, , 1–17
doi:

Fair and Faithful Explanation Selection in Credit Scoring

Nguyen Thanh Tung[1,*] **and Ly Linh Linh**[2]

Abstract

The integration of Explainable Artificial Intelligence (XAI) and fairness methods aims to improve both transparency and ethical accountability in AI systems. This report explores the concept of fair and faithful explanation selection, where the goal is to choose explanations that are not only accurate and faithful to the model but also fair to diverse groups. By examining current XAI techniques and fairness measures, this work highlights the challenges and proposes methods to reconcile these two goals for achieving reliable, unbiased, and interpretable AI.

Key words: Explainable AI, Faithfulness, Fairness

1. Introduction

(ML) Machine learning models are becoming increasingly complex, necessitating explanations that are accessible and understandable to users. From the literature it is evident that machine learning and its application is quickly evolving [1] (Lessmann et al., 2015). However, many AI systems struggle to provide clear explanations for their autonomous decisions and actions, creating a gap in transparency and understanding for human users [2]. While explanations may not be critical for certain applications—such as those involving routine or non-critical tasks—there is a growing debate among AI researchers about the importance of explainability. Some argue that prioritizing explanations is overly ambitious, technically challenging [3] and, in some cases, unnecessary. However, for high-stakes domains such as defense, healthcare, finance, and legal systems, the ability to provide meaningful and accurate explanations is indispensable. In these fields, explanations not only enhance user trust and comprehension but also play a pivotal role in ensuring accountability, ethical compliance, and informed decision-making. In this paper we will focus on credit scoring, understanding the rationale behind an AI's decision is crucial for addressing fairness concerns, meeting regulatory requirements, and empowering consumers. This highlights the need for AI systems to be both transparent and interpretable [4], particularly in applications where their decisions have significant societal impacts.

The management of risks, and especially the management of credit risk is one of the core challenges of financial institutions [5]. Therefore, credit scoring has been introduced in financial decision-making processes, enabling lenders to assess the creditworthiness of individuals and businesses effectively. These models leverage diverse data sources to predict the likelihood of repayment or default, supporting decisions such as loan approvals, interest rate determination, and credit limit assignments. Traditional credit scoring systems rely heavily on statistical models, but with advancements in machine learning (ML), credit scoring has evolved to incorporate complex algorithms capable of processing large-scale, high-dimensional datasets. However, the opacity of these advanced models has raised concerns about trust [6], accountability, and fairness in their application, several reports [7], [8] indicate future regulatory guidelines for this field. Explainable Artificial Intelligence (XAI) has emerged as a critical field for addressing these concerns. XAI techniques aim to make ML models more interpretable, providing insights into how decisions are made.

1.1. XAI in credit scoring

The goal of an explainable AI (XAI) system is to make its operations more understandable to humans by offering clear and meaningful explanations [9]. To achieve this, several general principles can guide the design of effective, human-centered AI systems. An XAI system should be capable of explaining its knowledge, actions, and future intentions. It should also disclose the key information influencing its decisions. However, the nature and quality of explanations depend heavily on the specific context, including the task at hand, the user's needs, and their expectations from the AI system [2].

Interpretability and explainability are inherently domain-specific concepts [10], and their definitions cannot be fully separated from the field of application. Explanations can be categorized as full or partial. Fully interpretable models provide complete transparency [11], offering detailed insight into every aspect of their reasoning process [12]. In contrast, partially interpretable models reveal critical components of their logic while maintaining some level of complexity. Interpretable models adhere to domain-specific "interpretability constraints," such as ensuring monotonicity with certain variables or maintaining expected relationships between correlated variables [13]. Black-box models, however, are not bound by these constraints and may lack transparency.

© The Author . Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Partial explanations [14] often involve tools like variable importance measures, local approximations of global models at specific points, and saliency maps, which highlight the most influential features in a decision. These elements provide users with a glimpse into the reasoning process of complex models [15], bridging the gap between full transparency and complete opacity. In the context of credit scoring, XAI allows stakeholders—such as regulators, institutions, and consumers—to understand and trust the model's outputs. Current applications of XAI in credit scoring focus on ensuring interpretability for compliance and customer transparency. However, the explanations provided often face challenges regarding their faithfulness, which refers to their fidelity to the underlying decision-making process of the model, and fairness, ensuring that the model does not perpetuate biases that disadvantage specific groups. With its latest discussion paper, EBA (2021) uncovered three main challenges related to the complexity of AI models: (i). The challenge of interpreting the results, (ii) the challenge of ensuring that management functions properly understand the models, and (iii) the challenge of justifying the results to supervisors (EBA 2021). The European Union, European Union, Parliament and Council (2016) and the European Commission (2021) is also working on AI-specific regulations. As a result, there is a need to utilize XAI models, uncovering how each variable or feature influences the prediction of credit default.

Among the emerging techniques, two frameworks have been widely recognized as the state-of-the-art in machine learning explainability and those are:

- the LIME framework, introduced by Ribeiro et al. in 2016 ([16])
- SHAPvalues, introduced by Lundberg et al. in 2017 ([17]).

One promising framework that might achieve this is the Shapley value (SHAP) framework, which has been applied successfully in other areas, such as disease detection [18] and surgery technique selection.

Jing Zhou from the Renmin University of China proposed a method for creating features for credit scoring focusing on the frequency, recentness and monetary value of the account information, using ML models [19]. For historical transaction data, new time-series approaches based on Recurrent Neural Networks and LSTM are showing outstanding results, even though they are still in a research phase [14]. On The other hand, more classical approaches such as Logistic Regression, Decision Trees and GradientBoosting Classifiers are also implemented for more complex data sets [15].

There may be inherent conflict between ML performance (e.g., predictive accuracy) and explainability. Often, the highest performing methods (e.g., DL) are the least explainable, and the most explainable (e.g., decision trees) are the least accurate.

Although the ML methods such as KNN and SVM mentioned above have achieved good results in CVD prediction, they belong to a single global optimization model, which has limited performance, poor robustness and poor fault tolerance. Compared with a specific classifier, the classifier based on ensemble learning has better advantages in complex classification tasks. Therefore, in the prediction of CVD risk, CatBoost has gradually gained the favor of scholars because of its high efficiency and reasonable processing of category features, strong generalization ability and high accuracy. Li, Zhang, Xiong, Hu, Liu, Tu, and Yao (2022) [22] used CatBoost to predict hospital mortality in mechanically ventilated patients with congestive heart failure (CHF). However, CatBoost model is not highly explanatory, its different parameters combinations may have different prediction results. For this reason, some scholars have started to try to optimize CatBoost using swarm intelligence algorithms. Zhang, Chen, Zhang, Liu, Yu, Zhang, and Gao (2021) [23] adopted difference-mutation brain storm optimization (DBSO) algorithm to optimize the CatBoost model and got better prediction results.

Since many attribution methods compute scores by approximating the prediction model locally by an interpretable model [24], faithfulness is often considered as the accuracy of the approximation in a neighborhood of the input [25].

1.2. Fairness in credit scoring

Fairness is a particularly pressing issue in credit scoring due to its societal impact. Biased models can disproportionately disadvantage underrepresented groups, leading to unequal access to financial services. Although fairness-aware approaches have been incorporated into ML pipelines, balancing fairness with accuracy and interpretability remains an ongoing challenge..

With the rise of explainability as a key requirement in AI, concerns about the fairness of explanations across different demographic groups have emerged. Fair explanations ensure that all groups receive equitable insights from model outputs, while faithful explanations ensure that explanations accurately reflect the true behavior of the model. Faithful explanations, which truly reflect a model's decision-making logic, are essential to maintain trustworthiness, but these explanations can inadvertently reveal or amplify biases within the model. This duality underscores the need for robust methodologies that consider both fairness and faithfulness in explanation generation

1.3. Our contribution

The motivation for this research arises from the limitations of existing methods in achieving robust fairness and faithfulness in credit scoring explanations. While progress has been made in both XAI and fairness research, there is a lack of comprehensive frameworks that integrate these dimensions in a balanced and practical manner. This gap has significant implications for the adoption and reliability of credit scoring models, as biased or unfaithful explanations can undermine the credibility of even the most accurate predictions.

This report investigates how fairness and faithfulness can be balanced to improve the overall quality of AI explanations. The significance of this research lies in its aim to bridge this critical gap. By proposing a framework that emphasizes both fairness and faithfulness in XAI for credit scoring, we seek to enhance the trustworthiness, accountability, and societal impact of these models. This study contributes to the fields of XAI and credit scoring by developing methodologies that not only explain model decisions but do so in ways that are equitable and reflective of true model behavior.

The primary contributions of this research are as follows:

1. A critical review of existing XAI methods in credit scoring with respect to fairness and faithfulness.

2. A novel framework for generating fair and faithful explanations tailored for credit scoring applications. 3. An empirical evaluation demonstrating the robustness and practical applicability of the proposed framework on real-world datasets.

4. Guidelines and insights for integrating fairness and faithfulness into the development of XAI systems in financial applications.

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on fairness and faithfulness in ML. Section 4 introduces the proposed framework and its methodology. Section 5 describes the experimental setup. Section 6 presents the results and discusses the findings and their implications. Finally, Section 7 concludes the paper with future research directions.

2. Literature Review

Back in 1981, [26] stated that the ability to explain decisions is the most highly desirable feature of a decision-assisting system. Since AI and ML become widespread especially in the real life application, and algorithmic decision, attention has shifted back from accuracy to explainability. Explanations of a classification system can take many forms, but should accurately reflect the classifier's inner workings [27].

