Universiteit Utrecht

Department of Information and Computing Science
Business Informatics

Evaluating Effects of XAI in Decision Making in Predictive Process Monitoring Domain: A Human-Grounded Approach

Master Thesis

Soobin Chae

Supervisors:
Suhwan Lee, MSc. PhD Candidate (Daily Supervisor)
Dr.ir. Xixi Lu (First Supervisor)
Dr. Hanna Hauptmann (Second Supervisor)

August 15, 2024

Abstract

Predictive Process Monitoring (PPM) leverages machine learning to forecast future behaviours of ongoing processes to support decision-making across various domains. However, the complexity and opacity of these machine learning models, often termed "black box" models, challenge user interpretability and trust, leading to the development of Explainable AI (XAI) techniques. Despite recent efforts to integrate XAI into the PPM field, it remains unclear how effectively these techniques explain predictions to users and support their decision-making processes, highlighting a gap in understanding the effects of explanations on user behaviour.

This thesis addresses this gap by investigating the impact of explanation styles and perceived AI accuracy on user decision-making within the PPM domain. An empirical user evaluation was conducted to assess the effectiveness of three explanation styles—Feature importance-based, Rule-based, and Counterfactual-based—in influencing task performance, agreement, and decision confidence in decision-making tasks related to loan application outcomes.

The results demonstrate that perceived AI accuracy significantly influences decision-making, with lower perceived accuracy linked to higher task performance across explanation styles. Counterfactual explanations were particularly effective in enhancing task performance and agreement, whereas Feature importance-based explanations resulted in the lowest agreement levels. Conversely, Rule-based explanations led to the highest satisfaction and decision confidence compared to Feature importance-based explanations. These findings show the importance of user evaluations in assessing the effectiveness of XAI explanations. This research contributes to the development of more user-centred and interpretable AI systems by providing insights into how explanation styles and perceived accuracy shape user trust and engagement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Predictive Process Monitoring (PPM) is a group of techniques that uses historical event logs to predict the future behavior of ongoing processes using machine learning [16]. This capability spans various predictive aspects, including remaining case duration, the next process step, and overall outcome prediction [44]. For example, in the loan application process, PPM can estimate the time remaining in the verification stage, predict the likelihood of approval, and anticipate the next process step, such as contacting the applicant. These predictions enable organizations to make informed decisions by incorporating them in a decision support system, thereby increasing efficiency and productivity across diverse domains, including healthcare, finance, and business operations.

Achieving high levels of prediction accuracy requires advanced machine learning techniques, resulting in complex predictive models with superior performance. However, the increased accuracy of these predictions often comes at the cost of model interpretability. Such models are frequently referred to as "black box" models because they are incomprehensible to human intuition. This makes it difficult for users (e.g., decision-makers, process owners) to understand the rationale behind the predictions, making them more hesitant to adopt the system [21]. This issue is particularly critical in the PPM domain, where model outputs can directly influence crucial decisions made by process owners regarding individual cases or even entire workflows [29].

Explainable AI (XAI) has emerged as a promising solution to address this challenge. XAI focuses on revealing the "how" and "why" behind specific predictions, aiming to provide insights while maintaining the strong predictive performance of complex models [3]. Recent advancements in PPM have applied existing XAI techniques such as SHAP [31] and LIME [41] to explain process predictions [19, 18, 24, 34, 56] or evaluate process predictive models [44]. By offering interpretable insights, XAI can enhance user trust

and facilitate the widespread adoption of PPM technologies across various sectors.

Following this realization, there has been a recent shift towards implementing XAI techniques in the PPM field. However, it remains unclear how well these techniques explain predictions to the users for decision-making processes. This is because the XAI evaluation still occurs mainly through predictive performance-based metrics, not accounting for the user side of the "interpretability" of these XAI explanations. Adopting the definition set by Miller [35], interpretability is the degree to which a human can understand the cause of a decision made by the AI method.

However, understanding is rarely an end goal in itself, and it is often more meaningful to measure the effectiveness of explanations in terms of a specific notion of usefulness or explanatory goals [9]. As highlighted in [52, 9, 67, 12], these goals include task performance, agreement, and decision confidence. "Task performance" evaluates how well the explanation helps users make informed and accurate decisions. "Agreement" measures the persuasiveness of the explanation in convincing users to take a particular action. "Decision confidence" assesses the level of confidence users have in their decisions after receiving the explanations. By setting clear explanation goals or effectiveness measures, researchers can effectively evaluate XAI explanations and ensure they are tailored to meet users' needs, therefore enhancing user experience and trust in AI predictions [12].

In the context of XAI evaluation, factors such as explanation styles and perceived AI accuracy can influence the effectiveness of explanations [9, 27]. This is because different explanation styles can support distinct logical reasoning as they determine how information is structured and presented to users [52, 9]. These styles are often defined by the algorithmic approach to generate explanations. On the other hand, perceived AI accuracy impacts user trust and reliance on AI predictions [10]. Research by Kenny et al. (2021) [27] suggests a direct relationship between AI prediction correctness and the information people use in decision-making. When AI predictions are perceived as accurate, users may consider the predictions sufficient to confirm their judgments or even alter their decisions based on the AI's output.

1.1 Research Motivation

Literature reviews indicate a general lack of high-quality user evaluations emphasizing a user-centred approach in XAI [13, 1]. This gap is evident in the PPM domain as well, where, to the best of our knowledge, only two studies have evaluated their XAI frameworks with a focus on user-centered

evaluation [18, 43]. As explanations serve a user's need, it is crucial to evaluate XAI explanations to gain insights into user requirements and the effects of different explanations on decision-making.

Therefore, in this thesis, we conduct an empirical user evaluation to investigate the impact of explanation styles on decision-making while considering the perceived accuracy (high or low) of AI predictions. We aim to enhance our understanding of how users react to AI predictions and explanations, allowing us to categorize decision-making situations based on prediction accuracy. We compare the effectiveness of three explanation styles: Feature importance-based, Rule-based, and Counterfactual-based as they have distinct logic in explaining. Feature importance-based explanations highlight the specific features impacting a prediction. Rule-based explanations translate the model's logic into clear rules, while Counterfactual-based explanations describe how a prediction might change with different inputs, such as "If X had not occurred, Y would not have occurred". To generate these explanation styles, we employ post-hoc XAI techniques: LIME [41] for Feature importance, Anchors [42] for Rule-based explanations, and DiCE [36] for Counterfactual explanations. We will use a mix of objective and subjective evaluation metrics: Task Performance and Agreement for objective metrics, and Decision Confidence for subjective metrics.

The evaluation will take place in a decision-making context where participants determine whether they agree with the AI's prediction of accepting or rejecting a loan application. To the best of our knowledge, no research has been conducted to compare the effectiveness of different explanation styles in decision-making while considering the impact of AI accuracy in the PPM domain. The experiment design for this user evaluation was adapted from previous research by F.M. Cau et al. [9] and Van der Waa et al. [52] with modifications applicable to our evaluation.

1.2 Research Questions

Following the introduction and context of this thesis, the main research question (MRQ) is structured as follows:

> **MRQ**: How do different explanation styles affect decision-making in Predictive Process Monitoring (PPM), taking into account the perceived accuracy of AI predictions?

This main research question leads to the following sub-research questions (RQs). The first research question (RQ1) focuses on understanding the current state-of-the-art in evaluating XAI explanations. This will help identify

gaps in the existing evaluations and determine what types of assessments are necessary. The second and third research questions aim to address these gaps by conducting thorough evaluations of XAI in decision-making contexts. Specifically, RQ2 explores how external conditions, such as perceived accuracy, influence decision-making. RQ3 examines the impact of different explanation styles on decision-making effectiveness.

> **RQ1**: What are the current XAI techniques applied in the PPM domain, and how are they evaluated?

To address this question, we will conduct a literature review of the XAI techniques employed in the PPM domain. Additionally, we will explore existing literature on the evaluation of XAI explanations in PPM. This review will help us identify current gaps in the evaluation of XAI explanations. This research question will be addressed in Chapter 3, Related Work.

> **RQ2**: How does the perceived level of accuracy in AI predictions influence decision-making in terms of effectiveness?

To answer this question, we will compare the effectiveness of decision-making, measured by task performance, agreement, and decision confidence, between high-accuracy and low-accuracy groups, irrespective of explanation styles.

> **RQ3**: How do different explanation styles vary in influencing decision-making in terms of effectiveness?

To answer this question, we will conduct an empirical user evaluation through an experimental study. We will measure effectiveness, specifically task performance, agreement, and decision confidence, across different explanation styles and analyze the differences.

1.3 Contribution

This thesis makes several key contributions to the field of PPM. First, it addresses the current gap of limited research on user evaluation by conducting empirical evaluation. Secondly, the effectiveness of three distinct explanation styles- Feature importance-based, Rule-based, and Counterfactual-based- will be compared within the PPM context. We expect that this comparative analysis will provide insights into differences in explanation styles. Third, the study examines whether the perceived accuracy of AI predictions

affects decision-making processes, providing empirical evidence on how users interact with and trust AI systems based on perceived reliability. Additionally, the thesis explores the reasons behind variations in explanation effectiveness, considering external subjective factors such as users' backgrounds and satisfaction levels. Finally, this research offers valuable insights into user requirements and preferences for XAI explanation, informing the design of more user-centred and interpretable AI systems that align with the needs of decision-makers in process domains.

1.4 Thesis Outline

This chapter has detailed our motivation to assess the impact of different explanation styles within the context of PPM. In Chapter 2 (Background), we outline an overview of PPM and XAI. We also address the evaluation approaches for XAI explanations. In Chapter 3 (Related Work), we discuss what the current XAI techniques are in the PPM domain and their evaluation approaches in the PPM domain. This will answer the first research question (RQ1). In Chapter 4 (Explanation Generation), we introduce how we generate explanations for each explanation style. Here, we describe the dataset and discuss the models & XAI techniques employed for generating explanations. Chapter 5 (Research Method) outlines the research method designed to assess these explanations, which includes experiment structures & surveys. Chapter 6 (Results) is where we report statistical results including descriptive and hypotheses testing. Following this, Chapter 7 (Discussion) focuses on presenting the findings from Chapter 6 data analysis, addressing the research questions (RQ2, RQ3) and the main research question (MRQ). We also discuss limitations, offering potential directions for future research. Finally, Chapter 8 (Conclusion) concludes the thesis, summarizing our conclusions and reflections on the study.

Chapter 2

Background

In this chapter, we present the main background concepts discussed in this paper which are PPM and XAI. First, we address the preliminaries of PPM to understand the prediction tasks. Next, we address the overview of XAI and the specific explanation styles we use in our user-study. Lastly, we introduce the overview of evaluating the explainability of explanations, more specifically for human-grounded evaluations.

2.1 Overview of PPM

PPM relies on the use of historic process data recorded in an *event log*, which is typically stored in the standard XES format. An event log is a collection of traces. Each trace represents a single execution of a business process, such as a loan application or an order fulfilment process within an information system. This execution is further broken down into individual *events*, showing the occurrence of a specific activity within the process. In each event there are three mandatory attributes:

1. **Case identifier**: Uniquely identifies the process instance to which the event belongs.
2. **Activity name**: Specifies the exact activity undertaken within the process.
3. **Timestamp**: Indicates the precise time at which the event occurred,

A trace is a sequence of events generated by executing activities in a process, sorted based on the timestamps of the events. All events within a trace belong to the same case identifier. The event log itself acts as a repository for all traces generated by various process executions within a particular business process. An example of an event log structure is visualized in Figure 2.1.

Case ID	Event ID	Activity	Timestamp	Resource	Requested Amount	Monthly Cost
C120	e_1	Create Application	2018/03/30 10:07:22	John Doe	38000	1281
C120	e_2	Validate Application	2018/03/30 14:12:29	Ben Markus	38000	231
C120	e_3	Decide	2018/04/12 11:15:30	Jill Adams	38000	342
C120	e_4	Close Application	2018/04/23 15:24:03	John Doe	38000	1213

Figure 2.1: Example Eventlogs [14]

Each event in the log is represented by a separate row with its associated attributes. Attributes with consistent values across all rows of a specific process instance are classified as static attributes. Examples of static attributes include case ID and customer name. Conversely, dynamic attributes possess values that change for each event within a process instance. In Figure 2.1, the Case ID serves as a unique identifier for the trace, while the event ID identifies a specific event within that trace. Fields representing dynamic attributes, such as timestamps, resource allocation, and monthly cost, exhibit varying values across each row associated with a process instance. In contrast, other fields, like Requested Amount, represent static attributes and maintain consistent values throughout the trace.

2.1.1 PPM Workflow

PPM approaches leverage past historical complete executions to provide predictions of an ongoing case, which is usually incomplete. They tend to have two phases: a training phase (*aka* offline-phase), in which a predictive model is learned from historical execution traces and a prediction phase (*aka* online phase), in which the predictive model is queried for predicting the future of an ongoing case [16]. Figure 2.2 and Figure 2.3 show the PPM workflow for the offline and online phases respectively.

During the offline phase the system first extracts prefixes (initial sequences of events) from a historical log to train the prediction model. These prefixes might be filtered based on specific criteria, such as retaining only those up to a certain length or certain state, to ensure efficient processing. Next, the identified prefixes are grouped into buckets. This can be done based on factors such as process state or similarity between prefixes. Features are then extracted from these buckets for classification purposes. Finally, each bucket of encoded prefixes is used to train a separate classifier. This creates a collection of specialized classifiers tailored to specific patterns within the data.

The online phase concerns the actual prediction for a running trace, leveraging the system that the elements built offline. Given a new event sequence (running trace) and the set of historical prefix buckets, the system first identifies the appropriate bucket based on the characteristics of the running trace.

Using the identified bucket, the system extracts features from the running trace relevant to that specific context. Finally, the system utilizes the pre-trained classifier associated with the chosen bucket to make a prediction based on the encoded features of the running trace. This allows the model to make predictions for new sequences by reusing the knowledge captured from the historical data.

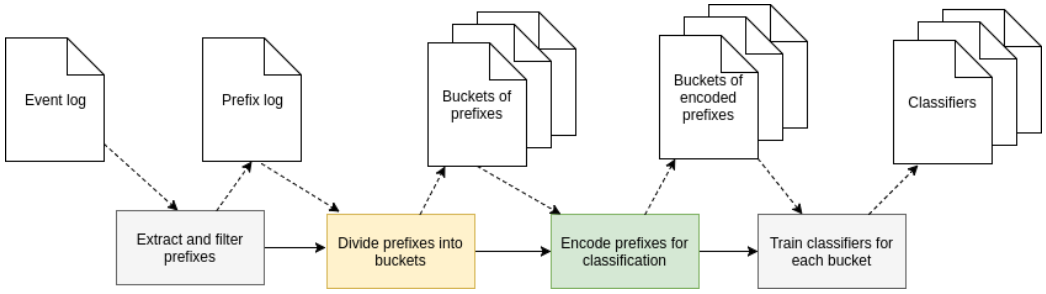


Figure 2.2: PPM workflow (offline phase) [51]

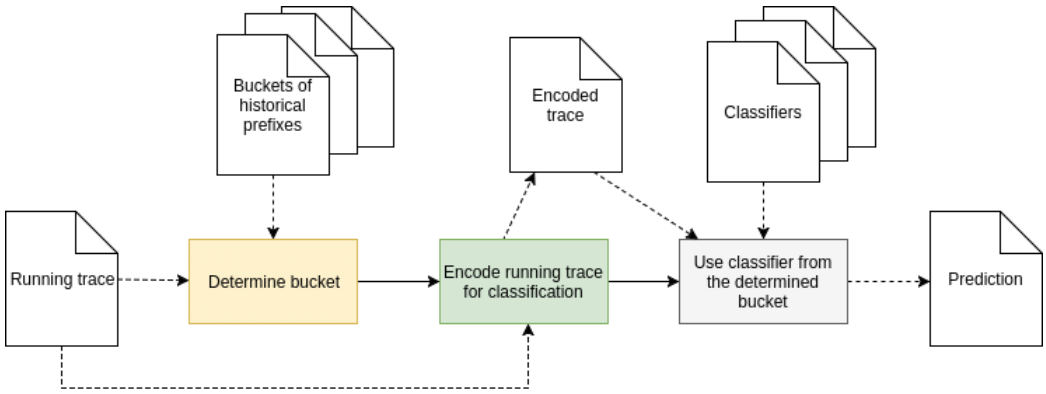


Figure 2.3: PPM workflow (online phase) [51]

Concerning the tasks of prediction, the existing prediction tasks are generally classified into three main categories [32, 51, 16]:

A Next Activity Prediction: Predictions related to the activities that are going to be executed next. For example, predicting the sequence of activities in a process from the current point until its completion.

B Outcome Prediction: Predictions related to categorical or boolean outcomes. Typical examples include classifying each ongoing case of a process, such as predicting whether a loan application will be accepted or not.

C **Remaining Time Prediction**: Predictions related to measures of interest that take numeric or continuous values. Typical examples include predicting the remaining time of an ongoing execution, its duration, or its cost.

In this thesis, we focus on outcome prediction, which will return the predicted class for ongoing cases. The prediction model will be trained to return the predicted outcome by training up to a specific state in the process.

2.2 Explainable AI

2.2.1 Taxonomy of explainability approaches

The field of XAI has shown significant growth in recent years, driven by the widespread adoption of machine learning, especially deep learning. In the literature, XAI techniques are commonly classified based on two primary criteria [58]: Scope and Stage.

1. **Scope**: This refers to the breadth of the explanation provided by a method and is composed of Global and Local explanations.
   (a) *Global* explanations aim to make the entire model's inferential process transparent
   (b) *Local* explanations focus on explaining individual model predictions
2. **Stage**: This refers to the point at which explanations are generated and is divided into *Ante-hoc* methods and *Post-hoc* methods. Additionally, *Post-hoc* methods are further categorized into *model-specific* and *model-agnostic* methods.
   (a) *Ante-hoc* methods integrate explainability into the model structure during training, resulting in inherently interpretable models (e.g., decision trees, linear regression) [23].
   (b) *Post-hoc* methods explain the behaviour of already trained models using external explainers at testing time. Post-hoc methods are further divided into:
      i. *Model-specific* methods are tailored to explain the behaviour of specific types of models.

ii. *Model-agnostic* methods can generate explanations for any prediction model, regardless of its underlying structure. The form of explanation remains the same even when the prediction model changes.

Furthermore, Markus et al. [33] classify XAI techniques based on the type of explanation, identifying three categories: model-based explanations, attribution-based explanations, and example-based explanations.

1. *Model-based explanations*: Utilise a model to explain the original task model. This may involve using the task model itself (e.g., decision tree) or generating more interpretable models to explain the task model (post-hoc explanations).

2. *Attribution-based explanations*: Rank or measure the explanatory power of input features to explain the task model. For example, Feature importance or influence-based explanation approaches fall into this category.

3. *Example-based explanations*: Explain the task model by selecting instances from the dataset or creating new instances. For example, identifying influential instances for the model parameters or output or creating a Counterfactual explanation.

In Table 2.1, XAI techniques are classified based on stage, explanation type, and scope. For this thesis, we selected explanation styles that represent each explanation type. For the model-based explanation type, we adopted a Rule-based style using Anchors [42] algorithm. For the attribution-based explanation type, we employed a Feature importance-based style derived from LIME algorithm [41]. Lastly, for the example-based explanation type, we utilised a Counterfactual-based style derived from DiCE algorithm [36], which generates Counterfactuals. All of these XAI techniques belong to post-hoc explanations and are within the local scope.

Table 2.1: Classification of XAI techniques with examples

Approach	Type of explanation	Scope	Examples of explainable AI techniques	Ref
Post-hoc explanation	Model	Global	Tree extraction	[4]
			Distill-and-compare	[50]
		Local	Anchors	[42]
	Attribution	Global	PDP	[17]
			ALE	[2]
			Feature importance	[15]
		Local	ICE	[20]
			SHAP	[31]
			LIME	[41]
	Example	Global	MMD-critic	[28]
		Local	DiCE	[36]
			Unconditional Counterfactual explanations	[60]

2.2.2 Explanation styles

Rule-based Style

Rule-based explanations resemble decision trees, presenting a set of logical "if-then" statements that mimic the model's decision-making process. These rules explicitly define the conditions under which specific outcomes are predicted. For example, in a loan application scenario, a Rule-based explanation might state: "If the applicant's credit score is above 700 and their debt-to-income ratio is below 35%, then approve the loan". This style provides users with a structured and transparent rationale, facilitating straightforward reasoning and making it suitable for those who prefer clear, rule-based logic.

Feature Importance-based Style

Feature importance-based explanations prioritize identifying the most influential factors contributing to a model's prediction. This style helps users gain valuable insights into the decision-making process by highlighting the key features in the process data that significantly impacted the outcome. For instance, within a loan approval system, a Feature importance-based explanation might reveal: "Credit score is the most significant factor in determining loan approval. Income level is another important determinant, particularly in assessing the applicant's ability to repay the loan". By emphasizing the relative importance of each feature, this style allows users to focus on the most important factors, thereby supporting a more targeted and analytical approach to reasoning.

Counterfactual- based Style

Counterfactual-based explanations utilise "what-if" analyses to illustrate how changes in input variables could alter the model's predictions. This style presents alternative scenarios that demonstrate the sensitivity of the model to specific factors. Continuing with the loan application example, a Counterfactual explanation might involve simulating a scenario where the applicant's credit score is increased by 50 points and analysing if it would lead to loan approval. Counterfactual analyses help users understand how modifications to the input data can potentially change the model's outcome.

2.3 Evaluating the Explainability of Explanations

2.3.1 Taxonomy of Evaluating Explanations

As a response to the wide adoption of XAI approaches, multiple XAI evaluation approaches have been proposed. A widely cited taxonomy of evaluation from Doshi-Velez and Kim [13] proposed a taxonomy for evaluating XAI explanations, highlighting three categories: the first is the *application-grounded* evaluation, which affects domain experts evaluated on actual tasks. The second is the *human-grounded* evaluation, which considers novice users evaluated on simplified tasks. The last is the *functionally grounded* evaluation, which requires no human experiments and uses properties of fidelity explanation quality. In this thesis, we consider *human-grounded* evaluation.

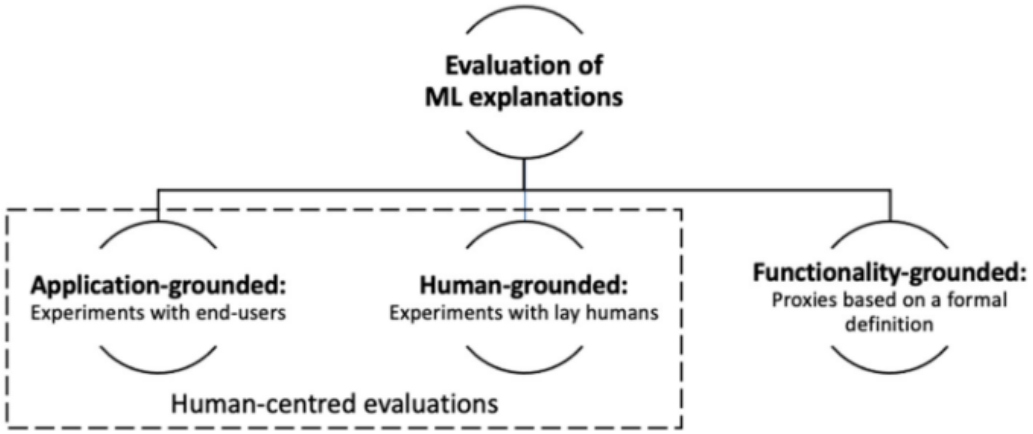


Figure 2.4: Categories of XAI evaluation approaches [67]

2.3.2 Metrics for Human-Grounded Evaluations

Besides evaluation tasks for XAI explanations, the choice of evaluation metrics plays a critical role in the correct evaluation of ML systems. Two types of evaluation metrics can be found in explainable AI research [67], which are subjective and objective metrics. Subjective measures consider the personal experience of the user on tasks and AI explanations, which have been largely embraced as the focal point for the evaluation of explainable systems [41, 65, 68]. For example, Zhou et al. [68, 66] investigated factors such as uncertainty and correlation that affect user confidence in ML-informed decision-making. Meanwhile, objective measures involve evidence measured on task and AI explanations, such as the task completion time and task performance, which refers to choosing correct answers.

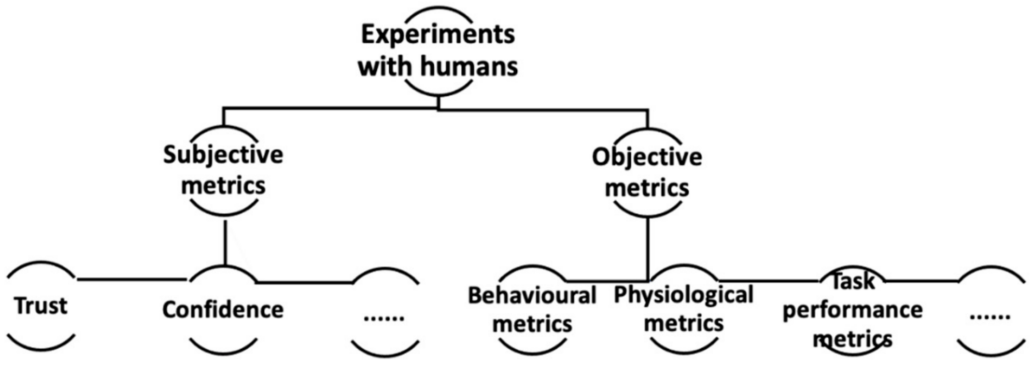


Figure 2.5: Metrics for Human-Ground Experiments [67]

In this thesis, we focus on decision confidence as a subjective measure,

assessing how confident one is with their choice. In Nourani et al. [37], they conducted experiments in an explainable video activity recognition tool in the cooking domain, exposing model weaknesses and strengths. They discovered that participants were more confident in their predictions when explanations were present, and users experienced overconfidence in their mental model either when explanations were present or when the model's strengths were observed earlier. Based on this, we aim to compare the decision confidence of the users before and after they are exposed to explanations. Subsequently, we will compare the decision confidence between the explanation types.

