Information Systems and Economics (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/infse.2026.070201
ISSN 2523-6407 Vol. 7 Num. 2

A Robust and Interpretable Framework for Cross-Market Credit Risk Prediction

Yu Bai, Yile Wang*

School of Economics, Beijing Technology and Business University, Beijing, 100048, China
**Corresponding author: wyl13393401611@126.com*

Keywords: Credit Risk Prediction; Cross-Market Analysis; Model Robustness; SHAP; Model Interpretability

Abstract: To address challenges in cross-market credit risk prediction—such as performance disparities across models, the instability of core risk factors, and a lack of sufficient interpretability—this paper proposes a research framework that integrates predictive performance comparison, model specification robustness testing, and interpretability analysis. Using three publicly available datasets—covering credit card defaults in Chinese Taiwan, credit risk in Germany, and credit approvals in Australia—we conduct predictive experiments employing methods such as logistic regression, random forests, and Extreme Gradient Boosting (XGBoost), evaluating classification performance via the Area Under the Curve (AUC). Furthermore, we examine the sensitivity of core variable estimates to model specifications through stepwise specification testing and specification curve analysis. We also analyze the contribution of risk factors—and the consistency across different interpretability methods—by utilizing Average Marginal Effects (AME) and SHapley Additive exPlanations (SHAP). Our experimental results indicate that the optimal predictive model varies by market, with the highest AUC values reaching 0.7732 for Chinese Taiwan, 0.7459 for Germany, and 0.9131 for Australia. Robustness analysis reveals that while some core variables maintain a consistent direction across various model specifications, others—such as loan amount in the German dataset—exhibit greater sensitivity to the chosen analytical setup. In our comparison of interpretability results, 74 out of 84 identified factors (approximately 88%) show consistent directional effects across the three markets; among variables with statistically significant Average Marginal Effects, this consistency rate rises to approximately 98%.

1. Introduction

Credit risk prediction is a critical component of financial risk management and credit decision-making [1]. Its primary objective is to leverage a borrower's historical records, repayment behavior, and other relevant characteristics to identify potential high-risk clients, thereby providing financial institutions with a basis for risk assessment and credit management. With the advancement of data analytics, both traditional statistical models and machine learning models have been increasingly applied to credit risk prediction; each approach offers distinct advantages in terms of predictive performance, scope of applicability, and interpretability [2]. While traditional statistical models

provide a clear analytical logic, machine learning models excel at extracting insights from complex data relationships, offering new perspectives for risk identification. However, because customer characteristics, variable distributions, and risk labels vary across different credit markets, a model that performs well in one market may not necessarily yield the same results in another [3]. Consequently, focusing solely on predictive performance within a single dataset is insufficient for a comprehensive understanding of a method's applicability; it is also essential to examine how risk factors behave—and how stable they remain—across different markets.

Although existing research on credit risk prediction has yielded a diverse array of models and analytical methods, there remains room for further refinement. On one hand, many studies evaluate model performance primarily through metrics such as accuracy or the area under the curve (AUC), often overlooking how key variables shift under different model specifications. When control variables or analytical conditions change, the estimated direction and magnitude of certain risk factors may also shift; relying on the results of a single model risks ignoring the inherent uncertainty of the conclusions [4]. On the other hand, traditional statistical analysis and machine learning interpretability techniques emphasize different aspects, and employing either approach in isolation may fail to fully capture the relationship between variables and predictive outcomes [5]. Furthermore, because variable definitions, sample sizes, and risk labels differ across markets, these variations must be accounted for when directly comparing model results. Given these challenges, it is necessary to integrate predictive performance comparisons, stability testing of model results, and the interpretation of predictions. By analyzing credit risk prediction from multiple angles, one can enhance the completeness and practical utility of research findings.

Addressing these issues, this paper investigates the robustness and interpretability of credit risk prediction across different markets. Using datasets—specifically, credit card default data from Chinese Taiwan, credit data from Germany, and credit approval data from Australia—this study conducts a comparative analysis of credit risk prediction tasks across these distinct markets. Regarding the modeling approach, this paper employs logistic regression as a representative traditional statistical method, alongside machine learning techniques such as random forests and gradient-boosted trees, to evaluate the predictive performance of various models. Building on this, the study systematically adjusts model specifications to assess the stability of the estimates for core variables and examines how results fluctuate across different analytical configurations. For the purpose of interpretation, the paper utilizes Average Marginal Effects (AME) to analyze the associations between variables and predicted risk, while employing SHapley Additive exPlanations (SHAP) to quantify the contribution of individual features to the predictions generated by machine learning models [6]. By integrating these two interpretive approaches, the study compares the directional effects of core variables and their performance across distinct markets, providing complementary analytical insights into the model predictions [7].

