L. Alfeo, A. and Cimino, M. (2024). Counterfactual-Based Feature Importance for Explainable Regression of Manufacturing Production Quality Measure. In Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods - ICPRAM; ISBN 978-989-758-684-2; ISSN 2184-4313, SciTePress, pages 48-56. DOI: 10.5220/001236960000365

Counterfactual-based feature importance for explainable regression of manufacturing production quality measure

Antonio L. Alfeo[1,2][a], and Mario G. C. A. Cimino[1,2][b]
[1]*Dept. of Information Engineering, University of Pisa, Pisa, Italy*
[2]*Research Center E. Piaggio, University of Pisa, Pisa, Italy*
{*luca.alfeo, mario.cimino*}@*unipi.it*

Keywords: eXplainable Artificial Intelligence, Expert-based validation, Feature Importance, Regression Problem

Abstract: Machine learning (ML) methods need to explain their reasoning to allow professionals to validate and trust their predictions, and employ those in real-world decision-making processes. To do so, explainable artificial intelligence (XAI) methods based on feature importance can be employed, even though those can be very computationally expensive. Moreover, it can be challenging to determine whether an XAI technique might introduce bias into the explanation (e.g., overestimating or underestimating the feature importance) in the absence of some reference feature importance measure or even some domain knowledge from which deriving an expected importance level for each feature. We address both these issues by (i) employing a counterfactual-based strategy, i.e. deriving a measure of feature importance by checking if some minor changes in one feature's values significantly affect the ML model's regression outcome, and (ii) employing both synthetic and real-world industrial data coupled with the expected degree of importance for each feature. Our experimental results show that the proposed approach (BoCSoRr) is more reliable and way less computationally expensive than DiCE, a well-known counterfactual-based XAI approach able to provide a measure of feature importance.

1 Introduction and Motivations

Machine learning technology has become ubiquitous, with unprecedented recognition performances (Alfeo et al., 2017) and applications spanning across every domain (Alfeo et al., 2019). However, since ML approaches can work as a black box, domain experts cannot easily validate and trust their outcomes (Alfeo et al., 2022a). This is especially important in real-world scenarios such as in smart manufacturing (Alfeo et al., 2022b). The adoption of AI technology can improve the manufacturing productivity only if the AI's outcomes can be understood and trusted enough to be integrated into decision-making processes (İç and Yurdakul, 2021). To address this challenge, Explainable Artificial Intelligence (XAI) methodologies can be used to provide some insights into the reasoning of ML models (Jeyakumar et al., 2020). Employing XAI techniques in smart manufacturing contexts can indeed lead to cost reduction, prediction error minimization, and enhanced debugging of AI-based systems (Ahmed et al., 2022).

The explanations provided via post-hoc XAI techniques can be organized according to their scope and form. An explanation can be *local* or *global*. Local explanations focuses on the ML's outcome for a specific instance. Global explanations offer insights into the decision process of the ML model as a whole. Moreover, according to the recent survey (Miller, 2019), the explanations' form can be organized into three main groups: (i) *Instance-based* explanations link a given instance to prototypes or counterfactual examples, triggering a similarity-based reasoning for end-users like domain experts. Given one data instance, its counterfactual is a similar instance that corresponds to a different ML model's outcome (Delaney et al., 2021); (ii) *Attribution-based* explanations unfold the AI model's decision process by evaluating the contribution of each input feature to the prediction. Attribution-based approaches can provide both local and global explanations (Afchar et al., 2021); and (iii) *Rule-based* explanations attempt to approximate the decision process of the algorithm by associating labels with input feature thresholds (van der Waa et al., 2021). Choosing a suitable explanation form is an application-dependent design choice. In the context of smart manufacturing, the end-users are

[a] https://orcid.org/0000-0002-0928-3188
[b] https://orcid.org/0000-0002-1031-1959

typically non-experts in AI. So, highly comprehensible explanations should be prioritized. In this regard, attribution-based (e.g., feature importance) and instance-based (e.g., counterfactual) explanations can be employed (Markus et al., 2021). Specifically, feature importance approaches are widely used due to the availability of model-agnostic techniques that generate feature rankings (Afchar et al., 2021). The widespread use of these approaches has exposed their limitations, including their great computational complexity (Kumar et al., 2020). To address this limitation, an increasing body of research is exploring innovative strategies that combine feature importance and counterfactual explanations (Alfeo et al., 2023b).

