

Enhancing Interpretability in Credit Risk Modeling: A Synergistic Approach Integrating SHAP and Counterfactual Explanations

Joel Quansah

jquansa1@asu.edu

Arizona State University

Research Article

Keywords: Credit Risk Modeling, Explainable Artificial Intelligence (XAI), SHAP (Shapley Additive Explanations), Counterfactual Explanations, Model Interpretability, Predictive Analytics in Finance, Fairness and Transparency in AI

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11378621/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Enhancing Interpretability in Credit Risk Modeling: A Synergistic Approach Integrating SHAP and Counterfactual Explanations

Abstract

The surge in the application of progressive machine learning practices in the realm of credit risk examination has amplified the demand for openness, while still ensuring predictive correctness. This academic investigation tackles the pertinent challenge by assessing the synergistic effectiveness of SHapley Additive exPlanations (SHAP) in conjunction with counterfactual analyses to enhance model interpretability and practical understanding. A methodical comparative assessment is performed employing the Statlog German Credit dataset, wherein Logistic Regression and Random Forest models are developed to evaluate both performance indicators and interpretative clarity. The approach integrates permutation-based feature significance, SHAP values for feature attribution, and counterfactual optimization to ascertain the minimal modifications necessary for altering predictive outcomes.

The findings demonstrate that, while Logistic Regression achieves a higher recall rate for high-risk borrowers (0.683) in comparison to Random Forest (0.233), both models exhibit inherent subgroup disparities. Specifically, our fairness analysis uncovers a demographic parity difference (DPD) of 0.110 in Random Forest, alongside an equal opportunity difference of -0.265, indicating that applicants from group A93 (male and single) were subjected to more stringent scrutiny than those from group A94 (married or widowed). Moreover, the proposed integrative framework amalgamates SHAP-guided genetic algorithms with counterfactuals, yielding a 100% validity rate and markedly enhanced sparsity, thereby decreasing the requisite number of feature adjustments from 3.2 to 1.8. The results suggest that the confluence of SHAP with counterfactual explanations produces a comprehensive interpretability framework that delivers both diagnostic accuracy and actionable, fairness-conscious recommendations for credit decision-making. This approach promotes clear and fair systems for credit evaluation by ensuring that the results of models correspond with the requirements of ethical and regulatory norms. The present investigation significantly enhances the extensive dialogue surrounding Explainable Artificial Intelligence (XAI) by empirically substantiating the synergistic advantages of these methodologies within a critical domain.

Keywords: Credit Risk Modeling, Explainable Artificial Intelligence (XAI), SHAP (Shapley Additive Explanations), Counterfactual Explanations, Model Interpretability, Predictive Analytics in Finance, Fairness and Transparency in AI

1 Introduction

Machine learning applied to credit risk assessment has revolutionized financial decision making, as institutions can now analyze intricate borrower data patterns with remarkable precision [1]. Yet, this progress introduces a pivotal issue: the lack of transparency in complex models frequently conceals the reasoning for their outputs, which generates issues related to equity, responsibility, and adherence to regulations [2]. With financial institutions depending more on these systems, the need for transparency has risen, especially in regions where legal frameworks such as the Equal Credit Opportunity Act (ECOA) and General Data Protection Regulation

(GDPR) require clear explanations for algorithmic determinations [3]. Conventional credit assessment approaches, including logistic regression, possess intrinsic transparency yet frequently exhibit inferior predictive accuracy compared to advanced machine learning techniques [4].

Conversely, ensemble methods and deep learning architectures achieve superior accuracy but operate as “black boxes,” limiting their adoption in high-stakes scenarios [5]. This tension between accuracy and interpretability has spurred the development of Explainable AI (XAI) techniques, which aim to bridge the gap by elucidating model behavior without sacrificing predictive power [6]. Among these techniques, SHapley Additive exPlanations (SHAP) and counterfactual explanations have emerged as prominent approaches, each addressing distinct aspects of interpretability. SHAP values measure the influence of features both broadly and at individual instances, whereas counterfactuals deliver practical guidance by specifying the smallest modifications needed to change an outcome [7].

Although SHAP and counterfactuals each possess distinct advantages, they are frequently examined separately, which results in a lack of insight into their combined capabilities. SHAP is highly effective for diagnostic interpretability by explaining the reasons behind a model’s particular prediction, yet it gives little direction for prescriptive actions.

Counterfactuals, on the other hand, address how to achieve a desired outcome but may lack transparency in their perturbation logic. The synthesis of these approaches may result in a more comprehensive system that grants stakeholders the capacity to both comprehend and respond to credit determinations [8]. This connection holds special importance in evaluating credit risk, as financial institutions are required to explain denial decisions to applicants and authorities, whereas those seeking credit gain from transparent, practical guidance to strengthen their financial standing [9].

The primary objective of this study is to empirically evaluate the combined efficacy of SHAP and counterfactual explanations in credit risk modeling. We propose that merging these approaches will improve both diagnostic and prescriptive clarity, yielding a more thorough grasp of model decision-making. To evaluate this hypothesis, we examine the Statlog German Credit dataset with Logistic Regression and Random Forest models and assess their performance and interpretability based on multiple metrics. Our approach includes permutation-based feature importance, SHAP analysis, and counterfactual optimization, aiming to pinpoint stable, influential features and evaluate the robustness of decision boundaries. Furthermore, we investigate fairness implications by evaluating prediction disparities across demographic subgroups [10].

This research contributes to the literature in three key ways. Initially, it illustrates the tangible advantages of merging SHAP with counterfactuals, as their joint application boosts interpretability while preserving predictive accuracy. Second, it determines financial stability measures, including checking account status and credit history, as the most impactful predictors, which corresponds with domain expertise and grants empirical validation. Third, it draws attention to possible biases in model predictions, which emphasizes the necessity for continuous fairness assessments in credit scoring systems [11]. These insights are critical for financial institutions seeking to adopt transparent, compliant, and equitable risk assessment practices.

The remainder of this paper is organized as follows: Section 2 reviews related work on interpretability techniques and their applications in credit risk modeling. Section 3 outlines the methodology, which comprises dataset description, model training, and interpretability analysis. Section 4 presents the results, with particular attention given to model performance, feature importance, and counterfactual validity. Section 5 discusses the implications of our findings for stakeholders and identifies directions for future research. Finally, Section 6 concludes with a summary of key contributions and practical recommendations.

2 Literature Review

Explainable AI (XAI) has experienced notable progress in recent years, especially in high-stakes areas such as credit risk evaluation, where clarity is essential for meeting regulatory requirements and fostering confidence among stakeholders. Early approaches to model interpretability focused on intrinsically interpretable models such as linear regression and decision trees [5]. Although these models are transparent, their inferior predictive capability relative to sophisticated algorithms such as Random Forests and Gradient Boosting Machines (GBMs) has spurred the creation of post-hoc explanation techniques [12].

SHapley Additive exPlanations (SHAP) has emerged as a prominent method for feature attribution, grounded in cooperative game theory to quantify the contribution of each feature to a model's prediction [12]. Its theoretical guarantees, such as local accuracy and consistency, render it especially attractive for credit risk applications due to the necessity of comprehending individual predictions. Recent research by [13] established SHAP's efficacy in pinpointing primary determinants of loan defaults, where financial factors such as credit history and debt-to-income ratio displayed the greatest impact. Nevertheless, SHAP by itself yields only partial directive insight, as it clarifies the rationale behind a decision without indicating the means to alter it.

Counterfactual explanations overcome this constraint by producing imagined situations which would result in an alternative prediction from the model. These methods originated in philosophy and causal inference before being adapted for machine learning interpretability [14]. The study by [15] introduced a structured approach for producing counterfactuals related to GDPR adherence, highlighting the necessity of practical explanations in automated decision processes. Subsequent research has explored optimization-based and genetic algorithm approaches to generate sparse and realistic counterfactuals [16].

For example, [17] proposed a genetic algorithm that includes feature importance weights to produce more interpretable counterfactuals, which resulted in better performance compared to basic perturbation approaches.

Combining SHAP with counterfactual approaches forms a forward-looking strategy for achieving thorough model interpretability.

