



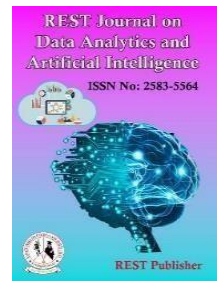
REST Journal on Data Analytics and Artificial Intelligence

Vol: 4(4), December 2025

REST Publisher; ISSN: 2583-5564

Website: <http://restpublisher.com/journals/jdaai/>

DOI: <https://doi.org/10.46632/jdaai/4/4/14>



How Stable Are Post-Hoc Explanations? A Robustness Study of Feature Attribution in Data Mining

Rahul Raj, *Ranjan Kumar Mishra

Netaji Subhas University, Jamshedpur (Jharkhand), India.

*Corresponding Author Email: ranjanlnmi4u@gmail.com

Abstract: Explainable artificial intelligence is increasingly used to justify the outputs of data mining models, but an explanation is only useful if it does not change every time the data, the model or the inputs change slightly. This paper studies the robustness of one widely used explanation, permutation-based feature importance, along three dimensions: resampling of the training data, choice of model family, and noise in the inputs being explained. Four model families were trained on two public datasets, a three-class wine classification task and a diabetes progression regression task, and their importance rankings were compared using Kendall's rank correlation. Rankings obtained from bootstrap-retrained random forests agreed only moderately (mean tau 0.52 on wine and 0.58 on diabetes), rankings from different model families agreed at 0.41 and 0.67, and Gaussian input noise at a quarter of a feature's standard deviation reduced agreement to 0.64 and 0.78. Linear models gave the most stable rankings. In both datasets, removing the most important feature caused a closely related feature's importance to rise sharply, showing that credit is redistributed among related attributes. We recommend that explanations be reported with uncertainty and that single-run feature rankings not be presented as definitive.

Keywords: explainable AI, data mining, feature importance, robustness, rank stability, permutation importance, trustworthy machine learning.

1. INTRODUCTION

A mining model that predicts well but cannot be questioned is hard to deploy where decisions carry consequences. Practitioners have therefore adopted explanation tools that rank which input attributes the model relied upon. These rankings are then used to justify decisions to clients, to select variables for later studies, and to argue that a model has learned sensible relationships. The original aim of knowledge discovery was to find patterns that are valid and understandable [1], and a feature ranking is often treated as the understandable part. The difficulty is that a ranking is itself the output of a computation with random and arbitrary elements: the sample the model was trained on, the learning algorithm, the data used for scoring, and the perturbation scheme. If these choices change the ranking substantially, then the ranking says as much about the analyst's procedure as about the world. Concerns of this kind have been raised for gradient-based and local explanations [12], [13], for adversarial manipulated explanations [14] and for the disagreement between methods applied to the same model [15], but practical mining guidance on how much variation to expect from the simplest and most commonly reported global measure is thin. This paper asks three questions about permutation feature importance. RQ1: how much does the ranking change when the training sample changes? RQ2: how much does it change when the model family changes while the data stay fixed? RQ3: how much does it change when noise is added to the data being explained? A fourth, exploratory question follows from the literature on correlated predictors: what happens to the ranking of other features when the top feature is removed? We answer them empirically on two public datasets and draw guidance for reporting explanations in mining projects.

2. RELATED WORK

Surveys of explainable AI organise methods by scope, access to model internals and form of output [5], [6], and Doshi-Velez and Kim call for evaluation that is rigorous and tied to the task [7]. Rudin argues that opaque models explained after the fact are a poor substitute for models that are interpretable by design [4]. Local attribution methods, including LIME and Shapley-value approaches, are well known [2], [3]; Miller reminds the community that human users expect explanations to be selective and contrastive [17]. Permutation importance has a long history in tree ensembles [8], and Fisher, Rudin and Dominici extended it to a model-class view, noting that a single model's importance may not reflect what is learnable from the data [9]. Strobl and colleagues showed that importance for correlated predictors is distorted and proposed conditional schemes [10], while later work has pointed out that permuting one feature independently forces the model to evaluate unrealistic combinations of inputs

