Bus Inf Syst Eng
https://doi.org/10.1007/s12599-026-00986-2

RESEARCH PAPER

# Rigorous Viability Assessment of Machine Learning Projects

## Example from the Domain of Predictive and Condition-Based Maintenance

**Domenique Zipperling · Lorenz Ott · Michael Vössing · Niklas Kühl**

Received: 14 March 2025 / Accepted: 21 November 2025


**Abstract** Machine learning offers significant potential for organizations, yet transitioning models from development to deployment remains challenging. Frameworks such as CRISP-ML(Q) and MLOps emphasize the need to integrate business, economic, and machine learning perspectives. However, a systematic literature review reveals a lack of methods that link machine learning perspectives with business objectives. To address this gap, the authors introduce a metric – called *profit-per-decision (ppd)* – for binary classification that incorporates both model performance and economic impacts. Further, the Viability Assessment Framework is proposed, which utilizes the metric and enables organizations to assess viability at different project stages: pre-development, post-development, and post-deployment. The authors evaluate the framework through expert interviews and a scenario-based evaluation with experts from eleven different companies and develop an open-source web application to support interaction during the case studies. Results confirm the framework's effectiveness in bridging technical and business perspectives, highlighting its industry relevance.

**Keywords** Viability assessment · Cost-sensitive performance estimation · Uncertainty estimation · ML life cycle

Accepted after 4 revisions by Natalia Kliewer

D. Zipperling · N. Kühl
University of Bayreuth, Universitätstraße 30, 95447 Bayreuth, Bavaria, Germany

D. Zipperling · N. Kühl
Fraunhofer FIT, Wittelsbacherring 10, 95444 Bayreuth, Bavaria, Germany

D. Zipperling (✉) · N. Kühl
FIM Research Center for Information Management, Wittelsbacherring 10, 95444 Bayreuth, Bavaria, Germany
e-mail: domenique.zipperling@fit.fraunhofer.de

L. Ott
TU Wien, 1060 Vienna, Austria

M. Vössing
Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe, Germany

## 1 Introduction

Machine Learning (ML) models used for binary classification tasks have the potential to improve processes (Susanto and Khaq 2024), products (Cooper and McCausland 2024) and services (Naeem et al. 2024) across domains. To successfully implement ML applications within organizations, capabilities and readiness factors such as identifying, evaluating, and prioritizing suitable applications have been outlined (Jöhnk et al. 2020; Weber et al. 2023). However, organizations still struggle to identify promising ML projects due to the difficulty of reliably assessing economic benefits (Benbya et al. 2021; Weber et al. 2023; van Giffen and Ludwig 2023). This presents a significant challenge for the Business and Information Systems Engineering (BISE) community.

Although organizations today must balance various objectives, such as economic, ecological, and social ones, a primary focus remains on ML projects that meet economic success criteria. We call such projects "viable". Therefore, organizations must assess economic viability at every stage of an ML project: before development (pre-development), after development (post-development), and during production (post-deployment). A central driver of economic

Published online: 10 February 2026

viability is the ML model's performance, which includes the costs and benefits of correct or incorrect decisions. Hence, two questions must be answered: What requirements concerning model performance (minimal performance) must be met to achieve viability? Is the performance of the developed/deployed model sufficient to meet these requirements? To ensure viability, organizations need to derive ML success criteria (Studer et al. 2021) by connecting the business, economic, and ML perspectives of the project (Weber et al. 2023; Duda et al. 2023). Here, involving domain experts is crucial, as they, rather than data scientists, identify valuable problems (van Giffen and Ludwig 2023). Moreover, their expertise is essential to estimating the costs and benefits of model decisions. In addition to connecting economic and ML perspectives, organizations must manage the "fear of the unknown" (-Merhi 2023, p. 3), which is driven by uncertainty in assessing viability (Kläs and Vollmer 2018). Since ML model performance determines viability, its estimation is the primary source of uncertainty. Therefore, we pose the following research questions (RQ):

RQ 1 How can an economic perspective be combined with the performance assessment of machine learning models?

RQ 2 What method can be employed to ensure the support of project decisions across the entire machine learning life cycle, while accounting for the uncertainty inherent in such decisions?

To begin our research, we examined existing methods for assessing the viability of ML projects by integrating model performance, uncertainty, and economic factors. We conducted a systematic literature review (SLR) following the frameworks of Webster and Watson (2002) and Wolfswinkel et al. (2013). Our analysis revealed a lack of methods that help decision-makers and developers to assess viability holistically. Current literature focuses on one specific phase, seldom considers costs, benefits, and uncertainty simultaneously, and does not integrate the proposed approaches into the ML life cycle.

To close that gap, we propose the Viability Assessment Framework. At the framework's center is a metric called profit-per-decision (*ppd*), which bridges the economic and ML perspectives, thereby addressing RQ1. To address RQ2, we adopt a statistical perspective on confusion matrices and derive the distribution for the metrics estimator, allowing for uncertainty quantification through confidence intervals. Furthermore, we incorporate the metric (*ppd*) into a comprehensive framework to determine viability requirements for a developed/deployed model. The framework connects economic and ML success criteria while factoring in uncertainty. Thus, we answer the previously stated question during pre-development: What requirements concerning model performance (minimal performance) must be met to achieve viability? To assess viability, the framework uses the estimator. It derives a minimum performance bound required to meet the economic success criteria with a given uncertainty, resulting in a viability statement: *The project is viable with a certain probability for a required model performance*. Further, the framework offers guidance to assess the probability with which a developed/deployed model will fulfill the economic success criteria. The result is a viability statement answering the question: Is the performance of the developed/deployed model sufficient to meet these requirements? The derived statement has the form: *The project is viable with a certain probability based on the actual estimated performance*. With these statements, decision-makers can identify promising projects during pre-development, make deployment decisions during post-development, or determine whether to continue using an application in production after its deployment. Especially for the post-development and post-deployment phases, it is essential to note that our framework is to be understood as a complement, not an alternative to the usual performance estimation process (Fig. 1).

To provide practical guidance, we demonstrate the framework's application using a predictive maintenance (PdM) and condition-based maintenance (CbM) use case. This domain is ideal due to the economic impact of model decisions, where false positives are costly and costs/benefits are quantifiable (Prytz et al. 2015; Florian et al. 2021). Further, maintenance datasets often have small sample sizes, especially in early production stages (Prytz et al. 2015; Bukhsh et al. 2019), increasing uncertainty of model performance (Hastie et al. 2009). We also evaluated the framework from a socio-technical perspective using semi-structured expert interviews. Additionally, we conducted a scenario-based evaluation in which experts apply the Viability Assessment Framework to a former use case of a world-leading German automotive manufacturer. In doing so, we enabled an evaluation from a practical perspective. A web application of the framework was implemented to ensure realism and to enable interaction between the framework and the experts. The results demonstrate the framework's necessity, emphasizing its effectiveness in bridging economic and ML perspectives, improving reporting, and supporting decision-making processes. Additionally, potential improvements to the framework are identified, which can serve as a roadmap for future research.

Our work bridges the gap between economic and ML perspectives by proposing a holistic ML project framework for viability assessment – the Viability Assessment Framework. As a result, several contributions are made.

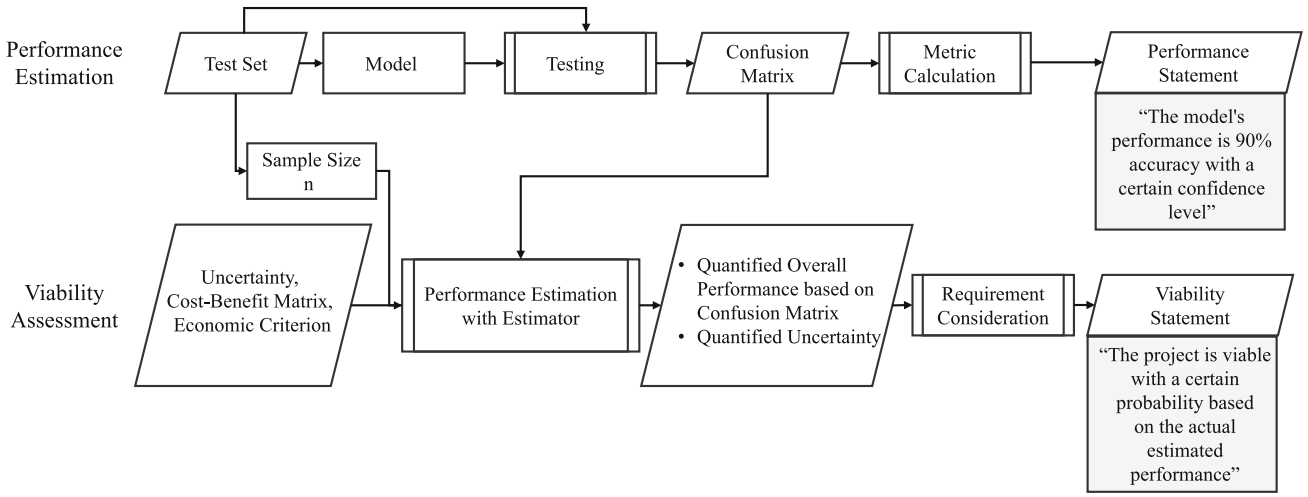


**Fig. 1** Interaction between performance estimation and viability assessment

We derive the distribution of the profit-per-decision estimator, which enables us to quantify uncertainty through confidence intervals. Embedding the estimator within the Viability Assessment Framework allows for the creation of formalized viability statements tailored to project-specific factors. The framework helps define ML requirements for viability before model development (pre-development) and provides a clear decision boundary after the model is developed (post-development) or deployed (post-deployment). Further, by integrating our work into the ML life cycle, we enable easy adoption in practice. In summary, our work offers a method to the BISE community that allows the identification of necessary ML requirements based on previously defined economic criteria while considering uncertainty. We thereby enable the analysis of a project's viability during pre-development, post-development, and post-deployment.

The remainder of the paper is structured as follows: Section 2 covers foundational concepts, and Sect. 3 presents the SLR's results emphasizing the research gap. We introduce the framework in Sect. 4 and apply it exemplarily in Sect. 5. Section 6 details its evaluation through interviews and a scenario-based evaluation involving a total of 16 PdM experts from eleven different companies. Section 7 discusses the findings from both practical and theoretical perspectives, highlighting limitations and future avenues of research. Section 8 concludes our work.

## 2 Foundations

This section conceptualizes economic viability and its assessment from an ML perspective. To this end, we define economic viability in reference to the net present value and connect it to ML projects' typical costs and benefits. Additionally, we outline the role of uncertainty quantification when assessing viability. Further, we introduce the concept of receiver operating characteristic (ROC) curves.

### 2.1 Economic Viability from a Machine Learning Perspective

A project is considered economically viable if it achieves a positive financial outcome, which is typically measured using net present value (NPV) (Lima et al. 2015; Myers and Majluf 1984; Archer and Ghasemzadeh 1999). The NPV is determined by the initial investment and the discounted sum of annual cash flows (*CF*), which can be divided into the annual benefits and costs (fixed or variable) (Gaspars-Wieloch 2017). Both components can have monetary (cost-savings vs. operational expenses) and non-monetary components (customer loyalty vs. reputation damage). Contingent valuation or willingness-to-pay approaches can be leveraged to transfer non-monetary factors to monetary values (Perni et al. 2021). As we consider ML projects in our work, the *CF* depends on the ML model's decisions, which links the economic viability to the model's performance. Therefore, it is sensible to divide the yearly *CF*s into two components: one that is dependent on the model decision ($CF_{model}(t)$) and one that is independent ($CF_{other}(t)$). In line with Gaspars-Wieloch (2017), Equation (1) formalizes the NPV of an ML project while $r$ denotes the discount factor, $T$ denotes the project duration, and $C_I$ denotes the initial investment.

$$\mathrm{NPV} = -C_I + \sum_{t=1}^{T} \frac{CF_{\mathrm{Model}}(t) + CF_{\mathrm{other}}(t)}{(1+r)^t} \tag{1}$$

### 2.2 Consideration of Model-Dependent Benefits and Costs

In the case of classification, the model-dependent *CF*, derived from the benefits and costs of model decisions, is influenced by the rates of correct and incorrect decisions – especially the cost of different error types (false positives (FP) and false negatives (FN)) and the benefits of correct predictions (true positives (TP) and true negatives (TN)).

