



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Explainable Artificial Intelligence in Data Science: Enhancing Transparency and Trust in Predictive Models

Chaya Lakshmi

HR Lecturer, Department of Computer Science and Applications, Vishvesvaraya Technological University, Karnataka, India, Email: Chayagowda1@gmail.com

Abstract

In today's data science world, the use of ensembles, deep neural networks and other higher capacity algorithms that have made it their business to make decisions are now a primary tool in predictive modelling. Today's data science world is focused on making predictions using algorithms which are ensembles, deep neural networks, and other high capacity algorithms whose internal workings it is difficult for the user to probe. Explainable Artificial Intelligence (XAI) helps solve this problem by generating evidence that is comprehensible to humans that allows for understanding how the input influences the output of the model. In this paper, we focus on XAI as a practical layer to help increase the transparency, debugging, accountability and calibrated trust of models. A critical review of major approaches to XAI is completed as well as a demonstration of an empirical study, using the Wisconsin Diagnostic Breast Cancer dataset included with scikit-learn. A fixed stratified train-test split is used to train the logistic regression, random forest and gradient boosting classifiers. The results from the tests that were not performed demonstrate very high predictive ability; logistic regression achieves an accuracy of 0.986 and ROC-AUC of 0.998, while random forest and gradient boosting both achieve an accuracy of 0.958. The advantages of using interpretable evidence to accompany global permutation importance and local coefficient based contribution analysis are illustrated. In the paper, it is suggested that explanations should not be accepted as evidence of the fairness, causality or safety of a model. What are the attributes of trustworthy XAI: fidelity, stability, appropriateness for the audience, uncertainty awareness and human oversight. The study ends with a practical approach to incorporating explanation into the data-science lifecycle.

Keywords: explainable artificial intelligence; XAI; data science; predictive modelling; interpretability; transparency; trust; SHAP; LIME; permutation importance.

1. Introduction

Machine learning has revolutionized data science by being able to predict from high dimensional data that is linear, nonlinear, and heterogeneous. However, the predictive power can lead to an operational issue: a model can have a score/classification result and not provide a decision-maker with a good explanation for why that result happens. This is particularly relevant if the forecasts impact on health, finance, employment or consequential settings, such as public services. Explainable artificial intelligence (XAI) is thus a set of techniques and design principles that are designed to increase the understandability of the behaviour of models for developers, domain experts, decision makers and individuals affected by the model.

The need for explanation isn't just a desire for pretty feature-importance graphics. Ribeiro, Singh and Guestrin proposed LIME, a simple local interpretable model that approximates a complex classifier locally by forming an interpretable model that explicitly connects with the question of when the complex model's prediction should be trusted. Lundberg and Lee then went on to propose a feature-attribution value based on a unified additive framework based on Shapley-value ideas. Other taxonomies differentiate between intrinsic interpretation and post hoc explanation, and local versus global explanation and between model-specific and model-agnostic approaches. These differences are important because they lead to different types of questions asked: a developer may be interested in spurious correlations for debugging, a regulator may desire a trail of auditability, and an end user may require a short explanation for a specific prediction.

XAI also has its boundaries. When it comes to the high-stakes applications, Rudin's view is that it's best to use models that are inherently interpretable, providing good predictive results, than to try to explain opaque systems. Adebayo et al. illustrate that some "apparently" plausible saliency methods are unable to pass simple sanity check,

and Slack et al. show that perturbation-based post-hoc explanations can be manipulated. Therefore, explanation needs to be assessed as an empirical object and not taken for granted simply because it seems intuitive.

There are two aims in this paper. First, it explains the role of XAI within modern data-science workflows, focusing on the importance of transparency, trust, accountability and model validation. Second, it offers a reproducible empirical example by showing how predictive performance and explanation can be reported together with a public benchmark data set. The main idea is that XAI can increase trust if it helps to verify or inform human judgment, yet can decrease it appropriately if it reveals instability, bias, data leakage or reliance on implausible features.

