Proceedings of ICEMGD 2026 Symposium: Data-Driven Business Management, Corporate Governance, and Organizational Performance
DOI: 10.54254/2754-1169/2026.LD35477

Decoding Machine Learning Performance in Credit Risk Assessment: A Comparative Analysis Based on the Home Credit Default Risk Challenge

Sisi Xia

School of Finance, Shanghai University of International Business and Economics, Shanghai, China
23067033@suibe.edu.cn

Abstract: Accurate credit risk assessment underpins modern financial risk management. Meanwhile, advanced machine learning algorithms have largely replaced traditional linear models in default prediction. However, real-world consumer credit data is scattered across complex relational tables and exhibits extreme class imbalance. As a result, conventional coarse-grained aggregations often fail to capture these nuances, leading to the loss of crucial micro-level default signals. Accordingly, this paper conducts a comparative analysis of the machine learning workflows that performed the best in the Home Credit default risk challenge. In particular, through qualitative deconstruction of pipeline architectures, three representative high-level parallel processing pipelines are analyzed and compared: one based on domain knowledge, one based on heterogeneous stacking framework, and one based on microscopic target aggregation scheme. Furthermore, the key differences among them in terms of feature engineering, handling of imbalanced data, and integration architecture are compared and evaluated. The results indicate that different architectures can effectively extract sparse default signals via their inherent mechanisms, yielding significant gains on imbalanced credit data. It further demonstrate the high effectiveness of feature dimension reconstruction and target dimension reduction in handling extremely imbalanced credit data, providing certain references for industrial credit risk modeling.

Keywords: Credit Risk Assessment, Machine Learning, Feature Engineering, Imbalanced Data, Default Prediction

1. Introduction

Credit risk assessment is the core of modern financial risk management and is directly related to the management of non-performing loans in financial institutions. In recent years, machine learning algorithms such as gradient boosting trees have gradually replaced traditional linear models and have become the mainstream method for default prediction [1]. However, in actual consumer credit business, data is usually scattered across multiple related tables, and the sample of defaults is extremely unbalanced. Traditional modeling methods, when dealing with such complex data, often rely on coarse-grained macro summaries, which easily causes the loss of micro-level default signals. Based on this, in the context of the public housing credit default risk challenge, this paper conducts a

DOI: 10.54254/2754-1169/2026.LD35477

comparative analysis and structural deconstruction study of the top-performing machine learning workflows. In particular, this study explores differences in feature engineering, data processing, and model architecture across high-performing approaches, including domain knowledge-based, deep temporal, and micro-target aggregation methods, in highly imbalanced credit data. It also analyzes how they capture and amplify sparse default signals. By using comparative methods, this study disassembles and analyzes three representative cutting-edge open-source workflows, highlighting their key feature engineering and data processing strategies. It provides practical guidance for building efficient and robust credit risk models in real-world applications.

2. Literature review

2.1. The evolution of credit risk models

The traditional credit scoring models have traditionally relied on logistic regression due to its strict linear interpretability and compliance with regulatory requirements [2, 3]. However, the influx of high-dimensional and non-traditional alternative data quickly exposes the limitations of these rigid linear assumptions [2]. Consequently, both competitive data science and industrial applications have largely shifted to Gradient Boosting Decision Tree (GBDT) architectures, particularly XGBoost and LightGBM [1, 4, 5]. Unlike merely delivering incremental performance gains, GBDT systems have become the core of modern default prediction pipelines because they can effectively handle the highly heterogeneous, sparse, and unnormalized tabular data typical of consumer credit profiles [1]. Its strong nonlinear fitting ability, robustness to missing values, and high computational efficiency make it the undisputed algorithmic foundation for risk modeling challenges, directly supporting the complex cross-tabulation feature engineering and complex integrated topological structure analyzed in the subsequent part of this study [6].

