Financial Innovation

RESEARCH Open Access

A framework to improve churn prediction performance in retail banking

João B. G. Brito[1*], Guilherme B. Bucco[2], Rodrigo Heldt[2], João L. Becker[3], Cleo S. Silveira[2], Fernando B. Luce[2] and Michel J. Anzanello[1]

*Correspondence: joao.brito@ufrgs.br; jbgb@uol.com.br

[1] Departamento de Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Av. Osvaldo Aranha, 99 – 5° Andar, Porto Alegre, RS 90035-190, Brazil
[2] Escola de Administração, Universidade Federal do Rio Grande do Sul, Washington Luiz, 855, Porto Alegre, RS 90010-460, Brazil
[3] Escola de Administração de Empresas de São Paulo, Fundação Getulio Vargas, Av. 9 de julho, 2029, Bela Vista, São Paulo, SP 01313-902, Brazil

Abstract

Managing customer retention is critical to a company's profitability and firm value. However, predicting customer churn is challenging. The extant research on the topic mainly focuses on the type of model developed to predict churn, devoting little or no effort to data preparation methods. These methods directly impact the identification of patterns, increasing the model's predictive performance. We addressed this problem by (1) employing feature engineering methods to generate a set of potential predictor features suitable for the banking industry and (2) preprocessing the majority and minority classes to improve the learning of the classification model pattern. The framework encompasses state-of-the-art data preprocessing methods: (1) feature engineering with recency, frequency, and monetary value concepts to address the imbalanced dataset issue, (2) oversampling using the adaptive synthetic sampling algorithm, and (3) undersampling using NEASMISS algorithm. After data preprocessing, we use XGBoost and elastic net methods for churn prediction. We validated the proposed framework with a dataset of more than 3 million customers and about 170 million transactions. The framework outperformed alternative methods reported in the literature in terms of precision-recall area under curve, accuracy, recall, and specificity. From a practical perspective, the framework provides managers with valuable information to predict customer churn and develop strategies for customer retention in the banking industry.

Keywords: Customer churn prediction, Imbalanced dataset treatment, Feature engineering, RFM

Introduction

The strategy of traditional banks has always been centered on their products and services, prioritizing internal practices and processes and considering customers as the commercial target (Lähteenmäki and Nätti 2013). However, the advancement of computing power and the reduction in its costs have led to a significant transformation in the industry by rapidly implementing new and highly competitive financial products and services (Feyen et al. 2021). Competitiveness has grown sharply as new financial technologies (e.g., FinTech) have risen (Murinde et al. 2022). Fintech is recognized as one of the most significant technological innovations in the financial sector and is characterized by rapid development (Kou et al. 2021a). Digital innovations coupled with abundant data have made it possible to devise new businesses and financial services, placing customers at the center of marketing decisions (Pousttchi and Dehnert 2018). In this scenario, managing customer retention is critical to avoid the defection of valuable customers (Mutanen et al. 2010). Reichheld and Sasser (1990) pointed out that reducing churn by 5% might increase a bank's profits from those customers by 85%.

Churn prediction is key in churn management programs (Ascarza 2018). Thus, customer churn prevention is a strategic issue that allows companies to monitor customer engagement behavior and act when engagement decreases (Gordini and Veglio 2017). Additionally, as the costs of acquiring new customers are typically higher than retaining existing ones, developing reliable strategies for churn management is critical for the sustainability of companies (Lemmens and Gupta 2020). However, managing customer churn is a challenging task for marketing managers (Ascarza and Hardie 2013).

Therefore, predictive models tailored to identify potential churners are fundamental in supporting managerial decisions (Zhao et al. 2022). Several models have been proposed to predict churn (Benoit and Poel 2012; He et al. 2014; Gordini and Veglio 2017), estimate the churn probability of each customer, and assess the accuracy of these predictions in the holdout sample. However, most of these methods focus on the type and mathematical properties of the model used to predict churn. While this concern is legitimate and important, less effort is devoted to data preprocessing, which comprises a set of tasks performed before training the models to improve the predictive performance (Jassim and Abdulwahid 2021). Furthermore, one of the main challenges of customer churn prediction (CCP) is training binary classification models when the churn event is rare (Zhu et al. 2018), as is the case of the retail banking industry (see Mutanen et al. 2010 and Keramati et al. 2016). The low proportion of the rare event compared with the majority class (e.g., retained clients) can create pattern gaps, making it hard for predictive models to identify a customer who is likely to churn. In extreme cases, a classification model may classify all customers as retained, mispredicting churns (Sun et al. 2009; Megahed et al. 2021). For instance, suppose there is a company with 98% of retained customers and 2% of churned customers and we want to predict churners. In this case, if we directly apply a commonly used learning algorithm, such as random forest or logistic regression, the outcome will probably be a high accuracy for the majority class (recall measure) and a poor specificity measure for the rare class.

