International Journal of Informatics and Data Science Research
ISSN 2997-3961 (Online)
Vol. 2, No. 7, Jul 2025,
Available online at: https://scientificbulletin.com/index.php/IJIDSR

Developing Explainable AI Scorecards for Retail Credit Risk: A Reproducible Framework Using Lending Data

Anvar Zokhidov
PhD Candidate in AI, the University of Texas at Dallas | Former Senior Manager in AI, Tenge Bank (Halyk Group)

Article information:
Manuscript received: 21 May 2025; **Accepted:** 22 Jun 2025; **Published:** 23 Jul 2025

Abstract: This article presents a step-by-step framework for developing an interpretable scorecard for retail credit-risk assessment in banks and financial institutions, and validates the approach on a fully synthetic consumer-lending dataset. The model is an integration of Weight-of-Evidence & Information Value (WoE & IV) binning, machine learning algorithm – Lasso (L1) Regularized Logistic Regression, model-risk supervision, statistical metrics such as Roc-auc & Gini and Confusion Matrix. The testing dataset demonstrates ROC-AUC of 86% and a 72% Gini uplift generalization over a rule-based benchmark. The final result is a ready credit scorecard illustrating the distribution of points assigned to the background features of a customer applying for a loan.

Keywords: credit scoring, scorecard, synthetic data, weight of evidence & information values, regularized logistic regression, explainable AI, machine learning, NPL, AR

1. Introduction and contribution

Banks, buy-now-pay-later (BNPL) organizations and financial institutions rely heavily on credit-scoring models to distinguish between low and high-risk applicants. High-risk is followed by a customer going default on the borrowed loan and low-risk assures a customer will pay-off. Therefore, credit-scoring models are the key component to decrease Non-performance loan (NPL) and increase Approval Rate (AR). While traditional rule-based approaches are still in use, scorecard-style models demonstrated superiority for their transparency and ease of governance that are crucial when trying to report on stakeholders the logic behind the decision-making of the model.

This article delivers two complementary contributions: (i) a reproducible, end-to-end framework for building professional credit scorecards, and (ii) a full analysis on the synthetic dataset that represents real-world consumer-financial features.

The framework was personally developed by the author based on prior experience as a Senior Manager in AI at Tenge Bank (Halyk Group), where similar scorecard models were designed and deployed in live risk-assessment pipelines.

2. Framework: building a transparent scorecard

2.1 Data requirements and preparation

The very first step in approving a customer's request for a loan is choose based on what criteria to score them. It strongly varies from a country to a country and from a region to a region. However, for this

survey features that are common almost everywhere had been chosen. That included: age, official income, credit history, number of registered real estates and number of registered vehicles. These feature data had been generated syntactically by generative AI based on the real behavior of the similar open source data available on the web. The next step was to prepare the data that included putting it into a dataset of rows (number of samples) and columns (number of features) as well as cleaning the dataset: missing-value imputation, outlier treatment, removing duplicates. The target of a value 1 were assigned to the datapoints with the days past due (DPD) over 90 and correspondingly 0 to the ones below 90. The final number of the datapoints were equal to 90000.

2.2 Binning, WOE & IV

The Weight of Evidence (WOE) gives the predictive ability of an independent variable X (input feature) in relation to the dependent variable Y (target). WOE demonstrates if a particular class of an independent variable has a higher distribution of good or bad customers. Good customers are the ones who paid back their loan on time and bad customers have defaulted on it. WOE is calculated using the following formula:

$$\mathrm{WoE}_i = \ln\left(\frac{P_i^{\mathrm{good}}}{P_i^{\mathrm{bad}}}\right)$$

In the survey, first, each not-categorical feature input was divided into bins that created their own class. Using the WOE formula values were calculated for each class. The positive WOE values denoted the distribution of good > distribution of bad and vice versa for the negative WOE values. Next, Information Value were assigned to each class by using the following formula:

$$\mathrm{IV} = \sum_{i=1}^{n}\left(\mathrm{WoE}_i \times \left(P_i^{\mathrm{good}} - P_i^{\mathrm{bad}}\right)\right)$$

The bins had been compressed until they reached a general trend that is if a monotonic trend of WOE values in ascending or descending order were observed. The categorical features were converted into WOE values as well having the number of categories and bins equal. Although, missing values had been dealt with in the previous steps, some might still occur during the next steps. In such cases WOE groups missing values into a class as well. Observing the derived Information Values, it had been made sure that all the features had high values enough to be used for modeling. Common IV range and their corresponding predictive power of the variable is illustrated in the next table:

Information Value (IV)	Predictability of Target
X < 0.02	No impact predictor
0.02 <= X < 0.10	Weak predictor
0.10 <= X < 0.30	Medium predictor
0.30 <= X < 0.50	Strong predictor
>= 0.50	Potential data leakage

Table 1. Information Value impact

Further step included transforming the original dataset into the WOE values obtained.