2.2. The trade-off between accuracy and interpretability in pipeline design

In the empirical-based loan assessment benchmark, the modern tree model ensemble method using XGBoost and LightGBM consistently outperforms the traditional logistic regression model in terms of the area under the Receiver Operating Characteristic Curve (AUC) metric [3]. Despite these breakthroughs, the inherent black-box nature of highly complex machine learning algorithms poses serious challenges to transparency, algorithm bias, and regulatory audit compliance [7, 8]. And this fundamental trade-off directly influences the architecture design of the risk prediction pipeline. In a highly competitive data science environment, prioritizing pure accuracy typically leads to highly opaque, multi-layered neural networks. However, the strict compliance requirements in real-world financial operations force practitioners to re-examine these structures, typically favoring single feature-centered models or feature engineering driven by business logic to maintain necessary transparency. Therefore, the strategies employed by top competitive workflows to address this limitation, either through the reduction of combination complexity or the integration of explainable artificial intelligence (XAI) frameworks, are crucial for translating theoretical prediction limits into practical credit risk models that meet institutional compliance requirements [9, 10].

2.3. Competition-driven machine learning

Large-scale data science competitions serve as practical benchmarks for credit default prediction. However, top-ranked solutions rarely rely on straightforward applications of standard algorithms;

DOI: 10.54254/2754-1169/2026.LD35477

instead, they exhibit highly complex and heterogeneous pipelines [1, 4]. Furthermore, even within a single competition, leading teams frequently implement substantially different approaches to feature engineering, data integration, class imbalance handling, and ensemble construction [1, 5]. As a result, leaderboard rankings alone cannot reveal which design choices are responsible for performance gains or whether they generalize beyond the competition setting. Despite the growing number of high-scoring solutions, systematic comparisons of their pipeline designs remain limited, leaving the relative effectiveness and transferability of different design choices unclear [6]. To address this gap, this study examines and contrasts selected high-ranking pipelines, focusing on their strategies for cross-table feature engineering and extreme class imbalance management. By examining the design choices underlying these pipelines, the study identifies techniques that consistently contribute to predictive performance and evaluates their applicability to real-world credit risk modeling [1, 6].

3. Methodology and comparative analysis framework

3.1. Selection of representative open-source pipelines

To understand the key technical determinants of top credit risk modeling techniques under similar production environment constraints, this study analyzes the pipeline designs of submissions to the Home Credit Default Risk Challenge. Notably, the dataset presents significant complexity, with approximately 310,000 training records and 50,000 test records spread across multiple relational database tables. It also faces severe challenges such as high-dimensional features and extreme class imbalance. In this context, it is crucial to systematically compare these leading solutions, as this not only helps in identifying transferable feature engineering strategies and robust data processing procedures, but also enhances the effectiveness of real-world risk management, rather than merely relying on the crude improvement achieved through algorithm integration.

In selecting pipelines for evaluation, emphasis was placed on top-ranked, well-documented, and methodologically distinct solutions. Accordingly, three representative top-performing open-source pipelines were selected for detailed examination:

The first pipeline, a Domain-Knowledge and Heterogeneous Stacking Paradigm (Rank #3), was developed by alijs and Evgeny. It benchmarks the integration of practical risk management into ML workflows. Its architecture emphasizes business-oriented data segmentation, manually engineered scorecard features, and a large decorrelated ensemble framework [11].

The second pipeline, a Deep Temporal Interaction and Nested Behavioral Paradigm (Rank #5), was developed by the Kraków-Lublin-Zabinka team. It represents a more advanced computational workflow, combining automated cross-source temporal feature extraction based on deep sequential models with nested modeling techniques designed to capture fine-grained risk dynamics [12].

The third pipeline, a Micro-Level Target Aggregation Paradigm (Rank #17), was developed by qchemdogdog and represents the highest-ranked individual gold-medal solution. Despite its simple architecture, the pipeline maps macro-level objectives to historical transactions via a bottom-up framework, achieving strong performance without deep learning or large ensembles [13].

3.2. Structural decomposition and analytical dimensions

This study uses a comparative deconstruction approach to examine the three selected open-source pipelines. Rather than focusing solely on predictive performance, the analysis investigates the

DOI: 10.54254/2754-1169/2026.LD35477

underlying design choices that define each pipeline. To ensure a consistent basis for comparison, a unified three-dimensional analytical framework is established.

