WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD
EQUIS
ACCREDITED

Bachelor Thesis

Advancing the Trustworthiness of AI: An Integrated Approach to Explainability

Dragomir Balan

Subject Area: Information Business

Student Number: 12023848

Supervisor: Dr. Sabrina Kirrane

Date of Submission: 1. August 2023

Institute for Information Systems and New Media, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

Contents

1 Introduction 7

1.1 Research questions 8

1.2 Thesis methodology 8

2 State of the art 10

2.1 Interpretability and machine learning models 10

2.1.1 Definitions and concepts 10

2.1.2 Classification of the machine learning models 11

2.2 XAI methods and techniques 13

2.2.1 Post hoc methods 13

2.2.2 Evaluation criteria 15

2.3 Regulatory frameworks and guidelines 17

2.3.1 GDPR and its impact on AI 18

2.3.2 EU AI Act and its provisions for XAI 18

3 XAI framework proposition 22

3.1 Literature on standard approaches 22

3.2 Purpose of the framework 23

3.3 Design and approach 25

3.3.1 Stage 1: Project Planning 26

3.3.2 Stage 2: Resource Assessment 28

3.3.3 Stage 3: System Development 31

3.3.4 Stage 4: Understanding & Compliance 34

3.3.5 Stage 5: System Deployment 36

4 Application of framework 38

4.1 Overview of XAI in the financial domain 38

4.2 Financial use case 39

4.2.1 Stage 1: Project Planning 39

4.2.2 Stage 2: Resource Assessment 40

4.2.3 Stage 3: System Development 42

4.2.4 Stage 4: Understanding & Compliance 43

4.2.5 Stage 5: System Deployment 46

5 Limitations & improvements 48

5.1 Framework suggestions 48

5.2 General remarks 49

6 Conclusion & future work 50

6.1 Conclusion 50

6.2 Future work 51

Appendix A: Dataset Information 59

List of Figures

1 Overview of the proposed framework 25
2 Missing values per column 41
3 Loan status distribution 41
4 Correlation matrix 41
5 Personal income x Loan amount x Loan status distribution 41
6 SHAP global explanation of model behaviour 43
7 SHAP global individualized explanation of model behaviour 44
8 SHAP local explanation of model behaviour 45

List of Tables

1 Model classification based on Lipton's (2017) criteria 12
2 Summarized overview of the post hoc methods 16
3 Comparison between project management frameworks 24
4 Common phases of the analysed frameworks with our framework 26
5 Overview of the *Project Planning* stage 28
6 Overview of the *Resource Assessment* stage 31
7 Overview of the *System Development* stage 33
8 Overview of the *Understanding & Compliance* stage 35
9 Overview of the *System Deployment* stage 37
10 Performance evaluation of the developed ML models 43
11 Counterfactual explanations for a specific client 46
12 Checklist for the application of the financial case study 47
A1 Use case dataset description, adapted from Tse (2020) 59
A2 Head of the dataset, Tse (2020) 60

Abstract

The increasing adoption of AI systems in the high-risk domains, such as finance, healthcare, and transportation implies the application of opaque (black-box) machine learning models. While these models are highly performant, their trustworthiness is hindered due to difficulties in explaining their behaviour. In an attempt to advance their reliance, the legal sector implements laws that target explainability and transparency, whereas the academic sector focuses on developing approaches that facilitate understanding and compliance. As its aim, this thesis explores the literature on explainable AI (XAI), highlighting the foundational concepts and elaborating on several well-known post hoc methods. In addition, we analyse in detail the current and upcoming regulations from the European Union, underlining their provisions for trustworthy AI. In the next part of our thesis, we identify research gaps in the explainability field of project management concepts, such as CRISP-DM. To contribute to this research area, we propose our own framework centred on XAI. In order to demonstrate its effectiveness, we conduct a case study that involves applying an opaque machine learning model involving credit defaulting data. In particular, we explain its behaviour using SHAP and counterfactual explanations. As a result, we outline the benefits for the sector when applying our framework, and our findings on XAI's contribution towards trustworthy AI systems.

1 Introduction

Artificial intelligence (AI) has experienced a rapid evolution and adoption in the recent years [55]. According to Adadi and Berrada [1], domains such as finance, healthcare, legal enforcement, and transportation have centred their attention towards the integration of AI systems in their processes. In particular, Javaid et al. [48] claim that AI is of significant interest to the companies as they strive towards more efficiency, scalability, and the broader opportunities of Industry 4.0. Several companies have already observed a gradual improvement in their workflows, especially in their service and manufacturing operations [56]. As a result, many other companies are considering implementing such systems on a higher level to cut down costs and increase innovation.

As the AI field is becoming more competitive, newer and better performing machine learning (ML) models — key components of the systems — are developed and embedded into people's daily tasks, in particular generative ones[1]. Academics like Strubell et al. [85] and Angelov et al. [4] argue that with more performative models, the complexity of the upcoming decision-making systems is also increasing. That is not only from the computational resources perspective (e.g., GPU consumption), but also in terms of their algorithmic and logical structures that are difficult to be interpreted by the humans. As such, a highly debated trade-off between performance and interpretability of the models is made, however a risky and costly one — especially in the critical domains [73]. Therefore, the rising concerns and need for trustworthy AI led to the emergence of explainable artificial intelligence (XAI), a field which aims to foster concepts that can produce more transparent models and system interfaces while enabling humans to understand, trust, and manage them [39].

One of the core motivations of this thesis is to contribute to the numerous efforts from academia and regulatory institutions that attempt to demystify the AI systems. While academics are focused on technical ways transparency can be ensured for these systems, regulators are implementing laws that specifically target the explainability, such as the European Union's (EU) General Data Protection Regulation (GDPR) [28], and the future Artificial Intelligence Act (EU AI Act) [29]. However, regulators are hardly keeping up with the rush for advancement, hence the creation of laws that provision boundaries and explainability requirements for the AI systems is lacking behind[2].

Moreover, according to Zhang et al. [97, p. 3], when newer models reach an improved performance, *"they must have found some unknown 'knowledge' "* — and one method to discover it is by better understanding them. Therefore, another motivation is to highlight that explainability plays an important role not only to appeal to the performance and regulatory requirements, but it can also help humans understand how they think and interpret material differently from the machines [40]. For instance, Gunning et al. [40] highlight that in the future, explainable systems may have impactful social roles in the society, and they may accelerate cross-disciplinary discoveries and applications of AI. Therefore, Dellermann et al. [23] mention that by leveraging the novel knowledge provided by the AI systems, humans will achieve far more superior results as opposed to neglecting AI.

[1] https://hbr.org/2022/11/how-generative-ai-is-changing-creative-work

[2] https://www.bloomberg.com/news/articles/2023-03-17/chatgpt-leaves-governments-scrambling-for-ai-regulations

As a result, we are aiming to provide a framework for how European businesses should assess and best deploy their tools to the wider society. According to the *AI Index Report 2023* by Stanford University [55], after cybersecurity, regulatory compliance, and individual privacy, lack of system understanding is the next relevant risk to organisations when considering the adoption of AI — with more than a third of the respondents answering so. However, only 22% of the respondents mention that they are taking steps to tackle this concern. This is not a surprise, since according to the *AI Impact Survey* by Liebl and Klein [50], nearly 33%-50% of the 113 European companies surveyed would classify their AI product as posing an increased risk. An additional 15% are uncertain of how their systems would be categorized under the upcoming EU AI Act. These results portray that AI system providers are clashing with several issues at once, those being the (1) unclear classification of the systems, (2) uncertain approaches to ensure the understanding of the models, and (3) blurred regulations.

1.1 Research questions

The thesis aims to explore the technical and regulatory requirements of explainability for AI systems. Therefore, we develop the following main question: *"How can XAI contribute towards ensuring trustworthy AI systems?"*. This question is then complemented by three others:

1. *Why is there an increasing necessity of XAI, especially in the high stakes domains?*
2. *What do providers need to consider in terms of technical and regulatory requirements when deploying their AI systems?*
3. *What is the degree of understanding that XAI methods facilitate when applied to opaque models?*

1.2 Thesis methodology

In the initiation of the thesis, we put forward several research questions of interest. To address them, we came up with a pragmatic research design that involves mixed methods from the qualitative and quantitative fields, which should not only describe the impact of our conducted study but also demonstrate it. Consequently, this means that we first explored secondary data from well-known research databases such as ACM Digital Library[3], arXiv[4], Frontiers in Big Data[5], IEEE Xplore[6], MDPI[7], and Science Direct[8]. We began by developing an understanding of the concepts revolving around the domain of trustworthy AI. Having filtered through different studies, we focused on peer-reviewed and trusted research surveys, papers, and experiments from the specified domains — elaborating on the definitions and notions they delve in. We then continued to review the state of the art, and the EU's regulations on AI together with their implications in the high-risk sectors.

[3] https://dl.acm.org/

[4] https://arxiv.org/

[5] https://www.frontiersin.org/journals/big-data

[6] https://ieeexplore.ieee.org/Xplore/home.jsp

[7] https://www.mdpi.com/

[8] https://www.sciencedirect.com/

In the second part of our thesis, we identified areas where our study can contribute, particularly in the field of explainability in project management concepts. Therefore, we developed an XAI framework to support businesses in deploying AI systems and academics in conducting research. A case study in the financial sector was conducted to put the framework into practice by implemeting a machine learning task, and we assessed the opportunities and results. To guarantee our study's rigorousness, we cross-reviewed our steps with advice from key players in the domain, and specified where our findings align. Moreover, recommendations for interested stakeholders to leverage the framework were provided, along with a discussion on limitations and potential improvements. In our conclusion, we outline our results as well as the future work that can be conducted.

2 State of the art

This section of the thesis aims to delve into the scientific literature revolving around the topic of trustworthy AI. First, we provide a clarification on key concepts and terms that will allow us to proceed through the classification of ML models. Next, we elaborate on the motivations behind the need for more trustworthy AI, which consists of the regulatory requirements. Finally, we provide insight into the instruments available that are applied to ensure a higher degree of transparency and understanding of the AI systems.

2.1 Interpretability and machine learning models

Burkart and Huber [15] mention that humans require an understanding, or at least an explanation, for specific decisions. This also applies in connection to automated decision-making systems. Understanding how these tools operate will contribute to more transparent models, uncover potential risks, and pave the path for trustworthy AI [73]. In the literature, there are various recurring terms that describe the requirements for understanding such systems better. Additionally, there are terms that help us distinguish certain ML models from others. In this subsection on the state of the art, we introduce the following definitions, and categorise the ML models based on their degree of understanding.

2.1.1 Definitions and concepts

The domain of XAI is still in its infancy [76]. As a result, it implies extensive research and the collection of as many perspectives as possible in order to arrive to a consensus on the definitions and concepts [73]. The surveys we analysed [1][10][11][26][38][47][73] contribute differently, however all have in common the fact that they refer to the most trusted papers of the domain, which are Doshi-Velez and Kim [25], Lipton et al. [51], and Rudin [71], as well as one of the most cited surveys by Barredo Arrieta et al. [10]. Therefore, throughout this section, we equally refer to the definitions provided by these authors.

According to Lipton [51], one of the key elements of interpreting the decision-making steps of the models is transparency. Transparency portrays the comprehension of the mechanism operating behind the model. The author elaborates that it is defined by three characteristics, those being *simulatability*, *decomposability*, and *algorithmic transparency.* Thus, transparency can be considered multi-levelled — starting from the training algorithm, then the individual components, and finally the complete overview of the model.

Interpretability is theorised as the *"ability to explain or to present in understandable terms to a human"* [25, p. 2]. In the field of XAI, Barredo Arrieta et al. [10] elaborate that the interpretability of a model is attributed to its design, meaning that a model's interpretability can be quantified. In addition, Rudin [71] conceptualises interpretability as a model being intrinsically interpretable (i.e., glass-box), meaning that the models do not require explanations in order to understand their behaviour.

Explainability, on the other hand, is linked with the concept of a bridging component between humans and a decision-making agent by the provision of insights through explanations [38]. Barredo Arrieta et al. [10] mention that explainability is connected to post hoc explainability, which comprises a set of techniques that explain hardly interpretable models after training takes place.

While there seems to be an agreement across the literature on what transparency entails, the definitions of interpretability and explainability are still up for debate. Therefore, many researchers prefer to use the terms interchangeably. However, there are also academics who try to further contribute to their distinguishment [72][73]. Nevertheless, one concept is almost always implied in each of the definitions, which is understandability. As such, we decided to categorise interpretability and explainability under a common objective — facilitating a higher degree of understanding of the models. Since this thesis focuses on the topic of post hoc provision of explanations, we adopt the overarching definition of explainability proposed by Barredo Arrieta et al. [10, p. 85]: *"given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand"*. Consequently, we also consider it as our main research topic to explore.

2.1.2 Classification of the machine learning models

To ensure explainability not only in critical domains but also wherever the understanding of an AI system is beneficial and expected, we need to understand how ML models behave and what lies behind them. ML is a subfield of AI that deals with using data and algorithms to simulate human learning, continuously improving its performance based on experience[9]. Consequently, the mathematical algorithms are referred to as models, and they are the core mechanism that drives AI forward. There are different kinds of models, each varying in their performance and degree of interpretability, however on a high level, models can be classified into transparent and opaque models [11].

Referring to Lipton's [51] model characteristics, models can be identified as transparent if they are (1) simulatable, in the sense that they can be simulated and reasoned by a human as a whole, (2) decomposable, as in they can be dissected into separate parts (e.g., input, variables, algorithm) and each part can be readily interpretable, and (3) algorithmically transparent, meaning that humans can comprehend the processes followed by the models to produce an output based on input data. A model that is simulatable denotes that it is also decomposable and algorithmically transparent, though if one of the latter two elements is absent, the statement does not hold true. Therefore, according to these three criteria, examples of transparent models are linear and logistic regressions, decision trees, k-nearest neighbours, rule based learners, generalized additive models (GAMs), and Bayesian models [10].

In their research, Burrell [16, p. 4] distinguishes between three forms of opacity, mainly *"opacity as intentional corporate or state secrecy"*, *"opacity as technical illiteracy"*, and *"opacity as the way algorithms operate at the scale of application"*. In our study, we are only referring to the third form, which relates to the opacity of the ML models from an algorithmic perspective. In this regard, opaque models utilise more complex decision boundaries that result in a higher model performance [11]. However, the models' lack of transparency may pose as an impediment to analyse how a prediction is made [26]. Examples of opaque models include tree ensembles, support vector machines (SVMs), and deep neural networks (DNNs) [11]. Particularly, DNNs have recently become an industry

[9] https://www.ibm.com/topics/machine-learning

		Criteria				
		Simulatability	**Decomposability**	**Algorithmic transparency**	**Post hoc analysis**	**Post hoc methods**
Model	**Linear / Logistic regression**	Predictors are readable by the user and variable interaction is kept minimal	Strengths of association between features and labels are represented, calculations are understandable	Shape of the error surface is understandable by users	Not required	Usually feature relevance, local explanations, visual techniques
	Decision trees	Users can understand the models with little mathematical background	Model maintains readability of rules (splits), and does not alter the data	Users can easily follow the tree nodes and obtain predictions	Not required	Usually feature relevance, local explanations, simplifications
	K-nearest neighbours	Model complexity matches user comprehension	Similarities and set of variables can be decomposed and analysed separately	Mathematical tools need to be applied to understand similarity issues	Not required	Usually class maps, local explanations
	Rule based learners	Rules are not too expansive, and variables can be easily read	If decomposed into small rule chunks, can be analysed	Mathematical tools required to understand the complexity of the rules	Not required	Usually local explanations, simplifications
	General additive models	Variables and interactions are constrained to user perception	Decomposition required if interactions become too complex to be simulated	Mathematical tools required for variables and interactions complexity	Not required	Usually feature relevance, local explanations, simplifications
	Bayesian models	Statistical relationships can be comprehended by target audience	If relationships involve too many variables, needs decomposition	Mathematical tools required for complex relationships and predictors	Not required	Usually feature relevance, local explanations, simplifications
	Tree ensembles	Not applicable	Not applicable	Not applicable	Required	Usually feature relevance, local explanations, simplifications
	Support vector machines	Not applicable	Not applicable	Not applicable	Required	Usually feature relevance, local explanations, simplifications
	Deep neural networks	Not applicable	Not applicable	Not applicable	Required	Usually feature relevance, visualization techniques

Table 1: Model classification based on Lipton's (2017) criteria

favourite [97]. Zhang et al. elaborate that models like DNNs, which comprise the Deep Learning (DL) sphere, are known to show great performance on large datasets due to their complex interconnected layer structure, which is capable of processing millions of parameters [97].

