Confident feature ranking

Anonymous Authors[1]

Abstract

Interpretation of feature importance values often relies on the relative order of the features rather than on the value itself, referred to as ranking. However, the order may be unstable due to the small sample sizes used in calculating the importance values. We propose that post-hoc importance methods produce a ranking and simultaneous confident sets for the ranking. Based on pair-wise comparisons of the feature importance values, our method is guaranteed to include the "true" (infinite sample) ranking with high probability and allows for a selection of top-k sets.

1. Introduction

Complex non-linear prediction models are widely used to augment or even replace human judgment; for example, in healthcare (1) and finance (2). Regulators, users, and developers of prediction models are all interested in understanding the relative contribution of the different inputs - the features - toward the model's predictions (3; 4). Commonly used feature importance (FI) methods such as Permutation Feature Importance (PFI, (5)) and SHAP (6; 7) measure the effect of removing or perturbing a feature on the predicted value or the loss. The specifics of this manipulation change across different implementations of the methods (8; 9; 10).

Studies have demonstrated that post-hoc feature importance methods, which are employed to explain trained prediction models, may not be stable and reliable due to multiple sources of uncertainty, such as the randomness of sampling and the hyperparameters of the methods (11; 12; 13). We focus on uncertainty in the sampling of the explanation set, affecting the aggregation of local examples into a single global value. Most methods for quantifying these types of uncertainty produce per-feature spread estimates (or confidence intervals) in the units of the FI method. These uncertainty measures are insufficient, because it is often the *rank of the feature importance*, rather than the value itself, which affects the decision-making and reporting on the features. In our analysis of Nature group articles from 2021 mentioning "feature importance" (35 articles with at least 20 citations), we found that all the articles either reported only the identity of the top-k ranked features (14) or implicitly used rank to choose which importance values to report (15). Considering the instability of these values, ranking the features based on the order of their importance values for a single sample might be misleading.

Therefore, the uncertainty for the rank of feature importance values should be estimated. We propose reporting a set of possible ranks for each feature that would include, in high probability, the true rank of every feature's importance value. Specifically, we estimate a simultaneous confidence set-rank for the true feature ranking. We deduce the set-ranks by testing all pairs of features for shifts in means and counting the number of rejections for each feature. Figure 1 displays an example of the results. These results can be used to derive feature sets that include the top-k features with high confidence, and can be used to approximate how many samples are needed for the feature importance methods.

2. Uncertainty in feature importance

Consider the supervised learning task of predicting a real-value outcome $Y \in \mathcal{Y}$ from a vector of p features $X = (X_1, ..., X_p) \in \mathcal{X}$. A prediction model $f : \mathcal{X} \to \mathcal{Y}$ is fit on a training set $\mathcal{D}_{train} = \{(x_i^*, y_i^*)\}_{i=1}^M$, and is found to fit the data well according to standard metrics (e.g., MSE or accuracy on external test sets). Researchers are then interested in quantifying each feature's predictive power to the model, the *feature importance value* (17).

Feature importance post-hoc methods measure an importance value for each feature, $\hat{\Phi}_1, \hat{\Phi}_2, \ldots, \hat{\Phi}_p \in \mathbb{R}$, based on a sample $\mathcal{D}_{explain} = \{(x_i, y_i)\}_{i=1}^N$, preferably independent of $\mathcal{D}_{train}$, and a trained model f. We assume a higher value of $\hat{\Phi}_j$ indicates higher importance. In many case, the features are ranked according to the FI values and only the top k features are considered. If feature j has the maximal value ($j = \arg\max \hat{\Phi}_k$), then its observed rank is $\hat{r}_j = p$.

In many cases, the FI values are calculated by averaging many independent runs:

- In SHAP (6) and PFI (8), the global FI is averaged over the absolute values assigned to each example (the *local* FI). Variability in the explanation set $\mathcal{D}_{explain}$ introduces the uncertainty into the global FI.
- Also in PFI, each local FI value is computed by aver-

