust and stable black box explanations.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, 6402–6413.

LeCun, Y. 1998. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 4765–4774.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence* 2(1):2522–5839.

Luss, R.; Chen, P.; Dhurandhar, A.; Sattigeri, P.; Shanmugam, K.; and Tu, C. 2019. Generating contrastive explanations with monotonic attribute functions. *CoRR* abs/1905.12698.

Melis, D. A., and Jaakkola, T. 2018. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, 7775–7784.

Politis, D. N.; Romano, J. P.; and Wolf, M. 2001. On the asymptotic theory of subsampling. *Statistica Sinica* 1105–1124.

ProPublica. 2017. Compas recidivism racial bias racial bias in inmate compas reoffense risk scores for florida.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144. New York, NY, USA: ACM.

Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; and Müller, K.-R. 2020. Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*.

Shapley, L. S. 1953. A value for n-person games. In *Contributions to the Theory of Games II*. 307–317.

Slack, D.; Hilgard, S.; Singh, S.; and Lakkaraju, H. 2020. How much should I trust you? modeling uncertainty of black box explanations. *arXiv preprint arXiv:2008.05030*.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, 3319–3328. JMLR.org.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (in)fidelity and sensitivity of explanations. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. 10967–10978.

Young, H. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14:65–72.

Zhang, Y.; Song, K.; Sun, Y.; Tan, S.; and Udell, M. 2019. "Why should you trust my explanation?" Understanding uncertainty in LIME explanations.

Zhang, Y.; Liao, Q. V.; and Bellamy, R. K. E. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 295–305. New York, NY, USA: Association for Computing Machinery.

Appendix

We start by discussing uncertainty attributions. We then provide a Proof for Theorem 1. We conclude with additional results.

A Uncertainty Attributions

Uncertainty attributions, similar to feature importance, provide "saliency" maps that allow us to visualize where input space uncertainty lies. We now compare feature importance with uncertainty attributions for certain and uncertain test points. Figure 5 shows examples of LIME and Kernel SHAP being applied to a BNN for high confidence MNIST test digits. We use the default LIME hyperparameters for MNIST: the "quickshift" segmentation algorithm with kernel size 1, maximum distance 5 and a ratio of 0.2. We plot the top 10 segments with weight greater than 0.01. We draw 1000 samples with both methods. Using the same configuration, we generate LIME and SHAP explanations for some MNIST digits to which our BNN assigns predictive entropy above a rejection threshold (that is, low confidence). The results are displayed in Figure 6. For feature importance in MNIST digits, the reference is an entirely black image. Note that alternative versions of SHAP exist that incorporate information about internal NN dynamics into explanations. However, they produce very noisy explanations when applied to our BNNs. We conjecture that this high variance might be induced by disagreement among the multiple weight configurations from our BNNs.

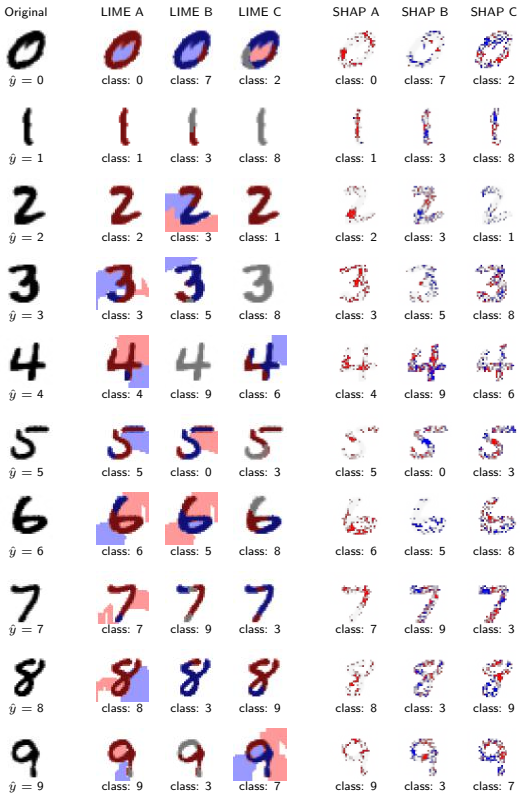


Figure 5: High confidence MNIST test examples together with LIME and SHAP explanations for the top 3 predicted classes. The model being investigated is a BNN. The highest probability class is denoted by $\hat{y}$.

