Cureus Journal of Computer Science
https://doi.org/10.7759/s44389-026-00104-z

Original Article

A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection

Janaki Raman Palaniappan [1]

1. *Software Engineer/Researcher, Brunswick Corporation, Sunrise, USA*

Received: April 16, 2026 | Review began: May 05, 2026 | Review ended: May 22, 2026 | Published: June 05, 2026

Abstract

The rise of digital payments has led to increased credit card fraud, causing significant financial losses. Due to the rarity of fraudulent transactions, datasets are highly imbalanced, making detection a major challenge for machine learning models, which often struggle to accurately detect these anomalies. Continuing research aims to improve the models. Therefore, selecting the right classifiers and sampling techniques is vitally important to build an effective model.

Typically, the original dataset is split into training and test sets. Sampling techniques are applied only to the training set to prevent data leakage and preserve the integrity of the test set. This study presents a comparative analysis of model's performance using both the original and resampled datasets split, evaluated across multiple classifiers and sampling strategies. The study utilizes a publicly available credit card fraud detection dataset.

While models trained and tested on resampled dataset show higher accuracy, this improvement often diminishes when evaluated against the original test set. Resampling techniques, though useful, come with limitations. Findings emphasize the importance of maintaining an untouched test set from the original dataset to ensure reliable model evaluation and produce trustworthy performance metrics. This study highlights that using an unaltered test set is essential for achieving realistic and generalizable model performance.

Categories: Machine Learning (ML), Security and Privacy

Keywords: credit card fraud, sampling techniques, original dataset, resampled dataset, artificial intelligence, random forest classifier, xgboost classifier, decision tree classifier, adaboost classifier

Introduction

Credit cards play an important role in society, and their use has increased significantly due to their advantages, such as the 'spend now, pay later' concept, cashback, cash withdrawals, etc. These benefits attract customers to apply for a credit card. Cardholders get one billing cycle to repay the amount. The credit limit of a credit card depends on the customer's usage history and credit score.

Credit card fraud is a criminal offense. This fraud is an easy target for fraudsters to withdraw significant amounts of money without the cardholder's knowledge. Fraudsters often attempt to make transactions appear legitimate, making it difficult for card issuers to detect them quickly. One of the easiest targets for fraudsters is Card-Not-Present (CNP) fraud. CNP is a credit card transaction where the payment card is not physically present during the transaction. Examples of CNP transactions are e-commerce purchases, internet bill payments, phone bill payments, food orders, manually entered card details, etc. It is an open gateway for fraudsters to make fraudulent transactions.

Globally, credit card fraud continues to rise, with projected losses reaching $41.06 billion by 2030 and $48.50 billion by 2034 [1]. As fraud techniques evolve, so must the methods used to detect them. Ongoing research focuses on developing real-time, accurate fraud detection systems using machine learning, especially in the face of highly imbalanced datasets.

Figure *1* presents the global losses from credit card fraud over the last five years [1-3].

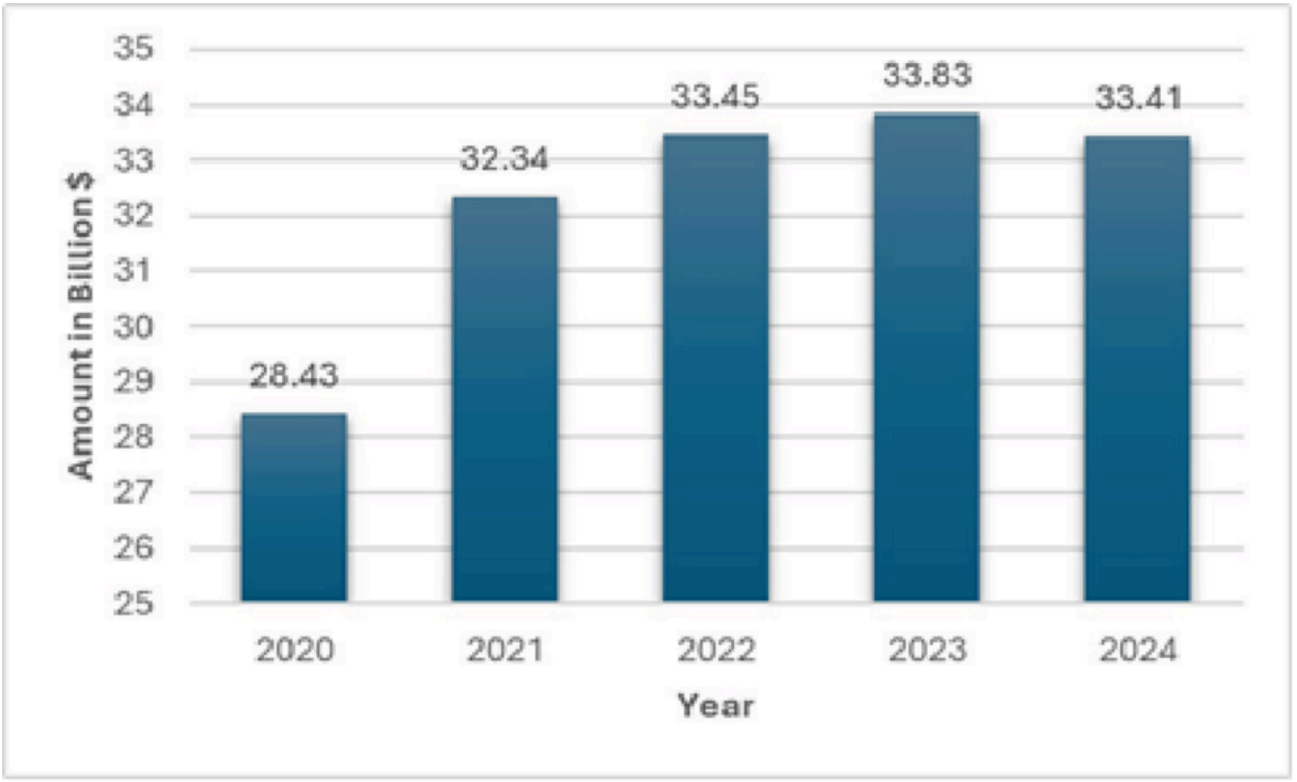

FIGURE 1: Global Losses From Credit Card Fraud

Chart created by the author using data compiled from publicly available industry and news reports.
Data sources: [1-3].