Regarding the objective measure, we focus on how *Task Performance* and *Agreement* between the user and the AI vary considering the different levels of AI accuracy. Task Performance is defined as the ability to make correct decisions. Buçinca et al. [8] explored the impact of explanation styles (inductive vs. deductive) on task performance using proxy tasks. Their findings suggest that users achieve similar performance levels regardless of explanation styles in these simplified scenarios. However, in real-world decision-making tasks, they discovered that users were more likely to choose the correct answer when presented with inductive explanations compared to deductive explanations, especially when the AI recommendation itself was incorrect. In our study, we aim to identify which explanation styles are most effective in leading users towards accurate decision-making in PPM, particularly when the AI model's accuracy varies.

Agreement refers to agreeing with the AI's prediction. Zhang et al. [64] investigated how confidence scores displayed alongside AI predictions influence user agreement. Their findings suggest that users are more likely to adopt the AI's prediction when presented with high confidence scores. Similarly, Van der Waa et al. [52] explored the persuasiveness of different explanation types (Rule-based vs. example-based) in the context of diabetes self-management. Their research demonstrates that explanations, regardless of style, are generally more persuasive than no explanation at all. In our study, we will utilize the agreement metric to assess user behavior in various experimental settings. We aim to understand when users are more likely to follow the AI's prediction in PPM tasks and identify potential persuasive influences depending on different explanation styles.

Chapter 3

Related Work

This chapter addresses the first research question (RQ1), which investigates the current state of the art in XAI evaluation and identifies gaps within the PPM domain. The chapter is structured as follows: first, we explore the XAI techniques currently employed in PPM. Next, we examine the evaluation approaches used for assessing explanations in PPM. Finally, we summarize the identified gaps and outline the contributions of this thesis.

3.1 XAI methods applied in PPM

Recently, explainability approaches have also been applied and investigated in the field of PPM. In Table 3.1, the relevant studies are classified based on the stage at which explanations are generated (Ante-hoc, Post-hoc), along with whether they are model-specific or agnostic for Post-hoc method. We further analyzed whether these studies encompassed explainability evaluation. In the PPM Tasks in this Table, NAP stands for "Next Activity Prediction", RTP for "Remaining Time Prediction", OOP for "Outcome Oriented Prediction", and TCP for "Total Cost Prediction".

Concerning the ante-hoc explainability, multiple approaches have been proposed. Maggi et al. [32] introduced a decision tree-based method, while Breuker et al. [7] proposed a probabilistic-based technique, and Senderovic et al. [46] presented a regression-based approach.

Post-hoc explainability can be further divided into model-specific and model-agnostic methods. Model-specific approaches offer explanations tailored to predictions from specific predictive models, such as attention-based LSTM models by Sindhgatta et al. [47], Layer-wise Relevance Propagation (LRP) into LSTM model by Weinzierl et al. [62], and gated graph neural networks (GGNN) by Harl et al. [22]. Pasquadibiceglie et al. [38] pro-

posed explaining process instance outcomes using *If-Then* rules learned by a neuro-fuzzy network.

Model-agnostic methods have received considerable attention in post-hoc explainability. For instance, Galanti et al. [19] applied SHAP [30] to explain local predictions from LSTM and Catboost models, addressing various prediction tasks like remaining time, activity occurrence, and case total cost. Rizzi et al. [44] utilised LIME [41] and SHAP to identify features leading to incorrect predictions, aiming to enhance predictive model accuracy.

Recent advancements include the development of Counterfactual explanations. For example, Huang et al. [25], Hsieh et al. [24], and Hundogan et al. [26] employ a Counterfactual approach, shedding light on alternative actions to achieve desired outcomes like loan approval. Verenich et al. [57] propose a two-step decomposition-based approach for predicting remaining time, followed by aggregation using flow analysis techniques. Mehdiyev and Fettke [34] introduce explainable DNN-based process outcome predictions through partial dependence plots (PDP) for causal explanations. Rehse et al. [40] employ connection weights to calculate Feature importance for outcome predictions.

Moreover, Elkhawaga et al. [14] and Stevens et al. [49] propose methodologies for quantitatively evaluating XAI methods. Elkhawaga et al. compare SHAP, Permutation Feature Importance [15], and ALE [2] using multiple process datasets, while Stevens et al. introduce metrics encompassing various transparent and non-transparent machine learning and deep learning models, along with post-hoc explainability techniques.

Table 3.1: XAI applied in PPM

Year	Ref.	PPM Task	Stage			iML / XAI Method	EXP Eval.
			Ante-hoc	Post-hoc			
				Model-specific	Model-agnostic		
2014	[32]	NAP	✓			Decision Tree	
2016	[7]	NAP	✓			Probabilistic based	
2017	[46]	RTP	✓			Regression-based	
2020	[47]	NAP		✓		Attention Based LSTM	
2020	[62]	NAP		✓		Layer-wise relevance propagation LSTM	
2021	[38]	OOP		✓		Neuro-Fuzzy Network	
2020	[22]	OOP		✓		Gated graph neural networks	
2019	[48]	OOP, RTP			✓	LIME(XGBoost)	
2023	[19]	OOP, RTP, TCP			✓	SHAP(Catboost)	✓
2020	[18]	OOP, RTP			✓	SHAP(LSTM)	
2020	[44]	OOP			✓	LIME, SHAP	
2020	[54]	OOP			✓	LIME, SHAP	✓
2021	[55]	OOP			✓	LIME, SHAP	✓
2022	[25]	OOP			✓	Counterfactual	✓
2021	[24]	NAP			✓	Counterfactual	✓
2023	[26]	OOP, NAP			✓	Counterfactual	
2019	[57]	RTP			✓	Decomposition-based	
2020	[34]	OOP			✓	PDP (DNN based)	
2018	[40]	OOP			✓	Connection-weight(DNN based)	
2024	[14]	OOP			✓	SHAP, Permutation Importance, ALE	✓
2023	[49]	OOP			✓	Attention-based (LSTM, CNN), iML(GLRM, LLM, LR), SHAP(XGB, RF)	✓

3.2 Evaluation of XAI methods in PPM

Unfortunately, only a few studies in the PPM domain are concerned with evaluating the explanations generated with XAI techniques. However, as the importance of explaining PPM results gains recognition, the evaluation of XAI explanations is expected to gain more interest. As explained in section 3.2, explainability evaluation can be categorized into three approaches: Application-, Human-, and Function-grounded evaluations. In Table 3.2, relevant studies are classified based on their evaluation approaches and whether they compare XAI methods.

Function-grounded evaluation approaches assess explanations based on their inherent characteristics. Velmurugan et al. [54, 55] focused on assessing stability and fidelity. According to Visani et al. [59], stability refers to the consistency of explanations generated for the same data sample under identical conditions. They proposed metrics to evaluate the stability of the top-K feature subset and their respective weights across multiple explanations for certain process instances [55]. Fidelity, on the other hand,

pertains to the ability of XAI methods to accurately mimic the behavior of the explained ML model in the vicinity of the explained process instance. Velmurugan et al. [54] introduced an approach to evaluate local XAI methods for their internal fidelity, which compares the decision-making process of the explainer proxy model with the explained complex black-box model. Huang et al. [25] also utilize a fidelity metric for Counterfactual explanations to evaluate the faithfulness of explanations. They additionally assess explanation quality by examining whether the most important attributes indicated by the explanations correspond to the domain knowledge. Among the Function-grounded evaluation studies in the PPM domain, Stevens et al. [49] and Elkhawaga et al. [14] proposed evaluation approaches for various XAI methods applied to PPM results. Stevens et al. introduced four out-of-the-box metrics from relevant XAI evaluation research, applied to different attributes in process mining data. They argue that these attributes should be studied separately due to differing characteristics and effects on predictions and explanations. Elkhawaga et al. proposed an approach for evaluating global model-agnostic XAI methods using feature attributions to explain ML model reasoning. They aimed to assess the consistency of explanations with basic concepts extracted from underlying data.

Only two studies [43, 19] focused on human evaluations, specifically application-grounded evaluation. Rizzi et al. [43] are among the first to investigate whether users understand the explanation plots. Instead of applying actual XAI methods, they generate three levels of different plots in event, trace, and event log levels. While the study involved participants from both the PPM and ML fields, comprehension and usage levels of explanations varied based on domain knowledge and experience. However, the study relied on qualitative evaluation with a limited number of participants without employing a consolidated user-interface evaluation methodology. Galanti et al. [19] proposed a framework to evaluate the understanding and comfort level of process analysts with results from an explainable predictive monitoring framework. Their evaluation focused on accuracy in task execution, perceived task difficulty, usability, and user experience dimensions. Regarding the comparison of XAI methods, no user study compares the effectiveness of different XAI methods in qualitative user evaluation in the PPM domain.

Table 3.2: Explainability Evaluation Approaches in PPM

Year	Ref.	Evaluation Metric	H. Evaluations		Function	XAI Methods
			Application	Human		Comparison
2020	[54]	Fidelity			✓	✓
2021	[55]	Stability			✓	✓
2022	[25]	Fidelity			✓	
		Domain Knowledge				
		Parsimony				
2023	[49]	Functional Complexity			✓	✓
		Importance Ranking Correlation				
		Level of Disagreement				
2024	[14]	Consistence			✓	✓
2022	[43]	Understandability	✓			
		Accuracy				
2023	[19]	Perceived Task Difficulty	✓			
		Usability				

3.3 Summary of Research Gaps and Thesis Contribution

In this chapter, we answered RQ2 by conducting current literature reviews. First, reviewed studies that apply XAI techniques in the PPM domain. Additionally, we explored existing research on evaluating XAI explanations, including the evaluation approaches and metrics used. Through this analysis, we identified two key research gaps that motivate our work.

First, there are limited comparative user studies in the PPM domain. Existing user studies in the PPM domain primarily focus on individual XAI frameworks or qualitative evaluations with limited generalizability. Second, there is a lack of research on the decision-making effects of different explanation styles. Existing research primarily focuses on the understandability and usability of explanations. There is no systematic investigation into how different explanation styles influence a user's decision process.

Given these research gaps, this thesis aims to address the aforementioned research gaps by conducting empirical user evaluation to compare the effectiveness of different explanation styles within the PPM domain. In addition, we will incorporate each representing a type of explanation: Model-based (Rule-based style), Attribution-based (Feature importance style), and Example-based (Counterfactual style).

Chapter 4

Explanation Generation

This chapter addresses how each XAI explanation is generated for the experiment. We will discuss the dataset, the chosen method for data preprocessing and the black-box model, and the specific approaches for generating each XAI explanation.

4.1 Data Set

The dataset used for this experiment uses the BPIC 2017 event log data [53]. This dataset concerns the loan application process of a Dutch financial institution and is based on real-life event logs. Figure 4.1 shows the code snippet of the BPIC 2017 event log before preprocessing. We selected this dataset for several reasons. First, it is a popular dataset widely used in the PPM domain (i.e. [25, 18, 51, 56]), which helps ensure the generalizability of our findings to other PPM applications. Second, the loan application process depicted in the data offers a familiar and relatable scenario for participants, even for those without prior knowledge of PPM. This is crucial for our human-centred evaluation, where the explanations are designed to be understandable by non-expert users.

This data includes information on 1) States of the loan application, 2) States of offers associated with the application, and 3) States of work items related to the application. The overall application process can be decomposed into three stages, receiving applications, negotiating offers, and validating documents respectively [25, 5]. At the end of the process, an application can be successful (*A_Pending*) or not (*A_Denied*). Additionally, an application set *A_Canceled* is set if the customer neither replied to the call nor sent the missing documents as requested.

	caseid	Activity	ts	LoanGoal	RequestedAmount
0	Application_652823628	A_Create Application	2016-01-01 09:51:15.304000+00:00	Existing loan takeover	20000.0
1	Application_652823628	A_Submitted	2016-01-01 09:51:15.352000+00:00	Existing loan takeover	20000.0
2	Application_652823628	W_Handle leads	2016-01-01 09:51:15.774000+00:00	Existing loan takeover	20000.0
3	Application_652823628	W_Handle leads	2016-01-01 09:52:36.392000+00:00	Existing loan takeover	20000.0
4	Application_652823628	W_Complete application	2016-01-01 09:52:36.403000+00:00	Existing loan takeover	20000.0
...	...	...	...	...	...
1202262	Application_1350494635	W_Call after offers	2017-01-06 06:33:02.212000+00:00	Home improvement	20000.0
1202263	Application_1350494635	W_Call after offers	2017-01-06 06:33:02.221000+00:00	Home improvement	20000.0
1202264	Application_1350494635	A_Cancelled	2017-01-16 09:51:21.114000+00:00	Home improvement	20000.0
1202265	Application_1350494635	O_Cancelled	2017-01-16 09:51:21.139000+00:00	Home improvement	20000.0
1202266	Application_1350494635	W_Call after offers	2017-01-16 09:51:21.146000+00:00	Home improvement	20000.0

Figure 4.1: Snippet of BPIC2017 eventlog

Figure 4.2 shows an example of a filtered process map that shows cases that end with *A_Pending*, *A_Denied*, *A_Canceled*.

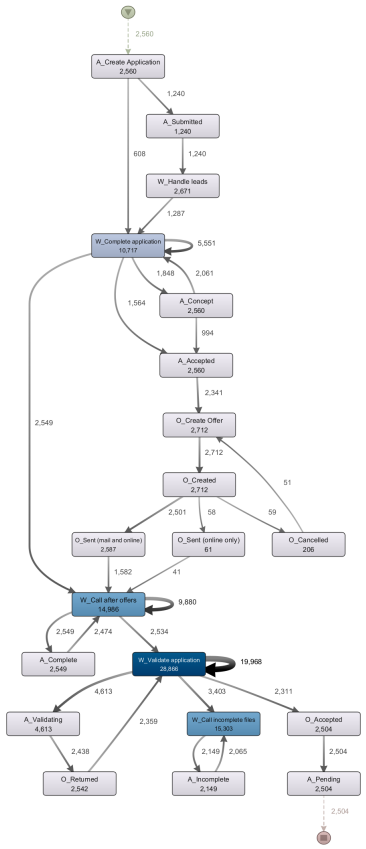


Figure 4.2: BPIC 2017 Process Map

4.2 Training Black-Box Model

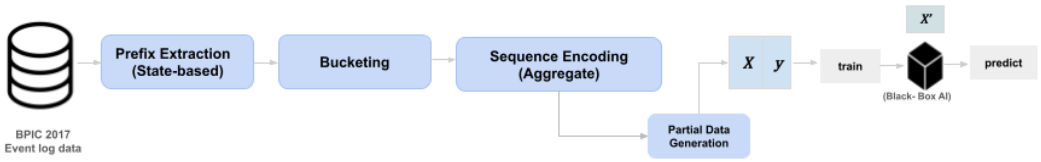


Figure 4.3: PPM workflow

4.2.1 Data preprocessing

Label Assignment for Cases

The first step for data preprocessing is assigning outcome labels to cases in event log data. Events are grouped by case ID to gather all events associated with each case. These events are then sorted chronologically to ensure that activities are in the correct order. The sequence of activities is scanned to identify key events that determine the case's outcome. Then labels are assigned based on predefined significant activities, ensuring that each case is categorized for subsequent analysis.

For this dataset, the process began by grouping all events by case ID to gather all events associated with each case. Within each group, events were sorted by their timestamp (ts) to ensure that the activities were in chronological order. Next, the sequence of activities for each case was scanned to identify key events that determine the case outcome. Specifically, the activities "*A_Pending*", "*A_Denied*", and "*A_Cancelled*" were used to assign labels. They were labelled as "Accept Application", "Reject Application", and "Cancel Application" respectively.

State-based Bucketing

When training the model with event log data, the initial sequences of events (prefixes) should be extracted from the historical event log. These prefixes can be filtered based on specific criteria, for example, by retaining only those up to a certain length (prefix-length-based bucketing) or certain state (state-based bucketing) to ensure efficient processing. Next, the identified prefixes are grouped into buckets.

For this dataset, state-based prefix extraction was used to create buckets, utilising *O_Returned* as the key state. We followed the approach of Wickramanayake et al. [63], which also used the BPIC 2017 dataset for the next

activity prediction. In addition, BPIC 2017 winner [39] analyzed *O_Returned* as the decision point.

Each case was sliced to include only the events up to and including the *O_Returned* activity. *O_Returned* was chosen because it signifies the applicant submitting documents for a selected credit offer. After this state, many cases proceed to be accepted (*A_Pending*), rejected (*A_Denied*), canceled (*A_Canceled*), or even validated again (*W_validate Application*) if there are missing documents. In essence, *O_Returned* was selected as the state for filtering because it serves as a pivotal decision point for predicting the outcome.

Aggregate Encoding

The prefixes within the same bucket must be represented as fixed-length feature vectors to be effectively used for classification. This transformation is achieved using abstraction techniques such as considering the last n events (last n-state encoding), calculating activity frequencies (aggregate encoding), or maintaining the order of events (index-based encoding).

For this dataset, aggregate encoding was employed to capture both the frequency of activities and the cumulative duration up to the *O_Returned* state, providing a comprehensive representation of each case. In the implementation, the cumulative duration for each case up to *O_Returned* was computed, and the frequency of each activity within the filtered event log for each case was calculated. Relevant case attributes (e.g., "LoanGoal", "RequestedAmount") were also included.

By aggregating all events and attributes up to each prefix, this encoding approach ensures that the model has a complete view of the past events that may influence the future state of each case. This comprehensive representation facilitates accurate and meaningful outcome predictions.

Partial Data Generation

After aggregate encoding, new attributes were generated based on the outcome label to enrich the dataset with additional features relevant to loan application data. The attributes were added to provide participants with more context beyond event logs, simulating real-world scenarios.

To introduce variability, half of the dataset was generated randomly, including attributes such as income, employment status, credit score, and age, independently of the outcome labels. This approach ensured that the generated rules did not disproportionately influence the features derived from the

event log data. The following rules were used to generate these additional features:

- **Income**: Generated using a normal distribution.
  - Accept: $40,000 (SD $7,000)
  - Reject: $25,000 (SD $5,000)
  - Cancel: $35,000 (SD $6,000)
- **Employment Status**: Probabilities for "Permanent" vs "Fixed-Term".
  - Accept: 70% Permanent, 30% Fixed-Term
  - Reject: 40% Permanent, 60% Fixed-Term
  - Cancel: 60% Permanent, 40% Fixed-Term
- **Credit Score**: Specific ranges, rounded to nearest 50.
  - Accept: 710-850
  - Reject: 500-750
  - Cancel: 550-800
- **Age**: Weighted sampling with higher density around 31-40 years.

Figure 4.4 shows the code snippet of the dataset after preprocessing including applying the above rule to half of the dataset.

O_Create Offer	O_Created	O_Sent (mail and online)	W_Call after offers	...	W_Shortened completion	caseid	duration	LoanGoal	RequestedAmount	label	Income	EmploymentStatus	CreditScore	Age
1	1	1	6	...	0	Application_703707186	670292.001	Home improvement	28000.0	Accept Application	21198	Fixed-Term	504	35
1	1	1	6	...	0	Application_1223117710	1958445.364	Car	9000.0	Reject Application	27062	Permanent	839	59
3	3	3	8	...	0	Application_52093579	1192192.038	Existing loan takeover	20000.0	Accept Application	44854	Fixed-Term	734	49
2	2	2	6	...	0	Application_738276182	933682.909	Home improvement	30000.0	Accept Application	28882	Permanent	587	50
1	1	1	6	...	0	Application_304416200	1092853.752	Not speficied	6900.0	Accept Application	27684	Permanent	525	51
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	1	1	8	...	0	Application_2053301194	604516.920	Existing loan takeover	10000.0	Accept Application	29927	Permanent	661	46
1	1	1	6	...	0	Application_2115042059	732008.785	Home improvement	5000.0	Reject Application	33539	Permanent	725	54
1	1	1	6	...	0	Application_1494188195	764601.330	Existing loan takeover	31000.0	Accept Application	35081	Permanent	510	50
1	1	1	4	...	0	Application_1081666766	562778.247	Car	38000.0	Accept Application	24252	Fixed-Term	556	53
1	1	1	6	...	0	Application_538816828	669659.407	Car	13500.0	Accept Application	42215	Permanent	833	24

Figure 4.4: Snippet of preprocessed data

Black-Box Model

We selected Random Forest [6] as the black-box model to generate explainable artificial intelligence (XAI) explanations. This model predicts the outcomes (Accept, Reject, or Cancel Application) based on activity frequencies and cumulative time duration. The event log was split into a training set (80% of the traces) and a testing set (20% of the traces). Given that the primary goal of this experiment was to produce XAI explanations for comparative analysis, we did not undertake exhaustive hyperparameter tuning or structural optimization of the model. Due to the high imbalance in the dataset (Accept: 55%, Cancel: 33%, Reject: 12%), we applied resampling techniques to address this issue. Following resampling, the model achieved an overall accuracy of around 0.85.

In addition, as we aimed to measure the influence of perceived AI accuracy on decision-making effectiveness, we considered using models with different levels of actual accuracy. However, testing with high and low-accuracy models resulted in significantly varied XAI outcomes, which could potentially bias the survey results. With differing content, we concluded that isolating the impact of different explanation styles would be challenging. Therefore, we trained a single model as previously mentioned, and used the model to generate XAI explanations. The only variation for participants was the presented accuracy level (high-accuracy group: 96%, low-accuracy group: 63%).

4.2.2 XAI Instance Generation

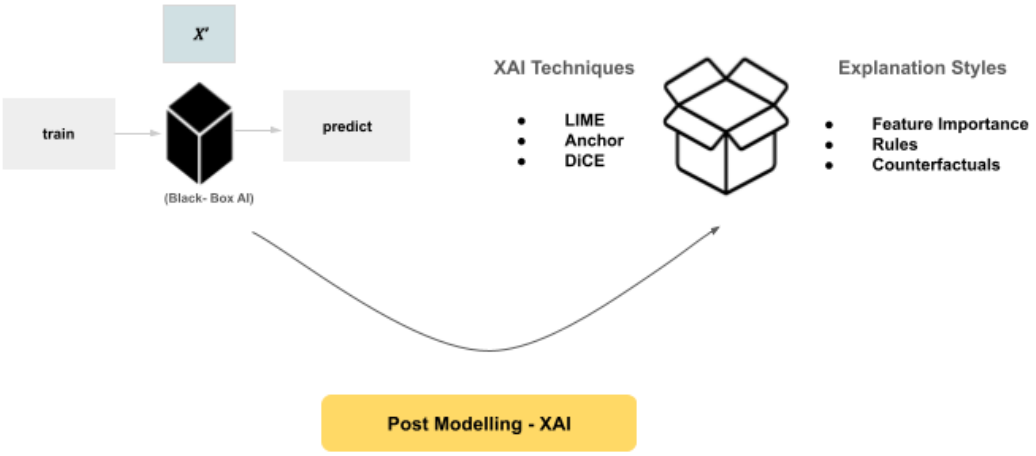


Figure 4.5: Post Modelling XAI

To generate local XAI explanations for individual predictions, we utilized local model-agnostic XAI techniques. Specifically, we employed LIME [41] for -based explanations, Anchor [42] for Rule-based explanations, and DiCE [36] for Counterfactual-based explanations. After generating the explanations, we refined them into a more human-friendly format while retaining their original content.

We initially generated six cases for each prediction category (accept, reject, cancel), including correctly classified and misclassified instances. However, to reduce the length of the survey, we later adjusted this to include only four cases by focusing on accept and reject application cases. In total, 12 instances were used for the survey, with four instances for each explanation style.

In the following sections regarding each XAI explanation, we include the "Cancel Application" cases, although these were not ultimately used in the survey due to the reduction in the number of instances.

Feature Importance-based: LIME

For Feature importance-based explanations, we used LIME [41], which identifies which variables influenced the prediction and highlights the important features of a particular prediction. LIME works by testing the effect on the predictions when variations of the data are input into the machine learning model. During the generation of instances, some parts were cut off due to long texts. These were revised using Illustrator to ensure clarity.

The figures below illustrate a prediction classified as "Reject Application". Figure 4.6 shows the original version and Figure 4.7 side displays the revised version. The right bars represent factors that support the AI's decision to reject the application, while the left bars indicate factors that oppose the decision, suggesting either "Cancel Application" or "Accept Application".

Figure 4.6: Example of Before Edit - LIME

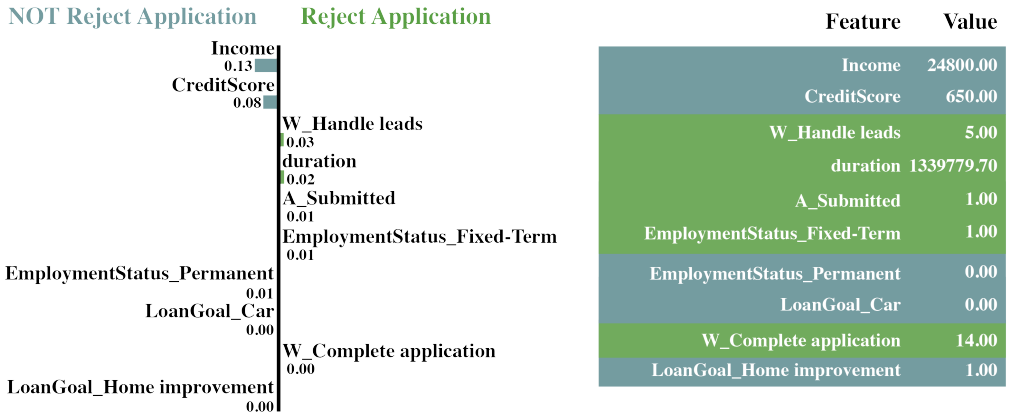


Figure 4.7: Example of After Edit - LIME

Rule-based: Anchor

We used the Anchor [42] technique for Rule-based explanations. Anchor generates rule explanations in a sequential manner, connecting conditions with "AND". For example, for a prediction to "Accept Application", the rule was:

"Income >44300.00" AND "W_Complete application <= 3.00" AND "Age >47" AND "LoanGoal_Home Improvement = 1" AND "duration <= 824966.04" AND "EmploymentStatus_Fixed-Term =0"

Since this format was difficult to read at a glance, we revised it into decision-tree-like figures for better clarity. Figure 4.8 below shows the revised version of the instance.