Therefore, proposing strategies to deal with the problem of highly imbalanced datasets is a relevant topic. According to Megahed et al. (2021), a challenging task is to increase the number of instances of the rare class by collecting additional samples from the real-world process, so rebalancing methods are important. These methods are part of the broad concept of data preprocessing, which is a set of strategies carried out on raw data before applying a learning algorithm. Recent studies aim to increase CCP performance by testing multiple alternative techniques along the predictive modeling process, including outliers treatment, missing values imputation (MVI), feature engineering (FE), imbalanced dataset treatment (IDT), feature selection (FS), hyperparameters optimization, and testing different classification models. Although such approaches address the challenges of modeling churn behavior, they do not deal with scenarios where the churn event is rare (Zhu et al. 2018). Therefore, data analysts should better understand the data preprocessing idea to ensure that data manipulation does not change the underlying data structure and favors the classification technique applied to such data (Megahed et al. 2021).

This study proposes a novel framework for CCP in the banking industry, where rare churn events are persistent (Gür Ali and Arıtürk 2014). Rarity tends to jeopardize the performance of traditional techniques aimed at binary classification. Our objectives are as follows:

(1) Propose and validate a data preprocessing phase that combines different approaches for data preprocessing (FE focused on the retail banking context, IDT oversampling (IDT-over), and IDT undersampling (IDT-under));
(2) Perform and validate a model training and testing phase with state-of-the-art classification techniques (XGBoost and elastic net) while using Bayesian approaches for hyperparameter optimization;
(3) Evaluate and discuss the impact of different data preprocessing approaches to improve the predictive power of classification models.

Our study makes a valuable contribution to the research on decision support systems. It sheds light on the importance of conducting data preprocessing steps in scenarios where a substantial class imbalance is observed. To address this issue, rigorous tests were applied, confirming the practical value of the model. The deliverables include the following:

(1) A sequence of steps applied to binary classification problems characterized by a highly imbalanced class;
(2) A set of features created with superior capacity to predict churn in the banking industry, using recency, frequency, and monetary value concepts (RFM) in the FE stage;
(3) A framework with high predictive performance to anticipate churn events;
(4) An evaluation of feature importance to assist managers in designing more effective measures to prevent churn.

The rest of the paper is organized as follows. "Literature review" section reviews the literature; "Proposed framework for CCP" section describes our framework; "Data and empirical context" section presents the tests with the proposed framework; "Results and discussion" section discusses the results; and, finally, "Conclusions" section concludes.

Literature review

In this section, we present a literature review on the fundamentals of the techniques employed in the proposed framework.

CCP

Financial institutions have been challenged to develop new methods for mining and interpreting customers' data to help them understand their needs (Broby 2021). Broby (2021) also reported that the advancement of technology has reduced information asymmetry between banks and their clients, boosting competition and making customer retention management increasingly important. Similarly, Livne et al. (2011) examined the link between customer relationships and financial performance at the firm level in the context of US and Canadian wireless (cellular) firms. They found a positive association between customer retention and future revenues. They also suggested that customer retention and the level of service usage play an important mediating role in the relationship between investments in customer acquisition and financial performance. These results indicate the importance of generating and retaining customer loyalty to drive financial performance.

Given the relevance of retaining customers, churn prevention practices have gained significance. The development of churn prediction models is an effective alternative to potentiate customer retention efforts. According to Broby (2022), statistical and computational models of machine learning such as CCP models are increasingly being integrated into decision support systems in the financial area. Thus, many studies have been conducted to determine effective classification models for churn prediction, especially when using data from financial sources (Lahmiri et al. 2020). However, modeling financial data is complicated due to the presence of multiple latent factors that can evolve and are usually correlated and even autocorrelated (Li et al. 2022). Table 1 presents some studies that focused on the banking industry to exemplify the methods commonly used to preprocess the dataset and the classification models adopted. In the table, we also compare our study with the other studies.