Credit card fraud datasets are heavily imbalanced due to the rare occurrence of fraudulent transactions. To mitigate heavily imbalanced data, researchers apply resampling techniques to the training dataset. This study compares models trained on resampled versus original datasets, emphasizing the importance of evaluating performance on an untouched test set.

Background

Credit card fraud creates a major strain on both cardholders and financial institutions. Victims face the inconvenience of disputing unauthorized charges, while financial institutions must investigate incidents, block compromised cards, and issue replacements to prevent further misuse. As fraudulent transactions are rare compared to legitimate ones, most fraud detection research emphasizes the dual challenge of class imbalance and accurate classification. To address this issue, studies commonly apply sampling techniques that adjust the training data to better represent minority fraud cases. Approaches such as oversampling, undersampling, and hybrid sampling are widely used to enhance minority class visibility and reduce bias toward the majority class. Together, these methods form the foundation of many fraud detection pipelines and play a crucial role in improving recall and overall model robustness under highly imbalanced conditions.

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

Building on these imbalance handling strategies, several studies have evaluated the effectiveness of different classifiers under various sampling conditions. Sundaravadivel et al. [4] applied the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance and reported that Random Forest achieved the highest accuracy and recall, demonstrating its robustness in fraud detection. Similarly, Gupta et al. [5] compared imbalanced and balanced datasets using oversampling, undersampling, and SMOTE, finding that eXtreme Gradient Boosting (XGBoost) performed substantially better when trained on balanced data, underscoring the importance of resampling in improving precision and accuracy. Other research has shown that the effectiveness of sampling techniques varies with the degree of imbalance. For example, Wongvorachan et al. [6] observed that random oversampling works well for moderately imbalanced datasets, whereas hybrid resampling is more suitable for extreme imbalance. Awoyemi et al. [7] further demonstrated that combining oversampling and undersampling enhances classifier performance, with k-nearest neighbors outperforming naive Bayes and logistic regression in their comparative study.

Beyond sampling, researchers have also explored behavior-based and ensemble-based approaches. Quah and Sriganesh [8] introduced a real-time fraud detection model using self-organizing maps to decipher, filter, and analyze customer spending patterns, illustrating the value of unsupervised learning in identifying anomalous behavior. Similarly, a study by Manohari et al. [9] demonstrated that modeling deviations in customer behavior using Support Vector Machine (SVM)-based anomaly detection can significantly improve the identification of subtle fraudulent patterns. On the ensemble side, Ranjan et al. [10] compared multiple imbalance handling strategies with ensemble classifiers such as AdaBoost, XGBoost, and Random Forest, concluding that oversampling followed by undersampling yields strong performance across these models. This aligns with the findings of Khalid et al. [11], who demonstrated that combining heterogeneous learners within an ensemble framework enhances robustness and stability, particularly in highly imbalanced fraud datasets. Farabi et al. [12] evaluated several machine learning algorithms using cross-validation and reported the relative strengths of each model in credit card fraud detection.

Across the literature, a common practice for dataset partitioning is to split the original dataset into a training set and a test set. The training set is further processed using sampling techniques, particularly when the dataset is highly imbalanced, and classifiers are then used to train the model, which is subsequently evaluated using the test set. While this approach is widely accepted, it leaves open the question of how model behavior changes when the resampled dataset is further split into new training and test sets. This study addresses that gap by comparing the performance of models trained on original dataset splits and resampled dataset splits, providing insights into how different splitting strategies influence classifier behavior under imbalanced conditions.

Problem statement

Credit card fraud detection is challenging due to the extreme imbalance between legitimate and fraudulent transactions, which limits the ability of machine learning models to learn meaningful patterns from the minority class and often shows a decrease in predictive power. Although sampling techniques such as oversampling, undersampling, and hybrid methods are widely used to mitigate this imbalance, challenges remain in how these techniques are incorporated into the evaluation process.

Even though the resampled dataset model (RDM) method achieves impressive performance (refer to Table *6*), these outcomes stem from a setup in which the original training set is first resampled, and both training and testing are performed on the resampled data (refer to Figure *2*). When applied in this manner, the protocol can unintentionally introduce overlap (data leakage) between the two sets, as synthetic or duplicated samples may appear in both, leading to artificially inflated performance metrics that do not reflect true generalization. This overlap can elevate performance scores beyond what would be expected in a real-world scenario, giving an overly optimistic impression of model effectiveness. The issue lies not in sampling itself, but in how the protocol is applied. A lack of awareness about this nuance can lead to overly optimistic conclusions about model capability.

Therefore, there is a need to systematically examine how different evaluation strategies, including those that may be unintentionally misapplied, affect fraud detection performance and to emphasize the importance of adopting rigorous, leakage-free protocols when working with highly imbalanced datasets. This study addresses this need by empirically

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

analyzing the impact of various evaluation protocols and highlighting the consequences of applying sampling before the train-test split.

Materials And Methods

Dataset description

This study uses the publicly available credit card fraud detection dataset from Kaggle [13], containing 284,807 transactions recorded over two days in September 2013. Among these, 492 transactions are fraudulent (0.1727%), resulting in a highly imbalanced dataset. Due to confidentiality constraints, most features (columns V1-V28) are transformed using principal component analysis, while 'Time' and 'Amount' remain untransformed. The feature 'Class' is the target variable, where a value of 1 indicates fraud and a value of 0 indicates non-fraud [14].

Sampling techniques

Sampling technique is the process of selecting subsets of data from a larger dataset to train the models. It helps reduce computational costs and time while maintaining meaningful model performance. Nine sampling techniques were applied to address class imbalance. These techniques represent diverse oversampling, undersampling, and hybrid strategies commonly used in fraud detection research.

