

Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction

Mohamed Fathimal P¹ , Nithisha SJ¹, Samyuktha S¹, Dharshana R¹

1. Computer Science and Engineering, College of Engineering, Guindy, Anna University, Chennai, IND

Received: February 28, 2026 | Review began: March 20, 2026 | Review ended: May 29, 2026 | Published: June 05, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

This study offers practical insights for developing successful customer retention strategies by highlighting the crucial role that data balancing techniques and advanced machine learning models play in improving churn prediction. This paper uses ensemble machine learning algorithms such as AdaBoost, Gradient Boosting, and XGBoost, along with baseline models like Logistic Regression, Random Forest, K-Nearest Neighbors, Multi-Layer Perceptron, Naive Bayes, and Decision Tree. The publicly available bank and telecom customer retention datasets are used. The Synthetic Minority Oversampling Technique (SMOTE) is applied to address imbalanced dataset issues and improve churn prediction accuracy by integrating machine learning algorithms with Particle Swarm Optimization (PSO) for feature optimization. Significant improvements in handling imbalanced data, reducing bias, and enhancing overall model performance are achieved using SMOTE, SMOTE-Edited Nearest Neighbour (SMOTE-ENN) preprocessing, and PSO-based feature selection. For telecom companies, this study demonstrates the role of machine learning algorithms and PSO in developing effective customer retention strategies and improving churn prediction performance. The paper also examines the impact of PSO for selecting optimal features in combination with SMOTE and SMOTE-ENN for data balancing. The findings show that ensemble approaches, especially XGBoost with PSO, perform better than conventional algorithms.

Categories: AI applications, AI/ML-based decision support systems, Machine Learning (ML)

Keywords: churn prediction, smote, customer retention, machine learning, pso

Introduction

One of the most competitive and dynamic industries, telecommunications, has witnessed a paradigm change in which retaining existing customers is now just as important as gaining new ones. Given the speed at which technology is developing and the growing variety of services available, telecom companies need to constantly innovate to draw in and keep consumers. Customer churn, or losing clients to rivals, has become a major issue in this fiercely competitive environment. Churn has serious repercussions, affecting not only direct revenue but also brand reputation and operational effectiveness. Telecom firms are under constant pressure as customers are exposed to appealing deals, enhanced services, and better network coverage from competitors, making it easier for them to switch providers. Therefore, churn is a complex problem that necessitates strategic intervention rather than just being a financial one. Dissatisfaction with service quality, expensive prices, unstable networks, or more alluring offers from rival firms are some of the causes of client attrition. Even while every client may have different reasons for leaving, churn is a structural issue

How to cite this article:

P M, SJ N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

that companies need to solve if they want to stay competitive. As per the Harvard Business Review article titled "The Value of Keeping the Right Customers" (<https://hbr.org/2014/10/the-value-of-keeping-the-right-customers>), the economic benefit of keeping current customers is five times less expensive than acquiring new ones.

Because of its affordability, churn reduction is essential to the long-term viability of any firm. Additionally, lowering turnover rates increases profitability, improves client lifetime value, and strengthens the company's overall stability. In addition to protecting their revenue streams, businesses that effectively handle customer attrition also establish a reputation for fostering client loyalty, which is a crucial difference in a competitive market.

Traditional tactics including improving customer service, implementing loyalty programs, and offering discounts have been frequently used to reduce attrition. Although these strategies have the potential to be successful, they are frequently reactive and only deal with churn after it has already happened. Furthermore, these tactics might not work effectively for broad, varied customer bases where it becomes difficult to identify at-risk clients.

Predictive analytics, especially machine learning, becomes essential in this situation. Large volumes of customer data can be analyzed using machine learning algorithms, which are well designed to learn from past trends and find hidden patterns that indicate the possibility of churn.

Despite machine learning's potential, data asymmetry remains a major obstacle to creating reliable churn prediction models. Because there are proportionally fewer customers that leave than stay loyal in typical churn datasets, the models are skewed and mostly predict the majority class, which consists of loyal customers. Because it is unable to correctly identify the minority class (churners), this imbalance compromises the model's efficacy. For predictive algorithms to be able to identify both devoted consumers and those who are at risk of churning, effective data balancing is essential.

This study assesses the effectiveness of different machine learning algorithms in addition to data balance to determine the best methods for churn prediction. The capacity of algorithms like Decision Trees (DT), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Weighted Random Forest (RF), as well as sophisticated ensemble techniques like Gradient Boosting and eXtreme Gradient Boosting (XGBoost), to manage complicated and high-dimensional datasets is evaluated. Because ensemble approaches use the characteristics of numerous models to achieve higher accuracy and resilience, they have proven particularly effective in prediction tasks. This study offers a thorough assessment of these algorithms' advantages and disadvantages in churn prediction by using them on balanced and unbalanced datasets.

This study lays the groundwork for next developments in churn prediction, such as the creation of real-time, adaptable models that can react to changing consumer behavior. These next-generation models have the potential to completely transform customer retention initiatives, strengthening bonds and guaranteeing sustained success in a field that is becoming more competitive.

Literature review

Businesses frequently struggle with customer churn, which is defined as the loss of customers over a specific time period. This is especially true in highly competitive industries like telecommunications. Accurately forecasting churn has therefore become crucial, allowing businesses to take proactive steps to keep clients and lessen the financial burden that attrition causes. For this reason, a variety of machine learning approaches have been investigated since they can spot patterns and trends that conventional approaches could miss.

There is a class imbalance because, in real-world scenarios, the number of customers who discontinue is usually substantially smaller than the number of consumers who remain faithful to the service. This discrepancy has the potential to significantly affect machine learning models' efficacy by causing them to bias predictions in favor of the majority class (non-churned consumers), which lowers the accuracy of churn forecasts.

Numerous studies have looked at various data balancing techniques to address this issue. To establish dataset balance, oversampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) have been widely used to generate synthetic examples of the minority class. SMOTE improves the model's ability to learn from the minority class by interpolating between current examples to create new cases for the underrepresented class. Furthermore, researchers

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. *Cureus J Comput Sci* 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

have also stressed the importance of feature engineering and selection in improving churn prediction models. To predict attrition, attributes such as customer demographics, transaction history, service usage trends, and customer feedback have been widely used. Principal component analysis and recursive feature elimination are two feature selection techniques that have been used to reduce dimensionality and find the most important information for churn prediction. To increase prediction accuracy, these hybrid models aim to use the benefits of many methods. While working with high-dimensional, complicated datasets - like those used in churn prediction - where basic models might not be able to identify complex patterns, these hybrid approaches can be quite helpful. Furthermore, understanding the efficacy of churn prediction models depends heavily on model evaluation.

The models' overall performance indicators, such as accuracy and precision, were greatly enhanced by this preprocessing phase, demonstrating the potential of combining supervised and unsupervised learning approaches to tackle challenging churn prediction problems. The study showed how hybrid techniques could offer reliable answers for prediction tasks, especially when dealing with datasets that cross class borders or have considerable variability. Social network analysis was used by Backiel et al. [1] to forecast turnover among prepaid mobile subscribers in the telecom industry. Their research made use of a sizable dataset from a well-known telecom company in Belgium, which included 200 million call logs and 1.4 million users during six months. From a network graph that depicted customer relationships, the researchers collected interaction parameters like the number of links to churn neighbors and the duration of calls with churners. This strategy produced better lift, accuracy, and profitability results than prior methods that only used individual consumer data.

Using a dataset from the UCI Machine Learning Repository, Chaubey et al. [2] carried out an extensive investigation evaluating the effectiveness of different supervised machine learning algorithms for forecasting consumer purchase habits. KNN, DT, Logistic Regression (LR), and sophisticated hybrid approaches like stacking-based classifiers were among the algorithms examined. With an impressive accuracy of 92.42%, the study demonstrated the efficacy of a hybrid model that combines KNN with Stochastic Gradient Descent. Although conventional models such as LR were praised for their ease of use and interpretability, they had trouble identifying non-linear relationships.

Research on attrition prediction was extended from the banking industry to the telecom industry by de Lima Lemos et al. [3].

To overcome class imbalances in prediction tasks, K-Means SMOTE, a unique hybrid technique that combines SMOTE with K-Means clustering, was presented by Douzas et al. [4]. K-Means SMOTE uses clustering to find minority-dense regions, guaranteeing that synthetic samples are generated in significant and pertinent sections of the feature space, in contrast to classic SMOTE, which may generate synthetic samples arbitrarily. K-Means SMOTE method increased machine learning models' capacity to recognize minority-class cases precisely by improving the quality of synthetic data.

Eria and Marikannan [5] used machine learning methods like Naïve Bayes for CCP, SVM, Neural Networks, and DT. Given their efficacy in both regression and classification tasks and their steady performance with sizable datasets, SVM and neural networks stood out among these for their high adoption rates. Significant issues with telecom churn prediction were brought to light by the study, such as high dimensionality, class imbalances, and enormous dataset sizes. Feature selection, normalization, and under sampling were among the many data preparation methods used to solve these problems.