How important each method is to evaluate explanations has been discussed in ([28]; [29]; [30]; [31]). There are various attempts to measure different aspects of an explanation: usefulness to humans ([32]; [33]; [34]); complexity [35]; difficulty of answering queries [36]; and robustness ([37]. [38]), and faithfulness [39].

There are many types of explanations [40]. Sanjoy Dasgupta separates these explanations into two groups: intrinsic explanations [41] and post-hoc explanations [42]. The issue of explainability is an open research question for some of the most successful (in terms of accuracy) forms of ML such as SVMs, DL, and many of the ANNs [43] .

In [44] research, they state that feature extraction (FE) and feature selection (FS) are the two most effective methods for reducing feature dimensionality. Popular FE techniques include Principal Component Analysis (PCA) [45], Linear Discriminant Analysis (LDA) [47], Independent Component Analysis (ICA) [47], and ISO-Container Projection (ISOCP) [48]. [49] proposed a sensitivity analysis based model, which analyses how much each feature contributes to the model's predictions by finding the difference between the prediction and expected prediction when the feature is ignored. Such explanations are given in the form of feature contributions.

SHapley Additive exPlanation (SHAP) [50] is a game-theoretic approach to explain ML predictions. SHAP seeks to deduce the amount each feature contributed to a decision by representing the features as players in a coalition game. The payoff of the game is an additive measure of importance, the so-called Shapley value, which represents the weighted average contribution of a particular feature within every possible combination of features. As such, local and global interpretations of a model are consistent and the average prediction is fairly distributed across all Shapley values, meaning that contrasting comparisons between explanations are possible. However, if the model is not additive then interpretation of the Shapley values is not always transparent, as predictive models may have non independent pay-off splits. Furthermore, while SHAP can be considered model agnostic, optimized implementations of the SHAP algorithm to all model types is not immediately straightforward or efficient.

Class activation maps (CAMs) are specific to CNNs. CAMs represent the per-class weighted linear sum of visual patterns present at various spatial locations in an image [51]. More formally, global average pooling is applied to the final convolutional feature map in a network, before the output layer. These pooled feature maps are then used as the input features to a fully connected layer and output through a loss function. By projecting the weights of the output back to the previous convolutional layer, the areas in the input image with greater influence over the CNNs' decision are highlighted per-class and visible through a heatmap representation. CAMs cannot be applied to pre-trained networks and networks that do not adhere to the specified fully convolutional network architecture.

Local interpretable model-agnostic explanations (LIME) [52] is a model-agnostic technique to create locally optimized explanations of ML models. LIME trains an interpretable surrogate model to learn the local behavior of a global "black box" model's predictions. For image classification, an input image is divided into patches of contiguous superpixels (i.e., an image object) and a weighted local model is then trained on a new set of permuted instances of the original image (i.e., some superpixels are turned to gray). The intuition is then that by changing aspects of the input data that are human understandable (spatial objects) and learning the differences between those perturbations and the original observations, one can learn what about the input contributed to each class score. However, these explanations are not always informative or reliable at a human level if the parameters that control the perturbations are chosen based solely on heuristics.

All these XAI techniques present their ability in making ML/AI models more interpretable. The last three mentioned techniques are credit scoring models that provide good classification performance as well as local and/or global explainability. The winner of FICO's Explainable Machine Learning Challenge in 2018, BRCG [53], is considered as a state-of-the-art of XAI in credit scoring and is therefore selected as the benchmark paper for this work.

In the context of attribution methods, faithfulness is often understood as the accuracy of a local approximation when predicting the output of a prediction model [54, 16, 24]. As we already discussed in the introduction, this can be misleading because a high accuracy does not guarantee that the explainer picked up the actual behaviour of the prediction model. Another problem with this definition is that it depends on the definition of a neighbourhood and a sampling strategy from which the test set is generated.

Recent attribution methods like LIME and MAPLE [16, 24] are partially based on the idea of approximating a black-box model locally by a linear model in order to use the coefficients of the linear models as feature scores. Let us note that in a setting with continuous features, the gradient is actually an analytic linear approximation of the classifier, providing that the classifier is differentiable. While there are good reasons to replace the gradient when the classifier is non-differentiable or features are discrete, it seems wasteful not to use it for differentiable classifiers over continuous features. In the latter setting, alternative attribution methods may be just a poor substitute for the gradient, not only giving inaccurate explanations, but also unnecessarily difficult to compute. For example, [55] showed that when the prediction model to be explained by LIME is linear, the expected coefficients of the approximating linear model are proportional to the partial derivatives of the prediction model. Since the partial derivatives exactly capture the behaviour of a linear function, this shows that LIME is to some extent faithful to linear models. However, the authors also found that the expected error of the linear approximation is bounded away from zero. Furthermore, as the approximation has to be computed based on perturbations of the input, the scores are noisy (as they depend on the sampled neighbours) and, compared to the gradient, relatively expensive to compute. More recently, [56] showed that some natural configurations of LIME converge to the same scores in expectation as a smoothed version of the gradient. While this can be desirable in the discrete setting, the original gradient seems to be a more accurate and more efficient explanation of the classifier's true behaviour in the continuous setting.

Research interest in the fairness of machine learning has increased over the last decade, with several fairness measures arising in the literature. As Barocas et al. [57] note, different measures are based on various intuitions regarding fairness. The following represent examples of the numerous attempts of defining a comprehensive fairness metric: individual fairness- advocates for similar treatment of similar cases [58]; statistical parity- considers an individual belonging to a protected group to bear the same risk score as all the other members[9]; accuracy fairness (or accuracy equity)- implies different treatments for different cases, assuming a perfect classifier is also fair [59]; threshold fairness- considers the same decision threshold should be applied for each group (protected or unprotected) [60]; and calibration- conditions the estimates to have the same effectiveness for individuals [12]. The majority of viewpoints concern equal treatment in general across different groups, as defined by protected attributes (also known as sensitive attributes), such as gender, nationality, and race (statistical parity), or, more specifically, equal treatment based on some constraints, such as predictive parity and threshold fairness.

Generally, the literature sources agree that not all fairness criteria can be used simultaneously to evaluate the fairness of an ML process [62], [63], [64], as some of them contradict others. As a result, while we review all of the criteria suggested by the literature in this field from a credit scoring standpoint, we do not expect to find methods that can satisfy all fairness criteria simultaneously.

The topic of Fair and Faithful Explanation Selection in Credit Scoring addresses a critical and underexplored area in the intersection of machine learning, ethics, and finance. While credit scoring models have been widely studied, existing literature often focuses on improving predictive accuracy or mitigating general bias. However, the dual challenges of ensuring fairness and maintaining faithful model interpretability remain inadequately addressed.

Most prior works on explainable artificial intelligence (XAI) in credit scoring primarily emphasize the development of interpretability frameworks without rigorously evaluating their fairness or their fidelity to the underlying model. This creates a significant research gap, as deploying explanations that are either unfair or misaligned with the model can exacerbate existing biases, undermine trust, and result in suboptimal decision-making. Moreover, the absence of holistic approaches that simultaneously address these dual objectives limits the practical utility and ethical deployment of credit scoring systems.

By focusing on the selection of fair and faithful explanations, this research provides a novel contribution to both academic discourse and practical applications in credit risk management. This topic is particularly timely, given the increasing scrutiny of AI systems for ethical and legal compliance, especially in high-stakes domains like finance. To the best of our knowledge, no comprehensive study to date has systematically addressed this dual challenge within the context of credit scoring, highlighting the unique contribution and potential impact of this work.

3. Methodology

In this section we will describe the methods used for this research, the reasoning or motivation for the choice of methods and the metrics employed. We will first introduce the baseline model that we will use as a current standard which is Catboost. Then we will elaborate on the method for inducing that multiple XAI techniques (such as LIME, SHAP, and counterfactual explanations) are applied to generate explanations for its predictions. After evaluating explanations and assessing faithfulness, we finally conduct fairness checks on the most faithful explanations.

3.1. Catboost

Ensemble methods combine several learners to obtain better predictive performance than a single constituent learning algorithm [65]. In this paper, we use a modification combination of a standard gradient boosting algorithm, which can avoid target leakage, and a new algorithm for processing categorical features, which is CatBoost (for "Categorical Boosting"). [66] proposed this algorithm in 2017 and compared it with the main two power-boosting algorithms, XGBoost and LightGBM, all of which are the same Gradient Boosting Decision Tree(GBDT) open-source learning algorithm. CatBoost is an implementation of gradient boosting, which uses binary decision trees as base predictors. CatBoost has two new methods that support its state-of-the art performance, the first is the algorithm to process classification features, and another is sort lifting algorithm [67].

CatBoost manages categorical values by utilizing category-based statistics. It assumes that the algorithm is better suited for encoding categorical features than manual efforts [68]. To achieve this, it transforms categorical features into numerical ones by leveraging the frequency of each category's occurrence in the dataset.

CatBoost employs oblivious decision trees, which differ from traditional decision trees by requiring that all nodes at the same depth use the same variable for splitting. This design helps reduce overfitting, making the model more robust to parameter changes [69]. Additionally, oblivious trees are highly parallelizable, enabling efficient GPU-based training and significantly speeding up the process of tuning the model.

The difference between CatBoost and Gradient Boosting algorithm is that Gradient Boosting estimates the gradient for all possible individuals in the leaf, which can lead to a bias. To overcome the overfitting problem, CatBoost computes the gradient for each individual separately [70]. Then, it trains the logarithm of the n models, which are trained at the same time. As a result, this will work well with small datasets as it is computationally expensive.