[11]. Evidence of fragility comes from studies of interpretation methods for neural networks [13], of robustness of explanations to input perturbation [12], and of saliency maps that are insensitive to the model they purport to describe [16]. A practitioner-oriented study reports that analysts frequently face disagreeing explanations and lack principled ways to resolve them [15]. The present work complements these by isolating three controlled sources of variation for one global explanation on small tabular datasets, which are typical of many applied mining projects.

3. SOURCES OF EXPLANATION VARIABILITY

We distinguish the sources of variability summarised in Table I. The first two arise before any explanation is computed, in the construction of the model. The third arises in the scoring data, and the fourth is the inherent redundancy of the attributes. Only the first three are varied systematically here; the fourth is examined through a case study.

TABLE I. Sources of Variability Studied

Source	Manipulation	Question
Training sample	Bootstrap resampling, 25 draws	RQ1
Model family	Four algorithms, same split	RQ2
Scoring data	Gaussian input noise	RQ3
Redundancy	Drop top feature, refit	Exploratory

4. EXPERIMENTAL DESIGN

Datasets: Two datasets distributed with scikit-learn [23] were used. The wine dataset [21] has 178 samples, 13 chemical measurements and three cultivar classes. The diabetes dataset [22] has 442 patients, 10 baseline variables and a continuous measure of disease progression one year later. Each was split 70/30 with a fixed seed, giving test sets of 54 and 133 cases. The small test sets are typical of applied work but add sampling noise, which we return to in the discussion.

Models and importance measure: For each dataset four model families were trained: a random forest, gradient boosting, a support vector machine with radial kernel (SVM for wine, SVR for diabetes), and a linear model (logistic regression for wine, ridge regression for diabetes), the SVM and linear models after standardisation. Importance was measured by permutation on the test set: the drop in accuracy (wine) or in R^2 (diabetes) when a feature's values are randomly shuffled, averaged over repeated shuffles [8], [9].

Stability measures: Agreement between two rankings is quantified by Kendall's rank correlation coefficient [20], which ranges from -1 to 1 and equals 1 for identical orderings. For RQ1, 25 bootstrap samples of the training data [19] were drawn, a random forest was refitted on each, importance was computed on the fixed test set, and agreement was averaged over all pairs of the 25 rankings. For RQ2, rankings from the four models fitted on the original training split were compared pairwise. For RQ3, Gaussian noise with standard deviation of 0.1 , 0.25 and 0.5 times each feature's training standard deviation was added to the test inputs, importance was recomputed for the already fitted model, and agreement with the clean ranking was averaged over eight noise draws. We also report the overlap of the three highest ranked features as a measure that matters when only a short list is communicated.

5. RESULTS

Predictive performance

Table II shows that all models were competent, so differences in explanations are not attributable to a model that failed to learn the task. The random forest was the best performer on both datasets, though the margins were small.

TABLE 2. Test Performance (Accuracy For Wine, R^2 For Diabetes)

Model family	Wine	Diabetes
Random forest	0.981	0.490
Gradient boosting	0.963	0.458
SVM / SVR (RBF)	0.944	0.417
Logistic / ridge regression	0.963	0.439

Stability across resampling (RQ1)

Bootstrap-retrained forests produced rankings that agreed only moderately: mean tau was 0.52 (standard deviation 0.17) on wine and 0.58 (0.14) on diabetes. The top-three feature sets had mean Jaccard similarity of 0.58 on wine and 0.69 on

diabetes, meaning that on wine a typical pair of runs shared fewer than two of their three leading features. The same model class trained on slightly different samples can therefore tell a noticeably different story about what matters.

