

BoXoR-C: Coherence and Directionality of Counterfactual Explanations

Antonio Maratea*, Domenico Lugubre

¹Department of Science and Technologies, University of Naples "Parthenope", Centro Direzionale Isola C4, 80143, Napoli (IT)

Abstract

Counterfactual explanations and feature selection can be seen as two different ways to recognize the most important variables influencing the outcome of a classifier, the former being valid on an instance scale, the latter on a global scale. They have been integrated recently in the the Bounday Crossing Solo Ratio (BoC-SoR) method, that gives a global relevance score to a feature considering how frequently it generates a counterfactual, that is how frequently it causes a class swap. In this paper a method to generate a local feature selection and to evaluate the stability of counterfactuals at a regional scale is proposed, based on BoC-SoR, with the aim of mitigating the Rashomon effect and highlighting the regions where counterfactuals are unreliable. The method exploits clustering in both the original feature space and in the explanation space, providing directional explanations. Tests on three real-data benchmarks, namely Diabetes, Adult Income and Credit Risk, confirm its viability and effectiveness.

Keywords

Counterfactuals, Rashomon effect, Feature selection, XAI

1. Introduction

With the relentless growth of Machine Learning model complexity and the continuously expanding domain of its applications, the eXplainable Artificial Intelligence has emerged as a way to render AI predictions comprehensible and justifiable to humans [1]. As soon as AI began to be deployed in high-stakes domains, such as healthcare, finance, law, and autonomous systems, the demand for transparency, accountability, and human trust has quickly become the elephant in the room. Fairness, bias, and legal rights to issue a recourse against an algorithmic decision urge explanations in an human understandable form [2].

Meanwhile, the EU has made a notable regulatory effort, first through the General Data Protection Regulation (GDPR), which focuses on privacy and pushes trustworthy AI, emphasizing the need for transparency and interpretability of Machine Learning models; second through the Artificial Intelligence Act in 2024, that introduces transparency requirements in Art. 13: systems should be "sufficiently transparent to enable users to interpret the system's output and use it appropriately" [3].

Needless to say, research on eXplainable Artificial Intelligence (XAI) is in its infancy and enforcing explainability by law without an established and sound scientific background implies the risk of producing plausible, convincing or convenient explanations, instead of reliable, faithful and trustable ones [1, 4].

Guidotti et al. [5, 6] proposed a taxonomy of eXplainable AI (XAI) methods, categorising them as either transparent by design (e.g. decision trees) — that is *intrinsically interpretable* or *ante hoc* — versus *post-hoc*, where explanations are generated after the training step. The latter can be further subdivided into *model explanation* (describing the overall logic of the model), *outcome explanation* (clarifying individual predictions), and *black-box inspection* techniques, whereas these last are furthermore distinguished as either *model-specific* — designed for specific model — or *model-agnostic*, that is universally applicable (see Maratea and Ferone [7]).

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

*Corresponding author.

✉ antonio.maratea@uniparthenope.it (A. Maratea); domenico.lugubre001@studenti.uniparthenope.it (D. Lugubre)

🌐 <https://www.uniparthenope.it/Portale-Ateneo/organigramma/1154> (A. Maratea)

🆔 0000-0001-7997-0613 (A. Maratea)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Given an instance s_i and its label l_i , *counterfactuals* are synthetic instances answering the question "what changes of the values of s_i would have led to a different label?"; they shed some light on the behaviour of the classifier and represent changes that could be addressed to alter the prediction, to the point of being used as a legit explanation for the prediction itself [8, 9, 10]. *Counterfactual explanations* can be classified as post-hoc, local explanations targeted to outcomes and they are interwound with feature importance measures. Whereas feature importance highlight features that are the most important on a global scale for a specific problem, a set of coherent counterfactuals can highlight the features that are the most important for reverting the prediction of the model in a specific region: counterfactual generation can be seen as a micro-scale feature importance determination. Indeed, as demonstrated by Mothilal et al. [11], features classified as highly important by attribution methods, being local as LIME or global as SHAP (when SHAP values are aggregated across all instances in a dataset), are often neither necessary nor sufficient to alter the model prediction. This misalignment raises concerns about the reliability of attribution scores, being true causality the unreachable silver bullet.

Notwithstanding criticalities, due to their simplicity, intuitive nature and recall of human causal reasoning, counterfactual generation elicited a steadily growing interest in recent years. Several major challenges remain open:

- **Granularity**, that is the local nature of the explanation obtained from the counterfactual. While intuitively close instances should have similar counterfactual explanations, each counterfactual is valid for a single instance and it is independent from its neighbours;
- **Actionability**, that is the actual feasibility of the suggested changes on the target variables. Certain counterfactuals may involve modifications that are unrealistic or ethically problematic, such as altering immutable attributes like age or gender (please see Lucic et al.[9]);
- **Directionality**, that is the difference involved in reversing the change from one class to the other: the variables and values to see approved a mortgage application that has been previously denied are different from the variables and values involved in denying a mortgage application that has been previously approved;
- **the Rashomon effect**, that is the presence of several different counterfactuals for the same instance, often contradicting each other. This issue strongly limits the human trust in the counterfactuals due to the conflicting explanations, ambiguous causality and consequent possible cherry picking.

In light of these considerations, here is BoXoR-C, a novel methodology that combines counterfactual generation with feature importance evaluation and clustering on a regional scale, aiming to several advantages:

1. to reduce the Rashomon effect through an aggregation in both the original feature space and the explanation space, so that the explanation validity is expanded from a single instance to the region surrounding the instance;
2. to characterize regions of the original space, where the explanation is stable and hence more likely to be reliable, safe and trustable;
3. to recognize regions of the original space, where the explanation is unstable and hence unlikely to be reliable, safe and trustable;
4. to cluster the explanations and to check whether the corresponding regions are consistently distributed in the original feature space.

The paper is organized as follows: first the counterfactual explanations are defined and their desirable properties briefly listed; then the baseline Boundary Crossing Solo Ratio method, that combines counterfactuals with feature importance, is described in detail; then the proposed method BoXoR-C is presented and its advantages are highlighted; finally the experiments on real data are reported and the conclusions are drawn.

2. Counterfactual Explanations

Building upon the previous discussions, *counterfactual explanations* can be characterised as post-hoc, local, model-agnostic explanation methods. They are meant to elucidate the probable causes behind the classification decision made by a pre-trained model, focusing on a specific instance of input and a minimal perturbation. As Guidotti et al. [6] note, counterfactual explanations fall under the category of active explanation methods, since they are not inherently provided by the classification model.

Thinking in counterfactual terms requires the imagination of a reality that contradicts the observed facts, hence the name "counterfactuals" [12]. From a cognitive psychology point of view, counterfactual reasoning can be regarded as a natural mechanism by which humans interpret cause-effect relationships [10, 13].

