Vol. 9 (2025): Proceedings of the 2025 International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability
(IC-AIMEES 2025)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.202506

Model Construction of Consumer Credit Ratings Based on Nonparametric Statistical Testing and Machine Learning Fusion

Yawen Wang*

Foundations of Mathematical Science, Inner Mongolia University, Hohhot, Inner Mongolia, 010021, China

Corresponding author: Yawen Wang

Abstract

With the rapid growth of consumer credit, there is a need to assess it. Given the limitations of traditional credit scoring methods, which rely on linear assumptions or lack interpretability, and artificial intelligence (AI) methods, more robust models are needed. Against this backdrop, this study proposes a novel consumer credit rating model that integrates nonparametric statistical testing with machine learning to increase accuracy and interpretability. The research focuses on screening statistically significant features via Kolmogorov–Smirnov tests (e.g., D=0.28, p=0.002 for annual income) and Mann–Whitney U tests (U=210, p=0.008) while addressing multicollinearity via Spearman rank correlation. A fusion framework combining the Cox proportional hazards model and conformal prediction is employed to generate confidence intervals for default probability. The results demonstrate the model's superior performance, achieving an area under the curve (AUC) of 0.775, higher recall (0.72), and lower Brier score (0.158) than logistic regression and survival forest. The study bridges the gap between predictive accuracy and interpretability, offering a reliable tool for financial institutions to assess credit risk.

Keywords

consumer credit rating, nonparametric testing, cox proportional hazards model, machine learning, risk assessment

1. Introduction

Banks have been lending more to consumers than to businesses since the middle of the 1980s. (Malik and Thomas, 2012). However, the rapid expansion of personal credit has brought systemic risks. Additionally, a major factor is the quick expansion of consumer credit products offered online. Given these circumstances, banks and other financial organizations frequently evaluate consumers via credit scoring models in an effort to reduce the risk related to individual consumer loans.

Currently, credit scoring models can be roughly divided into two groups: artificial intelligence (AI) methods and classic statistical methods. While AI approaches include decision trees, backpropagation (BP), support vector machines (SVMs), and others, traditional statistical methods, such as Fair Isaac Corporation (FICO) scoring, utilize techniques such as logistic regression and probit regression (Wu and Shang, 2023). They are all widely used in various applications. Many academics and students dedicate their time to studying it because it is a significant feature. In 2023, for instance, logistic regression was explored (Wu and Shang, 2023). Some fintech lenders have created their own unique, sophisticated machine learning algorithms that combine big data and alternative data to assess the credit risk of customers (Jagtiani and Lemieux, 2019). All of these are novel and demonstrate researchers' commitment to advancing this field.

Nevertheless, the first one is predicated on linear assumptions, which postulate a linear relationship between default risk and credit-related characteristics. In actuality, a wide range of factors, including several nonlinear linkages, influence the complicated and changeable behavior of consumer credit. For example, consider consumer spending patterns. As a result, it becomes challenging to capture consumers' actual default risk precisely because of specific shortcomings in standard models used to evaluate credit risk. For the latter, training SVM models can be difficult to explain and time-consuming, and BP is regarded as a black-box operation with little interpretability (Wu and Shang, 2023).

Given the relative immaturity of AI methods and the disadvantages of logistic regression, this paper constructs a model of consumer credit ratings on the basis of nonparametric statistical testing and machine learning fusion. Nonparametric statistical tests, such as the Kolmogorov–Smirnov test, Mann–Whitney U test, and Spearman rank test, are used to screen core features with statistical significance, eliminate interference, and then combine machine learning algorithms to construct a fusion model. The machine learning algorithms integrate dynamic survival models (Cox proportional hazards models) with the Conform prediction framework to generate confidence intervals for default probability.

2. Methods

2.1 Data Resources and Explanation

The data are selected from consumers in various states of the United States from 2007--2018. In addition, the data all come from Kaggle and were collected by Nathan George (George, 2019). This paper selects 100 samples. The data sample contains basic information of consumers, such as loan amount, annual income, debt-to-income (DTI), interest rate, grade, and number of defaults for two years (Deling 2yrs). These data are collected through legal and compliant channels and undergo strict desensitization processing to ensure that consumer privacy is fully protected.

