Leiden University

ICT in Business and the Public Sector

A Turnkey Explainable AI System for AI Compliance in the Financial Sector

Name:	Mitch Angenent
Student number:	s1287125
Date:	February 24, 2022
1st supervisor:	dr. G.J. Ramackers
2nd supervisor:	prof. dr. ir. J.M.W. Visser

MASTER THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1 — 2333 CA Leiden — The Netherlands

Acknowledgements

This thesis was written under the supervision of dr. Guus Ramackers, who I would like to thank for the inspiration to start this research, and his intensive support through the project. The numerous brainstorm and feedback sessions have helped tremendously to arrive at this point. In addition, I would like to thank prof. dr. ir. Joost Visser for the inspiration and constructive feedback on crucial elements in this research. Thank you both for your flexibility during the final parts of this thesis.

This thesis is a partial fulfillment of the requirements of the Master of Science degree 'ICT in Business and the Public Sector', a program at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University.

Abstract

Although AI-driven predictive models have numerous benefits, organizations of all sizes struggle with AI adoption. This is due to the legal, ethical, and regulatory concerns that arise from the black-box behavior of these techniques, and the lack of easy-to-implement tools to mitigate those risks on an enterprise scale. This study addresses this problem by developing, applying, and evaluating a turnkey explainable AI system that supports compliance with AI regulations in the financial sector - a heavily regulated industry. During an internship at a major Dutch insurance firm, we developed and deployed a system that combines multiple model-agnostic explainable AI techniques (including SHAP) under one interface, allowing the use of a single system to provide insights into the inner workings and fairness of any supervised learning model. In addition, the system is highly configurable and presents insights through an interactive report that is specifically tailored to serve a variety of stakeholders, from data scientists to compliance officers. We designed the system to support AI compliance and model refinement, and subsequently demonstrated these use cases and the system's broad applicability by applying it to four production models. We performed an extensive evaluation to quantify the system's effectiveness, and concluded that the system has a significant positive effect on participants' model understanding, internal communication, and ability to assess model fairness. Additionally, the participants perceived this tool to be significantly more effective and easier to use compared to previously used tools. Finally, we present a procedure that prescribes how the system can be quickly and effectively implemented in other organizations. By enhancing the ethical use of AI and supporting compliance with AI regulations, this system removes various dominant hurdles for widespread AI adoption.

Keywords: explainable AI, interpretable machine learning, AI compliance, AI regulations, responsible AI.

Contents

Chapter 1

Introduction

Over the past decades, the ongoing digitalization of our society has led to rapidly increasing data volumes. It is extremely beneficial for organizations to extract knowledge from this data for data-driven decision making, allowing them to adapt their strategies to the constantly changing environment and competition. Where traditionally descriptive analysis was used to comprehend past events, there currently is a paradigm shift taking place towards predictive analysis to make automated decisions in a split-second. This enables the automation of processes, leading to a more scalable business model with lower operational costs while also meeting customer demands in the 24/7 economy. These analyses can be performed by predictive models that consist of a set of hand-crafted rules. Defining rules by hand gets increasingly difficult as the process becomes more complex, to the point where the hand-crafted rules cannot properly capture the real world, hence excluding many applications.

A solution is offered by the field of artificial intelligence (AI), which includes techniques to allow machines to learn and execute specific tasks, without being explicitly programmed by a human to do so. A subfield of AI, machine learning (ML), employs algorithms to automatically extract patterns from historical data, making hand-crafted rules redundant. The resulting predictive model is able to make autonomous predictions and decisions for specific tasks, typically with a better performance than humans. Well-known examples are the recommendation systems in web shops or streaming platforms, which recommend products based on thousands of data points that capture the prior behavior of customers, enhancing customer experience and boosting profits. AI and ML also emerge in more crucial processes, such as loan acceptance. Credit scoring models predict the probability of default of the loan application based on hundreds of characteristics such as credit history and current income. The predictive model can instantly inform the applicant on their chances to get a loan, while also monitoring the risk for the lender.

In general, AI and ML serve both organizations and their customers by automating complex processes. With the tremendous amount of available historical data in combination with improved algorithms and ubiquitous computational power, AI and ML have the potential to be incorporated in a variety of applications. Effective adoption of these techniques therefore allows organizations to create and sustain a competitive advantage.

1.1 Research Problem

Despite the aforementioned benefits, organizations struggle with the adoption of AI and ML. During a 2021 survey, only one in three enterprises reported that they have deployed AI across their business processes [1]. Several barriers for adoption in business environments are reported in such studies [1–3], including limited AI expertise or knowledge, lack of tools or platforms for AI development, and legal, ethical, and regulatory concerns of executives. The latter is caused by the *black box* behavior of many ML models: only the predicted outcome is communicated, while the way the model establishes these results remains unclear. For example, the aforementioned credit scoring model only communicates whether or not the loan is accepted, without communicating the determining factors. Such a model could reject a loan based on attributes such as gender or race, which raises ethical objections and is prohibited by law. Given that ML models automatically extract patterns from data, they can find and employ all sorts of correlations that do not necessarily reflect a causal relationship in the real world. This lack of transparency raises concerns regarding bias and reliability of the results, preventing developers, end users, and managers from understanding and trusting these models.

These concerns are not ungrounded, given the frequent reporting of harmful AI systems. For example, the recidivism scoring model *COMPAS* turned out to be biased towards race [4], and an AI-driven recruitment tool used by Amazon was taken out of order as it disadvantaged women [5]. This sometimes even leads to intervention by regulators: in 2020, a judge banned the use of the predictive fraud model *SyRI* used by the Dutch government due to a lack of transparency and controllability [6]. With the number of incisive AI regulations on the rise, adopting AI and ML without precautions forms a significant risk.

A mitigation factor is the use of explainable artificial intelligence (XAI). This emerging field focuses on the techniques and best practices to enhance the transparency, interpretability, and explainability of AI solutions. XAI techniques improve human understanding of ML models by providing explanations of how a model arrives at a particular decision, e.g., by displaying the impact of each input feature on the model output. These explanations serve all kinds of actors with different purposes, with improved trust and understanding being recurring points. For end-users who interact with models to execute a task, the explanation of a single prediction allows the user to assess the prediction's correctness. This enhances trust and understanding, and thereby allows better human-computer interaction. Similar explanations of how the model arrived at the decision can be used to comply with a customer's 'Right to Explain' in the European Union and United States. Data scientists might use a high number of aggregated explanations for debugging purposes, such as inspecting the system for spurious correlations. Additionally, a better understanding of the inner workings of the model enables model refinement. For compliance officers and managers, a major concern is the accountability of AI systems. This concern can be reduced if the trust and understanding in these systems are enhanced with XAI techniques. XAI thus has the potential to lower the barriers for AI adoption by meeting the legal obligations surrounding AI employment, supported by the fact that nine out of ten IT professionals reported that the employment of explainable and trustworthy AI are crucial to widespread adoption of AI, and business success, in their organization [1].

Due to these advantages, scientific research has developed many XAI techniques in recent years, such as LRP [7], LIME [8] and SHAP [9]. These methods all have their own characteristics such as their applicability to different algorithms and the type of provided explanations. Therefore, no single technique is present that can serve all stakeholders and all use cases. Moreover, applying the techniques is manual work and requires a certain degree of knowledge of the field. The lack of skills in XAI and related technologies is reported as the biggest barrier for developing trusted AI [1], especially in large organizations where knowledge is spread across multiple departments. Some cloud computing platforms such as Microsoft Azure[1], Amazon Web Services[2], Dataiku[3], and Arthur AI[4] offer various out-of-the-box explainability features. However, leveraging these features requires a cumbersome migration to these platforms, accompanied with high costs and long lead times. Commercial software is therefore, for most organizations, not a feasible solution for the problem at hand. We can therefore state that, despite the advancements in the field of explainable AI, there currently is no solution that can directly be applied on an enterprise scale to comply with AI regulations.

In summary, the dominant hurdles of ML adoption are legal, ethical, and regulatory concerns that are caused by (i) the lack of model transparency and (ii) the lack of tools that can directly be implemented to address this on an enterprise scale. The importance of researching solutions to this problem is twofold. From an economical perspective, AI and ML have the potential to enable great economical growth. McKinsey Global Institute estimates that by 2030, AI-related technologies could deliver additional economical activity up to $13 trillion, or a 16% higher cumulative GDP, compared to 2018 [10]. They state that this AI-driven productivity growth is impacted by labor automation, innovation, and new competition. Since the size of this growth depends on several factors, including the pace of AI adoption, it is imperative to address the hurdles of AI adoption. For businesses, it is critical to apply AI-related technologies to keep up with competition. Second, from an ethical perspective, it is essential to address the ethical concerns that arise from the employment of AI-systems, such as potential biases and discrimination. All those involved in the creation and deployment of these applications, including data science practitioners, managers, and compliance officers, have moral obligations to address the ethical considerations of AI and prevent the deployment of harmful AI systems. Hence, solving this research problem would contribute to economical growth, while keeping a high ethical standard.

A potential solution direction would be to develop a new XAI system that addresses the aforementioned problems. This implies that such a XAI system should support compliance with AI regulations and be turnkey. By a turnkey system, we mean the following. First, such a system can directly be implemented in the organization without cumbersome migrations and configurations. On the other hand, such a system has such characteristics that it can be applied by the vast majority of the organization for supporting AI compliance.

[1] https://azure.microsoft.com/en-us/services/machine-learning/responsibleml/
[2] https://aws.amazon.com/sagemaker/
[3] https://www.dataiku.com/product/key-capabilities/explainability/
[4] https://trust.arthur.ai/explainable-ai

1.2 Scope and Objectives

In this thesis, we study this potential solution direction of developing a turnkey XAI system for AI compliance. Given the high variety of AI techniques and use cases - from simple linear regression on tabular data to deep learning on unstructured data such as text and images - we first define the scope and objectives in this section.

The XAI solution should serve an domain that is highly affected by the aforementioned barriers that prevent the widespread application of AI. This implies that this sector (i) has the prerequisites of applying AI, (ii) would highly benefit the adoption of AI, and (iii) has significant risks associated with the application of AI. The financial sector meets these requirements. First, the financial sector has traditionally been highly digitalized, resulting in a wealth of data available. This data is usually structured and of sufficient quality because it is already used for descriptive analysis, and can conveniently be fed into machine learning algorithms to establish predictive models. The resulting models can in their turn be incorporated in these processes. The sector thereby satisfies the second criterion, as predictive models can directly be employed to reduce operational costs. Moreover, machine learning models are well suited for complex, but narrowly-scoped tasks, such as assessing loan applications, predicting portfolio risks, and service optimization. These tasks, executed by opaque and potentially biased predictive models, can have a significant impact on citizens that need financial services such as a mortgage. This leads to the previously mentioned legal, ethical, and regulatory concerns that are associated with these models. Due to the high impact of these activities, combined with the deterioration of confidence after the financial crisis in 2008, the sector is subject to strict regulations. In these regulations, there is an increasing focus on the requirements of AI-driven systems, with a penalty or sanction if these requirements are not met. On top of the three criteria mentioned earlier, financial institutions are typically large and soloed organizations and thus need a turnkey solution to comply with these regulations without allocating excessive amounts of resources. All things considered, financial institutions would highly benefit from an explainable AI system to reduce the barrier of AI adoption. Besides, society as a whole would benefit from financial processes that are driven by ethical AI.

For those reasons, the goal of this thesis is to design, develop, deploy, apply, and evaluate an explainable AI system that aids compliance with AI regulations in the financial sector. These steps will be performed at one financial institution to keep this project manageable. Moreover, this research will address how the system can be applied in other organizations and domains, giving this solution the potential to make an impact far beyond this single institution.

The research question (RQ) central to this research project will be as follows:

***RQ**: How can an explainable AI system support compliance with AI regulations in a financial institution?*

This research question will be answered by addressing five angles in the form of five sub questions (SQ):

***SQ1**: What are the design criteria for such an XAI system?*

***SQ2**: How can a prototype of the system be developed and deployed in such a way that it satisfies the design criteria?*

***SQ3**: How can such an XAI system be used for compliance with AI regulations?*

***SQ4**: What are the main factors that determine the functional suitability of such an XAI system?*

***SQ5**: How can such an XAI system be used in other organizations and domains?*

Given that this is a design study, objectives are formulated which must be met in order to answer the sub questions and thereby the research question. One objective is formulated for each sub question:

O1: Define the design criteria for the system.

O2: Develop and deploy a prototype of the system.

O3: Describe the use cases for AI compliance of the system.

O4: Determine the main factors that determine the functional suitability of the system.

O5: Describe how the system can be used in other organizations and domains with a framework.

The contribution of this study is threefold. First, by designing, developing, and applying the XAI system, organizations in the financial industry are presented with both a tool and the associated recommendations to better comply with AI regulations. This removes the hurdle of widespread AI adoption, ultimately leading to economical growth for both organizations and economy as a whole. Second, by improving AI compliance with the XAI system, more responsible and ethical AI will be deployed. Third, the findings of this study, such as the main factors that determine the functional suitability of the system, contribute to the field of XAI and can be used to improve similar systems.

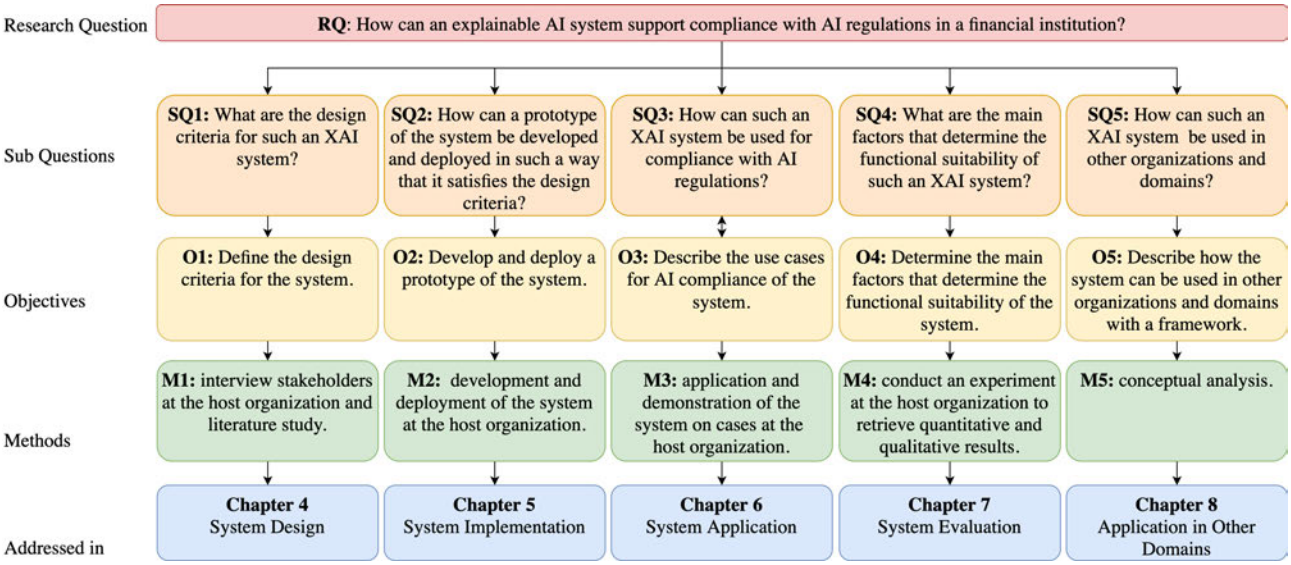


Figure 1.1: Overview of this research.

1.3 Methodology

This section describes the methodology that is applied to achieve the five objectives mentioned. Figure 1.1 presents how the research question, sub questions, objectives, and methods of this design study are connected. This first objective - defining the design criteria - will be achieved by interviewing stakeholders as the host organization regarding their needs, and by conducting a literature review. By developing and deploying the systems at the host organizations, the second objective will be met. To achieve the third objective, the use cases of the system will be described based on applying the system to predictive models at the host organization. The system will be evaluated in order to define the main factors that determine the functional suitability of the system. For this fourth objective, an experiment in the form of a survey is conducted at the host organization to gather both quantitative and qualitative results regarding the effectiveness and user experience of the system. Finally, a conceptual analysis is performed to define a framework and thereby meet the fifth objective. The research is conducted during an internship at a major Dutch financial services provider, Achmea.

1.4 Outline

The structure of this thesis is as follows. Chapter 2 provides background information regarding AI regulations and the field of explainable AI. The context in which this study is conducted, such as information about the host organization, how it employs machine learning, and under which regulations it operates, is presented in Chapter 3. Chapter 4 elaborates on the design criteria, XAI techniques, and the architecture used, while Chapter 5 describes the technical implementation of the system. The general use cases and the application of the system at the host organization are presented Chapter 6. Chapter 7 describes the setup and results of the system evaluation. How the system can be used in other organizations and domains is provided in Chapter 8. Finally, Chapter 9 concludes this work and proposes recommendations for future work.

Chapter 2

Background

This chapter provides relevant background information on the two main topics addressed in this research: AI regulations and explainable AI. Section 2.1 maps and describes the laws and regulations that financial institutions should comply with when employing data-driven applications. The field of explainable AI, from which we will use techniques to comply with these regulations, is outlined in Section 2.2.

2.1 AI Regulations

To get an understanding of the requirements for data-driven applications within the financial sector, this section maps most of the relevant laws and (self-)regulations that apply in this sector. This overview is not exhaustive, as the main goal is to indicate the variety of angles these regulations come from. Some are specifically tailored to applications of AI models, others are more focused on data-driven systems in general, and some address fundamental elements such as non-discrimination. The jurisdiction and legal biding of these laws and regulations are used as two dimensions to map them, which is further explained in the next paragraph. After this mapping, the content of these laws and regulations will be examined in more detail in the subsequent paragraphs.

Dimensions. Two primary dimensions are distinguished to create a mapping. Starting with the jurisdiction, laws and regulations can be applicable on an international level, on a national level, to specific sectors, or specific branches. In this research, we focus on the European Union, the Netherlands, the Dutch financial institutions, and the Dutch insurance industry, respectively. At each level of this hierarchy, the parent elements are also applicable, for example, the General Data Protection Regulation of the European Union is active in all member states and their sectors. The European Union imposes more regulations than other regions such as the USA (which has more specific regulations such as the Fair Credit Reporting Act and Equal Credit Opportunity Act), ensuring that the findings of this study can be utilized in other regions as well. Additionally, most laws in the next paragraph apply to non-European companies that generate revenue in the EU, a significant market.

The second dimension is how legally binding the requirements are. Legislation on an international and national level are legally binding. Non-compliance with these laws can lead to large fines and criminal investigations, even for natural persons, as some Dutch banks have experienced this after failing to comply with the Dutch Anti-Money Laundering and Anti-Terrorist Financing Act [11,12]. Less binding are self-regulations, which are typically guidelines that are drawn up by a regulator of a specific sector or an industry association. Underlying companies are imposed - or mutually agree - to comply with these guidelines. An example are the AI guidelines (SAFEST) proposed by the Dutch Central Bank, that affect all financial institutions. Non-compliance may lead to a warning and potentially more frequent and stricter inspections, but not to criminal investigations.

Table 2.1: Listing of regulations that are relevant for the application of AI.

#	Act	Legal binding	Jurisdiction	Note
1	GDPR	Law	European Union	
2	EU Artificial Intelligence Act	Law	European Union	Proposed
3	Anti-discrimination laws	Law	The Netherlands	
4	EU Ethics Guidelines for Trustworthy AI	Soft-law	European Union	Used in 2, 5, 6
5	DNB SAFEST	Soft-law	Financial institutions (NL)	
6	Ethical Framework	Soft-law	Insurance industry (NL)	

Laws and regulations. Table 2.1 lists several relevant laws and regulations that apply to data-driven techniques such as AI, which are discussed hereafter.

1. GDPR. The General Data Protection Regulations (GDPR) [13] is the first widespread regulation regarding data gathering and data use in the European Union (EU), which became effective in May 2018. In the Netherlands, the GDPR is known as the *AVG*. The philosophy behind the GDPR is to enforce the same standards for gathering, storing, and using Personal Identifiable Information (PII) of EU citizens, regardless of the location in the EU where this data is processed. The law uses a broad definition of personal data, as stated in Article 4: "any information relating to an identified or identifiable natural person", meaning that any information that can be used to distinguish a living person from others is covered by the GDPR. This includes names, identity numbers, addresses and other location data, online identifiers such as IP-addresses and cookies, and any information related to the "physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". This is regardless of the data format, and if the information is correct or objective. The GDPR stipulates that PII can only be stored with the permission of the individual, or if there is a reasonable ground to do so, on a basis in law. Information that is extra sensitive, such as race, ethnicity, religion, political preference, sexual orientation, and health-related or genetic data, may only be stored if there is a legal ground. Additionally, the law provides EU citizens with the rights to view, correct, transfer, and be informed about, and limit the use of their personal data. Non-compliance can lead to fines as high as 20 million euro or 4% of the total worldwide annual turnover of the breaching organization. The GDPR impacts the employment of AI and ML as it restricts the amount of data stored - a paramount resource for AI and ML applications to perform well.

2. EU Artificial Intelligence Act. The EU AI Act [14] has been proposed in April 2021 and will regulate the use of AI within the European Union. With this draft regulation, the European Commission is the first body that plans to regulate the use of AI and harmonize the laws across member states - an approach similar to that of the GDPR. The legislation will apply to any AI system that provides output in the European Union, meaning that it will impact organizations around the globe. The regulations will be based on the risk associated with the AI system, classifying AI systems into three tiers: limited or minimal risk, high-risk, and unacceptable risk. Systems that are classified as an unacceptable risk are those that manipulate human behavior or exploit vulnerable humans such as children, provide real-time biometric identification in public places for law enforcement, or predict a social score or personality traits based on social behavior. The use of these AI applications will be prohibited. The deployment of high-risk AI systems will also be regulated. This category includes (i) all AI systems that are part of products and safety systems that are already covered in EU legislation, such as medical devices, aviation, and motor vehicles, and (ii) AI systems that are classified as high-risk. A full list of the latter - which can be extended in the future - is listed in Annex III of the Act. Examples are:

1. Biometric identification and categorisation of natural persons in non-public places.
2. AI systems used for management and operation of critical infrastructure such as road traffic and the supply of water, gas, heating, and electricity.
3. AI systems used for determining access to, and assigning and assessing students in, education and vocational training.
4. AI systems used for recruitment, selection, and evaluation in employment.
5. AI systems used for evaluating access to to public and private benefits, including the establishments of credit scores for natural persons.
6. AI systems used by law enforcement to make risk assessments, detect the emotional state, and profiling of natural persons during investigations.
7. AI systems used by public authorities for migration, asylum, and border control management.
8. AI systems used for administration of justice and democratic processes.

Under the proposed regulations, providers of high-risk systems will be subject to extensive obligations:

- Registration: high-risk AI systems must be registered in a publicly accessible database that is managed by the European Commission.
- Conformity assessment: before an high-risk AI system can be used within the EU, it must undergo an assessment to ensure it conforms to the AI regulation. For most AI systems, this conformity assessment can be performed by the provider itself using self-assessment. AI systems listed under points A and B in the aforementioned list should be audited by an external party at least every five years.

- Information to users: end-users must be adequately informed about the characteristics, capabilities, and limitations of a system in order to be able to understand and interpret the system and the output.
- High quality data: datasets used for training, testing, and evaluating AI systems must be of high quality, i.e., must be relevant for the problem, representative for the population, free of errors, and complete. Sensitive personal information such as ethnicity and religion can only be used to monitor the AI system for potential bias.
- Robustness: providers must design AI systems that are both appropriately-performing and robust, thus resilient to errors and adversarial use.
- Risk and quality management: providers of high-risk AI systems are obliged to establish a risk management system to identify, evaluate, document, and mitigate the risks that are associated with the system. Additionally, they must establish quality management systems that address technical and regulatory standards, and automatically log the compliance with these standards.
- Technical documentation: the system conformity and other technical details of the AI system must actively be maintained in technical documentation.
- Monitoring and human oversight: the performance of high-risk AI systems must be continuously monitored by the provider as long as the system is in operation. The provider is obliged to inform the authorities in case of significant incidents. An additional requirement is that this monitoring can effectively be executed by humans that fully understand the AI system and have the authorization to disregard, override, or interrupt the system.

Albeit to a lesser extent, the draft regulation also prescribes obligations for importers, distributors, and users of the system. The third tier of the risk-based approach are minimum risk AI systems such as chatbots, spam filters, inventory management systems, and customer and market segmentation. Under the draft regulations, some of these systems are subject to transparency obligations, including systems that interact with humans such as chatbots, and emotion recognition software. Other minimum risk systems that do not fall under the aforementioned categories are unregulated, however, the regulation suggests that providers of these systems should regulate themselves with the standards that are imposed for high-risk AI systems. Non-compliance with the legislation can lead to fines as high as 30 million euro or 6% of the annual turnover. Due to the regulation's jurisdiction, span, and high amount of requirements, the EU AI Act will significantly impact businesses around the globe as it prohibits certain AI applications and will increase the costs for complying with the regulations for high-risk systems. This legislation reinforces the demand for a turnkey solution to control AI systems, especially for sectors with many high-risk AI systems such as the financial sector.