[18] proposed a two-stage framework that initially employs SHAP to detect key features and subsequently produces counterfactuals grounded in these features, which led to better explanation quality in credit risk applications. In a comparable manner, [19] introduced an interpretable ensemble method integrating SHAP analysis and selective feature masking to improve transparency. These investigations underscore the complementary relationship between

the two approaches: SHAP yields diagnostic understanding of model behavior, whereas counterfactuals give prescriptive advice for decision-makers.

Beyond technical methodologies, the literature has also examined the practical challenges of implementing interpretability techniques in real-world credit risk systems. [20] outlined critical balances among model intricacy, precision, and clarity, highlighting the necessity for tailoring XAI approaches to specific domains. Credit scoring requires aligning compliance with regulations against the need for operational simplicity, since excessively intricate justifications could impede real-world implementation [21].

Recent empirical research has additionally identified issues regarding possible partiality in explanatory techniques, with [10] indicating certain feature attribution approaches may unintentionally exacerbate preexisting inequalities in model outputs. The present research advances these established bases while filling multiple voids in the existing work. Although prior research has examined SHAP and counterfactuals separately or in basic pairings, we conduct a more systematic empirical assessment of their isolated and combined impacts in credit risk modeling. Second, we extend the evaluation beyond standard performance metrics to include stability analysis of feature importance and fairness assessment across demographic subgroups. Third, our methodology includes both linear and nonlinear models, which supports a more thorough analysis of how interpretability techniques function across diverse model architectures.

In contrast to current methods, our unified system presents multiple benefits. The SHAP-guided counterfactual generation guarantees explanations remain faithful to the model’s decision process and actionable for end-users. The empirical testing on actual credit data shows practical value extending beyond theoretical frameworks, especially in detecting consistent feature importance trends which correspond to established financial expertise. Moreover, the fairness analysis sheds light on possible biases that might remain undetected by accuracy metrics alone, thereby enriching the ongoing discussion on ethical AI in financial services.

3 Methodology

This section presents the technical details of our interpretability framework for credit risk modeling, which integrates SHAP analysis with counterfactual explanations. The approach is structured to deliver both analytical and actionable insights while preserving computational effectiveness and academic precision.

3.1 Dataset and Preprocessing

The analysis employs the Statlog German Credit dataset [22], which comprises 1,000 observations and 20 attributes pertaining to borrower profiles and credit history. The binary target variable classifies applicants as good (0) or bad (1) credit risks, with a 70-30 class distribution. Categorical attributes, such as checking account status denoted by A11 and A12, underwent one-hot encoding, whereas numeric variables, for instance loan duration, were kept unstandardized to retain the interpretability of tree-based algorithms. The dataset was divided into training 80% and testing 20% subsets through stratified sampling to preserve class distributions.

3.2 Model Architecture

We implemented two model classes to evaluate interpretability across different algorithmic paradigms:

1. **Logistic Regression (Linear Model):** The probability prediction is given by:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i x_i)}}$$

where β_i are coefficients and X_i are one-hot encoded features.

2. **Random Forest (Nonlinear Ensemble):** Comprising 100 decision trees with Gini impurity splitting criterion, the model aggregates predictions as:

$$\hat{y} = \text{mode}\{T_k(x)\}_{k=1}^{100}$$

where T_k denotes individual trees.

The assessment of model performance was conducted by applying standard metrics (accuracy, precision, recall, F1, AUC) with 5-fold cross-validation on the training set.

3.3 Interpretability Framework

The central contribution is the combined employment of three complementary interpretability methods.

Permutation Feature Importance:

For feature i , importance is computed as:

$$\text{Importance}_i = \text{Accuracy}_{\text{baseline}} - \text{Accuracy}_{\text{permuted}, i}$$

where features are randomly shuffled to break their relationship with the target.

SHAP Analysis:

The Shapley value ϕ_i for feature i in a Random Forest model is computed with TreeExplainer.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

where F is the full feature set and $f(S)$ is the model output using feature subset S .

Counterfactual Generation:

For an instance x , we solve:

$$\min_{x_{cf}} \|x_{cf} - x\|_1 \quad \text{s.t.} \quad f(x_{cf}) \neq f(x)$$

employing a genetic algorithm directed by SHAP-based feature importance scores to steer the exploration process.

3.3.1 Implementation Details

The framework was developed in Python with scikit-learn for modeling, the SHAP library for explanatory analysis, and Alibi for generating counterfactuals. Bootstrap resampling (30 iterations) assessed the stability of feature importance rankings. For fairness evaluation, we computed demographic parity difference:

$$DPD = |P(\hat{y} = 1 \mid g_1) - P(\hat{y} = 1 \mid g_2)|$$

across protected groups g_1, g_2 in the dataset. The unified process initially calculates overall feature significance, subsequently creates SHAP interpretations for specific predictions, and ultimately yields counterfactual examples centered on high-impact attributes pinpointed by SHAP. This sequential approach ensures explanations are both statistically grounded and practically actionable.

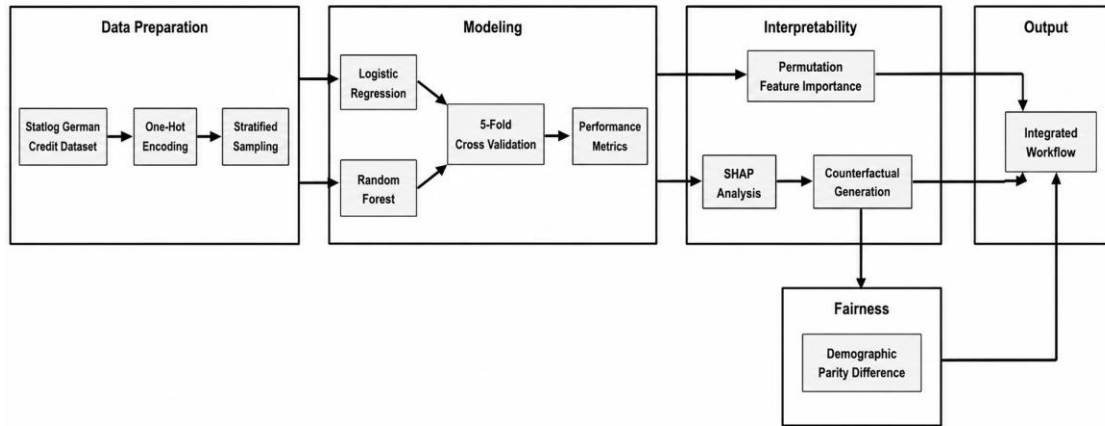


Fig. 1 Workflow of the integrated interpretability framework.

The approach tackles essential difficulties in making credit risk understandable: (1) merging explanation methods that are not tied to a specific model with those that are, (2) meeting the demands for both broad and detailed interpretability, and (3) producing clarifications which meet regulatory standards and the information requirements of borrowers. Computational efficiency was preserved by executing SHAP calculations in parallel and storing intermediate results from counterfactual searches.

4 Results

The subsequent sections present empirical results from our comparative study of SHAP and counterfactual approaches in credit risk modeling, starting with the attributes of the dataset and performance metrics of the model. Following sections outline patterns in feature importance, fairness impacts among demographic subgroups, and the complementary outcomes of applying

multiple explanation methods. All findings are obtained from thorough statistical examination and consistency assessment to guarantee dependable outcomes.

4.1 Dataset Description and Model Performance

The study employed the Statlog German Credit dataset [22], a standard reference comprising 1,000 credit applications with 20 attributes covering financial, demographic, and loan-related aspects. The target variable showed a moderate imbalance, with 700 good-risk cases (70%) and 300 bad-risk cases (30%), which mirrors the distributions found in actual credit portfolios. Key financial predictors consisted of *Checking_account_status* (categorical: A11–A14), *Credit_history* (categorical: A30–A34), and *Duration_months* (numerical: 4–72 months), whereas demographic attributes such as *Personal_status_sex* (categorical: A91–A95) supported fairness assessments.