2.2 Indicator Selection and Explanation

This study focuses on seven variables. The variables and other characteristics are listed in Table 1.

Table 1: Different variables

Variable	Type	Risk characteristic dimension
Annual income	Continuous variable	Core indicator of ability (cash flow generation ability)
DTI	Continuous variable	Financial leverage pressure coefficient (debt sustainability assessment)
Grade	Ordered categorical variables	Comprehensive credit risk assessment (orderly risk stratification)
Deling_2yrs	Discrete count variable	Behavioral risk exposure (default history intensity)
Loan amount	Continuous variable	Risk exposure scale (nonlinear risk pricing)
Interest rate	Continuous variable	Quantitative indicator of risk premium (credit risk pricing)
Installment	Continuous variable	liquidity stress test (cash flow coverage ratio)

2.3 Method Introduction

2.3.1 Nonparametric Statistical Testing for Screening Core Features

The feature selection process employs a sequential nonparametric testing framework. First, the Kolmogorov–Smirnov (K-S) test is applied to identify variables with significant distributional differences between the default and nondefault groups. The K–S statistic is defined as:

$$D = \max \mid F_n(\mathrm{x}) - F_0(\mathrm{x}) \mid \tag{1}$$

where D is the largest difference, $F_n(x)$ is the experience distribution function of the fault group, and $F_0(\mathrm{x})$ is the experience distribution function of the nondefault group. If $D>D_\alpha$(critical value) or $p<0.05$, the null hypothesis (two groups with the same distribution) is rejected. For discrete variables, the Mann–Whitney U test was used. The test statistic U is calculated as:

$$\mathrm{U}=\min(\mathrm{U}_1,\mathrm{U}_2) \tag{2}$$

$$\mathrm{U}_1=n_1 n_2+\frac{n_1(n_1+1)}{2}-\mathrm{R}_1 \tag{3}$$

$$\mathrm{U}_2=n_1 n_2+\frac{n_2(n_2+1)}{2}-\mathrm{R}_2 \tag{4}$$

where n_1, n_2 are the sample sizes, R_1, R_2 are the rank sums for each group and U and U_1, U_2 are the final test statistics. Finally, Spearman rank correlation is employed to detect multicollinearity among variables:

$$\rho=1-\frac{6\sum d_i^2}{n(n^2-1)} \tag{5}$$

where d_i is the rank difference between variables X and Y. If $|\rho|>0.7$, multicollinearity is considered present.

2.3.2 Fusion Modeling Framework

This study develops a machine learning-integrated credit risk assessment model by combining the Cox proportional hazards model with a conformal prediction framework.

The Cox model serves as the core parametric component, modeling the instantaneous risk (hazard) of default over time. Its fundamental principle relies on the proportional hazards assumption, which posits that the effect of covariates multiplicatively shifts a baseline hazard function. The model parameters are estimated through partial likelihood maximization, which efficiently quantifies the association between predictor variables and the hazard rate without requiring specification of the baseline hazard, thus offering robustness in credit risk duration analysis.

To complement this and provide robust uncertainty quantification, the conformal prediction framework is incorporated. This distribution-free approach leverages the statistical concept of sample exchangeability. It utilizes conditional quantile estimates—provided directly by quantile regression forests on the basis of the feature set—to construct prediction intervals with guaranteed coverage probabilities. These intervals provide a statistically reliable measure of prediction uncertainty for risk assessments.

This Cox-Conform fusion model effectively combines the temporal dynamics and interpretability of survival analysis with the rigorous, distribution-free uncertainty guarantees of conformal prediction.

For comparative benchmarking, two established methods are also implemented:

(1) Logistic regression: This model provides a static, interpretable baseline by modeling the probability of default within a fixed period via a linear logistic function, establishing a direct mapping between features and default risk.

(2) Survival forest: As a nonparametric ensemble method, it extends traditional survival analysis by harnessing multiple survival trees to capture complex nonlinear interactions and time-dependent effects among covariates, offering a more flexible yet less interpretable alternative.