Furthermore, it's well-known that the assumptions behind the reliability of feature importance measures may not hold in real-world scenarios. For instance, when the features are characterized by significant correlation and co-dependence, some measures of feature importance may become unreliable (Marcílio and Eler, 2020). In these cases, it's challenging to determine whether the XAI method is overestimating or underestimating the importance of one feature. A proper validation of a feature importance measure would need some reference importance measure or domain knowledge that provides an expected importance level for each feature. Unfortunately, this need is often neglected due to the high costs and the difficulties associated with obtaining an *a-priori* quantitative assessment of feature informativeness (Arras et al., 2022; Ali et al., 2023).

In this paper, we tackle this issue by leveraging both synthetic and real-world datasets that include a degree of expected importance for each feature. With the synthetic dataset, the expected importance of each feature is imposed by the data generation procedure (Guidotti, 2021). With the real-world dataset, we rely on the expertise of manufacturing domain professionals to obtain the expected feature importance level for each feature (Barr et al., 2020). We employ these datasets to evaluate a novel efficient counterfactual-based feature importance measure for regression problems (BoC-SoRr). The proposed method is compared to a well-established counterfactual-based feature importance measure from the state-of-the-art.

The structure of the paper is as follows. Section 2 presents the related works. Section 3 details the method we propose. Section 4 covers the case studies and the experimental setup. Section 5 discusses the results obtained, and finally, Section 6 outlines the conclusions drawn from this study.

2 Related Works

The majority of the research works using XAI approaches employ feature importance methods (Miller, 2019). Those methods assign importance scores to each feature based on some criteria, such as by estimating the Shapley value (Sundararajan and Najmi, 2020). In contrast, counterfactual explanations are minimally different versions of the sample whose predictions need to be explained, that results in a different prediction.

According to (Kommiya Mothilal et al., 2021), counterfactuals can offer an alternative means for deriving feature importance measures since both those approaches focus on the model's decision boundary. Counterfactual approaches aim to find the minimal change in the data instance that results in the crossing of the model's decision boundary (Schleich et al., 2021), whereas some feature importance measures attempt to approximate it (Ribeiro et al., 2016). This allows employing counterfactual approaches to generate new and improved procedures to measure the feature's importance and vice versa. Indeed, there is a growing body of research exploring these strategies (Alfeo et al., 2023b).

For instance, in (Wiratunga et al., 2021), the authors propose an approach for generating counterfactuals by modifying the most important features, as measured via Shapley values. Given an instance to be explained, the authors in (Vlassopoulos et al., 2020) obtain a local approximation of the model's decision boundary by generating counterfactuals via a variational autoencoder. Similarly, in (Laugel et al., 2018) the authors generate new instances in a hypersphere surrounding the sample to be explained to provide local decision rules that are consistent with the model's decision boundary. However, the high computational cost of this approach makes it non-feasible for datasets with a high number of features. DiCE (Mothilal et al., 2020) stands out as a renowned approach that employs counterfactual explanations to derive feature importance measures. DiCE generates a set of counterfactual explanations for a given prediction. By examining how the feature values change across the counterfactuals, DiCE provides insights into which features have the most significant influence on the prediction outcome. In short, the features that exhibit the greatest variation in the counterfactuals are considered more important. Considering its established reputation and its suitability both for classification and regression problems (Dwivedi et al., 2023), we select DiCE as the state-of-the-art benchmark for evaluating our proposed feature importance measure.

3 Design

In this section, the design of the proposed method is detailed.

We propose the Boundary Crossing Solo Ratio for Regression problem (BoCSoRr), a global feature importance measure obtained by aggregating local counterfactual explanations. This method is an adaption for the regression problem of the method presented in (Alfeo et al., 2023b). The method in (Alfeo et al., 2023b) is designed for the classification domain and is based on the concept of counterfactuals, i.e. samples characterized by minor features change but different classes. We employ this terminology in this study even if in this case there are no counterfactual classes. Indeed, since we address a regression problem the concept of counterfactual class is replaced by a more broad "significant difference in the model's outcome". Specifically, BoCSoRr evaluates the importance of one feature by considering the frequency with which a slight change in the value of that specific feature results in a significant change in the model's outcome (see Fig. 1).

