7-8-2023

A User-Centric Approach to Explainable AI in Corporate Performance Management

Oliver A. Vetter
Technical University of Darmstadt, oliver.vetter@tu-darmstadt.de

Alexander Efremov
Technical University of Darmstadt, alexander@efremov.de

A User-Centric Approach to Explainable AI in Corporate Performance Management

Completed Research Paper

Oliver A. Vetter
Technical University of Darmstadt
Hochschulstr. 1, 64289 Darmstadt
oliver.vetter@tu-darmstadt.de

Alexander Efremov
Technical University of Darmstadt
Hochschulstr. 1, 64289 Darmstadt
alexander.efremov@stud.tu-darmstadt.de

Abstract

Machine learning (ML) applications have surged in popularity in the industry, however, the lack of transparency of ML-models often impedes the usability of ML in practice. Especially in the corporate performance management (CPM) domain, transparency is crucial to support corporate decision-making processes. To address this challenge, approaches of explainable artificial intelligence (XAI) provide solutions to reduce the opacity of ML-based systems. This design science study further builds on prior user experience (UX) and user interface (UI) focused XAI-research, to develop a user-centric approach to XAI for the CPM field. As key results, we identify design principles in three decomposition layers, including ten explainability UI-elements that we developed and evaluated through seven interviews. These results complement prior research by focusing it on the CPM domain and provide practitioners with concrete guidelines to foster ML adoption in the CPM field.

Keywords: Explainable artificial intelligence, corporate performance management, UX

Introduction

Machine learning (ML) use in organizations has grown rapidly in recent years (Jordan and Mitchell, 2015). In a study by McKinsey (2021), 56% out of 1.843 participants from different industry fields reported using ML in at least one use case. One area in which organizations can stand to benefit from utilizing ML systems is corporate performance management (CPM), a company practice that deals with strategic and tactical activities, such as planning, budgeting, and forecasting. ML-based techniques have much potential for supporting humans in these tasks, especially in forecasting, as ML can learn from previous data and take outside effects into account (Makridakis et al., 2022). The opaque and abstract character of ML models is a significant obstacle, however, particularly for decision-making (Adadi and Berrada, 2018). This issue has given rise to the study area of explainable artificial intelligence (XAI), which aims to elaborate to users how ML models function or how a particular prediction was generated (Adadi and Berrada, 2018). Research has developed and evaluated XAI techniques to gain insights into ML models and algorithms in order to generate these explanations (Adadi and Berrada, 2018; Guidotti et al., 2019).

Despite the potential of ML in forecasting use cases, the technology is rarely used in the CPM domain. For instance, in a study by McKinsey (2021), ML had the lowest adoption rates in the CPM-related business functions of strategy and corporate finance at 7% and 6%, respectively. Despite the research advances in the field of XAI, 34% of respondents in emerging economies and 44% percent in developed economies perceived the topic of explainability as a relevant risk when using ML (McKinsey, 2021). A possible reason for this gap could lie in the design of ML-based systems. As shown by DARPAS's XAI program, a high-quality user experience (UX) and user interface (UI) for XAI systems are crucial to foster user

understanding (Gunning et al., 2021). Moreover, there is no one-size-fits-all strategy for XAI, as various users in different circumstances demand diverse solutions (Gunning et al., 2021). UX-/UI-centered research has therefore evaluated different user groups, motivations for explanations, and formulated design principles (see Laato et al., 2022; Liao et al., 2020). Yet, many studies focus on providing explanations to data scientists, which is why recent literature calls for matching XAI techniques with more lay users and their needs (Liao et al., 2020; Brennen, 2020). This lack of focus on the end users may explain the gap between the low current adoption of ML systems and their considerable potential in the CPM sector, as low usability might discourage non-data-scientists from adopting and using ML. In this study, we thus aim to provide CPM- and user-specific answers to the following research questions:

Research question 1 (RQ 1): What are the goals of the user groups of ML-based CPM systems, and what kind of explanations do they require to achieve them?

Research question 2 (RQ 2): How can an ML-based CPM system effectively and efficiently provide these explanations to the users?

To answer these research questions, we follow the design science research process proposed by Peffers et al. (2006). Thereby, our study builds on preceding research concerning XAI techniques and their UX-/UI-centered design by developing and evaluating a user-centric approach to provide explanations in an existing ML-based CPM information system (CPM system) which is utilized in a forecasting use case and being developed by a European provider of enterprise service management solutions. To this end, we identify the goals and requirements of the user groups of the CPM system. We subsequently derive design principles (DPs) to meet these requirements in the UX, which we apply, refine, and evaluate through six interviews with management and CPM experts. With this study, we thus make important contributions to UX-/UI-centered research for XAI systems by distinguishing and validating DP for distinct user groups of ML-based CPM systems. Furthermore, to our knowledge, this is the first study that elaborates DP for UX-/UI-design tailored specifically for users in the CPM context and their unique requirements. Thus, we provide scholars with fertile ground for future research on user-centric XAI approaches in the CPM or adjacent domains while offering guidelines to practitioners that they can utilize to bridge the gap between CPM experts unfamiliar with ML and the potential promised by ML-powered CPM systems.

Theoretical Background

This section first describes the business practice of CPM and then expands on the utilization of ML solutions for planning and forecasting, taking XAI approaches and UX/UI-centric research into consideration.

Corporate Performance Management

CPM has been described as a system combining management processes with corresponding business intelligence (BI) information systems (Miranda, 2004). BI information systems enable companies to collect and analyze the data enabling CPM practices (Miranda, 2004; Frolick and Ariyachandra, 2006). The objective of CPM is generally to support corporate decision-making in ensuring the company performs well in its success metrics, such as revenue or profit (Frolick and Ariyachandra, 2006). Because of its quantitative nature, CPM thus usually focuses on metrics that can be expressed in financial figures, such as revenue and profit, as overarching goals (Frolick and Ariyachandra, 2006). On a deeper level and depending on the company's business model, CPM practitioners can also look into figures such as specific or aggregated sales numbers for a specific product or the costs a particular department generates (Frolick and Ariyachandra, 2006). There are different frameworks and approaches to summarizing the processes that make up the CPM of a company: generally, they all include processes for the planning, analysis, and monitoring of the predefined performance metrics (Richards et al., 2019; Frolick and Ariyachandra, 2006).

In the context of CPM, planning can be described as the process of gathering relevant information for CPM decision-making, such as budget allocations (Rogers et al., 1999). This information is used to base strategies around them (Rogers et al., 1999). For this purpose, it is important to use internal and external information to predict how certain figures will perform in the future (Richards et al., 2019). This step of forecasting is essential to the planning process, and the correct execution can be a critical step to outperform the company's competition (Frolick and Ariyachandra, 2006). Extensive research has been done on different types of forecasting (Bontempi et al., 2013). This ranges from stock price prediction to the CPM-relevant

use case of sales prediction (Pavlyshenko, 2019). In this study, we focus exclusively on use cases entailing the forecasting of business figures in the context of CPM. In practice, forecasting is often done in the form of a time series problem (Bontempi et al., 2013). This means a collection of historical data, all in the form of a time series, is used to predict future values based on these variables (Bontempi et al., 2013). There are different models and algorithms to transform the input variables into the desired output variables. Simple or advanced statistical methods, like forecasts, are often robust methods and offer their own advantages (Adya and Collopy, 1998; Makridakis et al., 2022; Spiliotis et al., 2019): They do not heavily rely on the amount and the quality of their input data and are also not very computation heavy because of the simplicity of their algorithms, making them the dominant methods for forecasting in the past. They do, however, also have some weaknesses, as they only prescribe the input data and do not recognize causation, for instance (Barker, 2020). The rise of data availability and quality, together with an increase in readily and cheaply available computation power, laid the groundwork for ML-based approaches to these forecasting use cases (Makridakis et al., 2022).

