Transparent Credit Scoring Through Explainable AI

Utkarsh Dubey[a], Dushyant Bhardwaj[a], Kanav Singla[a]

[a]*Department of Computer Science and Engineering, Netaji Subhas University of Technology , Dwarka, New Delhi, India*

Abstract

Despite significant progress in credit-risk prediction, many machine learning models still lack interpretability, limiting their use in high-stakes financial decisions. To address this, we propose **X-FuzzyScore**, a transparent ensemble framework that integrates fuzzy logic, machine learning, and explainability techniques for credit-risk assessment. The system combines interpretable fuzzy rule-based reasoning with high-accuracy ensemble models **XGBoost** and **LightGBM** and employs **SHAP** and **LIME** for feature attribution. Trained and evaluated on benchmark datasets including the **UCI German Credit**, **UCI Taiwan Credit Card Default**, and **LendingClub Loan** datasets, the framework follows rigorous preprocessing, normalization, and feature alignment. Experimental results show that X-FuzzyScore achieves superior predictive performance (accuracy, F1-score, AUC) while maintaining interpretability through rule activations and SHAP-based explanations. An interactive visualization dashboard allows stakeholders to explore risk probabilities, linguistic rules, and feature contributions in real time. Overall, X-FuzzyScore offers a robust, explainable solution for credit-risk prediction, enhancing trust and regulatory compliance, with future work focusing on fairness-aware and large-scale real-time extensions.

Keywords: explainable AI, credit-risk, fuzzy logic, machine learning, SHAP, interpretability, XGBoost, LightGBM, credit scoring, transparency

1. Introduction

Credit risk analysis is a cornerstone of modern financial decision-making, directly impacting the stability and profitability of lending institutions. The ability to accurately predict the likelihood of loan default is essential for banks, credit card companies, and fintech organizations. In recent years, the integration of machine learning (ML) and fuzzy logic has enabled the development of more robust, interpretable, and human-centric credit risk models.

This research presents an explainable fuzzy credit-risk prediction framework (X-FuzzyScore) that combines fuzzy reasoning, ML ensemble methods, and explainability techniques (e.g., SHAP) to deliver transparent, actionable insights for credit scoring. The system is designed to:

- Predict credit-risk and loan-default probability for individuals and companies.
- Provide interpretable predictions in human language and visual formats.
- Integrate fuzzy reasoning (e.g., “high income”, “medium debt”) with ML accuracy.
- Offer a dashboard for model results, fuzzy rules, and SHAP explanations.

1.1. Context and Motivation

The project aims to address the lack of transparency in traditional credit scoring models by introducing explainable AI (XAI) techniques. Data preprocessing includes feature alignment, normalization, categorical encoding, and dataset integration.

1.2. System Architecture

The architecture consists of:

1. Data Preprocessing: Feature alignment, normalization, encoding, and integration.
2. Fuzzy Layer: Linguistic variable definition and fuzzy rule application.
3. ML Ensemble Layer: XGBoost, LightGBM, and Random Forest for prediction.
4. Explainability Layer: SHAP/LIME for feature attribution.
5. Visualization Frontend: Dashboard with risk gauge, SHAP bar plots, and fuzzy rule viewer.

1.3. Feature Description

The Taiwan Credit Card Default dataset includes features such as credit limit, sex, education, marriage, age, payment history, bill amounts, and payment amounts. Categorical variables are encoded and outliers are handled during preprocessing. For hybrid models, a fuzzy risk score is added as an additional feature.

1.4. Mathematical Formulation

Let X be the feature matrix, y the target (default status), and f_{ML} the machine learning model. The fuzzy risk score $f_{fuzzy}(X)$ is computed using fuzzy rules:

$$\text{Risk} = f_{ML}(X) + \lambda \cdot f_{fuzzy}(X) \tag{1}$$

where λ controls the influence of fuzzy logic.

1.5. Performance Comparison

Model	**Type**	**Accuracy**	**Precision**	**Recall**	**ROC-AUC**
LightGBM + Fuzzy	Hybrid	0.76	0.46	0.61	0.77
XGBoost + Fuzzy	Hybrid	0.76	0.46	0.61	0.77
RF + Fuzzy	Hybrid	0.78	0.51	0.56	0.77
Random Forest	Baseline	0.78	0.50	0.55	0.77
Logistic Regression	Baseline	0.77	0.49	0.55	0.75

Table 1: Model performance comparison on Taiwan Credit Card Default dataset.

1.6. Expected Outputs

- Probability: 0.87 → 87% chance of repayment
- Risk Label: “Low Risk”, “Medium Risk”, “High Risk”
- Fuzzy Rules Triggered: “IF income = high AND debt = low → risk = low (activation 0.82)”
- SHAP Explanation: income −0.18 → reduced risk; debt +0.07 → increased risk
- Visualization: Dashboard with gauge, SHAP bars, fuzzy memberships

2. Literature Overview

Credit risk assessment is foundational to financial decision-making, and over the decades, a variety of analytical and computational approaches have been developed to predict loan default probability. Traditional models such as Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Decision Trees were widely adopted owing to their interpretability and relatively simple implementation. However, their underlying assumptions—especially linearity and limited capacity to capture complex interactions—restrict their effectiveness on heterogeneous, real-world credit datasets.

With the advent of more powerful Machine Learning techniques, researchers began leveraging models such as Random Forests, Support Vector Machines, Gradient Boosting (particularly XGBoost and LightGBM), and Deep Neural Networks. These approaches have shown substantially better performance on standard credit datasets (e.g., German Credit, Taiwan Credit Card Default, LendingClub), thanks to their ability to automatically learn nonlinear patterns and feature interactions. Nevertheless, such models often operate as "black boxes," which is problematic in high-stakes domains such as credit, where interpretability, auditability, and regulatory compliance are crucial.

To mitigate the interpretability issue, fuzzy logic–based systems have been applied in credit scoring. Fuzzy Inference Systems (FIS) and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) provide rule-based reasoning using linguistic terms such as "low-medium income" or "medium-high risk," thereby increasing transparency. Yet, these systems typically require manually defined membership functions and rules, which limits their scalability and adaptability to large, evolving datasets. For example, Piramuthu's early exploration of neural and neuro-fuzzy methods in credit evaluation *Neural networks and neurofuzzy systems for credit risk evaluation* demonstrated advantages but also highlighted issues with training complexity and generalization.

Meanwhile, the field of Explainable AI (XAI) has matured significantly, offering tools that aim to render black-box models understandable after the fact. Model-agnostic approaches such as LIME and SHAP assign feature importance values and offer instance-level explanations. In particular, SHAP has gained traction due to its theoretically grounded additive feature attribution framework introduced by Lundberg & Lee in their paper *A Unified Approach to Interpreting Model Predictions*, and its growing use in credit contexts such as *Credit Risk Prediction Using Explainable AI*. Recent works like *SHAP and LIME: An Evaluation in Credit Risk* evaluate the discriminative power and stability of these explainability techniques in financial scenarios.