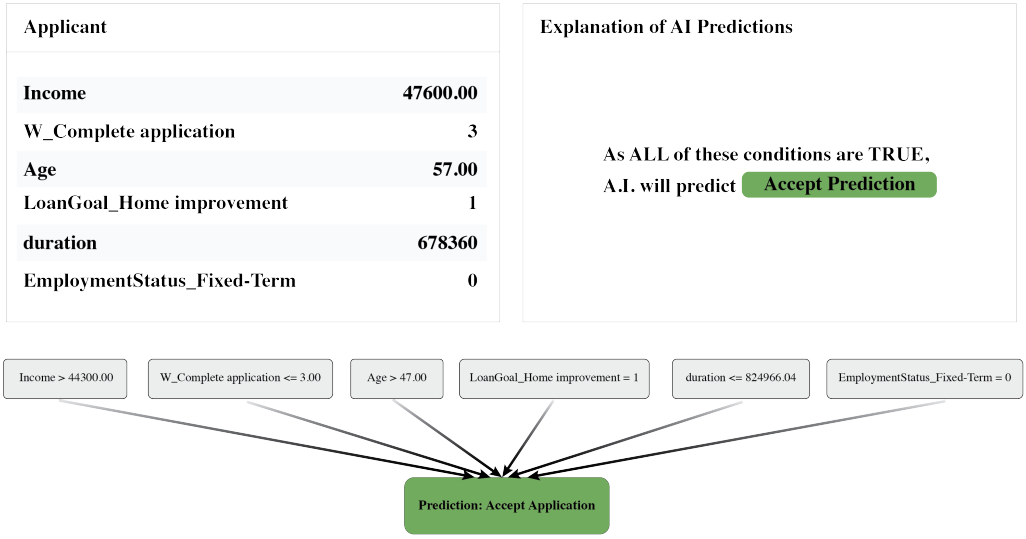


Figure 4.8: Example of After Edit - Anchor

Counterfactual-based: DiCE

We used DiCE [36] for Counterfactual-based explanations. DiCE allows customization of Counterfactuals by choosing the desired class, the number of Counterfactuals, and the method for generating them (general or random). The random method can be particularly useful when dealing with sparsity issues due to a large number of Counterfactuals or features. For instance, if the original prediction was "Reject Application", the desired class could be set to "Accept Application" or "Cancel Application".

In this experiment, we used the random method to generate two Counterfactuals for each desired outcome. Given that there were 38 features, including encoded categorical features, we selected nine categorical features for the instances shown to participants. This number was chosen to be consistent with the number of features shown in -based and Rule-based explanations. The criteria for selecting these features were based on their appearance in other explanations to ensure relevance.

Figure 4.9 shows the code snippet that DiCE generated for "Reject Application" prediction of the instances generated by DiCE and shows Counterfactuals or "Accept Application".

Query instance (original outcome : Reject Application)

W_Call mplete files	W_Shortened completion	W_Assess potential fraud	duration	RequestedAmount	Income	CreditScore	Age	EmploymentStatus_Fixed-Term	EmploymentStatus_Permanent	Lo
0	0	0	505875.375	15000.0	24800	650	23	0	1	

Diverse Counterfactual set (new outcome: Accept Application)

ıcome	CreditScore	Age	EmploymentStatus_Fixed-Term	EmploymentStatus_Permanent	LoanGoal_Boat	LoanGoal_Business goal	LoanGoal_Car	LoanGoal_Caravan / Camper
-	-	-	-	-	-	-	-	1
-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-
-	-	39.0	-	-	-	-	-	-
-	849.0	-	-	-	-	-	-	-

Figure 4.9: Example of Before Edit - DiCE

Figure 4.10 shows the explanation for "Reject Application" prediction of the instances generated by DiCE and shows Counterfactuals or "Accept Application" and "Cancel Application".

	W_Complete application	W_Call after offers	W_Validate application	W_Call incomplete files	Requested Amount	Income	Credit Score	Age	Loan Goal	Outcome
Original Input	14	6	2	0	15000	24800	650	23	Home improvement	➡ Reject
Different Input 1	-	-	1	-	-	-	849	-	Tax Payments	➡ Accept
Different Input 2	22	-	-	-	-	-	-	-	-	➡ Cancel

Figure 4.10: Example of After Edit - DiCE

Note that "Cancel Application" was later deleted due to the long length of the survey.

Chapter 5

Research Method

In this chapter, we outline the research methods for conducting the empirical user evaluation within the context of outcome prediction in a loan application process. The chapter is organized as follows: we introduce the evaluation objectives, describe the experimental design, define the hypotheses and analytical approach, and detail the overall experimental procedure.

5.1 Evaluation Objectives

This thesis aims to evaluate the impact of explanation styles on decision-making within the context PPM, considering variations in perceived AI accuracy. The experimental design for the user evaluation is structured around the following objectives, which are aligned with RQ2 and RQ3:

RQ2: How does the perceived level of accuracy in AI predictions influence decision-making in terms of effectiveness?

Obj. 1: Investigating the Influence of Perceived AI Accuracy Levels on User Decisions

RQ3: How do different explanation styles vary in influencing decision-making in terms of effectiveness?

Obj. 2: Assessing the Impact of Explanation Styles on Decision-Making

Obj. 3: Evaluating Changes between Pre- and Post-Explanation

Obj. 4: Investigating the Reasons Behind Effectiveness Differences

The first objective examines how perceived AI accuracy, categorized as high or low, influences decision-making in terms of task performance, agreement, and decision confidence. Understanding these influences will help categorize decision-making situations based on prediction accuracy. The second objective focuses on evaluating the effects of various explanation styles— Feature importance-based, Rule-based, and Counterfactual-based—on the aforementioned effectiveness metrics. The aim is to determine which styles best support informed decision-making, effectively persuade users to take specific actions and enhance their confidence in their decisions. The third objective involves assessing how decision-making metrics change after explanations are provided. The aim is to identify which explanation style influences the most in affecting task performance, agreement, and decision confidence. The final objective is to explore the reasons for differences in explanation effectiveness, focusing on external subjective factors such as users' backgrounds and satisfaction levels. This will help us to gain insights into the factors influencing the impact of explanations.

5.2 Experimental Design

The experiment is structured based on the evaluation objectives outlined in the previous section. Figure 5.1 illustrates the two-phase decision-making process, highlighting how accuracy and explanation styles influence the participants' decisions. Each "scenario" presents participants with an applicant profile and process map related to a loan application. These scenarios are described in more detail in Section 5.3.2.

Following the first objective, participants are initially divided into two distinct accuracy groups: High Accuracy and Low Accuracy, based on the displayed AI prediction accuracy for a set of control tasks. Participants in the high-accuracy group are presented with tasks where the AI model demonstrates high accuracy, such as 96%, while those in the low-accuracy group are shown tasks with a lower accuracy, around 63%. This setup allows us to isolate and assess the influence of perceived AI accuracy on decision-making. In Figure 5.1, we are comparing the effectiveness of "initial decisions".

Within each accuracy group, participants are further assigned to explanation style subgroups (Feature importance-based, Rule-based, and Counterfactual-based). This division helps evaluate the effect of different explanation styles on decision-making, addressing the second objective. Figure 5.1 shows how the explanation groups are divided. We compare the effectiveness from the post-explanation decision point. In this Figure, Feature- importance is abbreviated as "FI", and Counterfactual as "CF".

To assess the impact of the explanation provision, participants make decisions at two different points. In their initial decision-making phase, no explanations are provided. In the second phase (Post decision), participants are provided with one of the three explanation styles. This allows the assessment of within-group changes and between-group comparisons, highlighting the impact of explanations on user decisions.

To achieve the last objective, we also incorporate post-experiment questionnaires that assess subjective opinions about the decision-making tasks. Participants were asked to rank the information they relied on the most. We also evaluate their satisfaction and the perceived difficulty of the explanations they received. Additionally, we collect data on participants' educational backgrounds and experience with process mining/XAI to determine if these factors influence the effectiveness of the explanations. An open-ended question is also included to capture participants' suggestions for improving the explanations. This comprehensive feedback collection provides insights into the factors affecting explanation effectiveness and helps clarify the differences in the impact of explanations across various groups. This section corresponds to the post-questionnaire in Figure 5.1.

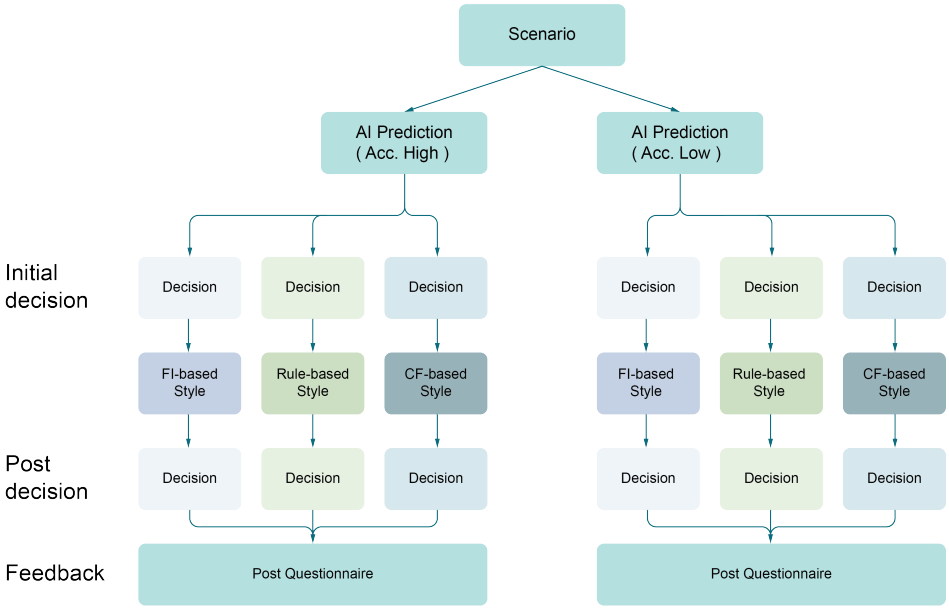


Figure 5.1: Experiment Setting

Independent Variables

1. **Perceived AI Accuracy**: Low or High (Participants will be divided into two groups based on AI prediction accuracy)
2. **Explanation styles**: Three styles (Participants will receive one of three styles of explanations: Rule-based, Counterfactual-based, or Feature importance-based. A control condition is also included where no explanation is provided.)

Dependent Variables

The dependent variables are the metrics to assess the impact of the styles of explanation and the accuracy level of AI on decision-making.

1. **Task performance**: The number of correct decisions the participants chose.
2. **Agreement**: This number of participants' decisions align with AI predictions.
3. **Decision Confidence**: 5-point Likert Scale ranging from 1 (completely unsure) to 5 (extremely confident), answering their confidence in their decision-making

5.3 Survey Procedure

Before distributing the survey, we conducted pilot testing with five individuals, representing a mix of those with and without process mining knowledge and those with and without a Business IT background. This pilot test aimed to estimate the average time required to complete the task and assess the task difficulty. Participants without these backgrounds took approximately 20 minutes to complete the survey. Recognizing the varying levels of knowledge and anticipating that a long survey duration might reduce the quality of responses, we decided to reduce the number of tasks. Initially, the survey included five scenario tasks that involved cancel predictions. Based on the pilot test feedback, we reduced this to four scenarios, focusing only on accept and reject predictions, each with correct and incorrect predictions. After these adjustments, the median completion time for the pilot survey was approximately 10 minutes.

The survey distribution was divided into multiple phases. The first batch consisted of Business Informatics master's students at Utrecht University enrolled in a Process Mining lecture. The second batch included other master's

students from backgrounds related to Business and IT, including Business Informatics. The third batch consisted of PhD researchers in the process mining field. The final batch was recruited from the Prolific platform[1] to fill in the required number of participants. In total, we gathered 222 participants. After excluding responses based on the quality of answers and completion time (under 5 minutes), we retained data from 181 participants. The median time to complete the survey for these qualified participants was 566.5 seconds, approximately 9-10 minutes.

In the actual survey, participants were randomly assigned to one of six groups using a randomizer implemented in the Qualtrics survey tool[2]. Figure 5.2 shows the survey structure made in Qualtrics and Table 5.1 shows the number of participants per accuracy and explanation styles groups. All participants, regardless of their group assignment, followed the same procedure throughout the experiment.

It is followed as:

1. **Introduction**: Participants first receive an overview of the study's objectives and context, specifically focusing on the loan application process. This initial step sets the stage for the experiment and ensures participants have a clear understanding of the context.

2. **Informed Consent**: Before participation, individuals are presented with a consent form detailing the study's scope, their rights as participants, and confidentiality measures.

3. **Experiment**: This involves the main decision-making tasks that participants complete. Each participant goes through 4 decision-making tasks. A detailed example of these tasks is discussed in Section 5.3.2.

4. **Post Questionnaire**: Concluding the experiment, this questionnaire gathers feedback on participants' satisfaction with the explanations (rated on a 1-5 scale) and the difficulty of the explanation interface. Open-ended answers are also collected for qualitative insights into the challenges participants faced in understanding the explanations and their suggestions for improving the interface.

5. **Demographic Questionnaire**: Participants complete a questionnaire to collect demographic information, including their highest education level, whether they have a STEM background, and their years of experience with XAI and Process Mining. This data helps in analysing

[1]Copyright © [2024] Prolific. https://www.prolific.com
[2]Copyright © [2024] Qualtrics. https://www.qualtrics.com

results across diverse groups and ensures a comprehensive understanding of the experiment's impact.

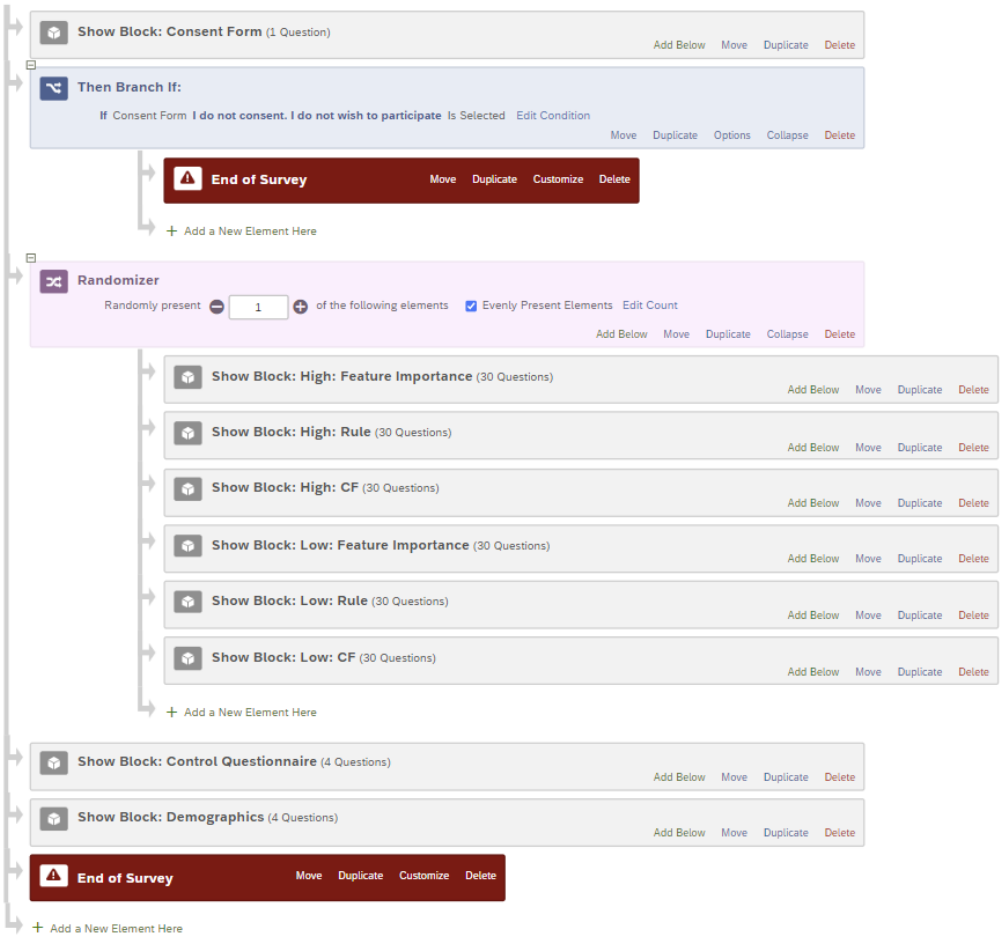


Figure 5.2: Experiment Setting

Table 5.1: Number of Participants per Explanation Groups

Accuracy	**Explanation Styles**	***N***
High	Feature importance	30
	Rule	30
	Counterfactual	27
Low	Feature importance	32
	Rule	31
	Counterfactual	31

5.3.1 Participants

Table 5.2 describes the demographics of the participants, including their education levels, fields of study, and experience with XAI and Process Mining. The total number of participants in the experiment was 181. Note that there are 2 missing entries in the demographic data; these correspond to participants who completed the experiment but did not answer the demographic questions.

Table 5.2: Participant demographics (N=181, missing = 2)

Characteristics		***N***	**%**
Education (highest completed)	High School	3	1.7
	Bachelor or equivalent	98	54.1
	Master or equivalent	69	38.1
	Ph.D. or higher	9	5
STEM Background	Yes	138	22.7
	No	41	76.2
Process Mining Experience	Never worked with it	107	59.1
	Less than 1 year	41	22.7
	1-2 years	17	9.4
	2-3 years	6	3.3
	3-5 years	4	2.2
	5+ years	4	2.2
XAI Experience	Never worked with it	107	59.1
	Less than 1 year	32	17.7
	1-2 years	24	13.3
	2-3 years	6	3.3
	3-5 years	8	4.4
	5+ years	2	1.1

5.3.2 Experiment Decision Making Task Example

This section outlines the structure of the decision-making tasks that participants engaged in, showcasing how scenarios were presented. The general process that a participant followed for each scenario is as follows:

1. Scenario: Applicant's Profile and Process Map
2. AI Prediction: Accept / Reject
3. Participant's Initial Decision: Agree / Disagree, Confidence Level
4. XAI Explanation
5. Participant's Post-XAI Decision: Agree / Disagree, Confidence Level

This process is repeated 4 times with different applicant scenarios. The whole survey material is also provided in Appendix A.

Pre-Scenario

Before the scenarios were shown, the participants were provided with the accuracy of the AI system. They were shown as: *You will now review a set of loan applications along with AI predictions indicating whether the loan will be Accepted or Rejected. The AI system used for these decisions has a ( )% accuracy rate.* For the high accuracy group, they were assigned as 96% and for the low accuracy group, they were assigned as 63%. This was shown only one time. As we aim to test the perceived accuracy, only the numbers were shown differently. The scenarios and tasks that were given to the applicants were the same.

Scenario

The scenarios included the sequence of activities (Directly Follows Graphs (DFGs) from Disco), and applicant characteristics. Figure 5.3 shows one of the scenarios, that the participants were shown. This scenario was the one that was correctly classified as a "Reject Application(Applicant 2)". Participants were also provided information on "Starting Point" and "Current Status", and the total duration time of the application process. Figure 5.4 shows the example of the information. With previous information on the process map, applicant info, and this additional information, participants conduct a personal assessment of whether they agree or disagree with the prediction.

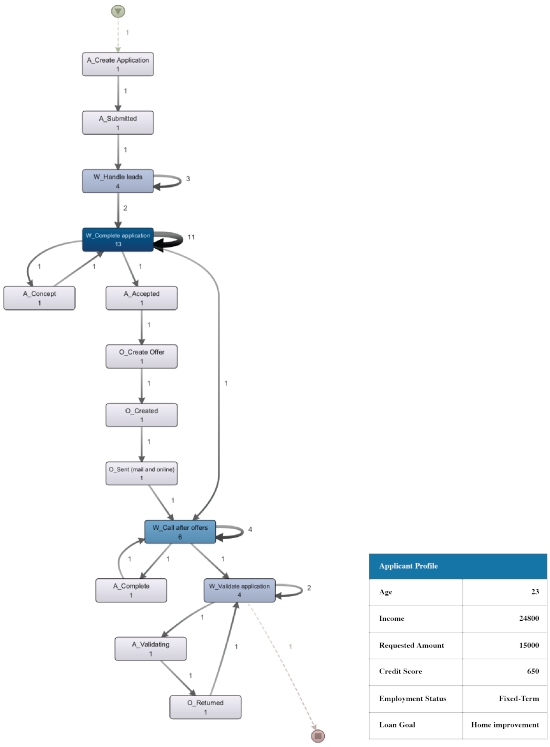


Figure 5.3: Scenario Example - Reject Application

Applicant 2 Process Map & Applicant Profile

* This is the 2nd applicant with the new information.

- **Starting Point**: A_Create Application
- **Current Status**: W_Validate Application
- **Duration:** 15 days

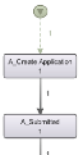


Figure 5.4: Introduction of Scenario

For the first scenario, participants were also provided with simple information on how to read the process map (Appendix A.1). In addition, they

were provided with a link that explains how to read the process map, in case they wanted to check the definitions of all activities (Appendix A.2).

Pre-Explanation Decision Stage

Initially, participants were presented with the AI prediction regarding the outcome of the loan application process. For example, the prediction was provided as: *"Based on this application process and the applicant's profile, the AI predicts that the application will be REJECTED after the last activity."* Participants were then asked whether they agreed with this prediction, responding with either Agree or Disagree to the question: *Do you agree with the AI Prediction of 'Reject Application'?"* Following this, participants rated their decision confidence on a scale of 1-5, based on their agreement or disagreement with the AI prediction. The question was phrased as: *"How confident are you with your decision on a scale of 1-5? (1- Not confident at all, 5- Very confident)"*

Explanation Styles

Depending on the group the participant is assigned to, participants are presented with different explanation styles. Figures 5.5, 5.6, 5.7 explanation that was provided to the participants. They show the Feature importance-based, Rule-based, and Counterfactual-based styles respectively.

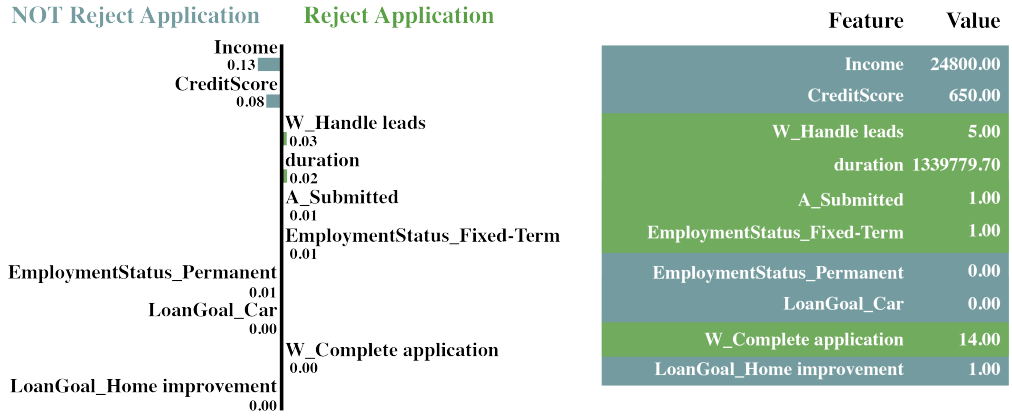


Figure 5.5: Feature importance-based style

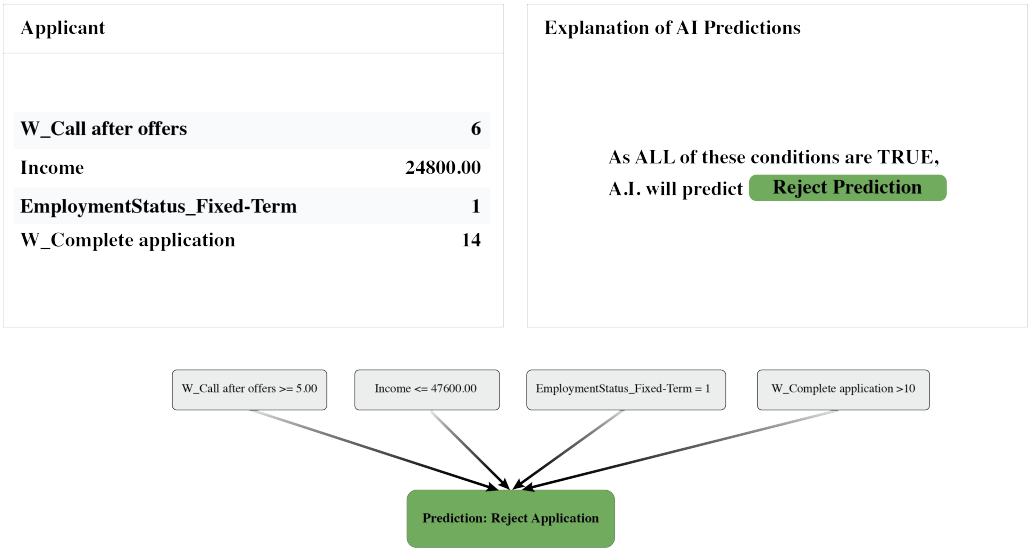


Figure 5.6: Rule-based style

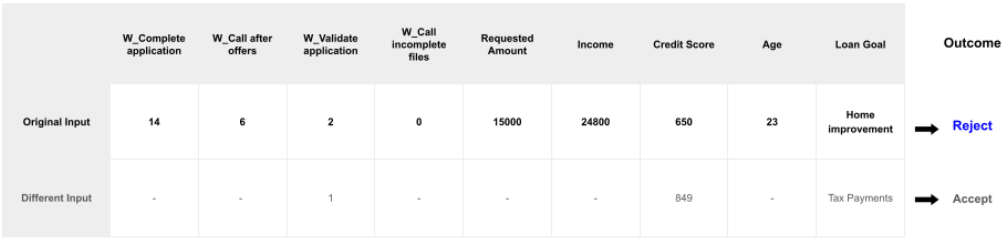


Figure 5.7: Counterfactual-based style

Post-Explanation Decision Stage

After being provided with the XAI explanations, participants were asked to reassess their agreement with the AI prediction and rate their decision confidence, following the same procedure as in the pre-explanation decision stage. The key difference was that this reassessment occurred after they had received the XAI explanations. This process aimed to measure the influence of the XAI explanations on their decisions. By comparing the pre- and post-explanation responses, we could evaluate the impact of different XAI explanations and determine how the provision of explanations affected participants' decisions.