The primary contributions of this paper are threefold. First, by conducting a systematic comparison across three credit markets with diverse data characteristics, the study evaluates the predictive performance of traditional statistical models versus representative machine learning models, offering empirical guidance for model selection in different market contexts. Second, the paper combines model specification stability analysis with specification curve analysis to examine the sensitivity of core risk factor estimates to changes in model configuration, thereby avoiding conclusions drawn from a single, rigid model specification. Finally, by synthesizing Average Marginal Effects with SHAP analysis, the study investigates the directional influence and relative importance of core variables from both traditional statistical and machine learning perspectives, further exploring the consistency and divergence between these interpretive outcomes [8].

2. Methodology

2.1. Overall Research Framework

This study develops a robustness and interpretability framework for cross-market credit risk prediction, as illustrated in Fig. 1. First, publicly available credit datasets from Chinese Taiwan, Germany, and Australia are selected, and their variables and prediction tasks are organized under a consistent analytical setup. Traditional statistical models and machine-learning models are then used to compare predictive performance [9-10]. Next, the stability of core risk factors is examined across alternative model specifications, while average marginal effects and SHAP are used to interpret model predictions. By integrating predictive performance, result robustness, and model interpretation into a unified workflow, the framework reduces the limitations of drawing conclusions from a single model or market.

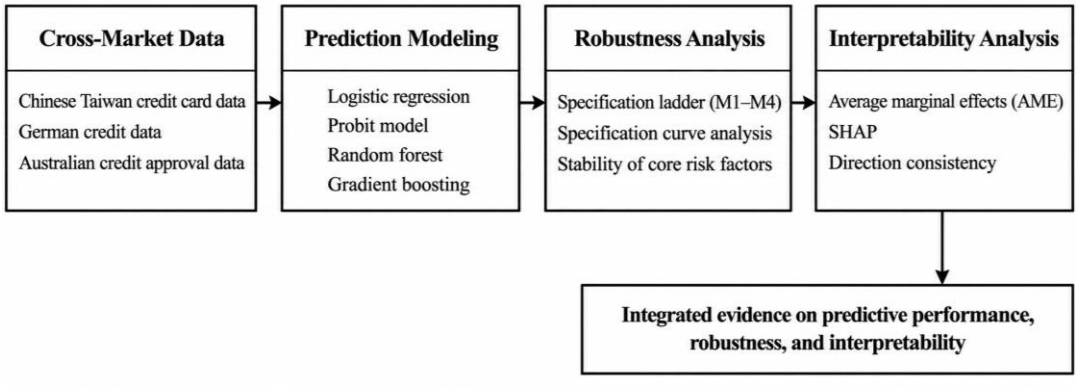

Figure 1: Overall Framework of the Proposed

2.2. Credit Risk Prediction Models

Building on the overall framework, this study employs logistic regression, the probit model, and representative machine-learning models for credit risk prediction [11-12]. For a binary prediction task, let y_i denote the risk-event label of sample i, and let x_i denote its feature vector. Logistic regression and the probit model can be expressed in a unified form as follows:

$$P(y_i = 1 | \mathrm{x}_i) = F(\mathrm{x}_i^T \beta) \tag{1}$$

Here, the cumulative distribution function $F(\cdot)$ is the logistic distribution function for logistic regression and the standard normal distribution function for the probit model. The parameter vector is estimated by maximum likelihood, and the model output represents the predicted probability of a risk event.

For the machine-learning component, random forests, gradient-boosting trees, and other models actually used in the original experiments are included in the comparison. Predictive performance is primarily evaluated using the area under the curve (AUC), which measures a model's ability to distinguish between risk classes. Comparing model performance across the three markets helps determine whether model rankings vary with market characteristics and provides a basis for the subsequent robustness and interpretability analyses.

2.3. Specification-Robustness Analysis

Because estimation results may depend on the choice of control variables and functional form, this study uses a stepwise specification test and specification-curve analysis to assess the stability of core

risk factors [13-14]. First, four progressively expanded specifications, M1–M4, are considered. M1 includes only the core variables; M2 adds demographic characteristics; M3 incorporates the full set of control variables; and M4 further adds a squared age term. The estimated direction, marginal effect, and statistical significance of core variables are compared across specifications to determine whether the findings depend on a particular model setup.