2. Literature Review

There are no agreed-upon definitions for interpretability. Doshi-Velez and Kim highlight the importance of rigorous evaluation and differentiate between application grounded evaluation, human grounded evaluation and functionally grounded evaluation. Lipton's critique on the “mythos” of interpretability also cautions against mixing-up transparency, decomposability, simulatability and post-hoc explanation. There are several ways of categorizing black-box explanation methods, as described by both Guidotti et al., who categorize them by the type of model, explanation target and explanatory representation, and Arrieta et al., who link XAI to the wider agenda of responsible AI.¹

Local surrogate methods like LIME model one prediction by sampling observations in the vicinity of the target point and query the black-box model and fit a simpler weighted surrogate. They are model agnostic and have local reasoning. They are sensitive to neighborhood definition, sampling and model nonlinearity. SHAP is a more formal additive attribution framework and it has gained wide usage due to the ability to sum up local attributions to obtain global summaries. SHAP values are not estimates of causal effects, however, they are explanations of how the model behaves. Interpretation can be significantly impacted by correlated predictors, background-data selection, and distribution shift.

Global explanation techniques include permutation importance, partial dependence, accumulated local effects, surrogate trees and rule extraction. Global explanations provide answers to questions about what the model is generally based on, while local explanations answer why a specific case got a specific prediction. In practice both levels are required. A global plot might show that a model is very sensitive to a plausible clinical variable, and a more local explanation might demonstrate that a particular prediction is strongly affected by an unusual measurement. The best workflows, then, take several explanation forms rather than relying on one particular plot to be the final one.

The ties between explainability and trust are more complex as well. Explanations can make users more confident, but confidence does not equal warranted trust. An unfaithful explanation, if convincing, can lead to overreliance. The quality of the explanation should thus be assessed based on factors like fidelity to the model, stability with minor variations, consistency with the domain knowledge, comprehensibility to the target group and usefulness to a specific decision. XAI is best used as a means to question, but not necessarily to defend a model.²

3. Conceptual Framework for XAI in Data Science

A data-science system may be seen as a series of activities beginning with data collection and preprocessing, followed by model training, model validation, model explanation and human action. Don't bolt XAI on as an afterthought after deployment. Explanation Requirements should affect feature engineering, model selection, design and documentation of validation. A brief summary of the workflow used in this paper is given in Figure 1.



Explanations should be evaluated for fidelity, stability, usefulness and context—not visual appeal alone.

¹ Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31. [Active link](#)

² Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. [Active link](#)

Figure 1 XAI-integrated predictive modelling workflow

4. Research Methodology

The empirical part is based on a dataset of Wisconsin Diagnostic Breast Cancer (WDBC) that is provided with scikit-learn. It consists of 569 observations and 30 continuous features derived from digitized images of breast-mass fine needle aspirates, along with a binary diagnostic target. The dataset was used here only as a methodological resource for modelling and explaining predictions, and is not offered as a clinical decision system.³

The data was divided into an 75/25 split between train and test, `random_state=42` to maintain class proportions. Three classifiers were tested: logistic regression with standardized predictors, a 500-tree random forest with class weighting and gradient boosting. The held-out test set was used for calculating accuracy, precision, recall, F1-score and ROC-AUC. There were no test observations to fit the models.⁴

Two complementary strategies were employed for explanation. First, permutation importance was calculated for the random forest by shuffling each feature in the test set and calculating the average ROC-AUC reduction over 100 repetitions. Permutation importance of the random forest was computed by shuffling each feature in the test set and calculating the average ROC-AUC reduction over 100 repetitions of the random forest model. The greater the reduction, the more influential that feature is for the discrimination made by the trained model. Second, an explanation was generated locally for one of the prediction holdouts in logistic regression by scaling each feature value by its fitted coefficient. The linear decision score is identical to this signed contribution, with the exception of the intercept, which may be different, and thus gives a clear, instance-level decomposition.⁵

