Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

A context-aware decision support system for selecting explainable artificial intelligence methods in business organizations

Marcelo I. Reis [a,b,*], João N.C. Gonçalves [c], Paulo Cortez [a], M. Sameiro Carvalho [d], João M. Fernandes [e]

[a] ALGORITMI/LASI, Department of Information Systems, University of Minho, Guimarães, Portugal
[b] Universidade Católica do Salvador, Salvador, Brazil
[c] Universidade Católica Portuguesa, CEGE – Research Centre in Management and Economics, Portugal
[d] ALGORITMI/LASI, Department of Production and Systems, University of Minho, Braga, Portugal
[e] ALGORITMI and CCG/ZGDV Institute, Department of Informatics, University of Minho, Braga, Portugal

ARTICLE INFO

Keywords:
Decision support system
Explainable AI
Machine learning
Business analytics
Supply chain

ABSTRACT

Explainable Artificial Intelligence (XAI) methods are valuable tools for promoting understanding, trust, and efficient use of Artificial Intelligence (AI) systems in business organizations. However, the question of how organizations should select suitable XAI methods for a given task and business context remains a challenge, particularly when the number of methods available in the literature continues to increase. Here, we propose a context-aware decision support system (DSS) to select, from a given set of XAI methods, those with higher suitability to the needs of stakeholders operating in a given AI-based business problem. By including the human-in-the-loop, our DSS comprises an application-grounded analytical metric designed to facilitate the selection of XAI methods that align with the business stakeholders' desiderata and promote a deeper understanding of the results generated by a given machine learning model. The proposed system was tested on a real supply chain demand problem, using real data and real users. The results provide evidence on the usefulness of our metric in selecting XAI methods based on the feedback and analytical maturity of stakeholders from the deployment context. We believe that our DSS is sufficiently flexible and understandable to be applied in a variety of business contexts, with stakeholders with varying degrees of AI literacy.

1. Introduction

1.1. Background and research motivation

Artificial Intelligence (AI) has contributed to significant changes in business and management processes within organizations, allowing the extraction of knowledge from data and improving decision-making processes in several contexts (Wamba-Taguimdje et al., 2020). Examples include a wide range of applications of AI techniques in manufacturing (Dengler et al., 2021), service quality (Guo et al., 2023), finance (Roeder et al., 2022) and supply chain management (Toorajipour et al., 2021), to name a few. The adoption of AI has grown considerably over the years, both in the public and private sectors, and recent studies (Gangwani and Zhu, 2024; Brem et al., 2021) have shown that organizations that take advantage of AI are more able to attract investment. Indeed, a Forbes article (Haan and Watts, 2023) published in 2023 reports that the AI global market size is expected to reach $407 billion by 2027, with expectations of compound annual growth rate of 37.3% until 2030.

While the range of applications for AI techniques is quite extensive, organizations still face major challenges when it comes to relying on intelligent systems for automated data-driven decision-making, as well as to understanding how such systems should be designed and implemented to generate value (Toorajipour et al., 2021; Jan et al., 2023; Enholm et al., 2022; Burger et al., 2023). These challenges stem from several factors. Firstly, business processes that impact an organization performance are usually complex and generally depend on exogenous information and subjective human factors that are difficult to be modeled by AI systems. Secondly, a large portion of the machine learning (ML) algorithms proposed in the literature tend to be "black-boxes" when applied to real predictive and prescriptive decisions. While these algorithms are often more suitable than classical parametric techniques for modeling nonlinear dynamics, they are also less explainable (James

[*] Corresponding author at: Universidade Católica do Salvador, Salvador, Brazil.
E-mail addresses: mindior@gmail.com (M.I. Reis), jngoncalves@ucp.pt (J.N.C. Gonçalves), pcortez@dsi.uminho.pt (P. Cortez), sameiro@dps.uminho.pt (M.S. Carvalho), jmf@di.uminho.pt (J.M. Fernandes).

https://doi.org/10.1016/j.compind.2024.104233
Received 17 July 2024; Received in revised form 13 December 2024; Accepted 15 December 2024

et al., 2023) (in the sense of their lack of understanding by humans), thereby undermining user's trust in their adoption (Vermeire et al., 2021). In business practice, this lack of model explainability also poses challenges in terms of providing insightful explanations to business stakeholders (Zhang and Chen, 2020), who increasingly seek to understand the rationale behind the general functioning of AI systems.

The problems related to the explainability of AI systems fall within the so-called explainable artificial intelligence (XAI) domain (Arrieta et al., 2020), which has received an increasing attention in recent years (Ali et al., 2023; Vilone and Longo, 2021; Adadi and Berrada, 2018; Abusitta et al., 2024; De Bock et al., 2024; Mersha et al., 2024; Angelov et al., 2021) and application contexts, such as Industry (Tchuente et al., 2024), Healthcare (Salih et al., 2024; Tjoa and Guan, 2020), and Finance (Weber et al., 2024). XAI can be defined as the use of ML techniques that, on the one hand, produce more explainable ML models while maintaining their ability to generalize and, on the other hand, enable humans to understand, trust and manage AI-based systems (Minh et al., 2022; Kostopoulos et al., 2024). Previous research studies on XAI have proposed a comprehensive set of methods with a well-defined explanation purpose and designed for a specific target audience, ranging from technical/domain experts to lay users (Arrieta et al., 2020; Adadi and Berrada, 2018; Brasse et al., 2023). These methods differ in terms of their dependence on the AI model developed (e.g., model-specific/model-agnostic techniques) and the scope of the explanation (e.g., local/global explanations). However, the vast majority of AI business studies fail to address the question of how to choose the most suitable XAI method for a given AI problem and business context. This decision becomes increasingly challenging as more XAI techniques emerge from the literature. In addition to the difficulty in selecting an appropriate explainability method (Amarasinghe et al., 2023, 2024), the process of quantifying and comparing the degree of explainability of XAI methods is a subject that has received scarce attention (Sovrano and Vitali, 2023; Islam et al., 2020). The same applies to XAI implementations in real contexts, where there are not only difficulties in their adoption by end-users who are not AI experts, but also the problem of ensuring that these methods address the desiderata (also known as requirements, needs, expectations) of the business stakeholders, in such a way as to promote the creation of XAI systems with practical and not just theoretical relevance (Vermeire et al., 2021; Kotriwala et al., 2021; Langer et al., 2021; Jesus et al., 2021).

1.2. Research contributions and organization

In this paper, we focus our attention on the problem of choosing, in a given business context, the most appropriate XAI methods for a given trained predictive ML model. We are particularly interested in investigating how useful the explanations given by different XAI methods are to the end-users according to their requirements. To this end, we propose a context-aware decision support system (DSS) that incorporates user knowledge and feedback into an application-grounded metric that estimates the degree of trust-satisfaction of any XAI method, enabling to select those most suitable to a particular business environment and research problem. Our primary goal is to build a metric with a simple mathematical formulation that allows users with varying degrees of analytical maturity to select the XAI technique that best suit their explainability needs and the requirements of the business problem. Research has shown the importance of building systems that on the one hand can adapt to different problems, users and business scenarios (Pawlicki et al., 2024; Haque et al., 2023), and on the other hand can include qualitative evaluation components to promote the adoption of XAI techniques by users of the application context (Aliyeva and Mehdiyev, 2024; Mohseni et al., 2021; Hoffman et al., 2023b). By including the human-in-the-loop, we intend to foster the selection of XAI methods that are aligned with the business stakeholders' desiderata and that can actually improve decision-making processes in real-world business operations. The proposed DSS was applied and tested in a focus company from an automotive supply chain, and was developed to be an open-source project for interested academics and practitioners.

In a nutshell, our paper contributes to the XAI research by:

(1) **Closing the gap between theory and practice in business organizations:** We propose a systematic process for choosing, from a set of XAI methods, the ones best suited to a given AI use case in a particular real-world business context. Our strategy follows a human-in-the-loop approach (Tsiakas and Murray-Rust, 2022), promoting XAI methods that actually meet the business stakeholders' desiderata and that prove useful in their decision-making processes.
(2) **Estimating the extent of suitability to the business context:** We take advantage of user feedback, in the form of ratings for trust and explanation satisfaction, to introduce a context-aware DSS comprising an analytical metric that estimates the degree of trust-satisfaction of any XAI method. In addition to including domain users' feedback, this metric is designed to also reflect the nature of the users' expertise, thereby enabling the selection of XAI methods according to the users' level of analytical maturity and/or business knowledge in that particular context. To the best of our knowledge, we are the first to propose this kind of metric.
(3) **Empirical validating the system with a real-world application context:** In sharp contrast to the majority of the existing literature, we test the practical relevance of our research in a real business task, with real data and real users from one of the world leading organizations in the automotive electronics sector. For the first time in the literature, the applicability of a DSS for selecting XAI methods in the context of Supply Chain Management (SCM) is explored, while comparing feedback from domain experts with that of academic researchers.

In what follows, we present an overview of related literature (Section 2) that motivates the proposed DSS (characterized in Section 3). We proceed with the empirical evaluation of our approach (Section 4), using a real-world case study. In Section 5, we present the results from the empirical evaluation, proving the value and utility of the proposed system in a real operational context. Finally, we discuss some practical implications of our work in Section 6 and conclude in Section 7, outlining possible avenues for future research.

2. Related work

Our research is particularly related to a relevant stream of literature in the field of XAI: metrics for evaluating the quality and the utility of explanations generated by XAI methods.

Using metrics to evaluate XAI methods has become a central topic of research in the ML community (Vilone and Longo, 2021; van der Waa et al., 2021; Nauta et al., 2023; Pawlicki et al., 2024; Zhou et al., 2021; Al-Ansari et al., 2024; Doumard et al., 2023), given its importance in facilitating the effective implementation of AI-based systems in operational contexts. Doshi-Velez and Kim (2018) were pioneers in introducing a concise taxonomy to categorize different metrics typically used to evaluate XAI methods. Such evaluation process can be conducted with or without involving the human-in-the-loop.

2.1. Non-human based evaluation

By definition, this type of evaluation of XAI methods does not benefit from any kind of human involvement. In such a setting, functionally-grounded metrics are used, where the quality of explanations is objectively evaluated merely in terms of algorithmic measures. Examples of this type of measures include, but are not limited to, *fidelity*, *accuracy* and *algorithmic complexity* of the ML model used to derive the explanations (for details see Schwalbe and Finzel, 2023). Although interesting from a computational point of view, by minimizing time and costs associated with human experimentation (Doshi-Velez and Kim, 2018), functionally-grounded metrics suffer from a limitation: the inability to fully satisfy the interests, needs and expectations of end-users (hereinafter referred to as *domain experts*) who use the ML model being explained to support various business operations on a day-to-day basis. Notwithstanding its limitations, this type of metrics may however be suitable whenever the XAI methods under evaluation have already been properly tested and validated in environments involving humans (Doshi-Velez and Kim, 2018).

2.2. Human subject-based evaluation

Human subject-based evaluation allows explanations to be evaluated on the basis of human-based experimental procedures. In this context, two types of evaluation metrics can be employed: human-grounded and application-grounded metrics. While the former take advantage of simple lay-human-based experimental procedures that mimic the real context in which the explanations are intended to be used, the latter consider real humans and real tasks throughout the evaluation procedure. In the case of application-grounded metrics, the quality of a given explanation is evaluated in the context of its end-task, using domain experts (Doshi-Velez and Kim, 2018).

Some efforts have been made to develop explainability metrics and methodologies involving the human-in-the-loop. Hoffman et al. (2023b) provide an interesting set of recommendations and measures for selecting XAI methods that produce meaningful explanations for end-users. Each domain expert is also provided with a scoring system that allows them to evaluate the explanations given by different XAI methods according to different indicators. A second example is the work of Vermeire et al. (2021), who explored an initial version of a practical methodology to support developers in providing useful explanations for domain experts. Although relevant in terms of offering instruments to promote the evaluation and selection of XAI methods according to the business stakeholders' desiderata, none of the aforementioned works evaluates their proposals with real users and real business tasks. While previous research widely acknowledges that domain experts should be an integral part of the evaluation of XAI systems (Vermeire et al., 2021; Langer et al., 2021; Riveiro and Thill, 2021), the evaluation of explanations according to their alignment with the business desiderata and their real impact on decision-making remains poorly explored (Langer et al., 2021; Jesus et al., 2021; Amarasinghe et al., 2023, 2024).

2.3. On the need to evaluate XAI methods in real-world business contexts

As reported in Table 1, the literature abounds with functionally-grounded and human-grounded XAI evaluation practices, whereas scarce attention has been paid to application-grounded evaluations. Recent research (Amarasinghe et al., 2023, 2024; Rong et al., 2024) has demonstrated the need for application-grounded evaluation studies as a way of fostering the adoption of XAI methods that actually reflect the needs of the business context in which they are applied. From the literature examined, the works Jesus et al. (2021), Amarasinghe et al. (2024) are, to the best of our knowledge, the only ones providing an application-grounded methodology, validated with real domain experts, to evaluate and compare the practical relevance of the explanations given by a different set of XAI techniques. Nevertheless, we failed to find any empirically validated work proposing an application-grounded methodology for choosing the most suitable XAI methods for a given business organization in the context of SCM, a field that lacks explainable data-driven models capable of supporting proactive and business-valuable decisions (Nimmy et al., 2022; Olan et al., 2024). As supply chains are the backbone of any organization (Barbosa-Póvoa et al., 2018), the evaluation of XAI techniques in this context is relevant in order to facilitate the adoption of AI-based approaches that prove useful in decision-making business processes.