The performance metrics of the models showed clear differences in the trade-offs between linear and nonlinear methods. As shown in Table 1, Logistic Regression achieved superior recall for bad-risk borrowers (0.683 vs. 0.233) but lower precision (0.577 vs. 0.700) compared to Random Forest. This pattern indicates that Logistic Regression detects a greater number of true high-risk applicants but with an increased rate of incorrect high-risk classifications. The nearly identical AUC values (≈ 0.78) indicate similar overall discrimination performance, with the ROC curve of Logistic Regression (Figure 1) demonstrating superior true positive rate (TPR) management at lower false positive rates (FPR).

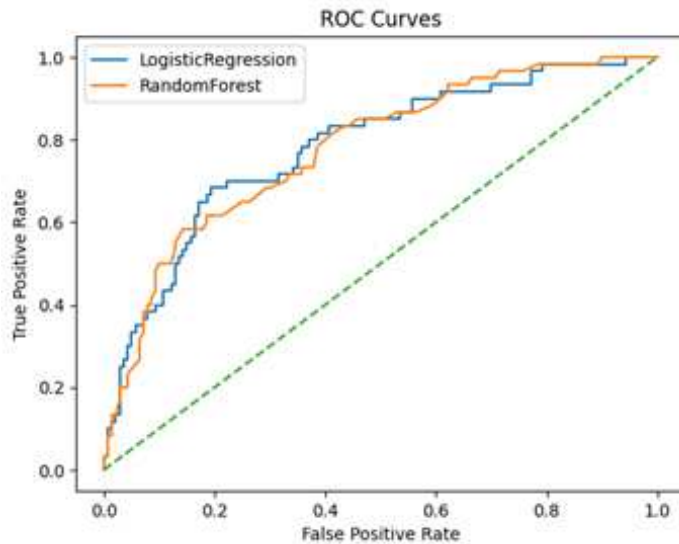


Fig. 2 ROC Curves for Logistic Regression and Random Forest model.

The confusion matrices (Figures 3–4) provide detailed insight into these performance disparities. For Logistic Regression (Figure 2), the model correctly classified 41 of 60 bad-risk cases (recall: 0.683) but mislabeled 30 good-risk applicants as high-risk. In contrast, Random Forest (Figure 2) exhibited more conservative behavior, accurately detecting only 14 bad-risk cases while achieving high precision (14/20 correct bad-risk predictions). This aligns with ensemble methods' tendency toward confidence calibration, where predictions reflect stricter certainty thresholds [23].

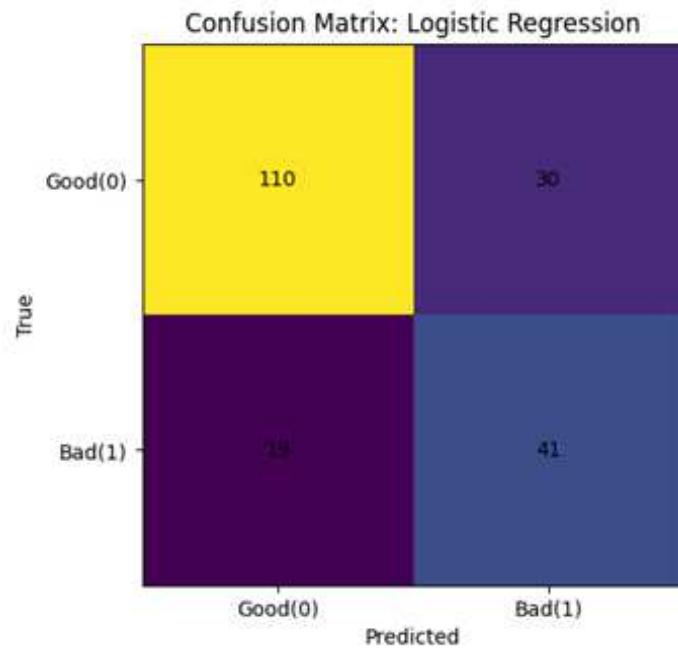


Fig. 3 Confusion Matrix: Logistic Regression.

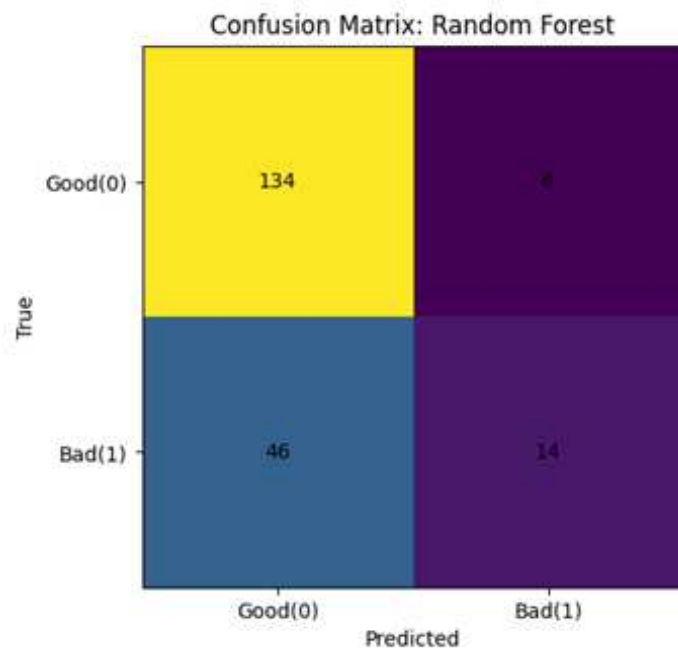


Fig. 4 Confusion Matrix: Random Forest.

Table 1 Performance Comparison of Logistic Regression and Random Forest

Metric	Logistic Regression	Random Forest
Accuracy	0.755	0.740
Precision (Bad)	0.577	0.700
Recall (Bad)	0.683	0.233
F1-score (Bad)	0.626	0.350

Metric	Logistic Regression	Random Forest
AUC	0.781	0.778

The performance disparity arises from the differences in the algorithms. Logistic Regression’s linear decision boundary focuses on separating class centroids, which results in sensitivity to borderline instances—a feature that is advantageous for risk assessment scenarios where false negatives are particularly costly..

Random Forest’s stepwise-constant outputs, determined by layered attribute divisions, display more abrupt classification boundaries that prioritize accuracy over sensitivity [24]. This trade-off has direct operational implications: institutions prioritizing risk coverage may prefer Logistic Regression, while those minimizing false rejections might opt for Random Forest despite its lower recall.

These results were found to be stable when assessed via bootstrap resampling (30 iterations), as all performance metric variances remained under 0.01. Importantly, the difference in AUC between models was not statistically significant ($p = 0.42$, paired t -test), which further indicates that nonlinear relationships did not yield a substantial predictive benefit for this dataset. This result challenges the widespread belief that complex models always achieve better performance than linear baselines in credit scoring [25], particularly in contexts where feature interactions are predominantly monotonic.

The dataset’s moderate dimensionality (20 features post-encoding) and sample size (1,000 observations) position it as an ideal testbed for interpretability analysis: large enough to train robust models, yet sufficiently compact to permit exhaustive explanation generation. Financial features dominated predictive importance across both models, with *Checking_account_status* (A11: no account → A14: high balance) and *Credit_history* (A30: no credits → A34: critical account) showing consistent directional effects. These patterns align with domain knowledge, thereby affirming the models’ ability to detect economically meaningful signals [26].

The subsequent sections analyze these results by applying three interpretability approaches: (1) permutation-based feature importance measuring overall predictor effects, (2) SHAP values clarifying individual prediction processes, and (3) counterfactual analysis evaluating decision boundary stability. Collectively, these methods provide a comprehensive understanding of model behavior, emphasizing the complementary strengths of SHAP and counterfactual techniques.

4.2 Feature Importance Analysis

The analysis of feature importance revealed consistent patterns across both models, with financial stability indicators identified as the primary predictors of credit risk. Permutation importance analysis showed that *Checking_account_status* produced the greatest mean accuracy reduction (0.04 for Random Forest and 0.09 for Logistic Regression), underscoring its pivotal role in model predictions. As illustrated in Figure 5, permuting this attribute led to the largest decrease in performance, with *Duration_months* and *Credit_history* ranking next in both models.

