Confident Feature Ranking

Bitya Neuhof
The Hebrew University of Jerusalem, Israel

Yuval Benjamini
The Hebrew University of Jerusalem, Israel

Abstract

Machine learning models are widely applied in various fields. Stakeholders often use post-hoc feature importance methods to better understand the input features' contribution to the models' predictions. The interpretation of the importance values provided by these methods is frequently based on the relative order of the features (their ranking) rather than the importance values themselves. Since the order may be unstable, we present a framework for quantifying the uncertainty in global importance values. We propose a novel method for the post-hoc interpretation of feature importance values that is based on the framework and pairwise comparisons of the feature importance values. This method produces simultaneous confidence intervals for the features' ranks, which include the "true" (infinite sample) ranks with high probability, and enables the selection of the set of the top-k important features.

1 INTRODUCTION

Complex nonlinear prediction models are widely used to augment or even replace human judgement in fields such as healthcare (Bhardwaj et al., 2017), finance (Rundo et al., 2019), and science (Deiana et al., 2022; Li et al., 2022). Regulators, users, and developers of such models are interested in understanding the relative contribution of the different inputs, i.e., features, to the model's predictions (Preece et al., 2018; Goodman and Flaxman, 2017). Feature importance (FI) methods such as permutation feature importance (PFI) (Breiman, 2001) and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017; Lundberg et al., 2019) measure the contribution of features by estimating the effect of removing, perturbing, or permuting the feature on the predicted value or prediction loss. The specifics of this manipulation vary depending on the method and implementation (Merrick and Taly, 2020; Covert et al., 2020a). These FI methods are employed to explain the predictions of models after they have been trained, and therefore they are called *post-hoc FI methods*. This paper focuses on global FI methods that explain the average model behavior rather than local FI methods that explain individual predictions.

Recently, studies have demonstrated that post-hoc FI methods can be unstable (Molnar et al., 2020; Marx et al., 2023) due to uncertainty stemming from the size and sampling of the data used to calculate the FI values (explanation set); randomness in the perturbations, permutations (Lakkaraju et al., 2020; Agarwal et al., 2022), or approximations (Merrick and Taly, 2020); hyperparameter selection (Slack et al., 2021; Ahn et al., 2023); and more. We focus on uncertainty in sampling the explanation set, which affects the stability of the FI values. Most methods for quantifying this type of uncertainty produce per-feature spread estimates (or confidence intervals) in the FI method's output units (Ishwaran and Lu, 2019; Covert et al., 2020b; Merrick and Taly, 2020; Slack et al., 2021; Ahn et al., 2023; Molnar et al., 2021).

Existing uncertainty measures are insufficient, because stakeholders often rely on the *rank of the FI value*, rather than the value itself, in their decisions. Feature rankings are unit-independent and are therefore easy to interpret and compare across FI methods (Jaxa-Rozen and Trutnevyte, 2021; Heldt et al., 2021). Instability in the global FI values can lead to instability in their ranking (Rising, 2021) (an example is provided in Figure 1). A simple ranking of the features based on the FI values cannot reflect this uncertainty. Moreover, due to the ranking's discrete nature, existing methods for quantifying uncertainty in FI values cannot easily be modified to work for ranking uncertainty. For example, we show that confidence intervals (CIs) produced by a naive bootstrapping method based on the estimation of the ranking distribution do not cover the true ranks. The previously mentioned challenges point to the need for a framework for defin-

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

ing, estimating, and reporting ranking uncertainty. To properly model ranking uncertainty, we first model the uncertainty of the global FI values and then infer the effect of this uncertainty on the rankings.

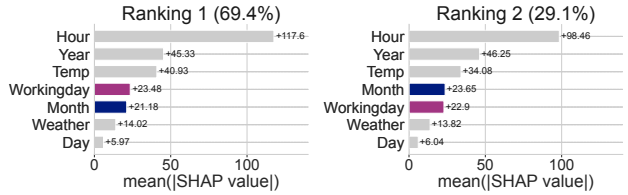


Figure 1: Bar plots of SHAP values for two samples of $n = 50$ observations from the bike sharing dataset using an XGBoost model. The ranking of the features is unstable for this sample size: ranking of the *Workingday* and *Month* features varies, depending on the sample. The chances of observing each of the rankings (69.4% and 29.1%) is estimated based on 1,000 independent samples of size $n = 50$.

In this paper, we present a *base-to-global* framework to quantify the uncertainty of global FI values. We define a two-level hierarchy of importance values, namely the *base* and *global* FI values, where the global FI values are the average of independent base FI values. Based on this framework, we propose a novel method for confidently ranking features. We define the *true rank* as a feature's ranking, obtained based on an infinite sample, for both a trained prediction model and an FI method. Our ranking method reports simultaneous CIs, ensuring, with high probability, that each feature's true rank is covered by the appropriate interval. We construct the intervals by examining all pairs of features, testing hypotheses regarding differences in means, and counting the number of rejections for each feature. The examination process tackles the multiple tests problem, which might result in the false discovery of a feature as relevant. The validity of our proposed method is demonstrated in a comprehensive evaluation on both synthetic and real-world datasets. Our findings confirm our method's effectiveness and highlight its potential in quantifying and enhancing ranking stability. Our base-to-global framework can be viewed as a generalization of the formulate, approximate, explain (FAE) (Merrick and Taly, 2020) framework for generating and interpreting Shapley-value-based FI methods. We extend the FAE concept in two respects: first, we generalize it to other post-hoc FI methods by defining the base values in a general way; and second, we address the uncertainty in the ranking of the global FI values.

Our main contributions in this paper are as follows: (1) We propose a novel ranking method for FI values. (2) We quantify the uncertainty of the ranking by providing simultaneous CIs for the features' ranks.[1] (3) We suggest an improved means of interpreting global FI values. We generalize confident ranking methods to accommodate correlations and potential departures from normality, which are common in FI values. To the best of our knowledge, our ranking method is the first to formally incorporate uncertainty control in the *ranking* of FI values.

2 QUANTIFYING UNCERTAINTY IN GLOBAL FI VALUES

2.1 Terminology

Consider the supervised learning task of predicting a real-value outcome $Y \; 2 \; Y$ from a vector of p features $X = (X_1; :::; X_p) \; 2 \; X$. A prediction model $f : X \; ! \; Y$ is trained on a training set $D_{train} = f(x_i; y_i) g_{i=1}^{M}$ and fits the data well according to standard metrics (e.g., MSE or accuracy on external test sets). Stakeholders are then interested in the extent to which a feature contributes to the model's performance or predictions – the FI value.

2.2 Base-to-Global Framework

Post-hoc global FI methods describe the average behavior of the model. These methods produce an importance value for each feature, $\hat{\Phi}_1; \hat{\Phi}_2; :::; \hat{\Phi}_p \; 2 \; R$, based on a trained model f and a sample $D_{explain} = f(x_i; y_i) g_{i=1}^{N}$, preferably independent of D_{train}. In most methods, the assumption is that a higher value of $\hat{\Phi}_j$ indicates greater importance. Generally, the features are ranked according to their FI values, and only the top-k features are considered.

In many cases, the FI values are calculated by averaging many independent runs. For example, in SHAP (Lundberg and Lee, 2017), the global FI value is an average of the absolute values assigned to each observation (the local SHAP values). Variability in the explanation set $D_{explain}$ introduces uncertainty into the global FI values.[2] In PFI (Breiman, 2001), the global FI value is the average obtained over multiple permutations. In this case, both variability in the explanation set and the randomized permutations introduce uncertainty into the global FI values.

In considering how these examples could be addressed in a single framework, we make the following observation: there is a two-level FI hierarchy in which the observed *global* FI value is an average of independent *base* FI values; in the first example (SHAP), the base FI values correspond to the local SHAP values, and in

[1] The paper's code is publicly available at: https://github.com/BityaNeuhof/confident_feature_ranking.

[2] This paper only considers the exact computation of SHAP values without approximation.

the second example (PFI), the base FI values correspond to the PFI values calculated for a single permutation on the full explanation set.

We set the following notations: matrix $\mathbf{v}_{n\ p}$ is defined as the matrix of *base* FI values, with rows $v_1;\dots;v_n\ 2\ \mathsf{R}^p$ representing the FI value for each feature.[3] $\mathbf{v}_j$ are the columns of the matrix referring to the base FI values for the j'th feature. The observed global FI value for the j'th feature is $\hat{\Phi}_j = \frac{1}{n}\ \sum\ v_{ij}$.

SHAP Example For a single observation $(x;y)$, the local SHAP value of a feature j is:

$$_j = \sum_{S\ \ [p]nfjg} \frac{jSj!(p\quad jSj\quad 1)!}{p!}$$

$$\mathsf{E}[f(X)jX_{S\,[fjg} = x_{S\,[fjg}]\quad \mathsf{E}[f(X)jX_S = x_S]$$

where $[p]$ is the set of all features, and S is a subset of features. The base FI value is:

$$v_j^{SHAP} = j\ _jj; \tag{1}$$

and the global FI value is: $\hat{\Phi}_j^{SHAP} = \frac{1}{n}\ \sum_{i=1}^{n} v_j^{SHAP}$.

PFI Example Let L be a loss function; the global PFI value of a feature j is:

$$\hat{\Phi}_j^{PFI} = \frac{1}{B} \sum_{b=1}^{B} L(f(X_{[j]}^b);Y)\quad L(f(X);Y);$$

where $X_{[j]}$ is a replication of X with a permuted version of the j'th feature, and B is the number of permutations. The base FI value can be defined either as a single permutation of the j'th feature: $v_j^{PFI} = L(f(X_{[j]});Y)\quad L(f(X);Y)$ (here the number of base FI values is the number of permutations $(n = B)$) or as the average of permutations for an observation:

$$v_j^{PFI} = \frac{1}{B} \sum_{b=1}^{B} L(f(x_{[j]}^b);y)\quad L(f(x);y); \tag{2}$$

where $x_{[j]}$ is a replication of observation x with a permuted version of the j'th feature. Here, the number of base FI values is the number of observations $(n = N)$. A detailed analysis of the sources of uncertainty in PFI is provided in Appendix A.

