Can Reduction of Bias Decrease the Need for Explainability? Working with Simplified Models to Understand Complexity

Pedro M. Marques[1,*]

[1]Univ Coimbra, CeBER, Faculty of Economics, Av Dias da Silva 165, 3004-512 Coimbra

Abstract

The research delves into the complexity of data analysis models, emphasizing the critical need for explainability, especially in AI-driven sectors. It aims to investigate the impact of bias reduction on explainability and explores how simplified models can be used to provide functionally equivalent results while being easier to understand and accept. By integrating statistical techniques and bias reduction methods, we expect to increase the acceptability of the models by lessening the fear of biased outcomes. Through quantitative and qualitative analysis, it evaluates the effectiveness of simplification techniques in promoting transparency and comprehension. Stakeholder involvement and ethical understanding are central to this approach. The research intends to contribute transparency in data analysis, addressing critical societal challenges.

Keywords

Bias, Explainability, Simplified models

1. Research Context and Motivation

The research addresses the complexity of data analysis models, with a focus on enhancing their explainability. This need for explainability is crucial as AI systems become more precise and are used in critical societal sectors, where understanding their decision-making process is essential for ethical considerations. The study explores two key aspects: first, whether reducing bias in models can diminish the need for explainability; and second, how simplified models can enhance the understanding of complex models. The goal is to integrate these aspects to improve transparency and understanding in complex systems, contributing to technological advancement.

The reduction of bias is crucial for ensuring fairness and justice in complex data analysis systems [1]. Studies have shown that reducing bias can lead to more intuitive and predictive models, thereby reducing the need for extensive explanations [2]. This suggests that less biased models can alleviate the burden of providing detailed justifications in situations where fairness is warranted. Additionally, the use of simplified models has been proposed to address the inherent

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

pedro.marques@student.fe.uc.pt (P. M. Marques)

https://www.uc.pt/feuc/ (P. M. Marques)

0009-0002-4849-0988 (P. M. Marques)

complexity of advanced AI models [3]. Simplified models enhance interpretability, making them more accessible to non-specialized professionals and stakeholders, thereby contributing to user confidence and acceptability.

Integrating the reduction of bias and the use of simplified models as complementary strategies can improve transparency and understanding in complex systems [4]. It is important to recognize that bias is not exclusive to machine learning but exists in other data analysis techniques as well [5]. Simplicity, explainability, and bias are relevant concepts across different levels of data analysis. Each context may require different explanations and interpretations, highlighting the importance of explainability. [1] stresses the importance of combating biases in data systems to ensure fairness and accuracy. Similarly, [5] emphasizes the need for explainability to make system decisions comprehensible. These measures aim to promote social justice and equality, necessitating a multidisciplinary approach to achieve these objectives.

2. Key related work that frames the research

This paper utilizes several research elements from the articles cited later where aspects of technology adoption, AI bias, model simplification, and validation of predictive models. [6] provides an overview of technology adoption models, crucial for understanding user acceptance of technologies in academic and business contexts. [7] explore prejudice and discrimination in AI, emphasizing interdisciplinary collaboration for effective solutions. [8] investigate model complexity and performance, demonstrating the ability of simplified models to capture ecosystem characteristics, emphasizing the importance of key processes. [9] assess the methodological quality of external validation studies of forecasting models, revealing poor communication, and handling practices that can hinder their use. Finally, [10] examine the effects of bias reduction on treatment effect estimates, challenging the assumptions of standard selection models and revealing complexities in bias reduction methods.

3. Specific research questions, hypothesis, and objectives

3.1. What influence does the reduction of bias have on the need to explain the results obtained?

Mitigating bias in data analysis can lead to more robust outcomes and reduce the need for clarification of underlying processes [11]. Researchers emphasize the importance of minimizing bias in scientific research to ensure the reliability of results. Ethical considerations require researchers to acknowledge study limitations and potential sources of bias. By reducing bias, models can be seen as more transparent and easier to understand, minimizing the effort required for interpretation [11]. The research aims to demonstrate the absence or reduction of bias in specific models, validating the reliability of results despite inherent complexity. Large data volumes do not guarantee representativeness or the absence of bias, as they may still retain biases that affect model precision [12, 10]. Therefore, models should be designed or improved to show the absence or reduction of specific types of bias.