The first dimension concerns feature engineering, focusing on how each pipeline transforms multi-table relational data into predictive features and incorporates domain-specific financial information. The second dimension examines class imbalance handling, analyzing how rare default signals are identified and amplified without introducing substantial sampling distortion. The third dimension evaluates model architecture and ensemble design, assessing the trade-offs between complex multi-layer stacking frameworks and more streamlined modeling approaches. Using this framework, the three selected pipelines are compared across the same set of analytical dimensions, enabling a structured evaluation of their different design philosophies and technical solutions.

3.3. Performance evidence and evaluation criteria

The performance evidence used is derived from the officially verified results reported on the public leaderboard of the Home Credit Default Risk Challenge. As all pipelines were evaluated under the same competition protocol, their scores provide a common basis for comparison. The competition primarily uses the Area Under the ROC Curve (ROC-AUC) as its evaluation metric. This metric is particularly suitable for credit default prediction because default events are relatively rare, making conventional accuracy measures sensitive to class imbalance and potentially misleading. ROC-AUC evaluates a model's ability to rank defaulted loans above non-defaulted loans across all possible classification thresholds, yielding a threshold-free measure of classification power. In this study, ROC-AUC scores from the leaderboard are treated as comparative evidence rather than as results from a controlled experimental setup. And they serve as a reference point for evaluating pipeline performance and relating architectural decisions to predictive outcomes.

4. Comparative analysis of high-scoring models

4.1. Feature engineering paradigms

Feature engineering represents the primary point of divergence among the three selected pipelines. Faced with the same multi-table relational dataset, each workflow adopts a fundamentally different strategy for transforming raw financial records into predictive representations, as shown in Table 1.

Table 1. Comparison of feature engineering strategies

Model Pipeline	Core Feature Engineering Paradigm	Target/Aggregation Method
Rank #3	Explicit business logic and tabular data reconstruction; integrates algebraic residuals of predicted external risk scores	Maintains unaggregated transactional tables; trains distinct sub-models on independent blocks
Rank #5	Automated neural sequence extraction; creates a 96-month multi-channel user history matrix	Uses 1D-Conv for monthly velocity and BiLSTM for long-term trends; infers latent loan terms via formulas
Rank #17	Micro-level historical record deconstruction; avoids macro-level global smoothing	Maps binary target directly to each historical entry; aggregates micro-predictions back to the user level

The Rank #3 pipeline adopts a domain-driven method, preserving transaction-level data through minimal aggregation. Following this strategy, distinct models are applied to separate data sources, such as baseline applications, historical bureau records, and installment tracks. It further strengthens

DOI: 10.54254/2754-1169/2026.LD35477

feature representation through an auxiliary model that reconstructs missing external risk indicators and integrates the resulting estimates back into the main workflow as additional predictive signals [11]. By contrast, the Rank #5 pipeline is fundamentally a conventional tabular model based on LightGBM with over 3,000 hand-crafted features [12]. As a supplementary module, it constructs a 96-month temporal matrix from customer histories across multiple data sources and processes it through a 1D-Convolution-BiLSTM network. The OOF predictions from this neural network are fed as a single additional feature into the main LightGBM model, yielding a marginal improvement of 0.001 in CV AUC on top of an already strong baseline. Besides, financial-domain knowledge is further incorporated by deriving latent contract characteristics, such as interest rates and repayment horizons, from historical loan records. Specifically, the implicit interest rate (ir) is approximated using a loan amortization proxy:

$$\frac{\text{Annuity} \times \text{CNT\_Payments}}{\text{Credit\_Amount}} = (1+ir)^{\text{CNT\_Payments}/12} \tag{1}$$

This enables the model to map historical loan contract terms onto current applications.

Alternatively, the Rank #17 pipeline focuses on micro-level target aggregation. By avoiding early aggregation, it assigns the default target to each historical transaction and predicts at the micro level. The resulting predictions are aggregated to the customer level, preserving rare distress signals that could be lost through conventional aggregation [13].

The three pipelines illustrate distinct feature engineering philosophies: business-driven feature reconstruction (Rank #3), neural temporal representation learning (Rank #5), and micro-level target aggregation (Rank #17). And these approaches demonstrate alternative solutions to the challenge of extracting predictive information from large-scale, multi-table credit datasets.