3.2. Explanability

Explanations of a classification system can have many forms. A common situation is where the predictive model is not understandable, either at a global or local level, and so a separate post-hoc explanation is needed. Over the past few years, many strategies for post-hoc explanation have emerged, such as LIME [25], Anchors [72], and SHAP [73].

In the Credit Scoring Field, this topic is relatively new and particularly important, given the strict regulation on the topic especially the European Community such as European General Data Protection Regulation (GDPR) and Ethics guidelines for trustworthy [74].

3.2.1. LIME

3.2.1.1. LIME Framework

Locally Interpretable Model-Agnostic Explanations (LIME) is a post-hoc, model-agnostic explanation technique designed to approximate any black-box machine learning model with a local, interpretable model to explain individual predictions [25].

LIME [25] provides an explanation of $f(x)$ by:

1. Using an interpretable representation $\psi : X \to X'$, such as the presence or absence of specific words in a document.
2. Approximating f near x with a simple model $g_x : X' \to Y$, typically a linear classifier.

However, LIME does not exhibit perfect consistency. For instance, points x with the same interpretable representation X' are assigned the same g_x, even though their predicted labels may differ.

To explain the output at a single data point, LIME perturbs this point to create several points in its local neighborhood and assigns weights to these perturbed points based on their distance to the original point. A linear model is then trained on the weighted perturbed points to provide an explanation, where the weights of the linear model correspond to the explanation for the original point.

In our experiments, we used the LIME implementation available at https://github.com/marcotcr/lime and applied the `LIMETabularExplainer`.

3.2.1.2. Kernel Width

In the LIME implementation, the default kernel width is defined as $0.75 \cdot \sqrt{n}$, which can be interpreted as 75% of the maximum value determined by Euclidean distances in an n-dimensional space. Given appropriate parameter settings, any function can be represented using this type of kernel [75]. The kernel width parameter controls the linearity of the induced model: the larger the width, the more linear the function becomes [76].

Following the same principle, we define a custom kernel width for the Manhattan distance as $0.75 \cdot n$. Similarly, a kernel width comparable to the performance of the Euclidean metric **?** (using the custom definition) for each L_k distance is given by:

$$\text{kernel width} = 0.75 \cdot n.$$

3.2.2. SHAP

SHAP is a method similar to LIME [73]. Since the Shapley value is a discrete concept, SHAP heavily relies on sampling when applied to continuous features. It uses a Boolean feature space X' and provides explanations as linear functions $g_x : X' \to Y$. However, in SHAP, the choice of g_x is inspired by Shapley values [78] from game theory. Specifically, the coefficients of g_x are guaranteed to sum to $f(x) - \phi_0$, where ϕ_0 is a constant for all x. A significant advantage of SHAP is its ability to reflect the influence of features in each sample while indicating the positive and negative effects of these influences.

Another method utilized in this study is TreeSHAP [79], which is designed specifically for tree-based machine learning models, such as LightGBM. TreeSHAP allows for the efficient computation of Shapley values for ensemble trees by propagating all subsets through each tree simultaneously, tracking their aggregate weights and subset counts, and achieving polynomial-time complexity [80]. Additionally, due to the additive property of Shapley values, the Shapley values of an ensemble tree model can be obtained efficiently.

For our experiments, we used the SHAP implementation available at https://github.com/slundberg/shap and reported results using the `KernelExplainer`.

3.2.3. Anchor

Introduced by [81], Anchors focus on generating high-precision rules, called anchors, which locally explain the behavior of a model for a specific prediction. These anchors serve as conditions or patterns in the input data that, when satisfied, guarantee the model will produce the same prediction with high confidence [9]. The Anchors method generates explanations in the form of decision rules expressed as IF-THEN statements, which define specific regions within the feature space. These rules, known as anchors, "anchor" predictions to the same class as the data point being explained, regardless of changes to feature values not included in the anchor [82].

High-quality anchors are characterized by two key properties: precision and coverage [83]. Precision measures the proportion of data points within the anchor's region that share the same class as the data point being explained. High precision ensures the reliability of the rule. Coverage quantifies the extent of the feature space encompassed by the anchor, with broader coverage indicating a more general and widely applicable rule.

This systematic approach allows Anchors to generate interpretable, high-precision rules that provide actionable insights into the behavior of complex machine learning models.

3.3. Fairness metrics

Another pre-processing technique is the *disparate impact remover*, proposed by [84]. The intuition behind this processor is to ensure independence by prohibiting the possibility of predicting the sensitive attribute x_a using the other features in X and the outcome y. This is achieved by transforming X into X', while preserving the rank of X within sensitive groups defined by x_a. By preserving the rank of X given x_a, the classification model $f(X')$ can still learn to choose higher-ranked credit applications over lower-ranked ones based on the other features.

The transformation is performed using an interpolation based on a quantile function and the cumulative distribution $F : P(X|x_a = a)$. This ensures that given the transformed X' at some rank, the probability of drawing an observation given $x_a = a$ is the same as for the entire dataset. Hence, x_a cannot be predicted with the other attributes, fulfilling the independence criterion.

Since ensuring perfect independence can have a strong negative impact on classifier utility, the transformation can be modified to only partially remove disparate impact. The meta-parameter $\lambda \in [0, 1]$ allows controlling the desired level of fairness-utility trade-off during transformation.

Statistical Parity Difference (SPD): This metric measures the difference between the probabilities of acceptance in the protected and unprotected groups. A value close to zero implies the same acceptance rate for both groups [85]. The fairness range for this metric is considered to be within the interval $(-0.1, 0.1)$:

$$\text{SPD} = \Pr(\hat{Y} = 1 \mid D = \text{unprivileged}) - \Pr(\hat{Y} = 1 \mid D = \text{privileged})$$

Disparate Impact (DI): Based on the principle of independence, this metric calculates the ratio between the probability of acceptance for unprivileged and privileged groups. A value close to 1 implies an ideal degree of fairness, while values lower than 1 indicate an advantage for the privileged group, and values higher than 1 indicate an advantage for the unprivileged group [84]. In a more flexible approach, the interval $(0.8, 1.25)$ is considered acceptable for a classifier to be fair:

$$\text{DI} = \frac{\Pr(\hat{Y} = 1 \mid D = \text{unprivileged})}{\Pr(\hat{Y} = 1 \mid D = \text{privileged})}$$

4. Experiment

To select explanations that are both fair and faithful and evaluate each method described above, we used the well-known German credit dataset available from the UCI Machine Learning Repository.

4.1. Data

We used the German credit dataset, which is one of the most popular datasets used for benchmarking in the field and has also been included in other research on faithfull AI. This dataset consists of samples of loan applications, with the target attribute being the outcome of the loan from each application. We considered the gender of the applicant as a protected attribute, as suggested in this study. There are 1000 consumer loans extended to 310 women and 690 men. For each client, we are being provided with their credit-risk type: good type or low risk (Y=1) and bad type or high risk (Y=0). In total, 300 borrowers are in default (Y = 0) among which 191 are men and 109 women. Moreover, there are 19 explanatory variables measuring socio demographic attributes, including information about age, education, profession, and customer history providing information regarding the relationship between the customer and the bank (e.g., other products owned, previous loans); and economic information such as client's income or loan amount.

In the initial database, the gender and the marital status of the applicants are specified in a common attribute with five categorical values (single male, married male, divorced male, single female, married or divorced female). Here, as we focus on gender discrimination whatever the marital status, we consider a binary variable representing the single, married, or divorced females (protected group) versus the single, married, or divorced males (unprotected group).

More information about the database can be found in Tables A1 and A2. We contrast in Figure A1 the distributions of each of the 20 variables for men and women. We see that the default rate is higher for women (35.16

4.2. Data Preprocessing

We place ourselves as a bank that seeks to assess the credit worthiness of our loan applications through the development of choosing the right explanations to explain the work of each model.

First, we transform the GoodCustomer column, indicating customer creditworthiness, to the target variable, then we remove the original one from the feature set. Extracting the target variable ensures that the rest of the dataset can be treated as input features, avoiding data leakage and maintaining a clear distinction between predictors and outcomes. Second, the binary values in this column are mapped to 1 for positive outcomes (good customers) and 0 for negative outcomes (bad customers). This step helps to create a uniform representation, which can simplify evaluation metrics (e.g., accuracy, precision, recall) and ensure consistency across different tools and algorithms.

The pre-processing of categorical attributes was finalized with one-hot encoding for the transformation into numerical values, a condition for being able to run all algorithms in the benchmarking phase, Each unique purpose is transformed into a separate binary column including in the PurposeOfLoan and Gender columns. Specifically, a binary matrix is constructed where each row corresponds to a record, and each column represents a unique loan purpose. This matrix is appended to the feature set, and the original PurposeOfLoan column is dropped. Also, the Gender column, which contains string values (Male and Female), is encoded into a binary variable where Male is mapped to 1 and Female to 0.

In simulating the model, 80% of the data is used as the training set of the training model, and the remaining data is used as the test set to evaluate the performance of the model, considering a stratification that assigns the instances to each set based on the target variable distribution. The processed dataset is represented as a pandas DataFrame, while the target variable is converted into a NumPy array. Categorical feature indices are identified and stored for model training, accounting for the newly encoded features. The preprocessing pipeline outputs a dictionary containing the processed feature set, the target variable, a list of column names, and the indices of categorical features. This ensures the dataset is ready for experiments while maintaining compatibility with algorithms sensitive to feature types.