When faced with an uncertain input, we posit that uncertainty attributions are more useful than feature importance. A positive uncertainty attribution means that the addition of that feature will make our model more certain. A positive feature importance means the presence of that feature serves as evidence towards a predicted class. A negative uncertainty attribution means that the the absence of that feature will make the model more certain. A negative feature importance attribution means the absence of that feature would serve as evidence for a particular prediction. While uncertainty attribution and feature importance solve similar problems and both provide "saliency" maps, uncertainty attributions highlight regions that need to be added or removed to make the input certain to a model. In some cases, we see that negative feature importance attribution aligns with negative uncertainty attribution, suggesting the features which negatively contribute to the model's predicted probability are the features that need to be removed to increase the models' certainty. The ability for uncertainty explanations to suggest the addition of unobserved features (positive uncertainty attribution) is unique. The feature importance methods under consideration are difficult to retrofit for uncertainty without a procedure like ours. They are unable to add features; they are limited to explaining the contribution of existing features. This may suffice if the input contains all the information needed to make a prediction for a specific class;

Figure 6: 10 MNIST test digits for which our BNN's predictive variance is above the rejection threshold. A single CLUE example is provided for each one (Antorán et al., 2021). The top scoring class is denoted by $\hat{y}$. LIME and SHAP explanations are provided for the 3 most likely classes.

otherwise, this results in noisy, potentially meaningless, explanations. Generative-model based explanation methods, like FIDO, can mitigate this, since they are flexible enough to deal with uncertain inputs.

B Proofs

Theorem 2. Let f_e be an ensemble predictor defined as $f_e(x) = \frac{1}{|\mathcal{B}|}\sum_{f\in\mathcal{B}} f(x)$. When ϕ is the Shapley value and v is defined per Equation 1, $\phi(v, f_e) = \mathbb{E}_f\left[\phi(v, f)\right]$

Proof. Let f_e be an ensemble predictor defined as $f_e(x) = \frac{1}{|\mathcal{B}|}\sum_{f\in\mathcal{B}} f(x)$. We now relate the Shapley values of f_e to the Shapley value of each model itself.

$$\begin{aligned}\phi_i(v, f_e) &= \frac{1}{|T|}\sum_{S\subseteq T\setminus\{i\}}\binom{T-1}{S}^{-1}\left(v(S\cup\{i\}, f_e) - v(S, f_e)\right)\\ &= \frac{1}{|T|}\sum_{S\subseteq T\setminus\{i\}}\binom{T-1}{S}^{-1}\left(\mathbb{E}\big[f_e(z)|z=\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_{s\cup i}=\boldsymbol{x}_{s\cup i}]}\big] - \mathbb{E}\big[f_e(z)|z=\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s=\boldsymbol{x}_s]}\big]\right)\end{aligned}$$

Note:

$$\begin{aligned}\mathbb{E}\big[f_e(z)|z=\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s=\boldsymbol{x}_s]}\big] &= \mathbb{E}\big[\mathbb{E}_f[f(z)]\ |z=\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s=\boldsymbol{x}_s]}\big]\\ &= \mathbb{E}\big[\frac{1}{|\mathcal{B}|}\sum_{f\in\mathcal{B}} f(z)|z=\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s=\boldsymbol{x}_s]}\big]\\ &= \frac{1}{|\mathcal{B}|}\sum_{f\in\mathcal{B}}\mathbb{E}\big[f(z)|z=\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s=\boldsymbol{x}_s]}\big]\\ &= \mathbb{E}_f\left[\mathbb{E}\left[f(z)|z=\bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s=\boldsymbol{x}_s]}\right]\right]\end{aligned}$$

As such:

$$\begin{aligned}
\phi_i(v, f_e) &= \frac{1}{|T|} \sum_{S \subseteq T \setminus \{i\}} \binom{T-1}{S}^{-1} \left(\mathbb{E}\big[f_e(z) | z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_{s \cup i} = \boldsymbol{x}_{s \cup i}]}\big] - \mathbb{E}\big[f_e(z) | z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s = \boldsymbol{x}_s]}\big] \right) \\
&= \frac{1}{|T|} \sum_{S \subseteq T \setminus \{i\}} \binom{T-1}{S}^{-1} \left(\mathbb{E}_f \left[\mathbb{E} \left[f(z) | z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_{s \cup i} = \boldsymbol{x}_{s \cup i}]} \right] \right] - \mathbb{E}_f \left[\mathbb{E} \left[f(z) | z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s = \boldsymbol{x}_s]} \right] \right] \right) \\
&= \frac{1}{|T|} \sum_{S \subseteq T \setminus \{i\}} \binom{T-1}{S}^{-1} \left(\mathbb{E}_f \left[\mathbb{E} \left[f(z) | z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_{s \cup i} = \boldsymbol{x}_{s \cup i}]} \right] - \mathbb{E} \left[f(z) | z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s = \boldsymbol{x}_s]} \right] \right] \right) \\
&= \mathbb{E}_f \left[\frac{1}{|T|} \sum_{S \subseteq T \setminus \{i\}} \binom{T-1}{S}^{-1} \left(\mathbb{E} \left[f(z) | z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_{s \cup i} = \boldsymbol{x}_{s \cup i}]} \right] - \mathbb{E} \left[f(z) | z = \bar{\boldsymbol{x}}_{[\bar{\boldsymbol{x}}_s = \boldsymbol{x}_s]} \right] \right) \right] \\
&= \mathbb{E}_f \left[\phi(v, f) \right]
\end{aligned}$$