A specification-curve analysis is then conducted. Holding the baseline risk-event definition fixed, the control variable set, treatment of continuous variables, and estimation method are varied, yielding 48 analytical specifications for each market. Estimates and confidence intervals across these specifications are summarized to identify directional stability and sensitivity to analytical choices. When a variable changes sign across specifications, the results are examined in relation to the alternative settings rather than treating the estimate from a single specification as a stable conclusion.

2.4. Interpretability Analysis

After evaluating predictive performance and specification robustness, this study combines average marginal effects (AMEs) with SHapley Additive exPlanations (SHAP) to provide complementary interpretations of core risk factors [15]. For a continuous variable, the AME represents the average change in predicted risk probability associated with a change in that variable. For a binary variable, it is calculated as the average discrete change in predicted probability when the variable changes from 0 to 1. The definitions are:

$$AME_j = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial P(y_i = 1 \mid \mathrm{x}_i)}{\partial x_{ij}} \tag{2}$$

$$AME_j = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{p}_i(d_j = 1) - \hat{p}_i(d_j = 0)\right] \tag{3}$$

Here, n denotes the number of observations, p_i denotes the predicted probability, and d_j denotes a binary variable. These measures are used to examine the direction and average magnitude of the association between core variables and predicted risk.

In parallel, SHAP is used to examine the contribution of each feature to the predictions of a machine-learning model. Its additive explanation takes the following form:

$$f(\mathrm{x}) = \phi_0 + \sum_{j=1}^{M}\phi_j \tag{4}$$

Here, (fx) denotes the model output, φ_0 is the baseline output, φ_j is the SHAP contribution of feature j, and M is the number of features. The distribution and relative importance of SHAP values are used to identify the features that contribute most to model predictions. Finally, the directions indicated by SHAP contributions are compared with the estimated directions of AMEs to examine agreement and differences in the interpretation of core risk factors, providing complementary evidence for cross-market credit risk prediction.

In summary, this study integrates predictive model comparison, specification-robustness tests, and complementary interpretation methods into a coherent analytical workflow. This workflow provides the methodological basis for the experimental design and empirical results discussed in the following chapters.

3. Experimental Design

3.1. Datasets and Cross-Market Characteristics

This study uses the Chinese Taiwan credit card dataset, the German Statlog credit dataset, and the Australian credit approval dataset to examine predictive performance and risk-factor characteristics across different credit markets. The three datasets correspond to credit card default prediction, credit risk classification, and credit approval classification, respectively, and provide a basis for comparing models under different data conditions. Their original sample sizes are 30,000, 1,000, and 690, respectively. Because the datasets differ in risk-label definitions, variable meanings, and sample sizes, their original task characteristics are retained during model comparison, and the variables are processed according to the corresponding variable mapping and preprocessing procedures. The original sample size may differ from the final number of valid observations after preprocessing; the effective sample size should therefore be determined from the actual analysis records.

3.2. Experimental Setup

After the datasets are selected, necessary preprocessing is performed, including data cleaning, missing-value handling, categorical-variable encoding, and feature organization. Input processing is adapted to the requirements of each model. Logistic regression, the probit model, random forests, gradient-boosting methods, and other applicable models are then used for credit risk prediction and performance comparison. For experiments using stratified five-fold cross-validation, the samples are divided into five subsets while preserving class proportions as far as possible. Each subset is used once for validation, while the remaining subsets are used for training, and the evaluation results across folds are aggregated. For experiments using an independent test set, model performance is evaluated according to the corresponding data-splitting procedure. Keeping the evaluation metrics and comparison criteria consistent where applicable enables an analysis of model performance across the three markets and provides a basis for subsequent robustness and interpretability analyses.

3.3. Evaluation Metrics

To evaluate predictive capability and the stability of core risk factors, this study mainly uses the area under the curve (AUC), average marginal effects (AMEs), confidence intervals, and directional consistency. AUC measures a model's ability to distinguish between risk classes and is defined as the area under the receiver operating characteristic (ROC) curve:

$$AUC = \int_0^1 TPR(u)du \tag{5}$$

Here, TPR denotes the true positive rate, and u denotes the false positive rate along the horizontal axis. A higher AUC generally indicates a stronger ability to distinguish between risk classes.

In addition, AME is used to examine the average association between core variables and predicted risk probabilities. Confidence intervals indicate the uncertainty of estimates, while directional consistency assesses whether the directions of core variables remain consistent across model specifications or interpretation methods.