At the first glance, it may appear that all transparent models satisfy the three prerequisites, however if a model, e.g., linear regression, contains too many variables and is rather dense than sparse, the model falls out of this characteristic [51]. This would not mean that the algorithm is not any more transparent by design, however its interpretability has significantly decreased. The author specifies that this also holds for models that are labelled as opaque, such as neural networks, where a single perceptron can classify as simulatable due to its comprehensibility.

Adhering to Lipton's [51] model categorisation framework, we present in Table 1 how each ML model is classified according to the three key characteristics, and which XAI methods may be applied. It is important to mention that while for transparent models the post hoc analysis is not required, certain XAI methods can still be applied to visualise in closer detail a model's decision-making process. The table is adapted and modified from Barredo Arrieta et al. [10, p. 90] and Belle and Papantonis [11, p. 7]. It represents the requirements for fulfilling the respective characteristic.

2.2 XAI methods and techniques

According to Zhang et al. [97], opaque models that deliver a higher performance, such as DNNs, are highly prone to abnormal behaviours. For instance, Eykholt et al. [31] observed in their study that a model's output can be altered due to the slightest (adversarial) change to the training data, such as a pixel on an input image of a road sign. Because of the fragile connection between a model's weights and the original problem, it is difficult to explain why DNNs behave a certain way [4]. Therefore, Burkart and Huber [15] mention that understanding how a model gets to a decision plays a crucial role, especially in processes that could lead to catastrophic results.

The need for more transparent models and the increasing strictness of the regulatory requirements called for the development of a new research trend, referred to as XAI [10]. The field has recently re-emerged with the aim to provide better *"cognitive support to users"* as they navigate through different decision-making systems [47, p. 1]. As its primary objective, XAI puts forward a collection of post hoc methods that produce more explainable models while retaining a high accuracy score and enhance humans' understanding of how to operate these systems [10].

However, Adamson [2] emphasizes that the simple provision of an explanation does not mean that one should intuitively trust a model in its complete sense. Adamson [2, p. 27] mentions there is a prominent danger that *"the success of an explanation will be measured by its ability to convince, rather than how true it is"*, which is where Barredo Arrieta et al. [10] suggest a standardization of evaluation metrics that quantify the effectiveness and level of understanding a post hoc method portrays. In the following section, we elaborate on the different types of XAI methods, as well as several evaluation criteria that measure their effectiveness.

2.2.1 Post hoc methods

Throughout the literature, there are two criteria that are attributed to ML models. The first one is referred to as intrinsic — where constraining a model's complexity can lead to an inherently interpretable model [72]. This kind of models are highly advocated by Rudin et al. [72], who demonstrate in their research how to make opaque models like DNNs intrinsically interpretable, such as through disentanglement and prototype learning approaches in the training phase.

The second criterion is explainable models, where by the use of post hoc methods, a model is explained how it achieves its results only after the training process is completed [11].

They are applied to models that are not readily interpretable, such as ensemble models or neural networks. This thesis focuses on exploring ways opaque models may be explained without the loss in performance, which is why we go in closer detail into the post hoc methods. The selection of the post hoc methods we discuss about was done in coordinance with the studies we explored, having chosen the most predominant and applied ones in the field [10][11][38][59][97]. A summarized overview of the methods, based on the aforementioned studies, may be observed in Table 2.

2.2.1.1 Model-agnostic methods

Having a closer look into the post hoc class, we can observe that it consists of two subgroups of explainability techniques. Belle and Papantonis [11] divide them into the model-specific and model-agnostic methods. The model-specific subgroup relates to techniques that produce explanations based on a model's underlying design, whereas model-agnostic methods can be adjusted to any architecture. In their research, Ribeiro et al. [69] highlight several advantages of model-agnostic methods, as opposed to model-specific. For instance, since the model-building cycle is an iterative process of testing multiple models before identifying the best performing one, it is much more feasible to compare the explanations across different architectures rather than focusing on only one. Additionally, with model-agnostic methods one is less likely to be limited to only one type of explanations [69]. For example, one may require both a textual and a visual explanation of how a model behaves. It is important to mention that the two groups are not mutually exclusive, meaning that some methods may be both model-agnostic and specific [59].

We can further distinguish the model-agnostic methods into global and local methods. Global methods relate the generalized behaviour of a model and base off their explanations on the complete distribution of the data [59]. Examples of global methods are partial dependence plots (PDPs), feature importance, and global surrogate models.

PDPs. PDPs represent the dependence of input variables over the predicted output while marginalizing across the values of the other input features [59]. For instance, PDPs can reveal whether the interaction of the predicted output and the input variables is linear or not. This proves useful when one wants to observe at which input value the model leads to an unexpected result [11].

Feature importance. Molnar [59] explains that feature importance represents the increase in error of a model once the values of a variable are permuted (i.e., shuffled). If an alternation in the values leads to no change, then the feature is classified as less important. This offers a global and highly compact overview of the variables' importance, since main feature effects and interaction effects are both considered in the calculation.

Surrogate models. The global surrogate method consists of constructing an alternative model that is easier to interpret in order to approximate its findings to the opaque algorithm. For instance, one can achieve comparable performance of SVMs with simple decision trees [59].

In contrast, local methods deal with explaining how a model behaves in a given instance that the system user would like to understand better. Examples of local methods comprise methods such as Local Interpretable Model-agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), and counterfactual explanations.

LIME. This method is similar to global surrogate models; however, the difference is that LIME aims to explain an individual data point [70]. Through dataset perturbation (i.e., adding noise), LIME can explain how a prediction changes based on a variation of the input data. Usually, the surrogate models that are used to locally approximate opaque ones are linear regressions with regularisation constraints, and decision trees.

SHAP. SHAP is a method that calculates the Shapley values, which are rooted in game theory, and attributes the payoffs (i.e., contributions) across all the input variables [53]. For instance, consider a model that results in a set of final predictions. However, the prediction for a specific data point is different than the average of total predictions. By implementing the SHAP method, we can generate explanations as to which features contributed to the different values of the respective data point in comparison to the average [59]. To note, SHAP can also be used to generate global explanations.

Counterfactuals. Counterfactual methods provide explanations to how a specific prediction changes given that the input is different [93]. Although similar to LIME, counterfactual explanations solely focus on varying the input for the respective data point, and not for the whole dataset. For instance, consider a bank that would like to know if a specific borrower would be predicted to default on a payment given a change in their salary or other personal financial circumstances [11].

2.2.1.2 Model-specific methods

Despite the wide-spread application of model-agnostic methods on opaque models, there are also model-specific methods that provide an advantage when wishing to look closer into how the inner components of a model behave, such as the activation units and hidden layers in neural networks [59]. Several methods that are most often applied are the analysis of the learned features, and gradient-based heat maps or saliency maps.

Learned features. In their book, Molnar [59] mentions that when implementing Convolutional Neural Networks (CNNs) to classification tasks of images, one can look into which features the network learned. Specifically, analysing the convolution channel (i.e., combination of neurons), rather than each convolution neuron, could lead to quicker and more optimized solutions. As a result, one can observe which features of the images were of highest significance for the classification tasks.

Gradient-based. Heat maps and saliency maps have a similar objective, albeit different visualization techniques. Barredo Arrieta et al. [10] describe them as methods that demonstrate the regions of the input data (e.g., images) that had the most significant weights when classifying. By backpropagating the gradient to the last convolutional layer, the user can observe which pixels and areas of the images were most important to the model when classifying [59].

2.2.2 Evaluation criteria

In their meta survey, Löfström et al. [54] mention that evaluating a model's performance is highly significant, since an explanation's quality depends on how the model operates

	Model-agnostic	**Model-specific**
Definition	Flexible techniques that can be adjusted to any architecture.	Techniques that produce explanations based on a model's underlying design.
Examples	**Global:** PDPs, Feature importance, Surrogate models, etc.	Learned features, Gradient-based (saliency, weights), Caption generation, Attention networks, etc.
	Local: LIME, SHAP, Counterfactuals, Rule-based learners, etc.	
Applicability	Any opaque model.	A given opaque model only.
Methodology	Relating the input of any model to its outputs.	Customised to the model.
Model comparison	Possible.	Not possible.
Explanation types	Simplification, visual, text, by example, network representation, etc.	
Evaluation criteria	Consistency, etc.	Domain knowledge, etc.
	Fidelity, stability, robustness, etc.	

Table 2: Summarized overview of the post hoc methods

and the outcomes it leads to. Known metrics that are most often applied to measure a model's performance are *precision*, *recall*, *F1-score*, and *AUROC* [66]. However, aside from the model aspect, Löfström et al. [54] additionally mention evaluation criteria that relate to the user and method aspects of an explanation. Below, we elaborate on a few of the most applied evaluation criteria regarding post hoc methods. The reason behind focusing on the following criteria is due to their widespread research and referral across the literature. However, it is important to mention that there are many other criteria mentioned in the literature, hence we refer the reader to the following research studies for a closer insight [3][19][22][54][70][80][95].

2.2.2.1 Method aspect

The criteria as part of the method aspect are classified as objective, since they are leaning on the technical side of the explanations [54]. An example of such evaluation criteria is *fidelity*, which stands for how closely an explanation method assimilates a system's opaque model [70]. Although closely intertwined with the *accuracy* metric of a model, Carvalho et al. [19] mention that *fidelity* can also be evaluated on both a global and local level. Consequently, a high *fidelity* indicates a high degree of model robustness, and vice versa.

Slack et al. [80] mention that another criterion which evaluates a method's effectiveness is *consistency*. According to Dai et al. [22], given that two different opaque models were trained on the same data and result in a similar output, an XAI method is characterized as consistent if it provides similar explanations. However, Carvalho et al. [19] points out that some opaque models may reach to a similar output through different methodologies (i.e., consider other features), therefore this evaluation metric should be applied with care.

According to Carvalho et al. [19], *stability* is an additional criterion that is of high importance to evaluate a method. Often encompassing the *identity*, *separability*, and *novelty* criteria underlined by Löfström et al. [54], *stability* refers to the comparison of explanations between similar data points for a fixed model. For example, Alvarez-Melis and Jaakkola [3] mention that if data points have similar feature values, then the explanations should also be uniform in this regard. Additionally, the authors specify that if the data points are slightly changed (i.e., perturbed) so that the model prediction stays almost the same, then the explanations should also remain similar. Therefore, while *consistency* relates to comparisons between models, *stability* relates to within-model assessment.

2.2.2.2 User aspect

Alternatively, Miller [57] characterizes the user aspect as subjective, since explanations are highly contextual. Referred by Löfström et al. [54, p. 10] as the *"user's mental model"*, the user aspect entails how a user interprets and assimilates the information from the explanations generated. In the literature, there are several criteria that are stated the most, one being *appropriate trust* [54]. According to Yang et al. [95], *appropriate trust* measures how much a user can rely on the model, given their previous experience and knowledge. By fostering this concept over time, users develop the ability to act accordingly when either a correct or incorrect recommendation is offered. Therefore, the user applies their domain insight when provided an explanation to how a model behaves.

Moreover, Miller [57] argues that humans require an explanation that does not only explain why an event simply happened, but also why it happened in comparison to other possible events. Therefore, another user aspect criterion is the *contrastiveness* of an explanation [19]. Stepin et al. [84] say that *contrastiveness* aims to measure if the explanation offers information about the factors needed to change that will lead to a different (preferred) result. Examples of such explanations are model-agnostic local counterfactuals [93].

Carvalho et al. [19] add that users' mental models have to equally be satisfied by an explanation, which is where Zhang and Chen [96] mention another criterion — *satisfaction.* The authors write that *satisfaction* stands for the extent that an intended purpose was fulfilled by the explanation provided, and a good indicator would be to observe how a user's *satisfaction* has changed. Löfström et al. [54] mention that explanation *satisfaction* can be measured through surveys, and one such example can be observed in the paper published by Dieber and Kirrane [24].

2.3 Regulatory frameworks and guidelines

In this thesis, the regulatory frameworks revolving around AI that are discussed are the EU's GDPR [28], adopted in 2018, and the upcoming EU AI Act [29]. While both legal regulations are contributing towards ensuring more trustworthy AI, they are doing so in different ways. In this subsection, we discuss the provisions of the regulations, as well as their advantages and drawbacks.

2.3.1 GDPR and its impact on AI

The introduction of the GDPR in the EU was a landmark. It is considered as one of the greatest implementations of the Union in its efforts to ensure user privacy and security across the data spectrum [36]. According to Goddard [36], the GDPR had a primary focus on data protection by design, i.e., the governance of the data and handling of it by any involved organisation. As such, the regulation laid down the basis of consent which has to be given by the parties subjected to data collection. Goddard [36, p. 2] further elaborates that consent has to be offered in a free manner, based on an *"informed and evidenced"* action. The information provided consists of extensive details about data recipients, the retention periods, and the users' varying rights [37].

The regulation additionally introduced several provisions that are related to the adoption of automated systems based on personal data [43]. Goodman and Flaxman [37] say the GDPR alluded to the fact that data subjects (i.e., system users) have the *right to explanation*, meaning that users are allowed to know the details of an automated decision taken in connection with them. The authors mention that, particularly, according to Articles 13 and 14 of the GDPR, the data controllers (i.e., companies) are supposed to provide *"meaningful information about the logic involved"* behind the system-based decision to the users. Therefore, the user could utilise the provided information to understand the algorithm and its output, as well as contest the decision should they consider that the automatic process was unfair [78].

However, Goodman and Flaxman [37] argue that the information provision requirement could take different forms. For example, *meaningful information* may be understood as technical details, such as the utilized training dataset, possible biases within the model parameters, and more specifically, the algorithmic design. Consequently, the authors are hesitant to elaborate whether the *right to explanation* is explicitly stated in the GDPR [78]. Other prominent authors in the field of GDPR, such as Ebers [27] and Wachter et al. [92], equally mention that while the EU may have emphasized on the *right to explanation* clause, it was not explicitly stated in the GDPR but only in the recitals, which are non-binding texts that reason the provision of regulations.

Selbst and Powles [78] state that *meaningful information* is considered meaningful in relation to the data subject, which is in line with their level of technical expertise and sufficient interpretation in order to exercise their rights. While a person may be shown the model behind a system, it may be hardly meaningful if it is provided alone as part of the information. Moreover, according to the authors, the *right to explanation* is believed to be embedded in Articles 13-15 of the regulation, and that this right should be functionally and flexibly used in specific contexts.

While a powerful regulation that, to this day, reinforces a sound set of laws in regard to data protection, the GDPR is still considered more vague in the context of AI – which is where the EU AI Act aims to build upon and lay the critical details [42].

2.3.2 EU AI Act and its provisions for XAI

The EU AI Act proposal is part of the European AI Strategy, which aims to ensure a human-centric and trustworthy approach towards AI[10]. The EU AI Act has several

[10] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

objectives in mind, such as developing successful AI systems in EU, ensuring AI has a positive impact in society, and facilitating a sandbox environment for system providers. In this subsection, we elaborate on the main components of the EU AI Act. However, it is important to mention that the Act is undergoing negotiations within the EU institutions, hence it is possible that specific sections we touch upon may have been amended.

2.3.2.1 Risk-based system

According to the EU AI Act, the scope of the regulation is to target all kinds of AI systems that, referring to Annex I [29], implement approaches involving ML, logic- and knowledge-based, and statistical methods. As such, the core suggestion of the proposal is the classification of the systems under a risk-based model, where the levels range from *unacceptable* and *high* to *limited* and *minimal* risk [91].