1. Oversampling [15]

· SMOTE [16]
· Adasyn (Adaptive Synthetic Sampling) [17]
· SVM SMOTE

2. Undersampling [15]

· Random Undersampling
· TL (Tomek Links)
· Cluster Centroids

3. Hybrid Sampling [18]

· SMOTE-Tomek [19]
· SMOTE ENN (SMOTE-Edited Nearest Neighbors) [19]
· Adasyn TL

Classifiers

Four machine learning classifiers were evaluated: Random Forest Classifier, XGBoost Classifier, Decision Tree Classifier, and AdaBoost Classifier.

To ensure a fair and reproducible comparison, each model was trained using a fixed set of hyperparameters appropriate to its classifier family. These configurations were kept consistent within each classifier group and were not tuned individually for different sampling techniques. The complete list of hyperparameter values is provided in Table *1*.

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

Classifier	Hyperparameter	Value
Random Forest	n_estimators	100
	random_state	42
XGBoost	n_estimators	100
	random_state	42
	objective	binary:logistic
Decision Tree	random_state	42
	criterion	entropy
AdaBoost	n_estimators	100
	random_state	42
	base_estimator	DecisionTreeClassifier(max_depth=3)
	algorithm	SAMME

TABLE 1: Hyperparameters Used for Classifiers in This Study

Implementation method

The dataset obtained from the Kaggle platform [13] is referred to as the original dataset (OD) throughout this study. Nine sampling techniques were applied to OD training set to generate a resampled dataset (RD). The following three models were designed to facilitate comparative analysis, as shown in Figure *2*:

Original Dataset Model (ODM) method: The OD is split into training and test sets. The model is trained on the OD training set and evaluated on the OD test set.

Resampled Dataset, Original Dataset Model (RDODM) method: Here, the model is trained on RD and evaluated on the OD test set, which is the method generally used to assess the model.

Resampled Dataset Model (RDM) method: In this method, RD is subsequently split again into training and testing subsets. The model is trained on the RD training set and evaluated on the RD test set. Because the RD test set contains synthetic or modified samples derived from the OD training data, it may introduce overlap between training and testing samples. This structure allows the study to compare correct and incorrect sampling practices.

Each model was trained and evaluated across various combinations of four classifiers and nine sampling techniques to determine the likelihood of fraudulent transactions. To maintain consistency across all experimental setups, the dataset was split into 80% for training and 20% for testing.

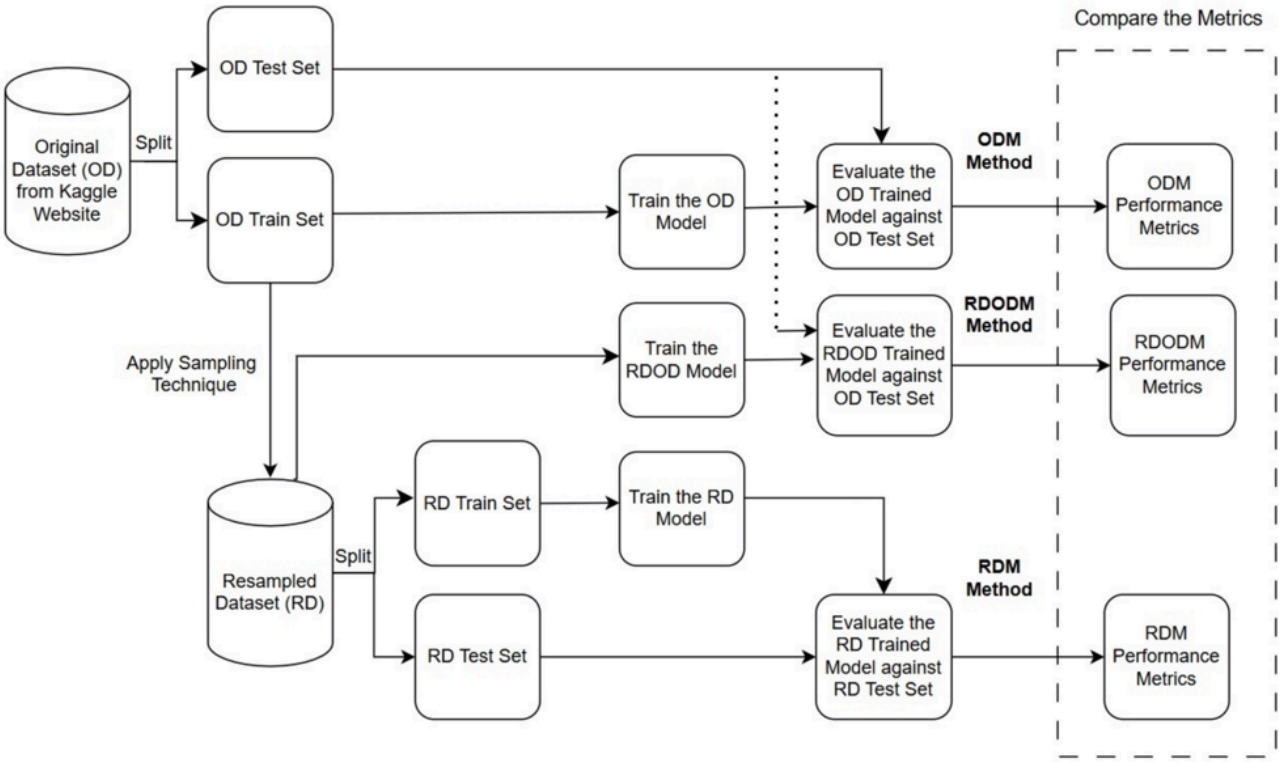

FIGURE 2: Three Model's Workflow Architecture Overview

Source: Designed by the author
ODM, Original Dataset Model; RDODM, Resampled Dataset, Original Dataset Model; RDM, Resampled Dataset Model

Evaluation metrics

Model performance was assessed using Accuracy, Precision, Recall, and Matthews Correlation Coefficient (MCC). Additionally, the Precision-Recall Area Under the Curve (PR-AUC) is presented as a robust evaluation metric, particularly effective for assessing binary classification models on highly imbalanced datasets.