4.2. Imbalance mitigation and signal amplification

Due to the extreme scarcity of default instances in real-world lending portfolios, models face severe minority class suppression. Thus, the selected pipelines completely reject conventional sampling methods, such as Synthetic Minority Over-sampling Technique (SMOTE) or random undersampling, as they tend to distort data distributions and introduce noise. Instead, the frameworks adopted distinct strategies: heterogeneous ensembling for Rank #3, and variations of target propagation with OOF transformations for Rank #5 and Rank #17, as detailed in Table 2.

Table 2. Comparison of imbalance handling techniques

Model Pipeline	Core Imbalance Strategy	Execution Mechanism
Rank #3	Ensemble-focused: Counteracting minority class prediction bias and high model variance	Parallelizes over 50 experiments; blends the seven most decorrelated models to offset errors
Rank #5	Representation shift-focused: Isolating sparse behavioral anomalies from macro-level noise	Uses nested modeling protocols on historical installments; calculates continuous risk weights for micro-actions
Rank #17	Label transformation-focused: Transforming label density to improve gradient descent efficiency	Converts sparse 0/1 hard labels into dense, continuous empirical probabilities via row-level OOF predictions

The Rank #3 approach mitigates minority class variance via heavy heterogeneous ensembling. Recognizing the instability induced by imbalanced objectives, the pipeline executed over 50 parallel experiments and selected the seven most decorrelated models via correlation analysis across the

DOI: 10.54254/2754-1169/2026.LD35477

prediction matrix. By blending models with diverse error patterns from different preprocessing and feature sets, the ensemble mitigates minority class bias [11]. By contrast, the Rank #5 pipeline employs a "nested model" approach to extract behavioral signals from auxiliary tables. It merges the binary target onto historical records (e.g., installment payments) and trains subordinate LightGBM models to predict default probability at the transaction level. The resulting OOF predictions are aggregated (min, max, mean, median) back to the customer level as additional features. Despite low sub-model AUCs (0.54-0.63) due to high noise, this technique captures micro-level behavioral patterns that macro-aggregations might miss [12]. Alternatively, the Rank #17 pipeline densifies labels. To address the inefficiency caused by sparse binary targets under heavy class imbalance, it maps targets to individual historical entries and generates micro-level predictions, converting them into dense, continuous empirical probabilities. These micro-level predictions are then aggregated to the user level via maximum and summation operators, producing dense risk factors and simplifying optimization for the master model, while preserving minority class signals [13]. The three pipelines differ fundamentally in imbalance handling: ensemble-based (Rank #3), nested anomaly isolation (Rank #5), and label transformation-based (Rank #17).

4.3. Algorithmic execution and ensemble topology

LightGBM underlies all three pipelines, yet their ensemble structures and meta-learning approaches differ, further highlighting trade-offs between model complexity and feature expressiveness. Table 3 outlines the stark contrast in their ensembling methodologies.

Table 3. Comparison of ensemble topologies

Model Pipeline	Structural Topology	Algorithmic Execution Details
Rank #3	Stacking-heavy: A highly sophisticated, four-tiered meta-stacking network	Integrates 15 base models, runs automated forward feature selection, and blends four diverse secondary algorithms
Rank #5	Feature-centric: Minimizes stacking layers; prioritizes deep sequential abstractions	Relies purely on a single LightGBM model powered by over 3,000 variables selected via strict Null Importance pruning
Rank #17	Embedding stacking hybrid: Shifts the stacking process into the feature phase	Uses OOF predictions from historical sub-models directly as structural features; executes a clean single master model