At the intersection of these paradigms, hybrid and integrated frameworks have been explored to strike a balance between accuracy and interpretability. Neuro-fuzzy or fuzzy-boosting models combine data-driven learning with human-readable rules, and some research layers in SHAP or other XAI methods for post-hoc explanation. For example, *A Two-Stage Fuzzy Neural Approach for Credit Risk Assessment in a Brazilian Credit Card Company* illustrates how fuzzy and neural techniques can cooperate. Also, in banking risk modeling, ANFIS has been used to rate risk components in relation to financial performance, as seen in *Rating the Impact of Risks in Banking on Performance*. However, many existing hybrids do not fully integrate modern XAI tools nor evaluate performance across a consistent set of benchmarks. Furthermore, studies like *Explainable Machine Learning in Credit Risk Management* outline ways to make black-box models auditable yet still highlight that achieving interpretability without sacrificing predictive power remains challenging.

In sum, while traditional statistical models offer interpretability and new machine learning methods offer accuracy, and while fuzzy systems introduce human-readability, none fully reconcile all three goals in one framework. Most hybrid works either neglect the integration of XAI, lack consistent benchmarking, or sacrifice one objective for another. This research therefore proposes the **X-FuzzyScore** framework, which unifies fuzzy rule-based interpretability, ensemble model accuracy, and SHAP-based explanation in a single cohesive system to deliver transparent, trustworthy, and high-performing credit risk predictions.

3. Data and Exploratory Analysis

3.1. Dataset Overview

We consider one primary dataset for credit risk modeling: (i)the Statlog (German) Credit dataset (1,000 samples; socio-economic attributes and loan characteristics). The German dataset serves as a compact, interpretable benchmark to validate generalization across data regimes and feature schemas. We follow the UCI conventions for feature definitions and target labeling ($y = 1$ indicates default; $y = 0$ non-default).

The German dataset. contains borrower level information at loan application time, with a binary risk label indicating "good" (non-default) or "bad" (default) credit. Attributes capture: (a) credit terms (duration, credit amount, installment rate), (b) financial standing (status of existing checking/savings accounts, employment length), and (c) personal/economic context (age, housing, job, dependents, telephone, foreign worker). Many predictors are categorical/ordinal codes reflecting risk-relevant categories used by lending institutions.

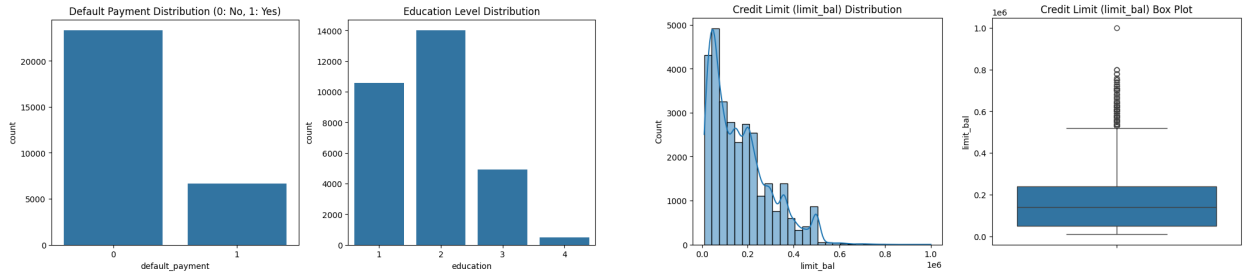
Attribute families.. For clarity, we group features into:

- **Credit terms:** *duration*, *credit amount*, *installment rate*, *number of existing credits*.
- **Banking status:** *checking account status*, *savings account*, *other installment plans*.
- **Demographics socio-economics:** *age*, *employment length*, *housing*, *job*, *telephone*, *foreign worker*, *personal status*.
- **Purpose history:** *credit history*, *purpose* (e.g., car, furniture), and property-related indicators.

In our pipeline, we align label semantics across datasets (`default = 1`), standardize naming, cap heavy-tailed numerics for robustness, and unify encodings (ordinal or one-hot as appropriate) to enable cross-dataset evaluation.

3.2. Exploratory Data Analysis

We summarize key findings from exploratory analysis of the German dataset and connect them to preprocessing choices.



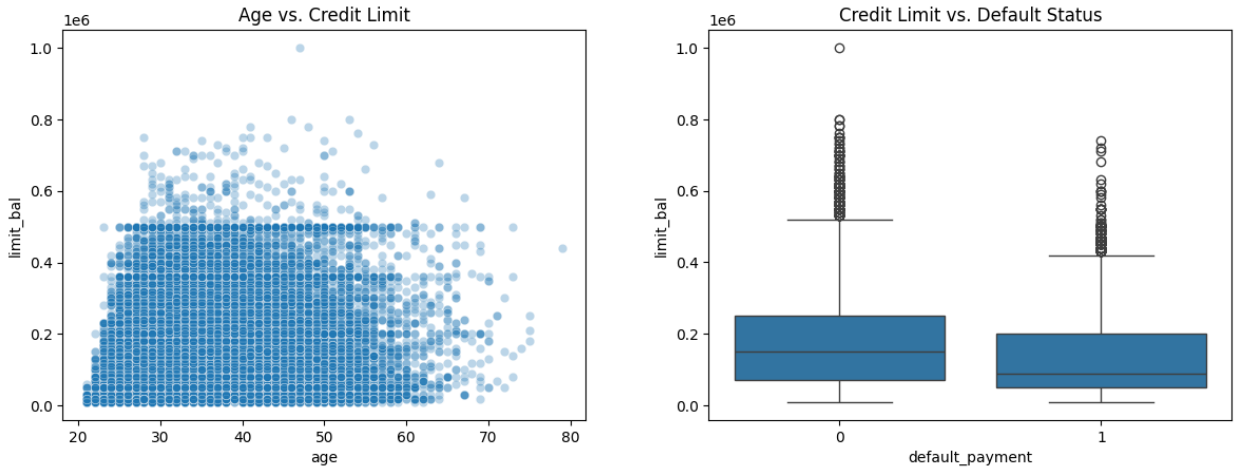
(a) Class balance (default vs. non-default)

(b) Distribution of a key categorical attribute

Figure 1: Target distribution and categorical composition inform thresholding and encoding strategies. Class balance indicates the degree of imbalance to be handled during training and evaluation.

The EDA. proceeds from univariate to bivariate analyses:

1. **Class balance (Fig. 1a).** We quantify the proportion of defaults vs. non-defaults to anticipate metric sensitivity and to guide threshold tuning or class-weighting.
2. **Categorical composition (Fig. 1b).** We inspect distributions of key categorical variables to detect sparse levels that may require grouping and to ensure encodings reflect ordinal structure where present (e.g., account status severity).



