Method Agnostic Model Class Reliance (MAMCR) Explanation of Multiple Machine Learning Models

Abirami Gunasekaran[1], Minsi Chen[1(✉)], Richard Hill[1], and Keith McCabe[2]

[1] School of Computing and Engineering, University of Huddersfield, Huddersfield, UK
M.Chen@hud.ac.uk

[2] Planning and Business Intelligence, University of Huddersfield, Huddersfield, UK

Abstract. Various Explainable Artificial Intelligence (XAI) methods provide insight into the machine learning models by quantitatively analysing the contribution of each variable to the model's predictions globally or locally. The contribution of variables identified as (un)important by one method's explanation may not be identified as the same by another method's explanation for the same machine learning (ML) model. Similarly, the important feature of many well performing ML models that fit equally well on the same data (which are termed as Rashomon set models) may not be the same as each other. While this is the case, providing the explanation based on a single model in the lens of a specific explanation method would be biased over the model/method. Hence, a framework is proposed to describe the consensus variable importance across multiple explanation methods for many almost-equally-accurate models as a method agnostic explanation for the model class reliance. Empirical experiments are carried out on the COMPAS dataset with six XAI (the Sage, Lofo, Shap, Skater, Dalex and iAdditive) methods for verifying whether an inadmissible feature becoming an (un)important feature is consistent across multiple explanation methods and getting the consensus explanation. The results demonstrate the efficiency of the method agnostic model class reliance explanation and its coverage to the model reliance range of all the almost-equally-accurate models of the model class.

Keywords: XAI · Ensembled explanation · Feature importance · Rashomon set

1 Introduction

The recent strategies under the XAI umbrella are mostly model agnostic. It means that irrespective of the ML model type and the internal structure, the explanation methods provide the explanation for the model's decisions. One such technique is the feature importance method [1]. These methods [2–8] can be plugged into any ML model to know the learning behaviour of the model in terms of feature importance. Here, the learning behaviour represents the order of important features on which the model takes its prediction. These model-agnostic methods require only the input and the predicted output of the model for providing the feature importance explanation.

K. K. Patel et al. (Eds.): icSoftComp 2022, CCIS 1788, pp. 56–71, 2023.
https://doi.org/10.1007/978-3-031-27609-5_5

The feature importance can be defined as a quantitative indicator that quantifies how much a model's output changes with respect to the permutation of one or a set of input variables [9]. The computation of these variable importance values is operationalized in different ways. The importance of the variables can be quantified by introducing them one by one, called feature inclusion [8] or by removing them one by one from the whole set of features, called feature removal [2]. The model can be retrained several times [11] for each of the input feature inclusions/removals or multiple retraining can be avoided [12] by handling the absence of removed features or the inclusion of new features. For that, any supplementary baseline input [13], conditional expectations [14], the product of marginal expectations [15], approximation with marginal expectations [3] or replacement with default values [2] can be used.

Though all these methods explain the feature importance behind the decisions of the model, the explanation obtained from a method may not be similar to the explanation of another method for the same model [17, 34]. This would confuse the analyst as which explanation should be trusted when different explanations are obtained. Unfortunately, there is no clear, standard principle to choosing the appropriate explanation method.

There may be many but different ML models that can fit equally well and produce almost similar accurate predictions on the same data. But the feature which is most important to one such model may not be an important feature for another well performing model [19].

In such a scenario, providing the explanation based on a single ML model using a specific explanation method would be biased (unfair) over the model/method. To this end, a novel explanation method is proposed to provide a method agnostic explanation across various method explanations of multiple almost-equally-accurate models. These near-optimal models [29] are termed as Rashomon set [19]. Instead of selecting a single predictive model from a set of well performing models and providing the explanation for it, the proposed method offers an explanation across multiple methods to cover the feature importance of all the well performing models in the model class.

The rest of the work is structured as follows: Sect. 2 reviews the related works, Sect. 3 deals with the proposed method, Sect. 4 speaks about the experiments, results and discussion, and Sect. 5 presents the conclusion.

2 Related Works

A plethora of strategies under XAI is developed for providing explanations for the black-box models. Among them, the major attention is being received by the feature importance methods. These methods [3–8, 11] aim to explain a single model's variable importance values by permutating the variables. The methods can give explanations as local feature importance [2] for a single instance or as global importance [4] for the entire data set.

Rashomon Effects: Initially, the problem of model multiplicity where multiple models fit on the data are equally good but different models was raised by [10]. There is no clear reason to choose the 'best' model among all those almost-equally-accurate models [22]. Moreover, the learning behaviour of the models varies among themselves. It means that the feature that is important for one model may not be important for another

model. Hence, to avoid a biased explanation of a single model, the comprehensive explanation for all the well performing models (Rashomon set models) is given as a range of explanation by [19].

In line with [19], the authors of [22] expanded the Rashomon set concept by defining the cloud of the variable importance (VIC) values for the almost-equally-accurate models and visualizing it with the variable importance diagram (VID). The VID informs that the importance of a variable gets changed when another variable becomes important.

Aggregating over a set of competing equally good models would reduce the non-uniqueness [10]. Based on this concept, the authors of [29] generated a set of 350 near-optimal logistic regression models on the COMPAS dataset, aggregated the models' feature importance values and presented the explanation a less biased importance explanation for the model class than a single model's biased explanation. Similarly, by ensembling the Rashomon set models using prior domain knowledge, the authors [30] correct the biased learning of a model. If the Rashomon set is large, the models contained in the set could exhibit various desirable properties [31]. Also, the authors observe that the model performance does not necessarily vary across different algorithms even though the ratio of Rashomon set models on the dataset is small.

