Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

Operationalizing Imbalance: Boundary vs. Density-Based Sampling in Extreme and Moderate Credit Risk Scenarios

Sai Prashanth Pathi

Independent Researcher
prashanthp.sai@gmail.com

Abstract:
Machine learning models in credit risk are frequently compromised by the class imbalance problem, where fraud cases or defaults represent a negligible fraction of the population. While the Synthetic Minority Over-sampling Technique (SMOTE) is the de facto standard for addressing this, recent literature suggests it may introduce noise and computational overhead without operational gain. This study conducts a rigorous comparative analysis of six strategies: Cost-Sensitive Learning (Baseline), Random Undersampling (RUS), Vanilla SMOTE, Borderline-SMOTE, ADASYN, and SMOTE-ENN applied to Gradient Boosted Decision Trees (XGBoost). The evaluation utilizes three datasets with varying imbalance ratios (from 0.17% to 6.9%) to test robustness across different financial contexts. Our experiments reveal two critical insights. First, we identify a "False Positive Trap" in extreme imbalance scenarios: while Vanilla SMOTE achieved the highest Area Under Precision-Recall Curve (AUPRC: 0.825), it degraded the F1-score to 0.282, rendering it operationally inviable. In contrast, Borderline-SMOTE maintained a comparable AUPRC (0.818) while achieving a superior F1-score (0.690). Second, we detect a "Threshold of Necessity": in scenarios with moderate imbalance (>5%), all sampling techniques failed to outperform the cost-sensitive baseline, suggesting that synthetic sampling is counter-productive when sufficient minority examples exist.

Keywords: Credit Risk, Class Imbalance, Fraud Detection, XGBoost, SMOTE, Anomaly Detection.

I. INTRODUCTION

The rapid digitization of financial services has necessitated the development of automated, high-velocity credit risk assessment models. Whether for real-time fraud detection or credit underwriting, these models must distinguish between the majority class (legitimate transactions/accounts) and the minority class (fraud/defaulters). However, real-world financial data is inherently imbalanced. In fraud detection, the positive class often constitutes less than 0.1% of the dataset [1].

Traditional classifiers, such as Logistic Regression and Support Vector Machines, aim to maximize overall accuracy, leading to a bias toward the majority class. In an extreme imbalance scenario, a trivial model predicting "legitimate" for every transaction achieves 99.9% accuracy but fails its primary objective: risk mitigation. To counter this, practitioners employ resampling techniques, most notably the Synthetic Minority Over-sampling Technique (SMOTE) [2].

While SMOTE is widely adopted, its "blind" generation of synthetic samples along the line segments joining k-nearest neighbors can introduce noise, particularly in high-dimensional financial data where class overlap is significant. Advanced variations such as Borderline-SMOTE [3] and ADASYN [4] attempt to resolve this by focusing on hard-to-classify examples.

Despite the proliferation of these techniques, there is a paucity of comparative studies that evaluate them through an operational lens. Most academic benchmarks rely heavily on ROC-AUC, a metric known to be overly optimistic in imbalanced domains [5], effectively masking the operational cost of False Positives. Furthermore, few studies differentiate between "Extreme Imbalance" (Fraud) and "Moderate Imbalance" (Default), assuming a one-size-fits-all efficacy for sampling.

This paper addresses these gaps by:

1. Benchmarking six handling strategies using XGBoost, the industry standard for tabular data.
2. Comparing efficacy across differing degrees of imbalance (0.17% vs. 6.9%).
3. Evaluating performance using AUPRC and F1-score to quantify the trade-off between risk capture and operational precision.

II. LITERATURE REVIEW

A. The Class Imbalance Problem

Class imbalance causes standard learning algorithms to underfit the minority class. Approaches to mitigate this fall into two categories: Data-level methods (resampling) and Algorithm-level methods (cost-sensitive learning).

B. Sampling Techniques

Random Undersampling (RUS) balances the distribution by randomly discarding majority samples. While computationally efficient, Liu et al. [6] note that it risks eliminating informative data, potentially degrading model generalization.

