PRE-PRINT: INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS (2023) 000–000

Human-AI Collaboration to Mitigate Decision Noise in Financial Underwriting: A Study on FinTech Innovation in a Lending Firm

ARTICLE INFO

Keywords:
FinTech
Lending
Underwriting
Artificial Intelligence
Explainable AI

ABSTRACT

Financial institutions have recognized the value of collaborating human expertise and AI to create high-performance augmented decision-support systems. Stakeholders at lending firms have increasingly acknowledged that plugging data into AI algorithms and eliminating the role of human underwriters by automation, with the expectation of immediate returns on investment from business process automation, is a flawed strategy. This research emphasizes the necessity of auditing the consistency of decisions (or professional judgment) made by human underwriters and monitoring the ability of data to capture the lending policies of a firm to lay a strong foundation for a legitimate system before investing millions in AI projects. The judgments made by experts in the past re-emerge in the future as outcomes or labels in the data used to train and evaluate algorithms. This paper presents Evidential Reasoning-eXplainer, a methodology to estimate probability mass as an extent of support for a given decision on a loan application by jointly assessing multiple independent and conflicting pieces of evidence. It quantifies variability in past decisions by comparing the subjective judgments of underwriters during manual financial underwriting with outcomes estimated from data. The consistency analysis improves decision quality by bridging the gap between past inconsistent decisions and desired ultimate-true decisions. A case study on a specialist lending firm demonstrates the strategic work plan adapted to configure underwriters and developers to capture the correct data and audit the quality of decisions.

1. Introduction

1.1 Background on Human-AI Collaboration for Augmented Decision-Making

Lending firms are actively seeking the application of artificial intelligence (AI) in financial technology (FinTech) to enhance the efficiency of loan processing and deliver responsible lending decisions to meet the urgent financial needs of their customers in a digitally dominated landscape (Kowalewski & Pisany, 2022). It can be achieved by a symbiotic merger of humans and AI, aiming to harness their respective strengths and prepare for the forthcoming wave of AI innovation (Wilson & Daugherty, 2019). The recent technological

leap in generative AI tools, such as OpenAI's Generative Pre-trained Transformer 4 (GPT-4) and Google Bard, has demonstrated a remarkable capability to reason, plan, and learn from experience at or above human-level capabilities in solving intricate and novel tasks across a broad range of domains (Akter, et al., 2023) (Dwivedi, et al., 2023). However, a comprehensive study conducted by Microsoft Research on GPT-4 uncovered inherent flaws: contradictory outputs of inputs within similar contexts and a lack of sufficient explanation for the generated decision (Bubeck, et al., 2023). These limitations coupled with the challenge of quantifying the reliability of explanations due to the black-box nature of generative AI, impede its application in high-stake financial decision-making tasks.

Studies proposed on black-box AI algorithms for lenders to provide credit decisions include deep neural networks (Luo, et al., 2017) (Zhao, et al., 2015) (Wu & Wang, 2000), ensemble classifiers (Xu, et al., 2018) (Abellán & Castellano, 2017), support vector machine (Harris, 2015) (Tomczak & Zięba, 2015), Bayesian networks (Anderson, 2019), (Leong, 2016), and genetic programming (Metawa, et al., 2017). In addition, three studies have introduced interpretable lending models. An interpretable knowledge-based system for retail and commercial loans is proposed to provide interpretable lending decisions through the activation weight of lending rules and attribute contributions (Sachan, 2022). Another study introduced the concept of globally-consistent rules to summarize a broad pattern of the classifier to provide local explanation and measure the quality of a decision by cardinality and quantity of data support for a given rule (Chen, et al., 2022). A hierarchical belief rule-based system is proposed to process factual and heuristic knowledge of financial underwriters by incorporating the collective knowledge of humans and data to provide reasoning behind a loan funding decision (Sachan, et al., 2020). This research highlights the importance of configuring the intelligence of human experts in transparent AI algorithmic systems to establish augmented decision-making processes in financial institutions.

The shortcomings of AI automation in advancing FinTech innovation within banking, investments, and microfinance can be traced back to the use of nonrepresentative data, inherent biases in the sampled data, choices in algorithmic methodologies, and human judgments influenced by their interpretations of AI outcomes (Ashta & Herrmann, 2021). Financial institutions have acknowledged that rushing into the opportunity of reshaping businesses by plugging data into AI algorithms with the prospect of immediate returns on investment from business process automation is a poor strategy (Fountaine, et al., 2019). In contrast, hybrid human-AI intelligence can mutually achieve a high-performance decision-support system (Dellermann, et al., 2019). However, the approach for creating an active learning AI system, which entails constant evaluation of the model output and assimilation of human-elicited knowledge is not well articulated by AI system developers to domain experts and organizational stakeholders. The disconnection between developers (or data scientists) and domain experts poses a considerable challenge in addressing the multifaceted concerns (Bellomarini, et al., 2022).

A study identified two primary concerns in AI deployment in FinTech startups. First, the adoption of AI for customer and employee support presents complexities, as technology demands appropriate resources and is found to be ineffective in solving specialized problems. Second, the integration of AI has led to a decline in

employee morale over concerns regarding the reduction in workforce requirements (Almansour, 2023). Human experts, such as financial underwriters, have a skeptical perspective on AI due to the difficulty of standardizing their subject matter expertise and cognitive thinking within a machine. This distrust drives them back to their familiar, unassisted manual tasks. It is widely recognized that machines may not surpass humans in applying heuristic knowledge to complex tasks, and humans may not necessarily excel over machines in executing repetitive and monotonous tasks (Glikson & Woolley, 2020). Automation does not eliminate the role of human underwriters (Fuster, et al., 2019). Therefore, augmenting lending decision tasks by configuring human-AI as an integrated unit would produce a high-performance system.

1.2 Role of Human Judgment in Generating Biased and Noisy AI Decisions

The performance of data-driven AI methods depends on involving meaningful features in the data (Rudin, 2019). Gathering high-quality features is challenging due to inevitable ambiguities resulting from missing information, a dynamic decision task environment (continuously evolving lending policies and regulations), and a lack of understanding of the data fed into the algorithm (Sachan, et al., 2021).

One significant source of ambiguity predominantly disregarded by many firms is the inconsistency in decisions or subjective judgment given by human experts, which gets recorded as labels in the dataset utilized to train AI algorithms (Kahneman, et al., 2016). The professionals often contradict their decision and deviate from their peers, especially when they stray from their expertise to intuition (Kahneman, D; Klein, G, 2009). Additionally, the quality of a decision is affected by the complexity of the task at hand. More complex tasks demand more time. The investment of time reflects the decision maker's confidence in their cognitive decision (Boundy-Singer, et al., 2023). The prevalence of variability in expert judgment is observed in many domains, such as medicine (Litvinova, et al., 2019) (Levi, 1989) (Koran, 1975), psychology (Garb, 1996), finance (Kahneman, et al., 2016), weather (Lusk, et al., 1990) (Stewart, et al., 1989), and human data annotators (Shan, et al., 2021). An experimental study assessed the quality of expert judgment by two metrics: learnability and ecological validity in various domains (Bolger & Wright, 1994). It concluded that the consistency of decisions given by bankers, weather forecasters, and research and development managers is better than that by professionals in other domains, such as clinical psychologists, physicians, and audit managers.

Financial underwriters rely on their cognitive abilities and heuristic knowledge to make decisions based on the ambiguous information in the documents submitted within a loan application pack. Human judgment is known to be noisy and biased. Consequently, an AI algorithm trained on past human judgments or manually annotated data might generate biased or noisy decisions. Judgment errors can be categorized into noise and biases. Bias in judgment can be attributed to social preconceptions, such as stereotyping of minorities, or cognitive biases, such as overconfidence and unfounded optimism (Kahneman, D., 2011). Noise is the scattered judgment which is not attributable to social or cognitive biases. For example, an AI algorithm would consistently produce the same decision, denoted as θ, for different borrowers with similar lending characteristics in their loan application. In contrast, humans could make varying decisions - θ,θ', and θ'' for different borrowers despite identical lending characteristics. AI decision-support systems are considered superior to humans in executing simple repetitive tasks. However, they are falling short in replicating human

judgments due to the inherent flaws in the data employed to train these systems. Figure 1 presents a visual example of extremely biased, noisy, and optimally accurate data space based on two features (2-dimensional) to estimate the decision to "fund" or "reject" loan applications.

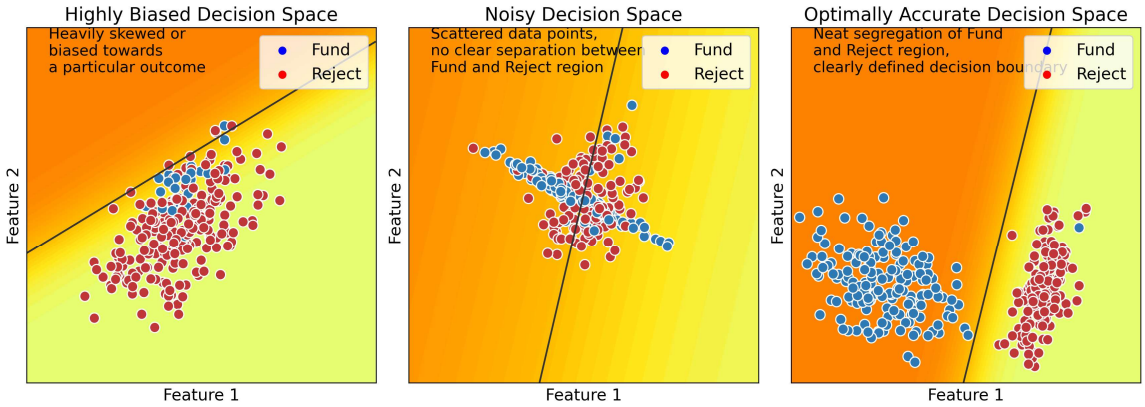

Figure 1: 2-dimensional decision space exhibiting highly biased, noisy, and optimally accurate decisions

1.3 Contribution

Consistency of loan underwriters' judgment is critical for a lending firm to spend time and money to build an augmented AI system by integrating an extensive dataset and allocating significant resources for tasks involving human intervention. An underwriter provides a decision by following the policy and regulations in a manual underwriting guideline. A financial institution's pre-defined rules and criteria revolve around credit history and affordability criteria (Sachan, et al., 2020). Lenders fall broadly into three categories: banks, FinTech firms, and non-bank lenders (Fuster, et al., 2019). Mainstream banks have strict lending criteria that enable them to provide AI-driven algorithmic decisions for many applications based on stringent rules. Some lending institutions follow a common-sense lending approach for the financial inclusion of borrowers rejected by mainstream banks due to less-than-perfect credit history. However, these firms face the challenge of inconsistent decisions and high processing time due to manual analyses of a considerable volume of dynamic information by underwriters (Peterson, 2017). These lending firms require a strategic initiative to leverage AI for digital FinTech transformation to offer efficient decisions without compromising their fundamental business values.

There is no question that the adoption of AI for the FinTech transformation of a lending firm can be accomplished primarily by utilizing new technologies and algorithms to expand their capacity for data-driven analysis. Most research addresses only the algorithmic approach. They failed to address the issues of data ambiguity and ways to control the variability in decisions by incorporating domain experts' opinions before initiating the development of an augmented AI system. This paper proposes an initial work plan to configure domain experts and developers to capture the correct data and assess the quality of decisions to get the most from AI implementation. It presents a technique to map ambiguous data from multiple sources to lending rules to maximize data coverage for the lending policy and measurement of decision inconsistency. The Evidential Reasoning-eXplainer (ER-X) is proposed to estimate the probability mass as the extent of support for a decision to "fund" or "reject" a loan application by jointly assessing multiple independent and highly

conflicting pieces of evidence. A piece of evidence is a criterion in a lending rule mapped with uncertain credit data. ER-X provides explainable decisions by weight and reliability of evidence.

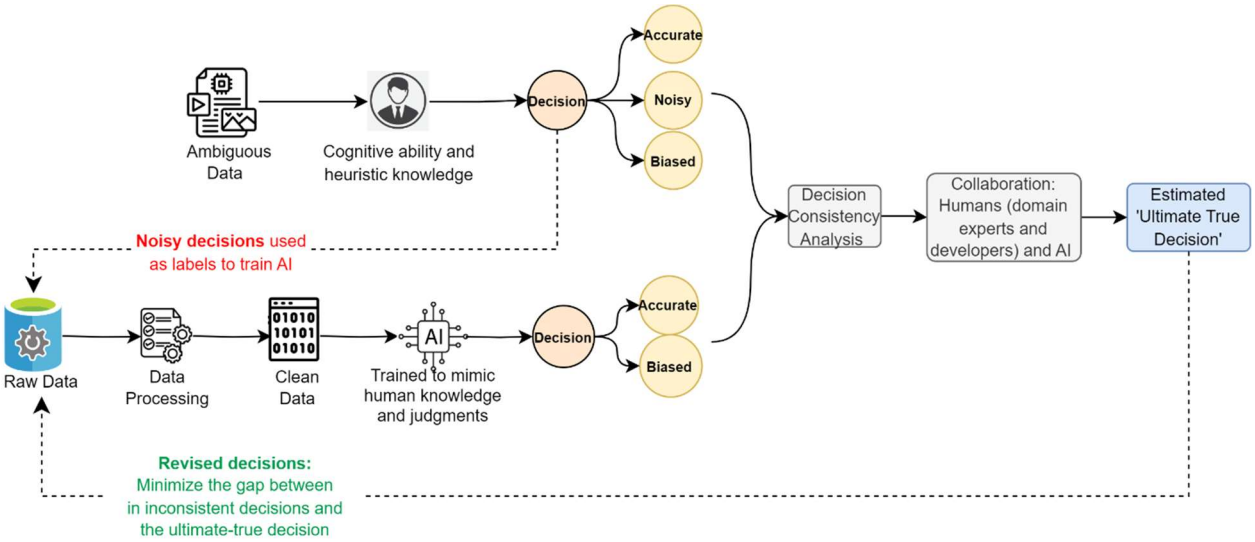

Figure 2: Humans and AI collaboration to reduce judgment noise

ER-X can infer from the missing values in numerical and categorical features in the data (missing pieces of evidence) without any requirement of data imputation. It introduces the concept of the "ultimate-true decision", a benchmark used to audit past human expert assessments that are later represented as labels in the data used to develop AI systems and to adjust the algorithm's output to enhance performance. The "ultimate-true decision" is derived by evaluating the variability between decisions estimated from the data, as a joint probability mass of multiple criteria in lending rules, and decisions provided by human experts in the form of subjective judgments. The subjective judgments of human experts, such as underwriters, are obtained through knowledge elicitation tasks. Figure 2 presents the main contribution of the paper as the process of reducing and detecting the judgment noise by consistency analysis to produce trustworthy decisions from an eXplainable AI (XAI) based decision-support system.

This study demonstrates a practical application of the proposed framework through a case study from a UK-based lending firm. The lending firm implemented this framework to investigate the decision consistency between data and human underwriters to explore their potential to streamline their lending process through the integration of AI, thereby marking their evolution into a FinTech firm. This study emphasizes the strategic task of establishing a connection between human underwriters and developers to lay a strong foundation for a trustworthy and high-performance augmented AI decision-support system. This research addresses the following questions:

Question 1: How to map uncertain data from multiple sources with lending rules?

Question 2: How to refine the decision quality by minimizing the gap between past-inconsistent decisions and the "ultimate-true decision"?

Question 3: How to detect variability in decisions?

Question 4: Are human underwriters inconsistent in their decision-making?

***Question 5**:* What types of collaborative tasks human experts (underwriters) and developers can perform to establish an augmented-AI decision-support system?