4.3. Machine Learning model

To achieve an interpretable model, it is necessary to reduce the number of features. Therefore, dimensionality reduction is performed using SHAP values to identify the most significant features. Since feature importance cannot be determined until a model is trained, CatBoost is selected as the benchmark algorithm to ensure consistency with model selection. The model parameters, such as learning rate and tree depth, are left at default, ensuring general applicability.

We run 5 models namely: The classical Logistic Regression is compared with random Forest, Gradient Boosting, CatBoost and the Neural Network models. Figure 1 shows how CatBoost is the best in class, Gradient Boosting with Logistic Regression and Neural Networks following close while Random Forest has the worst performance in terms of AUC.

4.4. Faithfulness and Fairness evaluation

The experiment was conducted to identify the most fair and faithful explanation method for a credit scoring model, leveraging a systematic approach combining interpretability, fairness, and faithfulness metrics. The German credit dataset was utilized for training, preprocessed by ensuring alignment between feature names and dataset columns while excluding irrelevant features such as the target variable and indices. A pre-trained gradient boosting model served as the predictive system, with its output probabilities wrapped in a callable function for subsequent

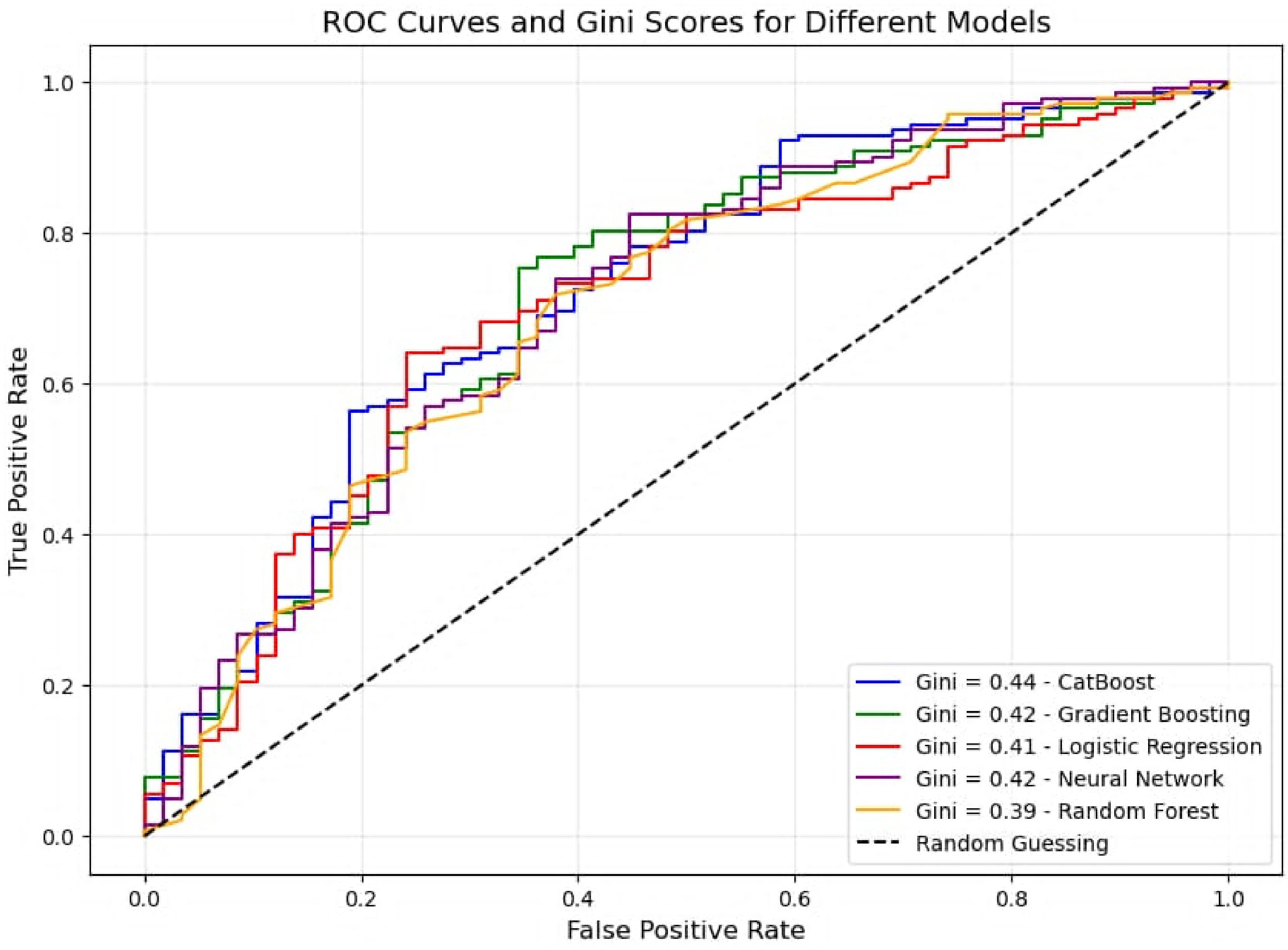


Fig. 1: ROC Curves and Gini Scores for Diferent Models

explanations. To generate explanations, an Explainer class was instantiated using the dataset, predictive model, feature names, and an indicator for discrete features. For a test instance, multiple explanation methods—such as Anchor, LIME, and SHAP—were applied to generate explanations. These were evaluated on their faithfulness, scored based on how well the explanations reflected the underlying model's behavior. The explanations were ranked, and the most faithful explanation was identified, highlighting the feature importances contributing to the decision. The fairness of the explanations was assessed using a disparate impact metric, with gender as the sensitive attribute. A Disparate Impact Calculator mapped gender attributes to binary values (Male: 1, Female: 0) and computed disparate impact ratios for each feature. Features with ratios below a threshold of 0.8 were flagged as potentially biased, offering insights into sources of unfairness within the explanations. Finally, the combined fairness and faithfulness of the explanations were evaluated using the FairFaithfulSelector. This module integrated the sorted explanations and fairness metrics to identify the optimal explanation, meeting predefined criteria for both aspects. If no explanation satisfied both conditions, the selection was rejected. Through this methodology, the experiment ensured a rigorous and comprehensive evaluation of interpretability methods in the context of credit scoring.

5. Results and Discussion

5.1. Faithfulness Score

After computing the explanations, we evaluate the faithfulness of each explanation generated by LIME with varying kernel widths, SHAP, and Anchors with varying kernel widths. Faithfulness is a critical metric in assessing the quality of these explanations, as it quantifies the extent to which an explanation aligns with the underlying model's behavior when making predictions in the local region around the data point of interest.

In this context, faithfulness is defined as the degree to which the explanation accurately reflects the model's decision-making process for perturbed instances in the neighborhood of the original input. This metric is essential for understanding whether the explanation truly represents the predictive logic of the model or if it is merely a plausible, yet inaccurate, narrative. A higher faithfulness score indicates that the explanation provides an accurate representation of the model's behavior, ensuring reliability and trustworthiness when interpreting the predictions.

Rule	Explanation
LoanDuration > 24.00	The loan term is longer than 24 months.
YearsAtCurrentJobGreaterOrEqualThan4 <= 0.00	The applicant has been at their current job for less than 4 years.
CriticalAccountOrLoansElsewhere <= 0.00	The applicant does not have critical accounts or loans elsewhere.
LoanRateAsPercentOfIncome > 3.00	The loan repayment exceeds 3% of the applicant's income.
loanpurposeFurniture > 0.00	The loan's purpose is related to furniture.
loanpurposeUsedCar <= 0.00	The loan is not for a used car.
loanpurposeElectronics <= 0.00	The loan is not for electronics.
LoanAmount > 3941.50	The loan amount is greater than 3941.50 units.
SavingsAccountBalanceGreaterOrEqualThan200 <= 0.00	The applicant's savings account balance is less than 200 units.
HasGuarantor <= 0.00	The applicant does not have a guarantor for the loan.

Table 1. Anchor Rules for Explanation

Conversely, a lower faithfulness score suggests a divergence between the explanation and the actual functioning of the model, which may lead to incorrect conclusions about the model's decision-making process.

Our experimental results provide a comparative ranking of faithfulness scores across the different methods and parameter configurations on the first instance. These findings reveal key insights into the performance of each explanation method, shedding light on how variations in kernel width influence the quality of explanations for LIME, SHAP, and Anchors.

The ranking of faithfulness scores is as follows:

- **Anchors**: 0.4690
- **LIME (kernel width 0.25)**: 0.4667
- **LIME (kernel width 0.75)**: 0.4252
- **LIME (kernel width 0.5)**: 0.4211
- **SHAP**: 0.4159

Anchors, as implemented in the *alibi* library, provide text-based explanations without built-in visualization tools such as SHAP or LIME. The explanation for the given instance is as follows.