Therefore by leveraging the linearity of the Shapley value and the linearity of our ensemble, we show that the Shapley value of an ensemble is equal to average Shapley value of the constituent models. □

C Additional Results

C.1 Evaluating Feature Importance Explanations

Similar to Table 2 for the Adult dataset in the main paper, Table 3 and Table 4 contain the results of various explanation evaluation criteria on the COMPAS and Cancer datasets respectively. We observe similar behavior to the Adult dataset: our variance metric identifies OOD data well and also suggests that explanation quality suffers for OOD data. In Figure 10, we notice that low model confidence leads to a degradation in quality of explanations. In particular, complexity of explanations decreases for samples on which model confidence was high. We also find that explanation quality suffers for models with low model confidence (here we define confidence as $|0.5 - f(x)|$ where $f(x) \in [0, 1]$ is the probability of the most likely label in our binary classification setting. In Figure 7 we notice that more confident examples has a *sharper*, more concentrated distribution of Shapley values. On the other hand, for the least confident sample, we notice that for some features the distributions can vary drastically. Figure 8 shows similar results for gradient-based explanation methods. We hope that future work can leverage our explanation variance metric to assist with model selection. We also hope that future work can study the connection between the variance in feature importance and fairness desiderata: in Figure 9, we show that altering one feature for an individual can drastically change the Shapley values for that feature alone; i.e., we decrease a person's age and then the distribution of Shapley values for age changes, while the distribution of Shapley values for the rest of the features is roughly the same.

		Variance	Complexity	Monotonicity	Efficiency	Faithfulness
Exact Shapley	non-OOD	0.00 ± 0.00	1.37 ± 0.23	0.95 ± 0.04	1.00 ± 0.00	-0.01 ± 0.22
	OOD	$\mathbf{0.01 \pm 0.00}$	1.37 ± 0.21	$\mathbf{0.92 \pm 0.04}$	1.00 ± 0.00	$\mathbf{-0.04 \pm 0.13}$
TreeSHAP	non-OOD	0.00 ± 0.00	1.38 ± 0.21	0.95 ± 0.04	0.98 ± 0.15	-0.01 ± 0.22
	OOD	$\mathbf{0.02 \pm 0.00}$	1.38 ± 0.20	$\mathbf{0.92 \pm 0.04}$	$\mathbf{0.86 \pm 0.34}$	$\mathbf{-0.04 \pm 0.13}$
KernelSHAP	non-OOD	0.00 ± 0.00	0.95 ± 0.25	0.95 ± 0.04	1.00 ± 0.00	-0.01 ± 0.21
	OOD	$\mathbf{0.01 \pm 0.00}$	$\mathbf{1.10 \pm 0.23}$	$\mathbf{0.92 \pm 0.04}$	1.00 ± 0.00	$\mathbf{-0.04 \pm 0.13}$
Integrated Gradients	non-OOD	1.56 ± 0.70	1.88 ± 0.18	0.83 ± 0.03	1.00 ± 0.06	0.00 ± 0.12
	OOD	6.20 ± 5.77	$\mathbf{1.94 \pm 0.05}$	0.82 ± 0.02	1.00 ± 0.00	$\mathbf{-0.11 \pm 0.07}$
Smoothgrad	non-OOD	0.81 ± 0.04	2.04 ± 0.14	0.78 ± 0.04	1.00 ± 0.06	0.00 ± 0.10
	OOD	$\mathbf{1.51 \pm 0.08}$	$\mathbf{2.06 \pm 0.05}$	0.78 ± 0.04	1.00 ± 0.00	0.05 ± 0.03

Table 3: Comparison of the explanation quality on OOD and in-distribution samples of the COMPAS dataset. Presented are mean and standard deviation of quality scores averaged across the ensemble of models. Values in **bold** indicate where the degradation in quality was statistically significant ($p<0.05$).

C.2 Hyperparameter Sensitivity

We now report how the empirical variance of feature importance as a function of ensemble size ($|\mathcal{B}|$) and subsample ratio ($\frac{k}{n}$). In Figure 12, we report heatmaps of the average variance of Integrated Gradient explanations across the test set for specific features,