Explainable Artificial Intelligence for Forecasting

ML can be described as the algorithmic generation of a model from provided data by extracting patterns within the data (Russell and Norvig, 2021). Most modern artificial intelligence (AI) systems are implemented using ML technologies (Brynjolfsson and Mitchell, 2017). Therefore, we use the term ML to refer to ML-based instances of AI in this study. Time series forecasting can be done with supervised learning, a subcategory of ML (Bontempi et al., 2013). This is due to the circumstance that historical data is used, which is regularly labeled (Bontempi et al., 2013). A good example is historical sales numbers, as the label would be the sales volume mapped to the date the sale was concluded (Ma and Fildes, 2021). There exists a lot of research on ML-based approaches to forecasting, mostly focusing on designing the right type of algorithm for the defined use cases. Prior research includes ML-based solutions for energy forecasting, sales forecasting, or other financial figures (Ghoddusi et al., 2019; Pavlyshenko, 2019; Wasserbacher and Spindler, 2022). This research shows that ML-based approaches can perform quite well in planning use cases. In recent years deep learning algorithms are starting to catch up or surpass traditional algorithms in terms of performance (Hewamalage et al., 2021). Deep learning refers to especially complicated ML techniques, often associated with models with multilayered circuit structures, which are referred to as artificial neural networks (Russell and Norvig, 2021). They also exacerbate one of the biggest disadvantages of ML: The actual or perceived lack of transparency and the lack of explainability (Gunning et al., 2021). This disadvantage has brought out another important aspect of ML, XAI, which will be a focus of our study (Adadi and Berrada, 2018).

ML models are often described and perceived as a black box, and XAI tries to address this (Brennen, 2020). Therefore, XAI can be summarized as approaches seeking to explain aspects of ML-based systems to their users and stakeholders (Langer et al., 2021). For terminological clarity, we first describe the concepts of interpretability and explainability as understood within this paper. Interpretability can be defined as the grade to which the model and its predictions are interpretable to the user (Russell and Norvig, 2021). This means that based only on the inspection of the model, a human can derive the reasoning behind a certain output and could also predict the output for a different input. Linear regressions or decision trees are interpretable because the human can simply go along the tree with the respective input or calculate according to the regression parameters (Russell and Norvig, 2021). Following this definition, interpretability can only be achieved by choosing an interpretable algorithm. Examples of non-interpretable algorithms are deep learning algorithms due to their complex structure and the vast number of parameters and layers they use (Castelvecchi, 2016). They and other uninterpretable ML algorithms are commonly referred to as black box models or algorithms (Castelvecchi, 2016). In comparison to interpretability, explainability is not something inherent to the algorithm. It can be provided by posthoc processes, e.g., by another algorithm being trained on top of the model to be explained. This new algorithm is in itself interpretable but performs similarly to the one to be explained. Because the explanation occurs after the model is already trained, methods designed to provide explainability can also be described as posthoc techniques (Arrieta et al., 2020; Russell and Norvig, 2021).

Most posthoc techniques are model-agnostic, i.e. they can be applied posthoc on every type of ML model. These posthoc techniques themselves can be divided into two categories: Global and local techniques (Murdoch et al., 2019): Global techniques operate on a dataset level, providing explainability into global relations and patterns the model has learned. Local techniques operate on a prediction level, providing

explainability into individual predictions. Arrieta et al. (2020) also provide further, not mutually exclusive, distinctions of posthoc techniques, such as explanations by simplification, feature relevance or simply visual explanations: Explanations by simplification generally try to extract rules based on how the model works and explain those rules to the users. Feature relevance refers to a wide array of explanations regarding the input variables of the model i.e. the features. They use different approaches like game theory to learn how to explain the features, and afterward provide information on e.g. the influence of the features on the model, the importance of the features to the model performance, or the interaction of the different features. Visual explanations in turn can be used to present the information gathered from the feature relevance techniques. As presented in this subsection, model-agnostic, posthoc techniques allow for abstraction from the utilized ML algorithms. Consequently, we focus the presented UX/UI approach on these to ensure the transferability of our results for most ML-based solutions.

UX-focused XAI Research

When designing a UX-focused approach, different aspects need to be kept in mind. First, it is important to provide explanations into adjacent fields in the ML-based systems, such as the model performance or the used data (Liao et al., 2020). This also motivates a more general approach in terms of looking at the whole UX of the ML-based system, instead of just the parts that are ML-related in the narrow sense. Liao et al. (2020) provide key factors that influence user requirements and design recommendations. Among others, these include the motivation of the users to obtain explainability, the type of users in terms of e.g. domain knowledge or prior experience with data science topics as well as their role in the ML-based system, and lastly the decision context in which the explanations are provided. Because of the focus on the forecasting use case and posthoc model-agnostic techniques, other factors identified, such as the data and algorithm type, are omitted for this study. In a data science-heavy domain, motivations for explainability can be divided into debugging the model, identifying biases, and building trust (Brennen, 2020). Liao et al. (2020) argue that the sheer volume of different contexts and user motivations makes it difficult to predefine general explainability needs. They, therefore, describe a question-driven approach to generating UX-Guidelines for ML-based systems. Focusing on strictly posthoc techniques, they aggregated explainability methods, which resulted in the addition of counterfactual explanations and example-based explanations to the list of explainability methods and techniques. They then built a question bank with questions that users of ML-based systems might have in terms of explainability (Liao et al., 2020). After considering a broader look of users on the ML-based system, they formulated ten question categories, which can be summarized as questions regarding input and output data of the ML-based system, performance of the ML model, “How”-questions concerning how the ML-based system and the ML model generate their predations, and questions concerning why specific values were or were not predicted as well as what could happen by differing parameters of the ML-based system.

As discussed by Liao et al. (2020) it is important to consider the roles of the users. Meske et al. (2022) define five stakeholder groups in XAI systems, which can be divided into three groups based on their interaction with the ML-based system: The first group includes the three stakeholder types of regulators, managers, and developers. It focuses on general regulatory certification, managing and controlling, and the responsibility for the development of the ML-based system. The second group, named users, relates to the actual end users of the ML-based systems. The third group, individuals affected by ML-based decisions, refers to stakeholders who may have no direct interaction with the ML-based system but still experience the consequences of its deployment. Another aspect to consider while designing the UX of an ML-based system utilizing XAI is the quality metrics such an approach should adhere to. Meske et al. (2022) further discuss different personalized quality criteria i.e. metrics that explanations should fulfill. They argue that explanations should generally be interpretable by the user, in a way that users should be able to comprehend them and that they seem plausible. Moreover, an important aspect is the effort required to get the explanations. Zhou et al. (2021) also discuss the clarity, broadness, simplicity, completeness, and soundness of explanations. Oh et al. (2018) evaluated their UX research via general usability metrics, such as the ease of use as well as the ease of learning the use, and more ML-specific metrics, such as the comprehensibility and the controllability of the ML-based system. Lastly, research also already provides several principles and guidelines for the design of the UX or UI in the context of XAI. Laato et al. (2022) recommend always considering visualizations. These types of recommendations were also provided in general UI research, to reduce the cognitive load of users and thereby increase usability satisfaction (Hu et al., 1999). Similarly, the use of coloring in the UI, when well combined with other UI elements, can be helpful (Hu et al., 1999). This

is further emphasized when presenting business information, such as in CPM use cases (Bačić and Fadlalla, 2016). Storytelling and the right use of symbols can also be beneficial (Bačić and Fadlalla, 2016). To summarize, there are metrics for the general perception of the UX or UI, e.g., concerning the visualizations and general usability, as well as metrics specific to the explainability goal and the ML domain that evaluate if the provided explanations are understandable and actionable. Our work builds on these suggestions, specifying and validating them for users in the CPM field.