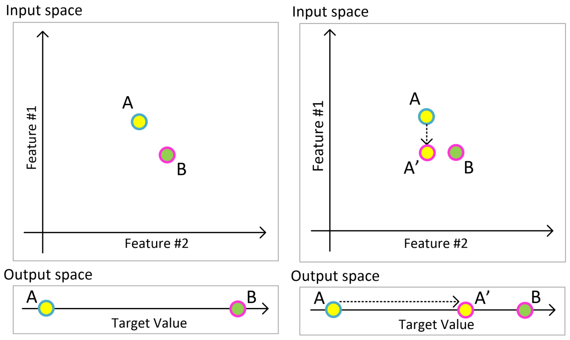

Figure 1: Representation of the idea behind BoCSoRr's feature relevance. Let's consider two samples, A and B, that are close in the input space and distant in the output space. Sample A' is generated using sample A and changing the value of Feature #1 with the one of sample B. Feature #1 can be considered *relevant* since, as a result of its change alone, A' is closer to B than A in the output space.

To find the samples that are close in the input space (i.e. the feature space) and distant in the output space (i.e. the regression target), we consider both (i) the cosine similarity (Abbott, 2014) between all the samples in the feature space, and (ii) the Euclidean distance of the labels (i.e. the target quantity of the regression problem). Via a min-max procedure (Abbott, 2014), we rescale between 0 and 1 both of these quantities and aggregate those (lines 2-4 of Algorithm 1), to obtain the $counterDist$, a distance bounded between 0 and 2. 0 corresponds to samples pair very far apart in the input space and with minimal difference in the output space. 2 corresponds to sample pairs very close in the input (whose values will therefore be characterized by slight differences) and with maximum difference in the output space. As in (Alfeo et al., 2023b), we aim at identifying sample pairs characterized by a great $counterDist$. Firstly, we select the n samples corresponding to the greatest average $counterDist$ with all the samples (line 5 of Algorithm 1), and then for each i instance among those, we select the k samples with the greatest $counterDist$ w.r.t. i (line 7 of Algorithm 1). As introduced at the beginning of this section, we call those samples *counterfactuals*. The above-presented two-step search procedure is preferred to a simpler search for the pairs of samples characterized by the greatest $counterDist$ to avoid populating the search result with outliers.

Then, we substitute (one at a time) the features' value of samples i with one of its counterfactuals. If this single value change shifts the prediction closer to the label of the counterfactual, the modified feature is considered *relevant*. The regression outcome change is measures as the absolute value of their Euclidean distance (lines 10-12 of Algorithm 1).

By taking into account all the samples i and their counterfactuals, the frequency with which a feature is considered relevant (line 17, Algorithm 1) is used as a proxy of the importance of that feature (Vlassopoulos et al., 2020). Algorithm 1 shows a high-level pseudocode of the above-described procedure.

4 Experimental setup

This section presents the experimental setup, the metrics used to evaluate the convenience of proposed method, and the datasets we employed in the experiments.

4.1 Synthetic dataset

To assess the reliability of the feature importance measure, we build a tabular dataset with predefined feature importance levels, following a methodology similar to the ones outlined in (Barr et al., 2020) and (Alfeo et al., 2023b).

Specifically, we employ the *make_regression* function from the Python library *scikit-learn* (Pedregosa et al., 2011). This function generates the data for a random regression problem. The target values are obtained as a random linear combination of the features, to which optional sparsity and noise can be included (Pedregosa et al., 2011). The *make_regression* function allows to explicit select how many samples to generate and how many features

Algorithm 1 Procedure to measure the feature importance (i.e., *BoCSoRr*).

```
Requires:
$M \Leftarrow$ trained machine learning model
$M(s) \Leftarrow$ the prediction of $M$ for instance $s$
$I \Leftarrow$ set of all the instances in the data
$F \Leftarrow$ set of all the features in the data
$L \Leftarrow$ set of all the labels in the data
$n \Leftarrow$ number of instances to query
$k \Leftarrow$ number of counterfactuals per instance to query
Procedure:
$relevantFeatures \Leftarrow emptyList()$
$instanceDist \Leftarrow computePairwiseDistance(I, I, Cosine\_similarity)$
$labelDist \Leftarrow computePairwiseDistance(L, L, Euclidean\_distance)$
$counterDist \Leftarrow minMax(labelDist) + minMax(instanceDist)$
$instancesToQuery \Leftarrow samplesWithTopAvgDist(counterDist, n)$
For each $i \in instancesToQuery$
    $counterfactuals \Leftarrow samplesWithTopDist(counterDist, i, k)$
    For each $c \in counterfactuals$
        For each $f \in F$
            $s_{tmp} \Leftarrow changeFeatureValue(i, c, f)$
            if ($|M(s_{tmp}) - M(i)| > |M(s_{tmp}) - M(c)|$ )
                $relevantFeatures.append(f)$
            End if
        End For
    End For
End For
$featureImportance \Leftarrow frequenceByFeature(relevantFeatures)$
return $featureImportance$
```