Starting with the unacceptable-risk, AI systems that undermine the safety, livelihood, and fundamental rights of citizens are completely prohibited by the Act [27]. Ebers [27] says that according to Title II of the Act, examples of unacceptable-risk systems include those that use subliminal techniques to distort a person's behaviour, assign a social score to people, biometrically identify citizens in real-time, and also exploit vulnerable groups of persons. Veale and Borgesius [91] say that, instead, a prominent focus was placed on the systems that are labelled as high-risk. Accordingly, Annex III of the Act lists eight areas where AI systems would classify as high-risk, such as biometric identification, employment, law enforcement, and migration control. The researchers highlight that some of the criteria that the providers of high-risk AI systems should fulfil involve model accuracy and robustness, documentation of the system processes, high quality data, and human oversight.

In addition to the high-risk systems, Sovrano et al. [83] mention that limited-risk systems would also have to abide by certain requirements, referred to as *transparency obligations.* According to Veale and Borgesius [91], these obligations are the disclosures that (1) a user interacts with a bot (i.e., AI system), (2) a user is being emotionally recognized and biometrically classified, and (3) a user may, by using the system, generate synthetic content that is not genuine (e.g., deepfakes).

Minimal-risk systems, on the other hand, are briefly touched upon in the Act, and hence may continue to be developed according to the existing regulations [27]. Nonetheless, Ebers [27] says that Article 69 of the Act does encourage that providers of systems other than high-risk voluntarily abide by its requirements.

2.3.2.2 Regulatory sandboxes

Borrowed as an idea from the fintech industry, regulatory sandboxes in AI are project environments where companies can innovate openly, and authorities can gain a closer insight into how they can shape future regulations [89]. In the EU AI Act, specifically Articles 53-54 [29], regulatory sandboxes are described as a place where AI system providers can optionally join and safely experiment with their technical solutions, and receive regulatory support from legal institutions before releasing their systems on the Union market[11]. Two countries known for implementing such regulatory sandboxes are

[11] https://www.eipa.eu/publications/briefing/sandboxes-for-responsible-artificial-intelligence/

the United Kingdom and Norway[12], where the latter followed a model based on the principles of the *Assessment List for Trustworthy AI (ALTAI)* [44], created as per European Commission's appointment.

According to the European Parliament's in-house think tank[13], this two-way relationship is expected to benefit both involved stakeholders, and foster a more secure AI sector. However, their study equally provides arguments for how the regulatory sandboxes may involve potential risks. For example, the misuse of sandboxes could imply that the regulators tolerate certain violations in order to attract innovators, which as a result may put consumers at risk. In addition, there could be a discrepancy between different EU Member States' regulatory sandboxes, where one is preferred than another.

2.3.2.3 Enforcement architecture

The Act proposes the instatement of an enforcement architecture which would grant system manufacturers that follow specific essential requirements the permission to offer their services on the EU market [91].

At first, as per Articles 56-58 of the EU AI Act [29], an European Artificial Intelligence Board will be established in order to *"to provide advice and assistance to the Commission"*, and act as a cooperational bridge for the Member States. In turn, each Member State will designate a national competent authority for the purpose of overseeing the enforcement of the regulation by the AI system providers. According to Smuha et al. [82], Article 59 of the EU AI Act states that the competent authority will act as a *notifying authority* and a *market surveillance authority* (MSA). Veale and Borgesius [91] say that the mission of the national competent authorities will be crucial, since the role of MSAs includes the power of acquiring information, applying penalties, and withdrawing the systems from the market.

However, when it comes to putting the requirements of the EU AI Act in motion, Veale and Borgesius [91] mention that this conformity is supposed to be carried out by the providers themselves. That is because the Act, which is closely based on the New Legislative Framework[14] regime, states that they are the best candidate to lead this procedure due to the extensive knowledge of their own product [91]. In addition, referring to explainability, Ebers [27] says that it is equally up to the providers of AI systems to make them explainable upon the request of the national competent authority.

As such, Smuha et al. [82] criticize the self-conformity approach. They mention that the regulation is lenient, since it is *"granting excessive discretion for AI providers"* [82, p. 39], and that independent bodies should instead be assessing the conformity of the systems. The authors recommend that the assessment should not be *"mere 'tick-box' exercises"* [82, p. 57], thus the Act would have to enable constructive discussions that justify the systems' compliance with the regulations between the stakeholders.

[12] https://dataethics.eu/sandbox-for-responsible-artificial-intelligence/

[13] https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2022)733544

[14] https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en

2.3.2.4 Explainability provision

In relation to explainability, Sovrano et al. [83] expand on two aims that are imposed by the Act, one being compliance-oriented explainability, and the other user-empowering.

From a compliance-oriented perspective, Sovrano et al. [83, p. 132] say that Annex IV(2) of the Act stipulates that an AI tool is required to be explainable through *"the design specifications of the system, namely the general logic of the AI system and of the algorithms"*, as well as *"the description of the system architecture explaining how software components build on or feed into each other"*. Annex IV supplies additional provisions that a system developer needs to comply before launching their product on the market [29]. According to Veale and Borgesius [91], the compliance-oriented explainability would be delivered mainly to regulatory organisations, hence the system users would have limited to no access to such details that are oftentimes trade secrets. However, this should still ensure a greater degree of explainability and security, since it is the same regulatory institutions that represent the interests of the system users [27].

Alternatively, user-empowering explainability stands for the user being eligible to use the systems in appropriate ways and interpret their outputs effectively [83]. For example, users may have complete access to the conformity assessment information, instructions for use, and the relevant standards of the system; and partial access to accuracy metrics, changes made to the system, and training dataset [91]. The provision of such explainability is of special significance when identifying how the AI systems contributed to a decision taken by a system user in a high stakes situation, thus determining the degree of liability for the parties involved [83].

An external opinion, highlighted by Patrick Grady in the following article[15], brings an interesting point of view regarding the role of explainability in the context of the EU AI Act. First, Grady mentions that the Act fails to underline if the AI systems would be required to be either intrinsically interpretable or explainable. Just like Rudin et al. [72], Grady mentions that this distinction may be of special significance for high-risk systems, since a model used in a critical situation would need to be interpretable by default rather than explained through instruments that may not always guarantee complete comprehension. However, Grady says that if the EU would impose a requirement for completely interpretable systems, that would mean the banning of using opaque algorithms and thus a stifling of AI progress. In contrast, the article mentions that there are high-risk domains where the performance requirement outweighs the need of model interpretability, such as the application of neural networks in traffic systems that ease road congestions [5]. In situations like this, Grady says that explainability would be a more preferred option instead of intrinsically interpretable models, so that the performance of the model is not affected and comprehension is still enhanced.

Panigutti et al. [64] conclude by mentioning that the EU AI Act's primary goal is to ensure that users utilize the systems appropriately. This is to be accomplished through two components: *transparency* and *human oversight*. Therefore, according to the authors, opaque models are not ruled out by the Act, hence system providers may develop their AI systems using such models.

[15] https://datainnovation.org/2022/08/the-eu-should-clarify-the-distinction-between-explainability-and-interpretability-in-the-ai-act/

3 XAI framework proposition

The following section is divided into three subsections. Firstly, we delve into the literature on the standard frameworks that are most often implemented in the industry [75]. Next, we introduce the motives and purpose of following a standardized framework that includes the discussed regulatory requirements and XAI in various business processes. Finally, we present our design methodology, the framework's aims and objectives, what differentiates it from the other frameworks, and how it may be best applied by the interested stakeholders.

3.1 Literature on standard approaches

With the rise of data accumulation and exploration, the need for robust methodologies to manage projects more efficiently has also increased [74]. In their study, Azevedo and Santos [6] mention three concepts that have an outstanding record of leading to high-quality data-oriented projects, and have been for long the industrial standard[16]. These concepts are the Cross Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Databases (KDD), and Sample, Explore, Modify, Model, Assess (SEMMA) [6]. Saltz and Krasteva [75] mention that they provide companies structured steps that are applicable to any industry, and ensure smooth processes in extracting useful knowledge from data. Therefore, we decided to focus on these three in closer detail throughout our framework conceptualization and discussion.

CRISP-DM. The CRISP-DM framework, which Schröer et al. [77, p. 1] refer to as the *"de-facto standard"*, is an iterative process consisting of six phases. In the first phase, *Business Understanding*, the business objectives are mapped and a problem statement which will proceed to be the main focus of the project is formulated. The second phase, *Data Understanding*, is where a business gets a closer insight into the data and discovers null hypotheses that will be tested. Next phase is the *Data Preparation*, which is where data is brought to a suitable state so that it can be inputted into the following algorithms. Once the data is pre-processed, the next phase is *Modelling*, where different hypothesis-testing techniques are applied to make predictions on the data. The results of the models are then assessed in the *Evaluation* phase, and an analysis is conducted to observe if the objectives have been met. Finally, the concluding phase, *Deployment*, consists of the system deployment to the client or market [20].

KDD. Fayyad et al. [33] refer to KDD as an iterative series of processes that lead to discovering useful knowledge from data. They describe it as a broader concept that also encompasses data mining, however it is different in its stages from the other methodologies. It consists of several phases, mainly *Pre-KDD*, *Selection*, *Pre-processing*, *Transformation*, *Data Mining*, and *Interpretation/evaluation* – all leading towards knowledge. The KDD framework does not include the deployment phase explicitly like CRISP-DM does, however Fayyad et al. [33] mention that as part of the *Post-KDD* stage, the acquired knowledge can be both used directly or further incorporated into different systems.

[16] https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

SEMMA. According to Azevedo & Santos [6], the SEMMA approach is much different than the other two concepts. While CRISP-DM and KDD begin with the *Business Understanding* and *Pre-KDD* stages, and end with *Deployment* and *Post-KDD* stages, respectively, SEMMA is mainly centred around the data mining aspect that is also included in the KDD framework. As stated in the name, SEMMA consists of the *Sample*, *Explore*, *Modify*, *Model*, and *Assess* phases[17], which describes a classic approach to data mining as the data scientists know and use.

3.2 Purpose of the framework

In a McKinsey article[18] about how businesses can deliver XAI, the authors recommend that organizations should not only provide explanations about how a system works, but also establish a framework that emphasizes on explainability for AI systems. According to the authors, the framework should encompass the adequate tools, AI governance, and guidelines for reviewing processes. By fostering new principles and integrating them in the work routine of developing and deploying such systems, companies will be able to trust the AI output more, and apply the knowledge from the explanations to rectify and improve parts of their operations.

However, according to Saltz et al.'s [74] study, despite the available and comprehensive frameworks, only 18% of the data scientists are implementing an approach in their projects. The authors argue that this may be due to two reasons, either the data scientists did not consider adopting such management methods, or they may not know which approach is most suitable for a specific project. Particularly, the second reason is attributed to several possibilities, such as increasing uncertainties in regard to fairness and bias, and stricter regulations around automated systems [79].

When analysing the frameworks in closer detail in Table 3, we can observe that none are explicitly touching upon post hoc explainability of opaque AI systems. Specifically, in the modelling phase of the CRISP-DM framework, it is recommended to *"report on the interpretation of the models..."* [20, p. 50]. The authors offer possible techniques to do so for opaque models, such as by providing technical information (e.g., neural network topology), and describing their behaviour (e.g., accuracy and sensitivity metrics). While the provision of information of how a model is algorithmically composed promotes greater model understanding, the topology of the opaque models hardly contributes to explaining why it resulted in a specific output [97]. Moreover, trying to describe the behaviour of a model using only the performance metrics can lead to overconfidence of the results and consequently detrimental outcomes [72].

In contrast, KDD acknowledges that the extraction of useful knowledge may be hindered by the application of more performant yet less understandable tools, such as by using neural networks instead of decision trees [33]. However, the framework does not recognize that the set of methods as part of XAI can explain opaque models without sacrificing performance. Further, while the framework does have *Interpretation/evaluation* as one of its final stages, the authors mention that this phase consists of visualizing the input

[17] http://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

[18] https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it

		Standard Frameworks			**Enhanced Frameworks**	
		CRISP-DM (Chapman et al., 2000)	**KDD** (Fayyad et al., 1996)	**SEMMA** (Azevedo & Santos, 2008)	**CRISP-ML(Q)** (Studer et al., 2021)	**Fair CRISP-DM** (Singh et al., 2022)
Concept	**Interpretability**	Referred to in the *Modelling* phase. Explicitly states that *"for opaque models, list any technical information about the model …"*.	Interpretability referred to making sense of the results through visualizations of mined patterns, and not of the opaque models' behaviour.	Does not refer to.	Mentions that the performance and interpretation trade-off does not hold in specific situations.	Does not refer to.
	Explainability	Mentioned in the *Business Understanding* phase. Recommended to list assumptions if a model must be explained.	Recognizes that neural networks (opaque) are relatively more difficult to understand than decision trees (transparent).	Does not refer to.	Referred to as a requirement and *"soft measure"* that needs to be evaluated.	Mentioned as a key component that needs to be considered in a system's design.
	Post hoc methods	Does not refer to.	Does not refer to.	Does not refer to.	Mentioned as options to explain models. However, does not elaborate on the kinds of methods.	Does not refer to.
	Regulatory requirements	Equivalent with *"Political requirements"*, and referred to as constraints in the *Modelling* phase.	Does not refer to.	Does not refer to.	Mentioned as *"Legal constraints"*. Authors also mention explainability may be increasingly demanded in specific domains due to regulations.	Mentions regulatory concerns in regard to fairness, and legal bias. Does not refer to explainability requirements.

Table 3: Comparison between project management frameworks

data and output predictions. As such, the KDD framework equally does not include the provision of explanations of the models.

The SEMMA framework is mostly oriented around the data mining part, disregarding any business understanding or deployment phase like the other two approaches [6]. Initially created for SAS' Enterprise Miner[19] data mining software, the framework still stands to be one of the most adopted approaches, however it equally does not provide guidance in taking adequate measures to facilitate explanations.

In addition to the standard frameworks, we identified two attempts at creating frameworks where the notion of explainability is integrated. The first one is a framework called CRISP-ML(Q). In this research paper [86], the authors are preserving the advantages of the standard CRISP-DM framework, while implementing quality assurance at each step of the ML processes. As an element of the *Evaluation* phase, the authors describe explainability of the models as an important component that confirms if the model's behaviour is in line with an expert's domain knowledge. In case that the explainability of a model does not reach a pre-set criteria, system providers are encouraged to revisit the previous phases (e.g., *Modelling* or *Data Preparation*).

The second framework is called Fair CRISP-DM, and the framework aims to incorporate fairness at each step of CRISP-DM [79]. The framework touches upon explainability only in the *Business Understanding* phase, suggesting that providers must address explainability issues in the AI system design. At the end of the article, the authors mention that the lack of explainability is one of the many other challenges aside fairness that hinder trust in AI. As such, they mention that the explainability concern will be included in their Fair CRISP-DM framework in additional future research.

[19] https://www.sas.com/en_us/software/enterprise-miner.html

Therefore, having identified a gap in the domain of project management approaches, and being motivated by the upcoming regulations and attempts at developing other frameworks that integrate explainability, we are proposing our own framework.