Methodology

For this study, we followed the design science research process proposed by Peffers et al. (2006), which is appropriate for research on applicable solutions to an existing problem. It includes the six phases problem identification and motivation (1), objectives of a solution (2), design and development (3), demonstration (4), evaluation (5), and communication (6). This process can be iterative and jump backs to any of the phases are possible. Design science in the IS domain can aim to deliver DPs that include prescriptive statements on how to perform activities to solve the problem (Gregor et al., 2020). They can have three different foci: First, they can aim to describe what users should be able to do with the artifact (principles about user activity). Second, they can aim to describe what features should be built into the artifact, such as requirements on a technical or functional level. And third, they combine the definitions of the first two fields, by describing, what users should be able to do with the artifacts as well as the features the artifact should possess (principles about user activity and an artifact) (Gregor et al., 2020). Further, DPs should address and define the actors involved in them, as well as the decomposition of the principle in a hierarchical manner due to the high complexity of IS (Gregor et al., 2020). The object of examination chosen for this study is an existing ML-based CPM system developed by a medium-sized European supplier of enterprise service management software. The system allows for the creation of ML models to forecast business figures based on historical data and data from external providers.

The first phase, problem identification, and motivation consisted of literature research and one qualitative interview. The goals of the literature research were to gain knowledge on ML-based use cases for the CPM domain with a focus on forecasting, the possibilities of XAI, and the existing user-centered research concerning the adoption of XAI in ML-based systems. The results of the literature research are presented in the previous section of this work. To extend the results of the literature in terms of the CPM context, an initial interview with a CPM expert (IIC) was conducted. Following the guidelines for qualitative interviews, the interview allowed for flexibility and was confidential (Myers and Mitchell, 2007). The semi-structured interview aimed to refine the collected knowledge on the different user groups of ML-based CPM systems, as well as on the requirements they have in terms of explainability. The interview was transcribed with a content-driven approach, excluding filler words and correcting obvious grammatical errors. Because assumptions regarding user groups and requirements were made before the interview, a directed content analysis approach was used (Hsieh and Shannon, 2005). The aforementioned assumptions were used as predefined codes in the following coding process as described by Saldana (2009).

The second phase, objectives of a solution, aims to define objectives, which the artifact in the form of the DPs will try to accomplish. To achieve this, the results of the literature research and the results of IIC were analyzed to first define the relevant user groups and their main goal in the ML-based CPM system. Based on that, user group-specific requirements for explainability were derived. These included the knowledge that the users should gain from the explanations and also exemplary actions that users should be able to perform with the help of the provided explanations, with the latter focusing on decision-making actions.

In the third phase, design and development, UX/UI-centric DPs for incorporating XAI into ML-based CPM systems were created. To later demonstrate and evaluate the principles in the next phases, prototypical artifacts in the form of UI mock-ups were designed following the DPs. This phase can thus be split into two parts: the development of the DPs and the design of mock-ups. First, derived from the main goals and objectives, one DP was created for each user group. This principle addressed the general need to enable the user group to achieve this goal. Focusing on three user groups, this produced three DPs (DP 1 – DP 3). Based on the research done in phase one, explainability requirements were derived and assigned to the appropriate DP. This resulted in six explainability requirements (XR 1 – XR 6). Then, based on explainability techniques, ten explainability elements were defined (XE 1 – XE 10), which were mapped according to their possibility of satisfying the explainability requirements. The definition of the XE contains the used explainability technique and some further meta-information. This includes the way in which the

XE should provide its explanations and what kind of information should be provided alongside the technique. In a separate step, three general UX/UI DP were formulated (UDP 1 – UDP 3) based on prior research and practical guidelines provided by the company. After the principles and associated definitions were created, representative UI mock-ups were designed. For their creation, we first considered which XR the respective XE is aiming to fulfill and, therefore, which information the CPM system and thus the mock-up must convey, before elaborating on the presentation and visualization of the information. To allow for a fast and flexible design, they were created in Microsoft PowerPoint. For each XE, 2-3 mock-up alternatives were created using applicable UDPs to examine how the XE is perceived by the user groups (e.g., see Figure 2). The mock-ups used representative dummy data and information to illustrate how actual explanations would be provided via the XE. During this phase, we collaborated closely with both the ML development lead as well as the product owner of the examined CPM solution, involving them in multiple feedback loops. Especially the product owner had valuable insights on the requirements and desires of the CPM clients of the company as well, which were incorporated into the DPs and the mock-ups.

The fourth phase, demonstration, and the fifth phase, evaluation, were conducted in the same interview setting, and are thus covered together. First, an interview approach to evaluate the DPs was developed. Based on the research results, generic and specific metrics for an evaluation of the XE were defined. These focused on more general quality (GQ) metrics as well as on XAI-specific quality (XQ) metrics (see section Theoretical Background). The first part, therefore, concentrated on general UI quality requirements with the latter focusing on quality requirements in regard to the provided explanations. The evaluated quality metrics are the following: Easy interpretation, meaning that the process of interpreting the representation is without much effort (GQ 1); Intuitive interpretation, meaning that the process of interpreting the representation is possible without a lot of knowledge required (GQ 2); Easy-to-learn interpretation, meaning that the knowledge to interpret the representation can be easily gained (GQ 3); Satisfactory interpretation, meaning that the process of interpreting the representation leaves the interpreter in a satisfied state regarding his goal with the interpretation (GQ 4); Transparent presentation of the explanations, meaning that the representation provides all the expected information (GQ 5); Complete explanations, meaning that all the relevant explanations for the regarded use case were provided (XQ 1); Plausible explanations, meaning that the occurrence of the explanations makes sense and can be understood (XQ 2); Trust-building explanations, meaning that the explanations help the user understand the ML-approach in such a way that the user gains trust in it (XQ 3). Additionally, as each XE should help the users achieve their main goal by fulfilling the assigned XRs, specific quality metrics are defined for each XE/XR combinations: Knowledge measures (KM) describe whether the XE presents the right knowledge i.e. the right explanations to achieve the XR. Actionability measures (AM) describe, whether based on the provided knowledge, decisions in connection with the XR can be made.

Interviewee ID	**Role in organization**	**User group**
IIC, IC 1	Management in a planning department	Model creator
IC 2	IT-Consulting for CPM solutions	Model creator
IC 3	Sales and presales for CPM solutions	Model creator
IM 1	Middle management	Model user, data consumer
IM 2	Upper management	Model user, data consumer
IM 3	Middle management	Model user, data consumer

Table 1. Overview of Interviewees for Mock-up Evaluation

The interviews were conducted with three CPM experts (IC 1 – IC 3) and three interviewees with management roles (IM 1 – IM 3). Table 1 shows their respective user role and their role in the organization. The CPM experts were asked to evaluate nine XEs and the management group eight XEs, according to the mapping of the XEs to their respective DPs (see Figure 1). Because of the similar XRs for the user group of model users and data consumers (see section Results), the second interview round was used to evaluate both DP and the associated XE to XR couplings. We conducted all interviews according to the following structure: 1) The general context of the interview was established. The interviewer explained the general approach for the interview and the goals of the study and the interviewee provided further background on

their knowledge regarding CPM-systems and ML if necessary. 2) The interviewees were presented the user groups of ML-based CPM-systems, as well as the goals of the user groups, and which explainability requirements were derived from them. 3) The interviewees were guided through a fictional case study for ML-based forecasting of ice cream sales. This included a fictional management role for the interviewee, fictional decisions for which the ML-generated forecasts were needed, and how the ML-based CPM system would be used to generate a ML model as wells as predictions from the model. 4) The interviewees were guided through the evaluation methodology. For this, all the generic quality metrics were presented in the form of hypothetical statements and all the other steps taken in one evaluation iteration were explained. 5) The interviewees were iteratively asked to evaluate each of the XE represented by the mock-ups. In a first step, two or three slightly different mock-ups for the respective XE were presented to the interviewee in random order. After an alternative was shown, the interviewee was asked to signal when the next can be presented. If needed, the interviewer provided explanations on what was shown in the representation. The alternatives served to evaluate some of UI-focused DPs, as the mock-up alternatives generally displayed the same information but in a different way or on different aggregation levels. The interviewee noted the mock-up alternative which they liked best. Furthermore, the alternatives were used to convey that there is not one unique solution to use the XE. This way, they could evaluate the XE independent from a case where they did not like a particular illustration even though the provided information was fitting. Next, the interviewee rated the respective XE according to the generic and specific quality metrics. For this, we used a balanced Likert scale with five possible expressions (1) Strongly disagree; 2) Rather disagree; 3) Partially agree; 4) Rather agree and 5) Strongly agree) and instructed our interviewees to rate the XE from 1 to 5, with 3 being the midpoint (Likert, 1932). Depending on the results, the interviewer asked follow-up clarifying questions. All other qualitative comments by the interviewee were also recorded and transcribed. The evaluation was conducted by analyzing the quantitative assessment of the mock-ups via the Likert items and the qualitative comments and feedback provided by the interviewees. The refined results of this study are presented in the following chapter.