should be informative (or not) in the dataset. Thus, we build a dataset consisting of 1000 samples, with five informative features followed by ten non-informative ones. Also, some additional Gaussian noise is introduced to each non-informative feature. The amount of noise progressively increases from the 6th feature to the 15th, such that those are supposed to be less and less informative. It's worth noting that since it is impossible to quantify how the addition of noise precisely diminishes the informativeness of one feature, the ground truth feature importance for this synthetic dataset consists of two importance levels, high (informative features) and low (non-informative features). Similar to other studies that generate data with known feature importance (Yang and Kim, 2019), the approach used to generate the synthetic dataset allows for comparison of the feature importance as measured via different XAI approaches on the same features and assess if the computed importance metrics is more or less aligned with the expected feature importance values.

4.2 Real-world dataset

This study employs a real-world dataset provided by Koerber Tissue, a company specialized in the production of industrial machines for tissue paper manufacturing. These machines are aimed at processing big reels of raw paper by pressing and gluing the paper layers while embossing a specific motif (e.g. the logo of the seller) onto the final product. Each machine is tested using various paper types and production settings, such as the machine speed or embossing pressure (i.e., the features of our analysis). For each production setting, multiple measurements are taken on the final product. These measurements encompass quality-related measures of the final product, such as paper resistance and bulkiness, i.e. the target of our analysis.

Like many real-world datasets, the company's data needs to be preprocessed to remove sensitive information and handle the missing values. To this aim, all of the columns with more than 66% missing values are removed. Then, we remove all the data instances with more than 50% features characterized by missing values. Finally, all the data instances are clustered according to the values of the most informative and sensitive features that do not present missing values. For each feature, the numerical missing value of one data instance is replaced with the median of its cluster. The sensitive features are then removed.

The resulting dataset consists of more than 440 instances and 12 attributes (Table 1), which are: (i) a unique identifier for each test measurement (ID), which is not considered an informative feature and thus it is removed from the analysis; (ii) the percentage of elongation of the raw paper (when dry) in the latitudinal (ELOLA) and longitudinal direction (ELOLO); (iii) the ratio of the raw paper resistance in the longitudinal and latitudinal directions (DRYRAT). (iv) the hardness of the rubber top (TRH) and bottom (BRH) roll used to imprint a motif on the paper, and measured in Shore A; (v) the strength of the raw paper in the latitudinal (STRLA) and longitudinal direction (STRLO); (vi) the thickness of the raw paper (THICK); (vii) the weight of the raw paper (WEIGHT); and (viii) the number of tissue layers in the final product (LAYERS); (ix) the bulkiness of the final product (BULK), i.e. the targets of the analysis.

In order to obtain the ground truth for the feature importance, we gathered both the experts of the tissue production process and the machine data analysts. The level of expected importance for a feature in the data results from their agreement on how critical and informative that feature could be for recognizing the bulkiness of the final product according to their experience and domain knowledge. For the purpose of this analysis, those levels are grouped into two categories, high and low.

Table 1: Expected feature importance level according to the domain experts.

Attribute	Units	Expected Imp.
ID	Integer	-
ELOLA	%	LOW
ELOLO	%	LOW
DRYRAT	Real	LOW
TRH	ShA	LOW
BRH	ShA	LOW
STRLA	N/m	LOW
STRLO	N/m	LOW
THICK	mm	HIGH
WEIGHT	gr/m^2	HIGH
LAYERS	Integer	HIGH
BULK	Real	-

4.3 Performance evaluation

As motivated in Section 2, the proposed method is compared with an established state-of-the-art method that derives measures of importance from counterfactuals, i.e. DiCE (Mothilal et al., 2020). All experimental results are provided via 10 repeated trials, the obtained performances are shown in aggregate form as mean or confidence interval. Each iteration includes the re-training of the ML approach and the measurements of the feature importance. This ensures that different measures of feature importance (e.g. BoCSoRr and DiCE) are actually evaluating the same trained ML model at each iteration.