A key work on integrating costs into ML projects is introduced by Elkan (2001), covering cost-sensitive learning (CSL), which integrates the mentioned costs into training, thereby guiding models to minimize expected costs or addressing imbalanced data sets (Araf et al. 2024). CSL typically uses a cost matrix $C(i,j)$, quantifying the cost of misclassifying class $j$. Other research also considers the benefits of correct decisions, providing cost-benefit matrices to quantify the cost-benefit structure of model decisions (Florian et al. 2021). Although cost/benefit values are not the primary focus of this study, they can be estimated based on domain knowledge (Vargas-Palacios et al. 2023; Jiang et al. 2008) or data (Ziegelmayer et al. 2022). The idea of cost sensitivity can extend to testing. This extension enables calculations of the expected model-dependent *CF* by linking the probability of a decision with its associated costs or benefits (see Equation (2)).

$$CF_{\text{Model}} = \sum_{i,j} P(i|j)C(i,j), \forall i,j \in \{positive, negative\} \tag{2}$$

### 2.3 Consideration of Uncertainty

Uncertainty and its quantification are complex and ongoing challenges in ML projects (Tyralis and Papacharalampous 2024). Various sources of uncertainty arise during the development and deployment, such as data acquisition and processing (e.g., label noise (Huang et al. 2022)), model training (e.g., hyperparameter uncertainty (Bergstra and Bengio 2012)), validation and performance estimation (e.g., statistical (Kläs and Vollmer 2018) or predictive uncertainty (Tyralis and Papacharalampous 2024; Darling and Stracuzzi 2018; Dewolf et al. 2022)), and deployment and maintenance (e.g., concept drift (Baier et al. 2019)). Especially statistical uncertainty is critical as it links the uncertainty in performance estimation to the uncertainty of the viability assessment. Performance estimation is usually based on a test dataset. In the case of binary classification, a confusion matrix is used to calculate metrics like accuracy, recall, or precision. As the test dataset only represents a subset of the population, the calculated metrics are estimations and inherently subject to uncertainty. Therefore, quantifying the statistical uncertainty in performance estimation is essential for deciding whether to proceed with a project. An established method to assess statistical uncertainty is the calculation of confidence intervals, which can be estimated through parametric methods (e.g., based on known or assumed distributions of the estimator) or non-parametric ones (e.g., bootstrapping). Confidence intervals, estimated through parametric methods for their robustness (Rodopoulos and Lemon 2014) and simplicity (Correa and Bellavance 2001), are a key approach to quantifying this uncertainty.

### 2.4 The ROC Curve

The ROC curve (Fawcett 2006) is widely used for evaluating the performance of binary classification models, providing a graphical representation of a classifier s ability to distinguish between two classes across varying decision thresholds. It is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR), which represents the proportion of false alarms. Mathematically, these quantities are defined as:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \tag{3}$$

The shape of the ROC curve provides insight into the trade-off between TPR and FPR at different thresholds. The Area Under the Curve (AUC) is commonly used to translate the ROC curve into a single metric, measuring the classifier's discriminative power. An AUC of 1.0 indicates perfect classification, whereas an AUC of 0.5 corresponds to a random classifier. By analyzing the ROC curve, researchers can compare different classification models, select optimal decision thresholds, and assess the robustness of predictive systems under varying conditions.

## 3 Related Work

To understand the current literature on viability assessment, we conducted an SLR based on Webster and Watson (2002) while adopting the five phases proposed by Wolfswinkel et al. (2013): define, search, select, analyze, and present. Our review focused on the intersection of uncertainty, benefits, costs, and finding a minimum/sufficient model performance as illustrated in Fig. 2. A detailed methodological description is provided in the Online Appendix A (available online via http://link.springer.com). We analyzed whether the identified literature provides answers to the following dimensions and related questions:

1. Pre-development viability: What minimal model performance must be met to achieve viability?


**Fig. 2** Our research is positioned at the intersection of the cost and benefits of the machine learning model decision, the uncertainty of estimating a model's performance, and of finding a minimum/sufficient performance regarding an economic success criterion

2. Post-development/deployment viability: Is the performance of the developed/deployed model sufficient to meet these requirements?
3. Life cycle integration: How are existing methods integrated into the machine learning life cycle?

Based on these questions, we derived three key dimensions that guide our analysis. The first dimension, pre-development viability, assesses the hypothetically minimal necessary model performance combined with a cost-benefit analysis to meet economic success criteria while accounting for uncertainty. The second dimension, post-development and post-deployment viability, evaluates whether the actual model performance, alongside the cost-benefit analysis, is sufficient to achieve economic success under uncertainty. Lastly, the third dimension, life cycle integration, focuses on embedding the viability assessment within the ML life cycle.

### 3.1 Overview of Literature and Assessment of the Research Gap

We reviewed existing literature on ML viability assessment methods and evaluated to what degree they satisfy the different dimensions and subdimensions. Therefore, methods were classified based on their applicability to pre-development or post-development/post-deployment viability assessment. We examined how they handle cost-benefit analysis, determine minimal or sufficient performance, consider uncertainty, and foster an integration into the ML life cycle. Each dimension was examined, and the results are presented in Table 1.

**Costs and benefits** Costs and benefits are considered to different degrees in existing literature. A significant number of the identified articles focus only on parts of the cost-benefit matrix. Most focus solely on the costs of FP and FN (Holte and Drummond 2008; Drummond and Holte 2006, 2000; Hernández-Orallo et al. 2013; Zhou and Liu 2012), whereas one article accounts for the benefits of TP and costs of FP (Hawkins 2020). Therefore, they partially fulfill the subdimension of cost and benefit. In contrast, DataRobot (2024) covers the full cost-benefit matrix, fulfilling the subdimension to three-quarters. Florian et al. (2021) additionally include investment and maintenance costs, thereby fulfilling the subdimension fully. All remaining literature ignores cost-benefit aspects (Renggli et al. 2019, 2023). As the consideration of costs and benefits does not change across phases, the evaluation applies to pre- and post-development as well as post-deployment alike.

**Minimal and sufficient performance** Regarding minimal performance determination during pre-development, the literature takes contrasting views, treating it as either an endogenous variable (e.g., Hawkins (2020)) or an exogenous one (e.g., Renggli et al. (2023)).

Hawkins (2020) and Renggli et al. (2023) are the only articles actively adressing the pre-development phase. Hawkins (2020) treats minimal performance as an endogenous variable and proposes a method to derive a hypothetical performance threshold. Synthetic ROC curves are simulated, and corresponding AUC values are calculated to assess whether a model can meet the viability criterion. The assessment is based on predefined project parameters, including sample size, baseline rate (positive class ratio), minimum return on investment (ROI), FP costs, and TP benefits. Thereby, AUC values are linked to the viability decision. Further, a ranking scheme identifies the minimum AUC required to meet the ROI threshold. This is the only study offering a method that fully satisfies the minimal performance subdimension. Regarding determining the sufficiency of the model performance during post-development/post-deployment, the project characteristics can be used to assess viability, thus also fulfilling this subdimension.

Renggli et al. (2023) adopt a data-centric approach. The method estimates the Bayes error rate for a given dataset and compares it to a predefined target performance, "performing a systematic and theoretically founded feasibility study before building ML applications" (Renggli et al. 2023, p. 218). The minimal performance subdimension is only partially fulfilled because it requires a predefined target performance rather than deriving one. Moreover,

**Table 1** The table shows the analysis results of the identified literature regarding viability assessment and integration into the ML life cycle

| Source | Pre-Development Viability | | | | Post-Development/ Post-Deployment Viability | | | | Integration into ML Life Cycle |
|---|---|---|---|---|---|---|---|---|---|
| | Cost and Benefit | Minimal Performance | Performance Uncertainty | Final Score | Cost and Benefit | Sufficient Performance | Performance Uncertainty | Final Score | Active Integration |
| Renggli et al. (2023) | ○ | ◑ | ○ | ◔ | ○ | ○ | ○ | ○ | ○ |
| Renggli et al. (2019) | ○ | ○ | ○ | ○ | ○ | ● | ● | ◕ | ○ |
| Hawkins (2020) | ◑ | ● | ◑ | ◕ | ◑ | ● | ○ | ◑ | ○ |
| Florian et al. (2021) | ● | ◑ | ○ | ◑ | ● | ● | ○ | ◕ | ○ |
| DataRobot (2024) | ◕ | ◑ | ○ | ◑ | ◕ | ● | ○ | ◕ | ○ |
| Drummond and Holte (2000) | ◑ | ◑ | ○ | ◑ | ◑ | ● | ○ | ◑ | ○ |
| Drummond and Holte (2006) | ◑ | ◑ | ● | ◕ | ◑ | ● | ● | ◕ | ○ |
| Holte and Drummond (2008) | ◑ | ◑ | ● | ◕ | ◑ | ● | ● | ◕ | ○ |
| Zhou and Liu (2012) | ◑ | ◑ | ○ | ◑ | ◑ | ● | ○ | ◑ | ○ |
| Hernández-Orallo et al. (2013) | ◑ | ◑ | ○ | ◑ | ◑ | ● | ○ | ◑ | ○ |
| Our approach | ● | ● | ● | ● | ● | ● | ● | ● | ● |

The circles indicate the degree to which the (sub-)dimension is addressed. An empty circle (○) means that the subdimension is not covered, while a full circle (●) illustrates that the subdimension is covered fully. Circles that are filled to some degree (◑) illustrate that the respective paper addresses the subdimension accordingly or takes a different perspective on a different subdimension. Based on how the subdimensions are met, we obtain a final score for the dimension itself. The color-coding indicates the dimensions covered based on the paper's scope (**black**) and potential additional dimensions the method may cover but are not explicitly stated (gray)

since it is not designed to assess a developed model, it does not contribute to sufficient performance determination.

In contrast, Renggli et al. (2019) do not address minimal performance during pre-development. However, it fully supports the assessment of sufficient performance during post-development/post-deployment by taking an ML-model-centered approach to deployment decisions in continuous integration pipelines. Specifically, it evaluates whether a new model outperforms a defined threshold or the previously deployed model.

The remaining studies focus primarily on evaluating developed models. While they cannot derive performance thresholds themselves, they can assess minimal performance relative to predefined targets, fulfilling that subdimension partially. However, they all support evaluating sufficient performance by linking model output to economic objectives. For example, DataRobot (2024) and Florian et al. (2021) use their respective cost-benefit models to optimize decision thresholds and quantify the economic impact. Other works can assess sufficient performance based only on cost considerations (Drummond and Holte 2006, 2000; Hernández-Orallo et al. 2013; Zhou and Liu 2012).

**Performance uncertainty** Hawkins (2020) quantifies uncertainty by simulating ROC curves, calculating their AUC values, and evaluating whether they meet a predefined ROI threshold. Repeating this process establishes a statistical link between AUC values and ROI, thereby enabling the calculation of confidence intervals. However, since AUC values can be ambiguous (Yu et al. 2015) and may fail to reflect error trade-offs (Lobo et al. 2008), different ROC curves with the same AUC may yield different economic outcomes. As a result, this subdimension is only partially fulfilled. Moreover, this method is limited to early-stage simulations and does not apply to post-development or post-deployment settings. In contrast, Drummond and Holte (2006) and Holte and Drummond (2008) use bootstrapping on a given confusion matrix to generate multiple cost curves and derive confidence intervals. This method directly quantifies uncertainty and fulfills the uncertainty subdimension fully across all phases – for the pre-development phase, assumed confusion matrices are necessary. Renggli et al. (2019) define an $(\epsilon, \delta)$ criterion under which a model is only deployed if it outperforms a baseline with a probability of at least $1 - \delta$ (e.g., 99%) and within an error tolerance $\epsilon$ (e.g., the width of the $(1 - \delta)$-confidence interval). This allows for rigorous uncertainty quantification in the post-development and post-deployment phases, resulting in full fulfillment of the subdimension in those contexts. All other studies do not explicitly account for uncertainty, regardless of project phase.

**Integration into the ML life cycle** Regarding integration into the ML life cycle, the identified literature primarily addresses either the pre-development or post-development phases, which leaves a gap in comprehensive, end-to-end integration. For example, Renggli et al. (2023) link data quality to model performance by estimating whether a dataset is sufficient to meet a predefined performance target – but do not address how to define such a target. Similarly, Hawkins (2020) focuses exclusively on pre-development, whereas DataRobot (2024) and Florian et al. (2021) target post-development, overlooking earlier stages. Renggli et al. (2019) is the only work that explicitly addresses two phases (post-development and post-

deployment) by evaluating whether a new model meets a target performance or outperforms an existing one. Yet, they overlook pre-development assessment. In summary, none of the reviewed articles actively integrates their methods across the entire life cycle, which leaves this subdimension unfulfilled.

**Research gap** The existing literature on viability assessment is scarce and fragmented, as most studies focus on isolated aspects rather than addressing these subdimensions simultaneously. For instance, many studies consider only the costs of false predictions (Holte and Drummond 2008; Drummond and Holte 2006, 2000; Hernández-Orallo et al. 2013; Zhou and Liu 2012), without accounting for true positive benefits or broader economic impacts. Only one article provides a method to derive a minimal performance boundary (Hawkins 2020), and none explicitly supports both minimal and sufficient performance assessment within a unified framework. Similarly, uncertainty quantification is inconsistently addressed: while a few articles offer robust methods (Drummond and Holte 2006; Holte and Drummond 2008; Renggli et al. 2019), most ignore it entirely.

Moreover, existing methods are typically confined to either the pre- or post-development phase, and no article spans the full ML life cycle. As a result, current methods lack a life cycle perspective and offer limited guidance on when and how to apply viability assessments in practice.