(a) Credit amount distribution

(b) Credit amount boxplot

Figure 2: Right-skew and outliers suggest winsorization/capping for numerical stability and to prevent tree-based models from over-partitioning extreme tails.

3. **Numerical distributions (Fig. 2).** Histograms and boxplots reveal right-skew and outliers in *credit amount* and related terms, motivating winsorization/capping to improve numerical stability and calibrate model sensitivity to extremes.
4. **Correlation/association (Fig. 3).** Heatmaps on numeric (and encoded) features identify correlated clusters (e.g., amount–duration) and generally low collinearity across categorical codes. This informs feature selection and regularization decisions.
5. **Target-conditioned checks (not shown).** Where relevant, we examine class-conditional distributions (e.g., default vs. non-default credit amount) to identify features with discriminative power and to validate monotonic tendencies for rule design.

Preprocessing implications.. Based on these observations, we (i) apply consistent categorical encoding with level consolidation where sparse, (ii) cap heavy-tailed numerical variables, (iii) harmonize labels and feature names across datasets, and (iv) retain a compact set of informative variables for cross-dataset comparability. These steps directly support robust training, better calibration, and improved interpretability of fuzzy rules and SHAP explanations in downstream modeling.

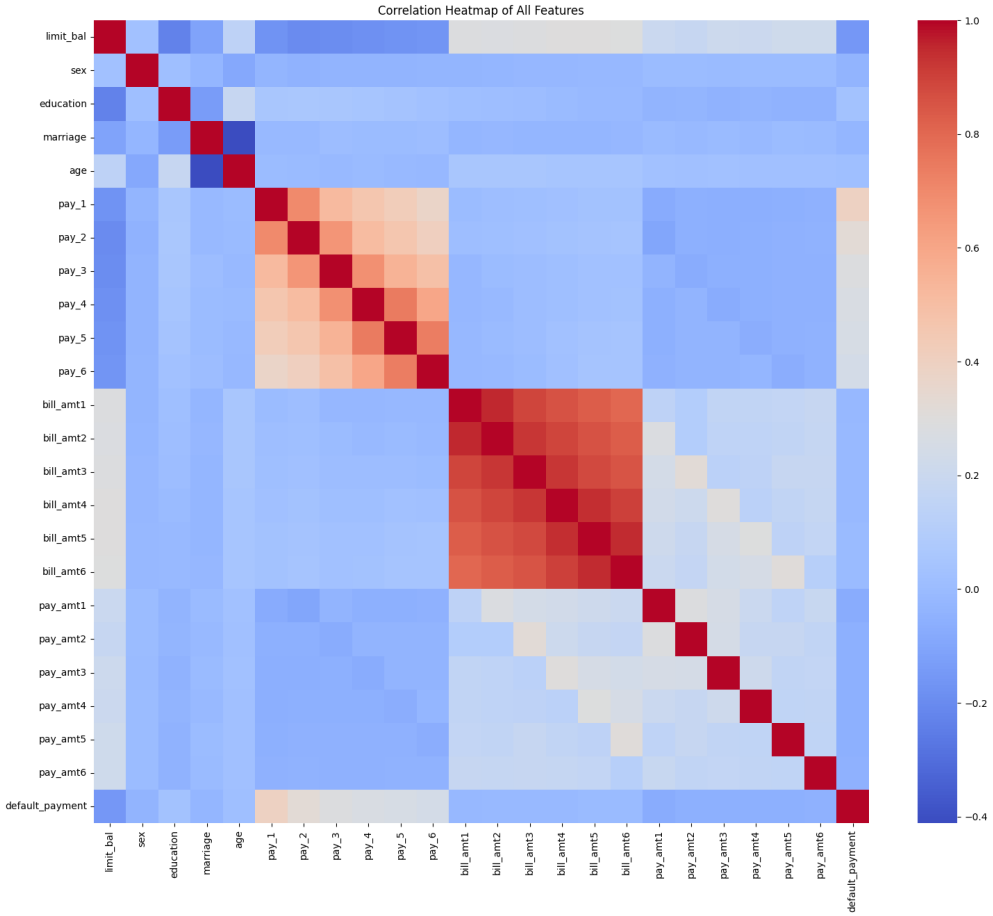


Figure 3: Feature correlation/association matrix (numeric or encoded features). We observe modest associations among credit amount and duration, and weak pairwise associations among most categorical codes, implying limited multicollinearity after encoding.

4. Methodology

This section details the end-to-end methodology of the Explainable Fuzzy Credit-Risk Prediction (X-FuzzyScore) framework, covering data preprocessing, fuzzy inference, machine learning models, hybridization strategies, explainability, and evaluation. We emphasize transparent and reproducible practices to ensure both predictive performance and interpretability.

4.1. Pipeline Overview

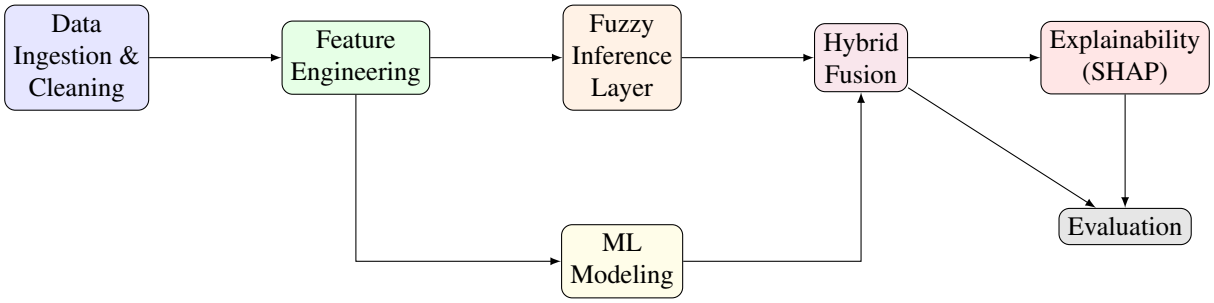


Figure 4: X-FuzzyScore Methodology Pipeline: Data preprocessing, feature engineering, fuzzy inference, ML modeling, hybrid fusion, explainability, and evaluation.

1. Data ingestion and cleaning (schema alignment, recoding, deduplication).
2. Feature engineering (winsorization/capping, scaling if required).
3. Fuzzy inference layer to compute a human-interpretable risk score.
4. ML modeling (Logistic Regression, Random Forest, XGBoost, LightGBM).
5. Hybridization: fuse fuzzy and ML via feature augmentation or late fusion.
6. Explainability: SHAP-based local and global attributions.
7. Evaluation: metrics, threshold selection, and calibration checks.

4.2. Data Preprocessing

We use the UCI Statlog German Credit Data Default dataset as the primary benchmark.[1]

Notation.. We consider a binary classification dataset

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \qquad x_i \in \mathbb{R}^d,\ y_i \in \{0, 1\},$$

where $y_i = 1$ denotes default and $y_i = 0$ non-default. Let

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix} \in \mathbb{R}^{N \times d}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in \{0, 1\}^N.$$

Preprocessing steps (from the EDA notebook):

- **Column normalization:** unify naming (e.g., `default payment next month` → `default_payment`; `PAY_0` → `PAY_1`) and lowercase all columns. *Rationale:* consistent naming prevents errors and improves code maintainability across team workflows.