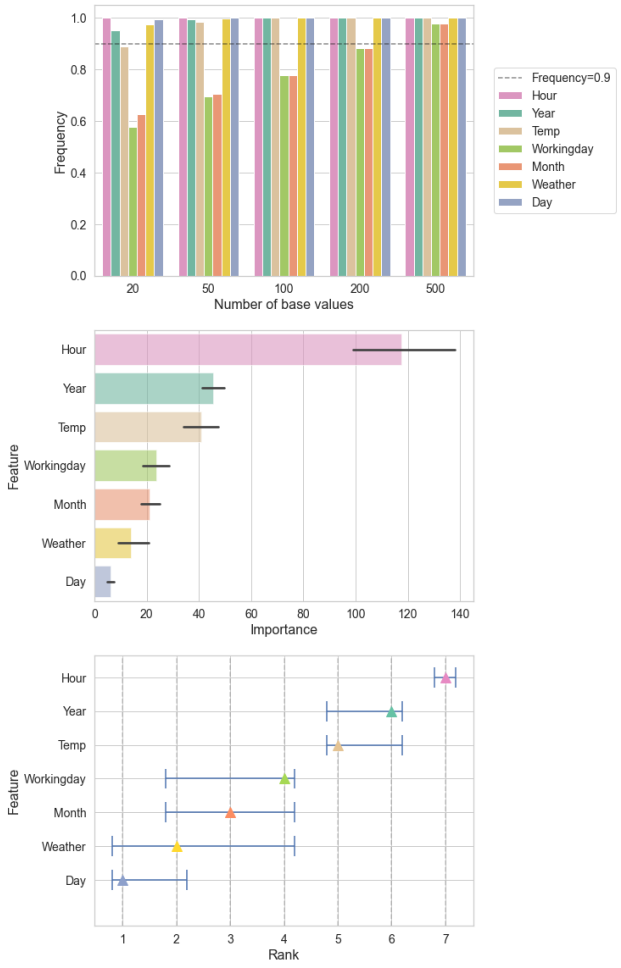

Figure 1. Frequency of the mode rank of each feature across repeated rankings (top), global feature importance values (middle), and ranking with confidence set (bottom) for the bike demand dataset (16), with an XGBoost model and TreeSHAP. The importance values for the middle and bottom figures were obtained from 50 observations. The confidence set identifies uncertain feature rankings in this sample size.

aging many randomized imputations (or permutations) for the feature value. Here, variability in the set of imputed values introduce uncertainty into the local FI.

Addressing these examples in a single framework, we identify a two level FI hierarchy: the observed *global* FI is an average of a set of independent *base* FI values. In the first example, the base FIs correspond to local values for multiple examples in the explanation set, and the global correspond to the usual global post-hoc values. In the second example, the base examples correspond to different imputations, and the global corresponds to a single local example.

Setting notations, define matrix $\mathbf{v}$ to be the matrix of *base* FI values, with rows $v_1, ..., v_n \in R^p$ representing FI value for each feature[1]. $\mathbf{v}_j$ are the columns of this matrix, referring to the base FI values for the j'th feature. Then the observed global importance is written $\hat{\Phi}_j = \frac{1}{n}\sum v_{ij}$.

Probability model and true global FI We model the rows v_i as independent samples from distribution F_v with mean vector $E[v_i] = (\Phi_1, ..., \Phi_p)$, the *true global FI values*. For finite n, $\hat{\Phi}_j$ would be an unbiased but noisy version of Φ_j; we are interested in understanding the effects of this deviation on the possible feature rankings.

3. Ranking confidence sets

In contrast to the observed noisy rankings, the *true ranking* are based on the true FI values $\Phi_1, ..., \Phi_p$. To allow for features with identical true importance, we follow (18) in redefining the true ranks to account for ties:

Definition 3.1 *(Set-rank) Define the lower-rank of* Φ_j *by* $l_j = 1 + \#\{k : \Phi_j > \Phi_k, j \neq k\}$ *and the upper-rank of* Φ_j *by* $u_j = p - \#\{k : \Phi_j < \Phi_k, j \neq k\}$. *Then the set-rank of* Φ_j *is the set of natural numbers* $\{l_j, l_{j+1}, \ldots, u_j\}$ *denoted here* $[l_j, u_j]$.

If there are no ties, the lower and upper ranks are identical and equal to the standard definition.

Our inferential goal is to construct a confidence set of ranks, estimated from the data, for each feature's importance such that with high probability, all sets cover the true set-ranks.

Definition 3.2 *(simultaneous coverage) The set of confidence intervals* $\{[L_j, U_j]\}_{j=1...p}$ *has simultaneous coverage at level* $1-\alpha$ *if*

$$\mathbb{P}\left([l_j, u_j] \subseteq [L_j, U_j], \quad \forall j \in \{1, \ldots, p\}\right) \geq 1 - \alpha.$$

L_j, U_j are functions of $\mathbf{v}$ the observed base FI. Different sets of observed base FI values would produce different confidence sets. In a simultaneous $1-\alpha$ confidence set, the probability that all intervals in the set cover the true set-ranks is at least $1-\alpha$. Note that simultaneous coverage is conservative and can result in relatively large intervals.