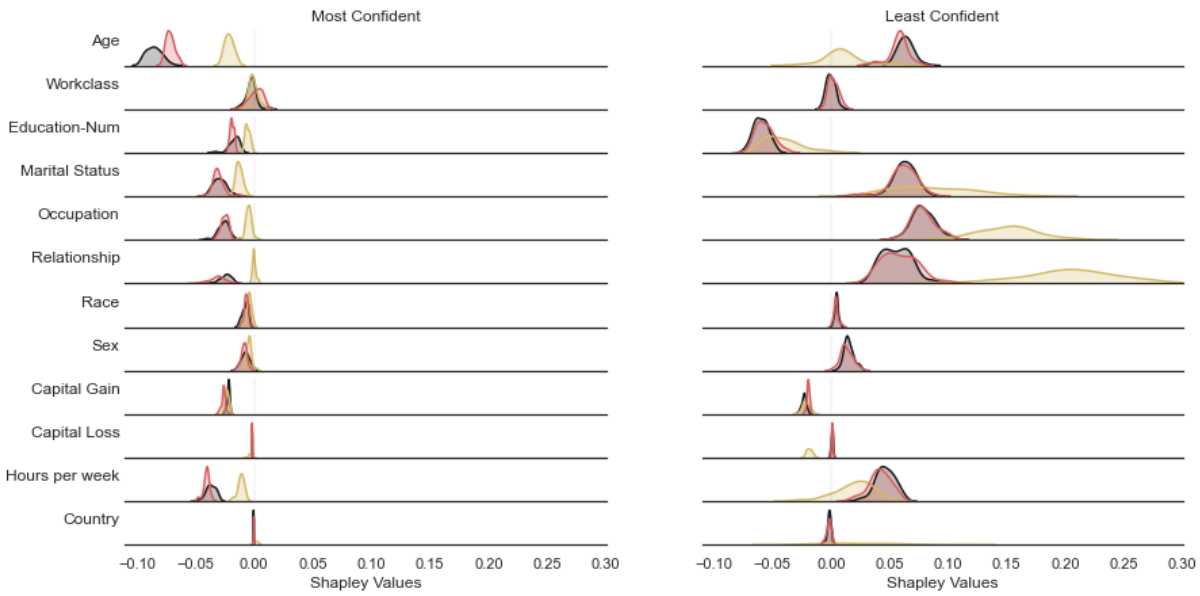


Figure 7: Feature importance scores of the most confident (left) and least confident (right) samples of the Adult dataset computed using Exact (red), Tree (black) and Kernel (yellow) SHAP. Each feature is presented as a distribution of its Shapley values associated with the ensemble of 30 Gradient Boosted Trees.