The analysis is deliberately intended to distinguish between predictive evaluation and explanation. There is a difference between having a high predictive performance and being a valid explanation method, and a plausible explanation and model generalizability. Algorithms are standard and publicly-available, and reproducibility is ensured by the use of a fixed random seed.⁶

5. Results

Table 1 Held-out predictive performance

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.986	0.989	0.989	0.989	0.998
Random Forest	0.958	0.957	0.978	0.967	0.995
Gradient Boosting	0.958	0.947	0.989	0.967	0.992

Both models showed good discrimination on the fixed test split for all three models. The highest overall scores were obtained for logistic regression with accuracy 0.986, F1 0.989 and ROC-AUC 0.998. The random forest model and the gradient boosting model both had an accuracy of 0.958 and a ROC-AUC of 0.995 and 0.992, respectively. These findings are relevant to the discussion of interpretability since a rather simple linear model can do at least as well on this dataset and split as can more complicated ensemble alternatives. This is an observation that concurs with the principle of not assuming you need a more complicated model when it isn't proven.⁷

³ Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. Active link

⁴ Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. Active link

⁵ Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42. Active link

⁶ Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. Active link

⁷ Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. Active link

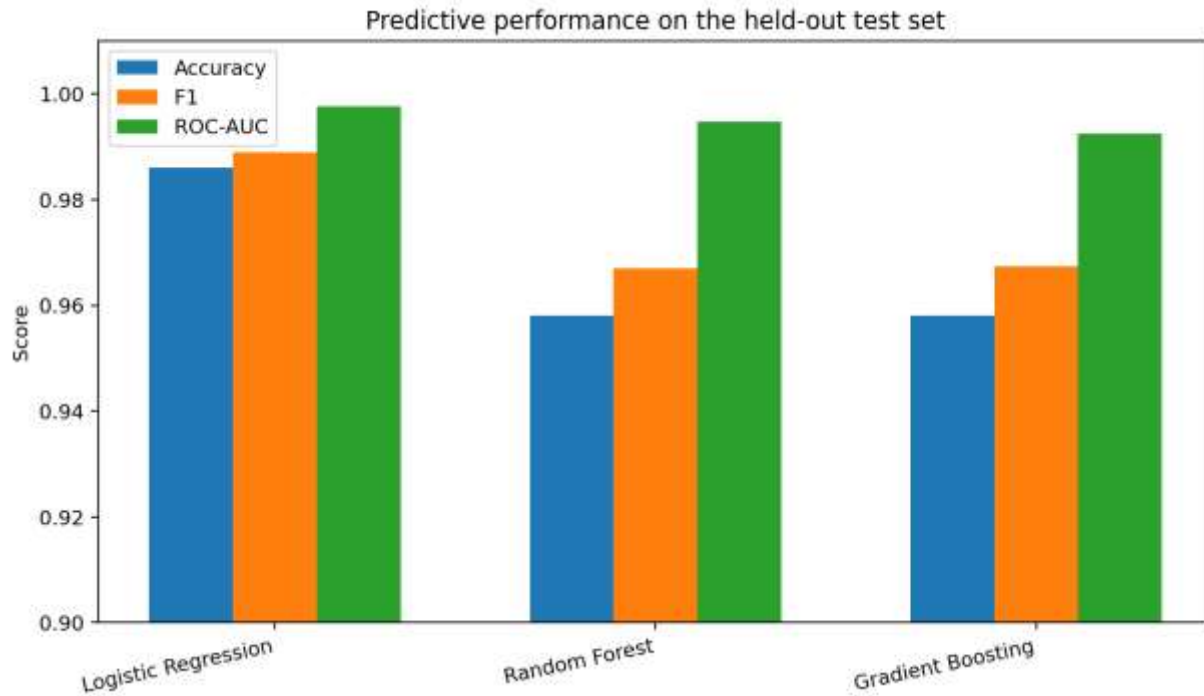


Figure 2 Accuracy, F1 and ROC-AUC of the three predictive models

5.1 Global Explanation

Permutation importance gives to the random forest a model-agnostic perspective on the variables on which it most heavily relies. The measure is based on the amount of degradation that occurs when one feature is disrupted at a time. It is thus directly related to predictive behaviour, but may underestimate or re-allocate importance if predictors are highly correlated. The ten largest average reduction of ROC-AUC in this test set analysis are reported in Figure 3.⁸