To support real-world adoption and unlock the full potential of these methods, viability assessment must be systematically embedded into the ML life cycle. This requires not only the integration and simultaneous consideration of all key dimensions – consideration of costs and benefits, determination of a minimal performance, determination of performance sufficiency, and consideration of uncertainty – but also phase-specific guidance for their effective application.

### 3.2 From Research Gap to Business Risks

The current literature does not provide a holistic framework to assess the viability of ML models across all phases. In the pre-development phase, the literature lacks a clear definition of the minimum hypothetical performance required to achieve a specific economic criterion with a given level of confidence. For example, Florian et al. (2021) introduce a cost model for ML in PdM using ROC curves, but their method ignores uncertainty. This theoretical gap leads to practical issues: Businesses relying on ML-based PdM models may base investment decisions on performance estimates without assessing their uncertainty. Without quantifying performance uncertainty, organizations struggle to set realistic profitability and ROI expectations before development. The only study that addresses uncertainty while enabling the generation of a minimal necessary performance during pre-development is Hawkins (2020). Although the method quantifies uncertainty by linking AUC values, ROI, and confidence intervals, it is limited by the inherent ambiguity of AUC values. In business terms, relying solely on AUC values to make investment decisions can misallocate resources, as AUC values ignore economic viability. Even with a high AUC value, the model's economic value depends on the exact shape of the ROC curve, as FP and FN influence costs and revenue differently. Organizations must assess viability beyond AUC values to ensure alignment with business goals and justify investment. The same applies to the post-development and post-deployment phases of an ML project. Not quantifying uncertainty, as in Florian et al. (2021), can lead businesses to misjudge future revenue streams and, as a result, to form unrealistic financial expectations. Conversely, considering uncertainty without integrating cost-benefit analysis, as in Renggli et al. (2019), prevents businesses from rigorously translating uncertainty into actionable economic insights, which limits their ability to make informed decisions.

### 3.3 Key Challenges and Addressing the Research Gap

Based on our assessment, we identify three key challenges. First, recent work predominantly focuses on post-development and post-deployment assessment, but either neglects cost-benefit considerations or fails to quantify uncertainty. Second, there is little to no literature on methods that actively derive performance boundaries before model development, and those that do lack a holistic consideration of costs and have limitations in their uncertainty quantification. These limitations also prevent a simple combination of different methods, as a change in approach is required to address them. Third, the identified articles propose standalone methods that are not directly integrated into the ML life cycle, limiting their adoption in practical settings and across industries.

The proposed Viability Assessment Framework addresses these challenges by identifying minimal performance for viability in pre-development and assessing if estimated performance achieves viability in post-development and post-deployment, all while considering uncertainty. Further, it integrates seamlessly into the ML life cycle.

## 4 The Viability Assessment Framework

In the previous sections, three challenges were identified: (a) methods do not connect the performance and its estimation uncertainty with project viability assessment,

(b) methods do not assess the hypothetical viability of a project before and actual viability after model development and deployment, and (c) methods are not integrated into the ML life cycle. We address issue (a) by deriving the distribution of an estimator for the *profit-per-decision (ppd)* metric, which accounts for both misclassification costs and the benefits generated from correct decisions. Thereby, confidence intervals of the metric's estimation can be derived, which provide an intuitive method for integrating uncertainty, thus resolving issue (a), answering RQ1, and parts of RQ2. Additionally, *ppd* can be used to address RQ2 by introducing a method to evaluate a project's viability both before and after model development, thereby addressing issue (b). As viability assessment does not end with an initial deployment, we further consider decision-making during post-deployment. To resolve issue (c), we propose a three-step framework – the Viability Assessment Framework – that integrates into the standard ML life cycle and thereby consolidates the mentioned solutions (see Fig. 3). All steps of the framework are centered around the *ppd* metric. The functionalities of the framework are also provided in an open-source repository.[1]

### 4.1 Overview

On the surface, the Viability Assessment Framework follows three steps, centered around the *ppd* metric, which we describe further in Sect. 4.2. The first step is referred to as pre-development assessment and is positioned before the model development. Utilizing *ppd*, the step evaluates the viability of an ML project by estimating a minimum performance boundary needed to meet a defined economic success criterion with a given probability. We refer to this performance as hypothetically minimal necessary performance. A minimum *ppd*, referred to as the profit requirement, is one possible economic criterion. This results in a statement that links the project's hypothetical viability to performance and uncertainty. It thereby supports the decision of whether to proceed with the project and start the development. Under the assumption that the viability assessment has a positive outcome, the model development process begins. The second step of the framework, post-development assessment, comes in during the model evaluation stage. Here, the *ppd* is used to assess the actual viability of the developed model and links viability, actual performance, and uncertainty. It results in a statement about the actual viability and supports the decision of whether to deploy the model. The framework's third step, post-deployment assessment, is a recurring reassessment of viability, either at fixed intervals or continuously. Since the post-deployment assessment in our framework is identical to the post-development assessment, we will only provide a detailed outline of the former to avoid repetition.

### 4.2 Profit-per-Decision

In Sect. 2, we conceptualize the NPV from an ML perspective, dividing annual *CF* into model-dependent and model-independent components. Since model-dependent *CF*, driven by the costs and benefits of correct or incorrect decisions, are key to determining viability (e.g., influencing feasible infrastructure costs), we focus on linking this component to an appropriate metric. We argue that such a metric must account for all potential outcomes of model decisions, incorporating both the costs of misclassifications and the benefits of correct classifications. Additionally, relative measures are more suitable since absolute measures depend on the size of the test set. Therefore, the expected profit per decision is a suitable metric, and the respective mean can be used for estimation. The expected profit per decision is generally defined as the sum of the costs for each possible outcome, weighted by its probability. This is formalized in Equation (4).

$$ppd = \sum_{k \in K} P(k) * C(k) \tag{4}$$

For a set $K$ of distinct outcomes, the function $P$ maps the event $k \in K$ to the respective true probability, while the function $C$ maps the event to the respective costs/benefits – positive values signaling benefits and negatives signaling costs. Since we focus on binary classification, there are only four possible outcomes: TP, FP, TN, and FN. The expected value of the true *ppd* for binary classification is defined in Equation (5).

$$ppd = c_{TP} * p_{TP} + c_{FP} * p_{FP} + c_{TN} * p_{TN} + c_{FN} * p_{FN} \tag{5}$$

To estimate the expected value of *ppd*, we use the mean profit per decision based on a given confusion matrix (see Equation (6)) and convert the actual counts to estimated probabilities (see Equation (7)). This estimator is unbiased and approximately normally distributed, which allows us to calculate one-sided confidence intervals (CIs) (see Equation (8)). For more information on the derivation as well as the distribution characteristics, see Online Appendix B.

$$\widehat{ppd} = \frac{1}{n} * \left( c_{TP} * TP + c_{FP} * FP + c_{TN} * TN + c_{FN} * FN \right) \tag{6}$$

$$\widehat{ppd} = c_{TP} * \hat{p}_{TP} + c_{FP} * \hat{p}_{FP} + c_{TN} * \hat{p}_{TN} + c_{FN} * \hat{p}_{FN} \tag{7}$$

[1] https://github.com/dozip/TheViabilityFramework.git

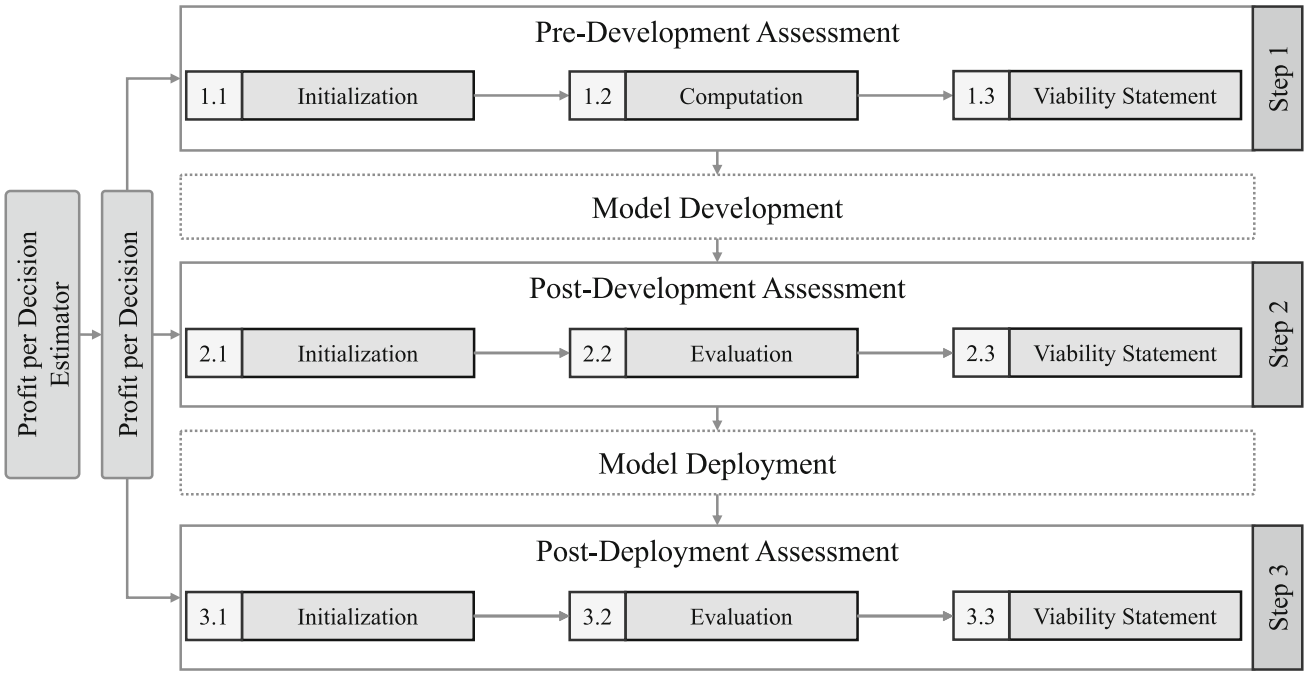


**Fig. 3** The Viability Assessment Framework consists of three steps that are integrated into the standard ML life cycle. The centerpiece of the framework is the proposed metric, profit-per-decision, and its estimator. These are used in the pre-development assessment to assess the viability before any development. If the assessment is positive, the standard ML model development starts. During post-development assessment, the developed model's viability is assessed. If viable, deployment is carried out. Post-deployment assessment is equal to the post-development assessment and offers a reassessment of viability during production

$$\left[\widehat{ppd} - z_{1-\alpha} * \sqrt{Var(\widehat{ppd})}, \infty\right] \tag{8}$$

A unique characteristic is that the expected/estimated value of *ppd* can be rewritten as a function of only the true/estimated true positive rate $\left(TPR/\widehat{TPR}\right)$ and true negative rate $\left(TNR/\widehat{TNR}\right)$ if we assume a class distribution a priori. The rewritten function for the expected value of *ppd* is formalized in Equation (9), while the parameter $p_{pos}$ denotes the proportion of positive samples in the considered dataset – referred to as base rate. By replacing the true values with the estimated ones, we can use the estimated rates to calculate the estimated *ppd*. This version of the true and estimated *ppd* is central to assessing a project's viability.

$$\begin{aligned} ppd\left(tpr, tnr\right) = p_{pos} * \left(c_{TP} * tpr + c_{FN} * (1 - tpr)\right) \\ +(1 - p_{pos}) * \left(c_{TN} * tnr + c_{FP} * (1 - tnr)\right) \end{aligned} \tag{9}$$

### 4.3 Pre-Development Assessment: Using Profit-per-Decision to Assess Project Viability Before Model Development

The first step of our framework aims to link a requirement for a project's viability, such as a minimum *ppd*, with a hypothetical minimal necessary estimated model performance while considering uncertainty. The consideration of uncertainty is characterized by the test sample size $n$ and the viability uncertainty $\alpha$. The step results in a statement about the project's hypothetical viability that contains the following information: *The project is viable with a probability of* $1 - \alpha$ *for the given profit requirement under the assumption that a hypothetical minimal necessary estimated model performance was achieved.*

Before we describe the method to generate this statement, we illustrate the idea more generally. Assume the performance of a model is estimated by a generic metric $P : (TP, FN, TN, FP) \rightarrow [0, 1]$ mapping the values of a confusion matrix to a value between zero and one. Before the development process and data gathering start, we want to assess what estimated performance $\hat{P}$ is needed (at least) to achieve a minimum required true performance $P_{req}$ with a given probability $(1 - \alpha)$. This performance represents the hypothetical minimal necessary estimated performance. As a solution, we can use a left-sided confidence interval of

the form $\left[\hat{P} - z_{1-\alpha} * SE, \infty\right]$ and try to find the lowest value of the estimated performance $\hat{P}$ such that the lower bound of the confidence interval is equal to or greater than $P_{req}$ $\left(P_{req} \leq \hat{P} - z_{1-\alpha} * SE\right)$. We denote this value as $P_{min}$. Assuming a test sample size of $n = 100$ and a minimum required true performance of $P_{req} = 0.9$ (see Fig. 4), we analyze which value of $\hat{P}$ is needed (at least) to achieve the true performance with a probability of at least 95% ($\alpha = 0.05$). Therefore, we consider the hypothetical performances of 0.90, 0.925, 0.95, 0.975, which will be referred to as the performance granularity. The line, representing the lower bound of the confidence interval $\left(\hat{P} - z_{1-\alpha} * SE\right)$ for an estimated performance of 0.95, is the first to cross the dotted line at $n = 100$. Therefore, a minimum estimated performance of 0.95 is required since it is the first considered value that leads to a lower bound above 0.9. We set $P_{min} = 0.95$.