2.3 Uncertainty in Feature Ranking

Global FI values are often interpreted as a ranking used to highlight or select the most relevant features. Since different FI methods produce FI values of varying scales, the ranking of the features is often used to compare the methods' output. The observed ranks $\hat{r} = (\hat{r}_1;\dots;\hat{r}_p)$, $\hat{r}_j\ 2\ f1;\dots;pg$ are typically derived directly from the observed global FI values, with the rank p assigned to the highest global FI value, and rank 1 assigned to the lowest.

The sampling of the base FI values introduces uncertainty into the global FI values. The global FI values are then ranked, propagating the uncertainty into the observed ranks. This process is summarized in Figure 2 which presents the framework's pipeline for quantifying the uncertainty in the observed ranks.

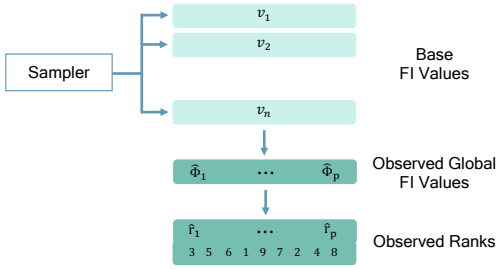


Figure 2: Base FI values are sampled as vectors, introducing uncertainty. The vectors are averaged to form the observed global FI values. Finally, the global FI values are ranked to produce the observed ranks.

Note that the definitions of the base and global FI levels specify the source of the uncertainty to be reflected in the CIs for the ranks; two options for base FI value definition for PFI are presented above.

3 STATISTICAL MODEL AND INFERENCE GOAL

In this section, we describe our statistical model for estimating ranking uncertainty. First, we define a feature's rank-set as a rank that considers information about ties. Then we define our inference goal – to provide simultaneous CIs for the rank-sets. Finally, we discuss the advantages of simultaneous CIs and provide an example in which the top-k features are highlighted.

3.1 True Global FI Values and Rank-Sets

Recall that in Section 2 we introduced the base FI values matrix $\mathbf{v}$, with rows $v_i\ 2\ \mathsf{R}_p$. Here, we model the rows v_i as independent samples from distribution F_v with mean vector $E[v_i] = (\Phi_1;\dots;\Phi_p)$; these are the *true global FI values*. Each observed global FI value $\hat{\Phi}_j$ is an unbiased but noisy version of Φ_j. We are interested in understanding the effects of this variability on the possible feature rankings.

In contrast to the observed noisy ranks, the *true ranks* $r_1;\dots;r_p$ are based on the true global FI values $\Phi_1;\dots;\Phi_p$. Whereas in the observed global FI values

[3] If the base FI values are the local values, $n = N$ is the size of $\mathcal{D}_{explain}$.

exact ties are unlikely, for the true global FI values we can imagine ties between equivalent features, or we may want to allow an indifference level. We follow Al Mohamad et al. (2022) in redefining the true ranks to account for ties:

Definition 1 *(Rank-Set) Define the lower rank of* Φ_j *as* $l_j = 1 + \#\{k : \Phi_j > \Phi_k;\ j \neq k\}$ *and the upper rank of* Φ_j *as* $u_j = p - \#\{k : \Phi_j < \Phi_k;\ j \neq k\}$. *Then the rank-set of* Φ_j *is the set of natural numbers* $\{l_j; l_{j+1}; \ldots; u_j\}$ *denoted as* $[l_j; u_j]$.

If there are no ties, the lower and upper ranks are identical and equal to the standard definition. In the remainder of the paper, the term 'true rank' will refer to the rank-set.

3.2 CIs for True Ranks

We propose quantifying the rankings' uncertainty using simultaneous CIs for the true ranks. Here, we define marginal and simultaneous CIs, and in Section 4 we propose a method for constructing valid simultaneous CIs for the true ranks.

Definition 2 *(CI for a True Rank) The ranks interval* $[L_j; U_j]$ *is an* $(1 - \alpha)$*-level CI for a true rank of the* j*'th feature if* $\mathsf{P}_{F_v}([l_j; u_j] \subseteq [L_j; U_j]) \geq 1 - \alpha$ *for any possible* F_v.

$L_j; U_j$ are functions of $\mathbf{v}$, the matrix of observed base FI values. We note that different sets of observed base FI values will produce different CIs.

The set of intervals $[L_1; U_1]; \ldots; [L_p; U_p]$ has marginal coverage if each interval is a valid CI of the corresponding true rank. For ranking and selection of features, marginal coverage is not sufficient, because it does not support selection after ranking (Benjamini and Yekutieli, 2005). Therefore, our ranking method constructs simultaneous CIs for the true ranks.

Definition 3 *(Simultaneous Coverage) The set of intervals* $[L_1; U_1]; \ldots; [L_p; U_p]$ $[1; p]$ *has simultaneous coverage at level* $1 - \alpha$ *if*

$$\mathsf{P}_{F_v}([l_j; u_j] \subseteq [L_j; U_j];\ \forall j \in \{1; \ldots; p\}) \geq 1 - \alpha.$$

In $1 - \alpha$ simultaneous CIs, the probability that all intervals cover the true ranks is at least $1 - \alpha$. Simultaneous CIs remain valid for any form of selection after ranking (for example, selection of the most important features). We note that simultaneous coverage is conservative and can result in relatively large intervals.

3.3 Top-K Set

Here we present an application of simultaneous CIs for the selection of the most important features (top-k) with a guarantee of coverage. Since the ranking is based on the observed FI values, the size of the set of possible top-k features might be greater than k.

Denote T_k $[p]$ as the set of features whose true FI value is ranked in the top-k $T_k = \{j : u_k$ p $k+1\}$. A simple selection method is to select features for which the upper bound of the CI is greater than $p - k$. With simultaneous coverage, the probability of an error for this selection is controlled (Hsu, 1996). Furthermore, the CIs for the features currently ranked among the top-k still have marginal coverage. These two properties are not guaranteed without simultaneous coverage (Ein-Dor et al., 2006).

Lemma 1 *Let* $\{[L_1; U_1]; \ldots; [L_p; U_p]\}$ *be* $1 - \alpha$ *simultaneous CIs for the true ranks. Define the* top-k set $\hat{T}_k = \{j : U_j$ p $k + 1\}$. *This set includes all features with an upper bound in the top-*k *ranks* $(p; p - 1; \ldots; p - k + 1)$. *Then* $\mathsf{P}(T_k$ $\hat{T}_k) \geq 1 - \alpha$.

To prove this, consider a case in which T_k $\hat{T}_k$ does not hold; then it must follow that there is some $j \in T_k$ that is not in $\hat{T}_k$. This means that the estimated upper bound U_j is less than the true upper rank u_j, so the CI $[L_j; U_j]$ does not cover $[l_j; u_j]$. Based on the definition of $1 - \alpha$ simultaneous CIs, the probability of any such event is at most α:

4 CONFIDENT SIMULTANEOUS FEATURE RANKING

In this section, we introduce our ranking method which is designed to rank FI values while taking into account the uncertainty associated with the post-hoc FI method and the sampling process. Using our base-to-global framework, we are able to quantify the uncertainty by calculating simultaneous CIs for the true ranks.

4.1 Feature Ranking

Our method uses pairwise hypothesis tests to estimate lower and upper bounds for the true rank of each feature. For each feature pair $j; k$, we perform two one-sided hypothesis tests:

(a) A test of $H^1_{jk} : \Phi_j < \Phi_k$ versus $H^0_{jk} : \Phi_j$ Φ_k;

(b) A test of $H^1_{kj} : \Phi_k < \Phi_j$ versus $H^0_{kj} : \Phi_k$ Φ_j.

Each test is composed of a p-value $p_{jk} = \mathit{pairCompare}(\mathbf{v}_j; \mathbf{v}_k)$ and a significance level $\alpha \in (0; 0.5]$; the test rejects H^0_{jk} if $p_{jk} < \alpha$. The test is *calibrated* if: $\mathsf{P}(p_{jk} \leq \alpha) \leq \alpha$ for any Φ_j Φ_k, meaning that the probability of rejecting H^0_{jk} when H^1_{jk} is correct is bounded by α. For the tests to be calibrated, they need to account for the possible dependence between $\mathbf{v}_j$ and $\mathbf{v}_k$. In our implementation, we use the paired-sample t-test (see Section 4.4).

There is a natural relation between the results of the one-sided hypothesis tests and the ranking of the global FI values. The rejection of a hypothesis $H^0_{jk} : \Phi_j \quad \Phi_k$, implies acceptance of a *partial ranking* of the global FI values, namely $\Phi_j < \Phi_k$. This partial ranking limits the global ranking: the feature j cannot be ranked highest ($u_j < p$), and the feature k cannot be ranked lowest ($l_k > 1$). We combine many partial ranking statements to improve the bounds on the ranks.

Combining many probabilistic decisions comes at a price. Setting the tests' rejection threshold to limits the marginal probability of making an error in each partial ranking to . However, when combined, the probability of making at least one error increases with the number of tests, and without proper adjustments it may greatly exceed . In the next subsection, we adjust the p-values to control this probability over multiple tests.

4.2 Controlling Partial Ranking Error

We define D as the set of partial rankings from all pairwise tests $D = f(j;k) : H^0_{jk}$ was rejected$g$. A partial ranking error is a pair $(j^0;k^0) \in D$ for which $\Phi_{j'} \quad \Phi_{k'}$. For simultaneous CIs, we want to control the probability of error over all the partial rankings.

