

# Interpreting Classification Models Using Feature Importance Based on Marginal Local Effects

Rogério Luiz Cardoso Silva Filho<sup>1,2</sup>(⊠) , Paulo Jorge Leitão Adeodato<sup>2</sup>, and Kellyton dos Santos Brito<sup>3</sup> .

<sup>1</sup> Instituto Federal do Norte de Minas Gerais – IFNMG, Montes Claros, Brazil rogerio.luiz@ifnmg.edu.br

<sup>2</sup> Centro de Informática, Universidade Federal de Pernambuco – UFPE, Recife, Brazil
 <sup>3</sup> Universidade Federal Rural de Pernambuco – UFRPE, Recife, Brazil

Abstract. Machine learning models are widespread in many different fields due to their remarkable performances in many tasks. Some require greater interpretability, which often signifies that it is necessary to understand the mechanism underlying the algorithms. Feature importance is the most common explanation and is essential in data mining, especially in applied research. There is a frequent need to compare the effect of features over time, across models, or even across studies. For this, a single metric for each feature shared by all may be more suitable. Thus, analysts may gain better first-order insights regarding feature behavior across these different scenarios. The  $\beta$ -coefficients of additive models, such as logistic regressions, have been widely used for this purpose. They describe the relationships among predictors and outcomes in a single number, indicating both their direction and size. However, for black-box models, there is no metric with these same characteristics. Furthermore, even the  $\beta$ -coefficients in logistic regression models have limitations. Hence, this paper discusses these limitations together with the existing alternatives for overcoming them, and proposes new metrics of feature importance. As with the coefficients, these metrics indicate the feature effect's size and direction, but in the probability scale within a model-agnostic framework. An experiment conducted on openly available breast cancer data from the UCI Archive verified the suitability of these metrics, and another on real-world data demonstrated how they may be helpful in practice.

Keywords: Feature importance · Explainable artificial intelligence · ALE plots

# 1 Introduction

Explainable artificial intelligence (XAI) is an emerging research area that enables blackbox models to become trustworthy for humans. With a growing interest in explaining machine learning (ML) models to fill the gap between interpretability and prediction performance, over the past few years, many techniques have been proposed, and explainability has become an essential subfield of ML [1]. This combination has helped the spread of ML in applied research areas even more, such as in education, healthcare, finance, and social media.

<sup>©</sup> Springer Nature Switzerland AG 2021

A. Britto and K. Valdivia Delgado (Eds.): BRACIS 2021, LNAI 13073, pp. 484–497, 2021. https://doi.org/10.1007/978-3-030-91702-9\_32

For instance, simply classifying a patient in a hospital into a particular health status is not particularly helpful. It is more desirable to investigate the conditions that have contributed to this [2] and would even become compulsory should any legal matters arise. Additionally, in the education domain, discovering why a student will drop out is more valuable than just predicting it [3, 4] because, as in medicine, the treatment depends on the probable cause. Similarly, auditing the behavior of machine learning bot detectors in social media is valuable in order to improve the models for new kinds of bots [5, 6].

Explanations may be expressed in many forms. For classification problems, specifically, feature importance is widely used [7, 8], and demonstrates the global impact of single features in the model predictions. A wide variety of different methods with different feature importance representations have been proposed for this purpose [9–11]. Despite these advances, there is still a lack of understanding as to how these methods are related and whether one method is preferable over another [12].

In applied research it is often necessary to track feature importance over time, across models or even across studies. Therefore, a method that enables the global feature contribution to be represented by a single metric is more suitable than multiple metrics or graphical representation, otherwise interpretability may be challenging to understand when handling several features in several models.

A standard single metric of feature importance is the coefficients of additive models such as linear and logistic regressions. This coefficient represents the weight of each feature in the additive function, which describes the relationship among features and outcome. However, for generalized linear models (GLM), which involve transformations of this linear predictor into other discrete outcomes, such as logistic regressions, the coefficient interpretability is not straightforward. Moreover, the coefficients are highly sensitive to unobserved heterogeneity [13] and data scale [14].