Results And Discussion

Experimental setup

All experiments were conducted using the publicly available credit card fraud dataset, which exhibits a severe class imbalance. Nine sampling techniques were applied to the OD training set to generate the RD. Four classifiers, namely Random Forest, XGBoost, Decision Tree, and AdaBoost, were evaluated under three configurations: ODM, RDM, and RDODM. In the ODM method, models were trained and tested on the OD. In the RDODM method, models were trained on the RD training set and evaluated on the OD test set. In the RDM method, the RD was again split into training and testing subsets, and models were trained and evaluated on these subsets. The dataset was split into training and testing sets using an 80:20 ratio. Performance was assessed using Accuracy, Precision, Recall, MCC, and PR-AUC. All experiments were executed using Python and scikit-learn in a standard computing environment.

Overview of the results

The output of the ODM method remains consistent across all sampling techniques in each classifier table, since no sampling is applied to this model (refer to Figure *2*). Any variation in results is solely due to differences in the classifiers used. The inclusion of the values of the ODM method serves as a baseline reference, enabling meaningful comparison

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

with the performance of other models that incorporate sampling techniques. In all four tables (Tables *2*-*5*), bold values indicate the best performance among the sampling techniques applied in the RDODM and RDM methods. The values of the ODM method are excluded from this comparison, as no sampling technique was applied in that case.

The italicized values in Tables *2*-*5* correspond to the Cluster-Centroids sampling technique, which produced very few samples in this dataset and therefore yielded unusually high accuracy and performance values. Cluster Centroids is an undersampling method, also referred to as the 'heart of the cluster,' that applies the k-means clustering algorithm to the majority-class samples and replaces each cluster with its centroid [20]. In highly imbalanced and high-dimensional fraud data, this process collapses many majority-class samples into a small number of centroids, resulting in an excessively reduced training set. Due to this limitation and its lack of effectiveness in our experiments, the results from this technique are excluded from the final comparative analysis.

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

Sampling Techniques		**Random Forest Classifier**											
		OD Train/OD Test Output (ODM)				RD Train/OD Test Output (RDODM)				RD Train/RD Test Output (RDM)			
		Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC
Oversampling Technique	SMOTE	99.96 %	0.97402	0.7653	0.8632	99.95 %	0.85567	0.84693	0.85103	**100.00%**	**0.99991**	**1**	**0.99991**
	ADASYN					99.95 %	0.84693	0.84693	0.84667	99.99 %	0.99982	1	0.99982
	SVM SMOTE					**99.96 %**	**0.95294**	**0.82653**	**0.88731**	99.98 %	0.99991	0.99964	0.99956
Undersampling Technique	TOMEK LINKS					99.96 %	0.97468	0.78571	0.87492	99.96 %	0.94029	0.84	0.88856
	CLUSTER CENTROIDS					5.98%	0.00182	1	0.0103	*100.00%*	*1*	*1*	*1*
	Random Undersampling					96.63 %	0.04495	0.91836	0.19911	93.04 %	0.9186	0.9518	0.86075
Hybrid Sampling Technique	SMOTE TOMEK					99.95 %	0.86597	0.85714	0.86131	99.99 %	0.99982	1	0.99982
	SMOTE ENN					99.94 %	0.82352	0.85714	0.83988	99.98 %	0.99974	0.99995	0.99969
	ADASYN TL					99.95 %	0.86458	0.84693	0.85546	99.99 %	0.99978	1	0.99977

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

TABLE 2: Comparative Results of Sampling Techniques With Random Forest Classifier

ADASYN, Adaptive Synthetic Sampling; ENN, Edited Nearest Neighbors; MCC, Matthews Correlation Coefficient; OD, Original Dataset; ODM, Original Dataset Model; RDODM, Resampled Dataset, Original Dataset Model; RDM, Resampled Dataset Model; SMOTE, Synthetic Minority Oversampling Technique; SVM, Support Vector Machine; TL, Tomek Links

Across the four classifiers, a consistent pattern is observed in how sampling techniques interact with each evaluation model. In Table *2*, the Random Forest classifier achieves the best performance when paired with SVM SMOTE under the RDODM method, while SMOTE produces the highest scores in the RDM method. This indicates that SVM SMOTE provides a stronger minority-class structure when evaluated on original data, whereas the RDM methods' results are inflated due to the use of a resampled test set.

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI
https://doi.org/10.7759/s44389-026-00104-z

Sampling Techniques		**XGBoost Classifier**											
		OD Train/OD Test Output (ODM)				RD Train/OD Test Output (RDODM)				RD Train/RD Test Output (RDM)			
		Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC
Oversampling Technique	SMOTE	99.96 %	0.96202	0.77551	0.86354	99.93 %	0.78095	0.83673	0.80802	**99.99 %**	**0.99986**	**1**	**0.99986**
	ADASYN					99.94 %	0.82178	0.84693	0.83397	99.99 %	0.99971	1	0.99971
	SVM SMOTE					**99.96 %**	**0.95238**	**0.81632**	**0.88154**	99.98 %	0.99995	0.99969	0.99964
Undersampling Technique	TOMEK LINKS					99.95 %	0.93827	0.77551	0.85279	99.96 %	0.92424	0.81333	0.86681
	CLUSTER CENTROIDS					22.97 %	0.0022	0.98979	0.02155	*100.00%*	*1*	*1*	*1*
	Random Undersampling					95.68 %	0.03613	0.93877	0.17962	93.67 %	0.91954	0.96385	0.87389
Hybrid Sampling Technique	SMOTE TOMEK					99.94 %	0.80392	0.83673	0.81984	99.99 %	0.99973	1	0.99973
	SMOTE ENN					99.93 %	0.75454	0.84693	0.79904	99.99 %	0.99972	1	0.99972
	ADASYN TL					99.94 %	0.82	0.83673	0.82802	99.98 %	0.99964	1	0.99964

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

TABLE 3: Comparative Results of Sampling Techniques With XGBoost Classifier

ADASYN, Adaptive Synthetic Sampling; ENN, Edited Nearest Neighbors; MCC, Matthews Correlation Coefficient; OD, Original Dataset; ODM, Original Dataset Model; RD, Resampled Dataset; RDODM, Resampled Dataset, Original Dataset Model; RDM, Resampled Dataset Model; SMOTE, Synthetic Minority Oversampling Technique; SVM, Support Vector Machine; TL, Tomek Links; XGBoost, eXtreme Gradient Boosting

In Table *3*, a similar pattern appears for the XGBoost classifier. SVM SMOTE yields the best performance under the RDODM method, as XGBoost benefits from well-defined minority-class boundaries created by this technique. In contrast, SMOTE produces near-perfect scores under the RDM method, confirming that these inflated results arise from evaluating on a resampled test set rather than reflecting genuine classifier capability.

