Context-aware SHAP for responsible feature selection

Sivapriya Vellaichamy
sivapriya.vellaichamy@jpmchase.com
J.P. Morgan AI Research
USA

Shubham Sharma
J.P. Morgan AI Research
USA
shubham.x2.sharma@jpmchase.com

Emanuele Albini
J.P. Morgan AI Research
UK
emanuele.albini@jpmorgan.com

Ivan Brugere
J.P. Morgan AI Research
USA
ivan.brugere@jpmchase.com

Ade Onigbanjo
J.P. Morgan AI Research
UK
ade.onigbanjo@jpmchase.com

Abstract

Artificial intelligence models are widely deployed for decision making in finance. Recently, emphasis on trustworthy and explainable methods have motivated adopting responsible feature selection practices, which yield simpler and more interpretable models. Responsible feature selection considers other aspects of model such as fairness and robustness, in addition to accuracy. Traditional methods like SHAP (SHapley Additive exPlanations) for feature selection help in obtaining models that have good accuracy but fail to take other criteria into consideration. In this paper, we introduce a novel context aware feature selection approach that integrates three criteria - performance, fairness and robustness. For each criteria, we propose varying 'foreground' and 'background' for computing SHAP and we select features accounting for all criteria. We demonstrate the proposed method by comparing against the standard SHAP based feature selection technique on logistic regression and XGBoost models.

CCS Concepts

• **Social and professional topics** → **Computing / technology policy**; • **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → *Machine learning*.

Keywords

SHAP, Feature Selection, Performance, Fairness, Robustness

ACM Reference Format:
Sivapriya Vellaichamy, Shubham Sharma, Emanuele Albini, Ivan Brugere, and Ade Onigbanjo. 2018. Context-aware SHAP for responsible feature selection. In *Proceedings of 4th Workshop on Explainable AI in Finance (XAI-FIN-2024).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recently, artificial intelligence models are increasingly a part of decision-making in finance. With large amounts of compute and data available, AI models regularly perform better than traditional statistical methods [21, 25]. Several studies [11, 22] highlight their wide use for finance applications, including credit line approval, targeted marketing and fraud detection. The widespread deployment of AI models has raised concerns regarding their trustworthiness and responsible use. Responsible AI requires that an approach achieves its primary objective e.g. accuracy in a way that is explainable, fair and robust [5, 19]. Several regulations and guidelines worldwide have emerged to ensure the models used in high-impact decision-making application can be trusted. Europen Union [27] emphasizes the right to obtain an explanation of the decision reached after an automated assessment and in the United States Algorithmic Accountability Act of 2019 requires companies to assess the impact of automated decision systems and ensure transparency and explainability [1].

In many finance applications, we encounter large–mostly tabular–datasets e.g. borrower's attributes on a loan application, company attributes on a securities report. As dimensionality increases, the number of data-points required to train a model with high fidelity is exponential–typically referred to the *curse of dimensionality* in machine learning. Feature selection is often utilized to retain the most important variables to achieve a certain criterion e.g. accuracy of the model, while reducing the effect of high dimensionality. In our methodology, we apply Shapley Additive Explanations (SHAP) for interpretable feature attribution after model training. It is a widely used technique for feature selection. Lundberg and Lee [16] describes the applicability of SHAP as a feature selection tool and it known to often outperform commonly used feature selection algorithms [17].

Computation of SHAP values for each features requires the user to choose a *background* dataset. The expectation over this background distribution is used for simulating the absence of a feature.

XAI-FIN-2024, November 15, 2024, Brooklyn, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

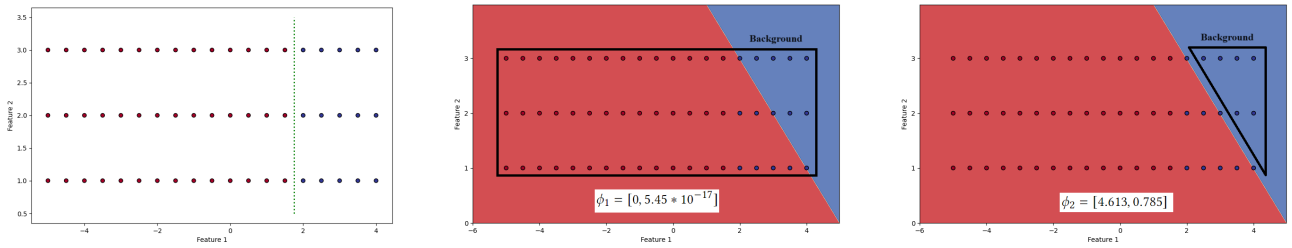

Figure 1: Impact of background distribution on SHAP-based feature selection. (a) An ideal classifier for the dataset with two features, highlighting the greater importance of Feature 1 for classifying red and blue classes. (b) SHAP values computed with a background of all data points indicate higher importance for Feature 2. (c) SHAP values using correctly classified blue data points as background reverse the feature importance rankings.