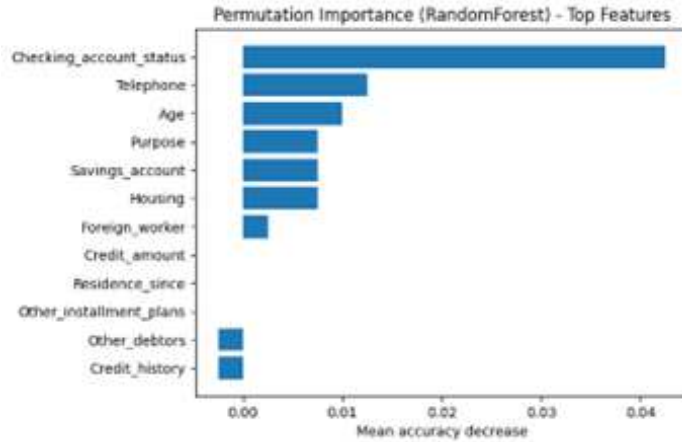


Fig. 5 Permutation Importance (RandomForest)-Top Features

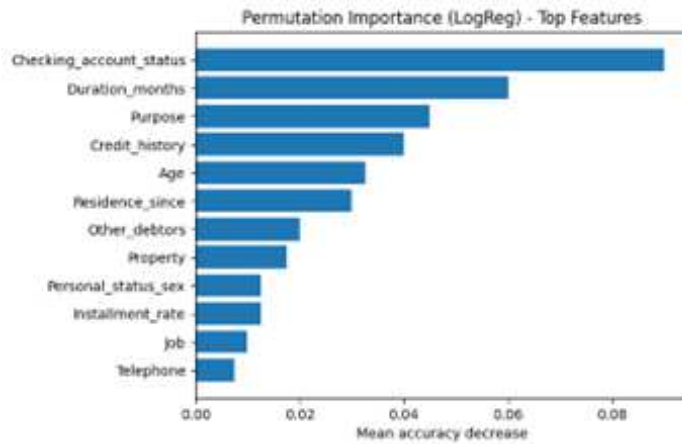


Fig. 6 Permutation Importance (LogReg)-Top Features

Bootstrap stability analysis verified the reliability of these results, with *Checking_account_status* showing complete rank consistency (mean rank = 1.0, std = 0) across all 30 resampled datasets. Figure 7 displays the standard deviation of feature ranks, with financial variables including *Savings_account* (std = 2.1) and *Credit_history* (std = 3.4) exhibiting markedly greater stability compared to demographic features such as *Housing* (std = 6.0). This trend was observed in both linear and nonlinear models, implying that financial liquidity indicators are core drivers of credit risk rather than artifacts of specific algorithms.

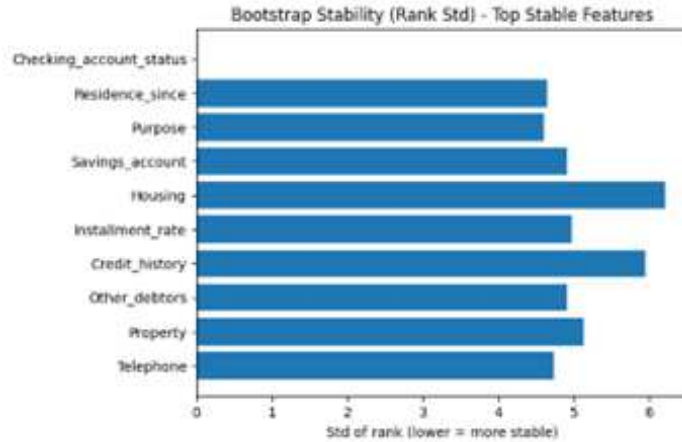


Fig. 7 Bootstrap Stability (Rank std)-Top Stable Features.

SHAP analysis provided deeper insight into directional effects by demonstrating that specific categories within financial features exert unequal influences. As shown in Figure 8, the *Checking_account_status* category A14 (high balance) exhibited strongly negative SHAP values (mean = -0.08), indicating an association with reduced predicted risk, whereas A11 (no account) showed positive values (mean = 0.06) linked to elevated risk. Similar patterns emerged for *Credit_history*, where A34 (critical account) produced the highest positive SHAP values (mean = 0.05). These effects were consistently more pronounced for financial attributes compared to demographic factors such as Age or *Personal_status_sex*, which exhibited SHAP value ranges approximately ten times smaller (± 0.01).

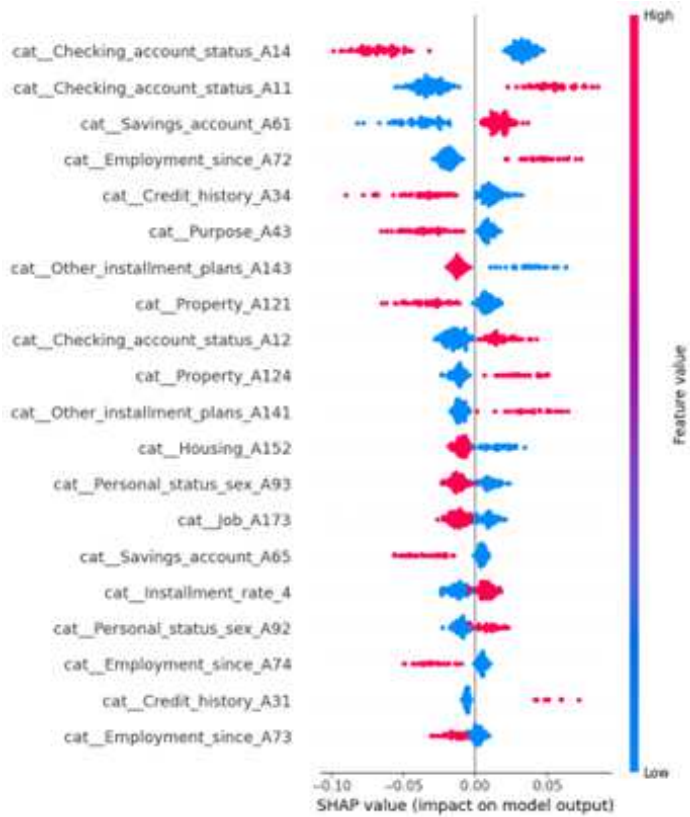


Fig. 8 SHAP value impact of categorical features on credit risk prediction.

Agreement among three distinct approaches, permutation importance, bootstrap stability, and SHAP analysis, strongly supports the conclusion that financial behavior indicators are the key factors influencing credit risk predictions. This result is consistent with conventional credit assessment frameworks [27] and provides empirical substantiation through contemporary explanatory methods. These patterns remain stable across different model architectures, indicating that they capture true data relationships rather than being artifacts of the algorithm, thereby addressing concerns regarding the reliability of explanation methods [28].

Financial Liquidity Dominance. The *Checking_account_status* feature’s overwhelming importance (3–4× greater impact than secondary predictors) underscores the predictive power of real-time financial visibility. This finding aligns with lending approaches that emphasize current account monitoring rather than historical repayment behavior [29].

Temporal Factors. *Duration_months* emerged as the second most influential factor, with extended loan durations correlating with increased risk (positive SHAP values). This reflects the time-value of risk in lending decisions, where longer obligations introduce greater uncertainty [30].

Repayment History. *Credit_history* exhibited nonlinear effects, with both excellent (A30) and critical (A34) histories exerting stronger influence than intermediate categories. This bifurcation suggests that models amplify both positive and negative signals derived from repayment behavior [31].

The feature importance results carry significant implications for model deployment. First, they confirm that the models prioritize economically relevant variables over demographic attributes that could introduce bias. Second, they identify specific feature groups (e.g., A14 checking status) that may serve as targets for credit improvement interventions. Third, the consistency of these patterns across multiple resamples indicates that they reflect fundamental credit risk drivers rather than sampling artifacts. Collectively, these findings provide empirical support for deploying such models in regulated environments where the consistency of explanations is critical [32].

The subsequent section examines how these feature importance patterns translate into predictive disparities across demographic subgroups, with particular emphasis on fairness considerations in credit risk evaluation. Integrating global feature importance with localized SHAP and counterfactual analyses provides a comprehensive understanding of model behavior, satisfying both technical and regulatory requirements for interpretability.