The study also highlighted how machine learning approaches may be applied to a variety of businesses, demonstrating how they can be used to solve churn-related issues in many fields. To maximize the financial gains from retention initiatives, a profit-centric approach was adopted for churn prediction in the telecom sector [6]. The study gave telecom companies a useful framework for setting priorities for their retention efforts by analyzing the cost-benefit trade-offs related to retaining customer segments. This study underlined how feature selection strategies, ensemble approaches, and data balancing tactics are becoming increasingly important for enhancing model performance. A complex ensemble-fusion model for churn prediction was presented by He and Ding [7]. To forecast churn among B2C e-commerce clients, SVM was used for prediction with k-means clustering for segmentation. Based on behavioral data, this method divided customers into discrete groups, allowing for more focused retention and prediction tactics. The study

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. *Cureus J Comput Sci* 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

showed that SVM outperformed LR when it came to managing multidimensional data variables, as reported by Xiahou and Harada [8]. To address the skewed distribution of churn and non-churn instances, the study also incorporated hybrid resampling techniques, such as SMOTE and SMOTE-ENN. RF was used as the main classifier in their machine learning framework to forecast customer attrition in the telecom sector. By combining machine learning algorithms with creative feature selection methods, a strong churn prediction system was created by Tongrao [9], specifically for the banking sector. Krishna et al. [10] used machine learning models to evaluate performance. The study investigated the application of algorithms known for their effectiveness in handling classification tasks, including AdaBoost, XGBoost, RF, DT, and LR. Lalwani et al. [11] used a machine learning approach for churn prediction. The Telecom dataset was developed by Macko [12].

A study on machine learning-based customer attrition prediction was carried out by Prabadevi et al. [13]. Using a telecom dataset, they tested four models: Stochastic Gradient Booster, RF, KNN, and LR. Stochastic Gradient Booster produced the best accuracy, at 83.9%. The model was the most successful of the algorithms examined because of its ability to manage nonlinear data and minimize bias mistakes.

A unique churn prediction framework was created by Rajamohamed and Manokaran [14] that integrated rough K-Means clustering with a variety of supervised machine learning models, including SVM, DT, and RF.

An advanced churn prediction algorithm was specifically designed for search ads using Gradient Boosting DT [15]. Both static features, such as account creation information, and dynamic features, such as user interaction history, including impressions, clicks, and search habits, were examined in their study. Advanced methods such as Fourier processing were used to improve the depiction of consecutive user activities, which allowed the researchers to record trends and patterns in client behavior over time. The study made sure the predictive model could successfully differentiate between churners and non-churners by optimizing the feature set.

A thorough investigation utilizing Orange Telecommunications data was carried out using 50,000 records and 230 variables, which offered a wealth of information for creating predictive models, as reported by Sagming et al. [16]. Undersampling, which entails shrinking the size of the majority class to match the minority class, was used in the study to address the problem of data imbalance. These cutting-edge characteristics improved the models' predictive capacity, reaching the highest possible level of accuracy.

Sagming et al. [16] improved their models' predictive power through these sophisticated features, allowing them to achieve a maximum accuracy, precision, recall, and F-measure of 98.5%. However, as other methods, such as SMOTE, were not investigated, the dependence on undersampling left room for improvement. Data from Denmark and the US were used to investigate customer retention and churn in the telecom industry [17]. Their study used cutting-edge machine learning models, such as RF and XGBoost, to examine several variables that affect customer churn, including service quality, network dependability, and customer satisfaction.

A novel ratio-based data balancing technique was introduced specifically tailored to address the issue of class imbalance in customer churn prediction in the telecom industry [18]. This strategy provided a novel approach by fine-tuning the representation of churners and non-churners in the training datasets, resulting in a more equitable distribution across classes. Unlike standard oversampling or undersampling strategies, this method attempted to balance the data more strategically, avoiding the hazards of overfitting and data loss. Machine learning models such as Gradient Boosting and XGBoost enhanced their predictive performance when using this balancing strategy, especially when working with datasets with skewed distributions. A customer attrition prediction was investigated in a rental service scenario, namely water purifier contracts. Their dataset consisted of around 84,000 contracts, and they created machine learning models that incorporated domain-specific feature engineering and Principal component analysis [19]. Table 7 highlights the benefits and limitations of existing research works.

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. *Cureus J Comput Sci* 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

References	ML Algorithms Used	Objectives	Benefits	Limitations
Backiel et al. [1]	Hybrid Methods for Social Network Analysis, Time-to-event	Predict the "time-to-churn" for prepaid mobile users using network influence.	Incorporates relational data (call patterns) beyond static demographics.	High computational cost for large networks; requires detailed call logs.
Chaubey et al. [2]	SVM, Random Forest, XGBoost	Predict customer purchasing behavior and churn risk via classification.	High predictive accuracy; compares multiple supervised models for best fit.	Sensitivity to feature selection; potential overfitting on niche datasets.
de Lima Lemos et al. [3]	Neural Networks, Random Forest, Gradient Boosting	Predict churn propensity in financial institutions using a large-scale ML approach.	Effectively handles complex, non-linear financial behavior patterns.	"Black box" nature of neural networks; requires massive historical data.
Verbeke et al. [6]	Profit-driven Data Mining, Logit-Ensemble Methods	Shift churn prediction from "accuracy-centric" to "profit-centric."	Directly aligns ML outcomes with business ROI and retention costs.	Complex implementation; requires detailed cost/benefit financial data.
He and Ding [7]	Hybrid Ensemble-Fusion Model	Improve prediction stability and accuracy through a multi-model fusion.	Reduces variance; more robust than single-algorithm approaches.	Increased model complexity and longer training times.
Krishna et al. [10]	Random Forest, XGBoost, CatBoost	Benchmark modern gradient boosting techniques for telecom churn.	Achieves state-of-the-art accuracy; handles missing data well.	Requires extensive hyperparameter tuning for optimal performance.
Wang et al. [15]	Large-scale Ensemble, Search Ads	Predict churn for advertisers in high-volume search advertising.	Scalable to millions of records; low latency in prediction.	Focused specifically on ad-tech; may not generalize to telco/finance.
Sagming et al. [16]	Gradient Boost for Topological Data Analysis (TDA)	Extract geometric features from data to enhance churn prediction.	Captures deep structural changes in behavior that standard ML misses.	Extremely high computational complexity; requires specialized TDA expertise.
Saleh and Saha [17]	Random Forest, AdaBoost, LogReg	Case study on Danish university student churn/retention.	Provides unique insights into the "student-as-customer" persona.	Geographically specific results; small survey sample size (311).

How to cite this article:

P M, SJ N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

Sikri et al. [18]	Ratio-based Balancing, XGBoost	Introduce a novel data balancing technique for skewed telecom data.	Outperforms SMOTE; yields higher F-scores for minority classes.	Novelty of ratio-balancing requires further validation across industries.
Xiahou and Harada [8]	K-means (Clustering) + SVM	segment B2C e-commerce customers before applying churn classification.	Better performance by training on homogeneous customer segments.	Two-stage process increases pipeline latency; cluster stability issues.
Lalwani et al. [11]	Sequential ML System, Random Forest, LogReg	Develop an end-to-end automated churn prediction system.	Practical, deployable architecture for real-time business use.	Performance is highly dependent on the quality of automated data pipelines.
Prabadevi et al. [13]	KNN, Decision Trees, Naive Bayes	Comparative analysis of lightweight ML models for churn.	High interpretability; low computational requirements for small firms.	Lower accuracy compared to ensemble methods like XGBoost.
SMOTE				
Douzas et al. [4]	K-Means, SMOTE, SVM, MLP	Address the class imbalance problem in churn datasets using oversampling.	Significantly improves recall for the minority (churner) class.	Risk of creating synthetic noise; K-means clustering is sensitive to outliers.
Suh [19]	ML for Home Appliance Rental	Predict churn in the recurring revenue home appliance rental sector.	Addresses the specific "rental care" lifecycle and contract status.	Rental behavior is very different from standard "subscription" services.
Rajamohamed and Manokaran [14]	Rough Clustering, Neural Networks	Use "Rough Set Theory" to handle uncertainty in credit card data.	Exceptional at managing ambiguous or incomplete customer records.	Mathematically intensive; less familiar to standard business analysts.
Eria and Marikannan [5]	Systematic Review (Meta-analysis)	Synthesize existing telecom churn literature to identify key predictors.	Comprehensive overview of features like "tenure" and "monthly charges."	Limited to existing studies; does not propose a new technical model.

TABLE 1: Critical Analysis of Literature Used

SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting; CatBoost, Categorical Boosting; KNN, K-Nearest Neighbors; ML, Machine Learning; MLP, Multilayer Perceptron; SMOTE, Synthetic Minority Over-sampling Technique

Table 2 classifies the literature based on the algorithms, feature engineering, data balancing, and specific applications used.