In response to these gaps, we developed a DSS that allows organizations to find the XAI methods that best suit a given AI use case and the needs of domain experts that regularly work on it. Our paper differs from the existing literature by proposing an application-grounded analytical metric that relies on the feedback and experience of domain experts to select the XAI method that best adapts and brings real utility to the target context. We test our system in a segment of a major supply chain with a real business task involving real users. The proposed system has the potential to strengthen the trust of domain experts in AI solutions, suggesting the selection of intelligible, trustworthy and satisfactory explanations that are aligned with the business stakeholders' desiderata.

3. Context-aware system for selecting XAI methods

The proposed system for selecting XAI methods in business organizations is presented in Fig. 1. Our system is context-aware in the sense that it is totally driven by the preferences of the business stakeholders (including developers and domain experts) who are going to use it in a real-world context, so as to meet their explainability interests, the business requirements of the context in which they operate, as well as their levels of satisfaction and trust in the overall XAI evaluation system. In line with other authors (Kotriwala et al., 2021; Schoonderwoerd et al., 2021; Amarasinghe et al., 2024), our interest lies in encouraging the adoption of XAI methods that actually meet the needs of the target users. We argue that the utility and application interest of an XAI method should be fully dependent on the application context, meaning that a given XAI method may be the most suitable in a given domain, characterized for instance by stakeholders with advanced analytical maturity, but this might not necessarily be the case for another domain where such maturity is lower.

As depicted in Fig. 1, the DSS consists of three main stages involving the human in the loop: (1) the construction of an *explanandum*, which acts as the computational object for which explanations are generated (orange box, Section 3.1); (2) the selection of a pool of XAI methods that meet the stakeholders' needs and which are responsible for generating the explanations for the constructed explanandum (blue boxes, Section 3.2); and (3) the XAISelector (Section 3.3), an application that facilitates the evaluation of explanations by stakeholders (red boxes, Section 3.3.1) and the estimation of their Degree of Trust-Satisfaction (DoTS) with respect to each selected XAI method (green boxes, Section 3.3.2), making it possible to select those that best fit the needs of the business context and problem.

In what follows, we characterize each of these stages together with their main components.

3.1. Construction of an explanandum

The first step towards applying the proposed DSS is the existence of a given ML model built and tested for a particular business problem. Hereafter, we refer to such a model as the *explanandum*.

In this work, we develop an *explanandum* for a specific business task within a real-world supply chain context. In this process (detailed in Section 4.2), we included domain experts in the loop, in order to ensure that the model meets business requirements. This contrasts with current practices in organizations, where the development of ML models is essentially focused on data scientists. Examples of collaboration between data scientists and domain experts include verifying that the model operates correctly and learns relationships between features genuinely relevant to the underlying problem, and that it does not exhibit significant biases during the learning stage (Dwivedi et al., 2023). This interpretation process also enables business stakeholders to conduct, if necessary, fine-tuning actions in order to train the best performing model to be deployed. Following this process, the goal is to ensure that the final model is coherent and free of any inconsistency that compromise its validity and the quality of the explanations generated

Table 1
Overview of different XAI evaluation practices in the literature.

Source	Type of evaluation	Desiderata	Real business task	Business context
Schmidt and Biessmann (2019)	Human-grounded	Trust	No	–
Islam et al. (2020)	Functionally-grounded	Comprehensibility	No	–
Hase and Bansal (2020)	Human-grounded	Simulatability	No	–
Vermeire et al. (2021)	Human-grounded	Utility	No	–
Rosenfeld (2021)	Functionally-grounded	Performance, Fidelity, Complexity, Stability	No	–
Jesus et al. (2021), Amarasinghe et al. (2024)	Application-grounded	Utility	Yes	Finance
Cugny et al. (2022)	Functionally-grounded	Stability, Infidelity, Complexity	No	–
Arias-Duart et al. (2022)	Functionally-grounded	Fidelity	No	–
Agarwal et al. (2022)	Functionally-grounded	Fidelity, Stability, Fairness	No	–
Hoffman et al. (2023b)	Human-grounded	Satisfaction, Trust, Comprehensibility, Effectiveness	No	–
Hoffman et al. (2023a)	Human-grounded	Effectiveness	No	–
Sovrano and Vitali (2023)	Human-grounded	Explainability	No	–
Miró-Nicolau et al. (2024)	Functionally-grounded	Fidelity	No	–
This paper	Application-grounded	Trust, Satisfaction	Yes	Supply chain management

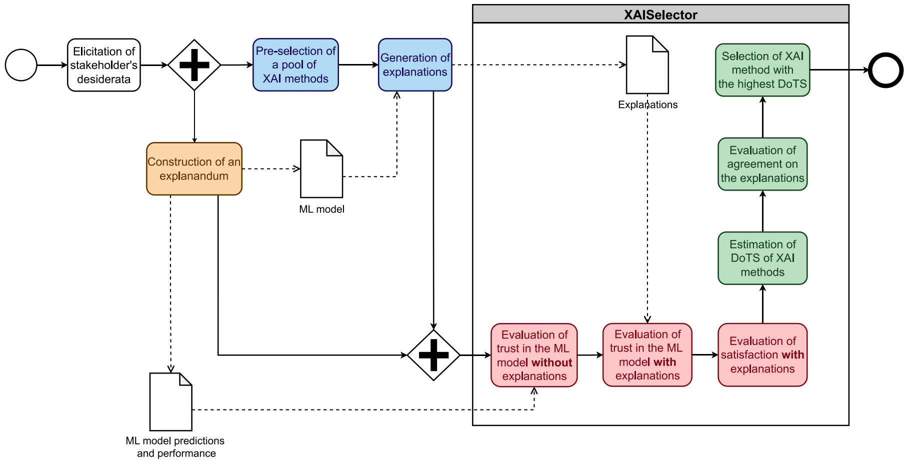

XAISelector
Elicitation of stakeholder's desiderata
Pre-selection of a pool of XAI methods
Generation of explanations
Explanations
Selection of XAI method with the highest DoTS
Construction of an explanandum
ML model
Evaluation of agreement on the explanations
Estimation of DoTS of XAI methods
Evaluation of trust in the ML model without explanations
Evaluation of trust in the ML model with explanations
Evaluation of satisfaction with explanations
ML model predictions and performance

Fig. 1. Business Process Model and Notation (BPMN) model of the workflow of the proposed DSS for selecting XAI methods in business organizations.

from it. As the explanations produced by the XAI methods rely on the deployed model, the engagement of the relevant stakeholders in the model-building process is an important factor to promote a better understanding of its general functioning and to increase the levels of satisfaction and trust in the overall XAI system.

3.2. Pre-selection of XAI methods and generation of explanations

The constructed explanandum is the basis for the production of explanations from a comprehensive pool of XAI methods, carefully selected to best suit the business stakeholders' desiderata. We propose to select this pool of methods by conducting interviews with relevant business stakeholders, led by the data scientist. The main purpose of the interviews is to understand and identify critical aspects, including the application problem for which the explanations are to be built, the profile and analytical maturity of the end-users, as well as their needs and constraints. In this process, we take advantage of the questionnaires proposed by Vermeire et al. (2021), which include questions that provide a general overview of the characteristics of the stakeholders and the context in which they operate — factors considered to influence the relationship between the explanatory information and understanding (Langer et al., 2021). Throughout this stage, we assume that the data scientist has sufficient technical knowledge to discard XAI methods that are too complex for the application context or that do not meet, for example, some of the stakeholders' visualization requirements.

Once the pool of XAI methods has been selected, explanations are generated for the explanandum, one for each individual method.

3.3. The XAISelector

In order to collect human feedback on the derived explanations, our DSS comprises the XAISelector, a web application that facilitates the stakeholders' evaluation regarding the utility of the explanations produced by XAI methods according to two main aspects: (1) the business stakeholders' trust in the ML model provided with explanations and (2) their satisfaction with the explanations produced. In our work, these aspects form the basis for determining the most suitable XAI method for a given business context, depending on the needs of the stakeholders operating within it (see Section 3.3.2). We publish the source code[1] of XAISelector, for the sake of reproducibility and for the primary

[1] https://github.com/mindior/xaiselector/.

purpose of making it a useful solution for interested researchers and practitioners. A Flask Python web application framework was used to develop a working system prototype, making it available online for wider visibility. The application architecture follows the Model-View-Controller (MVC) pattern (Leff and Rayfield, 2001), in which models, controls, and interfaces are separated into different components. The following processor and software were used for implementing the XAISelector: Intel(R) Core(TM) i7-10850H CPU-22.70 GHz 32 GB RAM, Windows 10. In what follows, we detail the main components of the XAISelector.

3.3.1. Evaluating explanations involving the human-in-the-loop

XAISelector facilitates the online implementation of two experimental stages to evaluate the utility of the explanations derived from a given pre-selected pool of XAI methods with business stakeholders. In the first stage, before having contact with the explanations, the stakeholders evaluate their trust in the developed ML model. For that, we propose to use the *Trust Scale Recommended for XAI* by Hoffman et al. (2023b) as evaluation instrument, consisting of a questionnaire with 8 items on a 5-point Likert-type scale. This first evaluation process takes place after presenting the results of the deployed ML model to the stakeholders, where the data scientist showcases its predictive performance and provides examples of model predictions supported by some relevant evaluation metrics (e.g., bias, prediction errors, coefficient of determination). In a second stage, the data scientist presents to the stakeholders the explanations produced from the different explainability methods so as to further evaluate their trust in the deployed model after having contact with such explanations, using the same evaluation instrument as before. Following this strategy, we are interested in comparing, for each XAI method, the stakeholders' levels of trust in the XAI-based ML pipeline prior to and after contact with the corresponding explanations. Finally, stakeholders are asked to rate their level of satisfaction with each explanation produced. To this end, we propose to employ the *Explanation Satisfaction Scale*, proposed by Hoffman et al. (2023b), which also consists of a questionnaire with 8 items on a 5-point Likert-type scale.

The two evaluation scales described above are implemented in the XAISelector application, in order to facilitate the collection of responses from business stakeholders. In what follows, each of these scales is used to build a quantitative metric for estimating the degree of trust-satisfaction of each individual XAI method.

3.3.2. Estimating the degree of trust-satisfaction of XAI methods

In Section 3.3.1, we propose evaluating explainability methods using a human-in-the-loop approach, i.e., exploiting stakeholder's trust and satisfaction as a way of gauging the quality and the utility of the explanations obtained. Yet, the existence of two independent evaluation measures built on Likert-type scales makes it difficult to find which XAI method provides the greatest practical utility for domain experts. This is particularly evident when the feedback obtained stems from different stakeholders with varying degrees of business expertise and analytical maturity, as is the case in most real-world organizations. In practice, XAI methods should be designed and selected according to their application context and the nature of the underlying problem (Amarasinghe et al., 2023, 2024). However, this task can become unfeasible, given the multitude of existing problems and business settings.

To address these challenges, we introduce a new analytical metric to establish an estimate of the Degree of Trust-Satisfaction (DoTS) of a pre-defined set of XAI methods. This metric can serve as a basis to select the methods that best suit the context, the problem and the business stakeholders operating therein. The rationale for creating this metric arises from the need to create evaluation strategies for XAI methods that reflect the actual needs of the context in which they are applied. So far, the literature has undervalued this issue, proposing metrics that do not take advantage of feedback from real target users and that are not defined to suit their business requirements. Consequently, it becomes difficult to evaluate XAI methods according to the specific characteristics of each application context, thus posing barriers to their adoption by the real-world users (see Amarasinghe et al., 2023, 2024, for a detailed discussion). We propose that DoTS depends on the stakeholders' feedback on the explanations generated, by measuring the two measures presented in Section 3.3.1: the stakeholders' level of trust in the XAI-based ML model, and the stakeholders' level of satisfaction with provided explanations. Previous studies (Kim et al., 2009) have demonstrated the importance of these two measures in building successful business relationships and in promoting the adoption of XAI techniques by users (Aliyeva and Mehdiyev, 2024; Hoffman et al., 2023a,b).

Next, we constructively describe the DoTS metric in a formal way, starting by defining some of its fundamental concepts.

Definition 1 (*XAI method*)**.** In a classical supervised learning setting, there is an input space $\mathcal{X} \subseteq \mathbb{R}^p$ and an output space $\mathcal{Y} \subseteq \mathbb{R}$. Let $\mathcal{F}$ be the space of predictive models that can map $\mathcal{X}$ into $\mathcal{Y}$. In such a setting, let $f \in \mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$ be the selected ML model that computes a predictive estimate $(\hat{y})$ when fed with a p-dimensional input vector $(\mathbf{x} \in \mathcal{X})$, resulting in $\hat{y} = f(\mathbf{x}) \in \mathcal{Y}$. An XAI method is a function $g : (\mathcal{X}, \mathcal{Y}, f) \rightarrow E$, where E is a set (or a singleton set) of explanations generated for the model f.

