Bridging AI and Ethics: An LLM-Based Framework for Transparent and Inclusive Credit Decisions

Marcelo Massashi Simonae[1 2], **Marlon Marcon**[1], **Dalcimar Casanova**[2]

[1]Programa de Pós-Graduação em Informática (PPGI-CP-DV),
Universidade Tecnológica Federal do Paraná (UTFPR) Dois Vizinhos – PR – Brasil
Estrada para Boa Esperança, Km 04 - Dois Vizinhos – 85660-000

[2]Programa de Pós-Graduação em Eng. Elétrica e de Computação (PPGEEC-PB)
Universidade Tecnológica Federal do Paraná (UTFPR) Pato Branco – PR – Brasil
Via do Conhecimento, Km 1 – Pato Branco – PR CEP 85503-390

msimonae@alunos.utfpr.edu.br, {marlonmarcon, dalcimar}@utfpr.edu.br

Abstract. ***1) Research Context:*** *The integration of advanced Machine Learning (ML) models into Intelligent Information Systems (IS) has created highly accurate but opaque "black-box" systems, especially in sensitive domains like credit scoring.* ***2) Scientific and/or Practical Problem:*** *This opacity undermines user trust, can perpetuate algorithmic bias, and challenges regulatory compliance (e.g., LGPD, GDPR). This creates a critical gap between AI's technical power and the socio-technical need for accountability in IS.* ***3) Proposed Solution and/or Analysis:*** *We propose and validate a two-layer framework that uses Large Language Models (LLMs) to translate technical outputs from Explainable AI (XAI) methods, like SHAP and LIME, into actionable, natural language narratives for non-expert users.* ***4) Related IS Theory:*** *Grounded in Decision Support Systems (DSS) theory, this work extends the classical DSS goal. It enhances decision quality not just via predictive accuracy, but by improving the transparency, trustworthiness, and interpretability of the system's reasoning for stakeholders.* ***5) Research Method:*** *We conducted an applied, experimental study on a public retail credit dataset. The methodology involved data preprocessing, XGBoost predictive modeling, quantitative evaluation of explanation fidelity with the MEMC metric, and developing a functional web prototype.* ***6) Summary of Results:*** *The framework effectively identified key credit denial factors with high fidelity, validated by the MEMC metric. The LLM-synthesis layer successfully transformed complex XAI data into clear, understandable, and practical explanations, enhancing the system's clarity and actionability.* ***7) Contributions and Impact to IS area:*** *This study contributes a validated framework for building more ethical, transparent, and socially inclusive intelligent systems. Its impact lies in bridging the gap between advanced AI and human-centric requirements, enabling responsible AI adoption and strengthening human-AI collaboration in decision-making.*

1. Introduction

Credit granting is a cornerstone of the modern economy, playing a crucial role in fostering consumption, investment, and business development. In the retail sector, credit analysis is a complex and strategic process to mitigate delinquency risks while optimizing sales volume and enhancing the customer experience. [Pedroso et al. 2019].

Historically, credit decisions employed traditional statistical methods and predefined heuristic rules. Although effective in specific contexts, these systems often showed significant limitations in their ability to process large volumes of data agilely and to identify the complex, non-linear patterns inherent in consumer risk profiles [Hand and Henley 1997].

The rise of Artificial Intelligence (AI) and, particularly, Machine Learning (ML), has revolutionized the approach to credit analysis. Advanced predictive models have become indispensable tools, including algorithms like logistic regression, decision trees, random forests, Support Vector Machines (SVM), and artificial neural networks. These algorithms can discern complex relationships among the variables that make up a credit applicant's profile, generating more accurate predictions about the probability of a customer honoring their financial commitments. [Bishop 2006, Faceli et al. 2011].

Modern AI-driven Information Systems (IS), especially those used in automated decision-making processes, have raised substantial concerns regarding their transparency and interpretability [Caires and Toledo 2022, Demajo et al. 2020, Jammalamadaka and Itapu 2022, Ribeiro et al. 2016, Rodrigues and Baranauskas 2021, Rothman 2020, Vilone and Longo 2021]. These systems, often characterized as "black boxes" due to their opacity, present significant challenges to understanding the factors that influence their decisions [Caires and Toledo 2022, Lundberg and Lee 2017, Rothman 2020, Vilone and Longo 2021]. Particularly in the context of credit assessment, the lack of explainability can hinder the understanding of the underlying reasons for a potential refusal, in addition to potentially introducing or perpetuating algorithmic biases [Demajo et al. 2020, Jammalamadaka and Itapu 2022].

Although AI models have shown a high accuracy rate, superior to traditional algorithms, interpreting their individual decisions can be complex [Caires and Toledo 2022, Demajo et al. 2020]. In some applications, this lack of interpretability may be irrelevant; however, in sectors where automated decision-making directly affects the interests of individuals, such as in credit assessment, transparency becomes fundamental [Demajo et al. 2020, Jammalamadaka and Itapu 2022]. In this regard, the Brazilian regulatory framework, through the General Data Protection Law (LGPD), establishes specific rights for personal data subjects [Brasil 2018]. According to Article 20 of the LGPD, data subjects can request a review of decisions made exclusively based on automated processing, including those related to credit assessment [Brasil 2018]. This principle aims to ensure that consumers can access clear information about the criteria used and challenge decisions considered detrimental [Demajo et al. 2020, Vilone and Longo 2021].

In the European context, Article 22 of the GDPR (General Data Protection Regulation) establishes rights and obligations regarding the use of automated systems in decision-making [European Union 2016]. One of the most relevant aspects of this article is the introduction of the right to explanation, which allows individuals to obtain detailed information about inferences produced automatically by a model. Furthermore, individuals can contest recommendations that may negatively impact them in legal, financial, mental, or physical aspects [European Union 2016]. This mechanism aims to mitigate problems related to the propagation of biases in computational models, i.e., learned from unbalanced historical data.

Thus, both the LGPD and the GDPR reinforce the importance of accountability in automated decision-making, seeking to ensure that opaque and potentially discriminatory processes do not harm individuals [Brasil 2018, European Union 2016]. Implementing these principles contributes to a more ethical and fair digital environment, allowing citizens greater control over the use of their personal data.

The search for greater transparency in AI has spurred the development of approaches within the field of Explainable Artificial Intelligence (XAI) [Caires and Toledo 2022, Demajo et al. 2020, Jammalamadaka and Itapu 2022, Rothman 2020]. For an algorithm to be considered part of this discipline, it must observe three fundamental principles: transparency, interpretability, and explainability [Rothman 2020]. These attributes ensure that automated decision-making processes are efficient, ethical, and aligned with regulatory requirements [Demajo et al. 2020, Jammalamadaka and Itapu 2022]. Techniques like LIME (Local Interpretable Model-agnostic Explanations) [Rodrigues and Baranauskas 2021] and SHAP (SHapley Additive exPlanations) [Lundberg and Lee 2017, Mendes 2023] are examples of tools developed in this field to provide this explainability.

Although robust from a technical standpoint, the traditional approach generates outputs such as feature importance charts (SHAP) and local variable contributions (LIME), which are often hermetic for non-specialized users, such as credit analysts or the final client.

Therefore, the general objective of this study is to develop and validate a two-layer framework that combines Explainable Artificial Intelligence (XAI) techniques with Large Language Models (LLMs), aiming to generate actionable and comprehensible explanations for automated credit refusal decisions in the retail sector.

To achieve this, the system collects the structured information and numerical weights generated by these multiple XAI tools and uses them as input context for an LLM. The language model, in turn, is tasked with synthesizing this technical data into a cohesive, contextual, and fluidly comprehensible natural language narrative.

This research directly addresses the Grand Challenges of Information Systems Research in Brazil (GranDSI-BR), particularly contributing to "Intelligent and Ubiquitous Systems" and "Software-Intensive Systems in Society." By proposing a framework for transparent and human-centric AI, this work tackles the challenge of developing intelligent systems that are not only technologically advanced but also ethically aligned and socially responsible, fostering trust and positive impact in their interaction with people and organizations.

2. Theoretical Foundation

We developed this research based on three pillars: the evolution of predictive models in credit analysis, the field of XAI, and the rise of Generative AI, and, following, we explain each one of them.

2.1. Credit Analysis and Predictive Models

Credit analysis uses the credit score, a statistical process that classifies individuals' payment behavior to predict their ability to repay debts. With the advancement of Machine Learning (ML), we can employ algorithms to induce functions capable of predicting customer behavior, whether in prediction (supervised learning) or description (unsupervised learning) tasks.

In the context of supervised classification, a range of algorithms is employed, including Logistic Regression, Decision Trees, Random Forest, XGBoost, Naive Bayes, K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP), each with a distinct balance between complexity and interpretability, as summarized in Table 1.