Definition 4 *(Family-Wise Error Rate) The family-wise error rate (FWER) is controlled at probability level on the set of partial rankings D if the probability of making any partial ranking error is less than : $P(\exists(j^0;k^0) \in D : \Phi_{j'} \quad \Phi_{k'})$.*

To control the FWER, we replace the original p-values with a set of adjusted p-values $\mathbf{p}^{adj} = FWERAdjust(\mathbf{v};pairCompare)$. After adjustment, the partial rankings are obtained by comparing p^{adj}_{jk} and p^{adj}_{kj} to the required FWER level :[4] Some examples of adjustment procedures in which the FWER is controlled are provided in Section 4.4.

4.3 Confident Simultaneous Feature Ranking

When the FWER is controlled for the partial rankings set, we can use the partial rankings set to derive simultaneous CIs for the true ranks:

Theorem 1 *(Al Mohamad et al., 2022) Let D be the set of partial rankings with FWER control at level . For $j = 1;:::;p$, define:*

$$L_j = 1 + \#fk : (k;j) \in Dg;$$
$$U_j = p \quad \#fk : (j;k) \in Dg:$$

Then the sets $f[L_j;U_j]$ for $j \in [p]g$ are (1) simultaneous CIs for the true ranks.

The construction naturally extends the definition of rank-set provided in Definition 1. The idea of the proof is that a coverage failure means that the set of true (one-sided) differences is smaller than the set of observed (one-sided) differences. This means that at least one partial ranking in D is false. Therefore the FWER upper bounds the probability of an error in the CIs (see proof in Appendix B.1). Our ranking method is based on ICRanks (Al Mohamad et al., 2022); the way in which the proposed method differs from ICRanks is discussed in Section 4.4.

Algorithm 1 summarizes our method for constructing simultaneous CIs for the true ranks. The algorithm works directly on the base FI matrix without requiring access to the trained model, the FI method, or the explanation set. The main assumption is that our paired test is calibrated for the possible distributions of base FI values.

Algorithm 1 Simultaneous CIs for Ranks

Require:
$\mathbf{v}$: base FI matrix;
1 > 0: level of confidence;
$pairCompare$: suitable paired test;
$FWERAdjust$: FWER adjustment procedure.
for $j;k \in [p]; j \neq k$ **do**
$p_{jk} \quad pairCompare(\mathbf{v}_j;\mathbf{v}_k)$.
end for
$\mathbf{p}^{adj} = FWERAdjust(\mathbf{v};pairCompare)$
$D \quad f(j;k) : p^{adj}_{jk} \quad g$
for $j \in [p]$ **do**
$L_j \quad 1 + \#fk : (k;j) \in Dg$
$U_j \quad p \quad \#fk : (j;k) \in Dg$
end for
return $[L_1;U_1];:::;[L_p;U_p]$.

4.4 Ranking Method Implementation

Paired Test We use a parametric paired t-test to compute p-values for the pairs of base FI values. Set $\mathbf{d} = \mathbf{v}_j \quad \mathbf{v}_k$ to be the vector of differences, and denote $\bar{d}$ as the sample average and s_d as the sample standard deviation. Then the one-sided level test rejects the null hypothesis if $\bar{d}=(s_d=\sqrt{n}) > T_{n\ 1}(1 \quad)$, where $T_{n\ 1}(1 \quad)$ denotes the 1 quantile of student-t (n 1 df). The paired t-test is fairly robust to departures from a normal distribution (Posten, 1979).

Adjustment for Multiple Tests We implement two sequential procedures to adjust (increase) the p-values:

- Holm's procedure (Holm, 1979). Assuming that the base FI values are normally distributed, the

[4] In practice, when both tests use the same data and the threshold is less than 0:5, none or just one of the null hypotheses will be rejected (there will not be a case in which both of the null hypotheses are rejected).

paired t-test is calibrated with this procedure. We implement Holm's procedure on the one-sided hypothesis tests, although for approximately normal data, the two-sided Holm would likely work well also. See Shaffer (1980, 1995) on using Holm's procedure for pairwise tests.

- Min-P (Westfall and Young, 1993). This procedure is based on bootstrapping. Therefore, no further assumptions are required.

The adjusted p-values are then compared to the predefined threshold of (details on the procedures are provided in Appendix B.2). With these procedures, if the p-values are calibrated, then the FWER for the rejected tests is controlled at the level, regardless of the dependence.

Comparison to ICRanks Similar to Algorithm 1, ICRanks (Al Mohamad et al., 2022) is based on Tukey's correction (Tukey, 1953) in order to control the differences between ranks simultaneously. Tukey's correction is designed for normal and independent data, which are distributional assumptions that would usually not hold for FI values. In contrast, our algorithm applies a test to each feature pair separately, and the Holm's or the Min-P procedure is performed on the resulting p-values; therefore, it can be utilized with robust or nonparametric tests (Wilcox, 2011).

5 EVALUATION

In this section, we evaluate our base-to-global framework and ranking method. We use synthetic data to assess our method's validity (simultaneous coverage) and efficiency. We analyze our ranking method by generating base FI values directly (Section 5.1). We note that feature ranking is an interpretability step at the end of an ML task, as shown in Figure 3; therefore, we simulate the entire process of training and explaining a model with simulated data (Section 5.2) and real data (Section 5.3).

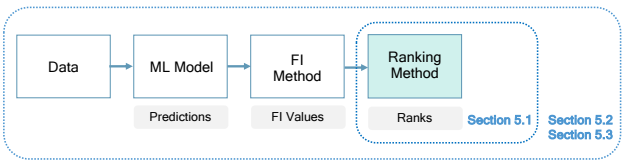


Figure 3: Feature ranking and evaluation process.

Metrics We use the metrics (ranking measures) suggested by Al Mohamad et al. (2022) to define simultaneous coverage and efficiency:

- *Simultaneous coverage* – the proportion of experiments where all true ranks are covered by their CIs: *one* if all $f\Phi_j \quad 2 \quad [L_j; U_j]g;$ *andzero* otherwise.
- *Efficiency* – the average relative size of the CIs: $\frac{1}{p\ (p\ \ 1)} \sum_{j=1}^{p} (U_j \quad L_j)$.

Higher coverage and lower efficiency are better.

5.1 Ranking Method Comparison

Ranking Methods We compare the ranking measures of four ranking methods: a naive ranking method based on empirical quantiles of bootstrap samples as a baseline (details are provided in Appendix D.1.1), ICRanks,[5] our ranking method with Holm's procedure, and our ranking method with the Min-P adjustment procedure.

We sample the base FI values from a multivariate-normal distribution $N_p(\ ; \Sigma)$ with predetermined vector of means and a covariance matrix Σ. The true global FI values are the means, and we control the correlation structure between the base FI values via the definition of the covariance matrix.

Vector of Means The structure of the vector is $(1^{\text{-exponent}}; 2^{\text{-exponent}}; \ldots; (p+1)^{\text{-exponent}})^T$, with -exponent 2 $[0{:}1; 0{:}25; 0{:}5]$. A lower value of -exponent results in a more dense vector of means. Ties between the means are allowed.

Covariance Matrix The covariance matrix structure is composed of a vector ^{2} of the variances of the base FI values sampled from the re-scaled chi-squared distribution ($^2_{(5)}$=5). The correlation matrix structure can be one of three structures: identity (no correlations), block-wise pairs, or equal correlations with 2 $[0{:}1; 0{:}5; 0{:}9]$. In addition, we vary the level of noise in the base FI values by scaling the vector by -factor 2 $[0{:}2; 1; 5]$.

We analyze the ranking measures for multiple conditions of the vector of means ( ) and the correlation matrix (Σ) (a total of 486 conditions). The number of features p is one of: $[10; 30; 50]$, and the number of base FI values n is one of: $[100; 300; 1000]$. We sample 100 independent explanation sets for each configuration and report the average ranking measures across the repetitions. Below, we present the results for $p = 30$, -exponent=0.25, and equal correlations. Additional results are presented in Appendix D.1.

Simultaneous Coverage In the naive ranking method, simultaneous coverage is not maintained in all conditions. All other methods maintain simultaneous coverage levels of almost 100%; this indicates that they are overly conservative compared to the nominal required simultaneous coverage of 90%.

Efficiency Without correlations, efficiency degrades as the -factor increases, with almost no difference ob-

[5] ICRanks package

served between the methods (Figure 4). With correlations, as increases our method becomes more efficient than ICRanks, with the Min-P adjustment seen to be slightly more efficient than Holm's procedure. The gap between the methods increases as the -factor increases (Figure 5).

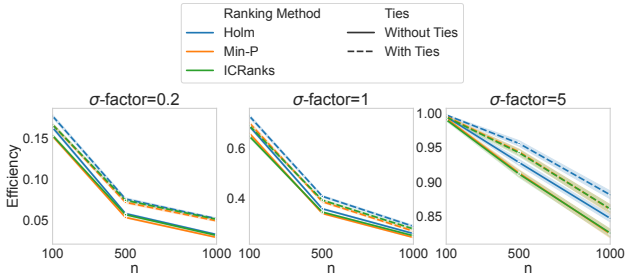
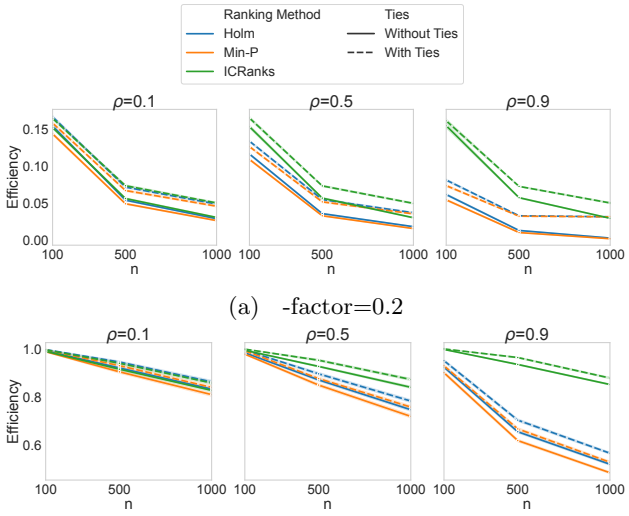


Figure 4: Ranking efficiency as a function of n for multiple -factors and three ranking methods. Low values mean smaller sets and are therefore better. The methods' efficiency is similar.