How to cite this article:

ML Methods Used	References
Ensemble Methods	
XGBoost	[2,3,6,14,15,17,18]
GBM	[9,16]
LightGBM and CatBoost	[7]
Large-Scale Ensembles	[1,13]
Standard Methods	
DT, RF, Logistic Regression, KNN, SVM	[2-4,6,9-12,16]
Neural Network-Based Method-MLP	[4,5,9,10,16,18]
Feature Extraction and Engineering Techniques	
Network Features	[1]
Topological Data	[16]
K-means Clustering	[8]
Rough Clustering	[12]
Data Balancing Methods	
SMOTE	[4]
Ratio-Based Data Balancing	[16]
Standard Resampling	[4-6]
Sector-Specific Methodological Trends	
Telecommunications	[1,7,14,16]
Financial Services	[3,12,18]
E-Commerce and Digital Services	[13]
Rental Services	[17]

TABLE 2: Literature Classification by Algorithms, Feature Engineering, Data Balancing Methods, and Methodological

How to cite this article:

Trends Used

XGBoost, eXtreme Gradient Boosting; GBM, Gradient Boosting Machine; LightGBM, Light Gradient Boosting Machine; CatBoost, Categorical Boosting; ML, Machine Learning; DT, Decision Tree; RF, Random Forest; KNN, K-Nearest Neighbors; SVM, Support Vector Machine; MLP, Multilayer Perceptron; SMOTE, Synthetic Minority Over-sampling Technique

Materials And Methods

Proposed method

The methodology for this research focuses on leveraging machine learning techniques to predict customer churn in the telecommunications and banking sectors. This study has four main stages: (1) data preprocessing and feature selection, (2) training a prediction model, (3) testing to predict test data, and (4) performance evaluation. The block diagram of the research design is presented in Figure 1.

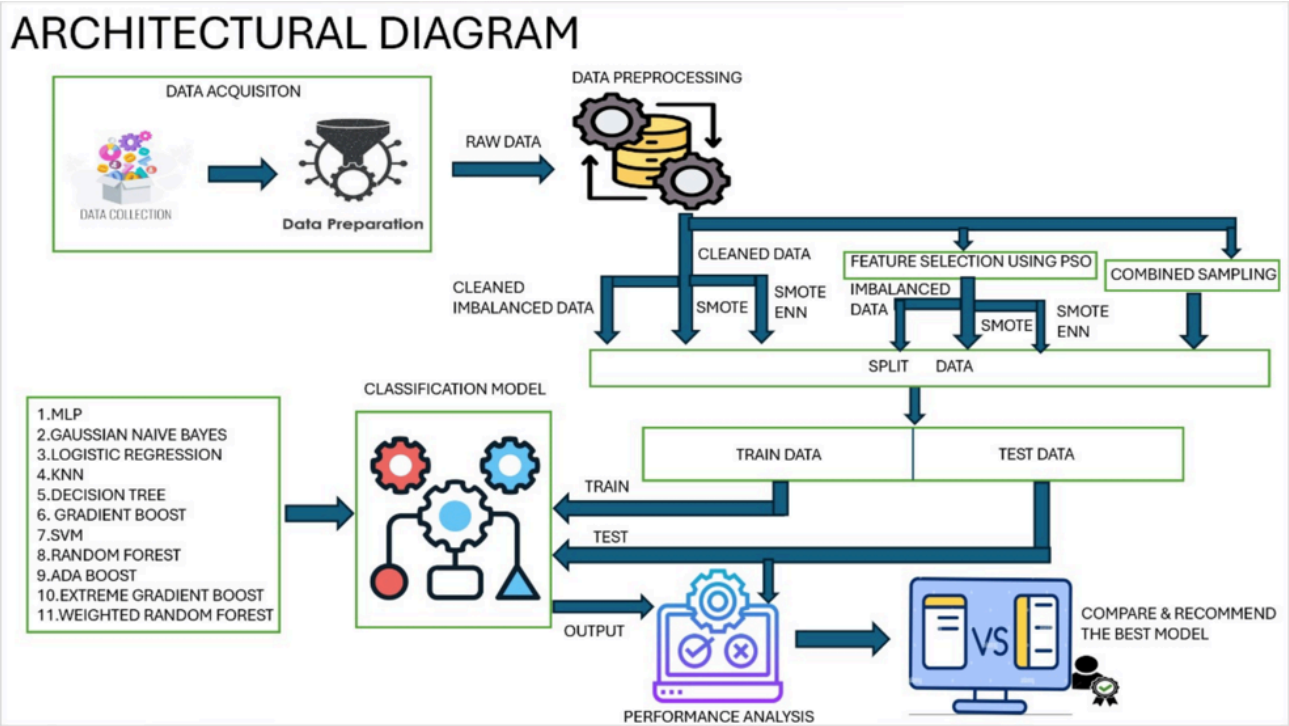


FIGURE 1: Architecture Diagram

SMOTE, Synthetic Minority Over-sampling Technique; ENN, Edited Nearest Neighbours; MLP, Multilayer Perceptron; KNN, K-Nearest Neighbors; SVM, Support Vector Machine

Datasets

This study used a publicly accessible telecom customer dataset that included 20 parameters pertaining to customer demographics, service consumption patterns, and interactions with the telecom provider. The primary goal of the analysis is to anticipate customer churn, with the target variable indicating whether a client has terminated service ("Yes") or is still an active subscriber ("No"). To ensure interoperability with machine learning classification models, the target labels were numerically encoded, with churned customers mapped to 1 and kept customers mapped to 0. In addition to the Telecom dataset, Macko's suggested dataset with 12 selected attributes and the Bank dataset were used. The inclusion of multiple datasets was done to assess the proposed methods' robustness, consistency, and generalization capability in heterogeneous data environments with different feature dimensions, customer behavior patterns, and data

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

distributions. This multi-dataset analysis allows for a more comprehensive evaluation of the approaches' behavioral performance under a variety of real-world scenarios. The comparative analysis framework was designed to evaluate the consistency and adaptability of the proposed methods across datasets with varying dimensionality, feature distributions, and domain-specific characteristics (Table 3).

Dataset	Domain	No. of Features	Preprocessing Applied	Target Variable	Purpose in Analysis
Telecom Dataset	Telecom Customer Churn	20	Label encoding, normalization	Churn (Yes/No)	Evaluate churn prediction under telecom customer behavior patterns
Macko Dataset	Telecom Variant	12 selected features	Feature selection, encoding	Customer Retention	Analyze model behavior under reduced feature space
Bank Dataset	Banking Marketing	35 (after one-hot encoding)	One-hot encoding, scaling	Subscription Outcome	Evaluate generalization across different domain characteristics

TABLE 3: Cross-Dataset Generalization

Preprocessing Data and Feature Selection

The datasets include a variety of features related to customer behavior and demographics and are first pre-processed to ensure the quality and consistency of the data. This involves cleaning the data by handling missing values, removing outliers, and converting categorical variables into numerical representations.

This research study analyses the prediction model using an imbalanced dataset. Therefore, the impacts of SMOTE, SMOTE-ENN (Edited Nearest Neighbor), undersampling of the majority class sample, and feature selection using Particle Swarm Optimization (PSO) were analyzed.

This binary encoding facilitates the use of a range of classification algorithms that require numeric data as input. Numerous attributes in the dataset offer a wealth of information for comprehending consumer behavior. Along with more transactional data like account length, phone charges, and customer service encounters, these elements also include demographic information like the state and area code. Statistics about client usage calls and fees for each service type are also available. However, not every feature in the dataset can be used to forecast churn, and some can add extra complexity or noise to the model. Some unnecessary or superfluous elements are eliminated from the data to improve its quality and the machine learning models' performance. Since the employed datasets do not include detailed financial indicators such as customer lifetime value or retention campaign cost, the present study primarily focuses on predictive performance and cross-dataset generalization analysis using standard classification metrics. Since they either do not directly affect churn prediction or have a strong correlation with other factors that provide more pertinent information,

features like state, area code, total day fee, and account length are not included. By concentrating the model's attention on the most informative variables, the elimination of these characteristics helps reduce dimensionality, simplify the analysis, and lower the chance of overfitting. The information is made more manageable and targeted by removing these less important elements, which enables the machine learning models to focus on the characteristics that are most suggestive of customer attrition. In the end, this refinement process increases the model's accuracy and efficiency, which strengthens the predictability and interpretability of the churn.

Thorough pre-processing step is applied to the dataset to guarantee its quality and machine learning applicability. This stage entails dealing with missing values, getting rid of duplicate entries, and getting rid of features that are not essential to predict customer attrition. Early resolution of these problems cleans and refines the data, lowering the possibility of biases or mistakes that could impair the study. Exploratory data analysis (EDA) is used to better understand the dataset after it has been pre-processed. The distribution of individual features is unbalanced during EDA. To ascertain the characteristics having the greatest impact on customer churn, the associations between the features and the goal variable (churn) are examined.