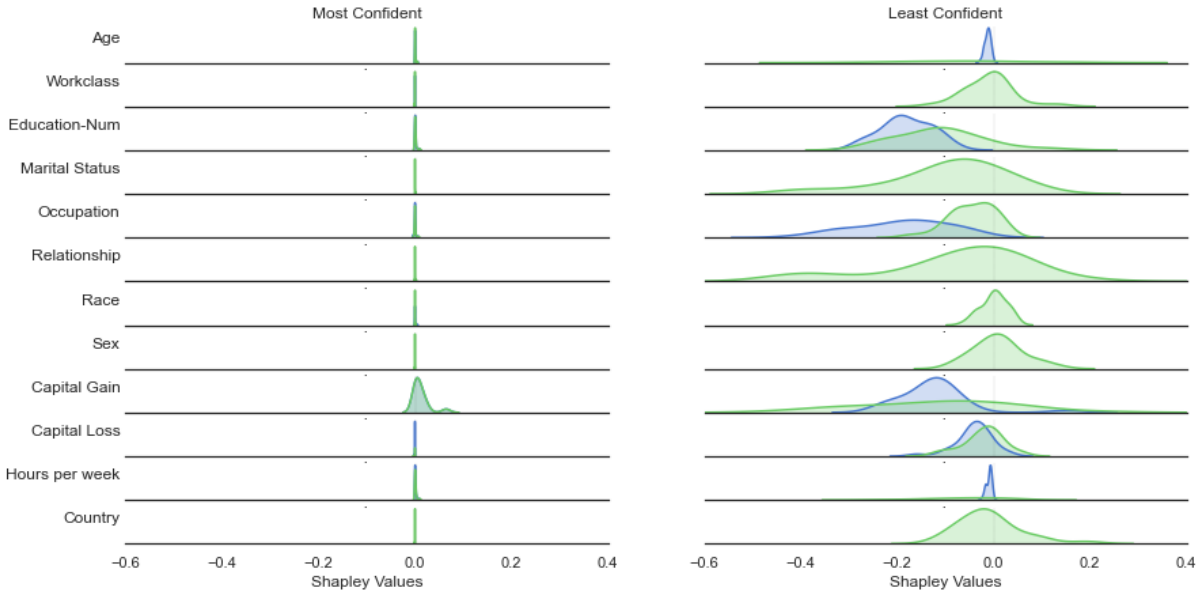
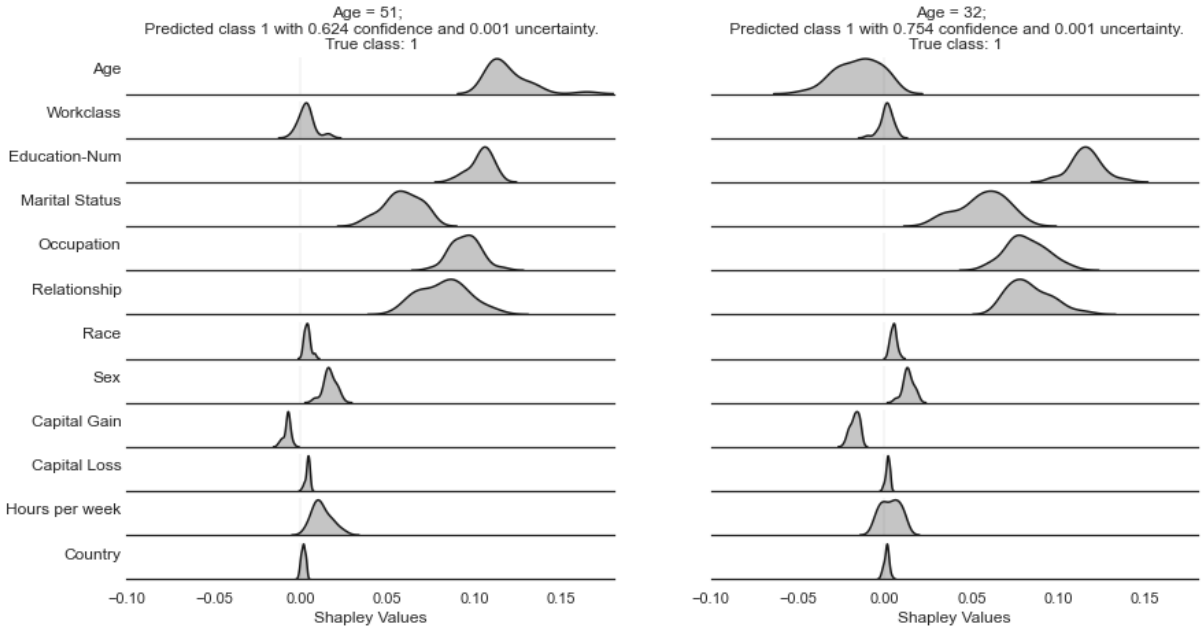
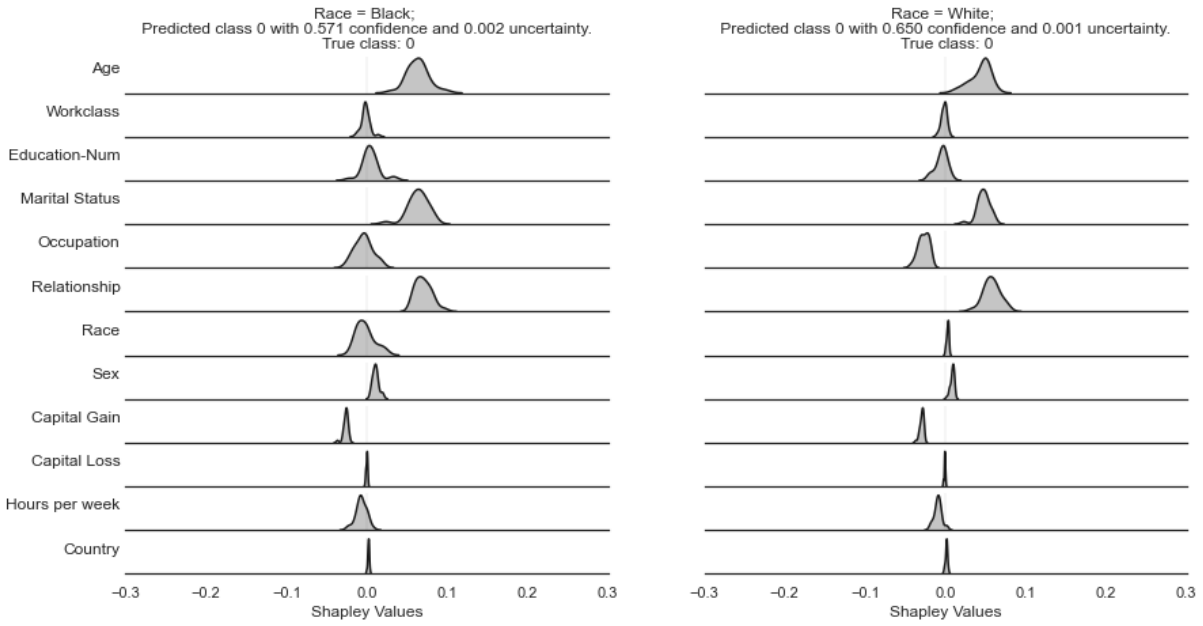


Figure 8: Feature attributions of the most confident (left) and least confident (right) samples of the Adult dataset computed using Integrated Gradients (blue) and SmoothGrad (green). Each feature is presented as a distribution of its explanations associated with the ensemble of 30 Multilayer Perceptrons.



(a) Adult Dataset: Age Differences



(b) Adult Dataset: Race Differences

Figure 9: On the top, we have two individuals with similar feature values except for their age; the Shapley distribution indicates that Age is more important for older individuals. On the bottom, we have two individuals with similar feature values except for their race; the Shapley distribution indicates the change in race affects the Shapley value for race and for occupation, while most other features are distributed similarly.