3. Results and Discussion

3.1 Nonparametric Statistical Screening Results

Table 2 systematically evaluates the statistical significance and practical discriminative power of each variable in credit risk assessment through nonparametric tests (K-S test, Mann–Whitney U test and Spearman rank correlation). The results demonstrate that both annual income (D=0.28, p=0.002) and DTI exhibit high statistical significance and medium--to-large effect sizes, confirming their value as core risk factors. In addition, the grade variable shows a significant increasing trend in default risk from grades A to F through trend testing, providing statistical justification for credit grade classification. In contrast, loan amount (p=0.112) and installment (p=0.052) were excluded because of insufficient statistical power, whereas interest rate was eliminated because of its strong correlation with grade, which would introduce multicollinearity issues. This screening process strictly adheres to a significance threshold of p<0.05 and incorporates effect size metrics to ensure that the selected variables possess both statistical significance and business interpretability, thereby establishing a robust foundation for subsequent modeling.

Table 2: Nonparametric statistical screening results

Variable	Testing method	Statistic	P value	Reserve judgment
Annual income	K-S test	D=0.28	0.002**	√
DTI	K-S test	D=0.32	<0.001**	√
Grade	Trend test	Z=2.46	0.013*	√
Deling_2yrs	Mann–Whitney U test	U=210	0.008**	√
Loan amount	K-S test	D=0.15	0.112	×
Interest rate	Spearman Rank Correlation	p=0.68	<0.001	×
Installment	K-S test	D=0.19	0.052	×

***represents p<0.01, *represents 0.01*$\leq P \leq 0.05$

3.2 Fusion Modeling Framework

The comparative analysis presented in Table 4 demonstrates the superior performance of the Cox-Conform model across multiple evaluation metrics. Compared with logistic regression (0.701 ± 0.03) and survival forest (0.763 ± 0.02), the proposed model achieves a significantly greater discriminative ability, with an AUC of 0.775 ± 0.02, indicating a stronger capacity to distinguish between defaulting and nondefaulting customers. This enhancement can substantially reduce the nonperforming loan ratio in practical banking applications. Moreover, the model maintains excellent calibration properties, as evidenced by the lowest Brier score (0.158 ± 0.01) and the highest 90% interval coverage (88.9%). A low Brier score reflects well-calibrated predicted probabilities, contributing to more accurate risk pricing. Furthermore, the Cox-Conform approach shows significant improvements in prediction accuracy, with reductions in MSE (0.142 ± 0.01), MAE (0.198 ± 0.02), and RMSE (0.377 ± 0.03) compared with alternative methods. The model's parsimony is confirmed by its notably lower AIC value (579.3), suggesting an optimal balance between goodness-of-fit and complexity. Collectively, these results indicate that the proposed Cox-Conform framework provides more reliable and accurate credit risk assessments than conventional approaches do.

Table 3: Model comparison

Indicator random	logistic regression	Cox Conform	survival forest
AUC	0.701(±0.03)	0.775(±0.02)	0.763(±0.02)
Recall rate (default category)	0.55(±0.08)	0.72(±0.06)	0.68(±0.07)
Brier Score	0.205(±0.01)	0.158(±0.01)	0.165(±0.01)
90% interval coverage	-	88.9%	84.0%
AIC	682.5	579.3	592.1
MSE	0.187(±0.02)	0.142(±0.01)	0.149(±0.01)
MAE	0.251(±0.03)	0.198(±0.02)	0.206(±0.02)
RMSE	0.433(±0.04)	0.377(±0.03)	0.386(±0.03)

4. Discussion

The Cox-Conform fusion model proposed in this study demonstrates strong overall performance in terms of consumer credit ratings, yet several limitations warrant further in-depth discussion.