This paper is organized as follows: Section 2 presents the literature review on the evidential reasoning approach. Section 3 describes the methodology to map uncertain data from multiple sources with lending rules outlined in the policy and ER-X framework to conduct consistency analysis to audit the assessment given by human experts (underwriters). It addresses Questions 1 to 3. Section 4 attempts to answer Questions 4 and 5 through a case study conducted on a lending firm. This firm was interested in exploring its potential to adopt AI to enhance lending decisions, and our study involved evaluating the quality of captured data and the consistency of underwriters' decisions for the most frequently funded and rejected loan applications. Section 5 discusses the results, and Section 6 concludes the paper. Appendix A defines the joint evidence, acronyms of lending criteria in the case study, and descriptive statistics of the data. Appendix B presents the accuracy metrics of the ER-X model and its comparison with deep learning, the decision tree model, and logistic regression.

2. Literature Review on Evidential Reasoning for Decision Analysis

ER-X is based on the Conjunctive-Maximum Likelihood Evidential Reasoning (C-MAKER) rule. The C-MAKER framework was proposed to infer from the ambiguous categorical data (missing values in input and output features) (Sachan, et al., 2021) (Liu, et al., 2019). It can preprocess the categorical features in data by numerical transformation and fusion of multiple pieces of evidence to reduce the cardinality of various categories for different machine learning models such as deep-learning, tree-based, and rule-based. It conceptualizes the notion of weight, i.e. the importance of evidence, and reliability, i.e. potential of evidence to point correctly to a decision for each subset of propositions in a power set of a frame of discernment. The frame of discernment is a mutually exclusive and collectively exhaustive set of expected outcomes (or decisions). The concept of weight and reliability of a piece of evidence in C-MAKER was an extension of the ER rule (Yang & Xu, 2013). However, it does not consider each proposition's distinct weight and reliability in a power set of a frame of discernment. The C-MAKER and ER rules have evolved from the Dempster-Shafer (DS) theory. The DS theory is a benchmark for information fusion and decision-making (Shafer, 1976). However, it provides counter-intuitive results when it is implemented to combine highly conflicting evidence. The rational probabilistic reasoning process of ER-X is rigorously compared to other evidence fusion rules, such as the proportional conflict redistribution rule (Smarandache, et al., 2010), Dempster's rule (Dempster, 2008), Smets' rule (Smets & Kennes, 1994), Dubois and Prade's rule (Dubois & Prade, 1988), and Yager's rule (Yager, 1987). These methods combine multiple pieces of evidence but do not compare the outcome of the same evidence from numerous sources. An approach to compare dissimilarity between probability for each proposition in a power set of a frame of discernment from pieces of evidence collected from multiple sources was presented by Yong (Yong, et al., 2004).

3. Methodology to Reduce Noisy Decisions by Evidential Reasoning-eXplainer (ER-X)

3.1 Mapping of Uncertain Credit Data with Lending Rules

A set of lending rules is formulated to capture the core aspects of the lending policy of a firm. Let, a lending institution has $J, j \in \{1, \dots, J\}$ number of standard lending rules. These rules mandate underwriters to pursue lending policies that enable them to provide individualized decisions for each loan application. Each rule has a set of criteria that contribute toward the rejection or acceptance of a loan. A criterion within a rule can represent a fact or a piece of heuristic information (Feigenbaum, 1980). For instance, consider a rule that assesses the "number of secured loan defaults for the last 24 months" has the following set of criteria: {"0_arrears", "1-2_arrears", "3+_arrears"}. In this set, "0_arrears" represents zero arrears of a borrower, "1-2_arrears" indicates one or two arrears, and "3+_arrears" denotes more than three arrears. Suppose the lending policy explicitly states the rejection of borrowers with more than three arrears, representing a clear-cut decline criterion based on factual information. On the other hand, the "1-2_arrears" criterion represents a heuristic piece of information because it does not point to a strict outcome. Underwriters jointly examine the criterion of multiple rules to provide a final decision. In ER-X, each criterion in a rule is treated as an independent piece of evidence.

Raw Data ($\mathcal{D}$)

Instances	Features					Outcome
	A_1	⋯	A_q	⋯	A_Q	$P(\Theta)$
1	$A_{1,v}$		$A_{q,v}$		$missing$	y
⋮	⋮		⋮		⋮	⋮
i	$missing$	⋯	$A_{q,v}$	⋯	$A_{Q,v}$	y
⋮	⋮		⋮		⋮	⋮
I	$A_{1,v}$		$A_{q,v}$		$A_{Q,v}$	y

Aggregated dataset ($\bar{\mathcal{D}}$) Mapped to Set of Lending Rule

Instances	Features mapped to each rule					Outcome
	e_1	⋯	e_j	⋯	e_J	$P(\Theta)$
1	$x_{1,1}$		$x_{1,j}$		$x_{1,J}$	y
⋮	⋮		⋮		⋮	⋮
i	$x_{i,1}$	⋯	$x_{i,j}$	⋯	$x_{i,J}$	y
⋮	⋮		⋮		⋮	⋮
I	$x_{I,1}$		$x_{I,j}$		$x_{I,J}$	y

Vertical coverage: sufficient instances for minority and majority cases, and decision

Horizontal coverage: Comprehensive data capture for criteria in a rule

Figure 3: Demonstrates a raw dataset (D) and an aggregated dataset ($\bar{\mathcal{D}}$)

- ***Definition of evidence:*** A piece of evidence refers to a criterion in a rule. A j^{th} rule has V_j number of criteria such that $v \in \{1, \dots, V_j\}$. A v^{th} evidence in a j^{th} rule is denoted by $e_{v,j}$. A set of evidence (or set of criteria) in a j^{th} rule can be represented as:

$$e_j = \{e_{v,j}, v \in \{1, \dots, V_j\}\}, \quad \forall j \in \{1, \dots, J\} \tag{1}$$

AI-enabled systems process borrowers' data integrated from various sources, such as credit bureaus, fraud intelligence, and digital loan applications, as a raw dataset (D). In Figure 3, the left table illustrates the structure of a raw dataset, where $Q, q \in \{1, \dots, Q\}$, represents the number of features (or columns), and $I, i \in \{1, \dots, I\}$ represents the number of instances (or rows). Each instance represents the data for a given borrower for a loan.

Merging multiple data sources into a single raw dataset can introduce ambiguities due to missing values in independent input features and missing labels in the output feature. Each instance in the captured historical dataset is expected to correspond to an outcome representing a past decision on a loan application made either

by a human underwriter or an AI algorithmic decision-support system deployed within a lending firm. Let N denote the total number of anticipated outcomes (or decisions) such that $n \in \{1, \dots, N\}$. The frame of discernment of a mutually exclusive and collectively exhaustive set of expected outcomes can be represented as:

$$\Theta = \{\theta_1, \dots, \theta_n, \dots, \theta_N,\ n \in \{1, \dots, N\}\} \tag{2.1}$$

In the context of lending decisions, the expected outcomes correspond to specific types of judgments made for loan applications. Expression (2.2) illustrates a set comprising two distinct lending decisions:

$$\Theta = \{\theta_1 = Fund, \theta_2 = Reject,\ n \in \{1,2\}\} \tag{2.2}$$

A power set of the frame of discernment is denoted by $P(\Theta)$. The cardinality of the power set, indicating the total number of subsets it contains, is denoted by $Y = 2^N$ such that $y \in \{1, \dots, Y\}$. Any subset indexed by y is a part of $P(\Theta)$, such that $y \in P(\Theta)$. The $P(\Theta)$ can be expressed as:

$$P(\Theta) = \{\emptyset, \{\theta_1\}, \dots, \{\theta_N\}, \{\theta_1, \theta_2\}, \dots, \{\theta_1, \theta_n\}, \dots, \{\theta_1, \dots, \theta_{N-1}\}, \Theta\} \tag{3}$$

Raw credit data have higher dimensionality than the cardinality of a set of lending rules, such that $Q \geq J$. Put simply, the datasets compiled from various sources have significantly more columns than the number of lending rules. Therefore, the Q number of columns in a raw dataset requires reshaping into J columns for each lending rule to process information for lending decisions. The right table in Figure 3 shows the aggregated dataset ($\bar{\mathcal{D}}$) extracted for a set of lending rules. It has $J, j \in \{1, \dots, J\}$ number of columns for each lending rule and $I, i \in \{1, \dots, I\}$ the number of rows for each borrower. A column for a rule is represented by e_j, where e denotes a piece of evidence and j denotes a rule. A single data point is referred as $x_{i,j}$. The dataset ($\bar{\mathcal{D}}$), aggregates information from various columns of the raw dataset and maps it into the lending rule-oriented dataset, $D^Q \rightarrow \bar{\mathcal{D}}^J$. Figure 4 illustrates selecting the data for a rule "secured loan worst status code in all addresses". The worst status code in row $i = 0$ is Default (DF), and $i = 2$ is Voluntary Repossession (VR). The data for each rule for i^{th} borrower is extracted from multiple columns of the raw dataset.

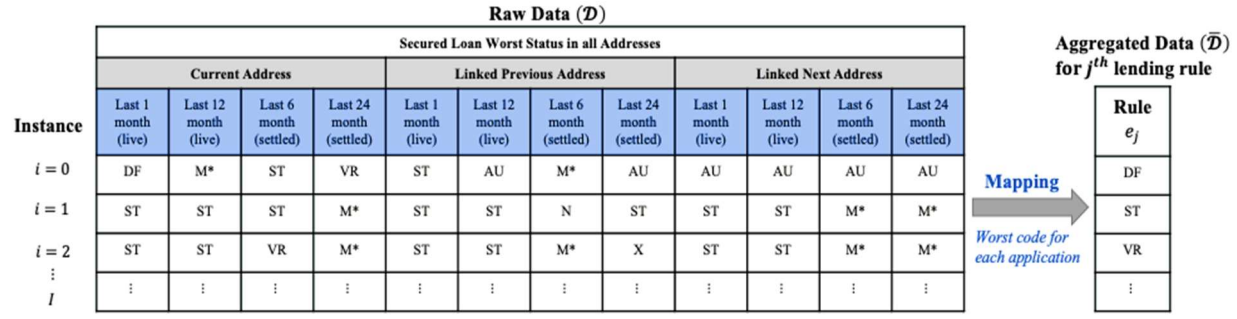

Figure 4: An example of aggregation of a raw dataset to map data to a lending rule. Acronym meanings can be referred to in Table 3, Appendix A.

Reliable data-driven algorithmic decisions require comprehensive vertical and horizontal data coverages. Horizontal coverage is the comprehensiveness of the data captured to ensure that every potential state or

condition stipulated by the criteria in the lending rules is represented. Exhaustive horizontal coverage can be achieved through the collaboration between developers and domain experts. The developers can easily pinpoint some straightforward columns in raw data for a given rule, and the rest of the columns can be identified with the assistance of domain experts. The vertical coverage is sufficient data support to represent the majority (most frequent type of borrowers) and minority cases (loan applications of rare kinds of borrowers) and consistency in outcome (past decisions or judgments). However, attaining exhaustive vertical coverage is typically unattainable because a dataset in a lending firm may only represent a limited group of borrowers, while other unconventional borrowers may be underrepresented or not represented at all.

3.2 ER-X

Preliminary definitions for ER-X:

- ***Definition of the probability mass**:* The degree of uncertainty for a piece of evidence is measured by its extent of support for a given subset of outcomes in the power set y, $y \in P(\Theta)$. It shows how strongly a piece of evidence supports y. A piece of evidence pointing to an outcome y is denoted by $e_{y,v,j}$, and its probability mass is denoted by $m_{y,v,j}$.

- ***Definition of the body of evidence (BOE):*** Focal elements are the non-zero probability masses of evidence for a given outcome within a power set $P(\Theta)$. The probability mass of each focal point can be obtained from multiple sources, such as estimation from data and judgment from domain experts, as their subjective degree of belief (Eriksson & Hájek, 2007). The BOE is the set of all focal elements. It can be represented as a distribution of probability mass as follows:

$$e_{v,j} = \{(e_{y,v,j}, m_{y,v,j}), \forall y \in P(\Theta), \textstyle\sum_{y \in P(\Theta)} m_{y,v,j} = 1\} \quad (4)$$

  where the pair $(e_{y,v,j}, m_{y,v,j})$ is the focal element of evidence if $m_{y,v,j} > 0$.

- ***Definition of the weight of evidence:*** The weight of evidence $(w_{y,v,j})$ is the importance of a piece of evidence in data, such that $0 \leq w_{y,v,j} \leq 1$.

- ***Definition of the reliability of evidence:*** The reliability of evidence $(r_{y,v,j})$ is a measure of how accurately the evidence indicates a subset of outcomes in the power set such that $0 \leq r_{y,v,j} \leq 1$.

- ***Definition of ultimate-true decision*:** The "ultimate-true decision" is a reference point for a piece of evidence against which an algorithmic decision and a subjective judgment are evaluated. The "ultimate-true decision" is estimated and validated by the knowledge elicitation of multiple human experts.

Data aggregated from various sources may suffer from limited vertical and horizontal coverage due to informational uncertainty and unforeseen uncertainty within a decision-task environment (Sachan, et al., 2021). Therefore, the BOE should not be a primitive estimation from the data. The subjective judgment of human experts can be incorporated to approximate the ultimate-true decision. Suppose, the BOE is collected from F number of sources such that $f \in \{1, \dots, F\}$. If multiple sources support a BOE, then the evidence is relatively

important in comparison to a highly conflicting BOE (Chen, et al., 2018). The BOE estimated from data is represented by f and acquired from domain experts f', such that $f, f' \in \{1, \dots, F\}$.

3.2.1 Estimate Probability Mass of Evidence from Data

3.2.1.1 Data Transformation by Similarity Distribution

(a) Numerical or Continous Data Transformation – Single Evidence

A sample $x_{i,j}$ in a dataset $(\overline{\mathcal{D}})$ for a j^{th} rule pointing to an outcome y can be represented as:

$$\left(x_{i,j}, y\right);\ \text{where } i \in \{1, \dots, I\}, j \in \{1, \dots, J\}, \text{and } y \in P(\Theta) \tag{5}$$

Each sample of a j^{th} rule is transformed into a similarity distribution across its set of criteria $e_j = \left\{e_{v,j} |\ v \in \{1, \dots, V_j\}\right\}$ to approximate the distribution between the sample and its recorded outcome. The transformed data $Z\left(x_{i,j}\right)$ can be represented in the following way:

$$Z\left(x_{i,j}\right) = \left\{\left(e_{v,j}, \alpha_{v,j}^{i}\right);\ v \in \left\{1, \dots, V_j\right\}, i \in \{1, \dots, I\}, \text{and } j \in \{1, \dots, J\}\right\} \tag{6}$$

In Expression (6), $\alpha_{v,j}^{i}$ is the degree of similarity to which a sample $x_{i,j}$ match to the values in a set of criteria of a j^{th} rule. The transformation of continuous data is summarised below:

$$\alpha_{v,j}^{i} = \begin{cases} \left(\alpha_{v,j}^{i} = \frac{e_{v+1,j} - x_{i,j}}{e_{v+1,j} - e_{v,j}}, \alpha_{v+1,j}^{i} = 1 - \alpha_{v,j}^{i}\right), & if \quad e_{v,j} \le x_{i,j} \le e_{v+1,j} \\ \left(\alpha_{v',j}^{i} = 0, \alpha_{v'+1,j}^{i} = 0\right), & otherwise \quad e_{v',j} \nleq x_{i,j} \nleq e_{v'+1,j} \end{cases} \tag{7}$$

In Expression (7), $v' \neq v$ and $v, v' \in \left\{1, \dots, V_j\right\}$.