In order to serve their purpose and provide actionable, understandable and reliable insights into model predictions, counterfactual explanations should satisfy several desirable properties [6].

- **Validity:** The original instance must have a different label than the counterfactual instance;
- **Minimality and proximity:** Among all valid counterfactuals, the selected one should be the instance that changes the fewest possible number of input features with minimal distance from the original one, while still achieving the desired prediction;
- **Actionability:** Counterfactual explanations must propose changes that are practically feasible and actionable. Indeed, some explanations may suggest unfeasible, unrealistic or ethically problematic changes, such as changing age or gender;
- **Plausibility:** A plausible counterfactual must be realistic and consistent with the observed data distribution. This property ensures that suggested explanations remain meaningful and aligned with real-world scenarios;
- **Diversity:** if that is the case, a set of counterfactual explanations should present diverse and distinct alternatives. This can be the normal consequence of a complex problem with multiple causes, or the side effect of an inadequate model or an overly intricate decision boundary;
- **Causality:** Counterfactual explanations should respect known causal relationships among features. Given that certain features can influence others (e.g., increasing the loan duration typically increases interest rates), a plausible and actionable counterfactual must preserve these established causal dependencies;
- **Discriminative Power:** An effective counterfactual should clearly highlight the features responsible for altering the classification outcome. By comparing the original instance to its counterfactual, users should intuitively understand why the prediction changed. However, discriminative power is inherently subjective and challenging to measure quantitatively without empirical validation involving human judgment or human-like approximation models.

Counterfactuals are one property away from adversarial examples: **imperceptibility**. An adversarial example is nothing more than a counterfactual engineered to be similar to the original instance to the point of being undetectable.

The satisfaction of all these properties, hard as it seems, is instrumental in ensuring theoretical soundness and practical value to counterfactual explanations, thereby enabling users to gain significant insights into the decision-making processes of AI systems.

3. Boundary Crossing Solo Ratio (BoC-SoR)

Feature importance methods, such as Shapley Additive Explanations (SHAP) when aggregated on all instances, are computationally-intensive and sensitive to feature correlation, whereas counterfactual explanations are limited to single predictions, lacking global interpretability. To overcome these

limitations, the *Boundary Crossing Solo Ratio (BoC-SoR)* approach was introduced by Alfeo et al. [14] as a novel explainability method that effectively integrates global feature importance and local counterfactual explanations, under the hypothesis that minimal modifications to the most relevant features significantly increase the likelihood of crossing the decision boundary.

More formally, given a binary problem, the original class O and the opposite class C , for each instance $o \in O$ its closest instance (nearest neighbour) of the other class is $c_{nn_o} \in C$.

The set B of boundary instances is defined as follows:

$$D = \{\text{dist}(o, c_{nn_o}) \mid o \in O\}, \quad B = \{b \in O \mid \text{dist}(b, c_{nn_b}) < \text{percentile}(th, D)\} \quad (1)$$

Where D is the set of all distances among pairs (o, c_{nn_o}) , while B is the set of instances considered boundary points, that is the pairs with a distance with respect to its closest instance of opposite class c_{nn_b} less than a given percentile. Here, dist refers to the Euclidean distance.

According to Alfeo et al. [14], an effective strategy to identify the closest counterfactual is through a k -Nearest Neighbor (k -NN) search: chosen a boundary instance, the k nearest instances from the opposite class are initially considered as potential counterfactual candidates for it; then the intermediate instances along the path between the original instance and its nearest neighbours are evaluated to finally find the minimally-different counterfactual instance, similarly to SMOTE [15].

For each boundary instance $b \in B$, its minimally-different counterfactual instance closestCF_b in class C is determined as the instance with minimal Euclidean distance from b according to the procedure described above:

$$\text{closestCF}_b \in C, \quad \text{dist}(b, \text{closestCF}_b) \text{ is minimal} \quad (2)$$

A feature f_i^b at index i is considered relevant if swapping its value in the counterfactual instance closestCF_b with the original value from instance b changes the classification outcome back to class O . Consequently, the global feature importance score (BoC-SoR) for feature i is quantified as:

$$\text{BoCSoR}_{f_i} = |\{b \in B \mid f_i^b \text{ is relevant}\}| \quad (3)$$

For further details on the algorithms please see [14].

4. The proposed method: BoXoR-C

To counteract the Rashomon effect and obtain insights into the regional validity of explanations, the proposed idea is based on clustering and a suitable modification of BoC-SoR. Clustering is here exploited to transform instance-level explanations into regions of validity and BoC-SoR is chosen to transform simple counterfactuals into feature importance measures at the cluster level.

Called F the original space and E the space of explanations, two different clusterings are performed: one in the original space F , to obtain clusters of similar instances and another in the counterfactual explanation space E , to obtain clusters of counterfactuals with similar explanations. The intuition is that similar instances should ideally present similar explanations, so the two clusterings should mostly agree.

In the first case ($F \rightarrow E$), from original instances to explanations, the purpose is to group closer instances in the original space and to characterize these regions according to the internal variability of explanations; whereas in the second direction ($E \rightarrow F$), from explanations to original instances, the purpose is to map similar explanations in the original space so to have an overview of their distribution and highlight critical regions, that is regions where the two clustering in F and E disagree the most.

In figure 1 a schematic view of BoXoR-C is shown.

4.1. From original features to explanations

Given the set of boundary points B defined in equation 1, first they are clustered according to a specific clustering algorithm and a suitable similarity measure in the original feature space. Once the boundary

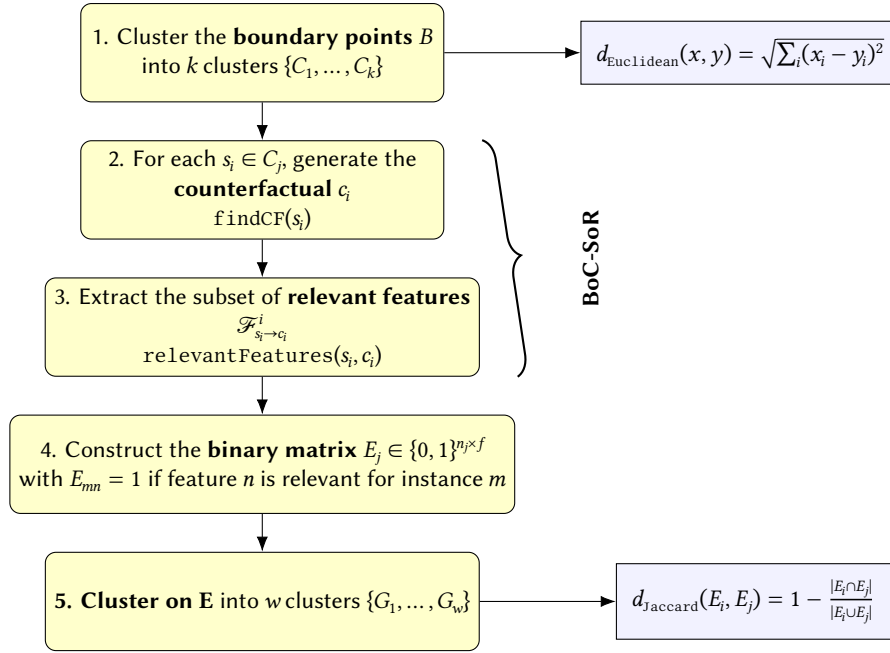


Figure 1: Schematic of BoXoR-C.