3.3 Design and approach

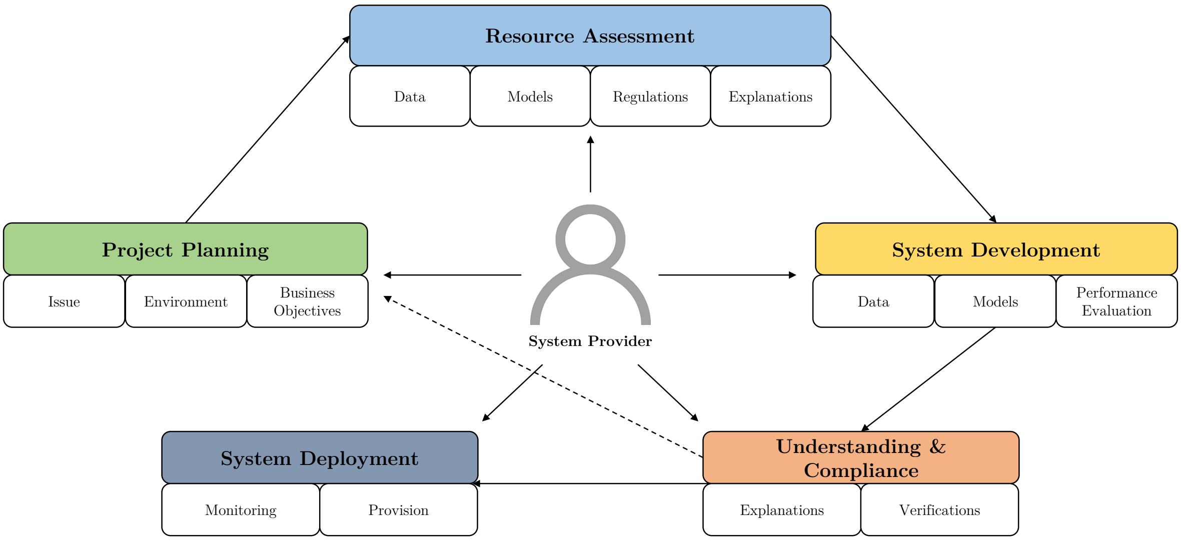

Figure 1: Overview of the proposed framework

Our proposed framework borrows ideas from the industry standard concepts, and introduces additional steps of explainability and conformity with the regulatory requirements. In our framework, we suggest a categorization of the well-known phases into five different overarching stages (Figure 1). Two propositions that are not found in the previous frameworks are the *Resource Assessment* and *Understanding & Compliance* stages. It is important to note that with our proposition, we do not suggest the complete replacement of the standard frameworks or not following them, but rather an enhancement that includes notions from the fields of trustworthy and explainable AI. As such, one may see different sections from concepts like CRISP-DM, CRISP-ML(Q), and Fair CRISP-DM overlapping with our own framework, which can be observed in closer detail in Table 4.

Each stage, and respective phase, is reinforced by general advice from the project frameworks, regulations from the legal sphere, and literature such as academic articles and guidelines revolving around XAI. *Project Planning* is the initializing stage, showing its continuation through an arrow to the *Resource Assessment* stage, and so on. The framework is structured in a circular motion, meaning that at any stage, the system provider is encouraged to return to any previous one (shown with a dashed arrow) should they need to do so or fail to meet certain requirements. It is important to note that our framework could be applied in different settings, one being for research purposes, or as the OECD [63, p. 21] says *"in the lab"*, and the second being in an industrial setting, also referred to as *"in the field"*. As such, in projects done for research purposes, stages such as *System Deployment* may not apply, whereas in an industrial setting this stage is strongly recommended to be followed.

Our XAI Framework		Standard & Enhanced Frameworks							
Stages	Phases	CRISP-DM		KDD	SEMMA	CRISP-ML(Q)		Fair CRISP-DM	
Project Planning	Issue	Business & Data Understanding	Background	Pre-KDD	-	Business & Understanding	Scope Definition	Business & Understanding	Defining Context
	Environment		-		-		-		-
	Business Objectives		Business Objectives		-		Success Criteria		Costs & Benefits
Resource Assessment	Data		Inventory Resources	Selection	Explore	Feasibility	Data Collection Quality Verification		Bias Analysis
	Models		Data Mining Plan	-	-		ML Applicability		-
	Regulations		-				Legal Constraints		-
	Explanations		-	-	-		-		Plan for Explainability
System Development	Data	Data Preparation		Preprocessing Transformation	Sample Modify	Data Preparation		-	
	Models	Modelling		Data Mining	Model	Model Training Assuring Reproducibility		-	
	Performance Evaluation	Evaluation		Interpretation/ evaluation	Assess	Evaluation	Determine Robustness	-	
Understanding & Compliance	Explanations	-		-	-		Increase Explainability	Fairness Evaluation	
	Verifications	-		-	-	-		-	
System Deployment	Monitoring	Deployment		Post-KDD	-	Monitoring		-	
	Provision				-	Deployment		Fairness Deployment	

Table 4: Common phases of the analysed frameworks with our framework

3.3.1 Stage 1: Project Planning

The *Project Planning* stage includes the phases *Issue*, *Environment*, and *Business Objectives*. In this stage, the system provider is encouraged to dedicate time, efforts, and resources to outlining the project they want to work on. In different situations, it is the case that the system provider has to consult with customers who are requesting the model to be completed for them, therefore it is important to receive as much input as possible from all stakeholders. As a deliverable, the system provider would have a clear trajectory set for the project, and be able to smoothly commence with it. While it is recommended to have a clear plan and a comprehensive understanding of the project's goals, the system provider is always welcome to return back to this stage and further elaborate. An overview of the stage is visible in Table 5.

3.3.1.1 Issue

In the *Issue* phase, the provider identifies the topic of interest that they would like to work on, and forms hypotheses around it. It is important to observe the motivation and purpose behind the issue, and consequently the ways the provider can contribute, improve, or mitigate it. In light with the previously enumerated frameworks, we can identify several similarities with this phase. Specifically, the CRISP-DM framework contains the *Business Understanding* phase, which partly overlaps with the *Issue* phase in our framework. According a Forbes article[20], in the last decade many businesses have plunged into amassing lots of data with the hope to extract valuable insights from it, however without having a clear purpose or overarching goals for doing so. Therefore, while CRISP-DM takes a business-oriented stance at almost every step of the concept, the *Issue* phase is supposed to have a deeper social meaning, oftentimes aligning with a provider's or company's mission and vision. The KDD, on the other hand, does bring attention to this requirement, mentioning that in practice, more time is required and spent on asking the appropriate questions, rather than finetuning the models to a specific objective that was not well thought out [33].

The ALTAI reiterates a need to identify purposes for the systems, which should be clearly communicated with system users [44]. In addition to that, the report generated by Information Commissioner's Office (ICO) and The Alan Turing Institute (TATI), titled *Explaining Decisions Made with AI*, recommends system providers to consider beforehand the domain they will be working in [46]. Moreover, the EU AI Act states in Article 7 that it is important to define the intended purpose behind the issue a company is willing to tackle, which will consequently contribute towards the decision of the risk classification according to the draft regulation [29].

3.3.1.2 Environment

This phase is interconnected with the first phase, since it consists of the provider transposing the issue into an environmental context. The provider should consider the stakeholders of the system, how they will interact with it, how they may be impacted by it, and how to increase acceptance and understanding of it.

In connection with the standard and enhanced frameworks, we did not observe a similar phase being incorporated. The reason we included it is because we consider this step is significant to follow, since according to Newman [60], many AI systems are not only in direct contact with the users, but also indirect — which can ultimately have different implications for the systems and society.

The XAI framework dedicated to the financial sector, proposed by van den Berg and Kuiper [90], puts forward a list of stakeholders that are involved in the processes of the AI system development and deployment. Examples are the end user, explainer (i.e., financial advisor, loan officer), AI developers, domain experts, management, and operational control. Therefore, depending on the domain and use case, in the *Environment* phase, providers should pinpoint the stakeholders of the systems and their respective role. The ALTAI additionally suggests considering if the system is designed to *"interact, guide or take decisions by human end-users that affect humans or society"* [44, p. 7]. In relation to the end user, Article 52 of the EU AI Act states that providers should design AI systems in such ways that users are aware that they interact with an AI system [29]. Therefore, these guidelines reinforced our decision to include the *Environment* phase under the *Project Planning* stage.

[20] https://www.forbes.com/sites/forbestechcouncil/2023/01/18/why-the-hoarder-mindset-is-limiting-data-value/

Stage 1: Project Planning				
Phase	**Purpose**	**Legal regulations**	**Framework advice**	**Guidelines**
1.1 Issue	Identify topic of interest to work on, listing the motivations behind it and how to contribute.	EU AI Act Article 7	CRISP-DM, KDD	AI HLEG (2020); ICO & TATI (2022)
1.2 Environment	Consider stakeholders, interactions, impacts, and future understanding of the system.	EU AI Act Article 52	van den Berg and Kuiper (2020)	AI HLEG (2020)
1.3 Business Objectives	Inquire business-related questions, introduce project success criteria, and consider non-profit and profit-oriented perspectives.	-	CRISP-DM; Fair CRISP-DM	Birkinshaw (2020)

Table 5: Overview of the *Project Planning* stage

3.3.1.3 Business Objectives

Once the environment surrounding the issue was identified, the system provider needs to reflect on the business objectives. We advise the providers to underpin the possible accomplishments and gains for their businesses, and the approaches they will take to achieve them. Additionally, it is important to consider methods that will track a project's efficiency and success, both from a non-profit and profit-oriented perspective.

The legal frameworks do not mention this as a requirement that a provider needs to follow; however it is a key component for the success of any company willing to operate in the AI field [13]. This phase is very similar to the one contained within the CRISP-DM framework, which states that companies should ask themselves business-related questions, such as how to increase marketing or sales profits given the initial intention. The framework additionally mentions the need to implement business success criteria that assess the outcome of a project. Simultaneously, Fair CRISP-DM recommends identifying the business objectives while considering possible trade-offs between accuracy, explainability, fairness, and transparency.

3.3.2 Stage 2: Resource Assessment

Part of the *Resource Assessment* stage are the *Data*, *Models*, *Regulations*, and *Explanations* phases. This stage serves the purpose for the provider to assess the conditions and stability of the resources at hand. In other frameworks, such as the KDD or SEMMA, assessing *ex ante* the resources a company will be working with is given less importance; or is briefly touched upon — such as in the CRISP-DM. We believe delving into the data governance requirements, available and required models for the project's tasks, compulsory regulations, and need for provision of explanations will offer the system provider a better understanding if the project may be proceeded with. A similar concept was identified in the CRISP-ML(Q) framework, called *Feasibility.* The subphase suggests conducting a feasibility study of the project before setting it up, by checking the *"availability, size, and quality of the training sample set"*, as well as *"legal constraints, and other requirements, such as robustness, explainability, resource demand..."* [86, p. 397]. An overview of the stage is visible in Table 6.

3.3.2.1 Data

In this phase, the system provider is encouraged to reason the purposes and methods they plan to apply when collecting their data. They should document the sources where they are collecting the data from, and if their actions comply with the regulatory requirements, such as the GDPR and EU AI Act. Best practices for cybersecurity, governance, privacy, and storage should be ensured at all times. In this phase, the provider should additionally conduct an Exploratory Data Analysis (EDA) to observe any errors or visible biases in the data, and plan how to mitigate them in the *System Development* stage.

In regard to the project management frameworks, CRISP-DM and KDD advise that for applications involving personal and sensitive data, system providers must consider privacy and legal issues, and confirm that they are allowed to use the data. Fair CRISP-DM additionally suggests using data visualization techniques to *"detect demographic disparities"* [79, p. 1536], which will consequently support the providers in taking the right course of action.

The regulatory guidelines and academic reports additionally reinforce the motive to consider the data requirements before proceeding with the development of the system. For instance, complying with GDPR's Article 12 requirements for transparent documentation and information provision, as well as Articles 16-18's rectification, erasure, and data retention requirements [28], involves high costs particularly for small businesses [21]. The upcoming EU AI Act, as stipulated in Article 10, will have even stricter data requirements in relation to training, validation, and testing datasets [29]. Everything involving processes such as data mining, labelling, aggregation, storage, and retention needs to be explicitly documented. In addition to that, ICO and TATI [46] recommend underlining the approaches that will be taken at the data collection step to improve the explainability of the systems, which proves the need to consider system explainability in advance based off the data you will be working with.

3.3.2.2 Models

The *Models* phase involves considering if the task requires the implementation of ML models. If a ML route is preferred, then the provider needs to document the methodology and design processes that are planned to be followed. Moreover, the provider should enumerate the list of models that they will test and implement, as well as state how the models classify under Lipton's [51] criteria represented in Table 1 — which would set the requirements for the *Explanations* phase later in this stage. Lastly, the provider is encouraged to conduct a literature research on studies that may be similar to their task, and do a cross-review with their plans and methodologies [86].

This phase borrows from CRISP-DM's *Modelling* phase, which, as early steps, lists the need to select appropriate modelling techniques, and generate a test design. Before building a model, providers are also told to list any model assumptions, and implement procedures which will verify a model's effectiveness and validity. The KDD framework additionally suggests considering the parameters that accompany the task and model, i.e., if the task is a classification or regression one. However, the frameworks do not mention the implications of using opaque or transparent models. Our proposed *Models* phase specifically has as its aim to focus on the type of models a provider intends to use.

The ICO report conditions that a provider needs to consider whether the project objectives lead to implementing an opaque model. They list as an example the financial sector, where the sensitivity of the task and the regulatory requirements *"dictate the need to use fully transparent and easily understandable AI decision-support systems* [46, p. 67]. In addition, the AI Now Institute mentions in their practical framework that opaque systems can be hardly subjected to accountability (e.g., due to adversarial attacks, decreased robustness), hence providers should carefully consider that the models they implement do not lead to unfavourable consequences [68]. However, according Panigutti et al. [64], the EU AI Act does not prohibit the use of opaque models, as long as the requirements listed in Article 15 for the systems to be accurate, robust, and cybersecure are met.

3.3.2.3 Regulations

While the previous two phases recommended considering the regulatory requirements at each respective process, this phase is encompassing the broader legal requirements. In this phase, the provider is encouraged to analyse the legal landscape regarding AI system development and deployment, consider which risk category their systems would classify, and additionally instate the mandatory measures such as those dictated by the GDPR.

The previous frameworks mention that analysing the legal constraints is of high significance, however neither of them goes into details of specific examples. The reason for that may be the ambiguous requirements for AI system development before recently [41]. Therefore, due to the existing GDPR and the future implementation of the EU AI Act, which will set stricter demands from providers, we felt compelled to pay closer attention to this phase. For example, referring to the GDPR's Articles 35-39, the ALTAI recommends instating measures such as a Data Protection Impact Assessment (DPIA), designate a Data Protection Officer (DPO), implement oversight mechanisms, and ensure privacy by design systems [28]. Moreover, AI system providers are advised to align their solution with relevant standards from ISO, specifically the ISO/IEC JTC 1/SC 42[21], or those proposed by IEEE[22] [44]. Additionally, once knowing the domain and which techniques will be applied in the system development process, providers will have to abide by specific requirements according to the risk categorization of their systems as indicated in the EU AI Act [29].

3.3.2.4 Explanations

The *Explanations* phase is meant to offer the provider the opportunity to consider whether they would require explaining the outcomes of their systems. If that happens to be the case, they are advised to brainstorm about possible post hoc methods that will have to be applied (e.g., surrogate models, counterfactuals, gradient-based methods).

From our analysis of the frameworks, we identified the CRISP-DM, and the Fair CRISP-DM having a similar objective as our proposed phase. CRISP-DM mentions that providers should evaluate the necessity to *"understand and describe or explain the model"* to specific stakeholders [20, p. 33]. The Fair CRISP-DM framework reiterates the importance for planning to address explainability concerns in the system's development stage, as

[21] https://www.iso.org/committee/6794475.html

[22] https://standards.ieee.org/industry-connections/ec/autonomous-systems/

Stage 2: Resource Assessment				
Phase	**Purpose**	**Legal regulations**	**Framework advice**	**Guidelines**
2.1 Data	Reason purposes and methods that will be applied for data collection. Conduct EDA to detect disparities.	GDPR Provisions, Articles 16-18; EU AI Act Article 10	CRISP-ML(Q); Fair CRISP-DM	AI HLEG (2020); ICO & TATI (2022)
2.2 Models	Consider models that will be implemented, and their degree of transparency.	-	CRISP-DM; KDD	ICO & TATI; Reisman et al., (2018)
2.3 Regulations	Examine the broader legal landscape prior to AI system development and deployment.	GDPR Provisions, Articles 35-39; EU AI Act Provisions	-	AI HLEG (2020); Reisman et al., (2018)
2.4 Explanations	Brainstorm about possible post hoc methods to be implemented for explaining the systems, if required.	GDPR Article 13; EU AI Act Provision 38	CRISP-DM; Fair CRISP-DM; van den Berg and Kuiper (2020)	ICO & TATI (2022)

Table 6: Overview of the *Resource Assessment* stage

well as planning for post hoc evaluation and auditing. Finally, the XAI framework proposed by van den Berg and Kuiper [90, p. 18], advises to *"know what type explanation you need prior to the design process"* of the system, and additionally to *"select priority explanations"* based on the domain, stakeholders, and objectives.