(a) -factor=0.2

(b) -factor=5

Figure 5: Ranking efficiency with low (a) and high (b) -factors, as a function of n for multiple levels of correlations ( ) and three ranking methods.

5.2 SHAP Ranking Measures

Here, we simulate the entire ML process as described in Figure 3 with simulated data. We analyze the ranking measures and runtime of our ranking method with Min-P and Holm's procedures compared to ICRanks.

Data Generating Process (DGP) We follow the DGP of Ishwaran and Lu (2019). We sample a data matrix X with independent uniformly distributed features and calculate Y as a function of X with noise. We define two functions:

(A) $y = 10\sin(\ x_1x_2) + 20(x_3 \quad 0{:}5)^2 + 10x_4 + 5x_5 + \ ;$
$fx_jg \quad U(0;1); \quad N(0;1)$:

(B) $y = (x_1^2 + [x_2x_3 \quad (x_2x_4) \ ^1]^2)^{0:5} + \ ;$
$x_1 \quad U(0;10); x_2 \quad U(\ ;2 \ ); x_4 \quad U(1;5);$
all other features $fx_jg \quad U(0;1); \quad N(0;1)$:

The definitions of $(X;Y)$ are for $p = 10$ features. We simulate a larger number of features by defining the functions for cycles of 10 features. For example, in function (a), $X_{11} \quad U(0;10)$ and is added to Y as X_{11}^2.

For this simulation, we sample a large data matrix $X_{M \ p}$ $(M = 500k)$, calculate Y as a function of X with noise, and train a prediction model on $D_{train} = (X;Y)$. We calculate the global FI values $\hat{\Phi}_1;:::\hat{\Phi}_p$ based on a sufficiently large sample $(n = 1M)$, making it a low variance estimator of $\Phi_1;:::\Phi_p$ (Slack et al., 2021). We generate multiple simulated datasets, varying the number of features (p) and base FI values (n), the DGP, and the prediction models. We sample 100 independent explanation sets for each evaluation configuration to measure the ranking efficiency, simultaneous coverage, and runtime.

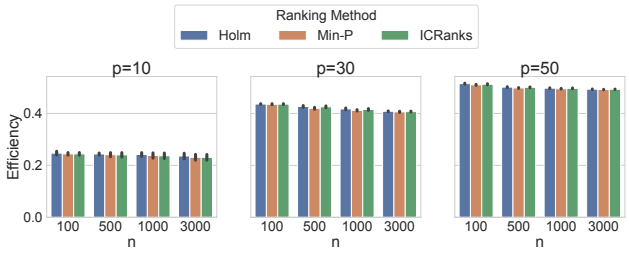
Below we present the results for the DGP-A with a random forest (RF) model (Breiman, 2001) and DGP-B with an XGBoost (XGB) model (Chen and Guestrin, 2016) (see Appendix D.2 for the complete results). We use TreeSHAP (Lundberg et al., 2019) to compute the base FI values, relying on the definition of base and global FI values presented in Section 2 (Equation 1). To calculate the ranking measures, we repeatedly sample $D_{explain}$ independent of D_{train}.

Simultaneous Coverage All of the examined methods maintain simultaneous coverage levels of almost 100% in all simulated conditions. However when the base FI values have an extremely long tail, simultaneous coverage is not guaranteed (an example is provided in Appendix D.2.3).

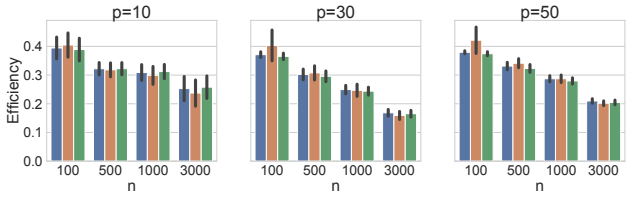
Efficiency We can see that ICRanks is comparable to our ranking method. The ranking efficiency improves as n increases (Figure 6). For the XGB model, we see that the efficiency of our method with the Min-P procedure is worse than that of our method with Holm's procedure for low n values; the Min-P procedure recalibrates the p-values based on resampled data, which is an inefficient process when n is low.

Runtime We analyze the runtime of TreeSHAP (computation of base FI values) and the ranking times (ICRanks, Holm's procedure, and the Min-P procedure). The runtime of ICRanks and Holm's procedure is ten times faster than the runtime of TreeSHAP. The Min-P procedure requires B repetitions (boot-

strap samples) of the pairwise tests, so the runtime increases with B. Details are provided in Appendix D.2.2.



(a) RF example with DGP-A



(b) XGB example with DGP-B

Figure 6: Ranking efficiency as a function of n for different numbers of features (p) and ranking methods. The efficiency improves as n increases.

5.3 Real Data Experiments

5.3.1 Ranking Stability

We demonstrate the use of our ranking method and present the simultaneous CIs produced with the bike sharing dataset (Fanaee-T and Gama, 2014), the TreeSHAP FI method, and Holm's procedure. We create 60/40 train/test splits, fit an XGB regression model (default hyperparameters) to the training set (R^2 = 0:98), and evaluate the performance on the test set (R^2 = 0:94). Then, we calculate the base FI values for n = 50 and n = 1000 by sampling from the test set. Presenting the CIs for the ranks enables us to compare the stability for different sizes of n (see Figure 7). The triangles within the CIs are the observed global FI values. The process of constructing the CIs for n = 50 base FI values is described in Appendix C.

5.3.2 Training Stability

Our base-to-global framework can also be used to quantify the uncertainty in training stemming from the sampling of the training set. Here, we use the COMPAS dataset (Angwin et al., 2016), an RF classification model (default hyperparameters), the PFI method, and the Min-P procedure. We define the base FI values as global PFI values. Each trained model produces a base FI vector v_i; the global FI values are obtained by resampling and training multiple equivalent models (with the same hyperparameters and size of D_{train} (M = 3K) and similar training accuracy (0:883 0:005)). We use the same explanation set

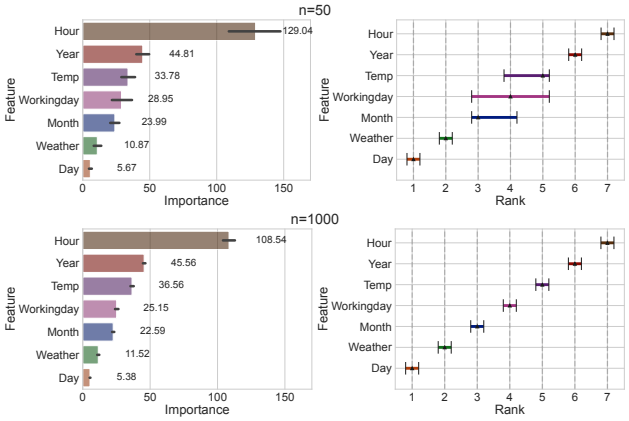


Figure 7: Global SHAP values (left) and CIs for the true ranks (right) for the bike sharing dataset. The importance values were obtained from 50 (top) and 1,000 (bottom) observations. The CIs point out uncertain feature rankings for a small sample size.

(N = 600) to calculate the importance values. Figure 8 presents the true ranks' CIs for two values of n. The observed uncertainty in the ranking for n = 10 indicates that the randomness in sampling can influence the learned mapping between the features and the target variable.

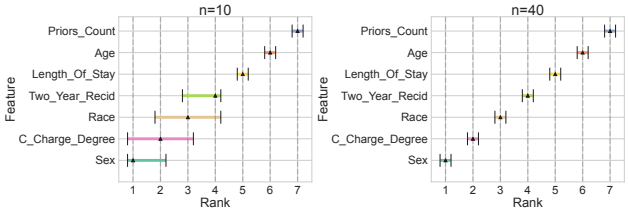


Figure 8: CIs for the true ranks for n = 10 (left) and n = 40 (right) trained models. The CIs present uncertainty in training a model based on M = 3K training observations.

5.3.3 High-Dimensional Data

In previous sections (Sections 5.1 and 5.2), we analyzed the validity of our ranking method in multiple settings, including with moderately high-dimensional data (p = 50), and showed that our method maintains simultaneous coverage. Now we demonstrate the use of our ranking method with high-dimensional data, utilizing the Nomao dataset (Candillier and Lemaire, 2012), which consists of 118 input features and a binary target variable. We create 60/40 train/test splits, fit an XGB classification model (with the default hyperparameters) to the training set ($accuracy$ = 0:99), and evaluate the performance on the test set ($accuracy$ = 0:97). Then, we calculate the base FI values with TreeSHAP; the distribution of the global FI values is shown on the left side of Figure 9. Thirty-

one features have importance values of zero.

We use our ranking method to rank all the features from the least important (1) to the most important (118); the full ranking is presented in Appendix D.3 (Figure 18). As 31 features have the same importance value of zero, the CI of each feature is [1; 31]. In such a case, if we measure the efficiency of the ranking across all features, the long CIs of the irrelevant features will affect it. The skewness of the CI length is presented on the right of Figure 9. A model will likely use some features, and the unused features will get a low importance value (or a value of zero); a filtering step is required to improve the ranking process by comparing only the relevant features.

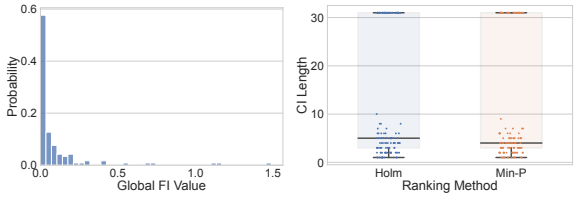


Figure 9: Global SHAP FI value distribution (left) and length of the CIs for the ranks (right) for the Nomao dataset features. The FI value of many features is zero, and the ranking is accordingly inefficient.