The Rank #3 architecture adopts a deep stacking-heavy design. And it builds on 15 base models, including LightGBM, XGBoost, and feed-forward neural networks. Their outputs pass through a forward feature selection layer that integrates raw features and first-stage predictions, followed by a secondary stack of diverse learners, whose results are combined via weighted averaging [11]. In contrast, the Rank #5 pipeline follows a feature-centric paradigm, relying on a single LightGBM model trained on approximately 3,000 features selected via Null Importance pruning. The team tested a minor ensemble (0.25:0.25:0.50) but observed no improvement over the best single model, confirming that performance is driven primarily by feature engineering rather than ensemble complexity [12]. By comparison, the Rank #17 pipeline integrates ensemble information directly into feature space. Instead of multi-layer stacking, it uses out-of-fold predictions from historical sub-models as features for a single final model. This design preserves the benefits of stacking while eliminating additional meta-model layers, resulting in a streamlined and efficient training pipeline suitable for deployment [13]. The three pipelines represent fundamentally different ensemble design

DOI: 10.54254/2754-1169/2026.LD35477

philosophies: stacking-heavy architecture (Rank #3), feature-dominant modeling (Rank #5), and embedding-based hybrid stacking (Rank #17).

5. Conclusion

This paper presents a structural deconstruction and empirical comparison of machine learning-based credit risk assessment pipelines in complex multi-table scenarios. In particular, evaluation of three top-ranked open-source solutions from the Home Credit Challenge indicates that each architecture employs distinct mechanisms to extract sparse default signals. The domain-knowledge approach reduces variance through unaggregated target mapping and heterogeneous ensembling, whereas the feature-centric approach captures cross-source dependencies through 1D-Convolution and BiLSTM while mitigating noise through nested modeling, and the micro-level target aggregation approach transforms sparse binary targets into dense empirical probabilities, reducing optimization complexity for downstream modeling.

These results highlight the effectiveness of feature space restructuring and target propagation to granular records for highly imbalanced credit data, providing practical guidance for credit risk modeling. However, the findings are derived from publicly documented pipeline architectures and competition-reported performance metrics. Future work will extend evaluation to real-world credit datasets across diverse business environments, develop standardized benchmarking frameworks, incorporate lightweight automated machine learning for feature selection, and integrate explainable AI techniques to improve interpretability and compliance in lending systems.

References

[1] Tang, Y. (2021). Research on loan default prediction models based on XGBoost and LightGBM algorithms. *Modern Computer, 27*(32), 33-37.

[2] Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767-2787.

[3] Zhang, L., & Wu, Z. (2023). Comparison of credit scoring card based on XGBoost machine learning model and logistic regression model. *Journal of South-Central Minzu University (Natural Science Edition), 42*(6), 72-78.

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[5] Zheng, Y. (2022). A default prediction method using XGBoost and LightGBM. In *2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* (pp. 1-6).

[6] Shi, S., Tse, R., Luo, W., D'Addona, S., & Pau, G. (2022). Machine learning-driven credit risk: A systematic review. *Neural Computing and Applications, 34*(17), 14327–14339.

[7] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. *Artificial Intelligence Review, 57*, 216.

[8] Mujo, A., Nikolla, S., Hoxha, E., & Pelivani, E. (2025). Explainable AI in credit scoring: Improving transparency in loan decisions. *Journal of Information Systems Engineering and Management, 10*, e-ISSN 2468-4376.

[9] Bhandari, R., Ramya, S. P., & Ebenezar, U. S. (2026). Explainability-guided credit risk prediction using machine learning model. In *2026 International Conference on Innovative Trends in Information Technology (ICITIIT)* (pp. 1–6).

[10] Chang, V., Xu, Q. A., Akinloye, S. H., Benson, V., & Hall, K. (2024). Prediction of bank credit worthiness through credit risk analysis: An explainable machine learning study. *Annals of Operations Research*.

[11] Alijs, & Patekha, E. (2018). 3rd place solution: Home credit default risk [Competition writeup]. *Kaggle*. https: //www.kaggle.com/c/home-credit-default-risk/discussion/64596

[12] Narsil, Meyk, & Alvor. (2018). Overview of the 5th solution: Home credit default risk [Competition writeup]. *Kaggle*. https: //www.kaggle.com/c/home-credit-default-risk/discussion/64625

[13] Ge, Q. H. (2018). 17th place mini writeup: Home credit default risk [Competition writeup]. *Kaggle*. https: //www.kaggle.com/c/home-credit-default-risk/discussion/64503