2.3 Model estimation and regularization

The dataset was split into 80% training and 20% test sets. Also, It had been made sure that the classes in each set were balanced, equally weighted and stratified. Then, an ℓ_2 (Lasso) regularized Logistic Regression, with the equation below, were trained on the training set with the hyper-parameters chosen via five-fold cross-validated ROC-AUC.

$$\hat{\beta} = \arg\min_{\beta} \left[-\frac{1}{n} \sum_{i=1}^{n} (y_i \log p_i + (1 - y_i) \log(1 - p_i)) + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

Having been trained, the testing set showed ROC-AUC of 0.86 and the 0.72 Gini score.

2.4 Score scaling and cut-off selection

The coefficients of the trained model were transformed for further score card building using a Points-to-Double-Odds (PDO) scheme: 20 points per odds doubling, base score 300 at 50:1 odds. This allowed to use these points to fill in the table of the features. Each input feature had its own binning and points assigned. The final score card model is illustrated in the following figure:

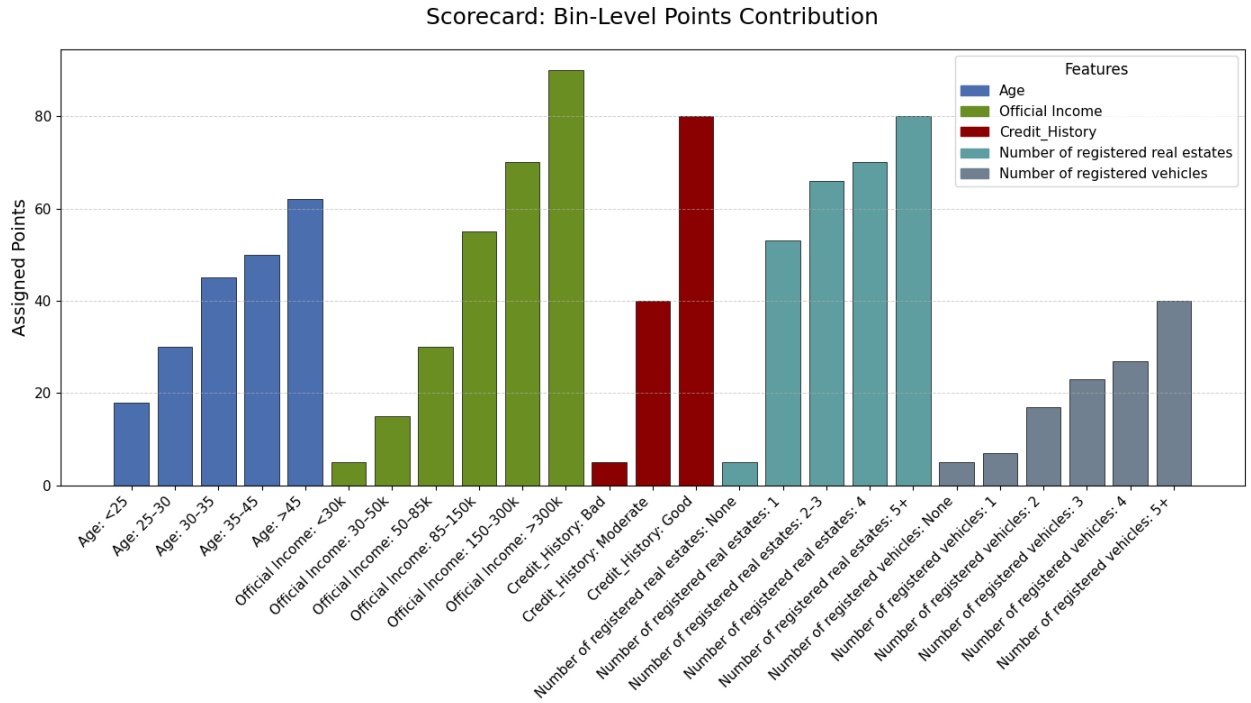

Figure 1. Credit score-card

When a customer comes to a bank for a loan their input features such as age, official income, credit history, number of registered real estates and number of registered vehicles go to the score-card and based on the binning receive their corresponding points. The points then add up resulting in a final score. If the final score exceeds a chosen cut-off (threshold), which is 185 in our case, the loan gets approved. However, in case the final score goes below the cut-off the requested loan gets rejected.