3. Anti-discrimination laws. Next to laws that specifically regulate the use of data or AI systems, there are many other laws on a national level that should be considered when using AI systems. Most apparent are anti-discrimination laws. As predictive models are typically trained on real-world data, they have the undesirable property to amplify potential biases that are encoded in the data. Therefore, compliance with anti-discrimination laws is not evident when employing AI systems. The most well-known anti-discrimination law

in the Netherlands is Article 1 of the Dutch Constitution. It dictates: "All persons in the Netherlands shall be treated equally in equal cases. Discrimination on the grounds of religion, belief, political opinion, race, gender, or any other grounds whatsoever is not permitted." [15]. This is a general statute that primarily prohibits the government from discriminating citizens. To better embed this right and ensure that it applies to both citizens and organizations, other laws such as the General Equal Treatment Act (*Algemene wet gelijke behandeling, AWGB*) from 1994 further elaborate on Article 1. The AWGB prohibits discrimination on the basis of nationality, race, origin, political belief, religion, gender and pregnancy, sexual orientation, and marital status. This act offers protection for those that are discriminated on the aforementioned grounds in the areas of labour (e.g., recruitment, selection, mediation, promotion, joining unions), social security (e.g., social benefits and student loans), and goods and services (e.g., housing, welfare, health care, culture, education, financial and insurance services) [16]. Mainly the latter affects businesses as they must ensure that their predictive models (i) do not use sensitive protected attributes such as race, (ii) do not use general protected attributes such as gender for their offerings, and (iii) are free of biases when using protected attributes such as gender for internal processes. However, this assumes differentiation based on clearly defined attributes. Legislation becomes more difficult when derived data points are involved, such as postal codes as a proxy for income. In a case from 2014, an insurance company based the premium of a life insurance on the postal code of a customer. Data from the Statistics Netherlands (*Centraal Bureau voor de Statistiek, CBS*) about the average income for each postal code was used. Since studies indicate that people with a higher income live longer - and therefore claim less for life insurance - the insurer charged a lower premium for customers that lived in a high-income neighborhood. Given the data from Statistics Netherlands that people with a non-Western migration background are more likely to live in a postal area with a lower average income, this policy also implied that customers with a non-Western migration background paid a higher premium. The Dutch College of Human Rights, which researches potential discrimination cases, substantiated with three reasons that direct differentiation based on income or indirect differentiation based on postal code is not prohibited for this insurance. First, the use of this data served a legitimate purpose, as it aligned the premium with the risk associated with the insurance. In addition, the approach was appropriate to reach the goal, since the differences in premiums were reasonable and no people were excluded from the insurance. Lastly, the means were necessary, as there was no alternative to the postal code that works equally well without discriminating. [17] This case study shows that the legislation for indirect differentiation is considerably more nuanced than that in the case of direct differentiation. Concluding, the anti-discrimination laws enforce that predictive models should be free of biases for protected attributes such as race, and that differentiation based on proxies such as postal code is only permitted if it has a valid ground.

4. Ethics Guidelines for Trustworthy AI. In 2019, the High Level Expert Group on Artificial Intelligence, an independent expert group formed by the European Commission, published the Ethics Guidelines for Trustworthy AI [18]. At the basis of the document are three complementing fundamental components (lawful, ethical, and robust) and four ethical principles (respect human autonomy, prevent harm, fairness, explicability) that the system should meet. Based on these components and elements, the expert group lists seven key requirements that AI systems should meet to be trustworthy:

- Human agency and oversight: AI systems should respect human autonomy and fundamental rights. For that reason, AI systems should be designed in such a way that allows for human oversight and intervention, e.g., with a 'human-in-the-loop' approach. Additionally, end-users must be able to comprehend AI systems and make autonomous decisions regarding the system.

- Technical robustness and safety: the results of AI systems should be accurate, reliable, and reproducible, especially when they highly affect users. The systems should be secure and resilient to attacks, both general software attacks and AI-specific. Lastly, fallback plans should be present in the AI systems in the event of problems.

- Privacy and Data governance: to protect users' fundamental rights, AI systems must protect user privacy throughout the entire lifecycle with a good data governance policy. The data used by the system should be of high quality and free of biases.

- Transparency: systems should be transparent to their users and thus inform them that they are interacting with an AI-driven system. Additionally, artifacts used for developing AI systems, such as data sets and algorithms, should be well-documented to enhance transparency, auditability, and traceability. Lastly, the explainability of an AI system should be taken into account during the development process. This implies that decisions made by high-impact AI systems should be traceable and be explained so that they can be understood by all stakeholders, including developers, regulators, and customers.

- Diversity, non-discrimination and fairness: inclusion and diversity must be taken into account during the entire lifecyle of an AI system. This includes the development of an accessible system, engaging stakeholders, and fostering a diverse organizational culture. Moreover, an essential requirement of a trustworthy AI system is that it is free of harmful biases, and for that reason there should be processes in place to analyze, address, and document the purpose, constraints, and requirements of the system.

- Societal and environmental well-being: an AI system's impact on society, democracy, social well-being, and environment should be considered during the entire lifecycle.

- Accountability: providers of AI systems should put mechanisms in place during the entire lifecycle to ensure responsibility and accountability for a system's outcomes. This includes identifying, assessing, reporting, and minimizing potential negative impacts of systems, for example, with frequent impact assessments. When encountering conflicts between the aforementioned requirements, these trade-offs should be documented and evaluated based on the fundamental components and ethical principles, or the development should be stopped if the system violates these principles. Finally, AI systems and their algorithms, data, and processes should be auditable by internal and external auditors to enhance the systems trustworthiness.

Additionally, the expert group presented an assessment list to aid the implementation of these requirements. These requirements are guidelines and therefore do not impact organizations directly. However, these guidelines are adopted by many other more binding documents, including the proposed EU Artificial Intelligence Act.

5. SAFEST AI Guidelines. In 2019, the Dutch central bank (*De Nederlandsche Bank, DNB)* presented general principles for the use of AI in the financial sector [19]. These general principles are grouped into seven areas, with the acronym SAFEST:

- Soundness: the central bank's primary concern is that AI systems should operate accurate, reliable, predictable, and lawful. Financial institutions should (i) ensure compliance with regulatory obligations, (ii) mitigate (financial) risks during the development of AI systems by involving domain experts, setting and documenting boundaries and fail criteria, and periodically retraining and reassessing systems, (iii) especially mitigate model risk for material AI systems using explainability, human oversight, and periodic evaluation of outcomes, (iv) safeguard and improve data quality and integrity by setting minimum requirements and putting constant efforts to ensure that data is correct, complete, representative and free of bias, and (v) be in control of procured and outsourced AI applications.

- Accountability: financial institutions should embed accountability for AI systems throughout the entire organization by assigning accountability and risks of AI systems at the board level, integrating the accountability in a risk management framework, and embedding accountability towards external stakeholders.

- Fairness: financial organizations should define a concept of fairness for their AI systems, take this into operation, and review the outcome of the system for unintentional biases using a human-in-the-loop process.

- Ethics: financial firms should a priori specify the ethical requirements, objectives, standards, and fallback procedures their AI systems should meet, and align these with their legal obligations, values, and principles.

- Skills: relevant skills, awareness, and understanding of AI should be present throughout the entire financial organization, including senior management, risk management, and compliance.

- Transparency: financial institutions should be transparent about their AI-related policies and decisions. This includes documenting shortcomings. Additionally, these organizations should constantly advance the traceability, reproducibility, and explainability of the outcomes of their AI applications. To improve the understanding of the internal working of a model, organizations could demonstrate how the input variables contribute to an individual outcome of the AI system (local explanation) or on an aggregated level (global explanation).

Although these principles share a common ground with the previously discussed international legislation and soft laws, these principles are specifically tailored for Dutch financial institutions and present the considerations and future direction of the central bank. It is therefore likely that the central bank, as a regulator of Dutch financial institutions, will actively regulate the use of AI in the Dutch finance industry based on these principles in the future.

6. Ethical Framework. The Dutch insurance sector has even more specific self-regulations given the Ethical Framework for Data-driven Applications (*Ethisch Kader Datagedreven Toepassingen*) [20]. This framework is proposed in 2020 by Dutch Association of Insurers (*Verbond van Verzekeraars*) and has been imposed on all members of the association as of January 1, 2021. The goal is that member insurers become more resilient to potential future laws and regulations through self-regulation. The framework unifies law, soft-law, and ethical aspects and has a broad scope, which must be further specified by the members themselves. It prescribes 30 standards that cover a broad and comprehensive scope that is in line with the operations of Dutch insurers, and covers all data-driven applications. These standards cover all aspects of the seven European Guidelines for Trustworthy AI.. Although the framework has a great overlap with the other discussed guidelines, it is a unique approach since it will be directly be enforced as it is only imposed on a small group of organizations. Another remarkable element of the framework is that it extensively incorporates ethical aspects. For those ethical standards, insurance firms commit not to perform certain actions, even if it is allowed by law or other sector-specific guidelines. An example is the acceptance of life insurance, for which insurance companies have the right to reject an application. This implies that they can decline high-risk applications such as those with serious diseases such as cancer. From a legal and risk management perspective, they can reject those applications, even if this means that the applicant can not get an insurance from any service provider. From an ethical standpoint, insurers should take their social responsibility as a financial services provider and make an effort to insure these vulnerable individuals as well. The challenge here is that ML algorithms rationalize a problem (e.g., minimize the risk) and that ethics is hard to define and quantify. For that reason, it is paramount that an AI system's outcomes are explained to the humans that oversee the process so that they can adjust the algorithm if deemed necessary. For ethical aspects in general, it is crucial for financial institutions to take this dimension into account as ethical violations can harm the reputation of the organization or the financial industry as a whole. Moreover, ethical standards are gradually shifting: self-regulatory guidelines often become laws in a matter of years, and ethical standards are incorporated in new self-regulatory guidelines. Direct action therefore has a great advantage. Lastly, financial service providers fulfill a pivotal role in society and should take their societal responsibilities accordingly. The Ethical Framework for Data-driven Applications will be further addressed in Section 3.3.

The aforementioned examples give an overview of the laws, soft-law, and ethical aspects that should be taken into account when employing AI applications. Implementing the regulations covers many legal risks that are associated with AI. However, note that explainability not only services AI compliance, but also benefits model evaluation and model refinement - highly relevant for high-impact systems.

Some handouts to implement data ethics in development processes are currently available. Examples are the Data Ethics Decision Aid [21] from Utrecht Data School, Utrecht University and the municipality of Utrecht. On behalf of the Dutch government, a team of researchers from the universities of Tilburg, Eindhoven, and Brussels, composed a handout with the steps both public and private organizations should undertake when developing AI-driven systems [22]. Despite these handouts, there is still a need for complete tools to execute the steps stipulated in the hand-outs, such as measuring biases and explaining the inner workings of models.

2.2 Explainable AI

This section provides background information on the field of explainable artificial intelligence with the goal to aid making design decisions that are grounded in literature. First, definitions and a taxonomy of the field are drawn up. Then, the typical actors of XAI are listed. This section is concluded with numerous examples of explainable AI insights and the techniques that generate these insights.

Definitions. The field of explainable AI focuses on researching, developing, and evaluating techniques and best practices to make the inner workings and outcomes of artificial intelligence understandable for humans. It is typically used to address the drawbacks of black-box models. Black-box models are AI systems for which the input and output can be observed, but the internal mechanisms that convert these inputs into an output remain unclear. These systems lack transparency, implying that even for the developer it is ambiguous how the system is established based on the development process and decisions such as parameters settings. Examples are neural networks, for which the developer cannot describe and motivate the inner workings of the system based on the input data, parameters used, and other design decisions. The opposite of black-box models are white-box, of which the internal working is transparent. An example is a linear model, where the outcome of the model is the weighted sum of the input features. The inner workings of the model can be easily examined by looking at the feature weights. White-box models are often referred to as interpretable models. Interpretability implies that the internal mechanisms of the model can be extracted and presented in such a way that it is understandable to humans, i.e. "the higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made." [23]. The third pillar of explainable AI is enhancing the explainability of a model. This revolves around communicating an explanation about the outcome in such a way that the reasoning of the model can be comprehended by the user. A commonly used approach is displaying which inputs have the most predictive value for a particular prediction, instead of just displaying the outcome. A crucial characteristic is that this explanation can easily be understood by the users - which varies between different actors. The requirements of a good explanation and how this differs per actor will be covered later. In academia, many definitions are given for transparency, interpretability, and explainability. The three definitions are often used interchangeably. However, the difference is that transparency mainly focuses on explaining the implementation process and how the algorithm establishes the model, interpretability focuses on explaining the technical and internal working of a complete model, and explainability focuses on elaborating on the outcome, often a single decision.

Taxonomy of XAI. The field of explainable AI includes many techniques. Figure 2.1 presents some criteria that can be used to classify these techniques:

- *Can it explain a particular model or many models?* Model-specific techniques access model parameters for interpretation. These methods therefore have the advantage of being optimised for specific algorithms and architectures (e.g., trees or neural networks), at the expense of broad applicability. On the other hand, model-agnostic techniques interpret models by analyzing their inputs and outputs instead of model internals, making them widely applicable yet less optimized.

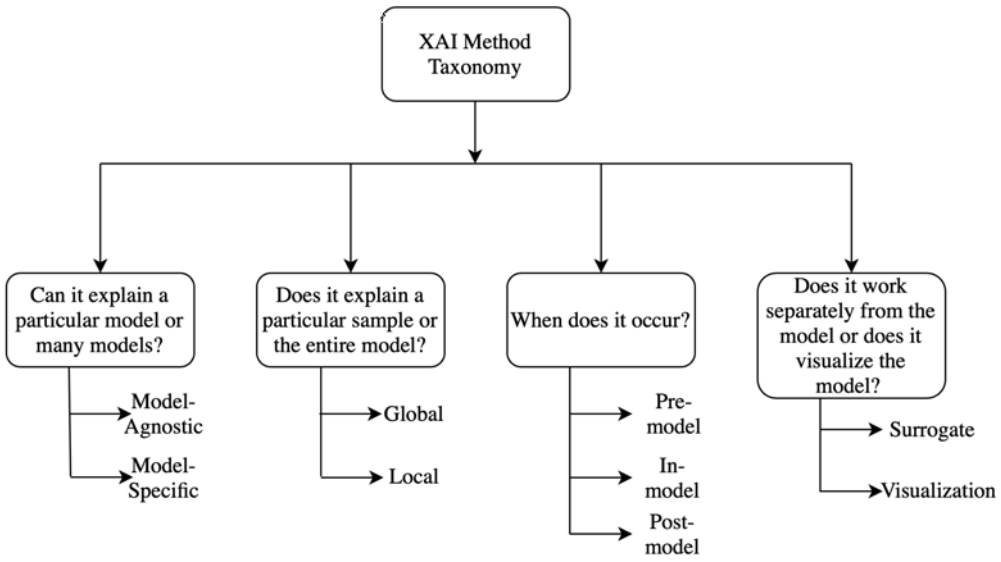


Figure 2.1: Taxonomy of XAI methods [24].

- *Does it explain a particular sample or the entire model?* Global explainability techniques interpret and explain the working of an entire model. How the model arrives at the decision of a particular instance can be explained using local explainability techniques. Examples are LIME [8] and Anchors [25] that display the most influential features. Aggregating a high number of local explanations can lead to a global understanding of the model, such as employed by SHAP [9].

- *When does it occur?* Model transparancy can be enhanced at multiple stages in the development process. Before training a model, pre-model techniques such as Principal Component Analysis [26] and t-Distributed Stochastic Neighbor Embedding [27] can be applied to get a better understanding of the data. When employing an algorithm that is interpretable by itself or has an built-in interpretability mechanism, we speak of in-model or ante-hoc techniques. After training, the learned patterns of models can be interpreted using post-hoc techniques. Model-agnostic techniques are typically post-hoc.

- *Does it work separately for the model or does it visualize the model?* Numerous XAI techniques can be separated based on the underlying methodology. Models can be interpreted by just visualizing their internal workings or by creating a new surrogate model that is more interpretable than the original model. An example of the latter is LIME [8], which builds a linear model locally around the data point of interest.

Target audience. Figure 2.2 presents the target audience (actors) of XAI systems and their typical needs:

- Data scientists and others involved in the development process want to understand, debug, and improve models. They are typically interested in the complete model, thus demand global explanations.

- Users that are affected by model decisions - typically customers - demand transparency about the decision-making process so that they understand their situation and can verify the fairness of the decision-making. Their concerns are mainly centered around the predictions that have an impact on their situation, so demand local explanations for these predictions.

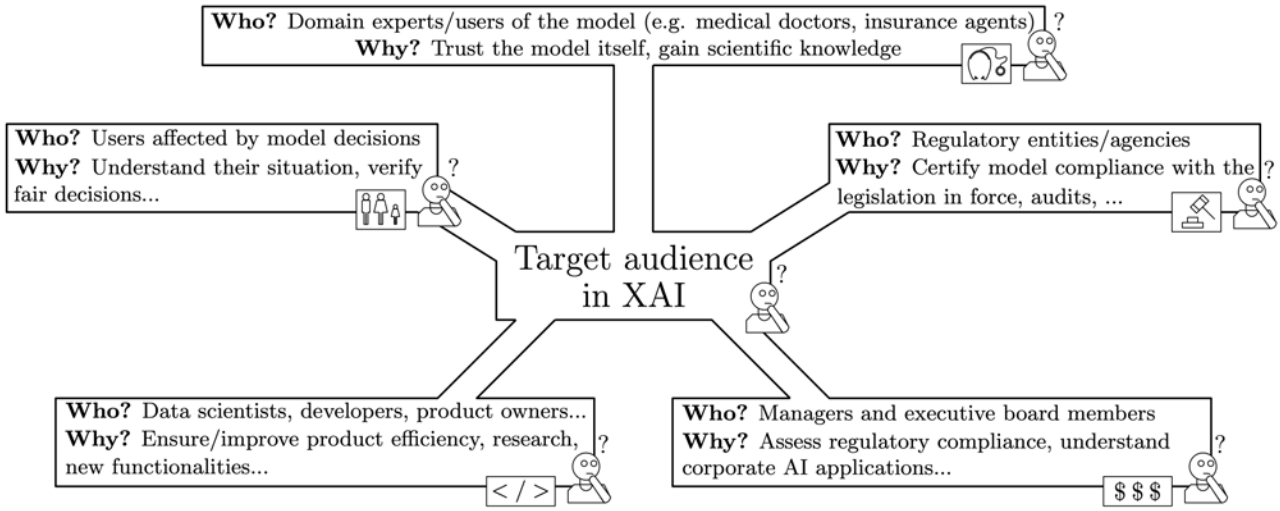


Figure 2.2: Target audience of XAI systems and their needs [28].

- End-users and domain experts need explanations to assess whether they can trust the model, which is crucial for good interaction with the system. Additionally, they can gain domain knowledge from explanations. This can either be on a local or global level.
- From a regulatory perspective, understanding of the model is needed to (i) verify compliance with regulations, (ii) inspect the reliability and robustness, and (iii) assess the impact on the customer. This can be executed both internally by business- (first line) and operational control (second line) or externally by auditors (third line) and regulators [29]. These actors are typically interested in the complete model, and hence demand global explanations.
- Managers and executive board members want to assess a system's regulatory compliance, and understand the corporate AI applications and how they align with the corporate strategy. Closer to the development, business owners require model understanding to verify if it fits the intended purpose and give approval for use [30]. These actors are typically interested in global explanations.

The aforementioned needs summarize the importance of model interpretability for high-impact AI systems.

Model-agnostic explainability techniques. Due to their wide applicability, post-hoc model-agnostic techniques fit the objective of this research. To get a better understanding of this field, some popular techniques for shallow (non-deep learning) ML-models will be addressed hereafter. Figure 2.3 presents a breakdown of the model-agnostic field into five explainability categories:

- **Explanation by example.** A simple way of explaining the inner workings of a ML model is by presenting representative examples ("prototypes"), e.g., the "average" instance for a specific class of a classification model or the most influential instances for training. This technique requires human interpretation to be informative and does not reveal the influence instance characteristics on the model outcome.
- **Explanation by simplification.** In the search for higher performance, the complexity of machine learning models can rapidly grow beyond the human comprehension. Examples are ensemble models such

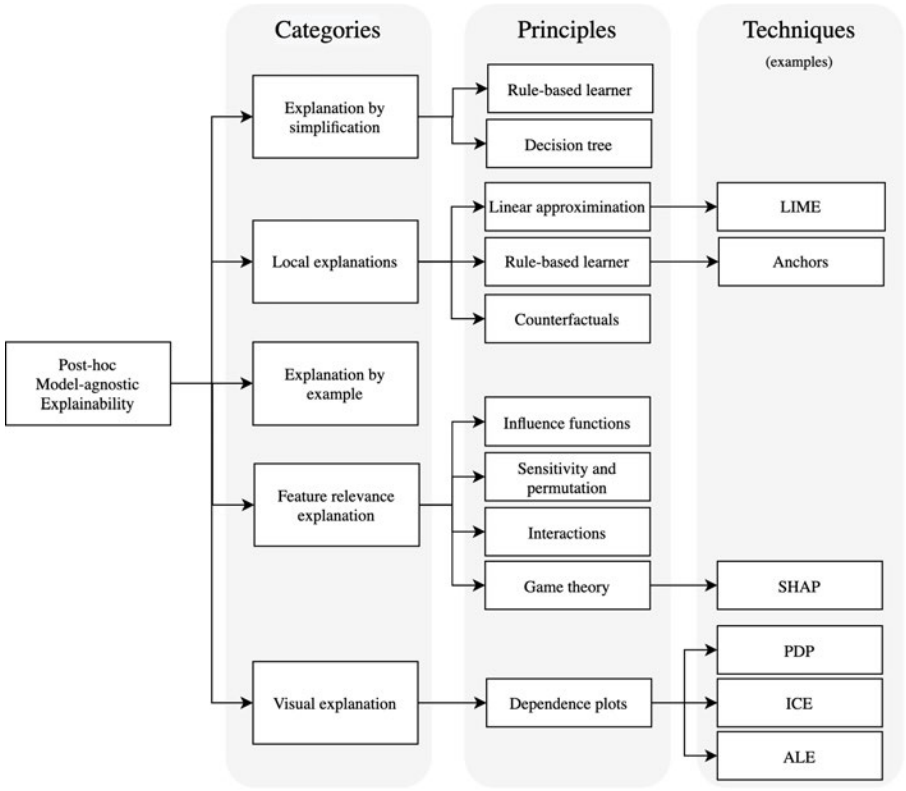


Figure 2.3: Model-agnostic explainability categories, principles, and techniques. Adapted from [28] and [30].

as Random Forest and Extreme Gradient Boosting, which construct hundreds of different decision trees that are collectively used to make predictions. The inner workings of these models can be explained by creating a less complex - and therefore easier to explain - model based on the original one. For example, an ensemble of decision trees can be reduced to a single decision tree or a rule set, which can then be visualized. The disadvantages of these techniques are that the simplified model can still be too complex to understand, or too simplified that it hardly captures the original model.

- **Local explanations.** Machine learning models are often complex due to the variety of examples the model should work with, also known as generalizability. It can therefore be helpful to limit the scope of an explanation to the behavior of the model on a single or for several instances. Local explanation techniques often use the concept of model simplification, which is more effective on the less complex local level.

  A well-known technique is Local Interpretable Model-Agnostic Explanations (LIME) [8]. LIME can explain single predictions of any classifier or regressor in a faithful way by approximating the opaque model locally with an interpretable model. Given the prediction of interest, LIME uses the original model to generate synthetic data around the prediction, on which interpretable models such as logistic regression and decision trees are trained. The model that is most faithful to the original model (i.e., has the smallest deviation in performance measures) can then be used to explain the prediction of interest. Since this resulting surrogate model is interpretable, model characteristics such as feature weights can be extracted to explain the most influential features for the given prediction. LIME is valued for easy-

to-understand and locally faithful explanations [23]. Disadvantages are, however, that models cannot always be captured linearly and that the explanations only hold for a specific instance and can therefore not be extrapolated to other instances.

The authors of LIME provided a solution in a later work, called Anchors [25]. The technique once again tries to approximate the model locally, but returns a rule set as an explanation instead of a linear model. Given a prediction, the returned "if-then" rules describe the instances for which this prediction (almost) always holds. For example, if the loan amount is below 50k and the applicant is full-time employed, then the loan is always accepted - all other feature values do not matter. This makes it easier for users to generalize the results. The disadvantage of Anchors is that the explanations can only be examined on a local level, and not aggregated into a global level such as LIME.

The third local explainability principle is that of counterfactuals. For a given instance, counterfactual techniques explain what the minimum change to that instance should be to fall into another (more desirable) category. For example, if a loan application of a customer is rejected, a counterfactual explanation shows which characteristics of the application should change in order for the application to be accepted. This principle provides insightful explanations, mainly for customers, but does not generalize.

Overall, local explainability techniques can provide insights into the model behavior for a specific instance and area of interest, but cannot be easily generalized to other instances or to a global level.

- **Feature relevance explanations.** A frequently used explainability category is that of feature relevance. It provides insights into the inner workings of a model by displaying the relevance of each feature for the model (globally) or for a single prediction (locally). How 'relevance' is defined differs per principle.

  Techniques based on influence functions compute the relevance of features by measuring the influence of a data point on the training parameters. A data point is influential if removing or upweighting the data point leads to a significant change in training parameters. Based on the characteristics of these influential data points, the most relevant features can be determined. Sensitivity and permutation based techniques compute the relevance of features by measuring the change in prediction uncertainty or error if the input features are changed. Changes of relevant features lead to bigger changes in the predictions than irrelevant features. Interaction based techniques measure the relationships and dependencies between features to compute their relevance, where features with strong relations have a higher relevance.