Table 1. Model Characteristics

Model	Type	Interpretability	Performance
Logistic Regression	Supervised	High	Moderate
Decision Tree	Supervised	High	Variable
Random Forest	Ensemble	Medium	High
XGBoost	Ensemble Boosting	Low	Very High

Before applying these models, Exploratory Data Analysis (EDA) is fundamental to examine and summarize the dataset's characteristics, identifying patterns, anomalies, and relationships between variables [Faceli et al. 2011]. Data preprocessing is equally crucial, involving sampling techniques, handling, transformation, cleaning, and data standardization. The goal is to ensure data quality, optimize model performance, and achieve acceptable results on validation metrics [Faceli et al. 2011].

2.2. Explainability in Machine Learning Models

The expansion of the use of artificial intelligence-based systems, especially in sensitive areas like the financial sector, raises relevant ethical concerns, including potential discriminatory practices, the absence of clear explanations, and a lack of transparency in automated decisions [Jammalamadaka and Itapu 2022]. In this context, the responsible AI approach emphasizes the importance of ensuring fairness and mitigating algorithmic biases, making the interpretability of models a fundamental aspect [Jammalamadaka and Itapu 2022].

To mitigate the opacity inherent in complex models, often described as "black boxes" by [Lundberg and Lee 2017], the field of Explainable Artificial Intelligence (XAI) has emerged. This area aims to make automated decisions more comprehensible, allowing financial institutions to identify the factors that motivate credit refusals and for customers to receive clear justifications for decisions that directly impact them. We can classify approaches in XAI into two main categories:

- **Interpretable models (intrinsic):** These are algorithms whose structure facilitates understanding, such as logistic regression and simple decision trees [Rothman 2020].

- **Post-hoc techniques:** Methods applied to complex models after training, to generate explanations. As illustrated in Figure 1, these techniques can offer local explanations (referring to a single prediction) or global explanations (referring to the overall behavior of the model) [Rothman 2020]. Among the most relevant post-hoc techniques for the explainability of machine learning models, the following stand out:

SHAP (SHapley Additive exPlanations) is an approach based on cooperative game theory that assigns an importance value to each input feature for a specific prediction. The great advantage of SHAP is its solid theoretical basis, which guarantees desirable properties such as consistency and additivity, making it a gold standard for attributing feature contributions [Lundberg and Lee 2017].

Complementarily, LIME (Local Interpretable Model-agnostic Explanations) was applied. This model-agnostic method operates by generating local perturbations around the instance to be explained and then fitting a simple (and therefore interpretable) linear model that approximates the behavior of the complex model in that vicinity. LIME is particularly effective in answering the question: "Which features were most important for this specific decision, assuming local linear behavior?" [Ribeiro et al. 2016].

Although powerful, interpreting output from these techniques requires technical knowledge to be interpreted correctly.

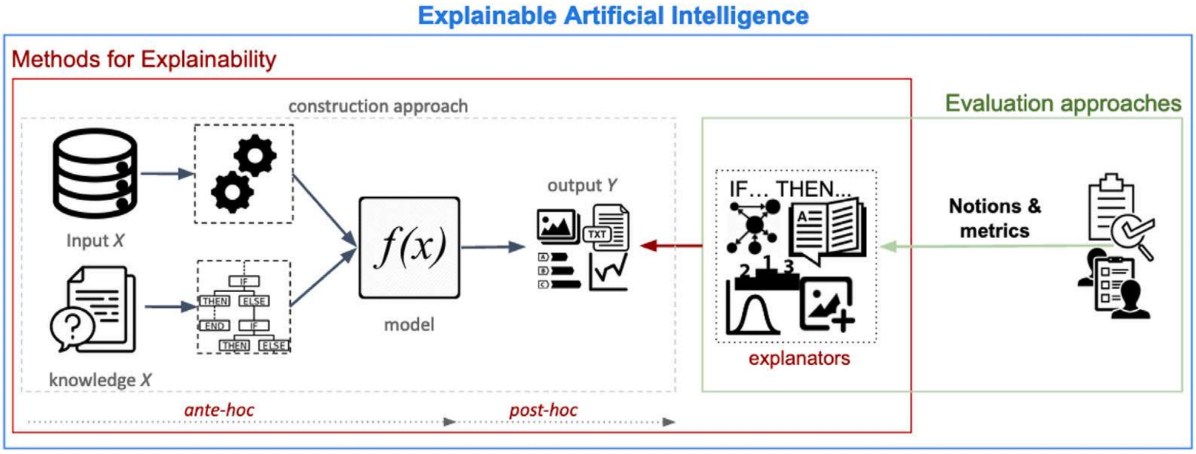


Figure 1. Diagrammatic view of Explainable Artificial Intelligence with interaction between explanation methods and their evaluation approaches.
Source: [Vilone and Longo 2021].

2.3. Evaluation of Explainability and Tools

For XAI models to be transparent, reliable, and valuable, the measurement of explainability is essential. The Mean Evaluation of Metrics Change (MEMC) metric, proposed by [M El-gezawy et al. 2023] offers an objective approach to evaluate the effectiveness of XAI techniques in identifying the most relevant variables for a classification model's decisions.

MEMC quantifies the impact of removing the important variables by XAI techniques on the model's performance, using metrics such as accuracy, precision, recall, and specificity. High MEMC values indicate that the XAI technique adequately represents the model's behavior, since excluding the explained variables causes a significant degradation in performance [M El-gezawy et al. 2023]. The authors define MEMC as:

$$\text{MEMC} = \frac{1}{n}\sum_{i=1}^{n}(f(\mathcal{M}_i) - f(\mathcal{M}_i^*))$$

Where:

- n represents the number of perturbations applied to the input data;
- $\mathcal{M}_i$ denotes the set of evaluation metrics for the original model for the i-th perturbation;
- $\mathcal{M}_i^*$ represents the set of metrics for the model with the explanation removed or perturbed for the same perturbation;
- $f(\cdot)$ is an aggregation function that combines the metrics into a single value, an arithmetic or weighted average.

The selection of algorithms for credit classification problems must, therefore, consider predictive performance and the capacity for explanation. Integrating explainability tools is fundamental to validating results, ensuring regulatory compliance, and promoting practices aligned with Responsible Artificial Intelligence.

2.4. Generative Artificial Intelligence and LLMs

Recently, the advancement of LLMs, such as architectures based on Transformers, has revolutionized the field of Natural Language Processing (NLP). These models are pre-trained on vast text corpora and can perform complex generation, summarization, and translation tasks with remarkable fluency [Vaswani et al. 2017].

Information Systems, explore LLMs to create conversational interfaces, automate report generation, and summarize complex information. Our research is at the forefront, exploring the capacity of an LLM not to create new content from scratch, but to perform a "domain translation" task: converting quantitative and technical XAI data into qualitative and comprehensible prose.

3. Related Work and Research Gap

The evolution of eXplainable Artificial Intelligence (XAI) in the financial sector has established a foundation for transparency, yet a significant gap remains between technical output and human actionability.

3.1. Foundational XAI in Credit Scoring

The theoretical basis for modern interpretability was solidified by [Lundberg and Lee 2017], who introduced SHAP (SHapley Additive exPlanations). Grounded in game theory, SHAP provides a unified framework that satisfies key properties—local accuracy and consistency—assigning a mathematical contribution to each feature. Building on this, [Rodrigues and Baranauskas 2021] explored LIME (Local Interpretable Model-Agnostic Explanations) specifically for credit data, utilizing the Jaccard coefficient to bridge the gap between classification models and human-readable explanations.

3.2. Comprehensive and Regional Approaches

Recent studies have moved toward multi-method architectures. [Demajo et al. 2020] proposed a "360-degree" explanation structure for XGBoost models, combining SHAP, LIME, Anchors, and ProtoDash to bolster user trust. In the Brazilian context, [Caires and Toledo 2022] applied SHAP to a massive dataset of 750,000 consumers, identifying critical variables like "Number of Restrictions" to mitigate operational biases. Parallelly, [Jammalamadaka and Itapu 2022] advanced the "Responsible AI" agenda by integrating bias mitigation techniques, such as Reweighing and Disparate Impact Remover, with automated model balancing.

3.3. Differentiation and Research Gap

Despite these advancements, existing literature reveals three primary limitations that this research aims to address:

- From Technical Data to Communicative Action: While [Demajo et al. 2020] offer robust technical explanations, their outputs (e.g., waterfall plots) are often designed for data scientists, imposing a high cognitive load on lay users. Our work differentiates itself by implementing a generative linguistic layer (LLM) that synthesizes these complex artifacts into personalized natural language narratives, democratizing access for the end applicant.

- Operationalizing the "Right to Explanation": Unlike [Jammalamadaka and Itapu 2022], who focus heavily on pre-processing bias mitigation, our framework focuses on the post-hoc communicative requirements of LGPD and GDPR. We prioritize the user's "Right to Explanation" by ensuring that the rationale for a credit denial is not only fair but also contextually justified and understandable.