This choice of background dataset is crucial since different choices can result in different attribute rankings. Typically, existing methods randomly sample data-points from the training set to constitute the background distribution.

However, the major limitation is that SHAP is primarily designed to support feature selection optimized for performance. It does not specifically accommodate feature selection based on other criteria.

This work includes several contributions:

- We measure how the choice of background distribution (i.e. *context*) affects the outcome of feature selection using SHAP.
- For criteria like fairness and robustness, we apply the contrastive properties of foreground against background in SHAP to derive a variable ranking.
- We demonstrate our methodology simultaneously achieving these multiple criteria.
- Our methodology is model-agnostic and we demonstrate its effectiveness on two classes of model, Logistic Regression and XGBoost.

2 Related Work

Feature selection with performance as the sole objective is a well researched problem and several methods to achieve that are highlighted in [7, 10].

Responsible Feature Selection involves consideration of ethical and social factors other than performance like fairness, adversarial robustness and interpretability [3]. Models with lower number of variables are considered more interpretable [20].

SHAP as an explainability technique was introduced in [16] where each feature is assigned an importance value for a particular prediction. Using SHAP for feature selection is a widely studied topic [9, 17] The paradigm has also been shown to extend to explaining fairness in machine learning by attributing a model's unfairness to each feature used to build the model [4].

Background distribution for SHAP: Calculating SHAP values requires the input of background distribution which significantly affects the results. The official SHAP documentation suggests taking 100 randomly drawn samples from the training dataset but limited study has been conducted on how the choice impacts the results. For instance, [28] experimentally evaluates the impact of background data size and concludes that the SHAP stability improves with background sample size. BalanceSHAP proposed in [15] finds increased predictability and less abnormal bee-swarm plot points when the background and foreground data points have equal distribution of classes.

The concept of attributions as contrastive explanations of an input relative to a reference or a background distribution is proposed in [18]. The paper highlights the need for conscious choice of references to obtain specific contrastive explanation. This idea plays a crucial role in the way we select the background and foreground data points for each of the criteria for our method. Generalized Shapley Additive Explanations (G-SHAP) introduced in [6] highlights how the SHAP value for a feature quantifies the difference between the model output of an instance and the background observation. This contrast is then used to answer questions related to why an instance belongs to one class over the other and model prediction differences between different groups of observations. The importance of baselines and their ability to significantly impact attributions is highlighted in [12].

3 Background

In this section, we highlight the significance of SHAP background in the context of feature selection, both from a mathematical perspective and through an illustrative example.

3.1 SHAP

Consider a binary classification model, denoted as M, which produces an output label of either 0 or 1. The inputs to the model M are defined as follows:

$$X = X_i, \quad i = 1, \ldots, N \tag{1}$$

Each instance i comprises V variables that contribute to the prediction probability P_i of being classified as 1. This can be expressed as:

$$X_i = v_{i,j}, \quad j = 1, \ldots, V \tag{2}$$

The prediction probability for instance i is given by:

$$P_i = M(X_i) \tag{3}$$

For this model M and instance i, SHAP (SHapley Additive exPlanations) provides the contribution of each variable towards the prediction deviation from a prior probability ϕ_0, which represents

the expected probability of samples in the background dataset *bg* [28].

$$P_i - \phi_0 = \sum_{j=1}^{V} \phi_{i,j,bg} \tag{4}$$

For model level importance of variable j for a background data $I_{j,bg}$, we sum the absolute SHAP values over all instances in D [16, 17].

$$I_{j,bg} = \sum_{i=1}^{N} |\phi_{i,j,bg}| \tag{5}$$

In this context, each variable is assigned an importance score by aggregating the absolute SHAP values across all instances in X. A higher aggregated value indicates greater importance, as it reflects a more significant role in the decision-making process of individual predictions on average. As demonstrated in Equation 4, the background set influences the SHAP computation for each instance, which subsequently impacts the variable ranking as outlined in Equation 5.

3.2 Illustration

We consider a dataset comprising two features, where each data point is to be classified into one of two classes: red and blue. An ideal classifier for this scenario, as depicted in Figure 1(a), would rely solely on Feature 1. This example clearly establishes the greater importance of Feature 1 over Feature 2.