4. Results and Discussion

4.1. Cross-Market Prediction Performance

Table 1 summarizes the out-of-sample AUC values of five representative models across the three

markets. For Chinese Taiwan, CatBoost achieves the highest AUC of 0.7732. For Germany, logistic regression obtains an AUC of 0.7459, outperforming the four selected machine-learning models. For Australia, XGBoost records the highest AUC at 0.9131. The best-performing model therefore differs across markets, suggesting that model performance depends on the characteristics of the data and that a single market cannot establish a universally optimal model.

In Chinese Taiwan, CatBoost improves the AUC by 0.0137 relative to logistic regression, while in Australia, XGBoost improves it by 0.0085. In contrast, the AUC values of random forest and XGBoost in Germany are 0.7209 and 0.7216, respectively, both below that of logistic regression. Greater model complexity therefore does not necessarily yield better predictive performance. Because fold-level standard errors and significance tests for AUC are not available in the reported results, these relatively small numerical differences should not be interpreted as statistically significant advantages.

Table 1: AUC Comparison of Five Representative Models across Three Markets

Model	Chinese Taiwan	Germany	Australia
Logistic Regression (L2)	0.7595	0.7459	0.9046
Random Forest	0.7682	0.7209	0.9092
XGBoost	0.7715	0.7216	0.9131
SVM (RBF)	0.6911	0.7023	0.8794
CatBoost	0.7732	0.7352	0.8987

The AUC ranges across the selected models are 0.0821, 0.0436, and 0.0337 for Chinese Taiwan, Germany, and Australia, respectively. Model performance varies more widely in Chinese Taiwan, whereas the five models produce more closely clustered AUC values in Australia. The three datasets do not share identical prediction tasks or label definitions; these AUC values should therefore be used primarily to compare models within each market, rather than to rank the intrinsic difficulty of risk prediction across markets.

4.2. Robustness of Core Risk Factors

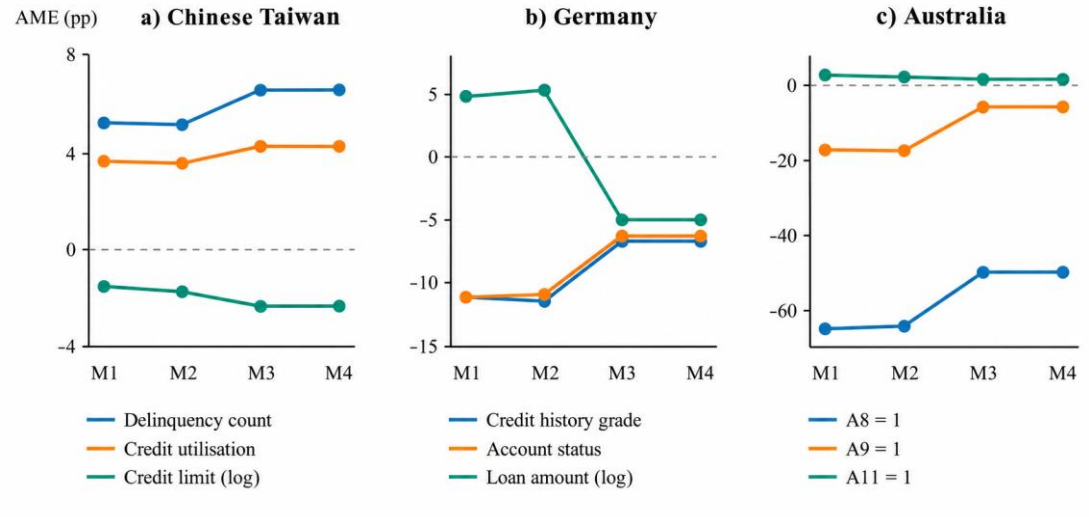

Figure 2: Specification-Ladder Stability of Core Risk Factors

Figure 2 presents the changes in average marginal effects for core variables across four specifications, M1–M4. In the Chinese Taiwan dataset, the marginal effect of the number of past-due payments increases from 5.26 percentage points in M1 to 6.59 percentage points in M3, while the

effect of credit utilization rises from 3.69 to 4.31 percentage points. The marginal effect of log credit limit remains negative, changing from −1.55 to −2.34 percentage points. All three variables retain the same direction across the four specifications, indicating directional stability under progressive changes to the control variables.

Germany exhibits a different pattern. The marginal effects of credit-history rating and account status remain negative across all four specifications, although their absolute magnitudes decrease after the full set of controls is added. By contrast, the marginal effect of log loan amount changes from 4.82 percentage points in M1 and 5.31 percentage points in M2 to −4.98 percentage points in M3, remaining negative in M4. This sign reversal indicates that the conditional association between loan amount and predicted risk is sensitive to the control-variable specification and should not be interpreted solely from a simpler model.