- Multi-Method Validation and Reliability: While studies like [Caires and Toledo 2022] rely on single-method interpretations, our research employs a dual-method approach (SHAP and LIME) validated by the MEMC metric [M El-gezawy et al. 2023]. Furthermore, we introduce a "Consistency Guardrail": the LLM synthesis is only triggered when global and local evidences converge, significantly reducing the risk of "hallucinations" or misleading technical interpretations.

3.4. Comparative Analysis

Table 2, presented below, summarizes the comparative analysis of the main related works, highlighting their central focuses, points of convergence with this study, and differentiation aspects.

Table 2. Comparative Analysis of Related Works

Work	Main Focus	Similarities with this study	Differences with this study
Lundberg and Lee (2017)	SHAP as a theoretical basis for XAI	Theoretical basis for the use of SHAP for explaining predictive models	Does not focus on practical application in credit or the Brazilian context
Rodrigues and Baranauskas (2021)	Explainability via LIME with financial data	Use of LIME and concern for comprehensible interpretations	Focus limited to one technique, without comparison with others or use of metrics like MEMC
Demajo et al. (2020)	360° explanation with XGBoost	Use of multiple XAI techniques, including SHAP, LIME, Anchors	Abroad international scope, without the context of Brazilian data
Caires and Toledo (2022)	SHAP with a Brazilian database	Direct applicability of SHAP on national data	Exclusive focus on SHAP, without discussion of fairness or user interfaces
Jammalamadaka and Itapu (2022)	Responsible AI and fairness in credit	Use of SHAP and discussion of algorithmic justice and bias	Greater emphasis on equity than on visual explanation for the user

4. Proposed Framework: ML-XAI-LLM for Humanized Explanations

As presented in previous sections, generating explainable outputs from an AI model is a challenging task, and to overcome such limitations and provide more natural outputs, in this section, we present our proposed framework, built in a two-layer architecture, as illustrated in Figure 2.

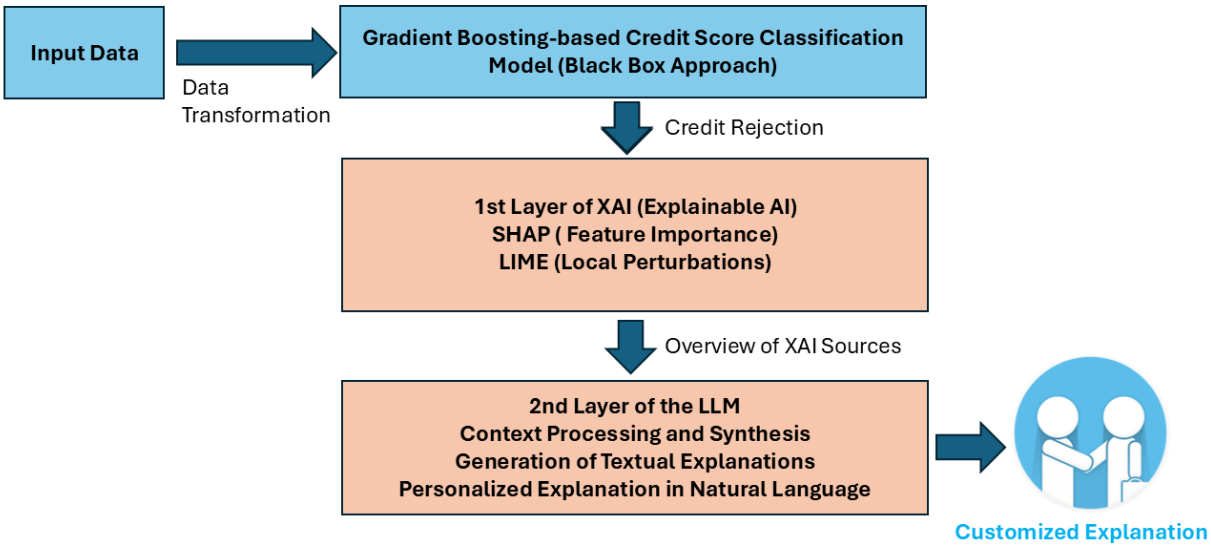

Input Data
Data Transformation
Gradient Boosting-based Credit Score Classification Model (Black Box Approach)
Credit Rejection
1st Layer of XAI (Explainable AI)
SHAP (Feature Importance)
LIME (Local Perturbations)
Overview of XAI Sources
2nd Layer of the LLM
Context Processing and Synthesis
Generation of Textual Explanations
Personalized Explanation in Natural Language
Customized Explanation

Figure 2. Levels of Tailored Explainability for Credit Rejection

4.1. Layer 1: Technical Analysis (Factors Extraction)

In this layer, a predictive ML model (e.g., XGBoost), trained for a specific task, has its decision for a particular instance analyzed by a portfolio of XAI techniques. Using multiple techniques (SHAP and LIME) provides a more robust and diverse view of the determining factors. The output of this layer is a set of structured and semi-structured data, such as SHAP values and weights of LIME's local model. The fidelity and robustness of these explanations were quantitatively validated through the MEMC (Mean Evaluation of Metrics Change) metric, which measures the impact on the model's performance when removing the most important variables identified by each XAI technique, as detailed in the methodology.

4.2. Layer 2: Natural Language Synthesis (Translation and Humanization)

The second layer represents the innovative core of the proposed framework. In it, the outputs generated by the XAI techniques in Layer 1, which are generally complex and difficult for non-technical users to understand, are organized and incorporated into a prompt directed to a Large Language Model (LLM). This prompt instructs the LLM to act as a communication specialist with three main objectives:

1. **Synthesize:** Consolidate the results from the different XAI techniques into a single, clear, and coherent narrative;
2. **Translate:** Explain the determining factors of the credit decision in natural language, avoiding technical terms and jargon;
3. **Advise:** Generate practical and personalized recommendations based on the identified negative factors, aiming to guide the user on improving their eligibility in the future.

The result is a humanized and user-centric explanation that can be directly integrated into the interface of an Information System, promoting greater transparency and trust in the automated decision-making process.

4.3. Quantitative Evaluation of the XAI Layer

The first layer of the framework, dedicated to Explainable Artificial Intelligence (XAI), was quantitatively validated to ensure the robustness and fidelity of the generated explanations. We applied established techniques such as SHAP and LIME to an XGBoost model trained with real retail credit data.

We evaluate qualitatively the output explanations using the MEMC (Mean Evaluation of Metrics Change) metric, which measures the consistency between the explanatory factors and the model's performance. The results confirm the scientific rigor of the approach and ensure that the technical explanations used as input for the LLMs are reliable, interpretable, and suitable for translation into accessible language.

4.4. Ethical AI and Regulatory Compliance

Regulatory Compliance and Hallucination Safeguards: The proposed framework translates abstract regulatory requirements into technical solutions, specifically addressing the "Right to Explanation."

LGPD and GDPR Alignment: The system operationalizes Art. 20 of the Brazilian LGPD [Brasil 2018] and Art. 22 of the GDPR [European Union 2016], which grant data subjects the right to request a review and explanation of automated decisions. By converting opaque probabilities into clear, actionable narratives, the framework ensures transparency and accountability for non-technical stakeholders.

Hallucination Prevention (The 3-Guardrail System): To ensure reliable LLM-generated explanations and prevent distortions ("hallucinations"), we implemented three strict safeguards:

1. Structured Prompt Engineering: The LLM acts strictly as a translator. It receives only validated numerical evidence from Layer 1 (e.g., "SHAP: Income = -0.34") and is explicitly forbidden via system prompts from inferring values not present in the input JSON.
2. Fidelity Guardrail: Only features validated by the MEMC metric (Section 6.2) are passed to the synthesis layer. If a feature's removal does not impact the model (MEMC approximately 0), it is filtered out, ensuring the narrative focuses only on causal factors [M El-gezawy et al. 2023].
3. Consistency Guardrail: The framework requires convergent evidence. A factor is highlighted in the final narrative only if both SHAP (Global) and LIME (Local) agree on its direction of influence, mitigating artifacts from single-method errors [Ribeiro et al. 2016, Lundberg and Lee 2017].

5. Theoretical Contribution

The Transparent Decision Support System (TDSS) This study extends the classical theory of Decision Support Systems (DSS). Traditionally, DSS architectures focus on improving the quality of the decision-maker's choices (e.g., the loan officer or the bank) [Hand and Henley 1997]. We propose a theoretical shift towards a Transparent Decision Support System (TDSS) architecture. In the era of ubiquitous AI, we argue that a DSS must not only support the agent of the decision but also empower the subject of the decision (the applicant). Our contribution to the Information Systems (IS) literature is twofold:

1. Operationalizing Actionable Explainability: We move beyond passive transparency (disclosing the algorithm) to "actionable explainability." Unlike standard XAI outputs which provide static feature weights, our framework generates semantic narratives that inform the user of specific, feasible actions to alter the outcome (e.g., "paying off debt X will increase score by Y").
2. Socio-Technical Governance: We demonstrate how abstract regulatory requirements (LGPD Art. 20, GDPR Art. 22) can be translated into implementable technical components [Brasil 2018, European Union 2016]. By coupling the predictive power of Gradient Boosting with the linguistic synthesis of LLMs, we bridge the semantic gap between statistical probability and human reasoning, contributing to the domain of Ethical AI Governance.