instances have been clustered into k clusters named $\{C_1, C_2, \dots, C_k\}$, a local BoC-SoR analysis within each cluster is performed ([14]):

- For each boundary instance $s_i \in C_j$, a class-changing counterfactual c_i is generated using Algorithm 1 (findCF).
- Each pair (s_i, c_i) is then processed using Algorithm 2 (relevantFeatures) to extract the set of features $\mathcal{F}_{s_i \rightarrow c_i}^i$ responsible for the class change.

At this stage, for each cluster C_j , a binary matrix $E_j \in \{0, 1\}^{n_j \times f}$ is built, where n_j is the number of instances in cluster C_j , $f = |F|$ is the total number of features and $E_{mn} = 1$ iff feature n was relevant for the m -th instance.

On each E_j , the Shannon entropy is computed as a global measure of diversity in explanations and for each feature F_n the following quantities are computed:

- The prevalence $p_n = \sum_m^{n_j} \frac{E_{mn}}{n_j}$ of each feature F_n ;
- The binary variance $\text{Var}_n = p_n(1 - p_n)$ of each feature F_n within the cluster.

By combining these three descriptors — prevalence, variability, and entropy — a rich understanding of the internal explanatory structure of each cluster can be obtained: features with high p_n and low Var_n are dominant and stable in the cluster, while high entropy signals random differences in the relevant features and an unstable cluster.

4.2. From explanations to original features

While clustering in the original feature space provides insight into geometric similarity, it does not guarantee consistency in the structure of the explanations. To complement the previous analysis, a second clustering is applied directly on the counterfactual explanations matrix $\mathbf{E} \in \{0, 1\}^{t \times f}$, where t is the total number of instances and each row encodes which features were modified in the counterfactual for a given instance.

To capture structural similarity between explanations, the pairwise **Jaccard distance** between binary vectors is used:

$$d_{\text{jaccard}}(e_i, e_j) = 1 - \frac{|e_i \cap e_j|}{|e_i \cup e_j|},$$

where e_i and e_j denote the binary attribution vectors of two instances. This yields an alternative partitioning based purely on the similarity of counterfactual patterns.

Once the explanations have been clustered into w clusters named $\{G_1, G_2, \dots, G_w\}$, again the three descriptors — prevalence, variability, and entropy — can be computed for each cluster to characterize it.

4.3. Alignment of clusterings

To assess quantitatively the agreement between the explanation-based clusters computed with the Jaccard similarity and those derived from Euclidean distances in the original space, standard agreement metrics can be used:

- **Adjusted Rand Index (ARI)**,
- **Normalized Mutual Information (NMI)**,
- **Homogeneity, Completeness, and V-measure**.

An high agreement would indicate that structurally similar explanations tend to emerge from geometrically close regions in feature space. On the other side a low agreement value may mask some regions where the agreement is qualitatively very good or may be consequence of only a few critical regions on an overall good qualitative performance.

4.4. Directionality

Explanations are directional: transitions from a negative to a positive class leverages different features with respect to transitions from a positive to a negative class: real-world scenarios often require reasoning in both directions.

The BoXoR-C pipeline should include **bidirectional counterfactual analysis**, generating explanations for both types of transitions: the schema in Figure 1 should be applied independently in both class transition directions ($0 \rightarrow 1$ and $1 \rightarrow 0$), allowing the identification of potential asymmetries in the behaviour of the model.

This bidirectional approach uncovers potential asymmetries in the decision boundary, where the importance or frequency of certain features may differ depending on the direction of class change. Understanding these differences provides a more complete and robust view of the decision rationale.

5. Experiments

In the experiments, **hierarchical agglomerative clustering** with Ward’s linkage and Euclidean distance has been tested in F , and with the Jaccard similarity measure and complete linkage has been tested in E . Of course, other clustering techniques (e.g., k -means, DBSCAN, spectral clustering) — not necessarily the same in both directions — can be chosen depending on the nature and geometry of the dataset. The advantage of hierarchical clustering is that it allows to control the granularity.

5.1. Data

The experiments in this study are based on three publicly available datasets: **Adult Census Income**¹, **Diabetes**² and **Credit Risk Dataset**³. The 20th percentile was considered for boundary points in the

¹<https://archive.ics.uci.edu/dataset/2/adult>

²<https://www.kaggle.com/datasets/mathchi/diabetes-data-set>

³<https://www.kaggle.com/datasets/laotse/credit-risk-dataset/data>

first and third datasets, while the 80th percentile was considered for Diabetes, due to its small size.

The **Adult Census Income** dataset, originally from the UCI Machine Learning Repository, has 48,842 instances and 14 attributes, categorical and numerical, related to the demographic and employment characteristics of individuals. The binary target class is 1 for an annual income that exceeds \$50,000, 0 otherwise.

First individuals older than 16 years with a positive value for hours worked, FNLWGT > 1, and adjusted gross income > 100 were filtered out; then missing values in workclass, occupation, and native.country were imputed using the mode. Age was binned into six categorical brackets, rare education levels were grouped into a single 'School' category, and infrequent races were merged into an 'Other' class. All categorical features were encoded using a combination of ordinal, when appropriate, or one-hot encoding; numerical and ordinal features were standardised to zero mean and unit variance.

The dataset was split randomly into 70% training and 30% testing and the LogisticRegression classifier was chosen from Scikit-learn (version 1.6.1) in Python 3.12 (parameters penalty='l2', C=1.0, solver='lbfgs' and max_iter=10000, all others as default).

Diabetes, is from the *National Institute of Diabetes and Digestive and Kidney Diseases*. The dataset contains 768 instances and 8 numeric attributes related to the health of patients. The binary target class is 1 if the patient tested positive for diabetes, 0 otherwise. The data were preprocessed by imputing missing values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI using the median, and then standardizing all features to zero mean and unit variance. The dataset was split into training and test sets with a 60/40 ratio. The training set was further balanced using SMOTE [15] (balanced_classes and k_neighbors=5) to obtain 299 samples per class, resulting in a total of 598 samples. Approximately 246 boundary points were identified in the training set as a baseline for generating counterfactuals.