6 RELATED WORK

6.1 Uncertainty in Feature Ranking

Ordering features in terms of their importance to the model's prediction is referred to as feature ranking. It is often infeasible to determine the "right" order or perfect subset of features, because it requires the examination of all possible feature subsets (Prati, 2012). Many studies suggested overcoming this limitation and obtaining a more stable ranking by applying a two-step procedure in which multiple rankings are generated and the outputs are combined into a single ranking (Saeys et al., 2008). Each of the rankings generated in the first step is obtained by ordering the global FI values, which are produced using different FI methods (Schulz et al., 2021) or by resampling the data and ranking the output FI values using a single FI method (Vettoretti and Di Camillo, 2021; Alaiz-Rodriguez and Parnell, 2020; Salman et al., 2022). The process of combining all of the rankings into a single ranking is sometimes based on voting (Vettoretti and Di Camillo, 2021; Schulz et al., 2021), pairwise comparisons of the rankings (Prati, 2012; Salman et al., 2022), or other techniques (Alaiz-Rodriguez and Parnell, 2020). In all of the techniques mentioned above, more stable ranking is achieved by aggregating multiple global scores or rankings, a process that is computationally expensive and requires many explanation sets or FI methods. In contrast, our ranking method produces a stable ranking based on a single FI method and explanation set.

6.2 Ranking and Selection

The problem of ranking and selection (R & S) of items has been well studied by researchers in the field of statistics, and various solutions have been suggested (Gupta, 1965; Boesel et al., 2003). Some studies focused on ranking items based on noisy data and pairwise comparisons (Wright et al., 2014; Valdeira and Soares, 2022). Other studies proposed methods that look for the best item (Eckman et al., 2020), select the set of top- or lowest-ranked items, or, most similar to our work, methods that rank all items based on the observed means (Zhang et al., 2014; Klein et al., 2020; Wright et al., 2014). After ranking all of the items, a subset of items might be selected. (Al Mohamad et al., 2022; Rising, 2021). Other researchers have proposed methods like ours that deal with the effect of multiple tests and examine how to control the FWER and increase the probability that the correct items are selected (Garcia and Herrera, 2008; Holm, 2013).

7 CONCLUSIONS

We propose a base-to-global framework and a method for constructing CIs for the true ranks of the global FI values. Because rankings are frequently used to summarize FI methods' output, it is crucial to consider the rankings' stability. Our method can be used with robust and nonparametric paired-tests to support non-standard FI distributions.

We present a rigorous criterion for quantifying uncertainty that can be explicitly modeled (e.g., the explanation set size). We view the proposed method as a step toward producing new forms of stability assessments for explainable ML. In future research, we aim to address other sources of instability, such as the difference between FI methods, although their effect is more challenging to quantify.

Finally, our current algorithm is conservative, as demonstrated in simulations where the coverage level surpasses the requested (1-)%. Future research will also be aimed at narrowing the CIs while maintaining nominal coverage and reducing the impact of the number of features on coverage by using a filtering step (e.g., eliminating non-important features).

Acknowledgements

We thank Noga H. Rotman for valuable discussions and advice; we thank the AISTATS anonymous reviewers and the area chair for their comments that helped improve this manuscript. This work was supported by the Israel Science Foundation and the Center for Interdisciplinary Research at Hebrew University.

References

Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., and Lakkaraju, H. (2022). Rethinking stability for attribution-based explanations. *arXiv preprint arXiv:2203.06877*.

Ahn, S., Grana, J., Tamene, Y., and Holsheimer, K. (2023). Local model explanations and uncertainty without model access. *arXiv preprint arXiv:2301.05761*.

Al Mohamad, D., Goeman, J. J., and van Zwet, E. W. (2022). Simultaneous confidence intervals for ranks with application to ranking institutions. *Biometrics*, 78(1):238–247.

Alaiz-Rodriguez, R. and Parnell, A. C. (2020). An information theoretic approach to quantify the stability of feature selection and ranking algorithms. *Knowledge-Based Systems*, 195:105745.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. propublica, may 23, 2016.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81.

Bhardwaj, R., Nambiar, A. R., and Dutta, D. (2017). A study of machine learning in healthcare. In *2017 IEEE 41st annual computer software and applications conference (COMPSAC)*, volume 2, pages 236–241. IEEE.

Boesel, J., Nelson, B. L., and Kim, S.-H. (2003). Using ranking and selection to "clean up" after simulation optimization. *Operations Research*, 51(5):814–825.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Candillier, L. and Lemaire, V. (2012). Nomao. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C53G79.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Covert, I., Lundberg, S., and Lee, S.-I. (2020a). Explaining by removing: A unified framework for model explanation. *arXiv preprint arXiv:2011.14878*.

Covert, I., Lundberg, S., and Lee, S.-I. (2020b). Understanding global feature contributions through additive importance measures. *arXiv preprint arXiv:2004.00668*.

Deiana, A. M., Tran, N., Agar, J., Blott, M., Di Guglielmo, G., Duarte, J., Harris, P., Hauck, S., Liu, M., Neubauer, M. S., et al. (2022). Applications and techniques for fast machine learning in science. *Frontiers in big Data*, 5:787421.

Eckman, D. J., Plumlee, M., and Nelson, B. L. (2020). Revisiting subset selection. In *2020 Winter Simulation Conference (WSC)*, pages 2972–2983. IEEE.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Ein-Dor, L., Zuk, O., and Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, 103(15):5923–5928.

Fanaee-T, H. and Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127.

Garcia, S. and Herrera, F. (2008). An extension on" statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of machine learning research*, 9(12).

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, 7(2):225–245.

Heldt, F. S., Vizcaychipi, M. P., Peacock, S., Cinelli, M., McLachlan, L., Andreotti, F., Jovanović, S., Dürichen, R., Lipunova, N., Fletcher, R. A., et al. (2021). Early risk assessment for covid-19 patients from emergency department data using machine learning. *Scientific reports*, 11(1):4200.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

Holm, S. (2013). *Confidence intervals for ranks*.

Hsu, J. (1996). *Multiple comparisons: theory and methods*. CRC Press.

Ishwaran, H. and Lu, M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in medicine*, 38(4):558–582.

Jaxa-Rozen, M. and Trutnevyte, E. (2021). Sources of uncertainty in long-term global scenarios of solar photovoltaic technology. *Nature Climate Change*, 11(3):266–273.

Klein, M., Wright, T., and Wieczorek, J. (2020). A joint confidence region for an overall ranking of populations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(3):589–606.

Lakkaraju, H., Arsov, N., and Bastani, O. (2020). Robust and stable black box explanations. In *International Conference on Machine Learning*, pages 5628–5638. PMLR.

Li, Z., Yoon, J., Zhang, R., Rajabipour, F., Srubar III, W. V., Dabo, I., and Radlińska, A. (2022). Machine learning in concrete science: applications, challenges, and best practices. *npj Computational Materials*, 8(1):127.

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2019). Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*.

Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Marx, C., Park, Y., Hasson, H., Wang, Y., Ermon, S., and Huan, L. (2023). But are you sure? an uncertainty-aware perspective on explainable ai. In *International Conference on Artificial Intelligence and Statistics*, pages 7375–7391. PMLR.

Merrick, L. and Taly, A. (2020). The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 17–38. Springer.

Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N., and Bischl, B. (2021). Relating the partial dependence plot and permutation feature importance to the data generating process. *arXiv preprint arXiv:2109.01433*.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B. (2020). General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pages 39–68. Springer.

Posten, H. O. (1979). The robustness of the one-sample t-test over the pearson system. *Journal of Statistical Computation and Simulation*, 9(2):133–149.

Prati, R. C. (2012). Combining feature ranking algorithms through rank aggregation. In *The 2012 international joint conference on neural networks (IJCNN)*, pages 1–8. Ieee.

Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018). Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*.

Rising, J. (2021). Uncertainty in ranking. *arXiv preprint arXiv:2107.03459*.

Rundo, F., Trenta, F., di Stallo, A. L., and Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. *Applied Sciences*, 9(24):5574.

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19*, pages 313–325. Springer.

Salman, R., Alzaatreh, A., and Sulieman, H. (2022). The stability of different aggregation techniques in ensemble feature selection. *Journal of Big Data*, 9(1):1–23.

Schulz, J., Poyiadzi, R., and Santos-Rodriguez, R. (2021). Uncertainty quantification of surrogate explanations: An ordinal consensus approach. *arXiv preprint arXiv:2111.09121*.

Shaffer, J. P. (1980). Control of directional errors with stagewise multiple test procedures. *The Annals of Statistics*, 8(6):1342–1347.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584.

Slack, D., Hilgard, A., Singh, S., and Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34:9391–9404.

Tukey, J. W. (1953). The problem of multiple comparisons. *Multiple comparisons*.

Valdeira, F. and Soares, C. (2022). Ranking with confidence for large scale comparison data. *arXiv preprint arXiv:2202.01670*.

Vettoretti, M. and Di Camillo, B. (2021). A variable ranking method for machine learning models with correlated features: in-silico validation and application for diabetes prediction. *Applied Sciences*, 11(16):7740.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic press.

Wright, T., Klein, M., and Wieczorek, J. (2014). Ranking populations based on sample survey data. *Statistics*, page 12.