To circumvent these limitations, marginal effects (MEs) have long been proposed [14] in the traditional statistical literature. Marginal effects use the prediction function to calculate the differences in probabilities of the outcome when the features partially change from one specified value to another. However, MEs fail to isolate the feature effect when data are correlated [15]. This problem arises when the computation of the feature effect uses conditional distribution. Thus, since correlated features move in tandem, it is unable to distinguish which feature value changes influence the model probabilities.

For black-box models however, the permutation feature importance (PFI) derived from tree-based algorithms is a standard single metric used to report feature contribution in classification problems. However, this kind of metric is linked to model error, which cannot be a metric of interest for analysts [15]. Furthermore, it does not report the direction of the feature effect, which may be critical for actionable research. Recently, SHapley Additive exPlanations (SHAP) [12] have been pointed out as the most common explainability technique in organizations [1]. This technique uses game theory (Shapley Values) to measure the contribution of each data point to each feature value. Thus, it can deliver explanations in fine grain, and by means of the average, report the global feature contributions, including their direction. Unfortunately, calculating the exact Shapley value is computationally expensive [15], and several approximations have been proposed [16].

Moreover, both PFI and SHAP are permutation-based techniques, and so, they are able to randomly sample from the marginal distribution considering unrealistic data points that are not present on training data. Therefore, they extrapolate in areas where the model was trained with either little or no training data, which may cause misleading explanations [17].

Recent advances in the interpretable machine learning field, such as accumulated local effects (ALE) plots, have put forward relevant contributions in this direction. They have shed light on detecting a more reliable feature effect with low computational cost when features are correlated. However, ALEs have only been used to visualize the feature effects across different values by plots, which are not visually friendly when the analyst compares the feature importance across multiple models.

Thus, this paper proposes new metrics of feature importance as a valuable option when compared to those already in existence. They allow direct interpretation in the probability scale, are more realistic when dealing with correlated data and are modeled in a model-agnostic framework. Although these metrics may be extrapolated for any class of supervised models, in this paper they are focused only on binary classification. Experiments use open-access data from the UCI Archive to introduce differences among provided metrics for logistic regression coefficients and random forest permutation feature importance. Lastly, real-world data are used to demonstrate how they may be helpful in a practical problem.

The remainder of this paper is organized as follows: Sect. 2 provides a background of the theory related to this work. Section 3 introduces and explains the proposed metrics. Section 4 presents the experiments, results, and interpretation, and lastly, Sect. 5 summarizes the main findings, future work, limitations, and conclusions.

# 2 Background

There are several goals for explaining prediction models. In this paper, the main goal of this paper is to support applied research providing single metrics that, in a more realistic scenario, are able to report the overall contribution of model features. Hence, this section defines feature importance, and reports the main existing metrics and methods for this goal, which are directly related to this work.

# 2.1 Feature Importance

The most common explanation for the classification model is feature importance. Also known as the feature-level interpretation or saliency method, the method is the most well-studied explainability technique. It explains the decision of an algorithm by assigning values that reflect the importance of input components in their contribution to the decisions. Regardless of the mechanism to calculate it, its common meaning is related to the individual contribution of the corresponding feature for a particular classifier [8].

This individual contribution may be derived from the global perspective, where the feature importance is related to the whole model, and from the local perspective, where the importance is derived for a specific data point [18]. Moreover, it may be internal to the model (intrinsically) as the coefficients of linear models, or by applying methods

that analyze the model after training (post hoc). Another criterion to classify these methods is related to their generalizability, whether they are model-specific or model-agnostic. While every intrinsic method is specific, all model-agnostic work is in a post hoc framework [15].

### 2.2 Marginal Effects

Marginal effects are a general concept and have different meanings depending on the discipline. This Subsection defines it according to econometrics, as the "additional" effect. On the other hand, in Subsect. 2.3, the word "marginal" is related to the probability distribution of a feature as well.

Given a features set  $X = (x_1, x_2)$ , and the predicted function  $\hat{y} = \hat{f}(X)$ , the marginal effects (ME) with respect to  $x_1$  at a specific value corresponds to the changes in the outcome  $\hat{y}$ , given that  $x_1$  changes in one unit. In other words, it is the first derivative of  $\hat{f}(X)$  with respect to a  $x_1$  at a specified value of the input space. However, if  $x_1$  is discrete or binary, the computation is more straightforward, and the finite difference is applied, rather than the derivative [4].