As performance measures, we first consider the *computational cost* of the proposed method. To do so we measure the time (in seconds) required to compute the feature's importance. The smaller the computation time needed to obtain the feature importance measure, the better. All the experiments are run on the same Google Colaboratory session, featuring an Intel Xeon CPU with 2 vCPUs, and 13 GB RAM. To ensure better comparability, the number of samples used by DiCE to compute the feature importance is constrained to those used by BoCSoRr, i.e. k times n.

Then, we consider the method's *fidelity*, i.e. if the feature importance measure correctly represents the ML model's decision process (Coroama and Groza, 2022). The fidelity can be measured by comparing the proposed feature importance measure with a reliable model-based reference measure. For instance, many ML approaches based on decision trees (Abbott, 2014) do provide a built-in measure of feature importance (i.e. the Gini index (Abbott, 2014)) that can be used as a reference measure when the ML model under analysis is indeed a decision tree. Since any feature importance measure provides different importance value ranges, those can be difficult to compare by simply using a distance measure. However, feature importance approaches are often used to derive a rank of the features, and two ranks can easily be compared using measures like the Spearman rank correlation coefficient (Zar, 2005).

The Spearman rank correlation coefficient, denoted as ρ, is a non-parametric measure of the strength and direction of the monotonic relationship between two variables. Spearman's ρ is calculated by first transforming the array of data points (i.e. the value of measured importance for each feature) into ranks and then computing the Pearson correlation coefficient on the ranked data. It ranges from -1 to 1, where $\rho = 1$ represents a perfect increasing monotonic relationship, and $\rho = -1$ represents a perfect decreasing monotonic relationship.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (1)$$

In (1), d_i represents the difference between the ranks of feature i, whereas n is the number of features. We employ Spearman's ρ between the reference feature importance measure and the one obtained via the other feature importance method as a measure of its fidelity.

Finally, we consider the *empirical correctness* of the proposed method, i.e. if there is an agreement between the obtained feature importance measure and the expected importance of each feature, i.e. the ground truth (Coroama and Groza, 2022). Such ground truth can be derived from domain experts' knowledge (Alfeo et al., 2022b) or available by design if the data are generated according to some predefined level of importance (Guidotti, 2021). Similarly to the proposal in (Alfeo et al., 2023a), to measure the agreement between the ranking of the features obtained via a feature importance measure and the ground truth, we group the ranked features according to the number of features with a given expected importance level. Since both the datasets used in this study feature two levels of expected importance, we can simply consider the highest features in the rank as important and the rest as non-important. For instance, if there are 5 important features according to the ground truth, the top 5 features according to the computed feature importance measure are labeled as "important", whereas the remaining features are considered "non-important". Then, as the measure of empirical correctness, we consider the percentage of the features that are correctly assigned to the ground truth importance level. This measure is bounded between 0 (worst case) and 1 (best case). As suggested in (Coroama and Groza, 2022), and we call it empirical correctness since it is based on a knowledge-driven *a-priori* assumption on the informativeness of each feature for the regression problem.

Given the different metrics considered in our analysis, we performed a manual exploration of BoCSoRr's hyperparameter space (i.e. k and n), aiming to strike a good trade-off among those. After extensive experimentation within the range of 3 to 20, the best trade-off is identified as k=19 and n=15 for the real data, and k=15 and n=10 for the synthetic data.

5 Results and discussion

Being a model-agnostic feature importance measure, BoCSoRr can be used with any ML method that handles tabular data. In this research, our primary objective is to explain the ML regression model rather than striving for optimal regression performance. As such, for all of our experimentation, we employed a shallow ML approach, i.e. the Decision Tree regressor, parametrized using the default values provided by the scikit-learn library (Pedregosa et al., 2011).

The Decision Tree regressor (Abbott, 2014) is an ML method used to predict a numerical value. This model creates a tree-like structure of rules by dividing the data into decision nodes. These splits are based on the values of one feature, and the "purity" (measured via the Gini Index) of the data groups resulting from the split. During tree construction, the Decision Tree keeps track of how variables influence the reduction of the Gini Index. Variables that significantly contribute to reducing the Gini Index are considered more important. Once trained, the Decision Tree provides a built-in measure of feature importance computed by summing the Gini Index reductions for a variable across all the splits it's involved in. This measure will be used to evaluate the fidelity of the proposed method.