Agreement across model families (RQ2)

Cross-model agreement was lower on wine (mean tau 0.41, range 0.25 to 0.53) than on diabetes (0.67, range 0.51 to 0.82). On wine the weakest pair was gradient boosting against logistic regression, and the strongest was the pair of tree ensembles. On diabetes the strongest agreement, 0.82, was between the random forest and the support vector regressor, a pairing one would not necessarily predict from their structure. The mean overlap of the top-three sets was 0.72 on wine and 0.83 on diabetes, so the leading attributes were often shared even when the full orderings were not. On diabetes the two dominant attributes, body mass index and the serum measure s5, led the random forest ranking by a wide margin, which explains why short lists agree more than full rankings.

TABLE 3. Summary Of Rank Agreement (Mean Kendall's T)

Condition	Wine	Diabetes
Bootstrap resampling	0.52	0.58
Across model families	0.41	0.67
Input noise, 0.1 SD	0.73	0.85
Input noise, 0.25 SD	0.64	0.78
Input noise, 0.5 SD	0.57	0.73

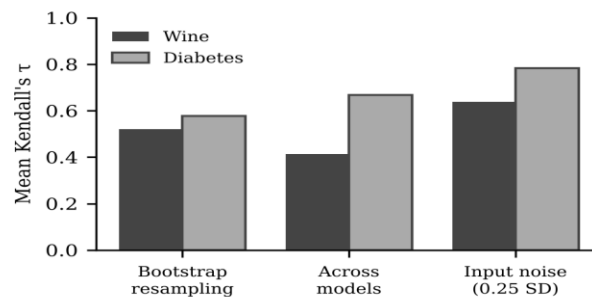


FIGURE 1. Mean rank agreement under resampling, model change and moderate input noise.

Sensitivity to input noise (RQ3)

Agreement with the clean ranking declined steadily as noise increased (Table III and Fig. 2). Linear models were the most robust: logistic regression retained tau of 0.66 and ridge regression 0.93 at the highest noise level. The random forest was the most sensitive on wine, falling from 0.65 to 0.48 between the lowest and highest noise levels, and gradient boosting and the random forest were the two most affected on diabetes. A plausible reading is that tree ensembles split on thresholds, so noise moves instances across split points, while a linear model's dependence on each feature is smooth. We did not test this explanation directly.

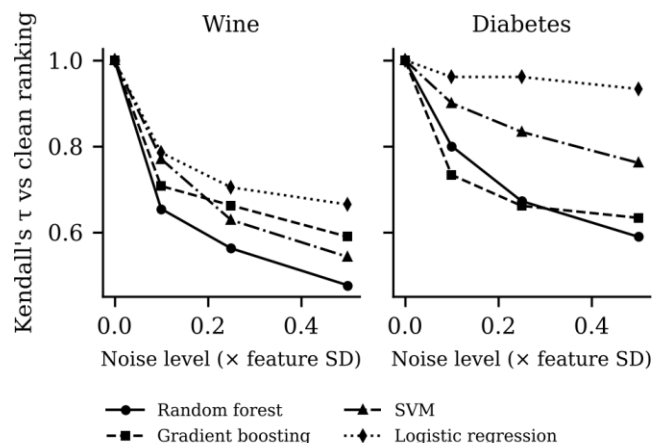


FIGURE 2. Rank agreement with the clean ranking as input noise grows, by model family (SVM and logistic regression on wine; SVR and ridge regression on diabetes)

Redundancy case study

Both datasets contain correlated attributes: wine has three feature pairs with absolute correlation above 0.7 and a maximum of 0.88, and diabetes has two pairs above 0.7 with a maximum of 0.90. For each dataset we removed the top-ranked feature from the random forest's inputs, refitted, and observed the importance of the feature most correlated with it (Table IV). On wine, alcohol content (correlation 0.55 with the removed feature, colour intensity) rose in importance from 0.038 to 0.159 while test accuracy was unchanged. On diabetes, the serum measure s5 (correlation 0.46 with body mass index) rose from 0.194 to 0.369 while R^2 fell by 0.13. The partners in these cases are only moderately correlated with the removed attribute, so the effect is not limited to near-duplicates. In both cases the credit assigned to the removed feature was partly taken up by another, which means that a feature's rank is conditional on which other features are present.