(b) Categorical Data Transformation – Single Evidence

If the data for j^{th} rule has categorical values, then the value in its set of criteria would be categorical. A categorical sample $x_{i,j}$ would have 100% similarity to one of the categorical criteria. The transformation of categorical data is summarised below:

$$\alpha_{v,j}^{i} = \begin{cases} 1, & if \quad x_{i,j} = e_{v,j} \\ 0, & otherwise \quad x_{i,j} \neq e_{v',j} \end{cases} \tag{8}$$

In Expression (8), $v' \neq v$ and $v, v' \in \left\{1, \dots, V_j\right\}$. Table 1 illustrates the transformation process of numerical and categorical data samples based on a set of criteria in a rule. For instance, consider a rule related to "Credit Score" has numerical criteria as {"0","100","300","600"}. A data sample, denoted as $x_{i,j} = 253$, falls within the range of 100 to 300. This numerical data point was then mapped to two membership values: $\alpha_{2,j}^{i} = 0.235$ and $\alpha_{3,j}^{i} = 0.765$. Similarly, the rule regarding the "Number of Secured Loan Defaults in the Last 24 Months" has categorical criteria, such as {"0_arrears", "1-2_arrears", "3+_arrears"}. A given data sample for this rule is $x_{i,j} = \text{"}1 - 2\_\text{arrears"}$, which belongs exclusively to the "1-2_arrears" category in the criteria set.

Table 1: Example of the transformation of numerical and categorical data

Data Type	Rule	$x_{i,j}$	$Z(x_{i,j})$
Numerical	Credit Score	253	{("0", 0.0), ("100", 0.235), ("300", 0.765), ("600", 0.0)}
Categorical	Number of Secured Loan Defaults in the Last 24 Months	1-2_arrears	{("0_arrears", 0), ("1-2_arrears", 1), ("3+_arrears", 0)}

(c) Generate Joint Transformation Data - Joint Evidence

The previous section has shown the method to transform numerical or categorical data for an individual rule with single independent evidence. A joint evidence space is obtained by combining a single piece of evidence from different rules. The data for a set of rules can have mixed data, i.e. combined realization of numerical and categorical data type or either. Suppose, we want to combine v^{th} criteria in a j^{th} rule and v'^{th} criteria in a j'^{th} rule such that $v \in \{1, \dots, V_j\}$ and $v' \in \{1, \dots, V'_{j'}\}$, respectively. The joint degree of similarity ($\alpha^{i}_{v_j, v'_{j'}}$) for each i^{th} instance approximates the distribution across a set of joint pieces of evidence. The sum of the joint degree of similarity for all joint pieces of evidence for each i^{th} instance is equal to 1, summarized below:

$$\alpha^{i}_{v_j, v'_{j'}} = \alpha^{i}_{v,j}\, \alpha^{i}_{v',j'} \tag{9}$$

here, $v \neq v'$ and $v \in \{1, \dots, V_j\}$,$v' \in \{1, \dots, V'_{j'}\}$

$$\sum_{v \in \{1,\dots,V_j\}} \left(\sum_{v' \in \{1,\dots,V'_{j'}\}} \alpha^{i}_{v_j, v'_{j'}} \right) = 1 \tag{10}$$

Figure 5 demonstrates the process of generating joint transformation data for two distinct rules, labeled j and j'. Table A1 corresponds to the column e_j associated with rule j and Table B1 depicts the column $e_{j'}$ associated with rule j' in an aggregated dataset ($\bar{\mathcal{D}}$). Tables A1 and B1 undergo numerical transformations based on the methods presented in Section 3.2.1.1, subsections (a) and (b). The resultant data from the transformations for the individual rules j and j' are displayed in Tables A2 and B2, respectively. Table C1 features the conjoined transformation of data representing joint evidence in rules j and j' by methods presented in Section 3.2.1.1, subsection (c).

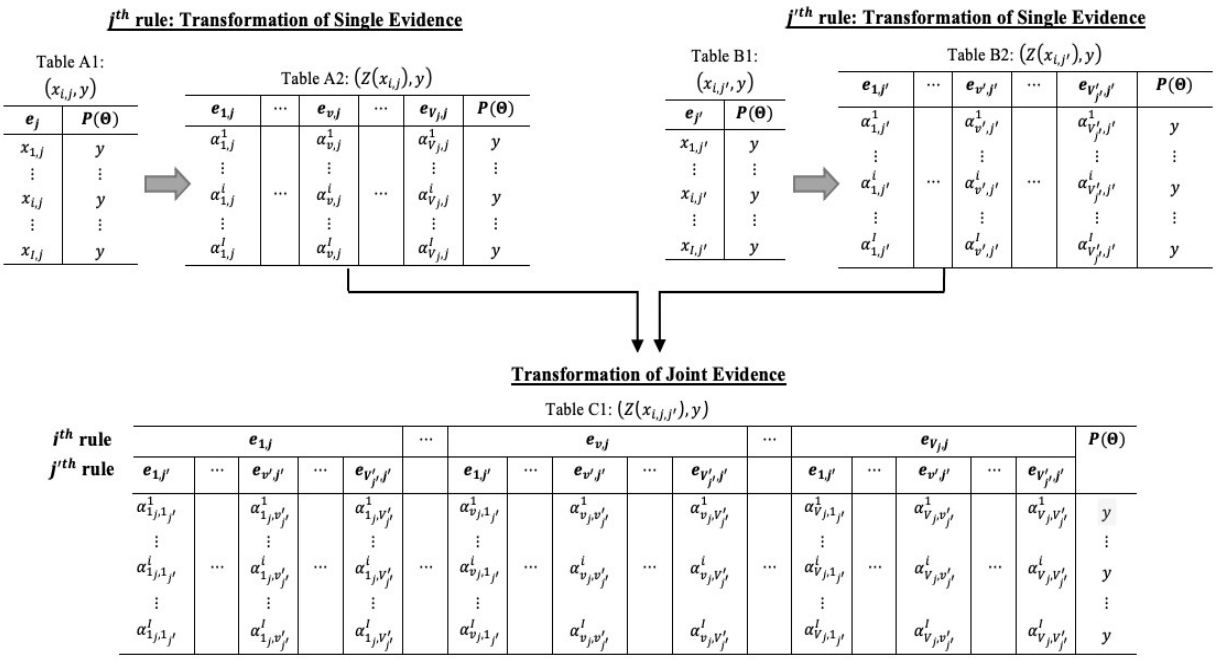

Figure 5: Data transformation process for joint evidence

3.2.1.2 Estimate Basic Probability of Evidence

The sample pair $(x_{i,j}, y)$ of each i^{th} instance of a j^{th} rule is transformed by the method described in Section 3.2.1.1 $(x_{i,j}, y) \rightarrow (Z(x_{i,j}), y)$. Table A and B in Figure 6 show the sample pair and the transformed sample pair, respectively. A contingency table is created from the transformed data table containing the sum of the matching degree of each v^{th} criteria paired with an outcome $y \in P(\Theta)$. Table C in Figure 6 is a contingency table, where $a_{y,v,j} = \left(\sum_{i=1}^{I} \alpha_{v,j}^{i} \mid \left(Z(x_{i,j}) = \alpha_{v,j}^{i}, y\right)\right)$ is the sum of matching degrees of a v^{th} criterion in a j^{th} rule which belongs to a class y.

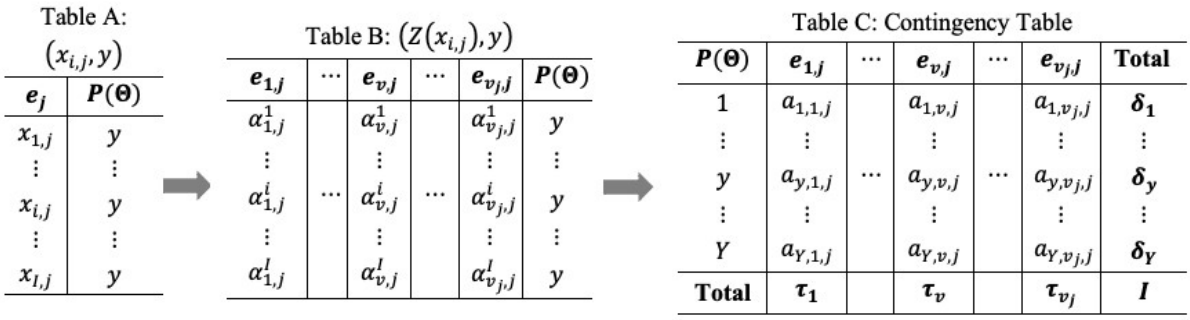

Figure 6: Illustrates the contingency table of a single piece of evidence

The sum of the matching degrees of all criteria for a given outcome is denoted by δ_y. It is the row-wise sum of the matching degrees in a contingency table, represented as $\delta_y = \sum_{v=1}^{v_j} a_{y,v,j}$. The sum of the matching degrees of all subsets of outcomes for a given criterion is denoted by τ_v. This is determined by summing the

matching degrees column-wise in a contingency table, represented as $\tau_v = \sum_{y \in P(\Theta)} a_{y,v,j}$. The overall summation of matching degrees across both rows and columns in the contingency table equates to the total number of instances in the dataset, denoted by I. Mathematically, this relationship is expressed as: $\sum_{v \in \{1,\ldots,V_j\}} \tau_v = \sum_{y \in P(\Theta)} \delta_y = I$. If any sample data is missing, the matching degree for all criteria in a rule is filled with zero. The missing samples are not included in estimating the support for a piece of evidence by probability mass.

Let $L^f_{y,v,j}$ be the likelihood of observing v^{th} criteria in a j^{th} rule for an outcome y (a piece of evidence $e_{y,v,j}$). Here, f represents estimation from data. It is calculated from the contingency table as follows:

$$L^f_{y,v,j} = \frac{a_{y,v,j}}{\delta_y}, \forall y \in P(\Theta) \tag{11}$$

The basic probability is calculated from normalized likelihood as follows:

$$p^f_{y,v,j} = \frac{L^f_{y,v,j}}{\sum_{y \in P(\Theta)} L^f_{y,v,j}}, \forall y \in P(\Theta) \tag{12}$$

Here, $p^f_{y,v,j}$ is the basic probability of evidence $e_{y,v,j}$. The sum of the probability mass of all V_j criteria in a rule is one.

3.2.1.3 Reliability and Weight of Evidence

The reliability ($r_{y,v,j}$) of evidence in a rule depends on the number of samples $x_{i,j}$ for the outcome y (Xu, et al., 2017). The reliability of v^{th} evidence in a j^{th} rule pointing to an outcome y can be obtained from the contingency table by the following Expression:

$$r_{y,v,j} = \frac{\alpha_{y,v,j}}{max_{y \in P(\Theta)} \alpha_{y,v,j}}, \quad \forall y \in P(\Theta) \tag{13}$$

The overall reliability of evidence $(r_{v,j})$ is the sum of the products of $r_{y,v,j}$ and $p^f_{y,v,j}$ for $\forall y \in P(\Theta)$, as shown below:

$$r_{v,j} = \sum_{y \in P(\Theta)} r_{y,v,j} \; p^f_{y,v,j} \tag{14}$$

The relative reliability of a rule can be quantified as the average reliability across all its individual pieces of evidence:

$$r_j = \frac{\sum_{v=1}^{V_j} r_{v,j}}{\sum_{j=1}^{J} \left(\sum_{v=1}^{V_j} r_{v,j} \right)} \tag{15}$$

Similarly, the relative weight, that is, the importance of a rule or its significance, is computed as the average weight of each piece of evidence:

$$w_j = \frac{\sum_{v=1}^{V_j} \left(\sum_{y \in P(\Theta)} w_{y,v,j} \; p^f_{y,v,j} \right)}{\sum_{j=1}^{J} \left[\sum_{v=1}^{V_j} \left(\sum_{y \in P(\Theta)} w_{y,v,j} \; p^f_{y,v,j} \right) \right]} \tag{16}$$

3.2.1.4 Probability Mass of Evidence

The probability mass of a piece of evidence is the weighted basic probability. The weight of a piece of evidence for a class is a parameter in ER-X. It is trained by data-driven optimization. The probability mass of each piece of evidence is given as follows:

$$m_{y,v,j}^{f} = w_{y,v,j}\, p_{y,v,j}^{f} \tag{17}$$

3.2.1.5 Combine Probability Mass of Multiple Evidence from Data

Finally, utilize the C-MAKER inference method to combine multiple pieces of evidence in all J rules to get inference by joint probability mass. The rules are combined iteratively. Each iteration combines two pieces of evidence. The interrelation index between two pieces of evidence, $e_{v,j}$ and $e_{v',j'}$ is estimated to combine multiple pieces of evidence pointing to the same outcome y, $y \in P(\Theta)$:

$$\psi_{y,v_j,v'_{j'}} = \begin{cases} 0\,, & \text{if } p_{y,v,j}^{f} = 0 \text{ or } p_{y,v',j'}^{f} = 0 \\ \dfrac{p_{y,v_j,v'_{j'}}^{f}}{p_{y,v,j}^{f}\, p_{y,v',j'}^{f}}, & otherwise \end{cases} \tag{18}$$

where, $\psi_{y,v_j,v'_{j'}}$ is the interrelation index of a joint evidence $e_{v_j,v'_{j'}}$ pointing to an outcome y, $y \in P(\Theta)$. The joint basic probability mass is calculated from joint transformation data by following the approach in Section 3.2.1.2. A joint contingency table between multiple pieces of evidence in different rules is created to obtain joint basic probability mass. The test to evaluate the feasibility of combining multiple pieces of evidence to generate joint probability mass can be conducted by sparse index (Sachan, et al., 2021). The sparse index ($\mathcal{S}$) is given by:

$$\mathcal{S} = \frac{3\mathcal{CRT}}{\mathcal{CT}+\mathcal{RT}+\mathcal{CR}} \tag{19}$$

where $\mathcal{C}\epsilon[0,1]$ is the proportion of non-empty columns in a joint contingency table, $\mathcal{R}\epsilon\{0,1\}$ indicate non-zero instances for singleton subsets in the power set of the frame of discernment, and $\mathcal{T}\epsilon[0,1]$ is the proportion of the number of complete samples in a joint contingency table. The highest value of the sparse index is one, which represents zero missing values in the joint contingency table, and its lowest value is zero, which represents an empty joint contingency table.

Let the joint support for a proposition y by two pieces of evidence $e_{y,v,j}$ and $e_{y,v',j'}$ be denoted by $\bar{m}_{y,v_j,v'_{j'}}^{f}$. The joint probability mass ($\bar{m}_{y,v_j,v'_{j'}}^{f}$) is given by:

$$\bar{m}_{y,v_j,v'_{j'}}^{f} = \begin{cases} 0 & , \quad y = \emptyset \\ \dfrac{m_{y,v_j,v'_{j'}}^{f}}{\sum_{y\in P(\Theta)} m_{y,v_j,v'_{j'}}^{f} + m_{P(\theta),v_j,v'_{j'}}^{f}} & , \quad \forall y \in P(\Theta), y \neq \emptyset \end{cases} \tag{20a}$$

$$m^{f}_{y,v_j,v'_{j'}} = \left[\left(1 - r_{v',j'}\right)m^{f}_{y,v,j} + \left(1 - r_{v,j}\right) m^{f}_{y,v',j'}\right] + \sum_{B\cap C=y} \psi_{B,C,v_j,v'_{j'}}\, m^{f}_{B,v,j} m^{f}_{C,v',j'} \quad (20b)$$

The residual support ($m^{f}_{P(\Theta),v_j,v'_{j'}}$) in Equation (20a) is earmarked to the power set as given by:

$$m^{f}_{P(\Theta),v_j,v'_{j'}} = m^{f}_{y,v,j}\, m^{f}_{\Theta,v_j,v'_{j'}} \quad (20c)$$

The weight of evidence is the training parameter. It is the variable that controls and estimates the characteristics of training data. The parameters are fine-tuned to bridge the gap between the expected and predicted probability masses for every data instance. The optimization problem aims to determine the optimal weight for each piece of evidence that minimizes an objective function. The objective function in ER-x quantifies the square of the difference between the expected and predicted probability masses, normalized by the total number of data instances.