In Australia, the marginal effect for $A_8 = 1$ remains negative, moving from −64.99 percentage points in M1 to −50.05 percentage points in M3. $A_9 = 1$ also remains negative, while its absolute magnitude declines from 16.98 to 5.51 percentage points. $A_{11} = 1$ remains positive, with its estimate decreasing from 2.80 percentage points in M1 to 1.82 percentage points in M3. Overall, specification-ladder analysis shows that directional stability and stability in magnitude are distinct: a variable may retain its sign while its estimated magnitude changes appreciably.

4.3. Specification-Curve Analysis

Adding control variables sequentially reveals how estimates change under selected specifications, but it does not cover all reasonable analytical choices. This study therefore enumerates alternative combinations of control variables, variable treatments, estimation methods, and risk-event definitions. The specification-curve analysis includes 192 specifications for Chinese Taiwan and 48 each for Germany and Australia. Table 2 summarizes the distribution of estimates, sign consistency, and statistical-significance rates for selected variables.

Table 2: Specification-Curve Results for Core Risk Factors

Market	**Core variable**	**No. of specs.**	**Median estimate (pp)**	**2.5%–97.5% quantiles**	**Sign consistency**	**Significance rate**
Chinese Taiwan	Number of past-due payments	192	7.82	[1.80, 20.02]	100.0%	100.0%
Chinese Taiwan	Credit utilization	192	5.93	[2.27, 13.42]	100.0%	100.0%
Chinese Taiwan	Log credit limit	192	−0.31	[−2.28, 0.89]	52.1%	73.4%
Germany	Credit-history rating	48	−11.05	[−18.67, −6.55]	100.0%	100.0%
Germany	Account status	48	−9.93	[−13.67, −6.31]	100.0%	100.0%
Germany	Log loan amount	48	0.91	[−8.72, 6.37]	50.0%	50.0%
Australia	$A_8 = 1$	48	−60.77	[−80.89,−49.92]	100.0%	100.0%
Australia	$A_9 = 1$	48	−13.62	[−18.21, −3.63]	100.0%	62.5%
Australia	$A_{11} = 1$	48	2.27	[1.75, 3.03]	100.0%	0.0%

The results for German loan amount provide a clear example of specification sensitivity. Among

the 48 estimates, 24 are negative and 24 are positive. The confidence intervals of 24 estimates lie entirely above zero, whereas the remaining 24 intervals cross zero. This agrees with the 50% sign-consistency and 50% significance rates in Table 3, indicating that both the direction and statistical significance of the loan-amount estimate are sensitive to analytical choices.

By comparison, credit-history rating and account status in Germany remain negative and statistically significant at the 5% level across all 48 specifications. The number of past-due payments and credit utilization in Chinese Taiwan also remain positive and significant across all 192 specifications. However, the sign-consistency rate for Chinese Taiwan's log credit limit is only 52.1%, indicating substantial sensitivity for this variable as well. These findings suggest that robustness should be assessed variable by variable rather than inferred from an overall market-level or model-level performance measure.

4.4. Cross-Market Analysis of Average Marginal Effects

After identifying the specification stability of individual variables, this section examines marginal effects and estimation uncertainty under the full-control M3 specification. Table 3 reports coefficient estimates, average marginal effects, standard errors, and standardized effects for selected variables. Figure 3 presents the point estimates and 95% confidence intervals, showing their direction and uncertainty.