The communication of the obtained results through this study concludes the design science research process to develop a user-centric approach to XAI in CPM systems.

Results

As literature has shown, XAI-based approaches to provide explainability should keep the users, their motivation, and the context in mind. This study, therefore, builds on the identified three user groups creators, end-users, and affected individuals of the ML-based systems or ML models (Meske et al., 2022). In order to account for the context of a forecasting process in a CPM system, we refine them as follows: By focusing the creator group onto ML models instead of the whole ML-based system, our study defines the user group of model creators. They are in charge of creating, validating, and administrating ML models for different use cases. The model can be a standard statistical model or an ML model. Exemplary company roles are administrators or data engineers/scientists in the planning departments. Secondly, this work splits the end-users group into model users and data consumers. Model users utilize the created and deployed model to generate forecast data. They validate the generated data, may make some adjustments, and report them in the CPM system. Exemplary company roles are planners or business analysts. By using the model they create data, which in turn is used by the data consumers. Data consumers use the reported forecast data to evaluate the company's plans and performance and make decisions based on them. They may also choose to relay information to non-user individuals affected by the ML-generated decisions. Exemplary company roles are managers or other decision-makers. As non-user individuals affected by the decisions of the ML model in practice typically do not gain access to the CPM systems or its explanations, we could not derive meaningful DPs for this group and thus omitted it from this study.

Design Principles and Derived Explainability Elements

The DPs identified in this study are structured in three decomposition layers. Incorporating the main goals of the user groups, we first create three overarching DPs, one for every user group for which the ML system is designed. These are shown in Figure 1 and describe which user group is addressed and what the explanations should be able to accomplish to help the users achieve their goals through the ML-based CPM system. Next, as discussed in the section Theoretical Background, research has identified different requirements that users can have based on the questions they ask in the context of ML-based systems (e.g., Liao et al., 2020). Among them are questions relating to the input and output data of the ML-based system,

the performance of ML models, how the ML models work, or how certain predictions are generated. Taking them, the CPM use case, and the IIC results into account, this study further identifies six explainability requirements (XR) that ML-based systems must meet in order to satisfy the questions posed by the three user groups and to thus help them reach their goals. This resulted in the following XRs, which we mapped to the overarching DPs according to which user group's demands they reflect, as shown in Figure 1:

XR 1: Provide explanations to help understand and evaluate the input data. Users should know what kind of data was used, how big the data sample is, and how good the quality of the data is. While this is important for any kind of data analytics, it is crucial for ML models due to their lack of transparency.

XR 2: Provide explanations to help understand and evaluate different ML models. Users should be able to see and understand the used ML algorithms, as well as the input and output data.

XR 3: Provide explanations to help assess the influence of internal and external drivers. Drivers denote features that contain information on driving factors crucial for the decision-making process and can stem from internal or external sources. Statistically, this distinction is irrelevant, but from a business perspective the drivers should be separated, as companies usually are only able to adjust internal drivers (such as running marketing campaigns as opposed to the external temperature).

XR 4: Provide explanations to help compare different models. This can include the comparison between different ML algorithms or other models, such as simple average functions.

XR 5: Provide explanations to help assess the applicability of the ML models to different use cases. Use cases can be the same type of forecasting done with other data (such as a different region for a sales forecast) or other figures that could benefit from a similar approach.

XR 6: Provide explanations to help understand and evaluate the predictions generated by the ML model. For this purpose, it should be explained which influencing factors were considered and how the prediction was generated.

Based on the research on explainability techniques and refinement through our interviews, we furthermore formulate XEs, which can be described as representations and visualizations of generated explanations or other information in the UI. These are based on XAI techniques, but can also provide non-XAI- or even ML-related information. The provided mapping of the XEs to the XRs completes the third decomposition layer of our DPs. The XE are described in the following together with the measures utilized to evaluate how well the XE provides explanations to satisfy their respective XRs:

XE 1: Provide insights on the size of the data sample i.e. the data amount. This includes information on the number of used time series of the actual figure (e.g. previous sales numbers) and of the internal and external drivers. Further information can be provided on the timeframe of the time series and various metrics concerning the amount of data points (e.g. average amount of data points per timer series). To fulfill XR 1, the XE should help in evaluating the data amount (KM 1.1) and enable decision-making concerning the data amount, such as deciding whether the amount is enough or more data needs to be acquired (AM 1.1). Only for DP 2 and DP 3, to fulfill XR 1, the XE should further help in evaluating the time frame (KM 1.2).

XE 2: Provide insights on the quality of the data. In this case, the quality can be evaluated by the amount of missing data points or anomalies per time series. To fulfill XR 1, the XE should help in evaluating the quality of the used data (KM 2.1) and enable decision-making concerning the data quality, such as deciding to improve the data quality further, exclude certain data series or data points, and many more possibilities (AM 2.1).

XE 3: Provide insights on the general context of how the ML model will be applied. High-level information concerning the input data, output data, and transformation process in the form of an ML algorithm is provided. General information on the ML algorithm can be provided (e.g. the advantages and disadvantages of the used algorithm and a short summary of its functionality). In order to satisfy XR 2, the provided information should help in evaluating the correctness of the context (KM 3.2) and in making decisions concerning the correctness of the context (e.g. changing some of the context or keeping it as is) (AM 3.2).