The objective function to train the weight for each piece of evidence is:

$$\begin{gathered} Minimize{:}f(w)=\frac{1}{2I}\sum_{i=1}^{I}\sum_{y\in P(\Theta)}(m^{o}-\bar{m}(w))^{2} \\ constraints:\ \ 0\leq w \leq 1 \end{gathered} \quad (21)$$

In Equation (21), the observed and estimated probability masses of each instance i are simply denoted by m^{o} and $\bar{m}(w)$, respectively. The objective function $f(w)$ minimizes the error between the observed (m^{o}) and estimated ($\bar{m}(w)$) probability masses across all instances in the dataset. Variable w simply denotes the weight of a piece of evidence $\left(w_{y,v,j}\right)$ and the joint piece of evidence $\left(w_{y,v_j,v'_{j'}}\right)$.

3.2.2 Probability Mass of Evidence from Domain Experts

The loan underwriters are domain experts. The subjective judgment of multiple underwriters can be acquired to understand the variability in decisions in a lending firm. An underwriter can provide judgment as a degree of belief by examining a single and a joint piece of evidence. The degree of belief reflects the probability mass of evidence for a given outcome.

The judgments by single piece of evidence can be trusted if it has direct power to provide a firm decision for a loan application regardless of other evidence. For example, a lending firm has strict decline policies for borrowers with "two or more bankruptcies in the past" and "six or more credit defaults in the past 24 months". Suppose BOE of both pieces of evidence is acquired from an underwriter, then it is expected that all underwriters would provide full support towards the reject decision and zero support for any other outcome. If the BOE of both pieces of evidence is estimated from data, then it is expected that the estimated probability mass of both pieces of evidence is one; $\bar{m}^{f}_{y,v,j} = 1$ (y is reject decision) because all samples in the dataset for pieces of evidence would have full support for rejection decision and zero support for other decisions. However, inconsistency in outcome from data could arise due to unintentional error by human underwriters, biases in an existing decision-support system, a large proportion of missing data, and recent amendments in lending policy and regulations. This type of inconsistency is high for evidence pointing to heuristic information.

Most decisions by underwriters are given by collective examination of multiple evidence instead of single evidence. For example, suppose a borrower's data has not activated clear-cut declinable criteria in the rules like in the previous example for single evidence. In that case, an underwriter jointly evaluates other remaining evidence, such as "credit score = 550", "unsecured loan defaults in last 24 months = 0", and "secured loan default in last 24 months = 1". In this case, some underwriters might reject this loan due to unconvincing credit history on account of two secured loan defaults. In contrast, some underwriters might approve funding due to a good credit score and few defaults in secured and unsecured loans. This indicates variability in decisions by domain experts. The ultimate-true decision for such cases can be approximated by combining the BOE from multiple underwriters to avoid noisy judgment by experts that would reflect in future training data.

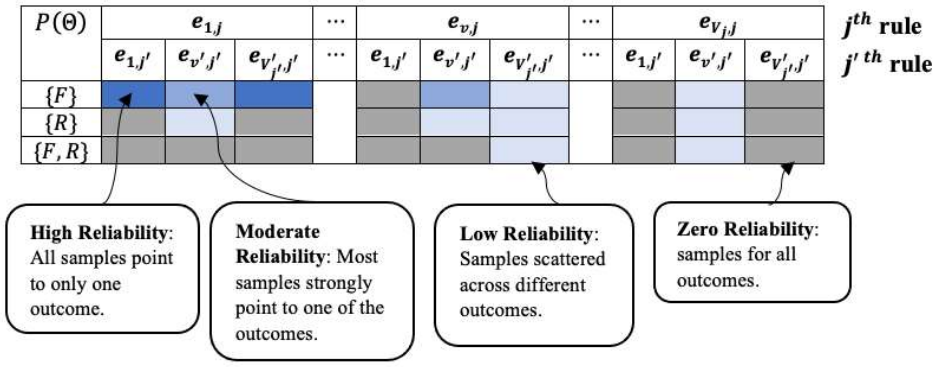

Figure 7: Example of the reliability of a piece of joint evidence. In this diagram, the powerset is the subset of two decisions, F (fund) and R (reject)

Figure 7 illustrates the reliability of joint pieces of evidence estimated by the amount of data support (number of samples) in the joint evidence space represented by the contingency table. For simplicity, a contingency table of two joint pieces of evidence is shown; however, more than two pieces of evidence can be combined. Usually, all possible joint pieces of evidence in the real-world data are not available, which results in empty columns in a contingency table. The piece of evidence reflecting the most recurrent characteristic of borrowers occurring in a lending firm has a large number of instances, whereas the characteristics of a rare type of borrowers may have a very small or almost zero number of instances.

Table 2: Consistency between estimated probability mass of evidence from data and subjective judgment from domain experts

Probability Mass of Evidence (Following applies to both single and joint evidence)	**Outcome** $y, y' \in P(\Theta)$	**Decision Consistency between data and domain experts**
All samples in the data (high reliability of evidence) and experts point to only one outcome. The probability mass of evidence estimated from data ($\bar{m}$) and obtained from	$y = y'$	Yes *Ultimate-true decision*

underwriters as a degree of belief is approximately equal to one. $\bar{m}^{f}_{y,v_j,v'_{j'}} \cong \bar{m}^{f'}_{y',v_j,v'_{j'}} \cong 1$	$y \neq y'$	No
Most samples for a piece of evidence (moderate reliability of evidence) and experts strongly point to one of the outcomes. The probability mass of evidence estimated from data ($\bar{m}$) and obtained from underwriters as a degree of belief are almost equal. $\bar{m}^{f}_{y,v_j,v'_{j'}} \approx \bar{m}^{f'}_{y',v_j,v'_{j'}}, \forall y, y' \in P(\Theta)$	$y = y'$	Yes
	$y \neq y'$	No
Samples for a piece of evidence scattered across different outcomes (low reliability of evidence). Unequal probability mass of a piece of evidence estimated from data and obtained from underwriters as a degree of belief. $\bar{m}^{f}_{y,v_j,v'_{j'}} \neq \bar{m}^{f'}_{y',v_j,v'_{j'}}, \forall y, y' \in P(\Theta)$	$y = y'$	No
	$y \neq y'$	No
A zero sample is used for the evidence for all outcomes in the dataset; therefore, the probability mass of a piece of evidence can be considered zero because it cannot be estimated without representative data. The information required to estimate the probability mass is absent from the data. $\bar{m}^{f}_{y,v_j,v'_{j'}} \neq \bar{m}^{f'}_{y',v_j,v'_{j'}}$ $\bar{m}^{f}_{y,v_j,v'_{j'}} = 0$ and $\bar{m}^{f'}_{y',v_j,v'_{j'}} \neq 0$	$y = y'$	No
	$y \neq y'$	No

Table 2 and Figure 7 demonstrate four scenarios to investigate the consistency between a decision by data and expert judgment. The human expert is denoted by f', and data is denoted by f. In the first scenario, the estimated probability mass of a joint piece of evidence derived from data is approximately equal to one for an outcome and is roughly equivalent to zero for other outcomes in a powerset because all evidence instances point to a single outcome (full support for an outcome). If the probability mass of evidence for an outcome estimated from data is equal to subjective judgment as a degree of belief from an expert, such that $\bar{m}^{f}_{y,v_j,v'_{j'}} \cong \bar{m}^{f'}_{y',v_j,v'_{j'}} \cong 1$ and $y = y'$; $y, y' \in P(\Theta)$, then the judgment consistency between data and an expert can be validated. This uniformity demonstrates the potential for automated decisions. These single or joint pieces of evidence are presumed to point to the "ultimate-true decision".

Similarly, in the second scenario, the probability mass of evidence for an outcome estimated from data is almost equal to that obtained from underwriters for all outcomes in a powerset. In this case, the judgment inconsistency between data and an expert is not very high, and the probability mass integrated from data and experts can be utilized to approximate the "ultimate-true decision". For other scenarios, human underwriters

are required to provide a manual decision on loan applications due to a lack of consistency in a decision by data and experts.

3.2.3 Relative Consistency to Estimate Ultimate-True Decision

The purpose of consistency analysis is to audit the assessment given by human experts in the past now reflected in the data, and manually adjust the output of the algorithm to improve the performance. This task establishes a connection between domain experts and developers.

The relative inconsistency between multiple experts and data for a given piece of evidence can be measured by the degree of credibility. The degree of credibility of a source of information for a given piece of evidence is measured by dissimilarity or lack of similarity between the probability mass of the evidence (Yong, et al., 2004). The Jousselme distance based on Cuzzolin's geometric interpretation of the evidence is used to measure the dissimilarity between two sources (Jousselme & Maupin, 2012). Suppose, two BOEs for a joint piece of evidence $e_{v_j,v'_{j'}}$ is obtained from different sources of judgment, a loan underwriter f' and data f, such that $f, f' \in \{1, \dots, F\}$. A vector of the basic probability of a joint piece of evidence $\left(\overrightarrow{p_f}\left(e_{v_j,v'_{j'}}\right)\right)$ under the power set of the frame of decrement from a source f is simply denoted by $\overrightarrow{p_f}$. The difference between BOEs from two sources $\left(d_{f,f'}\left(e_{v_j,v'_{j'}}\right)\right)$ for a piece of evidence is simply denoted by $d_{f,f'}$. The similarity measure $Sim_{f,f'}$ from two different sources for a piece of evidence $(e_{v,j})$ is:

$$S_{f,f'} = 1 - d_{f,f'} = \sqrt{\frac{1}{2}\left(\overrightarrow{p_f} - \overrightarrow{p_{f'}}\right)^T D\left(\overrightarrow{p_f} - \overrightarrow{p_{f'}}\right)} \tag{22}$$

In Expression (22), D is defined as $D(y, y') = \frac{|y \cap y'|}{|y \cup y'|}$, $y, y' \in P(\Theta)$ and $|.|$ represents the cardinality. The BOEs from different sources have high dissimilarity if the distance between BOEs is large and vice versa. If there are F sources for BOE acquisition, then the similarity between a pair of sources can be presented through the following similarity matrix SMM:

$$SMM = \begin{bmatrix} 1 & \cdots & S_{1,f'} & \cdots & S_{1,F} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{f,1} & \cdots & S_{f,f'} & \cdots & S_{f,F} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ S_{F,1} & \cdots & S_{F,f'} & \cdots & 1 \end{bmatrix} \tag{23}$$

The degree of support for a BOE from a f^{th} source for a given piece of evidence $(e_{v,j})$ is:

$$Sup_f = \sum_{\substack{f=1 \\ f \neq f'}}^{F} S_{f,f'} \tag{24}$$

The degree of credibility CRD_f is the normalized degree of support:

$$CRD_f = \frac{Sup_f}{\sum_{z \in \{1,\dots,F\}} Sup_z} \tag{25}$$

In Expression (25), CRD_f is the simple representation of the degree of credibility of a source for a joint piece of evidence $CRD_f(e_{v_j,v'_{j'}})$. The relative credibility of multiple sources is evenly distributed if the probability mass for different outcomes in the powerset is the same and vice versa. The probability mass in a BOE is information from multiple sources. It can be fused together to estimate the "ultimate-true decision" $\left(M_{y,v_j,v'_{j'}}\right)$ by considering the degree of credibility of each source to estimate weighted combined probability mass, as follows:

$$M_{y,v_j,v'_{j'}} = \sum_{z\in\{1,\dots,F\}} CRD_z \, \bar{m}^z_{y,v_j,v'_{j'}} \tag{26}$$

4. Study on a FinTech Transformation in Specialist Lending Firm

4.1 Lending Rules and Uncertain Lending Data

This study is conducted on a specialist lending firm based in the UK that provides mortgage loans to underserved communities. These communities often struggle with less than optimal credit scores, hindering their chances of securing a successful loan from a retail bank. This demographic requires a complex and time-intensive manual underwriting process. The firm intends to investigate the decision consistency between data and human underwriters to evaluate their potential to streamline their lending process through the integration of AI, marking its transition into a FinTech firm. The evaluation process is shown in Figure 8.

The lending firm established seven distinct lending rules, each featuring decline and referral criteria as the guideline for underwriters to provide mortgage loan decisions, as shown in Figure 9. The criteria in each rule are mutually exclusive. The subjective judgment by human underwriters on loan applications centres around these rules. They analyze borrower data captured from a digital application, data obtained from the Credit and Fraud Intelligence Bureau, and supporting documents (proof of ID, address history, and income) to satisfy the conditions in the rules. The data is captured by a document processing tool from digital applications and borrower-submitted supporting documents. However, the tool faced challenges in the accurate extraction of unstructured data from scanned documents through Optical Character Recognition (OCR), necessitating manual verification. A customer credit and fraud bureau provided information to verify identity, default, and fraud linked to an application's residence (current address, previous address, and previous linking address).

The data relevant to each rule is compiled following the procedure outlined in Section 3.1. Borrower affordability is then assessed by evaluating sustainable income, expenditure, and existing debts to estimate the Stressed-Maximum Affordability Monthly Repayment (Sachan, et al., 2020). The complete dataset merged from different sources had 5700 loan applications or instances. It had 26.8% rejected and 73.2% funded cases.

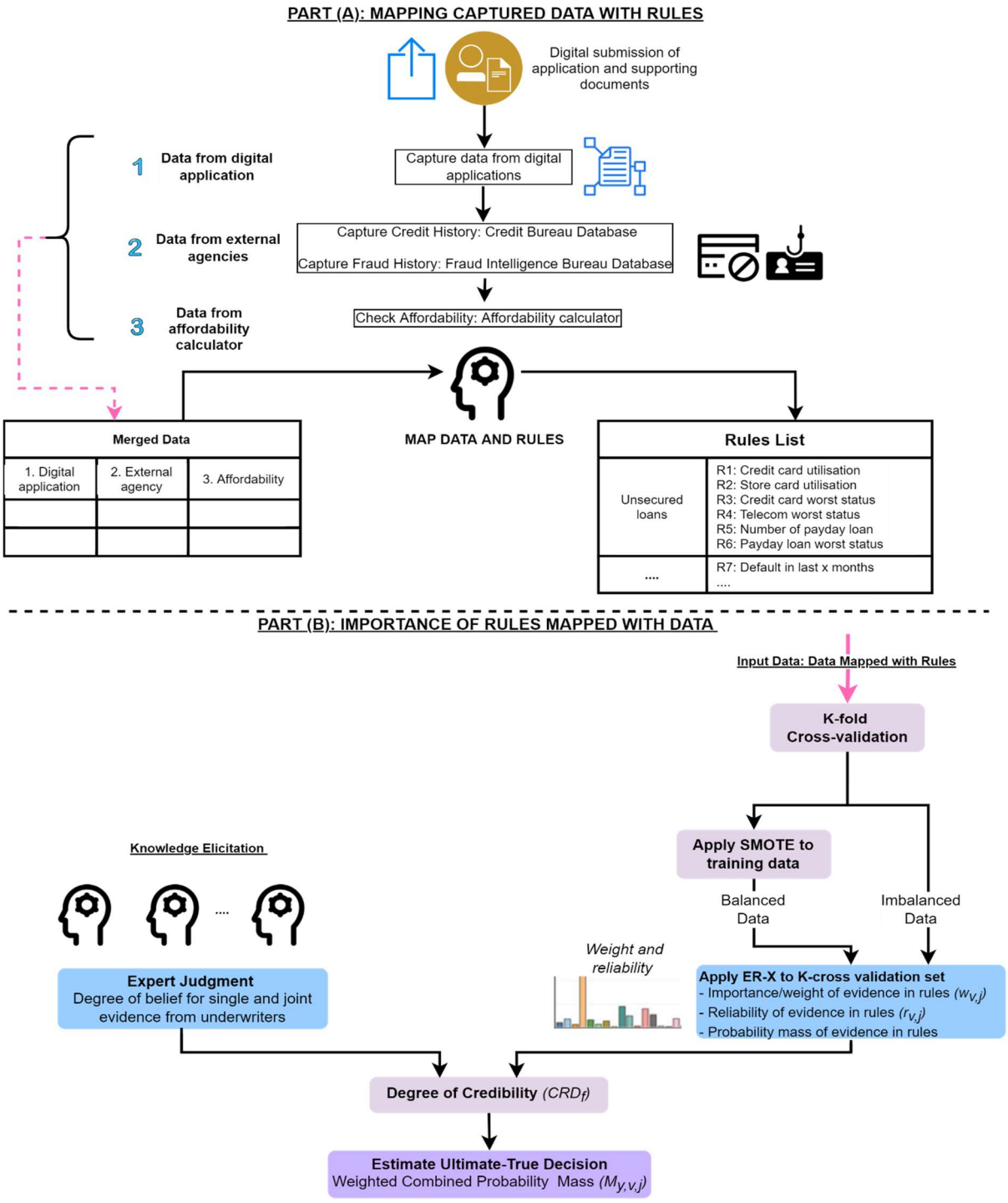

Figure 8: Process to evaluate decision consistency between data and domain experts