The XGBoost classifier was trained using the xgboost library (version 2.1.1) with Python 3.12 and scikit-learn 1.6.1. Hyperparameters were set using grid search and stratified k fold cross validation with k=5. The training set contained 460 samples (276 negative, 184 positive) before SMOTE and 598 samples (299 per class) after SMOTE.

Credit Risk, available on Kaggle, contains 32,581 instances with 12 attributes, including socio-economic information of loan applicants and variables simulating credit bureau data. The binary target class takes 1 if the loan was classified as risky, and 0 otherwise.

The data were preprocessed by first removing duplicate records. Missing values in the numerical variables were imputed using the median. Age was discretized into six ordinal brackets, and loan_grade was mapped to an ordinal scale from A (best) to G (worst). Categorical attributes were one-hot encoded. Finally, all numerical and ordinal features were standardized to zero mean and unit variance.

The dataset was split into training and test sets with a 80/20 ratio and the XGBoost classifier was trained using the xgboost library (version 2.1.1) in Python 3.12 with scikit-learn 1.6.1. Hyperparameters were set using grid search and stratified k fold cross validation with k=5. The best configuration was learning_rate=0.3, max_depth=7, and n_estimators=200.

5.2. Results and discussion

The analysis is directional. First BoXoR-C has been applied generating counterfactuals from Class 1 to Class 0, then the reverse.

5.2.1. Adult Income, from class 1 to 0

In this direction, the counterfactuals highlight the key factors associated with transitions to lower income levels, that is the most frequent variables found in counterfactuals generated for high-income people. In extreme synthesis, the most important feature among clusters results education, suggesting that education is a good investment and that an higher education protects from transitioning to lower income.

Figure 2 allows to characterize each cluster in the original feature space. Cluster 4 is the biggest, it shows low entropy and a dominant feature, that implies high stability and interpretability, with education consistently appearing in over 80% of cases. Cluster 3 is much smaller, but also exhibits a low entropy value, being clearly dominated by capital.gain, with age.group appearing rarely but consistently in the relevant features. Clusters 0 and 1 exhibit dispersed patterns, with variable roles of education, capital.gain, and hours.per.week, resulting in limited interpretability. These clusters give inconsistent explanations in terms of counterfactuals and the Rashomon effect should be expected. Despite its small size, Cluster 2 is consistently dominated by education.

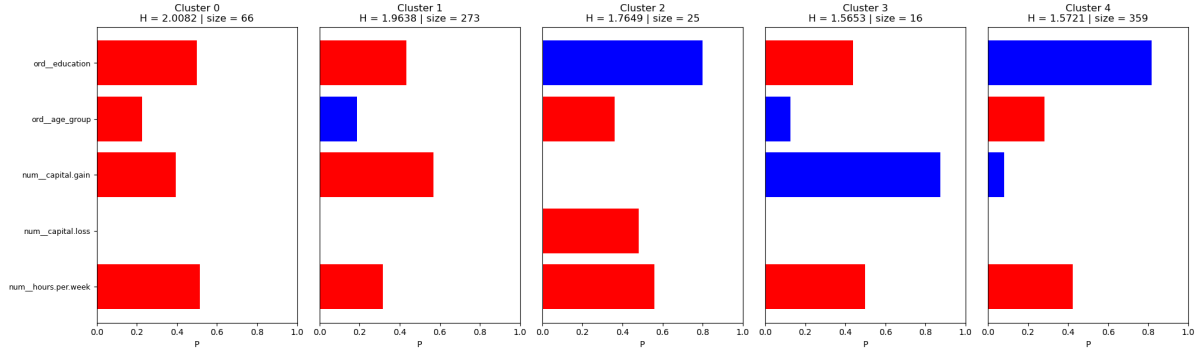


Figure 2: Adult Income, original space – from class 1 to Class 0: histograms representing the prevalence p_n of each feature in each cluster in the original feature space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

Figure 3 allows to characterize each cluster in the explanation space. The average entropy is much lower than before, and the clusters of explanations show more coherence, as there is a clear grouping of possible explanations. Notably, Clusters 3 and 4 primarily involve education, with Cluster 4 that highlights a correlation with age.group. Cluster 0 shows the dominance of capital.gain, with moderate contributions from education and age.group, indicating a partially stable explanation. Despite its smaller size, Cluster 1 is characterized by a sharp focus on occupation-related features, suggesting specific behavioural profiles, albeit with moderate internal variability. Cluster 2, the largest in terms of size and entropy, nonetheless exhibits a dominance of hours.per.week.

There is not a clear match among the clustering in the original space and the one in the explanation.

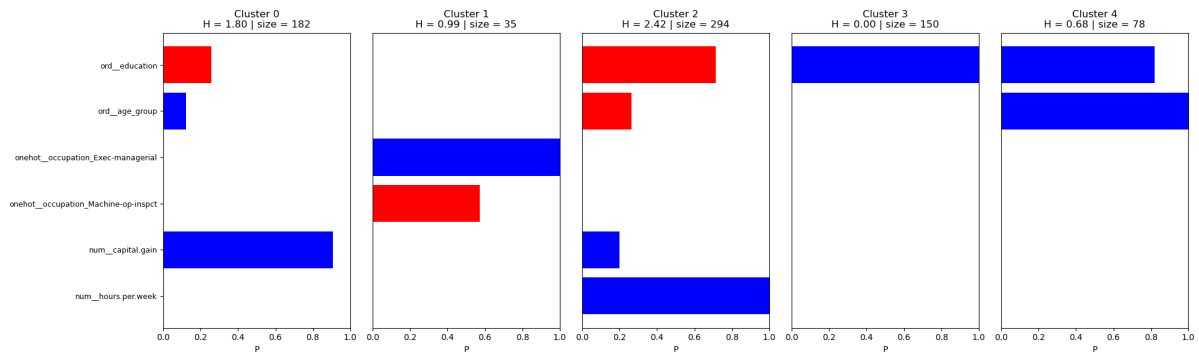


Figure 3: Adult Income, explanation space – from class 1 to Class 0: histograms representing the prevalence p_n of each feature in each cluster in the explanation space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

5.2.2. Adult Income, from class 0 to 1

In this direction, the counterfactuals highlight the key factors associated with transitions to higher income levels, that is the most frequent variables found in counterfactuals generated for low-income people. In extreme synthesis, the most important feature among clusters results `capital.gain`, suggesting that the investment capacity is a key factor for transitioning to higher income.