  Lastly, there are feature relevance techniques based on Shapley values from cooperative game theory. Shapley values, introduced by Lloyd Shapley in 1951 [32], can be used to fairly assign the contribution of each player to the total surplus of a cooperative game. A popular implementation of Shapley values to compute feature relevance is SHAP (SHapley Additive exPlanations) [9], proposed by Lundberg and Lee in 2017. SHAP models the prediction of a single instance as a game, where each feature of that instance is a player in the game. A player in the game can also be a group of features or a group of pixels in the case of image classification. By predicting the output for each possible feature combination

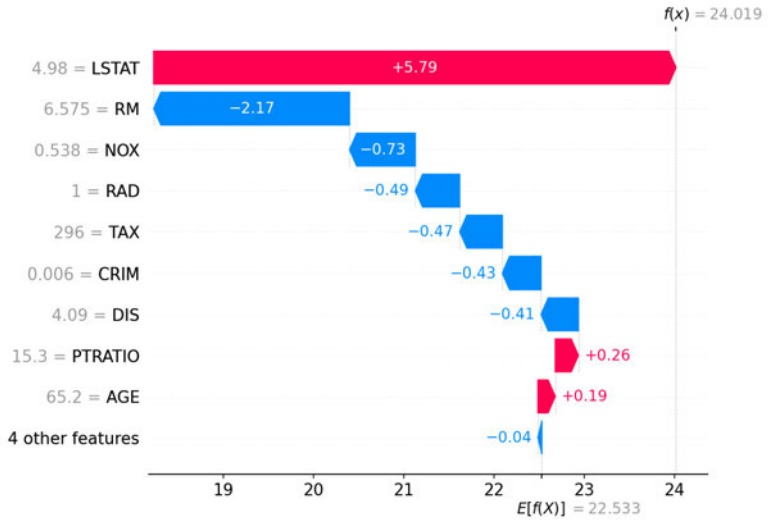


Figure 2.4: A feature contribution explanation using SHAP. 'LSTAT' has the largest positive (+5.79) and 'RM' the largest negative marginal contribution (-2.17) to the difference between the base value (22.533) and model output (24.019). [31]

(where features can be absent; the values for all absent features will be replaced with random values from the training data), the contribution of each feature to the output can be computed. This results in a *SHAP value* for each feature: the average marginal contribution of the feature value for all possible feature combinations. SHAP values have an additive nature, meaning that the sum of all SHAP values for a given prediction is the difference between the base value (the model output if all feature values are unknown, i.e., the average prediction in the training set) and the prediction output, as shown in Figure 2.4. This is the local accuracy axiom of Shapley values. In addition, SHAP inherited five other mathematical properties from Shapley values: missingness, consistency, linearity, dummy, and symmetry. The challenge of Shapley values is that exact computation requires predictions for $n!$ combinations (the ordering of features matters due to interactions) of features for a model with n features, which is infeasible in real-world applications. Therefore, other techniques that are based on Shapley values draw a sample from the $n!$ number of combinations to estimate the Shapley values, leading to stability issues. SHAP overcomes this challenge by approximating Shapley values using other feature relevance techniques. Six feature relevance techniques are unified using these mathematical properties by SHAP, including the model-agnostic LIME [8], and DeepLIFT [33] and Layer-Wise Relevance Propagation [7] for deep learning networks, ensuring that they have the same mathematical foundations. SHAP has both model-specific and model-agnostic implementations that leverage these techniques. The model-agnostic KernelSHAP uses the weighted linear regression of LIME to approximate Shapley values, leading to lower variance and higher computational efficient than other techniques. A key advantage of KernelSHAP over LIME is that local explanations can be compared to global explanations (aggregated local explanations) since both levels use the same atomic unit: Shapley values. However, SHAP is considerably more computationally heavy than LIME, as it scales exponentially with the number of features. To make the technique applicable to high-dimensional (real-world) models, the authors assumed feature independence, and ignored feature ordering. The SHAP implementation provides many plots to visualize the SHAP values, both on a local and global level.

In summary, there are many techniques to compute the relevance of features, both during training and predicting. The resulting explanations are intuitive, but lack details.

- **Visual explanations.** Techniques that explain complex relations in the model, such as the relationship between features and the model output, typically employ this category to make the explanation easier to understand. An example is the Partial Dependence Plot (PDP) [34], that shows the marginal effect a feature has on the predicted outcome of a model. It works as follows. For a given feature, it defines a grid of possible feature values. For each point in this grid, all instances in the dataset are forced to take this feature value, and then make predictions on these adjusted data points to capture the average model outcome. The relationship between the feature values and the average model outcomes are plotted. Partial Dependence Plots are intuitive and therefore easy to understand. One of the disadvantages of PDP is that heterogenous effects (such as two contrary clusters) might be hidden as it only shows the average marginal effects. This disadvantage is solved by Individual Conditional Expectation (ICE) curves [35]. The technique behind ICE is similar to PDP: it plots the relation between the feature values and model output, but plots a line for each instance individually instead of the average over all instances. ICE plots allow the identification of instances for which the model behaves similarly, but the detailing makes it harder to interpret the plots. The main limitation of ICE and PDP is that they assume feature independence, which is often not the case for real-world models. As an example, consider a model that uses a dataset with characteristics of persons such as height and age, for which we generate the PDP for the feature 'height'. For all instances in that dataset, PDP will replace the height with values from the grid, regardless of the other features such as the age of that instance. This can lead to the creation of artificial instances that do not exist in the real world, such as a person of age five with a height of 200 centimeters. Another visual explanation technique that does not assume feature independence is Accumulated Local Effects (ALE) [36]. For dependent features, ALE creates more realistic artificial instances by sampling feature values based on seen distributions (e.g. 80 - 120 centimeters for a person of age 5), and is therefore more reliable for models with dependent features. The main disadvantages of ALE are that it might suffer from stability issues and is harder to interpret, especially for laymen, since it displays the effect of a feature on the model output instead of the complete model output.

  In general, the main advantage of the aforementioned techniques is that it displays more relations than other explainability techniques such as feature relevance. The primary limitation of visual explanations is that the human brain can only comprehend three dimensions, and two dimensions are even more desirable.

As previously mentioned, there currently is not a single technique that can be applied to all cases. The selection of the most appropriate technique depends on the target audience, use cases, and models at hand. For our system, we will address this selection in Chapter 4.

Chapter 3

Context

This chapter describes the context in which this study is conducted. The research was performed during an internship at Achmea, a Dutch financial services provider. Achmea is an appropriate host organization for this research since it is: (i) a large organization, (ii) strongly digitized with an increasing focus on ML, and (iii) in a heavily regulated industry, and therefore in need of a turnkey XAI solution. Section 3.1 describes this organization, the business activities, and the department that hosts the internship. How AI and ML are applied at Achmea is outlined in Section 3.2. Sector-specific AI regulations and how Achmea complies with these regulations is covered in Section 3.3.

3.1 Organization

Achmea's roots go back to the Dutch province of Friesland in 1811, when a group of farmers decided to be jointly responsible for business risks such as fire. Over two hundred years - and several mergers and acquisitions - later, Achmea serves 10 million customers in the Netherlands with ten brands. With a revenue of 20 billion euro in 2020 [37] and a market share of approximately 25% [38], it is the largest insurance company in the Netherlands. While insurance remains the main activity, Achmea offers several other financial services, including banking and mortgage products, pension administration, and asset management. In 2020, it managed a total of 227 billion euros in assets [37]. The group also operates abroad - in Australia, Canada, Greece, Slovakia, and Turkey - where it employs 2,500 people. In the Netherlands, Achmea employs 13,300 FTE within various divisions. Some divisions are directly linked to one of the major brands, such as *Interpolis* (insurances), *Centraal Beheer* (insurances and other financial services), and *Zilveren Kruis* (health insurances). Other divisions work on groups of white label products, e.g. non-life insurance and income protection insurance, that are distributed through different Achmea brands. To reduce overhead and increase consistency between these divisions, generic services are provided by shared service centers, such as Achmea IT.

The division Achmea IT consists of numerous business units that handle the distribution of IT to specific brands, or focus on generic IT topics that are relevant to all divisions. An example of the latter is the Data Expertise Center, which consists of approximately 60 data experts that facilitate data use in the organization.

This section has been redacted due to company confidentiality reasons.

Figure 3.1: Examples of use cases of AI and ML in insurance.

3.2 Application of AI and ML

To adapt to the changing environment and stay ahead of competition, there is a program active within Achmea with the aim of becoming a more digital insurer by 2025. From the perspective of data analytics, part of this program is to improve the data maturity of the organization and make more use of predictive analysis. As previously stated, predictive analysis using machine learning allows the automation of complex tasks that can barely be captured with hand-crafted rules. To give an understanding of the types of tasks that can be automated in the insurance industry, the five use cases as presented in Figure 3.1 are explained hereafter. The remainder of the section describes how Achmea implements and controls AI.

First, a frequently given example of machine learning in the financial sector is automating the acceptance process. In the current 24/7 economy, customers want to be offered an insurance within minutes, even in the evenings and weekend. Automating this process is therefore essential to meet both customer demands and keep operational costs under control. Machine learning offers a solution for products for which it is hard to establish the criteria that applications should meet to be automatically accepted. Once a model has been trained and put into production, it can provide a decent risk assessment within seconds. A related example is the current trend in insurance of dynamic pricing. With dynamic pricing, the premium of the insurance is based on the associated risk, e.g., that the premium of home insurance is higher for properties in neighborhoods with an above-average number of burglaries. This gets more interesting when it is based on the individual actions of a customer, such as deriving the premium of car insurance from driving behavior. To apply this in a fair manner, a significant amount of data is required, where nuanced patterns may only be extracted with machine learning algorithms. The third use case of machine learning in insurance is applying them with the aim of fraud detection. ML models are more capable of identifying potential fraudulent claims than traditional rule-based models [39]. In 2020, Dutch insurers collectively detected 13,000 cases of fraud and thereby reduced the cost of claims by 88 million euro [40]. Machine learning also has great potential for accelerating claim management and other internal processes. Within Achmea, several ML-driven models are used for automatically routing claims and other notices to the right department, based on claim characteristics or patterns found in unstructured data such as text and images. Hereby, the file directly reaches the employees with the right competences and authorization, reducing the throughput time of the file without manual interference. The last use case of ML that is particularly interesting in the financial sector is to assess the risks of a complete portfolio. As models can predict the risk associated with one case, e.g., for automatic acceptance or dynamic pricing, it can also assess the risk of an entire portfolio by aggregating these results. This information can be leveraged for reporting purposes, or to make data-driven decisions on matters such as provisions. Once again, the benefit of ML-based models over rule-based models grows as the task gets more complex or more data is employed. Next to these five use cases that are particularly interesting for the insurance industry, there are numerous applications of AI and ML that do not only apply to insurance, such as targeted marketing. In summary, there are plenty of use cases of AI and ML in insurance to automate processes and thereby improve customer satisfaction and offerings, reduce throughput time and operational costs, and enhance reporting and risk management.

This section has been redacted due to company confidentiality reasons.

3.3 AI Compliance

The Dutch insurance industry - and thereby Achmea - is heavily-regulated, with an increasing focus on the regulation of AI and ML. This section describes the legislation Achmea is subject to, the current strategy to mitigate the risks associated with these legislations, and the current gaps.

Section 2.1 lists three legislations and guidelines that regulate the use of AI in the Dutch insurance industry. The anti-discrimination law and GDPR are active for a considerable amount of time, whereby Achmea is already compliant. Nevertheless, constant efforts are required to focus on the AI-side of these regulations. On the other hand, the guidelines - the Ethics Guidelines for Trustworthy AI, the SAFEST AI guidelines, and the Ethical Framework for Data-driven Applications - and the proposed EU AI Act are relatively new and hard to implement due to their broad scope. Even though these guidelines and legislation are currently not legally binding, it is advisable to implement before they come into force (EU AI Act) or before they are enacted into legislation (the three guidelines).

A remarkable example is the Ethical Framework for Data-driven Applications from the Dutch Association of Insurers [20]. As of January 1st, 2021, the Ethical Framework has become part of the self-regulation framework of the Association, along with nine other regulations. The document is thereby binding for all members of the Association. An independent foundation will perform audits on compliance with these regulations as of 2023. Although this approach is still not legally binding, it definitely is not non-committal. Non-compliance can even harm the reputation of insurers such as Achmea, the largest member of the association. From here on, we will further focus on this framework as it is concrete and intertwined with other regulations.

This broad framework unifies law, soft-law, and ethical aspects into 30 standards for all data-driven applications. This includes data security, application robustness, risk assessment methodologies, data quality, data governance, the use of sensitive data, inclusion, internal and external communication (transparency), training, and awareness. The following standards regarding the control and transparency of AI applications are listed in the framework [20]:

#15 In practice, the use of data-driven applications always takes place under adequate human supervision and responsibility, for example by retraining AI where necessary.

#16 New techniques will first be tested in a familiar setting, to see whether margins of error and other risks increase compared to alternative methods and processes.

#19 When violations of fundamental rights, including unjustified discriminatory bias, cannot be avoided or excluded in data-driven applications, insurers will not deploy an application.

#20 When opting to use data-driven systems, insurers pay attention to diversity and inclusiveness, especially for people at risk of exclusion or disadvantage due to special needs and/or a disability.

#23 Insurers provide an internal control and accountability mechanism for the use of data-driven applications and the data sources used.

This section has been redacted due to company confidentiality reasons.

To summarize, Achmea is fully committed to leverage AI and ML for numerous high-potential use cases, ultimately to become a more data-driven insurer, but is challenged to comply with new legislation in this area. To comply with the five aforementioned standards, a turnkey explainable AI toolkit is required that is grounded in the regulatory requirements to support governance processes. By standardizing model explainability, it should enable the decentralized data science teams to comply with the control and transparency standards of the Ethical Framework.

Chapter 4

System Design

This chapter describes the design of the XAI system. First, the design criteria of the system are defined in Section 4.1 to meet the first objective of this study. Section 4.2 then elaborates on the XAI techniques that are selected to generate the insights. Finally, Section 4.3 translates these design decisions into a conceptual architecture of the system.

4.1 Design Criteria

This section starts with defining the general properties of the system in Section 4.1.1. The stakeholders and their intended use of the system are then listed in Section 4.1.2. At last, Section 4.1.3 describes the types of model explanations and other information the system should provide to serve these stakeholders in executing their tasks.

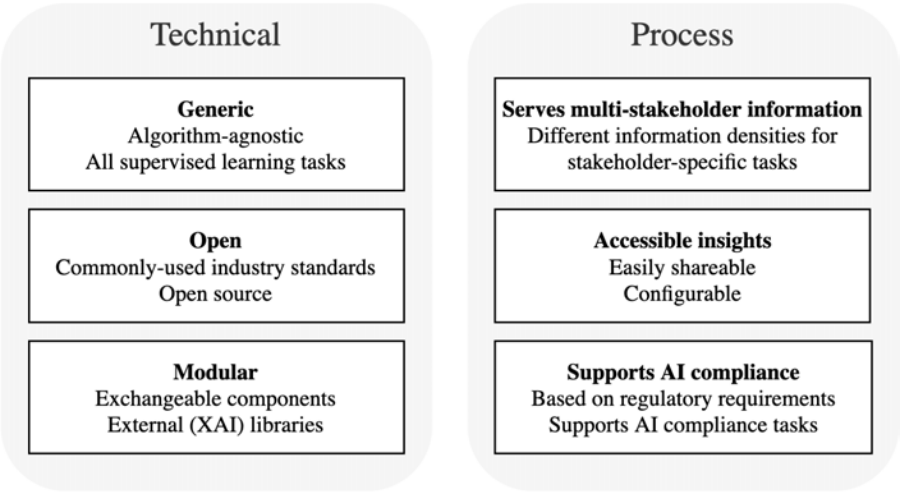


Figure 4.1: The general properties of the system ordered along two dimensions.

4.1.1 General Properties

Figure 4.1 displays the six general properties the system should have to be a turnkey system for AI compliance in the financial sector. Recall that by turnkey we mean that the system can directly be implemented in the organization without cumbersome migrations and configurations, and have such characteristics that it can be applied by the vast majority of the organization. We ordered the properties along two dimensions.

From a technical point of view, the system should be:

- **Generic.** The system should be generic to allow application to a wide range of models. It should work regardless of the algorithm or the supervised learning task (binary classification, multiclass classification, and regression) at hand, as long as the input data is structured. This scope ensures that the system can be applied to the vast majority of the models employed in the financial sector.

- **Open.** To be language- and vendor-independent, the system should support open source and commonly-used industry data and model formats such as Predictive Model Markup Language (PMML) [42]. This ensures that the system can be applied to the vast majority of models in an organization. Additionally, the system should eventually become an open source library by itself, so it can be easily implemented within other organizations by simply installing it to their development environment.

- **Modular**. The system design should be modular to ease maintainability. It must consist of exchangeable components, so that, for example, an obsolete XAI technique can easily be replaced with the state-of-the-art. The system should use proven external open source XAI libraries to improve user and managerial confidence in the system.

Given the aim of specifically supporting the AI compliance process, the system should:

- **Serve multi-stakeholder information.** To ensure that the system can be used by the majority of the organization, it should supply a variety of stakeholders with the information required for executing their specific task. As the stakeholders have different backgrounds, the served information should have the appropriate information density for each stakeholder. All stakeholders involved and their information needs are discussed in Section 4.1.2 and 4.1.3, respectively.

- Have **accessible insights.** To enhance acceptance and user satisfaction, the system's insights should be accessible, mainly for non-technical stakeholders. The insights produced by the system should be portable so that it can easily be shared between stakeholders. Additionally, the system should be configurable in an easy manner so that there is sufficient contextual information presented for stakeholders that are not involved in the development process. In general, we prioritize user friendliness for non-technical stakeholders over sophisticated technical solutions.

- **Support AI compliance.** In contrast to many XAI systems that are designed for data scientists, this system should be tailored to effectively support the AI compliance process. This means that the provided insights are based on the requirements of AI regulations (see Section 2.1) and support the AI compliance tasks.

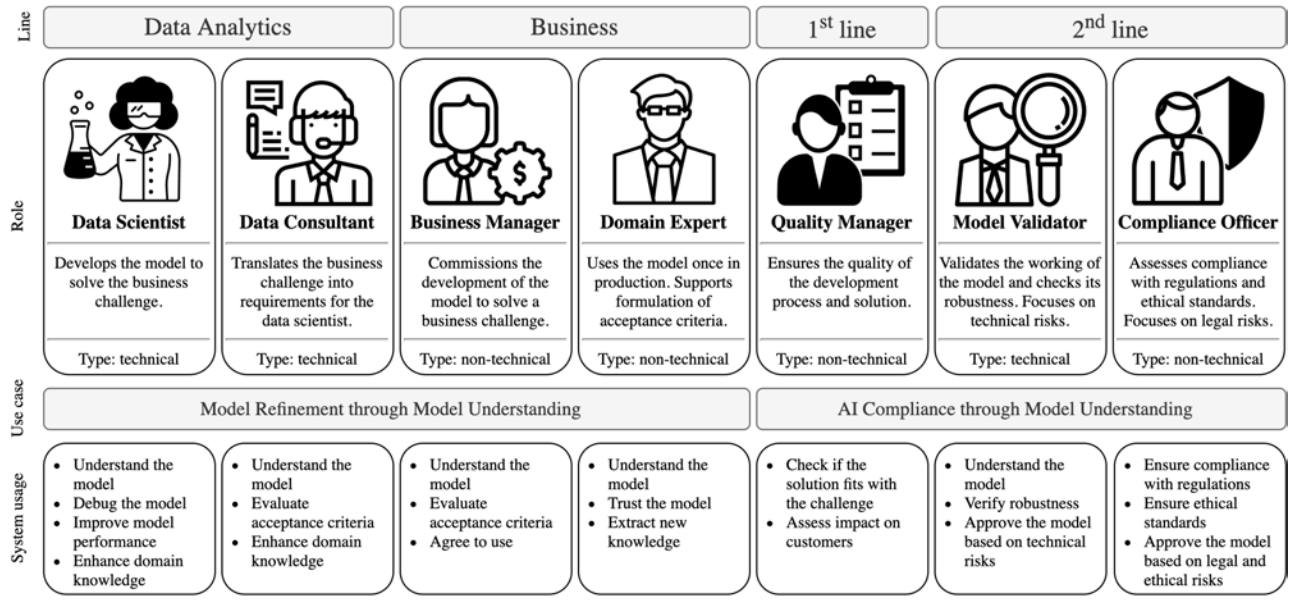


Figure 4.2: The stakeholders of the XAI system, arranged from left to right based on the sequence they interact with the model during the development. The boxes from top to bottom: the business line, the role description, the primary use case of the system, and how the system can support their tasks in the development process.

4.1.2 Stakeholders

This section describes the stakeholders of the XAI system to effectively select the explainability insights based on stakeholder needs. Figure 4.2 displays the seven stakeholders of the XAI system. These stakeholders and their needs are identified based on interviews and existing documentation on stakeholder needs for model management. As the system will be used internally, all stakeholders are internal. However, external stakeholders such as auditors, regulators, and customers will benefit from the system as internal compliance leads to better external compliance. See Section 2.2 for an overview of typical stakeholders of XAI systems, including external stakeholders.

The stakeholders of the system are:

- **Data Scientists** and other data science practitioners develop machine learning models to solve business challenges. After building the first version of the model, they want to understand the model's inner workings and use these insights to debug and improve the model. Additionally, by inspecting the learned patterns, data scientists want to enhance their domain knowledge, which in turn can help with tasks such as feature engineering. To do so, they need detailed insights into complex relationships.

- **Data Consultants** bridge the knowledge gap between the business manager and data scientist by translating the business challenge into technical requirements. They want to understand the model to properly present the solution to the business and evaluate whether it meets the acceptance criteria. Data consultants also seek for more domain knowledge to better fulfill their role. They need both detailed insights for their own knowledge, and a condensed version to present to non-technical stakeholders.

- **Business Managers** commission and fund the development of a model to solve a challenge in the business line they are responsible for. When data consultants present the proposed model, business managers need insights in the workings of the model to understand if it solves the business challenge at hand. Based on the insights, the model may be further refined by the data scientist. When a final version is presented, business managers require model insights to evaluate the acceptance criteria in order to decide whether to bring the model into production. For this, they need insights on a main level, including performance metrics.

- **Domain Experts** may be involved in the development process to provide input for the acceptance criteria since they will use the model once in production. Once the solution is presented, model understanding is paramount for domain experts to trust the system [8, 9, 43] and evaluate the acceptance criteria. Besides, domain experts want to extract knowledge from the model by interpreting the learned patterns. They demand detailed insights in complex relationships of the model, accompanied with background information on these explanations as they are non-technical stakeholders. The primary goal of the business manager and domain expert is to realize a better solution for their challenges by providing valuable feedback so that the data scientist can refine the model.

- **Quality Managers** oversee the quality of the development process and the proposed solution. They want a global understanding of the model to check if it fits the challenge. From a quality perspective, they value a transparent development process during which extensive efforts have been made to interpret the model's working. In addition, quality managers want to assess the impact of the model on the customer, and are thereby primarily interested in model bias. The quality manager is a non-technical stakeholder who demands concise insights and sufficient context.

- **Model Validators** and others involved in model risk challenge high-risk models before they can go to production. They validate the inner workings and therefore require a good understanding of the model. To verify the reliability and robustness of the model, model validators use specialized tools. They approve the use of the model if the technical risks are acceptable according to AI regulations. The model validator is a technical stakeholder that requires detailed insights in complex relationships with sufficient context about the model.

- **Compliance Officers** and privacy officers approve continued use of the model in production if the legal and ethical risks are manageable. For these assessments, they need insights regarding possible proxies (non-causal relationships) and potential biases. Compliance officers are non-technical stakeholders and therefore demand concise insights with sufficient context.

Based on the stakeholder interview, two desired primary use cases are identified. First, by understanding the inner workings of a model, it can be refined and further optimized - a key objective for data analytics and business. Second, by understanding the inner workings of a model, the compliance with AI regulations can be assessed - mainly relevant for the first and second line of compliance.

Table 4.1: The importance of each explanation type for each stakeholders to perform their tasks, expressed with tick marks (0 = no importance, 1 = low importance, 2 = medium importance, 3 = high importance).

Explanations	Data Scientist	Data Consultant	Business Manager	Domain Expert	Quality Manager	Model Validator	Compliance Officer
Feature relevance	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✓	✓✓✓	✓✓✓
Feature insights	✓✓✓	✓✓✓	✓	✓✓✓	✓	✓✓✓	✓✓
Feature interaction	✓✓			✓		✓✓	
Bias insights	✓	✓	✓	✓	✓✓	✓	✓✓✓

4.1.3 Explanations and Information Needs

The stakeholders, their role in the development process, and how the system can support their work are described in the previous section. This section elaborates on the necessary explanations and other information needs the system should provide to serve the stakeholders, resulting in a tangible list with design criteria.

Explanations From the stakeholder interviews, four desired types of explanations and their importance for each stakeholder were identified (Table 4.1):

- **Feature relevance**. The relevance of a feature for the model is a frequently used explanation in academic literature (see Section 2.2). All stakeholders have mentioned the feature relevance as a first means to enhance their understanding of the model as it is a straightforward explanation. These insights are relevant for both model refinement and AI compliance. For model refinement, feature relevance can be used for gaining a better understanding of the model and the domain, debugging the model by detecting non-causal relationships, and improving the model by finding inspiration for feature engineering. Feature relevance explanations help AI compliance by easing the detection of non-causal relationships, and identifying features that can be removed. Due to these use cases, we focus on global feature relevance explanations, thus the most relevant models for the entire model. This form of aggregated feature relevance is sufficient for business managers, quality managers, and compliance officers that prioritize simplicity over detail. Data scientists, data consultants, domain experts, and model validators desire additional details through aggregation at different levels (explanations for a region of the model, e.g., the most relevant features for a specific range of model outputs) and the preservation of instance-level to identify clusters and dispersions.

- **Feature insights**. All stakeholders demand some degree of feature insights to supplement the feature relevance insights. Frequently mentioned is the relation between the values of a feature and the output of the model, because this is not apparent from the feature relevance insights. The stakeholders are mainly interested in this insight for 5 to 10 most relevant features. These dependence plots enhance the understanding of the model, and thereby support both model refinement and AI compliance. For data scientists, data consultants, and model validators, detailed dependence plots are essential to be able to inspect complex relationships. Domain experts are particularly interested in enhancing their domain knowledge with these insights. Compliance officers are interested in these insights to identify non-causal relationships and potential biases. The latter two stakeholders are less technically inclined and therefore demand a clear representation of the relationships.

- **Feature interaction**. A few stakeholders have a minor interest in the interaction between features for a more comprehensive understanding of the model. Domain experts use this information to complement their domain knowledge. Data scientists and model validators use these insights to assess feature dependence in order to better interpret other results.

- **Bias insights**. Insights whether model predictions are biased are paramount to comply with AI regulations such as standards 19 and 20 of the Ethical Framework. The development of fair models is the responsibility of everyone involved in the process. However, the demand for these insights comes primarily from compliance officers, since they have responsibility for identifying and mitigating legal and ethical risks. Given that quality managers and compliance officers are non-technical stakeholders, the fairness insights should be concise.

Table 4.2: The importance of additional information needs for each stakeholder to perform their tasks, expressed with tick marks (0 = no importance, 1 = low importance, 2 = medium importance, 3 = high importance).

Information needs	Data Scientist	Data Consultant	Business Manager	Domain Expert	Quality Manager	Model Validator	Compliance Officer
Contextual info	✓	✓	✓✓	✓	✓✓	✓✓	✓✓✓
Data info	✓✓✓	✓✓		✓		✓✓✓	
Background info	✓✓✓	✓✓				✓✓	
In-depth insights	✓✓✓	✓✓	✓	✓✓		✓✓✓	✓

Additional information needs To ensure that the explanations are properly conveyed to stakeholders - the 'Serve multi-stakeholder information' property - four additional information needs are identified (Table 4.2):

- **Contextual information**. A requirement for understanding a model properly is to understand the model's context. This contextual information includes the goal of the model, the data source used, the employed algorithm and prediction task, and performance metrics. Additionally, the names of features and labels should be clearly described, in contradiction to the cryptic names often used in databases. This demand for contextual information increases the less involved the stakeholder is in the development process.