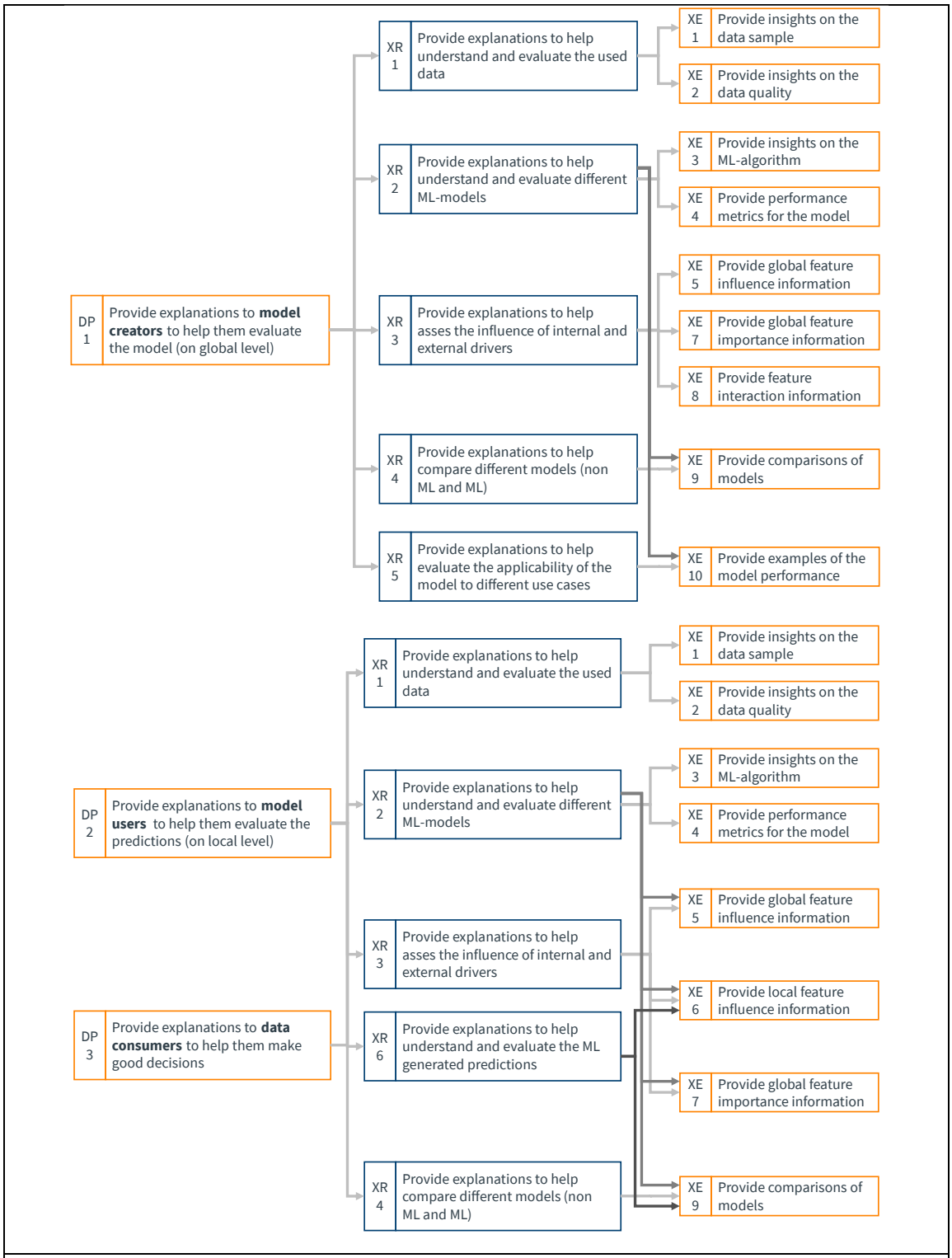

Figure 1. Mapping of XE to XR and DP

XE 4: Provide performance metrics for the model. These metrics can vary depending on the use case, but should generally give an indication of how well the model performs i.e. how well the model predictions match the actual data of testing data set. To satisfy XR 2, the metrics should help in evaluating the model quality (KM 4.2) and in deciding whether the model is usable or needs further refining/should be discarded completely (AM 4.2).

XE 5: Provide information on the influence of all internal and external drivers (features) used in the model on the predictions on a global level. This means that a selected feature's influence on all the results of the model is shown, depending on the expression of this feature. To satisfy XR 3, the presented driver influence should help in evaluating the influence of the drivers in the model (KM 5.3.1) and the driver-dependent biases (KM 5.3.2). Based on the explanations, decisions regarding the correctness of the feature influence can be made (e.g. use the driver, modify the feature selection, or discard the driver completely) (AM 5.3.1). Also, decisions can be made regarding the improvement of the features, such as paying external providers for better data on external drivers (AM 5.3.2). Only for DP 2 and DP 3, to fulfill XR 3, the XE should further help in evaluating the causal relationships of the drivers (KM 5.3.3) and enable decision-making regarding the usability of the model (AM 5.2).

XE 6: Provide information on the influence of all internal and external drivers (features) used in the model on a selected prediction (local) generated by the model. This can be accomplished via different means, but generally, it should help evaluate why a certain value was predicted based on how the features are expressed. Figure 2 depicts the alternatives shown to interviewees, which compare the predicted values with the actual values in a chart. In the lower half, it shows bar charts on how different drivers influenced the selected data point, as indicated by the shaded column in the above chart. In order to satisfy XR 6, the explanations should help evaluate the local influence of drivers on the selected value (KM 6.6) and enable decision-making regarding the use of the forecast data (AM 6.6). Further, to satisfy XR 2, decisions should be facilitated regarding the usability of the ML model (AM 6.2). To satisfy XR 4 decisions should be able to be made in regard to the correctness of the importance of the drivers (such as using or not using the drivers) (AM 6.3.1) and the improvement of the features, as described for XE 5 (AM 6.3.2).

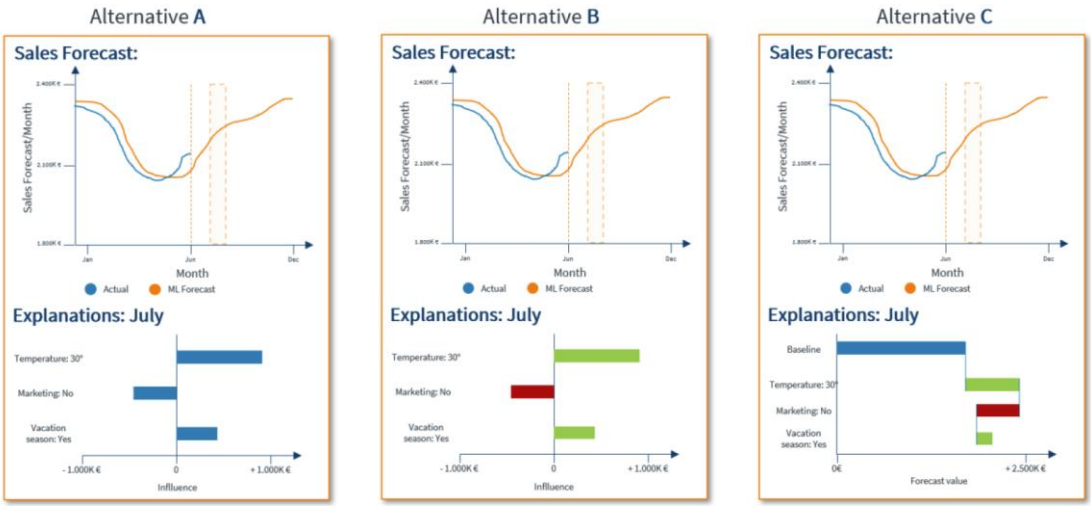

Figure 2. Shown Mock-ups for XE 6

XE 7: Provide information on the importance of the selected internal and external drivers (features) for the model's performance. High importance means that the model would perform worse if this feature would not have been considered in the model. Figure 3 depicts the alternatives for XE 7, which show the importance of different drivers in a bar chart, with the bar width indicating the loss of a performance metric if the driver were not to be included in the model. To satisfy XR 3, the explanations should help evaluate the importance of drivers to the model (KM 7.3). Decision-making concerning the correctness of the

importance of the drivers (such as using or not using the drivers) (AM 7.3.1) and the improvement of the features, as described for XE 5 (AM 7.3.2) should be possible. Only for DP 1 and DP 2, to satisfy XR 2, decision-making should be enabled in regard to the usability of the ML model (AM 7.2).