Faithfulness		Variance		Complexity	Monotonicity	Efficiency
TreeSHAP	non-OOD	0.00 ± 0.00	2.68 ± 0.08	0.99 ± 0.01	1.00 ± 0.00	-0.03 ± 0.13
	OOD	$\mathbf{0.01 \pm 0.00}$	$\mathbf{2.69 \pm 0.06}$	0.99 ± 0.01	$\mathbf{0.97 \pm 0.18}$	$\mathbf{-0.07 \pm 0.13}$
KernelSHAP	non-OOD	0.00 ± 0.00	2.47 ± 0.16	1.00 ± 0.01	1.00 ± 0.00	-0.03 ± 0.13
	OOD	$\mathbf{0.01 \pm 0.00}$	$\mathbf{2.50 \pm 0.19}$	$\mathbf{0.99 \pm 0.01}$	1.00 ± 0.00	$\mathbf{-0.05 \pm 0.11}$
Integrated Gradients	non-OOD	0.01 ± 0.00	2.85 ± 0.10	1.00 ± 0.00	1.00 ± 0.00	-0.02 ± 0.16
	OOD	$\mathbf{0.87 \pm 0.60}$	$\mathbf{2.88 \pm 0.10}$	1.00 ± 0.00	1.00 ± 0.00	$\mathbf{0.01 \pm 0.17}$
Smoothgrad	non-OOD	0.22 ± 0.02	2.83 ± 0.05	1.00 ± 0.01	$\mathbf{0.93 \pm 0.25}$	-0.02 ± 0.11
	OOD	$\mathbf{0.55 \pm 0.04}$	$\mathbf{2.84 \pm 0.05}$	$\mathbf{0.99 \pm 0.01}$	0.95 ± 0.23	0.03 ± 0.16

Table 4: Comparison of the explanation quality on OOD and in-distribution samples of the Cancer dataset. Presented are mean and standard deviation of quality scores averaged across the ensemble of models. Values in **bold** indicate where the degradation in quality was statistically significant ($p<0.05$).

(a) Adult dataset

(b) COMPAS dataset

Figure 10: Complexity of explanations (Exact Shapley values) by predictive confidence of a Gradient Boosted Tree ensemble. For this binary classification task, both correctly (in orange) and incorrectly (in blue) classified samples are displayed.

as we vary the hyperparameters of interest. In Figure 13, we report boxplots of how average variance for Integrated Gradients changes with ensemble size. In Figure 14, we report boxplots of how average variance for Integrated Gradients changes with the subsample ratio. For completeness, we include analysis of how varying ensemble size and subsample ratio affect the models' F1 score.