Zhang, S., Luo, J., Zhu, L., Stinchcomb, D. G., Campbell, D., Carter, G., Gilkeson, S., and Feuer, E. J. (2014). Confidence intervals for ranks of age-adjusted rates across states or counties. *Statistics in Medicine*, 33(11):1853–1866.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes, see Section 4.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **No. In this paper, we aim to improve the validity and efficiency of the ranking compared to other methods, not the complexity.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes, in the supplemental material.**
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes, in Sections 3 and 4.**
   (b) Complete proofs of all theoretical results. **Yes, in the supplemental material.**
   (c) Clear explanations of any assumptions. **Yes, in Sections 3 and 4.**
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes, in the supplemental material.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes, in Section 5 and in the supplemental material.**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes, in Section 5 and in the supplemental material.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes, in the supplemental material.**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes, in the references, footnotes, and supplemental material.**
   (b) The license information of the assets, if applicable. **Not Applicable.**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable.**
   (d) Information about consent from data providers/curators. **Not Applicable.**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable.**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

Confident Feature Ranking: Supplementary Materials

A PFI VARIANCE ANALYSIS

In Section 2, we present two options for defining the base FI values for PFI (Breiman, 2001): (1) a single permutation, and (2) a single observation. We obtain the same global FI values from the two base definitions by setting the values of the number of permutations (B) and the size of the explanation set (N) accordingly. However, the different decomposition of the global FI value to base FI values allows for the analysis of various sources of uncertainty – the variance of the permutations and the variance of the explanation set. Our framework is limited to quantifying only one source of uncertainty by aggregating base FI values to global FI values. Therefore, using it might be a problem when the uncertainty of the global FI values stems from multiple sources of uncertainty. Nevertheless, if most of the variability comes from one of the sources, it is reasonable to target it and disregard the other sources. In the case of PFI, we expect that the size of the explanation set introduces greater variance than the number of permutations. Our results clearly show this; therefore, we can use our framework to quantify the uncertainty of global PFI values.

A.1 Experimental Setup

In this experiment, we use the same DGPs (A and B) described in Section 5.2, including the definition of X, Y, and the functions.

A.1.1 Dummy Prediction Model

Instead of training a model, we create a *Dummy* model that predicts Y from X using the DGP's function. We use this approach to control the variability stemming from the training and focus on the variability of the permutations and explanation set.

A.1.2 Experiment Details

We sample the data as described above with various configurations of B, p, and N, and the two functions. For each configuration:

(1) We perform B permutations for each observation and calculate the loss difference for each permutation b: $L(f(x^b_{[j]}); y) \quad L(f(x); y)$.

(2) We average all of the permutations for each observation.

(3) We average all of the observations.

The result of steps 1-3 is a set of p global FI values $\hat{\Phi}_1^{PFI}; :::; \hat{\Phi}_p^{PFI}$. We repeat this process 100 times and calculate the average and standard deviation (SD) across the repetitions.

A.2 Results

For both functions we compare the SD of the global FI values for different values of B and N. In Figure 10, we can see that different features have different SDs, but in all conditions the SD is almost fixed with respect to B and decreases with N. This indicates that the number of observations introduces more variability to the global FI values than the number of permutations.

B DEFINITIONS AND PROOFS

B.1 Proof of Theorem 1

In this section, we present the detailed proof of Theorem 1.

Recall that D is the set of partial rankings and that we assume that the probability of any error in D is less than . To prove the theorem, we first show that any error in coverage, i.e., a CI that does not cover the true rank, must be caused by at least one partial ranking error in D:

Suppose that there is a coverage error. Without loss of generality, assume that the coverage error occurs for feature 1:

$$[l_1; u_1] \ * \ [L_1; U_1]:$$

The coverage error can be in one or both bounds:

(1) $l_1 < L_1$,

(2) $u_1 > U_1$.

If (1), then $L_1 > 1$, and there are $L_1 \quad 1 > 0$ pairs of the type $(1; k) \ 2 \ D$, meaning that there are $L_1 \quad 1$ features with a significantly lower observed global FI than the observed global FI of feature 1. However, according to Definition 1, $l_1 \quad 1 = \#fk : \Phi_1 > \Phi_k g$, meaning that there are only $l_1 \quad 1$ features with true global FI lower than the true global FI of feature 1. Combining these two statements together, there must be at least one feature $k = 2; :::; p$ for which $(1; k) \ 2 \ D$ but $\Phi_1 \quad \Phi_k$, meaning that there is a partial ranking error in D.

If (2), then the set $fk \ 2 \ 2; :::; p : (k; 1) \ 2 \ Dg$ is higher than the set $fk : \Phi_k > \Phi_1 g$. Again, this would mean that for at least one value of k there is a partial ranking error.

The event of at least one coverage error is contained in the event of obtaining a partial ranking error. Given that, the coverage error probability is bounded by $FWER =$.

B.2 FWER Adjustment Procedures

Here, we provide details on the two sequential procedures that we use in our implementation. After adjustments, the p-values are compared to a chosen level. Note that all p-values are inflated compared to their original level, making it less likely that the null hypothesis will be rejected. Furthermore, the p-values keep their relative order after adjustment. In the procedures below, this is governed by the max function, which assures that the order is maintained. The resulting process is sequential in that for a given level , after the first non-rejected value, all others would not be rejected. Let $p_1; :::; p_K$ be a set of K p-values obtained by testing a family on null hypotheses $H_1^0; : : : ; H_0^K$; below we demonstrate how the two FWER adjustment procedures are used to calculate $p_1^{adg}; :::; p_K^{adg}$, a set of adjusted p-values.

B.2.1 Holm's Procedure

We implement Holm's procedure (Holm, 1979) on one-sided hypothesis tests. The paired t-test is calibrated with this procedure for normally distributed base FI values.

Let $p_{(1)} \quad ; :::; \quad p_{(K)} \quad 1$ be the sorted set of p-values. Then:

$$\begin{aligned}
p_{(1)}^{adj} &= K \quad p_{(1)}; \\
p_{(2)}^{adj} &= \max fp_{(1)}^{adj}; (K \quad 1)p_{(2)}g; \\
&: : : ; \\
p_{(k)}^{adj} &= \max fp_{(1)}^{adj}; :::; p_{(k \ 1)}^{adj}; (K \quad (k \quad 1))p_{(k)}g; \\
&: : : ; \\
p_{(K)}^{adj} &= \max fp_{(1)}^{adj}; :::; p_{(K \ 1)}^{adj}; p_{(K)}g:
\end{aligned}$$

B.2.2 Min-P Procedure

Holm's procedure is highly conservative, since it is valid regardless of the structure of dependence between the p-values. To improve it, Westfall and Young (1993) suggested the Min-P procedure. The idea is to use

bootstrapping to model the structure of dependencies between p-values, obtain lower adjusted p-values, and reject more hypotheses. The details of the Min-P procedure described here are taken from Efron (2012).

Here, we also start with the sorted set of p-values $p_{(1)} \; ;:::; \; p_{(K)} \quad 1$. Let $i_1;:::i_K$ indicate the corresponding original indices, $p_{(k)} = p_{i_k}$, and define $I_k = fi_k; i_{k+1};:::;i_K g$ and $\quad_k = P_0 f\min_{j\,2\,I_k}(P_j) \quad p_{(k)}g$. Here, $(P_1;:::;P_K)$ indicates a hypothetical realization of the unordered p-values $p_1;:::;p_K$ obtained under the complete null hypothesis, meaning all H_k^0s are true. The adjusted p-values are then defined by:

$$p_j^{adj} = \max_k \; \; {}_k:$$

C Confident Feature Ranking: Step-By-Step

This section demonstrates how our ranking method constructs simultaneous CIs for the true ranks. We use the bike sharing dataset (Fanaee-T and Gama, 2014), the TreeSHAP FI method, and our ranking method with Holm's procedure for $n = 50$ base FI values. This demonstration is a detailed description of the example presented in Section 5.3 (Figure 7).

Base and Global FI Values We construct a TreeSHAP explainer (Lundberg et al., 2019) based on the trained XGB model. The base FI values are the absolute values of the local SHAP values the explainer produces for an explanation set of size $n = 50$. The global FI values are the average of the base FI values. The values and the order of the global base FI values for this explanation set are presented in Table 1.

Pairwise Differences The paired-sample t-test is based on the differences between the base FI values of two features $\mathbf{v}_j$ $\mathbf{v}_k$. The one-sided hypothesis $H^0_{jk} : \Phi_j$ Φ_k is rejected if the difference between the observed global FI values is significantly different from zero. In Figure 11a, we present the differences between *Workingday* and all other features. The average of the differences between *Workingday* and *Month* and *Temp* is near zero.

Partial Rankings We set the significance level to $= 0.1$, and for each pair of features, we run two paired one-sided t-tests; then, we adjust the p-values to multiple comparisons using Holm's procedure. In Figure 11b, gray and black indicate that the observed global FI value of the feature in row j is respectively less and greater than the observed global FI value of the feature in column k. White indicates that the difference is zero (neither H^0_{jk} nor H^0_{kj} were rejected). The set of partial rankings D is then obtained. For example, we can conclude that $(Month; Year)$ 2 D, $(Day; Month)$ 2 D, and $(Month; Workingday)$ $\not\in$ D.

Constructing Simultaneous CIs for the True Ranks For each feature, we initialize the lower bound of the CI to one and the upper bound to p. If there are no differences between the features, the CIs for all features are $[1; p]$. Otherwise, there are differences. Without loss of generality, consider the *Workingday* feature. By looking at the row for *Workingday* in Figure 11b, we can see that the observed global FI value of *Workingday* is significantly higher than the observed global FI values of *Day* and *Weather*, and it is significantly lower than the observed global FI values of *Year* and *Hour*. There is no significant difference between *Workingday* and *Month* and *Temp*. Therefore, we increase the lower bound by two, decrease the upper bound by two, and obtain the confidence set $[3; 5]$ for the true rank of *Workingday*.

We repeat the same process for all features and obtain 90% simultaneous CIs for the true ranks. See lower and upper bounds in Table 1 and a visualization of the CIs in Figure 11c.