For linear and additive models, MEs are constant across the feature values and are exactly the same as the regression estimated slopes (coefficients) [14]. However, for GLM models MEs take different values across feature distribution. For a logistic regression model, MEs reflects the logit shape, and are small when the probability is close to 0 or 1 and relatively large when close to 0.5 [18]. Thus, to summarize the MEs of  $x_1$  the average of all MEs (AMEs) is commonly used [14]. Moreover, there are other alternatives that may be more suitable depending on the researcher's questions.

#### **Summary Metrics of ME**

Average Marginal Effects (AMEs) are how much the outcome y changes on average when  $x_1$  varies in small changes. Thus, the derivative is computed for every small change on  $x_1$  for every data point and averaged. In practice, the numerical derivative is implemented across the observed values of  $x_1$ . A step (*h*) is defined for continuous features, and the MEs become close to their theoretical value on the limit, as *h* tends to 0. The Equations below demonstrate this beyond the simplified computation of AMEs.

$$ME = \hat{f}(x) = \lim_{h \to 0} \left( \frac{f(x+h) - f(x)}{h} \right)$$
(1)  
$$AMEs = \frac{1}{n} \sum_{i=1}^{n} ME_i$$

*Marginal Effect at the Mean* (MEM) is simply the computation of the MEs around the mean of the feature distribution. In practice, MEM is close to the AME if  $\hat{f}(x)$  is not too noisy and more feasible to compute, since evaluating the derivatives at the means is easier than taking the mean of each derivative [19].

*Marginal Effect at the Representative Value* (MER) is a simplification of MEM calculation for a value that could be an interesting operation point for the research domain. The marginal effect is calculated for each variable at a particular combination of

*X* values. Thus, MER provides a means to understand and communicate model estimates at theoretically important combinations of feature values [20].

These metrics were essential to shed light on those proposed in this paper since they are based on solid statistical theory [14]. In addition to XAI advances, it is possible to report a less-biased feature effect which could play an important role for ML applied researchers.

## 2.3 Marginal Local Effects

The concept of local effects was brought in [21] and is a fundamental part of their accumulated local effects (ALEs) plots. ALE plots were presented as an alternative for visualizing the effects of features in black-box supervised learning models instead of partial dependence plots (PDPs).

The PDPs, introduced by Friedman [22], are widely used to visualize the influence of features in supervised ML models, and have even been considered a causal interpreter for black-box models [23]. For a prediction function  $\hat{y} = \hat{f}(X)$ , where  $\hat{y}$  is a scalar response variable and  $X = (x_1, x_2)$ , PDPs illustrate the relationship between  $x_1$  and the outcome, marginalizing  $\hat{f}(X)$  over the distribution of  $x_2$ . Hence, the PDPs function at a particular value of  $x_1$  represent the average prediction from  $\hat{f}(X)$  if all data points take that value for  $x_1$ . This process takes into account unlikely combinations of X, building unrealistic plots when data are dependent.

As with MEs, ALEs use the conditional instead of marginal distribution. Thus, to overcome the intrinsic problems of MEs, as mentioned in Subsect. 2.2, ALEs use the averaged differences in  $\hat{f}(X)$  across intervals of the training data (local effect). This hack allows the extraction of isolated effects of features within the intervals. Lastly, it accumulates this averaged local effect and center subtracting the mean using the equations below.

$$\widehat{\tilde{f}_{J,ALE}}(x) = \sum_{k=1}^{k(x_1)} \frac{1}{n(k)} \sum_{i:x_j^{(i)} \in N(k)} \left[ f\left(z_{k,j}, x_1^{(i)}\right) - f\left(z_{k-1,j}, x_1^{(i)}\right) \right]$$
(2)

$$\widehat{f_{J,ALE\ Centered}}(x) = \widehat{\widetilde{f}_{J,ALE}}(x) - \frac{1}{n} \sum_{i=1}^{n} \widehat{\widetilde{f}_{J,ALE}}(x_j^{(i)})$$
(3)

where k is the interval of data and  $n_j(k)$  is the neighborhood. Hence, the ALE method calculates the differences in predictions, whereby the features of interest are replaced by grid values of z. The difference in prediction is the effect that features have for an individual instance in a specific interval. The sum on the right in (2) adds up the effects of all instances, which is divided by the number of instances in the interval k to produce the average. Finally, the ALE is vertically centered (3) in the sense that the mean ALEs of  $x_j$  and  $x_l$  on  $\tilde{f}_{j,lALE}(x_j, x_l)$  are both zero. Figure 1 presents a better insight into the computation of the local effect for the