All these works aim on solving the bias that arises from multiple models (Rashomon set) rather than considering the bias that comes for a model from multiple methods.

Explanation Evaluation and Ensembling: The common evaluating measures found in the literature for ensembling explainable approaches are stability [32, 34, 37], (in)fidelity [18, 37], consistency [32, 35], informativeness [33] and comparison metric [36]. Though the explanation methods provide varying explanations for the same model, no principled way could be found in the literature to get a consensus explanation across various methods. A framework [32] proposes the ensembled explanation of several model agnostic algorithms based on the consistency and stability scores with the aim to provide an ensembled explanation independent of the XAI algorithms. Similarly, a unifying framework for understanding the feature removal-based explanation methods is introduced in [7]. The authors showed the relationship of how the methods are related to one another in providing the explanation. It does not combine the explanations of the various methods into one explanation but offers comparable explanations of those methods. At the same time, by comparing the various method explanations for a model, the most representative knowledge of the data set is obtained through the common explanations from the various methods [34]. All the ensembling explanation works focus on the multiple explanations for a single model rather than model multiplicity.

A unified explanation across multiple methods has not been extensively studied and the research works related to the Rashomon set focus on the explanations that vary across the multiple models rather than across multiple methods. Hence, a framework is proposed to address the explanation bias happening across multiple methods for the multiple almost-equally-accurate models. The work is motivated to find the answer to the following research questions:

RQ1. while various explanation methods are applied on multiple well performing models to get the feature importance explanations, will the feature which is projected as (un)important by one explanation method be agreed by other methods?

RQ2. Is getting a consensus explanation that is consistent across the various applied methods for multiple almost-equally-accurate models possible?

3 Proposed Method

This section presents the proposed method for obtaining the method agnostic ensembled explanation of various almost-equally-accurate ML models. The processes involved in obtaining the model agnostic model class reliance range using the MAMCR framework are depicted in Fig. 1.

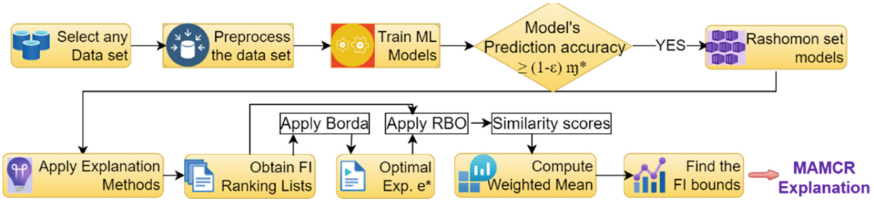

Fig. 1. The Process pipeline of Model Agnostic Model Class Reliance (MAMCR) framework

3.1 Models Building

Let $(X, Y) \in \mathbb{R}^{p+1}$, where $p > 0$, $X \in \mathbb{R}^p$ is the random vector of p input variables and $Y \in \mathbb{R}^1$ is the output variable. The process pipeline starts with the modelling of a class of multiple ML models on the pre-processed data of tabular type. As per the No Free Lunch theorem [21], there is no single ML model that is considered as best for solving the problems. Consequently, multiple ML models can be fitted on the same data set to verify the model's performance. This set of prespecified predictive models is referred to as model class [19].

$$Model\ class\ M = (\mathfrak{m}_t, \{t = 1, 2, \ldots, m\}) \tag{1}$$

where, M is a model class that consists of m models. Each model can take the input X and convert it to response Y. Each model's performance is assessed in terms of its prediction accuracy. The model class can be built with a set of regression algorithms. In that case, the model performance is assessed in terms of R^2 value.

3.2 Finding the Rashomon Set Models

From the multiple fitted models of the model class M, the almost-equally-accurate models form the Rashomon set (Ŗ). A Rashomon set is constructed based on a benchmark model $\mathfrak{m}^*$and a nonnegative factor ε as follows:

$$Ŗ(\varepsilon, \mathfrak{m}^*, M) = \{\, \mathfrak{m} \in M \mid \mathfrak{m}(X) \geq (1-\varepsilon)\, \mathfrak{m}^*(X) \,\} \tag{2}$$

Selection of $\mathfrak{m}^*$ with possible maximum accuracy and ε with a small positive value helps to search for the models whose prediction accuracy are not less than the $(1-\varepsilon)$ factor of $\mathfrak{m}^*$ accuracy and to construct the Ŗ models i.e., $Ŗ(M)$.

3.3 Obtaining Model Reliance Values and Ranking Lists

The model reliance [19] (or feature importance) indicates how much a model relies on a variable for making its predictions. The model reliance on the variable k (mr^k) is measured by the quantity of change in the model's performance with and without the variable k, where k = 1, 2, …, p. The more the change in the model performance, the higher the importance of that variable in the model's prediction contribution.

