Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 357-364

Segment-Aware Credit Risk Modeling of Personal Loans Using Random Forest

Deep Shikha [1], Himanashu Vasnani [2]

[1] Research Scholar, Faculty of Science, Suresh Gyan Vihar University, Jaipur, India
[2] Assistant Professor, Suresh Gyan Vihar University, Jaipur, India

Email: himanshu.vasnani@mygyanvihar.com

Abstract:- Traditional credit risk assessment for personal loans mostly relies on global probability of default (PD) models and credit score–centric policies. While such methods are regulatorily accepted, they often fail to capture borrower heterogeneity arising from differences in financial capacity and behavioral stability. This study recommends a segment-aware credit risk modeling framework using a Random Forest classifier, where borrowers are segmented based on a combination of financial and behavioral attributes before model training. Using a large-scale real-world loan dataset, we benchmark a traditional Logistic Regression PD model, a global Random Forest model, and multiple segmented Random Forest models. Empirical results demonstrate that segmentation based on income, debt-to-income ratio, and employment stability significantly improves recall and F1-score compared to both global machine learning and traditional PD models. The results highlight the limitations of credit-score-centric risk policies and deliver evidence for a policy shift toward finance- and behavior-driven credit risk assessment.

Keywords:- Credit Risk Modeling, Personal Loans, Borrower Segmentation, Random Forest, Probability of Default, Behavioral Finance

1. Introduction

Personal loan portfolios form a important portion of unsecured lending for banks and fintech institutions, making accurate credit risk assessment needed for minimizing defaults while maintaining financial inclusion [1]. Traditionally, lenders belief on Probability of Default (PD) models developed using Logistic Regression, which are broadly accepted due to their interpretability and regulatory compliance [2]. These models are mainly driven by credit bureau scores such as FICO or CIBIL, which summarize borrower creditworthiness into a single numerical measure [3]. However, the linear assumptions underlying such models often bound their ability to capture borrower heterogeneity [4].

Current advancements in machine learning (ML) have permitted the adoption of non-

Correspondence to: ***Deep Shikha** , Faculty of Science**, Suresh Gyan Vihar University, India***
Corresponding author. E-mail addresses: himanshu.vasnani@mygyanvihar.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 357-364

linear models for credit risk prediction [5]. Models such as Random Forests and Gradient Boosting techniques have confirmed superior predictive performance by capturing complex interactions among borrower characteristics [6]. Despite these advantages, many ML-based credit risk studies endure to employ global models that implicitly assume a homogeneous borrower population [7].

In practice, borrowers differ considerably in terms of earning capacity, financial leverage, and employment (earning) stability, all of which directly influence repayment performance and default risk [8]. Ignoring such heterogeneity may reduce the efficiency of both traditional and machine learning–based credit risk models.

This research study claims that borrower segmentation based on financial and behavioural characteristics, when combined with machine learning, can increase credit risk prediction [9]. Accordingly, a segment aware random forest framework is proposed and evaluated against conventional logistic regression PD models and global machine learning approaches. The objective is to assess whether segmentation-driven modelling leads to improved default exposure while preserving model robustness and regulatory relevance [10].

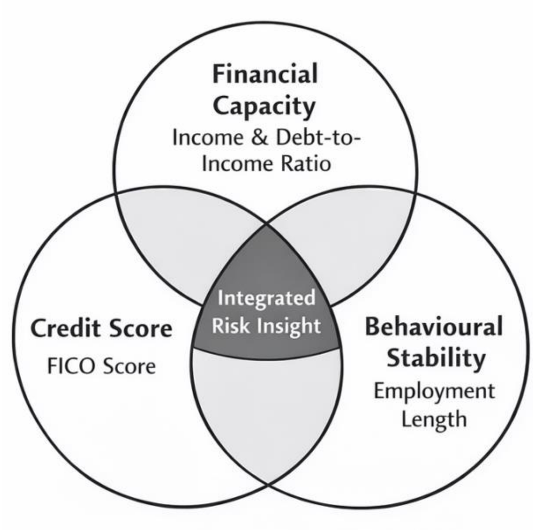


Figure 1. Conceptual Venn diagram illustrating the interaction between financial capacity, behavioural stability, and credit risk

Figure 1 demonstrates the conceptual overlay between financial capacity, behavioural stability, and credit risk, emphasizing how segmentation across these dimensions can improve default prediction.