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

Sampling Techniques		Decision Tree Classifier											
		OD Train/OD Test Output (ODM)				RD Train/OD Test Output (RDODM)				RD Train/RD Test Output (RDM)			
		Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC
Oversampling Technique	SMOTE	99.91 %	0.71698	0.77551	0.74521	99.80 %	0.45454	0.81632	0.6083	99.86 %	0.99767	0.99949	0.99716
	ADASYN					99.78 %	0.4278	0.81632	0.59005	99.86 %	0.99793	0.99918	0.99712
	SVM SMOTE					99.91 %	0.70642	0.78571	0.74455	**99.96 %**	**0.99951**	**0.99962**	**0.99914**
Undersampling Technique	TOMEK LINKS					**99.92 %**	**0.77319**	**0.7653**	**0.76884**	99.91 %	0.72	0.72	0.71953
	CLUSTER CENTROIDS					32.96 %	0.00253	0.98979	0.02809	*100.00%*	*1*	*1*	*1*
	Random Undersampling					91.29 %	0.01764	0.90816	0.11976	89.24 %	0.89285	0.90361	0.78418
Hybrid Sampling Technique	SMOTE TOMEK					99.79 %	0.44565	0.83673	0.6098	99.86 %	0.99784	0.99938	0.99722
	SMOTE ENN					99.74 %	0.38095	0.81632	0.55664	99.88 %	0.99816	0.99942	0.99755
	ADASYN TL					99.78 %	0.42391	0.79591	0.57993	99.86 %	0.99796	0.99931	0.99727

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

TABLE 4: Comparative Results of Sampling Techniques With Decision Tree Classifier

ADASYN, Adaptive Synthetic Sampling; ENN, Edited Nearest Neighbors; MCC, Matthews Correlation Coefficient; OD, Original Dataset; ODM, Original Dataset Model; RDODM, Resampled Dataset, Original Dataset Model; RDM, Resampled Dataset Model; SMOTE, Synthetic Minority Oversampling Technique; SVM, Support Vector Machine; TL, Tomek Links

In Table *4*, the Decision Tree classifier behaves differently from the ensemble models. TOMEK LINKS produces the best performance under the RDODM method because Decision Trees are highly sensitive to borderline majority samples, and removing these points helps the tree form cleaner splits. However, SVM SMOTE produces the highest score under the RDM method, reflecting the same optimistic bias caused by evaluating on a resampled test set.

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI
https://doi.org/10.7759/s44389-026-00104-z

Sampling Techniques		AdaBoost Classifier											
		OD Train/OD Test Output (ODM)				RD Train/OD Test Output (RDODM)				RD Train/RD Test Output (RDM)			
		Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC
Oversampling Technique	SMOTE	99.95 %	0.98648	0.74489	0.85702	99.53 %	0.25364	0.88775	0.47313	99.63 %	0.99592	0.99662	0.99252
	ADASYN					99.59 %	0.28196	0.87755	0.49615	99.59 %	0.99599	0.99588	0.99188
	SVM SMOTE					99.93 %	0.79207	0.81632	0.80376	**99.97 %**	**0.99973**	**0.99962**	**0.99936**
Undersampling Technique	TOMEK LINKS					**99.95 %**	**0.9375**	**0.7653**	**0.8468**	99.95 %	0.94915	0.74666	0.84162
	CLUSTER CENTROIDS					21.25 %	0.00217	1	0.02145	*100.00%*	*1*	*1*	*1*
	Random Undersampling					95.34 %	0.03357	0.93877	0.17279	91.77 %	0.89772	0.9518	0.83618
Hybrid Sampling Technique	SMOTE TOMEK					99.65 %	0.31407	0.88775	0.52688	99.63 %	0.99662	0.99594	0.99257
	SMOTE ENN					99.71 %	0.3625	0.88775	0.56627	99.73 %	0.99738	0.99724	0.99453
	ADASYN TL					99.47 %	0.2328	0.89795	0.45574	99.58 %	0.99522	0.99635	0.99154

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

TABLE 5: Comparative Results of Sampling Techniques With AdaBoost Classifier

ADASYN, Adaptive Synthetic Sampling; ENN, Edited Nearest Neighbors; MCC, Matthews Correlation Coefficient; OD, Original Dataset; ODM, Original Dataset Model; RDODM, Resampled Dataset, Original Dataset Model; RDM, Resampled Dataset Model; SMOTE, Synthetic Minority Oversampling Technique; SVM, Support Vector Machine; TL, Tomek Links

Table *5* shows that AdaBoost, which relies on sequentially reweighted Decision Trees, mirrors the behavior of its base learner. In the RDODM method, TOMEK LINKS yields the strongest and most reliable performance, as boundary cleaning reduces the number of misclassified majority samples that would otherwise be repeatedly up-weighted. In the RDM method, SVM SMOTE delivers the best outcome, but these results are again inflated due to the use of a resampled test set.