Figure 4 allows to characterize each cluster in the original feature space. Clusters 0, 2, and 4 demonstrate consistent patterns, particularly in the `capital.gain` field. This observation underscores a correlation between financial attributes and class transitions. Clusters 1, 3, and 5 in particular demonstrate high entropy and feature variability, resulting in more diffuse and less interpretable counterfactuals. This may, in turn, compromise the reliability of the explanations provided in these clusters. `capital.gain` is dominant in several clusters. Features such as `age.group` are non-actionable, while `hours.per.week` offers more practical options, though its explanatory impact is more variable and context-dependent.

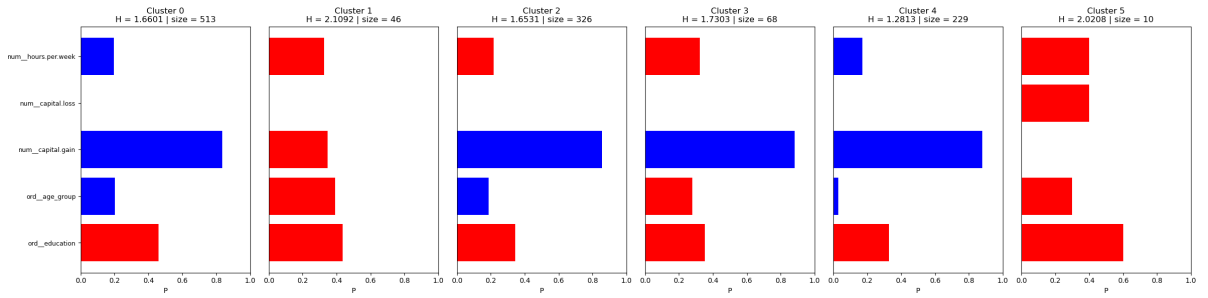


Figure 4: Adult Income, original space – from class 0 to Class 1: histograms representing the prevalence p_n of each feature in each cluster in the original feature space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

Figure 5 allows to characterize each cluster in the explanation space. Cluster 0, the largest, is dominated by `capital.gain`, with moderate contributions from education and hours.per.week, indicating a relatively stable but partially heterogeneous explanatory pattern. Cluster 1 and 5 are the most variable, with a broader spread across multiple features, including hours.per.week and education, suggesting unstable and less interpretable explanations in this region. Cluster 2 and 3 reflect a similar reliance on `capital.gain`, with the additional involvement of `age.group`, reflecting the correlation among the two.

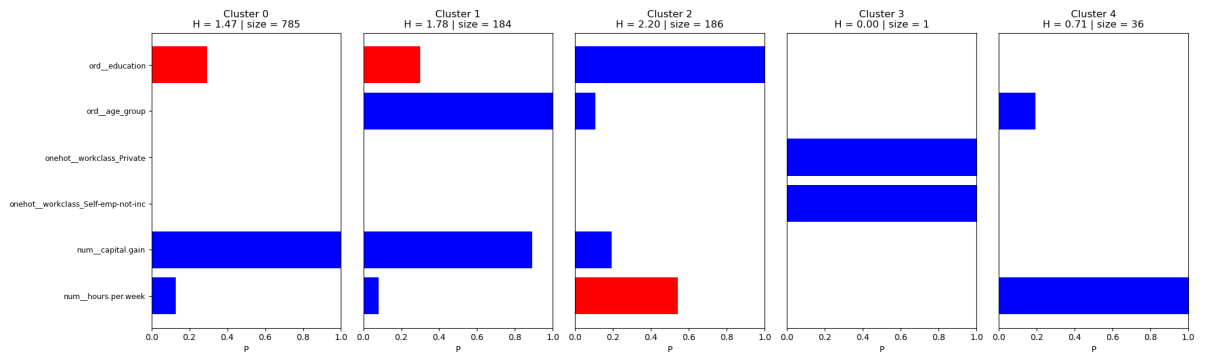


Figure 5: Adult Income, explanation space – from class 0 to Class 1: histograms representing the prevalence p_n of each feature in each cluster in the explanation space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

5.2.3. Diabete, from class 1 to 0

In this direction, the counterfactuals highlight the key factors associated with transitions from having a diabetes to being sane, that is the most frequent variables found in counterfactuals from class "diabetes" to "sane". While diabetes cannot be cured and most features are non-actionable, nonetheless the analysis allows to highlight the risk factors for Diabetes, that turned out to be coherent with medical literature. It must be stressed that the dataset is very small and that this result has a remarkable generality.

Figure 6 allows to characterize each cluster in the original feature space. Cluster 2 is clearly an outlier, while the only cluster with moderate entropy and a strong characterization is cluster 4, dominated by pedigree/familiarty. Triceps skin thickness, even if with low prevalence, appears consistently in 3 clusters and show to have a non-marginal influence.

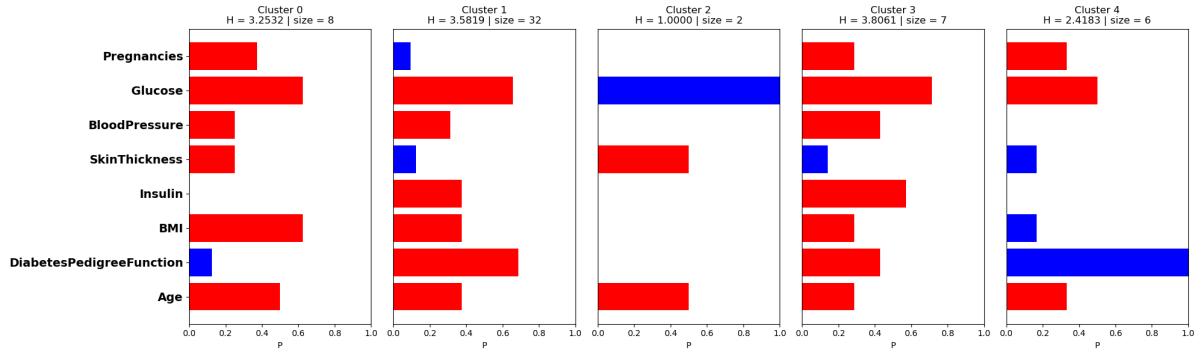


Figure 6: Diabete, original space – from class 1 to Class 0: histograms representing the prevalence p_n of each feature in each cluster in the original feature space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

Figure 7 shows very small and strongly characterized clusters: Clusters 2 and 3 are dominated by triceps skin thickness and age and pedigree/familiarty respectively. Cluster 0, that is the biggest, is dominated by glucose and pregnancies, while cluster 1 and 4 are outliers. Net of noise due to the small sample size, the key risk factors for Diabetes (familiarity, triceps skin thickness, age and pregnancies) can be recognized.