Statistical tests like the chi-square test are used to evaluate the relationships between the target variable and categorical data, such as state and area code. Certain tests help determine whether there is a statistically significant correlation between customer attrition and certain categorical variables. These characteristics might be considered superfluous for the prediction model and excluded from further examination if the association is not significant. It is also noted that several features show strong relationships with one another throughout the EDA process. Because they are all derived from the same underlying metric of service usage, namely call charges at various times of the day, features like total_day_charge, total_eve_charge, total_night_charge, and total_intl_charge are found to be closely connected. These highly correlated features are eliminated from the machine learning model to prevent redundancy and possible overfitting. By taking this precaution, the model is prevented from processing the same data repeatedly, which could result in inefficient learning and decreased model performance.

The dataset is made more efficient and instructive by carrying out these exploratory and cleaning procedures, which also help machine learning algorithms concentrate on the most pertinent aspects while removing extraneous noise or redundant data. This procedure raises the accuracy of churn predictions in addition to improving model efficiency.

Normalization

Normalization is used in the dataset to enhance the stability and performance of the machine learning models, as well as to rectify class imbalance. Larger-scale characteristics may dominate the learning process if they are employed without normalization, leading to skewed predictions and a slower rate of model convergence. A Min-Max scaler is used to normalize the range of the features.

For algorithms that depend on distance measurements, like KNN, or gradient-based techniques, including LR and neural networks, normalization is especially advantageous. Normalization facilitates faster and more efficient convergence of these methods, which can be sensitive to the magnitude of input features. By guaranteeing that every feature is handled equally, normalization can not only increase convergence but also produce more stable and accurate model predictions by freeing the algorithm from scale-based skews and enabling it to concentrate on the relationships between features. A more reliable and effective machine learning model is produced when SMOTE is used for data balance and normalization is used for feature scaling. This helps in normalizing the feature space and ensuring a balanced representation of both churned and non-churned clients.

Handling Class Imbalance Problem

The class imbalance in the dataset utilized for this study poses a serious problem because there are significantly more "No" (non-churn) occurrences than "Yes" (churn) instances. Due to their propensity to favor the majority class, machine learning models may perform poorly because of this imbalance, frequently producing poor predicted accuracy for the minority class (churn). The model's practical usefulness in real-world situations, where identifying crucial churned

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. *Cureus J Comput Sci* 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

customers is important, is compromised when it is skewed towards forecasting the majority class, as it may miss significant patterns linked to customer churn. Many data balancing strategies are used to lessen this problem and enhance model performance.

SMOTE: An innovative oversampling technique called SMOTE is used in this study to address the class imbalance. Instead of just copying existing minority class instances, which could result in overfitting and poor model generalization, oversampling aims to enhance the representation of the minority class - in this example, the churned consumers. Choosing an instance from the minority class and locating its closest neighbors in the feature space is how this method operates. A distance metric (like the Euclidean distance) that gauges how similar the data points are based on their feature values is used to identify the nearest neighbors.

One or more neighbors are chosen at random, and new data points between the original instance and these neighbors are generated to construct synthetic instances. The original data point and a randomly chosen neighbor are drawn as a vector, which is then multiplied by a random factor between 0 and 1 to form the new synthetic sample. The synthetic instance preserves the original instance's properties while adding additional variation based on the nearby data. It is located inside the convex hull, which is the area enclosed by the nearest neighbors. A more balanced dataset with an equal or more representative number of samples in the minority class than in the majority class is produced by repeating this method for every instance in the minority class. By adding more diversity to the dataset and making it more representative of real-world situations, the synthetic samples produced by SMOTE assist the model in discovering the underlying patterns of the minority class without overfitting. SMOTE helps to expand the decision border between the two classes by producing fresh, artificial samples, which enables the machine learning model to pick up more reliable features.

The new data points between the original instance and these neighbors are generated to construct synthetic instances. This is especially crucial for datasets that are unbalanced, since the model would otherwise be biased in favor of the majority class and miss significant patterns in the minority class. Furthermore, by giving machine learning models extra data to work with, SMOTE enhances their overall predicted performance, especially for algorithms that could otherwise have trouble with class imbalance. These models are now better able to forecast churn in future, unseen data by learning the distinctive characteristics of churned consumers. SMOTE improves the model's generalization capability and guarantees that it does not ignore the minority class by resolving the imbalance in this way, which improves the forecast accuracy for customer attrition. Thus, employing SMOTE as an oversampling strategy is crucial to guaranteeing that the machine learning model can precisely identify both kept and churned consumers, resulting in more trustworthy forecasts for business decision-making in the telecom sector.

SMOTE-ENN: The ENN approach first determines each observation's KNN before determining whether the majority class from the observation's KNN matches the observation's class. The observation and its KNN are removed from the dataset if their classes differ from the majority class of the observation's KNN. $K=3$ is the default number of nearest-neighbors used in ENN.

Step 1: Apply SMOTE to the Dataset.

Step 2: Calculate the number of nearest neighbors, K . $K=3$, if it is not found.

Step 3: Determine the observation's KNN among the other observations in the dataset and then use that neighbor to determine the majority class.

Step 4: The observation and its KNN are removed from the dataset if the observation's class and the majority class of the observation's KNN vary.

Step 5: Steps 2 and 3 should be repeated until the required percentage of each class is reached.

Feature selection using PSO: One popular swarm-based meta-heuristic optimization approach is PSO. Without a leader, it mimics the coordinated movements of flocks of birds seeking food; the birds follow the one closest to the food source. The possible solutions are represented as a collection of particles by the PSO algorithm. To accomplish specific goals in a

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. *Cureus J Comput Sci* 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

multidimensional area, this population-based meta-heuristic mimics social behavior, such as birds converging on a good location. PSO uses a population (a swarm) of individuals (particles) that are updated from iteration to iteration to conduct searches, much like evolutionary algorithms. The goal is to identify the ideal feature subset that gives a classifier the best performance.

The PSO algorithm was initialized with a swarm of particles randomly distributed across the search space. During each iteration, particle positions were updated based on both individual best experience (pBest) and global best experience (gBest) to balance exploration and exploitation. The fitness function was designed to maximize classification performance while minimizing redundant feature selection. Specifically, the fitness evaluation considered the classifier performance in terms of F1-score and recall due to the imbalanced nature of churn prediction datasets.

The velocity and position update process follows the standard PSO mechanism:

$$v_i^{t+1} = wv_i^t + c_1r_1(pBest_i - x_i^t) + c_2r_2(gBest - x_i^t)$$

where w represents the inertia weight, c_1 and c_2 denote cognitive and social learning coefficients, respectively, and r_1 and r_2 are random values in the range 0 to 1.

The optimal feature subset obtained through PSO was subsequently used for classifier training and evaluation under both imbalanced and balanced data scenarios. The integration of PSO with SMOTE/SMOTE-ENN preprocessing enabled the framework to analyze the influence of feature optimization on churn prediction performance across heterogeneous datasets.

PSO Algorithm

- Set up a population of particles, each of which stands for a subset of features.
- Use a classifier on the training and testing data to assess each particle's feature subset's performance.
- Based on the performance evaluation, update each particle's global best and personal best positions.
- Utilizing the global best and personal best positions, update each particle's velocity.
- For a set number of iterations or until a stopping requirement is satisfied (such as no performance gain), repeat steps 2-4.
- Choose the best feature subset that the global best particle represents.
- Analyze the classifier's performance on the test data using the chosen test dataset.

Thus, the two approaches to address data imbalance, (1) combined sampling and (2) cost-sensitive approach, were used.

The weighted RF classifier can handle the imbalanced data problem with a cost-sensitive approach by assigning weights to reduce misclassification. The combination of simple undersampling and the SMOTE algorithm may reduce the substantial loss of potential data from the majority class and reduce the probability of overfitting caused by oversampling using SMOTE.

Training phase

The models are built using machine learning algorithms such as LR, Perceptron, MLP, DT, SVM, KNN, Gaussian Naïve Bayes, and more advanced ensemble methods like RF, Gradient Boosting, AdaBoost, and XGBoost for the following scenarios.

Scenario 1: Classification with imbalanced dataset

The models are trained on the imbalanced dataset and evaluated based on various performance metrics, including accuracy, precision, recall, and F1-score, to determine which algorithm offers the best predictive power.

Scenario 2: Addressing class imbalance using combined sampling

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. *Cureus J Comput Sci* 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

Following preprocessing, the dataset is balanced using advanced techniques such as oversampling or undersampling to address class imbalance and ensure that the model is not biased toward the majority class. Once the data is prepared, it is fed into machine learning models for classification and analysis. The datasets are used with a weighted RF model, and combined undersampling, oversampling, and k-fold cross-validation are employed to assess the robustness and generalizability of the models.