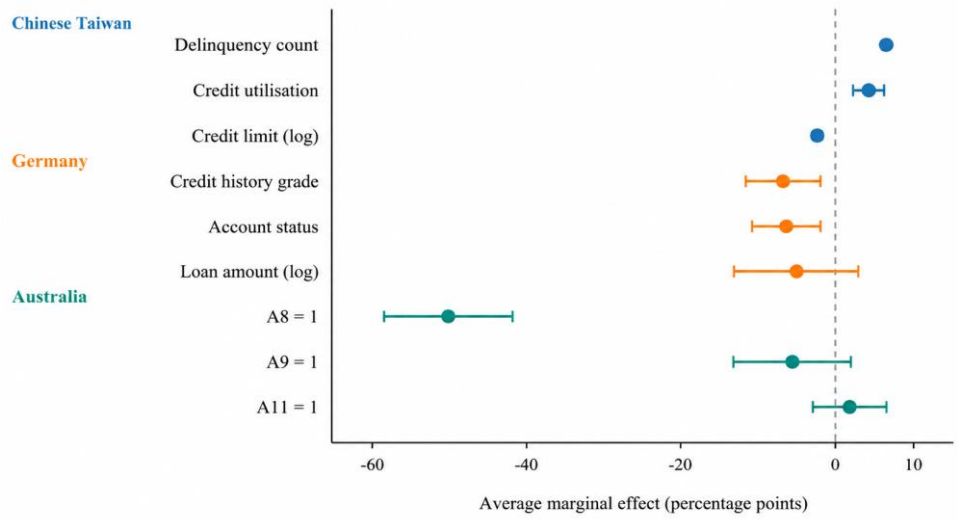

Figure 3: Average Marginal Effects of Core Variables Across Markets

Table 3: Average Marginal Effects of Core Variables Under Specification M3

Market	Variable	Coefficient	AME (pp)	SE (pp)	Standardized effect	Significance
Chinese Taiwan	Number of past-due payments	0.472	6.6	0.3	0.298	***
Chinese Taiwan	Credit utilization	0.308	4.3	1.0	0.195	***
Chinese Taiwan	Log credit limit	−0.167	−2.3	0.3	−0.106	***
Chinese Taiwan	Log average	−0.111	−1.5	0.2	−0.070	***

	repayment amount					
Germany	Credit-history rating	−0.456	−6.7	2.4	−0.224	***
Germany	Account status	−0.429	−6.3	2.2	−0.210	***
Germany	Log loan amount	−0.338	−5.0	4.0	−0.166	Not significant
Australia	$A_8 = 1$	−3.488	−50.1	4.2	−0.902	***
Australia	$A_9 = 1$	−0.565	−5.5	3.8	−0.099	Not significant
Australia	$A_{11} = 1$	0.199	1.8	2.4	0.033	Not significant

In Chinese Taiwan, the positive marginal effect of the number of past-due payments is 6.6 percentage points, exceeding the 4.3 percentage-point effect of credit utilization. Log credit limit and log average repayment amount are negatively associated with predicted risk, and both estimates are significant at the 1% level. In Germany, credit-history rating and account status have significantly negative estimates, whereas the standard error for log loan amount reaches 4.0 percentage points and its estimate is not statistically significant. In Australia, the estimated marginal effect for $A_8 = 1$ is −50.1 percentage points, markedly larger in absolute magnitude than those of the other displayed variables. However, neither $A_9 = 1$ nor $A_{11} = 1$ is statistically significant.

These results should be interpreted in light of variable coding and the task definitions of the respective datasets. In particular, German credit-history rating uses an ordinal encoding, while A_8, A_9, and A_{11} are anonymized variables in the Australian dataset. Their coefficient signs should not be translated directly into economic claims about higher or lower risk without considering the coding scheme. Similarly, the large marginal effect in Australia should not be used on its own to rank risk intensity across markets.

4.5. Model Interpretability Based on SHAP

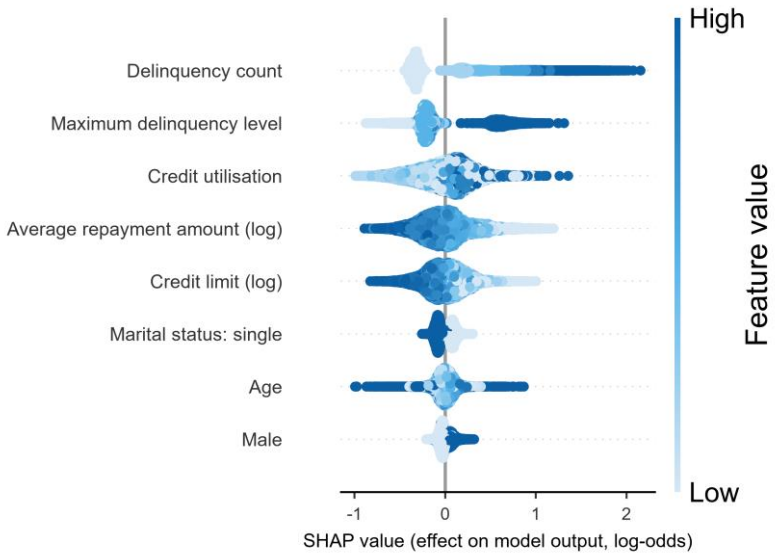

Figure 4: SHAP Summary Plot of Core Risk Factors

To examine how machine-learning models use different features to generate predictions, a SHAP beeswarm plot is used to show the distribution of feature contributions in the Chinese Taiwan dataset. In Figure 4, features are ranked by overall importance, the horizontal axis represents SHAP values, colors indicate whether feature values are relatively high or low, and the distribution of points reflects differences in contributions across individual samples.