An advantage of simultaneous coverage requirement is that coverage is also maintained for many functions of the observed ranks across features. One example is the top-k sets: instead of selecting the most important feature (or most important k features) based on the observed data, we may prefer selecting all features that could be ranked in the top-k as n increases based on the confidence sets. With simultaneous coverage, the error probability for this selection is controlled (19). Furthermore, the confidence intervals for the features currently ranked top-k still have marginal

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

coverage. These two properties are not guaranteed by the usual marginal coverage(20).

4. Confident simultaneous feature ranking

Our method uses pairwise hypothesis tests to form lower and upper bounds of the true rank of each feature. To see the relationship between tests and ranking, consider a feature pair j, k for which we can reject with high probability the hypothesis $H_{j\leq k} : \Phi_j \leq \Phi_k$. Then, with a high probability $\Phi_j > \Phi_k$, in that case, and its true rank is > 1. Our method is based on (21), and we compare in Section **??**.

Theoretically, we would like to form for each feature pair j, k two hypothesis tests:

$$H^0_{jk} : \Phi_j \leq \Phi_k \text{ versus } H^1_{jk} : \Phi_j > \Phi_k,$$
$$H^0_{kj} : \Phi_k \leq \Phi_j \text{ versus } H^1_{kj} : \Phi_k > \Phi_j.$$

In practice, when both tests use the same data and the probability of false rejection is small, the two null hypotheses will not be rejected together. We can therefore use a single two-sided test for the hypotheses:

$$H^0_{j,k} : \Phi_j = \Phi_k \text{ versus } H^1_{jk} : \Phi_j \neq \Phi_k.$$

If the null hypothesis is rejected, the sign of the observed difference can indicate which directional hypothesis was rejected. The paired test takes vectors $\mathbf{v}_j, \mathbf{v}_k$ and returns a p-value $p_{j,k}$ and a sign $s_{jk} = sign(\hat{\Phi}_j - \hat{\Phi}_k)$. Denote as

$$(s_{jk}, p_{j,k}) = pairedTest(\mathbf{v}_j, \mathbf{v}_k).$$

A decision is made by comparing $p_{j,k}$ to the threshold α, and then choosing the directional hypothesis based on s_{jk}. For example, if $p_{j,k} < \alpha$ and $s_{jk} = -1$, we conclude that $\Phi_j < \Phi_k$. Note that sign decisions are always for strict inequalities, as these narrow down the set of possible ranks.

4.1. Controlling the family wise error

Define D the set of sign decisions from all pairwise tests $D = \{(j, k) : H^1_{jk}$was accepted[2]$\}$. We make a sign error if there is a pair $(j', k') \in D$ for which $\Phi_{j'} \leq \Phi_{k'}$. A level α calibrated test[3] assures that the probability of this event is less than α. However, for simultaneous confidence intervals, we need a stronger condition.

Definition 4.1 *(Family wise error rate) The pairwise tests* $\{pairedTest(\mathbf{v}_j, \mathbf{v}_k)\}_{j<k}$ *control the family wise error rate (FWER) at level* α *if* P*(Making any sign error)* $\leq \alpha$.

[2]It is more common to use "reject the null" rather than "accept the alternative". However, here, we indeed have complimentary sets, hence rejecting H_0 means accepting H_1.

[3]See definition of calibrated test in Appendix B.

Indeed, the probability of making at least one error increases with more tests, and thus we need more conservative tests to control the FWER. We use the sequential Holm (22) procedure to adjust (increase) the p-values. The adjusted p-values are then compared to the requested level α; if the p-values are calibrated, the FWER for the rejected tests (after Holm) is controlled at level α, regardless of dependence. See (23) on using Holm for pair-wise comparisons.

4.2. Confident feature ranking

Once a family of pairwise tests is available with FWER control, we derive the simultaneous confidence rank set:

Theorem 4.2 *Let* D *be the set of sign decisions with FWER control at level* α*. For* $j = 1, ..., p$*, define*

$$L_j = 1 + \#\{k : (j, k) \in D\},$$
$$U_j = p - \#\{k : (k, j) \in D\}.$$

Then the sets $\{[L_j, U_j]$ *for* $j \in [p]\}$ *are simultaneous* $(1-\alpha)$ *confidence sets for the ranks of the true global FI values.*

The construction naturally extends the definition of ranks used in Definition 3.1. The idea in this proof is that a coverage failure means that the set of true (one-sided) differences was smaller than the set of (one-sided) sign decisions. This means that at least one sign decision was an error. Therefore the FWER upper bounds the probability of an error in the confidence sets. See proof in Appendix B.3.