TABLE 4. Effect Of Removing the Top Feature

	Wine	Diabetes
Removed feature	colour intensity	BMI
Most correlated partner	alcohol	s5
Correlation (abs.)	0.55	0.46
Partner importance before	0.038	0.194
Partner importance after	0.159	0.369
Change in test score	0.000	-0.131

6. DISCUSSION

The central observation is that a single permutation-importance ranking is a point estimate with considerable spread. With modest sample sizes, two equally reasonable analysts who resample the data, choose a different algorithm, or work with slightly noisier measurements can arrive at visibly different rankings. This does not make the explanation worthless: the top few attributes recurred across conditions, especially on diabetes, and short lists were more stable than complete orderings. It does mean that communicating the full order as if it were exact overstates what the analysis supports. Several recommendations follow. Explanations should be reported with an uncertainty band obtained by bootstrap or repeated fitting, as is routine for accuracy. When a ranking matters for a decision, it should be checked against at least one model from a different family, and disagreements should be disclosed. Where a short list is all that is needed, stability of the leading set should be preferred to the stability of the entire order. Finally, claims about the importance of any one attribute should be qualified by the presence of related attributes, since the redundancy case shows that importance migrates when features are removed or added. Limitations are plain. Only two small datasets and one importance measure were studied, test sets were small, and the correlation of partner features was only examined for the single most correlated attribute. Noise was added to the scoring data only, not to training data. Local explanation methods, Shapley-based attributions and conditional permutation schemes [10] were not evaluated and may behave differently. Results on high-dimensional or image data could differ markedly.

7. CONCLUSION

Permutation importance, one of the most commonly reported global explanations, was only moderately stable under resampling, change of model family and input noise on two small tabular datasets, with rank correlations typically between 0.4 and 0.8. Linear models gave the most stable rankings, short lists of leading features were more stable than full orderings, and removing one feature shifted importance to related ones. Explanations in data mining should therefore be reported with variability estimates and cross-checked across models. Future work will test conditional and Shapley-based measures, larger and higher-dimensional data, and noise in the training data, and will study how decision makers respond when explanations are presented with explicit uncertainty.

REFERENCES

- [1]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, pp. 37–54, 1996.
- [2]. M. T. Ribeiro, S. Singh, and C. Guestrin, "“Why should I trust you?”: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
- [3]. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.* 30, 2017, pp. 4765–4774.
- [4]. C. Rudin, "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.
- [5]. A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

- [6]. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, Art. no. 93, 2018.
- [7]. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
- [8]. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [9]. A. Fisher, C. Rudin, and F. 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.*, vol. 20, no. 177, pp. 1–81, 2019.
- [10]. C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, Art. no. 307, 2008.
- [11]. G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance," *Stat. Comput.*, vol. 31, Art. no. 82, 2021.
- [12]. D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," arXiv:1806.08049, 2018.
- [13]. A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3681–3688.
- [14]. D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2020, pp. 180–186.
- [15]. S. Krishna et al., "The disagreement problem in explainable machine learning: A practitioner's perspective," arXiv:2202.01602, 2022.
- [16]. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process. Syst.* 31, 2018, pp. 9505–9515.
- [17]. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019.
- [18]. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [19]. B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, no. 1, pp. 1–26, 1979.
- [20]. M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1–2, pp. 81–93, 1938.
- [21]. M. Forina et al., "PARVUS: An extendible package for data exploration, classification and correlation," *Inst. Pharmaceutical Food Anal. Technol.*, Genoa, Italy, 1988.
- [22]. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
- [23]. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.