Definition 2 (*Trust in the constructed explanandum*)**.** Let Q_T be a set of Likert-type questions, ranging from a minimum score value of m_T to a maximum score value of M_T, designed to measure the stakeholders' trust in a given predictive model f. For a given respondent $r \in R$, we define the overall trust level in the model f explained by an XAI method g $(T_{g,r})$ as the average of the normalized difference between the trust levels obtained with and without the explanations generated by g across all questions $q_t \in Q_T$. Formally, we have

$$T_{g,r} = \frac{1}{|Q_T|}\sum_{q_t \in Q_T}\frac{\left(T_{g,r,q_t} - T_{r,q_t}\right) - (m_T - M_T)}{(M_T - m_T) - (m_T - M_T)}, \tag{1}$$

with T_{g,r,q_t} and T_{r,q_t} denoting the trust levels obtained with and without the explanations generated by g for the respondent r in question q_t, respectively.

The overall T_g can then be calculated as the arithmetic mean of $T_{g,r}$ across respondents $r \in R$. In the definition of $T_{g,r}$, we employ a min–max normalization (James et al., 2023), where $\min = m_T - M_T$ and $\max = M_T - m_T$ are respectively the minimum and maximum values that the difference $T_{g,r,q_t} - T_{r,q_t}$ can take, for all $q_t \in Q_T$ and $r \in R$. All values of $T_{g,r}$ thus lie in the interval $[0, 1]$. Looking at the structure of Eq. (1), we seek to find an XAI method g such that $T_{g,r,q_t} > T_{r,q_t}$, for all $r \in R$ and $q_t \in Q_T$. Hence, from the perspective of trust, XAI methods with higher T_g are considered superior to those with lower T_g.

At this point, note that the sum of the differences of multiple Likert-type answers reflected in Eq. (1) is justified as we are interested in constructing a subjective evaluation index of the overall trust level in the explained model, rather than in interpreting the values of $T_{g,r}$ according to the underlying characteristics/labels of each Likert number (i.e., "I strongly disagree" to "I strongly agree") (Allen and Seaman, 2007; Norman, 2010; Batterton and Hale, 2017). We chose to calculate the average value (rather than the median) T_g across respondents $r \in R$ and questions $q_T \in Q_T$, because the trust index constructed is intended to be sensitive to small variations in respondents' answers, and not based on the most repeated value in each question. We do, however, acknowledge that this aggregation of scales should be evaluated for internal consistency between the answers (Allen and Seaman, 2007), e.g., using Cronbach's Alpha (Cronbach, 1951).

According to Definition 2, and following the rationale presented in Section 3.3.1, the use of the *Trust Scale Recommended for XAI* proposed by Hoffman et al. (2023b) implies that the value of the subjective index T_g, for a given g, is based on the answers to a questionnaire with eight $(|Q_T| = 8)$ Likert-type items ranging from $m_T = 1$ ("I disagree strongly") to $M_T = 5$ ("I agree strongly").

$$DoTS_g = \frac{\sum_{r\in R} w_{E,r}\left(w_{I,r}T_{g,r} + (1-w_{I,r})S_{g,r}\right)}{\sum_{r\in R} w_{E,r}} = \frac{\sum_{r\in R} w_{E,r}\left(\frac{w_{I,r}}{|Q_T|}\sum_{q_t\in Q_T}\frac{\left(T_{g,r,q_t}-T_{r,q_t}\right)-(m_T-M_T)}{2(M_T-m_T)} + \frac{1-w_{I,r}}{|Q_S|}\sum_{q_s\in Q_S}\frac{S_{g,r,q_s}-m_S}{M_S-m_S}\right)}{\sum_{r\in R} w_{E,r}}, \quad (4)$$

Box I.

Definition 3 (*Satisfaction with explanations*). Let Q_S be a set of Likert-type questions, ranging from a minimum score value of m_S to a maximum score value of M_S, designed to measure stakeholders' satisfaction with the explanations provided by an XAI method g. For a given respondent $r \in R$, we define the overall satisfaction level with the explanations generated by g ($S_{g,r}$) across all questions $q_s \in Q_S$ by

$$S_{g,r} = \frac{1}{|Q_S|}\sum_{q_s\in Q_S}\frac{S_{g,r,q_s}-m_S}{M_S-m_S}, \quad (2)$$

with S_{g,r,q_s} denoting the satisfaction level with the explanations generated by g for the respondent r in question q_s.

The overall S_g can then be calculated as the arithmetic mean of $S_{g,r}$ across respondents $r \in R$. Similarly to the construction of T_g, we normalize the values of S_{g,r,q_s} into the interval $[0,1]$ by considering $\min = m_S$ and $\max = M_S$ as the minimum and maximum values that S_{g,r,q_s} can take, for all $q_s \in Q_S$ and $r \in R$. The assumption of manipulating Likert scales as interval values follows the rationale given above when defining T_g. Analogously to the trust measure, we seek to find an XAI method g that maximizes S_g for all $r \in R$ and $q_s \in Q_S$.

In such a setting, the use of the *Explanation Satisfaction Scale* proposed by Hoffman et al. (2023b) implies that, for a given g, the value of the subjective index S_g is based on the answers to a questionnaire with eight ($|Q_S| = 8$) Likert-type items ranging from $m_S = 1$ ("I disagree strongly") to $M_S = 5$ ("I agree strongly"), enabling stakeholders to rate suitable and unsuitable explanations.

Using both definitions above, we are able to define the Degree of Trust-Satisfaction (DoTS) of an XAI method:

Definition 4 (*Degree of Trust-Satisfaction*). The basis of DoTS of an XAI method g is defined as the combination of the normalized measures $T_{g,r}$ and $S_{g,r}$, intended to be maximized across all respondents $r \in R$, i.e.,

$$\begin{aligned} T_g + S_g &= \frac{1}{|R|}\sum_{r\in R} T_{g,r} + S_{g,r} \\ &= \frac{1}{|R|}\sum_{r\in R}\left(\frac{1}{|Q_T|}\sum_{q_t\in Q_T}\frac{\left(T_{g,r,q_t}-T_{r,q_t}\right)-(m_T-M_T)}{2(M_T-m_T)}\right. \\ &\quad \left. + \frac{1}{|Q_S|}\sum_{q_s\in Q_S}\frac{S_{g,r,q_s}-m_S}{M_S-m_S}\right). \end{aligned} \quad (3)$$

This formulation assumes equal weights for both measures $T_{g,r}$ and $S_{g,r}$. Yet, in practice, stakeholders may be interested in giving more importance to one measure than another, depending on the problem and business context. On the other hand, Eq. (3) does not take into account the level of expertise of the stakeholder evaluating the explainability method g. Assuming different weights to be assigned to the trust and satisfaction components, we can thus write the DoTS metric as a weighted combination of the normalized measures $T_{g,r}$ and $S_{g,r}$ across respondents $r \in R$, as defined by Eq. (4) in Box I.

where $0 < w_{E,r} \leq 1$ is the weight quantifying the expertise of the stakeholder r, whereas $0 \leq w_{I,r} \leq 1$ and $1 - w_{I,r}$ reflect the relative importance assigned by stakeholder r to the evaluation components $T_{g,r}$ and $S_{g,r}$, respectively. At this point, note that $w_{I,r}$ is the only controllable parameter that comprises the DoTS metric. The effect of varying this parameter on the overall dynamics of the DoTS is studied in Section 5.3. The weight $w_{E,r}$ can be understood as a way of balancing the subjectivity of the evaluations, assuming that the ratings provided by more experienced stakeholders and/or those with greater analytical maturity should have a stronger impact when determining the DoTS of an XAI method (Doshi-Velez and Kim, 2018).

Of practical interest, the DoTS was conceived and designed to be a metric of easy mathematical interpretation in order to promote understandability for the domain users – a relevant aspect in the field of operational research (De Bock et al., 2024). On the other hand, the DoTS is an analytically tractable expression bounded in the interval $[0,1]$, making its interpretation more convenient for business stakeholders. If the value of $DoTS_g$ is close to 0, it means that the XAI method g reveals a low degree of trust-satisfaction and little utility to the context and stakeholders' needs, whereas values of DoTS close to 1 indicate a significant degree of trust-satisfaction for g. Following this intuition, given a set of XAI methods G, the one that should be selected as the method that best suits the stakeholders' needs according to their organizational context and analytical maturity is given by $g^* = \arg\max_g\{DoTS_g : g \in G\}$. As the DoTS is a metric that depends on subjective feedback, which can naturally vary from stakeholder to stakeholder, we suggest complementing it by also measuring the inter-rater agreement among the target stakeholders involved in the selection of g^*. For that, we consider the weighted kappa statistic (Cohen, 1968), appropriate for Likert scales.

3.3.3. On the adequacy and applicability of the DoTS

The DoTS metric has three important properties that make it suitable to be used in practical contexts. In what follows, we elaborate on such properties.

Sensitivity to the context and conditions of the problem. The DoTS metric, being dependent on human feedback, makes it possible for real users of the application context to select XAI methods that, on the one hand, yield satisfactory explanations for the explanandum constructed and, on the other hand, provide a better understanding of its outputs. In addition, the metric is sensitive to the analytical maturity of the stakeholders operating in the business context. To motivate the latter idea, we present the following example.

Example 1. Consider a set $R = \{r_1, r_2, r_3\}$ of three decision-makers who use a particular explanandum built to address a given business problem. Let $w_{E,r_1} = w_{E,r_2} = 0.25$ and $w_{E,r_3} = 1$ be the weights that reflect the degree of analytical maturity of the different decision-makers, measured, for instance, by work experience in the problem context. Take a scenario where all decision-makers consider that the measures of trust (T_g) and satisfaction (S_g) generated by a given XAI method g are equally relevant (i.e., $w_{I,r_1} = w_{I,r_2} = w_{I,r_3} = 0.5$). Suppose that while decision-makers r_1 and r_2 agree that the XAI method g is optimal in the sense of maximizing both the measures T_g and S_g (i.e., $T_{g,r} = S_{g,r} = 1, \forall r \in \{r_1, r_2\}$), decision-maker r_3 finds the method g inappropriate for the business problem in question (i.e., $T_{g,r_3} = S_{g,r_3} = 0$). In this context, considering the feedback from the three decision-makers, the overall score obtained using the DoTS metric for the XAI method g is $DoTS_g = 1/3$ ($\approx$ 33.3%). In other words, the negative feedback from a single experienced decision-maker heavily impacts the degree of utility of the XAI g method for the underlying problem. In contrast, let us now assume that $w_{E,r_1} = w_{E,r_2} = 1$ and $w_{E,r_3} = 0.25$. Preserving all other problem assumptions, this yields a score of $DoTS_g \approx 0.89$. This means that in a context operated by experienced users, the feedback from an inexperienced user, although valuable, does not impact strongly on the choice of a particular XAI method as the most adequate for the problem.

Consistency and comparability. The DoTS metric is deterministic, which guarantees the same scores given the same inputs collected from stakeholders, regardless the application context. This makes the metric generalizable and comparable between business environments.

Scalability and interpretability. Following the formulation (4), the construction of the DoTS metric involves three fundamental steps. The first step is the calculation of T_g, in which we start by computing, for each respondent $r \in R$, the normalized trust level obtained in each question $q_t \in Q_T$. This results in a complexity of $\mathcal{O}(|R| \times |Q_T|)$ for the computation of T_g. The second step is the computation of S_g which, following the same rationale as that used to compute T_g, results in a complexity of $\mathcal{O}(|R| \times |Q_S|)$. The third and final step in the construction of the DoTS consists of combining the measures T_g and S_g obtained for each respondent $r \in R$, considering the weights $w_{I,r}$ and $w_{E,r}$. The total complexity of the evaluation of the DoTS metric is therefore linear – expressed as $\mathcal{O}(|R| \times (Q_T + Q_S))$. The pseudo-code in Algorithm 1 summarizes the process of computing the DoTS metric. The DoTS is therefore interpretable and computational efficient, which makes it appealing for application in multiple business contexts.