Algorithm 1 describes the complete method. The algorithm works directly on the base FI matrix without requiring access to the trained model, FI method, or the explanation data set. The main assumption is that our paired test is calibrated for the family of possible distributions for base FI values.

Algorithm 1 Confident sets of ranks

Require:
- Base FI values $\mathbf{v}$
- Level of confidence $1 - \alpha > 0$;
- A suitable paired test;

for $j, k \in [p], j < k$ **do**
  $(s_{jk}, p_{j,k}) \leftarrow pairedTest(\mathbf{v}_j, \mathbf{v}_k)$.
end for
$\{p^{adj}_{j,k}\}_{j<k} \leftarrow Holm(\{p_{j,k}\}_{j<k})$
$D \leftarrow \{(j, k) : p^{adj}_{j,k} < \alpha, s_j k = 1\}$
$D \leftarrow D \cup \{(k, j) : p^{adj}_{j,k} < \alpha, s_j k = -1\}$
for $j \in [p]$ **do**
  $L_j \leftarrow 1 + \#\{k : (j, k) \in D\}$
  $U_j \leftarrow p - \#\{k : (k, j) \in D\}$
end for
$[L_1, U_1], \ldots, [L_p, U_p]$.

Paired test In the paper, we use the parametric paired t-test: Set $\mathbf{d} = \mathbf{v}_j - \mathbf{v}_k$ to be the vector of differences, write $\bar{d}$ for the sample average and s_d for the sample standard deviation. Then the two-sided α level test rejects the null hypothesis if $|\bar{d}/(s_d/\sqrt{n})| > T_{n-1}(1-\alpha/2)$, where the r.h.s. marks the $1-\alpha/2$ quantile of student-t ($n-1$ df). The paired t-test is fairly robust to departures from normality(24).

Comparison to ICranks Algorithm 1 is similar to the ICranks (21), with some differences. In ICranks, they used the Tukey correction (25) for the normal distribution to control all differences between ranks simultaneously. It assumes normality and independence between features and is hard to replace with non-parametric tests. In contrast, our algorithm applies tests to each feature pair and uses the Holm correction; hence, it can be used with robust or non-parametric location tests (26).

5. Evaluation

We briefly discuss our experimental setup and results. More details and more results are in Appendix C.

Synthetic data We use synthetic data to check the validity and efficiency of the method. For validity, we need to identify the true global importance value. We used the following two approaches: (a) *Mock importance values* - we simulated the base FI values from a multivariate normal distribution with predetermined means and covariance matrices; (b) *Low-variance importance values* - we built a prediction model and calculated $\hat{\Phi}_1, \ldots \hat{\Phi}_p$ based on a sufficiently large sample as a low variance estimator of $\Phi_1, \ldots \Phi_p$ (27).

Prediction model We used the XGBoost model (28) in all experiments, due to its computational efficiency and performance on high-dimensional tabular data.

Feature importance methods We selected two commonly used post-hoc feature importance methods TreeSHAP (6) and PFI (5). In both cases, we defined the base FI to be the local values associated with individual examples. Additional details are in Appendix A.

Metrics We use the the following metrics suggested by (18): *Simultaneous coverage:* The proportion of experiments where all true ranks are covered by their confidence sets. *Efficiency:* the average relative size of the confidence sets: $\frac{1}{p\cdot(p-1)}\sum_{j=1}^{p}(U_j - L_j)$. Lower efficiency is better.

5.1. Results

Simultaneous coverage In both the mock feature importancce and model feature importance, the ranking maintains simultaneous coverage levels of almost 1. This indicates that the method is overly conservative compared to the nominal required simultaneous coverage of $1-\alpha = 90\%$.

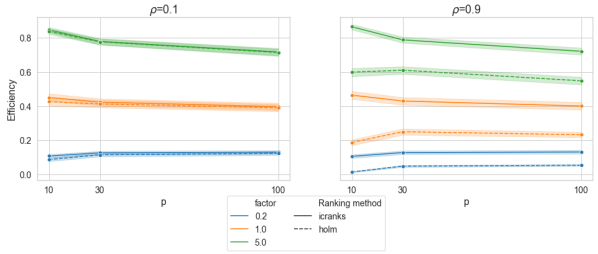

Figure 2. Ranking efficiency as a function of p for two levels of correlations ($\rho = 0.1$ and $\rho = 0.9$) and two ranking methods (our method is dashed, ICRanks is solid). SD of base FI are high (green), medium (orange), or low (blue). Low values mean smaller sets and are therefore better. Both methods maintain simultaneous coverage levels of $\tilde{1}$ in all conditions; our method is more efficient.