The extant studies typically focused on comparing the performance of predictive models using different fundamentals and the managerial benefits of churn prediction on customer management. However, data preprocessing methods and approaches for hyperparameter optimization are scarce in the CCP literature. Usually, authors that described data preprocessing adopted traditional practices already established in previous studies, including random undersampling (RU) (Benoit and Poel 2012; He et al. 2014; Gordini and Veglio 2017), MVI (Lemmens and Croux 2006), and outlier detection and elimination (Zhao and Dang 2008; Keramati et al. 2016). Most studies, except that by Geiler et al. (2022), have not analyzed how predictive performance can benefit from employing more robust data preprocessing techniques, especially when a substantial imbalance between classes is verified (a common issue in CCP).

There are noticeable differences between previous studies and our propositions. To the best of our knowledge, our study is the first to propose a complete framework for churn prediction that encompasses a sequence of preprocessing steps prior to the classification task. Additionally, our framework relies on more sophisticated algorithms for IDT-over and IDT-under steps, enabling a performance gain due to the synergy between these steps. Our study also differs from that of Lemmens and Croux (2006), He et al. (2014), Farquad et al. (2014) and Geiler et al. (2022) in terms of the distribution of the features—both IDT-over and IDT-under did not modify the sample distribution significantly (as proved by a KS test) because the number of added artificial instances was controlled. This is desirable because it will not change the relationship between features. Another interesting difference is that our framework did not devote a step to FS as seen in the studies by Farquad et al. (2014) and Keramati et al. (2016) because the classification techniques employed to perform this task use

Table 1 Previous research assessing churn prediction models (specifications of our study are presented in the bottom line for comparison purposes)

Authors	Context	Data preprocessing stages	Predictive models	Dataset size
Lemmens and Croux (2006)	Telecom	MVI IDT-Over IDT-Under	Logistic regression, bagging, and stochastic gradient boosting	Datasets 1 and 2: 51,306 customers Dataset 3: 100,462 customers
Xie et al. (2009)	Banking	OT	Support Vector Machines	2,382 customers
Zhao and Dang (2008)	Banking	–	Random forests, neural networks, decision trees, and Support Vector Machines	1,524 customers
Benoit and Poel (2012)	Banking	–	Random forest	244,787 customers
Huang et al. (2012)	Telecom	FE	Logistic regression, Naive Bayes, linear classification, C4.5, neural networks, Support Vector Machines, and data mining by evolutionary learning (DMEL)	827,124 customers
Farquad et al. (2014)	Banking	IDT-Over IDT-Under FS	Support Vector Machines, Naive Bayes trees	14,814 customers
He et al. (2014)	Banking	IDT-Over IDT-Under	Logistic regression and Support Vector Machines	46,406 customers
Datta et al. (2015)	TV subscription	MVI FE	Binomial probit model	16,512 customers
Keramati et al. (2016)	Banking	OT MVI IDT-Over FS	Decision trees	4,383 customers
Geiler et al. (2022)	Banking and others	IDT-Over IDT-Under	KNN, Logistic regression, Naive Bayes, Support Vector Machines, Decision trees, neural networks, Random Forest, XGBoost	16 datasets (average of 108,473 customers)
Tékouabou et al. (2022)	Banking	MVI FS	Chandy-Misra-Bryant	45,000 customers
Our study	Banking	MVI FE IDT-Over IDT-Under	XGBoost and Elastic Net	3,283,332 customers

an embedded approach. Finally, the FE step depends on managers' ability to analyze the context to create meaningful predictive features as a weak aspect of the proposed framework.

Data preprocessing methods

For Sammut and Webb (2010), data preprocessing aims at transforming raw data into useful information. The predictive performance of a modeling framework increases by adopting data preprocessing stages before proceeding to the model training. García et al. (2014) characterized this phase as performing data selection and cleaning tasks.

Additionally, for Pyle (1999), this phase comprises more than 80% of the modeling effort, encompassing activities beyond cleaning and selection. In our study, we adopted three preprocessing methods—FE, IDT-over, and IDT-under—which are described in more detail in the following sections.