For the SHAP calculation, we examine two different choices of background distribution. In the first case as shown in Fig. 1 (b), the background dataset consists of all data points, while in the second case in Fig. 1 (c), it comprises only the correctly classified data points of the blue class (true positives). In each scenario, we compute the absolute average SHAP values across all instances to determine the variable rankings as per equation 5.

$$\phi_1 = [0, 5.45 * 10^{-17}]$$

$$\phi_2 = [4.613, 0.785]$$

In the first case, the ϕ_1 values for Feature 2 are slightly greater than those for Feature 1, suggesting the importance of Feature 2. Conversely, in the second scenario, the rank ordering based on SHAP values (ϕ_2) is reversed. This example illustrates the significant influence of the background distribution when using SHAP for feature selection. Furthermore, it demonstrates that the commonly suggested method of choosing random data points can lead to an incorrect ranking of feature importance.

4 Methodology

This section highlights the core methodology proposed for deriving a feature selection list that incorporates multiple criteria like fairness and robustness, in addition to performance.The framework is designed to be extensible to other model characteristics such as privacy, relevance and additional considerations.

4.1 Feature selection

We consider the problem of feature selection using SHAP along three important criteria: performance, fairness and robustness. The aim of this paper is to use our methodology to derive a feature list that performs well across multiple criteria. We evaluate the proposed approach on various datasets and across different classes of models.

Throughout the paper, we use the following terms as defined below:

(1) **Foreground**: The data points for which the SHAP values are calculated
(2) **Background**: The data points that serves as the reference or background in the SHAP value calculation

Based on the discussion in Section 2 on contrastive explanation, we propose specific choice of background and foreground data points for each criterion. For a given instance, the resulting SHAP values explain the output in contrast to the background data points [2, 18]. The average absolute Shapley values indicate which features cause the model's output for the foreground points to differ from the output for the background points, on average, pairwise.

Figure 2 illustrates the overall flow of the methodology. For each criterion, relevant background and foreground data are selected to generate SHAP-based variable rankings. Each ranked variable list is then aggregated based on variable position to obtain a final ranked list. The effectiveness of this final list is compared against the SHAP-based ranking list derived from a randomly chosen training dataset for background and foreground data, which is the common practice. These sorted lists of variables form the basis for our feature selection process. Selecting the optimal set of features from the ranked list is achieved by cumulatively removing variables from the final ranked list.

4.2 Performance

Performance is a primary concern when selecting variables, as the accuracy of the model determines its utility. Although our methodology focuses on binary classification for this paper, it can be readily extended to multi-class classification.

Initially, we construct a model incorporating all variables under consideration. By predicting the output of the training dataset using this model, we identify correctly classified points: True Positives and True Negatives. Additionally, we identify points that are closer to the decision boundary in both the positive and negative classes, which we refer to as weak positives and weak negatives. These points lie within a defined threshold of the decision boundary. To derive a ranked list of variables optimized to enhance performance metrics such as the Area Under the Curve (AUC), which relies on the accuracy of both positive and negative classes, we employ the following methodology:

- *Positive Class Ranked List*: Using true positive data points from the training dataset as the background and weak negative data points as the foreground, we compute the SHAP values. The absolute values are then aggregated and sorted to generate the Positive Class Ranked List.
- *Negative Class Ranked List*: Using true negative data points from the training dataset as the background and weak positive data points as the foreground, we calculate the SHAP values. The absolute values are then aggregated and sorted to produce the Negative Class Ranked List.

Similar to traditional SHAP-based feature selection, we prioritize features with the largest absolute SHAP values. Intuitively, in the contrastive formulation of SHAP, these features are those that, on