Algorithm 1: The DoTS algorithm

```
Input: $G, R, Q_T, Q_S, T_{g,r,q_t}, T_{r,q_t}, S_{g,r,q_s}, m_T, M_T, m_S, M_S, w_{I,r}, w_{E,r}$
Result: $g^*$ // the XAI method with the highest utility
        and adaptability to the business context
for $g \in G$ do
    $DoTS_g \leftarrow 0$
    for $r \in R$ do
        $T_{g,r} \leftarrow 0, S_{g,r} \leftarrow 0$
        for $q_t \in Q_T$ do
            $T_{g,r} \leftarrow T_{g,r} + \frac{\left(T_{g,r,q_t} - T_{r,q_t}\right) - (m_T - M_T)}{(M_T - m_T) - (m_T - M_T)}$
        end
        $T_{g,r} \leftarrow \frac{1}{|Q_T|} T_{g,r}$                                ▷ Definition (1)
        for $q_s \in Q_S$ do
            $S_{g,r} \leftarrow S_{g,r} + \frac{S_{g,r,q_s} - m_S}{M_S - m_S}$
        end
        $S_{g,r} \leftarrow \frac{1}{|Q_S|} S_{g,r}$                                ▷ Definition (2)
        $DoTS_g \leftarrow DoTS_g + w_{E,r}(w_{I,r} T_{g,r} + (1 - w_{I,r}) S_{g,r})$
    end
    $DoTS_g \leftarrow DoTS_g / \sum_{r \in R} w_{E,r}$                        ▷ Definition (4)
end
$g^* = \arg\max_g \{DoTS_g : g \in G\}$
```

4. System evaluation

To test the effectiveness and practical utility of the proposed DSS, we resort to a real business problem from a leading focus company in the automotive sector: Bosch Automotive Electronics Portugal (AE/P). The problem falls within the area of supply chain demand estimation and was modeled using a multivariate ML approach. The constructed model serves as the *explanandum* for constructing explanations from a pool of pre-selected XAI methods and carrying out experiments with real business stakeholders. In this process, XAISelector is used as a way of collecting stakeholders' feedback on the practical utility of the explanations produced. The following software and libraries were used for implementing the explanations and carry out the experiments presented throughout this section: Python 3.9; R 4.1.0; Scikit-learn 0.23; SHAP 0.41.0; LIME 0.2.0.1; ALE 1.1.3 and rminer 1.4.6.

4.1. Case study

Our research context is the central logistics department of Bosch AE/P, operating on a daily basis with a multitude of customers, manufacturing components and suppliers. Given the complexity of its supply chain, the main focus is on the efficient management of the business processes. One of the processes with greatest impact on the organization is the demand management, in particular the proactive management of variations in the manufacturer's demand for components. In such a setting, the organization is interested in proactively estimate, over the planning horizon, the demand variation during the period in which no changes in supplier orders are allowed (also known as the frozen period (Lian et al., 2006)). This interest stems from the fact that the actual demand signals at the moment immediately after the end of the frozen period tend to differ from those planned at the beginning of the frozen period for that same moment. This is typically motivated by short-term changes in customer orders, which lead to a constant change in suppliers' order plans.

To address this problem, we develop a supervised learning strategy that makes it possible to estimate the variation in demand at the moment immediately after the end of the frozen period compared to that planned at the beginning of that period. A proprietary dataset of historical demand information provides the support for the proposed strategy, outlined in detail throughout Section 4.2.

4.2. On the construction of the explanandum

In the context of the business problem previously presented, we construct a specific explanandum, in the form of a ML model, to serve as an object for the generation of explanations and their subsequent evaluation using the DoTS metric. Nevertheless, note that the proposed DSS – including the DoTS metric – can be applied to any other explanandum.

4.2.1. Model definition and learning setup

Let $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a dataset with $n = |D|$ instances in the form of input–output pairs $(\mathbf{x}_i, y_i)$, where $\mathbf{x}$ is a p-dimensional feature vector of input (independent) variables from the space $\mathcal{X} \subseteq \mathbb{R}^p$ and y is the target (dependent) variable from the space $\mathcal{Y} \subseteq \mathbb{R}$. In our case, the target variable is the actual demand at the moment immediately after the end of the frozen period. Let $t : \mathcal{X} \to \mathcal{Y}$ be the target function that maps the input data encoded in $\mathbf{x}$ to the desired outputs values $y_i \in \mathcal{Y}$. Following the empirical risk minimization principle (Vapnik, 1999), we are interested in finding a predictive model $f \in \mathcal{F} : \mathcal{X} \to \mathcal{Y}$ that is close to the target function t on the training examples, where $\mathcal{F}$ is the space of predictive models (or hypotheses).

In the context of our business problem, we consider two groups of variables that form our input data $\mathbf{x} = [\mathbf{w}, \mathbf{z}]$, one describing the structure of the component ($\mathbf{w}$) and the other describing its demand dynamics ($\mathbf{z}$), which explain the variations in y during the frozen period. Table 2 presents a characterization of the variables, in terms of their type and the transformation employed for modeling purposes. The variables collected were selected with the assistance of business stakeholders, in order to promote the inclusion of meaningful explanatory information.

Table 2
Input variables used in the predictive model f.

Feature group	Variable	Description	Type of variable	Transformation
w	W_1	Type of component	Categorical	One-hot-encoding
	W_2	Distinct number of finished products using the component	Numerical	z-score
	W_3	Distinct number of customers depending on the component	Numerical	z-score
z	Z_1	Planned demand at the beginning of the frozen period	Numerical	z-score
	$Z_{2,l}$	Weekly lags ($l = 1, \dots, 4$) of planned demand	Numerical	z-score
	Z_3	Variance of the lags of planned demand	Numerical	z-score
	Z_4	Average of the lags of planned demand	Numerical	z-score
	Z_5	Amplitude ($\max - \min$) of the lags of planned demand	Numerical	z-score

Our modeling goal is to determine a predictive function $f(\cdot)$ that learns the relationships between the input variables encoded in $\mathbf{x} = [\mathbf{w}, \mathbf{z}]$ and the target variable, in order to further derive accurate estimates of the actual demand y at the moment immediately after the end of the frozen period, i.e.,

$$y_i = f(\mathbf{w}_i, \mathbf{z}_i) + \varepsilon_i, \tag{5}$$

where ε is the stochastic error process. Such predictive function $f(\cdot)$ can be derived by solving the following mathematical optimization problem:

$$f = \underset{h \in \mathcal{F}}{\arg\min} \ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(h(\mathbf{w}_i, \mathbf{z}_i; \theta), y_i\right), \tag{6}$$

where θ is a m-dimensional vector of model hyperparameters that shape the structure of the hypothesis h, whereas $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the loss function that measures the quality of the approximation of y_i by $h(\mathbf{w}_i, \mathbf{z}_i; \theta)$ over the training instances $i = 1, \dots, n$. We implement the regression strategy (5) by considering six predictive ML algorithms, as given in the SKLearn library (Pedregosa et al., 2011), namely Multilayer Perceptron (MLP) (Pinkus, 1999), Random Forest (Breiman, 2001), eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), Bagging (Breiman, 1996), k-Nearest Neighbors (kNN) (James et al., 2023) and Support Vector Regression (SVR) (Cortes and Vapnik, 1995). The grid-search optimization of the hyperparameters of the different models follows the configuration presented in Table 3. The choice of these algorithms stems from their widespread use in the context of demand estimation (Abolghasemi et al., 2020; Joseph et al., 2022; Bertolini et al., 2021), including in previous studies carried out in Bosch AE/P (Gonçalves et al., 2021; Barros et al., 2023) in the context of business analytics. During the model-fitting process, the target variable was standardized using a z-score transformation (James et al., 2023) over the training set. The model predictions are then post-processed by applying the inverse of the transformation.

Table 3
Hyperparameters of the learning regression models.

Algorithm	Parameters
Multilayer Perceptron	$\{100k \mid k = 1, 2, \dots, 9\}$ activation: Identity; ReLU
Random Forest	n_estimators $\in \{150, 300, 450\}$ max_depth $\in \{\text{“None”}, 50, 100\}$ max_features $\in \{0.6, 0.8, 1.0\}$ bootstrap $\in \{1, 0\}$
XGBoost	n_estimators $\in \{150, 300, 450\}$ max_depth $\in \{50, 100, 150\}$ learning_rate $\in \{0.0001, 0.001, 0.01\}$ subsample $\in \{0.6, 0.8, 1.0\}$
Bagging	n_estimators $\in \{150, 300, 450\}$ max_samples $\in \{0.6, 0.8, 1.0\}$ max_features $\in \{0.6, 0.8, 1.0\}$ bootstrap $\in \{1, 0\}$
kNN	n_neighbors $\in \{3, 5, 7\}$ weights $\in \{\text{“uniform”}, \text{“distance”}\}$ metric $\in \{\text{“Euclidean”}, \text{“Manhattan”}, \text{“Minkowski”}\}$
SVR	kernel $\in \{\text{“linear”}, \text{“poly”}, \text{“rbf”}\}$ degree $\in \{2, 3, 4\}$ C $\in \{\text{“scale”}, \text{“auto”}\}$ gamma $\in \{0.001, 0.1, 1\}$

Table 4
Summary of prediction performance (expressed in %NMAE and Adj. R^2) obtained from the different models over 20 rolling window iterations. The best values are highlighted in boldface. The values in round brackets represent the ranking within the column.

	%NMAE [Std. Err.]	Adj. R^2 [Std. Err.]
MLP	0.237 [0.117] (4)	0.973 [0.013] (4)
Random Forest	0.205 [0.113] (2)	0.975 [0.017] (2)
XGBoost	0.259 [0.163] (5)	0.974 [0.015] (3)
Bagging	0.206 [0.111] (3)	0.975 [0.016] (2)
kNN	**0.184** [0.112] (1)	**0.976** [0.023] (1)
SVR	0.274 [0.181] (6)	0.972 [0.019] (5)

4.2.2. Model evaluation and validation

Following the setup described in Section 4.2.1, the evaluation and analysis of the generalization capacity of the ML models are conducted on a dataset D comprising n = 63,203 records, captured in a weekly basis for 4342 different manufacturing components. Each record in the dataset is characterized by both the input features presented in Table 2 and the target variable y. We split such dataset into three time ordered portions: training data (50%, with the oldest records), validation data (20%) and test data (30%, with the most recent data samples). The validation data serves as the basis for tuning the model hyperparameters, following a grid-search optimization procedure. In this setup, the objective function to be minimized is the Mean Absolute Error (MAE) between the estimated and the actual demand. Once the hyperparameters are fixed, we then apply a rolling window evaluation scheme (Tashman, 2000), which produces 20 training and testing iterations through time. This is obtained by partitioning the original test data (the previously mentioned 30%) into 20 independent (non-overlapping) and equally sized test windows, each with 1.5% of the data. During this procedure, the training window (corresponding to 70% of the full data) rolls forward in each iteration by discarding the oldest 1.5% records and adding more recent 1.5% ones. Each iteration of the rolling window is then used to evaluate the performance of the predictive models. Following this strategy, the generalization ability of each model is tested with different and sequential test windows, thus making the evaluation robust and realistic in the sense of simulating the operation of a classical demand planning system.

To ensure the practical applicability of our explanandum and its alignment with the business objectives, several technical domain experts from the organization were included in the process of developing the predictive model. This process involved multiple rounds of testing on the modeling methodologies employed, enabling to gradually refine the final model obtained. Table 4 presents a summary of the average predictive performance of the different models over the 20 rolling window iterations, expressed in terms of the Normalized (by the amplitude of the full test interval in each iteration) MAE (NMAE) and of the Adjusted R^2 (James et al., 2023).

The NMAE is easily interpreted, as it expresses the error as a percentage of the target values, and it is scale-independent, enabling the aggregation of errors measured at different scales. We further evaluate Adjusted R^2 to extract insights regarding the goodness-of-fit of the different regression models.

The results show that the model with the best predictive performance is kNN. However, the kNN is explainable by nature and we are interested in evaluating the potential of explainability methods applied to a black-box regression model. For this reason, we selected the Random Forest as our explanandum, as it is commonly treated as a black-box algorithm (Guidotti et al., 2018; Zhao and Hastie, 2021) and proved to be the second most accurate predictive model for our data. Indeed, the use of the hyperparameter configuration shown in Table 3, together with the pre-processing and transformation strategies used to create the model inputs (see Table 2), make the Random Forest regression model less interpretable for business decision-makers. Nevertheless, it should be noted that the proposed DSS is agnostic to the ML algorithm used. Recalling that the main purpose of this system is to facilitate the selection of XAI techniques and not ML algorithms, the choice of the Random Forest (or another black-box algorithm) is, in the context of this research, justified.

4.3. XAI methods

Once the explanandum is defined, we select a pool of XAI methods to produce the explanations. For that, we conducted an interview with a senior domain expert to identify the business stakeholders' desiderata for implementing an XAI system to support the business problem presented in Section 4.1. In this process, we took advantage of the questionnaire proposed in Vermeire et al. (2021) (as described earlier in Section 3.2). For this particular problem and context, the only desideratum identified was the need for visual explanations. This finding is in line with the study by Kim et al. (2023), which showed that visual explanations tend to be preferred by users. Thus, we decided to test a set of explainability methods in the form of visual representations commonly used in practice, namely the Accumulated Local Effects (ALE) (Apley and Zhu, 2020), the Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2018), the Variable Effect Characteristic (VEC) Contour and Surface (Cortez and Embrechts, 2011, 2013), as well as the SHapley Additive Explanations (SHAP) (Dwivedi et al., 2023; Lundberg and Lee, 2017), including the Decision, Dependence, Summary and Waterfall plots. In total, eight distinct XAI methods were adopted to generate explanations of the explanandum.

Despite the fact that, in this specific context, we have selected a limited and non-exhaustive set of XAI methods with visual explanations, it is important to highlight that the DoTS metric is agnostic to the explainability method used. The DSS can therefore be adapted to test any XAI method, not just the ones addressed in this case study.