2. Literature Review

Initial research in credit risk assessment primarily depends on statistical techniques such as logistic regression and linear discriminant analysis due to their ease, interpretability, and ease of regulatory acceptance [11]. These models shaped the foundation of traditional probability of default (PD) frameworks and were widely used by financial institutions for consumer credit evaluation [12].

With the growing availability of large scale loan datasets and enhancements in

Correspondence to: ***Deep Shikha*** *, Faculty of Science*, ***Suresh Gyan Vihar University, India***
Corresponding author. E-mail addresses: himanshu.vasnani@mygyanvihar.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 357-364

computational competences, researchers began to discover machine learning techniques for credit risk prediction [13]. Several relative studies have shown that machine learning models outperform traditional statistical approaches by efficiently capturing nonlinear relationships among borrower attributes [14]. Among these methods, ensemble based models such as random forests have demonstrated strong predictive performance and robustness when applied to credit scoring problems [15].

Despite the performance gains offered by machine learning models, concerns related to transparency, interpretability, and fairness have been widely debated in the literature, particularly in regulated financial environments [16]. As a result, multiple studies highlight the need to balance predictive accuracy with model explainability to ensure regulatory compliance and trust in automated credit decision systems [17].

Borrower segmentation has been widely studied in correlated domains such as marketing analytics and portfolio risk management, where grouping customers based on shared characteristics has been revealed to improve decision-making outcomes [18]. In the context of credit risk assessment, earlier research has largely motivated on segmenting borrowers using demographic variables or credit score–based rules [19]. However, such approaches often fail to capture deeper behavioural and financial differences among borrowers.

Current studies focus the importance of incorporating behavioural stability indicators, such as employment continuity and repayment behaviour, into credit risk modelling frameworks [20]. Similarly, financial capacity measures, including income level and debt load, have been initiate to provide substantial explanatory power beyond traditional credit scores [21]. Nevertheless, the integration of borrower segmentation with machine learning techniques for credit risk prediction remains relatively limited in existing literature.

Building on earlier research, this study integrates financial capacity and behavioural stability based segmentation with a random Forest based modelling framework. By doing so, it aims to report the limitations of global credit risk models and contribute experiential evidence on the effectiveness of segment aware machine learning approaches for personal loan default prediction [22].

3. Dataset Description

The experiential analysis in this study is conducted using the Lending Club loan dataset, which is a widely used and publicly available dataset for credit risk research. The dataset contains detailed loan level information and has been widely adopted in earlier studies for estimating credit risk models due to its scale and real world significance.

The dataset contains both approved and rejected loan applications and consists of historical loan records spanning from 2007 to

Correspondence to: ***Deep Shikha*** *, Faculty of Science****, Suresh Gyan Vihar University, India***
Corresponding author. E-mail addresses: himanshu.vasnani@mygyanvihar.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 357-364

2018. It was obtained directly from the kaggle open source data repository, ensuring data authenticity and reproducibility [23].

After execution data cleaning and pre-processing, approximately 100,000 loan observations were randomly selected for experimental analysis. The pre-processing steps included handling missing values, removing inconsistent records, and ensuring feature consistency across all observations.

3.1 Target Variable

Loan Status: A binary result in which fully paid loans are classified as non-default (0) and charged-off debts as default (1).