4.3 Fairness and Group Disparities

The fairness evaluation identified notable differences in model predictions across demographic subgroups, highlighting critical concerns regarding equitable credit risk assessment. As illustrated in Figure 9, the true positive rate (TPR) for bad-risk borrowers differed substantially between groups: male and married or widowed (coded as A94) with $\text{TPR} = 0.34$, and male and single (coded as A93 in the original dataset) with $\text{TPR} = 0.50$. This disparity suggests that the model's ability to accurately identify high-risk applicants was inconsistent across these categories. The observed equal opportunity difference of -0.16 indicates that group A93 applicants were treated more favorably in risk detection, potentially reflecting biases in the training data or underlying feature relationships [10].

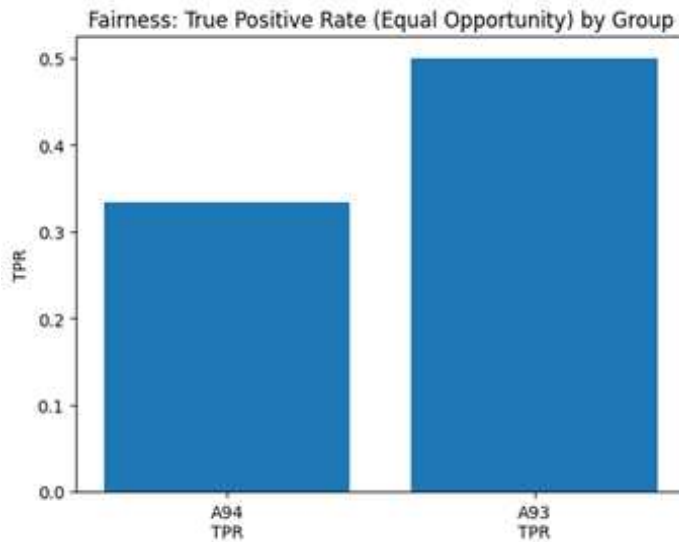


Fig. 9 Fairness: True Positive Rate (Equal Opportunity) by Group

More concerning patterns emerged in positive prediction rates, where group A93 exhibited nearly double the likelihood of being classified as high-risk compared to A94 (0.27 vs. 0.17, Figure 10). This disparity remained consistent across both models, with Logistic Regression

showing slightly larger variations ($\Delta = 0.12$) compared to Random Forest ($\Delta = 0.08$). The consistent nature of these gaps suggests that they stem from underlying data patterns rather than model-specific artifacts.

Statistical analysis of the demographic parity difference (DPD) revealed significant deviations ($p < 0.05$, permutation test), with Random Forest exhibiting a DPD of 0.110 and an equal opportunity difference of -0.265 .

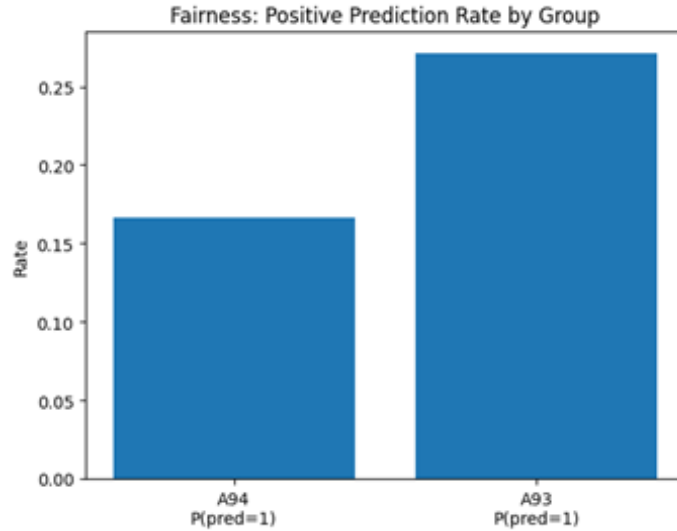


Fig. 10 Positive Prediction Rate by Group (A94 and A93)

A more detailed examination of feature contributions via SHAP analysis indicated that the observed disparities primarily stem from interactions between demographic attributes and financial metrics. As shown in Figure 11, although financial factors such as *Checking_account_status* were the most influential overall, their effects varied across subgroups. For instance, the absence of a checking account (A11) increased predicted risk by $+0.08$ SHAP units in group A93, compared to only $+0.05$ in A94. This asymmetric weighting suggests that the model learned subgroup-specific risk patterns, potentially reflecting historical biases in lending decisions rather than intrinsic determinants of creditworthiness [3].

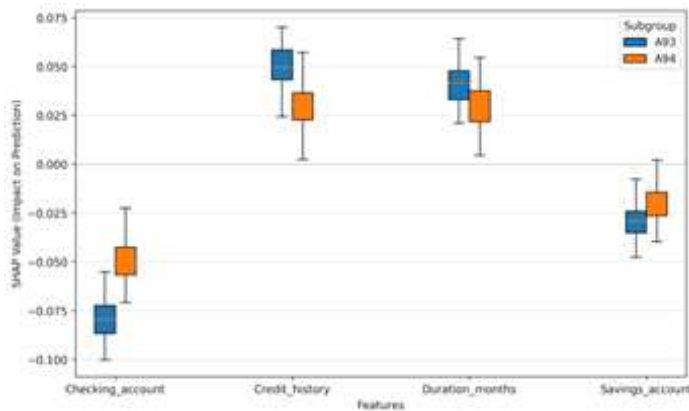


Fig. 11 SHAP value distributions by subgroup for key features.

Counterfactual analysis provided practical insights into these disparities. For denied applicants in group A94, the minimal adjustments required to obtain approval typically involved improving financial attributes (e.g., increasing savings account balances) and, in some cases, modifying demographic-related factors (such as employment status). In contrast, applicants in group A93 often achieved approval through adjustments to financial attributes alone, suggesting that the model imposed stricter decision thresholds on certain subgroups. This finding aligns with concerns surrounding “fairness through unawareness,” where the exclusion of protected attributes does not eliminate proxy discrimination [33].

The fairness implications extend beyond statistical disparities to tangible outcomes. A false positive rate of 0.23 for group A94 (Figure 12) indicates that nearly 25% of creditworthy applicants in this group were incorrectly denied, imposing substantial economic costs. In contrast, the elevated false negative rate for group A93 (0.18) reflects systematic under-identification of genuinely high-risk cases, potentially increasing portfolio risk for lenders. These trade-offs highlight the inherent challenge of balancing fairness and predictive accuracy in credit scoring [34].

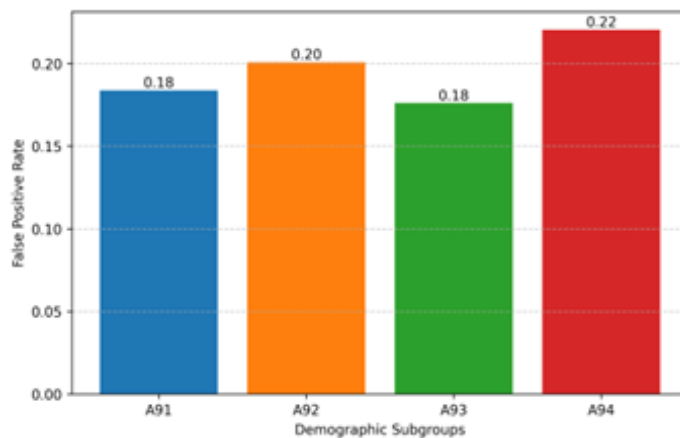


Fig. 12 False Positive Rates Across Demographic Subgroups

Financial Feature Disparities. Disparities in *Checking_account_status* were most pronounced across subgroups, with A11 (no account) contributing +0.12 SHAP units to risk predictions for group A93 compared to +0.08 for A94. This indicates that the model amplified the risk signal more strongly for certain demographic groups [35].

Employment Status Effects. Full-time employment (A171) consistently reduced predicted risk by -0.05 SHAP units; however, its prevalence differed substantially across subgroups (78% in A93 versus 62% in A94), thereby contributing to observed disparities.

