ASSESSING THE INCREMENTAL VALUE OF ALTERNATIVE DATA IN SME CREDIT SCORING

Xayot Khaitov
Independent Researcher
Management Development Institute of Singapore in Tashkent

Abstract. *This paper proposes a framework for assessing the additional value of alternative data in SME credit scoring. The approach separates the contribution of new information from that of a more complex algorithm. A hypothetical calculation illustrates the comparison of detection rates. The paper concludes that predictive improvement should be evaluated alongside incorrect warnings, borrower coverage, and implementation costs. The proposed framework is methodological and does not report empirical findings.*

Keywords: *SME lending, alternative data, credit scoring, incremental value, machine learning, model validation.*

Introduction

Alternative data can expand the information available for credit assessment. However, their usefulness depends on data quality, relevance, and appropriate application. International guidance identifies inaccurate information, privacy concerns, and fragmented sources among the challenges associated with their use.

The central question is whether additional information improves assessment beyond an existing approach. This paper aims to outline a simple comparison framework and demonstrate its interpretation through a hypothetical calculation.

Research Method

The paper uses comparative analysis and an illustrative calculation. Three model configurations are proposed to distinguish the contribution of additional data from that of a different algorithm. No actual borrower dataset is analyzed; all numerical values are hypothetical.

Research Analysis

The proposed framework begins with a baseline model and introduces changes in stages. This avoids attributing every improvement to alternative data when the algorithm has also changed.

Gambacorta et al. (2019) examine the contributions of machine learning and non-traditional data using information from a Chinese fintech firm. Their research

provides a methodological reference for separating these dimensions, without establishing that the same results apply to all SME portfolios.

The comparison includes:

- Model A: traditional financial and credit information, assessed using a conventional statistical model.
- Model B: the same approach with verified alternative data added.
- Model C: the same information as Model B, assessed using a selected machine-learning model.

Comparing A with B examines the contribution of additional information. Comparing B with C examines the contribution of the algorithm.

For a brief illustration, assume that the same test sample contains 100 SMEs, including 20 that subsequently experience a pre-defined repayment problem. The detection rate is:

$$DR = \frac{\text{Correctly identified problem borrowers}}{\text{All borrowers experiencing the defined problem}} \times 100\%.$$

Indicator	Model A	Model B
Borrowers experiencing repayment problems	20	20
Correctly identified problem borrowers	12	15
Detection rate	60%	75%
Other borrowers incorrectly flagged	8	8

The illustrative increase is $75\% - 60\% = 15$percentage points. Model B identifies three additional problem borrowers without increasing incorrect warnings. This suggests potential additional value within the example, not evidence of actual model performance.

For empirical testing, the models should use the same eligible borrowers, outcome definition, and testing period. Only information available at the assessment date should be included.

Data availability also requires attention. If transaction records exist only for digitally active enterprises, results should not be generalized automatically to all SMEs. Missing or incomplete information must remain visible rather than being treated as evidence of low risk. International guidance warns that inaccurate alternative data can distort creditworthiness assessments.

Finally, any predictive improvement should be considered alongside collection and verification costs. A more complex model is not necessarily more useful if its additional benefit is small or difficult to maintain.

Conclusion

The incremental value of alternative data should be measured against a consistent baseline. Additional information and algorithmic complexity should be evaluated separately.

The hypothetical calculation illustrates how detection rates can support this comparison, but practical evaluation must also consider incorrect warnings, borrower coverage, and implementation costs. Claims of improved credit assessment require testing on real borrower data.

References

1. Gambacorta, L., Huang, Y., Qiu, H., Wang, J. (2019). *How Do Machine Learning and Non-Traditional Data Affect Credit Scoring? New Evidence from a Chinese Fintech Firm.* BIS Working Papers No. 834. Bank for International Settlements.
2. International Committee on Credit Reporting (2018). *Use of Alternative Data to Enhance Credit Reporting to Enable Access to Digital Financial Services by Individuals and SMEs Operating in the Informal Economy.* Guidance Note. World Bank Group.
3. World Bank (2025). *The Use of Alternative Data in Credit Risk Assessment: Opportunities, Risks, and Challenges.*
4. Berger, A. N., Udell, G. F. (2006). A More Complete Conceptual Framework for SME Finance. *Journal of Banking & Finance*, 30(11), pp. 2945–2966. DOI: 10.1016/j.jbankfin.2006.05.008.