SMOTE, proposed by Chawla et al. [2], generates synthetic minority samples to prevent overfitting. However, it treats all minority samples equally, regardless of their proximity to the decision boundary.

To address this, Han et al. [3] proposed **Borderline-SMOTE**, which only oversamples minority instances near the decision boundary (the "Danger" zone), theorizing that these are the most critical for classification. Conversely, **ADASYN** (Adaptive Synthetic Sampling) [4] uses a weighted distribution to generate more samples for minority instances that are harder to learn, based on local density.

Hybrid Methods like **SMOTE-ENN** combine oversampling with undersampling (Edited Nearest Neighbours) to clean the decision boundary of overlapping noise [7].

C. Gradient Boosting

XGBoost [8] has emerged as the dominant algorithm for structured financial data due to its handling of missing values and regularization. While XGBoost includes a scale_pos_weight parameter for cost-sensitive learning, theoretical foundations suggest that reweighting examples can achieve optimal decision boundaries without the noise introduction inherent in resampling [9]. It remains empirically unclear, however, at what threshold of imbalance this algorithmic adjustment supersedes data-level resampling.

III. METHODOLOGY

A. Datasets

We utilize three datasets to simulate varying degrees of financial risk.

1. **Kaggle Credit Card Fraud (Extreme Imbalance):** Contains 284,807 transactions made by European cardholders in 2013. Only 492 are fraud (0.172%). Features are PCA-transformed (V1-V28) for privacy, plus Time and Amount.
2. **Give Me Some Credit (Moderate Imbalance):** Contains 150,000 borrowers. The target is SeriousDlqin2yrs (90-day delinquency). The default rate is approximately 6.9%, significantly higher than the fraud dataset.
3. **Synthetic Data:** A controlled dataset generated using sklearn.make_classification with 20,000 samples and 1% imbalance, used to validate theoretical boundary behavior.

B. Experimental Setup

We employ XGBoost as the base classifier. To prevent data leakage, all sampling techniques are applied strictly within a cross-validation loop (Stratified 5-Fold). The sampling is applied only to the training fold; validation folds remain untouched and imbalanced to reflect real-world distribution.

C. Algorithms Evaluated

1. **Baseline (Weighted):** XGBoost with scale_pos_weight = $\frac{Negatives}{Positives}$. No resampling.
2. **Random Undersampling (RUS):** Randomly removes majority samples.
3. **Vanilla SMOTE:** Standard implementation (k=5).
4. **Borderline-SMOTE:** 'Borderline-1' variant.
5. **ADASYN:** Density-based adaptive sampling.
6. **SMOTE-ENN:** Sequential SMOTE followed by ENN cleaning.

D. Evaluation Metrics

We prioritize AUPRC (Area Under Precision-Recall Curve) over ROC-AUC. Additionally, we report the F1-Score, which is the harmonic mean of Precision and Recall, to measure the stability of predictions.

IV. RESULTS

The experimental results, aggregated over 5-fold cross-validation, are presented below. "Std Dev" denotes the standard deviation across folds, indicating stability.

A. Performance on Extreme Imbalance (Credit Card Fraud - 0.17%)

Fig. 1 illustrates the Precision-Recall curves for all six methods on the Credit Card Fraud dataset. Both Vanilla SMOTE (Green) and Borderline-SMOTE (Red) exhibit a sharp decline in precision as recall approaches 0.85, a characteristic typical of classifiers on highly imbalanced data. While Vanilla SMOTE maintains a marginally higher precision across the curve (reflected in its higher AUPRC of 0.825), this global metric masks the specific operational instability revealed by the F1-score analysis.