We computed the feature importance for the synthetic data by employing both DiCE and BoCSoRr. In Fig. 2, the solid line represents the average feature importance measure obtained via BoCSoRr, whereas the dashed line represents the average feature importance measure provided by DiCE. Their values are rescaled via a min-max procedure to better compare them visually. The background color indicates the expected feature importance, with the initial five features being deemed significant (indicated by a green background), and subsequent non-informative features, resulting in less and less expected importance due to the introduction of noise (transitioning from green to red). The resulting average feature importance measures provided by BoCSoRr exhibit a better consistency with the expected feature importance since it does result in lower importance for the non-informative features (i.e. from the 6th to the 15th) and an overall greater measured importance with the informative ones. On the other hand, DiCE seems to provide very small importance to the 5th feature (an informative one) and greater importance to the 11th feature (a non-informative one).

We computed the performance measures presented in Section 4.3 with the synthetic data. As evident from the results in Table 2, compared to DiCE, BoCSoRr offers (i) significantly less computation time, i.e. less than half the one required for DiCE to compute the feature importance; (ii) greater fidelity, which means a greater agreement between the feature importance ranks obtained via the built-in feature importance measure of Decision Tree and

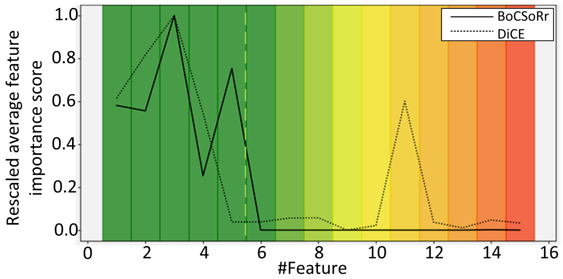

Figure 2: Feature importance obtained via DiCE and BoCSoRr with the synthetic dataset. Their values are rescaled via a min-max procedure to better compare them visually. The dashed line represent the separation between informative and non-informative features.

Table 2: 95% confidence interval of the performance evaluation measures obtained with ten repeated trials on the synthetic dataset. All the rank correlation (i.e. the measure of fidelity) values corresponds to p-values lower than 0.05.

Measure	BoCSoRr	DiCE
Fidelity [ρ]	**0.79** ± **0.03**	0.72 ± 0.08
Empirical c. [%]	**0.95** ± **0.05**	0.88 ± 0.03
Comput. time [*s*]	**31.45** ± **3.48**	64.33 ± 3.54

the ranks obtained via BoCSoRr, thus BoCSoRr better captures the decision process of the decision tree; and (iii) greater empirical correctness, which means a greater agreement between the feature importance ranks obtained via BoCSoRr and the expected feature importance as per the synthetic data construction procedure. The last result was anticipated by the qualitative evaluation provided with the results in Fig. 2.

With the real-world industrial dataset, the decision tree results in good regression performance, with an average Mean Square Error of 0.0065 while predicting the value of the paper's bulkiness. Then, we explain the trained decision tree with BoCSoRr and DiCE. By considering the performance measures described in Section 4.3, BoCSoRr provides better results than those obtained by DiCE, with each considered performance measure (see Table 3).

Compared to the results obtained with the synthetic data, between BoCSoRr and DiCE there is a greater difference in terms of fidelity and a smaller difference in terms of computation time. Since BoCSoRr and DiCE use the same number of samples to compute feature importance (see Section 4.3), the latter result can be interpreted as better scalability of BoCSoRr with respect to the number of features. In fact, as the number of features increases, BoCSoRr consistently outperforms DiCE in terms of percentage improvement. On the other hand, the difference in terms of fidelity requires further investigation. It is well-known from the literature that the correlation between features may affect the measurement of their importance for the ML model (Marcílio and Eler, 2020). Thus, any misalignment between feature importances should also be analyzed considering the correlation between the features. To analyze how the considered feature importance measures are affected by the features' correlation, we computed the maximum correlation (*MC*) between each feature and any other feature in the dataset. Then, we select the five features characterized by the most similar importance value according to BoCSoRr and DiCE. To ensure the comparability the importance values measured by BoCSoRr and DiCE are rescaled via a min-max procedure. We repeat this procedure to identify the five features characterized by the most dissimilar importance values according to BoCSoRr and DiCE. The violin plots in Fig. 3 illustrate the *MC* values obtained for these two groups of features.