[1] https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data

- **Categorical recoding:** group undocumented categories into "Others" (`education` $0, 5, 6 \to 4$; `marriage` $0 \to 3$). *Rationale:* sparse or undefined categories cause instability in tree splits and one-hot encodings; consolidation improves robustness and prevents overfitting to noise.
- **Payment status normalization:** map $-2, -1$ to 0 in `pay_1`–`pay_6` to represent no-consumption or paid-in-full as non-delinquency. *Rationale:* original codes -2 (no consumption) and -1 (paid in full) both indicate good standing and should not be treated as missing or distinct risk categories.
- **Deduplication:** drop duplicates after normalization to avoid leakage and inflated support. *Rationale:* duplicate rows artificially inflate training counts and can lead to overfitting or biased performance metrics.
- **Outlier handling (Winsorization):** for heavy-tailed features (`limit_bal`, `bill_amt`, `pay_amt`), apply percentile-based capping at the 1st and 99th quantiles:

$$x_i^{\text{capped}} = \begin{cases} p_1 & \text{if } x_i < p_1, \\ x_i & \text{if } p_1 \le x_i \le p_{99}, \\ p_{99} & \text{if } x_i > p_{99}, \end{cases} \tag{2}$$

  where $p_1 = \text{quantile}(x, 0.01)$ and $p_{99} = \text{quantile}(x, 0.99)$. This obtains features like `limit_bal_capped`, `bill_amt1_capped`, etc.

  Right-skewed distributions (observed in EDA histograms/boxplots) with extreme outliers distort tree-based partitioning and inflate variance estimates. Winsorization reduces skewness (e.g., from ≈ 5.0 to ≈ 0.07 for `bill_amt1`) while preserving rank order and avoiding artificial zero-inflation (unlike median imputation) or negative skew (unlike log transforms). Empirical comparison shows winsorization outperforms: (a) *median replacement*, which creates artificial spikes and loses distributional structure, and (b) *log transformation*, which over-corrects and produces negative skew, especially when data contains zeros.

We use an 80/20 train-test split ($n_{train} = 23971$, $n_{test} = 5993$) with stratification to preserve class proportions. Class imbalance is addressed during evaluation via threshold tuning and by optionally using class-weighted losses in ML models.

4.3. Fuzzy Inference Layer

. The fuzzy layer provides a human-interpretable risk score $s_{\text{fuzzy}} \in [0, 1]$ by encoding domain knowledge through linguistic rules. Unlike black-box ML models, fuzzy systems express risk assessment in natural language (e.g., "IF credit limit is low AND payment status is poor THEN risk is high"), making them auditable and explainable to regulators and stakeholders.

We define *linguistic variables* over selected features (e.g., credit limit, bill amounts, delinquency, payment ratios), each with *membership functions* for labels such as Low, Medium, and High. Each membership function $\mu_\ell : \mathbb{R} \to [0, 1]$ quantifies the degree to which a feature value belongs to the linguistic category ℓ.

4.3.1. Membership Functions

Triangular membership functions.. For a scalar feature x, a triangular membership function for label ℓ is defined as:

$$\mu_\ell(x; a, b, c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right), \ a < b < c, \tag{3}$$

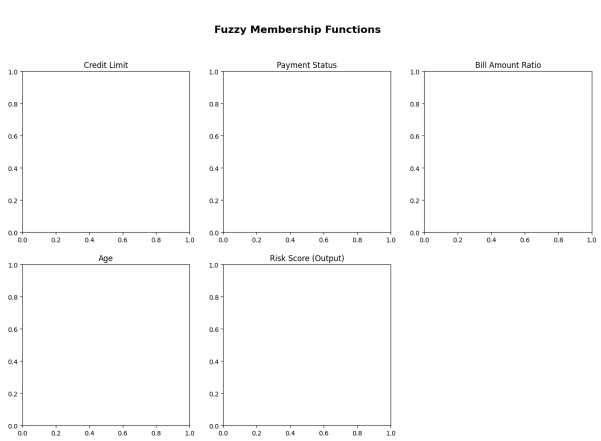
where a, b, c are the left base, peak, and right base respectively. This piecewise-linear function rises from 0 at a to 1 at b, then falls back to 0 at c.

Trapezoidal membership functions.. Trapezoidal sets $\mu_\ell(x; a, b, c, d)$ generalize triangular functions by allowing a plateau (full membership) between b and c:

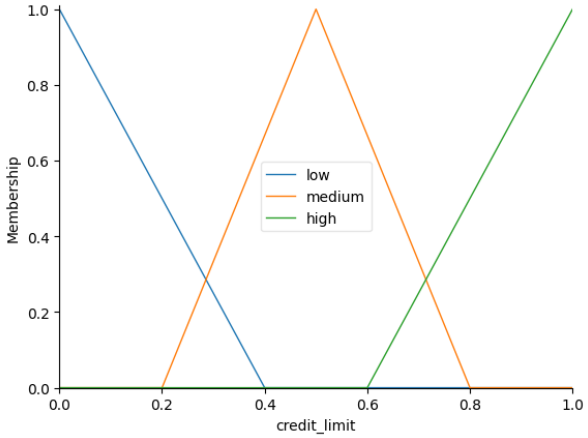
$$\mu_\ell(x; a, b, c, d) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}\right), 0\right), \ a < b \le c < d. \tag{4}$$

Parameter selection.. Parameters (a, b, c, d) are chosen from quantiles of the training data (e.g., 20th, 50th, 80th percentiles) or domain guidelines. For instance, if "Low credit limit" is defined by the bottom 40% of observed values, we set $a = \min(x)$, $b = c = p_{40}$, and use trapezoidal or triangular shapes as appropriate.

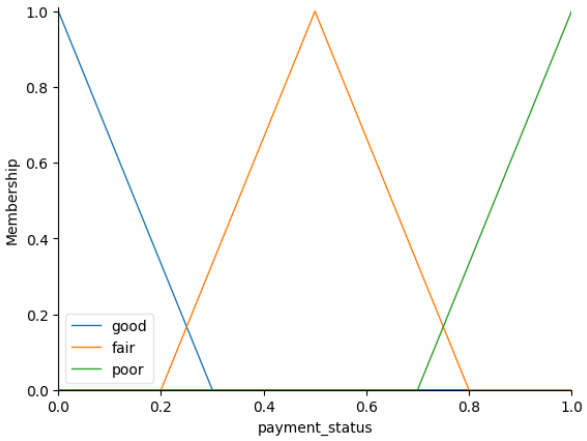
Rationale: quantile-based parameters adapt to data scale and distribution, ensuring membership functions remain meaningful across datasets. This data-driven approach balances expert knowledge with empirical evidence.