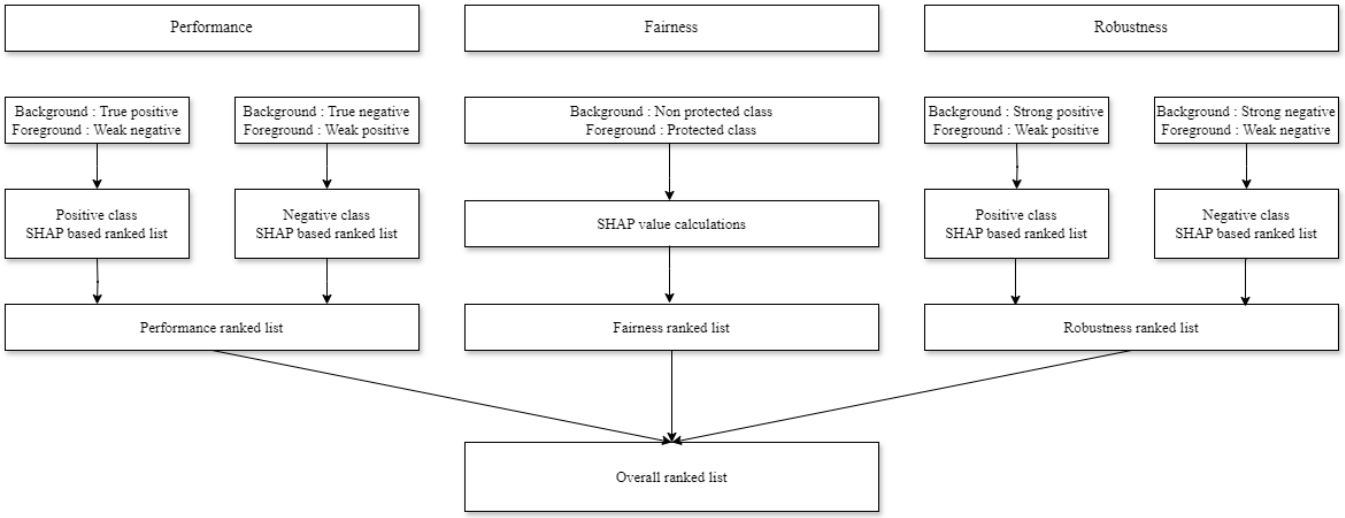

Figure 2: Flowchart describing the overall methodology

average, contribute the most to making the output of the points near the decision boundary on that side of the decision boundary. Consequently, these features are deemed most important for the model's ability to classify samples into one class rather than the other.

Final Performance Ranked List: We aggregate the Positive Class Ranked List and the Negative Class Ranked List to obtain the final list of variables for performance. This is achieved by summing the ranks from the two lists and then sorting the variables based on these aggregated ranks. Equal weight is assigned to both lists during this process, although this weighting is a hyperparameter that can be adjusted based on the specific application requirements.

4.3 Fairness

Ensuring fairness in decision-making is crucial for building trust in artificial intelligence models. As fairness gains increasing regulatory significance, it becomes an essential metric in the feature selection process for model development.

We use demographic parity difference metric to measure disparity. This metric is a measures of the selection rate disparity between different groups.

Mathematically,

$$DP = \mathbb{E}[p(X) \mid \text{class} = \text{protected}] - \mathbb{E}[p(X) \mid \text{class} = \text{non-protected}] \tag{6}$$

where p(X) denotes the predicted output of the model for each instance.

To obtain a fairness-based ranking list, we designate our background data as points belonging to the non-protected group and the foreground data as points from the protected group. For example, when assessing gender bias against females, we use males as the background and females as the foreground. Similarly, to investigate age-related bias, the background could consist of individuals younger than, say, 60 years old, while the foreground would include individuals older than 60 years.

Using these background and foreground choices, we compute the SHAP values with a model trained on all considered features. The absolute SHAP values for the foreground instances are aggregated to create the fairness-based ranking list.

Given the contrastive nature of SHAP explanations, this approach provides the average contribution of each feature to the adverse outcome (e.g., rejection) for the protected group relative to the non-protected group. Specifically, features with a high ϕ^F value are those that, on average, contribute more to the adverse outcome (e.g., higher model output) for the protected group compared to the non-protected group.

Consequently, we sort the variables in the order of prioritizing those with the lowest values as they are deemed most critical for retention in the fairness-based importance list.

4.4 Robustness

Robust models are characterized by their ability to maintain consistent decision-making in the face of minor perturbations, thereby avoiding significant deviations in outcomes. The robustness of a prediction is quantified by its distance from the decision boundary threshold, as defined in [23, 24]. Points situated further from the decision boundary are less susceptible to perturbations compared to those near the boundary. Models whose predictions are consistently well away from the decision boundary exhibit greater confidence and robustness than models with predictions clustered around the boundary.

The robustness of a prediction in the output space is defined as follows:

$$Robustness = |t - \hat{y}| \tag{7}$$

where $\hat{y}$ represents the prediction probability of the instance and t denotes the classification threshold separating the two classes. The overall robustness of the model is the aggregate of the robustness values of all instances.

We define a threshold ϵ which marks the distance beyond which we consider points to be located farther from the decision boundary. After constructing a model with all considered variables, the predictions from the training dataset fall into one of the following categories:

(a) Weak positive probabilities that lie in between t and t+ϵ
(b) Strong positive probabilities that lie above t+ϵ
(c) Weak negative probabilities that lie in between t and t-ϵ
(d) Strong negative probabilities that lie below t-ϵ

To evaluate robustness, the background and foreground choices are established for each class analogous to the performance:

- Positive Class Ranked List: The background dataset comprises data points with strong positive probabilities, while the foreground dataset includes data points with weak positive probabilities. SHAP values are computed for these data points, and the absolute values are aggregated and sorted to create the Positive Class Ranked List.
- Negative Class Ranked List: The background dataset consists of data points with strong negative probabilities, and the foreground dataset contains data points with weak negative probabilities. SHAP values are computed for these data points, and the absolute values are aggregated and sorted to form the Negative Class Ranked List.