There are various ways by which bias can be present in a data analysis approach, whether

statistical or ML-based. Among the types of bias we propose to analyze in this study, (which may vary depending on the context) we may consider: 1) Algorithmic bias; 2) Sample Bias; 3) Measurement bias; 4) Exclusion bias; 5) Selection bias; 6) Recall bias.

3.2. How can simplified models contribute to a better understanding and explainability in contexts where complex models are used?

The underlying hypothesis is that integrating simplified models results alongside complex models can be effective and make explainability more accessible. It is assumed that simplifying models can provide more comprehensible insights into the decisions made by the decision-maker relative to complex models, facilitating interpretation and increasing confidence in predictions.

Expanding explanations related to these models can increase acceptance of results and enhance confidence in decision-making. By demonstrating how simplified models represent the workings of general models, reliability and confidence in decisions can be bolstered [12].

The objectives defined for the study are: **Develop Simplified Models:** Create simplified models that concisely and accessibly represent the essential characteristics of the complex models used; **Analysing the Reliability of Results:** To compare the reliability and accuracy of the results obtained by models subjected to bias reduction processes with those obtained by conventional models; **Evaluation of the Need for Explainability:** Analyse whether the absence/reduction of particular kinds of bias is correlated with a reduction in the need for additional explainability techniques; **Identifying Correlations between Bias and Explainability:** Investigating whether there is a correlation between the presence of bias and the need to make models more explainable, seeking to understand how the perception of bias can influence this relationship. The work addresses the challenge of understanding models with many variables. It suggests the creation of simplified cases to elucidate model behaviour. Simplified models with fewer variables may present higher bias but can explain decisions effectively. The research explores how including simplified models can improve the explainability of complex models, making their decisions more transparent to stakeholders.

4. Research approach, methods, and rationale for testing the research hypothesis

With data-driven technological models significantly impacting our daily lives and influencing various critical aspects of society, we live through a crucial moment in human history. As [13], point out, business analytics involves three phases: describing, predicting, and prescribing to make data-driven decisions, considering both normative and descriptive approaches.

As technologies such as AI increase in complexity, so does the need for regulation. Various legislation emphasising AI systems' transparency, explainability and safety regulate the development and use of AI. The European Parliament's Artificial Intelligence Regulation [14] proposes that high-risk AI systems must be transparent, understandable, and explainable to suppliers and users, promoting trust and responsibility. Executive Order 14.110 of the United States [15] addresses security, privacy, transparency, and impact on the use of AI, emphasising the mitigation of bias and the need for interagency coordination. GDPR [16] is the mechanism

proposed by the European Union to protect citizens' personal data. It requires clear explanations of AI decisions, mandatory informed consent, data minimisation, and security through encryption and anonymisation. These regulations aim to ensure that AI is developed and used ethically and responsibly, maximising benefits and minimising risks. The explainability of AI is crucial to understanding its decision-making processes, be it deterministic or probabilistic, as stated by [17].

To select the most suitable simplification method for an AI model, various factors such as interpretability, performance, robustness, scalability, acceptability, and transparency must be considered. For example, [18] emphasizes balancing interpretability with performance to maintain user confidence. Also, in [19], techniques are discussed that preserve robustness and generalizability across different datasets. Strategies for scalable simplification methods handling large datasets are analyzed in [20]. The importance of user acceptability and trust is highlighted by [21], who discuss approaches that enhance transparency and explainability. Additionally, [22] addresses algorithm bias, stressing the need for practical and accessible explainability, while [23] underlines the ethical considerations in ensuring transparency through simplification methods.

When simplifying a complex model, it's crucial to consider performance metrics, interpretability, robustness, and bias reduction. [18] discusses performance metrics for evaluating simplified models, emphasizing the need to preserve relevant information. [20] highlights the importance of ensuring that simplified models remain interpretable for users. The robustness of simplified models against biases and the ethics of their decisions are covered by [22]. [20] also stresses the need to assess the generalizability and acceptability of simplified models, which are key to the validation process. According to [23], transparency evaluation requires collaboration between developers and researchers to build confidence in the model's predictions. Reducing bias and promoting fairness in decisions through simplification are supported by [24]. It is essential to balance accuracy, reliability, interpretability, and efficiency when simplifying complex models, as highlighted by [12]. This involves comparing the performance of simplified models with the original ones to ensure minimal accuracy loss. Careful feature selection, external validation, and stakeholder involvement are recommended to mitigate accuracy loss, according to [12]. Lastly, [25] warns that large data volumes don't guarantee representativeness or absence of bias, while [26] discusses how sparsity constraints can affect accuracy during complexity reduction. To enhance interpretability and applicability, approaches like sparsity and feature selection are key to reducing complexity without compromising accuracy. To improve the interpretability and applicability of ML techniques in various domains, it is necessary to assess the extent to which models can be simplified, and it is essential to consider approaches such as sparsity and feature selection, which play a key role in reducing complexity without compromising accuracy.