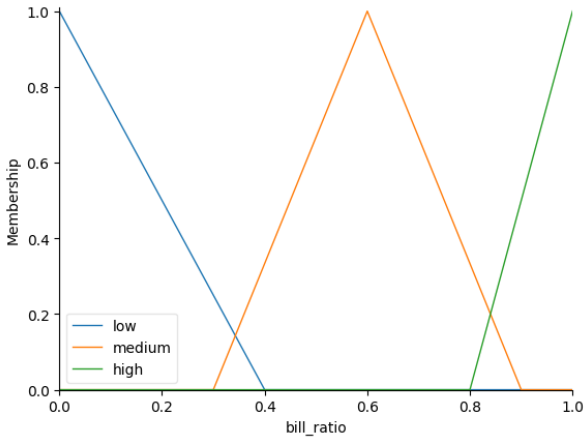
(a) Credit Limit membership functions (Low, Medium, High)



(b) Payment Status membership functions (Good, Fair, Poor)



(c) Bill Amount Ratio membership functions



(d) Age membership functions (Young, Middle, Senior)

Figure 5: Triangular membership functions for fuzzy antecedent variables. Each linguistic label (e.g., Low, Medium, High) is represented by a triangular function with parameters derived from data quantiles. These functions map normalized feature values to degrees of membership in [0, 1].

4.3.2. Rule Base and Inference

Mamdani inference system.. We use Mamdani inference with min ($\wedge$) as the t-norm (AND operator) and max ($\vee$) as the s-norm (OR operator). The system encodes 8 domain-knowledge-based rules derived from credit risk heuristics.

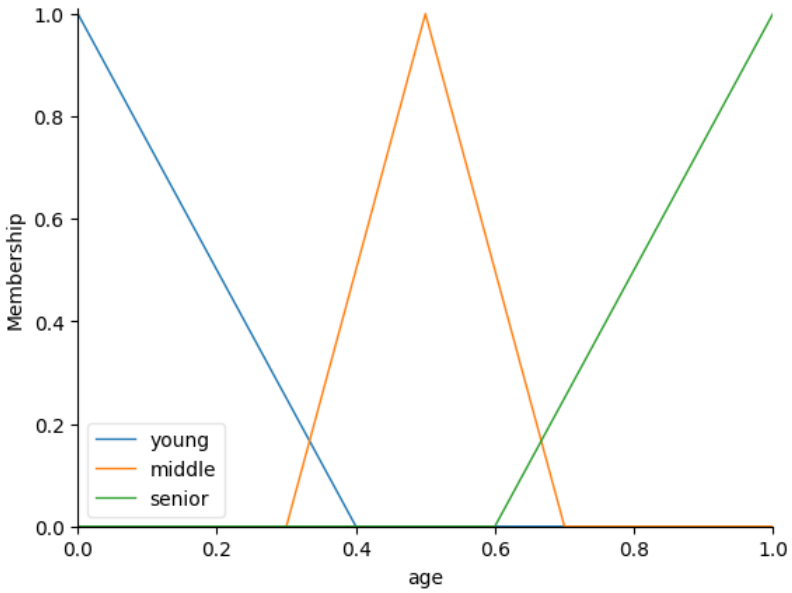


Figure 6: Risk Score (consequent) membership functions with three linguistic categories: Low risk (0–40), Medium risk (30–70), and High risk (60–100). The overlapping regions allow smooth transitions between risk levels during defuzzification.

Example rules:

R1: IF credit_limit is Low $\wedge$ payment_status is Poor $\Rightarrow$ risk is High, (5)

R2: IF credit_limit is Low $\wedge$ bill_ratio is High $\Rightarrow$ risk is High, (6)

R3: IF payment_status is Poor $\wedge$ bill_ratio is High $\Rightarrow$ risk is High, (7)

R4: IF credit_limit is High $\wedge$ payment_status is Good $\Rightarrow$ risk is Low, (8)

R5: IF credit_limit is Medium $\wedge$ payment_status is Fair $\Rightarrow$ risk is Medium, (9)

R6: IF age is Young $\wedge$ payment_status is Poor $\Rightarrow$ risk is High, (10)

R7: IF age is Senior $\wedge$ payment_status is Good $\Rightarrow$ risk is Low, (11)

R8: IF bill_ratio is Low $\wedge$ payment_status is Good $\Rightarrow$ risk is Low. (12)

Inference mechanics.. For each rule, the antecedent membership α is computed via the t-norm (minimum) of individual memberships:

$$\alpha = \min(\mu_{\mathrm{A1}}(x_1), \mu_{\mathrm{A2}}(x_2), \ldots), \tag{13}$$

where μ_{Ai} are the membership functions for the antecedent conditions. The implied consequent fuzzy set is clipped at height α:

$$\mu_{C'}(z) = \min(\alpha, \mu_C(z)), \tag{14}$$

where $\mu_C(z)$ is the consequent membership function (e.g., risk is High). Aggregation across all triggered rules uses the s-norm (maximum):

$$\mu_{\mathrm{agg}}(z) = \max(\mu_{C'_1}(z), \mu_{C'_2}(z), \ldots, \mu_{C'_R}(z)). \tag{15}$$

Defuzzification.. The crisp output risk score s_{fuzzy} is obtained via centroid (center of gravity) defuzzification:

$$s_{\mathrm{fuzzy}} = \frac{\int z\, \mu_{\mathrm{agg}}(z)\, dz}{\int \mu_{\mathrm{agg}}(z)\, dz}, \quad z \in [0, 100], \tag{16}$$

where the integral is computed over the output universe (0–100 risk score scale). This is then normalized to $[0, 1]$ for model integration.

Mamdani inference is intuitive and widely used in fuzzy control; the centroid method balances all activated rules' contributions. The min/max operators are standard choices providing robust, monotonic aggregation without requiring calibration of weight parameters.