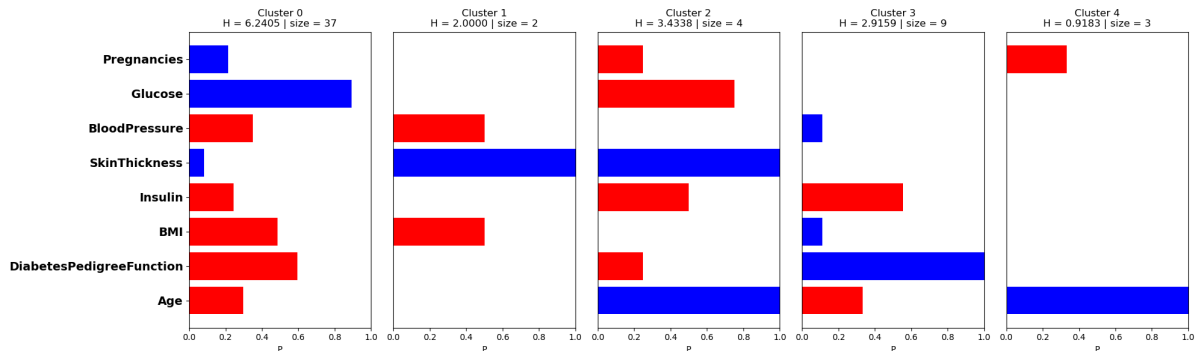


Figure 7: Diabete, explanation space – from class 1 to Class 0: histograms representing the prevalence p_n of each feature in each cluster in the explanation space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

What seems to emerge is that diabetes is more common in older people with familiarity, high level of glucose, previous pregnancies and thick triceps skin fold. BMI and high level of glucose appear randomly among the explanations in this direction.

5.2.4. Diabete, from class 0 to 1

In this direction, the counterfactuals highlight the key factors associated with transitions from being sane to having a Diabete, that is the most frequent variables found in counterfactuals from class "sane" to "diabete". While most features are non-actionable, for actionable variables the analysis allows to highlight the ones to keep under control, that again turned out to be coherent with medical literature, nonetheless the small sample size. In extreme synthesis, the most important feature among clusters results glucose, suggesting that controlling its level is a key protective factor for Diabetes.

Figure 8 allows to characterize each cluster in the original feature space. As can be read from entropy values, Cluster 1 has the minimum variability in terms of relevant features in the explanations, but all the features have such a variability that appear randomly distributed among instances. In cluster 2 the only variable that appears to be non randomly distributed has a borderline value of $Var_n = 0.16$ and a low prevalence $P_n = 0.2$. Cluster 3 appears to have the only really dominant feature in terms of prevalence and variability that is glucose level.

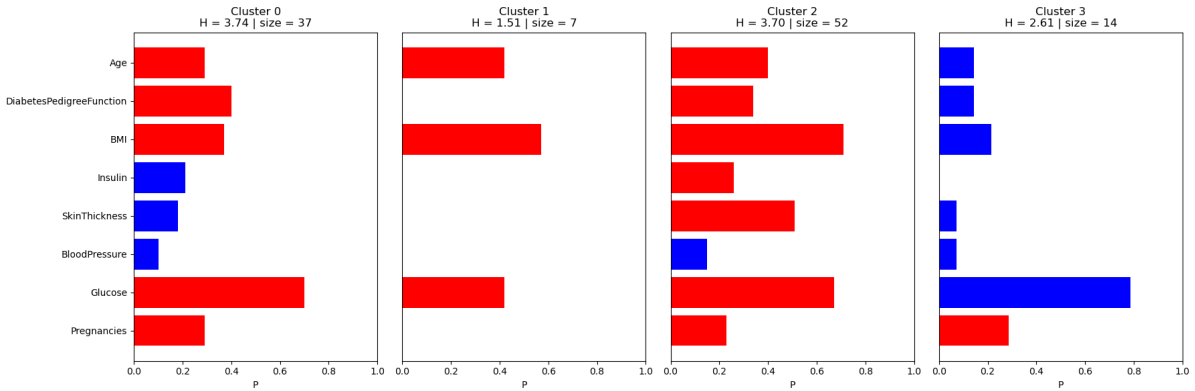


Figure 8: Diabete, original space – from class 0 to Class 1: histograms representing the prevalence p_n of each feature in each cluster in the original feature space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

Figure 9 shows three very small clusters (1,3 and 5) and two big clusters of equal size, with a different characterization: in Cluster 2 the dominant variables are BMI ad skin thickness, while Cluster 4 is dominated by glucose. Cluster 0 has an intermediate size and groups instances where BMI and Pregnancies are dominant.

What seems to emerge is that preventing diabete requires low glucose levels and that in the two subgroup of females with high BMI and previous pegnancies, and people with high BMI and thicker triceps skin fold, the risk is higher.

5.2.5. Credit Risk, from class 1 to 0

In this direction the counterfactuals highlight the factors that reduce the probability of loan default, that is the most frequent variables found in counterfactuals generated for people that most risk to fail their loan. In extreme synthesis, the failure risk is related to the income, the interest rate and the grade of the loan. Some variables like interest rate can be actioned by the loaner to reduce the risk of failure.

Figure 10 allows to characterize each cluster in the original feature space. Cluster 2, the largest, evinces heterogeneous changes primarily in person_income and loan_percent_income. This pattern is also observed in Cluster 3. In contrast, Clusters 0 and 1 are more stable, being dominated by person_income, loan_int_rate, and loan_grade_num. This observation indicates that these financial metrics consistently govern the transition to reduced default risk and potentially function as actionable mechanisms for borrowers aiming to enhance their credit performance.