Figure 1 presents a better insight into the computation of the local effect for the function  $\hat{f}(x_1, x_2)$ . The feature range of  $x_1$  was subdivided into k bins with roughly the same number of points indexed by N(k). Focusing on bin N(4), for each point falling into this bin,  $\hat{f}(x_1, x_2)$  have their  $x_1$  held by the left and right endpoints of the interval  $z_3$  and  $z_4$ . Next, the differences of the predictions of these points were averaged by dividing their sum by the number of points in N(4). The same was carried out for all intervals and summed up. Finally, the expectation over  $p(x_1)$  was subtracted.



Fig. 1. Illustration of the ALE estimation. Excerpt from [21]

#### 2.4 Summary of Metrics for Explainability

Marginal effects have been discussed in traditional statistical literature as an alternative for the coefficients of the GLM models due to their direct interpretability on the probability scale. In addition, they are less sensitive to the differences of data scale and unobserved heterogeneity than coefficients [10]. Summary metrics of MEs use the conditional distribution and are not robust against correlated data.

In the field of explainable machine learning field, PDPs have been widely used to report the feature effect. However, they extrapolated their results when using the density function and took into account an unreliable combination of data in their computation. ALE plots recently brought the new concept of local MEs and overcame both the aforementioned problems. Thus, this paper claims that the ALEs theory [21] is a good starting point together with the existing traditional metrics of MEs [14] to propose new global features importance which is able to fill the gap of robustness regarding the size and direction of feature effects on a binary classification model.

## **3** Proposed Metrics

This section proposes four new metrics that are suitable for applied research when comparing the feature effects across multiple models. These metrics are based on the ME and the shape of ALE plots that may be used to report the feature contribution on binary classification problems. Three of them possess explainability in terms of size and direction of the feature effect, while one accounts only for the contribution amount.

#### 3.1 Average Uncentered ALE (AUA)

Average is a natural metric to summarize a distribution (first-order momentum) and leverages great insights regarding the size and directions of the feature mean effects.

$$AUA_{x} = \sum_{k=1}^{k(x^{i})} \left( \widehat{f_{k,ALE}}(x^{i}) \right)$$
<sup>(4)</sup>

This is close to the MEs for linear models and AMEs for GLMs since it uses the average conditional distribution of the observed training data. However, it accounts for each local effect and may be somewhat different for noisy models.

## 3.2 Uncentered ALE at a Specific Value (UAS)

The UAS is an arbitrary choice of a specific value to calculate the uncentered ALE. In practice, it requires splitting the data into k intervals and calculating the uncentered ALE with (3) up to the interval where the choice values fall. This metric may be helpful when the analyst has sufficient domain knowledge or is verifying hypotheses.

$$UAS_{S,x} = \sum_{k=1}^{K_S(x^i)} \left( \widehat{f_{k,ALE}}(x^i) \right)$$
(5)

## 3.3 Maximum Uncentered ALE (MUA)

This metric is more actionable and consists of extracting the maximum change in predicting the outcome for the feature  $x_i$ . As the maximum may be positive or negative related to the outcome, it requires a previous absolute comparison in order to achieve the highest value.

$$MUA_{x} = MAX\left(\left|\widehat{\tilde{f}_{J,ALE}}(x)\right|\right)$$
(6)

## 3.4 ALE Absolute Average (AABSA)

This metric is a non-directional metric and highlights the overall feature importance. It measures, on average, how far the prediction changes away from the ALE average. Unlike the others, the centered ALE is considered, which has a mean zero.

$$AABSA_{x} = \frac{1}{k} \sum_{k=1}^{k(x^{i})} \left( \left| \widehat{f_{k,ALE}}(x^{i}) \right| - E(x) \right)$$

$$\tag{7}$$

## 3.5 Summary of the Novelty of the Proposal

In order to clarify the novelty and issues addressed by the proposal, Table 1 summarizes the main differences of each proposed metric and of those that already exist. In particular, the  $\beta$ - coefficients of logistic regression (LR) and permutation feature importance (PFI) from tree-based algorithms were considered, both widely used in the machine learning field as a global metrics and reported by a single parameter.