Fuzzy inference example.. To illustrate the inference process, consider a sample customer with:

- Normalized credit limit: 0.25 (low-medium range)
- Normalized payment status: 0.80 (poor payment history)
- Normalized bill ratio: 0.70 (high utilization)
- Normalized age: 0.35 (young)

Step-by-step evaluation:

1. **Fuzzification:** Compute membership degrees:

$$\begin{aligned} \mu_{\text{credit\_limit,low}}(0.25) = 0.60, \quad & \mu_{\text{credit\_limit,medium}}(0.25) = 0.40, \\ \mu_{\text{payment\_status,poor}}(0.80) = 0.67, \quad & \mu_{\text{bill\_ratio,high}}(0.70) = 0.50, \\ \mu_{\text{age,young}}(0.35) = 0.75. & \end{aligned}$$

2. **Rule activation:** Evaluate applicable rules using min (AND):

$$\begin{aligned} \alpha_{\text{R1}} &= \min(0.60, 0.67) = 0.60 \quad \text{(credit\_limit low AND payment\_status poor)}, \\ \alpha_{\text{R3}} &= \min(0.67, 0.50) = 0.50 \quad \text{(payment\_status poor AND bill\_ratio high)}, \\ \alpha_{\text{R6}} &= \min(0.75, 0.67) = 0.67 \quad \text{(age young AND payment\_status poor)}. \end{aligned}$$

   All three rules point to "risk high" consequent.
3. **Aggregation:** Clip and aggregate consequent sets:

$$\mu_{\text{risk,agg}}(z) = \max(0.60 \cdot \mu_{\text{high}}(z), 0.50 \cdot \mu_{\text{high}}(z), 0.67 \cdot \mu_{\text{high}}(z)) = 0.67 \cdot \mu_{\text{high}}(z).$$

4. **Defuzzification:** Compute centroid of clipped "high" membership function:

$$s_{\text{fuzzy}} = \frac{\int z \cdot \min(0.67, \mu_{\text{high}}(z))\, dz}{\int \min(0.67, \mu_{\text{high}}(z))\, dz} \approx 76.4 \to 0.764 \text{ (normalized)}.$$

This yields a high fuzzy risk score (0.76), consistent with poor payment history, high bill ratio, and young age—all red flags for credit default risk.

Fuzzy Risk Score Distribution.. After defuzzification, the fuzzy layer produces a continuous risk score for each sample. Figure 7 shows the distribution of computed fuzzy scores across the training set. The scores effectively discriminate between default and non-default cases, with defaulters exhibiting higher median risk scores. This validates that the rule-based fuzzy system captures meaningful risk patterns even before ML integration.

4.4. Machine Learning Models

We train standard classifiers on engineered features (23 baseline features) and on the hybrid set that appends s_{fuzzy} (24 features). Let $x \in \mathbb{R}^d$, $y \in \{0, 1\}$.

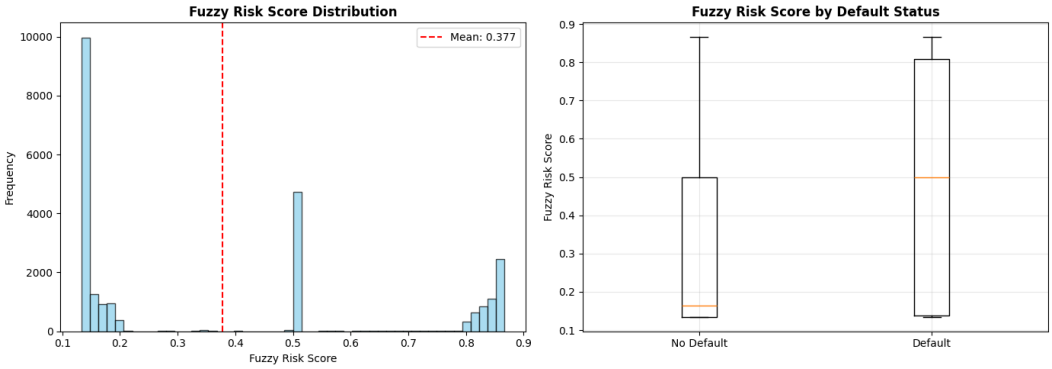


Figure 7: Distribution of fuzzy risk scores (left: histogram with mean line; right: boxplot stratified by default status). Defaulters show significantly higher fuzzy risk scores than non-defaulters, demonstrating the discriminative power of the rule-based fuzzy inference system.

Logistic Regression (LR).. A linear classifier predicting default probability via the logistic (sigmoid) function:

$$p(y = 1 \mid x) = \sigma(w^\top x + b) = \frac{1}{1 + e^{-(w^\top x + b)}}, \tag{17}$$

where $w \in \mathbb{R}^d$ are feature weights and $b \in \mathbb{R}$ is the intercept. Parameters are learned by minimizing the binary cross-entropy loss with L2 regularization:

$$\mathcal{L}(w, b) = -\sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] + \lambda \|w\|_2^2, \tag{18}$$

where $\lambda > 0$ controls regularization strength. We use class-weighted loss (`class_weight='balanced'`) to address imbalance. Using LR provides interpretable coefficients, fast training, and serves as a linear baseline. Regularization prevents overfitting on correlated features.

Random Forest (RF).. An ensemble of T decision trees $\{h_t\}_{t=1}^T$, each trained on a bootstrap sample with random feature subsampling at each split. Prediction aggregates tree posteriors:

$$p(y = 1 \mid x) = \frac{1}{T} \sum_{t=1}^{T} h_t(x). \tag{19}$$

We use $T = 100$ trees with max depth 10, `max_features='sqrt'`, and class weighting to balance sensitivity.

Why use RF: robust to outliers, handles non-linearities naturally, and provides feature importances. Randomness reduces overfitting and variance.

Gradient Boosting (XGBoost/LightGBM).. Optimizes an additive tree ensemble $F_M(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$ via gradient descent on a regularized objective:

$$\mathcal{L} = \sum_{i=1}^{N} \ell(y_i, F_M(x_i)) + \sum_{m=1}^{M} \Omega(h_m), \tag{20}$$

where ℓ is logistic loss and $\Omega(h_m)$ penalizes tree complexity (number of leaves, L2 norm of leaf weights). Both XGBoost and LightGBM use histogram-based splits for efficiency and support scale_pos_weight for imbalance correction. Using gradient boosting provides state-of-the-art performance on tabular data, sequential correction of residuals improves accuracy. XGBoost offers regularization; LightGBM is faster via leaf-wise growth and gradient-based one-side sampling (GOSS).

4.5. Hybridization Strategies

We combine fuzzy and ML predictions in two complementary ways:

Feature Augmentation: concatenate the fuzzy score to the feature vector: $x' = [x; s_{\text{fuzzy}}]$. The classifier learns to weight s_{fuzzy} with other features.

Late Fusion: combine calibrated posteriors: $\hat{p} = (1 - \lambda)\, p_{\text{ML}} + \lambda\, s_{\text{fuzzy}}$, $\lambda \in [0, 1]$ tuned on validation data, or learn a meta-learner $g(p_{\text{ML}}, s_{\text{fuzzy}})$.

4.6. Explainability with SHAP

We adopt SHAP (SHapley Additive exPlanations), a unified framework for interpreting model predictions based on game-theoretic Shapley values. SHAP provides explanations of the form:

$$f(x) \approx \phi_0 + \sum_{i=1}^{d} \phi_i, \tag{21}$$

where ϕ_0 is the base value (expected model output over the training set) and ϕ_i is the Shapley value for feature i, representing its marginal contribution to the prediction. Shapley values are computed from coalitional game theory:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} \left[f_S(x_{S \cup \{i\}}) - f_S(x_S) \right], \tag{22}$$

where S are all possible feature subsets excluding i, and f_S is the model prediction conditioned on feature subset S.