6. Case Study: Explainability in Credit Analysis

It was applied to an information system for retail credit scoring to validate the feasibility and value of our framework.

6.1. Experimental Methodology and Predictive Modeling

6.1.1. Dataset, Pre-processing and Feature Engineering

We obtained the dataset used in this research from a public repository provided by the "Nerd dos Dados" channel [dos Dados 2023], containing 10,476 records of credit applications with 17 attributes spanning socioeconomic profiles and financial history. Table 3 summarizes the main fields.

During data preparation, the following adjustments were implemented:

- **Outlier treatment:** Records with more than four children were removed to prevent distortion. High-value property and vehicle variables were retained due to their critical role in credit risk assessment.

Table 3. Data Description. Source: Elaborated for this research.

Attribute	Type	Description
CODIGO_CLIENTE	Numeric	Unique client identifier
UF	Categorical	Federative Unit of the client
IDADE	Numeric	Client's age in years
ESCOLARIDADE	Categorical	Level of education
ESTADO_CIVIL	Categorical	Client's marital status
QT_FILHOS	Numeric	Number of children
CASA_PROPRIA	Categorical	Indicates if the client owns a home (Yes/No)
QT_IMOVEIS	Numeric	Number of properties
VL_IMOVEIS	Numeric	Total value of properties
OUTRA_RENDA	Categorical	Indicates if the client has another source of income (Yes/No)
OUTRA_RENDA_VALOR	Numeric	Value of the other income
TEMPO_ULTIMO_EMPREGO_MESES	Numeric	Time at last job (in months)
TRABALHANDO_ATUALMENTE	Categorical	Indicates if the client is currently employed (Yes/No)
ULTIMO_SALARIO	Text/Numeric	Last salary (requires conversion to numeric)
QT_CARROS	Numeric	Number of cars
VALOR_TABELA_CARROS	Numeric	Total value of cars
SCORE	Numeric	Credit *Score* (Target Variable)

- **Imputation and Conversion:** Missing values in the *Last Salary* field were addressed using median imputation to preserve distribution. Textual salary data was converted to numeric (float) format.

- **Feature engineering:** Age was segmented into four categorical ranges (up to 30, 31–40, 41–50, and over 50 years) to better capture demographic patterns, removing the original continuous variable to eliminate redundancy.

- **Data encoding (Encoding):** Label Encoding was applied to all categorical variables (e.g., Marital Status, Education Level) to ensure compatibility with Machine Learning algorithms.

6.1.2. Experimental Configuration, Model Performance, Overfitting Analysis

The experimental configuration consisted of an 80/20 train-test split. A stratified 5-fold cross-validation was applied to the training set. Hyperparameter tuning was performed using Grid Search, optimizing the XGBoost algorithm with the following ranges: `learning_rate` (0.01–0.3), `max_depth` (3–10), and `n_estimators` (100–500).

The final model achieved perfect performance across all metrics: Accuracy = 1.0, Precision = 1.0, Recall = 1.0, F1-Score = 1.0, and AUC-ROC = 1.0. The final version of this study will provide detailed hyperparameter settings and a comparative Table 4 of the seven evaluated models.

Table 4. Performance of models on the test set.

Model	Acc.	Precision	Recall	F1	AUC-ROC	Best Hyperparameter
Random Forest	1.00	1.00	1.00	1.00	1.00	`{'max_depth': 10, 'n_estimators': 100}`
XGBoost	1.00	1.00	1.00	1.00	1.00	`{'learning_rate': 0.05, 'max_depth': 3, 'n_estimators': 100}`
KNN	1.00	1.00	1.00	1.00	1.00	`{'n_neighbors': 3, 'weights': 'uniform'}`
MLP	1.00	1.00	1.00	1.00	1.00	`{'alpha': 0.0001, 'hidden_layer_sizes': (50,)}`
Decision Tree	1.00	1.00	1.00	1.00	1.00	`{'max_depth': 10, 'min_samples_leaf': 1}`
Logistic Regression	0.92	0.98	0.80	0.88	0.97	`{'C': 10.0}`
Naive Bayes	0.85	0.82	0.72	0.77	0.89	`{}`

After pre-processing, we evaluated several classification models, including Logistic Regression, Decision Tree, XGBoost, KNN, MLP, Naive Bayes, and Random Forest. Among these, ensemble and tree-based models (Random Forest, XGBoost, KNN, MLP, and Decision Tree) achieved perfect accuracy (1.00) on the test set, correctly classifying all 1,362 approved and 733 rejected instances. Logistic Regression and Naive Bayes obtained slightly lower, yet still satisfactory, results, with Logistic Regression maintaining high precision (0.98).

Although perfect accuracy may indicate strong predictive capability, it also raises concerns about potential overfitting or dataset-specific characteristics. Confusion matrix analysis revealed that Logistic Regression produced 21 false positives and 132 false negatives, while Naive Bayes resulted in 119 false positives and 207 false negatives.

Tree-based models, particularly Random Forest and XGBoost, consistently achieved maximum scores across all metrics in 5-fold cross-validation (accuracy, precision, recall, F1-score, and AUC-ROC). Despite their superior performance, these models are complex to interpret, which poses challenges in sensitive domains such as credit approval, where transparency and justification of automated decisions are legal and ethical requirements.

For this reason, Random Forest and XGBoost were selected as priority candidates for Explainable Artificial Intelligence (XAI) analysis using SHAP and LIME. These techniques enable the identification of key decision factors and potential biases related to socioeconomic variables (e.g., employment duration, salary, and assets). Such interpretive analysis is essential to ensure regulatory compliance, promote fairness, and strengthen user trust in automated decision systems.

Overfitting and Limitations: While an F1-Score of 1.0 indicates perfect classification on this specific dataset, we acknowledge that in real-world scenarios with higher stochastic noise, this could suggest overfitting [Bishop 2006]. The limited dimensionality of the public dataset compared to proprietary banking data contributed to this ceiling effect. However, for the purpose of this study, the "perfect" model serves as an ideal controlled environment to rigorously validate whether the XAI techniques can correctly identify the deterministic logic learned by the model, separating predictive error from explanation error.

6.2. Analysis of Model Robustness via MEMC

Before applying XAI techniques, a comparative robustness analysis was conducted among the trained models using the MEMC (Mean Evaluation of Metrics Change) metric. Figure 3 presents a bar chart comparing the MEMC score among the different algorithms. It illustrates how sensitive each model's performance is to removing its most important features. The vertical axis represents the MEMC value, where a higher value means a greater drop in performance (in this case, in the AUC metric) after the perturbation.

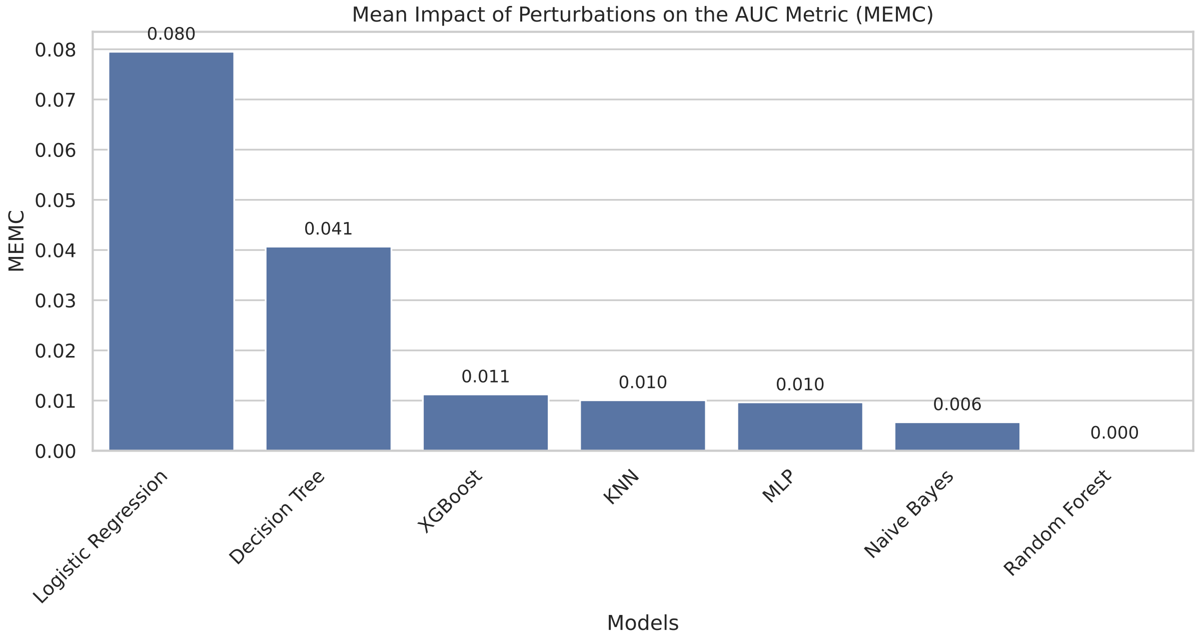


Figure 3. Results of the models by the MEMC metric

The analysis of the results reveals the following highlights:

- **Logistic Regression:** Showed the highest MEMC value (0.080). This indicates that the model suffered the most significant performance drop when removing its most important features, suggesting a strong dependence on a small set of features.
- **Decision Tree:** Was the second most impacted model, with an MEMC of 0.039.
- **XGBoost, KNN, MLP and Naive Bayes:** Showed considerably lower and similar MEMC values (0.011 for XGBoost, 0.010 for KNN, 0.010 for MLP, and 0.006 for Naive Bayes). This suggests that these models are more robust to removing their main features, with the predictive capability more distributed among the variables.
- **Random Forest:** Had the most distinct result, with an MEMC of 0.000. A zero value means that removing the most important features caused no drop in performance. This may indicate that the model is extremely robust or that the XAI technique used to identify the most important features for this specific model may not have been effective.