Different state-of-the-art explanation methods are selected to apply to each Rashomon set model to obtain their model reliance on p variables. Any global explanation method that returns the explanation in the form of feature importance can be chosen.

$$Explanations\ E = \left\{\mathfrak{E}_i\left(\mathfrak{m}_j\right) \middle| i = 1\ to\ n\ and\ j = 1\ to\ r\right\} \quad (3)$$
$$where\ \boldsymbol{n} = no.of\ explanation\ methods\ \{\mathfrak{E}_1, \mathfrak{E}_2, .., \mathfrak{E}_n\},\ \boldsymbol{r} = no.of\ models\ \text{in}\ \mathbb{R}(\mathfrak{M})$$

The obtained model reliance explanations E can be mapped to a model reliance vector as follows:

$$MRV_n(\mathfrak{m}) = (mr_n^1(\mathfrak{m}), (mr_n^2(\mathfrak{m}), \ldots, (mr_n^p(\mathfrak{m})) \quad (4)$$

where $mr_n^p(\mathfrak{m})$ represents the model reliance of the model $\mathfrak{m}$ on variable p that is obtained from the explanation method n. The model reliance vector values are also mapped to model reliance ranking lists as follows:

$$\begin{aligned} E_{MRR} &= \{[MRR_1, MRR_2, \ldots, MRR_r]\} \\ &= \{[e_1(\mathfrak{m}_1), e_2(\mathfrak{m}_1), \ldots, e_n(\mathfrak{m}_1)], [e_1(\mathfrak{m}_2), e_2(\mathfrak{m}_2), \ldots, e_n(\mathfrak{m}_2)], \ldots, \\ &\qquad [e_1(\mathfrak{m}_r), e_2(\mathfrak{m}_r), \ldots, e_n(\mathfrak{m}_r)]\} \end{aligned} \quad (5)$$

The explanation $E_{MRR}[1] = MRR_1 = [e_1(\mathfrak{m}_1), e_2(\mathfrak{m}_1), \ldots, e_n(\mathfrak{m}_1)]$is the set of model reliance ranking lists obtained for the 1^{st} model $(\mathfrak{m}_1)$ from $\boldsymbol{n}$ explanation methods. The $e_n(\mathfrak{m}_1)$ shows the feature ranking list for the model $\mathfrak{m}_1$ obtained from the n^{th} explanation method. For example, the order can be represented as follows,

$$e_n(\mathfrak{m}_1) = [f_1, f_3, f_4, f_p, \ldots, f_2]$$

where f_1 is the name of the input feature that has the highest importance value than all other variables $f_2, f_3, f_4, \ldots, f_p$. The model reliance ranking list follows the order $f_1 > f_3 > f_4 > f_p >, \ldots, > f_2$, where variable f_2 has the least importance among the p variables.

3.4 Finding the Reference Explanation $\mathfrak{E}^*$ and Consistent Explanations

Various methods that operationalize the feature importance computation may not produce the same explanation for a model [34]. The explanations not only differ in the ranking order but also in the computed model reliance values. Despite the variances, no clear reason could be found in the literature for selecting a specific explanation method. As pointed out by [16], if the results of different techniques point to the same conclusion, they very likely reflect the real aspects of the underlying data. Therefore, a reference

explanation reflecting the commonly found feature order among the different methods' explanations of a model should be discovered. This reference explanation captures the optimal feature order by aggregating all the explanations' feature ranking preferences using the modified Borda Count method [23].

$$e_j^* = Borda(E_{MRR}[j]), where\ j = 1\ to\ r\ models \tag{6}$$

The Borda function returns the result as an aggregated model reliance ranking order i.e., e_1^* captures the optimal ranking order of the features from the n explanations of the 1^{st} model. Likewise, for each model, a reference explanation is aggregated from the corresponding model's explanations from n methods. This leads to a totally $\boldsymbol{r}$ number of reference explanations for the Rashomon set models $\mathbb{R}(\mathbb{M})$.

To quantify the consistency of several methods in producing similar explanations to the model, the methods' explanations for the model are compared against the reference explanation. To find the consistency score, a ranking similarity method needs to be applied. The existing statistical method such as Kendall's τ [24] is considered inadequate to this problem because the ranking lists may not be conjoint. On the other hand, the Rank-Biased Overlap (RBO) [28] could handle the ranking lists even though the lists are incomplete. The RBO similarity between two feature ranking order lists R_1 and R_2 is calculated using the following equation as per [28].

$$RBO(R_1, R_2, p) = (1-p)\sum_{d=1}^{\infty} p^{d-1}.A_d$$
$$A_d = \frac{|R_1 \cap R_2|}{d} \tag{7}$$

The RBO similarity value ranges from 0 to 1, where 0 indicates no similarity between the feature ranking order lists and 1 indicates complete similarity. The p parameter ($0 < p < 1$) defines the weight for the top features to be considered. The parameter A_d defines the agreement of overlapping at depth d. The intersection size of the two feature ranking lists at the specified depth d is the overlap of those 2 lists (Refer to Eqs. 1–7 in [28]).

A similarity score is computed between the model's various explanations and the corresponding reference explanation and is referred to as *optimal similarity*. It is calculated as follows,

$$OPTIMAL\_SIM_{i,j} = RBO\left(e_{i,j}, e_j^*\right)$$
$$where\ i = 1\ to\ n\ methods;\ j = 1\ to\ r\ models \tag{8}$$

The $OPTIMAL\_SIM_{i,j}$ defines how much the explanation $(e_{i,j})$ from method i for the model j ($\mathfrak{m}_j$) is similar in complying with the reference explanation e_j^*, in terms of feature order. The *OPTIMAL_SIM* value is computed for all the method explanations of each model. Therefore, $\boldsymbol{n} \times \boldsymbol{r}$ similarity scores are obtained totally, that is, each explanation method gets a consistency score for each model.

3.5 Computing the Weighted Grand Mean (θ)

Among the various explanations of the Rashomon set models, the optimal similarity scores of the methods are calculated based on the method explanations' compliance

with the corresponding model's reference explanation. This score shows the degree of similarity that the method has, in explaining the model's optimal learning behaviour.