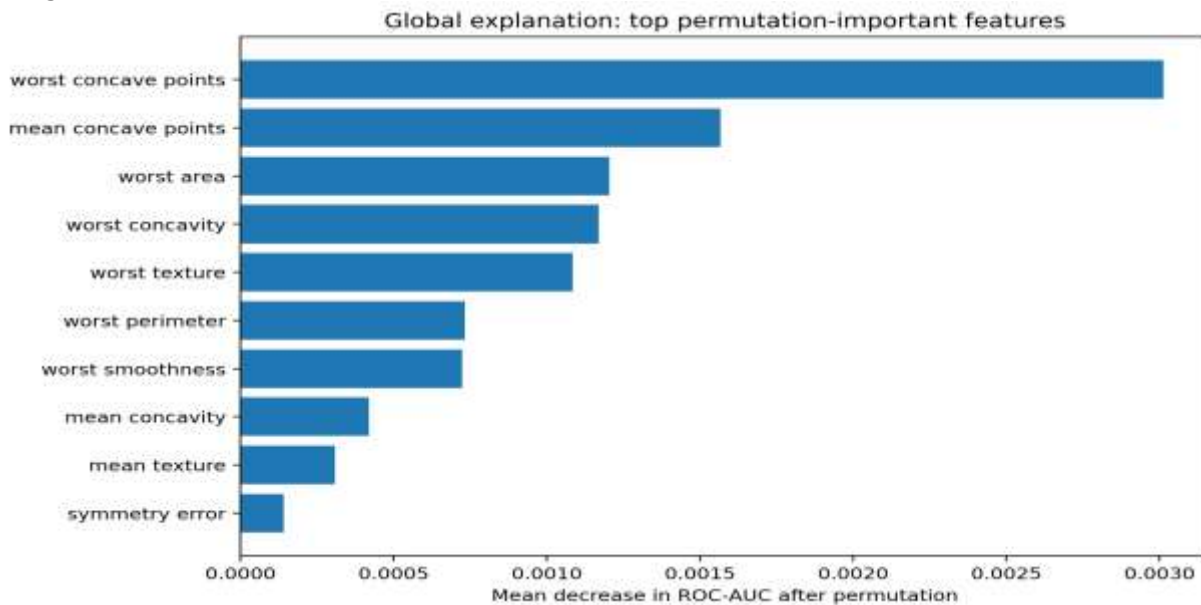


Figure 3 Top ten random-forest permutation importances

⁸ Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–186. Active link

5.2 Local Explanation

A "local explanation" is offering an explanation for one particular prediction. All standardized logistic regression models can be split into intercept + sum of feature-level products. Positive and negative scores go in opposite directions. The top 10 absolute contributions of one hold out observation are plotted in Figure 4. This explanation is instance-specific and thus can show when an otherwise secondary variable is influential for an individual case, as opposed to using a generic feature-ranking chart.⁹