Considering sparsity is a widely used technique that aims to reduce the number of active parameters in the model, making it simpler [27]. Through methods such as L1 regularisation (lasso), it is possible to promote simplification of the model while maintaining its effectiveness in the task at hand [28]. Feature selection is another important strategy that can be applied for simplifying complex models. By choosing only the most relevant features for the ML task, it is possible to reduce the dimensionality of the model and make it easier to interpret the results. Methods such as feature importance analysis and recursive feature elimination are commonly used to identify and select the most significant features [29] Finding the right balance between

interpretability and model performance is crucial. Oversimplification can lead to significant losses in accuracy, while excessive complexity can make it difficult to interpret the results, according to [30]. Therefore, constantly evaluating the balance between interpretability and performance is key to ensuring that the simplified model keeps its effectiveness in reducing bias and making decisions [30]. Integrating model simplification and explainability in AI is crucial for building trust in AI systems. According to [28], transparency and understandability are essential for users to trust automated decisions, especially in critical sectors like health, finance, and justice, where decisions have significant impacts. Simplifying complex models makes AI more accessible and understandable, enabling users to grasp how algorithms work and the rationale behind predictions. Explainability provides insights into decision-making, as noted by [28], allowing for a more informed assessment of AI recommendations. Reducing bias and discrimination through simplification and explanation is vital for ensuring fairness in AI decisions, as highlighted by [31]. Regular audits may be needed to identify and correct biases in AI explanations, preventing manipulative use and promoting fairness. Mathematical techniques to mitigate bias, as discussed by [32], are important for improving model accuracy and generalizability. Various strategies have been proposed to reduce bias, with [32] emphasizing the importance of systematic approaches involving hypothesis formulation, data collection, statistical analysis, and continuous improvement. Scientific integrity, as noted by [33], is key to ensuring reliable conclusions in bias mitigation efforts.

The bias can be multifaceted, as can be the tools to manage it. The aim is, therefore, to use an "ensemble" approach to bias reduction, which involves integrating methods to build a model that minimises unwanted bias characteristics. This technique involves combining and applying individual methods, such as resampling and weight adjustment, and evaluating the performance of each method before making a weighted combination. According to [30] ensemble methods, which combine multiple models to improve the accuracy and stability of forecasts, can simplify interpretation by providing a more robust view of the problem. However, the use of ensemble methods can increase computational complexity and require more processing resources. The approach is intended to be iterative, adjusting parameters and weightings to optimise effectiveness, while continuous ethical evaluation is essential to ensure the fairness of the resulting model. Thus, the effectiveness of this approach depends on the careful selection of individual methods and their constant adaptation to respond to the specifics of the data analysis scenario. The method adopted will integrate statistical and mathematical techniques to improve the models, ensuring they produce the desired results and are faithful to the original models. In addition, validation methods will be used, such as the root mean square error (RMSE), the coefficient of determination (R^2) and the Mann-Whitney U-test. In the context of increasing explainability, explanatory methods can be used, such as LIME (Local Interpretable Model-agnostic Explanations), which is a post-hoc interpretability technique that explains the predictions of complex models using simple local models, as presented in [34]. The SHAP Technique (SHapley Additive exPlanations), based on game theory that assigns importance values to each input feature to explain the predictions of the models [34]. We can also use decision trees as inherently interpretable ML models that can be used to simplify the decision logic of complex models [22]. Linear regression is also a simple and interpretable model often used to simplify the relationship between input and output variables in complex models [22].

As a complement to understanding the models, model visualisation techniques can also be

used, such as decision tree graphs, which show the decision rules adopted by the model; partial dependency graphs, which illustrate the relationship between a specific variable and the model's prediction; and heatmaps, which show the relative importance of each variable in the model's prediction.