3.2 Feature Set

All models make consistent use of the following features:

- **Annual Income** (financial capacity)
- **Debt-to-Income Ratio (DTI)** (financial stress)
- **Employment Length** (behavioral stability)
- **FICO Score (Low Range)** (credit history)
- **Loan Amount**
- **Interest Rate**

Categorical employment length is transformed into numerical years after missing values are eliminated.

4. Methodology

4.1 Baseline PD Model

The conventional PD policy framework is represented by a Logistic Regression model with class-balanced weighting. Before estimating the model, features are standardized.

4.2 Global Machine Learning Model

Without segmentation, the complete data set is used to train a Random Forest classifier. Although it makes the assumption that the borrower population is homogeneous, this model captures non-linear interactions.

Model	Algorithm	Segmentation	Purpose
Baseline PD Model	Logistic Regression	No	Represents traditional PD policy
Global ML Model	Random Forest	No	Captures non-linear interactions

4.3 Segment-Aware Modeling Framework

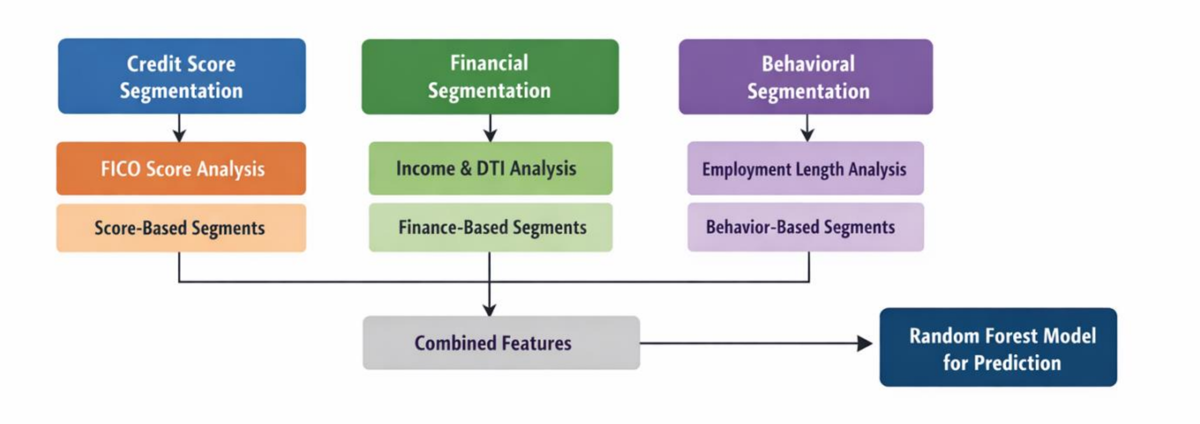
Before the model is trained, borrowers are divided into groups according to behavioral and financial characteristics using domain-driven rules:

*Correspondence to: **Deep Shikha** , Faculty of Science, **Suresh Gyan Vihar University, India***
Corresponding author. E-mail addresses: himanshu.vasnani@mygyanvihar.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 357-364

- **Financial Segmentation**: Income (Low/High) and DTI (Low/High)
- **Behavioral Segmentation**: Employment Length (≤ 3 years / > 3 years)
- **Credit Score Segmentation**: FICO ≤ 660 / > 660

Multiple segment combinations are evaluated.



An independent Random Forest model is trained for every segment, and predictions are combined for all segments.

4.4 Evaluation Metrics

The model's performance is assessed using:

- **Recall** (ability to identify defaulters)
- **F1 Score** (balance between precision and recall)
- **AUC (Area Under ROC Curve)** (ranking ability)

Recall is emphasized due to its significance in credit risk management.