The robustness-based lists are then sorted in order of SHAP values, such that the most important variables, which are to be retained, are those with the highest SHAP values. Intuitively, given the conservative nature of the SHAP explanations, we are retaining features that are contributing (on average) the most to make non-robust points more robust while suppressing features that are undermining the robustness of the outputs.

Final Robustness Ranked List: We aggregate the Positive Class Ranked List and the Negative Class Ranked List to obtain the final list of variables for robustness. The sum of the ranks from the two lists is sorted in ascending order. Equal weight is assigned to both lists in this process; however, this weighting is a hyperparameter that can be adjusted based on specific application requirements.

4.5 Final Feature List

From the preceding subsections, we derive three ranked lists, each focusing on a specific criterion. The variables in each list are assigned positional values for the purpose of aggregation, where the first variable in each list is valued at 1, the second at 2, and so forth. We then assign weights to each list based on the application and the relative importance of each criterion. These weighted lists are combined and sorted to produce the final ranked list:

$$\text{Final Ranked List} = \alpha \cdot \text{PL} + \beta \cdot \text{FL} + \gamma \cdot \text{RL} \tag{8}$$

where α, β and γ represent the weights assigned to the Performance List (PL), Fairness List (FL), and Robustness List (RL), respectively.

This final ranked list integrates considerations from performance, fairness, and robustness. It is compared against a commonly used method of feature selection, which involves randomly choosing examples for the foreground and background to generate a feature list.

5 Experiments and Results

5.1 Setup

To comprehend and assess the performance of the proposed methodology, we utilize three binary classification financial datasets. We conduct a comparative analysis with the existing feature selection method using SHAP. Our primary objective is to investigate the potential benefits of incorporating context-aware backgrounds and foregrounds in SHAP for responsible feature selection, while maintaining performance standards. The prevailing industry practice involves randomly sampling data points from the training dataset to establish the background distribution. This approach serves as our baseline for comparison.

Datasets: For our experiments, we select three publicly available datasets: **Adult**, **Bank Marketing** and **German Credit**. Each of these datasets contains sensitive attributes for which they are known to exhibit bias.

ADULT: This dataset comprises Census information and aims to predict whether an individual's income exceeds $50,000. The protected attribute in this dataset is gender. Additionally, the race variable is removed during pre-processing.

BANK MARKETING: This dataset, derived from the marketing efforts of a Portuguese banking institution between 2008 and 2013, aims to predict whether a customer will subscribe to a term deposit. The protected attribute in this dataset is marital status [13].

GERMAN CREDIT: This dataset consists of 20 features and 1,000 rows, with each row representing an individual who has applied for credit. The dataset classifies applicants as either good or bad credit risks based on a set of attributes. The protected attribute in this dataset is gender.

Models: For each dataset, we trained XGBoost and Logistic Regression models to illustrate the model agnostic nature of our methodology. Studies show that the for tabular data, tree based models like XGBoost [8] are the state of the art for performance [26] that has great relevance in the finance industry [14]. Protected attributes are removed before training of the model as is required by regulations in many places.

Metrics: The models generated using the list derived from our proposed methodology, as well as those based on SHAP-based feature selection, are evaluated according to three criteria: performance, fairness, and robustness. Performance is measured by the Area Under the Curve (AUC), fairness is assessed using the parity difference metric, and robustness is evaluated based on the distance from the decision boundary as discussed in the previous section. For both the baseline and our method, features are selected based on the AUC plots. Specifically, we select the model with the minimum number of features that surpasses a predetermined AUC threshold.

5.2 Results

Table 1 presents the consolidated results of comparing the Context-aware SHAP method to the conventional SHAP-based feature selection method. Figure 4 illustrates the plots of all models constructed, from which the final model for each method was selected for comparison. The values in bold represent the best result for each evaluation criterion. Across datasets, higher values of AUC and robustness indicate better performance, while lower values of disparity indicate better fairness. The thresholds are selected based

Table 1: Results comparing Context-aware SHAP to classical SHAP across models and datasets

Dataset	Model	Method	Number of Features	AUC	Fairness	Robustness
Adult	Logistic Regression	SHAP	**3**	0.810	0.098	0.306
Adult	Logistic Regression	Context-aware SHAP	4	**0.818**	**0.070**	**0.315**
Adult	XGBoost	SHAP	**5**	0.905	0.161	0.358
Adult	XGBoost	Context-aware SHAP	6	**0.922**	**0.160**	**0.373**
Bank	Logistic Regression	SHAP	8	**0.860**	0.015	0.398
Bank	Logistic Regression	Context-aware SHAP	**6**	0.852	**0.013**	**0.404**
Bank	XGBoost	SHAP	9	**0.952**	0.038	0.407
Bank	XGBoost	Context-aware SHAP	**8**	**0.952**	**0.028**	**0.415**
German Credit	Logistic Regression	SHAP	**1**	0.655	**0.00**	0.212
German Credit	Logistic Regression	Context-aware SHAP	2	**0.667**	**-0.043**	**0.220**
German Credit	XGBoost	SHAP	**4**	0.630	0.020	0.209
German Credit	XGBoost	Context-aware SHAP	5	**0.670**	**-0.038**	**0.211**

on a minimal drop in the AUC relative to the starting reference point of the all-feature model for each case.