5.3.3 Post-Questionnaire & Demographics

After completing the four decision-making scenarios, participants were asked to complete a post-questionnaire. The post-questionnaire consisted of four main sections:

1. Reliance on Information: Participants were asked to rank the information sources they relied on the most during the decision-making tasks: Process Map, Applicant's Profile, XAI Explanation, and AI Prediction. Figure 5.8 shows the example of the questions.

2. Satisfaction: Participants rated their satisfaction with the XAI explanation on a scale of 1 to 5 (1 - not satisfied at all, 5 - very satisfied).

3. Difficulty: Participants assessed the difficulty of understanding the XAI explanation on a scale of 1 to 5 (1 - extremely difficult, 5 - extremely easy).

4. Demographics: Participants provided information on their highest educational degree and education field.

5. XAI/Process Mining Experience: Participants indicated their level of experience with XAI or process mining, ranging from no experience to 5+ years.

Figure 5.8: Ranking Question Example

This data was collected to address the fourth evaluation objective: understanding the reasons behind the differences in effectiveness across different explanation styles.

5.4 Hypotheses

Based on the work of Kenny et al. [27], we anticipate a relationship between effectiveness measures and AI accuracy. Furthermore, we hypothesize that the styles of explanation provided can significantly influence effectiveness. These hypotheses are directly related to the evaluation objectives outlined earlier in Section 5.1. Briefly, Objective 1 focuses on the influence of perceived AI accuracy on effectiveness, Objective 2 examines the differences in effectiveness across various explanation styles, and Objective 3 investigates how the provision of explanations affects effectiveness and compares these changes across explanation styles. The relationships between these objectives and the corresponding hypotheses are summarized in Table 5.3.

Table 5.3: Evaluation Objectives and Hypotheses Alignment

Objectives	**Measure**	**Hypotheses**
Obj. 1	Task Performance	H1
	Agreement	H4
	Decision Confidence	H7
Obj. 2	Task Performance	H2
	Agreement	H5
	Decision Confidence	H8
Obj. 3	Task Performance	H3
	Agreement	H6
	Decision Confidence	H9

1. Task Performance

H1 (Perceived AI Accuracy Influence on Task Performance) focuses on the **initial** differences in task performance between the high-accuracy and low-accuracy groups before explanations are provided.

- $H1_0$: There is no difference in initial task performance between participants in the high-accuracy and low-accuracy groups before explanations are provided.
- $H1_1$: Participants in the high-accuracy group will have better initial task performance than those in the low-accuracy group before explanations are provided.

H2 (Explanation styles Influence on Task Performance) whether the styles of explanation (Rule-based, Counterfactual-based, Feature importance-based) affect task performance after explanations are provided.

- $H2_0$: The styles of explanation do not affect task performance after explanations are provided.
- $H2_1$: Task performance will vary significantly across different styles of explanations provided.

H3 (Explanation Provision Influence on Task Performance explores the impact of explanations on participants' ability to make correct decisions, considering potential differences in effectiveness across explanation styles.

- $H3_0$: The provision of explanations does not change the task performance of participants, regardless of the explanation styles.
- $H3_1$: Explanations will improve the task performance of participants on average, regardless of the specific explanation styles provided.

2. Agreement

H4 (Perceived AI Accuracy Influence on Agreement) focuses on the initial impact of accuracy before explanations.

1. $H4_0$: AI prediction accuracy does not affect the likelihood of participants agreeing with the AI's predictions (measured by the number of times participants' decisions align with AI predictions).
2. $H4_1$: Participants exposed to higher AI prediction accuracy are more likely to agree with the AI's predictions compared to those exposed to lower AI prediction accuracy.

H5 (Explanation styles Influence on Agreement) explores the effect of explanation styles on agreement after explanations are provided.

1. $H5_0$: The styles of explanation accompanying AI predictions do not influence participants' agreement with these predictions
2. $H5_1$: The styles of explanation provided will influence participants' agreement with AI predictions

H6 (Explanation Provision Influence on Agreement) This hypothesis investigates the overall impact of explanations on agreement, considering both before and after scenarios and potential variations by explanation styles.

1. $H6_0$: The provision of explanations does not influence participants' agreement with AI predictions, irrespective of the explanation styles.
2. $H6_1$: Participants' agreement with AI predictions increases after receiving explanations, with the degree of increase differing based on the styles of explanation provided.

3. Decision Confidence

H7 (Confidence Based on Perceived AI Accuracy) explores the influence of AI accuracy on participants' decision confidence.

- $H7_0$: There is no difference in decision confidence between participants in the high-accuracy and low-accuracy groups before explanations are provided.
- $H7_1$: Participants in the high-accuracy group will report higher decision confidence (on average) than those in the low-accuracy group before explanations are provided.

H8 (Confidence Based on Explanation styles) investigates whether the styles of explanation affect decision confidence after explanations are provided.

- $H8_0$: The styles of explanation provided do not influence participants' decision confidence after explanations are given.
- $H8_1$: Participants' decision confidence will vary significantly across different styles of explanations provided after explanations are given with certain explanation styles leading to higher confidence compared to others.

H9 (Confidence Change) Compare decision confidence before and after the explanations among the explanations groups.

- $H9_0$: Participants' decision confidence does not change before and after receiving an explanation.
- $H9_1$: Participants' decision confidence will increase (on average) after receiving an explanation compared to before.

5.5 Analytical Approaches

The variables used for testing the hypotheses varied depending on the specific hypothesis. Table 5.4 outlines the variables and statistical methods employed for each hypothesis. For all tests, we used G*Power to calculate the required sample size based on a mid-effect size to ensure valid results.

Hypotheses H1, H4, and H7, which examine the influence of AI accuracy, were tested using Independent Samples T-Tests. For these tests, G*Power calculations indicated that 88 participants were needed per accuracy group (high and low) with the following parameters: effect size = 0.5, $\alpha = 0.05$, and power (1-β err prob) = 0.95. Although one participant was missing from the high-accuracy group, we excluded them to maintain data quality.

Hypotheses H2, H5, and H8, which focus on outcomes after explanations are provided, were analyzed using One-Way ANOVA followed by Tukey HSD post-hoc tests. G*Power calculations with an effect size = 0.25, $\alpha = 0.05$, and power (1-β err prob) = 0.8 indicated that 26 participants were needed per explanation style group across both high and low accuracy, totalling 156 participants. For these hypotheses, we recruited sufficient participants for testing.

Hypotheses H3, H6, and H9, which assess changes before and after explanations, were tested using Paired Samples T-Tests. For these tests, G*Power calculations (effect size d = 0.5, $\alpha = 0.05$, power (1-β err prob) = 0.80) indicated that 26 participants per group were sufficient.

Apart from H1, H4, and H7, the remaining hypotheses were also analyzed separately for the high-accuracy and low-accuracy groups to explore the differential effects of perceived AI accuracy. Additionally, we conducted analyses combining explanation styles regardless of accuracy levels.

Table 5.4: Analytical Approaches per Hypotheses

Measure Task	**Hypotheses**	**IV**	**DV**	**Analysis Method**
Task Performance (TP)	H1	Perceived AI Acc.	Initial TP	Independent t-test
	H2	Exp. styles	Post TP	One-way ANOVA
	H3	Exp. styles	Initial TP - Post TP	Paired t-test
Agreement (Agg.)	H4	Perceived AI Acc.	Initial Agg.	Independent t-test
	H5	Exp. styles	Post Agg.	One-way ANOVA
	H6	Exp. styles	Initial Agg.-Post Agg.	Paired t-test
Decision Confidence (DC)	H7	Perceived AI Acc.	Initial DC	Independent t-test
	H8	Exp. styles	Post DC	One-way ANOVA
	H9	Exp. styles	Initial DC - Post DC	Paired t-test

Chapter 6

Results

This chapter presents the quantitative and qualitative results obtained from the participants and details the analyses performed. The analyses were conducted using IBM SPSS Statistics 27.

6.1 Descriptive Statistics

Tables 6.1 - 6.6 provide a summary of the results for *task performance*, *agreement*, and *decision confidence* across different explanation styles and accuracy levels.

Task performance is measured by the number of correct decisions made by participants, with a maximum possible score of 8 (2 points per scenario, across 4 scenarios). Agreement refers to the number of decisions where participants aligned with the AI's prediction, regardless of correctness, with a maximum score of 8 as well. Decision confidence is assessed on a 1-5 Likert scale, reflecting participants' confidence in their decisions, and is reported as an average score.

Task Performance

Table 6.1 summarizes the total task performance across explanation styles and accuracy groups. Participants in the high-accuracy group displayed varying levels of task performance depending on the explanation style provided. For those who received Feature importance explanations, the average task performance was 4.3 out of 8. Participants who received Rule-based explanations performed slightly better, with an average score of 4.5. However, those who received Counterfactual explanations had the lowest average task performance in the high-accuracy group, scoring 3.5 out of 8.

In contrast, participants in the low-accuracy group generally performed better across all explanation styles. Those who received Feature importance explanations had an improved average task performance of 4.5 out of 8, compared to their high-accuracy counterparts. Participants who received Rule-based explanations scored even higher, with an average of 4.9. The highest task performance was observed in the low-accuracy group receiving Counterfactual explanations, with an average score of 5.2 out of 8.

Table 6.1: Task Performance(avg.) by Accuracy, Explanation(Exp.) Styles

Accuracy	**Exp. Styles**	**Task Performance**	**SD**
	Feature importance	4.3	1.1
High	Rule	4.5	1.3
	Counterfactual	3.5	0.8
	Feature importance	4.5	0.9
Low	Rule	4.9	1.4
	Counterfactual	5.2	1.2

Table 6.2 shows task performance before and after explanations. The high-accuracy group showed minimal changes compared to the low-accuracy group. For Feature importance explanations, the average task performance slightly increased from 2.1 to 2.2. Those who received Rule-based explanations maintained a consistent score of 2.2 before and after explanations. Participants who received Counterfactual explanations experienced a minor improvement, with their average task performance rising from 1.7 to 1.8.

The low-accuracy group, however, showed more noticeable improvements. Participants who received Feature importance explanations saw their average task performance increase from 2.1 to 2.4. Those who received Rule-based explanations improved from 2.3 to 2.5. The most significant improvement was in the group receiving Counterfactual explanations, where the average task performance jumped from 2.3 to 2.8.

Table 6.2: Before, After Task Performance by Accuracy and Explanation(Exp.) Styles

Accuracy	Exp. Styles	Before Exp.	After Exp.
	Feature importance	2.1 (0.6)	2.2 (0.7)
High	Rule	2.2 (0.9)	2.2 (0.6)
	Counterfactual	1.7 (0.5)	1.8 (0.6)
	Feature importance	2.1 (0.6)	2.4 (0.6)
Low	Rule	2.3 (0.8)	2.5 (0.8)
	Counterfactual	2.3 (0.6)	2.8 (0.7)

Agreement

Table 6.3 summarizes the total agreement across explanation styles and accuracy groups. In the high-accuracy group, the Counterfactual explanations led to the highest agreement with AI predictions (7.0 out of 8), followed by Rule-based (6.4 out of 8) and Feature importance (6.1 out of 8).

In the low-accuracy group, Rule-based explanations led to the highest agreement (6.6 out of 8), with Feature importance and Counterfactual explanations both at 6.5 out of 8.

Table 6.3: Agreement(avg.) by Accuracy and Explanation(Exp.) Styles

Accuracy Level	Exp. Styles	Agreement avg.	SD
	Feature importance	6.1	1.9
High	Rule	6.4	1.5
	Counterfactual	7.0	1.3
	Feature importance	6.5	1.3
Low	Rule	6.6	1.6
	Counterfactual	6.5	1.3

Table 6.4 shows the average agreement before and after explanations. Rule-based explanations increased agreement in the high-accuracy group (from 3.1 to 3.3 out of 4), while Counterfactual explanations saw a slight decrease (from 3.5 to 3.4 out of 4). In the low-accuracy group, both Rule-based and Counterfactual explanations slightly increased agreement.

Table 6.4: Before, After Agreement by Accuracy and Explanation(Exp.) Styles

Accuracy	Exp. Styles	Before Exp.	After Exp.
	Feature importance	3.1(1.0)	2.9(1.2)
High	Rule	3.1(0.9)	3.3(0.9)
	Counterfactual	3.5(0.7)	3.4(0.7)
	Feature importance	3.4(0.8)	3.1(0.6)
Low	Rule	3.2(0.9)	3.2(0.9)
	Counterfactual	3.1(0.8)	3.4(0.7)

Decision Confidence

Decision confidence varied across explanation styles and accuracy levels. Table 6.5 shows the total decision confidence average. In the high-accuracy group, Rule-based explanations had the highest decision confidence with an average of 4.1, followed by Counterfactual (4.0), and Feature importance (3.6). Compared to the high-accuracy group, the low-accuracy group generally showed lower decision confidence. In the low-accuracy group, Feature importance-based explanations had the highest confidence (3.9), followed by Rule-based (3.8) and Counterfactual (3.7).

Table 6.5: Decision confidence(avg.) by Accuracy and Explanation Styles

Accuracy	Exp. Styles	Decision Confidence	SD
	Feature importance	3.6	0.9
High	Rule	4.1	0.7
	Counterfactual	4.0	0.6
	Feature importance	3.9	0.8
Low	Rule	3.8	0.9
	Counterfactual	3.7	0.6

Table 6.6 summarizes the decision confidence change between before and after explanations. In the high accuracy group, Rule-based explanations led to the highest decision confidence after explanations (4.1 out of 5), followed by Counterfactual (4.0 out of 5) and Feature importance (3.6 out of 5). For the low accuracy group, Feature importance explanations resulted in the

highest decision confidence after explanations (3.9 out of 5), followed closely by Rule-based explanations (3.8 out of 5) and Counterfactual explanations (3.7 out of 5).

Table 6.6: Before, After Explanation Decision Confidence by Accuracy and Explanation(Exp.) Styles

Accuracy	**Exp. Styles**	**Before Exp.**	**After Exp.**
	Feature importance	3.5(0.9)	3.6(0.9)
High	Rule	3.8(0.6)	4.0(0.6)
	Counterfactual	4.0(0.5)	3.9(0.5)
	Feature importance	3.7(0.9)	3.9(0.7)
Low	Rule	3.7(0.8)	3.8(0.8)
	Counterfactual	3.8(0.5)	3.6(0.6)

6.2 Hypotheses Result

All hypotheses were tested with a critical value of α =0.05. Table 6.7 provides a summary of the hypothesis results, with the "Validity" column indicating where the null hypothesis was rejected.

Hypotheses H1, H4, and H7 were tested by grouping participants based on accuracy level, independent of explanation styles. The remaining hypotheses were analyzed separately for the high and low-accuracy groups. Additionally, these hypotheses were tested by grouping participants according to explanation styles, regardless of accuracy levels.

In terms of task performance, H1 demonstrated that perceived AI accuracy significantly influenced initial task performance, with the low-accuracy group outperforming the high-accuracy group. H3 showed that task performance improved significantly after explanations were provided, particularly with the Counterfactual style in the low-accuracy group. This effect was also evident when participants were grouped by explanation styles. H5 revealed marginal differences in agreement between explanation styles, with Feature importance explanations leading to lower agreement rates, particularly when grouped by explanation styles. H6 demonstrated a significant decrease in agreement after explanations, especially with Feature importance-based styles in the low-accuracy group, as well as when grouped by explanation styles. Lastly, for decision confidence, H8 indicated that explanation styles influenced confidence levels, particularly in the high-accuracy group,

with Rule-based explanations resulting in higher confidence than Feature importance-based styles. The remaining sections provide detailed statistical analysis and results.

Table 6.7: Hypotheses Results Summary (Accuracy - Acc., Explanation - Exp.)

Measure	Hypotheses	Validity (Null Hypothesis Rejected)			
		High Acc.	Low Acc.	Grouped by Acc.	Grouped by Exp.
Task Performance	H1			✓	
	H2				
	H3		✓		✓
Agreement	H4				
	H5				✓
	H6		✓		✓
Decision Confidence	H7				
	H8	✓			
	H9				

6.2.1 Task Performance

First, we aimed to verify whether perceived AI accuracy, explanation styles, and explanation provision influence task performance, which measures the number of correct decisions.

H1: Accuracy Influence on Initial Task Performance

H1 investigated whether perceived AI accuracy levels influence initial task performance. The independent variable is the accuracy group (high vs. low), and the dependent variable is the initial task performance, measured before any explanations are provided. There was a significant difference in initial task performance between the high-accuracy and low-accuracy groups. Participants in the high-accuracy group had significantly different task performance compared to those in the low-accuracy group, with a mean difference of -0.24040 (df = 179, t = -2.137, p = .034). The effect size, measured by Cohen's d, was 0.75, indicating a medium to large effect.

As both task performance measurements were positive and calculated by subtracting low-accuracy task performance from high-accuracy task performance, the negative mean difference means that the low-accuracy group performed better. Thus, we reject the null hypothesis ($H1_0$), which states

that there is no difference in initial task performance between participants in the high-accuracy and low-accuracy groups before explanations are provided. The data supports the alternative hypothesis ($H1_1$), concluding that participants in the high-accuracy group have *lower task performance* before explanations are provided.

H2: Explanation Styles Influence on Task Performance

H2 investigated whether the styles of explanation affect task performance after explanations are provided. The independent variable is the explanation style, and the dependent variable is the task performance after explanations are given. However, the analysis revealed no significant difference in task performance across the different explanation styles for both the high ($F(2,85) = 2.971$, $p = .057$) and low-accuracy groups ($F(2,91) = 2.137$, $p = .124$) both indicating small effect sizes. This was the same when grouped with explanation styles as well. The Tukey HSD post hoc test also confirmed the ANOVA results, showing no significant differences in task performance between the different explanation styles in both groups

H3: Explanation Provision Influence on Task Performance

H3 examined whether providing explanations affects participants' task performance, with explanation provision as the independent variable and task performance as the dependent variable.

When testing the high and low-accuracy groups separately, the analysis showed a significant improvement in task performance for the Counterfactual-based style, particularly in the low-accuracy group. This is also shown in the descriptive statistics in Table 6.2. Specifically, in the low-accuracy group, task performance improved significantly from before (Mean = -0.4516, Std. Deviation = 0.7228) to after explanations, with a t-value of -3.478 (df = 30), $p = 0.002$, and an effect size of 0.72. No significant changes were observed in the high-accuracy group, nor in the Feature importance-based and Rule-based styles.

When grouped by explanation styles, task performance for the Counterfactual-based style increased significantly from before (Mean = -0.31034, Std. Deviation = 0.73046) to after explanations, with a t-value of -3.236 (df = 57), $p = 0.002$, and a medium effect size of 0.73.

Thus, we reject the null hypothesis ($H3_0$), which posits that explanations do not affect task performance. The data supports the alternative hypothesis ($H3_1$), indicating that Counterfactual-based explanations significantly

increase task performance, especially for the low-accuracy group, highlighting their effectiveness with a medium effect size.

6.2.2 Agreement

Second, we aimed to verify whether perceived AI accuracy, explanation styles and the provision of explanations influence agreement, measured by the number of decisions participants follow, regardless of whether those decisions are correct or incorrect.

H4: Accuracy Influence on Initial Agreement

H4 investigated whether perceived AI prediction accuracy affects the likelihood of participants agreeing with the AI's predictions. This follows the same analysis method as H1, except the dependent variable here is agreement instead of task performance. It indicated no significant difference in agreement between the high-accuracy and low-accuracy groups, with a t-value of -0.078 (df = 179) and a p-value of 0.938. The mean difference in agreement was very small (-0.01052), and the effect size, measured by Cohen's d, was -0.012, indicating a very small effect size. Thus, we fail to reject the null hypothesis ($H4_0$), suggesting that AI prediction accuracy does not significantly influence participants' agreement with AI predictions before explanations are provided.

H5: Explanation Styles Influence on Agreement

H5 examined whether the styles of explanation accompanying AI predictions influence participants' agreement with these predictions. The independent variable is the explanation style, and the dependent variable is agreement after explanations are provided. The analysis showed a marginally significant effect of explanation styles on agreement, with a p-value of 0.055, indicating a trend toward significance. Post hoc tests (Tukey HSD) revealed significant mean differences between the Feature importance and Counterfactual-based styles (mean difference = 0.39998, p = 0.043), suggesting differences in agreement rates between these styles.

The homogeneous subsets analysis indicated that the Counterfactual-based style had a higher mean agreement (mean = 3.4) compared to the other two styles (Feature importance-based style: mean = 3.04, Rule-based style: mean = 3.26).

However, when the high and low accuracy groups were analyzed separately, neither showed significant differences in agreement across explana-

tion styles. This suggests that while individual group effects may be subtle, the combined dataset reveals a trend toward significance, highlighting the Counterfactual-based style's potential impact.

Thus, we fail to reject the null hypothesis ($H5_0$), which states that explanation styles do not influence agreement. Nonetheless, the overall results suggest a marginally significant influence of explanation styles on agreement, particularly highlighting the impact of the Counterfactual-based style.

H6: Explanation Provision Influence on Agreement

H6 investigated whether the provision of explanations influences participants' agreement with AI predictions. The independent variable is the provision of explanations, and the dependent variable is agreement, measured before and after explanations are provided.

When testing the high and low-accuracy groups separately, the Feature importance-based style in the low-accuracy group showed a significant decrease in agreement, with a mean difference of 0.28 and a p-value of 0.027. This positive mean difference indicates that participants agreed more before the explanation was provided, as it was calculated by subtracting after-agreement from before-agreement. Conversely, in the high-accuracy group, there were no significant changes in agreement for any explanation styles.

When grouped by explanation styles, for the Feature importance-based style, the mean difference in agreement was 0.25, with a p-value of 0.028, indicating a significant decrease in agreement. The effect size was large, suggesting a substantial impact of explanations on agreement for this style. For the Rule-based and Counterfactual-based styles, there were no significant changes in agreement, as both p-values were above 0.05.

6.2.3 Decision Confidence

Lastly, we evaluate whether perceived AI accuracy, explanation styles, and the provision of explanations influence decision confidence, measured on a scale from 1 to 5.

H7: Accuracy Influence on Initial Decision Confidence

Same as H1 and H4, we measure whether decision confidence is influenced by the perceived AI Accuracy. The independent variable is the accuracy group (high vs. low), and the dependent variable is the initial decision confidence, measured before any explanations are provided.

The result indicated no significant difference in decision confidence between the high-accuracy and low-accuracy groups. The mean difference was 0.03800 (df = 179, t = 0.333, p = 0.739), with a very small effect size (Cohen's d = 0.050). Thus, we fail to reject the null hypothesis ($H7_0$), which states that there is no difference in decision confidence between participants in the high-accuracy and low-accuracy groups before explanations are provided. The data does not support the alternative hypothesis ($H7_1$). These results suggest that perceived AI accuracy does not significantly influence decision confidence before any explanations are provided.

H8: Decision Confidence Based on Explanation Styles

H8 investigates whether the explanation styles affect decision confidence after explanations are provided. The independent variable is the explanation style, and the dependent variable is the decision confidence after explanations are given.

When tested grouping by explanation styles, the results indicated no significant difference in decision confidence among the different explanation styles. The analysis showed a p-value of 0.329 (df = 2, F = 1.117), which is greater than the significance level of 0.05, suggesting that explanation styles do not significantly affect decision confidence. Thus, we fail to reject the null hypothesis ($H8_0$), which states that the styles of explanation do not influence participants' decision confidence after explanations are provided.

However, when tested separately, the results were significant (p = 0.031, F = 3.630), suggesting that explanation styles do have an impact on decision confidence in the high-accuracy group. Specifically, post hoc analysis revealed that Rule-based explanations resulted in higher confidence compared to Feature importance-based explanations, with a significant mean difference of 0.47796 (p = 0.033). This indicates that for participants in the high-accuracy group, Rule-based explanations may have greater confidence in decision-making than Feature importance explanations.

H9: Decision Confidence Based on Explanation Styles

H9 investigates whether decision confidence changes before and after receiving an explanation, considering different explanation styles. The independent variable is the explanation style, and the dependent variable is the change in decision confidence, measured before and after explanations.

When analyzing the high and low accuracy groups separately, the results were non-significant for both groups: high accuracy (Feature importance-based p=0.239, Rule-based p=0.076, Counterfactual-based p=0.557); low ac-

curacy (Feature importance-based p=0.372, Rule-based p=0.585, Counterfactual-based p= 0.285).