3. Results

The scorecard achieved ROC-AUC 0.86 (Gini 0.72) on the test set, outperforming the common bank benchmark of ROC-AUC 0.54. The cut-off of 185 made the score-card model less strict allowing for more loan approvals. The built credit score-card is a fully ready tool flexible enough in choosing a cut-off with Confusion Matrix that can be deployed by financial institutions and banks.

4. Conclusion

The article provides a reproducible guide for transparent credit-scorecard development with machine learning techniques and WOE & IV. It is valuable for financial institutions and banks to enhance risk management, increase Approval Rate, decrease NPL and it shows stakeholders the clear logic behind a made decision – approve or reject a loan.

Acknowledgments
The author would like to thank Tenge Bank for providing the professional environment that inspired the design of this framework. While the dataset used in this article is fully synthetic, the methods presented reflect practical experience gained during the author's tenure at the bank.

References

1. **Lin, A. Z., & Stokes, M. (2015).**
Entropy-based measures of weight of evidence and information value for variable reduction and segmentation for continuous dependent variables. *SAS Global Forum 2015 Proceedings*, Paper 3242-2015.
https://support.sas.com/resources/papers/proceedings15/3242-2015.pdf

2. **Lin, A. Z. (2014).**
Expanding the use of weight of evidence and information value. *Southeast SAS Users Group (SESUG) 2014 Proceedings*, Paper SD-20.
https://www.lexjansen.com/sesug/2014/SD-20.pdf

3. **Bhalla, D. (2015, March).**
Weight of evidence (WOE) and information value (IV) explained. *Listen Data*.
https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.html

4. **Anik, C. (2020).**
Weight of evidence (WoE) and information value (IV): How to use it in EDA and model building. *Medium*.
https://anikch.medium.com/weight-of-evidence-woe-and-information-value-iv-how-to-use-it-in-eda-and-model-building-3b3b98efe0e8

5. **Analytics Vidhya. (2021, June).**
Understand weight of evidence and information value.
https://www.analyticsvidhya.com/blog/2021/06/understand-weight-of-evidence-and-information-value/

6. **SAS Institute Inc. (n.d.).**
Computing weight of evidence and information value in PROC CASBINNING. *SAS Visual Statistics Documentation*.
https://documentation.sas.com/doc/en/vdmmlcdc/8.1/casstat/viyastat_binning_details02.htm

7. **Jurafsky, D., & Martin, J. H. (2025).**

Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.).
https://web.stanford.edu/~jurafsky/slp3/

8. **Divya, D., Shrivarshini, J., Karan, S., & Kanimozhi, R. (2024).**
Artificial intelligence in finance. *ResearchGate*.
https://www.researchgate.net/publication/380266484_ARTIFICIAL_INTELLIGENCE_IN_FINANCE

9. **Kelly, B. T., & Xiu, D. (2023).**
Financial machine learning (Working Paper No. 2023-100). *Becker Friedman Institute, University of Chicago*.
https://bfi.uchicago.edu/wp-content/uploads/2023/07/BFI_WP_2023-100.pdf

10. **Frontiers of Business Research in China. (2020).**
Toward understanding weight of evidence and information value. *FBR*, 14(1), Article 6.
https://fbr.springeropen.com/articles/10.1186/s11782-020-00082-6

11. **Journal of Financial Innovation. (2024).**

Understanding financial innovation using weight of evidence. *J. Financ. Innov.*, 10, Article 28. https://jfin-swufe.springeropen.com/articles/10.1186/s40854-024-00668-6

12. **van Wieringen, W. N. (2023).**
Lecture notes on ridge regression [arXiv preprint]. arXiv. https://arxiv.org/abs/1509.09169