# 4 Experiments

To introduce the new metrics, two experiments were conducted<sup>1</sup>. The goal of the experiments was two-fold. First, to compare the proposed metrics in this paper with intrinsic model metrics. More specifically, it considered the widespread  $\beta$ -coefficients and the permutation feature importance derived from LR and random forest (RF), respectively.

<sup>&</sup>lt;sup>1</sup> The implementation code can be found in this repository: https://github.com/rogerioluizsi/sum mary\_ale.git.

Metric	Report		Concern about		
	Size	Direction	Dependence data	Report in probability scale	Model-agnostic
AUA	v	v	v	v	ν
UAS	v	v	v	v	v
MUA	v	v	v	v	v
AABSA	v		v	v	v
$\beta - LR$	v	v			
PFI	v				

Table 1. Differences among proposed methods and similar existing methods in the ML field

Thus, it was possible to evaluate whether or not the metrics highlight the features in a similar manner. An open-access breast cancer dataset was used. This dataset is available ready for modeling binary classification and is a linearly separable problem with highly correlated data [24].

Second, a dataset was used from the National Brazilian Test for Secondary Education (ENEM), and the School Census from the 2009–2019 period. The goal here was to demonstrate how the proposed metrics may be helpful in a real-world problem. Thus, a data mining solution was developed to explore which and how the most important variables associated with school performance behave over the years. The report aimed at yielding valuable and actionable results for decision-makers through a single feature importance metric. These summary measures may enhance the analytical ability of the researcher when comparing the feature effects across supervised models, whichever the algorithm chooses to fit the data.

The ALIBI package [25], which has implemented the ALE plots, supports the computation of the metrics proposed in this paper. Therefore, all the mechanisms intrinsic to the ALEs theory, such as the interval definition, numerical derivation, and the computation of the local effects, follow the software implementation. In this paper, the performance of the models was not reported since the goals were limited to analyzing the model explicability.

#### 4.1 Breast Cancer Data

The breast cancer data included benign and malignant cell samples from 369 patients, 212 with cancer, and 157 with fibrocystic breast masses. Each sample contained thirty features, and LR and RF predicted the patient class in a 5-fold cross-validation setting with random sampling stratified by the target class. Therefore, both algorithms were applied for the same folds, and the mean was adopted as feature importance for each metric.

Figure 2 shows the LR coefficients in red and the four proposed metrics. For the metrics that illustrate the direction of the relationship, there is a high correlation and close magnitude. However, there were some discrepancies. This could have been due to highly paired correlated data (*e.g.* perimeter *vs.* radius), and so logistic regression

arbitrarily chose one of these (*e.g.* radius) to highlight the coefficient [26]. Also, the *p*-values were not checked in this experiment, and maybe some highlighted coefficients were statistically insignificant. However, neither of these are of concern for our metrics.



Fig. 2.  $\beta$ -coefficients and the proposed metrics of LR model for the breast cancer data

Figure 3 illustrates the permutation feature importance from the RF and the proposed metrics. The features set highlighted by all metrics is similar with a high correlation. In addition, surprisingly, the AABSA is fairly close to the permutation feature importance. Both are only positive and only account for the amount of feature contributions. However, they are built differently. While AABSA metrics the average change from the expected

ALEs mean (ALE 0), permutation is related to increasing the prediction error after permuting a feature. Hence, it may be due to the simplicity of the classification task on the cancer data [24].



Fig. 3. Feature importance and the proposed metrics of RF model for the breast cancer data

#### 4.2 Brazilian Secondary Educational Data

The second part of the experiment demonstrates how single feature importance may be helpful in practice. The data was taken from the 2009–2019 period of the largest test for secondary education in Brazil. The dataset contained demographic and socioeconomic information on students, and school characteristics over 32 features. The data included more than ten million students and was preprocessed to the school grain. The school was classified as good or bad in relation to the average scores achieved by their students in the test. To highlight the model-agnostic framework, two tree-based classifiers (RF and AdaBoost (AB)) were applied combined in a 10-fold cross-validation setting, and the overall mean was adopted for each metric.