Referring to the regulatory requirements, the GDPR sets the basis for assessing whether explanations will be required as part of the system. Particularly, Article 13 of the regulation mentions that if the model has the intention to *profile* people, system developers must provide access to personal data for the data subject, as well as *meaningful information* about the logic involved behind the models [28]. The reason for such is further described in the EU AI Act Provision 38, which states that if AI systems are not *"sufficiently transparent, explainable and documented"*, it could lead to hampering with the EU core values on *right of defence*, and *right to fair trial* [29]. In addition to that, ICO and TATI's [46] report highlights the importance of considering whether the potential post hoc methods that will be applied to explain an opaque system are appropriate to the context of the project.

3.3.3 Stage 3: System Development

The following stage refers to the development of the AI system. It requires the complete engagement and allocation of efforts from the providers in order to deliver high-performing systems. Part of the *System Development* stage are the *Data*, *Models*, and *Performance Evaluation* phases. In our framework, this stage implements similar concepts to the other standard and enhanced frameworks, such as the data-related phases from CRISP-ML(Q). However, in relation to model-building, we pay closer attention to the condition of opacity of the models in the *Models* phase. Based on the literature, we decided to divide the evaluation of the model into performance-based evaluation, which is discussed in the *Performance Evaluation* phase, and real setting evaluation, which will be discussed in the *Verifications* phase of the next stage. An overview of the stage is visible in Table 7.

3.3.3.1 Data

Once having acquired the data, the provider should pre-process it accordingly to the project's objectives. Initially, the provider should make sure that the data is complete, error-free, and there are no missing values that could highly affect the outcome of the project. If the data needs to be transformed (i.e., cleaned, feature engineered), providers should implement data versioning approaches to maintain transparency of how each key component changed and evolved. Most importantly, they should also ensure that there is no leakage occurring from the training and testing sets while splitting for the ML task. At all times, providers should document their processes.

For most of the procedures included in this phase, we refer to the CRISP-DM *Data Understanding* and *Data Preparation* phases. The framework encompasses all the necessary steps required to the pre-processing of data for an effective extraction of insights. However, an addition that we are suggesting is versioning of the data the provider is working with. According to Article 5 of the GDPR, providers should ensure the highest level of security of the data, and that it cannot be further pre-processed to alter its state of accuracy (i.e., it remains genuine even after changes) [28]. Similarly, ICO and TATI [46, p. 50] suggest the pre-processing of data in an *"explanation-aware manner"*. Therefore, Klump et al. [49] suggest several data versioning principles that providers should adopt, which are bound to contribute towards achieving higher system transparency and explainability. In addition, the EU AI Act Provisions stress the importance of high data quality to ensure high-performing AI systems, especially when it is used for the training of ML models [29]. This highlights the significance of also considering a more data-centric rather than model-centric approach in developing systems, as advocated by AI pioneer Andrew Ng[23].

3.3.3.2 Models

The *Models* phase is a very critical part of the system creation process. In this phase, the provider is encouraged to consider if the development process requires testing multiple models before finding the best performing one, or if they may already have a specific model in mind given past experience, domain knowledge, or research purpose. If the former happens to be the case, it is recommended to start with the baseline models (i.e., transparent), and constantly reassess the performance of the models. While indeed, a model needs to meet the regulatory requirements, it should also meet performance expectations and reach a task's goals. Therefore, the usage of opaque models is admissible if the models are explained. In addition, providers are instructed to make their systems reproducible, as well as document the architecture behind the system, in case that the client or authorized individuals would require to replicate the processes applied.

Our approach includes similar concepts to the *Modelling* phase of the CRISP-DM, CRISP-ML(Q), and Microsoft's The Data Science Process (TDSP)[24] frameworks. Throughout the model-building process, providers are supposed to incorporate steps such as hyperparameter optimization, cross-validation, and regularisation. The provider should pay additional attention if concepts such as transfer learning are used, and how the pre-trained models and data could interfere with the outcome of the task [86]. In terms of

[23] https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence

[24] https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview

Stage 3: System Development				
Phase	**Purpose**	**Legal regulations**	**Framework advice**	**Guidelines**
3.1 Data	Pre-process the data according to objectives and regulations, implement data versioning techniques, and document each process to ensure transparency.	GDPR Article 5; EU AI Act Provisions	CRISP-DM	ICO & TATI (2022); Klump et al. (2021)
3.2 Models	Start with baseline (transparent) models and reiterate through hyperparameter tuning, cross-validation, and regularisation techniques.	EU AI Act Articles 14-15	CRISP-DM, CRISP-ML(Q), TDSP	-
3.3 Performance Evaluation	Evaluate the technical performance of the model and observe metrics. Test model's reaction to noise and guided adversarial attacks.	EU AI Act Article 15	CRISP-DM; CRISP-ML(Q)	AI HLEG (2020); ICO & TATI (2022)

Table 7: Overview of the *System Development* stage

the regulations, we did not identify specific legal advice regarding the modelling phase, other than the EU AI Act's conditions in Articles 14-15 to fulfil the human oversight, robustness, and accuracy criteria [29]. This proves that there is still a liberty for the creators to *"develop their models following a transparent-by-design approach or by using XAI techniques"* — provided that the users can use the systems accordingly [64, p. 1144].

3.3.3.3 Performance Evaluation

The third phase from the *System Development* stage is *Performance Evaluation*, which deals with the assessment of how the models performed from a technical perspective. In this phase, the provider should consider which metrics they will evaluate based on the selected models. In addition, this phase should act as a testing ground for the providers to particularly evaluate an opaque model's reaction to noise, and potential adversarial attacks that could lead to detrimental outcomes. In case that a model underperforms and does not fulfil the initially set objectives, providers should return to the previous phases from this stage and pinpoint the affecting criteria.

This phase is similar to the *Assess Model* subphase from the CRISP-DM framework. It advises to summarise results of the precedent task, report on the values of the specific criteria, and rank the models in terms of their quality [20]. Since our framework is also an iterative process like the other standard approaches, we equally suggest the revision of parameter settings and model specificities. Moreover, considering the CRISP-ML(Q) framework, we suggest the incorporation of guided adversarial tests specifically for opaque models, such as DDNs [65]. This would help validate the robustness of a model, which would contribute to its error-free performance in a real setting. For conducting such adversarial attacks, we refer to the ART[25] package created by Nicolae et al. [61], and to additional research from Ballet et al. [8] and Cartella et al. [18].

Concurrently, the *Performance Evaluation* phase is supported by the legal literature. The EU AI Act has the dedicated Article 15, which stipulates in the second paragraph that the *"relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use"*. ALTAI provides a closer insight into the relevant metrics like

[25] https://adversarial-robustness-toolbox.readthedocs.io/en/latest/

"false positives, false negatives, F1 score..." [44, p. 10], and ICO and TATI [46] additionally recommend observing metrics like *precision*, *sensitivity*, and *specificity.* Having a well-performing and validated model will ensure easier extraction of explanations in the *Understanding & Compliance* stage, and lawful compliance with the regulatory requirements.

3.3.4 Stage 4: Understanding & Compliance

The following is the penultimate stage before a system's final deployment. It comprises two phases, the *Explanations*, and the *Verifications* phase. As we observed in the literature research, several authors such as Ebers [27], Sovrano et al. [83], and Gyevnar and Ferguson [41] highlight the importance of enhancing system explainability in relation to the legal environment. To our knowledge, there are no stages or phases similar to the ones contained within the well-known project management concepts. This absence underlines the evolution of how providers are mandated to construct the AI systems, and an urgent necessity for coming up with a framework that includes these guiding principles. An overview of the stage is visible in Table 8.

3.3.4.1 Explanations

In this phase of the *Understanding & Compliance* stage, the provider should assess the degree of interpretability of their systems at the current state. The model should undergo revision from involved stakeholders, and its initial understanding be evaluated with small groups of potential external users (e.g., focus groups). If understanding is not guaranteed due to model composition, then providers should apply post hoc XAI methods that were identified earlier in the *Resource Assessment* stage. When doing so, providers should consider the specific evaluation criteria and prerequisites of explanations, their target audience, and respective domains. By doing so, providers would ensure that users can comprehend and use the system appropriately, and consequently comply with the user-empowering aspect of explanation provision.

The standard frameworks do not explicitly suggest the explanation of systems and provision of post hoc methods to ensure system understandability. Our only discoveries were the enhanced frameworks CRISP-ML(Q) and TDSP. The former mentions that providers should increase explainability of the systems for end users, especially as *"explainability of a model helps to find errors and allows strategies"* [86, p. 403], whereas the latter recommends explaining the entire model behaviour, such as by using Microsoft's proprietary explanation package[26] for Python that includes methods like feature importance, PDPs, and counterfactuals.

The GDPR states in Provision 71 that if a person is subjected to profiling by a decision-making system, they have the right to *"obtain an explanation of the decision reached"*, which is further reinforced in Article 13 of the regulation [28]. Moreover, in Annex IV of the EU AI Act, it is mentioned that providers need to supply information about the *"general logic of the AI system and of the algorithms"*, the *"main classification choices"*, and the *"decisions about any possible trade-off made regarding the technical solutions adopted"* [29]. Authors like van den Berg and Kuiper [90] additionally recommend in their

[26] https://learn.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability-aml?view=azureml-api-1

Stage 4: Understanding & Compliance				
Phase	Purpose	Legal regulations	Framework advice	Guidelines
4.1 Explanations	Implement post hoc methods and provide explanations of model behaviour to enhance system understanding and compliance.	GDPR Provision 71, Article 13; EU AI Act Annex IV	CRISP-ML(Q); TDSP	van den Berg and Kuiper (2020)
4.2 Verifications	Engage in verifications with authorities through regulatory sandboxes. Employ third-party auditors and implement feedback.	EU AI Act Articles 9(7), 23, 53	OECD (2022)	ICO & TATI (2022); Reisman et al. (2018)

Table 8: Overview of the *Understanding & Compliance* stage

framework to consider contextual factors in delivering an explanation, specifically the domain, impact on stakeholder, data used, urgency of decision, and audience factors. For such, the authors list examples of post hoc methods and approaches that are mentioned in Barredo Arrieta et al. [10].

3.3.4.2 Verifications

This phase has the objective of presenting the provider with regulatory advice in order to understand whether their systems comply with the legal requirements. Usually in a regulatory sandbox, the provider would consult with the competent authorities if the compliance-oriented explainability is achieved. Simultaneously, providers can employ third-party auditors that will provide the necessary feedback on which criteria were not met, and monitor where the provider needs to deliver improved explanations.

Having assessed the standard frameworks, we did not observe this phase being mentioned under any form. However, we identified a different enhanced framework which is similar in our purpose, developed by the OECD [63]. The framework specifies the *Verify* phase, part of the *AI Model* section, which takes place before a model's deployment. In this phase, the authors of the framework emphasize on model inferencing — a step of the process where system providers should guarantee transparency and explainability to the stakeholders affected by the model. During this step, stakeholders such as data scientists, model engineers, and governance experts have to be involved.

Additionally, in their framework, Reisman et al. [68] mention that allowing an immediate access to researchers and auditors to review the systems is a very important prerequisite to system development. ICO and TATI [46] recommend the same course of action while referring to ICO's [45] comprehensive auditing framework. Since the EU AI Act has elaborated on regulatory sandboxes in Article 53, auditing and professional system reviewing is possible to be achieved in a secure and private manner under authority supervision [29]. Upon request from the competent authorities, providers are encouraged to cooperate by proving that their system is in line with the conformity demands, as per Article 23. Moreover, the EU AI Act states in Article 9(7), that a high-risk system needs to pass a testing prior to its release on the market [29], which is in line with Smuha et al.'s [82] paper, pointing to the fact that this assessment requires extensive dialogues and verifications with the authorities whether the legal conditions have been successfully met.

3.3.5 Stage 5: System Deployment

This stage is the final component of our framework. It deals with the finalisation of the system, and the provision of the project to the end target. However, even when the system is deployed, providers are expected to set in place monitoring mechanisms that continuously evaluate the performance of the model, and observe if it decays over time. Due to rapid technological advancements, it may be the situation that a system becomes outdated, which would require the redevelopment of the system to meet the standard requirements [68]. Moreover, according to Article 65(5) of the EU AI Act, after a system is deployed, market surveillance authorities reserve the right to issue penalties, or go as far as to withdraw the system from the market for non-compliance[27]. Therefore, we included two phases as part of this framework, *Monitoring* and *Provision*. An overview of the stage is visible in Table 9.

3.3.5.1 Monitoring

The *Monitoring* phase requires the implementation of monitoring mechanisms that will observe the usage of a provider's system, where it underperforms and needs more attention, and the constant (changing) feedback from the stakeholders. If a monitoring technique detects any anomalies within the system, the competent stakeholders should be immediately informed. Given that a provider is issued a warning and is advised to retrain their model, they should ensure that these objectives are fulfilled in a timely and adequate manner. If a model is retrained and readjusted to meet the updated criteria, providers are supposed to go over again through the *Performance Evaluation*, *Explanations*, and *Verifications* phases in order to ensure compliance.

Our suggested phase assimilates the ones contained within the CRISP-DM and CRISP-ML(Q) frameworks, where the provider is advised to come up with a comprehensive monitoring and maintenance plan before its final implementation. Once the system is put into production and the monitoring system identifies an issue, providers are advised not to necessarily retrain the model from ground-up, but instead *"fine-tune the existing model to new data"* [86, p. 406].

The regulatory guidelines highlight the need to *"monitor the impacts that the use of an AI system and its decisions has or may have on an individual, and on wider society"* [46, p. 21]. Moreover, the EU AI Act reiterates the requirement for creating a post-market monitoring plan, by mentioning in Article 61 that the system *"shall actively and systematically collect, document and analyse relevant data provided by users"* [29]. The GDPR has similar conditions in relation to data compliance in Article 41, where the authorities will monitor a provider's compliance to the GDPR's statute and review their operations [28].

3.3.5.2 Provision

This phase proves that the provider has met the technical and regulatory requirements for deploying their system. They may now deploy their system on the market, given that they provide continuous assistance and support where mandated by the regulations and required by the users. In this phase the provider, or competent stakeholder, should observe the system across its lifespan and retract the system once its purpose shall no longer be pursued.

[27] https://techcrunch.com/2022/04/01/ai-act-powers/

Stage 5: System Deployment				
Phase	**Purpose**	**Legal regulations**	**Framework advice**	**Guidelines**
5.1 Monitoring	Monitor how a system performs, and retrain or mitigate issues where required.	GDPR Article 41; EU AI Act Article 61	CRISP-DM; CRISP-ML(Q)	ICO & TATI (2022)
5.2 Provision	Release the system on the market while still providing continuous support and lifespan monitoring.	GDPR Article 17	IBM AI Model Lifecycle Management	Ginart et al. (2019); ISO/IEC 22989

Table 9: Overview of the *System Deployment* stage

The standard and enhanced frameworks contain this phase under the *Deployment* umbrella; however, they do not list a very specific provision laid down by the GDPR in Article 17, called the *right to be forgotten* [28]. In the GDPR, the *right to be forgotten* entails that the user may withdraw their consent to the processing of their data, thus the provider is compelled to guarantee that the respective user's data is not further handled. This brings attention to the fact that the provider must ensure a dynamic AI system, for which we refer to the paper published by Ginart et al. [35] and IBM's *Deploy* and *Monitor* phases from the AI Model Lifecycle Management[28] framework. In relation to system withdrawal from the market at the end of its mission, we refer to the ISO/IEC 22989 standard, which represents the AI System life cycle processes [94]. As part of the standard, the retirement step contains the data disposal, model disposal, and the decommissioning of the system processes, which are to be performed at the end of a system's running.