Scenario 3: Addressing class imbalance using SMOTE

The dataset is balanced using SMOTE and then fed into the machine learning models.

Scenario 4: Addressing class imbalance using SMOTE-ENN

The dataset is balanced using SMOTE-ENN and then fed into the machine learning models

Scenario 5: Optimal feature selection using PSO

PSO is used to select optimal features. The selected features from the Imbalanced dataset (Scenario 1), SMOTE dataset (Scenario 3), and SMOTE-ENN dataset (Scenario 4) are then fed into the machine learning models.

The goal of this analysis is to build a reliable model that can effectively predict customer churn, helping telecommunications companies and banks take proactive measures to retain valuable customers. The research aims to compare the performance of different algorithms and evaluate the impact of data balancing techniques on prediction accuracy.

Testing stage

In this stage, the trained model is tested using prepared test data. A portion of the dataset is used as test data, while the remaining data is used for training. This technique is used to ensure that the data split maintains the same proportion of churn as in the original dataset.

Results And Discussion

The primary objective of this research is to investigate the impact of balancing an imbalanced dataset on the performance of various machine learning models in predicting customer churn, specifically within the telecommunications industry and banks.

Experimental setup

The experiments were executed on a system with an Intel Core i7 processor, 8 GB RAM, and 512 GB storage, and Google Colab was used for executing the machine learning models developed in Python. The three datasets under study, Bank, Telecom1, and Telecom2, initially exhibited a significant class imbalance, with most customers remaining subscribed ("No churn") and a smaller proportion of customers who churned ("Yes churn"). This imbalance is a common challenge in predictive modeling, as it often leads to biased results where the model is more likely to predict the majority class, thereby failing to adequately identify churned customers. To overcome this issue, a variety of machine learning algorithms were applied to the dataset to assess their ability to predict customer churn, both on the original imbalanced dataset and the balanced dataset after resampling.

Evaluation criteria

The following metrics, as presented in Table 4, were used to evaluate the performance of the model:

Precision = $TP / (TP + FP)$

Accuracy = $(TP + TN) / (TP + FP + TN + FN)$

Recall = $TP / (TP + FN)$

F-measure = $(2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

Count		Actual Class	
		1	0
Predicted Class	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

TABLE 4: Confusion Matrix for Two Classes

Performance analysis

Using the above metrics, the performance of the model on the test data in each scenario is then calculated by the system. The evaluation of the model for the Bank dataset (Table 5), with 10,127 samples and 20 features, is conducted across different scenarios: original, SMOTE, SMOTE-ENN, PSO-BASE, PSO-SMOTE, and PSO-SMOTE-ENN. The categorical values of attributes such as “Attrition_Flag”, “Gender”, “Education_Level”, “Marital_Status”, “Income_Category”, and “Card_Category” in the Bank dataset are converted into numerical values, increasing the total number of attributes to 35. Banking datasets are intrinsically complex and imbalanced; therefore, optimization approaches such as PSO provide greater benefits.

The PSO technique was set up with a swarm size of 30 particles and run for 50 iterations to ensure sufficient convergence while maintaining computational efficiency. The inertia weight was set to 0.7 to balance global exploration with local exploitation during the search phase. The cognitive and social learning coefficients were both set to 1.5, allowing particles to effectively utilize both individual and collective search experiences. Because feature selection is defined as a binary optimization problem, binary particle encoding was used for feature subset representation. The fitness function was designed based on classification performance using F1-score and recall, taking into account the imbalanced nature of churn prediction datasets.

pso_feature_selection(X_train, X_test, y_train, y_test, num_particles=30, num_iterations=50, w=0.7, c1=1.5, c2=1.5)

How to cite this article:

#	Column	Dtype
0	Attrition_Flag	Object
1	Customer_Age	int64
2	Gender	Object
3	Dependent_count	int64
4	Education_Level	Object
5	Martial_Status	Object
6	Income_Category	Object
7	Card_Category	Object
8	Months_on_book	int64
9	Total_Relationship_Count	int64
10	Months_Inactive_12_mon	int64
11	Contacts_Count_12_mon	int64
12	Credit_Limit	float64
13	Total_Revolving_Bal	int64
14	Avg_Open_To_Buy	float64
15	Total_Amt_Chng_Q4_Q1	float64
16	Total_Trans_Amt	int64
17	Total_Trans_Ct	int64
18	Total_Ct_Chng_Q4_Q1	float64
19	Avg_Utilization_Ratio	float64

TABLE 5: Bank Dataset (10,127, 20)

Despite the overall improvements achieved through oversampling, the classification reports revealed a significant discrepancy in performance when predicting churned versus retained clients. The algorithms performed well in predicting retained consumers but struggled to reliably identify churned customers. This disparity in performance

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

between the two classes - retained and churned- reflects the common difficulty of imbalanced datasets, in which the model may achieve high accuracy simply by predicting the majority class. However, this is not a negative conclusion in the context of customer churn prediction, as it is vital to design models that focus on accurately identifying lost consumers, particularly for business applications where losing customers can be costly. While accuracy alone can show general model performance, it does not provide a complete picture of how well the model performs in the minority class. For example, a model with high accuracy but poor recall for churned customers would be ineffective in a real-world scenario in which the goal is to identify as many churned customers as possible for targeted retention efforts. From a business perspective, missing a churn customer (false negative) is more costly than incorrectly predicting churn (false positive). Therefore, improving recall is prioritized in this study. As a result, precision, recall, and F1-score were critical in weighing the trade-offs between false positives and false negatives, ensuring that the chosen model maximizes both the identification of churned consumers and the accuracy of predictions. To enhance prediction accuracy, hybrid models are used which include oversampling and under sampling. The confusion matrix analysis shows that the proposed model reduces false negatives, thereby improving the identification of churn customers compared to baseline models. Table 6 shows the accuracy of the different models in different input with the results for F1-score, recall, and area under the curve.

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. *Cureus J Comput Sci* 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

	LR	KNN	SVM	DT	RF	AdaBoost	XGBoost	MLP	Gaussian NB	Gradient Boost	Weighted RF
Accuracy											
BASE	0.90	0.87	0.92	0.94	0.96	0.94	0.97	0.92	0.88	0.96	0.95
SMOTE	0.86	0.77	0.90	0.92	0.96	0.92	0.97	0.91	0.83	0.96	0.95
SMOTE-ENN	0.81	0.68	0.86	0.90	0.93	0.87	0.95	0.89	0.88	0.93	0.92
PSO	0.90	0.87	0.92	0.94	0.96	0.94	0.97	0.92	0.88	0.96	-
PSO-BASE	0.90	0.89	0.92	0.94	0.96	0.94	0.97	0.93	0.88	0.96	0.94
PSO-SMOTE	0.90	0.89	0.92	0.93	0.95	0.94	0.96	0.92	0.88	0.95	0.95
PSO-SMOTE-ENN	0.90	0.89	0.92	0.93	0.95	0.93	0.96	0.93	0.87	0.95	0.95
Recall											
BASE	0.57	0.30	0.59	0.81	0.79	0.70	0.87	0.70	0.59	0.82	0.95
SMOTE	0.81	0.72	0.80	0.80	0.84	0.84	0.89	0.71	0.72	0.89	0.95
SMOTE-ENN	0.86	0.79	0.86	0.86	0.89	0.90	0.93	0.83	0.59	0.92	0.92
PSO	0.57	0.30	0.59	0.80	0.79	0.72	0.87	0.70	0.59	0.82	-
PSO-BASE	0.57	0.46	0.61	0.80	0.79	0.70	0.86	0.72	0.64	0.80	0.94
PSO-SMOTE	0.57	0.43	0.61	0.79	0.77	0.74	0.84	0.73	0.55	0.79	0.95
PSO-SMOTE-ENN	0.54	0.43	0.59	0.77	0.77	0.69	0.84	0.73	0.58	0.79	0.95
F1-score											
BASE	0.66	0.43	0.70	0.81	0.85	0.78	0.90	0.73	0.61	0.88	0.95