- **Data information**. Technical stakeholders, and to a lesser extent domain experts, demand information about the training data to better interpret the results. This includes information about data distributions and feature correlations.

- **Background information**. The insights are generated with XAI techniques whose operation is unknown for the vast majority of the users. Non-technical stakeholders take this for granted, while technical stakeholders appreciate background information about the establishment of the results. The system should provide this background information on demand for those interested.

- **In-depth insights**. The level of detail desired varies greatly between stakeholders. Business managers, quality managers, and compliance officers prefer an overview of the insights with aggregated results on a global level, to quickly comprehend the entire model. All other stakeholders - in particular data scientists and model validators - prefer to have additional in-depth insights that preserve details, e.g.,

to detect dispersions. The system should use methods and visualization techniques to handle both information densities.

In summary, we identified six general properties for the system to be a turnkey XAI system: 'Generic', 'Open', 'Modular', 'Accessible', 'Supports AI compliance', and 'Serves multi-stakeholder information'. By interviewing the various stakeholders, we formulated additional eight criteria to provides the right explanations to stakeholders to perform their tasks, and meet the other information needs of the stakeholders.

4.2 Methods

Section 4.1.3 lists the four explanation types that are identified based on stakeholder demands. This section addresses the selection of the appropriate method to generate each of these explanation types. Background information on the selected techniques is described in Section 2.2. First, several considerations are described.

General considerations. Three general considerations are taken into account during the selection of the methods. First of all, the selected techniques should be model-agnostic to ensure it can be applied to all algorithms ('Generic' property). In addition, model-agnostic techniques lead to the same explanations for each model type, easing the comparison between models. Second, we try to limit the number of techniques due to the limited time and cognitive capacity users have to understand the explanations. Besides, it is harder to compare different sections of the report if it uses different atomic units. Finally, the implementation of the technique is taken into account. It should support industry standards ('Open' property), be complete so it can directly be implemented ('Modular' property), have enough options for configuration ('Accessible' property), and be available as open-source library ('Open' and 'Modular' properties) for Python, the intended programming language due to the many data science and XAI implementations.

Feature importance. SHAP (SHapley Additive exPlanations) [9] is selected as feature relevance technique, based on four strong advantages of this technique. The primary reason for choosing SHAP is that the explanations are intuitive for a wide range of stakeholders. The computed SHAP value of each feature (the average marginal contribution of a feature value for all possible feature combinations) can easily be explained as the informativeness of the feature value (since it is compared with the 'average' feature value in the dataset), or how the feature value pushes the model outcome up or down from the expected outcome to the actual outcome. This relevance of a feature to the model output is easier to understand for non-technical stakeholders than, for example, the relevance of a feature to the model performance such as computed with other permutation feature relevance methods. Additionally, SHAP values are additive (the sum of all SHAP values is the difference between the expected and actual model prediction), contrastive (the actual prediction is compared to the average prediction), and fairly distributed among the feature values, which enhances the intuitiveness. The second reason for selecting SHAP is due to is strong solid theoretical foundations in game theory. As a result, it is a better technique from a legal perspective than techniques that are based on many model assumptions (e.g., LIME assumes linear behavior on a local level) [23], and is therefore preferred based on the 'AI Compliance' property. Third, SHAP values can be leveraged in many ways. The feature relevance can be

expressed for a single prediction (local level), or on a higher (regional or global) level by aggregating local explanations. At these higher levels, the relevance of each feature can be presented as a single SHAP value, or all individual SHAP values of the underlying explanations can be expressed. Due to this variety in information density, SHAP is suitable for serving the wide range of stakeholders ('Serve multi-stakeholder information' property). Additionally, SHAP can be employed for dependence plots and feature interactions. Since SHAP values are the atomic units of all these explanation types, the different explanations can easily compared by the user (the second general consideration). Lastly, the model-agnostic implementation, `KernelSHAP`, is available in the SHAP Python package [31] and supports any prediction functions. The resulting SHAP values can be used for a variety of visualizations that are shipped with this open-source package.

The disadvantages of `KernelSHAP` are twofold. `KernelSHAP` has a high computational complexity (scales exponential with the number of features) compared to model-specific implementations such as `TreeSHAP` (polynomial). As a result, it can take a significant time to compute the SHAP values for global explanations. Since the system will not be used real-time, we accept this drawback. Second, `KernelSHAP` ignores dependencies between features when it samples random values from the marginal distribution. In case of strongly correlated features, this can lead to unrealistic data points used for computing the SHAP values (see the example of partial dependence plots in Section 2.2). This is a common problem for permutation importance techniques, and techniques that solve this issue such as `TreeSHAP` (conditional sampling) can produce less intuitive results [23]. Despite these disadvantages, we conclude that SHAP is the most suitable feature relevance method for this system.

Feature insights. Dependence plots are selected as the method to generate feature insights because they present the insights desired by the stakeholders: how feature values impact the model output. Partial Dependence Plots and SHAP dependence plots - both model-agnostic techniques - are combined to meet the different information demands of the stakeholders. Partial Dependence Plots [34] are the most simple and intuitive representation of the relation between the values of a single feature and the model, and are therefore selected to serve all stakeholders. The distribution of the data will be added to prevent the user from putting too much weight on low-frequent feature values. The frequently-used machine learning package `scikit-learn` [44] has a solid implementation for generating partial dependence plots for both classification and regression models. Technical stakeholders demand more in-depth insights to identify dispersions - a shortcoming of partial dependence plots as heterogeneous groups can cancel each others effects. This is typically achieved by plotting Individual Conditional Expectation (ICE) curves [35], the instance-level version of partial dependence plots. However, we do not select this method since it can generate unrealistic data points for correlated features (see Section 2.2 for an example). Instead, we use SHAP dependence plots that uses real instances from the train or test data. For each instance in the the data set, the feature value (for the given feature) is plotted on the x-axis and the SHAP value of the feature of that instance is plotted on the y-axis (Figure 4.3). Since the real data points are used, it reveals possible dispersions for technical stakeholders and presents realistic relationships in the case of correlated features. In addition, this method uses the atomic SHAP values, allowing comparison with the feature relevance technique - an advantage that other feature-dependent techniques such as ALE do not possess.

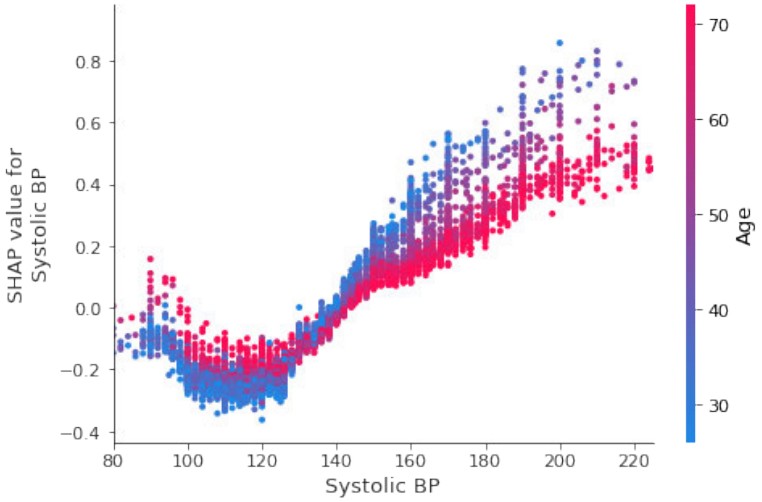


Figure 4.3: SHAP dependence plot between the values of systolic blood pressure (x-axis) and the corresponding SHAP values (y-axis) regarding the mortality rate. The vertical dispersion on the y-axis is mainly caused by the age (feature 'Age' has the strongest interaction effect with feature 'Systolic BP'). The feature value of the interacting feature ('Age') is encoded with the colors of the dots. Based on this plot, we can conclude that high systolic blood pressure on a low age leads to a higher mortality rate than on a high age. [31]

Feature interactions. In complex, non-linear models, features typically have interaction effects, meaning that the combination of features leads to different model outcomes than (just the sum of) the main effects of features. This is the dispersion on y-axis on SHAP dependence plots: when the value of one feature constant (x-axis), the interaction with other features leads to different model outcomes (y-axis). The SHAP package is able to compute SHAP interaction values, a generalization of SHAP values to higher order interactions. Due to the link with SHAP values it is desirable to use this method to visualize interactions. The interaction between any pair of two features can be plotted with SHAP interaction plot. However, this would lead to a high amount of additional plots, while the importance of feature interaction explanations is low (Table 4.1). Instead, we use SHAP dependence plots in combination with interaction values, as displayed in Figure 4.3, allowing the interaction effect to directly be employed to explain dispersions in these dependence plots.

Bias insights. Biased decisions made by predictive models or by interacting with predictive models may have originated in different parts of the business process. First, imbalanced sampling from the population or improper labeling of instances during the data gathering and selection phase can lead to biases in the data. Second, biases may occur in the learned patterns of a model. This algorithmic bias can arise from either biased or unbiased training data, although the algorithm often amplifies any existing bias in the data. In addition, design choices regarding algorithm selection, data encoding, and parameter settings affect the learned patterns and thus potential biases. Finally, bias can arise based on how model decisions are handled in the business processes, such as the selection of decisions to follow up on. This intervention biases can even occur in case of a perfectly unbiased model. Since the system focuses on interpreting predictive models, we focus on identifying potential algorithmic biases in these models.

We define algorithmic bias as the presence of larger systematic errors in model predictions for certain groups compared to other groups, leading to less desirable outcomes for these groups. The presence of algorithmic bias in a model does not automatically imply model unfairness - the essence is to ensure there are no excessive differences between groups. We therefore state that a model is fair if potential algorithmic biases fall within accepted, predefined, thresholds. Although model unfairness is not desirable, it is not prohibited by definition. A model is discriminatory only if algorithmic bias applies to groups that are based on attributes protected by law, such as gender, race, and sexual orientation (see Section 2.1). The system exclusively assesses algorithmic bias and fairness, given that compliance officers are in the best position to assess whether protected attributes are used.

Similar to feature importance, model fairness can be considered on multiple levels: for a single prediction or for a group of predictions. The latter is selected as it best fits the global scope in which the system will provide insights. Fairness on a group level is typically assessed by computing statistical metrics for groups. Numerous metrics have been proposed, which can be divided in three categories [45–47]. Distributional metrics concern the distribution of groups among the different outcomes. A frequently-used metric in this category is the demographic parity [48] (often refered to as statistical parity), which states that each group should have the same probability of being assigned to the positive outcome. These metrics do not take the actual outcome into account, in contrary to error-based metrics. Error-based metrics concern the errors of groups by comparing the predicted outcomes with the true outcomes (label), and state that error rates should be similar for each group. These metrics typically rely on a confusion matrix, which shows the number of true positive, true negative, false positive, and false negative predictions for a binary classification. The third category covers probability-based metrics that compare the probability of outcomes with the true outcomes. There is no agreement on what the best metric is, as this varies with regards to the use case and the stakeholders involved. We therefore formulate three criteria to select the appropriate metrics to be used by the system:

- The metrics should take the actual data labels into account to prevent that the differences between groups in the data are incorrectly marked as algorithmic bias.
- The metrics should be easy to understand to make it more useful to the stakeholders.
- The metrics should be generic to serve the wide-variety of use cases to which the system will be applied.

Distribution-based metrics are excluded based on the first criterion. Based on the second criterion, we prefer error-based metrics over probability-based metrics given their derivation of the easy-to-understand confusion matrices. The following three error-based measures are selected:

- **Predictive Parity** [49] is achieved if all groups have equal positive predicted values, which is the fraction of true positive predictions out of all positive predictions. It ensures that each group has the same probability of a correct positive prediction. The positive predicted value is also known as the precision, a commonly-used performance metric by data scientists, making this measure easier to explain and a good fit for precision-driven use cases.

- **Equal Opportunity** [50] is achieved if all groups have equal true positive rates, which is the fraction of true positives out of all positive instances. It ensures that for all groups the positive instances have an equal probability of being correctly classified as positive. The true positive rate is also known as the commonly-used performance metric 'recall', making this measure easier to explain and a good fit for recall-driven use cases.

- **Equalized Odds** [50] is achieved if all groups have equal true positive rates (the fraction of true positives out of all positive instances) and false positive rates (the fraction of false positives out of all positive instances). It ensures that for all groups the positive instances have an equal probability of being classified as positive, and for all negative instances to be incorrectly classified as positive. Equalized Odds has a stricter notion of fairness than Equal Opportunity as it takes both the true and false positive rates into account. We therefore select this measure to serve models that require a high certainty of fairness.

These measures satisfy our definition of algorithmic bias as they compare the error rates between groups. To make the link to acceptable bias and fairness, we follow by implementation by open-source libraries `Aequitas` [47] and `fairmodels` [51] and model the measures as parity between two groups. In this approach, the performance metrics (positive predicted values, true positive rates, and false positive rates) are computed for each group. By dividing the performance metric of the protected (minority) groups with the metric of the reference group, we get the (dis)parity between these groups. The majority group is selected as the reference group, as it typically has a privileged position, and of substantial size to reliably serve as reference group. The resulting parity of each group - 1.0 if it is perfectly equal to the reference group - is then compared with the fairness threshold τ. A model passes all criteria if all parities fall within the fairness thresholds:

$$\tau < Disparity < \frac{1}{\tau}$$

Following the 'Modular' criterion, we select the open-source Python library `Aequitas` [47] to easily implement the fairness measures. This library fits our approach as it supports the selected measures, computation of disparities using a custom reference group, and comparison with custom-defined fairness measures. Additionally, `Aequitas` enables the detection of indirect biases as it supports the formation of groups based on features that are present in the dataset but omitted in the model. This approach of leaving out protected attributes originates from legislation - you may only use protected attributes for predictive analysis if there is a legal ground - and is known as *fairness through unawareness*. However, it has been criticized [52, 53] since it does not guarantee fairness as protected attributes can be strongly correlated with other features. The indirect bias functionality allows us to detect biases in models that are trained though the *fairness through unawareness* approach, and thereby mitigate the risk of this commonly-used methodology. As a result, our system complies with the guidelines to only use protected attributes to detect bias in models, instead of using it in the model for scoring.

4.3 Architecture

This section summarizes the system design by presenting a system architecture, as shown in Figure 4.4. The architecture is modeled on an enterprise-level with Archimate to clearly illustrate the interaction with business processes, business actors, and the underlying technology. To reduce complexity, detailed system internals and business processes are omitted. The presented architecture is generic and does therefore not contain any organization-specific elements.

The Archimate model consists of three layers: a business layer, an application layer, and a technology layer. The business layer in yellow consists of business actors that are assigned to business processes. The application layer in blue includes application components and functions, realizing services that are used by business processes. The technology layer in green presents the computational resources, software, and services that are used to serve applications and business processes. More information regarding Archimate elements can be found in the Appendix.

In the following, Figure 4.4 will be elaborated on in more detail. Starting with the seven business actors that were identified in Section 4.1.2, the data scientist has a pivotal role in both developing and interpreting predictive models. The data scientist works on the development process from a technical perspective - consulting and other activities are omitted for simplicity purposes. This development process starts with inspecting and preprocessing business data, followed by training a predictive model on this data, ending with model evaluation. Model evaluation is typically done by observing model performance on a test set, which may retrigger the development process for further refinement. This development process can be executed on any development environment that is hosted on an enterprise server. In the new situation with the XAI system, the model is not only evaluated based on performance, but also interpreted to determine whether the results are constructed based on the correct grounds. To this end, the data scientist executes the process to generate an explainability report with insights regarding the inner workings of the model. This process can also be triggered by a compliance process, e.g., to periodically inspect production models. The first step is to load the data and model into the toolkit using the `Explainer API`, which automatically triggers functions to prepare the data and model for internal use. Using this API, other services can be triggered, such as the generation of a configuration file. The resulting configuration file can then be filled to define feature names, set parameters, and define the protected features, among other things. This configuration file is an Excel file, removing technical hurdles for other (non-technical) stakeholders than the data scientist to fulfill this step. With the populated configuration file, the API can be invoked - which automatically reads the configuration file and sets the parameters - to generate a report. This service calls an internal function that orchestrates the execution of other functions based on the set parameters. The majority of these functions generate a particular model insight, utilizing external libraries. These techniques access the data and model to create preliminary results such as SHAP values, which can then be employed in multiple ways. Generating XAI insights requires a fair amount of computing power that is delivered by the enterprise server or cluster, as well as the Python environment for the application. By separating the techniques, parallelization is possible to speed up the computation time. The preliminary results are then led to visualization techniques (grouped here for simplicity), which all use the same underlying techniques from an external graphing module to result in a consistent

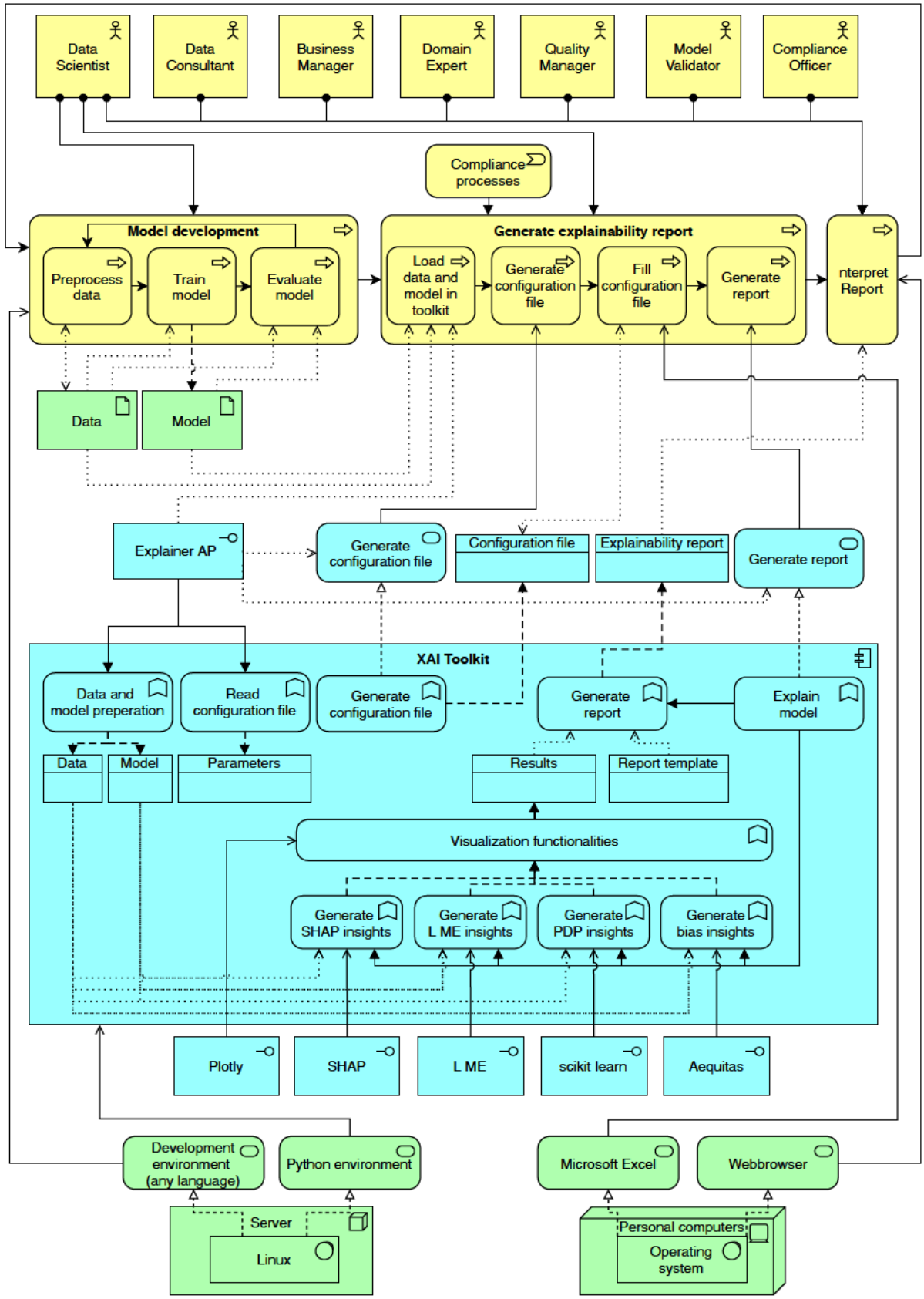


Figure 4.4: Overview of the XAI system.

user interface. This leads to one or more plots per insight that will then be merged into an existing report template that contains support functionalities such as navigation and background information. The system thereby delivers an interactive explainability report that can be viewed with any web browser on personal computers. This format enables easy sharing, usage, and storage of the results. Stakeholders involved can then interpret the results individually or in collaboration. The interpretation can retrigger the development process for model refinement, or - omitted here for simplicity - trigger other compliance processes or model deployment. In this architecture, we present the `XAI Toolkit` as a single component for clarity purposes, while in our implementation we split it into three components: `Explainer` with all explanations functions, `Configuration File Generator` with the function to generate configuration files, and `Report Generator` with the function to generate the report.

Chapter 5

System Implementation

This section describes the development and deployment of a system prototype to meet the second objective of this study. Section 5.1 describes how the design criteria are translated into system characteristics. The detailed technical implementation of these characteristics is described in Section 5.2. Section 5.3 elaborates on the organization-specific implementation and how the system is deployed at the host organization.

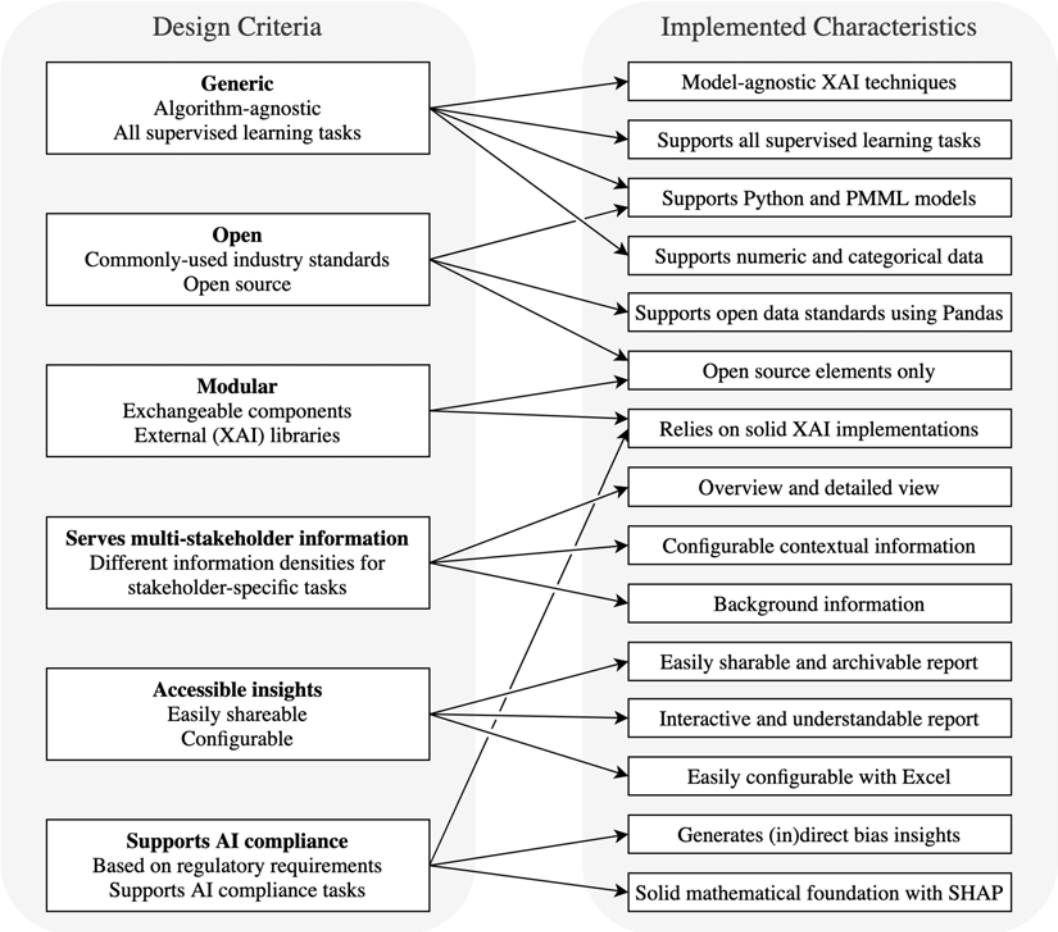


Figure 5.1: Mapping of design criteria to system characteristics.

5.1 Implementing the Design Criteria

This section describes how the design criteria for Section 4.1 are translated into system characteristics, as illustrated by Figure 5.1.