Table 3: 95% confidence interval of the performance evaluation measures obtained with ten repeated trials on the real-world dataset. All the rank correlation (i.e. the measure of fidelity) values corresponds to p-values lower than 0.05.

Measure	BoCSoRr	DiCE
Fidelity [ρ]	**0.73** ± **0.06**	0.23 ± 0.18
Empirical c. [%]	**0.66** ± **0.07**	0.58 ± 0.08
Comput. time [*s*]	**20.97** ± **0.87**	39.10 ± 0.75

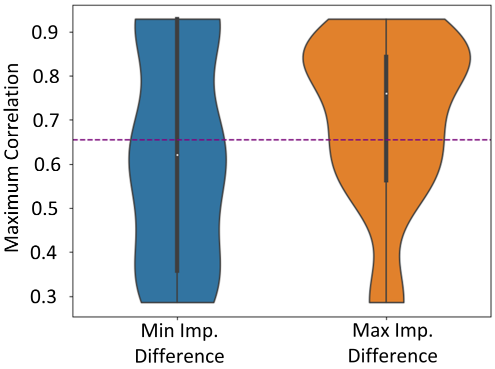

Figure 3: The *MC* of the 5 features characterized by the most similar and dissimilar importance scores provided by BoCSoRr and DiCE, with the real-world industrial dataset. The dotted line represent the average *MC* with all the features.

According to the results shown in Fig. 3, the group of features characterized by the greater importance difference as measured via DiCE and BoCSoRr is characterized by a greater *MC*. Overall, their average *MC* is even greater than the average *MC* among all the features of the whole dataset (the dashed line in Fig. 3). Vice versa for the features of the other group. In short, BoCSoRr is more reliable than DiCE, and they mostly disagree on the features that are more correlated with any other feature of the dataset. This may suggest that the difference both in terms of empirical correctness and fidelity can be motivated by the greater robustness of BoCSoRr to feature correlation.

6 Conclusion

We have introduced a novel model-agnostic measure of global feature importance for regression problems, namely BoCSoRr. BoCSoRr broadly utilizes the concept of counterfactuals and applies it to regression problems, to determine which features, if modified, are most likely to result in a significant change in the ML model's outcome. In our experiments, we employed both synthetic and real-world data and compared BoCSoRr performances against the ones obtained via DiCE, a well-known counterfactual approach able to derive a feature importance measure. With both datasets, BoCSoRr is more reliable and less computationally expensive than DiCE. The reliability of BoCSoRr is tested both in terms of fidelity, i.e. agreement with the model build-in feature importance measure, and by employing some human-driven domain knowledge about the expected importance of each feature for the regression problem.

As future research directions, BoCSoRr will be employed with other ML regression approaches with a built-in feature importance measure. This would provide a better understanding of whether the properties proven in this study are consistent despite the ML model used.

ACKNOWLEDGEMENTS

Work partially supported by (i) the company Koerber Tissue in the project “Data-driven and Artificial Intelligence approaches for Industry 4.0”; (ii) the University of Pisa, in the framework of the PRA 2022 101 project “Decision Support Systems for territorial networks for managing ecosystem services”; the European Commission under the NextGenerationEU programs: (iii) Partenariato Esteso PNRR PE1 - ”FAIR - Future Artificial Intelligence Research” - Spoke 1 ”Human-centered AI”; (iv) PNRR - M4 C2, Investment 1.5 ”Creating and strengthening of ”innovation ecosystems”, building ”territorial R&D leaders”, project ”THE - Tuscany Health Ecosystem”, Spoke 6 ”Precision Medicine and Personalized Healthcare”. (v) National Sustainable Mobility Center CN00000023, Italian Ministry of University and Research Decree n. 1033—17/06/2022, Spoke 10; and (vi) the Italian Ministry of Education and Research (MUR) in the framework of the FoReLab project (Departments of Excellence), in the framework of the ”Reasoning” project, PRIN 2020 LS Programme, Project number 2493 04-11-2021, and in the framework of the project ”OCAX -Oral CAncer eXplained by DL-enhanced case-based classification” PRIN 2022 code P2022KMWX3. The authors thank Tommaso Nocchi for his work on the subject during his master's thesis.