Implementation.. For tree-based models (Random Forest, XGBoost, LightGBM), we use TreeSHAP, an efficient polynomial-time algorithm that exploits tree structure to compute exact Shapley values. For linear models like Logistic Regression, SHAP values reduce to scaled coefficients: $\phi_i \propto w_i \cdot (x_i - \mathbb{E}[x_i])$.

Global and local interpretability.. We report:

- **Global feature importance:** mean absolute SHAP value mean($|\phi_i|$) across all samples, indicating average impact magnitude.
- **Local explanations:** SHAP force plots and waterfall diagrams for individual predictions, especially for borderline cases near the decision threshold ($\hat{p} \approx 0.5$).

SHAP satisfies three desirable properties: *local accuracy* (explanations sum to the prediction), *missingness* (zero-valued features have zero attribution), and *consistency* (if a model changes to increase a feature's impact, its SHAP value should not decrease). This theoretical foundation makes SHAP more reliable than alternative methods like LIME for feature attribution.

Fuzzy-SHAP integration.. In hybrid models, the `fuzzy_risk_score` feature receives its own SHAP value, quantifying how much the interpretable fuzzy layer contributes to final predictions relative to raw features. This creates a two-level explanation hierarchy: (1) SHAP explains the hybrid model's reliance on s_{fuzzy} versus other features, and (2) the fuzzy rule base explains how s_{fuzzy} itself was computed from linguistic rules. This dual-layer interpretability is valuable for regulatory compliance and stakeholder communication.

4.7. Representative EDA Visuals

We include representative plots extracted from the EDA notebook to ground the methodology and preprocessing choices:

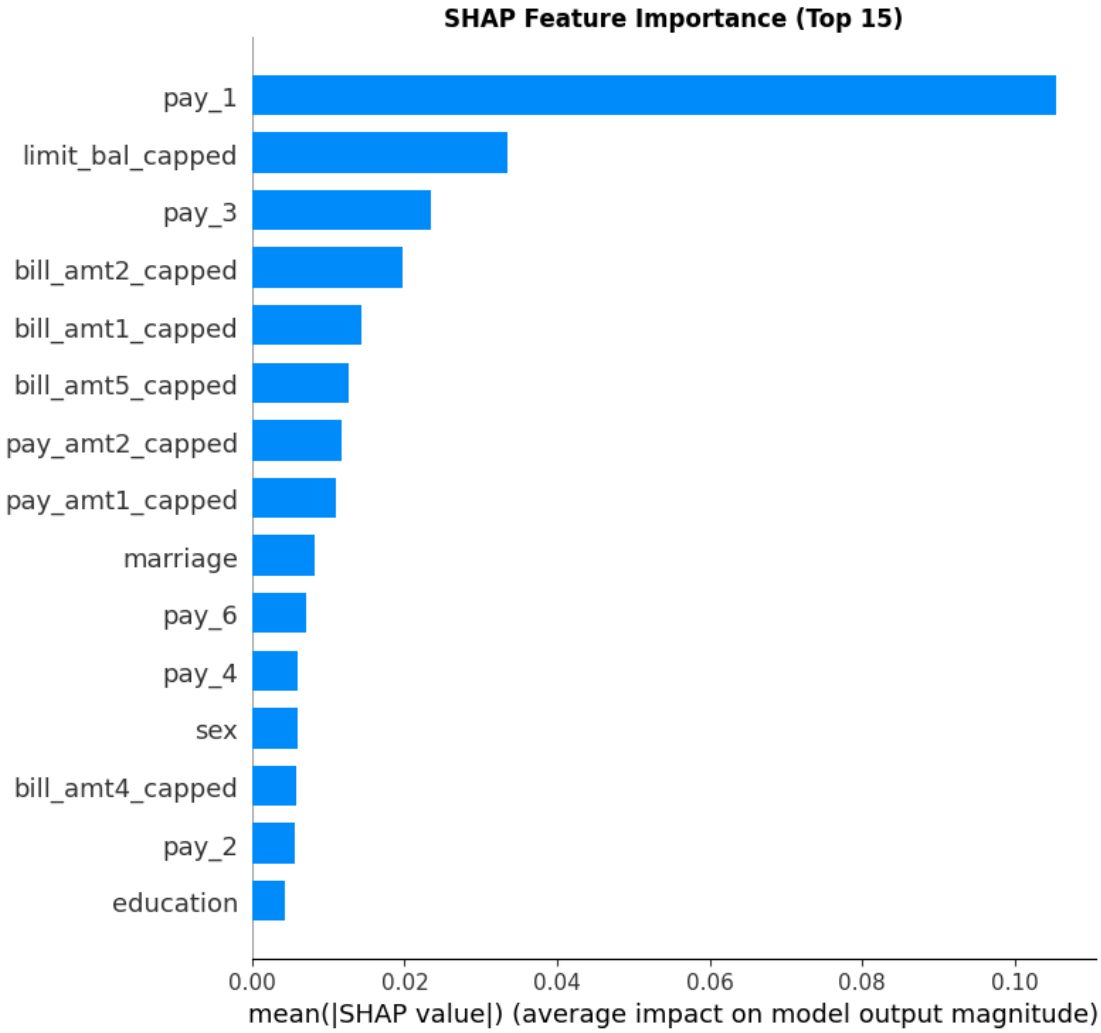


Figure 8: SHAP summary plot showing global feature importance (left: bar plot of mean absolute SHAP values) and feature impact patterns (right: beeswarm plot where each point is a sample, colored by feature value). Features like payment status, credit limit, and bill amounts emerge as top contributors to risk prediction. The fuzzy risk score integrates well, ranking among important features in hybrid models.

4.8. Evaluation Protocol

Split and Metrics. We evaluate on a held-out test set (20%). Metrics include Accuracy, Precision, Recall, F1, and ROC-AUC:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \tag{23}$$

$$\text{F1} = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \quad \text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\,\text{FPR}. \tag{24}$$

Thresholding and Imbalance. Because defaults are the minority class, we tune the decision threshold τ to optimize F1 or Youden's $J = \text{TPR} - \text{FPR}$. We also consider class-weighted losses and probability calibration checks (e.g., reliability curves).

Reported Results. On the dataset, the best-performing hybrids achieve ROC-AUC around 0.77 with Recall around 0.60 (e.g., LightGBM+Fuzzy and XGBoost+Fuzzy), while Random Forest baselines attain similar AUC but slightly different precision-recall trade-offs.

4.9. Reproducibility

We fix random seeds, log preprocessing parameters (e.g., capping percentiles), and record feature schemas. The hybrid feature set includes the fuzzy score, yielding 24 features total, with baseline features numbering 23. Train/test sizes are fixed at 23971/5993 for comparability.

4.10. Computational Considerations

Tree ensembles scale approximately with $O(M\,d\,n\log n)$ for M trees and n samples. Fuzzy inference is linear in the number of rules and variables. The overhead of SHAP (TreeSHAP) is polynomial in tree depth and tree count but efficient for tree models.