The system has four characteristics to satisfy the 'Generic' property. First of all, the system employs model-agnostic XAI techniques to support any algorithm that works with structured data. Both binary classification, multi-class classification, and regression models are supported. A requirement for classification models is including a `predict_proba()` function, and a `predict()` function for regression models. The system thereby supports the estimators of the frequently-used Python package `scikit-learn` [44], and models in the universal Predictive Model Markup Language (PMML) format, realizing a language-agnostic system. Models that are developed in other languages, or enterprise software such as SAS, can be converted to PMML with language-specific tools, such as `r2pmml` for models developed in R. In Python, the models can be imported with packages such as `sklearn-pmml-model`[1] or `PyPMML`[2]. These functionalities are not implemented in the system to keep it lightweight. Since other languages may support categorical data, the system both supports numeric and categorical data. A requirement for the data is that it is provided as a `pandas DataFrame`, a powerful package for data analysis in Python [54]. Pandas is shipped with functionalities to load data from common formats such as CSV and SQL, allowing to easily use these formats in conjunction with the system without implementation effort. The following snippet shows how the aforementioned libraries and the `Explainer` are leveraged to generate a report in just eight small steps:

```
# 1. Import third-party packages
import pandas as pd
from sklearn_pmml_model.ensemble import PMMLForestClassifier

# 2. Import the Explainer from the system
from xai_toolkit import Explainer

# 3. Load the model from the .pmml file with sklearn_pmml_model
model = PMMLForestClassifier(pmml='./path/to/model.pmml')

# 4. Load the train and test data as pandas DataFrames and store them in a tuple
data_train = pd.read_csv('./path/to/traindata.csv')
data_test = pd.read_csv('./path/to/testdata.csv')
data = (data_train, data_test)

# 5. Specify variables: the target column in the data and the location of the config file
target = 'model_output'
config_path = './path/to/config_file.xlsx'

# 6. Load the data, model, and variables into the Explainer
explainer = Explainer(data, target, model, classification=True, config_path=config_path)

# 7. Generate a configuration file, fill it in Excel, and re-upload it to the server
explainer.generate_config()

# 8. Explain the model
explainer.explain_model()
```

[1] https://pypi.org/project/sklearn-pmml-model/
[2] https://pypi.org/project/pypmml/

From the process perspective, a major criterium of the system is to serve multi-stakeholder information. Interviews identified that stakeholders demand insights in feature relevance, complex feature relations, feature interactions, and model bias - with different priorities and details depending on the specific stakeholders. To satisfy this difference in information density, we implemented an 'Overview' that is relevant for all stakeholders and an 'Detailed View' with more in-depth information for data scientists, data consultants, domain experts, and model validators. The 'Overview' consists of general and contextual information, feature relevance insights on an aggregated level, simplified feature insights with partial dependence plots, and bias and fairness insights. The feature relevance is presented for the complete model and per range of the model: per class for classification models, and for user-defined ranges for regression models. Additionally, features can be grouped to get a quick understanding of the relevance of different feature types. The 'Detailed View' presents the feature relevance and feature insights with more detail. It contains a SHAP summary plot to display the instance-level feature relevance, SHAP dependence plots revealing dispersion and feature interactions, and SHAP force plots that provide insights into local feature relevance. In addition to splitting the insights into two views, the following functionalities provide different information densities. All insights are displayed as interactive figures (mainly powered by Plotly [55]), allowing the user to get more information of a specific data point and to zoom in on regions of interest. Supporting information such as the sample sizes used and a brief summary is presented next to the results. More detailed background information can be opened with pop-up boxes, so it is unobtrusive for other stakeholders. Furthermore, feature and class names can easily be converted to descriptive terms using the configuration file. To enable these functionalities and satisfy the 'Accessible insights' property, the system outputs a HTML report, which can easily be shared between and opened by stakeholders. This interactive report provides the insights in an understandable way.

A key feature to enable understandable insights and wide-applicability is the ability to configure the system for a specific model, once again in such a way that it can be understood by all stakeholders. The configuration file in Excel can easily be shared with and understood by all stakeholders, and enables the configuration of most aspects of the report. This Excel sheet consists of five sheets:

- 'Info' enables the user to provide contextual information about the model, such as the goal of the model.
- 'Features' enables the user to configure understandable feature names and descriptions, which will be used instead of the column names from the dataset. Figure 5.2 displays a screenshot of this sheet. In addition, this sheet can be used to assign feature to custom-defined feature groups.
- 'Classes' (for classification models) or 'Ranges' (for regression models) enables the user to define understandable class descriptions (instead of the values used in the dataset) or define custom ranges of interest for regression models.
- 'Fairness' enables the user to define the protected features for which bias and fairness insights will be generated, and custom-defined groups and fairness thresholds. Figure 5.3 displays this sheet.
- 'Parameters' allows the user to easily define the parameters of the `Explainer`, such as whether feature grouping should be enabled.

This section has been redacted due to company confidentiality reasons.

Using a HTML report and configuration file in Excel have the additional benefit that it can be archived in most systems for compliance purposes. This is one of the examples how the system is tailored for the compliance process, next to how it serves multiple stakeholders. To serve the compliance process, the system also has bias insights that serve a wide variety of use cases. It allows the detection of direct and indirect model biases, in which it complies with future AI regulations. In addition, employing the mathematically-grounded SHAP fortifies the insights from a legal perspective. Overall, the system heavily relies on open source packages, significantly reducing the likelihood of errors - important for both user trust and from a risk perspective.

5.2 Technical Implementation

This section describes the technical implementation of the system in detail. We first address aspects that impact the whole system, and then the implementation of the different explainability insights.

General The system extensively employs `KernelSHAP` and the resulting SHAP values. To recap, `KernelSHAP` determines the relevance of features by setting features to 'missing' (excluding them from a coalition) and measuring the change in model output. Since most models do not properly function with omitted values, `KernelSHAP` requests a background dataset at initialization to impute 'missing' features with a random value from the background set. When computing the SHAP values for a singe instance, the entire background set - which is typically the training set - is used, which can lead to computational problems for large training sets. To tackle this potential issue, we summarize the background dataset with a weighted k-means sample, where each instance is weighted with the number of instances it represents. This significantly decreases the computation time required, while preserving the representativeness of the results. In the case of non-encoded categorical features, to which the k-means approach cannot be applied, we draw a random sample from the complete dataset. For both approaches, we set the default sample size to 25 to balance compute time and representativeness, which can be user-defined for specific needs. The user is encouraged to pass the train and test datasets of the model as two separate `DataFrames`, so that the train dataset is used as background and the test dataset is used for the explanations. If the user passes a single dataset, the system automatically performs a stratified split where 80% of the data as test set and 20% as train set. Besides this approach for the background, we use the default parameters of `KernelSHAP`.

This approach enables decent computation times for Python estimators, allowing report generation within 30 minutes for the Python models we tested. This lead time is acceptable as it is in proportion with the other steps in the development process, and it does not affect most stakeholders. PMML models can be converted to Python-based `scikit-learn` estimators with `sklearn-pmml-model`, which results in ten to hundredfold faster performing models than those converted with `PyPMML`. In the event that PMML models cannot be converted with `sklearn-pmml-model`, report generation can take hours to days, which is not acceptable. Therefore, we implemented a 'fast mode' that employs LIME to generate feature relevance insights, as LIME is less computationally-intensive then SHAP [56]. We selected LIME for this mode as it is strongly strongly connected with `KernelSHAP`. The downside of the fast mode is that is will only generate an 'Overview' of the report, lacks the solid foundation of SHAP, and does not support categorical features in our implementation

due to possible mapping issues with PMML. We disable the discretization of continuous features so that LIME displays the contribution for each feature, similar to SHAP. Other parameters remain at their default values.

The toolkit automatically handles numeric data, but requires user input to define non-encoded categorical columns. Those columns will then automatically be coded as numeric data for the functionalities that demand this format, such as the partial dependence plot and correlation matrix. Functions that handle categorical data internally, such as SHAP, get the original data to minimize the chance of mapping errors. For user convenience, the report always displays the data in the original format.

In general, for reproducability purposes, we set a random seed for an `Explainer` object and pass it to all dependencies.

Feature Relevance The 'Overview' section of the report presents the global feature relevance with a list of bars, where the size of each bar indicates the relevance. By default, these insights are established by generating and aggregating SHAP values for a sample of the test set. In line with the SHAP implementation, the global relevance of a feature is established by summing the absolute SHAP values of all instances for that feature, or in case of multiclass classification, where a SHAP value is computed for each class, by computing the mean of the absolute SHAP value of all classes for each instance and then summing these values for the entire sample. In fast mode, a high number of local LIME explanations of instances in the test set are generated. The global relevance of a feature is then established by taking the mean of all absolute contributions of all instances for that feature. We present the resulting feature relevances of both approaches in a relative manner where the relevance of each feature is expressed relative to the most influential feature, since the absolute feature relevances are not intuitive for the users of the overview. A summary of the most relevant features is presented in the report summary. As the default setting, the system uses a sample size of 1000 instances (or the maximum test set size) for generating these insights, which can be altered by the user.

We establish the global feature relevance - thus for the complete model - by drawing a random sample from the complete test set. To capture more specific behavior of the model, feature relevance insights are computed for specific data labels: each class for classification models, and certain user-defined ranges for regression models. The feature relevance for a specific output range is established by drawing a random sample from the instances in the test set with the corresponding label, then computing the feature relevance similar to the global insights. The sample size of the global feature relevance insight are distributed equally over the output ranges, for example, 500 instances will be used by default for each class of binary classification model.

Since real-world models can employ tens or even hundreds of features, the report presents three views of the global feature relevance: the relevance of all features, the most relevant features (20 by default), and the relevance of groups of features. The latter is achieved by allowing the user to assign features to custom-defined groups in the configuration file. The relevance of each group is then computed by taking the mean contribution over all features in the group. The report contains a view to compare different feature groups, and a view for each feature group to compare the features within that group with the group average.

The aforementioned feature relevance plots display contextual information on hover, such as user-defined feature names and description. In the case of strong feature correlations - feature correlations that exceed a configurable threshold that is set to 0.2 by default - the user is informed on which features are strongly correlated. All correlations between the 20 most influential features are displayed at the top of the report. The goal of these functionalities is to inform the user which features change jointly, in order to correct the feature independence assumption of `KernelSHAP`.

In the detailed view, the relevance of each feature is presented as a beeswarm plot - a one-dimensional scatter plot that displays for each instance what the feature relevance was for that feature. In SHAP's 'Summary plot', the beeswarm plots are stacked to present the feature relevance of the 20 most relevant features. We use the figure that is produced by SHAP and convert it to an interactive figure with Plotly. One plot is presented for regression models, and one per class for classification models as the SHAP values are computed for each model output.

Feature Insights The system utilizes `scikit-learn` to compute the partial dependence for the feature insight plots that are presented in the overview section, allowing the generation of this plot even in fast mode. This functionality can directly be applied to scikit-learn-based estimators - other models are first prepared by the system. The default yet customizable setting generates a plot for each of the 10 most relevant features, as computed with the feature relevance techniques. The plot, with the feature value on the x-axis and model output on the y-axis, consists of a partial dependence line for each model output, and a distribution of the data to inform the user of the likeliness of the feature values. All elements use the same number of data points: one for each unique value of categorical features, or one for each of the 40 intervals of continuous features. The sample size used to generate these insights can be set by the user and is set to 20,000 instances by default.

The 'Detailed View' presents SHAP dependence plots for the same number of most influential features. These scatter plots contain the feature value on the x-axis, the SHAP value on the y-axis, and a color bar on the off-diagional to color the feature value of the strongest interaction features, which is automatically determined by SHAP. To save computation time, we plot the instances and their SHAP values from the global feature relevance sample. A separate distribution is not added since the distribution becomes visible from the data dispersion. Similar to the SHAP summary plots, a static SHAP dependence plot is generated for each model output by the SHAP package, which is then converted by the system to an interactive format.

The feature insight plots in both the 'Overview' and 'Detailed View' are accompanied with a note of strong correlations, if any, to inform the user which feature values jointly change.

Bias and Fairness Open-source Python package `Aequitas` is employed for the computation of the bias measures. It requires a dataset with protected attributes of interest, and the actual and predicted labels in a binary classification format (0/1). Using the bias and fairness functionality therefore requires the user to explicitly specify that a test set is passed that contains the ground truth, since computing bias measure on the training set, or on a dataset without ground truth, will lead to unrepresentative results. Additionally, the system maps

the data as a binary classification problem. For multi-class classification tasks, the user should specify which classes should belong to the positive class and which to the negative class. The user can specify a threshold - or use the mean of the target - for regression tasks, where after all values below this threshold will be presented as the negative class and all above this threshold as the positive class. After this transformation, only the protected features that are specified by the user in the configuration file are kept in the dataset, for which the groups will be defined. For categorical features, each category is represented with a group. Continuous features are discretized into quartiles, or in user-defined bins. The system characteristic that bias measures can be generated for features that are present in the dataset but not used for scoring, enables the user to generate more fine-grained groups during preprocessing and subsequently passing the data to the system. After the groups are defined, the majority group is automatically selected as the reference group, as majority groups typically have a (unjustified) privileged position. In addition, the size of the majority group ensures that it is of sufficient size to be compared with.

After the preprocessing steps, `Aequitas` calculates the disparities for all groups and compares them to a user-specified threshold for accepted bias. To determine the default setting for this threshold, we examined available legislation on this topic. To the best of our knowledge, the only legislation that quantifies such selection rates is the 'four-fifths' rule from the Uniform Guidelines on Employee Selection Procedures (1978) [57]. It states: "A selection rate for any race, gender, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact." Although there are certain limitations to this regulation, including that it relates to demographic (statistical) parity and a different reference group, it provides a solid starting point. We therefore set the default fairness threshold to 0.8, similar to the default values of libraries `Aequitas` [47] and `fairmodels` [51]. Nevertheless, the optimal threshold differs per use case and therefore requires careful consideration. Besides, in our opinion, expressing this complex problem with a single conclusion is helpful for the stakeholder, but lacks nuance. The explainability report therefore states whether a model passed or failed the fairness criteria - on not whether it is fair or unfair - and leaves the more nuanced and case-specific interpretation of the fairness conclusion to the involved stakeholders. The output of `Aequitas` is used to create interactive plots that shows, for each of the three measures, the disparity, the margins of accepted bias, whether the model passed or failed the fairness criteria, and detailed information such as group sizes on hover. Such a plot is created for each protected feature and presented along with a textual summary.

This technical implementation is tested in two ways. First, some unit tests are written for core functionalities such as data validation. Second, the system is tested in a holistic manner by automatically explaining several dummy models with a variety of characteristics, such as used algorithm, programming language (R and Python), data type (numeric and categorical), and prediction task (binary classification, multiclass classification, and regression). This enhances the robustness and suitability of the system.

5.3 Organization-specific Implementation

This section describes the technical and functional implementation specific to the host organization. Implementation at other organizations would probably require similar steps, which will be further outlined in Chapter 8.

As described in Section 3, Achmea is already making efforts in the field of model management, including logging the inputs and outputs of production models for reproducibility purposes. For ease of use and user perception, we integrated the system with this framework by implementing a database connection. A separate component of the system enables connection to this SQL database and the retrieval of model runs of a specified model within a given time range. The returned `pandas DataFrame` can then be subjected to model-specific preprocessing by the data scientist and passed to the system. This provides insights into the behavior of the model on production data, and the underlying data distributions.

Achmea hosts multiple UNIX-based development environments that run on both on-premise servers and virtual machines. Each environment provides a sufficient amount of computational power and memory for the system, as some ensemble models in PMML require tens of gigabytes of memory - even though the environment is simultaneously used by multiple development teams. We deployed the system as a Python package on the development environments, enabling technical stakeholders to access the system with a single line (`from xai_toolkit import Explainer`) of code. The system and the dependencies can also easily be installed on a custom virtual environment.

To ensure that the stakeholders understand the system, two types of documentation have been drawn up that are tailored for groups of stakeholders. A work instruction is available in the enterprise work instruction repository that covers the motivation, general idea, and use cases of the system, along with a step-by-step instruction, and a reference to a short video that demonstrates how the insights should be interpreted. More detailed technical information is available for the technical stakeholders in the git repository. Next to a step-by-step instruction, it elaborates on how to configure the system's parameters, what considerations should be taken into account when setting the parameters, and how models that are developed in other languages than Python can be converted to the PMML format. Additionally, dummy data and models, scripts to prepare and load this data and models into the system, and configuration files are available for commonly-used model types in this git repository.

The integration with existing frameworks, the deployment to enterprise architecture, and various forms of documentation, removes the hurdles for widespread application within the organization to leverage it use cases, which will be further addressed in the next chapter.

Chapter 6

System Application

This chapter describes the application of the system at the host organization with the goal to further define and elaborate on the use cases. During the design (Chapter 4), we identified two primary use cases of the system: (i) AI compliance through model understanding, and (ii) model refinement through model understanding. Section 6.1 and 6.2 elaborate on those two use cases, respectively. Then, in Section 6.3, we describe how the system added business value when it was applied to four models of the host organization. Several screenshots of an explainability report - this dummy model predicts a claim value based on characteristics of a personal injury claim - are presented throughout this chapter to illustrate the use cases.

This section has been redacted due to company confidentiality reasons.

6.1 AI Compliance

This section explains the intended use case of the system, namely, how the use of the system leads to better compliance with AI regulations. This stems from two points. First, the use of the system leads to better model understanding, allowing a better assessment of whether the model meets all standards and guidelines. Second, having the system and corresponding procedures in place makes compliance with the regulations more tangible for third party auditors and regulators. From here on, we mainly focus how the system enables better assessment of the model and improves the communication of this assessment, mainly relevant for stakeholders in the 1st (Quality Managers) and 2nd line (Model Validators and Compliance Officers).

The explainability report enables:

- **Proxy detection.** The report eases the identification of proxies, i.e., inferred data points. Predictive models attempt to capture and predict real-world problems through features, which is a simplified representation of reality, based on which one can argue that models fully rely on proxies. In our opinion - and from a legal and ethical standpoint - this becomes problematic when models heavily rely on incorrect proxies: features that do have a correlation with the target variable, but do not have a direct causal relationship in the world. There typically is a third variable (a confounder, mediator, or moderator) involved that is not present in the dataset, which causes the algorithm to use the independent variable as a predictor of the dependent variable. This reduces the transparency of the model and might hide biases. A classic example is the postal code, which typically captures more information than just the geographic region, such as social-economic status or even descent. The toolkit aids the identification of influential proxies in two ways. First, the combination of feature relevance and a descriptive feature name challenges the stakeholders to assess whether a feature is a proxy. A higher feature relevance indicates a correlation between the feature and target variable, whereby the stakeholders should reason whether there is a causal relationship. As the plot indicates the feature relevance, it is a triage for this assessment. Second, the partial dependence plot reveals more regarding the relationship between a feature and the target. This allows the stakeholders to inspect the difference in model outcomes between feature values and reason whether this is a real-world relationship. Figure 6.1 displays the feature relevance of a model with multiple proxies, including 'Gender' and 'Employment' as a proxy for income. The difference between different types of employment is illustrated with the partial dependence plot presented in Figure 6.4. Stakeholders can inspect these relationships in greater detail in the 'Detailed View'. Identifying and addressing influential proxies increases the transparency of the model, potentially removes biases, and ensures more ethical establishment of predictions.

- **Bias detection.** The much-discussed model biases can be detected with two approaches. Most prevalent is using the bias and fairness section as displayed in Figure 6.2 to assess the disparity between groups of protected attributes. Figure 6.2 shows that the model performs less (in terms of precision, recall, and false positive rate) for claims of customers with unknown gender, compared to the reference group. For all measures, the disparity falls below the accepted fairness threshold of the four-fifths rule ($\tau = 0.8$). The model is also less-performing for woman-related claims. Using the Predictive Parity and Equal

Opportunity measures, the model biases fall within the acceptable range, which is not the case when relying on the more stringent Equalized Odds measure. The model should be retrained with this bias in mind, for example, by oversampling the minority groups in the trainset. Even better would be to ensure that gender no longer is a proxy for income by adding income as a feature. The other approach is to observe the differences between feature values for protected features using the partial dependence plot, which also reveals the difference between genders (Figures 6.3 and 6.4). Addressing algorithmic biases leads to better compliance with standards 19 and 20 of the Ethical Framework, and virtually every other regulation.

- **Data minimization.** Standard 12 of the Ethical Framework prescribes that the amount of data used should be limited to enhance user privacy. The report can aid the feature selection process and thereby reduce the number of features. Next to enhancing user privacy, lowering the number of features increases the maintainability of the development pipeline, reduces the computational complexity of the model, and improves model interpretability. The feature relevance plot in Figure 6.1 illustrates that for this model a minority of the features determine the majority of the predictions. The seventeen least relevant features (out of a total of twenty-four) can be removed without leading to a decrease in model performance. This cut-off is not quantified, as the feature selection requires a more nuanced interpretation by the stakeholder. For example, a feature might have limited relevance on a global scale, but be very informative for certain ranges of the output - which can be inspected using the feature relevance plots of different ranges of the outcome. Just the 'Overview' of the report provides a good starting point for the feature selection process to realize data minimization.

- **Internal communication.** The report enhances the communication between internal stakeholders by providing a tangible starting point of, and common ground during, the model interpretation process. This bridges the gap between different disciplines, thereby easing an interdisciplinary discussion about the model, especially desirable for complex discussion topics such as the acceptable model bias.

- **External communication.** The report is a tangible artifact for communication with external stakeholders, such as auditors and regulators. It can be archived along with the internal assessment of the previous point, which increases the auditability of the process.

To leverage the aforementioned use cases, the system is embedded in two of Achmea's compliance processes. The first is the Privacy Impact Assessment, which is mandatory for all applications that use personal data over a long period of time, such as many machine learning models in production. High-risk AI systems are, among other things, subject to an Ethical Impact Assessment. Both processes, that are executed by compliance officers, request information regarding the model or application from the data scientist and business, such as a justification of the data used. Although using the system is not mandatory, generating and interpreting an explainability report of the model is the most convenient way and thus the preferable way to deliver this information. Through incorporating the system, Achmea improves the compliance with standards 16, 19, 20, and 23 (see Section 3.3) by providing and internal control mechanism for AI-driven models.

This section has been redacted due to company confidentiality reasons.

6.2 Model Refinement

An additional use case of the system is that of model refinement, considering that understanding the model leads to insights into how the quality can be improved. This use case indirectly supports AI compliance, as better-performing and more robust models are desirable from a regulatory standpoint. Nevertheless, mainly the stakeholders within data analytics and business are interested in this use case. There are two primary triggers for this use case.

First, the system can be utilized for model validation. The goal is to understand the inner workings of the model to assess whether it works as intended. It provides more information than just the use of performance metrics, as it relieves on what grounds the predictions are established. This is relevant for models from any state of the model lifecycle, but especially for first prototypes that have not yet been thoroughly inspected. Model validation involves proxy and bias detection. The SHAP summary plot from the 'Detailed View', as shown in Figure 6.5, is very informative for model validation. It shows the homogeneity of the data and thereby reveals whether the feature relevance is based on outliers, strongly separated clusters, or gradual shifts. Any error that is identified during the model validation process can be resolved in the next iteration of the model. Second, a better model understanding can lead to new insights regarding model improvement. For example, inspecting the relevance of different feature types can form an inspiration for the feature engineering process, especially when looking at exceptions at a local level. In addition, inspecting data distributions, feature correlations (see Figure 6.6, and feature interactions can enhance domain knowledge, which, in turn, provides insights into how to capture it with a model.

To ensure that data scientists do not overlook this step, these use cases are embedded in the development process. The default development process, *StageGateManagement*, divides the development of models into multiple stages, with 'gates' in between. These 'gates' are checklists that should be checked before proceeding

This section has been redacted due to company confidentiality reasons.

to the next stage of development, with the aim to stop the project in time if it is not successful. The checklist of the gate between the 'creation phase', where a proof of concept is realized, to the 'growth phase', where a minimum viable product (MVP) is developed, lists model understanding as a requirement to proceed. The use of the system is not mandatory for this step, but recommended, as it is the most convenient way for data scientists to obtain the prescribed insights. In addition, it prepares the development team for future compliance assessments.

6.3 Cases

The system has been applied to four ML-based models during the internship at the host organization. These models have been selected to represent a variety of characteristics and use cases. In addition, the teams concerned were happy to cooperate and provide feedback. In this section, we describe the process of applying the system to these models, and the main added value of the enhanced model understanding.

This section has been redacted due to company confidentiality reasons.

This section has been redacted due to company confidentiality reasons.

This section has been redacted due to company confidentiality reasons.

This section has been redacted due to company confidentiality reasons.

Chapter 7

System Evaluation

By applying the system to four real-life models, we demonstrated the effectiveness of the system in a corporate setting. In this chapter, we describe the approach we employed to evaluate the effectiveness of the system in a quantitative manner. The goal is to define the main factors that determine the functional suitability - a software quality characteristic of ISO 25010 - of the system, and thereby answer the fourth subquestion of this study.

To satisfy this goal, we aim to answer the following questions:

1. *How effective is the system as a tool for executing the use cases of AI compliance?*
2. *How is the usage of the system perceived?*
3. *How are the different XAI techniques perceived?*

In this, the first two questions are particularly established to answer the third question, which is strongly connected to the fourth sub question of this study. In addition, the findings of this third question could make a valuable contribution to the field of XAI.

In Section 7.1 we discuss the setup of the experiment we conduct to answer these questions. Section 7.2 presents the results, followed by the discussion of these results in Section 7.3.

7.1 Experiment Setup

This section elaborates on the setup of the experiment. Since there is no consensus on the best evaluation method [43], we first examine the literature on evaluation methods to determine the rough outline of the experiment. We then formulate the different tasks of the experiment in detail, and finally elaborate on how we realized this setup.

Evaluation methods We select the most appropriate evaluation methods for our experiment from Figure 7.1, which organizes XAI evaluation methods along three dimensions [58].

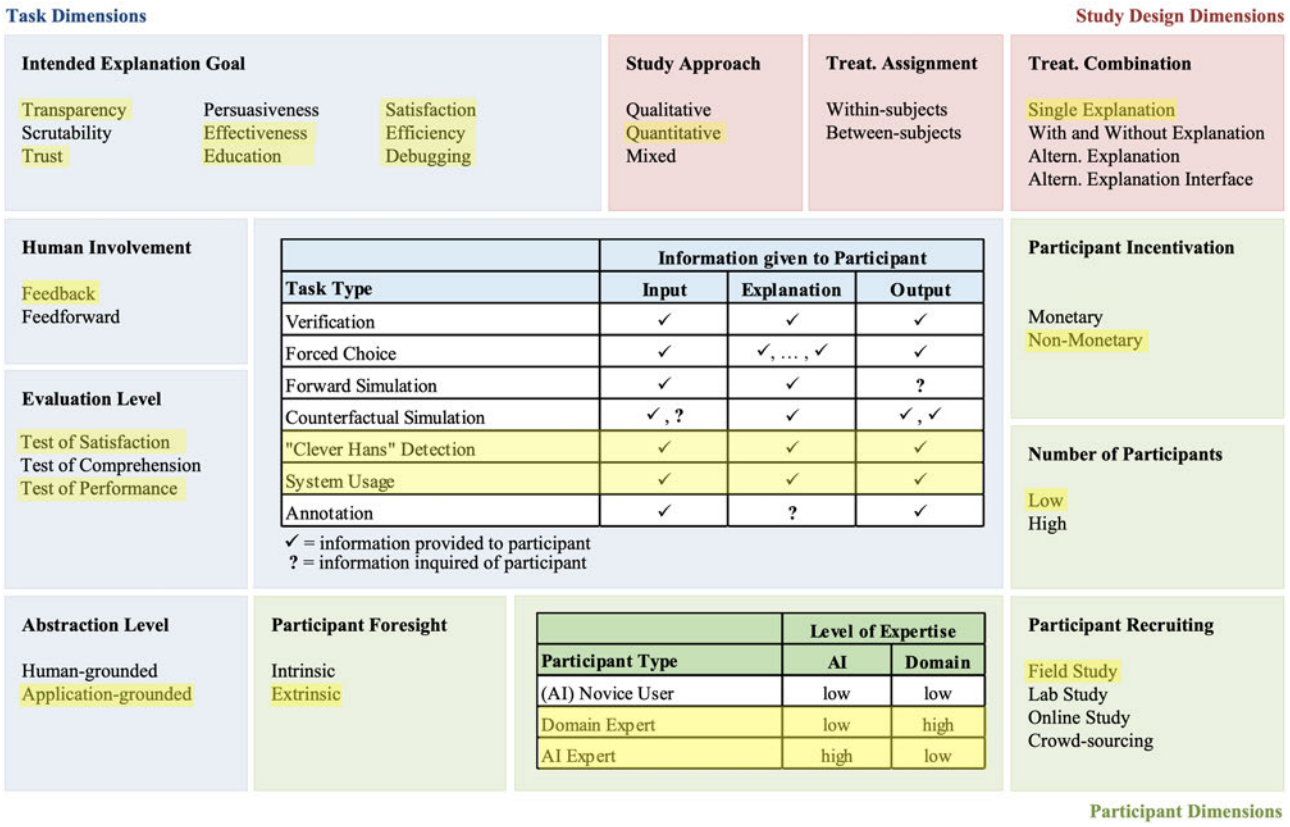


Figure 7.1: The dimensions of XAI evaluation methods from [58]. The highlighted concepts will be used in the experiment.