Since the different explanation methods produce different feature importance coefficients for each feature, the model has varying levels of reliance on a feature. Therefore, a grand mean (θ) across several methods should be estimated. For that, a weighted mean [38] is to be implemented. To weigh the feature importance values that are computed by each method for a model, the optimal similarity score is used. For each feature, the weighted mean of the feature importance values based on the methods' optimal similarity score as weight is calculated by,

$$\theta_{j,k} = \frac{\sum_{i=1}^{n} OPTIMAL\_SIM_{i,j} * mr_i^k(\mathfrak{m}j)}{\sum_{i=1}^{n} OPTIMAL\_SIM_{i,j}} \quad (9)$$
$$where\ k = 1\ to\ p\ features\ ;\ j = 1\ to\ r\ models$$

The grand mean of the feature k of the model j ($\theta_{j,k}$) is calculated by adding the product of the optimal similarity score of the 1 to n methods with its computed feature importance value for the k feature (mr_1^k *to* mr_n^k) and dividing the result with the sum of n methods' weights (i.e., optimal similarity scores of n methods). The grand mean is computed for all the p features for each Rashomon set model. Therefore, $p \times r$ weighted mean feature importance values are obtained.

3.6 Method Agnostic Model Class Reliance (MAMCR) Explanation

The method agnostic model class reliance explanation of the Rashomon explanation set is given as a comprehensive reliance range for each variable based on the reliance of all the well performing models under *n* explanation methods.

The model class reliance of all the p variables can be given as a range of lower and upper bounds of weighted feature importance values. The lower and upper bounds of the model class reliance for each variable can be defined as,

$$MCR^k = [MCR^{k-}, MCR^{k+}],\ \text{k} \in \text{p variables} \quad (10)$$

$$MCR^{k-} = \min_{\theta}\ {}^{\mathrm{r}}_{1}\mathrm{j}\,\theta_{j,k}\ \ ,\ \ MCR^{k+} = \max_{\theta}\ {}^{\mathrm{r}}_{1}\mathrm{j}\,\theta_{j,k} \quad (11)$$
$$\text{where}\ \ r = |\mathrm{R}(\mathfrak{M})|$$

The range [MCR^{k-}, MCR^{k+}] of variable k represents that if the MCR^{k-} value is low, the variable k is not important for any almost-equally-accurate models in the Rashomon set models [R(𝔐)]whereas if the MCR^{k+} is high, then the variable k is most important for every well performing model in R(𝔐). Thus, the MCR provides a method agnostic variable importance explanation for all the well performing models of the Rashomon set.

4 Experiments and Results

In this section, the concept of the proposed method is illustrated with the experiments on the 2-year criminal recidivism prediction dataset[1] which was released by ProPublica to study the COMPAS (Correction Offender Management Profiling for Alternative Sanctions) model that was used throughout the US Court system. The dataset consists of 7214 defendants (from Broward County of Florida) with 52 features and one outcome variable, which is 2-year recidivism. Among the 52 features, 12 are date type to denote jail-in and out, offence, and arrest dates, 21 are personal data identifiers such as first and last name, age, sex, case numbers and descriptions and other features are mostly numeric values such as no. of days in screening, in jail, from compas, etc. The framework is not limited to this data but is flexible enough to support any dataset.

In the analysis of the Race variable's contribution to predicting the 2-year recidivism, the authors [22] say that there are some well performing models which do not rely on inadmissible features like Race and gender. Additionally, for the same data set, the authors [29] report that the explanation based on a single model is biased over the inadmissible feature 'Race', whereas the grand mean of multiple models' feature importance values does not highlight the feature as an important feature for the majority of the models. To ensure whether these claims will be consistent across multiple methods' explanations and to answer the research questions as well, the same dataset used by [22, 29] with similar a setup (with 6 features - age, race, prior, gender, juvenile crime, and current charge - of all the 7214 defendants) is taken for the analysis.

To make the outcome prediction, the logistic regression model class is used in the analysis with 90% (6393) training data and 10% (721) test data as in [29]. The Stratified 5-fold cross-validation is used to train and validate the multiple models. The total trained models and the selected models to the Rashomon set are shown in Fig. 2. The reasonable sample of Rashomon set models (350) are obtained from the total trained (2665) models by filtering the models whose prediction accuracy are above the accuracy threshold $(1-\varepsilon)\mathfrak{y}^* = 0.6569$, where $\mathfrak{y}^*$ accuracy $= 0.6914$ and $\varepsilon = 5\%$. Those models form the Rashomon set.

To obtain the explanations for models' decisions, the iAdditive[2] and other 5 state-of-the-art XAI methods [3, 4, 7, 25, 26] based on the feature importance approach are applied to the Rashomon set models $[\mathfrak{R}(\mathfrak{y})]$. Normalization is applied to each method's computed importance values for each model. The model reliance rankings for each model are also obtained (E_{MRR}). Figure 3 shows the various methods' model reliance ranking range of the Rashomon set models grouped by each feature of the COMPAS dataset.

The distribution of feature importance ranks that are obtained from different methods illustrates the variation found in the various method explanations. Let's consider the 'Race' feature's rank explanations. The Shap [3], Skater [4] and iAdditive methods' ranks span from 1–6 for the models, whereas for the other 3 methods, the range is from 2–6. It means, as per the former methods' explanations, there are some models which consider the 'Race' feature as their most important (1st rank) feature. But in the view

[1] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

[2] iAdditive is an in-house XAI software tool.