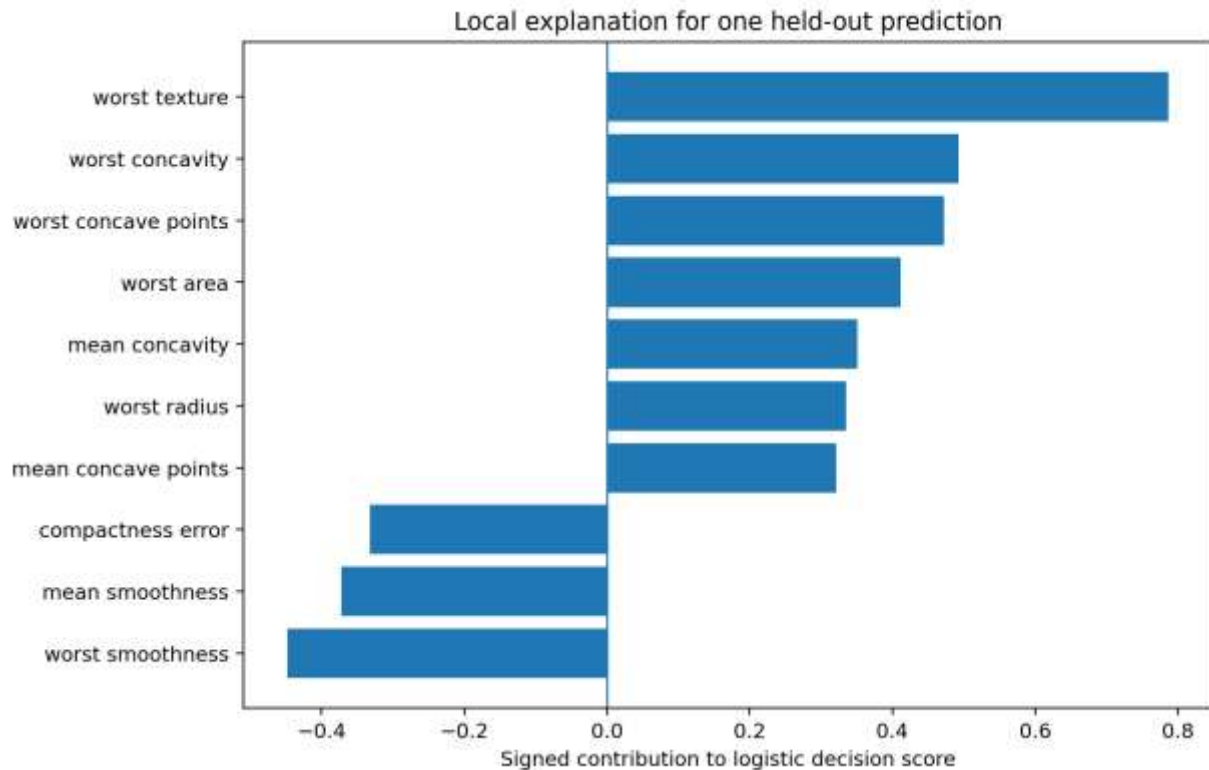


Figure 4 Signed local contributions for one held-out logistic-regression prediction

6. Discussion

The following three principles are illustrated in the empirical demonstration: First, explanation should be used in conjunction with, and not instead of, conventional validation. Out-of-sample tests were conducted on the models before interpretation, to compare them. Second, the choice of a model is a choice of an explainability. The logistic regression model has the best held-out performance and it is intrinsically decomposable in this experiment. Thus it would be necessary to have a specific reason to use an opaque ensemble, for example robustness across external datasets or better performance under a clinically relevant operating point.¹⁰

Third, local and global explanations are complementary. Permutation importance will tell you which variables are important to the overall performance of the model, while the local contribution plot will tell you the direction and magnitude of the evidence for one prediction in a linear model. Depending on the model class and the needs of stakeholders, a mature XAI system might include SHAP summaries, LIME explanations, counterfactual examples, or partial dependence or concept-based explanations. The question should guide the choice of the method and not popularity.¹¹

Trust can be strengthened by transparency in a number of ways. It can detect data leakage, dependence on the protected or proxy variables, improbable correlation, error analysis and communicate with domain experts. Also, it generates a record for post deployment audit. However, transparency can also be an appropriate reduction of trust. When explanations are unstable over small perturbations, change radically upon retraining, or stress features without domain plausibility, the appropriate answer is further investigation, not reassurance.