Starting with the task dimension, the experiment should reflect the intended use of the tool - supporting the AI compliance process - as faithfully as possible. We therefore choose an application-grounded evaluation method [43], whereby experts are asked to execute that actual task of the system. This is a more appropriate method for evaluating real-world applications than an human-grounded evaluation method, whereby inexperienced users are asked to execute a proxy of the real task [59]. In our case, executing the actual task implies that the participant uses the complete report to execute the AI compliance process, including proxy detection, bias detection, data minimization, and internal discussions regarding these topics. Due to the variety of these tasks, we use two kinds of evaluation levels and two different task types. The participants will perform a detection task where they try to detect proxies, biases, and irrelevant features, after which we assess the performance of each participant (a 'Test of Performance'). For hard-to-quantify aspects, such as enhanced trust and enabling communication, we ask the user for their satisfaction regarding these aspects (a 'Test of Satisfaction') after using the tool for executing the aforementioned tasks ('System Usage').

Regarding the participants, we focus on the intended stakeholders of the system, which are AI experts (data analytics) and domain experts (business and compliance lines). Because the stakeholders have different kinds and levels of expertise (extrinsic foresights), we provide an introduction video to ensure that all stakeholders have a certain understanding of the system. Hereby, we ensure that we do measure the first impression of the participants. Since we are tied to the host organization for this field study, we expect a low number of participants, and we will address them based on their intrinsic motivation.

Regarding the study design dimension, we focus on collecting quantitative results in order to quantify the effectiveness of the system. This quantitative focus allows us to use a survey, which eases the process of recruiting as many participants as possible. Participants will have the opportunity to leave remarks at the end of the survey, which we will process as qualitative results. Considering that we expect several dozens of participants, the number of participants will be too low to demonstrate statistical significance between different treatments and we will therefore give all participants the same treatment in form of the same report and questions. Due to the variety of stakeholders, we do ask the participants in which line they work, so we can compare the results between the different types of stakeholders. The provided report is an example of a 'Single combination' treatment, which we chose due to the lack of a representative benchmark - there currently is no explainability system in use within Achmea with this scope. We considered having the participants perform the Privacy Impact Assessment with and without the report, but this was not feasible due to the limited time of the participants, and would not be representative because we provide more information in the scenario with the report. To still be able to make a comparison between the scenario with and without the system, we ask the participant how they typically would address the detection task. If the participant indicates that they previously used a tool, we ask additional questions in order to compare the previously used tool and the system.

In summary, we will give the stakeholders an explanation report to execute AI compliance tasks, after which we ask them to fill in a survey so we can measure the system effectiveness, user satisfaction, and explanation usefulness. Figure 7.2 displays that the survey is divided into three parts, each of which we will now address in further detail.

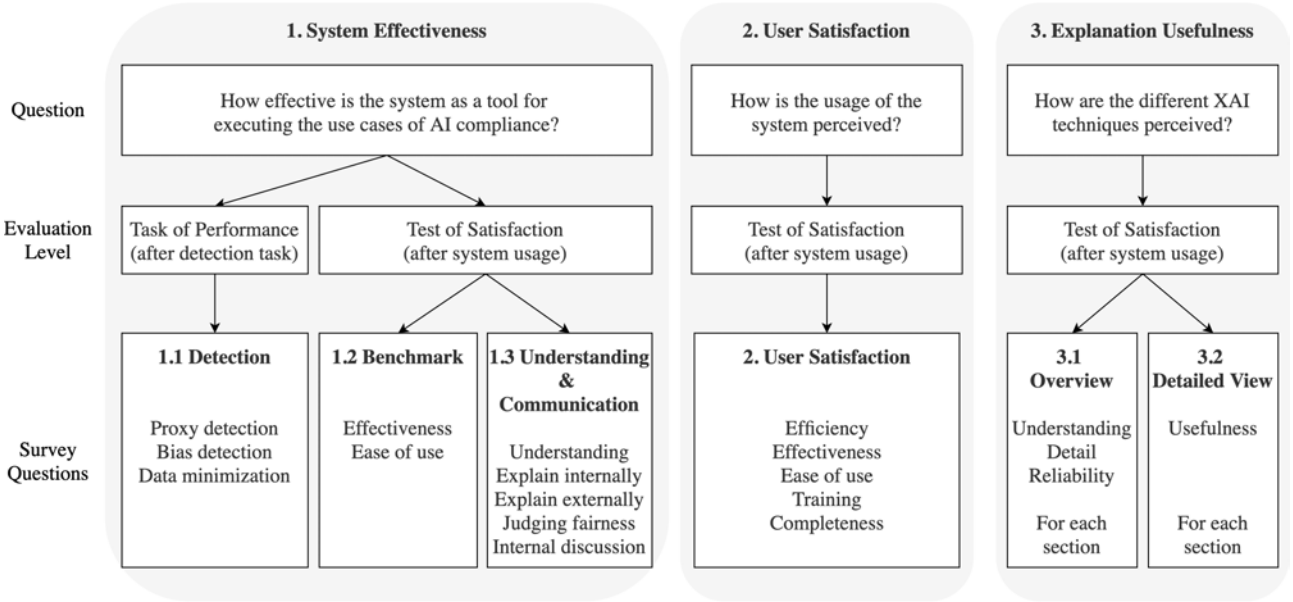


Figure 7.2: Overview of the three parts of the survey.

1. System Effectiveness As illustrated with Figure 7.2, the system effectiveness is determined by three parts of the survey. The detection task (1.1) consists of three tasks, of which the context is provided in the instruction video and in the survey:

- **Proxy detection**. *Are there input variables that may have a non-causal relationship with the output?* The participant can select none, one, or several variables from the list. In total, 5 of the 24 variables are proxies.
- **Bias detection**. *Are there any groups that the model performs less well for, and therefore disadvantages these groups?* The report presents bias insights for three different protected attributes (age, gender, and marital status), with a total of 12 groups. The participant can select none, one, or several groups from the list. One group (`Gender = unknown`) fails all criteria, one other group (`Gender = female`) fails the Equalized Odds criterion.
- **Data minimization**. *Are there any variables that could be omitted without affecting the performance of the model?* The participant can select none, one, or several variables from the list. A total of 17 out of the 24 variables can be omitted without increasing the mean absolute error of the model.

These tasks are an example of user task performance ('Human-AI Task Performance'), a frequently used evaluation method [60].

After the detection task, we establish a benchmark (1.2) by asking the participants how they normally would gain insight in the working of models. The participant can select that (i) this is the first time they have been involved with that, (ii) they have been working on this before, but could not measure it properly, or (iii) they have used other tools for model explainability. In the event that the participant has previously used a tool, the participant is asked to fill in the tool, and the following two statements on a five-point Likert scale:

- *This tool is more effective than the other tool.*
- *This tool is easier to use than the other tool.*

Finally, five questions are formulated to measure the systems effectiveness in enhancing model understanding and enabling communication (1.3). A good model understanding of the participants would indicate that the tool is effective, although we are aware that there are limitations to these user-reported results of the 'Test of satisifaction' approach [61]. The participant is asked to answer the following statements on a five-point Likert scale (strongly disagree - strongly agree):

- *I understand the working of the model.*
- *I can explain the working of the model to a colleague.*
- *I can explain the working of the model to a regulator or customer.*
- *The tool allows me to make a judgment about the fairness of this model.*
- *The tool allows me to have a discussion with colleagues about the fairness and workings of this model.*

2. User Satisfaction After the participant is forced to use the tool to answer the questions of the first part of the survey, we measure the user satisfaction on a Likert scale, a common evaluation method [60]. We base the statements on the dimensions 'Pperceived usefulness' and 'Perceived ease-of-use' of the Technology Acceptance Model [62], and consolidate them to reduce the number of questions for the participant:

- *Using this tool at work would help me work faster and more efficiently.*
- *Using this tool would increase my effectiveness at work by allowing me to make better decisions.*
- *I find this tool easy to use.*
- *Additional training on how the tool works would increase my effectiveness with the tool.*
- *From this tool, I can get all insights that I require to perform my work.*

3. Explanation Usefulness Using the Explanation Usefulness and Satisfaction evaluation method [60], we measure the user satisfaction with different report sections and thus the different explainability techniques. Our measurements are based on the Explanation Satisfaction proposed by Hoffman et al [61], from which we have picked three to reduce the time required for the survey. For each section of the 'Overview' - feature relevance section, feature insight section, and bias and fairness section - the participant is asked to answer the following statements on a 5-point Likert scale (strongly disagree - strongly agree):

- *This explanation helps me understand how the model works.*
- *This explanation of the model is sufficiently detailed for me.*
- *This explanation allows me to understand how reliable the model is.*

For the participants that have used the 'Detailed View', we only ask them to rate the usefulness of each section (SHAP force plot, SHAP summary plot, SHAP dependence plot) on a 5-point Likert scale (not useful at all - very useful) in order to limit the number of questions. The questions for each section are accompanied with a screenshot of the relevant section.

Experiment preparation The explainability report that is provided during the survey is based on the Injury Claim Provision Model (as described in Section 6.3), which we modify to inject proxies and biases - an approach inspired by the evaluation of LIME [8]. We select this case as it is a representative example of AI risks and possibilities within insurance, the intuitive model output (claim value in euro), and the clear predictive power of certain features and feature values. To protect customer data and respect internal compliance processes, we generate a new synthetic dataset using Python package `SDV` [63]. It leverages a Gaussian Copula to train a model on the distribution of the training data, and then uses this model to generate a new synthetic dataset with similar distribution. To improve the quality of the resulting dataset (measured with Chi-Squared and Inverted Kolmogorov-Smirnov D statistic tests), we support the model by custom-defining some distributions, and oversampling some less represented feature values.

This section has been redacted due to company confidentiality reasons.

We constructed the survey using Qualtrics[1], where we could put all necessary components. Participants are first shown information about the study, then give informed consent, input their demographic characteristics, watch a five-minute demonstration video about interpreting results from the report, and then download the HTML report from the browser. Then the aforementioned questions are asked which are accompanied by the necessary context to bring the non-AI experts up to speed. A draft of the survey was filled by a participant to evaluate the phrasing of the questions and the survey length. The final version was accessible to anyone within Achmea and we did not enforce that only certain roles could participate. We mainly targeted the stakeholders in the data analytics and compliance line (Figure 4.2) and strongly related roles, such as data engineers (related to data scientists) and data stewards (related to compliance officers), since we could more easily reach and enthuse these stakeholders than the business stakeholders who do not have a direct interest. To this end, we advertised the survey in internal communities and targeted the people and teams who were involved during the study. The survey was available for six weeks.

7.2 Results

This section presents the results of the survey in the order of question.

Participant demographics. A total of 30 insurance professionals participated in the experiment. We first examined whether the combination of the self-reported role and the line of participants matched our definition. For four participants, we corrected the line to our definition, as they reported their hierarchical line (e.g., a business or IT division) while they fulfill a compliance role (e.g. quality manager or data steward). We did not make any further corrections and used all responses. Out of the 30 participants, 16 have a data analytics role, 4 have a business role, and 10 have a compliance role. The data analytics roles include data scientist (6), data engineer (3), and data consultant (2). A manager, director, business consultant, and data analyst represent the participating business roles. The compliance roles include data steward (3), compliance officer (2), quality manager (2), and model validator (1). The working experience strongly varies between the different lines.

[1] `https://www.qualtrics.com/`

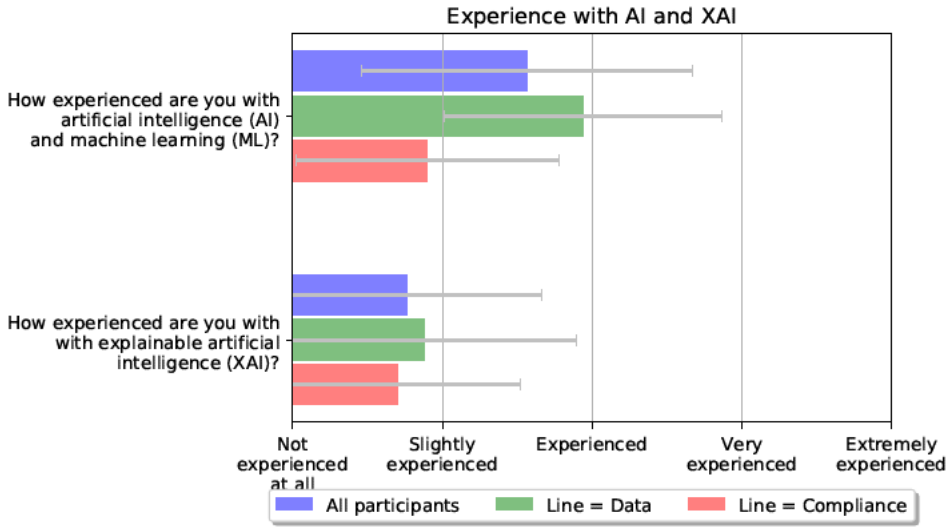


Figure 7.3: Average self-reported AI and XAI experience of the participants.

Participants in the data analytics line are mainly medior (years of experience = 4.44 ± 2.68), while the participating employees of the other lines are mainly senior (business = 16.75 ± 16.03, compliance = 12.10 ± 11.99). Figure 7.3 displays the average self-reported AI and XAI experiences. Data analytics participants are more experienced with AI and ML than the compliance participants (data = 2.94 ± 0.93, compliance = 1.90 ± 0.88, all participants = 2.57 ± 1.10). This difference is smaller for XAI, as most participants indicate to have little to no experience in this field (data = 1.88 ± 1.02, compliance = 1.70 ± 0.82, all participants = 1.77 ± 0.90).

Because of the different backgrounds and interests of the participants, we will make distinctions between the data and compliance lines when reporting certain results. We consider the sample size for the business line (4 participants) too small to arrive at representative insights for this group. The business line will therefore not be displayed separately, but the responses are included in the results of all participants.

1.1 Detection tasks For the proxy detection task, a majority of the participants identified the features 'Year of accident' and 'Marital status' as proxy (19 times), followed by 'Gender' (13 times), 'Country accident' (9 times), and 'Profession' (8 times). However, many participants also marked features that are not proxies in our opinion. We therefore calculated a 'net score' for each participant: the number of correctly identified proxies minus the the number of incorrectly selected features, with a minimum score of 0. The distribution of these scores is shown in Figure 7.4. Half of the participants was able to correctly identify one or multiple features as a proxy, corrected for incorrectly selected features. The average participant had a net score 0.93 ± 1.06 out of five (19%), meaning that roughly one additional proxy was correctly identified than incorrectly identified.

Twenty-six and twenty-three participants correctly identified the model unfairness towards groups `Gender = unknown` and `Gender = female`, respectively. We used the approach of the proxy detection task to compute the net score of the bias detection task. Figure 7.5 presents the distribution of the net scores: 19 participants correctly identified both biases without identifying an incorrect one, 7 identified both correct biases and one incorrect one, or just the two correct proxies, and 4 participants had a score of 0. On average, 1.50 out of two biases were identified correctly (75%), with a standard deviation of 0.72.

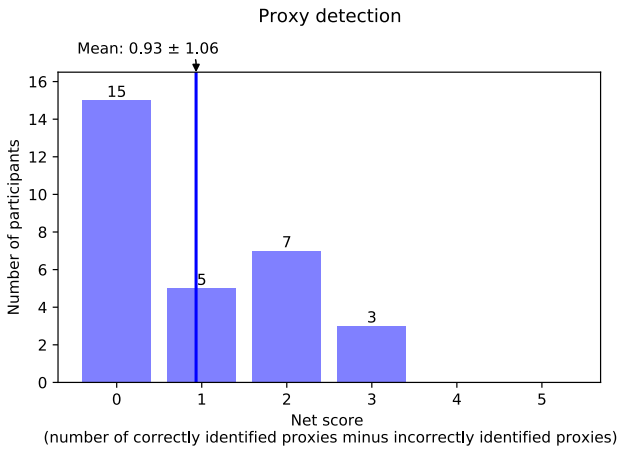


Figure 7.4: Score distribution for the proxy detection task.

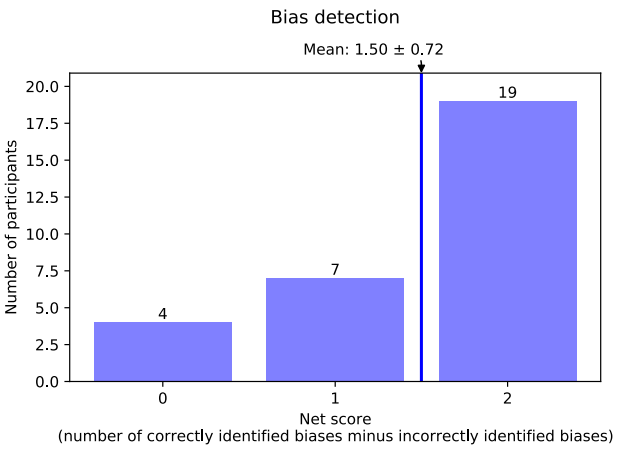


Figure 7.5: Score distribution for the bias detection task.

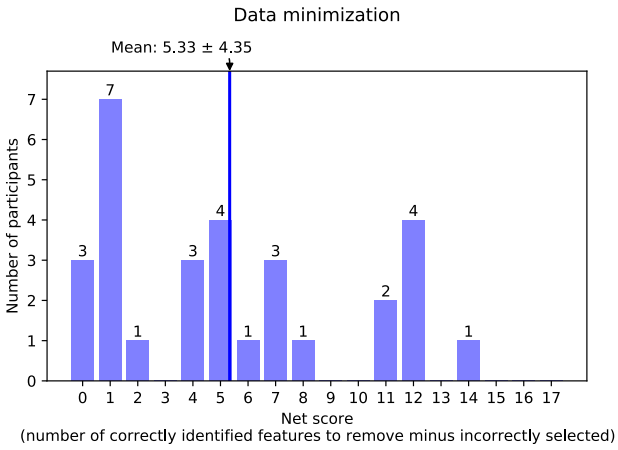


Figure 7.6: Distribution of the net scores for the data minimization task.

For the data minimization task, 28 out of 30 participants marked at least one feature for removal, with an average of 5.67. Using a similar approach to compute the net score, we observed that the average participant scored 5.33 ($\pm$4.35) out of 17 (31%), meaning that there was roughly 5 additional features correctly marked than incorrectly marked for removal. The distribution is displayed in Figure 7.6.

1.2 Benchmark After the detection tasks, the participants were asked to specify how they would normally gain insights in the working of models. Nine participants (30%) answered that this was their first experience with model interpretability, and five participants (17%) answered that they have been working on this before, but could not measure it properly. Sixteen participants (53%) answered that they used other tools for model interpretability, including SHAP (3), Azure Model Explainer (1), a custom-built XAI system (1), SAS (1), and unspecified Python packages (2). Several participants did not specify a tool, but mentioned that they inspected models for built-in feature importance functionalities of ensemble models (2), local effects (2), error plots and model fit (2), correlations (1) or distributions (1). The 16 participants were asked to compare the system with previously used tools in terms of effectiveness and ease of use. Figure 7.7 displays the results of this benchmark. The participants rated both the effectiveness and ease of use of the tool with a 4.06 ± 0.85 on a five-point Likert scale, 1.06 higher than we would expect our system performed similarly to the other tools.

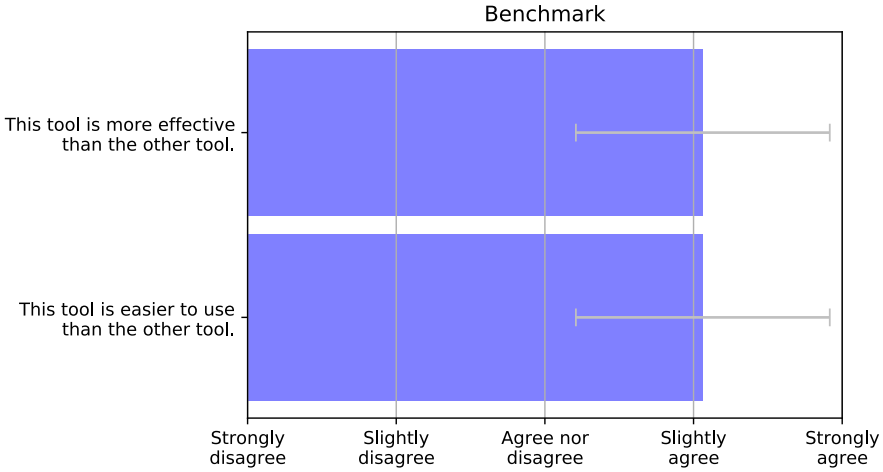


Figure 7.7: Average response of 16 participants that compared our system with a previously-used tool.

We perform a statistical test to determine whether this difference is significant. As a benchmark, we generate a sample of Likert-scale responses, equal to the length of our responses (16), with a mean of 3.000. The mean of 3.000 represents 'neutral' responses, thus no difference between our system and other tools. We choose the benchmark sample to be completely random, since we cannot assume that Likert-scale responses are normally-distributed [64]. We use the non-parametric Kruskal-Wallis H Test as the statistical test, as our responses are not normally distributed based on a Shapiro-Wilk test, on a ordinal scale, and too small to use a Mann Whitney U Test reliably. This non-parametric test is performed on our responses and the generated benchmark sample to determine whether there is a significant difference In addition, we compute *Cohen's d* to determine the size of the effect. For generalizablity, these results are averaged over 1000 iterations, in which we use different generated samples and keep our responses constant. Table 7.1 presents $p = 0.009$ and $d = 1.15$ for both the effectiveness and ease of use, indicating a significant ($p < 0.05$) and large effect ($d > 0.8$).

Table 7.1: Results of the statistical tests that compared our responses with a benchmark sample, averaged over 1000 samples.

Characteristic	Responses	Benchmark	p	Cohen's d
Effectiveness	4.063 ± 0.854	3.000 ± 1.050	0.009	1.148
Ease of use	4.063 ± 0.854	3.000 ± 1.054	0.009	1.151

1.3 Understanding and communication Figure 7.8 shows the average response of all participants (blue), participants with a data role (green), and participants with a compliance role (red) to the five questions that regard model understanding and communicating these insights. Following the aforementioned approach, we compared the responses of all participants to benchmark samples to determine the significance $p_{benchmark}$ and effect size $d_{benchmark}$. Additionally, we determined the significance p_{groups} of the difference between the response to a data role and a compliance role. The results are presented in Table 7.2.

Regarding model understanding, we observe an average response of 3.767 ± 0.935 on a 5-point Likert scale. This difference of 0.767 with the benchmark samples is significant ($p_{benchmark} = 0.002$) and medium-sized effect ($d_{benchmark} = 0.787$). Those with data roles indicated that they have a better understanding of the model (4.000

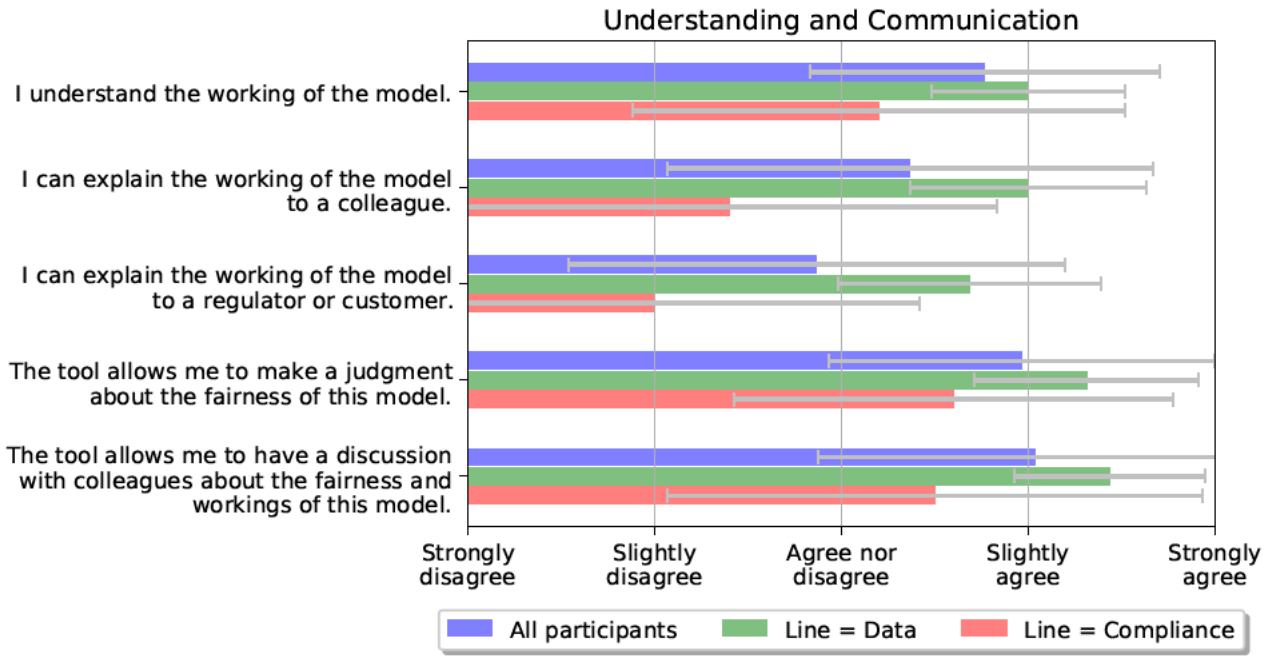


Figure 7.8: Average response of all participants (30), participants with data roles (16), and participants with compliance roles (10) to questions regarding understanding and communication.