4.4. Experimental design and evaluation of the XAISelector

The explanations generated by the XAI methods selected in Section 4.3 were presented to a set of stakeholders, in order for them to evaluate their quality and utility in practice. Given the structure of the DoTS metric, which is based on human feedback in a real context, we adopted the strategy of field testing to evaluate the proposed system. Such a strategy enables an XAI system to be evaluated according to three fundamental principles, namely in terms of its performance capacity, its degree of understandability and its responsibility component (for details see De Bock et al., 2024). Note that the field testing methodology matches one of the basic assumptions of the construction of an application-grounded metric, i.e., the evaluation of explanations in a real task with real stakeholders. We asked two different groups of stakeholders to take part in this evaluation process:

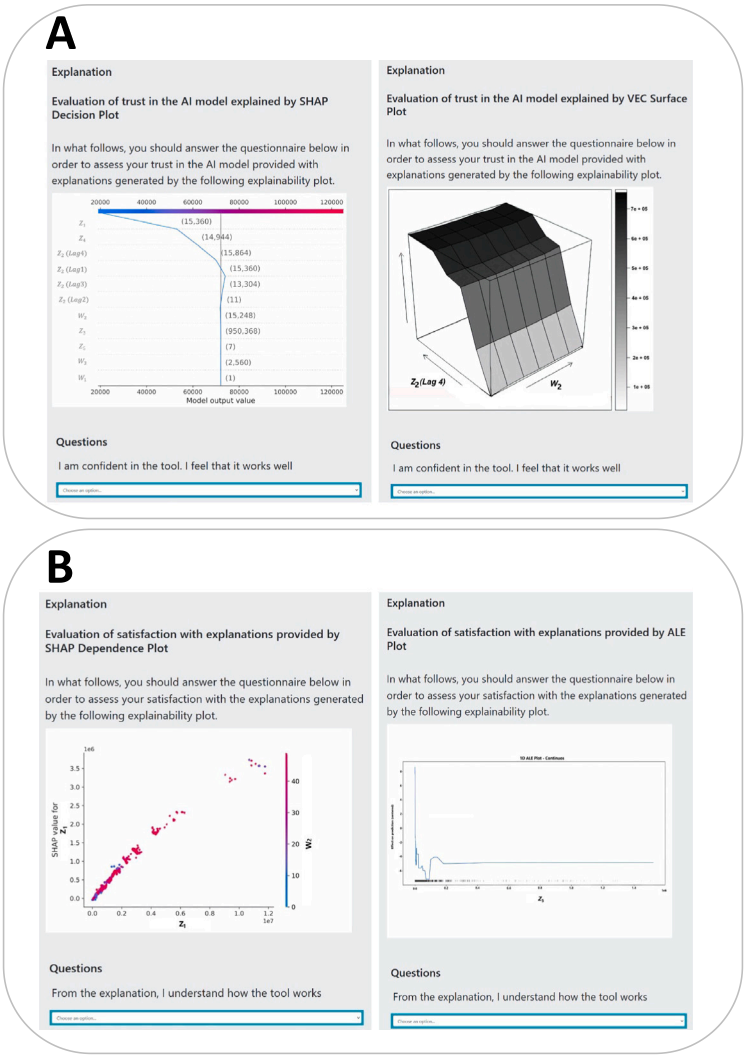
1. *Technical domain experts*. We selected five Bosch AE/P domain experts, mainly data science professionals with varying real-world experience in the SCM context. A demographic analysis of the participants revealed that three were between 25–30 years old and two were between 18–24. The highest academic qualification of the participants was "master's degree": 4, followed by "bachelor's degree": 1. In terms of work experience in SCM, one participant has between 4–6 years of experience, while the remaining have up to three years of experience. All the participants expressed knowledge in AI and Analytics.
2. *Researchers*. We also invited five researchers, other than the authors of this paper, to take part of the process of evaluating the practical utility of the explanations. This group of stakeholders may act as end users of the proposed system. We are interested in contrasting the perspectives of domain experts from the organization with those of other professionals working in the same context but from an academic and scientific perspective. An analysis of the age distribution of the researchers revealed that four were over 41 years old, while one was between 25–30 years old. Only one of the participants holds a master's degree, while all the others have a doctorate in areas related to SCM. A single researcher claimed to have no knowledge whatsoever of AI-based systems.

Table 5
XAI measures and corresponding questions (adopted from Hoffman et al. (2023b)) used for the general evaluation of the practical utility of the explanations provided. All the questions are Likert-type, with a response range ranging from 1 ("I strongly disagree") to 5 ("I strongly agree").

Measure	Question	Description
Trust	Q1T	I am confident in the ML model. I feel that it works well.
	Q2T	The outputs of the ML model are very predictable.
	Q3T	The ML model is very reliable. I can count on it to be correct all the time.
	Q4T	I feel safe that when I rely on the ML model I will get the right answers.
	Q5T	I am wary of the ML model.
	Q6T	The ML model can perform the task better than a novice human user.
	Q7T	I like using the ML model for decision making.
Satisfaction	Q1S	From the explanation, I understand how the ML model works.
	Q2S	This explanation of how the ML model works is satisfying.
	Q3S	This explanation of how the ML model works has sufficient detail.
	Q4S	This explanation of how the ML model works seems complete.
	Q5S	This explanation of how the ML model works tells me how to use it.
	Q6S	This explanation of how the ML model works is useful to my goals.
	Q7S	This explanation of the ML model shows me how accurate the ML model is.
	Q8S	This explanation lets me judge when I should trust and not trust the ML model.

As described in Section 3.3.1, the participants from both groups were asked to answer two questionnaires, one focused on evaluating the explanandum developed with and without explanations and the other on evaluating their satisfaction with the explanations produced. Table 5 summarizes the items in each questionnaire.

In this process, stakeholders were not interested in checking the efficiency of the XAI system. We have therefore removed one question regarding the assessment of the speed of the system from the trust scale. This leaves the trust questionnaire with seven questions rather than the original eight proposed in Hoffman et al. (2023b). The participants answered the questions directly in the XAISelector application, which implements both questionnaires for each XAI method in the pool and makes it possible to use the responses provided to determine the general levels of trust in the explained model (using Eq. (1)), the satisfaction levels with the explanations produced (using Eq. (2)), and consequently the DoTS metric (using Eq. (4)). Fig. 2 provides examples of trust and satisfaction questionnaires (panels A and B, respectively) implemented in the XAISelector for the different XAI methods, presented



Explanation

Evaluation of trust in the AI model explained by SHAP Decision Plot

In what follows, you should answer the questionnaire below in order to assess your trust in the AI model provided with explanations generated by the following explainability plot.

Questions

I am confident in the tool. I feel that it works well

Choose an option...

Explanation

Evaluation of trust in the AI model explained by VEC Surface Plot

In what follows, you should answer the questionnaire below in order to assess your trust in the AI model provided with explanations generated by the following explainability plot.

Questions

I am confident in the tool. I feel that it works well

Choose an option...

Explanation

Evaluation of satisfaction with explanations provided by SHAP Dependence Plot

In what follows, you should answer the questionnaire below in order to assess your satisfaction with the explanations generated by the following explainability plot.

Questions

From the explanation, I understand how the tool works

Choose an option...

Explanation

Evaluation of satisfaction with explanations provided by ALE Plot

In what follows, you should answer the questionnaire below in order to assess your satisfaction with the explanations generated by the following explainability plot.

Questions

From the explanation, I understand how the tool works

Choose an option...

Fig. 2. Screenshots of trust (panel A) and satisfaction (panel B) questionnaires implemented in XAISelector.

to stakeholders in the form of plots. Prior to collecting the answers to the questionnaires, each stakeholder was instructed on how each plot should be analyzed in the context of the business problem, thus ensuring that all plots were interpreted correctly.

5. Results and discussion

This section presents the results of applying the proposed DSS to our case study, according to the experimental design described in Section 4.4. Following Amarasinghe et al. (2024), we intend to highlight the interest of using the proposed DSS, in particular the DoTS metric, to select XAI methods with real-world practical utility, in line with feedback from stakeholders in the deployment context, so as to tackle different real-life case studies involving users with different levels of AI literacy. For each group of stakeholders, we analyzed the results in three dimensions, providing insights into how the XAI methods presented in Section 4.3 impact the levels of trust and satisfaction of the users and how the suggested context-aware system may be useful in selecting XAI methods with practical interest.

- Section 5.1 starts by examining how the different explanations provided by XAI methods impact stakeholders' levels of trust in the explanandum.
- Section 5.2 analyzes stakeholders' levels of satisfaction with the explanations produced by each of the different XAI methods, as well as how these explanations might contribute to their decision-making process.
- Lastly, Section 5.3 presents and discusses the results of the application of the DoTS metric, which combines the levels of trust and user satisfaction in the overall XAI system so as to provide estimates of both suitability and utility of each XAI method to each stakeholder group and to the problem under analysis.

5.1. *On the trust in the explanandum provided with explanations*

We start by analyzing and comparing the stakeholders' overall levels of trust (T_g) in the explanandum after introducing the explanations provided by each XAI method. Such levels are derived directly from Eq. (1). Fig. 3 shows the trust levels obtained for the different XAI methods on both stakeholder groups (the technical domain experts and researchers).

We found that the consistency between the answers to different items in the *Trust Scale Recommended for XAI* was satisfactory to high, as reflected by Cronbach's alpha values obtained across methods within each group ($\alpha \in [0.64, 0.98]$). The results show that the trust levels obtained are modest regardless of the XAI method used. This means that, for our population sample, the generation of explanations for the explanandum do not reflect in a significant increase in the trust levels of stakeholders. A comparison of the trust levels obtained for the different

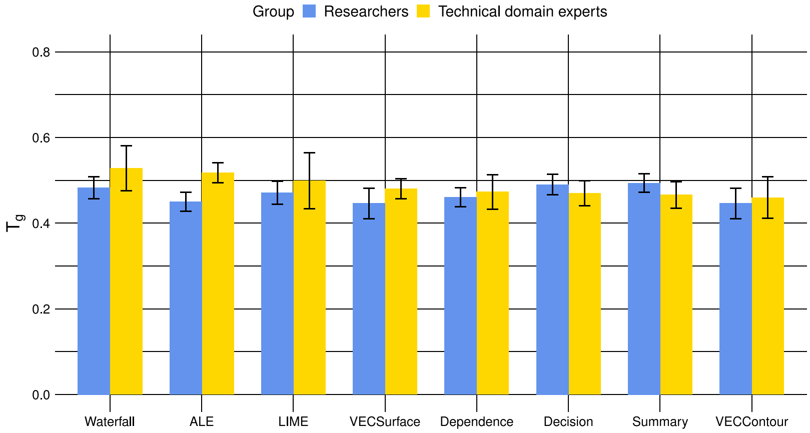


Fig. 3. Trust indexes in the explanandum provided with explanations. The error bars represent the standard error of T_g across stakeholders.

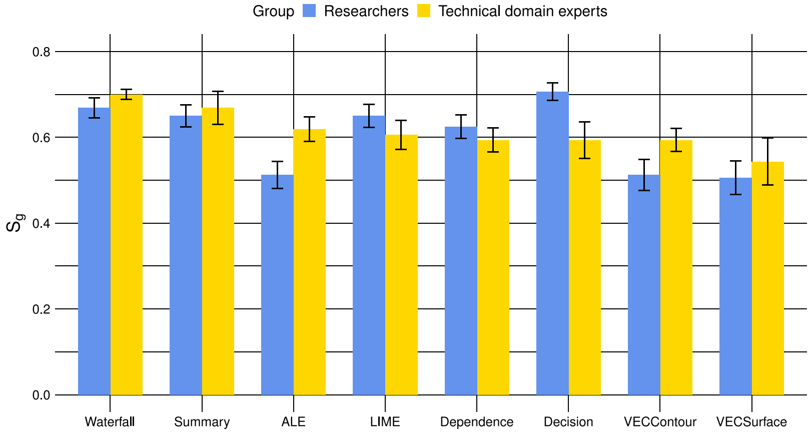


Fig. 4. Satisfaction indexes with the explanations provided by the XAI methods. The error bars represent the standard error of S_g across stakeholders.

XAI methods reveals that the Waterfall plot is the one that induces the highest level of trust in technical domain experts, followed by the ALE plot and the LIME plot. In contrast, the Summary plot and Decision plot methods are the ones that elicit the highest levels of trust in the group of researchers, albeit without significant differences between them. Here, it is interesting to note some disagreement and conflicting perspectives among the two groups. In particular, the two methods that induce the highest levels of trust in the group of researchers (Summary plot and Decision plot) are also those that generate high levels of mistrust among technical domain experts. Both groups seem to agree, however, that VEC contour plot is the least supportive of trust in the explanandum.

5.2. On the satisfaction with explanations

In a second experiment, we examined the stakeholders' levels of satisfaction with the explanations produced by the different XAI methods (S_g, Eq. (2)). Here, we also found that the consistency between the answers to different items in the *Explanation Satisfaction Scale* was satisfactory to high, as reflected by Cronbach's alpha values obtained across methods and groups ($\alpha \in [0.66, 0.97]$). Fig. 4 shows the different levels of satisfaction obtained for each explainability method in each of the test groups.