[28] https://www.ibm.com/cloud/blog/ai-model-lifecycle-management-overview

4 Application of framework

In this section of the thesis, we demonstrate an academic application of our proposed framework. Firstly, we elaborate on the employment of XAI in a high-risk domain, finance, by describing several studies from the literature sphere that have applied post hoc methods to explain the AI systems. Next, we show the approaches that we implemented on data related to credit defaulting, while following our proposed framework phases.

4.1 Overview of XAI in the financial domain

Having explanations that support a model's decision is of high significance, and that is continuously reiterated by economic supervisors, such as the German Federal Financial Supervisory Authority, BaFin. In their Big Data and Artificial Intelligence Principles document, BaFin insists how the decision of implementing a model should be explicitly documented [7]. Companies are supposed to observe how an implemented model's accuracy relates to its interpretability — and to deduce at which expense a trade-off may occur.

In addition, key players in the finance domain have also turned to the idea of implementing explainability in order to facilitate a higher understanding of opaque models, such as FICO — the company that invented the FICO credit score used by most USA lenders today[29]. In their research paper, Fahner [32] implemented post hoc methods such as PDPs to illustrate the effects of features part of a stochastic gradient boosting model. Similarly, the Bank of England's staff working paper, written by Bracke et al. [14], demonstrates an explainable approach to predicting mortgage defaults. The authors leveraged a Shapley-based and a surrogate-like method to explain the outcomes of a plausible situation that may lead to mortgage defaults.

From the academic sphere, studies like Misheva et al. [58] aim to demonstrate the advantages of applying post hoc explainable methods to better interpret opaque models' output. To elaborate, the authors implemented two post hoc model-agnostic methods, mainly LIME and SHAP, to a credit scoring task. With LIME explaining locally, and SHAP globally, the authors achieved a higher degree of understanding of how variables like loan amounts, payment amounts, and recoveries affect the defaulting probability of a borrower. A different application by Bussmann et al. [17] in credit scoring for P2P lending, using SHAP for local explanations, demonstrated a tailored description of most important variables that determined the defaulting probability of four companies.

Alternatively, an intrinsic method to increase system transparency in the financial sector is the implementation of inherently interpretable models [87]. Like Rudin et al. [72], Sudjianto and Zhang [87] advocate for the imposition of architecture constraints, such as *orthogonality*, *smoothness*, and *sparsity*. According to the authors, through constraints, models can achieve certain pre-set maximum requirements that guarantee their inherent interpretability. The authors achieved exactly that by unwrapping DNNs into a set of local linear models, which are more transparent, with little performance loss. As a result, they were able to predict the probability of borrowers defaulting on home loans by successfully identifying unique attributes and classes to them. Their interpretable model also highlighted variables of significance like delinquency, FICO, and loan to value ratio.

[29] https://www.fico.com/blogs/fico-score-research-explainable-ai-credit-scoring

4.2 Financial use case

In this subsection, we demonstrate the application of our framework on a credit defaulting use case. To enhance model understanding, we apply post hoc XAI methods and offer explanations as to why the model predicted a certain outcome. In order to rigorously audit our system, we implement Oxford researchers Floridi et al.'s [34] conformity assessment procedure (capAI) for AI systems, which is meant to act as an assessment tool for providers to check if their systems are in line with the upcoming EU AI Act. For the purpose of this application, we also developed a checklist which helps to understand our conformity with our framework and guidelines (Table 12). In addition, to observe in closer detail our code, we refer the reader to the notebook hosted on our repository[30].

4.2.1 Stage 1: Project Planning

Issue. Initially, we had to identify a domain and topic of interest that we wanted to discuss. Given our economic and financial expertise, as well as interest in the following area, we decided to focus on the financial sector. Accordingly, we identified that one of the most applied ML tasks in the field is the assessment of creditworthiness [62]. Our issue, thus, was to predict the likelihood of a person being classified as defaulting on their loan. In the *AI in Finance* section of their book, the OECD [62] elaborates that high risks arise around AI-based systems due to the lack of explainability in relation with their outcomes. As such, we set it as our mission to produce opaque ML models that we would explain with the help of post hoc methods. Therefore, according to the capAI assessment tool, we clearly defined our set of values and communicated them [34].

Environment. In this phase, we identified several stakeholders of the system. The first group of stakeholders is us, the authors, who are undertaking several roles — the system developer, domain expert, and explainer. The second group is the imaginable clients whom we deliver our explanations. The system we executed has one aim — to be utilized solely by the providers. This means that if this system had been implemented in a real setting, the clients would not be able to directly interact with it, and only the competent stakeholders may leverage it (i.e., a bank's employees). This concretely refers to Newman's [60] emphasis on indirect systems, where even though the clients do not directly interact with the AI system, the outcomes may affect their ability to borrow money from the bank.

As per capAI [34], we managed to identify the stakeholders, the kind of interaction between them and the system, and the impact as a result of the contact. In addition, we are in line with the EU AI Act's requirements for environmental considerations.

Business objectives. We set it as our goal to achieve a high performing and explainable model, which should fulfil capAI's goals and metrics requirements [34]. Given that the model we developed could be applied in an industrial setting, we also considered the OECD's [63] criteria for this phase, such as the industrial sector, business function, and business model for our system. For instance, this AI system would be used in the fintech industry, where a bank may use it to identify and minimize credit-related risks. The business function would be to successfully avert offering the loan to customers who may likely default, and to offer the loan to those that may have been considered as defaulting

[30] https://github.com/dragomiru/BachelorThesis

but instead may not be. This approach should benefit both the customers and the system users.

4.2.2 Stage 2: Resource Assessment

Data. For this case study, we decided to explore the internet for open datasets. We accessed the platform Kaggle[31] and looked for datasets that are related to credit defaulting, which is where we identified the dataset by Tse [88]. In order to comply with the regulatory requirements and ensure that we can use the data, we looked at the license type, which is listed under the *CC0: Public Domain*. This means that the provider has dedicated the dataset to the public domain, and we can modify and perform work on it without the need for permission[32].

After making sure that we are allowed to use the dataset, we observed the dataset description, where the author mentions that it contains information simulating credit bureau data. This means that the dataset is synthetic, and the data points approximate real situations. Synthetic data is known for being a robust substitution to real data due to the accurate simulation of factual data points in a regulatory-compliant manner [67]. In addition to that, we observed the metadata, where the features related to loan classification are listed. For further information regarding the dataset features, we refer the reader to Appendix A.

To follow the advice of our framework, we conducted an EDA phase, and observed the data distribution through histograms, scatter-plots, and correlation plots (Figures 2-5). As such, we realized that the dataset contains missing values for a couple of variables, and planned that in the *System Development* stage, we will mitigate the issue by imputing them with the median value. Moreover, we observed the distribution of the data, and identified several outliers which we also had to take care of in the following stage. However, the most important observation we discovered was that our target class *loan_ status* was highly imbalanced, with nearly 80% of the data points being classified as non-defaulting. This issue had to be alleviated so that our models would perform well and that it would not interfere with our model explainability, which is why we considered that we will either apply an under- or oversampling technique. Lastly, we also identified that the dataset contains several categorical features, which we intended to dummy-code so that they can be included in our model-building part.

In relation with our framework, we reasoned and documented each process of our data acquirement and exploration, which should serve as proof for compliance with the GDPR and EU AI Act regarding transparency requirements [34]. However, we would have to investigate further whether the synthetic data provided publicly by Tse [88] is also completely anonymous (i.e., it does not contain traces of the data it was trained on).

Models. When it comes to the *Models* phase of the *Resource Assessment* stage, we identified that a ML route is preferred to be followed. Having looked at our task, we understood that we will implement classification models. Since the data is in a tabular format and models outside of the DL sphere tend to perform equally well, we specifically focused on linear and tree-based models. The models we had in mind were the logistic

[31] https://www.kaggle.com/

[32] https://creativecommons.org/publicdomain/zero/1.0/

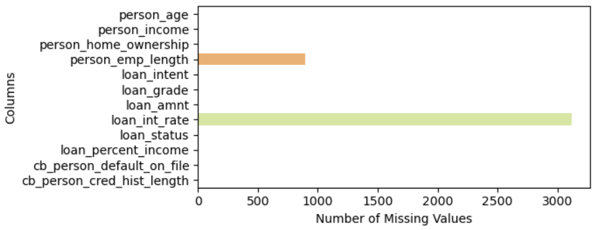

Figure 2: Missing values per column

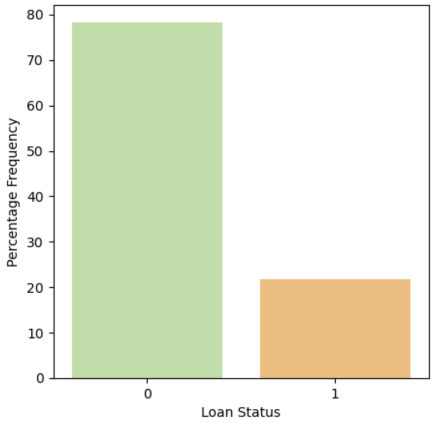

Figure 3: Loan status distribution

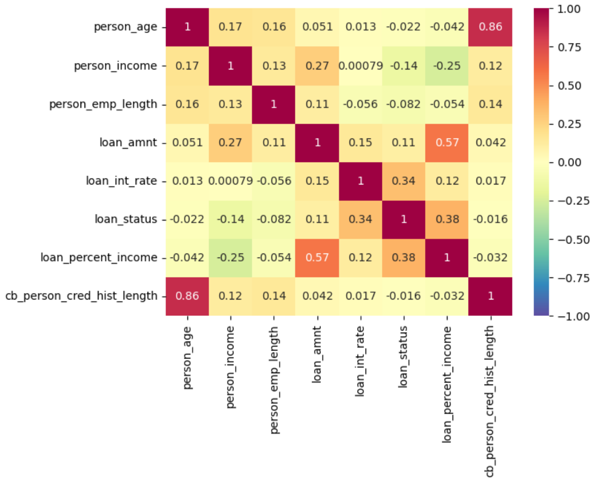

Figure 4: Correlation matrix

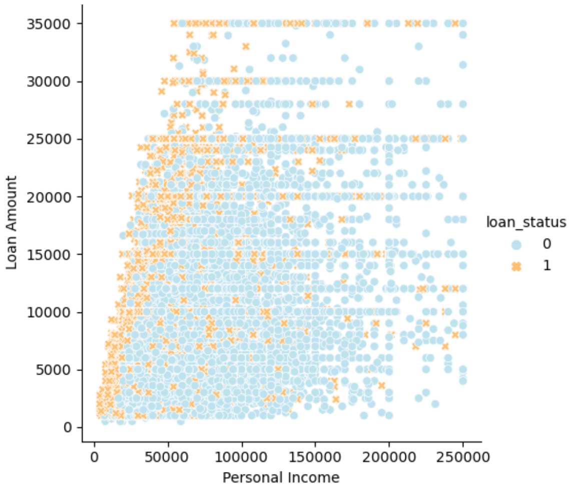

Figure 5: Personal income x Loan amount x Loan status distribution

regression, random forest, and gradient boosting decision trees (specifically from the LightGBM[33] package). According to the criteria listed in Table 1, the first model is classified as transparent, whereas the other two are opaque, which indicated the necessity to consider post hoc methods in the *Understanding & Compliance* stage. In this regard, we believe that our actions are compatible with the framework, since they are reinforced by the XAI literature in the finance domain which we elaborated in subsection 4.1 of our thesis.

Regulations. In this phase, we considered the broader regulations of the GDPR and the EU AI Act. In regards to the EU AI Act, we envision that if this system would be released on the market, it could be classified as a high-risk system as indicated in Provision 37 of the proposal [29]. Referring to the GDPR, having conducted this project in an academic setting for research purposes, we did not consult with the overseeing authorities. However, we abided by the data requirements as stipulated by the regulation, which is a successful compliance according to the capAI [34].

Explanations. Part of this phase, we reflected on our domain knowledge and personal

[33] https://github.com/microsoft/LightGBM

expertise from previous conducted projects. Since we are familiar with game theoretic concepts, we decided that we will apply the SHAP method on our opaque models to better understand model behaviour. Next, referring to Miller [57], we considered that providing a mere explanation of why a potential client was predicted to default would not suffice, which is why we decided to apply counterfactual methods that demonstrate a contrastive situation for the said client. For this purpose, we discovered the DiCE[34] package.

We believe that the methods we selected would provide sufficient and satisfying insight for the potential client, as well as for the loan provider. As a result of our explanations, the client may utilize the *meaningful information* to contest the decision in a legal setting as per GDPR's Article 13 [28].

4.2.3 Stage 3: System Development

Data. Having assessed the data resource in the previous stage, we commenced with the pre-processing of the data. In this phase, we removed the outliers and duplicated data points, and dummy-coded the categoric features. Then, we had to make sure that the data is complete and there are no missing values that would interfere with our model-building phase. As such, we split the dataset into the training and testing set, with 80% and 20% cohorts, respectively. The reason we did it before remedying the null values and imbalanced issues is to ensure that there is no leakage from the training set into the testing set, which is kept away from the training of the model. We, then, imputed the missing values from the specific columns using the median, and oversampled the lower class (i.e., loan_status = 1).

According to capAI, without a qualified reason, our imputation process may have led to propagating more bias into the data [34]. Moreover, the median technique may not be the most optimal choice in our situation, hence studies like Bennett [12] recommend hot-deck or regression imputation approaches. In addition, we did not find implementing a data versioning technique a necessity for this academic case study. That is because we did not perform substantial feature engineering or pre-processing that would highly alter the state of the data. However, we would consider this important step in a real setting in order to ensure greater data reproducibility.

Models. We initiated the *Models* phase by performing the transparent logistic regression model. To ensure that we achieve the best performance by the respective algorithm, we implemented a grid search hyper-parameter optimization technique, and adjusted it according to the increasing performance. In addition, we performed a 5-fold cross-validation in order to mitigate the possibility of overfitting our model on new data. While the logistic regression has resulted in satisfactory results and performance, we continued with training a random forest and a LightGBM model. This process was highly iterative and required testing multiple combinations of parameters in order to identify the most appropriate ones. At each stage that involved randomisation, we made sure that the processes could be replicated. We also added comments and documented each line accordingly to inform the reader where necessary. According to capAI's *Development* stage, we conformed to each specified requirement [34].

[34] https://interpret.ml/DiCE/

Performance Evaluation. To test which model performs best, we analysed several performance indicators. Firstly, after a model finished training, we analysed the confusion matrix of the respective model, and the corresponding *precision*, *recall*, *F1-score*, and *AUROC* values. In this setting, false negatives (i.e., a client is predicted that they would not default but in truth they do) are costly, however falsely rejecting a client who would not have defaulted is also in a lender's interest, hence false positives equally play an important role. Therefore, the metrics which are most important for us to consider are the *F1-score* and *AUROC*. As such, we identified that the LightGBM model is the best performing model out of the three, which happens to be opaque. The results for the following performance metrics can be observed in Table 10.

		Performance Metric			
		Precision	**Recall**	**F1-score**	**AUC Score**
Model	**Logistic Reg.**	0.845748	0.809472	0.820242	0.871822
	Random Forest	0.880217	0.878895	0.879512	0.902146
	LightGBM	0.938685	0.936748	0.933467	0.941018

Table 10: Performance evaluation of the developed ML models

In this phase, we managed to categorize the models, identify and analyse the most important evaluation metrics, and observe the best performing model from the ones we developed [34].

4.2.4 Stage 4: Understanding & Compliance

Explanations. To explain the model behaviour, we applied the post hoc methods we considered in the *Resource Assessment* stage: the model-agnostic SHAP and counterfactuals. With SHAP, we looked at the global explanation of the model, which shows the attribution based on the average effect of the individualized attributions (i.e., each data point's contribution in an ensemble) [52]. This method is known to represent the feature importance far more accurately than Gini importance or permutation methods. We may observe in Figure 6 the most important features contributing to a model's predictions.