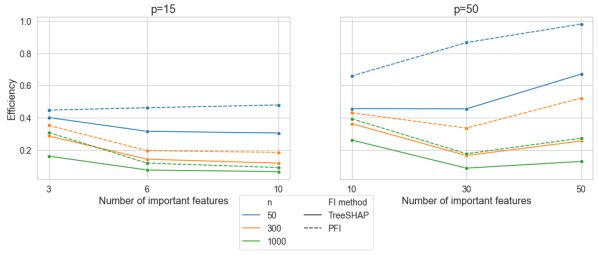

Figure 3. Ranking efficiency for $p = 15$ (left) and $p = 50$ (right) features. The efficiency is presented as a function of the number of important features, the number of base FI values (n), and the FI method: TreeSHAP (solid) and PFI (dashed). Efficiency decreases with n as expected. The efficiency of TreeSHAP is better than the PFI method, but it may depend on the number of permutations.

Mock data: Efficiency compared to ICRanks We compared our efficiency to the ICRanks (18) method, for different number of features (p) and base FI values (n). The results are in Figure 2. When FI are highly correlated ($\rho = 0.9$), our method is more efficient than ICRanks.

Simulated FI: Efficiency comparing FI methods Here we study the efficiency under multiple simulation conditions, varying the FI algorithm (TreeSHAP and PFI), p, number of relevant features, and n. Figure 3 shows the results. In general, the ranking is inefficient with a small number of important features, since the noisy features may be indistinguishable inducing large true rank-sets.

5.2. Concluding remarks

We view this method as a step to producing new types of stability checks in explainable machine learning. Our current algorithm is conservative, as demonstrated in simulations where the coverage level surpasses the requested 90%. Future research aims to narrow the confidence sets while maintaining nominal coverage. We seek to reduce the impact of the number of features on coverage and enhance practitioner confidence through the implementation of robust tests, which currently do not appear to affect coverage.

References

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

[2] Francesco Rundo, Francesca Trenta, Agatino Luigi di Stallo, and Sebastiano Battiato. Machine learning for quantitative finance applications: A survey. *Applied Sciences*, 9(24):5574, 2019.

[3] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*, 2018.

[4] Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016, 2016.

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

[6] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[7] Scott Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*, 2019.

[8] Luke Merrick and Ankur Taly. 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, 2020.

[9] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *arXiv preprint arXiv:2011.14878*, 2020.

[10] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

[11] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

[12] Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *International Conference on Machine Learning*, pages 5628–5638. PMLR, 2020.

[13] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358*, 2021.

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

[15] Gioele La Manno, Kimberly Siletti, Alessandro Furlan, Daniel Gyllborg, Elin Vinsland, Alejandro Mossi Albiach, Christoffer Mattsson Langseth, Irina Khven, Alex R Lederer, Lisa M Dratva, et al. Molecular architecture of the developing mouse brain. *Nature*, 596(7870):92–96, 2021.

[16] Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127, 2014.

[17] Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions through additive importance measures. *arXiv preprint arXiv:2004.00668*, 2020.

[18] Diaa Al Mohamad, Jelle J Goeman, and Erik W van Zwet. Simultaneous confidence intervals for ranks with application to ranking institutions. *Biometrics*, 78(1):238–247, 2022.

[19] Jason Hsu. *Multiple comparisons: theory and methods*. CRC Press, 1996.

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

[21] Diaa Al Mohamad, Erik van Zwet, Aldo Solari, and Jelle Goeman. Simultaneous confidence intervals for ranks using the partitioning principle. 2021.

[22] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[23] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1):561–584, 1995.

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

[25] John Wilder Tukey. The problem of multiple comparisons. *Multiple comparisons*, 1953.

[26] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

[27] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34:9391–9404, 2021.

[28] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[29] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

A. Unifying global feature importance methods

In this section, we will use our framework to define the base and global levels of importance values for TreeSHAP and Permutation Feature Importance.