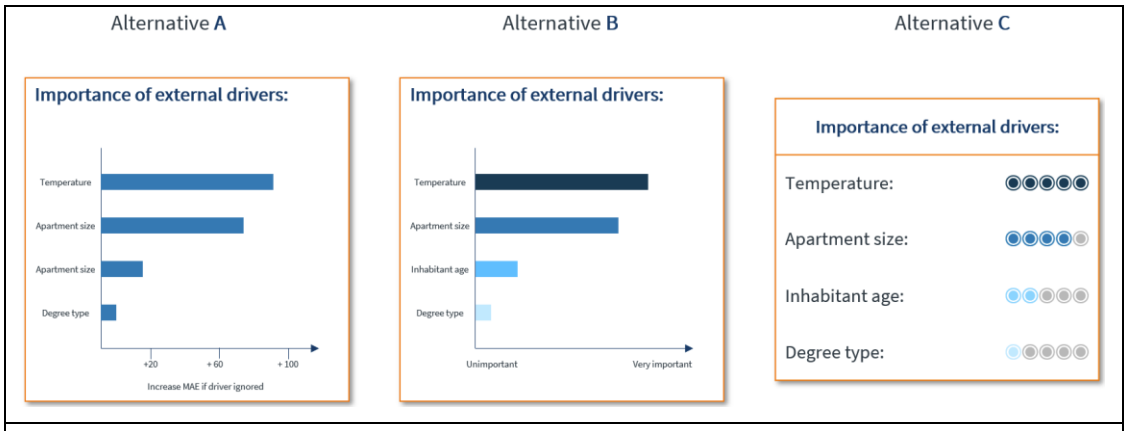

Figure 3. Shown Mock-ups for XE 7

XE 8: Provide information for the interaction of the selected internal and external drivers (features). A high overall interaction suggests that the features correlate in such a way that they can describe patterns that would not be recognized if the drivers were used standalone. High interactions can therefore indicate the right combination of the features so that the model makes up "more than the sum of its parts". Figure 4 displays the alternatives for XE 8, with the overall interaction of different drivers shown in the upper half as well as the interaction of a selected driver with the others in the lower half. The interaction is visualized via different variations of bar charts, with the bar width indicating the degree of interaction. To fulfill XR 3, the explanations should help in evaluating the relationship between different drivers (KM 8.3), facilitate decisions regarding the correctness of the drivers (AM 8.3.1) as described for XE 7, and the improvement of drivers (AM 8.3.2) as described for XE 5 and XE 7.

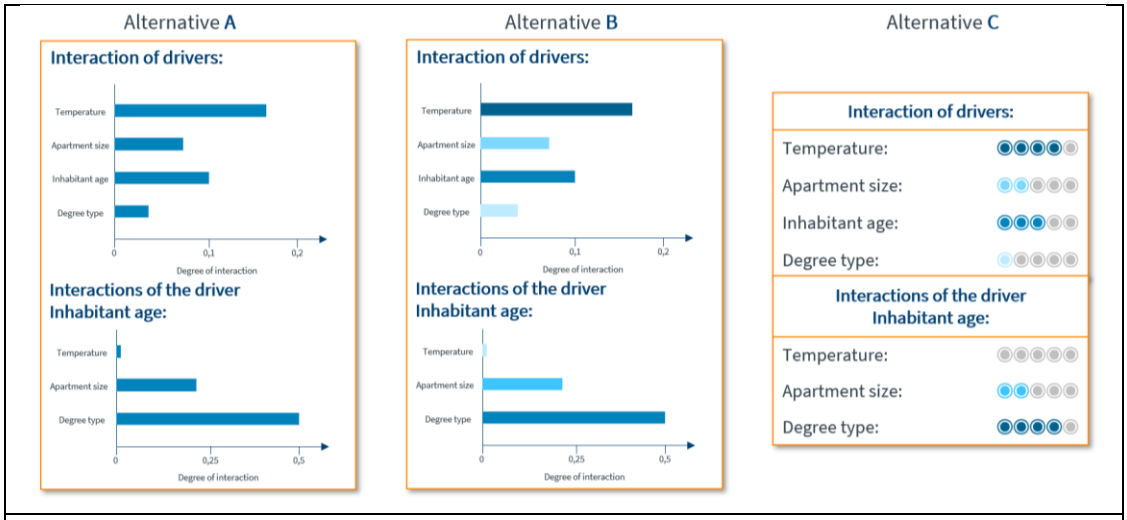

Figure 4. Shown Mock-ups for XE 8

XE 9: Provide comparisons of different models. ML models could be compared with other models or other means of achieving a forecast, such as a simple moving average, which averages the actual data in a moving time window. Also, the actual data can be included to better assess the accuracy of the models. For a better comparison, the focus is on visualizing the forecasts of different models. To satisfy XR 2, the explanations should help evaluate the model quality (KM 9.2) and facilitate decisions regarding the usability of the model (AM 9.2) as described for XE4. To satisfy XR 4, the provided explanations should help compare different models in terms of quality (KM 9.4) and choose the appropriate model (AM 9.4). Only for DP 1 and DP 2, to satisfy XR 6, decision-making should be enabled in regard to the usability of forecast data (AM 9.6).

XE 10: Provide examples of the model's predictions based on different performance classifications. Classifications can be done based on the performance metrics and could show the best, medium, and worst time series in terms of performance. A time series in this case means a sub-dataset of the overall use case where the model is applied, e.g. different regions of a sales forecast. To satisfy XR 2, the explanations should help evaluate the model quality (KM 10.2) and make decisions regarding the usability of the model (AM 10.2) as described for XE 4 and XE 9. To achieve XR 5, the explanations should help evaluate the applicability of the model in the intended area (KM 10.5) and enable decision-making regarding the selection of areas where the model will be applied (AM 10.5).

The following general UX/UI-DPs were derived to analyze different guidelines based on the user needs:

UDP 1: Use visualizations where necessary and possible. Especially if data can be illustrated in charts of various forms, it should be.

UDP 2: Utilize color coding to better differentiate numbers or graphics. A good example is the indication of good or bad performance metrics based on an intuitive "street light" classification (Red for bad, yellow for medium, green for good). Other options include intensifying the display color of bars in bar charts.

UDP 3: Abstract and aggregate information if necessary. Examples are the aggregation of big amounts of data into a standardized value or simply displaying continuous values in a normalized five-point illustration.

These were regarded as guidelines subordinate to the importance of the information provided by the XE, so they were only applied when they seemed fitting. The main focus of these principles is to provide good usability. Based on the evaluation of prior research (see section Theoretical Background), data can be better understood when visualized. Further colors have the potential to make decision makings easier. Lastly, deriving inspiration from lots of web services or e-commerce tools used by lay users, this study proposes the use of aggregated five-point scales, similar to those used in rankings from e.g. products on shopping sites. One point on the scale indicates a poor expression of the considered value, and five points indicate the opposite.

Evaluation

Each of the XE was discussed and evaluated with the experts using mock-ups of slightly different implementations of the XE as the basis for the discussion (e.g., with and without color-coding of the displayed metrics). For instance, the results of the evaluation of XE 7 (which aims to provide insights concerning the importance of the drivers for the model) are shown in Table 2, whereby the results from the CPM-expert group are in blue colors and the results from the Management group are red. IC 1, IM 2, and IM 3 chose the option with bar charts and no color coding, IC 2 with bar charts and color-coding, and IC 3 as well as IM 1 with five-point scales and color-coding. When asked to evaluate XE 7, IC 2 noted that the measurement of MAE increase when withholding the driver in the model as an indicator for the driver's importance was interesting. All of the general metrics were evaluated with at least "rather agree", with the exception of IC 1, who only partially agreed with GQ 4. The explainability-specific metrics all had a majority of at least "rather agree" and in the case of XQ 2 and XQ 3 a majority of "strongly agree". IC 1 and IM 1 only partially agreed with the completeness of the explanations. All interviewees strongly agreed that the explanations help in evaluating the drivers. A majority strongly agreed that decisions based on these explanations could be made in terms of selecting the correct and important drivers. In terms of deciding on driver improvement actions a majority rather agreed. For the decisions concerning the model usability most of the Management-interviewees only partially agreed. IC 2 suggested a different color coding and was missing the data source for plausibility. IM 3 wanted a tooltip that explains what the MAE is. IM 1 mentioned the feature interaction effects that were shown in XE 8 but not to the Management-group.

General feedback to the XE 7 mock-ups included that the color coding was helpful, but red, yellow, and green coloring could be a problem in terms of accessibility, as color-blind people would not be able to interpret them (IC 3). IC 2, as mentioned in most of the XEs individual evaluations, was missing the data sources for more plausibility. IC 2 also stated that because of their statistics and mathematics background, their evaluations could be a little "nitpicky". IC 2 suggested providing more explanations for the terms and metrics displayed in the XEs, as people with no statistics or data science background may not be able to understand them. IM 1 generally found all the information helpful but suggested layering it for example in the form of tooltips, so to allow the users to get the detailed information only if they need it. They stated that they mainly want to see an indication if something is good or bad at first glance. IM 2 also liked the XEs and emphasized the potential of combining them. IM 3 was positively surprised by the possibilities of what XAI technologies can accomplish. They also stated that it was difficult to evaluate the data amount and quality XEs because they are hard to understand for lay users. When they saw the following XEs they could also retroactively better understand the data-centric XEs. In order to answer the research questions, the general and XE-specific feedback and the overall evaluation results are discussed in the following.