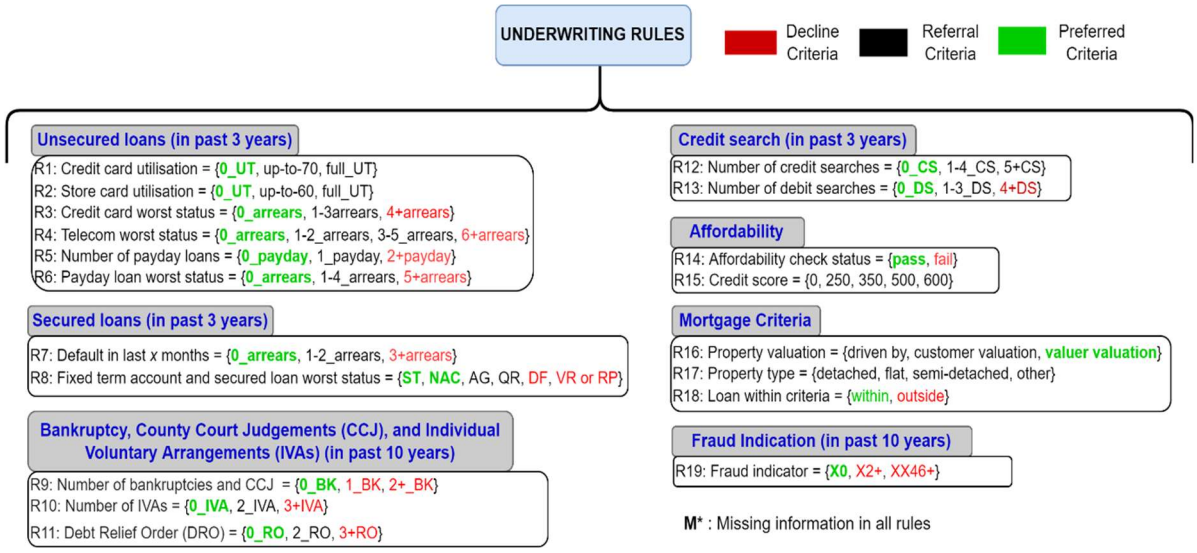

Figure 9: Decline and referral rules for loan underwriting. A full definition of acronyms for criteria can be seen in Table 4, Appendix A.

4.2 Evaluating Criteria in Lending Rules and Processing Time

The augmentation of human intelligence in AI starts by first getting insight into data serving the AI algorithm. A rule weight can be interpreted as the global importance of a feature for the rule in a dataset. Reliability can be interpreted as the capability of the feature to point correctly to a particular decision. Figure 10 shows the relative weight and reliability of rules arranged in decreasing order for balanced and imbalanced data. The rules R16, R15, and R14 on property valuation, credit score, and affordability status have the highest weight and reliability in aggregated lending data. This implies that data for these rules has the caliber to point strongly towards a decision, and in general, they were critical in estimating decisions.

Certain criteria within these rules, such as "R14 = pass", "R15 = 500", "R15 = 600", and "R16 = valuer valuation", have a high probability mass for the funding decision. Conversely, criteria such as "R14 = fail" and "R15 = 0", have a high probability towards a reject decision, as shown in the left column of Figure 11. It shows the probability mass for two outcomes in the powerset of frame of discernment $\Theta = \{\{F\}, \{R\}, \emptyset = \{F, R\}\}$, where F stands for the fund, R stands for the reject, and the set $\{F, R\}$ is empty due to the absence of nondeterministic recorded decisions and no missing outcome in the output feature in the dataset. The performance of the ER-X framework was compared with that of a deep neural network (DNN), a decision tree, and logistic regression.

Like any other machine learning algorithm, ER-X is susceptible to overfitting on the training data, which can result in poor performance on validation or unseen data. The k-fold cross-validation can be used as a regularization technique to suppress overfitting. In this approach, an independent test set is initially set aside for the final model evaluation; it is not used during the k-fold cross-validation process. The k-fold cross-validation of the remaining training dataset is partitioned into k equally sized folds or subsets. An AI algorithm is trained on $k - 1$ folds and evaluated on a one-fold left out. This procedure is repeated k times, each time with a different fold serving as the validation set.

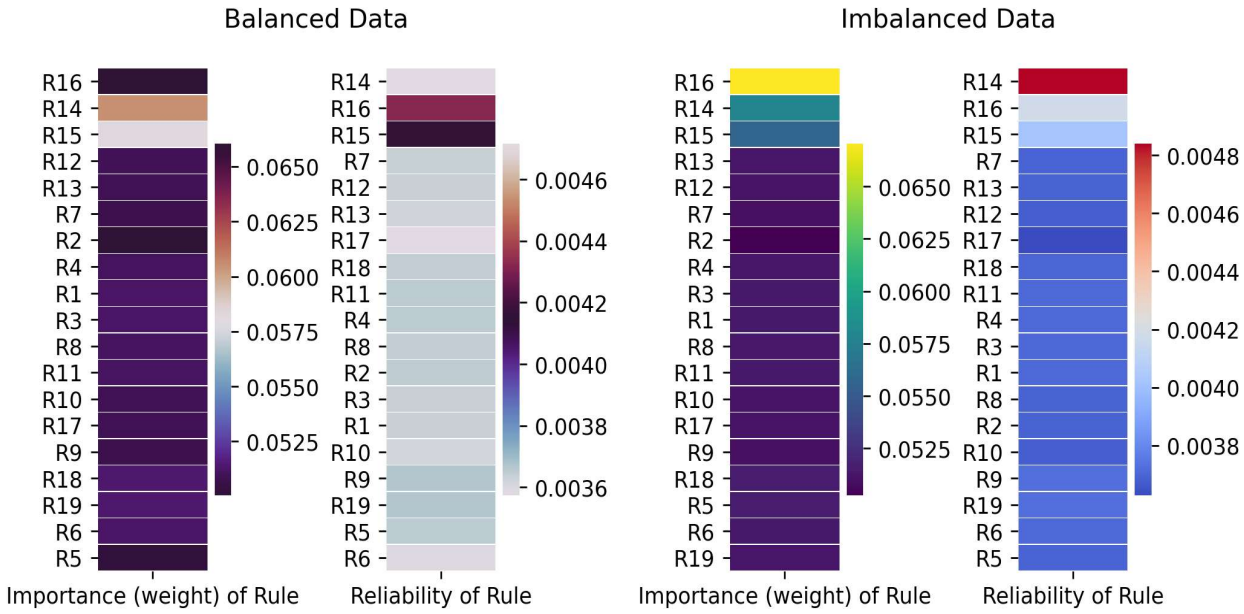

Figure 10: Importance (w_j) and reliability (r_j) of lending rules

The ER-X was applied to the k cross-validation set ($k = 5$ could be an arbitrary choice). Both the weight and reliability of the evidence are affected by imbalanced classes in a dataset. The Synthetic Minority Over-sampling Technique (SMOTE) technique has been implemented to balance minority classes by generating synthetic instances $\left(x_i = \left(x_{i,j},\ j \in \{1, \dots, J\}\right)\right)$ around the decision region of the minority class (Almaghrabi, et al., 2021). After balancing the classes using SMOTE, ER-X was applied to k different cross-validation sets to data balanced by SMOTE and imbalanced data. After cross-validation, the model version that exhibited the best performance on the validation folds was selected for the final assessment on an independent test set that was previously set aside. The dataset consisted of 5700 loan applications. A stratified one-sixth portion of the data was kept aside for testing purposes. The stratification process ensures that the proportion of samples from each class in the subset is representative of the entire dataset (Sechidis, et al., 2011). The remaining data were utilized for 5-fold cross-validation.

Table 8 in Appendix B presents the average accuracy metrics of the 5-fold cross-validation set and an independent test set for balanced (by SMOTE) and imbalanced datasets across four models: ER-X, DNN, Decision Tree, and Logistic Regression. In both balanced and imbalanced datasets, ER-X's performance is relatively close to DNN and is better than the decision tree and logistic regression; however, the DNN is not inherently interpretable and cannot incorporate expert knowledge. A similar trend was observed in the evaluation of the independent test set. However, model-agnostic and model-specific methods exist to interpret decisions by DNNs. The task of decomposing non-linearly transformed decisions made by DNNs is an active area of research (Zhang, et al., 2023). The details of the hyperparameters utilized for the DNN, Decision Tree, and Logistic Regression models are shown in Tables 9, 10, and 11 in Appendix B, respectively.

In this paper, the performance of AI models was measured using metrics such as precision, recall, and F1-score. Precision is a measure of the model's accuracy in predicting that a loan application should be funded; it is calculated as the proportion of loan applications correctly identified as suitable for funding (true positives) out of all the applications predicted to be funded (positive instances) by the AI model. Recall, on the other hand, evaluates the model's ability to correctly identify all loan applications that are genuinely suitable for funding, which is determined by the proportion of loan applications correctly identified as suitable for funding (true positives) out of all actual fundable applications (true positives and false negatives). F1-score strikes a balance between precision and recall.

In Figure 11, the right column shows the normal probability density function of time to process the information for a given rule, $\mathcal{N}(\mu = mean\ time, \sigma^2 = variance)$. The processing time is the time for information/data to arrive and get processed by the lending firm. The processing time is in hours; it is assumed that each working day has 8 hours. The information required for all rules does not arrive with the loan application pack. Additional documents are requested if the information in the application pack is unsatisfactory, especially previous and linking residential addresses to search credit history. An algorithm or human underwriters can provide a final decision after receiving complete information (or data). It suggests that an algorithm may improve the quality of lending decisions but may not effectively reduce the processing time of a loan application due to delays in the advent of complete information.

The mean processing time of information for rules for "Secured loans", "Bankruptcy, Individual Voluntary Arrangements (IVA), CCJ, and DRO", and "Affordability" is zero and has significantly less variance compared to other rules. The affordability rules (R14 and R15) and bankruptcy, IVA, CCJ, and DRO rules (R9, R10, and R11) have clear-cut decline criteria. These criteria have a high probability mass for rejection decisions and a short processing time. For example, "R14 = fail" explicitly rejects a loan application. Its mean information processing time is zero, with low variance (only 1 to 2 hours) across different loan applications. A part decision reached by decline criteria could be transformed into a hardcoded heuristic. However, this analysis alone is insufficient to confirm the trustworthiness of a decision derived from singular evidence (decline criteria) for task augmentation. This is due to the potential susceptibility of the algorithm's output to data noise and reflecting only a subset of borrowers. To mitigate this, human experts should be engaged in the auditing process, revisiting both their previous decisions and those made by the AI algorithms to reduce noise in judgment through the integration of high-quality data.

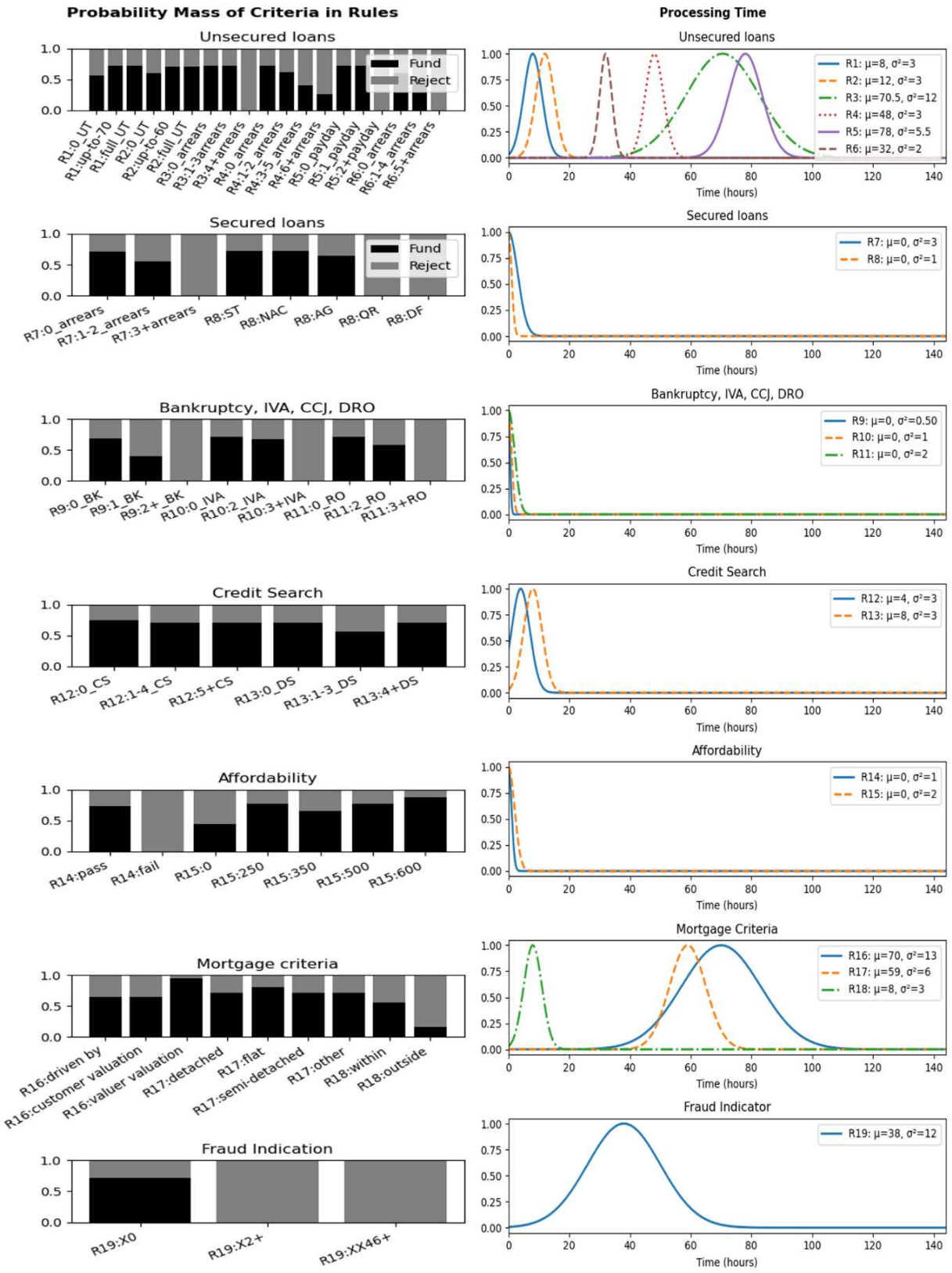

Figure 11: Probability mass of criteria in rules and processing time of information for each rule

4.3 Knowledge Elicitation to Evaluate the Decision Consistency Between Data and Experts

Knowledge elicitation (KE) is a resource-intensive procedure that requires substantial investment in time, active engagement, and specialized expertise of subject matter experts. The domain experts collaborate closely with the AI system development team on carefully structured tasks. In this study, KE was conducted to map data against a pre-defined rule set and acquire judgments from underwriters for the most recurrently funded and rejected cases that emerge in the firm to compare consistency between decisions derived from data and judgment from domain experts. Two PhD researchers acted as knowledge engineers to transfer domain-specific insights from two junior underwriters and two senior underwriters at the lending firm.