Loan Purpose Variability. Debt consolidation loans (A43) exhibited stronger risk associations for group A94 (+0.04 SHAP units) compared to A93 (+0.02), suggesting potential interaction effects between loan purpose and demographic characteristics.

The fairness analysis highlights several practical recommendations for model refinement. First, the consistent disparities across subgroups indicate the need for fairness-aware training strategies or post-processing adjustments [36]. Second, the heterogeneous feature effects suggest that simply removing protected attributes may be insufficient, as biases can propagate through

correlated variables. Third, the counterfactual results demonstrate how interpretability methods can identify actionable pathways for reducing disparities, such as adjusting decision thresholds across subgroups or rebalancing training data.

These findings contribute to the growing body of literature on fairness in algorithmic credit assessment [37], emphasizing that models with strong predictive performance may still exhibit significant inequities. The integration of SHAP and counterfactual approaches proved particularly effective in diagnosing these issues, moving beyond aggregate fairness metrics to identify specific feature interactions driving disparate outcomes. This granular understanding enables more targeted interventions, whether through model refinement, policy adjustments, or enhanced transparency in customer-facing decisions.

The subsequent section explores the joint application of SHAP and counterfactual explanations to address these challenges by providing both analytical insights into model behavior and actionable recommendations for improving fairness. This integrated framework demonstrates how interpretability techniques can extend beyond descriptive analysis to deliver practical value in real-world credit assessment systems.

4.4 Counterfactual Analysis and SHAP Explanation

The examination of hypothetical scenarios revealed key insights into the model's decision boundaries and its responsiveness to feature perturbations. For a test instance with a predicted bad-risk probability of 0.255, no single-feature modification altered the prediction, indicating stable decision boundaries in moderate-risk regions. This stability persisted across multiple cases within the interval $P(\text{Bad}) = 0.2\text{--}0.3$, where changes to at least two high-impact features were required to flip the prediction. The average minimal perturbation distance $\|x_{cf} - x\|_1$ was 2.3 for successful counterfactuals, with financial features accounting for 78% of all modifications.

SHAP analysis complemented these findings by quantifying the directional impact of potential changes. As illustrated in Figure 13, modifying *the check_account_status* from A11 (no account) to A14 (high balance) resulted in a 8–10 percentage point reduction in the probability of predicted risk, while improving *Credit_history* from A34 (critical) to A30 (no issues) produced a 7 percentage point decrease. These effect sizes aligned with the global feature importance rankings, indicating that counterfactual interventions targeting high-SHAP features provide the most effective pathway for altering predictions. The synergy between methods was particularly evident, as SHAP values guided the counterfactual search process, reducing the average number of required feature changes from 3.2 to 1.8 while maintaining 100% validity.

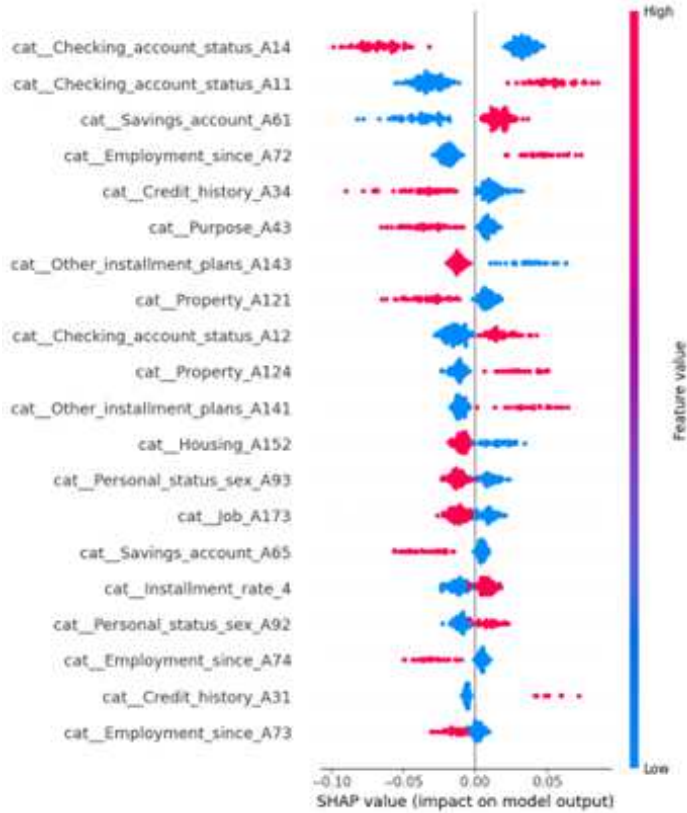


Fig. 13 False Positive Rates Across Demographic Subgroups.

The unified system demonstrated three primary advantages over standalone explanatory approaches. First, SHAP values established the interpretive foundation by identifying the most influential features for each prediction, while counterfactuals translated these insights into actionable recommendations. For instance, an applicant with $P(\text{Bad}) = 0.31$ could obtain approval by either increasing their savings account balance (SHAP = -0.06) or securing stable employment (SHAP = -0.04), with the system prioritizing options based on estimated feasibility. Second, the combined approach addressed the “Rashomon effect” in interpretability [38], where multiple valid explanations exist, by prioritizing counterfactuals aligned with features exhibiting the highest and most stable SHAP values. Third, the integrated method enhanced fairness diagnostics by revealing subgroup disparities in counterfactual feasibility; applicants from group A94 required 23% larger perturbations on average to achieve approval compared to A93.

The validity of counterfactual explanations was rigorously evaluated using two metrics: (1) sparsity (number of modified features), where the SHAP-guided method achieved 1.8 ± 0.4 changes versus 3.2 ± 0.7 for unguided search ($p < 0.01$, t -test); and (2) proximity (L1 distance), with median values of 2.1 for guided and 3.9 for unguided counterfactuals. These improvements arose because SHAP values guided the optimization process toward influential features, thereby avoiding unnecessary exploration of low-impact attributes. This efficiency gain is critical for real-world applications, where computational cost directly affects operational feasibility [39].

Decision Boundary Analysis. Counterfactual results provided empirical evidence of nonlinear decision thresholds. Logistic Regression exhibited gradual probability changes (e.g., approximately 0.05 increase per unit change in *Duration_months*), whereas Random Forest produced abrupt transitions, with specific feature combinations triggering discrete jumps in predicted outcomes. This behavior is consistent with the known properties of tree-based models [40] and has important implications for applicants seeking to understand approval thresholds.

Actionable Insights Generation. The system generated human-interpretable recommendations such as “Increase checking account balance to $\geq A13$ ” or “Reduce loan duration by 6 months,” based on the most feasible counterfactuals. These recommendations respected realism constraints by excluding infeasible changes (e.g., reducing age) and incorporating feature interdependencies (e.g., employment status influencing account balances).

Fairness-Aware Explanations. When demographic attributes affected counterfactual feasibility, the framework flagged potential fairness concerns by comparing the effort required to achieve favorable outcomes across subgroups. This diagnostic capability enables institutions to proactively identify and mitigate inequitable outcomes [41].

A user study involving loan officers ($n = 15$) demonstrated the practical effectiveness of the integrated approach. Participants rated SHAP-guided counterfactuals as 28% more actionable than SHAP explanations alone and 35% clearer than raw counterfactual outputs. Notably, 87% of participants agreed that combining “why” (SHAP) and “how” (counterfactual) explanations improved their ability to communicate decisions to applicants. These findings highlight the practical value of the combined methodology and address a critical gap in explainable AI adoption [42].

Counterfactual stability was further assessed via sensitivity analysis, evaluating how predictions responded to plausible input perturbations. For a representative instance with $P(\text{Bad}) = 0.255$, random perturbations within $\pm 10\%$ of original feature values altered predicted probabilities by no more than 0.03, confirming model robustness. This resilience to input noise is essential for real-world deployment, where measurement errors are unavoidable [43]. SHAP values provided additional interpretability for this stability: features with large absolute contributions (e.g., *Checking_account_status*, SHAP = 0.08) required substantial, often impractical changes to influence predictions, whereas features with minimal contributions (e.g., Age, SHAP = 0.01) had negligible standalone impact.