While Random Forest proved to be the most stable (null MEMC) and Logistic Regression the most sensitive, XGBoost, with its intermediate MEMC score, emerged as an interesting choice as it represents a practical compromise between stability and the ability to detect variations, justifying its selection for the subsequent steps.

6.2.1. Comparative Analysis of the Fidelity of Explainability Methods via MEMC

To validate the fidelity of our explanations, we utilized the Mean Evaluation of Metrics Change (MEMC) metric [M El-gezawy et al. 2023].

Where M represents the evaluation metric (e.g., F1-Score), and n is the number of perturbed instances.

This metric quantifies the performance drop when the most important features identified by an XAI method are masked.To evaluate the robustness of the model and the consistency of the explainers, we applied the following:

1. **Model robustness:** The XGBoost classifier obtained an original F1-Score of 1.0 on the test set, which indicates a perfect fit. As observed in Table 5, this demonstrates high predictive capability, but also raises the possibility of overfitting, an aspect that deserves critical discussion.

2. **Validation by MEMC:** By masking the five most important variables from each explainability technique(see Table 5),the F1-Score dropped from 1.0 to 0.0 in all cases. This behavior reveals that:

   - The identified variables are essential for the model's functioning.

   - Regardless of the technique used, removing these attributes causes the performance to collapse.

   - The four methods (SHAP and LIME) consistently identify key variables.

3. **Differences between explainers:** Despite the identical performance drop (Δ=1.0), each method highlighted slightly different variables. SHAP identified key variables such as Time at last job, last salary, and total value of properties. LIME, in addition to these, brought in attributes such as Indicates if the client has another source of income, the client's marital status, and the Number of children, indicating a more localized and sensitive view of specific instance profiles.

4. **Synthesis:** The results show that different XAI techniques can converge on the central attributes and offer complementary nuances. This diversity enriches the interpretation, allowing for an understanding of both the global structure of the model and relevant local variations.

The validation with the MEMC metric showed that the four analyzed explainability techniques consistently identify the most determinant variables for the credit model. The result demonstrated that the selected variables are indeed crucial for the prediction, validating the fidelity of the explainers.

Table 5. Validation of XAI Techniques by the MEMC Metric.

Techniques	F1 Drop	Original F1	Masked F1	Top 5-Features
SHAP	1.0	1.0	0.0	Time at last job, Last salary, Total value of properties, Total value of cars, Indicates if the client owns a home
LIME (Aggregated)	1.0	1.0	0.0	Last salary, Total value of properties, Time at last job, Total value of cars, Indicates if the client owns a home

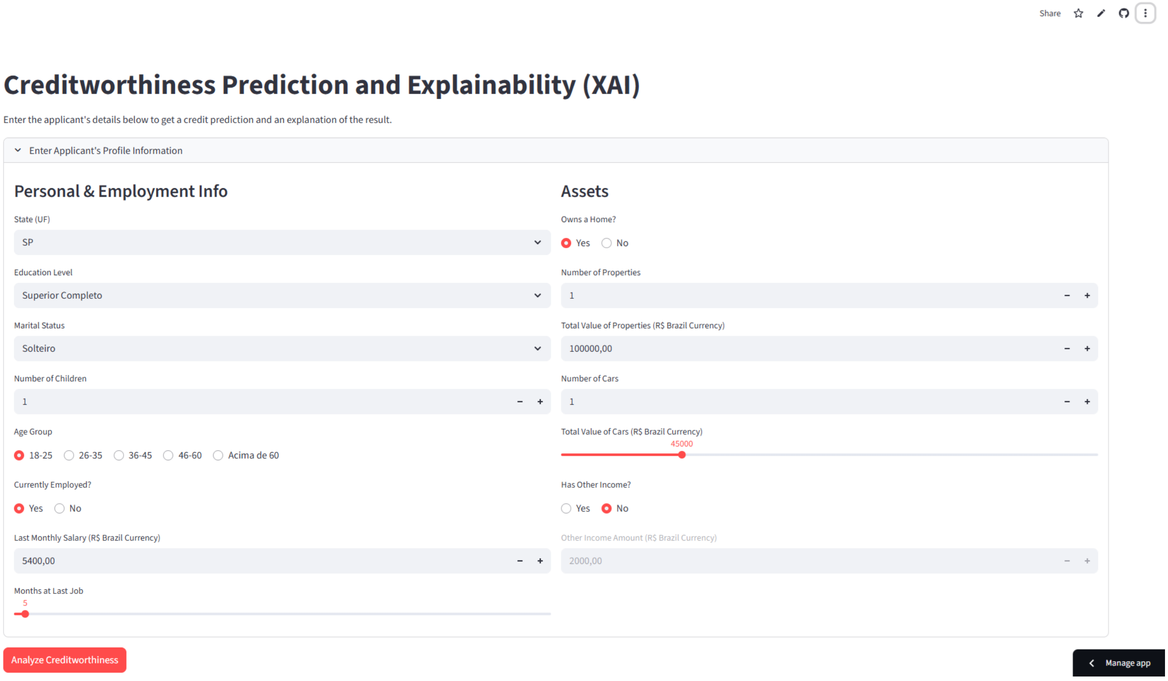
6.3. Implementation and Results of the XAI-LLM Framework

To demonstrate the framework's practical application and validate its results, we conducted a series of simulations using a purpose-built, interactive web interface. This visual tool, developed for this research, allows for the dynamic input of applicant data and provides real-time credit decisions and corresponding explanations. The interface is publicly accessible for exploration and academic review[1].

For this study, we conducted 10 distinct simulations of credit analysis requests. In each simulation, a unique client profile was created by entering 15 distinct input attributes into the interface, reflecting the full scope of data collected by the system, as detailed in Table 5. The underlying XGBoost model then processed these profiles to obtain a decision. We automatically applied the model's prediction displayed on the interface, the two-layer explanation framework.

- **Layer 1 Execution (Technical Analysis):** Execution of the portfolio of XAI tools (SHAP and LIME), for each decision. This layer produced a set of structured outputs, such as feature importance scores and high-precision decision rules, summarizing the technical rationale behind each prediction.
- **Layer 2 Application (Natural Language Synthesis):** The technical outputs from Layer 1 were systematically compiled into a detailed prompt for a Large Language Model (in this prototype, the OpenAI API). The prompt instructed the LLM to synthesize these factors into a coherent, non-technical narrative and provide actionable advice for the applicant, rendered on the user interface.

The results of this process, including the final model decision and the outputs from Layer 1 and Layer 2 for each simulation, are presented in in Figure 5 and Table 7, respectively. At the same time, Figure 4 displays the web application's interface. Figure 5 shows the SHAP feature impact plot and LIME explanations based on a traditional XAI framework, and Figure 6 presents the humanized explanation of the LLM. Table 6 illustrates all simulations that were effectively the converter framework's ability to transform complex technical data into personalized and actionable explanations, fulfilling this research's central objective. The application is publicly available in ***https://sbsi2026xai.streamlit.app/*** and the source code is hosted on Github, in ***https://anonymous.4open.science/r/SBSI2026-7ACF***.



Creditworthiness Prediction and Explainability (XAI)

Enter the applicant's details below to get a credit prediction and an explanation of the result.