The task was executed in four stages. The first stage initiated an open dialogue between knowledge engineers and underwriters. This stage served dual purposes. It introduced underwriters to the concept of AI tools designed to support them in decision-making tasks and simultaneously allowed knowledge engineers to understand the basic aspects of the problem domain. In the second phase, knowledge engineers conducted non-intrusive observations of the underwriters while processing loan applications. This allowed the engineers to gain first-hand insights into their day-to-day operations and decision-making processes. In the third phase, a series of questions were prepared for underwriters based on the tasks performed during the observational stage (in phase two) and the lending guidelines of the firm as read by the knowledge engineer. In the fourth phase, the underwriters were presented with an online assessment sheet containing various scenarios (technically termed 'multiple evidence'). They were prompted to make heuristic decisions based on these scenarios, facilitating a richer understanding of the experiential knowledge embedded in their judgments.

Figure 12 illustrates an example of a window in an online assessment tool used for capturing the judgments of underwriters. The assessment consists of multiple windows, each requiring underwriters to indicate their decision by moving a radio button along a slider. This decision is transferred into probability mass by the technique shown in Section 3.2.1.1 for each outcome in the powerset; $P(\Theta) = \{\{F\}, \{R\}, \{F, R\}\}$. Here, the numerical values -1, 0, and 1 correspond to reject $\{R\}$, not sure $\{F, R\}$, and fund $\{F\}$, respectively. A radio button placed entirely towards -1, 0, or 1 indicates full support for an outcome. In the system, complete certainty for a given decision equates to a probability mass being equal to 1, $\bar{m}^{f'}_{y',v_j,v'_j} = 1$. Conversely, if the radio button is positioned between two outcomes, it denotes a degree of uncertainty, translating to a probabilistic decision.

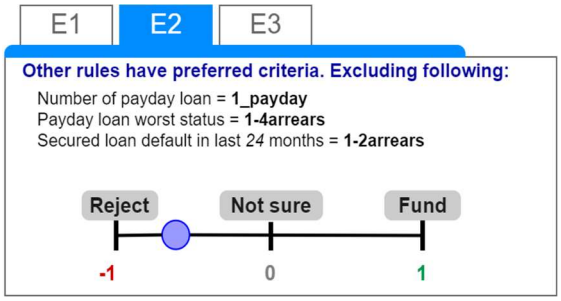

Figure 12: Illustration of online assessment

A loan application is expected to be declined if it triggers one or more clear-cut decline criteria (single pieces of evidence). Conversely, if a loan application does not activate any clear-cut criteria, multiple pieces of evidence are jointly assessed to reach a decision on whether to fund or reject the loan application. The boundary of inconsistency analysis was set to the subset of the most frequent cases encountered by underwriters in the firm. Each loan application type is defined by a combination of individual or multiple pieces of evidence. The multiple evidence for recurrently funded cases and rejected cases can be seen in Table 5 and Table 6 (Appendix A), respectively. A joint piece of evidence within these tables can be identified by its unique group number. For example, evidence labeled E34131 represents the characteristics of a funded case (the last case in Figure 14). Here, E means evidence, 3 is the third evidence in the first group, 4 is the fourth evidence in the second group, and so on. This structured naming convention aids in a systematic and organized representation of the various pieces of evidence associated with each loan application case.

Figures 13, 14, and 15 demonstrate the probability mass through line plots and degree of credibility (relative consistency) among four domain experts and data by heat maps for the clear-cut decline, recurrently funded, and rejected cases, respectively. In these figures, the junior underwriters are designated with labels '1' and '2', while senior underwriters are marked as '3' and '4'. The scattered judgment for single pieces of evidence in clear-cut cases and multiple pieces of evidence in recurrent cases is visually represented by variations in color intensity. The degree of credibility does not vary much among underwriters and data for clear-cut decline criteria and recurrently rejected loan applications compared to funded loan applications. The probability mass for an outcome by data reflects the collective judgment by multiple underwriters in the past at time t, whereas the subjective judgment by underwriters through online assessment was obtained at the time t', such that $t < t'$.

All underwriters fully agreed to reject any loan applications that displayed more than six telecom arrears, acknowledging this as an unambiguous decline criterion (R4: 6+arrears); however, this consensus among underwriters did not align with the data, which indicated a slight probability of funding ($\overline{m} = 0.268$). Similarly, data assigned a small probability mass to the decision to fund when the information regarding the worst status of a secured loan was missing (R8: M*). This discrepancy led to a discussion with the quantitative risk team. They concluded that the R4:6+arrears and R8: M* were referral criteria three years ago and later added as decline criteria. In alignment with the existing and previous decline policies, borrowers lacking credit scores (R15: M*) are subject to immediate decline. Surprisingly, this policy is not mirrored in the subjective judgment of the senior underwriters, compared to the data and junior underwriters. A loan that falls outside the established criteria (R18: outside) is generally declined outright, but exceptions can be made with managerial approval. Data and senior underwriters reflected this awareness. However, junior underwriters gave full support for explicit rejection. The consistency of collective decisions in the past at t and recently at t' indicates an "ultimate-true decision" for a given set of evidence representing a loan application. However, this does not imply that the decision will remain constant in the future, as it could change due to shifts in lending policy.

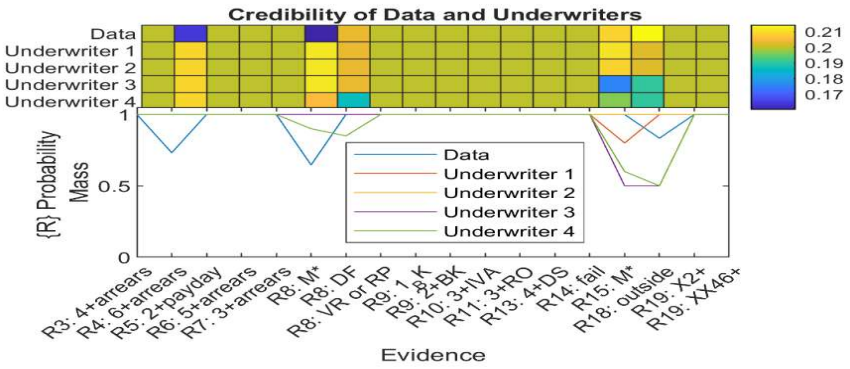

Figure 13: Consistency analysis of 16 clear-cut decline criteria. The line plot shows the probability mass for the reject {R} decision. The remaining probability mass is assigned to fund {F} decision by data and not-sure {F, R} by underwriters. Refer to Table 4, Appendix A for the full definition of evidence.

The variability in the degree of credibility does not show the low confidence of individual experts. Instead, it validates the existence of noisy judgment or disagreement among loan underwriters. The inconsistency between experts and past decisions by their peers exhibited by data is high for ambiguous loan applications. The most inconsistent cases for funding — [E31111, E32111, E31121, E32121, E34131, E32131, E33131] — are primarily associated with the third piece of evidence in group 1 in Table 5 (Appendix A). This subgroup has the highest number of missing (M*) criteria. Additionally, six out of seven of these cases have a credit score of less than 250 (belong to the first and second evidence of group 2 in Table 5). The most inconsistent cases for rejection — [E111211, E111311, E111321, E111361, E111331] — are characterized by missing debt relief order, credit scores exceeding 350, and in four out of five cases, a single payday loan in the past three years with three-to-five worst Telcom status. Complete elimination of inconsistency might be impractical. However, it's crucial to recognize the potential for underwriters to display inconsistencies in their future decisions on loan cases, as this can have significant implications for decision-making reliability.

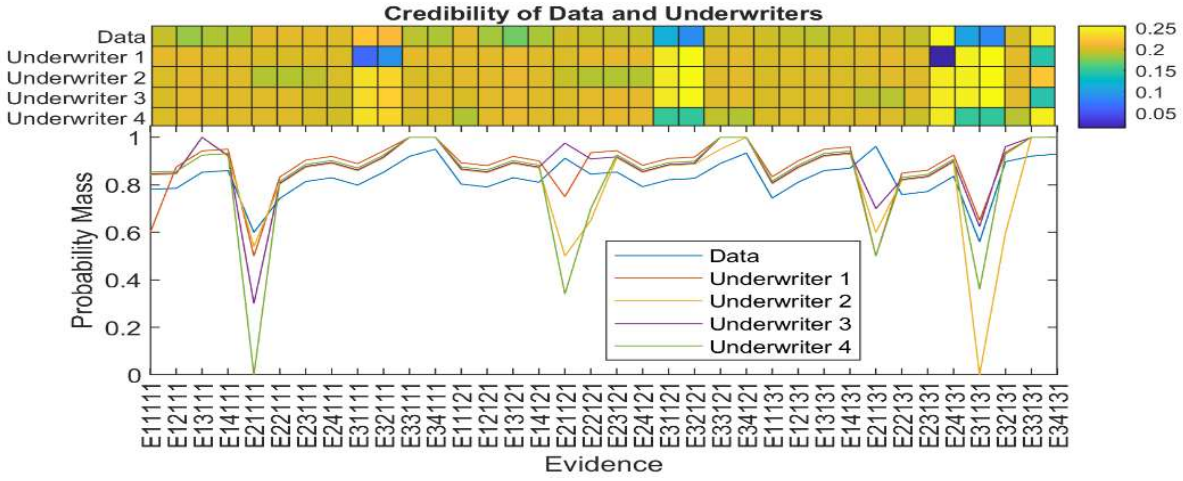

Figure 14: Consistency analysis of 36 recurrently funded cases. The line plot shows the probability mass for the fund {F} decision. The remaining probability mass is assigned to reject {R} decision by data and not-sure {F, R} by underwriters. Refer to Table 5, Appendix A for the full definition of joint evidence.

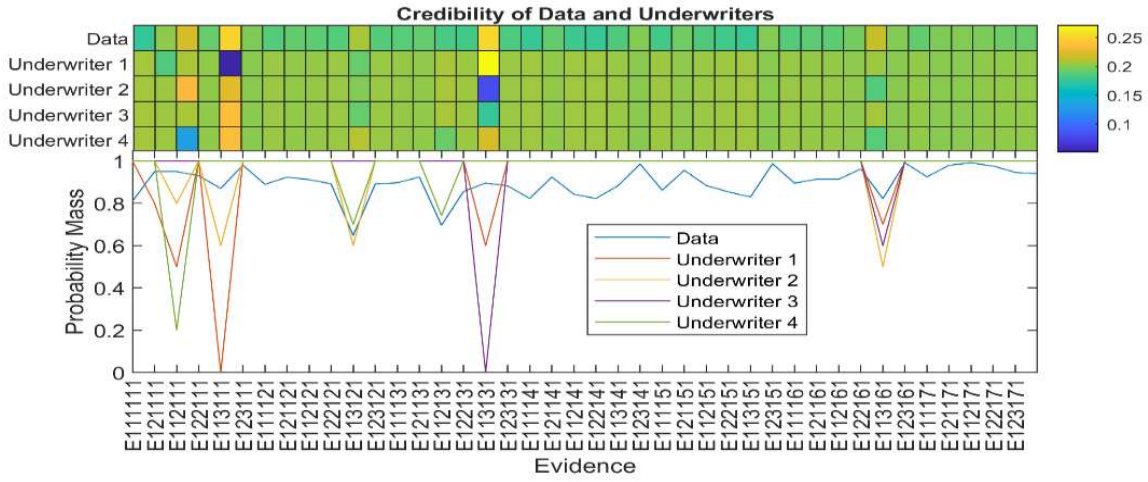

Figure 15: Consistency analysis of 42 recurrently rejected cases. The line plot shows the probability mass for the reject {R} decision. The remaining probability mass is assigned to fund {F} decision by data and not-sure {F, R} by underwriters. Refer to Table 6, Appendix A for the full definition of joint evidence.

4.4 Impact of Auditing Noisy Decisions by Consistency Analysis

This study aimed to maintain complete homogeneity in decision-making conditions and the quality of evidence assessed by underwriters. To maintain uniformity, knowledge elicitation of the most recurrent cases was conducted by a collective presentation of evidence captured from various ambiguous data sources on a computer screen, as illustrated in Figure 12. In practice, underwriters obtain some evidence from multiple documents arriving in an asynchronous time and some from external databases. As a result, the underwriter may encounter variations in the source and quality of evidence when assessing seemingly identical loan applications, which can lead to inconsistent decisions. A system for consistency analysis was implemented using the ER-X approach to standardize the assessment of information and conduct noise audits.

Figure 16 illustrates the impact of auditing ambiguous lending decisions using the ER-X model over four iterations. The auditing focused on three specific types of lending decisions: 16 clear-cut decline criteria, 36 recurrently funded cases, and 42 recurrently rejected cases. These selected categories constitute 40% of the mortgage loan application dataset, which comprises 5700 cases.

In the first iteration, ER-X was executed without underwriter input to establish a baseline performance. During this initial run, the average area under the curve (AUC) and F1-score calculated using 5-fold cross-validation were 0.86 and 0.80, respectively. AUC metric indicates the model's proficiency in distinguishing between loans that should be funded and those that should be rejected. A high AUC score, for instance, 0.86 achieved here, suggests that the ER-X model is adept at distinguishing loans that should be funded and rejected. An AUC score nearing 1 indicates a high level of model precision in discerning fundable from nonfundable applications. On the other end of the spectrum, an AUC score of 0.5 would mean the model's decision-making is essentially random, making its decisions no better than a coin toss. The F1-score is the balance between precision (accuracy of positive predictions) and recall (ability to capture all actual positive instances). An F1-score of 1 represents perfect precision and recall, while a score closer to 0 indicates poor precision and recall.

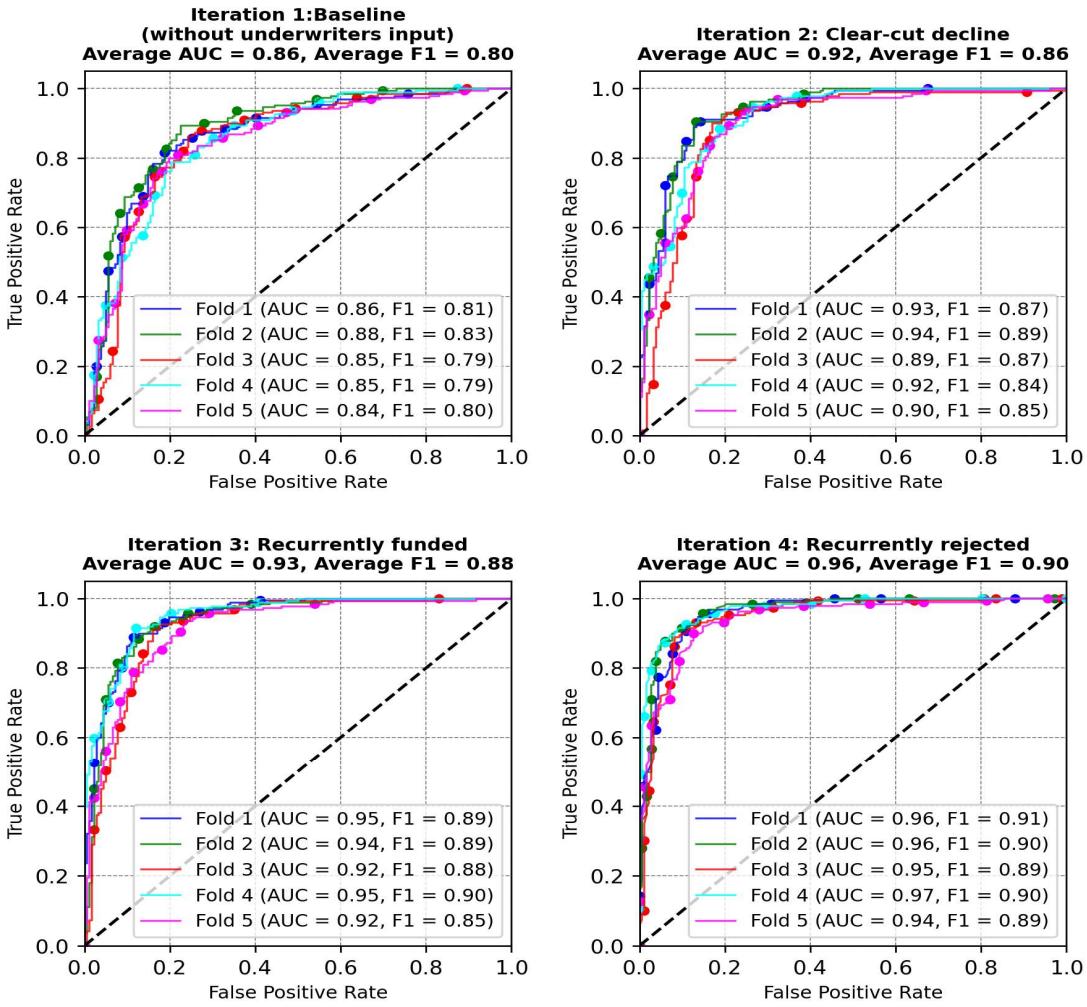

Figure 16: ER-X Performance After 4 Iterations of Consistency Analysis

The second iteration introduced improvements to the training dataset by incorporating refined feedback from underwriters and performing a consistency analysis on the clear-cut decline criteria to infer an "ultimate-true decision". This process significantly enhanced the performance of the ER-X model. The AUC rose to 0.92, and the F1-score increased to 0.86, marking a noticeable improvement from the initial baseline metrics.