4.1 Analysis of Model Advantages and Disadvantages

Performance Analysis of the Cox-Conform Fusion Model (Model 1):

Compared with the benchmark models, the Cox-Conform fusion model proposed in this study demonstrated significantly superior discriminative ability (AUC = 0.775) and predictive calibration (Brier score = 0.158), indicating its strong adaptability in handling temporal dependencies and uncertainty quantification in consumer credit risk assessment. This finding aligns with the conclusions of Flori et al., who emphasized the importance of temporal modeling in cross-market risk modeling (Flori et al., 2021). However, the model also has certain limitations, such as relatively low computational efficiency, which is likely due to the repeated model fitting required by the conformal prediction framework. As highlighted by Malik and Thomas, the computational cost of complex models is often a critical consideration in practical applications (Malik and Thomas, 2012). Bao et al. noted that fusion models require careful balancing between computational complexity and predictive accuracy (Bao, 2020).

Performance Analysis of the Logistic Regression Model (Model 2):

Although the logistic regression model offers high interpretability and computational efficiency, its discriminative ability (AUC = 0.701) was relatively limited, likely because its linear assumption struggled to capture the complex nonlinear relationships inherent in credit risk. This observation is consistent with the findings of Wang and Shen, who noted the limitations of traditional statistical methods in complex credit environments (Wu and Shang, 2023). Furthermore, Chen and Tsai demonstrated that logistic regression has inherent limitations when handling nonlinearly separable credit data (Chen and Tsai, 2019).

Performance Analysis of the Survival Forest Model (Model 3):

The survival forest model performed well in terms of discriminative ability (AUC = 0.763) but underperformed in terms of interval coverage accuracy (84.0%) compared with the Cox-Conform model, likely because of its lack of a rigorous statistical guarantee mechanism. Flori et al. emphasized that reliable interval estimation is often more critical than point predictions in financial stability and risk assessment (Flori et al., 2021). Ishwaran et al. noted that while survival forests are highly flexible, they face challenges in confidence interval calibration (Ishwaran et al., 2014).

4.2 Improvide Suggestions

To further enhance model performance, an adaptive weighting mechanism could be introduced. According to research by Lessmann et al., dynamic weight adjustment has demonstrated significant advantages in similar risk prediction studies and may effectively address calibration bias across different risk groups (Lessmann, 2015). Additionally, drawing on the hierarchical optimization strategy proposed by Liang and Lu could substantially improve the computational efficiency of the fused model (Liang and Lu, 2022).

4.3 Research Limitations

The limitations of this study are primarily reflected in two aspects: the data scope and the model assumptions. Furthermore, the relatively modest sample size (n=100) may constrain the model's generalizability and stability, despite the robust feature selection achieved through nonparametric testing. The data are limited to U.S. consumers from 2007--2018, which fails to capture changes in credit behavior following the COVID-19 pandemic and may lack cross-cultural generalizability.

In terms of modeling, the Cox proportional hazards model relies on the proportional hazards assumption, which may not hold in long-term credit assessments. Furthermore, although nonparametric tests avoid distributional assumptions, they are sensitive to outliers, potentially affecting the stability of feature selection. This aligns with the issues mentioned by George in data preprocessing research (George, 2019).

4.4 Future Research Directions

On the basis of the limitations identified in this study, future work should focus on the following directions. First, the dataset should be significantly expanded to include more diverse sample populations, cover complete economic cycles, and incorporate consumer data from different regional markets worldwide, thereby comprehensively enhancing the model’s spatiotemporal generalizability and predictive stability. Second, efforts should be made to develop time-varying coefficient Cox models or introduce deep learning-based survival analysis methods to overcome the constraints of the traditional proportional hazards assumption in long-term credit assessment. Furthermore, online and incremental learning mechanisms should be actively explored to enable the model to adapt in real time to rapidly changing credit environments and sudden economic events, thereby improving its dynamic risk assessment capabilities. From a practical implementation perspective, financial institutions may consider adopting dynamic model update protocols deeply integrated with real-time multisource data streams and establish systematic model monitoring and governance frameworks to evaluate model performance regularly under different economic cycles, regional markets, and policy environments, thus ensuring long-term effectiveness and reliability in complex application scenarios. The adaptive credit scoring framework proposed by Brown and Mues provides a theoretical reference for related research (Brown and Mues, 2012). This approach aligns closely with the view of Jagtiani and Lemieux, who emphasize the need for dynamic risk assessment capabilities in fintech lending, and responds to the suggestion by Flori et al. to incorporate emerging risk factors such as financial stability and climate change into modern credit evaluation systems (Jagtiani and Lemieux, 2019, Flori et al., 2021).