The number of past-due payments and the maximum delinquency level occupy the first two positions in the feature ranking. High values of these features are generally associated with positive SHAP values, indicating relatively strong positive contributions to the model output. SHAP values for credit utilization extend to both sides of zero, suggesting that its contribution is not identical across all samples. By contrast, high values of average repayment amount and credit limit are generally associated with negative SHAP values, indicating contribution patterns different from those of the delinquency-related features.

The SHAP distributions for age and gender are relatively concentrated around zero, and their overall influence on model output is weaker than that of core behavioral variables such as past-due payments. Contributions from marital status are also mainly concentrated near zero. These patterns indicate that predictions depend not only on relative feature importance but also on feature values and their associated contribution directions. However, SHAP values measure feature contributions to predictions from a particular model rather than causal effects estimated after controlling for other factors. For features whose distributions overlap substantially, a uniform monotonic relationship should not be inferred from the beeswarm plot alone.

4.6. Consistency between SHAP and AME Interpretations

After examining feature contributions from an individual machine-learning model, the directions of SHAP contributions are compared with the average marginal effects from logistic regression. Figure 5 presents representative variables under specification M2 for Chinese Taiwan, using the sign of the logistic-regression AME as the reference direction. For the number of past-due payments, credit utilization, credit limit, and average repayment amount, the SHAP directions from LightGBM, XGBoost, and Stacking all agree with the AME directions. This indicates that these core variables provide consistent directional information across the interpretation methods.

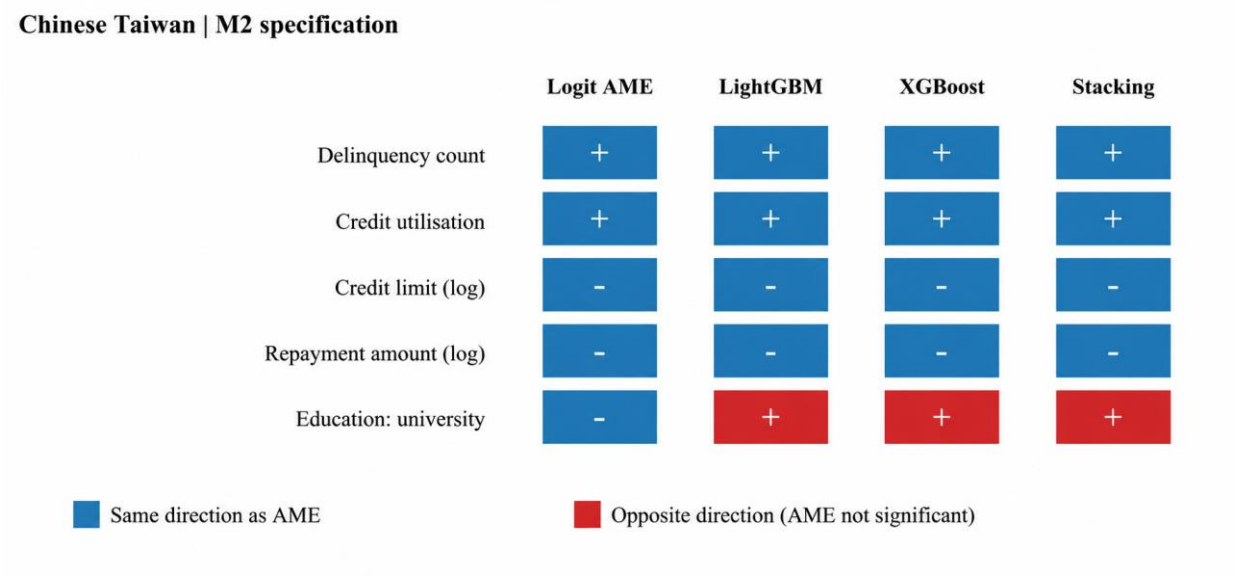

Figure 5: Consistency between SHAP and AME Interpretations

In contrast, the AME for the university-education variable is not statistically significant, while the

SHAP directions from all three machine-learning models differ from the direction of its point estimate. This suggests that when statistical evidence is weak, the two interpretation approaches may provide different directional signals, and conclusions should not rely on the interpretation of a single model alone.

Across all variables, 37 of 42 directional comparisons agree in Chinese Taiwan, corresponding to 88%. All 33 comparisons with statistically significant AMEs are directionally consistent. Across the three markets, 74 of 84 comparisons agree, also corresponding to 88%. Among the 57 comparisons with statistically significant AMEs, 56 are directionally consistent, or 98%. Overall, AME and SHAP exhibit high directional consistency for variables with statistically significant AMEs. This agreement does not mean that the two methods share the same numerical scale, nor does it independently establish a causal relationship between a variable and credit risk.