Anchors, as implemented in the *alibi* library, provide text-based explanations without built-in visualization tools such as SHAP or LIME. The explanation for the given instance is summarized in Table 1:

Metrics:

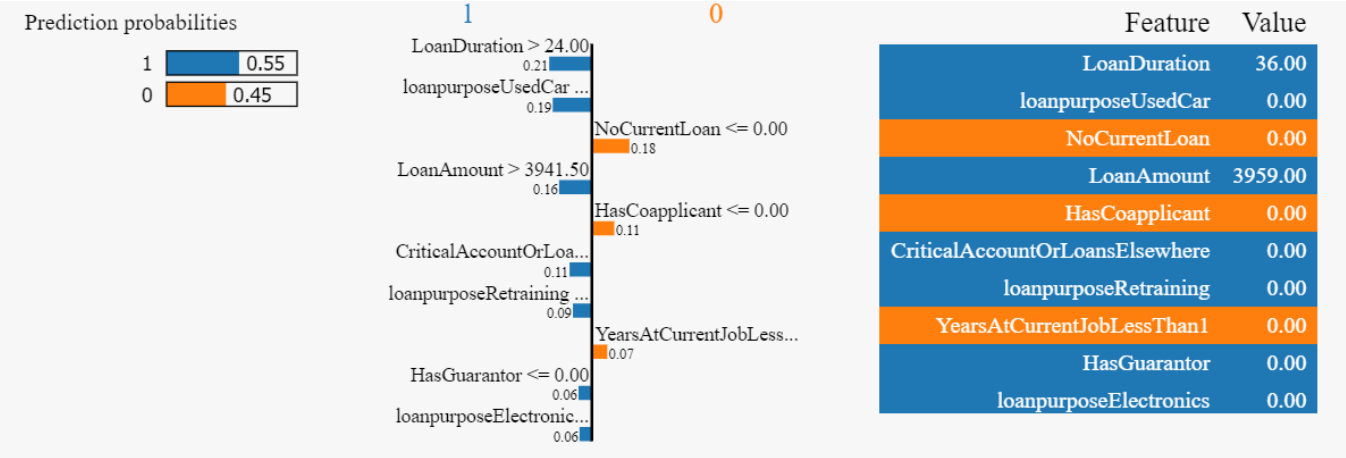
- **Precision (96%)**: 96% of the instances in the dataset that meet these rules are classified in the same way as this instance.
- **Coverage (0.0012)**: Only 0.12% of the instances in the data set meet all these rules.

The results indicate that Anchors performed slightly better than all other methods in terms of faithfulness, achieving the highest score of 0.4690. This suggests that Anchors may offer more interpretable and locally faithful explanations for the underlying model. Its performance reflects its ability to identify conditions (or ”anchors”) that guarantee consistent predictions in a localized feature space. Among the LIME variants, the method with a kernel width of 0.25 exhibited the highest faithfulness score (0.4667). This indicates that smaller kernel widths, which weigh closer instances more heavily in neighborhood sampling, might better capture the local decision boundaries of the model. However, the kernel width of 0.75 (0.4254) outperformed the medium width of 0.5 (0.4211), suggesting that fine-tuning this hyperparameter can significantly influence LIME's performance. Interestingly, SHAP scored the lowest in faithfulness (0.4159) among all methods. While SHAP provides globally consistent explanations due to its game-theoretic foundation, its performance here may reflect a relative inability to align with the model's localized decision-making for the specific instance being evaluated. This highlights potential trade-offs between global consistency and local faithfulness in explanation techniques. The findings underline the importance of selecting explanation methods based on the intended use case. For applications prioritizing local fidelity and interpretability, Anchors and LIME (with appropriate kernel widths) may offer superior results. SHAP, while powerful for global interpretability, might require additional modifications or hybrid approaches to improve its local faithfulness.

5.2. Fairness Score

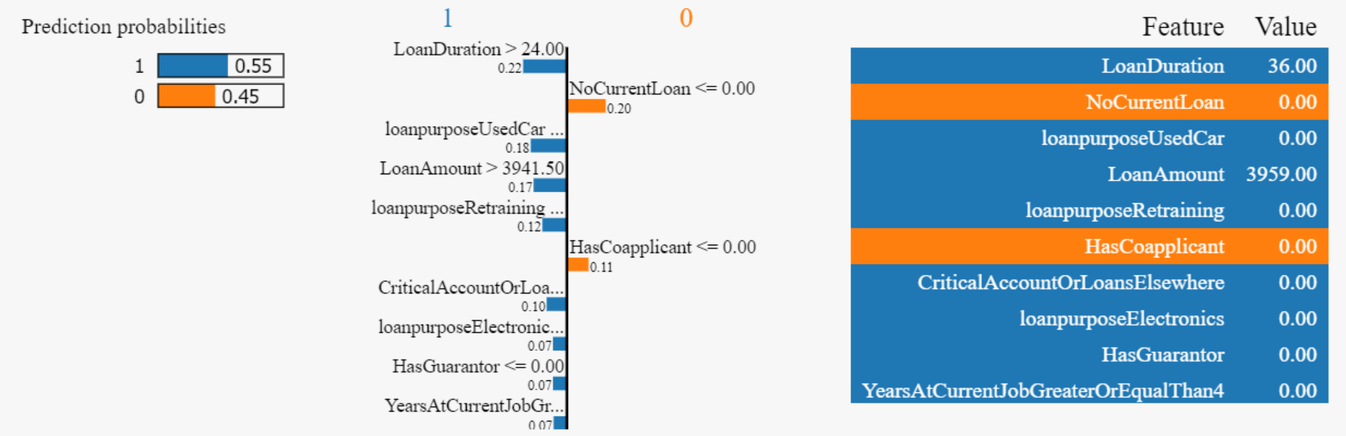
We assess the Disparate Impact Ratio (DIR) for the top three explanation methods: Anchors, LIME (kernel width 0.25) and LIME (kernel width 0.75) in Table 2. These methods were selected from an initial set of five explanation techniques based on their higher faithfulness scores in previous evaluations. The DIR serves as a critical metric for evaluating fairness in machine learning explanations, specifically with regard to gender representation. It quantifies the potential for bias in the feature selection process by measuring the disproportionate impact on different gender groups.

The DIR value for the Anchors method is 0.7523, which is slightly below the commonly accepted threshold of 0.8 for indicating fairness. This suggests that the Anchors method may exhibit a mild gender bias, favoring one gender group over



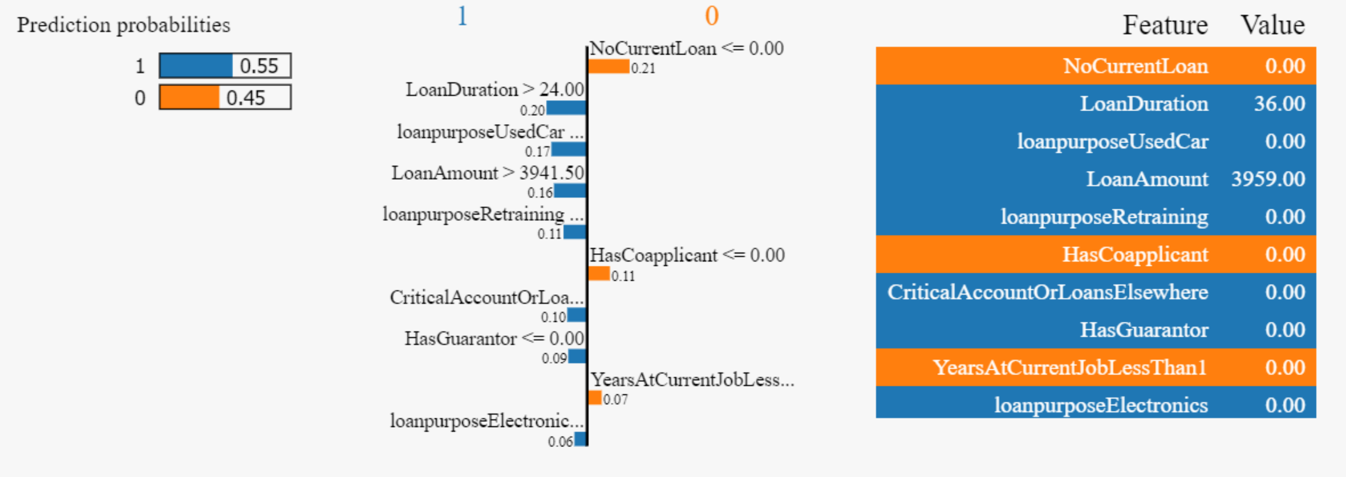
Feature	Value
LoanDuration	36.00
loanpurposeUsedCar	0.00
NoCurrentLoan	0.00
LoanAmount	3959.00
HasCoapplicant	0.00
CriticalAccountOrLoansElsewhere	0.00
loanpurposeRetraining	0.00
YearsAtCurrentJobLessThan1	0.00
HasGuarantor	0.00
loanpurposeElectronics	0.00

Fig. 2: LIME (kernel width 0.25) explanation on the first instance.



Feature	Value
LoanDuration	36.00
NoCurrentLoan	0.00
loanpurposeUsedCar	0.00
LoanAmount	3959.00
loanpurposeRetraining	0.00
HasCoapplicant	0.00
CriticalAccountOrLoansElsewhere	0.00
loanpurposeElectronics	0.00
HasGuarantor	0.00
YearsAtCurrentJobGreaterOrEqualThan4	0.00

Fig. 3: LIME (kernel width 0.75) explanation on the first instance.



Feature	Value
NoCurrentLoan	0.00
LoanDuration	36.00
loanpurposeUsedCar	0.00
LoanAmount	3959.00
loanpurposeRetraining	0.00
HasCoapplicant	0.00
CriticalAccountOrLoansElsewhere	0.00
HasGuarantor	0.00
YearsAtCurrentJobLessThan1	0.00
loanpurposeElectronics	0.00

Fig. 4: LIME (kernel width 0.5) explanation on the first instance.

the other in the process of selecting relevant features. While the DIR value indicates some level of disparate impact, it does not fall drastically below the threshold, implying that the bias is relatively limited. However, this result merits further investigation to determine which gender group is being favored and to explore whether any fairness mitigation techniques are

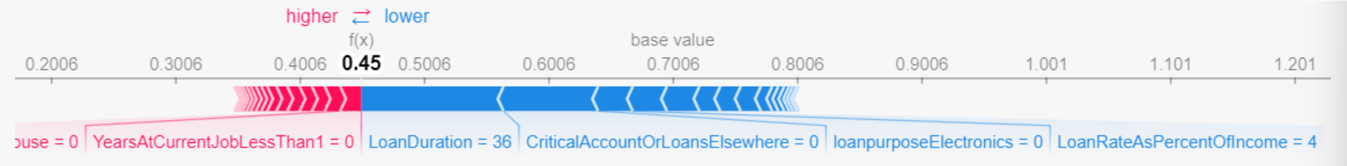


Fig. 5: SHAP explanation on the first instance.