Interviewees and chosen alternatives:		**IC 1:** A **IC 2:** B **IC 3:** C		**IM 1:** C **IM 2:** A **IM 3:** A		
Statement ID	Statement summary	Strongly disagree	Rather disagree	Partially agree	Rather agree	Strongly agree
GQ 1	Easy interpretation				●●●●	●●
GQ 2	Intuitive interpretation				●●	●●●●
GQ 3	Easy to learn interpretation				●	●●●●●
GQ 4	Satisfactory interpretation			●		●●●●●
GQ 5	Transparent presentation				●●●●	●●
XQ 1	Complete explanations			●●	●●●●	
XQ 2	Plausible explanations				●●	●●●●
XQ 3	Trust-building explanations				●●	●●●●
KM 7.3	Evaluate driver importance					●●●●●●
AM 7.3.1	Decisions driver importance				●●	●●●●
AM 7.3.2	Decisions driver improvement		●	●	●●	●●
AM 7.2.	Decisions model usability (IM)			●●		●

Table 2. Evaluation Results From the Interviews for XE 7

The general and explainability quality metrics were asked for every XE together with the XE-specific knowledge and actionability metrics and were used to validate and refine both every XE on its own (as described in detail for XE 7 above) as well as the overarching UDP. In the following, we first qualitatively discuss the overall evaluation results and then elaborate on further insights that were not directly integrated into the XE, XR, and UDP detailed above. Regarding the evaluation of the explainability elements, it is important to note that because of the small sample size (n=3 in each interview group), the results should only be interpreted on an indicative level. Nonetheless, the XEs were generally well perceived in terms of helping to fulfill the explainability requirement that they were mapped to. In the first interview group, representing the model creators, in at least 85% of the cases the knowledge and actionability metrics, which aimed to evaluate, whether the XE would help satisfy the XR, were evaluated as at least "rather agree". In

over 50% they were even evaluated with "strongly agree". Even when XEs were not as well evaluated concerning the generic metrics, the interviewees still felt that they provide the right knowledge and help make decisions. The generic metrics were also rated with at least "rather agree" in over 80% of the cases and "strongly agree" in around 50% of the cases. The second interview group, representing the model users and data consumers, evaluated the XE with similar results. The knowledge metrics were evaluated with at least "rather agree" in over 90% of the cases (over 60% strongly agreed), and the actionability metrics in over 80% of the cases (over 40% strongly agreed). The general metrics also showed good evaluation results of over 85% "rather agree" (over 30% strongly agreed).

In the case of low-rated XEs, interviewees often felt that some information to understand the ML-related metrics or terms was missing, complicating the use by lay users (e.g., IM 1 and 2 regarding XE 1 or IM 1, 2, and 3 in regards to XE 4). In some cases, the visualization of the charts was not optimal (e.g., IC 2 and IM 1 in regards to XE 5 or IC 1 regarding XE 8). One interviewee (IC 3) also emphasized the importance of providing data sources for all calculated metrics to make the explanations plausible. This was also proposed in research by Laato et al. (2022). Because this was not mentioned by the other interviewees, it should be implemented with caution, so as not to overload lay users with information. The possibility to "drill down" as proposed by Laato et al. (2022) to get the requested information on demand should be evaluated. In terms of the chosen mock-up alternatives the conclusion is not as clear: For XE 5, regarding the global driver influence, all interviewees chose the same option, and for XE 9, regarding the model comparisons, 5 out of 6. The rest only had majorities in their respective interviewee groups in 7 out of 9 cases (IC) or 7 out of 8 cases (IM). This further supports the assumption that the design of the XE should be user-group specific. Based on the qualitative comments of the interviewees during the selection and afterward, it should be evaluated how the best features of the different options could be combined. Again, the often-suggested option of tooltips or "drill downs" should be evaluated, to accommodate different levels of details in one UI. Alternatively, things like color-coding which were better received by the Management interviewees could be made configurable in the ML-based system, depending on the user utilizing them. The UX/UI-DPs should therefore be evaluated further, in a context where they could be applied depending on the user's personal preference. In conclusion, the conceptualized DPs provide the right XEs to the identified XRs, as indicated by the generally favorable evaluation results of the KMs and AMs. While the evaluation results for the general and explainability-specific UX metrics were almost equally as well-rated, there is still some improvement potential for the UI design. More interactable explanations, such as tooltips for ML-related metrics and terms, could offer much potential, as well as the combination of the different XEs at the appropriate steps in the user journey.

Discussion

In order to contribute relevant knowledge to IS-research, design science studies should fall into one of three categories (Gregor and Hevner, 2013): *Exaptation* means that they extend existing knowledge to new problems. *Improvement* describes studies that develop new solutions to existing problems. *Invention* studies develop novel solutions for new problems. The user-centric approach developed in this work combines, refines, and focuses prior research to develop a solution to provide explainability in ML-based CPM systems. Therefore, it provides a new solution to a known problem, thus classifying it as an *improvement* work.

This study makes two major **theoretical contributions**. First, this work refines prior stakeholder and user groups identified by, e.g., Langer et al. (2022) and Meske et al. (2022) to a more ML model-focused level. In particular, it specializes the user groups for the CPM context, connects them with their respective XR, and thus invites future research in the CPM field to evaluate XAI approaches using our framework for CPM users. Within our framework, we follow Brennen (2020) and take the lay user groups model users and data consumers into account. Additionally, we encourage further studies to refine user groups of XAI systems for other contexts to lay the groundwork for research on the user-centric design of respective XAI approaches. Second, we provide DPs representing UX/UI building blocks that describe what kind of information should be displayed to the user, which can be used by scholars to inform future user-centric XAI research. Our DPs relate directly to our identified user groups following the advice of Liao et al. (2020) to consider user requirements to derive DPs. Through this study, we thus validate existing research on the UX/UI design of XAI approaches (e.g., Zhou et al., 2021; Oh et al., 2018; Laato et al., 2022) for the CPM field and refine CPM-specific design by centering our DPs around the CPM users and their requirements in particular. The positive evaluation by our lay user groups indicates that our XEs are suitable for providing

explainability to these users in particular, following the suggestion of Brennen (2020). Moreover, our derived XR can be matched to the XAI questions in the UX-centered research of Liao et al. (2020) with the exception of XR 6, which may therefore be specific to the CPM domain. However, our DPs contribute to XAI research not only in the CPM context. A plethora of similar forecasting use cases exist in other domains such as sales forecasting for supply chain management (e.g., Bi et al., 2022) or wind speed forecasting for power grid balancing (e.g., Yang and Chen, 2019) with model users similar to planners in the CPM context for instance. As our DPs are tailored to provide explainability specifically to CPM users, we argue that they can apply to any domain with use cases sufficiently similar to the forecasting of business figures and user groups whose characteristics and goals align with those of the CPM user groups (e.g. data consumers that utilize the model predictions to support their decision-making processes, for instance for planning or performance monitoring purposes). In this regard, we invite future research to adopt our evaluation framework to validate and extend our DPs for contexts other than CPM.

Additionally, we make two primary **practical contributions** with our study. First, organizations can utilize our developed evaluation framework to assess their ML-based CPM systems currently in use for their ability to provide explainability to their users. We invite practitioners to rethink whether their CPM systems are truly understood by users, especially lay users, and offer considerations on the requirements and evaluation criteria to be fulfilled to enhance user understanding. Second, our DPs can be used as concrete guidelines to foster explainability in practice when developing ML-based CPM systems by incorporating an XAI approach. Getting more explainability into ML-based solutions may help organizations build trust in the technology in CPM, as they provide knowledge that without them would be lost in the "black box". By utilizing our user-centric approach, practitioners may therefore address the perceived risk of low explainability (see McKinsey, 2021) and thus foster the use and adoption of ML in the CPM domain.