This generic concept, in combination with the proposed metric, enables the generation of a hypothetical minimum necessary estimated performance and is the essence of the pre-development assessment step. Here, $P_{min}$ is not based on an arbitrary minimum required true performance $P_{req}$ but on a minimum required *ppd*, establishing a link between economic and ML perspectives. The step consists of three substeps: initialization (1.1), computation (1.2), and the viability statement (1.3), which are illustrated in Fig. 5.

In step 1.1, different parameters need to be set, such as the cost-benefit matrix $C$, which depends on the project, and the viability uncertainty, which will most likely be set by management. The parameters necessary for the computation, such as the considered sample sizes, assumed base rate, and considered performances (performance granularity), need to be set as well. Especially the definition of the cost-benefit matrix is a central challenge and essential for linking economic and ML perspectives, with approaches varying by domain. In cancer detection, Vargas-Palacios et al. (2023) derive costs from domain experts, while Ziegelmayer et al. (2022) use medical data and literature. In churn prediction, studies show that acquiring new customers is five times more expensive than retaining current ones, guiding the cost estimates (Wagh et al. 2024; Gui 2017). For predictive maintenance, downtime, part, and labor costs can be used to define both costs and benefits (Prytz et al. 2015). After the definition of all required parameters, the hypothetical minimal necessary estimated performances are generated for the considered sample size (step 1.2). Similar to the generic concept, we want to find a minimum estimated performance to achieve a minimum required true *ppd* with a given probability. In this case, *ppd* is set to zero. As demonstrated in Equation (9), the value of $ppd/\widehat{ppd}$ can be calculated using the true/estimated $TPR/\widehat{TPR}$ and $TNR/\widehat{TNR}$, which means that the minimum estimated performance is not only a scalar but a tuple of $\widehat{TPR}$ and $\widehat{TNR}$. Therefore, we are looking for all combinations of $\widehat{TPR}$ and $\widehat{TNR}$ that will lead to a true *ppd* greater than zero with a given probability. With a performance granularity of 0.01, all values between zero and one in 0.01 steps are considered: (0, 0.01, 0.02, ..., 0.99, 1). Formally, we are interested in all combinations of $\widehat{TPR}$ and $\widehat{TNR}$, such that the condition below holds (see Equation (10)). In essence, we again try to find the smallest value $P_{min}$ such that $P_{req} \leq \hat{P} - z_{1-\alpha} * SE$.

$$0 \leq \widehat{ppd}\left(\widehat{TPR}, \widehat{TNR}\right) - z_{1-alpha} * \sqrt{Var\left(\widehat{ppd}\left(\widehat{TPR}, \widehat{TNR}\right)\right)} \tag{10}$$

As the main interest lies in the minimum performance, only the first tuples that fulfill the condition are considered. For example, if $\left(\widehat{TPR}, \widehat{TNR}\right) = (0.8, 0.9)$ and $\left(\widehat{TPR}, \widehat{TNR}\right) = (0.81, 0.9)$ both fulfill the condition, we only consider the first one. Due to a trade-off between errors, more than one single tuple will usually be identified. By trade-off, we mean that one more FP error can be equalized by a certain number of fewer FN errors. The set of all identified combinations is used to perform a linear regression, which describes the performance bound functionally, as illustrated in Equation (11).

$$TNR = c - m * TPR \tag{11}$$

The variable $m$ estimates the trade-off between FP and FN model decisions regarding the minimal required true *ppd* (e.g., $m = 2 \Rightarrow$ a one percentage-point decrease in TPR can be equalized by a two percentage-point increase in TNR). The extracted linear relationship generates a viability statement regarding the viability uncertainty and sample size. Online Appendix C contains a more detailed description of the method. In step 1.3, the results are reported visually by plotting the found set of tuples together with the respective linear regressions, described in more detail in Sect. 5. The results are reported textually by generating the viability statement:

*The project is viable with a probability of 1- α for a required true ppd of zero if $TNR \geq c - m * TPR$ holds for the estimated model performance if it was tested with n samples. For every percentage-point decrease in TNR, m percentage-points in TPR are to be gained to conserve viability.*

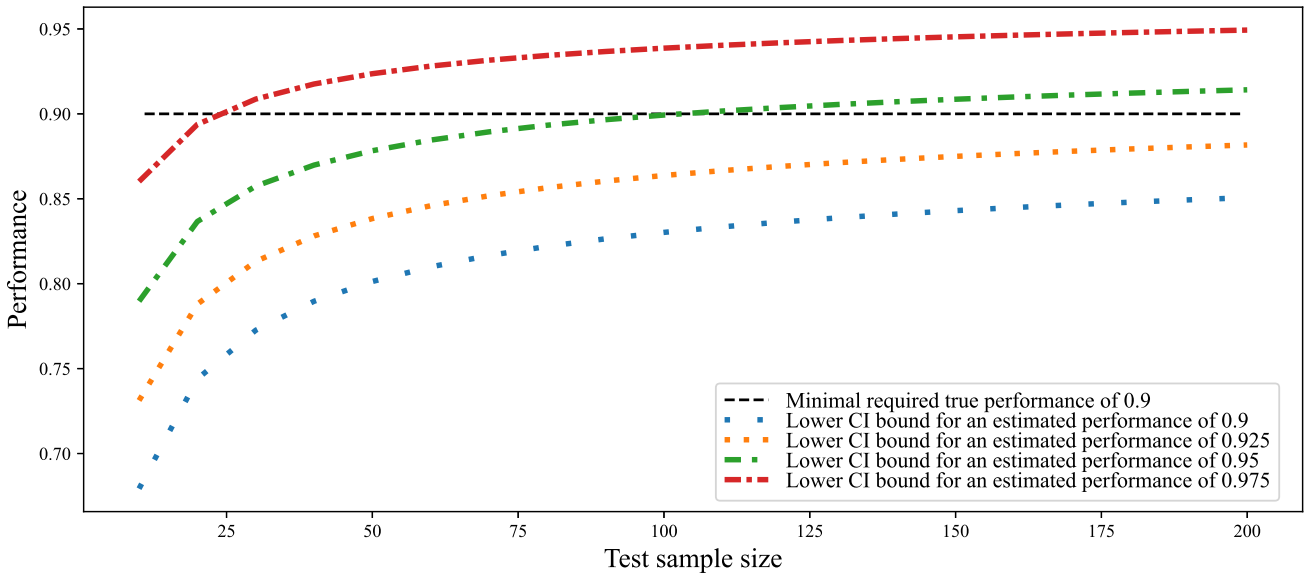


**Fig. 4** Visual illustration of generating a minimal performance $P_{min}$ for a minimal required $P_{req}$ one. The vertical line represents the minimum required performance, while all other lines represent the values of a confidence interval's lower bound over sample size for a given value of $P$. For example, the lowest line describes the development of the lower bound over sample size for $P = 0.9$.

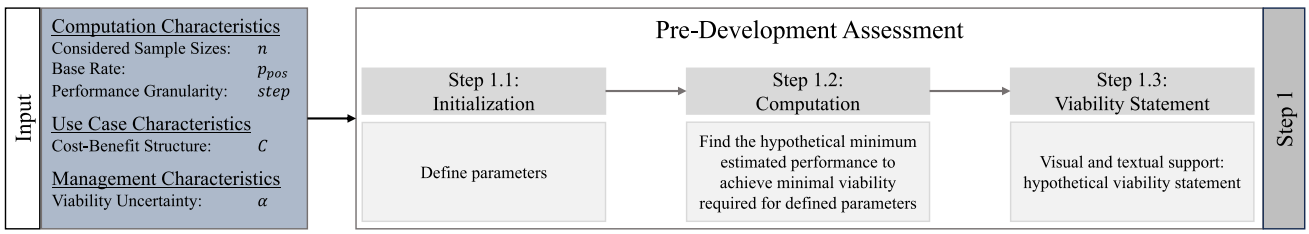


**Fig. 5** The first step, pre-development assessment, is divided into three substeps. During the initialization (step 1.1), different parameters are defined. During the computation (step 1.2), the assessment is carried out. During the viability statement (step 1.3), the results are presented visually and textually in the form of a statement

### 4.4 Post-Development Assessment: Using Profit-per-Decision to Assess Project Viability After Model Development

Post-development assessment is the second step of the framework and introduces project viability to the typical evaluation process. In contrast to step 1, we link a project's viability to an actual estimated model performance under uncertainty considerations by applying the proposed metric *ppd* to a confusion matrix. This matrix is the result of evaluating the model on a final global holdout set. To offer more guidance during the application, the step is also divided into three substeps (Fig. 6). The three substeps are initialization (2.1), evaluation (2.2), and the viability statement (2.3).

During initialization (step 2.1), the viability uncertainty $\alpha$, the cost-benefit matrix $C$, and the used test sample size need to be defined. Then, a model is evaluated, which results in a confusion matrix. The matrix is used to estimate *ppd*, and the relevant left confidence interval is calculated (step 2.2). Lastly, during step 2.3, the CI is used to generate the actual viability statement:

*The lower bound of the* $1 - \alpha\%$ *confidence interval of the profit-per-decision is X. Based on the actual estimated profit-per-decision* $\widehat{ppd}$*, there is a* $p\%$ *chance to estimate* $\widehat{ppd}$ *or larger if ppd is zero. Therefore, the project is viable with* $(1 - p)\%$ *probability.*

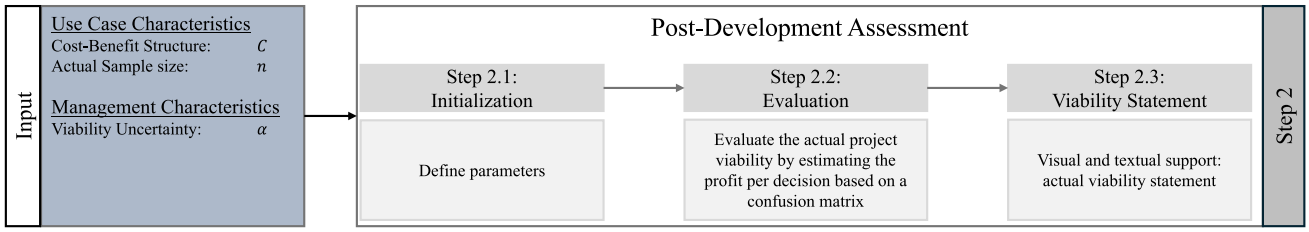


**Fig. 6** The second step, post-development assessment, is divided into three substeps. During the initialization (step 2.1), different parameters are defined. During the evaluation (step 2.2), the project's actual viability is assessed. During the viability statement (step 2.3), the results are presented textually in the form of a statement. The post-deployment assessment is conducted in the same way

## 5 Exemplary Application

This section applies the Viability Assessment Framework to a real-world use case, demonstrating its application from the initial concept to model monitoring. We emphasize its three steps (pre-development, post-development, and post-deployment), which thereby complement frameworks like MLOps (Kreuzberger et al. 2023). The section aims to provide detailed practical guidance on the framework's application based on a public real-world dataset. The described workflow aligns with the workflows and decision-making processes outlined by industry experts in interviews and the scenario-based evaluation, having experience in managing and developing models for ML-based PdM applications.

### 5.1 Example Use Case: Air Pressure System of Scania Trucks

To illustrate the Viability Assessment Framework, a common dataset in PdM and CbM is used, containing real-world data from Scania trucks. The dataset[2] provides real-world operational data related to the air pressure system (APS), which controls braking and gear shifting functions. Positive samples indicate failures in a specific APS component, while negative samples represent normal operation. The classification task is to predict maintenance needs (class 1: failure, class 2: no failure). We use the dataset as an example and take, if available, parameters from other articles and repositories that utilize the same data. Further, the dataset is used to train a model and showcase the utilization of the post-development and post-deployment assessment.

### 5.2 Pre-Development Assessment

Imagine a logistics company managing a large fleet of trucks, striving to reduce downtime and improve operational efficiency. A business stakeholder is tasked with developing an ML-based PdM application to anticipate failures before they happen, minimize costly breakdowns, and optimize fleet performance.

Before committing resources, she wants to assess how project characteristics (e.g., costs for an FP), managerial guidelines (e.g., viability uncertainty), and business directives (e.g., profit requirements) are connected to the performance of a potential ML model. She aims to determine the performance needed to meet the profit requirement and whether such performance is achievable.