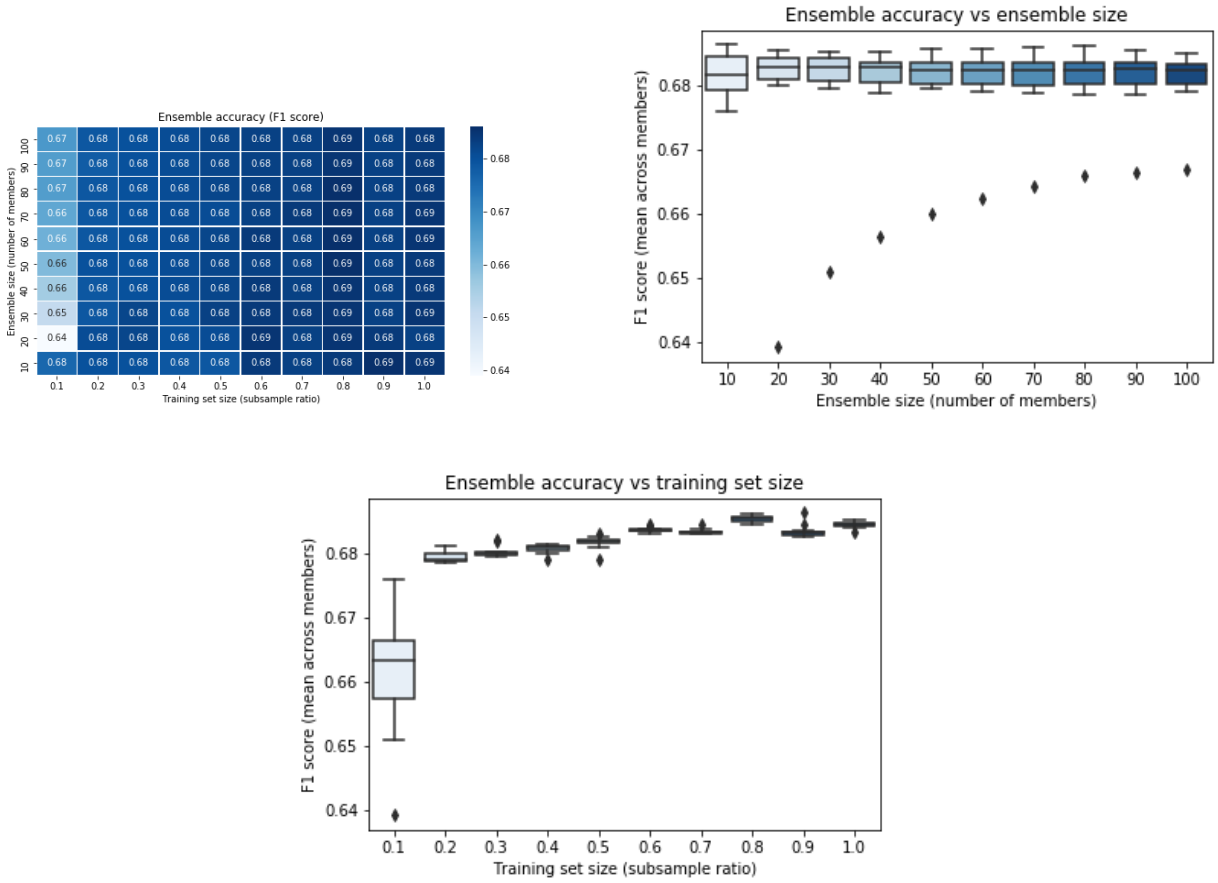


Figure 11: We report how the models' F1 scores vary with ensemble size and subsample ratio.

Figure 12: Average variance in feature importance scores across the Adult test set reported as a function of ensemble size and subsample ratio. Notice that subsample ratio seems to matter more for the variance estimates to converge. Ensemble size does not need to be too large for variance estimates for most features to converge.

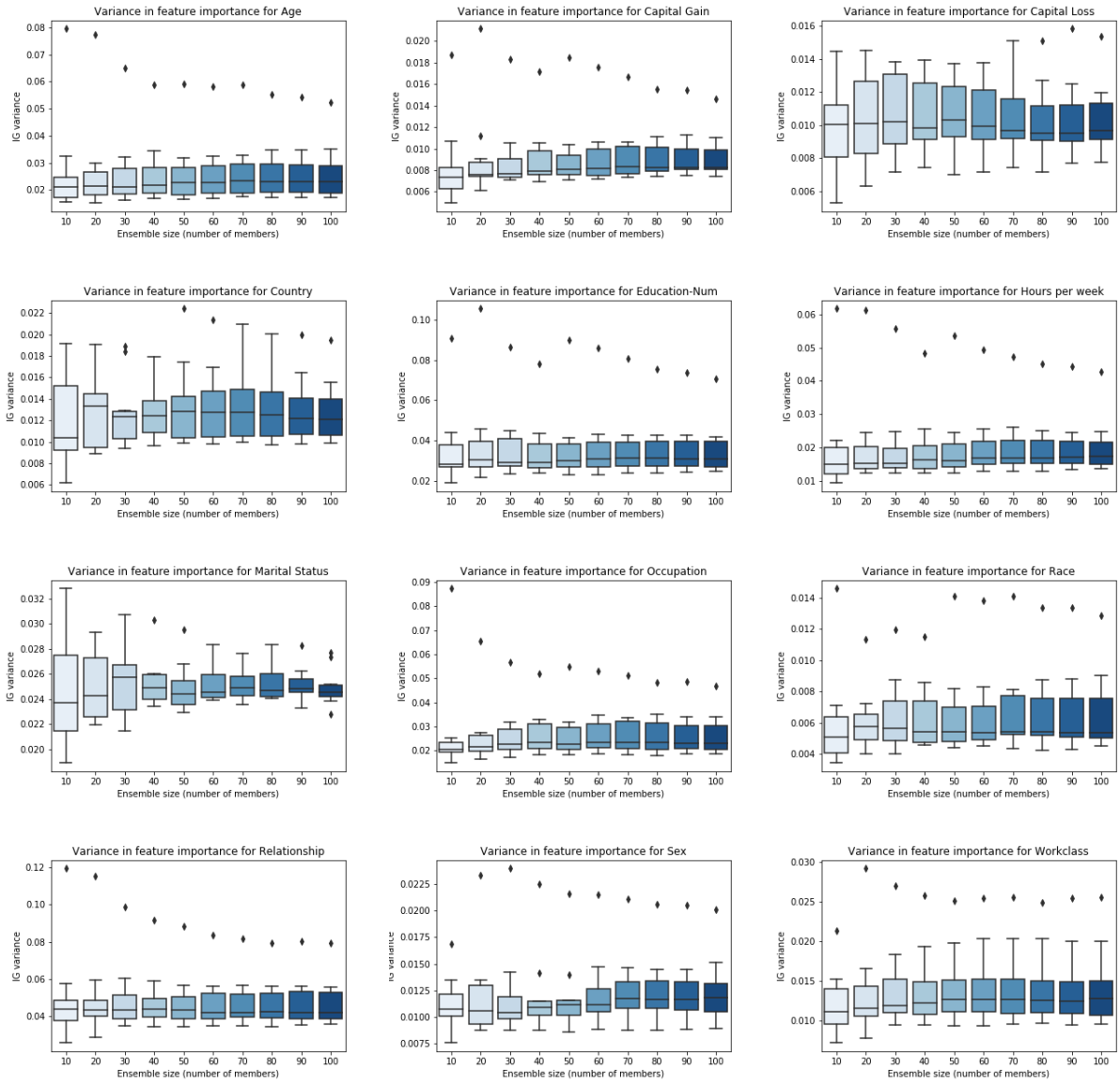


Figure 13: Average variance in feature importance scores across the Adult test set reported as a function of ensemble size. By 50 model ensembles, variance estimates for most features seem to converge. Here we have set the subsample ratio to $\frac{1}{2}$.