Explanation Method	DIR
Anchors	0.7523
LIME (kernel width 0.25)	0.8224
LIME (kernel width 0.75)	0.8412

Table 2. Disparate Impact Ratio (DIR) for Different Explanation Methods

warranted. In contrast, the LIME method with a kernel width of 0.25 demonstrates a DIR of 0.8224, exceeding the 0.8 threshold. This value suggests that LIME (kernel width 0.25) performs well in terms of fairness, exhibiting minimal gender bias in its feature selection. A DIR above 0.8 is generally considered to reflect an equitable distribution of feature importance between male and female groups. This result positions LIME (kernel width 0.25) as a promising explanation method in terms of fairness, as it appears to avoid substantial gender bias in its feature selection process. The DIR for LIME with a kernel width of 0.75 is 0.812, also exceeding the threshold for fairness. Similar to LIME (kernel width 0.25), this method demonstrates relatively balanced gender representation in its feature selection. Although the DIR for LIME (kernel width 0.75) is slightly lower than that for LIME (kernel width 0.25), the difference is minimal and suggests that both methods exhibit comparable fairness. This finding implies that LIME (kernel width 0.75) also provides fair and unbiased explanations, with only a marginal difference in gender representation compared to LIME (kernel width 0.25). To summarize, all three selected explanation methods — Anchors, LIME (kernel width 0.25), and LIME (kernel width 0.75) — exhibit DIR values above the critical threshold of 0.8, indicating that they do not display significant disparate impact with respect to gender. Among the methods, LIME (kernel width 0.25) demonstrates the highest DIR of 0.8224, making it the most equitable of the three. LIME (kernel width 0.75) and Anchors follow closely, with DIR values of 0.812 and 0.7523, respectively. Although the disparity between these methods is small, the results suggest that LIME (kernel width 0.25) provides the most balanced gender representation. The slight differences in DIR values highlight the potential for subtle biases in feature selection across explanation methods, with LIME (kernel width 0.25) offering the most equitable outcome. Further investigation into these disparities and the application of fairness-enhancing techniques could be beneficial to ensure that no unintended biases persist in the explanations provided by these methods.

5.3. Selecting the Most Faithful and Fair Explanation

The optimal explanation method was chosen based on two primary metrics: faithfulness and the Disparate Impact Ratio (DIR). Faithfulness reflects how well the explanation aligns with the model's behavior, while DIR measures the fairness of the explanation in relation to gender disparity. Both metrics are essential for determining the most appropriate method for model interpretability in fairness-sensitive applications.

Explanation Method	Faithfulness	DIR
Anchors	0.4690	0.7523
LIME (kernel width 0.25)	**0.4667**	**0.8224**
LIME (kernel width 0.75)	0.4211	0.8412

Table 3. Faithfulness and Disparate Impact Ratio (DIR) for Different Explanation Methods

The **Anchors** method, with a faithfulness score of 0.4690, offers a relatively moderate alignment with the model's behavior. While this score suggests that Anchors can capture key features influencing the model's predictions, it is slightly lower than the faithfulness scores of the other methods considered. Furthermore, the DIR for Anchors was calculated to be 0.7523, which is below the commonly accepted fairness threshold of 0.8. This result indicates a potential gender bias, as the method appears to favor one gender over another in its explanations. Such a result raises concerns about the method's fairness and its potential impact on decision-making processes, particularly in sensitive applications such as hiring, lending, or legal judgments.

In contrast, **LIME (kernel width 0.25)** demonstrates a higher faithfulness score of 0.4667, indicating that it provides a more accurate reflection of the model's decision process compared to Anchors. More importantly, **LIME (kernel width 0.25)** also achieves a DIR of 0.8224, which is above the fairness threshold of 0.8. This suggests that LIME (kernel width 0.25) offers not only a more faithful explanation but also a fairer one, minimizing gender disparity in the model's predictions. The higher DIR indicates that this explanation method is less likely to introduce or perpetuate biases, making it a more suitable choice when fairness is a key concern.

LIME (kernel width 0.75), while showing the highest faithfulness score of 0.4211, has a DIR of 0.8412, which, although above the fairness threshold, is slightly lower than that of **LIME (kernel width 0.25)**. The marginal reduction in fairness compared to LIME (kernel width 0.25) suggests that while this explanation method remains fair, it may still exhibit some degree of bias in comparison to the alternative. Furthermore, the increased kernel width in LIME (kernel width 0.75) suggests a more flexible local approximation, which can lead to explanations that better capture model complexity, but at the cost of potentially increasing the variance in fairness across different subsets of data.

Given the trade-offs between these methods, **LIME (kernel width 0.25)** was ultimately selected as the optimal explanation method. Despite its slightly lower faithfulness score compared to LIME (kernel width 0.75), its higher DIR of 0.8224 makes it the fairest choice. In scenarios where fairness is of paramount importance, such as when explaining decisions in sensitive applications, ensuring that the explanation method exhibits minimal bias is crucial. **LIME (kernel width 0.25)** strikes the best balance between providing an accurate, faithful explanation and maintaining fairness across different gender groups.

6. Conclusion

While the current study has made significant strides in identifying an optimal explanation method through a combined evaluation of faithfulness and fairness, several avenues for future research remain. The findings highlight important aspects of explainability and fairness in machine learning, but they also expose limitations and open questions that merit further exploration.

First, while this study focused on gender as the primary demographic attribute for assessing fairness through the Disparate Impact Ratio (DIR), future work should explore additional sensitive attributes, such as age, race, socioeconomic status, or intersectional combinations of these factors. The inclusion of these attributes will provide a more comprehensive understanding of potential biases in explanation methods and their broader implications. This expansion would also align with fairness guidelines in real-world applications, where multiple dimensions of discrimination are often at play.

Second, the evaluation of explanation methods was conducted under specific configurations of the LIME and Anchors algorithms. While kernel width and other parameters were varied in this study, future research could systematically explore a broader range of hyperparameter configurations to identify optimal settings for different datasets and problem contexts. Additionally, adapting parameter tuning techniques to balance faithfulness and fairness dynamically, rather than treating them as independent metrics, could lead to more robust and universally applicable explanation frameworks.

Third, the current study assumes a fixed definition of fairness based on the DIR threshold (0.8). However, fairness is a context-dependent concept, and the acceptable thresholds may vary across domains and stakeholders. Future research could involve the integration of domain-specific fairness constraints or user-defined fairness objectives, which could allow for personalized or application-specific explanation selection processes. Similarly, the exploration of alternative fairness metrics, such as equality of opportunity, equalized odds, or individual fairness, would help generalize the findings to different ethical and legal frameworks.

Another promising area for future work is the incorporation of causal inference techniques into explanation evaluation. The DIR metric used in this study measures correlations between gender and feature importance rankings, but it does not account for potential confounding factors or causal relationships. By leveraging causal models, future studies could provide a deeper understanding of the origins of bias and develop explanation methods that actively mitigate causal disparities.

Furthermore, while the study demonstrated the trade-offs between faithfulness and fairness, the interaction between these two metrics warrants closer examination. Specifically, future research could investigate whether improvements in fairness inherently lead to a loss in faithfulness or vice versa. Understanding this relationship could lead to the development of hybrid metrics or multi-objective optimization techniques that balance these competing goals effectively.

Lastly, a practical extension of this work would involve deploying the selected explanation methods in real-world applications to evaluate their impact on decision-making processes. For instance, understanding how end-users interpret and act upon explanations generated by LIME (kernel width 0.25) or other methods could provide valuable feedback for refining these techniques. Human-centered evaluations, including user studies and qualitative analyses, could shed light on the usability, interpretability, and trustworthiness of explanations from the perspective of non-expert stakeholders.

In summary, future work should prioritize expanding the scope of fairness analysis to multiple sensitive attributes, optimizing explanation parameters dynamically, incorporating causal inference, and examining the interplay between faithfulness and fairness. Moreover, practical deployments and user-centric studies will be instrumental in bridging the gap between theoretical advancements and real-world impact, paving the way for fair and faithful explanations in machine learning systems.