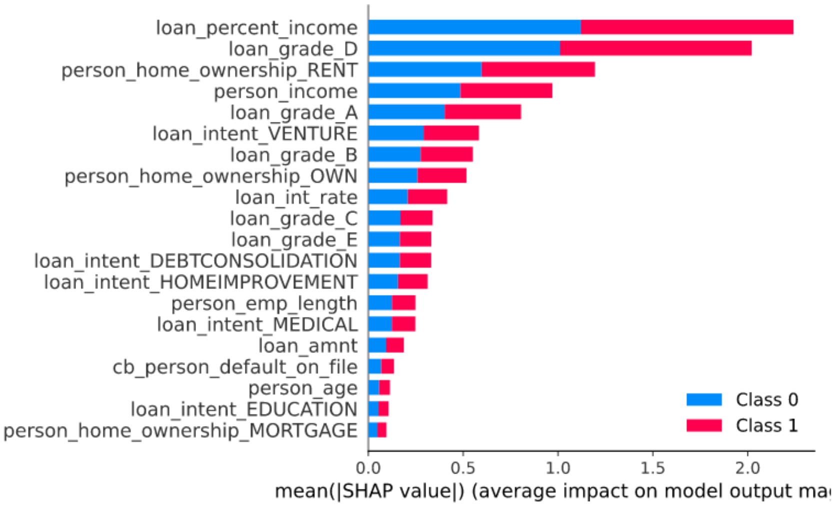

Figure 6: SHAP global explanation of model behaviour

To observe each individualized contribution, we also generated a global individualized plot (Figure 7). In the individualized plot, the most significant variables contributing to the model are sorted in descending order of importance just like in the global summary plot, starting with the most important ones. Moreover, each data point is represented as a point on the x-axis of the respective variable, and a collection of data points around the same SHAP value are piled on top of each other and represented as a swarm. In addition to that, if a data point is coloured red, then it means that this data point is higher on the continuous or binary scale of the respective variable, and vice versa if it is coloured in blue. Lastly, if a data point has a negative SHAP value, then it means that this variable had a negative impact on the probability that a client is predicted as defaulting, and vice versa if it has a positive SHAP value.

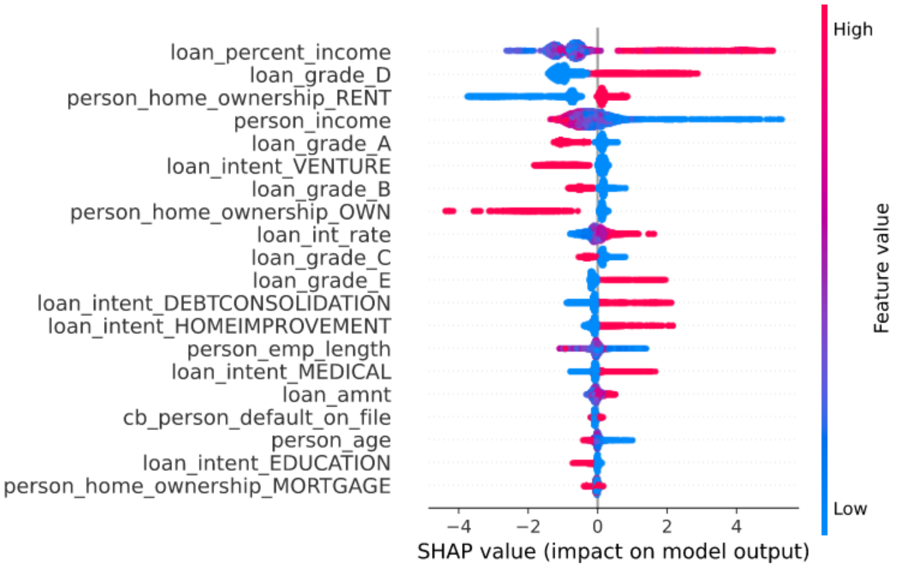

Figure 7: SHAP global individualized explanation of model behaviour

We can thus interpret the following from the first few features by observing Figure 7:

- *loan_percent_income*: If the loan as a percentage of a client's income increases, then the probability that they may be predicted to default is also increasing. The collection of the points on the negative side of the x-axis denotes that there are more data points that had a low *loan_percent_income* value.

- *loan_grade_D*: If this variable has value 1 instead of 0 (i.e., a client's loan is graded D), then there is a higher probability that a client will be classified as defaulting on their loan.

- *person_home_ownership_RENT*: If a person is more likely to rent their home (rather than own or through mortgage), there could be a higher probability of being classified as defaulting, however the data samples are showing that majority tend not to be classified as such.

- *person_income*: The higher a person's income, the less likely are they to be predicted to default on their loan.

- *loan_grade_A*: If a person's loan is graded A, then they are less likely to be predicted to default on the loan. Compare with *loan_grade_D*.

When it comes to observing the model behaviour locally, we can visualize the waterfall plot for a specific client (Figure 8). We selected a client at random from the testing set that was predicted to default, with 52% probability, and which in reality has also truly defaulted. Thus, in Figure 8, we can identify two key components – the first one being the base value, and second the variable arrows. The base value represents the mean of the log-odds of all the generated trees in a model like LightGBM, which essentially represents the probability of a client defaulting. The variable arrows are red (higher) and blue (lower), where the red arrows can be understood as pushing the prediction towards defaulting, while the blue arrows are features that push the prediction towards not defaulting on the loan.

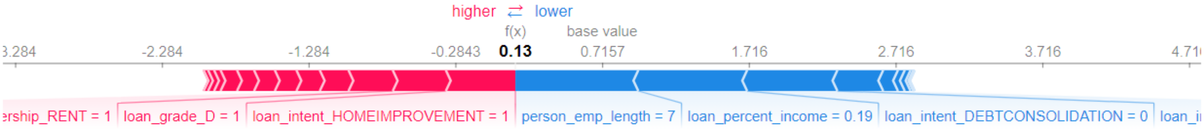

Figure 8: SHAP local explanation of model behaviour

Therefore, for the respective client, we could interpret the following variables that increase the probability to be predicted as defaulting:

- *loan_intent_HOMEIMPROVEMENT*: It appears that the client took the loan for the purpose of improving their home, which contributed the most to being predicted to default.

- *loan_grade_D*: The client was scored D by the loan providers, meaning that based on the client's credit history, quality of collateral, etc., the model did not consider them as trustworthy, which contributed towards their prediction of defaulting.

- *person_home_ownership_RENT*: Since this client is renting their home, there may be a higher likelihood that they are predicted to default on their loan.

And additionally, the following features contributed to decreasing the probability to be predicted as defaulting:

- *person_emp_length*: This client has been working for 7 years, which lowers their probability of defaulting on their loan.

- *loan_percent_income*: The income percentage of the loan is rather low, at a value of 19%, as opposed to majority of other loans. Therefore, this value lowered the probability for the client to be predicted as defaulting.

- *loan_intent_DEBTCONSOLIDATION*: This variable is equal to 0, which is in fact contributing to lowering the probability that this client would default.

However, since the SHAP values do not identify causality but instead demonstrate the variables that contributed to a model's prediction, we also implemented counterfactual methods to visualize which features a client may change in order to be predicted as non-defaulting. We allowed the features *person_income*, *person_emp_length*, *loan_percent_income* and *cb_person_cred_hist_length* to be varied because we believe that the other features can be hardly adjusted in a way that could lead to a non-defaulting situation (i.e. cannot tell a person to own or not rent their housing anymore, or to take a loan for a different reason). The minimum and maximum values we specified for variance were selected by analysing the client's own values in comparison to the rest of the samples, and applying our own domain knowledge. Observing Table 11, we can deduce that this person's yearly income may have to be higher (and possibly having been employed for longer) so that they are not predicted to default on their loan. However, one should look at this result merely as an additional support, and in all cases use their domain knowledge, as well as professional and moral judgement when providing advice.

		Feature			
		person_income	**person_emp_length**	**cb_person_cred_hist_length**	**loan_status**
Situation	**Original**	48,000	7	3	1
	Counterfactual #1	63,909.20	12.6	3	0
	Counterfactual #2	57,747	7	3	0

Table 11: Counterfactual explanations for a specific client

Reflecting on our framework's guidelines, we managed to provide explanations that may simultaneously satisfy the clients and abide by the legal requirements. We specifically followed the recommendations listed in van den Berg and Kuiper's [90] XAI framework destined for the financial sector. Moreover, cross-reviewing with the OECD [63] framework advice, we complied with the transparency and explainability characteristics part of the *AI Model* stage.

Verifications. Since we developed the system *in the lab*, we did not consult with the competent authorities nor performed the system in a regulatory sandbox. However, this is something we would consider doing in our future research in a real setting (see Table 12).

4.2.5 Stage 5: System Deployment

We did not deploy our model, hence we do not elaborate in closer detail on this stage. However, this is a stage which we would closely follow in our future research in a real setting (see Table 12).

Our XAI Framework		Case Study Action	Guideline Cross-check	Deliverable	Status
Stage	Phase				
Project Planning	Issue	Aligned professional experience in finance with interest, and identified a topic to explore.	**capAI:** Clearly defined the set of values and communicated them.	Highlighted issue and objective which can be further explored in an environmental context.	Completed.
	Environment	Identified stakeholders' responsibilities, and mapped their interactions.	**capAI:** In line with EU AI Act's requirements for environmental considerations.	Underlined potential impacts of the AI system, and if it (in)directly interacts with the end-users.	Completed.
	Business Objectives	Set success criteria in terms of model performance, and considered business-related perspectives.	**OECD:** Followed the industrial sector, business function, and business model criteria.	Business and economic objectives are considered throughout the duration of the project, and their achievement is aimed.	Completed.
Resource Assessment	Data	Identified dataset, checked usability, conducted an EDA phase, and planned steps to mitigate issues.	**capAI:** Reasoned and documented each process of data acquirement; compliance with GDPR and EU AI Act is ensured. Further investigation required if synthetic data is completely anonymous.	Verified data resource and guarantees that work can be conducted on it.	Partially completed.
	Models	Identified transparent and opaque models, and considered linear and tree-based models to be implemented.	**capAI:** Model sources and strategies for validating models was outlined.	Outlined design process; ensured familiarity with models architectures.	Completed.
	Regulations	Considered broader regulations of GDPR and EU AI Act. Identified that our system would be classified as high-risk.	**capAI:** Abided to data requirements as stipulated by the regulations.	Discover implications of the regulations and if it applies to own system. Ensure compliance and their following at each consecutive step.	Completed.
	Explanations	Reflected on domain knowledge and expertise. Identified post hoc methods to be applied to deliver explanations. Reasoned their usage in terms of purpose and audience.	**capAI:** GDPR's *meaningful information* requirement is fulfilled so that the affected stakeholders can take the appropriate legal measures.	Carefully selected post hoc methods which will appropriately explain model behaviour and outcomes, ensuring higher degree of trustworthiness.	Completed.
System Development	Data	Cleaned the data, split into training/test sets, remedied null values, and oversampled the imbalanced class.	**capAI:** Appropriate pre-processing measures taken. However, need to reason choice of imputation method to avoid bias.	Clean dataset which can be used in the *Models* phase.	Partially completed.
	Models	Initiated with transparent models, moved to opaque models. Implemented grid search, hyperparameter tuning, and cross-validation steps. Additionally, ensured reproducibility.	**capAI:** Each specified requirement as part of the *Development* stage is conformed with.	High-performing model which provides accurate predictions, and can be reproduced for technical and legal purposes.	Completed.
	Performance Evaluation	Expressed which performance metrics are of highest significance for our task. Then categorised the models based on these metrics, identifying the best performing one.	**capAI:** The model fulfills the selected performance criteria.	Best performing model is selected based on specific criteria, and its robustness is ensured.	Completed.
Understanding & Compliance	Explanations	Provided comprehensive explanations for model behaviour, involving global and local methods, and counterfactuals. With this information, stakeholders can understand decisions better.	**OECD:** Complied with transparency and explainability characteristics as part of the *AI Model* stage.	Adequate explanations which ensure a greater understanding of model behaviour and a higher trust for relying on them.	Completed.
	Verifications	**To do:** Engage in verifications with authorities through regulatory sandboxes. Employ third-party auditors and implement feedback.	**To comply: capAI:** Take part in the auditing ecosystem proposed by the EU AI Act. **OECD:** Check compliance with legacy systems and regulatory requirements**.**	**To achieve:** Legal compliance and attained AI system provision standards.	Future research.
System Deployment	Monitoring	**To do:** Monitor how a system performs, and retrain or mitigate issues where required.	**To comply: capAI:** Monitoring required in order to log issues and record remedying actions, which is in line with the EU AI Act.	**To achieve:** A continuously well performing system which remedies any arising issues, and ensures post-market compliance throughout lifetime.	Future research.
	Provision	**To do:** Release the system on the market while still providing continuous support and lifespan monitoring.	**To comply: capAI:** Provide the system to the end-target, collect feedback, and retire system if lifetime reached.	**To achieve:** Conclusion of the project and its provision, as well as system protection and overseeing across its entire lifespan.	Future research.

Table 12: Checklist for the application of the financial case study

5 Limitations & improvements

Here we state the limitations that we encountered with the application our framework and potential improvements to it based on additional literature from the domain, as well as the limitations we faced with during the complete duration of the thesis.

5.1 Framework suggestions

Resource Assessment. In the *Data* phase of the *Resource Assessment* stage we found ourselves working with synthetic data. Since we are not the creators of the dataset, we cannot be certain that the data is fully anonymised and contains no traces of personal information. Should that be the case, according to Recital 26[35] of the GDPR, the GDPR would not be applicable to our system hence it may be disregarded. However, if the data is only pseudonymized, meaning that it may in fact be attributed to natural persons, then the GDPR would apply in its entirety. Therefore, a potential improvement to the framework would be adding the clause of considering the type of data the provider is working with, which is especially important as the application of synthetic data in the AI industry is rapidly rising[36].

Another limitation we encountered was the inability to access the ISO standards related to AI, which requires a fee for each retrieved standard document. The ISO standards are a strong basis of ensuring that systems are developed in line with up-to-date techniques and requirements, hence an improvement would be to include them in the framework in order to refer to more concrete examples. In addition, while reviewing the regulatory literature after our framework development, we discovered that we overlooked other European legal texts that would come either in direct or indirect contact with AI systems. In their publication, the European Union Agency for Cybersecurity (ENISA) highlights different legislative initiatives that should be considered concomitantly with the GDPR and EU AI Act while developing AI systems [30]. Examples are the Cybersecurity Act, and the upcoming Cyber Resilience Act, which are two legal documents that should reinforce the guarantees of a system's robustness.

System Development. Having explored the literature on adversarial attacks, we were led to identify that adversarial attacks are highly coupled with system explainability. For instance, Baniecki and Biecek's [9] survey demonstrates that attacks may also be conducted on post hoc explanation methods in order to evaluate their reliability. A study that specifically stands out from the survey is Slack et al. [81], which reveals the vulnerabilities of the LIME and SHAP methods on a similar application like ours. Therefore, an improvement to our proposed framework would be to not only consider adversarial attacks on the ML models, but also on the post hoc methods.

Understanding & Compliance. After the application of the framework, we realised that in the *Explanations* phase, we executed post hoc methods only on the final best-performing model. Since each model has its own internal algorithm and method of arriving to the predictions [10], the framework should recommend that providers execute the same post hoc methods on all previously developed ML models and assess comparison criteria such as *fidelity*, *consistency*, and *stability*.

[35] https://gdpr.eu/recital-26-not-applicable-to-anonymous-data/

[36] https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/

System Deployment. Even though we did not proceed with the *System Deployment* stage in our application, we identified in other frameworks different kind of provisions for the *Monitoring* stage. For example, the capAI framework by Floridi et al. [34] includes the necessity to test the systems exhaustively *ex ante*, and also implement automated anomaly detection systems after the system is deployed. As such, we consider that the following provisions should also be included in our proposed framework, and highlighted in closer detail.