Using the Viability Assessment Framework, she aligns economic success criteria with ML success criteria, leveraging pre-development assessment functionalities that provide additional guidance regarding the necessary parameters and information. This structured decision support provides clarity on whether the investment is justified before full-scale development is initiated. Therefore, she collaborates with the after-sales team to define a cost-benefit matrix, which quantifies the costs and benefits of model decisions, a profit requirement that sets the minimum expected profit based on model outcomes, and the viability uncertainty that defines the maximum probability of failing to meet the profit requirement. Additionally, she coordinates with engineering teams to assess data availability, determine possible train and test dataset sizes, and estimate failure rates.

As a result of her consultations, a cost-benefit matrix is defined as $C = (\,15.0, \quad -16.5, \quad 0.0, \quad -1.5\,)$ (in 1,000 currency units (CU)) (Prytz et al. 2015), where $c_1$ represents the benefit of correctly predicting a failure, $c_2$ the cost of missing a failure, $c_3$ the neutral gain of correctly predicting no failure, and $c_4$ the cost of a false failure prediction. Furthermore, the viability requirement is set at 80%, the profit requirement is 0 CU, the base rate is assumed to be 1%, and the assumed future test dataset consists of 1,000 samples. To determine the hypothetical minimum required performance, she runs the pre-development assessment step of the Viability Assessment Framework, obtaining the following visual results containing the exact performance bounds (see Fig. 7a) as well

[2] https://www.kaggle.com/datasets/uciml/aps-failure-at-scania-trucks-data-sethttps://www.kaggle.com/datasets/uciml/aps-failure-at-scania-trucks-data-set (Accessed 30 Dec 2025)

as textual results building on the linear regression as described in Sect. 4.3.

*A viable model tested on 1,000 samples requires a minimum TPR of* 64%. *For every* 1%*-point decrease in the TNR, the TPR must increase by* 4.3 *percentage-points to maintain viability.*

Based on these results, she consults domain experts, reviews past projects, analyzes reported performance in research papers, and adds identified benchmarks (Costa and Nascimento 2016) to the results from the pre-development assessment (see Fig. 7b). Then, she presents the results to a steering committee in her company consisting of other business stakeholders. The committee deems the necessary performance achievable as the identified benchmarks are above the performance boundary. Consequently, she initiates the model development phase.

### 5.3 Model Development

Before model training, the necessary data is collected and divided into training, validation (optional), and test sets. Here, the pre-development assessment guides the split by analyzing how the test dataset's size influences the viability boundary. If increasing the test sample size has a negligible impact on the performance boundary, the test dataset's size is finalized (Fig. 8a). A developer sets up an XGBoost classifier and fine-tunes its hyperparameters using Grid Search with 10-fold cross-validation on the training set. He uses the test set as a single data fold to evaluate model performance. Aware of the highly imbalanced class distribution, he integrates SMOTE (Chawla et al. 2002), applying it dynamically to each fold during the GridSearch to improve class balance.

### 5.4 Post-Development and Post-Deployment Assessment

After model development, its viability must be assessed before deployment. The developer evaluates the model using 2,000 test samples and optimizes the decision threshold by leveraging ROC curves to balance sensitivity and specificity. The resulting confusion matrix[3] is used in the post-development assessment step of the framework, assessing the viability based on a 20% viability uncertainty. The post-development step produces visual decision support (Fig. 8b) that directly indicates whether viability is achieved, supplemented by textual information offering further details and context on the viability assessment:

[3] TP: 43, FN: 18, TN: 1867, FP: 72.

*The lower bound of the 80% confidence interval of the profit-per-decision is 68.96 CU. Based on the actual estimated profit-per-decision of 120.00 CU, there is a* 2.0% *chance to estimate 120.00 CU or larger ifppd is zero. Therefore, the project is viable with 98% probability.*

The development team reports the viability statement alongside the figure to the business stakeholder, presenting it again to the steering committee. The members of the committee can use the provided information to decide whether the current model has sufficient performance to be deployed. The confidence of viability (98.0%) is above the required 80%.

As the model has a 98% chance of achieving the defined profit requirement, the ML model is deployed. After deployment, performance should be continuously monitored and evaluated. The Viability Assessment Framework can be utilized by the performance-monitoring team to reassess viability continuously based on new field data and decide whether to maintain the model in production or initiate retraining. Furthermore, the team can use the estimator ($\widehat{ppd}$) to compare the profit of a new model against the current one, ensuring data-driven decision-making. Additionally, when planning a new project, other business stakeholders can leverage the results of this project to benchmark the achievable performance of their project. This is particularly valuable during the pre-development assessment, helping them to determine whether the hypothetical minimum necessary performance is realistic before committing further resources.

### 5.5 A Final Remark

We would like to add a final remark on the exemplary application. This section presents a simple workflow demonstrating how the Viability Assessment Framework can support viability assessments throughout different phases of an ML project. For instance, before presenting results to the committee, the framework can be used for scenario analysis, exploring various profit and viability requirements. Similarly, in post-development or post-deployment, the model can be evaluated to estimate the probability of incurring a specific loss or achieving a certain profit, providing insights into both potential benefits and risks.

## 6 Evaluation

We evaluated the Viability Assessment Framework using two methods to conduct a practice-oriented evaluation

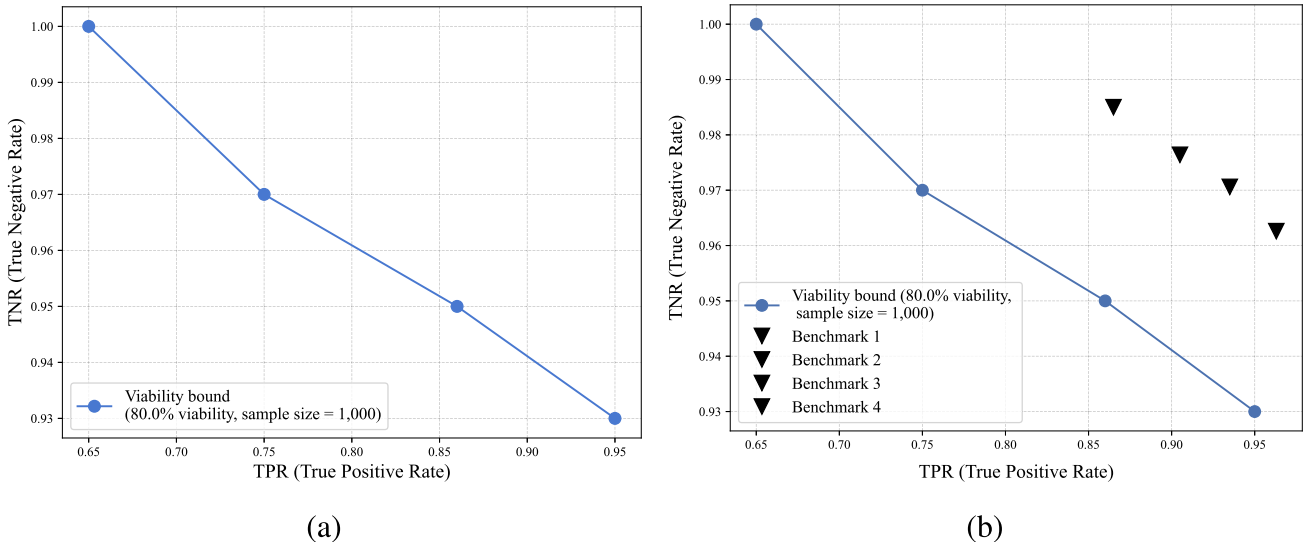


**Fig. 7** The figure shows how the Viability Assessment Framework can be utilized. Subfigure (**a**) shows the results of the pre-development assessment, while (**b**) complements the plot with benchmarks

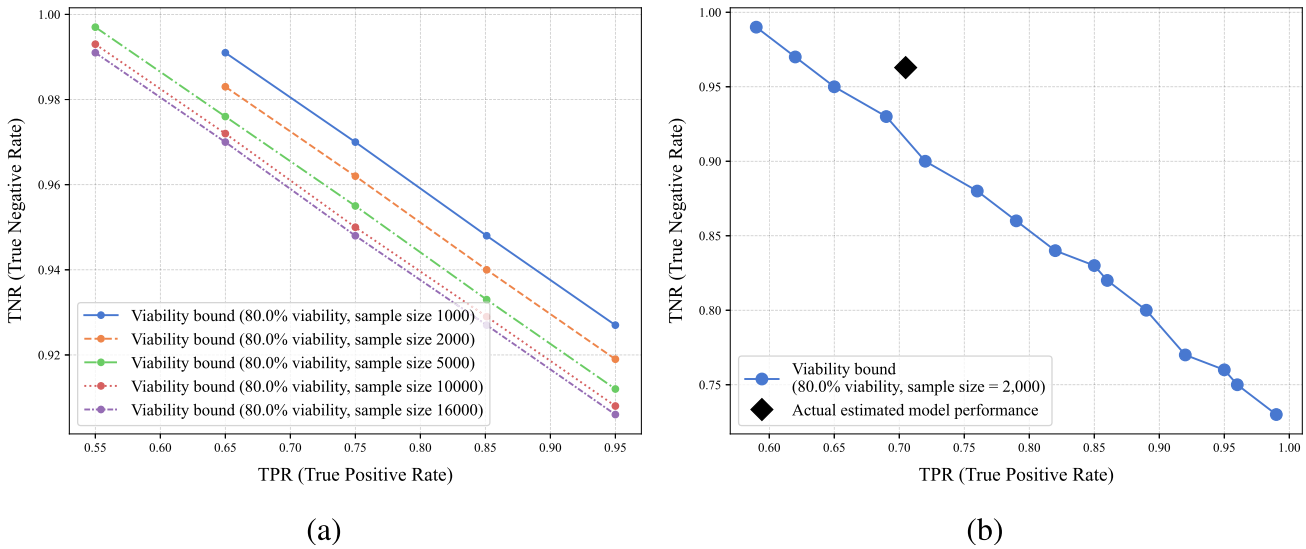


**Fig. 8** The figure shows how the viability assessment framework can be utilized. Subfigure (**a**) explores boundaries for various test dataset sizes, thereby supporting the data split decision, while (**b**) compares the boundary to the actual model performance

from a socio-technical perspective: semi-structured interviews and a scenario-based evaluation. The interviews evaluated the framework from multiple perspectives by engaging industry experts who advise companies on ML use case identification, development, deployment, and monitoring, as well as data scientists working on PdM applications – ranging from early proof-of-concept (PoC) stages to production-ready solutions. The interviews were carried out in three rounds with a total of eight experts from six different companies. The scenario-based evaluation adopted a practice-oriented approach, which enabled experts to interact with the framework's functionalities through a dedicated web application, facilitating hands-on evaluation in a realistic setting. The scenario-based evaluation was conducted in two rounds using a past use case from a world-leading German car manufacturer. Overall, eight experts from six different companies took part in the scenario-based evaluation. In total, 16 experts from 11 companies evaluated the framework.

### 6.1 Interview Study

To evaluate the necessity and suitability of the Viability Assessment Framework, we conducted semi-structured expert interviews. Semi-structured interviews offer the benefit of acquiring opinions (Hammarberg et al. 2016)

while enabling a more adaptable interview situation (Adams 2015).

For the initial interviews, we selected the initial interviewees purposely, inspired by Glaser and Strauss 's (1967). We selected the first three interviewees (I1-I3) for their expertise in consulting organizations in identifying, implementing, and maintaining ML applications, allowing us to evaluate the framework for business decision-makers. Based on the pre-screening of the data, subsequent interviewees were chosen. The second set of interviewees (I4-I6) consisted of specialists with expertise in ML techniques, in order to gain deeper insights into the actual usefulness of our framework for developers. The final round of interviews (I7-I8) aimed to acquire more insights from different fields and domains while increasing the collected data. Furthermore, the last round was used to analyze theoretical saturation (Glaser and Strauss 1967). In total, we conducted eight interviews with eight different participants from six different companies, with an average duration of thirty-seven minutes. A detailed description of the interview structure is provided in Online Appendix D while an overview of the interviewees can be found in Table 2.

### 6.2 Scenario-Based Evaluation with Industry Experts

To complement the interview results, we conducted a scenario-based evaluation with eight experts from six different companies over two rounds, applying the Viability Assessment Framework to a past use case of a world-leading German car manufacturer. Below, we provide an overview of the general approach and process.

*Use case* We considered a former supervised learning problem for diagnosing APS failures in premium vehicles, aiming to classify the system as operational or at risk of failure. Here, $p \in [0, 1]$ represents the continuous, normalized probability of APS failure, derived from sensor readings and operational data. We defined a binary failure label $y \in \{0, 1\}$, mapped from $p$ based on a threshold $tr_{fail}$ such that:

$$y = \begin{cases} 1, & p > tr_{fail} \\ 0, & \text{otherwise} \end{cases} \tag{12}$$

Repair shops labeled the ground truth based on workers' APS assessments, identifying necessary repairs (TP) or unnecessary ones (FP). FN arise from breakdowns due to APS failure or repairs during routine inspections. The ML-based PdM model was trained on the dataset $D = \{(x^{(i)}, p^{(i)}, y^{(i)})\}_{i=1}^{N}$ where $N$ is the number of vehicle records. Features included APS sensor readings (e.g., pressure or temperature), operational parameters (e.g., engine load or acceleration), and vehicle-specific data (e.g., age or maintenance history). Due to disclosure rules, in-house data was replaced with public data, though the use case remains aligned.