Adult: Table 1 gives us the results for Logistic Regression and XGBoost models on Adult data. Gender is the protected variable. We choose model based on them reaching a minimum of 0.9 AUC for XGBoost and 0.8 for Logistic Regression. We give equal weighting for α, β and γ while arriving at the final list. For logistic regression and XGBoost we notice that the Context-aware SHAP method gives better results on all fronts with just choosing one more variable than the classic SHAP. For the logistic regression case especially, we see a massive improvement in fairness and a good improvement in robustness. In XGBoost we see an improvement in AUC and robustness with comparable disparity value.

Banks: For the Bank Marketing dataset, as referenced in Table 1, the protected variable is marital status. Models are selected based on achieving a minimum AUC of 0.95 for XGBoost and 0.85 for Logistic Regression. Equal weights are assigned to α, β, and γ for XGBoost, while weights of 1.5, 2, and 1 are assigned respectively for Logistic Regression. For Logistic Regression, the Context-aware SHAP method selects a smaller number of variables and performs better in terms of fairness and robustness, although it exhibits a slightly lower AUC. For XGBoost, both Context-aware SHAP and conventional SHAP achieve the same level of AUC; however, the proposed methodology selects fewer features and demonstrates superior performance in fairness and robustness.

German Credit: Table 1 summarizes the results for the German Credit dataset, where the protected variable is gender. Models are selected based on achieving a minimum AUC of 0.60 for XGBoost and 0.65 for Logistic Regression. Equal weights are assigned to α, β, and γ for XGBoost, while weights of 1, 2, and 2 are assigned respectively for Logistic Regression. For both Logistic Regression and XGBoost, the Context-aware SHAP method yields better results across all evaluation criteria, despite selecting only one additional variable compared to the classic SHAP method. Any disparity value equal to or less than zero indicates a non-discriminatory model against the protected class.

Overall, the results illustrate how Responsible Feature Selection can help achieve better or similar levels of performance while improving fairness and robustness metrics, all with a comparable number of variables.

6 Conclusion

This work proposes a SHAP-based methodology for responsible feature selection under multiple criteria. We demonstrate the significance of the background and foreground data used for SHAP computations and its impact on feature selection. The method is applied to three datasets, using both Logistic Regression and XGBoost. Its efficacy is compared to the conventional method of a background using random sampling. We demonstrate that we can identify models with similar or better performance in most cases, while simultaneously improving fairness and robustness metrics. The proposed method is extendable to other models and can be applied to different criterion with the appropriate choice of background and foreground data.

7 DISCLAIMER

in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

[1] 116th Congress (2019-2020). 2019. Algorithmic Accountability Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231 Accessed on August 2 2024.

[2] Emanuele Albini, Jason Long, Danial Dervovic, and Daniele Magazzeni. 2022. Counterfactual shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1054–1070.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[4] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389* (2020).

[5] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 648–657.

[6] Dillon Bowen and Lyle Ungar. 2020. Generalized SHAP: Generating multiple types of explanations in machine learning. arXiv:2006.07155 [cs.LG] https://arxiv.org/abs/2006.07155

[7] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & electrical engineering* 40, 1 (2014), 16–28.

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[9] Ian Covert, Scott M Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems* 33 (2020), 17212–17223.

[10] Pradip Dhal and Chandrashekhar Azad. 2022. A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence* 52, 4 (2022), 4543–4581.

[11] Matthew F Dixon, Igor Halperin, and Paul Bilokon. 2020. *Machine learning in finance*. Vol. 1170. Springer.

[12] Johannes Haug, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. 2021. On Baselines for Local Feature Attributions. arXiv:2101.00905 [cs.LG] https://arxiv.org/abs/2101.00905

[13] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 3 (2022), e1452.

[14] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* 247, 1 (2015), 124–136.

[15] Mingxuan Liu, Yilin Ning, Han Yuan, Marcus Eng Hock Ong, and Nan Liu. 2022. Balanced background and explanation data are needed in explaining deep learning models with SHAP: An empirical study on clinical decision making. *arXiv preprint arXiv:2206.04050* (2022).