Enter Applicant's Profile Information

Personal & Employment Info

State (UF): SP

Education Level: Superior Completo

Marital Status: Solteiro

Number of Children: 1

Age Group: ● 18-25 ○ 26-35 ○ 36-45 ○ 46-60 ○ Acima de 60

Currently Employed? ● Yes ○ No

Last Monthly Salary (R$ Brazil Currency): 5400,00

Months at Last Job: 5

Assets

Owns a Home? ● Yes ○ No

Number of Properties: 1

Total Value of Properties (R$ Brazil Currency): 100000,00

Number of Cars: 1

Total Value of Cars (R$ Brazil Currency): 45000

Has Other Income? ○ Yes ● No

Other Income Amount (R$ Brazil Currency): 2000,00

Analyze Creditworthiness

Figure 4. Web User Interface

Table 6. Client Profile Inputs for Credit Analysis Simulations (with increased row spacing)

ID	State	Education Level	Marital Status	Age Range	Currently Employed?	Owns Home?	Other Income?	Value Other Income (R$)	Children	Properties	Cars	Last Salary (R$)	Property Value (R$)	Car Value (R$)	Time at Job (m)
1	SP	College Degree	Single	18-25	Yes	Yes	No	0	1	1	1	5,400	100,000	45,000	5
2	MG	High School	Divorced	36-45	Yes	No	No	0	1	0	1	4,200	0	35,000	24
3	RJ	College in progress	Married	46-60	Yes	Yes	Yes	2,500	2	2	2	8,500	400,000	75,000	60
4	SP	College Degree	Single	26-35	Yes	Yes	Yes	5,000	0	3	2	15,000	800,000	150,000	12
5	SC	College Degree	Married	46-60	Yes	Yes	No	0	3	2	0	3,100	250,000	0	120
6	RJ	College Degree	Married	36-45	Yes	Yes	No	0	3	1	1	5,500	200,000	40,000	30
7	PR	High School	Single	26-35	Yes	Yes	Yes	4,000	0	1	1	6,000	380,000	60,000	78
8	SP	College Degree	Married	Over 60	Yes	Yes	Yes	3,000	1	2	2	10,200	600,000	120,000	84
9	MG	College Degree	Divorced	36-45	Yes	Yes	No	0	2	1	1	7,000	400,000	25,000	84
10	SP	College Degree	Single	18-25	Yes	No	Yes	4,000	0	0	1	4,500	0	50,000	18

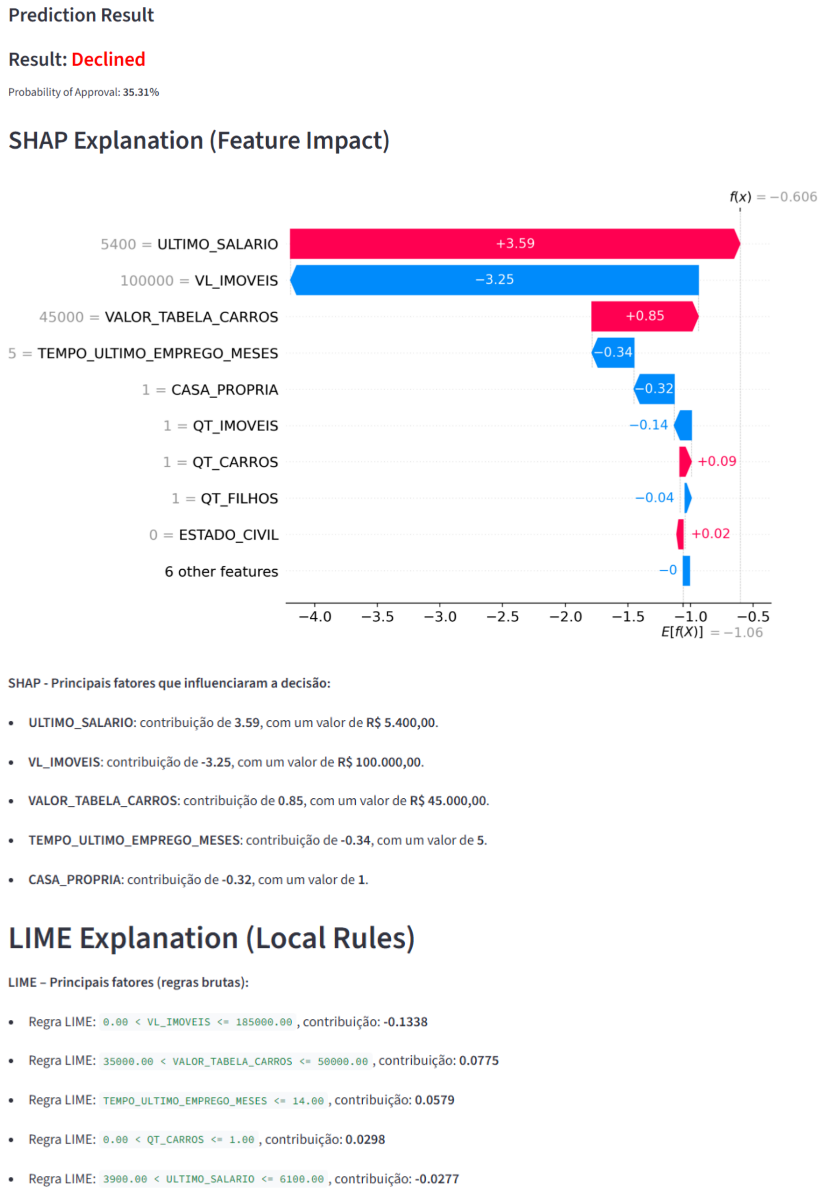
Table 7. Analysis and Explanation Results of the XAI-LLM Framework

ID	Decision	Layer 1 Results (Technical Analysis Summary)	Humanized Explanation Results (LLM Output Summary)
1	Rejected	**SHAP - Key Negative Factors:** Last salary, Total value of properties, Total value of cars, Time at last job, Indicates if the client owns a home. **LIME Rule:** IF Total value of properties <= 185000 and Total value of cars <= 50000 and Time at last job <= 14 and Number of cars <= 1 and and Last salary <= 6100 THEN Rejected.	"**Result Analysis:** The credit model predicted a 'Declined' outcome for your application." "**SHAP Analysis:** ULTIMO_SALARIO: contributed positively with a value of R$ 5.400,00. VL_IMOVEIS: contributed negatively with a value of R$ 100.000,00. VALOR_TABELA_CARROS: contributed positively with a value of R$ 45.000,00. TEMPO_ULTIMO_EMPREGO_MESES: contributed negatively with a value of 5. CASA_PROPRIA: contributed negatively with a value of 1." "**LIME Analysis:** VL_IMOVEIS: Having a property valued between R$ 0,00 and R$ 185.000,00 had a negative impact on the decision. VALOR_TABELA_CARROS: Having a car valued between R$35.000,00 and R$ 50.000,00 had a positive impact on the decision.TEMPO_ULTIMO_EMPREGO_MESES: Having been employed for less than 14 months had a positive impact on the decision. QT_CARROS: Owning between 0 and 1 cars had a positive impact on the decision.ULTIMO_SALARIO: Having a salary between R$ 3.900,00 and R$ 6.100,00 had a negative impact on the decision." **Recommendations:** Consider providing additional information or documentation to support your application, especially regarding the value of your property and salary. If possible, try to increase the time you have been employed at your current job to improve your creditworthiness. Review your financial situation and consider addressing any factors that may have negatively impacted the decision.
2	Rejected	**SHAP - Key Negative Factor:** Time at last job, Total value of properties, Last salary and Total value of cars and Not owning a property . **LIME Rule:** IF Total value of properties <= 0 and Time at last job <= 75 and Total value of cars <= 35000 and Last salary <= 6100 and Number of cars <= 1 THEN Rejected.	"**Result Analysis:** The credit decision for your application has been predicted as 'Declined'." "**SHAP Analysis:** Time Since Last Job: Major negative impact due to extended unemployment (-3.32) No Real Estate Assets: Property value of R$ 0.00 significantly reduced creditworthiness (-2.67). Last Salary: R$ 4,200.00 — below ideal range (-1.63). Car Value: R$ 35,000.00 — modest, with a negative impact (-1.09).No Home Ownership: Slight negative effect (-0.27)." "**LIME Analysis:** No Property: Strongest negative contributor (-0.2480). Time Since Last Job (22–75 months): Negative impact (-0.1610). Car Value (<= R$ 35,000.00): Negative impact (-0.0855). More Than One Car: Slight positive impact (+0.0377). Additional Income Below R$ 0.00: Minor positive impact (+0.0268)."**Recommendations:** Consider improving your employment stability by reducing the time since your last job.If possible, acquire real estate assets to strengthen your financial profile.Increase your income level or consider additional sources of income.Reapply for credit after addressing these factors to improve your chances of approval.
3	Approved	**Key Positive Factors:** All financial indicators are strong, including Salary, which offsets the shorter tenure at the previous job. Assets such as the total value of properties, vehicles, and other income sources also contribute positively to the overall financial profile.	"**Result Analysis:**The credit model has predicted that your loan application will be approved. **Key Positive Contributions**: Last Salary: R$ 8,500.00, Property Value: R$ 400,000.00, Car Value: R$ 75,000.00 and Number of Cars: Between 1 and 2 **Negative Contributions:** Time at Last Job: 60 months and Home Ownership: Value of 1 (negatively impacted the model) **Recommendations:** Maintain employment stability Consider diversifying your assets to enhance long-term financial health.