*Web application* To enhance the realism, we deployed the framework as a web application. The web app has three main functions: (1) introducing the Viability Assessment Framework within a typical project workflow (e.g., CRISP-DM or MLOps), (2) enabling intuitive pre-development assessments via an input form for use-case parameters and result analysis, and (3) supporting post-development and post-deployment assessments. A detailed overview is provided in Online Appendix E.

*Process of the scenario-based evaluation* The overall process is outlined in Fig. 9. After a brief introduction, the use case was illustrated. Participants went through the use case twice, progressing through the pre-development and post-development steps – once on their own and once using the framework. In the first iterations, experts were asked to describe their approach to answer the central question for each step. Then, the use case was revisited utilizing the framework's web application. Participants could modify key parameters, such as the test dataset's sample size or viability uncertainty, to explore different scenarios. At the end, we presented a standardized set of questions assessing the framework's usefulness, the relevance of the provided information, and the likelihood of integration into participants' workflows.

*Participants and analysis* For the scenario-based evaluation, we selected eight additional experts from six different companies possessing extensive industry experience and expertise in ML-based PdM for two rounds. In the first round, we worked with experts from manufacturers who were closely familiar with the use case. In the second round, we extended the evaluation to PdM experts from various companies, who were not fully familiar with the specific use case but had strong expertise in ML development and ML-based PdM. Each expert is experienced in developing early ML-based proof-of-concept solutions and transitioning them to production. An overview of the experts is provided in Table 3. The analysis followed the same structured approach as for the interviews, provided in Online Appendix D. Additionally, responses to the closing questions were analyzed.

### 6.3 Synthesis of Results

In this subsection, we present the findings from the interviews and scenario-based evaluation in a synthesized and aggregated form, identifying overarching themes and supporting them with expert quotes. The source of each quote is indicated by the prefixes: "I" for interviews and "C" for the scenario-based evaluation.

**Table 2** Overview of interview experts

| ID | Professional Title | Related Industry | Company ID | Experience (Years) | Duration (Min.) |
|---|---|---|---|---|---|
| **Interview Round 1 - Focus on Decision-Makers from the Business Perspective** | | | | | |
| I1 | Consultant | Production & Operations Consulting | A | 2 | 40 |
| I2 | Senior Manager – Data Science | Data Science & Consulting | B | 7 | 30 |
| I3 | Data Scientist | Data Science & Consulting | B | 7 | 34 |
| **Interview Round 2 - Focus on Decision-Makers from the Development Perspective** | | | | | |
| I4 | Product Owner PdM | Automotive | C | 7 | 35 |
| I5 | Machine Learning Expert | Construction Industry | D | 8 | 48 |
| I6 | Data Scientist Pre-Development | Automotive | C | 4 | 26 |
| **Interview Round 3 - Focus on Additional Data Collection and Saturation** | | | | | |
| I7 | Project Manager & Data Scientist | Data Science & Consulting | E | 4 | 43 |
| I8 | Consultant | IT & Strategy Consulting | F | 2 | 40 |

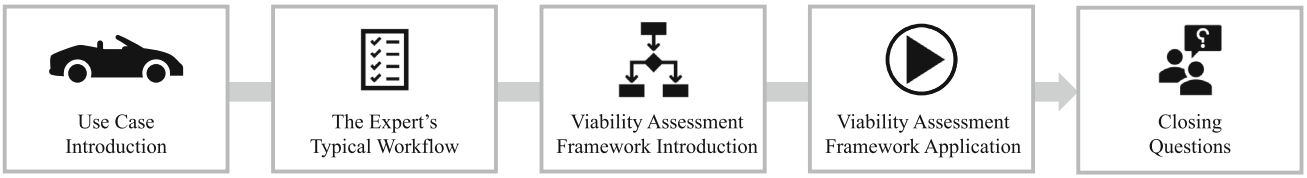


**Fig. 9** The figure illustrates the different stages of the scenario-based evaluation from (1) the use case introduction, (2) experts describing their usual approach, (3) introducing the Viability Assessment Framework, (4) the application of the Viability Assessment Framework, to (5) closing questions

We identify five overarching themes: (1) the *necessity* of our Viability Assessment Framework, (2) the framework's *usefulness from an organizational perspective*, (3) the framework's *usefulness from an operational perspective*, (4) prioritizing *cost vs. uncertainty* in viability assessment, and (5) *improvements and challenges*. Although the data provides other insights, such as maintenance schedules and the ability to replace components or modules, we focus on the dimensions and subdimensions relevant to evaluating our framework.

*Necessity* During the interviews and scenario-based evaluation, experts emphasize the importance of considering two perspectives where "one perspective is technical and the other one is economic" (C2). However, experts agree that there is a lack of methods that align the economic perspective of a project with its ML perspective – especially before model development – constituting the necessity of the Viability Assessment Framework. C8 states that "linking business and ML metrics is a crucial and important, yet unresolved issue in industry". Further, I2 emphasizes that not deriving ML success criteria from an economic perspective is "exactly the blind flight [developers are] doing today. They say 'best-can-do', but it's not clear what the target [performance] is". This statement is supported by I5: "What is often unclear to colleagues is what the target metrics should be. They tend to say there shouldn't be any errors, so I find [the Viability Assessment Framework] very helpful". I8 describes that in his experience "organizations frequently do not know what they want and define [...] requirements arbitrarily [...] leading to irrational decisions". Further, C4 critiques the current process, noting that it starts with model development before linking model performance to the business case, whereas he would prefer to define performance criteria first. Additionally, the business case is usually not properly calculated, leading to a misalignment between economic and ML perspectives (I2, C2). A tool to support the alignment by connecting business, economic, and ML perspectives is "currently unavailable" (I2). C6 states that "there should be more tools [like the Viability Assessment Framework] because tools in this area are missing." Most experts who applied the framework via the web app (C1-C7) showed strong interest in integrating the framework into their workflows. One expert (C2) was especially enthusiastic, requesting to share the application internally with product owners, underlining the need for and necessity of the proposed framework.

*Usefulness from an organizational perspective* During the interviews, the benefits of the Viability Assessment Framework from an organizational perspective become evident, particularly in project management. The framework is seen as a potential way to *enable portfolio*

**Table 3** Overview of experts for the scenario-based evaluation

| ID | Professional Title | Related Industry | Company ID | Experience (Years) | Duration (Min.) |
|---|---|---|---|---|---|
| **Round 1 - Companies Use Case Experts** | | | | | |
| C1 | Data Scientist | Automotive | C | 10 | 46 |
| C2 | Data Scientist | Automotive | G | 9 | 56 |
| C3 | Data Scientist | Automotive | C | 5 | 45 |
| C4 | Data Scientist | Automotive | C | 4 | 50 |
| **Round 2 - Experts from Other Companies with Partial Use Case Familiarity** | | | | | |
| C5 | ML Engineer | Software Development | H | 7 | 40 |
| C6 | Data Scientist | Automotive | I | 2 | 50 |
| C7 | Data Scientist | Automotive | J | 6 | 57 |
| C8 | Data Scientist | Consulting | K | 9 | 61 |

*management* as it puts a profit uncertainty tag next to each project – during pre- and post-development alike (I2). This enables project managers to prioritize possible applications, adding additional value as project managers are driven by two things: "risk and profit expectations" (I5). Further, the integration of the framework into the organization's project management can ensure *consistency* with broader goals. For example, costs and uncertainty thresholds (e.g., 99%) "must not be arbitrary" (I5) and are ideally aligned with sales and marketing. Additionally, the framework would *standardize, objectify, and clarify* the decision-making process across an organization, increasing its transparency. C2 explains how "there is currently no standardized framework for the decision-making process [in ML projects], leading to subjective decisions based on rule-of-thumb estimates".

*Usefulness from an operational perspective* The framework's usefulness for operational tasks encountered during ML projects is a key aspect repeatedly emphasized. Experts highlight the framework's usefulness in aligning goals across disciplines by encouraging stakeholders to contribute their expertise. Below, we outline key aspects to illustrate how the framework supports goal alignment and interdisciplinarity.

Some experts categorize the Viability Assessment Framework as an "expert tool" for different reasons, influencing the *understandability* of the decision support (I1-I5, C5, C7, C8). For one, the framework increases the overall complexity of the decision by introducing uncertainty (I1-I3, I5, C7, C8). I4 emphasizes that the value of the framework during pre-development is especially given if an expert can judge which performance is common and what current benchmarks are. Another reason mentioned is that the provided information "is not suited for management" (I5). C5 states that "the tool would likely be best handled by someone operating at the intersection [of business and technical perspectives.]" Yet, experts agree that the provided information facilitates reporting towards management, emphasizing the *reporting and translation capabilities* of the framework (I2-I6, C1-C3, C5-C7). I6 states that the visual and textual information is "highly appropriate for management meetings [and is a nice] selling slide" which is supported by C2 stating that the framework acts as "a translation tool between the developer and business stakeholder". Further, C3 emphasizes that integrating boundaries with actual measured performance enhances reporting effectiveness, while C8 sees benefits in "raising awareness for different, important components among non-technical stakeholders or junior developers when analysing a use case, [...] especially during early ideation phases to identify use cases that are doomed from the start." The translation capabilities complement the framework's ability to *support finding a target performance* (I2, I4, I5, I7, C3, C4), which directly aligns with the experts' rationale for its necessity. I2 describes that "if the business side of an organization could define a requirement and simply translate it, it would be super, super helpful". Additionally, understanding "when a model is a good model" (I4) and identifying "where I need to go with my performance based on a theoretical threshold would have been helpful during past discussions" (I5). These *benchmarking capabilities* are highlighted by multiple experts (I7, I8, C2, C5-C7). For example, C2 states: "If I've done a similar project with a comparable component, I can use its performance to assess my current project's viability early on." Furthermore, I7 mentions that, based on the visualization, he considers the exemplary project to be ambitious, as it requires achieving a minimum TPR of 80% with an FPR of 10%. The interaction between the interviewee and the framework illustrates how the provided information can be used to assess the project's viability. Additionally, I7 explains how he would "refer to literature, review past projects, or consult experts to determine how well models typically perform for this application". I8 would use the framework in a similar trajectory, by comparing past projects within the same domain and determining if the derived ML success criterion is achievable. C7 emphasizes that when multiple

models are developed, the resulting performance boundary plays a key role in distinguishing between good and poor models. In addition, experts highlight the framework's *quality gate capability* throughout the project, particularly during the iterative model development process, where usually multiple PoCs are developed (C1-C3, C7). By aligning goals early on, preliminary results like PoCs can be assessed against defined targets, enabling business stakeholders to guide developers while providing developers with clear performance goals, and reducing the need for self-set benchmarks (C1, C4). Further, the framework provides a clear decision boundary based on pre-defined and aligned economic success criteria and uncertainty thresholds (I5, C3, C4).

Moreover, the framework stands out positively by enhancing the *availability of relevant information* (I1, I3, I4, I5, C1-C7). I3 states that the framework "provides additional information" while I1 emphasizes that the visualization aids in capturing the illustrated information. Furthermore, the experts agree that the framework supports decision-making by enhancing the ability to decide (I2), increasing trust in the decision (I3), and providing a greater sense of security during the decision-making process (I2). I2 also states that the framework might help to solve decision vacuums, as developers and management can more clearly define who makes which decisions.

While most experts view the framework's operational usefulness positively overall, I6 expresses skepticism that, while the additional information is relevant, it is not clear how the information would help to make a decision more confidently. C8 takes a more critical stance, arguing that the current version of the framework relies on too many assumptions (e.g., available cost data) and oversimplifies real-world scenarios. It is argued that "the cost-performance relationship is either too trivial, where no tool is necessary, or too complex for the framework" (C8). However, a potential value for non-technical stakeholders or junior developers is acknowledged, especially if its functionality is expanded (C8). Nevertheless, most experts deem the Viability Assessment Framework highly useful from an operational perspective.

*Cost vs. uncertainty* The experts' opinions diverge on which information is more critical: costs or uncertainty. Some argue that *cost is more important*, others believe that *uncertainty is more important*, while some emphasize that *the combination is key*. I6 states that the information about uncertainty is nice to have and "important for the management to deal with possible risks [...], but developing a feeling for how much a wrong prediction costs" is central for decision-making. I3 takes an opposing stance by emphasizing "that the uncertainty is the more important information". Other experts express that "the combination is highly beneficial". I4 describes that "costs are key because if the model performs poorly, it would never be implemented. However, security also plays a significant role, as uncertainty should be minimized". This perspective is shared by I7, who states: "I need both, but if I had to choose one, I would prioritize the costs." C7 also stresses that the additional consideration of uncertainty is highly beneficial, while C8 explains that uncertainty must always be considered.