⁹ Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. [Active link](#)

¹⁰ Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. [Active link](#)

¹¹ Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *KDD*. [Active link](#)

This separation from trust and trustworthiness is a key one. A reliable predictive system should be trusted following evidence of performance, robustness, fairness, security and appropriate governance. XAI provides evidence of model reasoning, but can not independently ensure these properties. In particular, feature attribution is associational, not causal, with respect to the fitted model, not causal to changing the real-world outcome. Similarly, when an explanation is missing a sensitive feature, this does not mean that discrimination is absent, for a proxy variable and interaction effects could still be present.

The literature also shows that explanation methods can also go wrong. An examination of the literature confirms that saliency maps that are convincing to the human eye might not react well to randomization of the model or the label. When adversarial work is applied to LIME and SHAP, it is shown that they can be tricked with carefully crafted conditions. These results serve as motivation for explanation validation. Helpful tests: repeated explanations under resampling, sensitivity to small input perturbations, comparison across explanation methods, known-feature recovery on synthetic benchmarks, human-grounded evaluation with the actual intended users.¹²

7. Practical XAI Framework for Predictive Modelling

A practical implementation can be arranged in five controls. The first is model proportionality: Employ the most simple model that meets the predictive and operational needs, particularly with respect to consequential decisions. The second involves explanation matching: identify global, local, counterfactual or example-based explanations, based on the type of question asked by the stakeholder. The third is explanation validation: test fidelity, stability and sensitivity; don't assume validity. The fourth is uncertainty communication: present probabilities, confidence or prediction intervals as applicable; do not turn uncertain predictions into categorical statements of certainty. The fifth is governance: document data provenance, model version, explanation method, known limitations, and explanation human decision process.

In the context of data science teams, this implies an XAI report should be reproducible and versioned, along with the model itself. Training data, explanation methods, evaluation metrics, monitoring thresholds, subgroup checks, training authorizations and use exemptions can be documented on a model card or other technical document. If the data distribution changes, the performance of the predictions and the structure of explanations should be reviewed. Stable accuracy score combined with a significant change in the features used to achieve the accuracy may be a bad sign of different operating conditions.¹³

8. Limitations

The empirical study is a methodological demonstration, on a relatively small and clean benchmark data set. Findings not to be extrapolated to clinical use. To the extent that the fixed train-test split can be reproduced, that is the case, but only repeated cross-validation or external validation will replace it. Permutation importance may be misleading if the predictors are correlated, and the local linear contribution analysis is only applicable to the fitted logistic model. The study is not a user experiment and so cannot conclude that the explanations offered in the study increase human trust in a psychological sense, as measured. Rather, trust is talked about as warranted reliance, which is accompanied by clear evidence.¹⁴

9. Conclusion

EAI is becoming more crucial as predictive performance is no longer enough when models inform consequential decision-making. XAI can provide greater insight into how a system works, by displaying global dependencies, local decision drivers and potential failure modes, making predictive systems more inspectable. The results are an empirical illustration of how a typical data-science paper can report predictive metrics, alongside global and local explanations, and how a transparent model may achieve as good or better performance on a task as more complex models. The key takeaway is to value explanation as part of the model's validation and governance process rather than as a pretty picture. Transparency enhances trust only when explanations are faithful, stable, appropriately scoped and open to challenge. Predictive accuracy and interpretability, uncertainty awareness, domain expertise and human accountability are, therefore, a cornerstone of responsible data science.¹⁵

¹² Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv. Active link

¹³ Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology. Active link

¹⁴ Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. KDD. Active link

¹⁵ Molnar, C. (2022). Interpretable Machine Learning (2nd ed.). Active link

REFERENCES

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31. [Active link](#)
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. [Active link](#)
3. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. [Active link](#)
4. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42. [Active link](#)
5. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. [Active link](#)
6. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. [Active link](#)
7. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. [Active link](#)
8. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. [Active link](#)
9. Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43. [Active link](#)
10. Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. [Active link](#)
11. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *KDD*. [Active link](#)
12. Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv*. [Active link](#)
13. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*. [Active link](#)
14. Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. *KDD*. [Active link](#)
15. Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). [Active link](#)