TABLE I. PERFORMANCE ON EXTREME IMBALANCE (CREDIT CARD FRAUD - 0.17%)

Method	Mean AUPRC	Std Dev	Mean F1-Score	Training Time (s)
Vanilla SMOTE	**0.825**	0.033	0.282	5.59
Borderline-SMOTE	0.818	0.043	**0.690**	4.78
SMOTE-ENN	0.814	0.022	0.275	360.28
Baseline (Weighted)	0.811	0.030	0.520	2.11
ADASYN	0.800	0.031	0.147	5.77
Random UnderSampling	0.704	0.041	0.084	**0.69**

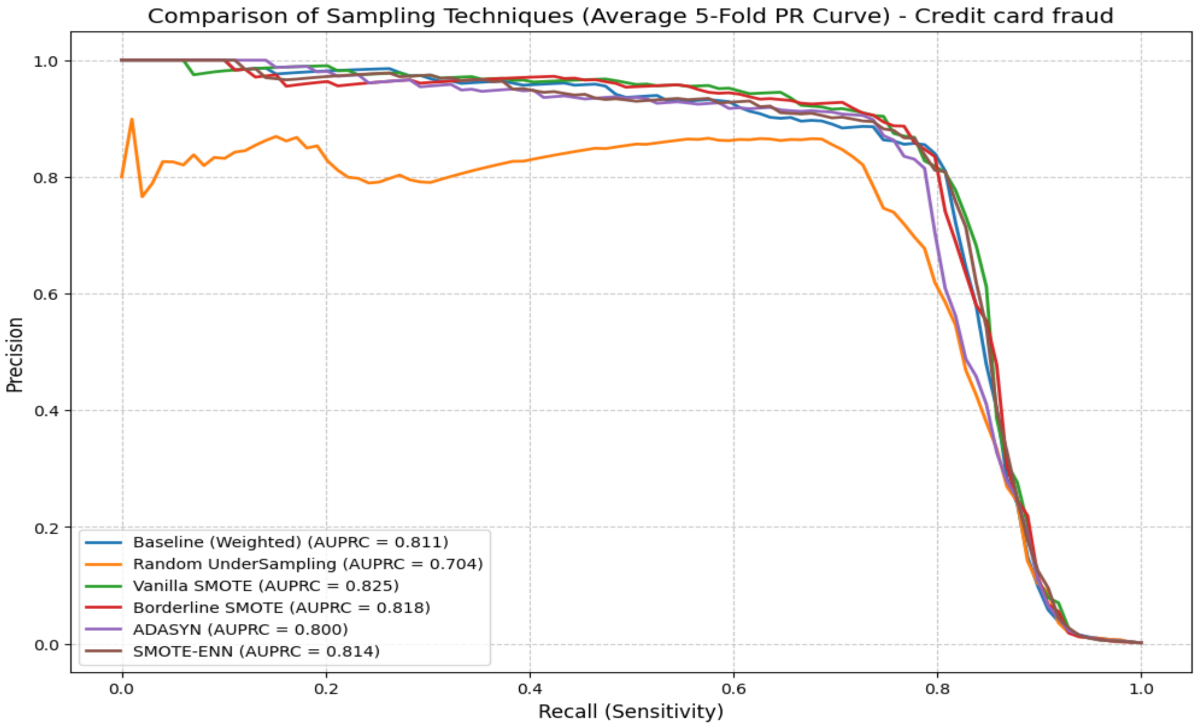

Fig. 1. Comparison of Precision-Recall Curves on the Credit Card Fraud dataset (Extreme Imbalance). Vanilla SMOTE (Green) and Borderline-SMOTE (Red) show similar global performance characteristics with a sharp precision drop-off at high recall.

B. Performance on Moderate Imbalance (Give Me Some Credit - 6.9%)

In the moderate imbalance scenario (Table II), the Cost-Sensitive Baseline achieved the highest AUPRC (0.392), outperforming all synthetic sampling techniques. Notably, Random UnderSampling (RUS) provided a competitive approximation (0.382) at a fraction of the computational cost (0.50s), whereas complex methods like SMOTE-ENN yielded diminishing returns (0.334) despite significantly higher training times.