*Improvements and challenges* During the interviews and scenario-based evaluation, several improvements to the framework are proposed. I7, C8, and I8 emphasize that *consideration of project costs* in "absolute numbers" (I7, C8) would be interesting as well as "the expected net profit to prioritize projects" (I8). C3 also emphasizes that additional information about "which [maximal] profit can be achieved" would be helpful. C7 notes that treating costs as non-deterministic could enable sensitivity analysis, enhancing overall usefulness. Further, the *extension of the visual decision-support* is emphasized. During pre-development, the framework generates boundaries and uncertainty. Experts suggest adding multiple lines to show hypothetical performance, different uncertainty thresholds, or fixed profits, thereby enabling scenario analysis (C4, C7, C8). This idea also applies to post-development, where I7 proposes using ROC curves to refine decision thresholds. C7 suggests using color-coding to distinguish preferable and non-preferable areas, and adding arrows next to the axis titles to help non-technical users understand how the metrics should be interpreted. Additionally, experts emphasize the need to show not just the probability of achieving a defined profit but also the risk of zero or negative outcomes, thereby focusing also on the *consideration of risk* (I5, I6). I5 notes that decision-makers need to compare probabilities of, for example, making \$50 versus losing \$20 per decision. This would better address both benefits and risks in decision-making. Further, applying the framework poses challenges on its own. C2 highlights difficulties in quantifying the costs and benefits of model decisions and mitigating the risk of subjective bias or external influences in value selection. On the matter of quantifying costs, C4 takes an opposing view: "If we correctly identify a failure, we aim to bring the customers into our repair shop. Each task or maintenance in our repair shop generates direct financial benefits, allowing us to quantify the value of a true positive, depending on the use case." Further, I2 mentions that developers, as well as management, must trust the framework in the first place, emphasizing the necessity to define *requirements for usage*. C1 and C7 comment on the risk of manipulating results, highlighting the need for defined requirements when using the framework, as well as clear guidelines for validation. Lastly, C5 mentions that for a real-world

implementation, *firm-specific particularities* must be considered.

## 7 Discussion

Organizations still struggle to identify promising machine learning (ML) projects due to the difficulty of correctly assessing the economic benefit (Benbya et al. 2021; Weber et al. 2023). Therefore, rigorous viability assessment is crucial, as it links the estimation of the ML model's performance to potential economic success while considering estimation uncertainty.

Through a systematic literature review (SLR) on viability assessment, we found that existing literature predominantly addresses either the costs and benefits or the estimation uncertainty, but rarely integrates both, while typically not being explicitly embedded into the ML life cycle. To close this gap, we proposed the Viability Assessment Framework, which shifts the focus to an economics-centered assessment of ML projects. The framework consists of three steps: *Pre-development assessment*, which estimates the *hypothetical* minimum necessary model performance for project viability, *post-development assessment*, and *post-deployment assessment*, which both evaluate whether the *actual* performance is sufficient for project viability. All phases are centered around an estimator for the expected profit-per-decision of the underlying ML model. We showcased the framework's capabilities by applying it to a public, real-world predictive maintenance dataset, demonstrating how it can inform decision-making. Further, we evaluated the framework through expert interviews from management and development perspectives, as well as by means of a scenario-based evaluation applying its web-based implementation to a real-world use case from a leading German automotive manufacturer. In total, 16 experts from eleven different companies participated in the evaluation.

In the remainder of the discussion, we explore potential extensions to our framework, summarize theoretical as well as practical contributions, and highlight limitations.

### 7.1 Potential Framework Extensions

During the evaluation, experts suggested different possibilities for extending and refining the Viability Assessment Framework.

*Visual decision support* Extensions to the visual decision support could integrate multiple economic and uncertainty criteria, adding risk perspectives. Showing multiple lines for different profit and viability uncertainties during pre-development assessment could offer a broader view of benefits and risks. The requested lines can already be generated within the current framework, as it allows for dynamic adjustments of viability uncertainties and economic criteria. During post-development assessment, combining the graphs from the post-development assessment with ROC curves could help refine thresholds and more accurately assess risks and benefits. Since ROC curves are standard in performance evaluation, they could be easily integrated by adjusting the axes to show TPR versus FPR, where $FPR = 1 - TNR$.

*Additional cost and benefit consideration* An extension of our framework could include additional costs and benefits. Non-monetary factors could also be integrated by translating them into monetary values (Perni et al. 2021). The different factors could be integrated either deterministically or stochastically. Deterministic integration would assume fixed values for costs and benefits. No additional adaptations of the method would be necessary, as they would simply translate to the mean of the estimator's distribution, enabling a direct estimation of the NPV. Stochastic integration would treat these factors as random variables, accounting for uncertainty. If normality and additivity were assumed, both the deterministic and stochastic integration would not alter the framework's general functionality, as distributions for the NPV estimator could be readily derived. Both approaches would allow for a more comprehensive analysis that assesses whether the model's performance would be sufficient to achieve a positive NPV while accounting for uncertainty.

*Scenario analyses* The current version of the framework enables scenario analysis both within one project and across multiple projects at different phases. For instance, during the pre-development phase, the framework facilitates the analysis of various cost-benefit structures related to model decisions, minimum economic criteria, uncertainty thresholds, and test sample sizes. When applied across projects, the framework supports scenario analysis by deriving the hypothetical minimum model performance necessary for different projects – given identical economic and uncertainty criteria. This enables the selection of projects that can achieve the same economic outcome with the same probability but with lower model performance. As a result, the framework provides a method for prioritizing projects during both pre- and post-development, addressing a critical challenge in real-world applications (Weber et al. 2023). However, the current implementation is time-consuming. Introducing a more intuitive and direct approach to scenario analysis would enhance the framework's usefulness.

The extensions for visual decision support and additional cost-benefit considerations mentioned by the experts could enhance the scenario analysis capabilities of our framework. The visual enhancements would allow for a better risk and profit assessment across scenarios for one

application, while the integration of additional costs and benefits would provide a more thorough evaluation across ML projects.

*Other applications* Our framework was exemplarily applied and also evaluated in connection with predictive maintenance. Yet, it can be easily adapted to other binary classification use cases, like for predicting product failures in manufacturing (Frumosu et al. 2020) or credit scoring (Yotsawat et al. 2021). Furthermore, the framework could also be extended to a multi-class problem, as the resulting confusion matrix also follows a multinomial distribution. Given a cost-benefit matrix (Wang et al. 2020), while acknowledging the challenge of estimating reliable values, the method remains applicable to determine a performance boundary.

### 7.2 Contributions to Theory and Practice

Our work offers valuable contributions to both theory and practice from a conceptual as well as an operational perspective. The SLR identified a clear gap, as most articles focus on assessing developed or deployed models, lack a holistic view of costs and benefits (Drummond and Holte 2006; Holte and Drummond 2008; Hawkins 2020; Renggli et al. 2019), or neglect uncertainty considerations (Florian et al. 2021; DataRobot 2024). From a theoretical perspective, we addressed this gap by introducing the profit-per-decision estimator and deriving its theoretical distribution to link model performance, economic impact, and uncertainty. Additionally, we proposed a consistent framework for assessing project viability across the entire life cycle, particularly enhancing pre-development assessment. This extends the scarce literature (Hawkins 2020) by offering a more holistic view on costs and benefits while addressing limitations regarding uncertainty quantification through deriving a theoretical distribution rather than an empirical one. Moreover, we complement previous work on ML life cycle management (Duda et al. 2023; Kreuzberger et al. 2023; Chapman et al. 1999; Amershi et al. 2019; Studer et al. 2021), enabling a more detailed analysis of high-level phases such as business understanding.

From a conceptual yet practice-relevant perspective, the Viability Assessment Framework can serve as a blueprint for transitioning from a generic NPV model as described in Sect. 2.1, Equation (1) to a concrete cost model tailored to specific use cases. By incorporating cost and uncertainty factors, the framework enables a comprehensive assessment across pre-development, post-development, and post-deployment phases, ensuring informed decision-making throughout the ML life cycle.

An example of such an instantiation is the cost model proposed by Florian et al. (2021), which focuses on total unitary expected cost. This model accounts for yearly investment costs and decision-related costs and benefits (e.g., costs of FP), determined by PdM-specific parameters like mean time to failure, scoring frequency, and FP rejection rate. However, this approach lacks mechanisms for pre-development assessment and uncertainty consideration. If the Viability Assessment Framework were applied, the proposed method could be more effective by enabling uncertainty quantification and pre-development assessment. Therefore, beyond its practical benefits, the framework's theoretical foundations can support designing use-case-specific cost models for ML projects – particularly for the parts of the cost model influenced directly by the ML model.

From an operational and practitioner perspective, the scenario-based evaluations and interview results highlight the framework's relevance at both operational and strategic levels. Experts stress the lack of methods linking economic and ML perspectives, particularly in defining performance boundaries based on economic success criteria while considering uncertainty, underscoring the industry's need for the Viability Assessment Framework. They substantiate the framework's value in addressing this gap by generating a clear performance boundary. As business experts oftentimes identify valuable problems, our framework translates their perspective into an ML context, addressing key challenges in implementation (van Giffen and Ludwig 2023). It allows comparison of hypothetical model performance with benchmarks from experience, literature, or expert input during pre-development. Further, experts note that the reporting capabilities can increase trust in decision-making. Moreover, our framework can be used to complement other work like Renggli et al. (2023). The method evaluates whether a given dataset is sufficient to achieve the target performance. The performance boundary derived from our framework can be used as an input parameter, offering additional insights into the viability of the project from a data-centric perspective. Additionally, the results of the post-development assessment phase provide a way to assess the probability of a model leading to a viable project, simplifying reporting and justifying project decisions to stakeholders. Furthermore, we integrate the framework into the standard ML lifecycle, which offers clearer implementation guidance and fosters a more standardized, objective, and transparent decision-making process. In summary, our work contributes to the BISE community by providing a structured approach to identifying, prioritizing, and transitioning promising ML projects from development to deployment through viability assessment.

### 7.3 Limitations

Our work is subject to limitations. Limitations regarding the estimation approach and the statistical foundations are

twofold. One limitation comes from how confidence intervals are calculated. The estimator's variance is approximated by using estimated probabilities, which introduces a discrepancy between the defined and actual coverage probabilities. However, this effect diminishes as the sample size increases. An illustrative example of this limitation is provided in Online Appendix B. Another limitation of the approach is the potential for biased estimates, which is a common issue when evaluating performance using holdout sets (Raschka 2018; Kohavi 1995; Blum et al. 1999). Another well-known challenge in any estimation process is the assumption that the data – particularly the training and validation datasets in our case – faithfully represent the true underlying distribution. Our framework also encounters this issue, specifically concerning the validation and test datasets, where we assume that they reflect the actual data-generating distribution.

Limitations in the framework's applicability are largely influenced by the reliability of its input parameters, particularly the estimated costs and benefits. Here, the reliability of the results depends on the reliability of the cost-benefit matrix. Estimating costs and benefits is not an exact science. It varies by domain and remains an ongoing challenge in research and practice, as noted during the framework's evaluation. In domains such as PdM with actual costs (e.g., part costs) and benefits (e.g., reduction of downtime), the cost-benefit matrix can be estimated more easily (Prytz et al. 2015). Other domains, like cancer detection, do not have a direct link, and the values must be estimated through data, literature, or domain knowledge (Vargas-Palacios et al. 2023; Ziegelmayer et al. 2022). The same is true in cases like the one of Jiang et al. (2008), where software developers use their domain knowledge to estimate the trade-off between FP and FN errors (e.g., an FP being 20-times less desirable than an FN), which can serve as a proxy for actual cost values. However, appropriate methods for estimating cost values are necessary to realize the full potential of our framework.

Further limitations concern the evaluation of the proposed framework. While our framework offers a structured approach to assessing an ML project's viability, its evaluation remains limited by the lack of widespread implementation and empirical validation across multiple organizations. Specifically, its impact on KPIs – both objective (e.g., reduction in ML project failure rates) and subjective (e.g., improved decision confidence and perceived model reliability) – has yet to be rigorously measured. Without real-world adoption and longitudinal studies, the generalization of our results is limited. Future research should focus on benchmarking the framework against industry practices, conducting additional scenario-based evaluations, and assessing its influence on business outcomes to enhance generalizability.

## 8 Conclusion

Applications of machine learning models improve products, processes, and services, as in maintenance planning to predict maintenance activities. A key element of developing such models should be a rigorous assessment of the economic viability to ascertain whether the model's performance is sufficient to add business value. The viability assessment is particularly important when model decisions are linked to widely varying costs and benefits, as the cost of incorrect decisions can outweigh the benefits of correct ones. In maintenance planning, for example, false negatives are associated with significantly higher costs than possible benefits from true positives. Integrating viability assessment into the development process requires considering decision-specific costs, benefits, and performance estimation uncertainty. Not attending to all aspects can lead to unnecessary development costs when it can be demonstrated in advance that the project will not be viable. This also applies to incorrect deployment decisions. To address these challenges, we proposed the Viability Assessment Framework, which supports reasonable, viability-centered decision-making for development and deployment. Future studies can, for instance, investigate how small sample sizes can best be split between training/validation samples and test samples to achieve the best trade-off between learning performance and certainty in viability assessment. Moreover, deterministic or stochastic costs for development, deployment, or maintenance of the application can be integrated into the estimator. Additionally, the sensitivity of the framework concerning the values of the cost-benefit matrix may be analyzed. Further, more in-depth empirical studies can strengthen the qualitative results.