FE

FE is performed to transform the raw data into a new format that favors modeling and contributes to performance gains (Khoh et al. 2023). It involves the initial discovery of features and their stepwise improvement based on domain knowledge to adjust the data representation (Kuhn and Johnson 2019). Mathematical operations, interactions between attributes, matrix decomposition, discretization, and binarization are typical FE operations performed regardless of the business context. However, based on content assumptions, we can create new features using several practices, e.g., business context and customer behavior (Zheng and Casari 2018). According to Ascarza et al. (2018), customer retention involves the sustained continuity of customer transactions by a company. In this sense, RFMs are constituted by transactional data and can enhance prediction performance by expressing customer behavior and enabling churn predictors (Fader et al. 2005). Similarly, Kou et al. (2021b) used transaction data to enhance bankruptcy prediction in the banking industry.

IDT-over

In binary classification modeling, it is common to have a rare class (Megahed et al. 2021). In extreme cases, a class rarity does not convey a sufficiently delimited decision boundary between two groups (Weiss 2004). Classification problems such as bank fraud detection, infrequent medical diagnoses, and oil spill recognition in satellite images are well-known examples of rare classes (Galar et al. 2012). The imbalance between categories can make predictive classification modeling unfeasible. IDT-over methods try to circumvent this by generating synthetic instances that reinforce the boundary between classes (Triguero et al. 2012).

Researchers have used IDT-over methods in two ways—to create a new representation of the data (dismissing the original data completely) and to oversample (creating artificial instances of the rare class) (Sun et al. 2009). One popular oversampling algorithm is the synthetic minority oversampling technique (SMOTE), which assembles synthetic examples of the infrequent category (Fernandez et al. 2018). With similar purposes, the ADASYN algorithm searches for cases that are more difficult to discriminate between classes and generates synthetic instances to reinforce them (He et al. 2008). The algorithm uses a k-nearest neighbor (KNN) algorithm to add new artificial examples related to the minority class, making it more distinct from the majority class. The objective of ADASYN is to identify a weighted distribution of the rare class, considering its learning difficulty (He et al. 2008). It uses rare class instances very close to the majority class to generate other cases, reinforcing points in the boundary between the two groups. Recent studies have demonstrated that ADASYN outperforms SMOTE (Dey and Pratap 2023).

These findings highlight ADASYN as a robust oversampling technique to handle imbalanced data scenarios.

IDT-under

IDT-under methods remove instances from the majority class to balance the groups (Fernandez et al. 2018). The techniques range from simple RU to more refined algorithms. However, procedures such as RU do not guarantee the retainment of instances that maximize the identification of patterns, failing to select significant examples (He and Ma 2013). Therefore, Lin et al. (2017) proposed the CLUS algorithm to maximize the heterogeneity of the majority class instances by reducing redundancies. Similarly, Zhang and Mani (2003) proposed the NEARMISS algorithm.

The NEARMISS algorithm uses a KNN-based method to identify and remove instances with noisy patterns or those already represented in the data (redundant) (Zhang and Mani 2003). The procedure selects majority class instances with large average distances to the KNNs, keeping the majority class instances at the decision boundary (Bafna et al. 2023).

Classification algorithms

Researchers have developed several classification algorithms to predict a churn event. As either a customer will continue to be a customer or will churn, binary classification models are a promising alternative to address this problem. We now present the fundamentals of the two state-of-the-art classification techniques we test in our framework—XGBoost and elastic net.

XGBoost

The XGBoost is a tree-based ensemble method under the gradient-boosting decision tree framework developed and implemented in the R programming language in the package "xgboost" (Chen et al. 2022). This method uses an ensemble of classification and regression trees (CARTs) to fit the training data samples, utilizing residuals to calibrate a previous model at each iteration toward optimizing the loss function with a performance metric [e.g., precision-recall area under curve (PR-AUC)] (Chen and Guestrin 2016). The XGBoost adds one CART when it identifies a subset of hard-to-classify instances. To avoid overfitting in the calibration process, XGBoost adds a regularization term into the objective function, controlling the complexity of the model. It also combines first- and second-order gradient statistics to approximate the loss function before optimization (Chen and Guestrin 2016). Equation 1 presents the objective process of the XGBoost model at each iteration.