Thus, systematized temporal data mining evaluated how the main features related to the performance of the school had behaved over the years. Due to space limitations, only the Max Uncentered ALE - MUA was reported by three plots presenting the outputs, as discussed below.

Figure 4 presents a box plot with the MUA for each feature. The clarity of colored dots illustrates the evolutionary directions of the variable over the years. The income per



Max Uncentered ALE - MUA

Fig. 4. Feature importance (MUA) during the period (RF and AB means) (Color figure online)

capita of students is the feature with the highest importance during the whole period. This is followed by parent's education and the students' race (brown students' negative effects). The importance of student computers seems to be a growing tendency over the years.

Figure 5 separates a selected set of features to obtain a better understanding of their behavior during the period. The computer lab has a higher positive slope, while faculty jobs (the number of schools where teachers work) have a higher negative.

Lastly, in Fig. 6, temporality was disregarded, and the features were organized for an overview of their importance in the following groups: non-actionable features (race and gender), school features (infrastructure), student features (parent's education and income) and teacher features. In general, the group of features related to students demonstrated more potential to improve the quality of schools than others. Additionally, non-actionable features had a strong influence, both negative and positive.

# 5 Discussion and Conclusion

This paper has proposed new model-agnostic metrics of feature importance in an attempt to circumvent the drawbacks and constraints of the existing methods, such as the  $\beta$ - coefficients of additive models and feature importance from tree-based algorithms, widely



Fig. 5. Evolution of the feature importance (MUA) of a selected features set (RF and AB means)



Fig. 6. Feature importance (MUA) by groups during the period (RF and AB means)

used for this purpose. This paper has proposed other options with a number of advantages, such as being agnostic to models, interpretable in the probability scale, and more reliable under correlated data.

The accumulated local effects are the key trick for isolating the main effect of the variable even in correlated data. The four proposed metrics were validated in two experiments that illustrated their suitability to actual data mining applications.

In the first experiment, breast cancer data were used to compare the proposed metrics with the coefficients of logistic regression and the permutation feature importance of random forest. The results illustrated that the proposed metrics are robust when facing correlated data and did not suffer the effects experienced by logistic regression (LR). All the proposed metrics captured the desired aspects of the attributes and were highly correlated. The AABSA proposed metric, which is directly comparable to the random

forest (RF), since the permutation feature importance is only positive, captured very similar attribute importance.

Nevertheless, the comparisons were limited, and more tests are required to evaluate the metric behavior in other situations. For example, the LR coefficients are known to be sensitive to unobserved heterogeneity. Hence, when features are added to the model and improve the predictions, the remaining coefficients may change, even if this feature is not correlated with others [13]. Despite the marginal effects (a key to the proposed metrics) being more robust in this situation [26], there was no empirical evidence of this in our context. Additionally, an empirical test of the robustness in a scenario of correlated data compared to existing metrics is required since this paper only considered the theory inherited from ALE plots. Thus, we intend to make more analyses on a large scale in future work, including other XAI approaches.

It should be mentioned that this paper has not yet compared the proposed metrics across algorithms, since the algorithms are able to use the input features in a totally different manner to achieve similar results, it was already known as the "Rashomon" effect [27]. Thus, even though the metrics proposed here are model-agnostic, the comparisons across different algorithms must be interpreted with caution, even on the same data.

In addition, despite the motivation to compare feature importance across models, the results must be interpreted with care, and validation by domain experts is required. For example, in the second experiment, the data set was scaled equally for each year, and the set of variables was the same with the same values. Even after these careful transformations, the comparison may be inappropriate, and a piece of domain knowledge may help to reduce the risks of misinterpretation.