Classifier	Sampling	Model	Accuracy	Precision	Recall	MCC
Random forest	SVM SMOTE	RDODM	99.96%	0.95294	0.82653	0.88731
Random forest	SMOTE	RDM	100.00%	0.99991	1	0.99991
XGBoost	SVM SMOTE	RDODM	99.96%	0.95238	0.81632	0.88154
XGBoost	SMOTE	RDM	99.99%	0.99986	1	0.99986
Decision Tree	TOMEK LINKS	RDODM	99.92%	0.77319	0.7653	0.76884
Decision Tree	SVM SMOTE	RDM	99.96%	0.99951	0.99962	0.99914
AdaBoost	TOMEK LINKS	RDODM	99.95%	0.9375	0.7653	0.8468
AdaBoost	SVM SMOTE	RDM	99.97%	0.99973	0.99962	0.99936

TABLE 6: Consolidated Results of Top Outcomes From Tables 2-5

MCC, Matthews Correlation Coefficient; RDODM, Resampled Dataset, Original Dataset Model; RDM, Resampled Dataset Model; SMOTE, Synthetic Minority Oversampling Technique; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting

Table *6* presents a consolidated summary of the top outcomes from Tables *2*-*5*, for both RDODM and RDM methods. Among these, the highest overall performance is achieved by the Random Forest classifier using the SVM SMOTE technique in the RDODM method and the SMOTE technique in the RDM method.

Upon reviewing the table contents, it is evident that across all classifiers, the RDM method consistently delivers higher performance compared to the RDODM method. The reason is primarily attributed to the structure of the RDM method, which uses both the training and test sets derived from RD. Since the data has been modified either increased or reduced, based on the applied sampling technique, the test set may contain synthetic or repeated samples that closely resemble the training data. This can artificially inflate performance metrics, making the RD test set less reliable for evaluating true model generalization.

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

In contrast, the RDODM method trains on the resampled dataset but evaluates using the original dataset's test set, which remains untouched by sampling. This approach provides a more realistic and trustworthy assessment of the model's ability to generalize unseen, real-world data. Considering these factors, the most reliable and highest performing configuration is the SVM SMOTE technique applied to the Random Forest classifier within the RDODM method. This setup demonstrates strong predictive capability while maintaining evaluation integrity.

In summary, while the RDM method may show superior metrics due to favorable test data conditions, the RDODM method offers a more accurate reflection of the model's true performance. Evaluating with the original test set ensures that the results are not biased by sampling artifacts and better represent the model's behavior in practical scenarios.

A PR-AUC curve is a graphical tool used to evaluate binary or multi-classification models, particularly on imbalanced datasets. It represents the area under the precision recall curve. It plots precision (accuracy of positive predictions) against recall (model's ability to find all true positive instances). PR-AUC curve focuses only on the positive class, making it more informative when positives are rare.

Figures *3*-*4* present the PR-AUC curves corresponding to the classifiers and sampling techniques listed in Table *6*. This visualization facilitates a comparative understanding of how each model performs across different sampling strategies and classifiers. Each graph includes three curves, representing the performance of the three models: ODM, RDODM, and RDM. The labels used in the graph are described as follows:

· ODM refers to the PR-AUC value of the ODM method
· RDODM refers to the PR-AUC value of the RDODM method
· RDM corresponds to the RDM method

Although the RDM method exhibits higher PR-AUC curve values across several classifiers, this can be attributed to the nature of its test set, which is derived from the resampled dataset. As discussed previously, while the RDM method may show higher performance, its test set drawn from resampled data can introduce bias due to overlapping with training samples. In contrast, the RDODM method, tested on untouched original data, provides a more reliable measure of true model generalization.

Notably, the PR-AUC score for the RDODM method reaches 0.89 for both XGBoost-SVM SMOTE and Random Forest-SVM SMOTE combinations. However, a closer examination of Table *6* and the PR-AUC graph suggests that the Random Forest classifier with SVM SMOTE in the RDODM method deliver the most reliable and overall best performance.

Across all classifiers, the PR-AUC curves for the RDM variant rise sharply and remain near-perfect (AUC = 1.00), reflecting the inflated separability caused by evaluating on a resampled test set that overlaps with training data. In contrast, the RDODM curves show a smoother and more realistic progression, maintaining higher precision at moderate recall levels without the artificial plateau seen in the RDM method. The ODM curves remain consistently lower and flatter, indicating the difficulty of distinguishing minority class instances when no resampling is applied. This visual pattern reinforces that the RDODM method provides the most reliable and unbiased representation of true model performance.

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI
https://doi.org/10.7759/s44389-026-00104-z