5. Conclusion

This paper addresses the issues of model robustness and interpretability in cross-market credit risk prediction by constructing an analytical framework that encompasses predictive performance comparison, model specification testing, and multi-method interpretation. The research demonstrates that credit risk prediction should not only focus on a model's classification capability but also examine the sensitivity of the model specification to core risk factors, as well as the extent to which different interpretation methods provide mutually reinforcing insights. By integrating predictive evaluation with result interpretation, this paper offers a systematic approach for comparing model performance across diverse credit markets, providing a reference for shifting credit risk prediction research from a narrow focus on performance metrics toward a more comprehensive evaluation framework.

This study is subject to certain limitations. First, the datasets employed vary in terms of sample size, variable definitions, and risk labels—and are primarily drawn from publicly available benchmarks—meaning the applicability of these findings to other markets or real-world operational environments requires further validation. Second, while the robustness analysis considers various model specifications, it cannot encompass every possible combination of variables, model configuration, or market scenario. Furthermore, although average marginal effects and SHAP values—which interpret predictions from the perspectives of statistical models and machine learning models, respectively—may show consistent directional alignment, this provides complementary evidence rather than establishing direct causal relationships.

References

[1] Montevechi A A, de Carvalho Miranda R, Medeiros A L, et al. Advancing credit risk modelling with Machine Learning: A comprehensive review of the state-of-the-art[J]. Engineering Applications of Artificial Intelligence, 2024, 137: 109082.

[2] Nikolaidis D, Doumpos M. Credit scoring with drift adaptation using local regions of competence[C]//Operations research forum. Cham: Springer International Publishing, 2022, 3(4): 67.

[3] Chen D, Ye J, Ye W. Interpretable selective learning in credit risk[J]. Research in international business and finance, 2023, 65: 101940.

[4] Hashimoto R, Miura K, Yoshizaki Y. Application of Machine Learning to a Credit Rating Classification Model: Techniques for Improving the Explainability of Machine Learning[R]. Bank of Japan, 2023.

[5] Chen Y, Calabrese R, Martin-Barragan B. Interpretable machine learning for imbalanced credit scoring datasets[J]. European Journal of Operational Research, 2024, 312(1): 357-372.

[6] Li H, Wu W. Loan default predictability with explainable machine learning[J]. Finance Research Letters, 2024, 60: 104867.

[7] Famà A, Myftiu J, Pagnottoni P, et al. Explainable machine learning for financial risk management: two practical use cases[J]. Statistics, 2024, 58(5): 1267-1282.

[8] Hlongwane R, Ramabao K, Mongwe W. A novel framework for enhancing transparency in credit scoring: Leveraging

Shapley values for interpretable credit scorecards[J]. Plos one, 2024, 19(8): e0308718.
[9] Atteba R G, Shwe T, Mendonça I, et al. Credit Evaluation Through Integration of Supervised and Unsupervised Machine Learning: Empirical Improvement and Unsupervised Component Analysis[J]. Applied Sciences, 2025, 15(24): 13020.6
[10] Uddin M S, Chi G, Al Janabi M A M, et al. Modeling credit risk with a multi-stage hybrid model: An alternative statistical approach[J]. Journal of Forecasting, 2022, 41(7): 1386-1415.
[11] Mita T A B. Machine Learning-Driven Credit Risk Modeling: Transforming Loan Default Prediction and Portfolio Management in US Commercial Banking[J]. American Journal of Data Science and Analytics, 2025, 6(12): 01-42.
[12] Machado M R, Chen D T, Osterrieder J R. An analytical approach to credit risk assessment using machine learning models[J]. Decision Analytics Journal, 2025, 16: 100605.
[13] Stern J, Krämer M D, Schumacher A, et al. Differences between lifelong singles and ever-partnered individuals in Big Five personality traits and life satisfaction[J]. Psychological Science, 2024, 35(12): 1364-1381.
[14] Girardi P, Vesely A, Lakens D, et al. Post-selection inference in multiverse analysis (PIMA): An inferential framework based on the sign flipping score test[J]. psychometrika, 2024, 89(2): 542-568.
[15] Neupane K, Griva I, Axtell R, et al. Detecting and Explaining Unlawful Insider Trading: A Shapley Value and Causal Forest Approach to Identifying Key Drivers and Causal Relationships[J]. Computational Economics, 2026: 1-33.