The results show that the levels of stakeholders' satisfaction with the explanations produced are, in general, higher than those related to stakeholders' trust in the explanandum provided with explanations, regardless of the XAI method considered. This suggests that in spite of the fact that the explanations provided fail to induce sizable increases in the trust levels with the explanandum, stakeholders are satisfied with the quality of the explanations produced. In particular, the results reveal that the Waterfall plot is the best-performing method in terms of satisfaction levels among technical domain experts and the second-best among researchers, followed by the Summary plot. However, as with the evaluation of trust, there are also some contrasting evaluations in this case. While the ALE plot is the third method with the highest satisfaction index among technical domain experts, it is the second worst method among researchers together with the VEC contour plot. By contrast, the Decision plot is the best explainability method for researchers in terms of satisfaction with the explanations derived therefrom, but it ranks in the bottom three for technical domain experts.

In order to better understand the utility of the explanations produced by each XAI method in the stakeholders' decision-making process, we focused our attention on the answers to the question *"Q6S: This explanation of how the ML model works is useful to my goals"* (Fig. 5).

The results agree with those shown in Fig. 4, with Summary, Waterfall and Decision plots being leading methods for producing explanations with practical utility for carrying out business tasks.

5.3. On the degree of trust-satisfaction of XAI methods

When looking at the measures of trust and satisfaction, we can grasp interesting insights into the value of the different XAI methods to both groups of stakeholders. However, the use of these measures in isolation can hamper the process of deciding which of these methods best suits the context and the business problem at hand. This holds true particularly when the number of explainability methods increases. Hence, we propose to take advantage of the DoTS metric (Eq. (4)),

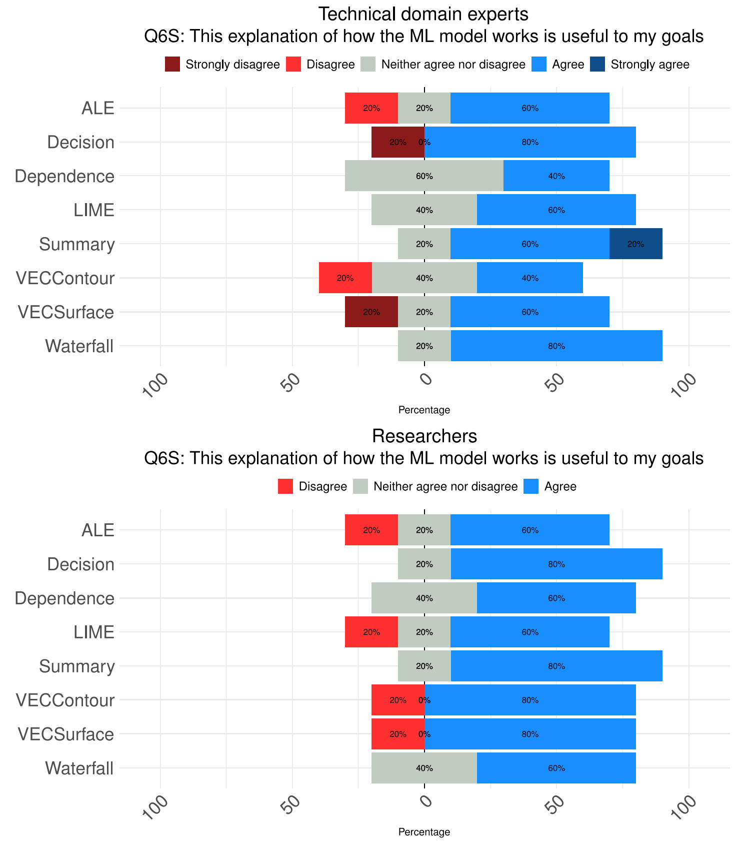


Fig. 5. Distribution of satisfaction scores across stakeholders within each test group for the question Q6S.

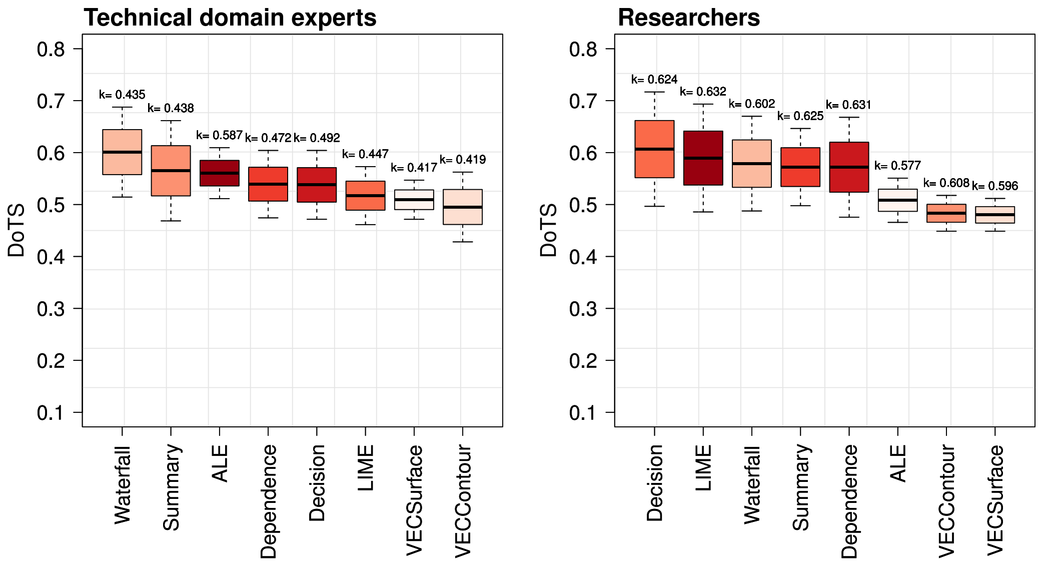


Fig. 6. Distribution of DoTS for each stakeholders' group using different combinations of weights $(w_{I,r}, 1 - w_{I,r})$. The color gradient illustrates the agreement, via weighted kappa statistic (κ), between all the $r \in R$ stakeholders within each group when evaluating each XAI method. Darker red colors reflect greater agreement. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

which combines these two measures so as to facilitate the selection of the XAI methods with the greatest utility for business stakeholders. Fig. 6 shows the distribution of the DoTS values for each method and stakeholders' group, assuming different weights for the level of trust in the explained model (w_I) and for the level of satisfaction with the explanations produced ($1 - w_I$), ranging from 0 to 1 in steps of 0.025.

We consider that $w_E = 0.25$ if the stakeholders have less than 3 years' experience; $w_E = 0.5$ for stakeholders' experience within $[4, 6]$ years; $w_E = 0.75$ for stakeholders' experience within $[7, 10]$ years; and $w_E = 1$ if stakeholders' experience is greater than 10 years.

To probe deeper into the stakeholders' preferences, we also evaluate the agreement between the responses in each group for each XAI method. To do this, we compute the pairwise agreement between all stakeholders $r \in R$ within each group using the weighted kappa statistic (Cohen, 1968). Then, for each explainability method, we calculate the average agreement (κ), consisting of the average pairwise

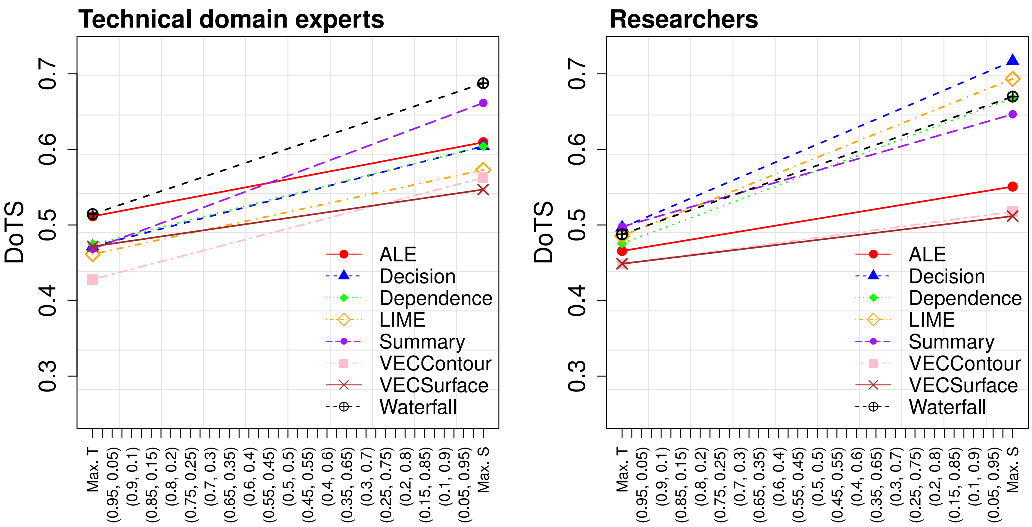


Fig. 7. Sensitivity analysis of the DoTS metric for different combinations of $(w_{I,r}, 1 - w_{I,r})$.

agreement among stakeholders. This is illustrated in Fig. 6 by a color gradient scheme, where darker red colors reflect greater agreement among stakeholders.

A first comparison between both groups reveals that the overall levels of DoTS obtained for researchers are slightly higher than those obtained for technical domain experts, regardless the XAI method. Comparing the XAI methods ranked according to the levels of trust and satisfaction obtained from the interaction with the different stakeholders (Figs. 3 and 4, respectively), we can observe that the DoTS metric reflects their actual feedback. This validates the DoTS metric in the sense that it adequately quantifies human inputs. Focusing the analysis on each of the groups separately, we observe that, according to the technical domain experts, the Waterfall, Summary and ALE plots are those with the highest degree of trust-satisfaction, with the ALE plot proving to be the method with the highest levels of agreement within the group. From the researchers' perspective, the Waterfall plot also ranks in the top three methods with the greatest DoTS, but after the Decision plot and the LIME plot, which turns out to be the one with the highest levels of intra-group agreement. Interestingly, from our contact with both groups, we found that researchers were more interested in understanding the outputs of the explanandum rather than the explanandum as a whole. In sharp contrast, the technical domain experts sought an XAI technique that was able to provide them with an overall understanding of the explanandum and not just its outcomes. This provides a rationale for the fact that the best performing methods for researchers, in terms of maximizing DoTS, offer local explanations, while two of the three top-performing methods for technical domain experts offer global explanations. This consistency between the obtained results and the stakeholders' needs validates the practical interest of DoTS in selecting XAI methods with real utility for the stakeholders in light of their explainability needs.

A more detailed breakdown of the DoTS obtained for the different values $w_{I,r}$ (Fig. 7) shows that the selection of an XAI method g generally depends on the stakeholders' preferences regarding the measures of trust (T_g) and satisfaction (S_g).

In particular, for the group of technical domain experts, we observe in the left panel of Fig. 7 that the Waterfall is the only technique that shows the highest values of DoTS regardless of the weight combination considered. This dominance does not hold for most of the remaining methods. For instance, we observe that the Summary method only exhibits higher DoTS whenever greater importance is given to the measure S rather than to the measure T. If the goal is to maximize the levels of trust in the explanandum provided with explanations, the ALE method is preferable to the Summary method. Another example is the case of VEC-based methods, where VEC surface only demonstrates better trust-satisfaction levels compared to those of VEC contour when the goal is to maximize T. If the goal is to maximize S, VEC contour outperforms VEC surface. Examining the results from the perspective of the researchers (right panel of Fig. 7), we observe that the Decision plot is the one with the highest degree of trust-satisfaction, either in the sense of maximizing T or in the sense of maximizing S. Conversely, VEC-based methods are those which, from the standpoint of the group, exhibit the lowest DoTS.

Regardless of the combination of weights adopted, it is still important to evaluate the extent of agreement among stakeholders in selecting the XAI method that is most adaptable to the problem and business context. Fig. 8 illustrates the relationship between the average levels of agreement of the stakeholders within each group and the DoTS, assuming a scenario where the measures T and S are equally relevant.

We found that maximizing the DoTS and maximizing the average degree of agreement between stakeholders (κ) may prove to be conflicting objectives. For instance, in the group of technical domain experts (left panel of Fig. 8), we found that although the Waterfall method had the highest degree of trust-satisfaction it is also one of the methods showing the most disagreement between the stakeholders who evaluated it. This inflation of DoTS is due to the positive feedback from a more experienced stakeholder that contributed to valuing the Waterfall method over the others. At this point, if the agreement between stakeholders prevails, we favor the selection of the ALE method without incurring in a sharp drop in the DoTS. In the group of researchers (right panel of Fig. 8), we found no significant disagreements between the methods evaluated, with the XAI methods presenting the highest DoTS also revealing the highest levels of agreement between stakeholders. These findings suggest the use of an agreement metric as a complement to the DoTS metric whenever evaluating the utility of a given XAI method.