Tree SHAP TreeSHAP is a post-hoc local feature importance method for tree-based ML models. It is a model-specific version of SHapley Additive exPlanation (SHAP) values (6), that exactly computes the Shapley values from cooperative game theory (29). SHAP is a method to compute Shapley values for the conditional expectation of the predictions of the model $\mathbb{E}[f(X)|X_S = x_S]$. For a tree model the conditional expectation can be estimated recursively based on the tree structure and therefore allow to compute the SHAP values in polynomial instead of exponential time. The TreeSHAP method is exact and not stochastic.

For a single observation (x, y) the local TreeSHAP value of a feature j is:

$$\tilde{v}_j^{TreeSHAP} = \sum_{S \subseteq [p]} \frac{|S|!(p-|S|-1)!}{p!} \Big(\mathbb{E}[f(X)|X_S = x_S] - \mathbb{E}[f(X)|X_{S\setminus\{j\}} = x_{S\setminus\{j\}}] \Big)$$

Where $[p]$ is the set of all features and S is a subset of features.

The *global* TreeSHAP averages over the absolute-value of all observations in a sample of size N:

$$\hat{\Phi}_j^{TreeSHAP} = \frac{1}{M} \sum_{i=1}^{M} |v_j^{TreeSHAP}|$$

Formally, the *base* explanation is $v_i j^{TreeSHAP} = |\tilde{v}_i j^{TreeSHAP}|$.

Permutation Feature Importance Permutation Feature Importance (PFI) (5; 10) is also commonly used as a model-agnostic post-hoc global feature importance method. The feature importance is defined as the increase in a model loss when a single feature value is randomly permuted.

Let L be a loss function, we defined the *base* level of PFI for the j'th feature as the mean over B permutations:

$$v_j^{PFI} = \frac{1}{B} \sum_{b=1}^{B} L(f(x^b_{[j]}), y) - L(f(x), y)$$

. Where $x_{[j]}$ is a replication of an observation x with a permuted version of the j'th feature.

Respectively, the *global* level is defined as an average over M base level importance values:

$$\hat{\Phi}_j^{PFI} = \frac{1}{M} \sum_{m=1}^{M} v_j^{PFI} = \frac{1}{B} \sum_{b=1}^{B} L(f(X^b_{[j]}), Y) - L(f(X), Y)$$

Where $X_{[j]}$ is a replication of the data matrix X with M rows, and a permuted version of the j'th feature. The right term is the common definition of PFI.

B. Definitions and proofs

B.1. Calibration

Let $X \sim F_\theta$ be a family of distributions parameterized by $\theta \in \Theta$, and let $\Theta_0 \subset \Theta$.

Definition B.1 *(A calibrated test) The p-value $p(X)$ for the test $H_0 : \theta \in \Theta_0$ is calibrated if for any $\theta \in \Theta_0$, and $\alpha < 1$ the test $p(X) < \alpha$ controls the type 1 error at level α. Specifically:*

$$P_{\theta \in \Theta_0}(p_{j,k} < \alpha) < \alpha, \text{ for any } \alpha < 1, \theta \in \Theta_0.$$

Whether a test is calibrated depends on the assumptions of the test and the family of distributions associated with the null hypothesis (H_0).

B.2. Holm procedure

We describe the Holm procedure interms of the the adjustment of the p-values.

Take a set of K p-values $p_1, ..., p_K$, and sort them $p_1 \leq , ..., \leq p_{(K)} \leq 1$.

Then
$p^{adj}_{(1)} = K \cdot p_1$,
$p^{adj}_{(2)} = \max\{p^{adj}_1, (K-1)p_{(2)}\}$,
so that the k'th p-value
$p^{adj}_{(k)} = \max\{p^{adj}_{(1)}, ..., p^{adj}_{(k-1)}(K-k+1)p_{(k}\}$,
and
$p^{adj}_{(K)} = \max\{p^{adj}_{(1)}, ..., p^{adj}_{(K-1)}, p_{(K)}\}$.

Notes:

- After adjustments, the p-values are compared to a chosen level α. Note that all p-values are inflated compared to their original level, making it less likely to reject the null hypothesis.
- Furthermore, the p-values keep their relative order after adjustment. This is governed by the $\max$ function,

which assures the order is kept. The resulting process is sequential, in that for a given level α, after the first non-rejected value, all others would not be rejected.