**Funding** Open Access funding enabled and organized by Projekt DEAL.

**Supplementary Information**The online version contains supplementary material available at https://doi.org/10.1007/s12599-026-00986-2.

## References


Adams WC (2015) Conducting semi-structured interviews. Handbook of Practical Program Evaluation p 492-505. https://doi.org/10.1002/9781119171386.ch19

Amershi S, Begel A, Bird C, DeLine R, Gall H, Kamar E, Nagappan N, Nushi B, Zimmermann T (2019) Software engineering for machine learning: a case study. In: IEEE/ACM 41st international conference on software engineering: software engineering in practice (ICSE-SEIP), pp 291–300, https://doi.org/10.1109/ICSE-SEIP.2019.00042

Araf I, Idri A, Chairi I (2024) Cost-sensitive learning for imbalanced medical data: a review. Artif Intell Rev 57(4):80. https://doi.org/10.1007/s10462-023-10652-8

Archer N, Ghasemzadeh F (1999) An integrated framework for project portfolio selection. Int J Project Manag 17(4):207–216. https://doi.org/10.1016/s0263-7863(98)00032-5

Baier L, Kühl N, Satzger G (2019) How to cope with change? Preserving validity of predictive services over time. In: Hawaii international conference on system sciences (HICSS-52), Hawaii, Jan 8–11, University of Hawai'i at Manoa/AIS, pp 1085–1094

Benbya H, Pachidi S, Jarvenpaa S (2021) Special issue editorial: artificial intelligence in organizations: Implications for information systems research. J Assoc Inf Syst 22(2):10

Bergstra J, Bengio Y (2012) Random search for hyper-parameter optimization. J Mach Learn Res 13:281–305

Blum A, Kalai AT, Langford J (1999) Beating the hold-out: bounds for k-fold and progressive cross-validation. In: Annual conference computational learning theory

Bukhsh ZA, Saeed A, Stipanovic I, Doree AG (2019) Predictive maintenance using tree-based classification techniques: a case of railway switches. Transp Res Part C Emerg Technol 101:35–54

Chapman P, Clinton J, Khabaza T, Reinartz T, Shearer C, Wirth R (1999) CRISP-DM: towards a standard process model for data mining. In: Proceedings of the 4th international conference on the practical application of knowledge discovery and data mining, Springer, London, pp 29–39

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Int Res 16(1):321–357

Cooper RG, McCausland T (2024) AI and new product development. Res Technol Manag 67(1):70–75. https://doi.org/10.1080/08956308.2024.2280485

Correa J, Bellavance F (2001) Parametric versus non-parametric methods in the estimation of confidence intervals for health data. Stat Methods Med Res 10(4):339–356

Costa CF, Nascimento MA (2016) IDA 2016 industrial challenge: using machine learning for predicting failures. Springer International, Cham, pp 381–386. https://doi.org/10.1007/978-3-319-46349-0_33

Darling MC, Stracuzzi DJ (2018). Toward uncertainty quantification for supervised classification. https://doi.org/10.2172/1527311

DataRobot (2024) Profit curve. https://docs.datarobot.com/en/docs/modeling/analyze-models/evaluate/roc-curve-tab/profit-curve.html. Accessed 23 Oct 2024

Dewolf N, Baets BD, Waegeman W (2022) Valid prediction intervals for regression problems. Artif Intell Rev 56(1):577–613. https://doi.org/10.1007/s10462-022-10178-5

Drummond C, Holte RC (2000) Explicitly representing expected cost: an alternative to ROC representation. In: Proceedings of the sixth ACM SIGKDD international conference on knowledge discovery and data mining, Association for Computing Machinery, NY, USA, pp 198–207. https://doi.org/10.1145/347090.347126

Drummond C, Holte RC (2006) Cost curves: an improved method for visualizing classifier performance. Mach Learn 65(1):95–130

Duda S, Hofmann P, Urbach N, Völter F, Zwickel A (2023) The impact of resource allocation on the machine learning lifecycle: bridging the gap between software engineering and management. Bus Inf Syst Eng. https://doi.org/10.1007/s12599-023-00842-7

Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the seventeenth international conference on artificial intelligence, 4–10 Aug, Seattle 1

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Florian E, Sgarbossa F, Zennaro I (2021) Machine learning-based predictive maintenance: a cost-oriented model for implementation. Int J Prod Econ 236(108):114. https://doi.org/10.1016/j.ijpe.2021.108114

Frumosu F, Khan AR, Schiøler H, Kulahci M, Zaki M, Westermann-Rasmussen P (2020) Cost-sensitive learning classification strategy for predicting product failures. Expert Syst Appl 161(113):653. https://doi.org/10.1016/j.eswa.2020.113653

Gaspars-Wieloch H (2017) Project net present value estimation under uncertainty. Central Europ J Oper Res 27(1):179–197. https://doi.org/10.1007/s10100-017-0500-0

van Giffen B, Ludwig H (2023) How siemens democratized artificial intelligence. MIS Q Exec 22:1–21. https://doi.org/10.17705/2msqe.00072

Glaser BG, Strauss AL (1967) The discovery of grounded theory: strategies for qualitative research. De Gruyter, New York

Gui C (2017) Analysis of imbalanced data set problem: the case of churn prediction for telecommunication. Artif Intell Res 6(2):93. https://doi.org/10.5430/air.v6n2p93

Hammarberg K, Kirkman M, de Lacey S (2016) Qualitative research methods: when to use them and how to judge them. Hum Reprod 31(3):498–501. https://doi.org/10.1093/humrep/dev334

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer, Heidelberg

Hawkins J (2020) Minimum viable model estimates for machine learning projects. In: Computer science/information technology (CS/IT), AIRCC, CSEA 2020, pp 37–46. https://doi.org/10.5121/csit.2020.101803

Hernández-Orallo J, Flach P, Ferri C (2013) Roc curves in cost space. Mach Learn 93(1):71–91. https://doi.org/10.1007/s10994-013-5328-9

Holte RC, Drummond C (2008) Cost-sensitive classifier evaluation using cost curves. In: Washio T, Suzuki E, Ting KM, Inokuchi A (eds) Advances in knowledge discovery and data mining. Springer, Heidelberg, pp 26–29

Huang Y, Bai B, Zhao S, Bai K, Wang F (2022) Uncertainty-aware learning against label noise on imbalanced datasets. Proceedings of the AAAI Conference on Artificial Intelligence 36(6):6960–6969. https://doi.org/10.1609/aaai.v36i6.20654

Jiang Y, Cukic B, Menzies T (2008) Cost curve evaluation of fault prediction models. In: 2008 19th international symposium on software reliability engineering (ISSRE), IEEE, pp 197–206

Jöhnk J, Weißert M, Wyrtki K (2020) Ready or not, AI comes - an interview study of organizational ai readiness factors. Bus Inf Syst Eng 63(1):5–20. https://doi.org/10.1007/s12599-020-00676-7

Kläs M, Vollmer AM (2018) Uncertainty in machine learning applications: A practice-driven classification of uncertainty. In: Safecomp workshops

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th international joint conference on artificial intelligence - volume 2, Morgan Kaufmann, San Francisco, IJCAI'95, p 1137-1143

Kreuzberger D, Kühl N, Hirschl S (2023) Machine learning operations (MLOps): overview, definition, and architecture. IEEE Access 11:31866–31879. https://doi.org/10.1109/ACCESS.2023.3262138

Lima JD, Trentin MG, Oliveira GA, Batistus DR, Setti D (2015) A systematic approach for the analysis of the economic viability of investment projects. Int J Eng Manag Econ 5:19–34. https://doi.org/10.1504/ijeme.2015.069887

Lobo J, Jiménez-Valverde A, Real R (2008) AUC: a misleading measure of the performance of predictive distribution models. Global Ecolog Biogeogr 17:145–151. https://doi.org/10.1111/J.1466-8238.2007.00358.X

Merhi MI (2023) An evaluation of the critical success factors impacting artificial intelligence implementation. Int J Inf Manag 69(102):545. https://doi.org/10.1016/j.ijinfomgt.2022.102545

Myers SC, Majluf NS (1984) Corporate financing and investment decisions when firms have information that investors do not have. J Financ Econ 13(2):187–221. https://doi.org/10.1016/0304-405x(84)90023-0

Naeem R, Kohtamäki M, Parida V (2024) Artificial intelligence enabled product service innovation: past achievements and future directions. Rev Manag Sci. https://doi.org/10.1007/s11846-024-00757-x

Perni A, Barreiro-Hurlé J, Martínez-Paz JM (2021) Contingent valuation estimates for environmental goods: validity and reliability. Ecolog Econ 189(107):144. https://doi.org/10.1016/j.ecolecon.2021.107144

Prytz R, Nowaczyk S, Rögnvaldsson T, Byttner S (2015) Predicting the need for vehicle compressor repairs using maintenance records and logged vehicle data. Eng Appl Artif Intell 41:139–150. https://doi.org/10.1016/j.engappai.2015.02.009

Raschka S (2018) Model evaluation, model selection, and algorithm selection in machine learning. arxiv:1811.12808

Renggli C, Karla B, Ding B, Liu F, Schawinski K, Wu W, Zhang C (2019) Continuous integration of machine learning models with ease.ml/ci: towards a rigorous yet practical treatment. arxiv:1903.00278

Renggli C, Rimanic L, Kolar L, Wu W, Zhang C (2023) Automatic feasibility study via data quality analysis for ML: A case-study on label noise. In: IEEE 39th International Conference on Data Engineering, Anaheim, pp. 218–231. https://doi.org/10.1109/ICDE55515.2023.00024

Rodopoulos A, Lemon M (2014) Parametric models in reliability: point and interval estimation under progressively type-i censored samples. Commun Statist Theor Methods 43(10):2079–2093

Snyder H (2019) Literature review as a research methodology: an overview and guidelines. J Bus Res 104(333):339. https://doi.org/10.1016/j.jbusres.2019.07.039

Studer S, Bui TB, Drescher C, Hanuschkin A, Winkler L, Peters S, Müller KR (2021) Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology. Mach Learn Knowl Extract 3(2):392–413. https://doi.org/10.3390/make3020020

Susanto E, Khaq ZD (2024) Enhancing customer service efficiency in start-ups with ai: A focus on personalization and cost reduction. J Manag Inform 3(2):267–281. https://doi.org/10.51903/jmi.v3i2.34

Tyralis H, Papacharalampous G (2024) A review of predictive uncertainty estimation with machine learning. Artif Intell Rev 57(4):94. https://doi.org/10.1007/s10462-023-10698-8

Vargas-Palacios A, Sharma N, Sagoo GS (2023) Cost-effectiveness requirements for implementing artificial intelligence technology in the women s UK breast cancer screening service. Nat Commun 14(1):6110. https://doi.org/10.1038/s41467-023-41754-0

Wagh SK, Andhale AA, Wagh KS, Pansare JR, Ambadekar SP, Gawande S (2024) Customer churn prediction in telecom sector using machine learning techniques. Results Control Opt 14(100):342. https://doi.org/10.1016/j.rico.2023.100342

Wang H, Kou G, Peng Y (2020) Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. J Oper Res Soc 72:923–934. https://doi.org/10.1080/01605682.2019.1705193

Weber M, Engert M, Schaffer N, Weking J, Krcmar H (2023) Organizational capabilities for AI implementation - coping with inscrutability and data dependency in ai. Inf Syst Front 25(4):1549–1569

Webster J, Watson R (2002) Webster and watson literature review. MIS Q 26

Wolfswinkel J, Furtmueller E, Wilderom C (2013) Using grounded theory as a method for rigorously reviewing literature. Europ J Inf Syst 22. https://doi.org/10.1057/ejis.2011.51

Yotsawat W, Wattuya P, Srivihok A (2021) A novel method for credit scoring based on cost-sensitive neural network ensemble. IEEE Access 9:78521–78537. https://doi.org/10.1109/ACCESS.2021.3083490

Yu W, Park E, Chang Y (2015) Comparison of paired ROC curves through a two-stage test. J Biopharm Stat 25:881–902. https://doi.org/10.1080/10543406.2014.920874

Zhou B, Liu Q (2012) A comparison study of cost-sensitive classifier evaluations. In: Zanzotto FM, Tsumoto S, Taatgen N, Yao Y (eds) Brain informatics. Springer, Heidelberg, pp 360–371

Ziegelmayer S, Graf M, Makowski M, Gawlitza J, Gassert F (2022) Cost-effectiveness of artificial intelligence support in computed tomography-based lung cancer screening. Cancers 14(7):1729. https://doi.org/10.3390/cancers14071729