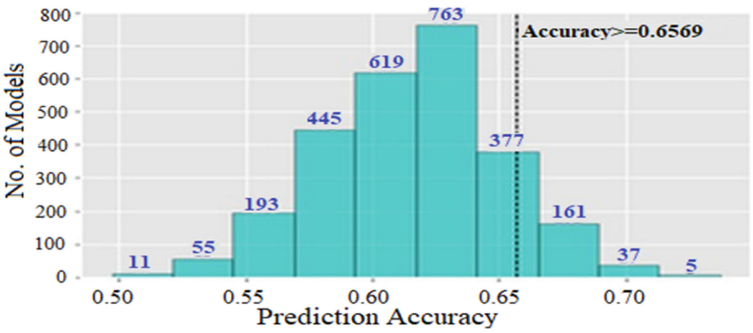

Fig. 2. The prediction accuracy frequency of all the trained models. The accuracy threshold $(1-\varepsilon)\mathfrak{y}^* = 0.6569$, where $\mathfrak{y}^* = 0.6914$ and $\varepsilon = 5\%$ is used to search for the Rashomon set models (Ɍ). Models with an accuracy level above the threshold value are only included in the Ɍ.

of the latter methods, for none of the models, 'Race' is the 1st priority feature. Let's take the 'Juvenile crime' feature. As per the Sage [7] method explanations, the'crime' feature is the most important feature for most of the models, whereas, for the Shap and iAdditive methods, the median ranks lie in 4th and 5th positions, respectively. The Skater and Lofo [26] methods have similar 3rd rank position to the feature and the Dalex [25] method stood in between the Sage and Skater rank positions by giving 2nd rank.

From this, it could be observed that for the same models, these methods provide different feature importance explanations (in the form of computed values and ranks as well). If any one of the methods is selected to provide the explanation for a well performing model, it could end up in a method-dependent explanation of that model. It means that the explanation would be biased over the specific method. Therefore, to get a consensus explanation for the almost-accurate models over all the applied explanation methods, the model agnostic model class reliance (MAMCR) explanation method is to be implemented.

Firstly, a reference explanation e* is aggregated from the corresponding explanations of 6 methods for each model to reflect the common feature ranking order. These reference explanations reflect the optimal learning for all the models in the Rashomon set (see Fig. 4). To quantify the consistency of various explanations obtained from multiple methods, the corresponding reference explanation (e*) is compared against each model's method-wise explanation.

Next, for each model of the Rashomon set, the weighted average is computed for all the features based on the method's consistency score. The method explanation which complies well with the optimal explanation will contribute more to the average model reliance value. For each of the six variables of the 350 models, the grand means ($\theta_{j,k}$) are computed using Eq. 9 based on the concern method's consistency/optimal similarity scores.

The method agnostic model class reliance explanation (MAMCR) for the multiple almost-accurate models based on multiple methods' explanations is presented as a range. The lower and upper bounds [MCR^-, MCR^+] of each variable's grand mean are selected as the model class reliance for all the models in the Rashomon set. The method agnostic

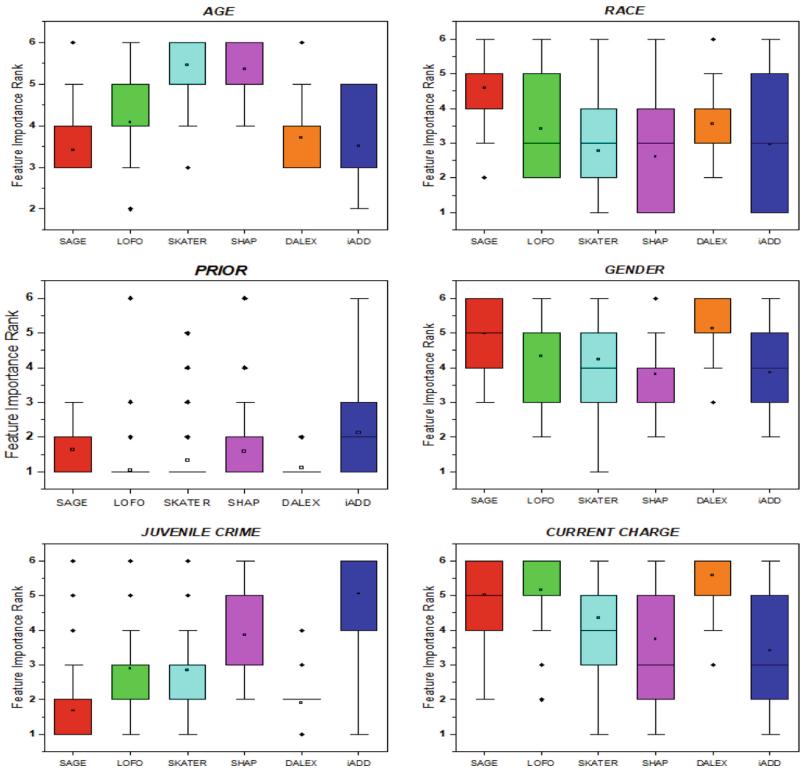

Fig. 3. Model reliance/feature importance rankings obtained from the 6 explanation methods for the COMPAS dataset. A box plot showing the range of ranks allocated by each method for the 350 Rashomon set models for a feature is shown in each panel. The difference in the feature rankings illustrates the variations found in the various method explanations.

MCR is shown in Table 1. In that, the high MCR^{-} value (e.g., 0.08) indicates that the *Prior* feature is used by all the models and the low MCR^{+} value (e.g., 0.10) indicates that the *Age* feature is least used by all the models.