This study is not without limitations, however. First, although the technical possibilities of posthoc techniques and other aspects of ML or XAI were kept in mind, they were not evaluated in detail. Completely interpretable models were omitted from the start, and their potential for specific and small use cases should be determined, as the development of the conceptualized XEs could potentially come with a technical overhead. Secondly, while this study made a first step toward user-centric XAI design in CPM, its scope is too small for empirical validation of the identified XEs. Further quantitative studies could thus build on our evaluation framework to refine and validate the XEs with a larger sample of CPM users from organizations of various sizes and industries. Lastly, the evaluation of the results was done via UI mock-ups for standalone XEs. As mock-ups are static, they cannot truly show the UX with interactive elements. Therefore, interactive explanations were largely omitted and could hold the potential for additional explanatory power to be unlocked by future research.

Conclusion

In this study, we followed the design science research process as presented by Peffers et al. (2006) to create UX/UI designs for the ML-based CPM system developed by a medium-sized supplier of enterprise service management software, which we evaluated and refined through six interviews with CPM and management experts from the firm. As results, we were able to derive DPs, which aim to provide explainability according to the user's goals and requirements in order to facilitate knowledge exchange between users and ML-based CPM systems. To create our DPs, we, therefore, specified the user groups of ML-based CPM systems as model creators, model users, and data consumers and identified their respective goals. The DPs themselves consist of a three-layer decomposition structure. The highest levels (DP 1 – DP 3) describe the goals of the user groups on which the explanations should focus. The second layer derives more specific explainability requirements (XR 1 – XR 6), such as the need to describe data sample size and quality that can be connected to different DPs. Lastly, the explainability elements (XE 1 – XE 10) that deploy techniques from the field of XAI such as feature importance or simply present metadata, such as model performance metrics, to fulfill different XRs constitute the third level. Further, more general UX/UI-DPs are provided (UDP 1 – UDP 3). They suggest the use of visualization, color-coding, and aggregations to point scores. The DPs are validated on an indicative level from interviews, for which we elaborate an evaluation framework. Our framework includes generic metrics concerning the UX/UI design, as well as XE-specific metrics, evaluating whether the right knowledge was provided and the right actions can be taken based on the explanations. The evaluation results show a high agreement of users with the defined quality metrics. Another finding from the interviews suggests that the same XE could be deployed with a user-group-specific design to further enhance usability. Therefore, the DPs developed in our study offer first concrete guidelines for designing

XAI approaches in CPM to practitioners while providing scholars with both a CPM-specific evaluation framework and user-specific DPs for future XAI approaches to be refined and expanded on through research in the CPM domain.

Acknowledgement

This research and development project is/was funded by the German Federal Ministry of Education and Research (BMBF) within the "Innovations for Tomorrow's Production, Services, and Work" Program (funding number 02L19C150) and implemented by the Project Management Agency Karlsruhe (PTKA). The authors are responsible for the content of this publication.

References

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

Adya, M. P., and Collopy, F. L. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting, 17*, 481-495.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A. et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115.

Bačić, D., and Fadlalla, A. (2016). Business information visualization intellectual contributions: An integrative framework of visualization capabilities and dimensions of visual intelligence. *Decision Support Systems, 89*, 77-86.

Bi, X., Adomavicius, G., Li, W., and Qu, A. (2022). Improving sales forecasting accuracy: A tensor factorization approach with demand awareness. *INFORMS Journal on Computing*, *34*(3), 1644-1660.

Bontempi, G., Ben Taieb, S., and Le Borgne, Y. A. (2013). Machine learning strategies for time series forecasting. In *Lecture Notes in Business Information Processing* (Vol. 138, pp. 62-77). Springer.

Brennen, A. (2020). What do people really want when they say they want "explainable AI?" we asked 60 stakeholders. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*, 1-7.

Brynjolfsson, E. and Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science 358*(6370), 1530-1534.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature, 538*, 20-23.

Frolick, M. N., and Ariyachandra, T. R. (2006). Business performance management: One truth. *Information Systems Management, 23*(1), 41-48.

Ghoddusi, H., Creamer, G. G., and Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. *Energy Economics*, *81*, 709-727.

Gregor, S., Chandra Kruse, L., and Seidel, S. (2020). Research perspectives: The anatomy of a design principle. *JAIS*, *21*(6), 1622-1652.

Gregor, S., and Hevner, R. A. (2013). Positioning and presenting design science. *MISQ, 37*(2), 337-355.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Comput. Surv., 51*(5), 1-42.

Gunning, D., Vorm, E., Wang, J. Y., and Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters, 2*(4), pp 1-11.

Hewamalage, H., Bergmeir, C., and Bandara, K. (2021). Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting, 37*(1), 388-427.

Hsieh, H.-F., and Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277-1288.

Hu, P. J.-H., Ma, P.-C., and Chau, P. Y. (1999). Evaluation of user interface designs for information retrieval systems: A computer-based experiment. *Decision Support Systems*, *27*(1), 125-143.

Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Laato, S., Tiainen, M., Najmul Islam, A.K.M., and Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research*, *32*(7), 1-31.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E. et al. (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.

Liao, Q. V., Gruen, D., and Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1-15.

Likert, R (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5-55.

Ma, S., and Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, *288*(1), 111-128.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., Semenoglou, A.-A., Mulder, G., and Nikolopoulos, K. (2023). Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward. *Journal of the Operational Research Society*, 74(3), 840-859.

McKinsey. (2021). *The state of AI in 2021*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-thestate-of-ai-in-2021 (visited on October 11, 2022).

Meske, C., Bunde, E., Schneider, J., and Gersch, M. (2022). Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. *Information Systems Management*, *39*(1), 53-63.

Myers, M. D., and Mitchell, T. M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, *17*(1), 2-26.

Miranda, S. (2004). Beyond BI: Benefiting from corporate performance management solutions. *Financial Executive, 2*(20), 58-61.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Interpretable machine learning: Definitions, methods, and applications. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(44), 22071-22080.

Oh, C., Song, J., Choi, J., Kim, S., Lee, S., and Suh, B. (2018). I lead, you help but only with enough details. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* 1-13.

Pavlyshenko, B. (2019). Machine-learning models for sales time series forecasting. *Data*, *4*(1), 1-15.

Peffers, K., Tuunanen, T., Gengler, C., Rossi, M., Hui, W., Virtanen, V., and Bragge, J. (2006). The design science research process: A model for producing and presenting information systems research. *Proceedings of First International Conference on Design Science Research in Information Systems and Technology*, 84-106.

Richards, G., Yeoh, W., Chong, A. Y. L., and Popovič, A. (2019). Business intelligence effectiveness and corporate performance management: An empirical analysis. *Journal of Computer Information Systems*, *59*(2), 188-196.

Rogers, P. R., Miller, A., and Judge, W. Q. (1999). Using information-processing theory to understand planning/performance relationships in the context of strategy. *Strategic Management Journal*, *20*, 567-577.

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Saldana, J. (2009). *The coding manual for qualitative researchers*. SAGE Publications.

Spiliotis, E., Nikolopoulos, K., and Assimakopoulos, V. (2019). Tales from tails: On the empirical distributions of forecasting errors and their implication to risk. *International Journal of Forecasting, 35*(2), 687-698.

Wasserbacher, H., and Spindler, M. (2022). Machine learning for financial forecasting, planning and analysis: Recent developments and pitfalls. *Digit Finance, 4*(1), 63-88.

Yang, H.-F., and Chen, Y.-P. P. (2019). Representation learning with extreme learning machines and empirical mode decomposition for wind speed forecasting methods. *Artificial Intelligence, 277*, 103176.

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