[16] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:1705.07874 [cs.AI] https://arxiv.org/abs/1705.07874

[17] Wilson E Marcílio and Danilo M Eler. 2020. From explanations to feature selection: assessing SHAP values as feature selection mechanism. In *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*. Ieee, 340–347.

[18] Luke Merrick and Ankur Taly. 2020. The Explanation Game: Explaining Machine Learning Models Using Shapley Values. arXiv:1909.08128 [cs.LG] https://arxiv.org/abs/1909.08128

[19] Raha Moraffah, Paras Sheth, Saketh Vishnubhatla, and Huan Liu. 2024. Causal Feature Selection for Responsible Machine Learning. *arXiv preprint arXiv:2402.02696* (2024).

[20] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[21] Lukas Ryll and Sebastian Seidens. 2019. Evaluating the performance of machine learning algorithms in financial market forecasting: A comprehensive survey. *arXiv preprint arXiv:1906.07786* (2019).

[22] Jaydip Sen, Rajdeep Sen, and Abhishek Dutta. 2022. Introductory chapter: machine learning in finance-emerging trends and challenges. *Algorithms, Models and Applications* (2022), 1.

[23] Shubham Sharma, Sanghamitra Dutta, Emanuele Albini, Freddy Lecue, Daniele Magazzeni, and Manuela Veloso. 2023. REFRESH: Responsible and Efficient Feature Reselection guided by SHAP values. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 443–453.

[24] Shubham Sharma, Alan H. Gee, David Paydarfar, and Joydeep Ghosh. 2020. FaiR-N: Fair and Robust Neural Networks for Structured Data. arXiv:2010.06113 [cs.LG] https://arxiv.org/abs/2010.06113

[25] Sheojung Shin, Peter C Austin, Heather J Ross, Husam Abdel-Qadir, Cassandra Freitas, George Tomlinson, Davide Chicco, Meera Mahendiran, Patrick R Lawler, Filio Billia, et al. 2021. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC heart failure* 8, 1 (2021), 106–115.

[26] Ravid Shwartz-Ziv and Amitai Armon. 2021. Tabular Data: Deep Learning is Not All You Need. arXiv:2106.03253 [cs.LG] https://arxiv.org/abs/2106.03253

[27] European Union. 2018. EU Privacy Regulation. https://www.privacy-regulation.eu/en/recital-71-GDPR.htm Accessed on August 2 2024.

[28] Han Yuan, Mingxuan Liu, Lican Kang, Chenkui Miao, and Ying Wu. 2022. An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models. *arXiv preprint arXiv:2204.11351* (2022).

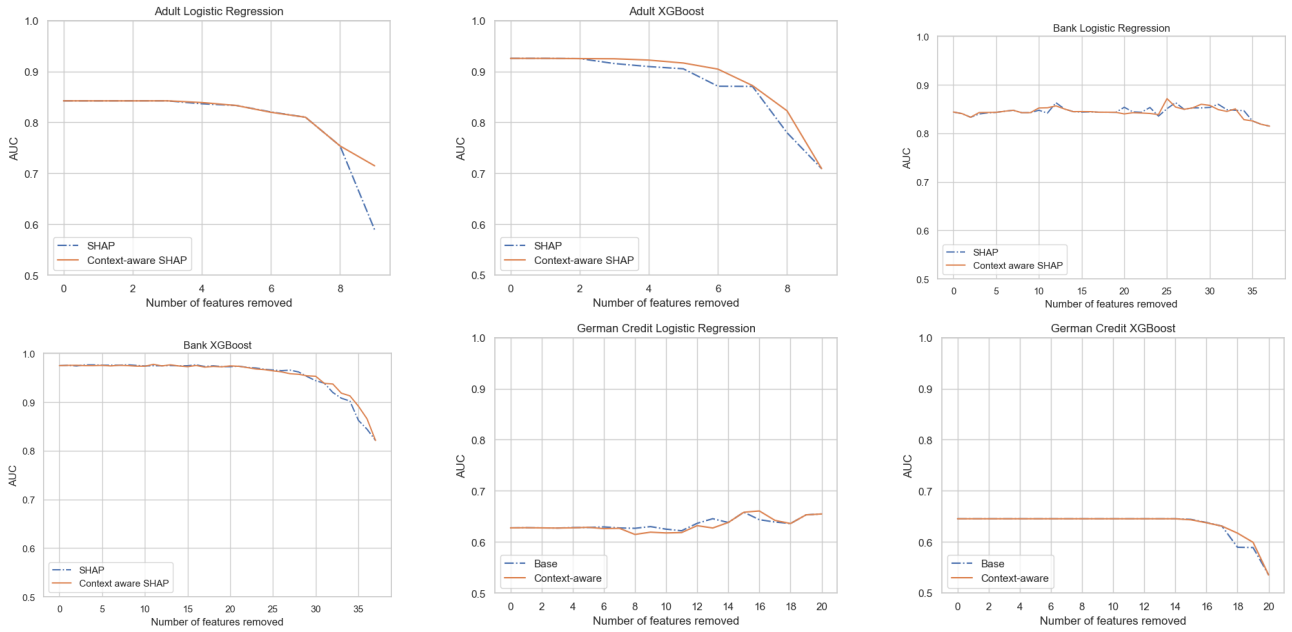

Figure 3: Comparison of Performance ranked list from Context-aware SHAP and SHAP based ranked list