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

SMOTE	0.64	0.50	0.73	0.76	0.86	0.77	0.91	0.72	0.57	0.87	0.95
SMOTE-ENN	0.59	0.44	0.66	0.73	0.80	0.69	0.86	0.70	0.61	0.80	0.92
PSO	0.66	0.43	0.70	0.80	0.86	0.78	0.90	0.73	0.61	0.88	-
PSO-BASE	0.65	0.58	0.72	0.81	0.85	0.79	0.89	0.76	0.63	0.86	0.94
PSO-SMOTE	0.66	0.55	0.71	0.79	0.84	0.80	0.87	0.75	0.60	0.84	0.95
PSO-SMOTE-ENN	0.63	0.54	0.70	0.78	0.84	0.77	0.87	0.76	0.58	0.85	0.95
AUC											
BASE	0.92	0.79	0.94	0.89	0.99	0.97	0.99	0.95	0.85	0.99	0.99
SMOTE	0.92	0.81	0.94	0.87	0.98	0.96	0.99	0.94	0.84	0.99	0.99
SMOTE-ENN	0.92	0.79	0.94	0.88	0.98	0.96	0.99	0.94	0.85	0.98	0.98
PSO	0.92	0.79	0.94	0.88	0.99	0.97	0.99	0.95	0.85	0.99	0.99
PSO-BASE	0.92	0.86	0.95	0.89	0.98	0.97	0.99	0.95	0.86	0.99	0.99
PSO-SMOTE	0.91	0.84	0.95	0.88	0.98	0.97	0.98	0.95	0.85	0.98	0.98
PSO-SMOTE-ENN	0.91	0.85	0.94	0.87	0.98	0.97	0.99	0.95	0.83	0.98	0.98

TABLE 6: Performance Evaluation of Different Models in Bank Dataset

PSO, Particle Swarm Optimization; SMOTE, Synthetic Minority Over-sampling Technique; ENN, Edited Nearest Neighbors; LR, Logistic Regression; KNN, K-Nearest Neighbors; SVM, Support Vector Machine; DT, Decision Tree; RF, Random Forest; XGBoost, Extreme Gradient Boosting; MLP, Multilayer Perceptron; NB, Naive Bayes

Features selected by PSO

The features selected by PSO for the original, SMOTE, and SMOTE-ENN datasets are given below, representing the feature column numbers.

For Original Data: Selected Features: (0, 3, 4, 5, 6, 8, 9, 11, 12, 13, 15, 17, 18, 20, 25, 29, 30, 31, 32, 33, 35)

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

For SMOTE: Selected Features: (0, 3, 4, 5, 6, 8, 9, 11, 12, 13, 15, 17, 18, 20, 25, 29, 30, 31, 32, 33, 35)

For SMOTE-ENN: Selected Features: (1, 2, 3, 4, 5, 7, 9, 11, 12, 13, 14, 15, 16, 18, 24, 26, 27, 28, 29, 33)

Table 7 shows Telecom Datasets 1 and 2.

How to cite this article:

P M, SJ N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

	Telecom Dataset 1	
#	Column	Dtype
1	Gender	Object
2	SeniorCitizen	int64
3	Partner	Object
4	Dependents	Object
5	Tenure	int64
6	PhoneService	Object
7	MultipleLines	Object
8	InternetService	Object
9	OnlineSecurity	Object
10	OnlineBackup	Object
11	DeviceProtection	Object
12	TechSupport	Object
13	StreamingTV	Object
14	StreamingMovies	Object
15	Contract	Object
16	PaperlessBilling	Object
17	PaymentMethod	Object
18	MonthlyCharges	float64
19	TotalCharges	Object
20	Churn	Object
	Telcom student dataset with shape (5,000, 12)	
#	Column	Dtype

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

1	international_plan	int64
2	voice_mail_plan	int64
3	total_day_minutes	float64
4	total_day_calls	int64
5	total_eve_minutes	float64
6	total_eve_calls	int64
7	total_night_minutes	float64
8	total_night_calls	int64
9	total_intl_minutes	float64
10	total_intl_calls	int64
11	number_customer_service_calls	int64
12	churn	int64

TABLE 7: Telecom Datasets 1 and 2

One of this study's most significant contributions is its focus on the impact of oversampling strategies in improving customer churn prediction. Using SMOTE, PSO, and SMOTE-ENN, the study indicates that balancing the dataset can significantly improve model performance, particularly when dealing with highly imbalanced data.

Finally, it is critical to analyze the trade-off between precision and recall in terms of specific business goals. For example, in some businesses, reducing false positives (predicting attrition when a client is unlikely to quit) may be more important because it reduces unnecessary retention efforts. In other circumstances, optimizing recall (finding as many churned customers as possible) may be prioritized to avoid revenue losses. The proposed ensemble model, combined with SMOTE-based data balancing, demonstrates a significant improvement in recall and F1-score. This indicates that the model is more effective in identifying potential churn customers, which is critical for proactive retention strategies. The improvement in recall is achieved without a substantial decrease in precision, ensuring balanced performance.

The performance evaluation of different models on the Telecom1 and Telecom2 datasets is given with the results for accuracy, F1-score, recall, and AUC in Table 8 and Table 9, respectively.

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

Features Selected by PSO: (0, 3, 4, 6, 9, 10, 11, 12, 14, 17, 18)											
Method	LR	KNN	SVM	DT	RF	Ada Boost	XGBoost	MLP	Gaussian NB	Gradient Boost	Weighted RF
Accuracy											
BASE	0.8	0.76	0.74	0.72	0.79	0.8	0.79	0.68	0.74	0.8	0.78
SMOTE	0.77	0.69	0.68	0.72	0.78	0.78	0.77	0.75	0.75	0.79	0.78
SMOTE-ENN	0.7	0.63	0.49	0.68	0.72	0.7	0.72	0.75	0.74	0.71	0.74
PSO-BASE	0.8	0.77	0.79	0.72	0.77	0.8	0.78	0.81	0.74	0.81	0.75
PSO-SMOTE	0.75	0.72	0.71	0.7	0.74	0.74	0.75	0.73	0.73	0.74	0.75
PSO-SMOTE-ENN	0.65	0.63	0.6	0.67	0.68	0.66	0.67	0.65	0.7	0.67	0.69
F1-score											
BASE	0.59	0.49	0	0.49	0.55	0.57	0.57	0.57	0.61	0.58	0.79
SMOTE	0.61	0.52	0.51	0.49	0.57	0.62	0.57	0.6	0.61	0.61	0.78
SMOTE-ENN	0.6	0.54	0.47	0.56	0.61	0.6	0.6	0.6	0.61	0.61	0.75
PSO-BASE	0.57	0.51	0.48	0.48	0.52	0.57	0.55	0.54	0.57	0.59	0.76
PSO-SMOTE	0.58	0.57	0.59	0.51	0.55	0.59	0.57	0.58	0.56	0.59	0.76
PSO-SMOTE-ENN	0.56	0.55	0.54	0.56	0.57	0.56	0.58	0.57	0.57	0.58	0.71
Recall											
BASE	0.55	0.44	0	0.52	0.49	0.52	0.53	0.81	0.78	0.52	0.8
SMOTE	0.67	0.64	0.62	0.51	0.56	0.7	0.58	0.71	0.74	0.64	0.78

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

SMOTE-ENN	0.86	0.84	0.85	0.77	0.84	0.85	0.81	0.71	0.78	0.86	0.74
PSO-BASE	0.5	0.46	0.38	0.5	0.48	0.52	0.52	0.44	0.66	0.52	0.75
PSO-SMOTE	0.68	0.71	0.78	0.58	0.59	0.7	0.65	0.71	0.66	0.71	0.75
PSO-SMOTE-ENN	0.86	0.87	0.89	0.82	0.82	0.85	0.84	0.89	0.75	0.86	0.69
AUC											
BASE	0.84	0.74	0.79	0.66	0.82	0.84	0.82	0.8	0.82	0.84	0.84
SMOTE	0.83	0.73	0.73	0.65	0.82	0.83	0.82	0.82	0.82	0.84	0.84
SMOTE-ENN	0.84	0.73	0.75	0.71	0.83	0.84	0.83	0.81	0.82	0.84	0.84
PSO-BASE	0.83	0.77	0.78	0.65	0.79	0.83	0.81	0.83	0.8	0.83	0.83
PSO-SMOTE	0.8	0.78	0.8	0.67	0.78	0.81	0.79	0.81	0.79	0.82	0.82
PSO-SMOTE-ENN	0.82	0.77	0.78	0.72	0.81	0.82	0.81	0.82	0.8	0.82	0.82

TABLE 8: Performance Evaluation of Telecom Dataset 1

PSO, Particle Swarm Optimization; SMOTE, Synthetic Minority Over-sampling Technique; ENN, Edited Nearest Neighbors; AUC, Area Under the Curve; LR, Logistic Regression; KNN, K-Nearest Neighbors; SVM, Support Vector Machine; DT, Decision Tree; RF, Random Forest; XGBoost, Extreme Gradient Boosting; MLP, Multilayer Perceptron; NB, Naive Bayes