5.2 General remarks

One limitation we encountered while writing the thesis was the lack of formalism in the XAI domain. Since the attention towards XAI has shifted only in the recent decade, there are a plethora of definitions and concepts introduced by known researchers. While their propositions are vastly contributing to the field, there is a need for more testing and evaluation in order to arrive to a common basis on which the domain may continue building upon. Several authors, such as Saeed and Omlin [73] and Samek and Müller [76], agree with our remark.

Another limitation we were faced with was during our attempt to perform an adversarial attack on our model using the ART package [61]. To elaborate, the data we worked with is tabular, and according to Cartella et al. [18], adversarial attacks have been mainly implemented on image recognition tasks using DNNs. Since the tabular domain is yet to be further explored, the authors advise adapting the attacking algorithms to fit tabular instead of image data, which we propose to accomplish in our future research.

6 Conclusion & future work

In this chapter, we conclude our findings and highlight the potential future work that could be done in relation with our thesis.

6.1 Conclusion

Our main research topic of the thesis was to explore the explainability of AI systems. Therefore, we set one overarching question, followed by three sub-questions. To answer our first research sub-question *"Why is there an increasing necessity of XAI, especially in the high stakes domains?"*, we began by delving into the available literature on XAI and highlighting the domain's mission to enable a higher understanding of the AI systems. In addition to that, we elaborated on the regulatory requirements such as the GDPR and the upcoming EU AI Act, and their stricter explainability demands for high-risk systems. Specifically, even though the providers may not be explicitly barred from using ML algorithms that are not intrinsically interpretable, they will have to conform to the compliance-oriented and user-empowering explainability aspects as reinforced by the regulations. Providers are additionally encouraged to increase their interaction with the legal authorities by engaging in regulatory sandboxes and perform verifications in order to make sure that the systems do not pose a significant risk for the society.

Next, we classified the ML models based on the criteria of opaque and transparent models, identifying whether a post hoc analysis of the models is required. For opaque models, we discovered that this is often a necessity, which is why we described the post hoc methods for the delivery of explanations. We enumerated several examples from the model-agnostic and model-specific subgroups, and detailed their application and types of information provided. In addition, we elaborated on the evaluation criteria from a model, method, and user aspect, and discussed how they enable the assessment of the methods given a specific situation.

For the second sub-question *"What do providers need to consider in terms of technical and regulatory requirements when deploying their AI systems?"*, we turned to one of the core components of AI systems development — project management frameworks. Such frameworks provide a defined structure of processes that should be implemented and followed, enabling teams to be better informed and prepared for the delivery of their systems. Therefore, we explored the most implemented frameworks, such as CRISP-DM, KDD, and SEMMA, and their respective enhancement propositions. Having identified specific gaps regarding explainability in these frameworks, we proposed our own project management framework that focuses on the discussed regulations and explanation requirements. Our two novel propositions were the *Resource Assessment* and *Understanding & Compliance* stages, where the former focuses on an *ex ante* assessment of the technical and legal resources the provider should focus on, and the latter includes the provision of explanations and verifications with the stakeholders. Each respective phase of the stages was backed by the regulations and advice from key institutional players in the sphere of AI.

In regard to the third sub-question *"What is the degree of understanding that XAI methods facilitate when applied to opaque models?"*, we demonstrated an application of our proposed framework on a use case from the financial domain. Focusing on the task of predicting the probability of a client defaulting on a loan, we developed an opaque model which outperformed other transparent and opaque models. Consequently, we also provided explanations for the model behaviour, by applying the SHAP global and local methods, as well as counterfactuals. As a result, we were able to identify the contributing features to the model outcomes, and the contrastive situations that a client may conform to in order to not be predicted as defaulting. We identified the framework to be a useful instrument, since being introduced to the regulatory and technical requirements early on during the project allowed us to carefully analyse in advance whether it is feasible to proceed with the next step. In a real setting this would translate to cutting down additional time, efforts, and costs when failing to comply with such provisions.

Overall, considering the above evidence and the application of our research methodology, we conclude that we managed to answer our main question *"How can XAI contribute towards ensuring trustworthy AI systems?"*.

6.2 Future work

The work that we conducted implies additional future research directions. We begin by elaborating on the ways we would expand on our performed study, followed by how other stakeholders can further the work of the explainability domain.

From our perspective, it would be interesting to reproduce the studies of Ballet et al. [8] and Cartella et al. [18] by conducting an adversarial attack on the opaque models, but also on the post hoc methods accompanying them. As a result, we could analyse in closer detail the robustness of our ML models and the explanations we provide. Another significant future work would be to analyse the application of our suggested framework in a real situation. By seeing how the framework manifests in the daily processes of system providers, we could better deduce which parts of the framework need to be adjusted and improved. In addition, in the thesis we aimed to elaborate on a generalized framework which can be applied in any given AI development situation. However, we realize that various domains have different necessities and requirements, which is why we consider that designing a tailored framework to specific fields, such as healthcare, mobility, logistics, and e-commerce, would provide catered and better adapted provisions for the system providers.

In addition, the domain of XAI would benefit of further research and development from key players such as academics, advisory institutions, and authorities. A close-knit cooperation between the development and regulatory realms would imply that the innovation of high-performing AI systems is not stifled, and that they are developed in a trustworthy and responsible manner under legal supervision. Moreover, we consider that dedicating more attention and effort towards the domain of interpretable ML would also further the trustworthiness of AI, especially as the researchers in this domain focus on developing intrinsically interpretable models that are simultaneously high-performing.

References

[1] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[2] Greg Adamson. Explainable Artificial Intelligence (XAI): A reason to believe? *Law in Context. A Socio-legal Journal*, 37(3), 2021.

[3] David Alvarez-Melis and Tommi S. Jaakkola. On the Robustness of Interpretability Methods. June 2018.

[4] Plamen P. Angelov, Eduardo A. Soares, Richard Jiang, Nicholas I. Arnold, and Peter M. Atkinson. Explainable artificial intelligence: an analytical review. *WIREs Data Mining and Knowledge Discovery*, 11(5):e1424, 2021.

[5] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions, February 2023.

[6] Ana Azevedo and M F Santos. KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. (Instituto Politécnico do Porto. Instituto Superior de Contabilidade e Administração do Porto), 2008.

[7] BaFin. Big data and artificial intelligence: Principles for the use of algorithms in decision-making processes, 2021.

[8] Vincent Ballet, Xavier Renard, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, and Marcin Detyniecki. Imperceptible Adversarial Attacks on Tabular Data. December 2019.

[9] Hubert Baniecki and Przemyslaw Biecek. Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A Survey, June 2023.

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020.

[11] Vaishak Belle and Ioannis Papantonis. Principles and Practice of Explainable Machine Learning. *Frontiers in Big Data*, 4, 2021.

[12] Derrick A. Bennett. How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5):464–469, October 2001.

[13] Julian Birkinshaw. What Is the Value of Firms in an AI World? In Jordi Canals and Franz Heukamp, editors, *The Future of Management in an AI World: Redefining Purpose and Strategy in the Fourth Industrial Revolution*, IESE Business Collection, pages 23–35. Springer International Publishing, Cham, 2020.

[14] Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. Machine Learning Explainability in Finance: An Application to Default Risk Analysis. *SSRN Electronic Journal*, 2019.

[15] Nadia Burkart and Marco F. Huber. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 70:245–317, January 2021.

[16] Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. 3(1), 2016.

[17] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. Explainable AI in Fintech Risk Management. *Frontiers in Artificial Intelligence*, 3, 2020.

[18] Francesco Cartella, Orlando Anunciacao, Yuki Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. Adversarial Attacks for Tabular Data: Application to Fraud Detection and Imbalanced Data. January 2021.

[19] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, August 2019.

[20] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. CRISP-DM 1.0. *SPSS*, 2000.

[21] Chinchih Chen, Carl Benedikt Frey, and Giorgio Presidente. Privacy Regulation and Firm Performance: Estimating the GDPR Effect Globally. 2022.

[22] Jessica Dai, Sohini Upadhyay, U. Aïvodji, Stephen H. Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

[23] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems, May 2021.

[24] Jürgen Dieber and Sabrina Kirrane. A novel model usability evaluation framework (MUsE) for explainable artificial intelligence. *Information Fusion*, 81:143–153, May 2022.

[25] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, March 2017.

[26] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, May 2018.

[27] Martin Ebers. Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s), August 2021.

[28] EC. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016.

[29] EC. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS, 2021.

[30] ENISA. *Standardisation in support of the cybersecurity of AI.* Publications Office, LU, 2023.

[31] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust Physical-World Attacks on Deep Learning Models, April 2018.

[32] Gerald Fahner. Developing Transparent Credit Risk Scorecards More Effectively: An Explainable Artificial Intelligence Approach. pages 7–14, 2018.

[33] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From Data Mining to Knowledge Discovery in Databases. 17(3), 1996.

[34] Luciano Floridi, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act, March 2022.

[35] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making AI Forget You: Data Deletion in Machine Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[36] Michelle Goddard. The EU General Data Protection Regulation (GDPR): European Regulation that has a Global Impact. *International Journal of Market Research*, 59(6):703–705, November 2017.

[37] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, October 2017.

[38] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A Survey Of Methods For Explaining Black Box Models, June 2018.

[39] David Gunning and David Aha. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58, June 2019.

[40] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120, December 2019.

[41] Balint Gyevnar and Nick Ferguson. Aligning Explainable AI and the Law: The European Perspective, February 2023.

[42] Philipp Hacker and Jan-Hendrik Passoth. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML*

2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, Lecture Notes in Computer Science, pages 343–373. Springer International Publishing, Cham, 2022.

[43] Ronan Hamon, Henrik Junklewitz, and MARTIN Jose Ignacio Sanchez. Robustness and Explainability of Artificial Intelligence, 2020.

[44] AI HLEG. Assessment List for Trustworthy Artificial Intelligence (ALTAI). Technical report, 2020.

[45] ICO. Guidance on the AI Auditing Framework: Draft Guidance for Consultation. 2020.

[46] ICO and TATI. Explaining decisions made with AI. Technical report, 2022.

[47] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable Artificial Intelligence Approaches: A Survey, January 2021.

[48] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, and Rajiv Suman. Artificial Intelligence Applications for Industry 4.0: A Literature-Based Study. *Journal of Industrial Integration and Management*, 07(01):83–111, March 2022.

[49] Jens Klump, Lesley Wyborn, Mingfang Wu, Julia Martin, Robert R. Downs, and Ari Asmi. Versioning Data Is About More than Revisions: A Conceptual Framework and Proposed Principles. *Data Science Journal*, 20:12, March 2021.

[50] Andreas Liebl and Till Klein. AI Act Impact Survey. Survey, appliedAI Institute for Europe, 2022.

[51] Zachary C. Lipton. The Mythos of Model Interpretability, March 2017.

[52] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, January 2020.

[53] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[54] Helena Löfström, Karl Hammar, and Ulf Johansson. A Meta Survey of Quality Evaluation Criteria in Explanation Methods, March 2022.

[55] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. The AI Index 2023 Annual Report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, 2023.

[56] McKinsey. Cost decreases from adopting artificial intelligence (AI) in organizations worldwide as of fiscal year 2020, 2022.

[57] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019.

[58] Branka Hadji Misheva, Joerg Osterrieder, Ali Hirsa, Onkar Kulkarni, and Stephen Fung Lin. Explainable AI in Credit Risk Management, March 2021.

[59] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* 2 edition, 2023.

[60] Jessica Newman. A Taxonomy of Trustworthiness for Artificial Intelligence. Technical report, Center for Long-Term Cybersecurity, UC Berkeley, UC Berkeley, 2023.

[61] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian M. Molloy, and Ben Edwards. Adversarial Robustness Toolbox v1.0.0, November 2019. arXiv:1807.01069 [cs, stat].

[62] OECD. AI in Finance. In *OECD Business and Finance Outlook 2021: AI in Business and Finance*, OECD Business and Finance Outlook. OECD, 2019.

[63] OECD. OECD Framework for the Classification of AI systems. 2022.

[64] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, and Emilia Gomez. The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1139–1150, New York, NY, USA, 2023. Association for Computing Machinery.

[65] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning, March 2017.

[66] David M. W. Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2011.

[67] Trivellore E. Raghunathan. Synthetic Data. *Annual Review of Statistics and Its Application*, 8(1):129–140, 2021.

[68] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability. Technical report, AI Now Institute, 2018.

[69] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-Agnostic Interpretability of Machine Learning, June 2016.

[70] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, February 2016.

[71] Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, September 2019.

[72] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges, July 2021. arXiv:2103.11251 [cs, stat].

[73] Waddah Saeed and Christian Omlin. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, March 2023.

[74] Jeff Saltz, Nicholas Hotz, David Wild, and Kyle Stirling. Exploring Project Management Methodologies Used Within Data Science Teams. *Americas Conference on Information Systems*, 2018.

[75] Jeff Saltz and Iva Krasteva. Current approaches for executing big data science projects—a systematic literature review. *PeerJ Computer Science*, 8:e862, February 2022.

[76] Wojciech Samek and Klaus-Robert Müller. Towards Explainable Artificial Intelligence. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 5–22. Springer International Publishing, Cham, 2019.

[77] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, 181:526–534, January 2021.

[78] Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, November 2017.

[79] Vivek Singh, Anshuman Singh, and Kailash Joshi. Fair CRISP-DM: Embedding Fairness in Machine Learning (ML) Development Life Cycle. 2022.

[80] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. In *Advances in Neural Information Processing Systems*, volume 34, pages 9391–9404. Curran Associates, Inc., 2021.

[81] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods, February 2020.

[82] Nathalie A. Smuha, Emma Ahmed-Rengers, Adam Harkens, Wenlong Li, James MacLaren, Riccardo Piselli, and Karen Yeung. How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act, August 2021.

[83] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. Metrics, Explainability and the European AI Act Proposal. *J*, 5(1):126–138, March 2022.

[84] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, 9:11974–12001, 2021.

[85] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics.

[86] Stefan Studer, Thanh Binh Bui, Christian Drescher, Alexander Hanuschkin, Ludwig Winkler, Steven Peters, and Klaus-Robert Müller. Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology. *Machine Learning and Knowledge Extraction*, 3(2):392–413, April 2021.

[87] Agus Sudjianto and Aijun Zhang. Designing Inherently Interpretable Machine Learning Models, November 2021.

[88] Lao Tse. Credit Risk Dataset, 2020.

[89] Kristin Undheim, Truls Erikson, and Bram Timmermans. True uncertainty and ethical AI: regulatory sandboxes as a policy tool for moral imagination. *AI and Ethics*, November 2022.

[90] Martin van den Berg and Ouren Kuiper. XAI in the Financial Sector: A Conceptual Framework for Explainable AI (XAI). *Hogeschool Utrecht, Lectoraat Artificial Intelligence*, 2020.

[91] Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, August 2021.

[92] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, December 2016.

[93] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, March 2018.

[94] Wei Wei, Milan Patel, and Paul Cotton. The foundational standards for AI. AI Workshop 24-25 May, 2022. ISO / IEC.

[95] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? 2017.

[96] Yongfeng Zhang and Xu Chen. Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020.

[97] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, October 2021.

Appendix A: Dataset Information

Feature Name	Description
person_age	Age
person_income	Annual income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non-default, 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default (Y is yes, N is no)
cb_preson_cred_hist_length	Credit history length

Table A1: Use case dataset description, adapted from Tse (2020)

		Feature											
		person_age	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
Index	0	22	59000	RENT	123	PERSONAL	D	35000	16.02	1	0.59	Y	3
	1	21	9600	OWN	5	EDUCATION	B	1000	11.14	0	0.1	N	2
	2	25	9600	MORTGAGE	1	MEDICAL	C	5500	12.87	1	0.57	N	3
	3	23	65500	RENT	4	MEDICAL	C	35000	15.23	1	0.53	N	2
	4	24	54400	RENT	8	MEDICAL	C	35000	14.27	1	0.55	Y	4
	...	...	...	...	...	...	...	...	...	...	...	...	...

Table A2: Head of the dataset, Tse (2020)