Continued on next page

Table 7 – continued from previous page

ID	Decision	Layer 1 Results (Technical Analysis Summary)	Humanized Explanation Results (LLM Output Summary)
4	Approved	**Key Positive Factors:** Very high impact from all income sources and asset values, which compensated for the Home Ownership: Value of 1 (negatively impacted the model).	"**Result Analysis:** AThe credit model has predicted that your application will be approved. **Key Factors:** High property value (R$ 800,000), Car value (R$ 150,000), Strong salary (R$ 15,000), Recent employment (12 months), Home ownership had a slight negative impact. **Recommendation:** Your assets and income were decisive. To maintain a strong credit profile, focus on stable employment and responsible asset management.
5	Rejected	**SHAP - Key Negative Factor:** Last salary, Total value of properties, Number of cars, Indicates if the client owns a home, . **LIME Rule:** IF Last salary <= 3900 and Time at last job >75 and Total value of cars <= 0 and Total value of properties <= 370000 and Number of cars <= 0 THEN Rejected.	"**Result Analysis:**Your credit application was declined based on the evaluation of your financial profile. Your last recorded salary of R$ 3,100, the total value of your properties at R$ 250,000, and not owning a car were the main factors that negatively impacted your credit score. Additionally, having no car value and a long tenure at your last job (120 months) were considered less favorable in this analysis." **Recommendations:** Your credit was not approved. To improve your chances in the future, consider:Increasing your salary closer to R$3,900.00 Reducing employment duration below 75 months. Raising the value of your cars or properties. These are suggestions to help strengthen your credit profile. We're here to support you however we can.
6	Rejected	**SHAP - Key Negative Factor:** Time at last job, Total value of properties, Total value of cars, Last salary, and Number of cars. **LIME Rule:** IF Last Time at last job >= 75 and Total value of properties <= 370000 and Total value of cars <= 50000 and Last salary <= 6100 and Number of cars <= 1 THEN Rejected.	"**Result Analysis:**Your credit application was declined based on the evaluation of your financial profile. Time at last job: 30 months, slightly below the ideal range of 22 to 75 months. Total value of properties: R$ 200,000, below the favorable bracket of R$ 185,000 to R$ 370,000.Total value of cars: R$ 40,000, at the lower end of the positive range (R$ 35,000 to R$ 50,000). Last salary: R$ 5,500, within but close to the minimum of the adequate range (R$ 3,900 to R$ 6,100). Number of cars: 1, at the lower limit of the favorable range." **Recommendations:** Your credit was not approved. To improve your chances in the future, consider:Extending your current employment duration to fall within the ideal range.Increasing the value of your properties and vehicles. If possible, reducing the number of cars to align with the preferred range. These are suggestions to help strengthen your credit profile. We're here to support you with whatever you need.
7	Approved	**Key Positive Factors:** Very high impact from all income sources and asset values, which compensated for the Home Ownership: Value of 1 (negatively impacted the model).	"**Analysis Result:** Your credit has been approved based on the evaluation of your financial profile. Property value of R$ 380,000, indicating financial stability. Employment duration of 78 months, showing professional consistency. Last salary of R$ 6,000, which contributed positively. According to decision rules, owning property above R$ 380,000, working more than 75 months at your last job, and having a car valued over R$ 60,000 were all favorable. While your salary and number of car were slightly below ideal ranges, they did not prevent approval.**"Recommendations:** Your strong salary, stable employment history, valuable assets, and moderate car ownership have positively influenced the credit decision. It is advisable to maintain or improve these factors to continue securing favorable credit opportunities in the future.
8	Approved	**Key Positive Factors:** Very high impact from all income sources and asset values.	"**Analysis Result** Your credit has been approved based on the evaluation of your financial profile. Car value of R$ 120,000 Property value of R$ 600,000 Employment duration of 84 months Salary of R$ 10,200, within a favorable range Ownership of 2 cars, which also supported the decision These indicators reflected financial stability and responsibility, leading to a positive outcome in your credit evaluation."**Recommendations:** Your strong employment history, valuable assets, and stable income were key factors in the approval decision.Maintain your current financial stability and continue to manage your assets responsibly to strengthen your credit profile further.If you have any questions or need further assistance, please don't hesitate to reach out to discuss your financial goals and how we can support you.
9	Approved	**Key Positive Factors:** Very high impact from all income sources and asset values, which compensated for Total value of cars and the Home Ownership: Value of 1 (negatively impacted the model).	"**Analysis Result** Your credit has been approved based on the evaluation of your financial profile. Property value of R$ 400,000 showed solid asset ownership. Employment duration of 84 months reflected professional stability. Last salary of R$ 7,000 fell within a favorable income range." According to decision rules, having property above R$ 370,000, a salary between R$ 6,100 and R$ 11,500, and long job tenure all contributed positively. Although your car value was modest and you own only one vehicle, these had minimal impact on the final decision. **Recommendations:** Your strong income, stable employment history, and valuable real estate holdings were key factors in the approval decision.Consider reducing the value of your car or increasing the number of cars owned to potentially improve your creditworthiness in future applications.
10	Rejected	**SHAP - Key Negative Factor:** Last salary, Total value of properties, Time at last job and the Home Ownership. **LIME Rule:** IF Total value of properties <= 0 and Time at last job <= 22 and Total value of cars <= 50000 and Number of cars <= 1 and Indicates if the client has another source of income >0 THEN Rejected.	"**Result Analysis:** We understand that a credit denial can be frustrating, so here's a clear summary of the key factors behind the decision: Your last salary of R$ 4,500 had a strong negative impact. Property value reported as R$ 0 also contributed negatively. Employment duration of 18 months was below the ideal range. Additional factors included a car value of R$ 50,000, owning only one car, and having other income, which were all considered unfavorable in this case." **Recommendations:** Based on the analysis, the main factors contributing to the decline are the low value of VL_IMOVEIS and ULTIMO_SALARIO.To improve your credit application in the future, consider increasing your income and assets to meet the lender's criteria. You may also want to provide additional documentation or explanations to support your application and address any concerns raised by the model.



Prediction Result

Result: Declined

Probability of Approval: **35.31%**

SHAP Explanation (Feature Impact)

SHAP - Principais fatores que influenciaram a decisão:

- **ULTIMO_SALARIO**: contribuição de **3.59**, com um valor de **R$ 5.400,00**.
- **VL_IMOVEIS**: contribuição de **-3.25**, com um valor de **R$ 100.000,00**.
- **VALOR_TABELA_CARROS**: contribuição de **0.85**, com um valor de **R$ 45.000,00**.
- **TEMPO_ULTIMO_EMPREGO_MESES**: contribuição de **-0.34**, com um valor de **5**.
- **CASA_PROPRIA**: contribuição de **-0.32**, com um valor de **1**.

LIME Explanation (Local Rules)

LIME – Principais fatores (regras brutas):

- Regra LIME: `0.00 < VL_IMOVEIS <= 185000.00` , contribuição: **-0.1338**
- Regra LIME: `35000.00 < VALOR_TABELA_CARROS <= 50000.00` , contribuição: **0.0775**
- Regra LIME: `TEMPO_ULTIMO_EMPREGO_MESES <= 14.00` , contribuição: **0.0579**
- Regra LIME: `0.00 < QT_CARROS <= 1.00` , contribuição: **0.0298**
- Regra LIME: `3900.00 < ULTIMO_SALARIO <= 6100.00` , contribuição: **-0.0277**

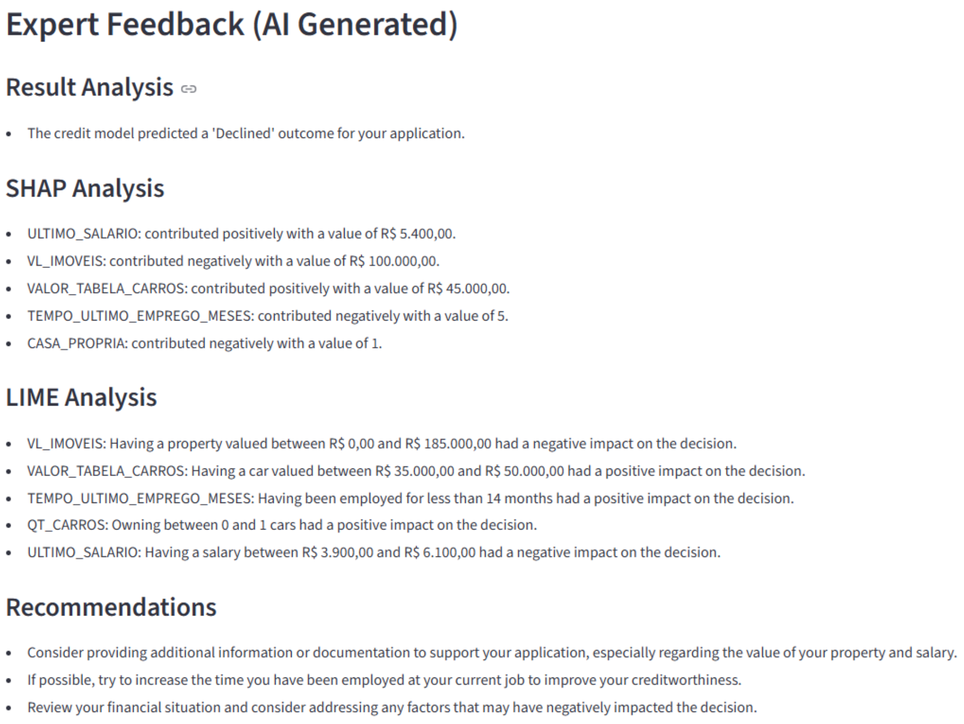
Figure 5. SHAP feature impact plot - SHAP and LIME Explanations

7. Discussion and Critical Analysis of Results

7.1. Synthesis of Results and Domain Translation

The results of the proposed ML-XAI-LLM framework represent a substantial step toward bridging the historical divide between technical explainability and user-centered interpretability. By integrating SHAP and LIME, the system captures a multi-perspective view of the decision rationale, encompassing both global trends and local instance-specific nuances.