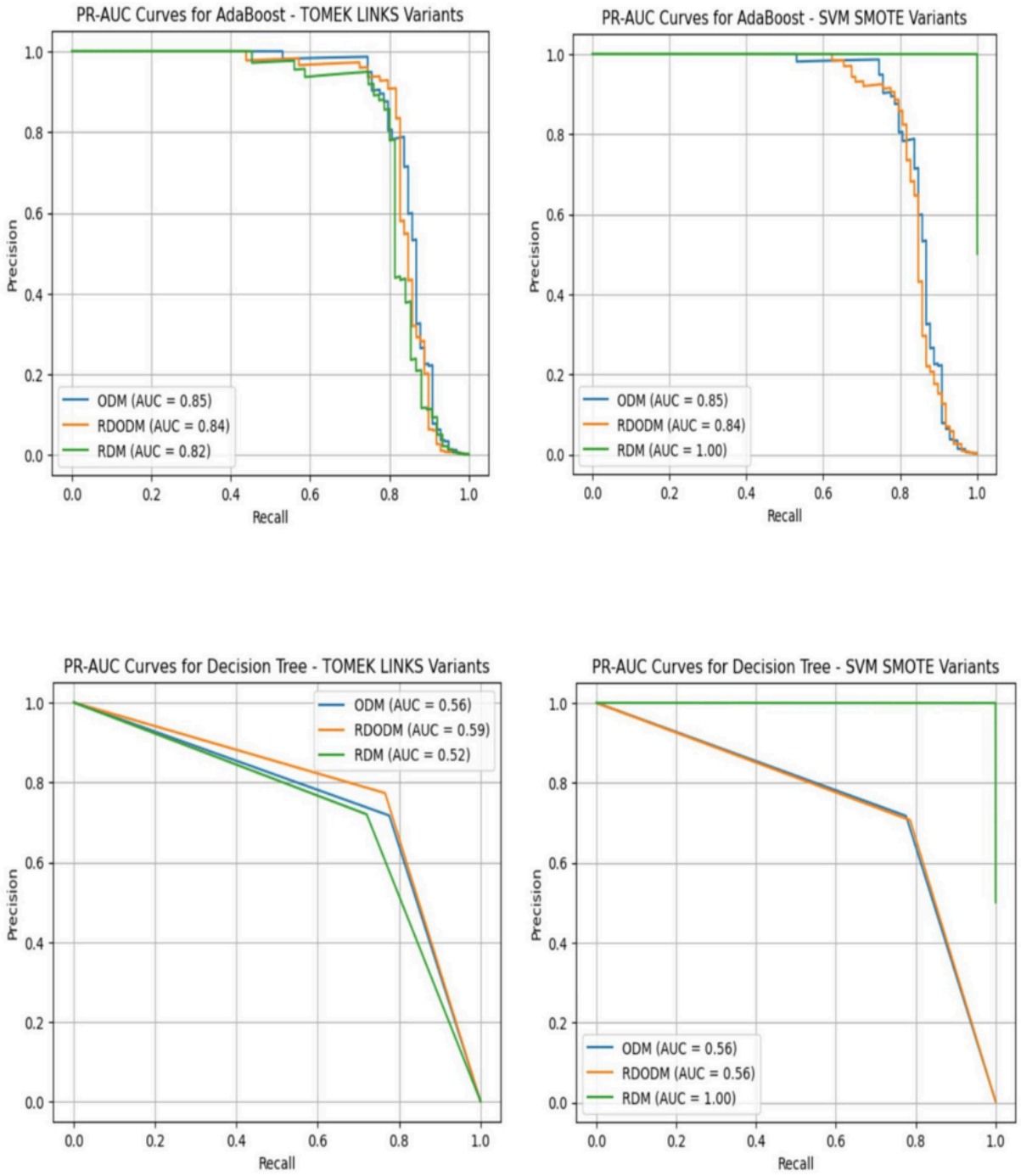

FIGURE 3: PR-AUC Curves for AdaBoost and Decision Tree Classifiers Across Three Methods (ODM, RDODM, and RDM)

Source: Author

ODM, Original Dataset Model; PR-AUC, Precision–Recall Area Under the Curve; RDODM, Resampled Dataset, Original Dataset Model; RDM, Resampled Dataset Model; SMOTE, Synthetic Minority Oversampling Technique; SVM, Support Vector Machine

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI
https://doi.org/10.7759/s44389-026-00104-z

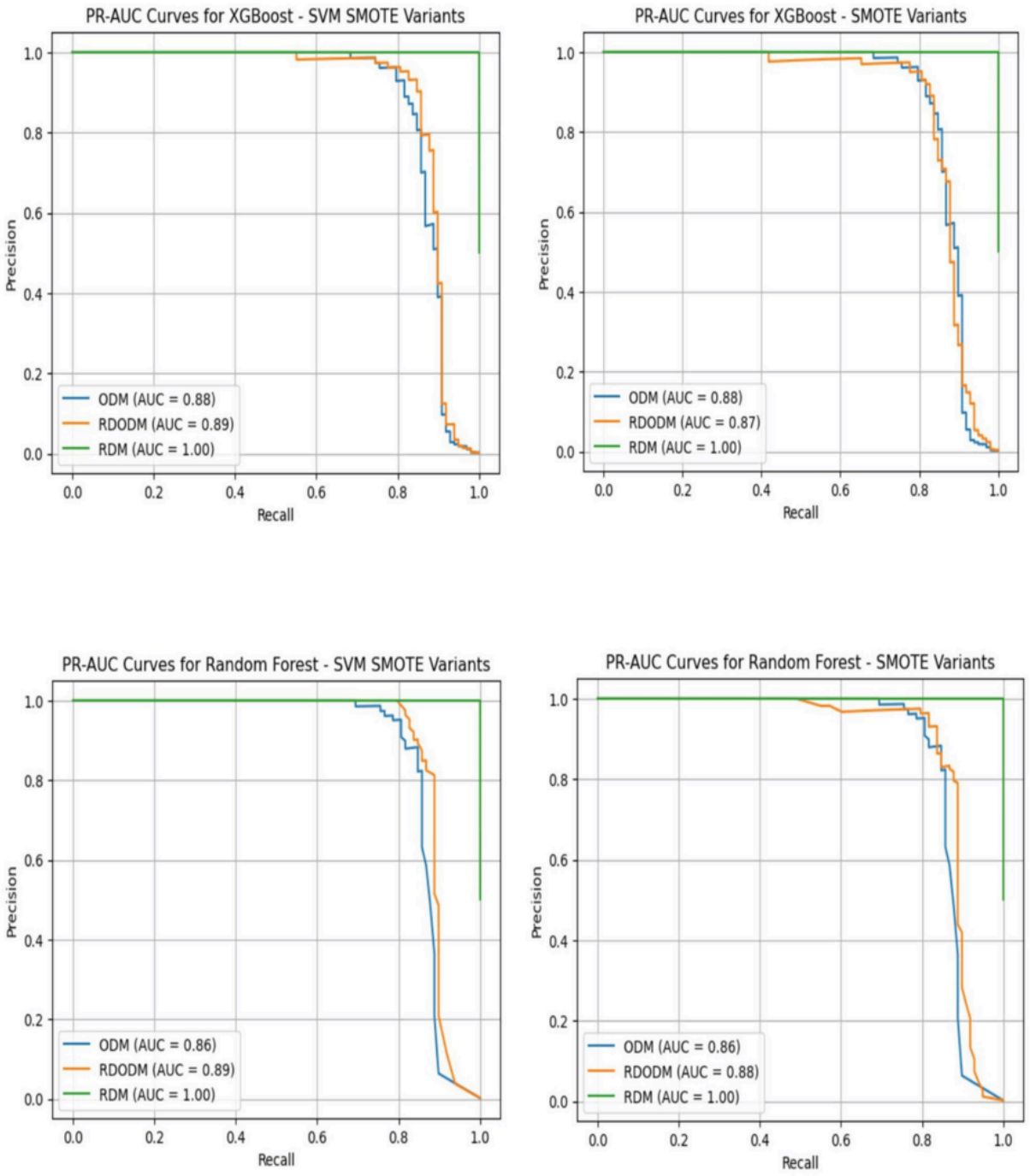

FIGURE 4: PR-AUC Curves for XGBoost and Random Forest Classifiers Across Three Methods (ODM, RDODM, and RDM)