4.1 Discussion

Various methods' explanations are compared against the 'Race' feature's importance. The distribution of many models' model reliance is shown in Fig. 5. The number of models that falls on the feature importance range is displayed on each bar in the histogram. As per the Sage [7] explanation, the 'Race' feature is not at all an important feature used by most of the models. It could be observed from Fig. 5a that 324 models out of 350 are given the feature importance value as less than 0.1. This informs that the Race feature is not an important feature for the 324 models. It complies well with the claim of [29]. On the other hand, it is not true based on other methods' explanations. From Figs. 5b–5e, it could be observed that there are many models that rely on the 'Race' feature from the moderate to high range, whereas Fig. 5f is consistent with Fig. 5a. It alerts us that

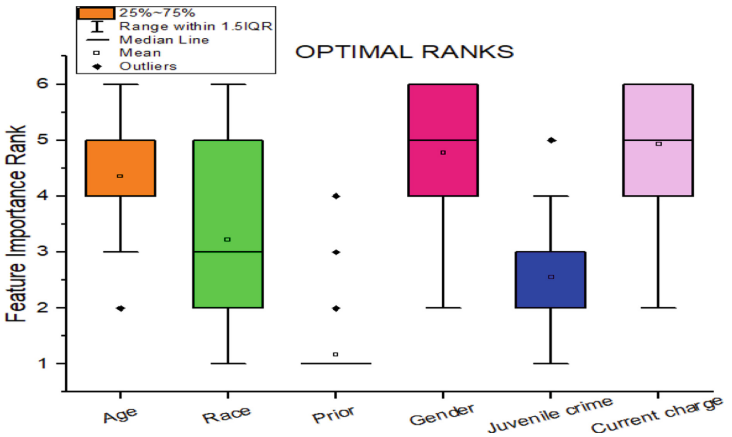

Fig. 4. The feature-wise rank distribution of optimal reference explanations (e*) for 350 Rashomon set models.

Table 1. The method agnostic model class reliance explanation of the Rashomon set models for the six features of the COMPAS dataset.

Features	[MCR$^-$	MCR$^+$]	STD
Age	0.023774	0.103612	0.015621
Race	0.021176	0.33566	0.089296
Prior	0.08947	0.698398	0.090259
Gender	0.017584	0.188301	0.039289
Juvenile crime	0.074106	0.426745	0.054915
Current charge	0.017144	0.236485	0.041713

the explanation obtained from a method is not necessarily the same as the one obtained from another method for the model.

This addresses the first research question (RQ1) that while multiple explanation methods are applied on multiple well-performing models for getting the feature importance explanations, the feature which is projected as (un)important by one explanation method is not necessarily agreed by another method. Therefore, the identified importance of the feature depends completely on the method that is applied for obtaining the explanation.

While comparing the method explanations of each feature (see, Fig. 3), no two methods could be identified in producing a similar explanation pattern in all the feature explanations. For example, the Skater and Shap method explanations for the Age feature resemble the same pattern except for the outlier. Similarly, the Sage and Dalex are in a similar pattern on the same variable. The same methods could not be found with similar patterns in other feature explanations. For example, the Skater and Shap methods have contrasting explanation patterns in Juvenile crime feature, whereas the Skater and Lofo

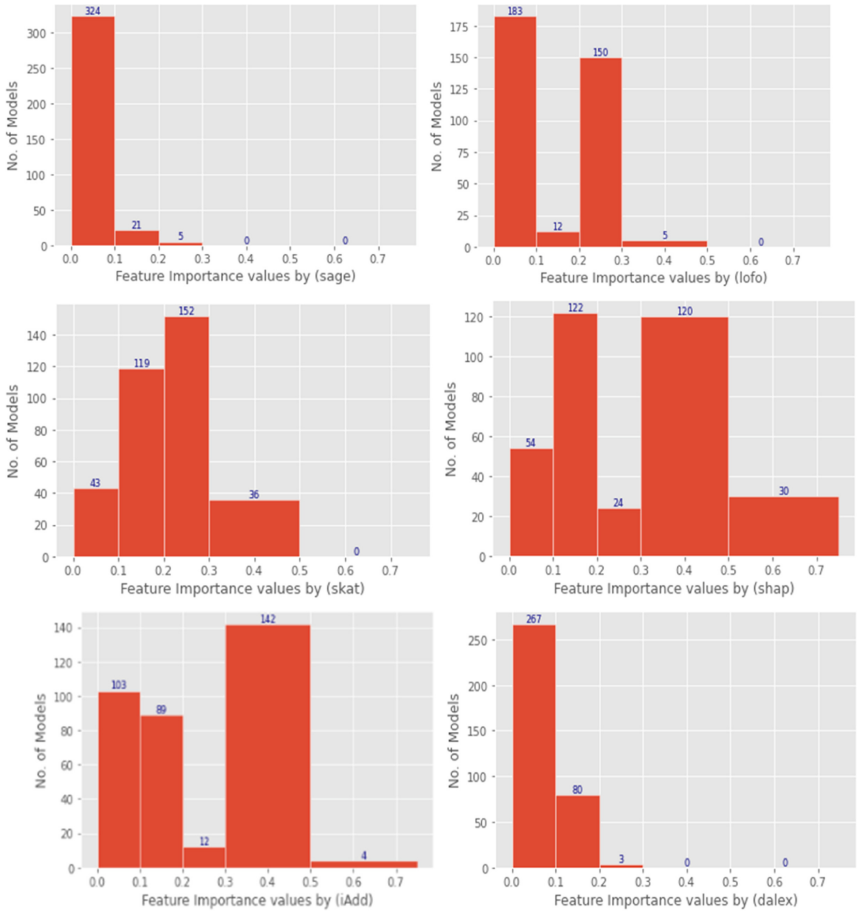

Fig. 5. The feature importance values of the 'Race' feature for 350 Rashomon set models, grouped by each method (5a. Sage, 5b. Lofo, 5c. Skater, 5d. Shap, 5e. iAdditive and 5f. Dalex). The data label of each bar shows how many models lie within the feature importance bin range.

methods exhibit a similar pattern. One of the possible reasons observed for the variation could be that a feature becomes the most important when another variable becomes the least important [22]. It is illustrated in Fig. 6.