6. Practical and managerial implications

We highlight several practical implications arising from our work for the use and management of XAI-based systems. First, the proposed DSS promotes the interaction of stakeholders not only in the development of the explanandum but also in the assessment of trust and satisfaction with the overall XAI system. Given the myriad of existing XAI methods, and the fact that each AI problem may have different business specifications, the proposed system makes it possible to test any type of explanation with different stakeholders who directly

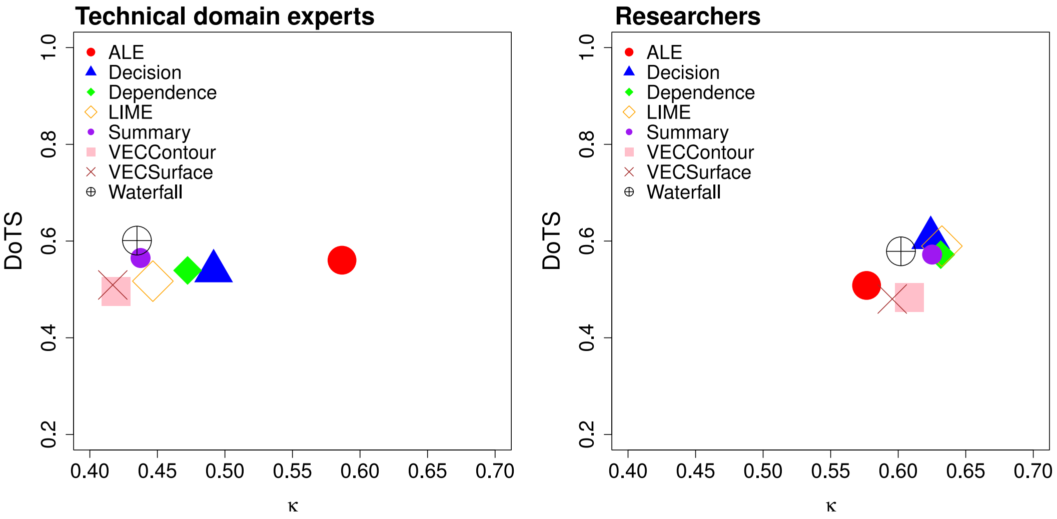


Fig. 8. Illustration of the DoTS for the different methods according to the average agreement level of the stakeholders within each group (κ), assuming $w_{I,r} = 1 - w_{I,r} = 0.5$.

interact with it in their daily business. We believe that our system is sufficiently understandable and flexible to be applied to any group of stakeholders, with different AI literacy levels, in any organizational context. This is a relevant aspect of our work, as current literature tends to neglect how different explanations should be presented and evaluated by less proficient AI users (Haque et al., 2023).

Second, the feedback provided by the stakeholders' assessments is used to formulate a flexible analytical metric – the DoTS – able to estimate the degree of trust-satisfaction of each XAI method and to quantify its utility to the business context. The proposed metric is simple to manipulate and it is sensitive to the degree of experience and/or knowledge of the stakeholders. Moreover, the DoTS is designed to be scalable. In fact, although the DoTS metric defined in Eq. (4) only covers two measures, i.e., stakeholders' trust and satisfaction in the XAI system developed, other indicators can easily be included in the formulation if they prove useful for the context and research problem under analysis. The DoTS metric also enables stakeholders to define the degree of importance that each measure should have when selecting explainability methods. This is particularly relevant as explainability needs can vary across stakeholders, depending, for instance, on how often the XAI system is used in their decision-making processes or how they perceive the risk associated with the task being explained.

Third, we tested the applicability of the proposed system in a focal company in the context of supply chain demand management. Our empirical results show the relevance of the proposed system in promoting the selection of XAI methods that meet the explainability needs of different stakeholders with different degrees of expertise and analytical maturity. For our data, our results demonstrated that academic stakeholders (end-users) tend to prefer local explanations, whereas technical domain experts tend to favor XAI methods that provide global explanations.

6.1. Implications for understandability and justifiability in business operations

This work offers further implications for understandability and justifiability in real-world business environments, two features that an XAI system needs to offer to effectively impact the decision-making process (De Bock et al., 2024). Following De Bock et al. (2024), understandability is the system's ability to allow users to understand how a given AI model operates and how the solutions generated from it were obtained. On the other hand, justifiability refers to the capacity of the system to help the users judge whether the results of the AI model match their intuition, based on their business knowledge. In short, it allows the user to trust in the model. Recalling that the DoTS metric embedded in the proposed DSS takes advantage of measures of trust and user satisfaction as a way of assisting the selection of XAI methods, we provide some examples of real-world cases that could benefit from the application of our approach.

Healthcare. The implementation of our DSS in healthcare can offer valuable insights into reducing resistance to adopting AI systems for decision support in clinical contexts. For example, given an AI system designed to help detect a specific pathology, the DoTS metric can be used to select the XAI techniques that best suit the explanatory needs of the interested clinicians and enable them to understand, trust and, most importantly, use that system as a decision-making tool during their medical practice. Even the patients themselves can benefit from using our system by being able to understand, for instance, what factors led the AI system to suggest such diagnosis.

Legal operations. AI models, especially those based on natural language processing, have been used to automatically generate business contracts so as to reduce the intensive human effort involved in these tasks. Yet, recent studies (see, e.g., Giampieri, 2024) have identified numerous vulnerabilities in these models, pointing to the need for human supervision as a way of minimizing the introduction of misleading or even legally unfounded contractual clauses. The use of XAI techniques has proved useful for legal stakeholders to understand the algorithmic rationale behind the resulting output and to identify potential gaps in AI models (Stathis and van den Herik, 2024). The main contribution of the proposed DSS in this context is to help legal stakeholders who want to take advantage of AI tools in the contract drafting/revision process, by suggesting XAI methods that provide such stakeholders with a clear understanding of how the AI model generates contractual information.

Supply chain demand forecasting. In supply chain management, it has been common practice to use machine learning models to capture non-linear demand patterns. The "black box" nature of this type of models might result in users not trusting the models developed, which in turn motivates them to combine statistical forecasts with judgmental forecasts. However, it is well known that this strategy can be biased (Fildes and Goodwin, 2007). The DoTS metric can be relevant in this context, allowing decision-makers to select suitable XAI methods that help them to understand the nature and results of forecasting models, while avoiding excessive recourse to manual and subjective adjustments of forecasts.

Overall, the proposed DSS can help decision-makers in various organizational environments to find the XAI method that best suits the needs of the context and the profile of the users operating in it, thereby fostering the understanding and implementation of AI systems in practice.

6.2. Implications for decision-making using contextual factors

From a practical perspective, the proposed DoTS metric has a particularly relevant feature, namely its sensitivity to contextual factors that may differ from one organization to another. Since it is defined in a context-agnostic fashion, when applied in different business environments involving users with different profiles and different explanatory needs, the same XAI method may be classified as the most adequate for one context but fail to be so in others. This provides important flexibility to the DoTS metric, making it suitable for users seeking different types of explanations, depending on the business context in which they operate. Examples of these explanations could be (1) explanations by influence; (2) visual explanations; (3) explanations by simplification; (4) explanations based on examples or even (5) textual explanations (for details see Gerlach et al., 2022).

Given the fact that it is entirely driven by human feedback and characteristics, the proposed DSS can be tailored to various types of business and, given a set of XAI techniques, suggest those that best suit the users of that particular context. For instance, in small-sized organizations, which typically do not operate with highly qualified AI personnel, our DSS may suggest XAI techniques that are easy to implement and that offer global explanations for the AI models constructed. In contrast, in large-sized organizations, we expect the users to exhibit greater analytical maturity and, therefore, provide feedback in order to select more granular explanation techniques that impact decision-making at various organizational levels.

7. Conclusion and directions for future research

This work investigates how the business stakeholders' desiderata and feedback can be included in the process of choosing suitable XAI methods for a given AI-based problem and context. Frequently, XAI methods tend to be evaluated solely by means of objective metrics, which overlook the utility and potential impact that these methods might have on the decision-making processes of business stakeholders. We propose a context-aware DSS to deal with the problem of deciding which XAI methods, from a pre-defined set, are most adapted to the problem and to the users of the underlying business context. The proposed approach is driven by the preferences of the business stakeholders, so as to meet their explainability interests and business desiderata. We introduce an analytical metric that estimates the degree of trust-satisfaction of a given XAI method. This metric is fundamentally based on the stakeholders' trust in the explanandum provided with explanations, as well as their satisfaction with the quality and utility of those explanations. We tested the proposed system with real supply chain data, in a real AI-based business task involving real users from a major automotive electronics organization.

From a theoretical perspective, while most of the XAI literature fails to employ application-grounded methodologies to evaluate the utility of explainability methods with end-users (cf. Langer et al. (2021), Jesus et al. (2021), Amarasinghe et al. (2023, 2024)), we follow a human-in-the-loop approach, focusing the evaluation of XAI methods on business stakeholders who can take advantage of these methods to support their decision-making processes. Our work is consistent with recent recommendations to promote evaluations of XAI methods in an application-grounded environment (Amarasinghe et al., 2023) and to measure the degree of usefulness and quality of explanations for users (Saeed and Omlin, 2023). Other relevant works (e.g., Schmidt and Biessmann, 2019) take advantage of trust measures as a way of evaluating the quality of XAI methods, but do not incorporate the subjective feedback nor the degree of expertise of real users into the process. From a practical perspective, published studies with application-grounded evaluations are typically focused on the context of finance (as in Jesus et al., 2021; Amarasinghe et al., 2024) or healthcare (as in Lundberg et al., 2018). However, to the best of our knowledge, this is the first paper to evaluate the practical utility of different explainability methods in the context of supply chain management, a field with extensive research and significance in the context of operations research and operations management (De Bock et al., 2024).

Despite the theoretical and practical interest of our work, we stress some important limitations that could serve as a motivation for future work in this context. Firstly, we acknowledge that the number of participants involved in the experiment is relatively small and could have an impact on the correct interpretation of our results. We expect to test the proposed system with a wider range of users, ideally working in different business contexts, so as to facilitate the generalization of our results and to improve statistical power. Yet, it is well-known (Amarasinghe et al., 2023, 2024) that the process of evaluating XAI systems with real users is logistically challenging and may require several iterations between all the stakeholders involved, making it difficult to engage their interest and acceptance. Secondly, our work evaluates the explanations provided by each XAI method under analysis in a static environment based on preconstructed explanations for the explanandum. A promising research avenue is the development of interactive and explainable intelligent systems that dynamically explore the inferences of predictive models with, for instance, counterfactual explanations. A further relevant step for future research includes adding other measures to the proposed DoTS metric, in addition to those used in this study. In this respect, if more dimensions of analysis are introduced into the DoTS metric, the use of metaheuristics (Khanduja and Bhushan, 2021) could be explored as a way of optimizing the weights to be assigned to each dimension according to one or more objective functions of interest to users in the business context.

CRediT authorship contribution statement

Marcelo I. Reis: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **João N.C. Gonçalves:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Paulo Cortez:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **M. Sameiro Carvalho:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **João M. Fernandes:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by: European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project nº179826; Funding Reference: SIFN-01-9999-FN-179826]. The authors would like to thank Bosch Car Multimedia, Braga, Portugal, for providing the materials and support needed to carry out this research. In particular, we would like to thank Tiago M. Martins for his collaboration in the interviews to collect the explainability requirements; José M. Silva for his suggestions on the presentation of model results; and José M. Silva, Maria Pires and Luís Marques for the source code revisions. Finally, we would like to thank the anonymous reviewers for their helpful suggestions.

Data availability

The authors do not have permission to share data.

References

Abolghasemi, M., Beh, E., Tarr, G., Gerlach, R., 2020. Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. Comput. Ind. Eng. 142, 106380.

Abusitta, A., Li, M.Q., Fung, B.C., 2024. Survey on explainable AI: techniques, challenges and open issues. Expert Syst. Appl. 255, 124710.

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 52138–52160.

Agarwal, C., Saxena, E., Krishna, S., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H., 2022. OpenXAI: Towards a transparent evaluation of post hoc model explanations. In: Advances in Neural Information Processing Systems. Vol. 35, pp. 15784–15799.

Al-Ansari, N., Al-Thani, D., Al-Mansoori, R.S., 2024. User-centered evaluation of explainable artificial intelligence (xai): A systematic literature review. Hum. Behav. Emerg. Technol. 2024 (1), 4628855.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., 2023. Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Inf. Fusion 99, 101805.

Aliyeva, K., Mehdiyev, N., 2024. Uncertainty-aware multi-criteria decision analysis for evaluation of explainable artificial intelligence methods: A use case from the healthcare domain. Inform. Sci. 657, 119987.

Allen, I.E., Seaman, C.A., 2007. Likert scales and data analyses. Qual. Prog. 40 (7), 64–65.

Amarasinghe, K., Rodolfa, K.T., Jesus, S., Chen, V., Balayan, V., Saleiro, P., Bizarro, P., Talwalkar, A., Ghani, R., 2024. On the importance of application-grounded experimental design for evaluating explainable ML methods. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38, pp. 20921–20929, 19.

Amarasinghe, K., Rodolfa, K.T., Lamba, H., Ghani, R., 2023. Explainable machine learning for public policy: Use cases, gaps, and research directions. Data Policy 5, e5.

Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M., 2021. Explainable artificial intelligence: an analytical review. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 11 (5), e1424.

Apley, D.W., Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. J. R. Stat. Soc. Ser. B Stat. Methodol. 82 (4), 1059–1086.

Arias-Duart, A., Parés, F., Garcia-Gasulla, D., Giménez-Ábalos, V., 2022. Focus! Rating XAI methods and finding biases. In: 2022 IEEE International Conference on Fuzzy Systems. pp. 1–8.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Barbosa-Póvoa, A.P., da Silva, C., Carvalho, A., 2018. Opportunities and challenges in sustainable supply chain: An operations research perspective. European J. Oper. Res. 268 (2), 399–431.