Figure 11 allows to characterize each cluster in the explanation space. Cluster 0, the largest, is dominated by person_income, indicating a relatively stable and interpretable explanatory pattern. Cluster 3

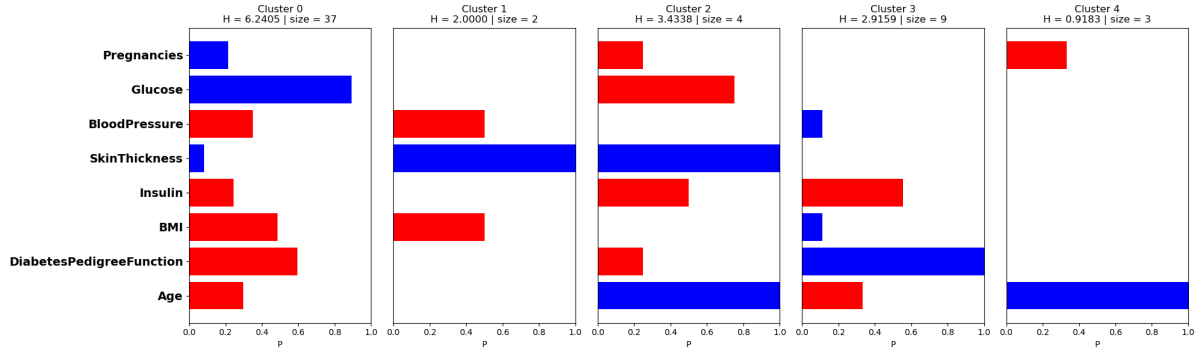


Figure 9: Diabete, explanation space – from class 0 to Class 1: histograms representing the prevalence p_n of each feature in each cluster in the explanation space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

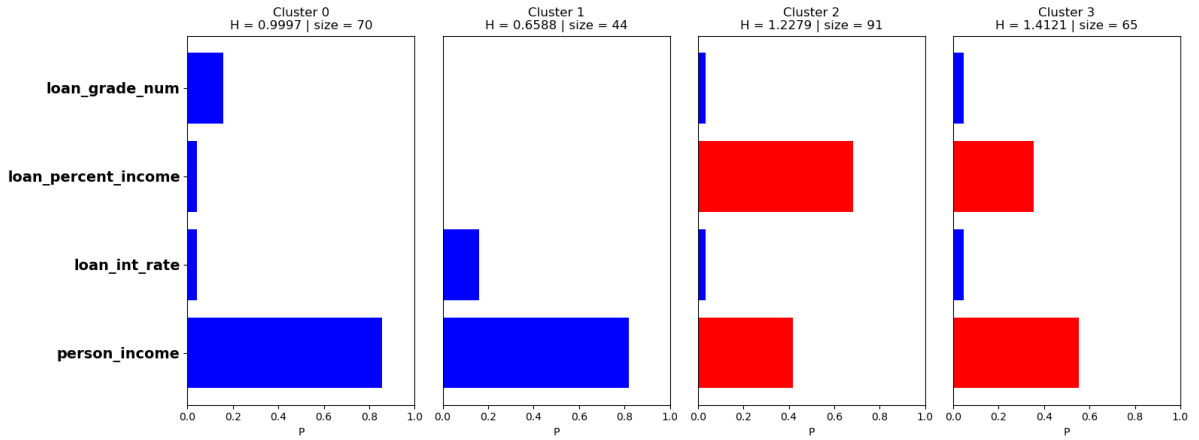


Figure 10: Credit Risk, original space – from class 1 to Class 0: histograms representing the prevalence p_n of each feature in each cluster in the original feature space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

also shows a focused and consistent explanation, being primarily driven by `loan_percent_income`. Clusters 1 and 2, smaller with 16 and 17 instances respectively, have higher entropy and heterogeneous feature contributions, particularly `loan_int_rate` in Cluster 1 and `loan_grade_num` in Cluster 2, indicating less stable and more diffuse explanatory patterns. Clusters 4 and 5, the smallest clusters with 12 and 5 instances, show low entropy, each dominated by a single feature (`loan_amnt` in Cluster 4 and `person_emp_length` in Cluster 5), reflecting highly focused but very specific explanations that may have limited generalization capacity.

5.2.6. Credit Risk, from class 0 to 1

In this direction, the counterfactuals highlight the factors that increase the probability of loan default, that is the most frequent variables found in counterfactuals generated for people that less risk to fail their loan. In extreme synthesis, the most important feature among clusters results income, suggesting that an high income is the best guarantee against the risk of failure. It must be noted that the people with lowest income are also the most likely to apply for a loan, so there is an intrinsic bias in the data.

Figure 12 allows to characterize each cluster in the original feature space. Clusters 1 and 2, the largest ones, are stably dominated by `person_income`, while other features appear mostly random, limiting interpretability. Clusters 0 and 3 show moderate entropy, with `person_income` again emerging as the main driver, supported by loan-related variables (`loan_int_rate`, `loan_amnt`). Cluster 4

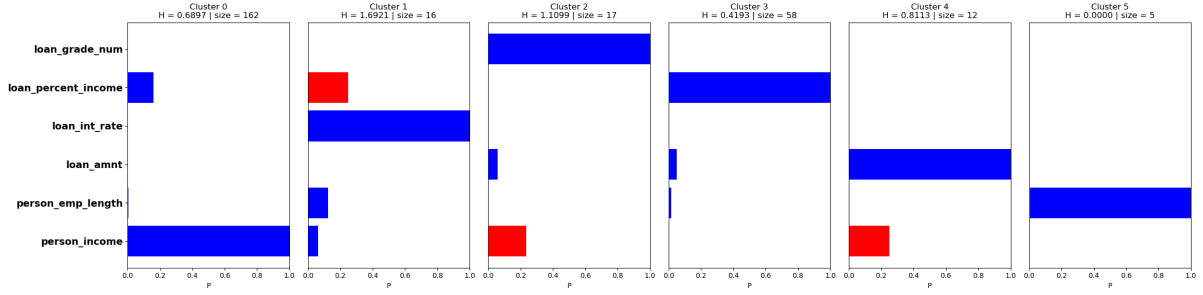


Figure 11: Credit Risk, explanation space – from class 1 to Class 0: histograms representing the prevalence p_n of each feature in each cluster in the original feature space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

highlights a stronger role of loan_percent_income, still in combination with person_income. Finally, Cluster 5, although very small, exhibits a highly stable structure around person_income and loan_percent_income.

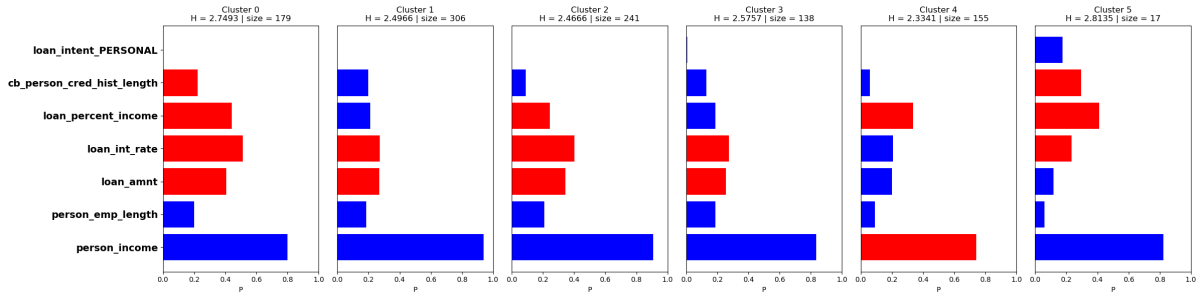


Figure 12: Credit Risk, original space – from class 0 to Class 1: histograms representing the prevalence p_n of each feature in each cluster in the original feature space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

Figure 13 allows to characterize each cluster in the explanation space. Cluster 1, the largest, shows the lowest entropy and is clearly dominated by person_income, yielding a stable and interpretable profile. Cluster 2 is also strongly characterized by person_income and loan_int_rate, with limited variability. Clusters 0 and 3, by contrast, exhibit high entropy values and a more dispersed pattern, with multiple features contributing randomly, which reduces interpretability. Cluster 4, despite its small size, presents a coherent structure dominated by person_income and loan_amnt. Finally, Cluster 5 highlights the joint role of person_income and cb_person_cred_hist_length, ensuring moderate stability.