Appendix

Table A1. Database description

Short name	Complete name	Variable type	Domain
Age	Age	Numerical	$\mathbb{R}^+$
CreditAmount	Credit amount	Numerical	$\mathbb{R}^+$
CreditDuration	Credit duration	Numerical	$\mathbb{R}^+$
AccountStatus	Status of existing checking account	Categorical	#4
CreditHistory	Credit history	Categorical	#5
Purpose	Credit Purpose	Categorical	#10
Savings	Status of savings accounts and bonds	Categorical	#5
EmploymentDuration	Employment length	Categorical	#5
InstallmentRate	Installment rate	Numerical	{1, 2, 3, 4}
Gender&PersonalStatus	Personal status and gender	Categorical	#4
Guarantor	Other debtors	Categorical	#3
ResidenceTime	Period of present residency	Numerical	{1, 2, 3, 4}
Property	Property	Categorical	#4
OtherInstallmentPlan	Installment plans	Categorical	#3
Housing	Residence	Categorical	#3
NumberOfCredit	Number of existing credits	Numerical	{1, 2, 3, 4}
Job	Employment	Categorical	#4
NumberLiablePeople	Dependents	Numerical	{1, 2}
Telephone	Telephone	Binary	#2
ForeignWorker	Foreign worker	Binary	#2
CreditRisk	Credit score	Binary	#2

Table A2. Feature overview

Complete name	**Description**
Age	Age in years
Credit amount	Credit amount
Credit duration	Duration in months
Status of existing checking account	A11: . . . ¡ 0 DM, A12: 0 $\leq$ ... ¡ 200 DM, A13: $\geq$ 200 DM / salary assignments (1 year), A14: no checking account
Credit history	A30: no credits taken/ all credits paid back duly, A31: all credits at this bank paid back duly, A32: existing credits paid back duly till now, A33: delay in paying off in the past, A34: other credits existing (not at this bank)
Credit Purpose	A40: car (new), A41: car (used), A42: equipment, A43: radio/television, A44: domestic appliances, A45: repairs, A46: education, A48: retraining, A49: business, A410: others
Status of savings accounts and bonds	A61: ¡ 100 DM, A62: 100 ¡ ... ¡ 500, A63: 500 ¡ ... ¡ 1000, A64: $\geq$ 1000 DM, A65: unknown / no savings account
Employment duration	A71: unemployed, A72: ¡ 1 year, A73: 1 $\leq$ ... ¡ 4 years, A74: 4 $\leq$ ... ¡ 7 years, A75: $\geq$ 7 years
Installment rate	Installment rate in percentage of disposable income
Personal status and gender	A91: male: divorced/separated, A92: female: divorced/separated/married, A93: male: single, A94: male: married/widowed
Other debtors	A101: none, A102: co-applicant
Period of present residency	Present residence since
Property	A121: real estate, A123: car or other, A122: building society savings agreement, A124: unknown / no property
Installment plans	A141: bank, A142: stores, A143: none
Housing	A151: rent, A152: own, A153: for free
Number of existing credits	Number of existing credits at this bank
Employment	A171: unemployed, A172: unskilled - non-resident, A173: skilled employee, A174: management/ self-employed/ highly qualified
Dependents	Number of people being liable to provide maintenance
Telephone	A191: none, A192: yes
Foreign worker	A201: yes, A202: no
Credit score	1: Good, 2: Bad

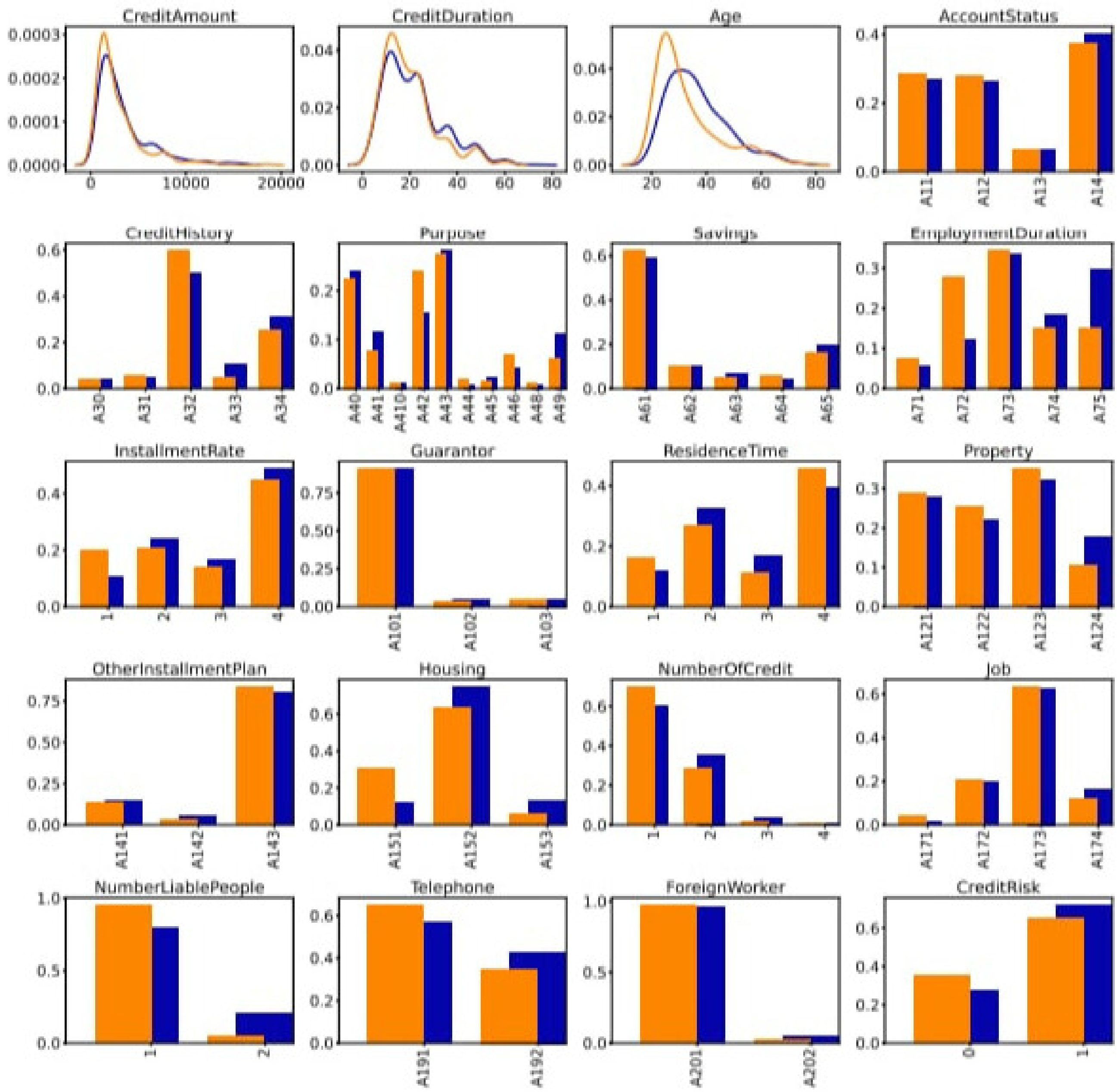


Fig. 6: Feature distributions

Notes: These figures display the feature distributions by gender, using kernel density estimation for continuous variables. Blue color refers to men and orange to women.

References

[1] Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L. C. (2015). Benchmarking state-of the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

[2] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G. Z. (2019). XAI—Explainable artificial intelligence. Science robotics, 4(37), eaay7120.

[3] Das, A., Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.

[4] Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., Lin, S. F. (2021). Explainable AI in credit risk management. arXiv preprint arXiv:2103.00949.

[5] Covello, V. T., Mumpower, J. (1985). Risk analysis and risk management: an historical perspective. Risk analysis, 5(2), 103-120.

[6] Moldovan, D. (2023). Algorithmic decision making methods for fair credit scoring. IEEE Access, 11, 59729-59743.

[7] White Paper on Artificial Intelligence: A European Approach to Excellence and Trust, Eur.Commission, Brussels, Belgium, 2020.

[8] Human Rights Council, "The right to privacy in the digital age," Hum. Rights Council, Geneva, Switzerland, Tech. Rep. U.N. Doc. A/HRC/48/31, 2021.

[9] Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. R., Samek, W. (2020, July). xxAI-beyond explainable artificial intelligence. In International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 3-10). Cham: Springer International Publishing.

[10] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

[11] Barceló, P., Monet, M., Pérez, J., Subercaseaux, B. (2020). Model interpretability through the lens of computational complexity. Advances in neural information processing systems, 33, 15487-15498.

[12] Lecouat, B., Ponce, J., Mairal, J. (2020). Fully trainable and interpretable non-local sparse models for image restoration. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16 (pp. 238-254). Springer International Publishing.

[13] Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., Wang, T. (2018). An interpretable model with globally consistent explanations for credit risk. arXiv preprint arXiv:1811.12615.

[14] Hall, R. J. (1988). Learning by failing to explain: Using partial explanations to learn in incomplete or intractable domains. Machine Learning, 3, 45-77.

[15] Keil, F. C. (2006). Explanation and understanding. Annu. Rev. Psychol., 57(1), 227-254.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

[17] Scott M. Lundberg and Su-In Lee A Unified Approach to Interpreting Model Predictions. In 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[18] El-Sappagh, S., Alonso, J. M., Islam, S. R., Sultan, A. M., Kwak, K. S. (2021). A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. Scientific reports, 11(1), 2660.

[19] Danyang Huang, Jing Zhou, and Hansheng Wang. Rfms method for credit scoring based on bank card transaction data. Statistica Sinica, 28(4):2903–2919, 2018.

[20] C. Wang, D. Han, Q. Liu, and S. Luo. A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism lstm. IEEE Access, 7:2161–2168, 2019.

[21] Shenghui Yang, Haomin Zhang, et al. Comparison of several data mining methods in credit card default prediction. Intelligent Information Management, 10(05):115, 2018.

[22] Li, L., Zhang, Z., Xiong, Y., Hu, Z., Liu, S., Tu, B., Yao, Y. (2022). Prediction of hospital mortality in mechanically ventilated patients with congestive heart failure using machine learning approaches. International Journal of Cardiology, 358, 59-64.