Table 1: Ranks and Simultaneous CIs

Feature	Observed Global FI	Observed Rank	CI
Hour	129.042	7	[7, 7]
Year	44.805	6	[6, 6]
Temp	33.777	5	[4, 5]
Workingday	28.95	4	[3, 5]
Month	23.987	3	[3, 4]
Weather	10.865	2	[2, 2]
Day	5.673	1	[1, 1]

D EXPERIMENT DETAILS AND ADDITIONAL RESULTS

D.1 Ranking Method Comparison

D.1.1 Baseline Ranking Method

We implement a naive ranking method to construct CIs for the features' true ranks based on bootstrap samples. For each sample, we rank the global FI values. We report lower and upper bounds $(L_j; U_j)$ by taking the =2 and 1 =2 quantiles of the ranks of the j'th feature over the bootstrap distribution.

D.1.2 Additional Results

Equal Correlations In Figure 12, we present the ranking efficiency of three ranking methods for $p = 10$ and $p = 50$ as a function of n, with multiple levels of correlations. We use the same configuration as in Figure 5: low and high -factors, -exponent=0.25, with and without ties. We can observe the same efficiency trends seen for $p = 30$: the efficiency increases as increases, the gap between the methods increases as the -factor increases, the efficiency improves as the n increases, and our ranking method is more efficient than ICRanks.

Number of Features We compare the ranking efficiency for different numbers of features and multiple values of -exponent, with $n = 500$, -factor=1, equal correlations, and = 0:5 (see Figure 13). The efficiency degrades as the means become more dense (-exponent decreases), , and the number of features has almost no effect on the efficiency.

Correlation Structure We compare the equal correlation structure with the block-wise pairs structure of the correlation matrix. In Figure 14, we present the results for different numbers of features, $n = 500$, -factor=1, -exponent-0.25, and low (0.1) and high (0.9) values of . The differences between the structures of the correlation matrix are more substantial for = 0:9.

D.2 SHAP Ranking Measures

Here, we sample the training and explanation sets using the DGP described in Section 5.2. For each configuration of $(X; Y)$, we train an XGB (default hyperparameters) or RF (number of estimators=1,000) regression model for this experiment. We follow the XGB tutorial[6] to train both the XGB and RF models (see the train and test R^2 of the model in Table 2).

Table 2: Prediction Models' Performance

Model	DGP	p	Train R^2	Test R^2
RF	DGP-A	10	0.786	0.783
		30	0.523	0.515
		50	0.346	0.336
	DGP-B	10	0.863	0.862
		30	0.836	0.836
		50	0.745	0.744
XGB	DGP-A	10	0.956	0.952
		30	0.963	0.958
		50	0.954	0.945
	DGP-B	10	0.875	0.868
		30	0.951	0.946
		50	0.964	0.959

D.2.1 Additional Results

In the paper, we present an example of the efficiency of RF with DGP-A and XGB with DGP-B. Here, we present the complementary efficiency results for all configurations (see Figure 15). The simultaneous coverage for all configurations is almost one (0:997 0:009). In addition, we compare the efficiency of our CIs (using our method with the Min-P procedure) with $n = 1000$ base FI values, to the efficiency of the true FI ranks as an upper bound on the efficiency of the observed values (see Table 3); as can be seen, the efficiency of our ranking method with $n = 1000$ is not ideal, even in the case of perfect true ranking.

[6] XGB tutorial

Table 3: Ranking Efficiency

Model	DGP	p	True Efficiency	Mean Efficiency
RF	DGP-A	10	0.222	0.238
		30	0.393	0.412
		50	0.486	0.495
	DGP-B	10	0.222	0.386
		30	0.634	0.641
		50	0.772	0.777
XGB	DGP-A	10	0	0.204
		30	0	0.293
		50	0	0.32
	DGP-B	10	0	0.299
		30	0	0.247
		50	0	0.287

D.2.2 Ranking Runtime Analysis

We also analyzed TreeSHAP's runtime and our method's runtime. In Figure 16, we present the runtime of TreeSHAP and the different ranking methods. TreeSHAP's runtime clearly depends on the sample size n, since it is a local FI method. In contrast, the runtime of the ranking methods depends on the number of features (p). The runtime of our method with Holm's procedure is comparable to that of ICRanks but with lower variance. The runtime of our method with the Min-P procedure is much higher, because it is based on a bootstrap process and increases with the number of repetitions (B).

D.2.3 Non-Normal Base FI Value Distribution

We use the paired-sample t-test to compare base FI values and adjust the p-values with the Min-P or Holm's procedure. Our primary assumption is that the paired test is calibrated for the possible distributions of base FI values (note that the paired-sample t-test is calibrated even when the base FI values are not normally distributed). However, we found that our method does not always maintain simultaneous coverage when the base FI values have an extremely long tail. In this example, we sample the data from:

$$
\begin{aligned}
y &= x_1x_2 + x_3^2 \quad x_4x_7 + x_8x_{10} \quad x_6^2 \\
&\quad + x_{11}x_{12} + x_{13}^2 \quad x_{14}x_{17} + x_{18}x_{20} \quad x_{16}^2 \\
&\quad + x_{21}x_{22} + x_{23}^2 \quad x_{24}x_{27} + x_{28}x_{30} \quad x_{26}^2 + \ ; \\
X &\quad N_{30}(\mathbf{0}; \Sigma); \qquad N(0; 1); \\
&\text{where } \Sigma \text{ is an equal correlation matrix} \\
&\text{with} \quad = 0{:}3 \text{ and } f \ {}_j^2 g \qquad {}^2{:}
\end{aligned}
$$

We train an RF model and calculate the base FI values with TreeSHAP. Table 4 summarizes the average coverage and simultaneous coverage. As can be seen, for all sizes of n (the number of base FI values) our method does not maintain simultaneous coverage; the marginal coverage is almost 90% for small sizes of n, and the simultaneous coverage of the Min-P procedure is better.

We further analyze the CIs of the features for $p = 30$, an RF model, and a single explanation set of size $n = 100{;}000$. In Figure 17, the base FI values distribution of two features, for which we found coverage errors for multiple explanation sets. The true global FI values of the two features are almost identical, and the variance is relatively large. More importantly, the distributions of the base FI values of both features display extremely long tails, and the observed global FI values are influenced by the rare values at the tails. In such cases, we recommend replacing the paired t-test with a robust alternative (Wilcox, 2011).

D.3 High-Dimensional Example

Displaying the global FI values or the CIs for the ranks for many features may be difficult to interpret. Therefore, typically only the top-k features are presented. For example, in SHAP's global bar plot API[7] the default number

[7] Global bar plot API.

Table 4: Ranking Coverage

n	Ranking Method	Coverage	Simultaneous Coverage
100	Holm	0.839	0.08
	Min-P	0.96	0.63
500	Holm	0.75	0.01
	Min-P	0.888	0.56
1000	Holm	0.7	0.01
	Min-P	0.85	0.37
3000	Holm	0.634	0.01
	Min-P	0.782	0.14

of features to present is 10. In Figure 18, we present the complete ranking for the Nomao dataset(Candillier and Lemaire, 2012) features . Our ranking method makes it easier to interpret which features are irrelevant and determine how to select a threshold k for the most important features to display.

E IMPLEMENTATION DETAILS

E.1 Availability of Data

The bike sharing dataset (Fanaee-T and Gama, 2014) contains 10,886 records of bike rentals between 2011 and 2012 from the Capital Bikeshare program in Washington, D.C. The regression task is forecasting demand for bike rentals based on time and environmental measures such as month and weather. The data is publicly available at Kaggle.

The COMPAS dataset (Angwin et al., 2016) contains 6,172 records of criminal history from Broward County from 2013 and 2014. The classification task is assessing a criminal defendant's likelihood to re-offend based on jail and prison time, and demographics. The data is publicly available at Propublica's Github.

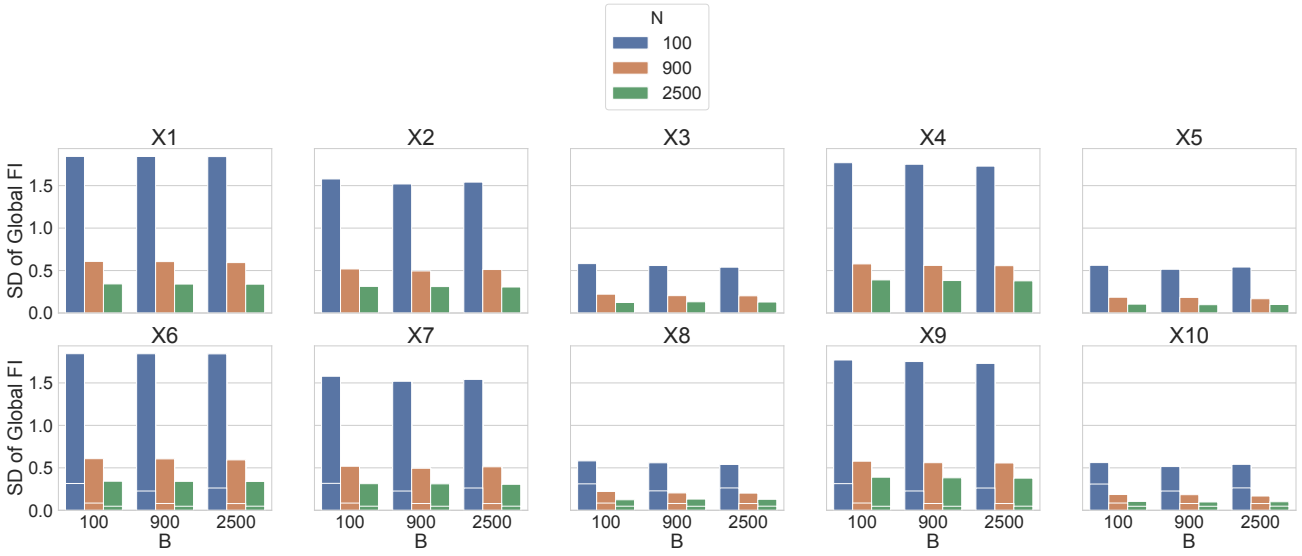
Nomao is a search engine of places. The Nomao dataset (Candillier and Lemaire, 2012) contains 34,465 records with comparison features about places, such as name and localization, from different sources. The classification task is detecting whether two sources of information refer to the same place. The data is publicly available at OpenML.