B.3. Proof of Theorem 4.2

Recall that D is the set of sign decisions, and that we assume that the probability of any error in D is less than α. To prove the theorem, we first show that any error in coverage of a rank confidence interval must be caused by at least one sign error in D:

Suppose that there is a coverage error for the set of confidence intervals, and without loss of generality assume that the coverage error occurs for feature 1,

$$[\ell_1, u_1] \nsubseteq [L_1, U_1].$$

The coverage can break on one or both sides:

If $\ell_1 < L_1$, then $L_1 > 1$. Therefore, in D, there are $L_1 - 1 > 0$ pairs of the type $(1, k) \in D$. (Meaning, there are $L_1 - 1$ features that the test found to be significantly smaller in FI compared to feature 1). At the same time, according to Definition 3.1, $\ell_1 - 1 = \#\{k : \Phi_1 > \Phi_k\}$, meaning there are only $\ell_1 - 1$ which are truely smaller than feature 1. Combining these two statements together, there must be at least one feature $k = 2, ..., p$ for which $(1, k) \in D$ but $\Phi_1 \not> \Phi_k$, meaning there is a sign error in D.

If instead $u_1 > U_1$, this means that the set $\{k \in 2, ..., p : (k, 1) \in D\}$ is larger than the set $\{k : \Phi_k > \Phi_1\}$. Again, this would mean that at least for one value of k we have a sign error.

Then the event of at least one coverage error is contained in the event of getting a sign error. So the probability of coverage error is bounded by the $FWER = \alpha$.

C. Experiments Details and Results

C.1. Synthetic data

C.1.1. Feature ranking - mock importance values

Data description For this experiment, we generate multivariate-normal data with off-diagonal correlations between the features and two levels of correlation: low ($\rho = 0.1$) and high ($\rho = 0.9$). The variance of the features was sampled from chi-squared distribution and was scaled by different factors: 0.2, 1, 5. The structure of the means vector is: $\mu = (1, \ldots, p + 1)^{\beta}$, where β gets different values that affects the spread of the means: 0.5 (most spread), 0.25, 0.1 (most compact). Figure 4 is a detailed version of Figure 2, separated by the different values of the exponent.

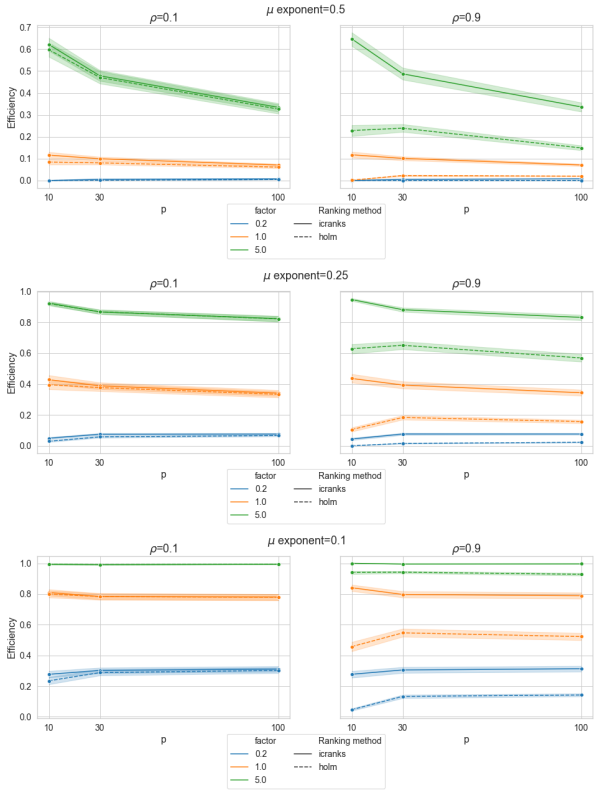

Figure 4. Ranking efficiency as a function of p for two levels of correlations ($\rho = 0.1$ and $\rho = 0.9$) and different values of μ exponent.

In addition, we analyzed the efficiency as a function of the number of observations (n), see summary in Figure 5 and the breakdown to different values of μ exponent in Figure 6. To evaluate the efficiency and simultaneous coverage of the ranking methods, we sampled 100 independent explanation sets for each configuration and reported the average across repetitions.

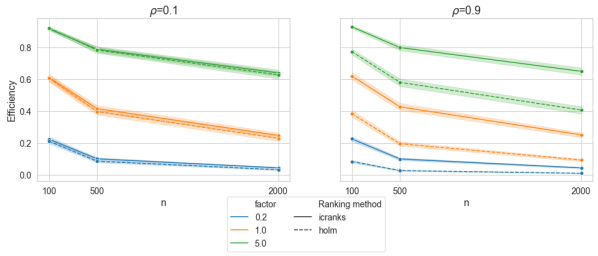

Figure 5. Ranking efficiency as a function of n for two levels of correlations ($\rho = 0.1$ and $\rho = 0.9$).