TABLE II. PERFORMANCE ON MODERATE IMBALANCE (GIVE ME CREDIT - 6.9%)

Method	Mean AUPRC	Std Dev	Mean F1-Score	Training Time (s)
Baseline (Weighted)	**0.392**	0.012	0.337	1.84
Random UnderSampling	0.382	0.011	0.330	**0.50**
SMOTE-ENN	0.334	0.015	0.343	37.82
Vanilla SMOTE	0.309	0.012	0.360	1.70
Borderline-SMOTE	0.308	0.012	**0.364**	3.66
ADASYN	0.303	0.012	0.357	3.73

C. Performance on Synthetic Control (Theoretical Imbalance - 1.0%)

Experiments on the synthetic control dataset reinforced the superiority of boundary-aware methods in clean feature spaces. Borderline-SMOTE achieved the highest AUPRC (0.202) and F1-score (0.189), significantly outperforming Vanilla SMOTE (AUPRC 0.107) and ADASYN (AUPRC 0.083), confirming that density-based sampling struggles even in controlled environments with lower noise.

TABLE III. PERFORMANCE ON SYNTHETIC DATASET (IMBALANCE - 1%)

Method	Mean AUPRC	Std Dev	Mean F1-Score	Training Time (s)
Borderline-SMOTE	**0.202**	0.080	**0.189**	0.911
Baseline (Weighted)	0.184	0.064	0.172	4.723
Random UnderSampling	0.111	0.018	0.064	**0.520**
Vanilla SMOTE	0.107	0.040	0.104	0.825
SMOTE-ENN	0.086	0.030	0.093	2.957
ADASYN	0.083	0.030	0.094	0.875

V. DISCUSSION

A. The False Positive Trap in Extreme Imbalance

A superficial analysis of Table I suggests Vanilla SMOTE is the superior method based on AUPRC (0.825). However, a deeper inspection of the F1-Score reveals a critical operational flaw.

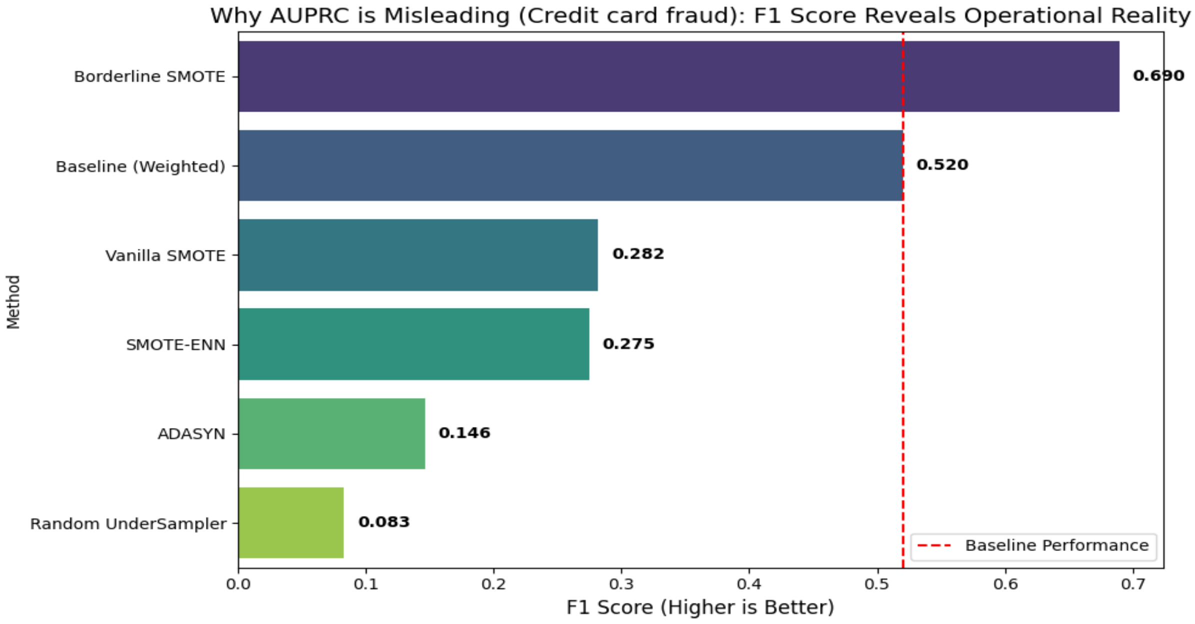

Fig. 2. The "False Positive Trap": Comparing F1 Scores reveals that Borderline-SMOTE is the only operationally viable sampling method, outperforming the Baseline. Vanilla SMOTE and ADASYN degrade performance despite high AUPRC scores.