Figure 6 shows the feature importance values computed for Juvenile crime and Prior features by the 6 methods for the 350 almost-accurate models. Each point in the plot represents a model's reliance on those variables. When the Prior feature importance (y-axis) of the models reaches its maximum value such as above 0.6, the crime feature importance (x-axis) of them is below ≈0.35 (shown within a box). When the crime feature's importance of a model reaches above 0.8 or around 1, its Prior importance is very low such as less than ≈0.15. It indicates that the feature Prior is the most important feature of a model when the Juvenile crime is less important than the Prior feature. So, if a method allocates a feature with high importance in its explanation obviously another

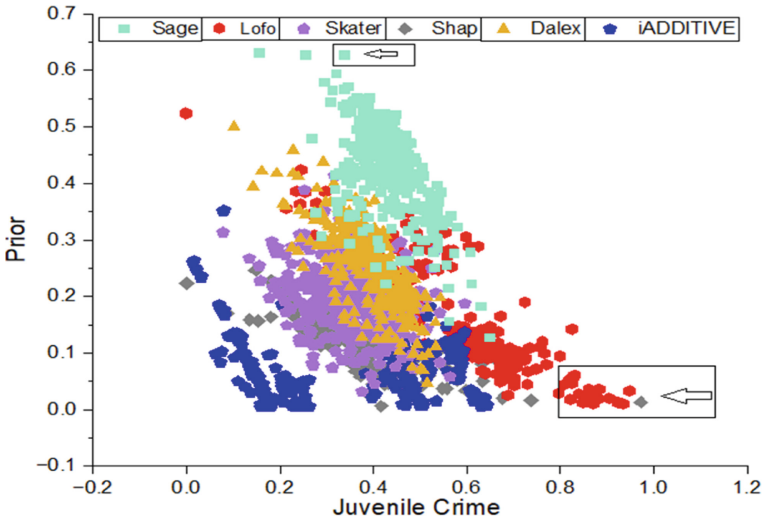

Fig. 6. The feature importance values of Prior and Juvenile crime features computed by 6 methods. While the importance values of the Juvenile crime feature increase, the prior feature importance decreases and vice versa, which is emphasised with a box.

feature gets reduced importance which may make the explanation vary from another method's explanation.

Despite the variations, the methods and their explanations can be compared based on their computational dependency on the feature permutation [27] function. Identifying the commonalities in the explanations [20] of multiple methods which point to similar feature-wise explanations is considered as revealing the true importance of the underlying data [16]. Hence, the MAMCR method finds the weighted mean for the feature explanations based on the method's consistency in producing similar explanations and through which it provides a comprehensive range for the multiple almost-equally-accurate models. It represents the feature-wise model reliance bounds for all the well-performing models of the pre-specified model class that are computed by the pre-specified methods.

To validate the MAMCR explanation bound suitability to all well performing models, a new, almost-equally-accurate test model is created using the same model class (i.e., Logistic regression) algorithm with random sampling data. This model's accuracy is verified against the Rashomon set threshold (0.6569). The explanations from the six methods are obtained for the model and the grand mean of each variable is found. The test model's feature importance which is plotted along with the MAMCR bounds is displayed in Fig. 7. It elucidates that the test model's feature importance of all the variables lies within the MAMCR boundary values. Thus, the second research question (RQ2), finding the consistent explanation across multiple explanation methods for the almost-equally-accurate models, is addressed through the MAMCR framework by obtaining the method agnostic MCR bounds.

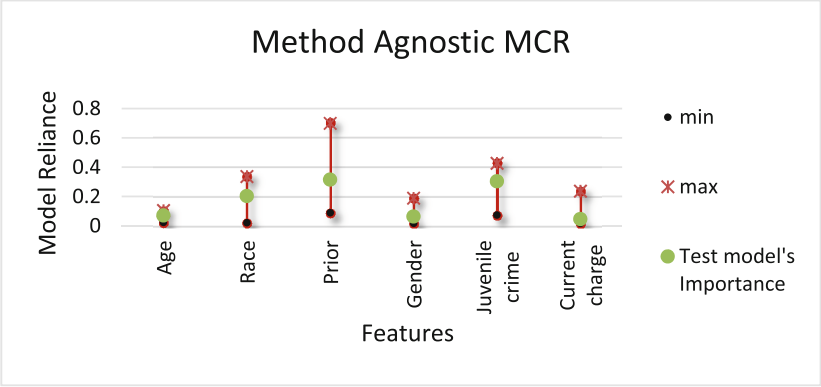

Fig. 7. The feature importance values of a Test model's features along with the MAMCR bounds. The test model's importance values lie within the MAMCR explanation range.

5 Conclusion

The experiments conducted on the COMPAS data set alert us that the method's explanation which highlights a feature as most important may not be projected as such by another method. These inconsistencies in the generated explanations by different explanation methods for the Rashomon set models motivated the proposal of a novel framework for discovering consistent explanations across multiple explanation methods. It provided a method agnostic explanation as a model class reliance for the multiple almost-equally-accurate models. The efficiency of the method agnostic MCR explanation is illustrated by describing the comprehensive variable importance value range for all the well performing models of the pre-specified model class across multiple explanation methods.