$\pm$ 0.516) than those in a compliance role (3.200 $\pm$ 1.317), although this difference is not significant (p_{groups} = 0.120). This difference increases and becomes significant when asked whether they can communicate their understanding of the model internally (data = 4.000 $\pm$ 0.632, compliance, = 2.400 $\pm$ 1.430, p_{groups} = 0.003) or externally (data = 3.688 $\pm$ 0.704, compliance, = 2.000 $\pm$ 1.414, p_{groups} = 0.003). Although the data roles have no difficulty with the internal communication of their model understanding, they do report a lower confidence for communication model understanding externally, leading to a mean of all participants of 2.867 on the 5-point scale. The participants report higher agreement with the statement that the tool allows to make judgments about the fairness of the model (3.967 $\pm$ 1.033), with a significant and large effect compared to the benchmark. The average response of a data role is higher than that of a compliance role (4.312 $\pm$ 0.602 compared to 3.600 $\pm$ 1.174), although this difference is not significant. We observe similar results regarding discussion with co-workers (all participants = 4.033 $\pm$ 1.159, data = 4.438 $\pm$ 0.512, compliance = 3.500 $\pm$ 1.434). Considering all points, compliance participants reported lower agreement than those with data roles, with a higher standard deviation.

Table 7.2: Average response regarding understanding and communication of all participants, data roles, and compliance roles. $p_{benchmark}$ and $d_{benchmark}$ indicate the significance and Cohen's d between all participants and a benchmark sample (averaged over 1000 samples). p_{groups} indicates the significance in the difference between the data and compliance roles.

Statement	All participants	Line=Data	Line=Compl.	$p_{benchmark}$	p_{groups}	$d_{benchmark}$
Understanding	3.767 $\pm$ 0.935	4.000 $\pm$ 0.516	3.200 $\pm$ 1.317	0.002	0.120	0.787
Explain internally	3.367 $\pm$ 1.299	4.000 $\pm$ 0.632	2.400 $\pm$ 1.430	0.155	0.003	0.316
Explain externally	2.867 $\pm$ 1.332	3.688 $\pm$ 0.704	2.000 $\pm$ 1.414	0.768	0.003	-0.113
Judging fairness	3.967 $\pm$ 1.033	4.312 $\pm$ 0.602	3.600 $\pm$ 1.174	$<$ 0.001	0.083	0.949
Internal discussion	4.033 $\pm$ 1.159	4.438 $\pm$ 0.512	3.500 $\pm$ 1.434	$<$ 0.001	0.056	0.953

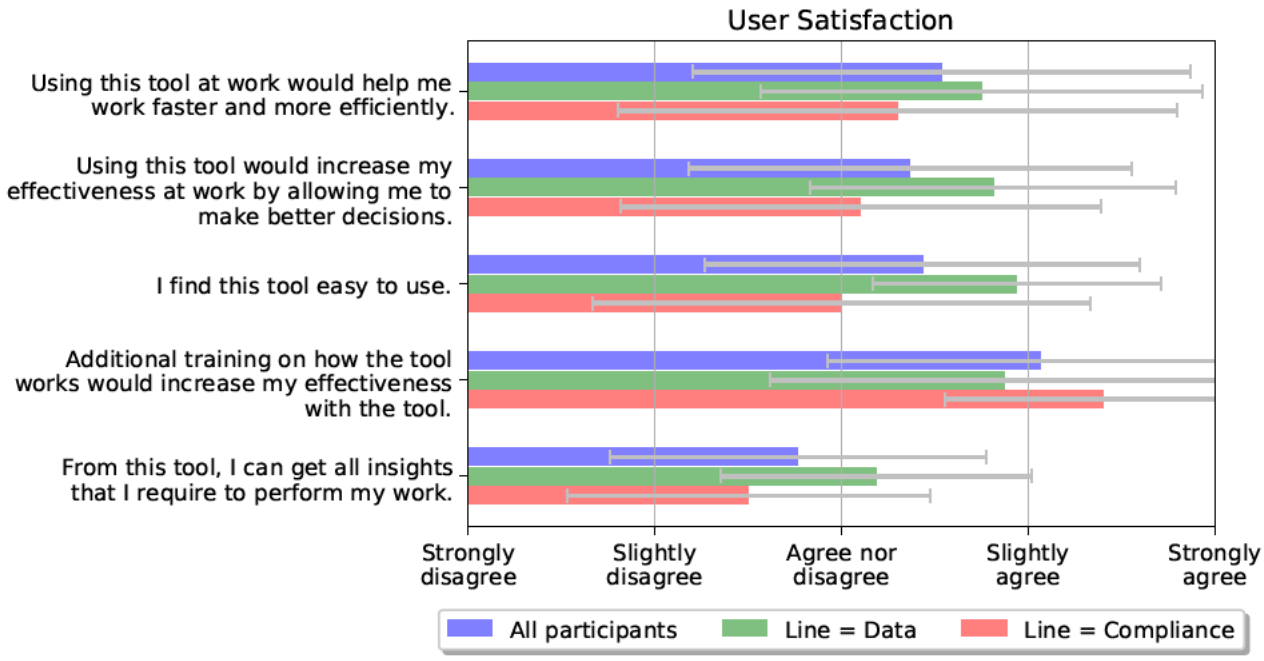


Figure 7.9: Average response of all participants (30), participants with data roles (16), and participants with compliance roles (10) to questions regarding their satisfaction with the system.

2. User Satisfaction The results of the user satisfaction questions are reported in a similar format with Figure 7.9 and Table 7.3. The participants reported a slight agreement that the tool would help them in terms of efficiency (3.533 ± 1.332); a small but not significant effect compared to the benchmark. The system's effectiveness and ease of use are scored similarly (3.367 ± 1.189 and 3.433 ± 1.165) but lack statistical significance with the benchmark. Participants with data roles reported a higher agreement with the statements than those with compliance roles for all of the three aforementioned statements, although this difference is not significant. All participants, but mainly the participants from the compliance field, indicated that additional training would increase their effectiveness with the tool (4.067 ± 1.143 and 4.400 ± 0.843). The participants are the least congruent in that the system provides all the insights that they need to perform their work, with a score of 2.767. Overall, we observe that participants with data roles have a higher satisfaction than participants with compliance roles, although this difference is not significant.

Table 7.3: Average user satisfaction of all participants, data roles, and compliance roles. $p_{benchmark}$ and $d_{benchmark}$ indicate the significance and Cohen's d between all participants and a benchmark sample (averaged over 1000 samples). p_{groups} indicates the significance in the difference between the data and compliance roles.

Statement	All participants	Line=Data	Line=Compliance	$p_{benchmark}$	p_{groups}	$d_{benchmark}$
Efficiency	3.533 ± 1.332	3.750 ± 1.183	3.300 ± 1.494	0.070	0.444	0.453
Effectiveness	3.367 ± 1.189	3.812 ± 0.981	3.100 ± 1.287	0.179	0.103	0.333
Ease of use	3.433 ± 1.165	3.938 ± 0.772	3.000 ± 1.333	0.132	0.068	0.398
Training	4.067 ± 1.143	3.875 ± 1.258	4.400 ± 0.843	< 0.001	0.305	0.992
Completeness	2.767 ± 1.006	3.188 ± 0.834	2.500 ± 0.972	0.317	0.068	-0.231

3. Explanation Satisfaction Figures 7.10, 7.11, and 7.12 and Table 7.4 display the responses regarding the explanation satisfaction of the different sections of the 'Overview'. The participants report that the feature importance section has a significant ($p_{benchmark} = 0.003$) and large ($d_{benchmark} = 0.848$) effect on their under-

Table 7.4: Average response regarding explanation satisfaction of all participants, data roles, and compliance roles. $p_{benchmark}$ and $d_{benchmark}$ indicate the significance and Cohen's d between all participants and a benchmark sample (averaged over 1000 samples). p_{groups} indicates the significance in the difference between the data and compliance roles.

Statement	All participants	Line=Data	Line=Compl.	$p_{benchmark}$	p_{groups}	$d_{benchmark}$
FI - Understanding	3.833 ± 0.950	4.000 ± 0.816	3.700 ± 1.059	0.003	0.506	0.848
FI - Detail	3.433 ± 1.305	3.688 ± 1.138	3.200 ± 1.398	0.123	0.414	0.373
FI - Reliability	2.300 ± 1.119	2.438 ± 1.153	2.000 ± 1.054	0.013	0.335	-0.658
PDP - Understanding	3.567 ± 1.223	4.000 ± 0.966	2.900 ± 1.449	0.020	0.033	0.507
PDP - Detail	3.567 ± 1.305	4.062 ± 0.998	3.100 ± 1.449	0.039	0.075	0.487
PDP - Reliability	2.667 ± 1.295	3.000 ± 1.265	2.500 ± 1.179	0.312	0.286	-0.288
B&F - Understanding	3.900 ± 0.923	3.750 ± 1.000	4.000 ± 0.816	0.001	0.495	0.930
B&F - Detail	3.933 ± 1.081	3.938 ± 1.181	3.700 ± 1.059	0.001	0.473	0.893
B&F - Reliability	3.700 ± 1.119	3.625 ± 1.204	3.900 ± 0.738	0.008	0.798	0.658

standing of how the model works (3.833 ± 0.950). This section scores lower on the sufficiency of details (3.433 ± 1.305) and ability to assess reliability (2.300 ± 1.119). We observe no significant differences between the groups. Comparable to the feature importance section, the partial dependence plot has a significant ($p_{benchmark}$ = 0.020) and medium ($d_{benchmark}$ = 0.507) effect on model understanding (3.567 ± 1.223), in which there is a significant difference (p_{groups} = 0.033) between data roles (4.000 ± 0.966) and compliance roles (2.900 ± 1.449). Participants report a similar agreement with the statement that this section is sufficiently detailed (3.567 ± 1.305). Despite a higher score than the feature importance section, the partial dependence underperforms on the basis of presenting insights into model reliability (2.667 ± 1.295). The bias and fairness section has a significant positive effect on the participants ability to understand the model (3.900 ± 0.923, $d_{benchmark}$ = 0.930) and assess it reliability (3.700 ± 1.119, $d_{benchmark}$ = 0.658). The participants also report that this section is sufficiently detailed with an agreement of 3.933 on the 5-point Likert scale. The aforementioned applies for all participants, as there is no significant difference between the groups.

Figure 7.13 displays the averaged score for the sections of the 'Overview' and 'Detailed View'. We observe that the bias and fairness section has the highest rating of the 'Overview' section. The sections of the 'Detailed View' are all similarly useful, according to 16 participants that used the 'Detailed View' during the survey.

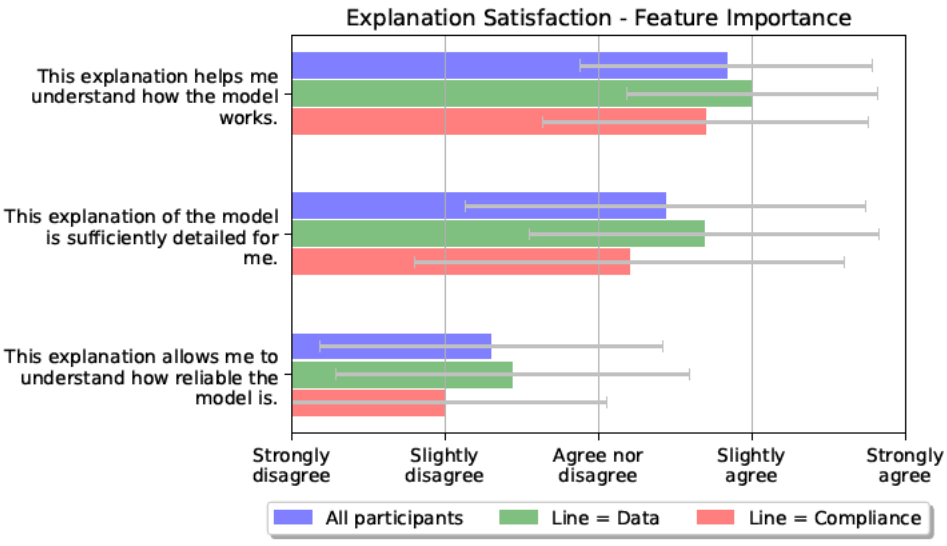


Figure 7.10: Explanation satisfaction of the feature importance plots in the 'Overview'.

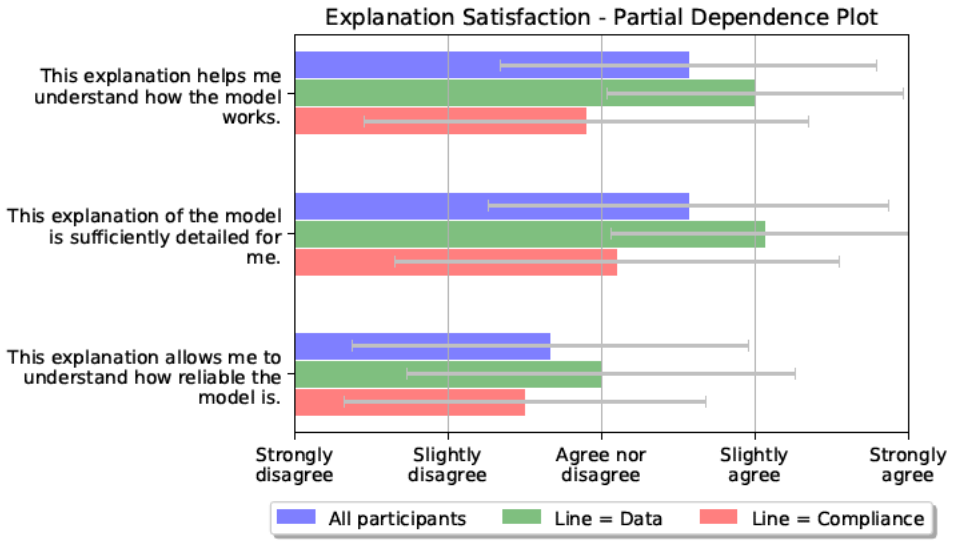


Figure 7.11: Explanation satisfaction of the Partial Dependence Plot.

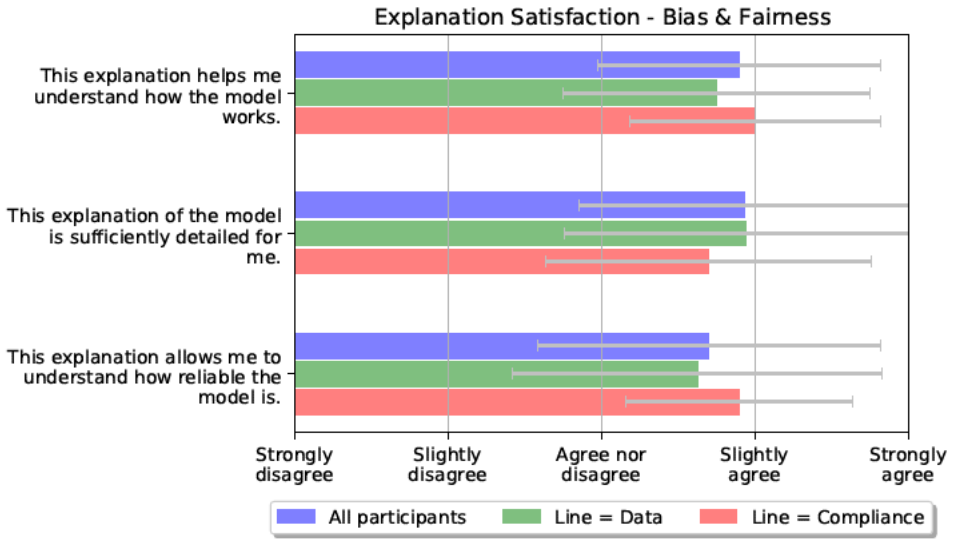
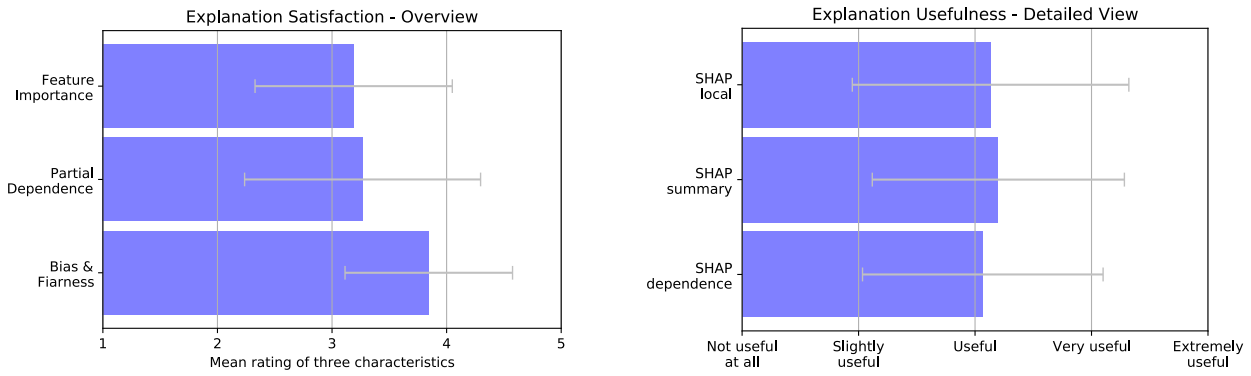


Figure 7.12: Explanation satisfaction of the Bias & Fairness section.

Participant remarks Finally, the participants were free to leave a remark about the survey, which was done by 17 participants. We interpreted the remarks and attached one or multiple labels to the remarks, resulting in 12 types of labels that together were mentioned 32 times.

Seven participants left positive remarks regarding the appearance and usefulness of the report, and the benefits of the system for Achmea. Four participants remarked that the insights and explanations in the report are too technical for some stakeholders, and one participant remarked that the tool has a learning curve. Three participants questioned why the report was in English, while the main language within Achmea is Dutch. Eight remarks were left regarding improving and supplementing the insights in the report: three participants would like to have an indication of the reliability and significance of the insights, three others would like to see more insights in the model performance (more than just the Mean Absolute Error that is presented at the top of the report), one participant would like the option to specify the reliability and trustworthiness



(a) Overview

(b) Detailed View

Figure 7.13: Explanation satisfaction summarized for the 'Overview' and 'Detailed View'

of the model beforehand, and a model validator would like to assess the impact of interactions and model correlations on the model performance.

Regarding the survey, two participants found that a clear definition of trustworthiness was lacking - which could also be presented in the report. One participant commented that the survey was long, and one commented that the survey was difficult. Two participants remarked that they normally do not work with models, two participants found it difficult to understand the model, and one participant mentioned both. Four of these five participants have a compliance role.

7.3 Discussion

This section elaborates on the results and limitations of the system evaluation.

During the detection tasks, the average participant scored 19%, 75%, and 30% for the proxy detection, bias detection, and data minimization, respectively. Based on these results, we state that the system is most effective for bias detection. Note that these scores are based on the interpretation of a single untrained stakeholder within a short amount of time, while in the real-world scenario multiple more experienced stakeholders would jointly interpret the results and do follow-up investigations on the findings. We therefore expect even better performance in the real application.

Due to the survey, one-third of the participants have received some experienced with model interpretability. One out of five participants now have a system to interpret models, where before they could not. Half of the participants compared the system with a previously used tool, and indicated that this system was more effective and easier to use (both 4.063 ± 0.854 on a 5-point Likert scale) - a significant and large effect. We can therefore conclude that, on average, the system is beneficial for all types of stakeholders.

The participants self-reported that the system has a significant medium-sized effect on their model understanding (3.767 ± 0.935), and a significant large effect on their ability to judge a model's fairness (3.967 ± 1.033) and ability to discuss the working and fairness with co-workers (4.033 ± 1.159). The reported effectiveness for judging fairness aligns with the participants performance on the bias detection task. The enhanced

model understanding and improvement of internal discussions were not specially tested during the experiment, but were noted by involved stakeholders during the cases (Section 6.3). Participants in the line of compliance reported to have more difficulty than participants in the line of data to explain the working of the model internally (-1.600) and externally (-1.688).

We observe a positive but not significant effect regarding efficiency (3.533 $\pm$ 1.332) and effectiveness (3.367 $\pm$ 1.189). We believe that this is due in part to the fact that model interpretability is not an equal role for every stakeholder, which was also evidenced by the remarks of some participants. The same goes for the completeness of the report (2.767 $\pm$ 1.006), for which we received several useful additions. However, we do note that the formulation of this question might have been misleading, as it questioned whether the report was complete enough to execute a participants' activities, which clearly is much broader than just interpreting models. The participants indicated that a training would be useful to increase their effectiveness with the system (4.067 $\pm$ 1.143). As the participants had hardly any experience in using the system, we believe that a decent training before the survey would increase the participants' performance in the detection tasks and their satisfaction. Despite the fact that there is not always a significant difference, we do note that participants in data roles have higher satisfaction than those in compliance roles. A potential solution is making the explanation less technical (as proposed by three participants), and improve the knowledge of compliance roles by means of training.

All explanations in the 'Overview' have a significant effect on the participants model understanding. This is a large effect for the feature importance and bias and fairness sections, and a medium effect for the partial dependence plot - this latter is mainly caused by a significant lower agreement for compliance roles (2.900 $\pm$ 1.449) compared to data roles (4.000 $\pm$ 0.966). Regarding the amount of detail, the feature importance section and partial dependence plot perform similarly with 3.433 $\pm$ 1.305 and 3.567 $\pm$ 1.305, respectively, while the level of detail is highly appreciated for the fairness section (3.933 $\pm$ 1.081). The feature importance section and partial dependence plot mainly underperform in the area of enhancing reliability (2.300 $\pm$ 1.119 and 2.667 $\pm$ 1.295), while the bias and fairness section has a medium effect on the participants ability to assess the reliability of a model (3.700 $\pm$ 1.119). This is in line with our expectations, and shows that multiple techniques in an XAI system can effectively complement each other. The system's significant ability to support judging model fairness is inextricably linked to the bias and fairness section. Of all sections in the 'Overview', the participants are most satisfied with the bias and fairness section. The three sections in the 'Detailed View' are similarly useful.

This evaluation approach has three main limitations. First of all, this application-grounded evaluation mimiced several conditions of actual system applications, but did not account for the joint interpretation of the results, joint establishment of definitions, and the opportunity to further investigate the insights that are presented by the report. Because of this, the results of the detection tasks are unsubtle and miss nuances that would normally occur. Nevertheless, we believe this was the best suitable approach with the resources we had at our disposal. Secondly, the sample size of 30 participants limits our ability to statistically prove the significance of some results. We therefore did not include the business line in our comparisons. Grouping the participants based on their line results in heterogeneous groups, which might cause the loss of possible

characteristics of certain participants. More participants would enable more granular grouping of the participants, for example, on their actual role. Finally, apart from the detection task, all results were self-reported by the participants, making it infeasible to verify the results, e.g., whether a participant genuinely understands the model. As a result, we cannot exclude that some participants suffer from an 'Illusion of Confidence', where they blindly trust the outcome of the system instead of reasoning by themselves. Since we were aware of this concept, we deliberately added as little value judgment as possible to the system, such as marking the features that could be omitted without impacting model performance.

Chapter 8

Application in Other Domains

During this research, we focused on an XAI system for AI compliance in the financial sector, and further scoped by conducting the study at an insurance firm. This raises the question whether and how the system can be deployed and used in other organizations and domains - the fifth subquestion of this research. To this end, this chapter briefly discusses the generalizability of the system and provides guidance how the system can be applied in other organizations and sectors.

Sector Regardless of the sector, the main purpose of the system is supporting compliance with AI regulations. In this study, we focused on the standards of the Ethical Framework, which are applied by the majority of Dutch insurers. Despite this narrow scope, the concerns that we addressed with the system, such as having a control framework in place and checking models for potential biases, recur in virtually every AI regulation. With the proposed Artificial Intelligence Act by the European Commission, these standards are enforced in any organization that leverages AI models that have an impact in the European Union, whereby this law also affects organizations outside this region. We therefore believe that our system supports compliance with AI regulations for any organization that applies AI within the EU, regardless of the sector. This relevance is more significant for sectors where models typically have a high impact on users, such as in the financial, medical, and public domain, as these high-risk systems are subject to more regulations in the EU AI Act.

More broadly, the system aims to stimulate the use of responsible and ethical AI, regardless of the presence of legal frameworks. Although regulations provide some guidance on the desirable use of AI, ultimately organizations themselves are responsible for taking their social responsibilities. In other words, responsible and ethical business processes do not revolve around what is legally allowed, but what is desirable for society. We believe this system is well-suited for supporting this vision. Organizations should determine the society impact of their AI-driven systems and the adverse effects it may have. For example, a retail organization may have an AI-driven recommendation engine that recommends nutrition based on personal characteristics such as gender, age, and shopping patterns. During a risk analysis, the organization may identify that such a system can cause certain groups of people to be recommended with less healthy food, which is not prohibited by law but undesirable from an ethical point of view. Based on this risk assessment, the organization should

formulate a specific set of rules for AI-driven systems that mitigate the risk, e.g., that gender should not have a decisive role in the predictions and therefore not among the 50% most influential features. We recommend that this analysis and formulation of rules are carried out by an interdisciplinary group of users before predictive models are developed. After model development, our system can be used to validate whether the model meets the formulated criteria in a quantitative manner. Note that it is also conceivable that there are AI-driven systems that have hardly any societal impact, and thus that little or no rules need to be drawn for these systems.

Regardless of the regulations a model is subject to, or the social impact it has, it is typically desirable to understand and validate the internal working, and ultimately improve the model's performance - the system's use case of model refinement. As this use case does not significantly differ between sectors, we believe that this use case is applicable to any sector. In summary, we believe our system can be applied to any domain, and is particularly beneficial in heavily-regulated areas such as the financial, medical, and public domain.