Further refinements were applied to the training dataset during the third and fourth iterations by integrating improved underwriter feedback. A consistency analysis was performed on the recurrently funded and recurrently rejected cases, allowing the model to infer "ultimate-true" decisions. These adjustments resulted in a continued rise in the AUC and F1-scores. In the third iteration, the AUC and F1-scores increased to 0.93 and 0.88, respectively. During the fourth iteration, these metrics further escalated to 0.96 and 0.90, respectively. This progression indicates the continuous improvement of the ER-X model's performance over four iterations of consistency analysis.

5. Discussion

This study provides compelling evidence that the development of an AI-based lending decision-support system requires groundwork on input-data quality evaluation and assessment of the accuracy of decisions made by human experts. Past judgments by experts subsequently emerge in the future as outcomes in training data. Features derived from consolidated data must be aligned with lending policies to lay a solid foundation for a legitimate and ethical system.

The integration of AI in lending decision-making processes has profound ethical implications for both individuals and broader society. It demands a balance between technological advancement and ethical considerations to prevent discriminatory practices and uphold fairness, accountability, and transparency. The ethical use of AI has the potential to revolutionize financial accessibility, provide better access to credit, and empower individuals from diverse communities. However, ethical lapses can magnify societal inequalities, undermine trust, and violate individual rights. Therefore, lending institutions must adhere to ethical guidelines and regulatory frameworks, conduct regular audits, and foster open dialogue between stakeholders to address ethical concerns regarding the responsible use of AI before allocating resources and investments to develop AI decision-support systems. In the wake of FinTech transformation, regulators have initiated proposals to set legal guidelines for the safe implementation of AI in financial services firms (Burri & von Bothmer, 2021). Therefore, a lending firm is responsible for ensuring the integrity of every loan application decision on diverse profiles of borrowers.

The concept of consistency analysis presented in this paper has the potential to raise the quality of decisions by offsetting the gap between past inconsistent decisions and the "ultimate-true" decisions. There are no universally "correct" decisions for ambiguous cases. The significant practical expertise of underwriters is accumulated as heuristic knowledge over years of learning and from receiving critical feedback from their managers and peers.

Maintaining decision consistency is challenging, especially for loan applications with ambiguous information. For example, in the case study, loan applications with missing data on bankruptcy, IVA, and DRO had the most irregularities between outcomes from the data and the four underwriters compared to other cases. Estimating the "ultimate-true" decision for the most frequent type of loan applications in a lending firm would allow the lending firm to allocate the collaborative responsibilities between human underwriters and an AI system. Ideally, an interpretable AI system (to understand the reasoning), such as a heuristic rule-based or data-driven algorithm, would automate specific manual and cognitive tasks required for frequently occurring loan applications. As a result, underwriters can focus on non-routine tasks and exceptional loan applications that require specialist knowledge and occasional collaboration with a development team consisting of developers and data scientists. A consistency analysis is not a one-off process. It should be regularly scheduled to monitor the evolution of irregularities. Table 3 synthesizes the tasks for establishing collaboration between the underwriters and developers.

Table 3: Collaborative Tasks

Task	Underwriters-Developers Collaborative Tasks
Transform data into value	Identify attributes in raw data to achieve exhaustive coverage of the lending policy.
	Recognize minority and majority groups of borrowers in the data.
	Detect ambiguity in the input data.
Improve the quality of decisions	Maintain consistency in heuristic decisions among underwriters.
	Estimate ultimate-true decisions.
	Identify and then minimize the gap between past-inconsistent decisions and the "ultimate-true decisions".
	Identify any recent amendments to policies and regulations.
Audit and alter algorithmic decisions	Identify and then reduce the gap between inaccurate and biased algorithm-generated decisions and the ultimate-true decisions.

The process of Knowledge Elicitation (KE) plays a pivotal role in validating individual decisions on various types of loan applications, necessitating the examination of multiple pieces of evidence. However, this intricate process can be time-consuming, leading to a perception among human experts that such activities are unproductive and a "waste of time" (Forsythe & Buchanan, 1989). Despite these challenges, the notion of human involvement in refining and augmenting AI systems remains undisputed, fostering the advent of augmented work patterns. These emergent patterns encapsulate a range of activities, including the monitoring and refinement of algorithmic outputs as well as the annotation of data previously unidentified by the algorithms (Grønsund & Aanestad, 2020) (Sachan, et al., 2023).

6. Limitations and Future Research

One limitation of the current study is the absence of a thorough examination of the cost and resource implications associated with the refinement and maintenance of the ER-X model proposed for auditing the consistency of decisions and input data in AI decision-support systems. A comprehensive understanding of financial and human resources is essential to assess the sustainability of augmented work patterns facilitated by ER-X or other similar proposed techniques. A prospective direction for future research could be to undertake a comprehensive cost-benefit analysis, evaluating the time, costs, and resources required to integrate expert knowledge into the AI decision-support system. The scope of ER-X could be expanded to provide generalized audits for AI-driven decisions across various sectors, such as law, insurance, and healthcare. The proposed methodology is focused on its application to ambiguous structured data; however, the potential of ER-X can be extended to unstructured data.

Additionally, an unaddressed limitation of this study is the potential shift in the composition of teams conducting AI systems data auditing. Financial experts engaged in collaborative human-AI workflow

patterns might need a diverse range of expertise within their teams. Subsequent research should explore the dynamics of forming multidisciplinary teams in such a context, aiming to understand how diversifying skills and knowledge can refine the auditing process and elevate the overall performance of human-AI collaborative systems.

7. Conclusion

This paper contributes to the emerging research on the development of augmented human-AI collaborative workflows to manage noisy decisions and ensure the accurate embodiment of lending policies within data utilized by AI systems. It conceptualizes collaborative tasks between human underwriters and developers to perform value-added analysis, such as consistency analysis and knowledge elicitation exercises. These exercises pre-assess the quality of data and decisions to lay a legitimate foundation for a high-performing augmented AI decision-support system through a series of iterative refinements and the integration of underwriter feedback. The findings of this study demonstrate the intricate interrelationship between past judgments made by experts and the subsequent manifestation of decisions by experts in the training data of AI algorithms. It emphasizes the importance of meticulous alignment of lending policies with features derived from consolidated data to forge an ethical and legitimate system.

We introduced the ER-X methodology, a novel approach designed to evaluate the capability of evidence aggregated from multiple sources to provide a trustworthy data-driven decision using AI. Central to the ER-X methodology is consistency analysis, which is a cornerstone strategy to enhance decision quality. This was accomplished by bridging the gap between past inconsistencies and "ultimate-true" decisions.

The variability inherent in past decisions for different types of loan applications, represented by a set of single (clear-cut cases) or multiple pieces of evidence, was determined by comparing outcomes estimated from data and expert judgments derived from four human underwriters during the knowledge elicitation process. A business case study on a lending firm demonstrates guidance on a procedure to capture human expert knowledge through an online assessment. This study addresses the strategy of remodelling the end-to-end loan decision-making process to incorporate AI by establishing underwriters' verification feedback to prepare for transformation into a FinTech firm without defecting back to manual decisions and relying on hard-coded heuristics. The paper signifies that an augmented system workflow must exceed technical boundaries through multidisciplinary collaborations in lending firms to deploy a continual learning system.

Acknowledgments (To be added after double-blind review. Anonymized with XXX)

We express our appreciation to all the underwriters at XXX for their valuable insights into the underwriting process and for generously dedicating their time to participate in the knowledge acquisition tasks. This work is funded by the XXX (Grant code: XXX). We wish to express our sincere gratitude to the two anonymous reviewers whose constructive feedback significantly enhanced the quality of the manuscript.

Bibliography

Abellán, J. & Castellano, J. G., 2017. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert systems with applications,* pp. 1-10.

Akhavan, M., Sebt, M. V. & Ameli, M., 2021. Risk assessment modeling for knowledge based and startup projects based on feasibility studies: A Bayesian network approach. *Knowledge-Based Systems,* Volume 222, p. 106992.

Akter, S. et al., 2023. A framework for AI-powered service innovation capability: Review and agenda for future research. *Technovation,* p. 102768.

Almaghrabi, F., Xu, D. L. & Yang, J. B., 2021. An evidential reasoning rule based feature selection for improving trauma outcome prediction. *Applied Soft Computing,* p. 107112.

Almansour, M., 2023. Artificial intelligence and resource optimization: A study of Fintech start-ups. *Resources Policy,* p. 103250.

Anderson, B., 2019. Using Bayesian networks to perform reject inference. *Expert Systems with Applications,* pp. 349-356.

Ashta, A. & Herrmann, H., 2021. "Artificial intelligence and fintech: An overview of opportunities and risks for banking, investments, and microfinance. *Strategic Change,* pp. 211-222.

Bellomarini, L. et al., 2022. Data science with Vadalog: Knowledge Graphs with machine learning and reasoning in practice. *Future Generation Computer Systems,* pp. 407-422.

Bijak, K. & Thomas, L. C., 2012. Does segmentation always improve model performance in credit scoring?. *Expert Systems with Applications,* pp. 2433-2442.

Bohanec, M. & Rajkovic, V., 1988. *Knowledge acquisition and explanation for multi-attribute decision making..* France: Avignon, s.n., pp. 59-78.

Bolger, F. & Wright, G., 1994. Assessing the quality of expert judgment: Issues and analysis. *Decision support systems,* pp. 1-24.

Bonarini, A. & Maniezzo, V., 1991. Integrating expert systems and decision-support systems: principles and practice.. *Knowledge-Based Systems,* 4(3), pp. 172-176.

Boundy-Singer, Z. M., Ziemba, C. M. & Goris, R. L., 2023. Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour,* pp. 142-54.

Bubeck, S. et al., 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303,* p. 12712.

Burri, T. & von Bothmer, F., 2021. The new EU legislation on artificial intelligence: a primer. *Available at SSRN 3831424.*

Chen, C. et al., 2022. A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. *Decision Support Systems,* p. 113647.

Chen, L., Diao, L. & Sang, J., 2018. Weighted evidence combination rule based on evidence distance and uncertainty measure: An application in fault diagnosis. *Mathematical Problems in Engineering.*

Dellermann, D., Ebel, P., Söllner, M. & Leimeister, J. M., 2019. Hybrid intelligence. *Business & Information Systems Engineering,* pp. 637-643.

Dempster, A., 2008. *Upper and lower probabilities induced by a multivalued mapping.* Berlin, Heidelberg: Springer.

Dubois, D. & Prade, H., 1988. Representation and combination of uncertainty with belief functions and possibility measures. *Computational intelligence,* pp. 244-264.

Dwivedi, Y. et al., 2023. So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management,* pp. 71, p.102642.

Eriksson, L. & Hájek, A., 2007. What are degrees of belief?. *Studia Logica,* pp. 183-213.

Feigenbaum, E. A., 1980. *Knowledge engineering: The applied side of artificial intelligence,* s.l.: STANFORD UNIV CA DEPT OF COMPUTER SCIENCE. (No. STAN-CS-80-812)..

Forsythe, D. E. & Buchanan, B. G., 1989. Knowledge acquisition for expert systems: Some pitfalls and suggestions. *IEEE Transactions on Systems, Man, and Cybernetics,* pp. 435-442.

Fountaine, T., McCarthy, B. & Saleh, T., 2019. Building the AI-powered organization. *Harvard Business Review,* pp. 62-73.

Fuster, A., Plosser, M., Schnabl, P. & Vickery, J., 2019. The role of technology in mortgage lending. *The Review of Financial Studies,* pp. 1854-1899.

Garb, H. N. &. S. C. J., 1996. Judgment research and neuropsychological assessment: A narrative review and meta-analyses. *Psychological Bulletin,* p. 140.

Glikson, E. & Woolley, A. W., 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals,* pp. 627-660.

Goertzel, B., 2014. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence,* p. 5(1):1.

Grønsund, T. & Aanestad, M., 2020. Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems,* p. 101614.

Harris, T., 2015. Credit scoring using the clustered support vector machine. *Expert Systems with Applications,* pp. 42(2), 741-750.

Jousselme, A. L.; Grenier, D.; Bossé, É., 2001. A new distance between two bodies of evidence. *Information fusion,* pp. 91-101.

Jousselme, A. L. & Maupin, P., 2012. Distances in evidence theory: Comprehensive survey and generalizations. *International Journal of Approximate Reasoning,* pp. 118-145.

Kahneman, D., 2011. *Thinking, fast and slow.* s.l.:macmillan.

Kahneman, D; Klein, G, 2009. Conditions for intuitive expertise: a failure to disagree.. *American psychologist,* p. 515.

Kahneman, D., Rosenfield, A. M., Gandhi, L. & Blaser, T., 2016. Hidden Cost of Inconsistent Decision Making. *Harvard Business Review.*

Koran, L. M., 1975. The Reliability of Clinical Methods, Data and Judgments: (First of Two Parts). *New England Journal of Medicine,* pp. 642-646.

Kowalewski, O. & Pisany, P., 2022. Banks' consumer lending reaction to fintech and bigtech credit emergence in the context of soft versus hard credit information processing. *International Review of Financial Analysis,* p. 102116.

Leong, C. K., 2016. Credit risk scoring with bayesian network models. *Computational Economics,* pp. 423-446.

Leong, C. K., 2016. Credit risk scoring with bayesian network models. *Computational Economics,* pp. 423-446.

Levi, K., 1989. Expert systems should be more accurate than human experts: evaluation procedures from human judgement and decision making. *IEEE Transactions on Systems, Man, and Cybernetics,* pp. 647-657.

Litvinova, A., Kurvers, R., Hertwig, R. & Herzog, S., 2019. When experts make inconsistent decisions. *PsyArxiv Preprint* .

Liu, X., Sachan, S., Yang, J.-B. & Xu, D.-L., 2019. *Maximum Likelihood Evidential Reasoning-Based Hierarchical Inference with Incomplete Data.* s.l., IEEE, pp. 1-6.

Luo, C., Wu, D. & Wu, D., 2017. A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence,* pp. 465-470.