The proposed research consists of a quantitative and a qualitative phase. In the quantitative phase, mathematical models are investigated, while in the qualitative phase, methodologies are used to help decision-makers understand existing biases, as highlighted by [35] and [20]. This approach acknowledges the need for decision-makers to make informed choices, even when biases are present in models. It's important to note that the presence of bias does not automatically correlate with the model's utility or the degree to which the bias is decisive. Bias affects the model's outcomes without necessarily impacting decision-maker utility. Describing bias is easier than assessing its utility since utility assessments are inherently subjective. Therefore, care should be taken when discussing the favourability or unfavourability of a bias-reduced model, as it suggests a relationship between bias and utility. A qualitative analysis should be conducted to evaluate the acceptability of models with reduced bias.

To operationalise this analysis, adoption models are suggested. To use Technology Adoption Models (TAM), such as Rogers' Diffusion of Innovations Model and the Unified Theory of Acceptance and Utilisation of Technology (UTAUT) in research, it is necessary to consider the scope and objectives of the research [36].The research proposes using TAM and UTAUT to validate whether decision-makers will be willing to accept the simplified models' results as a general description of how the complex model works. [6] presents a wide range of models and warns of the need for careful identification of the key variables and the proposed questions (e.g. "Perceived Usefulness" or "Ease of Use"), carried out through the theoretical models, which inform the methodological design of the research, and the data collected by questionnaire. The data will be analysed and interpreted considering the principles of TAM and UTAUT. This research aims to address societal challenges by promoting transparency in data analysis by offering a path to informed decision-making and accountability. It intends to offer a sound theoretical framework for assessing the acceptance and utilisation of technology, including how users perceive and adopt the results of artificial intelligence models. As discussed by [20], transparency in data analysis is key to promoting trust and understanding of the results generated by those models.

5. Expected next steps

The use of simplified models alongside complex models can be a valuable strategy for improving understanding in complex contexts. Simplification techniques, such as reducing dimensionality, facilitate human interpretation and allow more effective identification and mitigation of bias. To address research gaps, specific simplification techniques should be investigated. The interpretability of complex and simplified models in different domains should be compared. Operationalization involves the careful selection of complex models, the precise definition of simplification criteria, and the development of interpretability metrics for objective assessment. This research can contribute to the understanding of complex models and simplification techniques and provide practical guidance for researchers and practitioners in several domains,

but our focus will be particularly business and finance, where interpretability is critical for decision-making.

In addition, we expect that the creation of practical tools and resources, guided by the active participation of users, will enable the efficient implementation of these techniques in practice. Continuous feedback from users, combined with the systematic publication of results in specialized scientific forums, will foster the dissemination of knowledge and contribute to the ongoing evolution of this area of research.

References

[1] E. Ferrara, Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies, Sci 6 (2023) 3.

[2] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, 2021.

[3] K. Fauvel, V. Masson, E. Fromont, V. M. andÉlisaand, andÉlisa Fromont, A performance-explainability framework to benchmark machine learning methods: Application to multivariate time series classifiers, 2021.

[4] G. B. Mensah, Artificial intelligence and ethics: A comprehensive review of bias mitigation, transparency, and accountability in ai systems (2023).

[5] J. Sreerama, G. Krishnamoorthy, Ethical considerations in ai: Addressing bias and fairness in machine learning models, Journal of Knowledge Learning and Science Technology 1 (2023) 2959–6386.

[6] H. Taherdoost, A review of technology acceptance and adoption models and theories, volume 22, Elsevier B.V., 2018, pp. 960–967.

[7] X. Ferrer, T. V. Nuenen, J. M. Such, M. Cote, N. Criado, Bias and discrimination in ai: A cross-disciplinary perspective, 2021.

[8] C. Raick, K. Soetaert, M. Grégoire, Model complexity and performance: How far can we simplify?, Progress in Oceanography 70 (2006) 27–57.

[9] G. S. Collins, J. A. D. Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L. M. Yu, K. G. Moons, D. G. Altman, External validation of multivariable prediction models: A systematic review of methodological conduct and reporting, 2014.

[10] T. A. Diprete, H. Engelhardt, Estimating causal effects with matching methods in the presence and absence of bias cancellation, Sociological Methods and Research 32 (2004) 501–528.