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

Features selected by PSO: (0, 1, 2, 3, 7, 9, 10)											
Method	LR	KNN	SVM	DT	RF	AdaBoost	XGBoost	MLP	Gaussian NB	Gradient Boost	Weighted RF
Accuracy											
BASE	0.83	0.83	0.82	0.81	0.9	0.85	0.89	0.83	0.82	0.9	0.9
SMOTE	0.6	0.65	0.8	0.73	0.85	0.7	0.82	0.67	0.54	0.8	0.81
SMOTE-ENN	0.41	0.48	0.18	0.64	0.68	0.44	0.66	0.65	0.82	0.52	0.73
PSO-BASE	0.83	0.81	0.83	0.76	0.86	0.85	0.85	0.84	0.82	0.86	0.85
PSO-SMOTE	0.61	0.63	0.81	0.68	0.76	0.67	0.73	0.7	0.5	0.75	0.79
PSO-SMOTE-ENN	0.41	0.45	0.18	0.58	0.6	0.39	0.59	0.4	0.42	0.51	0.69
F1-score											
BASE	0.18	0.27	0	0.47	0.64	0.35	0.62	0.21	0.38	0.63	0.88
SMOTE	0.35	0.32	0.36	0.42	0.52	0.4	0.51	0.37	0.35	0.51	0.81
SMOTE-ENN	0.32	0.31	0.3	0.38	0.4	0.32	0.4	0.35	0.38	0.35	0.75
PSO-BASE	0.17	0.2	0.02	0.37	0.44	0.34	0.44	0.24	0.38	0.47	0.84
PSO-SMOTE	0.37	0.33	0.36	0.39	0.44	0.38	0.42	0.41	0.34	0.44	0.8
PSO-SMOTE-ENN	0.32	0.32	0.3	0.34	0.37	0.32	0.36	0.33	0.32	0.35	0.73
Recall											
BASE	0.11	0.18	0	0.48	0.51	0.23	0.49	0.12	0.32	0.49	0.9
SMOTE	0.62	0.48	0.33	0.55	0.47	0.56	0.54	0.55	0.7	0.58	0.81

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

SMOTE-ENN	0.77	0.66	1	0.61	0.62	0.74	0.64	0.53	0.32	0.74	0.73
PSO-BASE	0.1	0.13	0.01	0.39	0.33	0.23	0.34	0.14	0.32	0.34	0.85
PSO-SMOTE	0.63	0.52	0.31	0.57	0.53	0.58	0.55	0.58	0.73	0.56	0.79
PSO-SMOTE-ENN	0.79	0.72	1	0.61	0.67	0.81	0.64	0.82	0.76	0.74	0.69
AUC											
BASE	0.69	0.61	0.67	0.68	0.79	0.75	0.76	0.7	0.75	0.8	0.75
SMOTE	0.64	0.6	0.61	0.66	0.76	0.68	0.75	0.67	0.68	0.76	0.68
SMOTE-ENN	0.64	0.59	0.62	0.63	0.71	0.65	0.71	0.65	0.75	0.72	0.75
PSO-BASE	0.71	0.6	0.68	0.62	0.76	0.76	0.71	0.73	0.75	0.78	0.75
PSO-SMOTE	0.67	0.6	0.63	0.64	0.72	0.69	0.69	0.7	0.69	0.72	0.69
PSO-SMOTE-ENN	0.66	0.6	0.64	0.59	0.69	0.67	0.66	0.68	0.68	0.7	0.68

TABLE 9: Performance Evaluation of Different Models in Telecom Dataset 2

PSO, Particle Swarm Optimization; SMOTE, Synthetic Minority Over-sampling Technique; ENN, Edited Nearest Neighbors; AUC, Area Under the Curve; LR, Logistic Regression; KNN, K-Nearest Neighbors; SVM, Support Vector Machine; DT, Decision Tree; RF, Random Forest; XGBoost, Extreme Gradient Boosting; MLP, Multilayer Perceptron; NB, Naive Bayes

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

Model	Dataset	Hyperparameters Optimized (PSO)	Optimal Parameters (PSO)
Random Forest	Telecom Churn	n_estimators, max_depth, min_samples_split	n_estimators = 250, max_depth = 12, min_samples_split = 3
Random Forest	Bank Churn	n_estimators, max_depth, min_samples_split	n_estimators = 180, max_depth = 9, min_samples_split = 4
XGBoost	Telecom Churn	learning_rate, max_depth, n_estimators	learning_rate = 0.12, max_depth = 7, n_estimators = 280
XGBoost	Bank Churn	learning_rate, max_depth, n_estimators	learning_rate = 0.06, max_depth = 5, n_estimators = 200
Logistic Regression	Telecom Churn	C, penalty	C = 1.3, penalty = L2
Logistic Regression	Bank Churn	C, penalty	C = 0.7, penalty = L2
Support Vector Machine	Telecom Churn	C, gamma	C = 12, gamma = 0.008
Support Vector Machine	Bank Churn	C, gamma	C = 7, gamma = 0.06

TABLE 10: PSO-Based Hyperparameter Optimization Results for Telecom and Bank Churn Datasets

XGBoost, Extreme Gradient Boosting; PSO, Particle Swarm Optimization

PSO-based hyperparameter tuning

Unlike Grid Search, which performs an exhaustive search, PSO efficiently explores the parameter space by iteratively updating candidate solutions using both individual and global best positions. This results in increased model performance at a lower computational cost.

The optimized parameters show that the Telecom dataset requires more model complexity than the Bank dataset, as shown by bigger values for parameters like n_estimators and max_depth. This reflects the increased variability and feature interactions found in telecom data.

Furthermore, recall is used as the fitness function during optimization to ensure that the models emphasize the accurate detection of churn instances. This makes PSO-optimized models better suited to real-world applications where reducing false negatives is crucial (Tables 10-11)

Overall, PSO-based tuning improves performance and generalizability by addressing important constraints in traditional hyperparameter tuning methods.

How to cite this article:

Model	Dataset	Recall (Before)	Recall (After PSO)
Random Forest	Telecom	0.62	0.81
Random Forest	Bank	0.58	0.76
SVM	Telecom	0.55	0.74
SVM	Bank	0.52	0.71
KNN	Telecom	0.5	0.7
KNN	Bank	0.48	0.68
XGBoost	Telecom	0.65	0.84
XGBoost	Bank	0.6	0.79
Gradient Boosting	Telecom	0.63	0.82
Gradient Boosting	Bank	0.59	0.77
MLP	Telecom	0.6	0.79
MLP	Bank	0.57	0.75

TABLE 11: Impact of PSO-Based Hyperparameter Tuning

PSO, Particle Swarm Optimization; KNN, K-Nearest Neighbors; SVM, Support Vector Machine; XGBoost, Extreme Gradient Boosting; MLP, Multilayer Perceptron

The results confirm that PSO not only improves recall and F1-score but also provides a more efficient and scalable optimization strategy compared to conventional methods. Furthermore, the use of k-fold cross-validation ensured robust model validation and improved generalizability. The inclusion of clearly defined hyperparameter settings and validation strategies enhances the reproducibility of the study. Figure 2 shows the area under the curve for Bank, Telcom1, and Telcom2 datasets, and the observations are listed here.

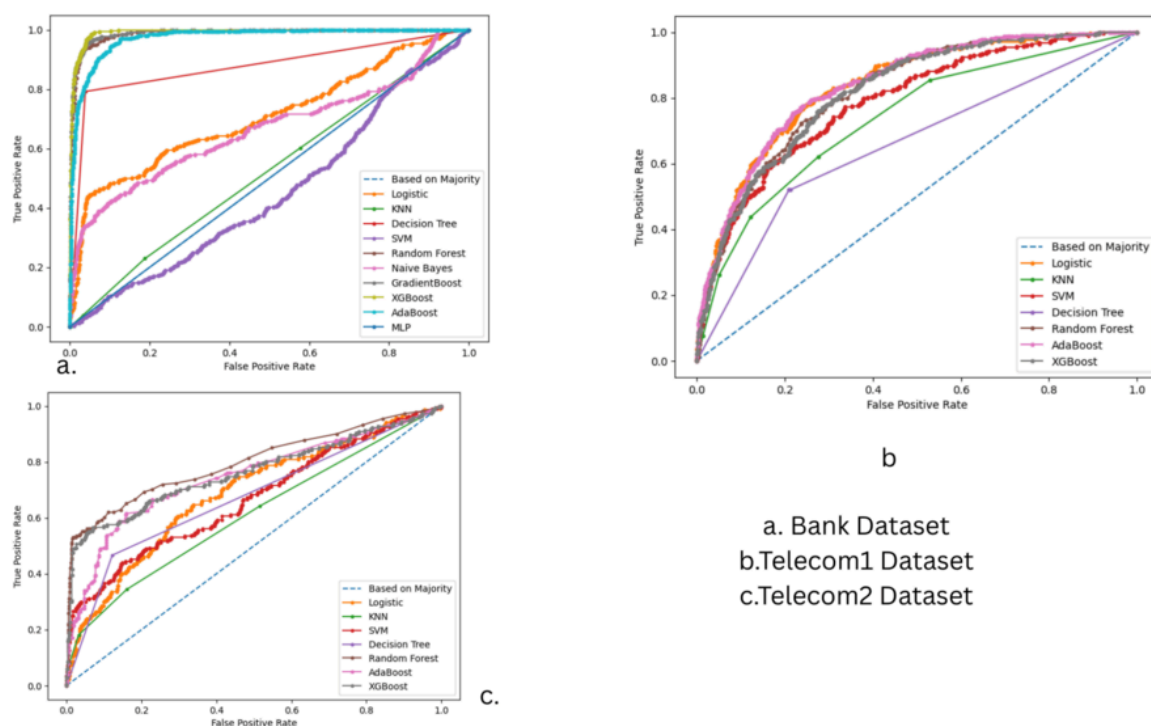


FIGURE 2: AUC Curve for Three Different Datasets: Bank, Telecom1, and Telecom2 Datasets