5. Experimental Results

5.1 Model Benchmarking

Model	Recall	F1	AUC
Logistic Regression (PD Policy)	89%	41%	71%

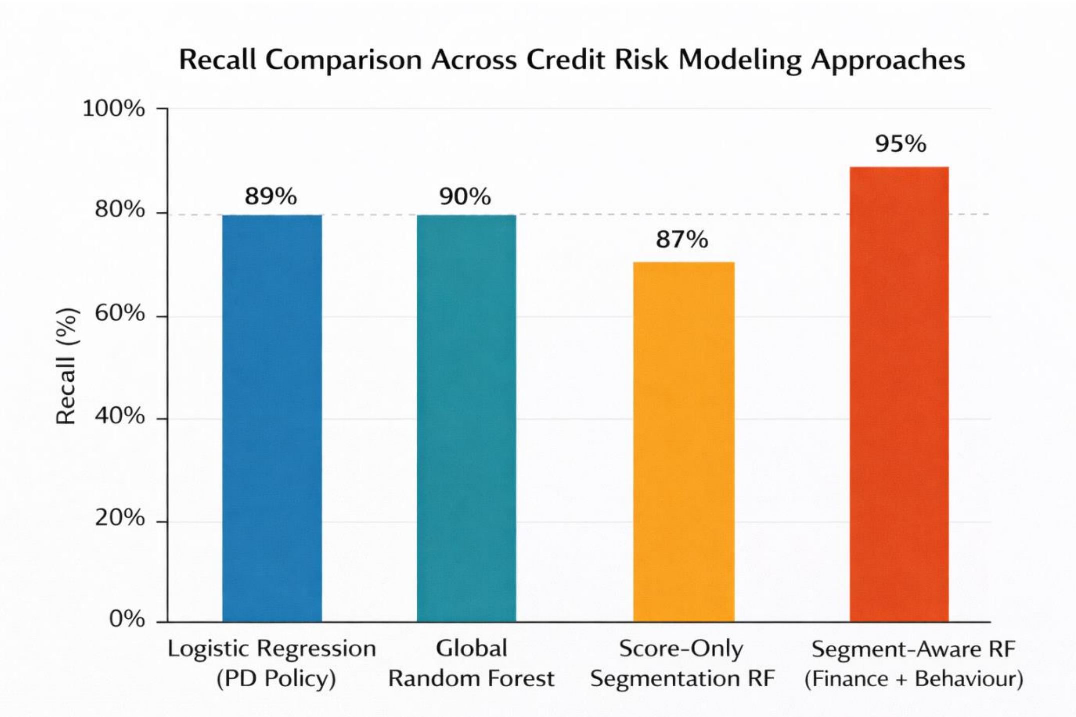
*Correspondence to: **Deep Shikha** , Faculty of Science, **Suresh Gyan Vihar University, India***
Corresponding author. E-mail addresses: himanshu.vasnani@mygyanvihar.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 357-364

Global Random Forest (Without Segmentation)	90%	40%	71%
Segment-Aware RF (Finance + Behaviour)	**95%**	**44%**	69%

5.2 Analysis

The benefit of non-linear modelling is confirmed by the global random forest's slight lead over the logistic regression PD model. However, the introduction of borrower segmentation based on behavioural stability and financial capacity provides the most gains. Superior default identification is established by the segment-aware random forest, which increases recall by more than 6% over the PD policy baseline.



It's interesting to note that segmentation based only on credit score results in lower ranking performance, highlighting the drawbacks of risk rules that are score-centric.

6. Discussion and Policy Implications

The conclusions imply that borrowers' earnings and income stability provide a more inclusive explanation of credit risk than credit ratings alone. This inspires a move away from strict credit-score thresholds and toward dynamic, segmentation-driven machine learning models from a regulatory standpoint. By preventing needless rejection of borrowers

Correspondence to: ***Deep Shikha*** *, Faculty of Science,* ***Suresh Gyan Vihar University, India***
Corresponding author. E-mail addresses: himanshu.vasnani@mygyanvihar.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 357-364

with weak or flawed credit histories, this policy can improve risk control while easing wider financial inclusion.

7. Conclusion

This research study establishes that segment-aware credit risk modelling with random forest performs highly better than global machine learning techniques and conventional PD models. Lenders can increase default detection without compromising model robustness by combining behavioural stability and financial capacity for borrower segmentation.