In this work, the explanation methods that return the feature importance values as a global explanation are only considered for the explanation ensembling. The future work can be extended for the instance-wise explanations and for other explanation output formats as well.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access **6**, 52138–52160 (2018)
2. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)
3. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)
4. Choudhary, P., Kramer, A.: datascience.com team: datascienceinc/Skater: Enable Interpretability via Rule Extraction (BRL) (v1.1.0-b1). Zenodo (2018). https://doi.org/10.5281/zenodo.1198885
5. Mateusz, S., Przemyslaw, B.: Explanations of model predictions with live and breakdown packages. R J. **10** (2018). https://doi.org/10.32614/RJ-2018-072

6. Gosiewska, A., Biecek, P.: iBreakDown: Uncertainty of Model Explanations for Nonadditive Predictive Models. arXiv preprint arXiv:1903.11420 (2019)
7. Covert, I., Lundberg, S., Lee, S.I.: Feature Removal Is a Unifying Principle for Model Explanation Methods. arXiv preprint arXiv:2011.03623 (2020)
8. Horel, E., Giesecke, K.: Computationally efficient feature significance and importance for machine learning models. arXiv preprint arXiv:1905.09849 (2019)
9. Wei, P., Lu, Z., Song, J.: Variable importance analysis: a comprehensive review. Reliab. Eng. Syst. Saf. **142**, 399–432 (2015)
10. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
11. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L.: Distribution-free predictive inference for regression. J. Am. Statist. Assoc. **113**(523), 1094–1111 (2018)
12. Robnik-Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Trans. Knowl. Data Eng. **20**(5), 589–600 (2008)
13. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328, PMLR (2017)
14. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inform. Syst. **41.3**, 647–665 (2014)
15. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: 2016 IEEE Symposium on Security And Privacy (SP), pp. 598–617 (2016)
16. Gifi, A.: nonlinear multivariate analysis (1990)
17. Kobylińska, K., Orłowski, T., Adamek, M., Biecek, P.: Explainable machine learning for lung cancer screening models. Appl. Sci. **12**(4), 1926 (2022)
18. Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D.I., Ravikumar, P.K.: On the (in) fidelity and sensitivity of explanations. In: Proceedings of the NeurIPS, pp. 10 965–10 976 (2019)
19. 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**(177), 1–81 (2019)
20. Jamil, M., Phatak, A., Mehta, S., Beato, M., Memmert, D., Connor, M.: Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football. Sci. Rep. **11**(1), 1–7 (2021)
21. Wolpert, D.H.: The supervised learning no-free-lunch theorems. In: Roy, R., Köppen, M., Ovaska, S., Furuhashi, T., Hoffmann, F. (eds.) Soft Computing and Industry, p. 2542. Springer, London, U.K (2002). https://doi.org/10.1007/978-1-4471-0123-9_3
22. Dong, J., Rudin, C.: Exploring the cloud of variable importance for the set of all good models. Nature Mach. Intell. **2**(12), 810–824 (2020)
23. Lin, S.: Rank aggregation methods. Wiley Interdiscipl. Rev. Comput. Statist. **2**(5), 555–570 (2010)
24. Kendall, M.G.: Rank correlation methods (1948)
25. Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: dalex: responsible machine learning with interactive explainability and fairness in python. J. Mach. Learn. Res. **22**(1), 9759–9765 (2021)
26. Erdem, A.: https://github.com/aerdem4/lofo-importance. Accessed 22 July 2022
27. Covert, I., Lundberg, S.M., Lee, S.I.: Explaining by removing: a unified framework for model explanation. J. Mach. Learn. Res. **22**, 209–211 (2021)
28. Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. **28**, 4 (2010)
29. Ning, Y., et al.: Shapley variable importance cloud for interpretable machine learning. Patterns 100452 (2022)

30. Hamamoto, M., Egi, M.: Model-agnostic ensemble-based explanation correction leveraging rashomon effect. In: 2021 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 01–08. IEEE (2021)
31. Semenova, L., Rudin, C., Parr, R.: A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. arXiv preprint arXiv: 1908.01755 (2019)
32. Bobek, S., Bałaga, P., Nalepa, G.J.: Towards model-agnostic ensemble explanations. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A. (eds.) ICCS 2021. LNCS, vol. 12745, pp. 39–51. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-77970-2_4
33. Nguyen, T.T., Le Nguyen, T., Ifrim, G.: A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In: Lemaire, V., Malinowski, S., Bagnall, A., Guyet, T., Tavenard, R., Ifrim, G. (eds.) AALTD 2020. LNCS (LNAI), vol. 12588, pp. 77–94. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-65742-0_6
34. Fan, M., Wei, W., Xie, X., Liu, Y., Guan, X., Liu, T.: Can we trust your explanations? Sanity checks for interpreters in Android malware analysis. IEEE Trans. Inf. Forensics Secur. **16**, 838–853 (2020)
35. Ratul, Q.E.A., Serra, E., Cuzzocrea, A.: Evaluating attribution methods in machine learning interpretability. In: 2021 IEEE International Conference on Big Data (Big Data) pp. 5239–5245 (2021)
36. Rajani, N.F., Mooney, R.J.: Ensembling visual explanations. In: Escalante, H.J.,et al. (eds.) Explainable and Interpretable Models in Computer Vision and Machine Learning. TSSCML, pp. 155–172. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-98131-4_7
37. Velmurugan, M., Ouyang, C., Moreira, C., Sindhgatta, R.: Evaluating Explainable Methods for Predictive Process Analytics: A Functionally-Grounded Approach. arXiv preprint arXiv:2012.04218 (2020)
38. Bland, J.M., Kerry, S.M.: Weighted comparison of means. BMJ **316**(7125), 129 (1998)