REFERENCES

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.

Afchar, D., Guigue, V., and Hennequin, R. (2021). Towards rigorous interpretations: a formalisation of feature attribution. In *International inproceedings on Machine Learning*, pages 76–86. PMLR.

Ahmed, I., Jeon, G., and Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, 18(8):5031–5042.

Alfeo, A. L., Cimino, M. G., Lepri, B., Pentland, A. S., and Vaglini, G. (2019). Assessing refugees' integration via spatio-temporal similarities of mobility and calling behaviors. *IEEE Transactions on Computational Social Systems*, 6(4):726–738.

Alfeo, A. L., Cimino, M. G., and Vaglini, G. (2022a). Degradation stage classification via interpretable feature learning. *Journal of Manufacturing Systems*, 62:972–983.

Alfeo, A. L., Cimino, M. G. C., Egidi, S., Lepri, B., Pentland, A., and Vaglini, G. (2017). Stigmergy-based modeling to discover urban activity patterns from positioning data. In *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*, pages 292–301. Springer.

Alfeo, A. L., Cimino, M. G. C., and Gagliardi, G. (2023a). Matching the expert's knowledge via a counterfactual-based feature importance measure. In *Proceedings of the 5th XKDD Workshop - European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2023)*.

Alfeo, A. L., Cimino, M. G. C. A., and Gagliardi, G. (2022b). Concept-wise granular computing for explainable artificial intelligence. *Granular Computing*, pages 1–12.

Alfeo, A. L., Zippo, A. G., Catrambone, V., Cimino, M. G., Toschi, N., and Valenza, G. (2023b). From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks. *Computer Methods and Programs in Biomedicine*, 236:107550.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, page 101805.

Arras, L., Osman, A., and Samek, W. (2022). Clevr-xai: a benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40.

Barr, B., Xu, K., Silva, C., Bertini, E., Reilly, R., Bruss, C. B., and Wittenbach, J. D. (2020). Towards ground truth explainability on tabular data. *arXiv preprint arXiv:2007.10532*.

Coroama, L. and Groza, A. (2022). Evaluation metrics in explainable artificial intelligence (xai). In *International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability*, pages 401–413. Springer.

Delaney, E., Greene, D., and Keane, M. T. (2021). Instance-based counterfactual explanations for time series classification. In *International inproceedings on Case-Based Reasoning*, pages 32–47. Springer.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. (2023). Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33.

Guidotti, R. (2021). Evaluating local explanation methods on ground truth. *Artificial Intelligence*, 291:103428.

İç, Y. T. and Yurdakul, M. (2021). Development of a new trapezoidal fuzzy ahp-topsis hybrid approach for manufacturing firm performance measurement. *Granular Computing*, 6(4):915–929.

Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., and Srivastava, M. (2020). How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33:4211–4222.

Kommiya Mothilal, R., Mahajan, D., Tan, C., and Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. (2020). Problems with shapley-value-based explanations as feature importance measures. In *International in proceedings on Machine Learning*, pages 5491–5500. PMLR.

Laugel, T., Renard, X., Lesot, M.-J., Marsala, C., and Detyniecki, M. (2018). Defining locality for surrogates in post-hoc interpretablity. In *Workshop on Human Interpretability for Machine Learning (WHI)-International Conference on Machine Learning (ICML)*.

Marcílio, W. E. and Eler, D. M. (2020). From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347. Ieee.

Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " 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*, pages 1135–1144.

Schleich, M., Geng, Z., Zhang, Y., and Suciu, D. (2021). Geco: quality counterfactual explanations in real time. *Proceedings of the VLDB Endowment*, 14(9):1681–1693.

Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR.

van der Waa, J., Nieuwburg, E., Cremers, A., and Neerincx, M. (2021). Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404.

Vlassopoulos, G., van Erven, T., Brighton, H., and Menkovski, V. (2020). Explaining predictions by approximating the local decision boundary. *arXiv preprint arXiv:2006.07985*.

Wiratunga, N., Wijekoon, A., Nkisi-Orji, I., Martin, K., Palihawadana, C., and Corsar, D. (2021). Actionable feature discovery in counterfactuals using feature relevance explainers. CEUR Workshop Proceedings.

Yang, M. and Kim, B. (2019). Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*.

Zar, J. H. (2005). Spearman rank correlation. *Encyclopedia of biostatistics*, 7.