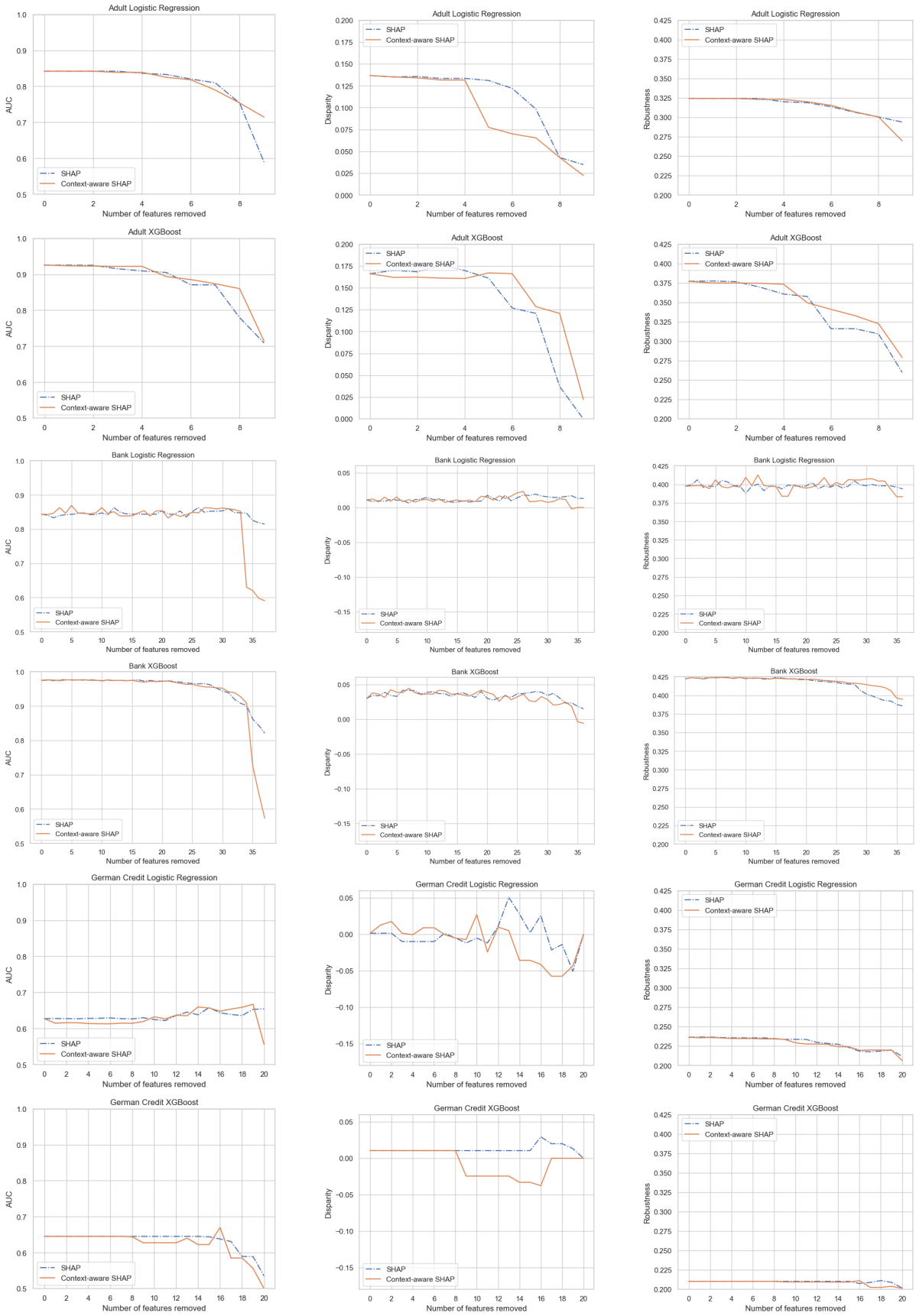

Figure 4: Effect of sequential features dropping/removal on the performance, fairness and robustness of the model. The X axis shows the number of features removed from all features. We plot the performance, fairness and robustness along the Y axis. The orange line indicates the data of Context-aware method and the dotted blue line represents the conventional method.