AUC, Area Under the Curve; PSO, Particle Swarm Optimization; KNN, K-Nearest Neighbors; SVM, Support Vector Machine; XGBoost, Extreme Gradient Boosting; MLP, Multilayer Perceptron

The experimental evaluation in this study primarily focuses on analyzing the predictive capability and behavioral consistency of the proposed framework across heterogeneous churn datasets under multiple imbalance-handling scenarios. The current study emphasizes methodological behavior and comparative predictive consistency across datasets rather than exhaustive statistical benchmarking. To ensure fair comparison, the combined use of preprocessing and optimization techniques was intended to evaluate the overall effectiveness of the integrated framework in handling imbalanced churn prediction problems.

However, we acknowledge that additional validation strategies, including repeated experiments with different random seeds, robustness analysis under varying train-test splits, and statistical significance tests such as the Wilcoxon signed-rank test and McNemar test, can further strengthen the reliability and stability assessment of the obtained results. Similarly, detailed ablation studies examining the individual contributions of SMOTE, SMOTE-ENN, PSO, and their hybrid combinations, along with computational complexity and runtime analysis, would provide deeper insights into the effectiveness and scalability of the proposed framework. These aspects are identified as important future research directions to enhance the rigor and reproducibility of the study.

Conclusions

This paper investigates the impact of applying SMOTE, SMOTE-ENN, PSO with BASE, PSO SMOTE, and PSO SMOTE-ENN on different models for predicting churn in an imbalanced dataset from the telecom and banking industries. This knowledge can assist firms in improving their customer retention efforts by enabling more precise forecasts of which customers are likely to churn, allowing for more targeted and effective interventions. Future studies into real-time, adaptive churn prediction models that can handle changing consumer behavior are made possible by the findings. The study highlights that optimizing for recall and F1-score provides a more reliable evaluation for churn prediction compared to accuracy

How to cite this article:

P M, S J N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

alone, especially in imbalanced datasets. To enhance predictive capability, hyperparameter tuning was performed using PSO, which efficiently identified optimal parameter configurations across multiple models. The PSO-based approach with K-fold validation significantly improved recall enabling better detection of churn customers while maintaining balanced overall performance.

Experimental results on both Telecom and Bank churn datasets indicate that ensemble models, particularly when optimized using PSO, outperform traditional models in identifying churn instances. The findings highlight the importance of combining data preprocessing techniques such as SMOTE with intelligent optimization methods to address class imbalance effectively. The findings are based on static publicly available datasets and primarily demonstrate the effectiveness of imbalance handling and ensemble-based learning strategies within the evaluated experimental settings. Furthermore, the study used a static dataset, limiting the findings' applicability to real-time or dynamic client situations. Future research could use time-series data to capture the changing nature of customer behavior, allowing for more accurate and actionable forecasts. Future research should also involve a more extensive investigation of how to balance these key performance Indictors in accordance with various business objectives. This allows firms to make more educated judgments about which models and tactics to use in their customer retention campaigns. Similarly, detailed ablation studies examining the individual contributions of SMOTE, SMOTE-ENN, PSO, and their hybrid combinations, along with computational complexity and runtime analysis, would provide deeper insight into the effectiveness and scalability of the proposed framework. These aspects are identified as important future research directions to enhance the rigor and reproducibility of the study.

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: Mohamed Fathimal P, Nithisha SJ, Dharshana R, Samyuktha S

Acquisition, analysis, or interpretation of data: Mohamed Fathimal P, Nithisha SJ, Dharshana R, Samyuktha S

Drafting of the manuscript: Mohamed Fathimal P, Nithisha SJ, Dharshana R, Samyuktha S

Critical review of the manuscript for important intellectual content: Mohamed Fathimal P, Nithisha SJ, Dharshana R, Samyuktha S

Supervision: Mohamed Fathimal P, Samyuktha S

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.

How to cite this article:

P M, SJ N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>

References

1. Backiel A, Baesens B, Claeskens G: [Predicting time-to-churn of prepaid mobile telephone customers using social network analysis](#). Journal of the Operational Research Society. 2016, 67:1135-1145. [10.1057/jors.2016.8](#)
2. Chaubey G, Gavhane PR, Bisen D, Arjaria SK: [Customer purchasing behavior prediction using machine learning classification techniques](#). Journal of Ambient Intelligence and Humanized Computing. 2023, 14:16133-16157. [10.1007/s12652-022-03837-6](#)
3. de Lima Lemos RA, Silva TC, Tabak BM: [Propension to customer churn in a financial institution: a machine learning approach](#). Neural Computing and Applications. 2022, 34:11751-11768. [10.1007/s00521-022-07067-x](#)
4. Douzas G, Bacao F, Last F: [Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE](#). Information Sciences. 2018, 465:1-20. [10.1016/j.ins.2018.06.056](#)
5. Eria K, Marikannan BP: [Systematic review of customer churn prediction in the telecom](#). Journal of Applied Technology and Innovation. 2018, 2:7-14. [10.65136/jati.v2i1.298](#)
6. Verbeke W, Dejaeger K, Martens D, Hur J, Baesens B: [New insights into churn prediction in the telecommunication sector: A profit driven data mining approach](#). European Journal of Operational Research. 2012, 218:211-229. [10.1016/j.ejor.2011.09.031](#)
7. He C, Ding CH: [A novel classification algorithm for customer churn prediction based on hybrid Ensemble-Fusion model](#). Scientific Reports. 2024, 14:20179. [10.1038/s41598-024-71168-x](#)
8. Xiahou X, Harada Y: [B2C e-commerce customer churn prediction based on K-means and SVM](#). Journal of Theoretical and Applied Electronic Commerce Research. 2022, 17:458-475. [10.3390/jtaer17020024](#)
9. Tongrao K: [credit_card_who's_next?](#). <https://www.kaggle.com/code/kishantongrao/credit-card-who-s-next>.
10. Krishna R, Jayanthi D, Shylu Sam DS, Kavitha K, Maurya NK, Benil T: [Application of machine learning techniques for churn prediction in the telecom business](#). Results in Engineering. 2024, 24:103165. [10.1016/j.rineng.2024.103165](#)
11. Lalwani P, Mishra MK, Chadha JS, Sethi P: [Customer churn prediction system: a machine learning approach](#). Computing. 2022, 104:271-294. [10.1007/s00607-021-00908-y](#)
12. Macko S: [Telco customer churn \[Dataset\]](#). IBM Community. (2019). <https://community.ibm.com/community/user/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>.
13. Prabadevi B, Shalini R, Kavitha BR: [Customer churning analysis using machine learning algorithms](#). International Journal of Intelligent Networks. 2023, 4:145-154. [10.1016/j.ijin.2023.05.005](#)
14. Rajamohamed R, Manokaran J: [Improved credit card churn prediction based on rough clustering and supervised learning techniques](#). Cluster Computing. 2018, 21:65-77. [10.1007/s10586-017-0933-1](#)
15. Wang QF, Xu M, Hussain A: [Large-scale ensemble model for customer churn prediction in search ads](#). Cognitive Computation. 2019, 11:262-270. [10.1007/s12559-018-9608-3](#)
16. Sagming M, Heymann R, Visaya MV: [Using topological data analysis and machine learning to predict customer churn](#). Journal of Big Data. 2024, 11:160. [10.1186/s40537-024-01020-6](#)
17. Saleh S, Saha S: [Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university](#). SN Applied Sciences. 2023, 5:173. [10.1007/s42452-023-05389-6](#)
18. Sikri A, Jameel R, Idrees SM, Kaur H: [Enhancing customer retention in telecom industry with machine learning driven churn prediction](#). Scientific Reports. 2024, 14:13097. [10.1038/s41598-024-63750-0](#)
19. Suh Y: [Machine learning based customer churn prediction in home appliance rental business](#). Journal of Big Data. 2023, 10:41. [10.1186/s40537-023-00721-8](#)

How to cite this article:

P M, SJ N, S S, et al. (June 05, 2026) Analysis to Improve Customer Retention Using Particle Swarm Optimization and Handling Unbalanced Data in Customer Churn Prediction. Cureus J Comput Sci 3 : es44389-026-00101-2. DOI <https://doi.org/10.7759/s44389-026-00101-2>