The main limitation of the proposed metrics would be the extrapolation of the local effect out of the interval where it was computed. The local effect is averaged across the conditional distribution and may only hold when the predictors X jointly fall within the same bin. Thus, there may be a problem if bin widths are too small and the predictive function is too noisy. Furthermore, local effects may be unreliable if the quantity of data points into the underlying bin is very small on the training data. Although this paper has used deciles to equally subdivide the feature distribution in order to minimize this risk, together with cross-validation for more reliable results, caution must be always taken in the interpretation with the validation of the domain expert.

# References

- Bhatt, U., Xiang, A., Sharma, S., et al.: Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 648– 657. ACM, New York (2020)
- Razavian, N., Blecker, S., Schmidt, A.M., et al.: Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data 3, 277–287 (2015). https://doi.org/10. 1089/big.2015.0020
- Pellagatti, M., Masci, C., Ieva, F., Paganoni, A.M.: Generalized mixed-effects random forest: a flexible approach to predict university student dropout. Stat. Anal. Data Min., 1–17 (2021). https://doi.org/10.1002/sam.11505
- Berens, J., Schneider, K., Görtz, S., et al.: Early Detection of Students at Risk-Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods (2019)
- Yang, K.C., Varol, O., Davis, C.A., et al.: Arming the public with artificial intelligence to counter social bots. Hum. Behav. Emerg. Technol. 1, 48–61 (2019). https://doi.org/10.1002/ hbe2.115

- Leite, M.A.G.L., Guelpeli, M.V.C., Santos, C.Q.: Um Modelo Baseado em Regras para a Detecção de bots no Twitter, pp. 37–48 (2020). https://doi.org/10.5753/brasnam.2020.11161
- Barredo Arrieta, A., Díaz-Rodríguez, N., del Ser, J., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
- Saarela, M., Jauhiainen, S.: Comparison of feature importance measures as explanations for classification models. SN Appl. Sci. 3(2), 1–12 (2021). https://doi.org/10.1007/s42452-021-04148-9
- Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: 34th International Conference on Machine Learning, ICML 2017, vol. 7, pp. 4844–4866 (2017)
- Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 41(3), 647–665 (2013). https://doi.org/10.1007/s10 115-013-0679-x
- Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-August, pp. 1135–1144 (2016). https://doi.org/10.1145/ 2939672.2939778
- Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777. Curran Associates Inc., Red Hook (2017)
- 13. Mood, C.: Logistic regression: uncovering unobserved heterogeneity, pp. 1-25 (2017)
- Long, J.S., Long, J.S.: Regression Models for Categorical and Limited Dependent Variables. Sage, New York (1997)
- 15. Molnar, C.: Interpretable Machine Learning (2019)
- Bhatt, U., Ravikumar, P., Moura, J.M.F.: Towards aggregating weighted feature attributions (2019)
- Hooker, G., Mentch, L.: Please stop permuting features: an explanation and alternatives, pp. 1–15 (2019)
- Guidotti, R., Monreale, A., Ruggieri, S., et al.: A survey of methods for explaining black box models. ACM Comput. Surv. 51 (2018). https://doi.org/10.1145/3236009
- Bartus, T.: Estimation of marginal effects using margeff. Stata J. 5, 309–329 (2005). https:// doi.org/10.1177/1536867x0500500303
- 20. Leeper, T.J.: Interpreting Regression Results using Average Marginal Effects with R's margins (2021). https://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf32
- Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. J. R. Stat. Soc. Ser. B Stat. Methodol. 82, 1059–1086 (2020). https://doi. org/10.1111/rssb.12377
- Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451
- 23. Zhao, Q., Hastie, T.: Causal interpretations of black-box models. J. Bus. Econ. Stat. **39**, 272–281 (2021). https://doi.org/10.1080/07350015.2019.1624293
- 24. Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences (2017). http://archive.ics.uci.edu/ml
- Klaise, J., Van Looveren, A., Vacanti, G., Coca, A.: Alibi explain: Algorithms for explaining machine learning models. J. Mach. Learn. Res. 22(181), 1–7 (2021). http://jmlr.org/papers/ v22/21-0017.html
- Mood, C.: Logistic regression: why we cannot do what we think we can do, and what we can do about it. Eur. Sociol. Rev. 26, 67–82 (2010). https://doi.org/10.1093/esr/jcp006
- Fisher, A., Rudin, C., Dominici, F.: 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, 1–81 (2019)