E.2 Reproducibility Instructions

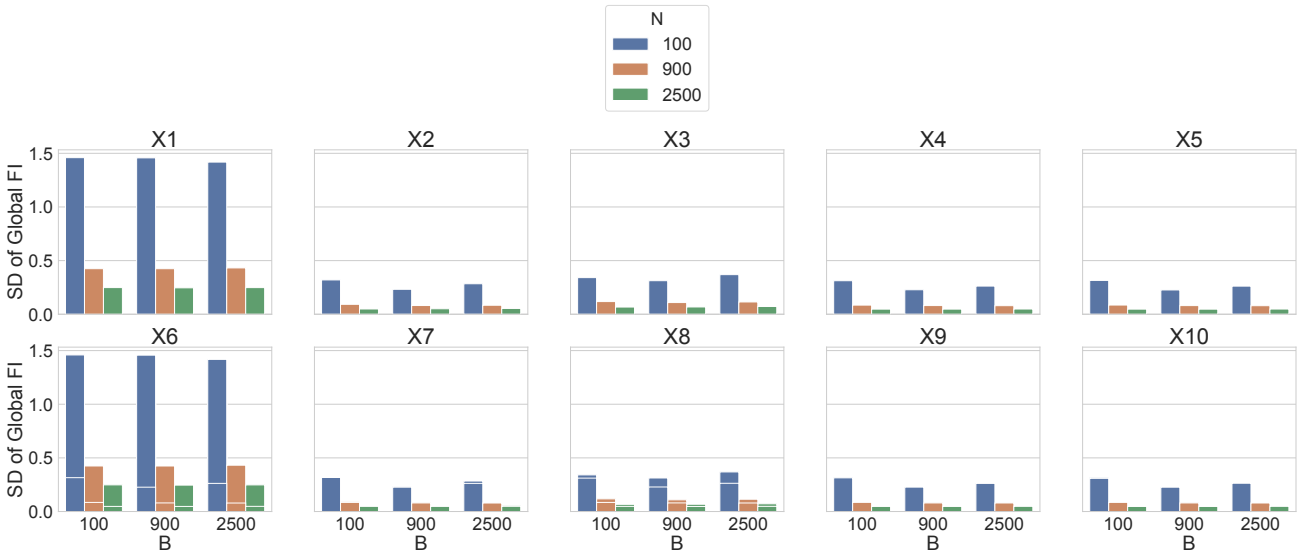
The code for our ranking method, experiments, and visualizations was written in the Python programming language (Python version 3.10.3) and can be found in this Git repository. The ICRanks R package was imported to Python through the rpy2 package.

E.3 Computing Infrastructure

The experiments were performed on a server with 64G RAM and 16 CPUs.

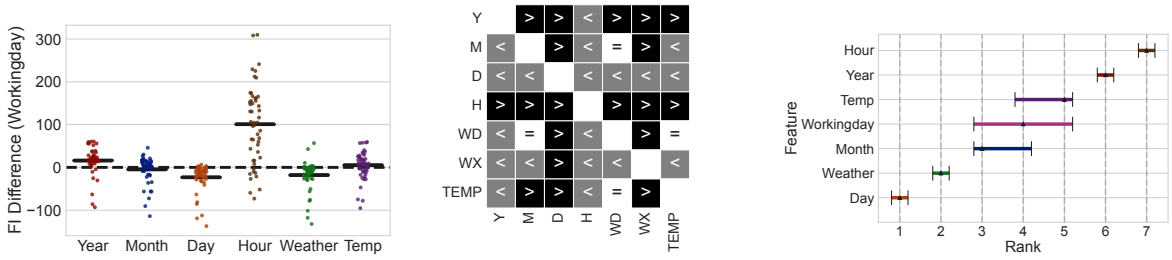


(a) DGP-A



(b) DGP-B

Figure 10: SD of global FI values of $p = 10$ features for two functions; with respect to the number of permutations (B) and the number of observations (N).



(a) Base-FI differences (b) Partial rankings (c) CIs for the true ranks

Figure 11: Visualization of Algorithm 1: first, test all one-sided pairs of hypotheses for the difference between base FI values (a), then adjust the p-values and obtain the partial rankings from the rejected hypotheses (b), and finally construct the CIs for the true ranks (c).

(a) p = 10: -factor=0.2

(b) p = 50: -factor=0.2

(c) p = 10: -factor=5

(d) p = 50: -factor=5

Figure 12: Ranking e ciency for p = 10 and p = 50.

Figure 13: Ranking e ciency as a function of p for multiple values of -exponent and three ranking methods.

(a) =0.1

(b) =0.9

Figure 14: Ranking e ciency with low (a) and high (b) values of , as a function of p for two correlation structures and three ranking methods.

(a) RF with DGP-B

(b) XGB with DGP-A

Figure 15: Ranking e ciency as a function of n for di erent numbers of features (p) and ranking methods.

(a) RF with DGP-A

(b) RF with DGP-B

(c) XGB with DGP-A

(d) XGB with DGP-B

Figure 16: TreeSHAP and ranking times (in seconds) as a function of the number of features (p) for di erent sizes (n) of $D_{explain}$.

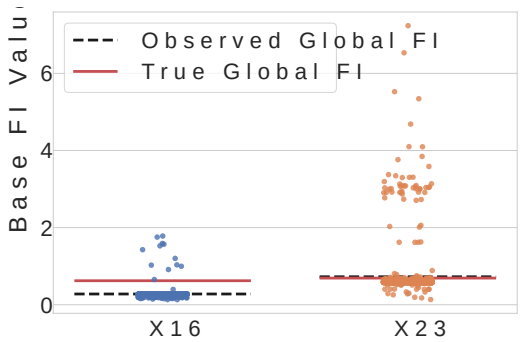


Figure 17: Base FI values' distributions for features $X16$ and $X23$.

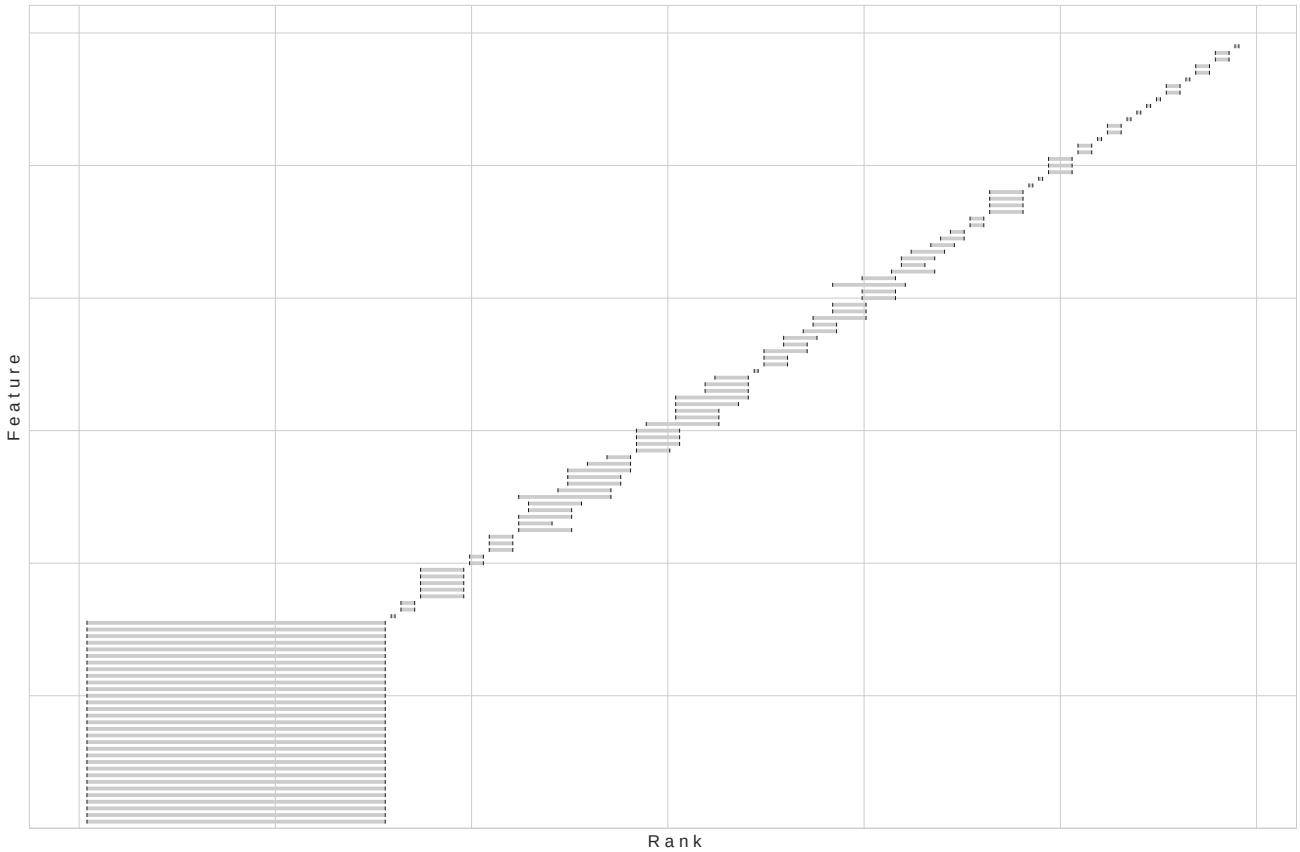


Figure 18: CIs for the true ranks of the Nomao dataset features. The features are ordered by their observed global FI value. There are 31 irrelevant features and many intersections between CIs.