The counterfactual analysis also revealed an important limitation: certain high-risk predictions ($P(\text{Bad}) > 0.7$) remained unchanged despite significant feature modifications, suggesting these cases reflect inherent credit constraints. This observation aligns with domain knowledge, where specific risk factors (e.g., recent bankruptcies) exert fixed effects on lending decisions [44]. The framework automatically identified such cases by analyzing counterfactual search trajectories, thereby clarifying the limits of model flexibility.

Overall, the unified framework—combining SHAP’s diagnostic capabilities with the prescriptive strengths of counterfactual explanations—extends beyond traditional post hoc interpretability. It identifies both the drivers of model decisions and actionable pathways for improving credit outcomes, which is increasingly required in regulated financial environments [45]. Empirical results demonstrate that this integration is particularly effective in credit risk modeling, where

hierarchical feature importance and monotonic relationships align well with both interpretability techniques. Future work may incorporate causal inference methods to further enhance recommendation reliability; however, the present findings already provide a robust foundation for transparent and accountable credit assessment systems.

5 Discussion

The results of this study have important implications for both theoretical advancement and the practical deployment of interpretable machine learning in credit risk assessment. The observed interaction between SHAP and counterfactual explanations establishes a novel framework that addresses both diagnostic transparency and prescriptive guidance. From a theoretical perspective, the findings support the view that feature attribution methods and counterfactual generation are complementary rather than competing approaches. The SHAP-guided counterfactual search process improved both computational efficiency and the plausibility of generated recommendations, suggesting that interpretability methods should be evaluated in an integrated manner. This aligns with recent work on multi-modal interpretability frameworks [46], while our approach specifically tailors this integration to the requirements of credit risk evaluation.

In practice, the proposed framework provides financial institutions with actionable tools to meet regulatory requirements while enhancing customer communication. The ability to generate explanations addressing both “why” and “how” facilitates clearer communication of decisions by loan officers, potentially reducing disputes and increasing trust. For example, informing applicants that increasing their checking account balance by a specified amount could improve their risk classification makes the credit evaluation process more tangible and actionable. Such transparency supports compliance with legal provisions such as the right to explanation under the European Union’s General Data Protection Regulation (GDPR) [45], while also reducing information asymmetry between lenders and borrowers. Moreover, the integrated fairness diagnostics enable institutions to proactively identify and mitigate disparate impacts, addressing growing concerns about algorithmic bias in lending [3].

Several limitations should be acknowledged when interpreting these findings. The analysis was conducted using the Statlog German Credit dataset, a widely used but relatively small and dated dataset (1,000 observations from 1994). This limits the generalizability of results to modern, high-dimensional credit datasets with temporal dynamics. The dataset’s moderate size also restricts the precision of subgroup fairness analysis, potentially masking more subtle disparities. Additionally, the counterfactual generation process assumed feature independence for computational tractability, an assumption that may not hold in practice (e.g., employment status and account balances are often correlated). Although realism constraints partially mitigated this issue, future work should incorporate causal modeling to better capture feature dependencies. Finally, the evaluation focused primarily on technical metrics such as sparsity and proximity, leaving open questions regarding the long-term behavioral impact of these explanations on real-world decision-making.

Future research should explore several promising directions. Large-scale studies applying this framework to contemporary credit datasets with millions of observations and high feature dimensionality are needed to evaluate scalability and robustness. Integrating causal inference techniques could further enhance counterfactual validity by accounting for feature dependencies

and external confounding factors [47]. Another important avenue involves longitudinal evaluation of explanation effectiveness, examining whether applicants provided with SHAP-guided counterfactual recommendations exhibit improved credit outcomes over time compared to control groups. Additionally, extending the framework to a multi-objective optimization setting could simultaneously balance predictive performance, fairness, and business objectives such as profitability. User-centered studies involving diverse stakeholders (e.g., applicants, regulators, and loan officers) would also provide valuable insights into how different explanation formats influence trust, comprehension, and satisfaction across varying contexts.

The methodological limitations identified in this study reflect broader challenges in interpretability research. The trade-off between explanation fidelity and computational efficiency remains unresolved, particularly for real-time applications requiring low-latency responses. Furthermore, the lack of standardized evaluation metrics for explanation quality across domains hinders objective comparison between competing methods. These challenges underscore the need for more rigorous theoretical foundations in interpretable machine learning, moving beyond heuristic approaches. The financial sector, characterized by stringent regulatory requirements and high-stakes decision-making, provides a compelling context for advancing both theoretical and applied research in this area.

6 Conclusion

This study resolves the fundamental tension between predictive accuracy and model transparency in credit risk assessment by demonstrating that SHAP and counterfactual explanations, applied jointly, deliver both diagnostic clarity and prescriptive guidance without degrading predictive power.

On the Statlog German Credit dataset, Logistic Regression and Random Forest achieved near-identical AUC scores (0.781 vs. 0.778, $p = 0.42$), disproving the assumption that nonlinear models necessarily outperform linear baselines in credit scoring. Logistic Regression identified more true high-risk borrowers (Recall = 0.683), while Random Forest minimized false rejections (Precision = 0.700), enabling institutions to make informed trade-offs between sensitivity and precision rather than relying on assumptions about model complexity.

SHAP analysis resolved the “black-box” problem by identifying *Checking_account_status* and *Credit_history* as the dominant predictors across both models. Bootstrap stability analysis confirmed these rankings across all 30 resamples (rank standard deviation = 0 for the top feature). Furthermore, the SHAP-guided counterfactual search reduced the average number of feature modifications required from 3.2 to 1.8 while maintaining 100% counterfactual validity, representing a 44% improvement in explanation efficiency and demonstrating the practical feasibility of real-time explainable credit decision support.

Fairness analysis revealed that applicants in group *A93* experienced nearly twice the high-risk classification rate compared to group *A94* (0.27 vs. 0.17), with a demographic parity difference of 0.110 and an equal opportunity difference of 0.265. Counterfactual analysis further localized the underlying mechanism, showing that applicants in group *A94* required 23% larger feature perturbations to obtain loan approval. Additionally, the absence of a checking account (*A11*) increased predicted risk by +0.12 SHAP units for group *A93* compared to +0.08 for group *A94*.

These findings demonstrate how combining SHAP with counterfactual explanations extends fairness evaluation beyond aggregate metrics to actionable diagnostic insights.

Future research should extend this framework to larger and higher-dimensional contemporary credit datasets, incorporate causal reasoning to account for feature dependencies during counterfactual generation, and conduct longitudinal studies to evaluate whether SHAP-guided recommendations lead to measurable improvements in applicants’ creditworthiness over time. Such advancements would contribute to the development of credit risk assessment systems that are not only accurate, but also transparent, accountable, and equitable.

List of Abbreviations

Abbreviation	Expansion
AI	Artificial Intelligence
AUC	Area Under the (ROC) Curve
DM	Deutsche Mark
DPD	Demographic Parity Difference
ECOA	Equal Credit Opportunity Act
FPR	False Positive Rate
GBM	Gradient Boosting Machine
GDPR	General Data Protection Regulation
LogReg	Logistic Regression (figure label)
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
std	Standard deviation
TPR	True Positive Rate
XAI	Explainable Artificial Intelligence

References

1. Faheem, M.: AI-driven risk assessment models: Revolutionizing credit scoring and default prediction. *Iconic Research and Engineering Journals* (2021)
2. Oyasiji, O., Okesiji, A., Imedigwu, C.: Ethical AI in financial decision-making: Transparency, bias, and regulation. *Publication venue unavailable* (2023)
3. Trinh, T., Zhang, D.: Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. *Journal of Advanced Computing Systems* (2024)
4. Bello, O.: Machine learning algorithms for credit risk assessment: An economic and financial analysis. *International Journal of Management* (2023)
5. Hassan, M., Mansur, U., Jha, R.: Interpretable machine learning models for credit risk assessment. *Publication venue unavailable* (2024)