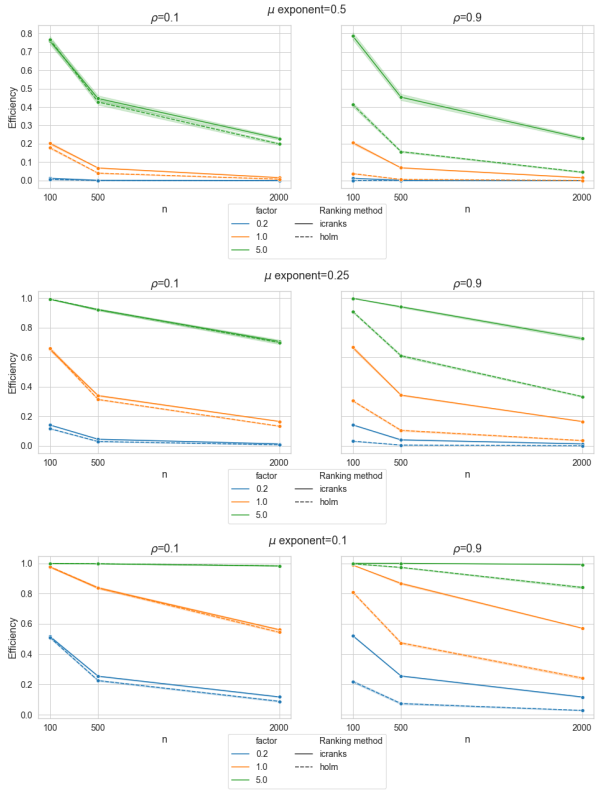

Figure 6. Ranking efficiency as a function of n for two levels of correlations ($\rho = 0.1$ and $\rho = 0.9$) and different values of μ exponent.

C.1.2. Feature ranking - TreeSHAP and PFI importance values

Here we simulate a regression task with X sampled from the following synthetic data: a multivariate normal distribution, with $\mu = (1, \ldots, p+1)^{0.5}$, a block-wise pairs correlation matrix with $\rho = 0.3$, and a chi-squared variance. In addition, we analyzed a different number of total features ($p = 15$ and $p = 50$), with varying numbers of important features. The value of Y was calculated based on the important features as a linear combination of the features with interactions or a trigonometric transformation of the sum of the important features. Finally, we added noise to Y and standardized it.

For each configuration of (X, Y), an XGBoost model was trained and tested (70/30) on 500,000 observations. Then, the true importance values were computed with TreeSHAP and PFI using 50,000 observations for $p = 50$ and 10,000 for $p = 15$.

To measure our ranking method as a function of the number of base importance values (n), we repeatedly (100 times) sample data from the same distribution for each value of n, compute the local TreeSHAP and PFI values, and ranked the features with our method. Then, we measure the efficiency and coverage. The reported results are the average across repetitions.

In Figure 3 we presented the results for a linear Y with interactions. Here (Figure 7) you can see the results for the trigonometric Y.

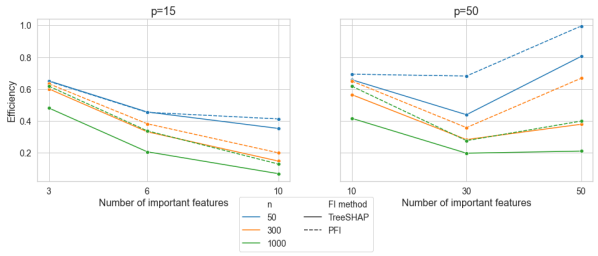

Figure 7. Ranking efficiency for $p = 15$ (left) and $p = 50$ (right) features with a trigonometric Y.

C.2. Real data

We demonstrated the motivation to use our confident feature ranking method with the bike sharing demand dataset (16).

We created a 60/40 train and test sets and trained an XGBoost model (sklearn default implementation; train and test R^2 of 0.98 and 0.94, respectively). We used TreeSHAP to build an explainer model and compute local SHAP values for each observation in the test set. We define the base importance values as the absolute values of the local SHAP values.

Based on a single sample, the *Naive* ranking of the features is the order of the bases' means, where the feature with the lowest mean is ranked as 1.

To measure the stability of the ranking as a function of the number of base values, we re-sampled with replacements from the test set. We evaluated the stability of the aggregation from base to global importance values. We define the stability of a rank as the frequency of the most common rank of the feature across repeated rankings (*Naive* rankings).