$$J(f_t) \cong \sum_{i=1}^{n} \left[L\left(y_i, \hat{y}_i^{t-1}\right) + g_i f_t\left(\vec{x}_i\right) + \frac{1}{2} h_i f_t^2\left(\vec{x}_i\right) \right] + \Omega(f_t) \tag{1}$$

where $\vec{x}_i$ represents the ith instance in the training set ($\vec{x}_i \in R^m$, where m is the number of features); y_i denotes the ith observed instance in the data set ($y_i \in R$); $\hat{y}_i^t$ symbolizes the prediction of the ith instance at the tth iteration; f_t is the CART added in the

tth iteration; g_i and h_i are the first and second derivatives of the loss function L, respectively; and Ω characterizes the regularization term described in Eq. 2 (Chen and Guestrin 2016):

$$\Omega(f_t) = \gamma T_t + \frac{\lambda}{2}\sum\nolimits_{j=1}^{T} {\omega_j}^2 \tag{2}$$

here γ and λ are constants controlling the regularization process; T denotes the number of leaves in the tree; and ω_j is the score of each leaf. Representing the structure of a CART f as a function $q : R^m \to \{1, 2, 3, \ldots, T\}$ mapping an observed data instance to the corresponding leaf index, we derive $f\left(\overrightarrow{x}_i\right) = \omega_{q(\overrightarrow{x}_i)}$, with $\omega_{q(\overrightarrow{x}_i)} \in R^T$. Next, plugging Eq. 2 into Eq. 1, removing the constant terms $L(y_i, {\widehat{y}_i}^{t-1})$ that do not impact the optimization process, and defining $I_j = \left\{q\left(\overrightarrow{x}_i\right) = j\right\}$ as the instance set of leaf j, we can rewrite the objective function at each iteration as Eq. 3 (Chen and Guestrin 2016).

$$J(f_t) \cong \sum\nolimits_{j=1}^{T} \omega_j \left[\sum\nolimits_{i \in I_j} g_i + \frac{\omega_j}{2}\left(\lambda + \sum\nolimits_{i \in I_j} h_i\right)\right] + \gamma T \tag{3}$$

For a fixed tree structure q, the optimal weight of leaf j (${\omega_j}^*$) is obtained simply by deriving Eq. 3 with respect to ω_j and equating it to zero, leading to Eq. 4 (Chen and Guestrin 2016).

$${\omega_j}^* = -\frac{\sum_{i \in I_j} g_i}{\lambda + \sum_{i \in I_j} h_i} \tag{4}$$

The optimal value of the objective function is Eq. 5 (Chen and Guestrin 2016).

$$J^*(q_t) = -\frac{1}{2}\sum\nolimits_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\lambda + \sum_{i \in I_j} h_i} + \gamma T \tag{5}$$

Equation 6 is used as a scoring function to gage the tree structure quality. High gain scores denote tree structures that better discriminate observations. As it is impossible to enumerate all possible tree structures, a greedy algorithm starting from a single leaf and iteratively adding branches to the tree is used. Letting I_L and I_R represent the instance sets of left and right nodes, respectively, after a split (i.e., $I = I_L \cup I_R$ is the instance set of the split node before splitting it), the new tree has $T + 1$ nodes with the same structure, except around the split node.

$$Gain = \frac{1}{2}\left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\lambda + \sum_{i \in I_L} h_i} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\lambda + \sum_{i \in I_R} h_i} - \frac{\lambda + \sum_{i \in I} h_i}{\lambda + \sum_{i \in I} h_i}\right] - \gamma \tag{6}$$

This process quantifies how a given node split compares to the previous one. Once the *Gain* is negative, the algorithm will stop the cotyledon depth growth. The XGBoost algorithm prunes the features that do not contribute to the prediction (embedded FS) and generates feature importance ranking (*Gain*) based on the relative contribution of each attribute to the model, as depicted in Eq. 6.