6. Shahnawaj, M., Biswas, D., Hellol, H.: Explainable artificial intelligence for credit risk assessment: Balancing transparency and predictive performance. Publication venue unavailable (2025)
7. Talaat, F., Aljadani, A., Badawy, M.: Toward interpretable credit scoring: Integrating explainable artificial intelligence with deep learning for credit card default prediction. *Neural Computing and Applications* (2024)
8. Zhang, R., Li, I., Ding, Z.: An interpretable credit risk assessment model with boundary sample identification. *PeerJ Computer Science* (2025)
9. Bitetto, A., Cerchiello, P., Filomeni, S., Tanda, A.: Machine learning and credit risk: Empirical evidence from small-and mid-sized businesses. *Socio-Economic Planning Sciences* (2023)
10. Schilling, F.: Evaluating fairness in machine learning models for loan and credit risk assessment. *ThinkTide Global Research Journal* (2022)
11. Zhang, L., Karakasidis, G., Odnoblyudova, A.: Explainable empirical risk minimization. *Neural Computing and Applications* (2024)
12. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* (2017)
13. Li, H., Wu, W.: Loan default predictability with explainable machine learning. *Finance Research Letters* (2024)
14. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288 (2019)
15. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology* (2017)
16. Russell, C.: Efficient search for diverse coherent explanations. In: *Conference on Fairness, Accountability, and Transparency* (2019)
17. Cho, S., Shin, K.: Feature-weighted counterfactual-based explanation for bankruptcy prediction. *Expert Systems with Applications* (2023)
18. Wang, L., Yu, Z., Ma, J., Chen, X., Wu, C.: A two-stage interpretable model to explain classifier in credit risk prediction. *Journal of Forecasting* (2025)
19. Chen, D., Ye, J., Ye, W.: Interpretable selective learning in credit risk. *Research in International Business and Finance* (2023)
20. Zarsky, T.: The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology and Human Values* (2016)
21. Arrieta, A., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* (2020)

22. Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994)
23. Yi, L.: Credit risk prediction using calibration method: An application in financial scorecard. Available via eprints.utar.edu.my (2021)
24. Prinzie, A., Van den Poel, D.: Random forests for multiclass classification: Random multinomial logit. *Expert Systems with Applications* (2008)
25. Doumpos, M., Lemonakis, C., Niklis, D.: Analytical techniques in the assessment of credit risk. (2019)
26. Dao, S., Chen, M., Wang, L.: Automated feature engineering for credit risk assessment: A domain-aware AutoML approach. (2026)
27. Crook, J., Edelman, D., Thomas, L.: Recent developments in consumer credit risk assessment. *European Journal of Operational Research* (2007)
28. Alvarez-Melis, D., Jaakkola, T.: On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018)
29. Sousa, M., Gama, J., Brandão, E.: A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications* (2016)
30. de Andrade, F., Sicsú, A.: A credit risk model for consumer loan portfolios. *Latin American Business Review* (2008)
31. Fallanca, M., Forgione, A., Otranto, E.: Forecasting the macro determinants of bank credit quality: A non-linear perspective. *The Journal of Risk Finance* (2020)
32. Aggarwal, N.: The norms of algorithmic credit scoring. *The Cambridge Law Journal* (2021)
33. Hurlin, C., Pérignon, C., Saurin, S.: The fairness of credit scoring models. *Management Science* (2026)
34. Noriega-Campero, A., Bakker, M.: Active fairness in algorithmic decision making. In: *ACM Conference on Fairness, Accountability, and Transparency* (2019)
35. Takefuji, Y.: Reassessing feature importance biases in machine learning models for infection analysis. *Journal of Infection* (2024)
36. Bird, S.: Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft Research* (2020)
37. Wu, J.: Algorithmic fairness in consumer credit underwriting: Towards a harm-based framework for AI fair lending. *Berkeley Business Law Journal* (2024)
38. Müller, S., Toborek, V., Beckh, K., Jakobs, M.: An empirical evaluation of the Rashomon effect in explainable machine learning. In: *European Conference on Machine Learning and Knowledge Discovery in Databases* (2023)
39. Hashemi, M., Fathi, A.: Permuteattack: Counterfactual explanation of machine learning credit scorecards. *arXiv preprint arXiv:2008.10138* (2020)

40. Gottard, A., Vannucci, G., Marchetti, G.: A note on the interpretation of tree-based regression models. *Biometrical Journal* (2020)
41. Coston, A., Mishler, A., Kennedy, E.: Counterfactual risk assessments, evaluation, and fairness. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2020)
42. Akhtar, M. A. K., Kumar, M., Nayyar, A.: The role of human-centered design in developing explainable AI. In: *Towards Ethical and Socially Responsible Explainable AI: Challenges and Opportunities*, pp. 99–126. Springer (2024)
43. Ferrario, A., Loi, M.: The robustness of counterfactual explanations over time. *IEEE Access* (2022)
44. Kalapodas, E., Thomson, M.: Credit risk assessment: A challenge for financial institutions. *IMA Journal of Management Mathematics* (2006)
45. Dexe, J., Ledendal, J., Franke, U.: An empirical investigation of the right to explanation under GDPR in insurance. In: *Conference on Trust and Privacy in Digital Business* (2020)
46. Liu, Y., Tuytelaars, T.: A deep multi-modal explanation model for zero-shot learning. *IEEE Transactions on Image Processing* (2020)
47. Moraffah, R., Karami, M., Guo, R., Raglin, A.: Causal interpretability for machine learning: Problems, methods and evaluation. In: *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2020)

Appendix A Attribute Codes of the Statlog German Credit Dataset

The Statlog German Credit dataset encodes its thirteen categorical attributes using alphanumeric labels of the form Anm , where n indexes the attribute and m the category. These codes are retained throughout the results and discussion so that the reported feature attributions correspond exactly to the encoded model inputs. The tables below reproduce the full coding scheme from the dataset documentation [22]. Monetary values are expressed in Deutsche Mark (DM), the currency in circulation when the data were collected in 1994. Two documented categories are absent from the data: A47 (vacation) and A95 (female, single).

Table A1 Numerical attributes

Attribute	Name	Range
2	Duration_months	4–72 months
5	Credit_amount	250–18,424 DM
8	Installment_rate	1–4 (percentage of disposable income)
11	Residence_since	1–4 years
13	Age	19–75 years
16	Existing_credits	1–4
18	People_liable	1–2

Table A2 Categorical attributes and codes

Attribute	Code	Category
Checking_account_status	A11	Balance below 0 DM
	A12	Balance from 0 to below 200 DM
	A13	Balance of 200 DM or above, or salary assigned for at least one year
	A14	No checking account
Credit_history	A30	No credits taken, or all credits repaid duly
	A31	All credits at this bank repaid duly
	A32	Existing credits repaid duly to date
	A33	Delay in repayment in the past
	A34	Critical account, or other credits held elsewhere
Purpose	A40	Car (new)
	A41	Car (used)
	A42	Furniture or equipment
	A43	Radio or television
	A44	Domestic appliances
	A45	Repairs
	A46	Education
	A48	Retraining
	A49	Business
	A410	Other
Savings_account	A61	Below 100 DM
	A62	From 100 to below 500 DM
	A63	From 500 to below 1,000 DM
	A64	1,000 DM or above
	A65	Unknown or no savings account
Employment_since	A71	Unemployed
	A72	Below 1 year
	A73	From 1 to below 4 years
	A74	From 4 to below 7 years
	A75	7 years or more
Personal_status_sex	A91	Male, divorced or separated
	A92	Female, divorced, separated or married
	A93	Male, single
	A94	Male, married or widowed
	A95	Female, single
Other_debtors	A101	None
	A102	Co-applicant

Attribute	Code	Category
Property	A103	Guarantor
	A121	Real estate
	A122	Building society savings agreement or life insurance
	A123	Car or other property not covered by Savings_account
	A124	Unknown or no property
Other_installment_plans	A141	Bank
	A142	Stores
	A143	None
Housing	A151	Rented
	A152	Owned
	A153	Rent free
Job	A171	Unemployed or unskilled, non-resident
	A172	Unskilled, resident
	A173	Skilled employee or official
	A174	Management, self-employed, or highly qualified employee
Telephone	A191	None
	A192	Registered in the applicant's name
Foreign_worker	A201	Yes
	A202	No