Overall, the analysis indicated no significant change in decision confidence before and after explanations for any of the explanation styles, with p-values above 0.05: Feature importance-based (p=0.253), Rule-based (p=0.135), Counterfactual-based (p=0.218). The effect sizes were small, suggesting a negligible impact of explanations on decision confidence. Thus, we fail to reject the null hypothesis ($H9_0$), which states that participants' decision confidence does not change before and after receiving an explanation.

6.3 Ranking Analysis

After completing the decision-making task, participants were asked to rank the information they relied on the most, with the choices being the Applicant's Profile, AI Explanation, AI Prediction, and Process Map. Rankings were measured on a scale from 1 to 4, with 1 indicating the highest reliance and 4 the lowest. Table 6.8 shows a summary of the average rankings across different accuracy levels and explanation styles. In Table 6.8 - 6.10 and Figures 6.1 - 6.4, the Applicant's Profile is abbreviated as "Profile," and "AI Explanation" refers to the explanation provided to participants after the pre-decision stage. Overall, participants most frequently relied on the Applicant's Profile, followed by the AI Explanation, AI Prediction, and Process Map. This trend is all presented in Figures 6.1 - 6.4.

As shown in Figure 6.1, participants in the low-accuracy group tended to rely more on the instances provided in the scenario (Process Map and Applicant's Profile). For Process Map, the low accuracy group ranked as 3.11 while the high accuracy group ranked as 3.24 on average. For the Applicant's Profile, the low-accuracy group ranked as 1.40 while the high-accuracy group ranked as 1.48. In contrast, participants in the high-accuracy group placed greater reliance on AI Explanation (high-accuracy: 2.35, low-accuracy:2.45) and AI Prediction (high-accuracy:2.92, low-accuracy:2.45). Note that a *lower* number means *higher reliance*.

Table 6.8: Ranking Avg. Summary by Accuracy and Explanation(Exp.) Styles (1- highest reliance, 4- lowest reliance)

Accuracy	Exp. Styles	Profile	AI Exp.	AI Prediction	Process Map
	FI	1.54	2.37	2.96	3.14
High	Rule	1.49	2.18	2.93	3.40
	CF	1.42	2.51	2.87	3.19
	FI	1.31	2.47	2.86	3.37
Low	Rule	1.44	2.30	3.31	2.95
	CF	1.46	2.56	2.95	3.03

Table 6.9: Ranking by Accuracy

Accuracy	Profile	AI Explanation	AI Prediction	Process Map
High	1.48	2.35	2.92	3.24
Low	1.40	2.45	3.04	3.11

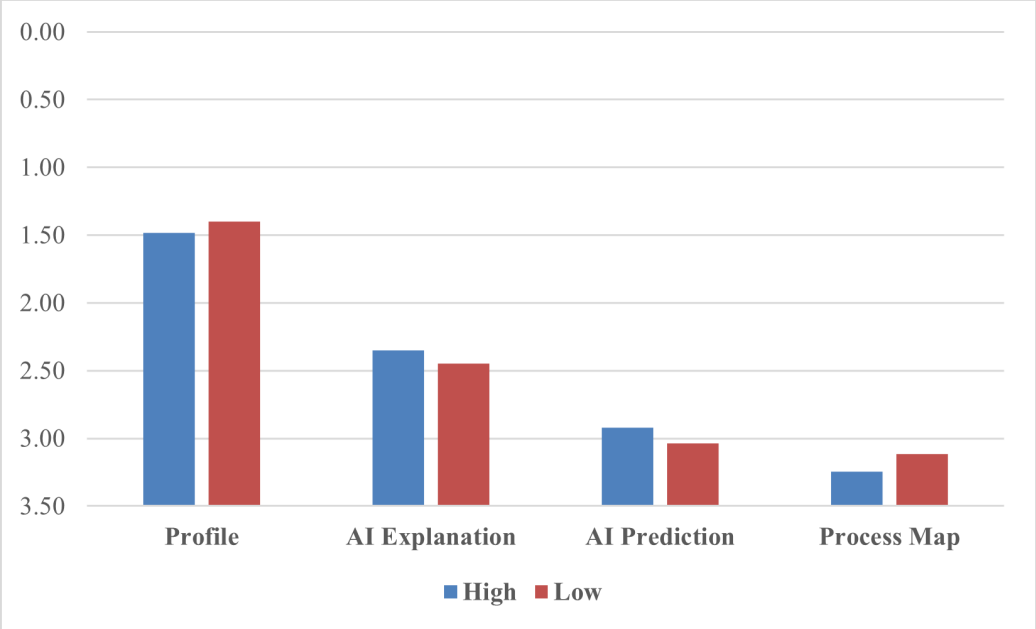


Figure 6.1: Ranking Comparison by Accuracy Level

When examining reliance by explanation style (Table 6.10 and Figure 6.2), all three explanation styles showed similarly high reliance on the Applicant's Profile. However, participants who received Rule-based explanations

relied the most on the AI Explanation compared to other groups (Rule-based: 2.24, Feature importance-based: 2.42, Counterfactual-based 2.54). Conversely, those who received Rule-based explanations relied the least on AI Prediction (Rule-based: 3.12, Feature importance-based: 2.91, Counterfactual-based 2.91). Regarding the Process Map, participants who received Counterfactual-based explanations showed a marginally higher reliance than other explanation styles (Counterfactual-based:3.11, Rule-based: 3.18, Feature importance-based: 3.25).

Table 6.10: Ranking by Explanation (Exp.) Styles

Exp. Styles	**Profile**	**AI Explanation**	**AI Prediction**	**Process Map**
Feature Importance	1.42	2.42	2.91	3.25
Rule	1.46	2.24	3.12	3.18
Counterfactual	1.44	2.54	2.91	3.11

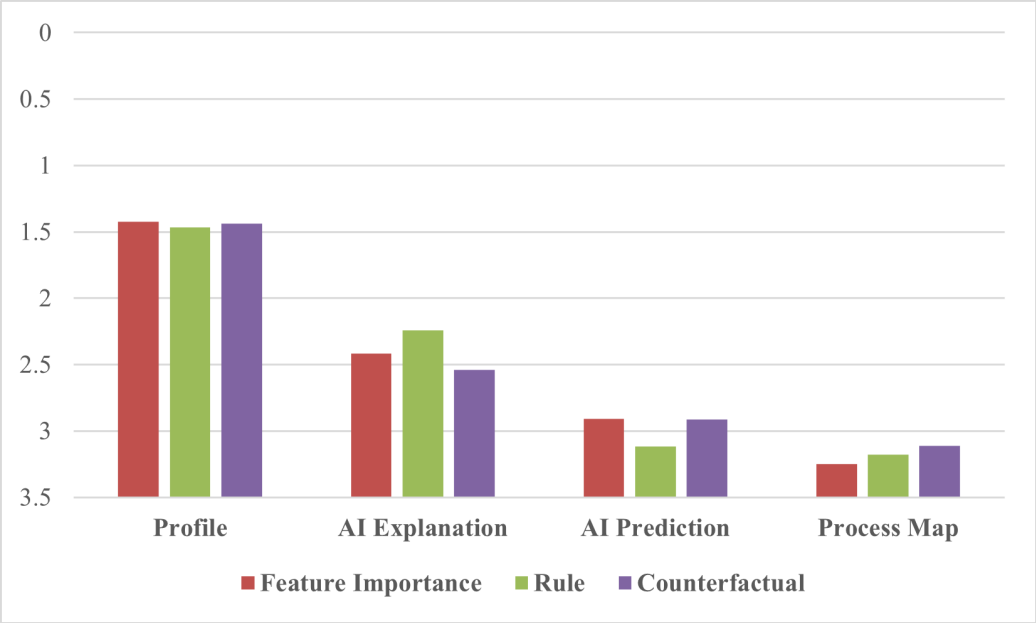


Figure 6.2: Ranking Comparison by Explanation Styles

A more detailed comparison of the high-accuracy (Figure 6.3) and low-accuracy group (Figure 6.4) revealed a significant difference in the reliance on AI Prediction and Process Map within the Rule-based explanation group. In the low-accuracy group, participants who received rule-based explanations tended to rely more on the Process Map than on AI Prediction. In contrast, those in the high-accuracy group showed greater reliance on AI Prediction

over the Process Map. Across both groups, participants who received Rule-based explanations generally relied more on the explanation than those in other groups. Additionally, participants who received Counterfactual-based explanations in the low-accuracy group relied more on the Process Map compared to their high-accuracy counterparts.

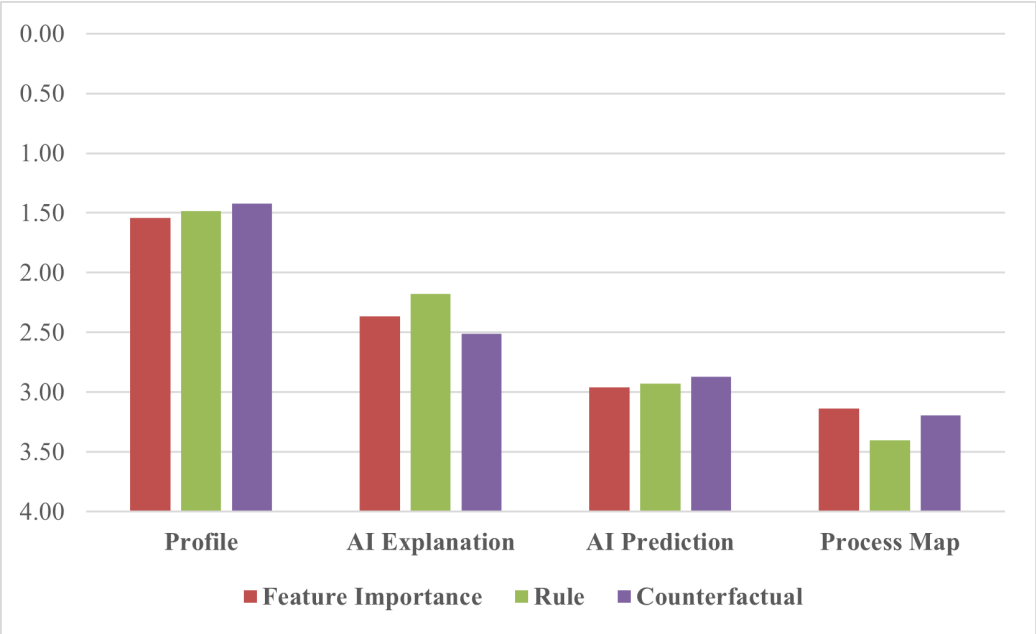


Figure 6.3: High Accuracy Group Ranking

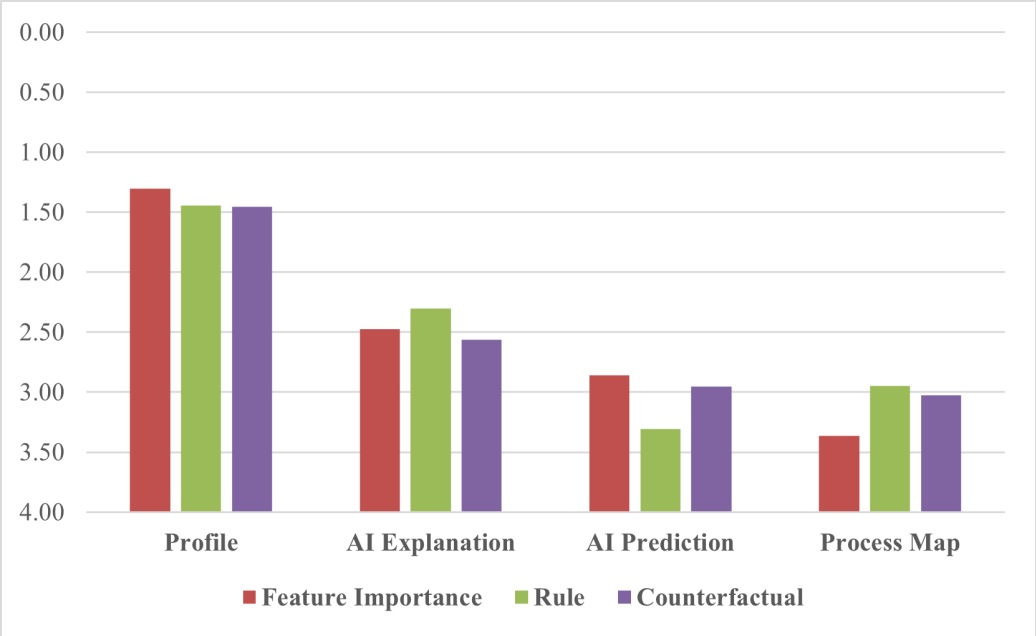


Figure 6.4: Low Accuracy Group Ranking

6.4 Satisfaction and Difficulty

Based on the explanation styles and accuracy levels received, we expected different levels of satisfaction with the explanation and difficulty in understanding. Satisfaction was measured on a 1-5 Likert scale, where 1 indicates "Not Satisfied" and 5 indicates "Very Satisfied." Difficulty was also measured on a 1-5 Likert scale, where 1 indicates "Extremely Difficult" and 5 indicates "Extremely Easy." Table 6.11 and Figure 6.5 show the average satisfaction scores, and Table 6.12 and Figure 6.6 show the average difficulty scores. The results indicate that the satisfaction and difficulty levels varied across different explanation styles and accuracy levels.

For high accuracy, the Rule-based explanation style had the highest average satisfaction score of 3.80, followed by Feature importance with a score of 3.66, and Counterfactual with a score of 3.41. In terms of difficulty, the Rule-based explanation style was rated as the easiest with an average score of 3.30, followed by Counterfactual at 3.04, and Feature importance at 2.97.

For low accuracy, the Rule-based explanation style again had the highest average satisfaction score of 3.61, followed by Feature importance at 3.39, and Counterfactual at 3.19. Regarding difficulty, the Rule-based explanation style was rated as the easiest with an average score of 3.29, followed by Feature importance at 3.23, and Counterfactual at 2.84.

Table 6.11: Satisfaction Average by Explanation Styles

Accuracy	**Explanation Styles**	**Satisfaction Avg. (1-5)**
	Feature Importance	3.66
High	Rule	3.80
	Counterfactual	3.41
	Feature Importance	3.39
Low	Rule	3.61
	Counterfactual	3.19

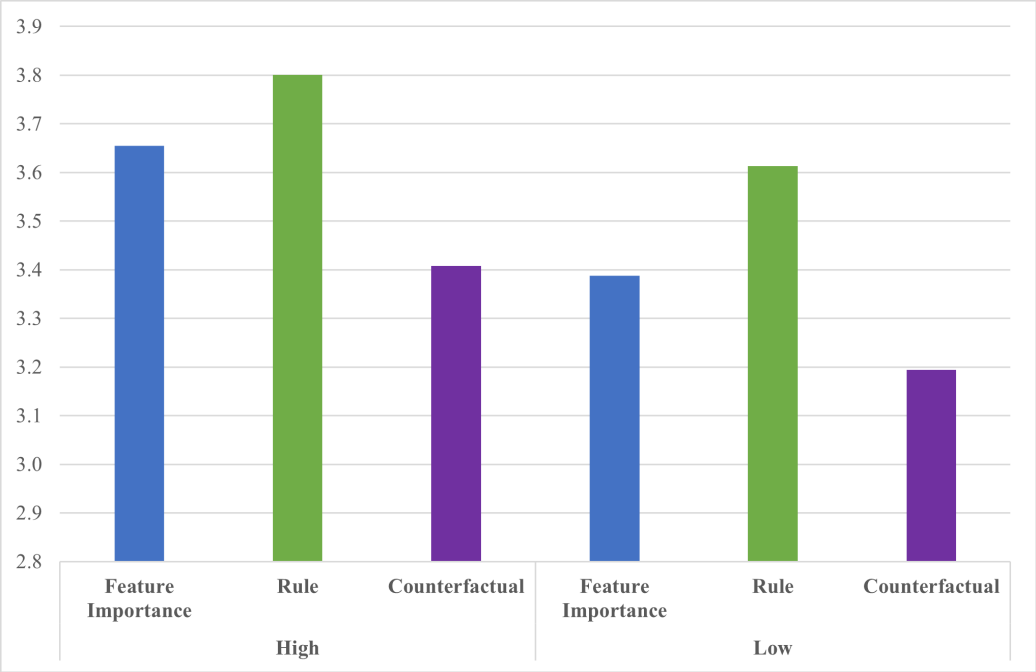


Figure 6.5: Satisfaction Average by Explanation Styles

Table 6.12: Difficulty level by Explanation Styles

Accuracy	**Explanation Styles**	**Difficulty Avg. (1-5)**
	Feature Importance	2.97
High	Rule	3.30
	Counterfactual	3.04
	Feature Importance	3.23
Low	Rule	3.29
	Counterfactual	2.84

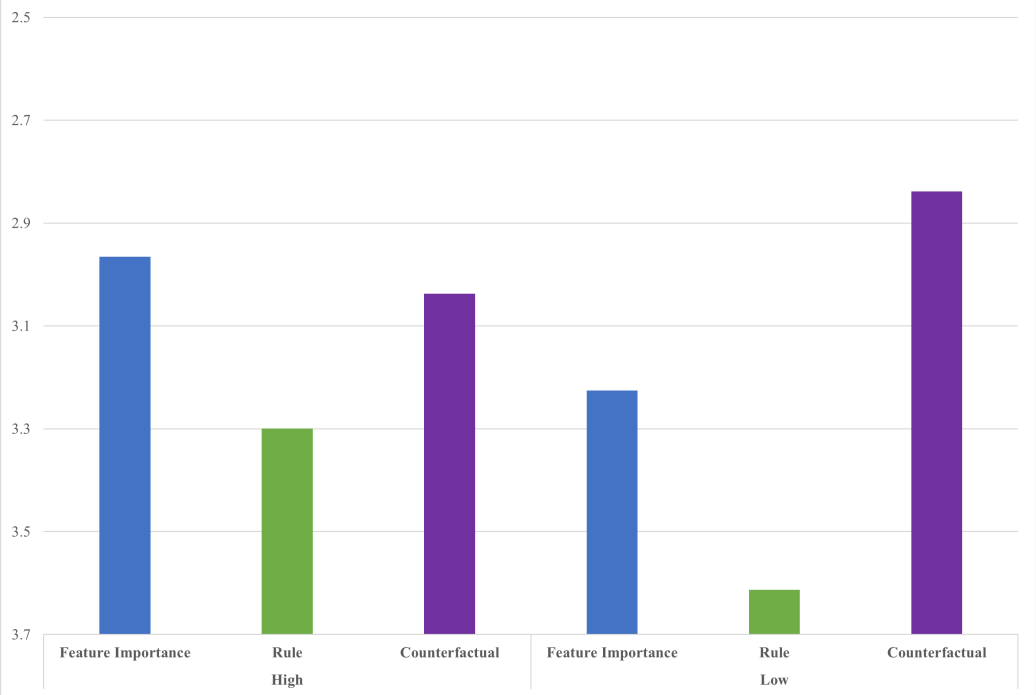


Figure 6.6: Difficulty Average by Explanation Styles

These results suggest that the Rule-based explanation style generally led to higher satisfaction compared to the other styles, regardless of the accuracy level. Additionally, it was perceived as easier to understand. The Feature importance explanation style had relatively lower satisfaction and higher difficulty ratings, particularly in the high accuracy condition. The Counterfactual explanation style tended to have lower satisfaction and higher difficulty ratings than the other styles.

6.5 Background-based Statistics

As the experiment scenario was in the process domain and related to XAI, we anticipated different results based on participants' backgrounds, including their STEM background and experience in Process Mining and XAI. Tables 6.13, 6.14- 6.17 summarize the *task performance*, *agreement*, and *decision confidence* with different backgrounds, all measured after explanations. Overall, there were no significant differences between the backgrounds in STEM and experience in XAI.

The differences in task performance based on the background were relatively small, with a process mining background showing a slightly more noticeable positive impact, particularly for participants with less than 1 year of experience, as shown in Table 6.15. This might be due to the first batch distribution where we tested with Process Mining course enrolled students. Differences in agreement rates are also minor, with process mining experience slightly reducing agreement rates (6.14, 6.15). Regarding decision confidence, a background in STEM, Process Mining, and XAI generally increases decision confidence, with the most noticeable increase observed in participants with XAI experience. However, task performance among participants with more than 5 years of XAI experience was the lowest, though this group consisted of only two individuals. Thus, these findings should be interpreted with caution due to the skewed data distribution in certain groups.

Table 6.13: Task Performance, Agreement, Decision Confidence Avg. & SD by STEM Background

STEM Background	**Task Performance**	**Agreement**	**Decision Confidence**
No	2.4(0.8)	3.3(0.8)	3.6(0.7)
Yes	2.3(0.7)	3.2(0.9)	3.8(0.8)

Table 6.14: Task Performance, Agreement, Decision Confidence Avg. & SD by Process Mining(PM) Experience

PM Experience	**Task Performance**	**Agreement**	**Decision Confidence**
No	2.2 (0.7)	3.3 (0.8)	3.8 (0.7)
Yes	2.4 (0.8)	3.1 (1.0)	3.9 (0.7)

Table 6.16: Task Performance, Agreement, Decision Confidence by XAI Experience

XAI Experience	Task Performance	Agreement	Decision Confidence
No	2.3 (0.7)	3.2 (0.8)	3.7 (0.8)
Yes	2.3 (0.7)	3.2 (0.9)	3.9 (0.6)

Table 6.15: Task Performance, Agreement, Decision Confidence by Process Mining(PM) Experience (Detail version)

PM Experience	Task Performance	Agreement	Decision Confidence
No experience	2.2 (0.7)	3.3 (0.8)	3.8 (0.7)
Less than 1 year	2.6 (0.8)	3.0 (0.9)	3.7 (0.8)
1 - 2 years	2.2 (0.7)	3.2 (1.2)	4.0 (0.4)
2 - 3 years	2.0 (0.6)	3.3(0.8)	3.8 (0.7)
3 - 5 years	2.5 (1.0)	3.5 (1.0)	4.3 (0.4)
5 + years	2.2 (0.5)	3.2(0.9)	3.9 (0.8)

Table 6.17: Task Performance, Agreement, Decision Confidence by XAI Experience (Detail version)

XAI Experience	Task Performance	Agreement	Decision Confidence
No experience	2.3 (0.7)	3.2 (0.8)	3.7 (0.8)
Less than 1 year	2.3 (0.7)	3.2 (1.0)	3.7 (0.7)
1 - 2 years	2.2 (0.5)	3.3 (0.8)	4.0 (0.5)
2 - 3 years	2.5 (1.0)	3.1 (1.4)	3.9 (0.7)
3 - 5 years	2.6 (0.9)	2.8 (0.9)	4.0 (0.5)
5 + years	1.5 (0.7)	3.5 (0.7)	4.6 (0.1)

6.6 Qualitative Analysis

To gain qualitative insights into XAI explanations, participants were asked to provide feedback on potential improvements for the explanations they received after the experiment section in the survey. The following are the summarized points from the participants who expressed similar opinions. More points are also available in the Appendix A.4

Feature Importance-based Explanations

Participants who received Feature importance-based explanations mentioned several areas for improvement:

1. **Weighting of Reasons:** Users expressed a desire for clear explanations on the weighting of each reason in the decision-making process (5 participants).

2. **Specific Details on Variables:** Participants requested detailed explanations on why certain thresholds are set, how the duration factors in, and the significance of percentage gaps between income and loan amounts (7 participants).

3. **Contextual Scaling:** There is a need for scales to provide context (e.g., credit score ranges) and to explain the relative values of data used in decision-making (6 participants).

Rule-based Explanations

Participants who received Rule-based explanations highlighted the following improvements:

1. **Step-by-Step Process:** Participants called for step-by-step explanations of the decision-making process, ideally presented in a flowchart format (8 participants).

2. **Detailed Feature Explanations:** Participants wanted more emphasis on explaining what each rule or feature means and how it impacts the decision. They also requested specific conditions under which decisions are made to help understand the decision logic (6 participants).

Counterfactual Explanations

Participants who received Counterfactual explanations suggested the following enhancements:

1. **Focus on Alternatives:** Users wanted to see several different alternatives that could affect the decision in combination(9 participants).

2. **Understanding Hierarchies:** There was a specific interest in understanding the hierarchy of importance factors and how different combinations of factors influence decisions (4 participants).

3. **Comparative Examples:** Participants desired examples of accepted cases in the event of rejection. They wanted to see comparisons between their case and similar accepted cases to better understand the reasons behind the AI's decision (7 participants).

In the feedback from participants who received Feature importance-based explanations, a common request was for information on the relative importance of different factors influencing the AI's prediction and broader contextual information to help interpret the AI's decisions. For Rule-based explanations, participants expressed a preference for a flowchart format to make the decision logic more accessible and easier to follow. They also wanted explanations for why specific rules influenced the AI's decision. Participants who received Counterfactual explanations wanted to see more alternatives and a hierarchy of important factors. They also wanted to see the comparisons between their case and similar accepted cases to better understand the reasons behind the AI's decision.

Chapter 7

Discussion

In this chapter, we discuss the findings of the experiment's results that eventually answer the research questions (RQ2, RQ3) and related evaluation objectives in Section 5.1 in this thesis. Furthermore, we address the limitations and future works related to the findings.

7.1 Research Questions & Findings

In the experiment, we evaluated the effectiveness of three different explanation styles (Feature importance-based, Rule-based, and Counterfactual-based) in decision-making by considering the level of perceived AI accuracy in the PPM domain. We used metrics such as "Task Performance," "Agreement," and "Decision Confidence" to conduct a comprehensive user evaluation that includes both objective and subjective measures in a human-grounded evaluation. We investigated subjective opinions about each explanation style through a post-questionnaire assessing information reliance (ranking), satisfaction, difficulty, and open-ended questions. This was done to potentially find the reasons behind the hypothesis results and trends observed in the descriptive statistics. Furthermore, as this study was conducted in the PPM domain, we also aimed to determine if there were varying effects based on participants' educational backgrounds or experiences in Process Mining and XAI.