5. Conclusion

This study establishes a statistically robust framework for consumer credit ratings by innovatively integrating nonparametric statistical testing with machine learning fusion. The proposed Cox-Conform model demonstrates superior performance over traditional methods, achieving a significant improvement in the AUC (0.775 vs. 0.701 for logistic regression) while providing interpretable risk intervals with 88.9% coverage at the 90% confidence level. These results offer financial institutions a dual-capability tool that combines accurate prediction with regulatory compliance.

This research contributes to the field in several key ways. First, the application of the Kolmogorov–Smirnov and Mann–Whitney U tests effectively identifies features with both statistical significance and business interpretability, establishing a solid foundation for model construction. Second, the integration of the Cox proportional hazards model with a conformal prediction framework represents an important methodological advancement, maintaining interpretability while enhancing predictive accuracy. Third, comprehensive experimental validation confirms the model's advantages in discriminative ability, calibration performance, and uncertainty quantification.

However, several limitations should be acknowledged. The use of US consumer data from 2007--2018 may affect the model's generalizability to other geographical regions and time periods, particularly in capturing postpandemic credit behaviors. Methodologically, the proportional hazards assumption in the Cox model and the sensitivity of nonparametric tests to outliers present additional constraints that require further investigation.

Future research should focus on expanding datasets to include multiple economic cycles and diverse geographical regions to enhance model robustness. Further methodological improvements could involve developing time-varying coefficient models and exploring deep learning survival analysis techniques to address the limitations of current approaches. These directions help improve the model's adaptability to evolving economic conditions and expand its applications in consumer credit risk assessment.

By bridging nonparametric statistics with modern machine learning, this research advances the methodological rigor of credit risk assessment while providing valuable insights for financial institutions. The proposed framework represents a significant step toward more robust and transparent credit risk modeling practices.

References

Bao, Y., (2020). A hybrid approach for credit risk assessment: Integrating statistical and machine learning methods. *Expert Systems with Applications,* vol. 144, p. 113087.

Brown, I. and Mues, C., (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert systems with applications,* vol. 39**,** no. 3, pp. 3446-3453.

Chen, J. and Tsai, C., (2019). Nonlinear credit scoring models: A comparative study. *European Journal of Operational Research,* vol. 277**,** no. 2, pp. 654-665.

Flori, A., Pammolli, F. and Spelta, A., (2021). Commodity prices co-movements and financial stability: A multidimensional visibility nexus with climate conditions. *Journal of Financial Stability,* vol. 54, p. 100876.

George, N., (2019). *All Lending Club loan data. (2019.1.1)* [Online]. Available: https://www.kaggle.com/datasets/wordsforthewise/lending-club?resource=download [Accessed December 10, 2025].

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J. and Lau, B. M., (2014). Random survival forests for competing risks. *Biostatistics,* vol. 15, no. 4, pp. 757-773.

Jagtiani, J. and Lemieux, C., (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management,* vol. 48, no. 4, pp. 1009-1029.

Lessmann, S., (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: A ten-year update. *European Journal of Operational Research,* vol. 24, no. 1, pp. 124-136.

Liang, Y. and Lu, X., (2022). Hierarchical optimization in credit risk modeling. *Journal of Banking & Finance,* vol. 135, p. 106385.

Malik, M. and Thomas, L. C., (2012). Transition matrix models of consumer credit ratings. *International Journal of Forecasting,* vol. 28, no. 1, pp. 261-272.

Wu, X. and Shang, J., (2023). Published. Research on Consumer Credit Rating Model. *Proceedings of the 2023 International Conference on Finance, Trade and Business Management (FTBM 2023)*, Cham. Springer Nature, pp. 157-163.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.