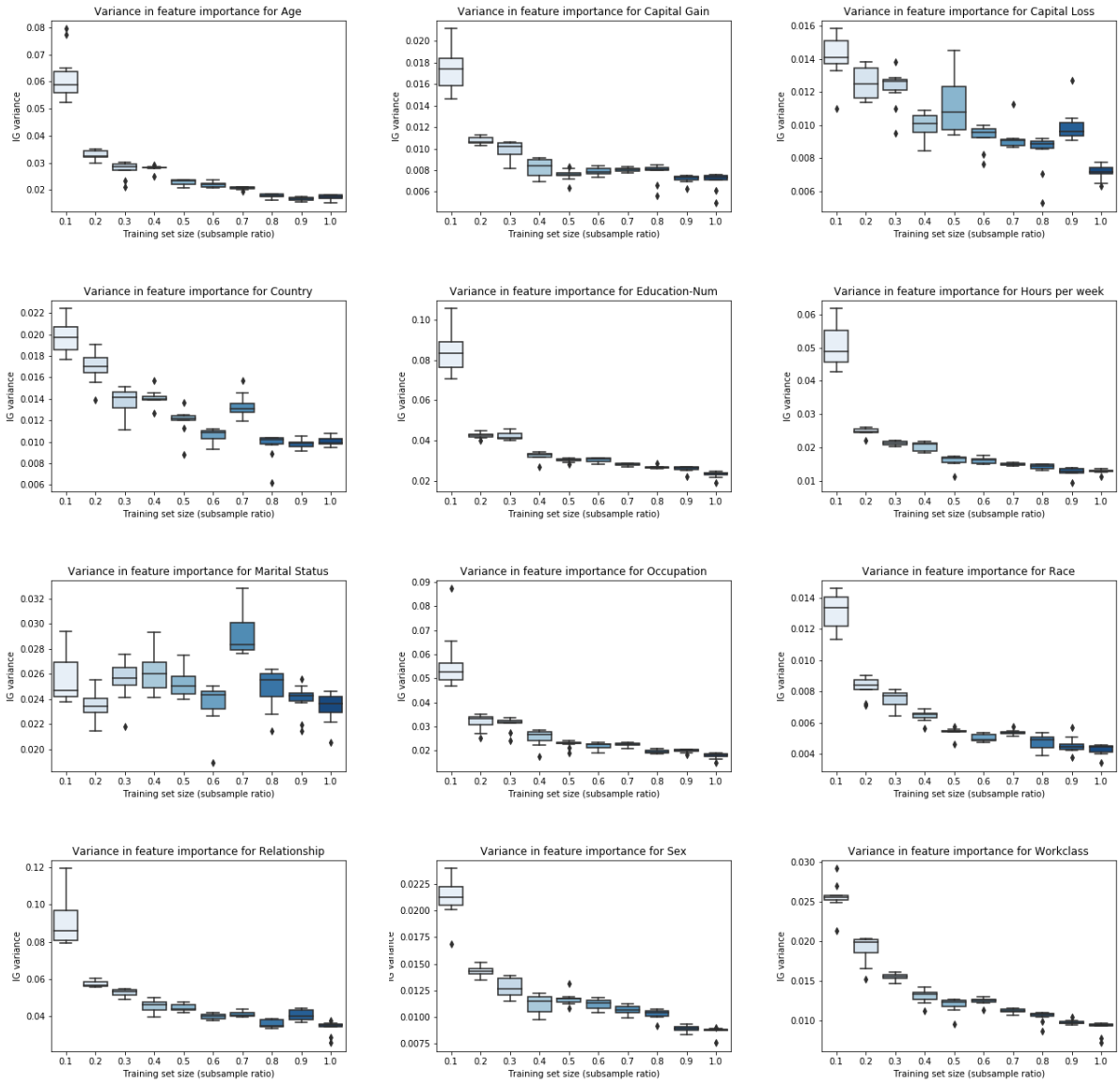


Figure 14: Average variance in feature importance scores across the Adult test set reported as a function of subsample ratio. For most features, variance estimates see to converge by the time the subsample ratio passes $\frac{1}{2}$; however, Martial Status and Capital Loss have erratic behavior likely due to outliers appearing in the subsampling procedure. Note here we use 100 models in our ensemble.