[23] Chen, M. X., Lu, C. C., Sun, P. C., Nie, Y. X., Tian, Y., Hu, Q. J., ... Liu, Y. G. (2021). Comprehensive transcriptome and proteome analyses reveal a novel sodium chloride responsive gene network in maize seed tissues during germination. Plant, Cell Environment, 44(1), 88-101.

[24] Plumb, G., Molitor, D., and Talwalkar, A. Model agnostic supervised local explanations. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (2018), pp. 2520–2529.

[25] Ribeiro, M. T., Singh, S., and 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 (2016), pp. 1135–1144

[26] Teach, R. L., Shortliffe, E. H. (1981). An analysis of physician attitudes regarding computer-based clinical consultation systems. Computers and Biomedical Research, 14(6), 542-558.

[27] Dasgupta, S., Frost, N., Moshkovitz, M. (2022, June). Framework for evaluating faithfulness of local explanations. In International Conference on Machine Learning (pp. 4794-4815). PMLR.

[28] Leavitt, M. L. and Morcos, A. Towards falsifiable interpretability research. arXiv preprint arXiv:2010.12016, 2020.

[29] Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics, 10(5):593, 2021.

[30] Kim, B. and Doshi-Velez, F. Machine learning techniques for accountability. AI Magazine, 42(1):47–52, 2021.

[31] Pruthi, D., Bansal, R., Dhingra, B., Soares, L. B., Collins, M., Lipton, Z. C., Neubig, G., and Cohen, W. W. Evaluating explanations: How much do explanations from the teacher aid students? Transactions of the Association for Computational Linguistics, 10:359–375, 2022.

[32] Jesus, S., Belem, C., Balayan, V., Bento, J., Saleiro, P., ´ Bizarro, P., and Gama, J. How can i choose an explainer? an application-grounded evaluation of post-hoc explanations. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 805–815, 2021.

[33] Mohseni, S., Block, J. E., and Ragan, E. D. A humangrounded evaluation benchmark for local explanations of machine learning. arXiv preprint arXiv:1801.05075, 2018.

[34] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–52, 2021

[35] Poppi, S., Cornia, M., Baraldi, L., and Cucchiara, R. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2299–2304, 2021.

[36] Barcelo, P., Monet, M., P ´ erez, J., and Subercaseaux, B. ´ Model interpretability through the lens of computational complexity. arXiv preprint arXiv:2010.12265, 2020.

[37] Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049, 2018.

[38] Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., and Lakkaraju, H. Rethinking stability for attribution-based explanations. arXiv preprint arXiv:2203.06877, 2022.

[39] Dasgupta, S., Frost, N., Moshkovitz, M. (2022, June). Framework for evaluating faithfulness of local explanations. In International Conference on Machine Learning (pp. 4794-4815). PMLR.

[40] Lipton, Z. C. The mythos of model interpretability. Queue, 16(3):31–57, 2018.

[41] Rana, A., Bridge, D. (2018, July). Explanations that are intrinsic to recommendations. In Proceedings of the 26th conference on user modeling, adaptation and personalization (pp. 187-195).

[42] Slack, D., Hilgard, A., Singh, S., Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. Advances in neural information processing systems, 34, 9391-9404.

[43] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer google schola, 2, 1122-1128.

[44] Malekipirbazari, M., Aksakalli, V., Shafqat, W., Eberhard, A. (2021). Performance comparison of feature selection and extraction methods with random instance selection. Expert Systems with Applications, 179, 115072.

[45] Pierce, K. M., Hope, J. L., Johnson, K. J., Wright, B. W., Synovec, R. E. (2005). Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis. Journal of Chromatography A, 1096(1-2), 101-110.

[46] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179–188.

[47] Stone, J. V. (2004). Independent component analysis: a tutorial introduction. MIT press.

[48] Zheng, Z., Chenmao, X., Jia, J. (2010). Iso-container projection for feature extraction. In Proceedings of IEEE international symposium on intelligent signal processing and communication systems.

[49] Štrumbelj, E., Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, 41(3), 647-665.

[50] Chen, H., Lundberg, S., Lee, S. I. (2021). Explaining models by propagating Shapley values of local components. Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability, 261-270.

[51] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).

[52] Dieber, J., Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. arXiv preprint arXiv:2012.00093.

[53] Demajo, L. M., Vella, V., Dingli, A. (2020). Explainable ai for interpretable credit scoring. arXiv preprint arXiv:2012.03749.

[54] Sanchez, I., Rocktaschel, T., Riedel, S., and Singh, S. Towards extracting faithful and descriptive representations of latent variable models. AAAI Spring Syposium on Knowledge Representation and Reasoning (KRR): Integrating Symbolic and Neural Approaches 1 (2015), 4–1.

[55] Garreau, D., and Luxburg, U. Explaining the explainer: A first theoretical analysis of lime. In International Conference on Artificial Intelligence and Statistics (2020), PMLR, pp. 1287–1296.

[56] Agarwal, S., Jabbari, S., Agarwal, C., Upadhyay, S., Wu, S., and Lakkaraju, H. Towards the unification and robustness of perturbation and gradient based explanations. In International Conference on Machine Learning ICML (2021), M. Meila and T. Zhang, Eds., vol. 139 of Proceedings of Machine Learning Research, PMLR, pp. 110–119.

[57] S. Barocas, M. Hardt, and A. Narayanan. (2019). Fairness and Machine Learning. [Online]. Available: https://www.fairmlbook.org

[58] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fair ness through awareness," in Proc. 3rd Innov. Theor. Comput. Sci. Conf. New York, NY, USA: Association for Computing Machinery, Jan. 2012, pp. 214–226.

[59] W. Dieterich, C. Mendoza, and T. Brennan, "Compas risk scales: Demon strating accuracy equity and predictive parity," Northpointe Inc., Traverse City, MI, USA, vol. 7, no. 4, Tech. Rep., 2016.

[60] C. Simoiu, S. Corbett-Davies, and S. Goel, "The problem of infra marginality in outcome tests for discrimination," Ann. Appl. Statist., vol. 11, no. 3, pp. 1193–1216, Sep. 2017.

[61] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in Proc. 8th Innov. Theor. Comput. Sci. Conf. (ITCS), vol. 67. Wadern, Germany: Schloss Dagstuhl–Leibniz Zentrum fuer Informatik, 2017, p. 43.

[62] B.Paaßen,A.Bunge,C.Hainke,L.Sindelar,andM.Vogelsang,"Dynamic fairness—Breaking vicious cycles in automatic decision making," in Proc. 27th Eur. Symp. Artif. Neural Netw. (ESANN), 2019, pp. 1–16.

[63] T. Bono, K. Croxson, and A. Giles, "Algorithmic fairness in credit scor ing," Oxford Rev. Econ. Policy, vol. 37, no. 3, pp. 585–617, Sep. 2021.

[64] D. C. Parkes and R. V. Vohra, “Algorithmic and economic perspectives on fairness,” 2019, arXiv:1909.05282

[65] de Lange, Petter Eilif, Borger Melsom, Christian Bakke Vennerød, and Sjur Westgaard. 2022. Explainable AI for Credit Assessment in Banks. Journal of Risk and Financial Management 15: 556. https:// doi.org/10.3390/jrfm15120556

[66] Yan, J., Yu, W. and Zhao, J. L. (2019)FinBrain: when finance meets AI 2.0. Frontiers Inf Technol Electronic Eng, 20: 914–924.

[67] Wei, X., Rao, C., Xiao, X., Chen, L., Goh, M. (2023). Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model. Expert systems with applications, 219, 119648.

[68] Torrent, N. L., Visani, G., Bagli, E. (2020). PSD2 explainable AI model for credit scoring. arXiv preprint arXiv:2011.10367.

[69] Ron Kohavi. Bottom-up induction of oblivious read-once decision graphs. In European Conference on Machine Learning, pages 154–169. Springer, 1994.

[70] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In Advances in neural information processing systems, pages 6638–6648, 2018.

[71] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical fea tures support. arXiv preprint arXiv:1810.11363, 2018.

[72] Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High precision model-agnostic explanations. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[73] Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pp. 4765–4774, 2017

[74] HLEG. AI. Ethical guidelines for trustworthy ai. European Commission, Apr. 2019.

[75] Cristianini, N., Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press

[76] Burges, C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2, 121–167.

[77] Barrera Vicen, A., Paluzo Hidalgo, E., Gutiérrez Naranjo, M. Á. (2023). The metric-aware kernel-width choice for LIME.

[78] Shapley, L. S. A value for n-person games. Contributions to the Theory of Games, 2(28):307–317, 1953.

[79] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. 2019. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv arXiv:1802.03888.

[80] Molnar, Christoph. 2019. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. SHAP (Shapley Additive Explanations): chap. 9.6. Available online: https://christophm.github.io/interpretableAI-book/shap.htAI (accessed on 6 November 2022).

[81] Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)

[82] Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (2018)

[83] Hernandez-Leal, P., Kartal, B., Taylor, M.E.: A survey and critique of multia gent deep reinforcement learning. Auton. Agent. Multi-Agent Syst. 33(6), 750–797 (2019). https://doi.org/10.1007/s10458-019-09421-1

[84] Feldman, M. , Friedler, S. A. , Moeller, J. , Scheidegger, C. , Venkatasubrama- nian, S. (2015). Certifying and removing disparate impact. In ACM SIGKDD in- ternational conference on knowledge discovery and data mining (pp. 259–268)

[85] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fair ness through awareness,” in Proc. 3rd Innov. Theor. Comput. Sci. Conf. New York, NY, USA: Association for Computing Machinery, Jan. 2012, pp. 214–226.