[11] J. Smith, H. Noble, Bias in research, Evidence-Based Nursing 17 (2014) 100–101.

[12] A. Aldoseri, K. N. Al-Khalifa, A. M. Hamouda, Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges, 2023.

[13] G. Mentzas, K. Lepenioti, A. Bousdekis, D. Apostolou, Data-driven collaborative human-ai decision makinkg, Springer, 2021, pp. 139–151.

[14] P. Europeu, Lei da ue sobre ia: primeira regulamentação de inteligência artificial x (2023).

[15] J. Biden, Safe, secure, and trustworthy development and use of artificial intelligence, 2023.

[16] RGPD, General data protection regulation, 2016.

[17] J. Chaquet-Ulldemolins, F. J. Gimeno-Blanes, S. Moral-Rubio, S. Muñoz-Romero, J. L. Rojo-álvarez, On the black-box challenge for fraud detection using machine learning (i): Linear models and informative feature selection, Applied Sciences (Switzerland) 12 (2022).

[18] D. L. Kitane, Sparsity in machine learning: Theory and applications (2022).

[19] Y. Zhang, C. Wu, Y. Tian, X. Zhang, A co-evolutionary algorithm based on sparsity clustering for sparse large-scale multi-objective optimization, Engineering Applications of Artificial Intelligence 133 (2024).

[20] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, B. Sikdar, A review of trustworthy and explainable artificial intelligence (xai), IEEE Access 11 (2023) 78994–79015.
[21] T. Lazebnik, S. Bunimovich-Mendrazitsky, A. Rosenfeld, An algorithm to optimize explainability using feature ensembles, Applied Intelligence 54 (2024) 2248–2260.
[22] G. M. Johnson, Algorithmic bias: on the implicit biases of social technology, Synthese 198 (2021) 9941–9961.
[23] J. S. de Souza, J. M. Abe, L. A. de Lima, N. A. de Souza, The brazilian law on personal data protection, International Journal of Network Security & Its Applications 12 (2020) 15–25.
[24] C. S. Withers, S. Nadarajah, Bias reduction when data are rounded, Statistica Neerlandica 69 (2015) 236–271.
[25] H. H. J. Senetaire, D. Garreau, J. Frellsen, P.-A. Mattei, Explainability as statistical inference, arXiv v3 (2022).
[26] S. Dolgikh, Sparsity constraint in unsupervised concept learning, CEUR-WS:, 2022.
[27] T. Hastie, R. T. Martin, W. Hastie, Tibshirani, Wainwright, Statistical Learning with Sparsity - The Lasso and Generalizations Statistical Learning with Sparsity, 1 ed., CRC Press, 2016.
[28] N. Tiwary, S. A. M. Noah, F. Fauzi, T. S. Yee, Max explainability score–a quantitative metric for explainability evaluation in knowledge graph-based recommendations, Computers and Electrical Engineering 116 (2024).
[29] Y. Yao, Y. Pan, J. Li, I. W. Tsang, X. Yao, Sanitized clustering against confounding bias, Machine Learning 113 (2024) 3711–3730.
[30] C. Kotrachai, P. Chanruangrat, T. Thaipisutikul, W. Kusakunniran, W. C. Hsu, Y. C. Sun, Explainable ai supported evaluation and comparison on credit card fraud detection models, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 86–91.
[31] G. Vilone, L. Rizzo, L. Longo, A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence, 2020.
[32] R. Nishant, D. Schneckenberg, M. N. Ravishankar, The formal rationality of artificial intelligence-based algorithms and the problem of bias, Journal of Information Technology 39 (2024) 19–40.
[33] E. Prem, From ethical ai frameworks to tools: a review of approaches, AI and Ethics 3 (2023) 699–716.
[34] P. E. Love, W. Fang, J. Matthews, S. Porter, H. Luo, L. Ding, Explainable artificial intelligence (xai): Precepts, models, and opportunities for research in construction, 2023.
[35] R. González-Sendino, E. Serrano, J. Bajo, P. Novais, A review of bias and fairness in artificial intelligence, International Journal of Interactive Multimedia and Artificial Intelligence In press (2023) 1.
[36] V. Venkatesh, J. Y. Thong, X. Xu, Unified theory of acceptance and use of technology: A synthesis and the road ahead, Journal of the Association for Information Systems 17 (2016) 328–376.