Barros, J., Gonçalves, J.N., Cortez, P., Carvalho, M.S., 2023. A decision support system based on a multivariate supervised regression strategy for estimating supply lead times. Eng. Appl. Artif. Intell. 125, 106671.

Batterton, K.A., Hale, K.N., 2017. The likert scale what it is and how to use it. Phalanx 50 (2), 32–39.

Bertolini, M., Mezzogori, D., Neroni, M., Zammori, F., 2021. Machine learning for industrial applications: A comprehensive literature review. Expert Syst. Appl. 175, 114820.

Brasse, J., Broder, H.R., Förster, M., Klier, M., Sigler, I., 2023. Explainable artificial intelligence in information systems: A review of the status quo and future research directions. Electron. Mark. 33 (1), 26.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brem, A., Giones, F., Werle, M., 2021. The AI digital revolution in innovation: A conceptual framework of artificial intelligence technologies for the management of innovation. IEEE Trans. Eng. Manage..

Burger, M., Nitsche, A.-M., Arlinghaus, J., 2023. Hybrid intelligence in procurement: Disillusionment with AI's superiority? Comput. Ind. 150, 103946.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: 22nd ACM International Conference on Knowledge Discovery and Data Mining. pp. 785–794.

Cohen, J., 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. 70 (4), 213.

Cortes, C., Vapnik, V., 1995. Support vector machine. Mach. Learn. 20 (3), 273–297.

Cortez, P., Embrechts, M.J., 2011. Opening black box data mining models using sensitivity analysis. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining. CIDM, pp. 341–348.

Cortez, P., Embrechts, M.J., 2013. Using sensitivity analysis and visualization techniques to open black box data mining models. Inform. Sci. 225, 1–17.

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16 (3), 297–334.

Cugny, R., Aligon, J., Chevalier, M., Roman Jimenez, G., Teste, O., 2022. AutoXAI: A framework to automatically select the most adapted XAI solution. In: 31st ACM International Conference on Information & Knowledge Management. pp. 315–324.

De Bock, K.W., Coussement, K., De Caigny, A., Słowiński, R., Baesens, B., Boute, R.N., Choi, T.-M., Delen, D., Kraus, M., Lessmann, S., et al., 2024. Explainable AI for operational research: A defining framework, methods, applications, and a research agenda. European J. Oper. Res. 317 (2), 249–272.

Dengler, S., Lahriri, S., Trunzer, E., Vogel-Heuser, B., 2021. Applied machine learning for a zero defect tolerance system in the automated assembly of pharmaceutical devices. Decis. Support Syst. 146, 113540.

Doshi-Velez, F., Kim, B., 2018. Considerations for evaluation and generalization in interpretable machine learning. In: Explainable and Interpretable Models in Computer Vision and Machine Learning. pp. 3–17.

Doumard, E., Aligon, J., Escriva, E., Excoffier, J.-B., Monsarrat, P., Soulé-Dupuy, C., 2023. A quantitative approach for the comparison of additive local explanation methods. Inf. Syst. 114, 102162.

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

Enholm, I.M., Papagiannidis, E., Mikalef, P., Krogstie, J., 2022. Artificial intelligence and business value: A literature review. Inf. Syst. Front. 24 (5), 1709–1734.

Fildes, R., Goodwin, P., 2007. Against your better judgment? how organizations can improve their use of management judgment in forecasting. Interfaces 37 (6), 570–576.

Gangwani, D., Zhu, X., 2024. Modeling and prediction of business success: A survey. Artif. Intell. Rev. 57 (2), 1–51.

Gerlach, J., Hoppe, P., Jagels, S., Licker, L., Breitner, M.H., 2022. Decision support for efficient XAI services - a morphological analysis, business model archetypes, and a decision tree. Electron. Mark. 32 (4), 2139–2158.

Giampieri, P., 2024. AI-powered contracts: a critical analysis. Int. J. Semiot. Law-Rev. Int. Sémiot. Juridique 1–18.

Gonçalves, J.N., Cortez, P., Carvalho, M.S., Frazão, N.M., 2021. A multivariate approach for multi-step demand forecasting in assembly industries: Empirical evidence from an automotive supply chain. Decis. Support Syst. 142, 113452.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D., 2018. A survey of methods for explaining black box models. ACM Comput. Surv. 51 (5), 1–42.

Guo, Y., Li, Y., Liu, D., Xu, S.X., 2023. Measuring service quality based on customer emotion: An explainable ai approach. Decis. Support Syst. 114051.

Haan, K., Watts, R., 2023. Top AI statistics and trends. https://www.forbes.com/advisor/business/ai-statistics/, [Online; Accessed 10 March 2024].

Haque, A.B., Islam, A.N., Mikalef, P., 2023. Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technol. Forecast. Soc. Change 186, 122120.

Hase, P., Bansal, M., 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In: 58th Annual Meeting of the Association for Computational Linguistics. pp. 5540–5552.

Hoffman, R.R., Jalaeian, M., Tate, C., Klein, G., Mueller, S.T., 2023a. Evaluating machine-generated explanations: a "scorecard" method for XAI measurement science. Front. Comput. Sci. 5, 1114806.

Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J., 2023b. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Front. Comput. Sci. 5, 1096257.

Islam, S.R., Eberle, W., Ghafoor, S.K., 2020. Towards quantification of explainability in explainable artificial intelligence methods. In: 33rd International Florida Artificial Intelligence Research Society Conference. AAAI Press, pp. 75–81.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al., 2023. An Introduction to Statistical Learning. Vol. 2, Springer.

Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, M., Kuusk, A., 2023. Artificial intelligence for industry 4.0: Systematic review of applications, challenges, and opportunities. Expert Syst. Appl. 216, 119456.

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., Gama, J., 2021. How can I choose an explainer? An application-grounded evaluation of post-hoc explanations. In: 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 805–815.

Joseph, R.V., Mohanty, A., Tyagi, S., Mishra, S., Satapathy, S.K., Mohanty, S.N., 2022. A hybrid deep learning framework with CNN and bi-directional LSTM for store item demand forecasting. Comput. Electr. Eng. 103, 108358.

Khanduja, N., Bhushan, B., 2021. Recent advances and application of metaheuristic algorithms: A survey (2014–2020). In: Metaheuristic and Evolutionary Computation: Algorithms and Applications. pp. 207–228.

Kim, D.J., Ferrin, D.L., Rao, H.R., 2009. Trust and satisfaction, two stepping stones for successful e-commerce relationships: A longitudinal exploration. Inf. Syst. Res. 20 (2), 237–257.

Kim, D., Song, Y., Kim, S., Lee, S., Wu, Y., Shin, J., Lee, D., 2023. How should the results of artificial intelligence be explained to users?-research on consumer preferences in user-centered explainable artificial intelligence. Technol. Forecast. Soc. Change 188, 122343.

Kostopoulos, G., Davrazos, G., Kotsiantis, S., 2024. Explainable artificial intelligence-based decision support systems: A recent review. Electronics 13 (14), 2842.

Kotriwala, A., Klöpper, B., Dix, M., Gopalakrishnan, G., Ziobro, D., Potschka, A., 2021. XAI for operations in the process industry-applications, theses, and research directions. In: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering. pp. 1–12.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., Baum, K., 2021. What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 296, 103473.

Leff, A., Rayfield, J.T., 2001. Web-application development using the model/view/controller design pattern. In: 5th IEEE International Enterprise Distributed Object Computing Conference. pp. 118–127.

Lian, Z., Deshmukh, A., Wang, J., 2006. The optimal frozen period in a dynamic production model. Int. J. Prod. Econ. 103 (2), 648–655.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.

Lundberg, S.M., Nair, B., Vavilala, M.S., Horibe, M., Eisses, M.J., Adams, T., Liston, D.E., Low, D.K.-W., Newman, S.-F., Kim, J., et al., 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. 2 (10), 749–760.

Mersha, M., Lam, K., Wood, J., AlShami, A., Kalita, J., 2024. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. Neurocomputing 128111.

Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N., 2022. Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. 1–66.

Miró-Nicolau, M., Jaume-i Capó, A., Moyà-Alcover, G., 2024. Assessing fidelity in XAI post-hoc techniques: A comparative study with ground truth explanations datasets. Artificial Intelligence 335, 104179.

Mohseni, S., Zarei, N., Ragan, E.D., 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. (TiiS) 11 (3–4), 1–45.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C., 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Comput. Surv. 55 (13s), 1–42.

Nimmy, S.F., Hussain, O.K., Chakrabortty, R.K., Hussain, F.K., Saberi, M., 2022. Explainability in supply chain operational risk management: A systematic literature review. Knowl.-Based Syst. 235, 107587.

Norman, G., 2010. Likert scales, levels of measurement and the "laws" of statistics. Adv. Health Sci. Educ. 15, 625–632.

Olan, F., Spanaki, K., Ahmed, W., Zhao, G., 2024. Enabling explainable artificial intelligence capabilities in supply chain decision support making. Prod. Plan. Control 1–12.

Pawlicki, M., Pawlicka, A., Uccello, F., Szelest, S., D'Antonio, S., Kozik, R., Choraś, M., 2024. Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. Neurocomputing 602, 128282.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12 (85), 2825–2830.

Pinkus, A., 1999. Approximation theory of the MLP model in neural networks. Acta Numer. 8, 143–195.

Ribeiro, M.T., Singh, S., Guestrin, C., 2018. Anchors: High-precision model-agnostic explanations. In: 32nd AAAI Conference on Artificial Intelligence. pp. 1527–1535.

Riveiro, M., Thill, S., 2021. That's (not) the output I expected! On the role of end user expectations in creating explanations of AI systems. Artificial Intelligence 298, 103507.

Roeder, J., Palmer, M., Muntermann, J., 2022. Data-driven decision-making in credit risk management: The information value of analyst reports. Decis. Support Syst. 158, 113770.

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., Kasneci, E., 2024. Towards human-centered explainable AI: A survey of user studies for model explanations. IEEE Trans. Pattern Anal. Mach. Intell. 46 (4), 2104–2122.

Rosenfeld, A., 2021. Better metrics for evaluating explainable artificial intelligence. In: 20th International Conference on Autonomous Agents and Multiagent Systems. pp. 45–50.

Saeed, W., Omlin, C., 2023. Explainable AI (XAI): Core ideas, techniques, and solutions (XAI): A systematic meta-survey of current challenges and future opportunities. Knowl.-Based Syst. 263, 110273.

Salih, A.M., Galazzo, I.B., Gkontra, P., Rauseo, E., Lee, A.M., Lekadir, K., Radeva, P., Petersen, S.E., Menegaz, G., 2024. A review of evaluation approaches for explainable AI with applications in cardiology. Artif. Intell. Rev. 57 (9), 240.

Schmidt, P., Biessmann, F., 2019. Quantifying interpretability and trust in machine learning systems. In: AAAI 2019 Workshop on Network Interpretability for Deep Learning.

Schoonderwoerd, T.A., Jorritsma, W., Neerincx, M.A., Van Den Bosch, K., 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. Int. J. Hum.-Comput. Stud. 154, 102684.

Schwalbe, G., Finzel, B., 2023. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data Min. Knowl. Discov. 1–59.

Sovrano, F., Vitali, F., 2023. An objective metric for explainable AI: how and why to estimate the degree of explainability. Knowl.-Based Syst. 278, 110866.

Stathis, G., van den Herik, J., 2024. Ethical and preventive legal technology. AI Ethics 1–18.

Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. Int. J. Forecast. 16 (4), 437–450.

Tchuente, D., Lonlac, J., Kamsu-Foguem, B., 2024. A methodological and theoretical framework for implementing explainable artificial intelligence (XAI) in business applications. Comput. Ind. 155, 104044.

Tjoa, E., Guan, C., 2020. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. 32 (11), 4793–4813.

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., Fischl, M., 2021. Artificial intelligence in supply chain management: A systematic literature review. J. Bus. Res. 122, 502–517.

Tsiakas, K., Murray-Rust, D., 2022. Using human-in-the-loop and explainable AI to envisage new future work practices. In: 15th International Conference on PErvasive Technologies Related to Assistive Environments. pp. 588–594.

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

Vapnik, V., 1999. The Nature of Statistical Learning Theory. Springer Science & Business Media.

Vermeire, T., Laugel, T., Renard, X., Martens, D., Detyniecki, M., 2021. How to choose an explainability method? towards a methodical implementation of XAI in practice. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 521–533.

Vilone, G., Longo, L., 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion 76, 89–106.

Wamba-Taguimdje, S.-L., Fosso Wamba, S., Kala Kamdjoug, J.R., Tchatchouang Wanko, C.E., 2020. Influence of artificial intelligence (AI) on firm performance: The business value of AI-based transformation projects. Bus. Process Manag. J. 26 (7), 1893–1924.

Weber, P., Carl, K.V., Hinz, O., 2024. Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. Manag. Rev. Q. 74 (2), 867–907.

Zhang, Y., Chen, X., 2020. Explainable recommendation: A survey and new perspectives. Found. Trends Inf. Retr. 14 (1), 1–101.

Zhao, Q., Hastie, T., 2021. Causal interpretations of black-box models. J. Bus. Econom. Statist. 39 (1), 272–281.

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