6. Conclusions

Feature selection and counterfactual explanations act on a different scale. Feature selection on a global scale is sometimes too general, while counterfactual explanations are often too specific, being valid only for one instance. An intermediate-scale feature importance measure based on counterfactuals has been proposed in this paper: it allows to characterize "safe" regions, where the counterfactuals explanations and the feature involved are stable and with a limited Rashomon effect; at the same time it allows to recognize "unsafe" regions, where the counterfactuals explanations and the feature involved change randomly, so that very close points may end up with very different counterfactuals. Clustering has been used in both the original feature space and the explanation space, accounting for label directionality and more insights into the possible explanations an causal relationships. Results on real data are promising,

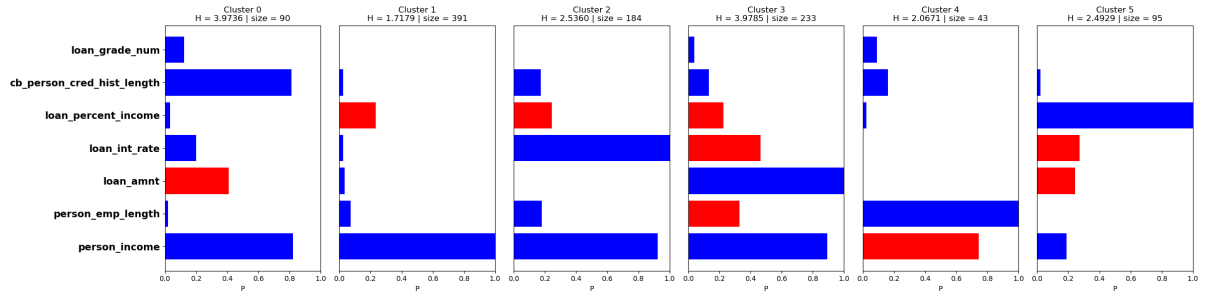


Figure 13: Credit Risk, explanation space – from class 0 to Class 1: histograms representing the prevalence p_n of each feature in each cluster in the original feature space. The entropy H is reported on top of each cluster. Only features with $p_n > 0.10$ in at least one cluster are shown. The red bars have $Var_n > 0.17$ and are considered randomly distributed in the cluster.

even on small datasets, and confirm the viability and effectiveness of the proposed method.

Acknowledgments

This work was supported by: the Digital Twin and Fintech services for sustainable supply chain (SmarTwin) project (Fondo per la Crescita Sostenibile – Accordi per l’innovazione di cui al D.M. 31 dicembre 2021e D.D. 18 marzo 2022 - CUP B69J23000500005) Ministero dello Sviluppo Economico (MISE); the context-AwaRe deCision-making for Autonomus unmmmaneD vehicles in mArine environmental monitoring (ARCAD-IA) project (PE00000013_1 - CUP E63C22002150007) cascade call of the Future Artificial Intelligence Research (FAIR) project Spoke 3 - Resilient AI, within the National Recovery and Resilience Plan (PNRR) of the Italian Ministry of University and Research (MUR).

Declaration on Generative AI

During the preparation of this work, the author(s) used Deep-L in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

- [1] A. Maratea, A. Ferone, Deep Neural Networks and Explainable Machine Learning, 2019, pp. 253–256. doi:10.1007/978-3-030-12544-8_23.
- [2] J. Burrel, How the machine ‘thinks’: Understanding opacity in machine learning algorithms, Big Data & Society 3 (2016). URL: <https://doi.org/10.1177/2053951715622512>. doi:10.1177/2053951715622512, original work published 2016.
- [3] E. Parliament, C. of the European Union, Regulation (eu) 2024/1689 of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>, accessed: 2025-04-01.
- [4] H. Chung, J. Frankle, Z. C. Lipton, F. Doshi-Velez, False sense of security in explainable artificial intelligence (xai), arXiv preprint (2024). URL: <https://arxiv.org/abs/2405.03820>. arXiv:2405.03820.
- [5] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, 2021. URL: <https://arxiv.org/abs/2102.13076>. arXiv:2102.13076.
- [6] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery 38 (2022) 1–55. doi:10.1007/s10618-022-00831-6.
- [7] A. Maratea, A. Ferone, Pitfalls of local explainability in complex black-box models, in: A. Ciaramella, C. Mencar, S. Montes, S. Rovetta (Eds.), Proceedings of WILF 2021, the 13th International Workshop

on Fuzzy Logic and Applications (WILF 2021), Vietri sul Mare, Italy, December 20-22, 2021, volume 3074 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: <https://ceur-ws.org/Vol-3074/paper13.pdf>.

- [8] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, *Harvard Journal of Law & Technology* 31 (2017) 841–887. URL: <https://arxiv.org/abs/1711.00399>.
- [9] A. Lucic, H. Oosterhuis, H. Haned, M. de Rijke, Focus: Flexible optimizable counterfactual explanations for tree ensembles, 2021. URL: <https://arxiv.org/abs/1911.12199>. arXiv:1911.12199.
- [10] R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization*, 2019, pp. 6276–6282. URL: <https://doi.org/10.24963/ijcai.2019/876>. doi:10.24963/ijcai.2019/876.
- [11] R. K. Mothilal, D. Mahajan, C. Tan, A. Sharma, Towards unifying feature attribution and counterfactual explanations: Different means to the same end, *CoRR abs/2011.04917* (2020). URL: <https://arxiv.org/abs/2011.04917>. arXiv:2011.04917.
- [12] C. Molnar, *Interpretable Machine Learning*, Lulu.com, 2020.
- [13] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.
- [14] A. L. Alfeo, A. G. Zippo, V. Catrambone, M. G. Cimino, N. Toschi, G. Valenza, From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks, *Computer Methods and Programs in Biomedicine* 236 (2023) 107550. URL: <https://www.sciencedirect.com/science/article/pii/S0169260723002158>. doi:10.1016/j.cmpb.2023.107550.
- [15] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* 16 (2002) 321–357.