As visualized in Fig. 2, there is a massive disparity between the methods. Vanilla SMOTE achieved an F1-Score of only 0.282, while Borderline-SMOTE achieved 0.690. This discrepancy indicates that Vanilla SMOTE effectively increases Recall (catching frauds) but at the expense of generating excessive False Positives. In a production banking environment, this "trap" renders the model operationally expensive due to the high volume of manual reviews required for false alarms.

Borderline-SMOTE emerged as the operational superior. By generating samples only near the decision boundary, it maintained high AUPRC (0.818) while doubling the F1-Score (0.690) compared to Vanilla

SMOTE. This validates the hypothesis that in extreme imbalance, the definition of the boundary is more critical than the volume of the minority class.

B. The "Threshold of Necessity"

In the "Give Me Some Credit" dataset, where the default rate is moderate (~6.9%), we observe a reversal in performance. As shown in Fig. 3, the Baseline (Weighted) curve consistently dominates the sampling techniques across the entire recall range.

This supports our hypothesis of a "Threshold of Necessity." When the minority class is moderately represented (>5%), synthetic sampling introduces noise that confuses the gradient boosting algorithm. The model learns more effectively from the raw, weighted data than from synthetically altered distributions. Notably, Random Undersampling (Orange line in Fig. 3) performed competitively while being the fastest method.