This research successfully bridges the gap between complex AI mathematics and human understanding through a process called "domain translation". **The Problem:** Standard AI models (like XGBoost) provide "black-box" decisions and abstract mathematical weights that are impossible for a regular person to understand. **The Solution:** The framework takes technical data from SHAP (which calculates the impact of each variable) and LIME (which identifies local decision rules) and feeds them into a Large



Expert Feedback (AI Generated)

Result Analysis

- The credit model predicted a 'Declined' outcome for your application.

SHAP Analysis

- ULTIMO_SALARIO: contributed positively with a value of R$ 5.400,00.
- VL_IMOVEIS: contributed negatively with a value of R$ 100.000,00.
- VALOR_TABELA_CARROS: contributed positively with a value of R$ 45.000,00.
- TEMPO_ULTIMO_EMPREGO_MESES: contributed negatively with a value of 5.
- CASA_PROPRIA: contributed negatively with a value of 1.

LIME Analysis

- VL_IMOVEIS: Having a property valued between R$ 0,00 and R$ 185.000,00 had a negative impact on the decision.
- VALOR_TABELA_CARROS: Having a car valued between R$ 35.000,00 and R$ 50.000,00 had a positive impact on the decision.
- TEMPO_ULTIMO_EMPREGO_MESES: Having been employed for less than 14 months had a positive impact on the decision.
- QT_CARROS: Owning between 0 and 1 cars had a positive impact on the decision.
- ULTIMO_SALARIO: Having a salary between R$ 3.900,00 and R$ 6.100,00 had a negative impact on the decision.

Recommendations

- Consider providing additional information or documentation to support your application, especially regarding the value of your property and salary.
- If possible, try to increase the time you have been employed at your current job to improve your creditworthiness.
- Review your financial situation and consider addressing any factors that may have negatively impacted the decision.

Figure 6. Human-Centered Explanations Enabled by LLMs

Language Model (LLM). **The Result:** The LLM translates these abstract numbers—such as the -3.25 penalty for property value—into a simple, human-centered explanation: "Your assets do not correspond to the income necessary for credit approval". **The Impact:** This makes credit decisions transparent, provides applicants with actionable advice, and ensures the system follows responsible AI principles.

7.2. Societal Impact and the Global South

This framework holds particular significance for the Global South, specifically Brazil. In a landscape characterized by high interest rate spreads and where credit access is a vital engine for social mobility, the opacity of traditional scoring systems often exacerbates existing inequalities.

By aligning with the Brazilian Grand Challenges in Information Systems (GranDSI-BR)—specifically the challenge of "Software-Intensive Systems in Society"—this framework promotes financial inclusion. It empowers users with limited financial literacy to understand credit barriers, shifting the paradigm from "black-box exclusion" to transparent, informed engagement.

7.3. Limitations and Threats to Validity

Despite the promising results, we acknowledge certain limitations:

- **Data Dependency:** The LLM's performance is inherently tied to the quality of the XAI inputs. Sparse or ambiguous technical data can lead to less informative narratives.

- **Semantic Mapping:** The framework assumes stable semantic relationships between indicators and real-world implications, which may vary across different institutional or cultural contexts.

- **Human Evaluation:** A primary threat to validity is the absence of qualitative testing with human subjects (e.g., loan officers or applicants). Conducting such studies requires strict adherence to ethical protocols (CEP/CONEP in Brazil), which fell outside the current technical validation cycle.

- **Mitigation Strategy:** To address the lack of subjective evaluation, we employed the MEMC (Mean Evaluation of Metrics Change) metric as a rigorous quantitative proxy. The observed collapse in the F1-Score (from 1.0 to 0.0) upon masking key features confirms that the explanations are mathematically faithful to the model. This statistical fidelity is a necessary precondition for future studies focused on perceived trust and user satisfaction.

7.4. Human Oversight and Over-reliance

Finally, it is critical to emphasize that while LLM-generated narratives enhance transparency, they are intended to augment—not substitute—human oversight. Especially in high-stakes financial decisions, these explanations should serve as a complementary tool alongside expert reviews and deeper audits to prevent over-reliance on automated narratives.

8. Conclusions and Future Works

The ML-XAI-LLM framework effectively bridges the gap between technical model outputs and user-friendly explanations in credit decisions. Combining SHAP and LIME captures both global and local factors that influence predictions. The MEMC metric confirms the reliability of these insights.

A key achievement is the LLM's ability to translate complex data into clear, actionable narratives for non-experts. However, its effectiveness depends on the quality of input data, and cultural or institutional differences may affect interpretation. Significantly, while the framework improves transparency, it should complement, not replace, expert analysis in critical financial decisions.

Acknowledgments

The authors thank the funding agencies and the professors who contributed to the realization of this work. It is also important to acknowledge the role of technological tools that assisted in the formal improvement of this document. Tools such as Grammarly, for grammatical and stylistic review; DeepL, for precision in translations between Portuguese and English; and ChatGPT, as an assistant for the reformulation and consistency of ideas, were employed with the intent of refining the linguistic presentation. It is emphasized that such technologies were used as support for revision, maintaining the full integrity and intellectual authorship of all content presented here.

Data, Web Interface and Code Availability

Data, Web Interface and Code Availability To support reproducibility and open science, all artifacts generated in this study are available:

- **Source Code**: The complete Python implementation, including the XGBoost pipeline, XAI configurations (SHAP/LIME), and the MEMC validation script, is hosted at: https://github.com/msimonae/SBSI2026.

- **Dataset:** The utilized dataset is publicly available via the "Nerd dos Dados" repository [dos Dados 2023] or https://github.com/msimonae/SBSI2026.

- **Web Interface:** https://simonaecreditdecision.streamlit.app/

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Brasil (2018). Lei Geral de Proteção de Dados Pessoais (LGPD), Lei nº 13.709/2018. `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709compilado.htm`. Acesso em: 23 maio 2025.

Caires, D. d. O. and Toledo, C. F. M. (2022). Técnicas de interpretabilidade para aprendizado de máquina: um estudo abordando avaliação de crédito. In *Workshop de Matemática, Estatística e Computação Aplicadas à Indústria - WMECAI*. Galoá.

Demajo, L. M., Vella, V., and Dingli, A. (2020). Explainable ai for interpretable credit scoring. Papers, arXiv.org.

dos Dados, C. N. (2023). Canal nerd dos dados. `https://www.youtube.com/watch?v=lK8ANM7VkNU`. Acesso em: 23 maio 2025.

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`. Acesso em: 23 maio 2025.

Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. d. L. F. d. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the royal statistical society: series a (statistics in society)*, 160(3):523–541.

Jammalamadaka, K. R. and Itapu, S. (2022). Responsible ai in automated credit scoring systems. *AI and Ethics*, pages 1–11. Acesso em: 29 maio 2025.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. Acesso em: 29 maio 2025.

M El-gezawy, A. M., Abdel-Kader, H., and Ali, A. H. (2023). A new xai evaluation metric for classification. *IJCI. International Journal of Computers and Information*, 10(3):58–62.

Mendes, M. P. (2023). Análise de explicabilidade em modelos de regressão aplicados a dados imobiliários.

Pedroso, F. S., da Silva, G. E., and Brescian, S. A. T. (2019). Concessão de crédito no setor de varejo: estudo aplicado em quatro supermercados no norte matogrossense. *Observatorio de la Economía Latinoamericana*, (11):9.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. Acesso em: 30 maio 2025.

Rodrigues, R. B. and Baranauskas, J. A. (2021). Explicabilidade utilizando lime: um estudo de caso para o mercado financeiro. Acesso em: 30 maio 2025.

Rothman, D. (2020). *Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps*. Packt Publishing Ltd.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106. Acesso em: 30 maio 2025].