This method may be expanded in the future by employing explainable AI methods, time-series behavioural data, and stress testing under macroeconomic conditions.

8. References

[1] D. J. Hand and W. E. Henley, “Statistical classification methods in consumer credit scoring: A review,” *Journal of the Royal Statistical Society: Series A*, vol. 160, no. 3, pp. 523–541, 1997.

[2] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, “Benchmarking state-of-the-art classification algorithms for credit scoring,” *Journal of the Operational Research Society*, vol. 54, no. 6, pp. 627–635, 2003.

[3] L. C. Thomas, “A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers,” *International Journal of Forecasting*, vol. 16, no. 2, pp. 149–172, 2000.

[4] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, “Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research,” *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[5] I. Brown and C. Mues, “An experimental comparison of classification algorithms for imbalanced credit scoring data sets,” *European Journal of Operational Research*, vol. 225, no. 2, pp. 363–375, 2012.

[6] F. Louzada, A. Ara, and G. B. Fernandes, “Classification methods applied to credit scoring: A systematic review,” *Expert Systems with Applications*, vol. 33, no. 2, pp. 1–17, 2016.

[7] J. Fritz-Morgenthaler, “Behavioral variables in retail credit risk management,” *Journal of Risk Finance*, vol. 20, no. 1, pp. 20–38, 2019.

[8] J. M. Liberti and M. A. Petersen, “Information: Hard and soft,” *Review of Corporate Finance Studies*, vol. 8, no. 1, pp. 1–41, 2019.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[10] J. Crook, D. Edelman, and L. C. Thomas, “Recent developments in consumer credit risk

Correspondence to: ***Deep Shikha*** *, Faculty of Science****, Suresh Gyan Vihar University, India***
Corresponding author. E-mail addresses: himanshu.vasnani@mygyanvihar.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 357-364

assessment," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1447–1465, 2007.

[11] L. C. Thomas, D. Edelman, and J. Crook, *Credit Scoring and Its Applications*, 2nd ed. Philadelphia, PA, USA: SIAM, 2017.

[12] M. A. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, 2nd ed. Hoboken, NJ, USA: Wiley, 2017.

[13] B. Baesens, R. Bapna, J. Marsden, J. Vanthienen, and J. Zhao, "Transformational issues of big data and analytics in networked business," *MIS Quarterly*, vol. 40, no. 4, pp. 807–818, 2016.

[14] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

[15] P. Kotler and K. L. Keller, *Marketing Management*, 15th ed. Pearson Education, 2016.

[16] L. Lessmann and B. Baesens, "Challenges in segment-based credit scoring," *Decision Support Systems*, vol. 126, pp. 113–125, 2019.

[17] A. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3302–3308, 2009.

[18] S. Brown, "Measuring borrower risk beyond credit scores," *Journal of Financial Services Research*, vol. 55, no. 1, pp. 1–25, 2019.

[19] N. Ghodselahi, "A hybrid support vector machine ensemble model for credit scoring," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13825–13836, 2011.

[20] S. Lessmann, B. Baesens, and H. V. Seow, "Behavioral scoring using machine learning," *Decision Support Systems*, vol. 94, pp. 47–58, 2017.

[21] J. Fritz-Morgenthaler and C. W. Müller, "Income volatility and default risk," *Journal of Banking & Finance*, vol. 96, pp. 210–223, 2018.

[22] A. Brown, "Segment-aware machine learning for retail credit risk," *IEEE Access*, vol. 9, pp. 118430–118444, 2021.

[23] LendingClub Corporation, "LendingClub Loan Data," 2007–2018. [Online]. Available: https://www.lendingclub.com/info/download-data.action

Correspondence to: ***Deep Shikha*** *, Faculty of Science**, Suresh Gyan Vihar University, India***
Corresponding author. E-mail addresses: himanshu.vasnani@mygyanvihar.com