Luo, C., Wu, D. & Wu, D., 2017. A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence,* pp. 465-470.

Lusk, C. M., Stewart, T. R., Hammond, K. R. & Potts, R. J., 1990. Judgment and decision making in dynamic tasks: The case of forecasting the microburst.. *Weather and Forecasting,* pp. 627-639.

Metawa, N., Hassan, M. K. & Elhoseny, M., 2017. Genetic algorithm based model for optimizing bank lending decisions. *Expert Systems with Applications,* pp. 75-82.

Murphy, C., 2000. Combining belief functions when evidence conflicts. *Decision Support Systems,* pp. 1-9.

Peterson, D., 2017. Maximize efficiency: How automation can improve your loan orig- ination process. *Moody's Analytics.*

Rubin, D. B., 1976. Inference and missing data. *Biometrika,* pp. 581-592.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence,* pp. 206-215.

Sachan, S., 2022. *Fintech Lending Decisions: An Interpretable Knowledge-Base System for Retail and Commercial Loans.* s.l., Cham: Springer International Publishing, pp. 128-140.

Sachan, S., Almaghrabi, F., Yang, J. B. & Xu, D. L., 2021. Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: An application on healthcare and finance. *Expert Systems with Applications,* p. 115597.

Sachan, S. et al., 2023. A Blockchain Framework in Compliance with Data Protection Law to Manage and Integrate Human Knowledge by Fuzzy Cognitive Maps: Small Business Loans. *IEEE International Conference on Blockchain and Cryptocurrency (ICBC) ,* pp. 1-4.

Sachan, S. et al., 2020. An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications,* p. 113100.

Sachan, S. et al., 2020. An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications,* p. 113100.

Sechidis, K., Tsoumakas, G. & Vlahavas, I., 2011. On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD,* pp. 145-158.

Shadbolt, N. R., Smart, P. R., Wilson, J. & Sharples, S., 2015. Knowledge elicitation. *Evaluation of human work,* pp. 163-200.

Shafer, G., 1976. *A mathematical theory of evidence.* s.l.:Princeton university press.

Shafer, G., 1976. A mathematical theory of evidence. *Princeton university press.*

Shan, G., Zhou, L. & Zhang, D., 2021. From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems,* p. 113513.

Smarandache, F., Dezert, J. & Tacnet, J. M., 2010. *Fusion of sources of evidence with different importances and reliabilities.* s.l., IEEE, pp. 1-8.

Smets, P. & Kennes, R., 1994. The transferable belief model. *Artificial intelligence,* pp. 191-234.

Soldatenko, D. M., 2020. Artificial Intelligence: Past, Present and Future. *Russian Foreign Economic Journal,* pp. 127-134.

Stewart, T. R. et al., 1989. Analysis of expert judgment in a hail forecasting experiment. *Weather and forecasting,* pp. 24-34.

Tomczak, J. M. & Zięba, M., 2015. Classification restricted Boltzmann machine for comprehensible credit scoring model. *Expert Systems with Applications,* pp. 42(4), 1789-1796.

Van Someren, M. W. & Verdenius, F., 1998. Introducing inductive methods in knowledge acquisition by divide- and-conquer. *AAAI,* pp. 20-28.

Wagner, W. P., Otto, J. & Chung, Q. B., 2002. Knowledge acquisition for expert systems in accounting and financial problem domains. *Knowledge-Based Systems,* pp. Knowledge-Based Systems.

Wilson, H. J. & Daugherty, P. R., 2019. Creating the symbiotic AI workforce of the future. *MIT Sloan Management Review,* pp. 1-4.

Wu, C. & Wang, X. M., 2000. A neural network approach for analyzing small business lending decisions. *Review of Quantitative Finance and Accounting,* pp. 259-276.

Wu, W. W., 2011. Improving classification accuracy and causal knowledge for better credit decisions. *International Journal of Neural Systems,* pp. 21(04), 297-309.

Xu, D., Zhang, X. & Feng, H., 2018. Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model.. *International Journal of Finance & Economics.*

Xu, D., Zhang, X. & Feng, H., 2019. Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model. *International Journal of Finance & Economics,* pp. 903-921.

Xu, X. et al., 2017. Data classification using evidence reasoning rule. *Knowledge-Based Systems,* Issue 116, pp. 144-151.

Yager, R. R., 1987. On the Dempster-Shafer framework and new combination rules. *Information sciences,* pp. 93-137.

Yang, J. B. et al., 2006. Belief rule-base inference methodology using the evidential reasoning approach-RIMER. *IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans,* pp. 36(2), 266-285.

Yang, J. B. & Xu, D. L., 2013. Evidential reasoning rule for evidence combination. *Artificial Intelligence,* pp. 1-29.

Yang, J.-B. & Xu, D.-L., 2013. Evidential reasoning rule for evidence combination. *Artificial Intelligence,* Issue 205, pp. 1-29.

Yong, D., WenKang, S., ZhenFu, Z. & Qi, L., 2004. Combining belief functions based on distance of evidence. *Decision support systems,* pp. 489-493.

Zhang, B., Dong, Z., Zhang, J. & Lin, H., 2023. Functional network: A novel framework for interpretability of deep neural networks. *Neurocomputing,* pp. 94-103.

Zhao, Z. et al., 2015. Investigation and improvement of multi-layer perceptron neural networks for credit scoring. *Expert Systems with Applications,* pp. 3508-3516.

Zhu, H., Beling, P. A. & Overstreet, G. A., 2002. A Bayesian framework for the combination of classifier outputs. *Journal of the Operational Research Society,* pp. 53(7), 719-727.

Appendix A:

Table 4: Definition of code in criteria in lending rule in Figures 4 and 9

Linking Rule	Status Code	Definition
R1, R2	0_UT	Zero utilization. Missing credit or store card data.
R2	up-to-60	Up to 60% utilization
R1	up-to-70	Up to 70% utilization
R1, R2	full_UT	Full utilization
R3, R4, R6, R7	0_arrears	Account zero payment in arrears
R4	1-2_arrears	Account one to two payments in arrears
R3	1-3_arrears	Account one to three payments in arrears
R6	1-4_arrears	Account one to four payments in arrears
R4	3-5_arrears	Account three to five payments in arrears
R7	3+arrears	More than two payments in arrears

R3	4+arrears	More than three payments in arrears
R6	5+arrears	More than four payments in arrears
R4	6+arrears	More than five payments in arrears
R5	0_payday	Zero number of payday loans
R5	1_payday	One number of payday loan
R5	1+payday	More than one payday loan
R9	0_BK	Zero number of bankruptcies
R9	1_BK	One number of bankruptcies
R9	2+_BK	More than two bankruptcies
R10	0_IVA	Zero individual voluntary arrangements
R10	2_IVA	One or two individual voluntary arrangements
R10	3+IVA	More than two individual voluntary arrangements
R11	0_RO	Zero debit relief order
R11	2_RO	One or two debit relief order
R11	3+RO	More than two debit relief order
R12	0_CS	Zero credit searches
R12	1-4_CS	One to four credit searches
R12	5+CS	More than four credit searches
R13	0_DS	Zero debit searches
R13	1-3_DS	One to three debit searches
R13	4+DS	More than three debit searches
R8	DF	Account in default
R8	AG	Account at the state of agreed payments
R8	AU	Account up-to date
R8	QR	Account in query
R8	RP	Repossession
R8	ST	Account settled
R8	VR	Voluntary repossession
R19	X0	No suspected fraud activity
R19	X2+	More than 2 suspected fraud activity in last 36 months
R19	XX46+	four to five suspected fraud activity in last 48 months
M*	All rules	Missing information

Table 5: Recurrently funded cases

Evidence Group 1	**[1] -**R14 = pass and R18 = within and R19 = X0 and R9 = 0_BK and R10 = 0_IVA and R11 = 0_RO **[2] -** R14 = pass and R18 = within and R19 = X0 and R9 = M* and R10 = M* and R11 = 0_RO **[3] -** R14 = pass and R18 = within and R19 = X0 and R9 = M* and R10 = M* and R11 = M*			
Evidence Group 2	**[1] -** R15 = 250	**[2] -** R15 = 350	**[3] -** R15 = 500	**[4] -** R15 = 600

Evidence Group 3	**[1] -** R1 = (0_UT or up-to-70) and R2 = (0_UT or up-to-60) and R3 = 0_arrears and R4 = 0_arrears and R5 = 0_payday and R6 = 0_arrears and R7 = 0_arrears and R8 = (ST or NAC) and R12 = (0_CS or 1-4_CS) and R13 = (0_CS or 1-3_CS)		
Evidence Group 4	**[1] -** R16 = driven by	**[2] -** R16 = customer valuation	**[3] -** R16 = valuer valuation
Evidence Group 5	**[1] -** R17 = (detached or flat or semidetached)		

Table 6: Recurrently rejected cases

Group 1	**[1] -** R14 = pass and R18 = within and R19 = X0 and R9 = 0_BK and R10 = 2_IVA		
Group 2	**[1] -** R11 = M*	**[2] -** R11 = 2_RO	
Group 3	**[1] -** R1: (M* or 0_UT or up-to-70) and R2: (M* or 0_UT or up-to-60) and R12:(0_CS or 1-4_CS M* or) and R13:(0_CS or 1-3_CS or M*)		
Group 4	**[1] -** R15 = 0	**[2] -** R15 = 250	**[3] -** R15 = 350
Group 5	**[1]** R3: 1-3_arrears and R4: 0_arrears and R5: 0_payday and R6: 0_arrears and R7: 1-2_arrears and R8: (ST or NAC)		
	[2] - R3: 0_arrears and R4: 0_arrears and R5: 1_payday and R6: 0_arrears and R7: 1-2_arrears and R8: (ST or NAC)		
	[3] - R3: 0_arrears and R4: 3-5_arrears and R5: 1_payday and R6: 0_arrears and R7: (0_arrears or 1-2_arrears) and R8: (ST or NAC)		
	[4] - R3: 0_arrears and R4: 0_arrears and R5: 0_payday and R6: 1-3_arrears and R7: 1-2_arrears and R8: AG		
	[5] - R3: 0_arrears and R4: 1-2_arrears and R5: 1_payday and R6: 1-3_arrears and R7: 0_arrears and R8: QR		
	[6] - R3: 0_arrears and R4: 0_arrears and R5: 1_payday and R6: 1-4_arrears and R7: (0_arrears or 1-2_arrears) and R8: QR		
	[7] - R3: 0_arrears and R4: 0_arrears and R5: 0_payday and R6: 0_arrears and R7: 0_arrears and R8: (ST or NAC)		
Group 6	**[1] -** R16: (customer valuation or driven by) and R17: (detached or flat or semidetached)		

Table 7 provides a detailed overview of the number of samples associated with each lending rule. Each lending rule encompasses data for 5700 loan applications segregated into "Samples (Fund)" and "Samples (Reject)".

Table 7: Descriptive Statistics of Samples per Lending Rule

Linking Rule	Status Code	Samples (Fund)	Samples (Reject)
R1	0_UT	3192	2508
R1	up-to-70	4109	1590
R1	full_UT	4075	1624
R2	0_UT	3448	2251
R2	up-to-60	4018	1681
R2	full_UT	4018	1681
R3	0_arrears	4115	1584
R3	1-3arrears	4115	1584
R3	4+arrears	0	5700
R4	0_arrears	4104	1710
R4	1-2_arrears	3534	2166
R4	3-5_arrears	2280	3420
R4	6+arrears	1527	4172
R5	0_payday	4126	1573
R5	1_payday	4126	1573
R5	2+payday	0	5700
R6	0_arrears	3420	2280
R6	1-4_arrears	2850	2850
R6	5+arrears	0	5700
R7	0_arrears	4075	1624
R7	1-2_arrears	3135	2565
R7	3+arrears	0	5700
R8	ST	4104	1596
R8	NAC	4104	1596
R8	AG	3705	1994
R8	QR	0	5700
R8	DF	0	5700
R9	0_BK	3932	1710
R9	1_BK	2280	3420
R9	2+_BK	0	5700
R10	0_IVA	4104	1596
R10	2_IVA	3876	1824
R10	3+IVA	0	5700
R11	0_RO	4104	1596
R11	2_RO	3305	2394
R11	3+RO	0	5700
R12	0_CS	4275	1425
R12	1-4_CS	3989	1710
R12	5+CS	3989	1710
R13	0_DS	3989	1710
R13	1-3_DS	3192	2508
R13	4+DS	3989	1710
R14	pass	4172	1527

R14	fail	0	5700
R15	0	2565	3135
R15	250	4446	1254
R15	350	3705	1994
R15	500	4446	1254
R15	600	5016	684
R16	driven by	3705	1994
R16	customer valuation	3705	1994
R16	valuer valuation	5424	275
R17	detached	4092	1607
R17	flat	4560	1140
R17	semi-detached	4092	1607
R17	other	4092	1607
R18	within	3135	2565
R18	outside	946	4753
R19	X0	4098	1601
R19	X2+	0	5700
R19	XX46+	0	5700

Appendix B:

Table 8: Average accuracy metrics of 5-fold validation set and test set after four iterations of consistency analysis

Data	**Model**	Average accuracy metrics of the 5-fold validation set			Test Set (Independent test set for the final evaluation)		
		Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Balanced (SMOTE)	ER-X	0.930	0.920	0.925	0.928	0.918	0.923
	DNN	0.939	0.908	0.923	0.937	0.905	0.921
	Decision Tree	0.893	0.883	0.888	0.890	0.880	0.885
	Logistic Regression	0.868	0.858	0.863	0.865	0.855	0.860
Imbalanced	ER-X	0.877	0.926	0.901	0.875	0.923	0.899
	DNN	0.880	0.893	0.886	0.877	0.890	0.883
	Decision Tree	0.869	0.859	0.864	0.866	0.856	0.861
	Logistic Regression	0.848	0.838	0.843	0.845	0.835	0.840

Table 9: Hyper-parameters of deep neural network (DNN)

Hyper-parameters	**Imbalanced Data (Original Data)**	**Balanced Data (SMOTE)**
Number of hidden layers (L)	5	4

	$\{L_1 = 62,$ L_2 to $L_4 = 80, L_5 = 2\}$	$\{L_1 = 62,$ L_2 to $L_3 = 80, L_4 = 2\}$
Activation function	-*ReLu*: $L_1 to\ L_4$ -*SoftMax* in output layer	-*ReLu*: $L_1 to\ L_3$ -*SoftMax* in output layer
Dropout rate	10% at L_5	15% at L_4
Batch size	100	100
Epoch	100	100
Regularization strength	L^2regularization strength = 0.01 in each layer	L^2regularization strength = 0.01 in each layer
Learning rate	0.001	0.001

* L denotes a layer in deep neural network. Layer (L_1) is input layer and number of units in first layer. The missing data was imputed by missForest an approach based on random forest algorithm. Default values were set for other hyper-parameters.

Table 10: Hyper-parameters of decision tree

Hyper-parameters	**Imbalanced Data (Original Data)**	**Balanced Data (SMOTE)**
Maximum depth of the tree	8	5
Measure the quality of a split	gini	gini
Minimum number of samples to split node	2	2
Maximum number of leaf nodes	9	7

*Default values were set for other hyper-parameters

Table 11: Hyper-parameters of logistic regression

Hyper-parameters	**Imbalanced Data (Original Data)**	**Balanced Data (SMOTE)**
Regularization Type	L2 (Ridge)	L2 (Ridge)
Regularization Strength (λ)	0.01	0.02
Optimization Algorithm	Stochastic Gradient Descent (SGD)	Stochastic Gradient Descent (SGD)
Convergence Tolerance	0.0001	0.0001
Solver	L-BFGS	L-BFGS

*Default values were set for other hyperparameters. The optimization algorithm is based on a quasi-Newton method that approximates the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm, which belongs to quasi-Newton methods. The regularization strength (λ) was selected using 5-fold cross-validation on a validation set.