Source: Author

ODM, Original Dataset Model; PR-AUC, Precision–Recall Area Under the Curve; RDODM, Resampled Dataset, Original Dataset Model; RDM, Resampled Dataset Model; SMOTE, Synthetic Minority Oversampling Technique; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

Conclusions

Detecting credit card fraud remains challenging due to the extreme class imbalance inherent in real-world datasets. This study compared three evaluation strategies to assess how different data preparation methods influence model reliability. training/testing on the original dataset (ODM), training/testing on a resampled dataset (RDM), and training on resampled data while testing on the original dataset (RDODM). Results show that the RDM method consistently achieves higher performance, but this inflated accuracy is a consequence of overlap between resampled training and testing data, which makes the evaluation unreliable. In contrast, the RDODM method provides a more accurate measure of generalization. Under this protocol, both XGBoost and Random Forest combined with SVM SMOTE performed similarly, with Random Forest demonstrating slightly greater consistency and reliability. Overall, this work highlights the risks of evaluating models on resampled test data, providing a practical warning for researchers working with highly imbalanced datasets and emphasizes the importance of using untouched test sets to obtain unbiased performance estimates suitable for real-world deployment.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Janaki Raman Palaniappan

Acquisition, analysis, or interpretation of data: Janaki Raman Palaniappan

Drafting of the manuscript: Janaki Raman Palaniappan

Critical review of the manuscript for important intellectual content: Janaki Raman Palaniappan

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

All data generated or analyzed during this study are included in this published article and/or its appendices.

References

1. Global card fraud losses at $33 billion. (2026). Accessed: May 3, 2026: https://www.globenewswire.com/news-release/2026/01/07/3214821/0/en/Global-Card-Fraud-Losses-at-33-Billion.html.
2. Payment card fraud losses approach $34 billion. (2025). Accessed: May 3, 2026: https://www.globenewswire.com/news-release/2025/01/06/3004931/0/en/Payment-Card-Fraud-Losses-Approach-34-Billion.html.
3. Card industry's fraud-fighting efforts pay off: Nilson Report. (2023). Accessed: May 3, 2026: https://www.paymentsdive.com/news/card-industry-fraud-fighting-efforts-pay-off-nilson-report-credit-debit/639675/.
4. Sundaravadivel P, Isaac RA, Elangovan D, KrishnaRaj D, Rahul VVL, Raja R: Optimizing credit card fraud detection with random forests and SMOTE [RETRACTED]. Scientific Reports. 2025, 15:17851. 10.1038/s41598-025-00873-y

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z

5. Gupta P, Varshney A, Khan MR, Ahmed R, Shuaib M, Alam S: Unbalanced credit card fraud detection data: A machine learning-oriented comparative study of balancing techniques. Procedia Computer Science. 2023, 218:2575-2584. 10.1016/j.procs.2023.01.231

6. Wongvorachan T, He S, Bulut O: A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. Information. 2023, 14:54. 10.3390/info14010054

7. Awoyemi JO, Adetunmbi AO, Oluwadare SA: Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI), Lagos, Nigeria. 2017, 1-9. 10.1109/ICCNI.2017.8123782

8. Quah JTS, Sriganesh M: Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications. 2008, 35:1721-1732. 10.1016/j.eswa.2007.08.093

9. Manohari KG, Sravya S, Nandini V, Joy KV, Kumar CR, Pagadala PK: Mitigating credit card fraud through behavior-based classification and anomaly elimination using support vector machine. Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 1. Kumar A, Gunjan VK, Senatore S, Hu YC (ed): Springer, Singapore; 2025. 1273:481-488. 10.1007/978-981-97-8031-0_51

10. Ranjan RK, Singh A, Tiwari A: Credit card fraud detection under extreme imbalanced data: A comparative study of data-level algorithms. Journal of Experimental & Theoretical Artificial Intelligence. 2021, 34:571-598. 10.1080/0952813x.2021.1907795

11. Khalid AR, Owoh N, Uthmani O, Ashawa M, Osamor J, Adejoh J: Enhancing credit card fraud detection: An ensemble machine learning approach. Big Data and Cognitive Computing. 2024, 8:6. 10.3390/bdcc8010006

12. Farabi SF, Prabha M, Alam M, et al.: Enhancing credit card fraud detection: A comprehensive study of machine learning algorithms and performance evaluation. Journal of Business and Management Studies. 2024, 6:252-259. 10.32996/jbms.2024.6.13.21

13. Credit card fraud detection dataset. Accessed: May 3, 2026: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?resource=download.

14. Credit card fraud detection. Accessed: May 3, 2026: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data.

15. Carvalho M, Pinho AJ, Brás S: Resampling approaches to handle class imbalance: a review from a data perspective. Journal of Big Data. 2025, 12:71. 10.1186/s40537-025-01119-4

16. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002, 16:321-357. 10.1613/jair.953

17. He H, Bai Y, Garcia EA, Li S: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong. 2008, 1322-1328. 10.1109/IJCNN.2008.4633969

18. Salehi A, Khedmati M: Hybrid clustering strategies for effective oversampling and undersampling in multiclass classification. Scientific Reports. 2025, 15:3460. 10.1038/s41598-024-84786-2

19. Batista GEAPA, Prati RC, Monard MC: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explorations Newsletter. 2004, 6:20-29. 10.1145/1007730.1007735

20. Bhattacharya R, De R, Chakraborty A, Sarkar R: Clustering based undersampling for effective learning from imbalanced data: An iterative approach. SN Computer Science. 2024, 5:386. 10.1007/s42979-024-02717-4

How to cite this article:
Palaniappan J (June 05, 2026) A Comparative Analysis of Original and Resampled Dataset Splits in Machine Learning Models for Credit Card Fraud Detection. Cureus J Comput Sci 3 : es44389-026-00104-z. DOI https://doi.org/10.7759/s44389-026-00104-z