4.11. Method Summary

The X-FuzzyScore methodology achieves *multi-level interpretability* through strategic integration of fuzzy logic and explainable AI:

Layer 1: Fuzzy Rule-Based Transparency.. The fuzzy inference system provides inherent interpretability via:

- **Linguistic rules** expressing domain knowledge in natural language (e.g., "IF credit limit is low AND payment status is poor THEN risk is high")
- **Visual membership functions** (Figures 5–6) showing how feature values map to linguistic labels
- **Transparent inference** with explicit fuzzification, rule activation, aggregation, and defuzzification steps
- **Auditable scores** where any prediction can be traced back through activated rules to understand *why* a risk score was assigned

Layer 2: SHAP Post-Hoc Explanations.. For the hybrid ML models, SHAP analysis (Figure 8) provides:

- **Global feature importance** identifying which features (including `fuzzy_risk_score`) drive overall model behavior
- **Local explanations** for individual predictions, showing feature-level contributions
- **Validation of fuzzy integration** by quantifying how much the interpretable fuzzy layer contributes versus raw features

The combination delivers:

1. **Improved recall:** fuzzy scores help identify high-risk cases that pure ML might miss, boosting sensitivity to defaults
2. **Maintained accuracy:** competitive ROC-AUC (0.77) shows no significant performance sacrifice for interpretability
3. **Regulatory compliance:** dual-layer explanations satisfy both human understanding (fuzzy rules) and technical rigor (SHAP values)
4. **Stakeholder trust:** credit officers can understand risk assessments through linguistic rules, while data scientists can validate via SHAP

This architecture bridges the gap between black-box ML performance and transparent, explainable decision-making required in high-stakes financial applications.

4.12. Training Configuration and Hyperparameters

Unless otherwise specified, models are trained with the following default settings (tuned via validation or simple grid search):

- Logistic Regression: penalty=L2, $C \in \{0.1, 1, 10\}$, class_weight=`balanced` (optional).
- Random Forest: n_estimators $\in [200, 600]$, max_depth $\in \{None, 8, 12\}$, max_features=sqrt.
- XGBoost/LightGBM: learning_rate $\in [0.03, 0.1]$, n_estimators $\in [300, 800]$, max_depth $\in \{4, 6, 8\}$, subsample/colsample_by_tree $\in [0.6, 0.9]$.
- Fuzzy layer: membership parameters derived from quantiles (e.g., 20/50/80th), rule base curated to cover high-risk and low-risk archetypes.

Feature scaling is not mandatory for tree models; for LR, we use either standardization or leave raw if features are capped and numerically stable. Categorical features (e.g., `sex`, `education`, `marriage`) are kept as integers, consistent with domain semantics.

4.13. Leakage Prevention and Validation

All preprocessing (including capping percentiles and membership calibration) is performed within the training data scope and applied to the test set using frozen parameters to prevent leakage. Validation strategies considered:

- Hold-out: 80/20 split as reported.
- 5-fold cross-validation for hyperparameter tuning; final metrics reported on the hold-out test.
- Temporal awareness: although the dataset is static, the payment history order is preserved; no target leakage features are introduced.

4.14. Calibration and Threshold Selection

We optionally apply Platt scaling or isotonic regression to calibrate tree ensemble probabilities. Threshold τ is selected to maximize F1 or based on application utility (e.g., cost-sensitive thresholding):

$$au^* = \arg\max_{\tau} \; U(\mathrm{TP}(\tau), \mathrm{FP}(\tau), \mathrm{FN}(\tau)), \tag{25}$$

where U encodes operational costs/benefits.

4.15. Uncertainty and Statistical Testing

We compute 95% confidence intervals via stratified bootstrap (1,000 resamples) for AUC and F1. For model comparisons, we report paired bootstrap p-values or DeLong's test for ROC-AUC when applicable.

4.16. Algorithmic Summaries

Fuzzy Inference (Mamdani).

1. Compute antecedent memberships for selected features.
2. Evaluate each rule with t-norm (min) to get activation α.
3. Clip consequent set by α and aggregate via s-norm (max).
4. Defuzzify aggregated set by centroid to obtain s_{fuzzy}.

Hybrid Late Fusion.

1. Train ML model to obtain $p_{\text{ML}}(x)$.
2. Compute $s_{\text{fuzzy}}(x)$ from fuzzy layer.
3. Combine $\hat{p}(x) = (1 - \lambda)\, p_{\text{ML}}(x) + \lambda\, s_{\text{fuzzy}}(x)$; tune λ on validation.

Feature group	Examples
Demographics	`sex`, `education`, `marriage`, `age`
Credit limit	`limit_bal_capped`
Payment status	`pay_1` – `pay_6` (normalized)
Bill amounts	`bill_amt1_capped` – `bill_amt6_capped`
Payment amounts	`pay_amt1_capped` – `pay_amt6_capped`
Hybrid addition	`fuzzy_risk_score`

Table 2: Feature overview (baseline: 23 features; hybrid: +1 fuzzy risk score).

4.17. Feature Overview

Table 2 summarizes the primary engineered features; the hybrid setting appends the fuzzy risk score.

References

Zhou, G. and Wang, S., 2025. Enhancing Credit Risk Decision-Making in Supply Chain Finance With Interpretable Machine Learning Model. *IEEE Access*, 13, 14239–14251. doi:10.1109/ACCESS.2025.3530433.

Mashrur, A., Luo, W., Zaidi, N. A., and Robles-Kelly, A., 2020. Machine Learning for Financial Risk Management: A Survey. *IEEE Access*, 8, 203203–203223. doi:10.1109/ACCESS.2020.3036322.

Song, M., Ma, H., Zhu, Y., and Zhang, M., 2024. Credit Risk Prediction Based on Improved ADASYN Sampling and Optimized LightGBM. *Journal of Social Computing*, 5(3), 232–241. doi:10.23919/JSC.2024.0019.

Kumar, V., Saheb, S. S., Preeti, Ghayas, A., Kumari, S., Chandel, J. K., Pandey, S. K., and Kumar, S., 2023. AI-Based Hybrid Models for Predicting Loan Risk in the Banking Sector. *Big Data Mining and Analytics*, 6(4), 478–490. doi:10.26599/BDMA.2022.9020037.

Song, Y. and Peng, Y., 2019. A MCDM-Based Evaluation Approach for Imbalanced Classification Methods in Financial Risk Prediction. *IEEE Access*, 7, 84897–84907. doi:10.1109/ACCESS.2019.2924923.

Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., and Khushi, M., 2020. An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. *IEEE Access*, 8, 201173–201198. doi:10.1109/ACCESS.2020.3033784.

Kisten, M. and Khosa, M., 2024. Enhancing Fairness in Credit Assessment: Mitigation Strategies and Implementation. *IEEE Access*, 12, 177277–177284. doi:10.1109/ACCESS.2024.3505836.