Organization types Given that the system is relevant for most sectors, we now address the types of organizations that are best suited for the use of the system. Overall, the system can be applied to any organization that employs AI, although there are some types of organizations where the system would be particularly effective compared to other XAI solutions. It concerns organizations with (i) large compliance departments, (ii) many roles involved in the development process, and (iii) several data science teams scattered across the organization - exactly the conditions we took into account when establishing the design criteria. These types of organizations benefit the 'multi-stakeholder information' property, the standardization of XAI, and the easing of internal discussion and communication, the most. In addition, this turnkey system can be highly beneficial for smaller organizations that do not have the technical resources to develop such a system by themselves. Note that in smaller organizations, roles should remain separate for trustworthy compliance, i.e., that a person other than the model developer should evaluate the model. However, smaller organizations with large (centralized) technical teams might want a system that is fully tailored for data scientists. Besides, certain organizations may only work with unstructured data, which is currently not supported by the tool. Regardless of the aforementioned points, a prerequisite is that an organization has a good data maturity.

System Deployment An organization that meets the aforementioned criteria might want to adopt this system. Figure 8.1 displays the eight steps that we formulated as guidelines for adoption of the system. These steps are largely based on the best practices of deployment at the host organization and system evaluation. First of all, organizations that already have a framework for model management can first consider how to integrate the system with this framework, so that all applications to control models are communicated and offered in a uniform manner. Second, the system should be deployed as a Python package in the default development environment, with the goal to remove the hurdles for using the system. To this end, the development environment should have sufficient computational resources, which is typically present in organizations that employ machine learning. To ensure that the system is used by data scientists, it should be adopted in the default development process, preferably as an acceptance criterion for the first proof of concept and all following versions. Then, an interdisciplinary team should perform a societal impact assessment on the use of

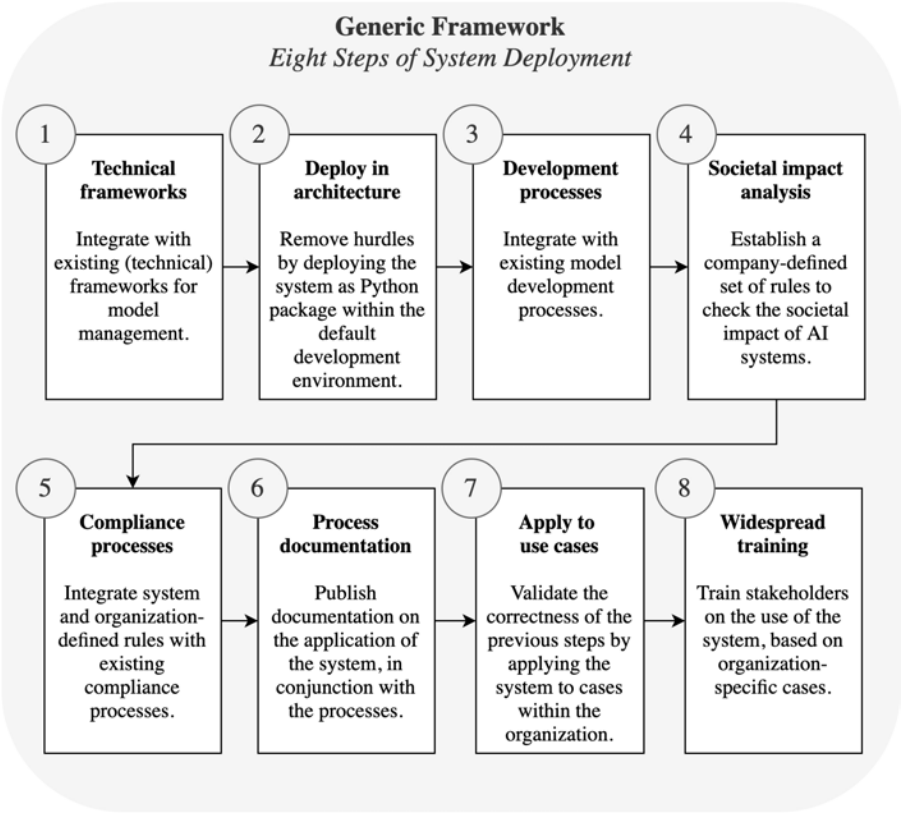


Figure 8.1: The generic framework that describes the eight steps to deploy the system at any organization.

AI within the organization, with the aim to define a set of organization-specific rules that can be used to check models for potential harmful effects. These organization-specific rules can be supplemented with rules specific for certain use cases. Subsequently, these rules should be integrated with the existing compliance processes, resulting in one framework of rules a model should adhere to. The compliance processes should refer to the report that is produced by the system that can be used to evaluate whether the model meets these criteria. Note that these criteria are not part of the report, but that the insights that can be extracted from the report ease the assessment process. The outcomes of these assessments should be stored for auditing purposes. The sixth step is to document the aforementioned steps and make it accessible for the entire organization. Apply the system to organization-specific cases to evaluate the document steps, and as a representative example of how the system works within this domain - beneficial for the last step. Finally, based on the outcomes of the system evaluation, we recommend setting up a training for the stakeholders about how to apply the system. The organization-specific cases can be used in this training to engage the stakeholders. By executing these eight steps, we expect that the system will optimally support the ethical use of AI within any organization.

Chapter 9

Conclusion

Organizations of all sizes report that they struggle with AI adoption due to the legal, ethical, and regulatory concerns that are caused by the lack of model transparency and the lack of tools that can directly be implemented to address this on an enterprise scale. This study addresses this problem by designing, developing, deploying, applying, and evaluating a turnkey explainable AI system that supports compliance with AI regulations in the financial sector. Using existing model-agnostic XAI techniques, the system developed in this study generates a understandable, highly configurable, and easily shareable report with model insights for any supervised learning model. System application enhances model understanding, enabling two primary use cases: AI compliance and model refinement. First, the system supports AI compliance as it enables proxy detection, bias detection, data minimization, and improves internal communication between, and decision making by, stakeholders. Second, the insights might be used for validating and refining models, leading to more accurate and more ethical predictive models. The aforementioned use cases and the system's broad applicability have been demonstrated by applying the system to four machine learning models at the host organization. The system can easily be deployed by installing it as a Python package, as demonstrated at the host organization. Following the eight implementation steps of the generic framework, this turnkey system also guarantees quick implementation in other organizations.

To quantify the effectiveness, we performed an application-grounded 5-point Likert scale survey with 30 insurance professionals. The experiment indicated that the system has a significant effect on participants' model understanding (3.767 ± 0.935, a medium effect), internal communication (4.033 ± 1.159, a large effect), and ability to assess model fairness (3.967 ± 1.033, a large effect) - for the latter, the participants also demonstrated their ability to identify biases. The functionalities to assess the model for biases and fairness are most valued by participants in terms of detail, enhancing understanding, and assessing a model's trustworthiness. Additionally, the participants perceived this tool to be significantly more effective and easier to use (both 4.063 ± 0.854, a large effect) than previously used tools.

Overall, we conclude that the system delivers business value in three ways. First and foremost, the system provides better AI compliance as it improves model understanding, communication, and the ability to assess model fairness. Second, the application of the system adds immediate value to the business through model refinement. Finally, using this system instead of other tools enables the standardization of the AI compliance process, enhancing effectiveness and ease-of-use.

During the project, we found two interesting and additional benefits of the system. First, with our aim to serve non-technical stakeholders, we found that our methods to serve comprehensible insights, such as configurable feature names, grouping of features, and an interactive report, were highly valued by both technical and non-technical stakeholders. Additionally, we observed that the implementation of this system has a major effect on the awareness on the ethical use of AI, even beyond our target group. With the aforementioned benefits, this study offers a turnkey system for more ethical use of AI, removing regulatory hurdles for AI adoption.

There are several limitations to our system and approach. First of all, the system supports traditional supervised machine learning models that are trained on tabular data. Although the vast majority is currently of this type, this is rapidly shifting to deep learning models that are trained on unstructured data, for example language models that perform predictions based on texts such as customer inquiries. Near-future modifications are therefore required to keep the system future-proof. Second, the system explains the inner workings of models, but does not automatically detect or fix incorrect models. Therefore, merely applying the system is not sufficient for AI compliance, and requires proper interpretation and follow-up, which is more difficult for organizations to implement than just a technical system. Finally, there are limitations to the evaluation of our system, as (i) the sample size limited our ability to identify significant differences between stakeholders, (ii) the self-reported results did not allow the validation of the true understanding of participants, and (iii) our participants had little to no experience with the system.

Our recommendations for future work are fourfold. First of all, the evaluation of the system could be repeated after the users have received training and gained more experience with the tool. This will ensure more faithful results regarding the system's effectiveness, and insights on the impact of system training. Additionally, future work could deepen our understanding of how different types of explanations can best be explained to non-technical stakeholders such as compliance officers, in order to increase the effectiveness of XAI systems. Another direction is to study the difference between model-specific and model-agnostic XAI techniques on both the technical performance and functional suitability, with the aim to identify the best approach for future XAI systems. Finally, additional research in the field of XAI is required to enable generic and global insights for the next generation of models, such as deep learning and reinforcement learning on unstructured data, to continue facilitating responsible use of AI.

Bibliography

[1] IBM, "Global AI Adoption Index 2021," tech. rep., 2021. https://filecache.mediaroom.com/mr5mr_ibmnews/190846/IBM's%20Global%20AI%20Adoption%20Index%202021_Executive-Summary.pdf.

[2] J. Bughin, E. Hazan, S. Ramaswamy, M. Chui, T. Allas, P. Dahlström, N. Henke, and M. Trench, "Artificial Intelligence: The Next Digital Frontier?," tech. rep., McKinsey Global Institute, 2017.

[3] S. Ransbotham, D. Kiron, P. Gerbert, and M. Reeves, "Reshaping Business With Artificial Intelligence," *MIT Sloan Management Review*, 2017.

[4] S. Corbett-Davies, E. Pierson, A. Feller, and S. Goel, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.," *Washington Post*. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

[5] J. Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women," *Reuters*, Oct. 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[6] "SyRI legislation in breach of European Convention on Human Rights," tech. rep., 2020. https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, pp. 1–46, July 2015.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *CoRR*, vol. abs/1602.04938, 2016. http://arxiv.org/abs/1602.04938.

[9] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *CoRR*, 2017. http://arxiv.org/abs/1705.0787.

[10] J. Bughin, J. Seong, J. Manyika, M. Chui, and R. Joshi, "Modeling the global economic impact of AI." https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy.

[11] Fiscale inlichtingen- en opsporingsdienst, "ING betaalt 775 miljoen vanwege ernstige nalatigheden bij voorkomen witwassen," Sept. 2021. https://www.fiod.nl/ing-betaalt-775-miljoen-vanwege-ernstige-nalatigheden-bij-voorkomen-witwassen/.

[12] Netherlands Public Prosecution Service, "ABN AMRO pays EUR 480 million on account of serious shortcomings in money laundering prevention," https://www.prosecutionservice.nl/latest/news/2021/04/19/abn-amro-pays-eur-480-million-on-account-of-serious-shortcomings-in-money-laundering-prevention.

[13] European Parliament and the Council of the European Union, "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." http://data.europa.eu/eli/reg/2016/679/2016-05-04, Apr. 2016.

[14] European Commission, "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." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[15] "Grondwet voor het Koninkrijk der Nederlanden." https://wetten.overheid.nl/BWBR0001840/2018-12-21#Hoofdstuk1_Artikel1, Aug. 1815.

[16] College voor de Rechten van de Mens, "Gelijkebehandelingswetgeving." https://mensenrechten.nl/nl/gelijkebehandelingswetgeving.

[17] College voor de Rechten van de Mens, "Geen discriminatie bij Finvita overlijdensrisicoverzekering." https://mensenrechten.nl/nl/nieuws/geen-discriminatie-bij-finvita-overlijdensrisicoverzekering, Jan. 2014.

[18] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," tech. rep., European Commission, Apr. 2019. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

[19] J. van der Burgt, "General principles for the use of Artificial Intelligence in the financial sector," tech. rep., De Nederlandsche Bank, 2019.

[20] Verbond van Verzekaars, "Ethisch Kader Datagedreven Toepassingen," tech. rep., 2020. https://www.verzekeraars.nl/branche/zelfreguleringsoverzicht-digiwijzer/ethisch-kader-datatoepassingen.

[21] Utrecht Data School, "Data Ethics Decision Aid (DEDA)." https://dataschool.nl/en/deda/.

[22] Tilburg University, "Handreiking voor niet discriminerende algoritmes." https://www.tilburguniversity.edu/nl/over/schools/law/departementen/tilt/onderzoek/handreiking.

[23] C. Molnar, *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2019. https://christophm.github.io/interpretable-ml-book/.

[24] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, p. 52, June 2020.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 1527–1535, 2018.

[26] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37–52, Aug. 1987.

[27] L. van der Maaten and G. E. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[28] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Dec. 2019. http://arxiv.org/abs/1910.10045.

[29] M. Berg, van den and O. Kuiper, "A Conceptual Framework for Explainable AI (XAI)," tech. rep., Hogeschool Utrecht, 2020. https://www.hu.nl/onderzoek/projecten/uitlegbare-ai-in-de-financiele-sector.

[30] V. Belle and I. Papantonis, "Principles and Practice of Explainable Machine Learning," Sept. 2020. http://arxiv.org/abs/2009.11698.

[31] S. Lundberg, "SHAP." https://github.com/slundberg/shap.

[32] L. S. Shapley, "A Value for n-Person Games," in *Contributions to the Theory of Games (AM-28), Volume II* (H. W. Kuhn and A. W. Tucker, eds.), pp. 307–318, Princeton University Press, 1953. https://doi.org/10.1515/9781400881970-018.

[33] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2019.

[34] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, pp. 1189–1232, Oct. 2001.

[35] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," 2014.

[36] D. W. Apley and J. Zhu, "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," Aug. 2019. http://arxiv.org/abs/1612.08468.

[37] Achmea, "Achmea jaarresultaten 2020," tech. rep., Mar. 2021. https://nieuws.achmea.nl/jaarresultaten-2020/.

[38] Verbond van Verzekeraars, "Dutch Insurance Industry in Figure," tech. rep., 2016.

[39] L. Chandradeva, T. Amarasinghe, M. Silva, A. Aponso, and N. Krishnarajah, "Monetary transaction fraud detection system based on machine learning strategies," pp. 385–396, Jan. 2020.

[40] Verbond van Verzekeraars, "In 2020 88 miljoen bespaard door aanpak verzekeringsfraude." https://www.verzekeraars.nl/publicaties/actueel/in-2020-88-miljoen-bespaard-door-aanpak-verzekeringsfraude.

[41] S. J. Edgett, "Idea-to-launch (Stage-Gate®) model: An overview," *Stage-Gate International*, pp. 1–5, 2015.

[42] R. Grossman, S. Bailey, A. Ramu, B. Malhi, P. Hallstrom, I. Pulleyn, and X. Qin, "The management and mining of multiple predictive models using the predictive modeling markup language," *Information and Software Technology*, vol. 41, no. 9, pp. 589–595, 1999.

[43] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," Mar. 2017. http://arxiv.org/abs/1702.08608.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[45] S. Verma and J. Rubin, "Fairness definitions explained," in *Proceedings of the International Workshop on Software Fairness*, FairWare '18, (New York, NY, USA), pp. 1–7, Association for Computing Machinery, 2018.

[46] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," 2018.

[47] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A Bias and Fairness Audit Toolkit," Apr. 2019. http://arxiv.org/abs/1811.05577.

[48] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, "Fairness through awareness," *CoRR*, vol. abs/1104.3913, 2011. http://arxiv.org/abs/1104.3913.

[49] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," 2016.

[50] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3315–3323, 2016. http://arxiv.org/abs/1610.02413.

[51] J. Wiśniewski and P. Biecek, "Fairmodels: A Flexible Tool For Bias Detection, Visualization, And Mitigation," Apr. 2021. http://arxiv.org/abs/2104.00507.

[52] P. Gajane and M. Pechenizkiy, "On Formalizing Fairness in Prediction with Machine Learning," May 2018. http://arxiv.org/abs/1710.03184.

[53] B. Ruf and M. Detyniecki, "Active Fairness Instead of Unawareness," Sept. 2020. http://arxiv.org/abs/2009.06251.

[54] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference* (S. van der Walt and Jarrod Millman, eds.), pp. 56–61, 2010.

[55] Plotly Technologies Inc., "Collaborative data science." https://plot.ly.

[56] D. Garreau and U. von Luxburg, "Explaining the Explainer: A First Theoretical Analysis of LIME," Jan. 2020. http://arxiv.org/abs/2001.03447.

[57] Equal Employment Opportunity Commission, "Uniform Guidelines on Employee Selection Procedures." https://www.govinfo.gov/content/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml, Aug. 1978.

[58] M. Chromik and M. Michael, "A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI," in *IUI Workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20)*, (Cagliari, Italy), ACM, 2020.

[59] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems," *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 454–464, Mar. 2020.

[60] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," Aug. 2020. http://arxiv.org/abs/1811.11839.

[61] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for Explainable AI: Challenges and Prospects," Feb. 2019. http://arxiv.org/abs/1812.04608.

[62] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989. http://www.jstor.org/stable/249008.

[63] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410, Oct. 2016.

[64] C.-H. Wu, "An empirical study on the transformation of Likert-scale data to numerical scores," *Applied Mathematical Sciences*, vol. 1, no. 58, pp. 2851–2862, 2007.

Appendix

A.1 Archimate components

The four following pages display more information about the components of the Archimate modelling language:

- Figure A.1 lists the elements of the Archimate business layer.
- Figure A.2 lists the elements of the Archimate application layer.
- Figure A.3 lists the elements of the Archimate technology layer.
- Figure A.4 lists Archimate relationships.

These figures are extracted from the Archimate documentation: `https://pubs.opengroup.org/architecture/archimate3-doc/`.

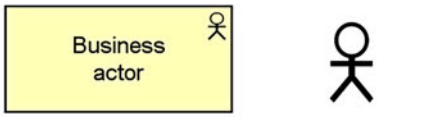
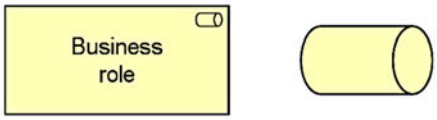
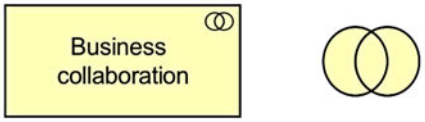
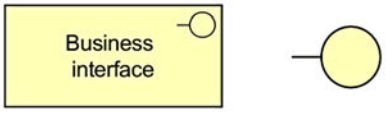
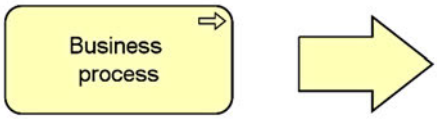
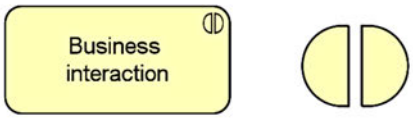
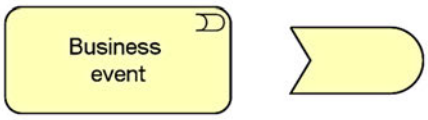
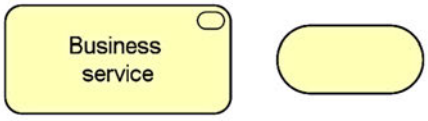
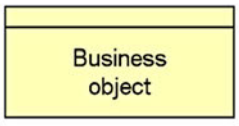
Element	Description	Notation
Business actor	Represents a business entity that is capable of performing behavior.	
Business role	Represents the responsibility for performing specific behavior, to which an actor can be assigned, or the part an actor plays in a particular action or event.	
Business collaboration	Represents an aggregate of two or more business internal active structure elements that work together to perform collective behavior.	
Business interface	Represents a point of access where a business service is made available to the environment.	
Business process	Represents a sequence of business behaviors that achieves a specific result such as a defined set of products or business services.	
Business function	Represents a collection of business behavior based on a chosen set of criteria (typically required business resources and/or competencies), closely aligned to an organization, but not necessarily explicitly governed by the organization.	
Business interaction	Represents a unit of collective business behavior performed by (a collaboration of) two or more business actors, business roles, or business collaborations.	
Business event	Represents an organizational state change.	
Business service	Represents explicitly defined behavior that a business role, business actor, or business collaboration exposes to its environment.	
Business object	Represents a concept used within a particular business domain.	

Figure A.1: Elements of the Archimate business layer.

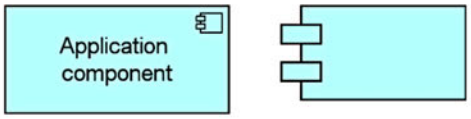
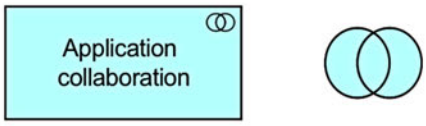
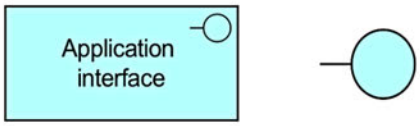
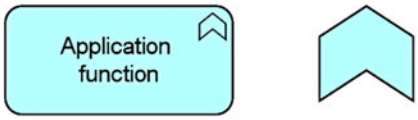
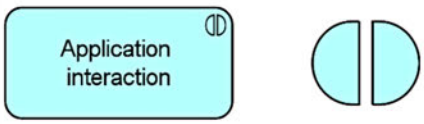
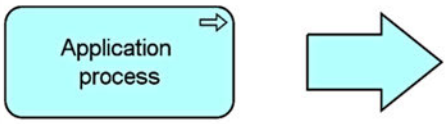
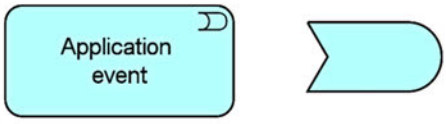
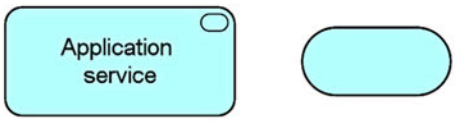
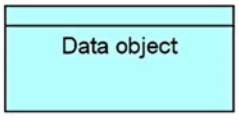
Element	Definition	Notation
Application component	Represents an encapsulation of application functionality aligned to implementation structure, which is modular and replaceable.	
Application collaboration	Represents an aggregate of two or more application internal active structure elements that work together to perform collective application behavior.	
Application interface	Represents a point of access where application services are made available to a user, another application component, or a node.	
Application function	Represents automated behavior that can be performed by an application component.	
Application interaction	Represents a unit of collective application behavior performed by (a collaboration of) two or more application components.	
Application process	Represents a sequence of application behaviors that achieves a specific result.	
Application event	Represents an application state change.	
Application service	Represents an explicitly defined exposed application behavior.	
Data object	Represents data structured for automated processing.	

Figure A.2: Elements of the Archimate application layer.

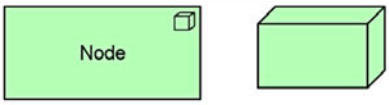
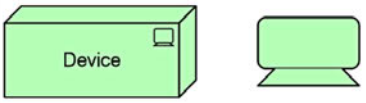
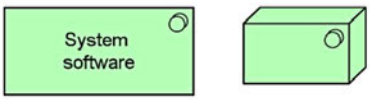
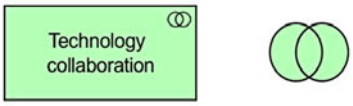
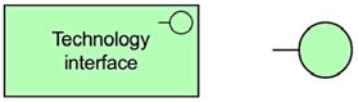
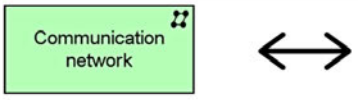
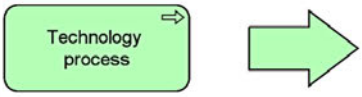
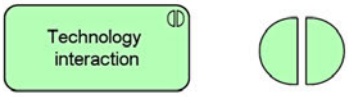
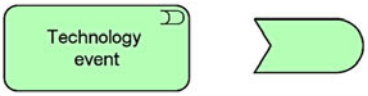
Element	Definition	Notation
Node	Represents a computational or physical resource that hosts, manipulates, or interacts with other computational or physical resources.	
Device	Represents a physical IT resource upon which system software and artifacts may be stored or deployed for execution.	
System software	Represents software that provides or contributes to an environment for storing, executing, and using software or data deployed within it.	
Technology collaboration	Represents an aggregate of two or more technology internal active structure elements that work together to perform collective technology behavior.	
Technology interface	Represents a point of access where technology services offered by a node can be accessed.	
Path	Represents a link between two or more nodes, through which these nodes can exchange data, energy, or material.	
Communication network	Represents a set of structures that connects nodes for transmission, routing, and reception of data.	
Technology function	Represents a collection of technology behavior that can be performed by a node.	
Technology process	Represents a sequence of technology behaviors that achieves a specific result.	
Technology interaction	Represents a unit of collective technology behavior performed by (a collaboration of) two or more nodes.	
Technology event	Represents a technology state change.	
Technology service	Represents an explicitly defined exposed technology behavior.	
Artifact	Represents a piece of data that is used or produced in a software development process, or by deployment and operation of an IT system.	

Figure A.3: Elements of the Archimate technology layer.

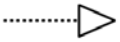
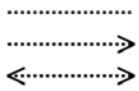
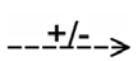
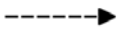
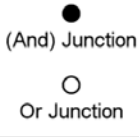
Structural Relationships		**Notation**	**Role Names**
Composition	Represents that an element consists of one or more other concepts.		← composed of → composed in
Aggregation	Represents that an element combines one or more other concepts.		← aggregates → aggregated in
Assignment	Represents the allocation of responsibility, performance of behavior, storage, or execution.		← assigned to → has assigned
Realization	Represents that an entity plays a critical role in the creation, achievement, sustenance, or operation of a more abstract entity.		← realizes → realized by
Dependency Relationships		**Notation**	**Role Names**
Serving	Represents that an element provides its functionality to another element.		← serves → served by
Access	Represents the ability of behavior and active structure elements to observe or act upon passive structure elements.		← accesses → accessed by
Influence	Represents that an element affects the implementation or achievement of some motivation element.		← influences → influenced by
Association	Represents an unspecified relationship, or one that is not represented by another ArchiMate relationship.		associated with ← associated to → associated from
Dynamic Relationships		**Notation**	**Role Names**
Triggering	Represents a temporal or causal relationship between elements.		← triggers → triggered by
Flow	Represents transfer from one element to another.		← flows to → flows from
Other Relationships		**Notation**	**Role Names**
Specialization	Represents that an element is a particular kind of another element.		← specializes → specialized by
Relationship Connectors		**Notation**	**Role Names**
Junction	Used to connect relationships of the same type.		

Figure A.4: Archimate relationships.