7.1.1 RQ 2 - Obj. 1

RQ2 - Obj. 1 was about investigating the influence of perceived AI accuracy levels on users' decision-making. We have found that there were differences in effectiveness between high and low-accuracy in decision-making.

Difference in Task Performance

As shown in the task performance measure in Table 7.1 and Figures 7.1, 7.2, it has shown that participants in the low-accuracy group generally performed better than the high-accuracy group. Specifically, participants in the low-accuracy group outperformed those in the high-accuracy group when comparing the same explanation styles.

Moreover, hypothesis H1 demonstrated that participants in the low-accuracy group exhibited better initial task performance than those in the high-accuracy group. This initial performance was assessed before any explanations were provided, indicating that participants relied solely on process maps, application profiles, and the initial prediction and accuracy rate presented to them. These results suggest that the lower perceived accuracy of AI predictions may drive participants to engage more critically with the information. In addition, ranking analysis in Figures 6.1 shows that the higher accuracy group focused more on AI prediction than the low accuracy group did. This addresses the second research question (RQ2 - Obj. 1), showing that differences in task performance exist between high and low-accuracy groups, regardless of the explanation styles used.

Table 7.1: Task Performance, Agreement, Decision Confidence across Accuracy and Explanation(Exp.) Styles

Accuracy	**Exp. Styles**	**Task Performance**	**Agreement**	**Decision Confidence**
	Feature Importance	4.30	6.10	3.58
High	Rule	4.50	6.43	4.07
	Counterfactual	3.50	7.00	3.97
	Feature Importance	4.50	6.50	3.88
Low	Rule	4.90	6.60	3.81
	Counterfactual	5.20	6.50	3.66

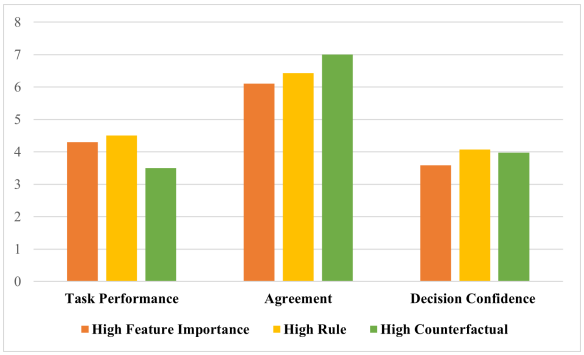


Figure 7.1: Task Performance, Agreement, Decision Confidence in High Accuracy Explanation Styles

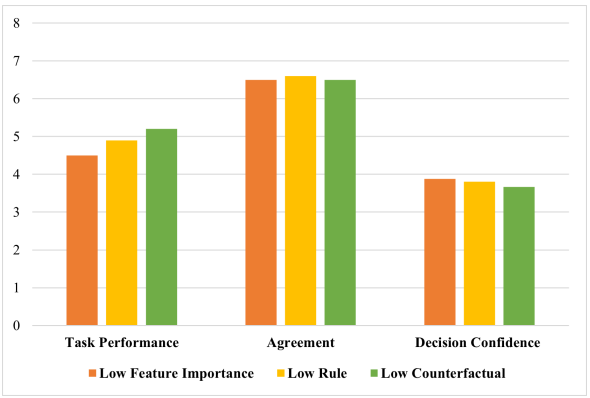


Figure 7.2: Task Performance, Agreement, Decision Confidence in Low Accuracy Explanation Styles

Different results in High and Low Accuracy Groups

While only H1, H4, and H7 specifically examined differences between high and low-accuracy groups irrespective of explanation styles, other hypotheses revealed notable distinctions between these groups when considering explanation styles. For instance, hypothesis H3, which focused on changes in task performance, was significant only in the low-accuracy group for the Counterfactual explanation style. This was not observed in the high-accuracy group. The descriptive results further indicated that participants in the low-accuracy group outperformed the high-accuracy group after explanations were provided. This trend is reflected in Table 7.2 in the task performance measure, showing that an AI with a perceived low accuracy might encourage more user engagement.

Table 7.2: Before, After Explanation (Exp.): Task Performance, Agreement, Decision Confidence among Accuracy and Exp. Styles (Feature importance - FI. Counterfactual-CF)

Accuracy	Exp. Styles	Task Performance		Agreement		Decision Confidence	
		Before Exp.	After Exp.	Before Exp.	After Exp.	Before Exp.	After Exp.
High	FI	2.1	2.2	3.1	2.9	3.5	3.6
	Rule	2.2	2.2	3.1	3.3	3.8	4.0
	CF	1.7	1.8	3.5	3.4	4.0	3.9
Low	FI	2.1	2.4	3.4	3.1	3.7	3.9
	Rule	2.3	2.5	3.2	3.2	3.7	3.8
	CF	2.3	2.8	3.1	3.4	3.8	3.6

Similarly, hypothesis H6, which examined changes in agreement, showed a significant decrease only in the low-accuracy group for the Feature importance style. In contrast, the high-accuracy group showed a smaller trend of agreement decrease. On the other hand, differences in decision confidence between the groups were significant only in the high-accuracy group, where participants who received Rule-based explanations reported higher decision confidence (H8). These findings show that even with identical scenarios and explanations, perceived AI accuracy can lead to varying outcomes.

7.1.2 RQ 2 - Obj. 2

Obj. 2 investigated the difference in effectiveness between the explanation styles. The results showed that there were differences in *agreement* measure between Counterfactual-based and Feature importance-based explanations. In addition, there was a significant difference in *decision confidence* between Feature importance and Rule-based styles.

Counterfactual & Feature Importance- Agreement

Hypothesis H5 highlighted differences in the persuasiveness of various explanation styles, particularly in terms of agreement. The counterfactual explanation style was significantly more persuasive than both the Rule-based and Feature importance styles, with notable differences compared to the Feature importance-based style.

Counterfactual explanations, which present alternative scenarios and "what if" options, might have demonstrated stronger persuasive power. This finding is consistent with previous research, where counterfactual explanations

were perceived as more helpful than causal explanations [11]. Additionally, another user study suggests that the natural human tendency toward "Counterfactual thinking may enhance the persuasive impact and effectiveness of these explanations [61]. These results also align with the descriptive statistics in the agreement measure, as shown in Table 7.1 where the Counterfactual-based style showed a higher agreement rate compared to all other explanation styles which are close to causal explanations.

Feature Importance & Rule-based - Decision Confidence

Regarding decision confidence, Hypothesis H8 indicated significant differences between Feature importance-based and Rule-based explanation styles. Participants who received Rule-based explanations reported significantly higher decision confidence than those who received Feature importance-based explanations. The ranking analysis (Section 6.3) shows that participants who received Rule-based explanations relied more on AI Explanation than Feature importance, which may have resulted in these statistics. In addition, this result aligns with findings on satisfaction and difficulty levels (Table 7.3), as discussed further in RQ3. Rule-based explanations' clarity and straightforward nature may have contributed to higher decision confidence by providing users with a more accessible understanding of AI predictions.

Table 7.3: Satisfaction and Difficulty across Accuracy and Explanation(Exp.) Styles Satisfaction (1 - Very Dissatisfied, 5- Very Satisfied, Difficulty (1 - Extremely difficult, 5 - Extremely easy)

Accuracy	**Explanation Styles**	**Satisfaction**	**Difficulty**
High	Feature Importance	3.7	3.0
	Rule	3.8	3.3
	Counterfactual	3.4	3.0
Low	Feature Importance	3.4	3.2
	Rule	3.6	3.6
	Counterfactual	3.2	2.8

7.1.3 RQ 2 - Obj. 3

Obj. 3 focused on understanding how the provision of explanations influences effectiveness measures and identifying any differences between the explanation styles. We conducted a separate analysis for high and low accuracy groups, while also examining whether similar trends emerged across explanation styles within both accuracy groups. Initially, we assessed the impact of providing explanations on all effectiveness measures, followed by investigating patterns across explanation styles in these measures.

Improved Task Performance: Counterfactual Style

For task performance, most explanation style groups showed an increase after explanations were provided. The descriptive statistics in Table 7.2 revealed that task performance improved across both accuracy groups. As expected, the provision of additional information generally led to better task performance.

When comparing task performance improvements between high and low-accuracy groups, the low-accuracy group showed greater improvements than the high-accuracy group, as visualized in Figure 7.3, which illustrates the improvements across all accuracy and explanation groups. This suggests that perceived accuracy may play a significant role in influencing task performance. Especially, the Counterfactual style in the low-accuracy group showed a 19% improvement rate. This finding was further supported by H3, which showed a significant increase in task performance for the Counterfactual-based style. These results suggest that Counterfactual explanations might be particularly sensitive to the influence of perceived AI accuracy, leading to greater effectiveness in this context compared to other explanation styles.

Decreased Agreement: Feature importance- based style

When comparing agreement across different explanation styles, we observed a consistent trend in the Feature importance style for both accuracy groups. Participants who received Feature importance explanations experienced a decrease in agreement, a trend that was statistically significant as indicated by H6. This decrease in agreement was notable compared to the initial agreement before any explanation was provided. Figure 7.4 illustrates the agreement trends across the different explanation styles.

Meanwhile, for Rule-based explanations, agreement increased after the explanation was provided in the high-accuracy group, while it remained consistent in the low-accuracy group. In contrast, Counterfactual explanations

showed different trends: agreement decreased slightly in the high-accuracy group but increased in the low-accuracy group.

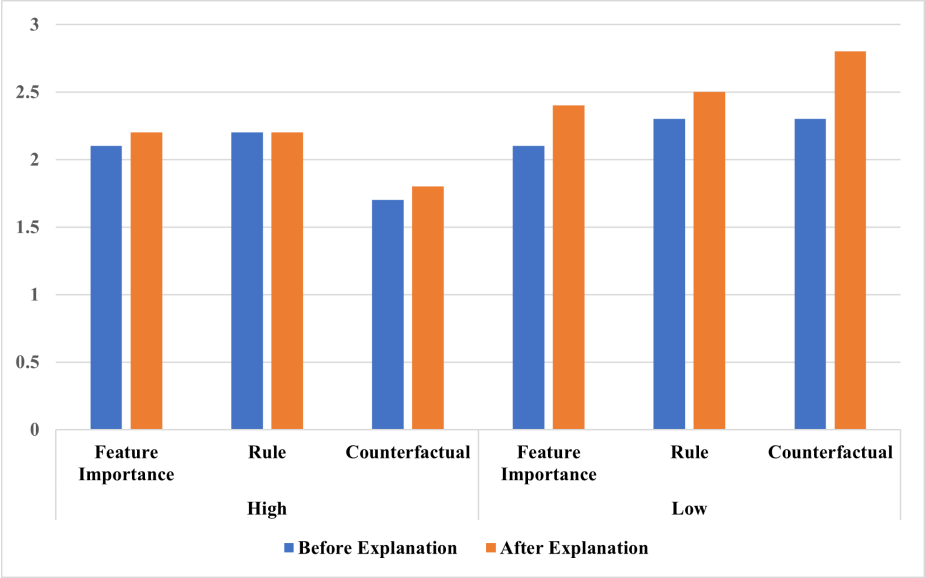


Figure 7.3: Task Performance Before, After Explanations

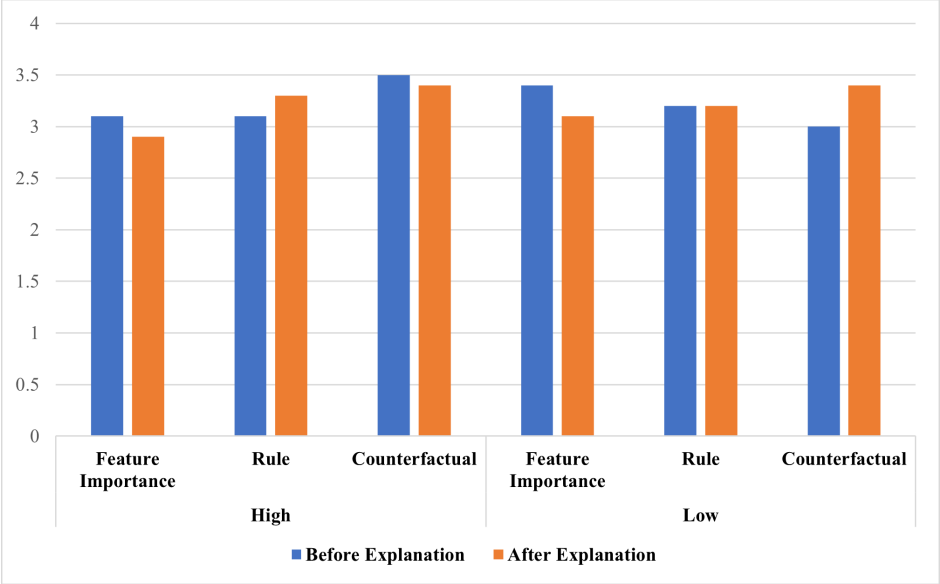


Figure 7.4: Agreement Before, After Explanations

Interaction between Perceived AI Accuracy and Explanation Styles on Task Performance and Agreement

We initially expected a high persuasive power, meaning higher agreement would affect task performance negatively. This follows the argumentation that persuasive explanations can cause harm as they may convince users to over-trust a system [45]. Based on this, we analyzed the trend between agreement and task performance across all explanation styles, grouped by the same explanation styles. Figures 7.5 - 7.7 illustrate the trends.

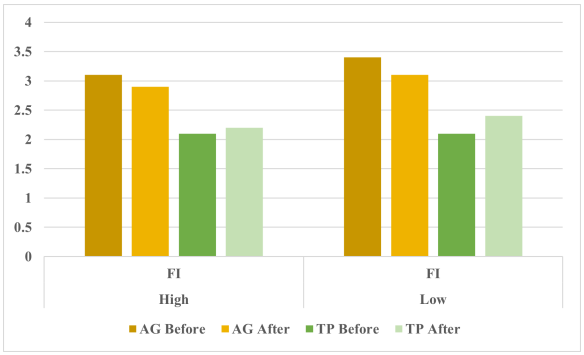


Figure 7.5: Before, After Task Performance and Agreement for Feature importance-based Explanations (FI) in High and Low Accuracy group

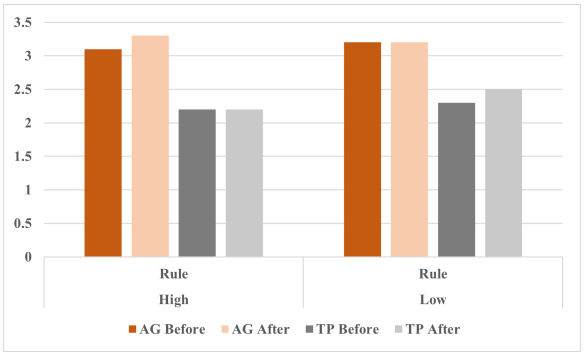


Figure 7.6: Before, After Task Performance and Agreement for Rule-based Explanations in High and Low Accuracy group

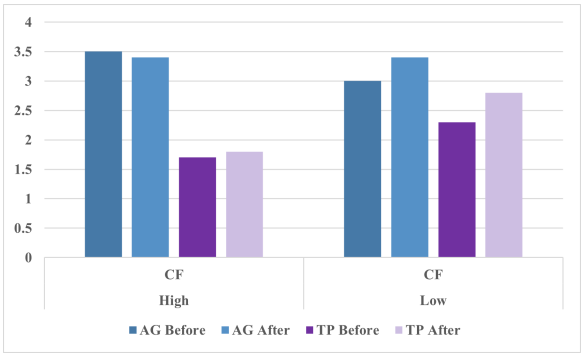


Figure 7.7: Before, After Task Performance and Agreement for Counterfactual-based Explanations in High and Low Accuracy group

For the Feature importance-based style (Figure 7.5), the data shows that when agreement decreases, task performance increases in both high and low-accuracy groups. In the Rule-based style (Figure 7.6), in the high-accuracy group, when agreement increased slightly, task performance remained stable. Conversely, in the low-accuracy group, when agreement remained consistent, task performance improved slightly.

However, for the Counterfactual style, the relationship varied depending on the perceived AI accuracy. The expected effect of agreement and task performance was observed in the high-accuracy group, when agreement lowered slightly, task performance increased. In contrast, in the low-accuracy group, when agreement increased, task performance increased together, suggesting that perceived low AI accuracy may alter the typical relationship between these metrics. This result indicates that perceived AI accuracy might influence the interaction between task performance and agreement differently depending on the explanation style.

7.1.4 RQ 3 - Obj. 4

The main research question of this thesis focused on how different explanation styles impact decision-making effectiveness, particularly when considering varying levels of perceived AI accuracy. As discussed in Sections 7.1.1, 7.1.2, and 7.1.3, perceived AI accuracy significantly influences decision-making, and distinct patterns and trends emerge between explanation styles when accounting for this factor.

In summary, Feature Importance and Counterfactual explanations were most effective in the objective measures of *task performance* and *agreement*, while Rule-based explanations excelled in *decision confidence*. Specifically, Counterfactual explanations led to increased task performance (H3) and

achieved the highest agreement rates (H5) among the different styles. In contrast, Feature Importance explanations resulted in a decrease in agreement (H6) when provided and had the lowest overall agreement rate (Tables 7.1, 7.2). Additionally, participants who received Rule-based explanations exhibited higher decision confidence than those who received Feature importance explanations (H8).

To explore the reasons behind these outcomes, we analyzed post-questionnaire responses (including ranking analysis, satisfaction, and difficulty), along with qualitative feedback, which addresses RQ 3 - Obj. 4.

Explanation Difficulty and Satisfaction influencing Effectiveness

Feature importance-based explanations were rated as the most difficult and received the lowest satisfaction scores, as shown in Table 7.3. This is also related to the result of the lowest decision confidence shown in the descriptive statistics in Table 7.1. The decrease in agreement (H6) and low decision confidence suggest that Feature importance explanations might not provide sufficient context or clarity to convince users of the AI's predictions. Participant feedback supported this, with many indicating a desire for more context to understand the relative values of data used in decision-making.

Counterfactual explanations, on the other hand, presented a different scenario. Although they were also rated as high in difficulty and low in satisfaction (Table 7.3), they had high persuasive power, demonstrated by the highest agreement rate and improved task performance. Despite being considered difficult and having low satisfaction, users were more likely to trust and agree with AI predictions when given detailed, scenario-based explanations that explore different outcomes.

Rule-based explanations did not significantly affect objective measures such as task performance and agreement. However, they were rated as the easiest to understand and received the highest satisfaction scores, which may explain the highest decision confidence observed in the descriptive statistics (Tables 7.1, 7.2). While Rule-based explanations clearly explain the logic of model predictions, we expected a positive impact on task performance and agreement, but this was not the case. This suggests that high satisfaction or simplicity does not necessarily translate to better objective effectiveness. However, higher satisfaction and lower difficulty were associated with higher decision confidence, which was evident in H8. The ranking analysis in Section 6.3 and Figure 6.2 also show that participants who received Rule-based explanations tended to rely on the most on AI explanations than other explanation groups.

These results suggest that balancing complexity and satisfaction is cru-

cial, as it involves comparing the need for more information with user satisfaction. This balance can be achieved by incorporating feedback from participants, which could potentially enhance both effectiveness and satisfaction, for example by incorporating hierarchies in Counterfactual-based explanations or more detailed feature explanations for Rule-based explanations or giving more contextual information on Feature importance-based explanations.

Background Influencing Effectiveness

We also anticipated different results based on participants' backgrounds, given that the experiment scenario was based on the process domain and related to XAI. However, as mentioned in the Results chapter, there were no significant differences overall based on background. This is related to the limitations that will be addressed in the next section regarding skewed backgrounds, which might obscure the reasons for effectiveness.

However, among other backgrounds, task performance and agreement seemed to be influenced by the level of experience in process mining. Participants having a process mining background showed a slightly more noticeable positive impact, particularly those with less than one year of experience, as shown in Table 6.15. As mentioned in the Results chapter, this might be due to the first batch distribution where we tested with students enrolled in a Process Mining course, where they can interpret process maps. Additionally, participants with process mining experience slightly reduced agreement rates (Tables 6.14, 6.15). This might indicate a more critical assessment of AI predictions by those with process mining experience.

7.2 Limitations

One of the primary limitations of this study is the difficulty participants faced in solving process related tasks without a background in process mining. The aim of this thesis was to investigate the effectiveness of different XAI explanation styles in the context of PPM by conducting a human-grounded experiment using simplified tasks and lay participants. We selected participants with business or IT backgrounds, anticipating that this would help them solve the process-related tasks in the scenarios. However, despite the simplified design and detailed descriptions, many participants expressed difficulty in completing the tasks without prior process mining knowledge.

This challenge is evident in our analysis, which revealed that participants with a background in process mining, particularly those from the first batch enrolled in a Process Mining course, performed better than others. This

suggests that the tasks may have been too complex for participants without a process mining background, potentially skewing the results. Furthermore, the ranking analysis in Section 6.3 showed that most participants relied the least on the process map. While this was anticipated, given the relative simplicity of other information, providing a more accessible PPM context could have helped to generalize the results more effectively within the PPM domain.

Therefore, one of the limitations is that the results may not be fully generalizable within the PPM context, and the participant pool may have been skewed toward those with process mining experience.

Another limitation is the choice of machine learning models used in the study. While our experiment aimed to compare different XAI techniques using a black-box model, we opted for Random Forest due to its simplicity and configurability as a black-box model. However, using more complex models like neural networks or deep learning could have provided more accurate predictions and insights into how different XAI styles affect decision-making.

A closely related limitation is the use of XAI techniques that are not specifically designed for event log data. In this experiment, we employed XAI techniques commonly used for tabular data. However, event log data have unique characteristics, including timestamps, activities, and various attributes. Using XAI techniques developed specifically for event log data could have generated more process-related explanations and closer approximations to real-life PPM scenarios.

7.3 Future Works

This thesis has provided insights into how different explanation styles impact user decision-making in the context of Predictive Process Monitoring (PPM), considering the level of perceived accuracy. However, several areas for future research could expand our understanding and address the limitations outlined in the previous section.

Firstly, as mentioned in the limitations, this thesis involved lay participants with business and IT backgrounds, which potentially resulted in skewed results for those with process mining backgrounds. For human-grounded research, more simplified process-related tasks could encourage greater user engagement and allow participants to approach tasks more proactively. This, in turn, would enable a more critical comparison between performance with and without explanations. Additionally, involving domain experts and using real-life scenarios through application-grounded experiments could provide a more comprehensive understanding of explanation styles' effectiveness in

PPM. This approach would allow for a more realistic assessment of explanation effectiveness in complex, real-world PPM scenarios. Domain experts can offer valuable insights into the challenges and requirements of decision-making within their specific domain.

Furthermore, this study primarily focused on outcome-oriented prediction tasks. Future research can explore the impact of different explanation styles on "next activity prediction" tasks, where the goal is to predict the subsequent step in a process based on the current stage. This would provide a deeper understanding of how explanations can support decision-making throughout the process life cycle.

Thirdly, while this thesis utilized Random Forest as the black-box model for producing post-hoc XAI explanations, future research could explore the impact of different model complexities on explanation effectiveness. As mentioned in the limitations, employing more complex algorithms like neural networks or deep learning models might generate different prediction patterns and subsequently necessitate distinct explanation strategies.

This thesis focused on Rule-based, Feature importance, and Counterfactual explanations, which were selected from model-based, attribution-based, and example-based XAI explanation types. To further enhance comparisons within the PPM context, using XAI techniques specifically developed for event log data could generate more process-related explanations and provide closer approximations within the PPM context.

Lastly, a potential addition to the explanation styles comparison could be human language-generated explanations using Large Language Models (LLMs), which are becoming increasingly popular for their ability to produce easily interpretable text without requiring specialized background knowledge. This aligns with feedback received in the open-ended questions, where some participants expressed difficulty in understanding visual-based explanations alone. Incorporating LLM-based XAI explanations could help address this issue and allow for a comprehensive analysis of the difficulty in understanding XAI, as well as its effectiveness in decision-making.

Chapter 8

Conclusion

The field of Predictive Process Monitoring (PPM) is currently characterized by a lack of user evaluations in XAI, leaving the effectiveness of these XAI explanations in supporting users' decision-making processes unclear. This thesis aimed to evaluate the impact of Feature importance-based, Rule-based, and Counterfactual-based explanations on decision-making within the PPM domain. To further explore this, we introduced the condition of perceived AI accuracy to better categorize decision-making situations based on prediction accuracy. The effectiveness of these explanation styles was measured using three different metrics: task performance, agreement, and decision confidence. The evaluation was conducted in a decision-making context where participants determined whether they agreed with the AI's prediction of accepting or rejecting a loan application.

The results indicated that perceived AI accuracy significantly affects decision-making, with lower accuracy leading to higher task performance across all explanation styles. This suggests that perceived accuracy plays a crucial role in shaping user engagement and trust in AI systems. Additionally, Counterfactual explanations were particularly effective in improving task performance and agreement, where we believe that scenario-based explanations might enhance user engagement and understanding. Conversely, Feature importance explanations showed the least agreement among all explanation styles. Rule-based explanation styles received the highest satisfaction and lowest difficulty ratings, resulting in the highest decision confidence, particularly compared to Feature importance, which was rated the lowest in satisfaction and highest in difficulty. These findings suggest that there is a relationship between satisfaction and effectiveness, which is not always linear. Instead, lower satisfaction and higher difficulty appear to correlate with more significant differences in effectiveness. This indicates that tailoring explanation styles to user preferences and needs may enhance overall effectiveness.