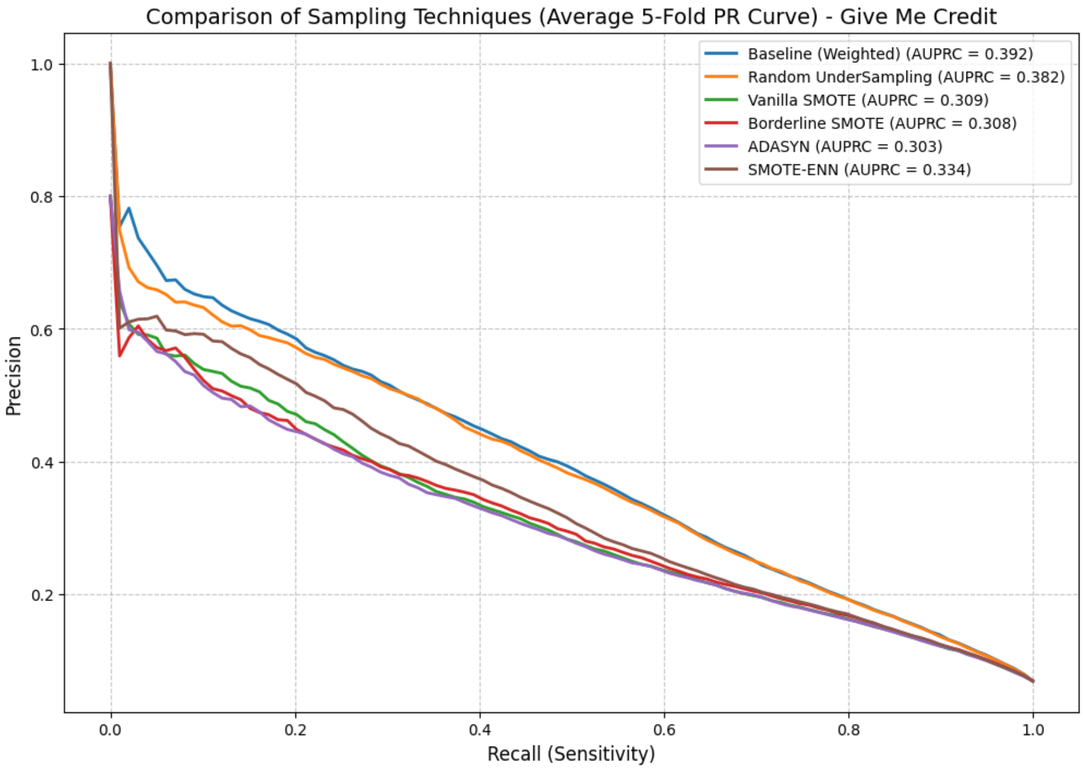

Fig. 3. Performance on Moderate Imbalance (6.9%). The Baseline (Blue) and Random UnderSampling (Orange) outperform complex oversampling methods, suggesting a "Threshold of Necessity" for synthetic data generation.

C. The Failure of ADASYN

ADASYN consistently underperformed, particularly in the Fraud dataset (Fig 2, Light Green Bar). Its mechanism focuses on "hard-to-learn" examples, which, in financial transaction data, are often outliers or noise. By adaptively oversampling these outliers, ADASYN likely amplified the noise in the dataset, leading to model degradation. This aligns with findings by García et al. [10], who observed that density-based oversampling is highly sensitive to class overlap and can deteriorate classifier performance in noisy feature spaces.

D. Computational Efficiency

SMOTE-ENN proved computationally expensive (360s vs 4.7s for Borderline-SMOTE) without yielding performance gains. Given the need for frequent retraining in fraud detection systems, the 75x latency increase makes SMOTE-ENN difficult to justify in production environments compared to Borderline-SMOTE.

VI. CONCLUSION

This study challenges the "one-size-fits-all" approach to handling class imbalance in credit risk. By benchmarking boundary-based versus density-based techniques across varying imbalance regimes, we demonstrate that operational viability often diverges from theoretical performance.

Our findings reveal a "False Positive Trap" in extreme fraud scenarios (0.17% imbalance). While Vanilla SMOTE inflated the AUPRC to 0.825, its low F1-score (0.282) indicated a prohibitive rate of false alarms. In contrast, Borderline-SMOTE proved superior, focusing synthetic generation on the decision boundary to achieve a 2.4x improvement in F1-score (0.690), thereby balancing risk capture with operational precision.

Additionally, we identify a "Threshold of Necessity" at approximately 5% minority class prevalence. In moderate imbalance scenarios, complex sampling failed to outperform the Cost-Sensitive Baseline, suggesting that modern algorithms like XGBoost handle moderate skew more effectively through native reweighting than through synthetic data manipulation. Furthermore, the consistent underperformance of ADASYN confirms that density-based oversampling is ill-suited for noisy financial data, as it tends to amplify outliers.

Implications for Practice: We advocate for a regime-specific strategy: adopt Borderline-SMOTE for high-velocity, low-prevalence fraud detection to minimize customer friction, and adhere to Cost-Sensitive Learning for regulatory credit scoring where data integrity is paramount. Future work will extend this evaluation to model interpretability, examining the impact of synthetic boundaries on SHAP value stability.

REFERENCES:

[1] Dal Pozzolo, A., et al. (2014). "Learned lessons in credit card fraud detection from a practitioner perspective," *Expert Systems with Applications*, 41(10), 4915-4928.

[2] Chawla, N. V., et al. (2002). "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 16, 321-357.

[3] Han, H., Wang, W. Y., & Mao, B. H. (2005). "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," *International Conference on Intelligent Computing*.

[4] He, H., et al. (2008). "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *IEEE International Joint Conference on Neural Networks*.

[5] Saito, T., & Rehmsmeier, M. (2015). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, 10(3).

[6] Liu, X. Y., Wu, J., & Zhou, Z. H. (2008). "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics*.

[7] Batista, G. E., Prati, R. C., & Monard, M. C. (2004). "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*.

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

[9] C. Elkan, "The foundations of cost-sensitive learning," *International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

[10] V. García, J. S. Sánchez, and R. A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, 2012.