Building on this result, future work may focus on application-grounded evaluations of these explanation styles with domain experts. Additionally, it can also incorporate more sophisticated black-box models or advanced XAI techniques for improved performance. Additionally, new explanation styles, including human language-based explanations using Large Language Models (LLMs), could be explored. Future studies could also consider using "next activity prediction" tasks to provide insights that are more relevant to the process life cycle.

Bibliography

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In Edith Elkind, Manuela Veloso, Noa Agmon, and Matthew E. Taylor, editors, *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[2] Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):1059–1086, 2020.

[3] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115, 2020.

[4] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting blackbox models via model extraction. *CoRR*, abs/1705.08504, 2017.

[5] Liese Blevi, Lucie Delporte, and Julie Robbrecht. Process mining on the loan application process of a dutch financial institute. *BPI Challenge*, pages 328–343, 2017.

[6] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[7] Dominic Breuker, Martin Matzner, Patrick Delfmann, and Jörg Becker. Comprehensible predictive models for business processes. *MIS Q.*, 40(4):1009–1034, 2016.

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In Fabio Paternò, Nuria Oliver, Cristina Conati, Lucio Davide Spano, and Nava Tintarev, editors, *IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020*, pages 454–464. ACM, 2020.

[9] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. Effects of AI and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Trans. Interact. Intell. Syst.*, 13(4):22:1–22:42, 2023.

[10] Julia Cecil, Eva Lermer, Matthias FC Hudecek, Jan Sauer, and Susanne Gaube. Explainability does not mitigate the negative impact of incorrect ai advice in a personnel selection task. *Scientific reports*, 14(1):9736, 2024.

[11] Lenart Celar and Ruth MJ Byrne. How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition*, 51(7):1481–1496, 2023.

[12] Michael Chromik and Martin Schuessler. A taxonomy for human subject evaluation of black-box explanations in XAI. In Alison Smith-Renner, Styliani Kleanthous, Brian Y. Lim, Tsvi Kuflik, Simone Stumpf, Jahna Otterbacher, Advait Sarkar, Casey Dugan, and Avital Shulner-Tal, editors, *Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies co-located with 25th International Conference on Intelligent User Interfaces (IUI 2020), Cagliari, Italy, March 17, 2020*, volume 2582 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2020.

[13] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[14] Ghada Elkhawaga, Omar M Elzeki, Mervat Abu-Elkheir, and Manfred Reichert. Why should i trust your explanation? an evaluation approach for xai methods applied to predictive process monitoring results. *IEEE Transactions on Artificial Intelligence*, 2024.

[15] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[16] Chiara Di Francescomarino and Chiara Ghidini. Predictive process monitoring. In Wil M. P. van der Aalst and Josep Carmona, editors, *Process Mining Handbook*, volume 448 of *Lecture Notes in Business Information Processing*, pages 320–346. Springer, 2022.

[17] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[18] Riccardo Galanti, Bernat Coma-Puig, Massimiliano de Leoni, Josep Carmona, and Nicolò Navarin. Explainable predictive process monitoring. In Boudewijn F. van Dongen, Marco Montali, and Moe Thandar Wynn, editors, *2nd International Conference on Process Mining, ICPM 2020, Padua, Italy, October 4-9, 2020*, pages 1–8. IEEE, 2020.

[19] Riccardo Galanti, Massimiliano de Leoni, Merylin Monaro, Nicolò Navarin, Alan Marazzi, Brigida Di Stasi, and Stéphanie Maldera. An explainable decision support system for predictive process analytics. *Eng. Appl. Artif. Intell.*, 120:105904, 2023.

[20] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[21] Pascal Hamm, Michael Klesel, Patricia Coberger, and Hermann Felix Wittmann. Explanation matters: An experimental study on explainable AI. *Electron. Mark.*, 33(1):17, 2023.

[22] Maximilian Harl, Sven Weinzierl, Matthias Stierle, and Martin Matzner. Explainable predictive business process monitoring using gated graph neural networks. *J. Decis. Syst.*, 29(Supplement):312–327, 2020.

[23] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable AI systems for the medical domain? *CoRR*, abs/1712.09923, 2017.

[24] Chihcheng Hsieh, Catarina Moreira, and Chun Ouyang. Dice4el: Interpreting process predictions using a milestone-aware counterfactual approach. In Claudio Di Ciccio, Chiara Di Francescomarino, and Pnina Soffer, editors, *3rd International Conference on Process Mining, ICPM 2021, Eindhoven, The Netherlands, October 31 - Nov. 4, 2021*, pages 88–95. IEEE, 2021.

[25] Tsung-Hao Huang, Andreas Metzger, and Klaus Pohl. Counterfactual explanations for predictive business process monitoring. *CoRR*, abs/2202.12018, 2022.

[26] Olusanmi Hundogan, Xixi Lu, Yupei Du, and Hajo A. Reijers. CREATED: generating viable counterfactual sequences for predictive process analytics. In Marta Indulska, Iris Reinhartz-Berger, Carlos Cetina, and Oscar Pastor, editors, *Advanced Information Systems Engineering - 35th International Conference, CAiSE 2023, Zaragoza, Spain, June 12-16, 2023, Proceedings*, volume 13901 of *Lecture Notes in Computer Science*, pages 541–557. Springer, 2023.

[27] Eoin M. Kenny, Courtney Ford, Molly S. Quinn, and Mark T. Keane. Explaining black-box classifiers using *post-hoc* explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artif. Intell.*, 294:103459, 2021.

[28] Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. Examples are not enough, learn to criticize! criticism for interpretability. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2280–2288, 2016.

[29] Suhwan Lee, Marco Comuzzi, and Nahyun Kwon. Exploring the suitability of rule-based classification to provide interpretability in outcome-based process predictive monitoring. *Algorithms*, 15(6):187, 2022.

[30] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *CoRR*, abs/1802.03888, 2018.

[31] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[32] Fabrizio Maria Maggi, Chiara Di Francescomarino, Marlon Dumas, and Chiara Ghidini. Predictive monitoring of business processes. In Matthias Jarke, John Mylopoulos, Christoph Quix, Colette Rolland, Yannis Manolopoulos, Haralambos Mouratidis, and Jennifer Horkoff,

editors, *Advanced Information Systems Engineering - 26th International Conference, CAiSE 2014, Thessaloniki, Greece, June 16-20, 2014. Proceedings*, volume 8484 of *Lecture Notes in Computer Science*, pages 457–472. Springer, 2014.

[33] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics*, 113:103655, 2021.

[34] Nijat Mehdiyev and Peter Fettke. Prescriptive process analytics with deep learning and explainable artificial intelligence. In Frantz Rowe, Redouane El Amrani, Moez Limayem, Sue Newell, Nancy Pouloudi, Eric van Heck, and Ali El Quammah, editors, *28th European Conference on Information Systems - Liberty, Equality, and Fraternity in a Digitizing World, ECIS 2020, Marrakech, Morocco, June 15-17, 2020*, 2020.

[35] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.

[36] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 607–617. ACM, 2020.

[37] Mahsan Nourani, Chiradeep Roy, Jeremy E. Block, Donald R. Honeycutt, Tahrima Rahman, Eric D. Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In Tracy Hammond, Katrien Verbert, Dennis Parra, Bart P. Knijnenburg, John O'Donovan, and Paul Teale, editors, *IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021*, pages 340–350. ACM, 2021.

[38] Vincenzo Pasquadibisceglie, Giovanna Castellano, Annalisa Appice, and Donato Malerba. FOX: a neuro-fuzzy model for process outcome prediction and explanation. In Claudio Di Ciccio, Chiara Di Francescomarino, and Pnina Soffer, editors, *3rd International Conference on Process Mining, ICPM 2021, Eindhoven, The Netherlands, October 31 - Nov. 4, 2021*, pages 112–119. IEEE, 2021.

[39] Elizaveta Povalyaeva, Ismail Khamitov, and Artyom Fomenko. Bpic 2017: density analysis of the interaction with clients. *BPI Challenge*, 2017, 2017.

[40] Jana-Rebecca Rehse, Nijat Mehdiyev, and Peter Fettke. Towards explainable process predictions for industry 4.0 in the dfki-smart-lego-factory. *Künstliche Intell.*, 33(2):181–187, 2019.

[41] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[42] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1527–1535. AAAI Press, 2018.

[43] Williams Rizzi, Marco Comuzzi, Chiara Di Francescomarino, Chiara Ghidini, Suhwan Lee, Fabrizio Maria Maggi, and Alexander Nolte. Explainable predictive process monitoring: A user evaluation. *CoRR*, abs/2202.07760, 2022.

[44] Williams Rizzi, Chiara Di Francescomarino, and Fabrizio Maria Maggi. Explainability in predictive process monitoring: When understanding helps improving. In Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas, editors, *Business Process Management Forum - BPM Forum 2020, Seville, Spain, September 13-18, 2020, Proceedings*, volume 392 of *Lecture Notes in Business Information Processing*, pages 141–158. Springer, 2020.

[45] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[46] Arik Senderovich, Chiara Di Francescomarino, Chiara Ghidini, Kerwin Jorbina, and Fabrizio Maria Maggi. Intra and inter-case features in predictive process monitoring: A tale of two dimensions. In *Business Process Management: 15th International Conference, BPM 2017, Barcelona, Spain, September 10–15, 2017, Proceedings 15*, pages 306–323. Springer, 2017.

[47] Renuka Sindhgatta, Catarina Moreira, Chun Ouyang, and Alistair Barros. Exploring interpretable predictive models for business processes. In Dirk Fahland, Chiara Ghidini, Jörg Becker, and Marlon Dumas, editors, *Business Process Management - 18th International Conference, BPM 2020, Seville, Spain, September 13-18, 2020, Proceedings*, volume 12168 of *Lecture Notes in Computer Science*, pages 257–272. Springer, 2020.

[48] Renuka Sindhgatta, Chun Ouyang, Catarina Moreira, and Yi Liao. Interpreting predictive process monitoring benchmarks. *CoRR*, abs/1912.10558, 2019.

[49] Alexander Stevens and Johannes De Smedt. Explainability in process outcome prediction: Guidelines to obtain interpretable and faithful models. *European Journal of Operational Research*, 2023.

[50] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In Jason Furman, Gary E. Marchant, Huw Price, and Francesca Rossi, editors, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 303–310. ACM, 2018.

[51] Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi. Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Trans. Knowl. Discov. Data*, 13(2):17:1–17:57, 2019.

[52] Jasper van der Waa, Elisabeth Nieuwburg, Anita H. M. Cremers, and Mark A. Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artif. Intell.*, 291:103404, 2021.

[53] Boudewijn van Dongen. Bpi challenge 2012, 2012.

[54] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Evaluating fidelity of explainable methods for predictive process analytics. In Selmin Nurcan and Axel Korthaus, editors, *Intelligent Information Systems - CAiSE Forum 2021, Melbourne, VIC,*

Australia, June 28 - July 2, 2021, Proceedings, volume 424 of *Lecture Notes in Business Information Processing*, pages 64–72. Springer, 2021.

[55] Mythreyi Velmurugan, Chun Ouyang, Catarina Moreira, and Renuka Sindhgatta. Evaluating stability of post-hoc explanations for business process predictions. In Hakim Hacid, Odej Kao, Massimo Mecella, Naouel Moha, and Hye-young Paik, editors, *Service-Oriented Computing - 19th International Conference, ICSOC 2021, Virtual Event, November 22-25, 2021, Proceedings*, volume 13121 of *Lecture Notes in Computer Science*, pages 49–64. Springer, 2021.

[56] Ilya Verenich, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, and Irene Teinemaa. Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. *ACM Trans. Intell. Syst. Technol.*, 10(4):34:1–34:34, 2019.

[57] Ilya Verenich, Marlon Dumas, Marcello La Rosa, and Hoang Nguyen. Predicting process performance: A white-box approach based on process models. *J. Softw. Evol. Process.*, 31(6), 2019.

[58] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *CoRR*, abs/2006.00093, 2020.

[59] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. Statistical stability indices for lime: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society*, 73(1):91–101, 2022.

[60] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[61] Greta Warren. Counterfactual explanations for explainable ai (xai). In *ICCBR Doctoral Consortium*, pages 34–39, 2022.

[62] Sven Weinzierl, Sandra Zilker, Jens Brunk, Kate Revoredo, Martin Matzner, and Jörg Becker. XNAP: making lstm-based next activity predictions explainable by using LRP. In Adela del-Río-Ortega, Henrik Leopold, and Flávia Maria Santoro, editors, *Business Process Management Workshops - BPM 2020 International Workshops, Seville, Spain, September 13-18, 2020, Revised Selected Papers*, volume 397 of *Lecture Notes in Business Information Processing*, pages 129–141. Springer, 2020.

[63] Bemali Wickramanayake, Zhipeng He, Chun Ouyang, Catarina Moreira, Yue Xu, and Renuka Sindhgatta. Building interpretable models for business process prediction using shared and specialised attention mechanisms. *Knowl. Based Syst.*, 248:108773, 2022.

[64] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 295–305. ACM, 2020.

[65] Jianlong Zhou, Syed Z. Arshad, Kun Yu, and Fang Chen. Correlation for user confidence in predictive decision making. In Henry B. L. Duh, Christopher Lueg, Mark Billinghurst, and Weidong Huang, editors, *Proceedings of the 28th Australian Conference on Computer-Human Interaction, OzCHI 2016, Launceston, Tasmania, Australia, November 29 - December 2, 2016*, pages 252–256. ACM, 2016.

[66] Jianlong Zhou, Constant Bridon, Fang Chen, Ahmad Khawaji, and Yang Wang. Be informed and be involved: Effects of uncertainty and correlation on user's confidence in decision making. In Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo, editors, *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, Seoul, CHI 2015 Extended Abstracts, Republic of Korea, April 18 - 23, 2015*, pages 923–928. ACM, 2015.

[67] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

[68] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. Effects of influence on user trust in predictive decision making. In Regan L. Mandryk, Stephen A. Brewster, Mark Hancock, Geraldine Fitzpatrick, Anna L. Cox, Vassilis Kostakos, and Mark Perry, editors, *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*. ACM, 2019.

Appendix A

Survey Materials

A.1 Extra Description for Scenarios

A.1.1 Simple Version

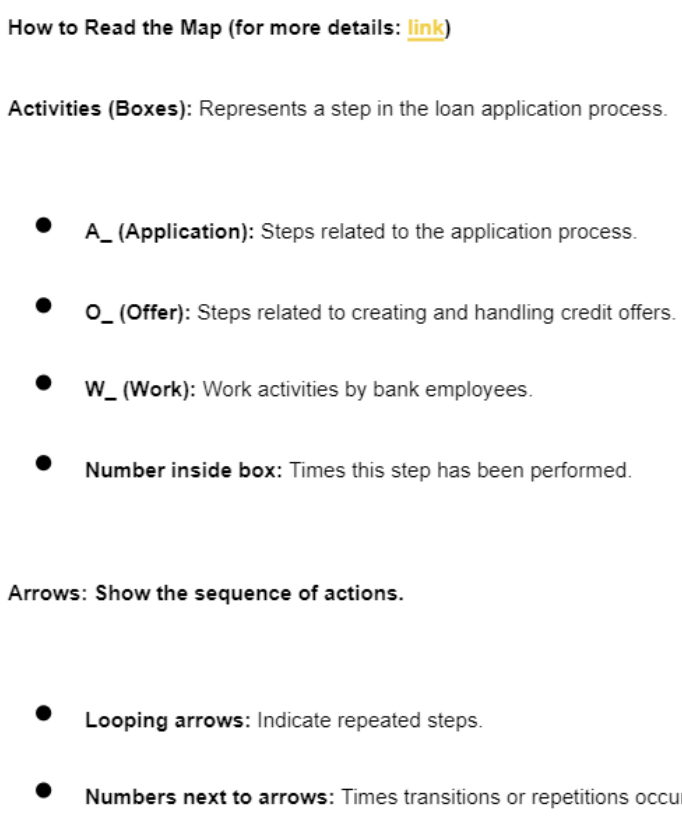
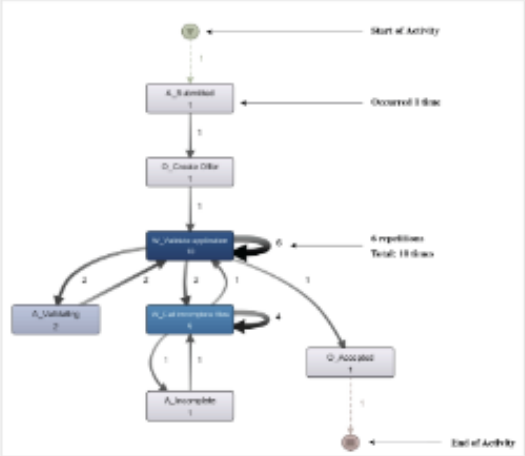


Figure A.1: Additional Information for Process Maps(Simple version)

A.1.2 Detailed Version

Reading Process Map (example)



Activities (Boxes): Represents a step in the loan application process.

- **A_ (Application):** Steps related to the application process.
- **O_ (Offer):** Steps related to creating and handling credit offers.
- **W_ (Work):** Work activities by bank employees.
- **Number inside box:** Times this step has been performed.

Arrows: Show the sequence of actions.

- **Looping arrows:** Indicate repeated steps.
- **Numbers next to arrows:** Times transitions or repetitions occur.

<Example>

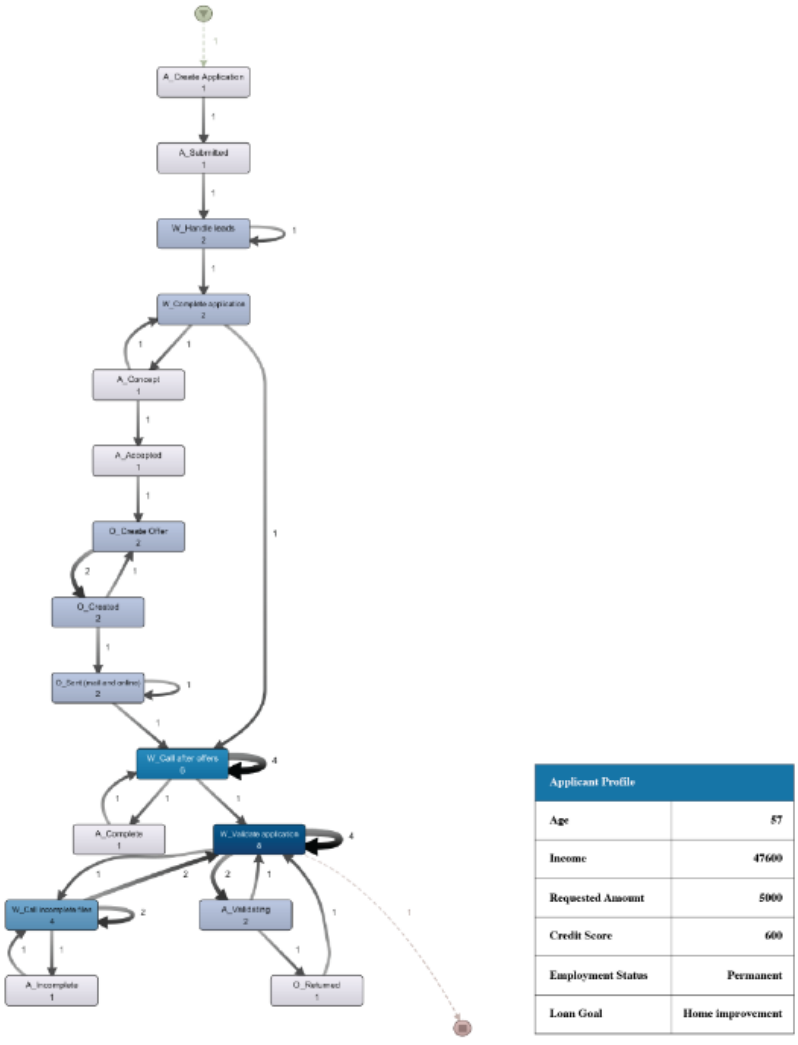
1. **A_Submitted:** The application is submitted.
2. **O_Create Offer:** A credit offer is created.
3. **W_Validate application:** The application is validated. After the first W_Validate application step, it can loop back up to 6 times to ensure all documents are correct and completed.

Activity	Description	Activity	Description
A_Create Application	The process of starting a loan application.	**W_Call after offers**	The bank calls the customer after sending the loan offer.
A_Submitted	The loan application has been submitted by the customer.	**A_Complete**	The loan offer and required documents have been sent to the customer.
W_Handle leads	The initial review of the application by the bank.	**W_Validate application**	The bank validates the received documents.
W_Complete application	The bank employee finishes processing the application.	**A_Validating**	The bank checks and verifies the received documents. (State changes)
A_Concept	First check by the bank, often involving a call to the customer.	**W_Call incomplete files**	The bank calls the customer to resend missing documents.
A_Accepted	The application has passed initial checks and a loan offer can be made.	**A_Incomplete**	The document verification failed; the customer needs to resend documents.
O_Create Offer	Creating a loan offer for the customer.	**O_Returned**	The customer has submitted documents for the selected loan offer.
O_Created	The loan offer has been created.	**A_Validating**	The bank checks and verifies the received documents. (State changes)
O_Sent (mail and online)	The loan offer has been sent to the customer by mail and online.		

Figure A.2: Additional Information for Process Maps (Link version)

A.2 Scenarios

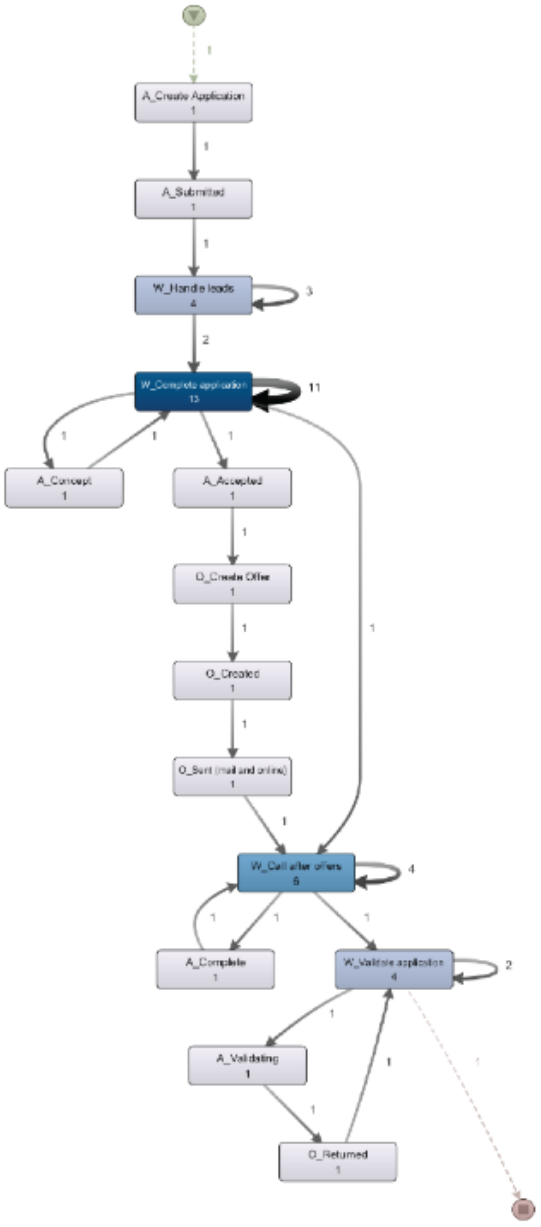
- **Starting Point**: The process begins at "**A_Create Application**," where the applicant submits their application.
- **Current Status**: The last step was "**W_Validate Application**," where the bank checks the documents received from the applicant.
- **Duration:** This process has taken around **8 days.**



Applicant Profile	
Age	57
Income	47600
Requested Amount	5000
Credit Score	600
Employment Status	Permanent
Loan Goal	Home improvement

Figure A.3: Applicant 1 (Correctly Predicted: Accept)

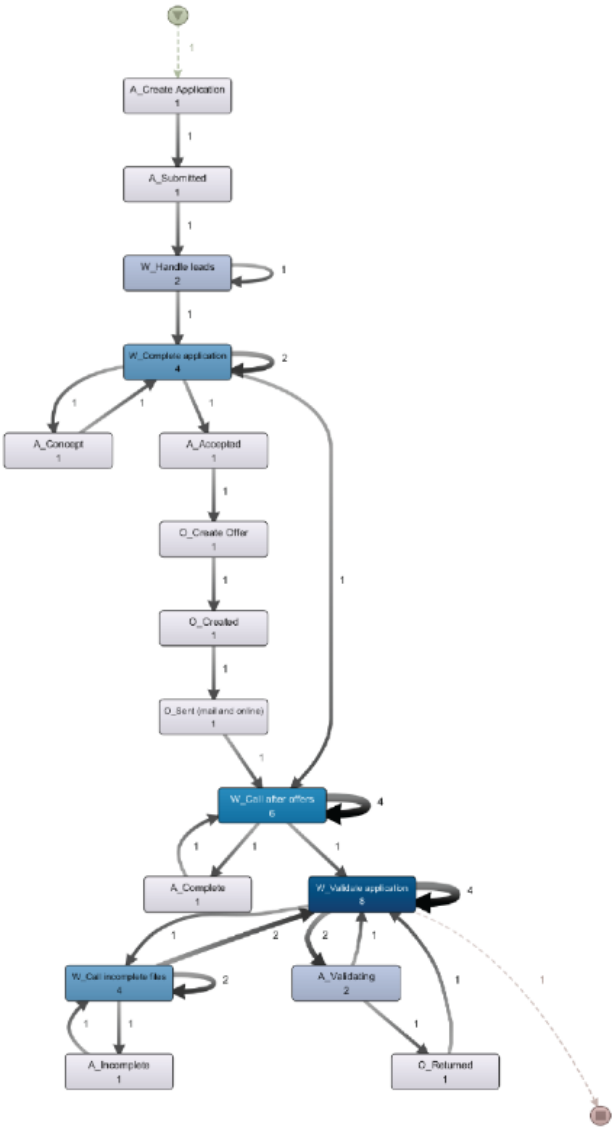
- **Starting Point**: A_Create Application
- **Current Status**: W_Validate Application
- **Duration:** 15 days



Applicant Profile	
Age	23
Income	24800
Requested Amount	15000
Credit Score	650
Employment Status	Fixed-Term
Loan Goal	Home improvement

Figure A.4: Applicant 2 (Correctly Predicted: Reject)

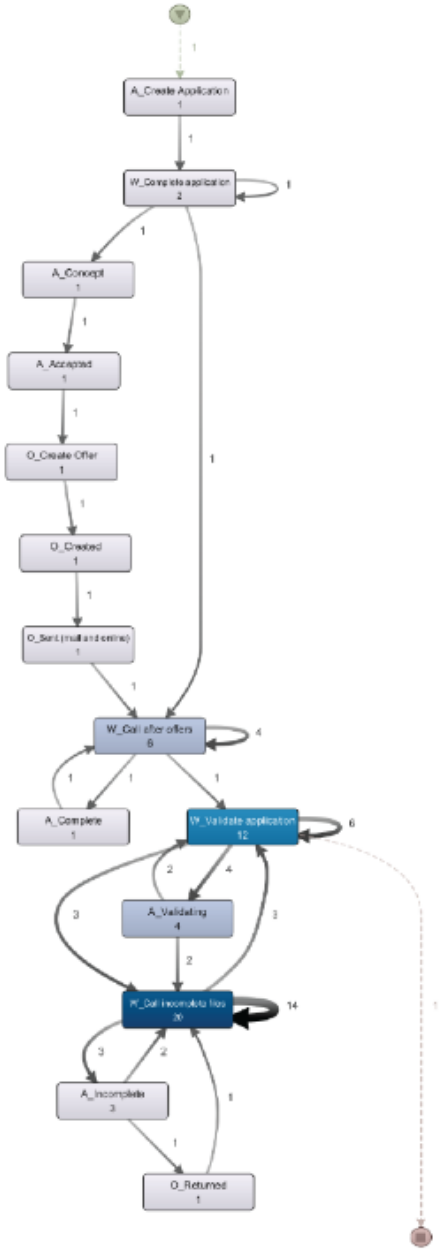
- **Starting Point**: A_Create Application
- **Current Status**: W_Validate Application
- **Duration:** 37 days



Applicant Profile	
Age	27
Income	44400
Requested Amount	42500
Credit Score	700
Employment Status	Permanent
Loan Goal	Existing loan takeover

Figure A.5: Applicant 3 (Misclassified: Reject)

- **Starting Point**: A_Create Application
- **Current Status**: W_Validate Application
- **Duration:** 25 days



Applicant Profile	
Age	55
Income	34500
Requested Amount	5000
Credit Score	680
Employment Status	Fixed Term
Loan Goal	Extra spending limit

Figure A.6: Applicant 4 (Misclassified: Accept)

A.3 XAI Explanations

A.3.1 Feature Importance-based Explanation

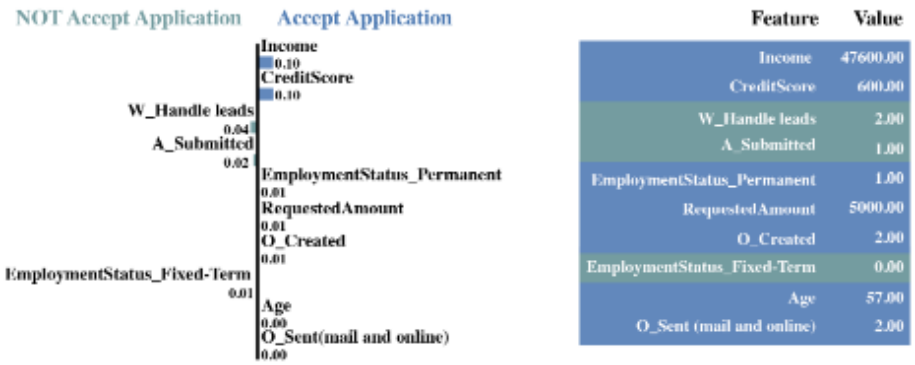


Figure A.7: Feature Importance Explanation for Applicant 1

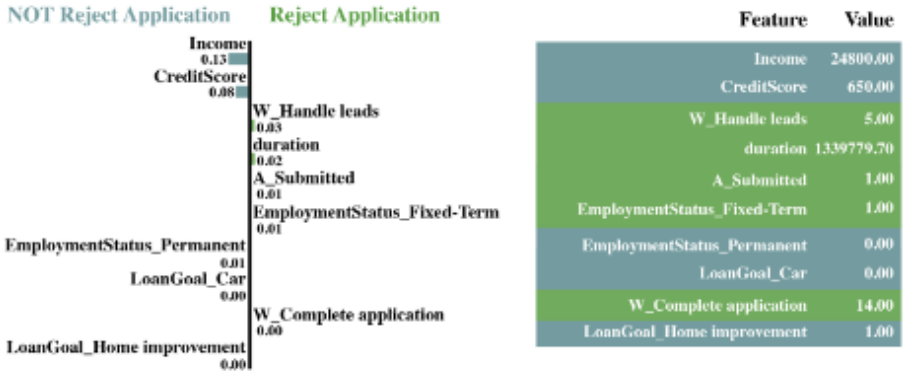


Figure A.8: Feature Importance Explanation for Applicant 2

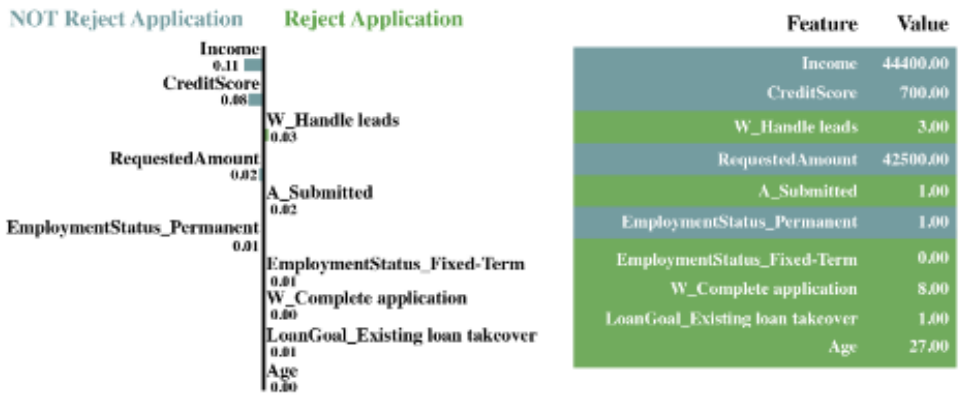


Figure A.9: Feature Importance Explanation for Applicant 3

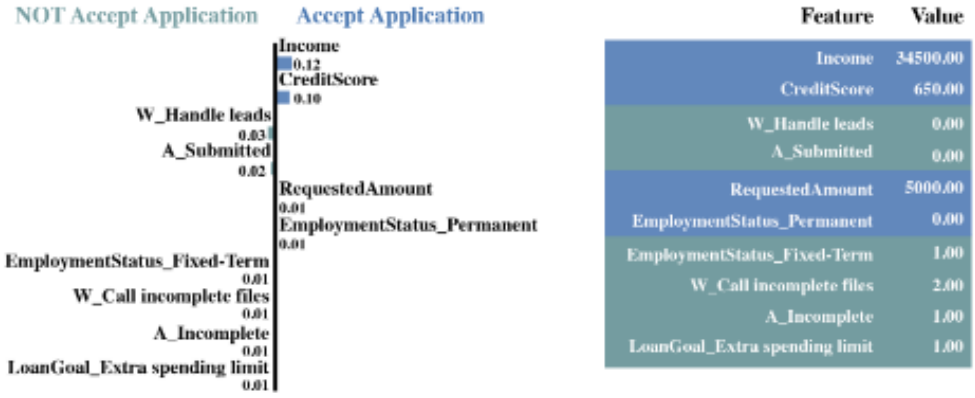


Figure A.10: Feature Importance Explanation for Applicant 4

A.3.2 Rule-based Explanation

Figure A.11: Rule-based Explanation for Applicant 1

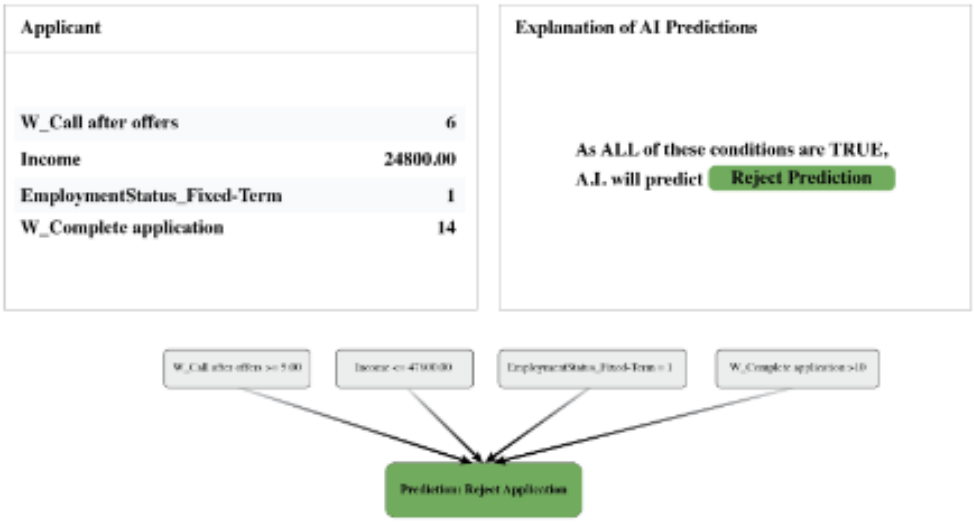


Figure A.12: Rule-based Explanation for Applicant 2

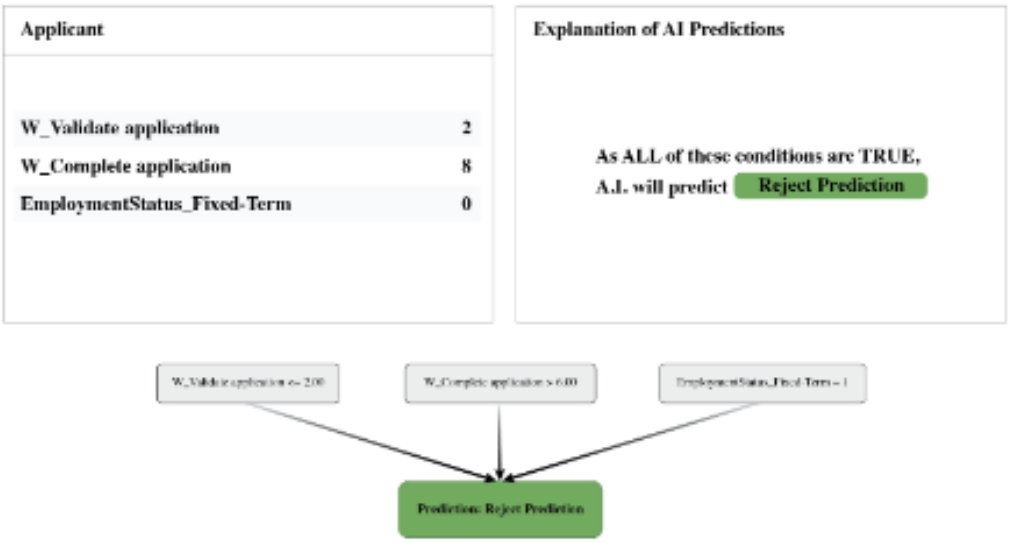


Figure A.13: Rule-based Explanation for Applicant 3

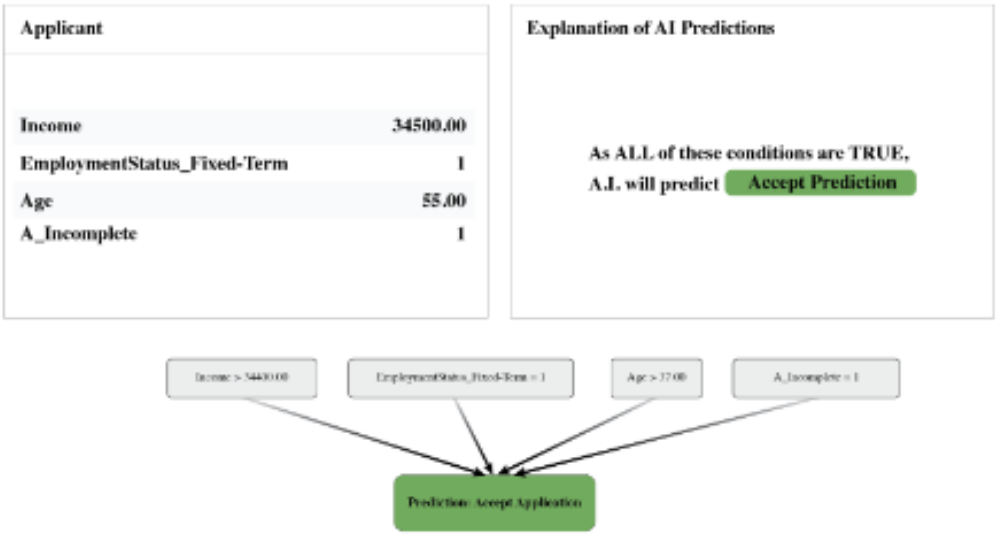


Figure A.14: Rule-based Explanation for Applicant 4

A.3.3 Counterfactual-based Explanation

	O_Sent (mail and online)	W_Validate application	W_Call after offers	W_Call incomplete files	Requested Amount	Income	Credit Score	Age	Loan Goal	Outcome
Original Input	2	2	6	0	5000	47600	600	57	Home Improvement	→ Accept
Different Input	-	5	-	-	-	-	-	-	Unknown	→ Reject

Figure A.15: Counterfactual Explanation for Applicant 1

	W_Complete application	W_Call after offers	W_Validate application	W_Call incomplete files	Requested Amount	Income	Credit Score	Age	Loan Goal	Outcome
Original Input	14	6	2	0	15000	24800	650	23	Home improvement	→ Reject
Different Input	-	-	1	-	-	-	849	-	Tax Payments	→ Accept

Figure A.16: Counterfactual Explanation for Applicant 2

	W_Complete application	W_Call after offers	W_Validate application	W_Call incomplete files	Requested Amount	Income	Credit Score	Age	Loan Goal	Outcome
Original Input	8	6	2	0	42500	44400	700	27	Existing loan takeover	→ Reject
Different Input 1	-	-	-	-	-	-	800	-	-	→ Accept

Figure A.17: Counterfactual Explanation for Applicant 3

	O_Sent (mail and online)	W_Validate application	W_Call after offers	W_Call incomplete files	Requested Amount	Income	Credit Score	Age	Loan Goal	Outcome
Original Input	1	3	6	2	5000	34500	650	55	Extra spending limit	→ Accept
Different Input	-	-	-	-	-	-	-	-	Tax Payments	→ Reject

Figure A.18: Counterfactual Explanation for Applicant 4

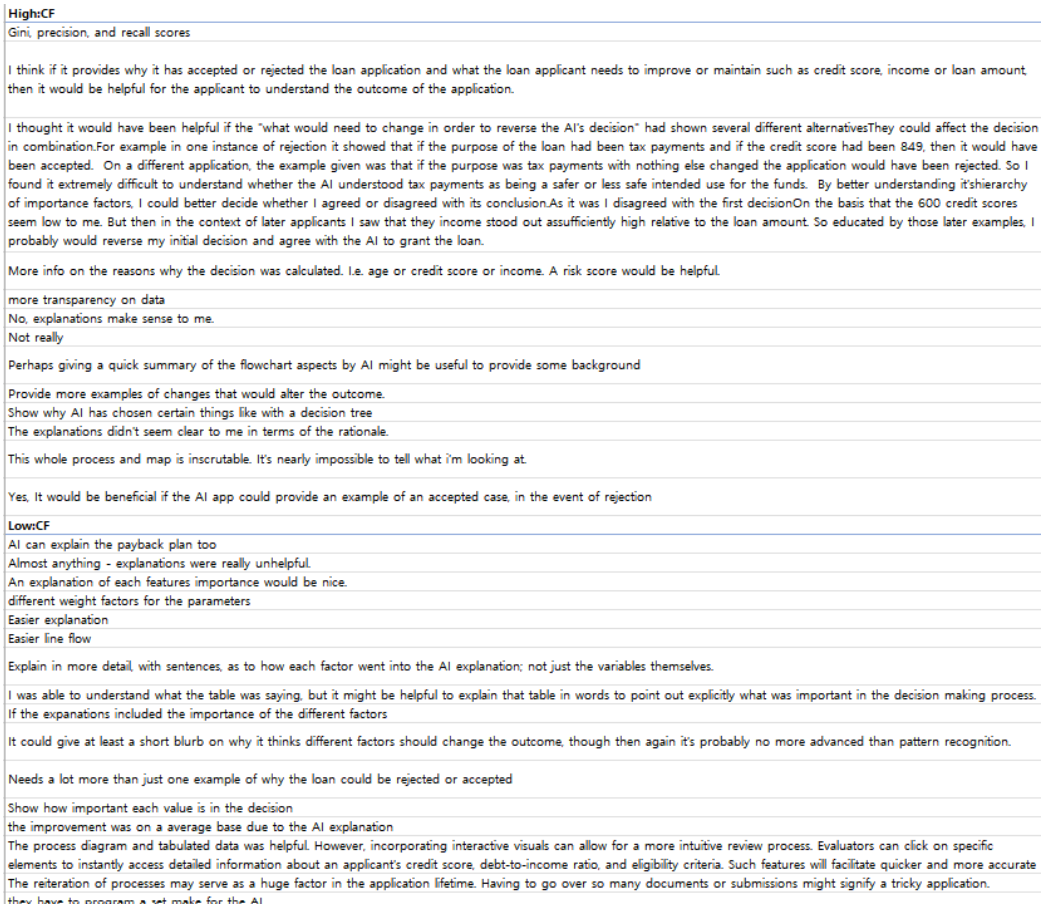
A.4 Feedbacks about XAI Explanations

High:FeatureImportance
Continuous feedback and improvement, transparency in AI-driven decisions,
honestly that process map was hard to understand or even look at, that was the only difficult part
I think information should be expressed using process chats often to improve AI Explanation understandability.
I would like to know what the AI system thinks about the relationship between income and the requested amount. I believe these two are interrelated, but after reading the explanations, it seems the AI system disagrees with this statement. So, I would like to see an explanation of what the AI system thinks about this matter.
It's important to take into account the duration of fixed term contract and also the duration in which the person has to give back the money in order to better evaluate what to do
Perhaps a more simplistic diagram with easier jargon.
Regarding to the total count for each node, perhaps it helps if the size of the total count is elaborated a bit more. For example, why is it a good/bad feature if the total count is big/small.
Short descriptions on why duration or other variables can potentially add decimals points to sway a decision in either way.
Simplify process map
the percentage gap between the income and loan amount should also be considered.
There is no scale for applicants feature. I wonder whether credit score 600 is high or low
they should improve in the information display
use of an LLM to provide clear English explanation
Low:FeatureImportance
additional explanation as to what is in the hard-to-read flow chart
Explain if a combination of factors has a certain influence. Example: a certain income to loan request ratio is always accepted, but depending on a certain factor (e.g., fixed/permanent job) this will differ
Explain why some factors have a positive influence (on reject/accept) and explain why some have a negative influence
I think the AI should also look at the percentage of the yearly income somebody request and take this into the score
Make it clearer in laymans terms for a reader who may not be familiar with loan applications
Make it more readable instead of showing statistics and charts
More details about why the loan was accepted or rejected would make it easier to understand.
Perhaps the weighting each reason holds
strategies to arrive at possible answers should be programmed on it
The AI explanation was not consise across the models, it was very confusing.
The AI should explain how risk preferences are set, that is, how the bank's risk control model determines its tolerance for risk.
The steps are systemic
would be easier to understand if the AI used full sentences, not just keywords

Figure A.19: Feature Importance-based Explanation Feedbacks

High:Rule
additinal details for rejection
By adding the probability score of the applicant.
Colour coding
documentation verification
explain what the rankings and relative values of data used in decision making is
Explains why AI set this threshold, for example, why set 30,000 as the threshold at income, and above 30,000 will be one of the factor to be accepted?
good enough
I guess some more motivation concerning why some treshold have been chosen, but i guess that is more to clarify why that decision might be the right one
I'm not a domain expert, so the effect of a variable (positive indicator for acceptance or negative), and a scale (like credit score 600-800) would be more clarifying
it could use natural language to explain the decision to make it easier to understand
It should at least have more explanations as to why
Make the explanations morg human friendly.
Maybe an explanation of fixed term versus permanent employment could be helpful
Maybe take into account the proportions of the loan they need and how much they earn.
More user friendly interface to understand how the decision was made.
Low:Rule
1. Simplify language: Use clear and concise language, avoiding technical jargon and complex financial terms. 2. Visual aids: Incorporate diagrams, flowcharts, or infographics to illustrate the decision-making process and key factors influencing loan approval. 3. Concrete examples: Provide hypothetical scenarios or real-life examples to demonstrate how the AI model evaluates loan applications. 4. Feature attribution: Explain which specific features (e.g., credit score, income, employment history) contribute most to the loan decision. 5. Transparency in data usage: Clearly explain what data is used, how it's collected, and how it's processed.
A deeper layer for understanding why the value of for instance 47.000€ income is enough for a loan?
A key should exist below the AI Explanation that spells out any abbreviations, especially when there are underscores or other symbols present.
An explanation of how the AI works would be nice. I don't know why the evaluation criteria seem to change for every applicant.
change the layout and the font size
Explain meaning of variables in more detail
For the accept case, it is ACCEPTED when all conditions hold true. For the REJECTED case, i am still unsure whether all conditions should hold true, or if one of the conditions is enough.
Length of time for repayment of the loan
Make it more like a actual programming flowchart
Maybe include flow charts (visual aids) that map out decision making process step by step. Additionally I'd probably want to see more information provided like similiar cases to see and understand rationale better.
maybe use a universal set of data points in the explanation
more detail in the explantion when relating to profile of candidate
Refund plans and duration would be an improvement
Seems like it didn't take much into account about the applicant's profile.
should be more interactive
Show how duration factors in
The AI can provide more detailed explanation about the loan application
We see the names of the features that contributed to the final decision of the model, but it is not always clear what these features entailed e.g. A_incomplete=1 was a reason for the acceptance of one of the applications, which sounds like because an application was incomplete it was more likely to be accepted, which sounds intuitively wrong.
well i dont think there should be any improvement because it all look good
What weight has each items (variable) when application makes the decisions.

Figure A.20: Rule-based Explanation Feedbacks



High:CF
Gini, precision, and recall scores
I think if it provides why it has accepted or rejected the loan application and what the loan applicant needs to improve or maintain such as credit score, income or loan amount, then it would be helpful for the applicant to understand the outcome of the application.
I thought it would have been helpful if the "what would need to change in order to reverse the AI's decision" had shown several different alternativesThey could affect the decision in combination.For example in one instance of rejection it showed that if the purpose of the loan had been tax payments and if the credit score had been 849, then it would have been accepted. On a different application, the example given was that if the purpose was tax payments with nothing else changed the application would have been rejected. So I found it extremely difficult to understand whether the AI understood tax payments as being a safer or less safe intended use for the funds. By better understanding it'shierarchy of importance factors, I could better decide whether I agreed or disagreed with its conclusion.As it was I disagreed with the first decisionOn the basis that the 600 credit scores seem low to me. But then in the context of later applicants I saw that they income stood out assufficiently high relative to the loan amount. So educated by those later examples, I probably would reverse my initial decision and agree with the AI to grant the loan.
More info on the reasons why the decision was calculated. I.e. age or credit score or income. A risk score would be helpful.
more transparency on data
No, explanations make sense to me.
Not really
Perhaps giving a quick summary of the flowchart aspects by AI might be useful to provide some background
Provide more examples of changes that would alter the outcome.
Show why AI has chosen certain things like with a decision tree
The explanations didn't seem clear to me in terms of the rationale.
This whole process and map is inscrutable. It's nearly impossible to tell what i'm looking at.
Yes, It would be beneficial if the AI app could provide an example of an accepted case, in the event of rejection
Low:CF
AI can explain the payback plan too
Almost anything - explanations were really unhelpful.
An explanation of each features importance would be nice.
different weight factors for the parameters
Easier explanation
Easier line flow
Explain in more detail, with sentences, as to how each factor went into the AI explanation; not just the variables themselves.
I was able to understand what the table was saying, but it might be helpful to explain that table in words to point out explicitly what was important in the decision making process.
If the expanations included the importance of the different factors
It could give at least a short blurb on why it thinks different factors should change the outcome, though then again it's probably no more advanced than pattern recognition.
Needs a lot more than just one example of why the loan could be rejected or accepted
Show how important each value is in the decision
the improvement was on a average base due to the AI explanation
The process diagram and tabulated data was helpful. However, incorporating interactive visuals can allow for a more intuitive review process. Evaluators can click on specific elements to instantly access detailed information about an applicant's credit score, debt-to-income ratio, and eligibility criteria. Such features will facilitate quicker and more accurate
The reiteration of processes may serve as a huge factor in the application lifetime. Having to go over so many documents or submissions might signify a tricky application.
they have to program a set make for the AI

Figure A.21: Counterfactual-based Explanation Feedbacks