Elastic net

Friedman et al. (2010) developed and implemented fast algorithms for fitting generalized linear models with elastic net penalties in the R programming language package "glmnet." We adopted the two-class logistic regression. It considers a response variable $Y \in \{-1, +1\}$ and a vector of predictors $\vec{x} \in R^m$ (m is the number of features), representing class-conditional probabilities through a linear function of the predictors (see Eq. 7):

$$P\left(Y = -1|\vec{x}\right) = \frac{1}{1 + e^{-\left(\beta_0 + \vec{x}\,\prime\beta\right)}} = 1 - P\left(Y = -1|\vec{x}\right), \tag{7}$$

which is equivalent to stating that

$$log\frac{P\left(Y = -1|\vec{x}\right)}{P\left(Y = +1|\vec{x}\right)} = \beta_0 + \vec{x}\,\prime\beta \tag{8}$$

$$P\left(Y = -1|\vec{x}\right) + P\left(Y = +1|\vec{x}\right) = 1 \tag{9}$$

The model fit occurs by regularized maximum (binomial) likelihood. Let $P\left(\vec{x}_i\right)$ be the probability for the ith instance of the training set at a particular value of parameters β_0 and β. $\vec{x}_i \in R^m$ represents the ith instance sample of the training set, and y_i is the ith observed example in the data set ($y_i \in \{1, +1\}$). We maximize the penalty log-likelihood function as follows:

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left\{I\left(y_i = -1\right)loglogP\left(\vec{x}_i\right) + I\left(y_i = +1\right)loglog\left[1 - P\left(\vec{x}_i\right)\right]\right\} - \lambda PF_\alpha(\beta) \tag{10}$$

where $I(\cdot)$ is the indicator function (note that $I\left(y_i = -1\right) + I\left(y_i = +1\right) = 1$), and $PF_\alpha(\beta) = \sum_{k=1}^{m}\left[\frac{1}{2}(1-\alpha){\beta_k}^2 + \alpha|\beta_k|\right]$ is the elastic net penalty factor (Zou and Hastie 2005), representing a compromise between the ridge regression penalty ($\alpha = 0$) and the lasso penalty ($\alpha = 1$). We can rewrite Eqs. 8 and 9 more explicitly as follows:

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left\{I\left(y_i = -1\right)\left(\beta_0 + \vec{x}_i\prime\beta\right) - loglog\left[1 + e^{\beta_0 + \vec{x}_i\prime\beta}\right]\right\} - \lambda\sum\nolimits_{k=1}^{m}\left[\frac{1}{2}(1-\alpha){\beta_k}^2 + \alpha|\beta_k|\right] \tag{11}$$

Bayesian hyperparameters optimization

In both algorithms, hyperparameter setup configures a critical stage to avoid overfitted models. The process may rely on optimization routines, such as the Bayesian hyperparameter optimization (Victoria and Maragatham 2021). This method uses a probabilistic model to find hyperparameter values that maximize the adopted performance metric (e.g., PR-AUC). The process is iterative, meaning the algorithm learns to reach optimal or suboptimal hyperparameter values at each interaction (Snoek et al. 2012; Kuhn 2022).

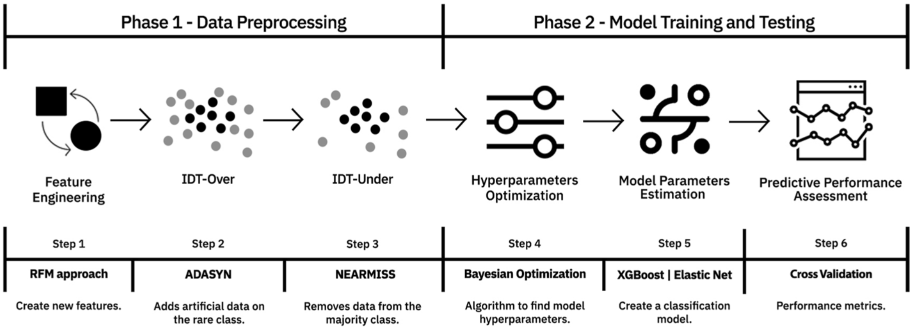
Phase 1 - Data Preprocessing
Phase 2 - Model Training and Testing
Feature Engineering
IDT-Over
IDT-Under
Hyperparameters Optimization
Model Parameters Estimation
Predictive Performance Assessment
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
RFM approach
ADASYN
NEARMISS
Bayesian Optimization
XGBoost | Elastic Net
Cross Validation
Create new features.
Adds artificial data on the rare class.
Removes data from the majority class.
Algorithm to find model hyperparameters.
Create a classification model.
Performance metrics.

Fig. 1 Framework for CCP

Proposed framework for CCP

The proposed framework for CCP comprises two phases—data preprocessing and model training and testing (Fig. 1). We explain each one in the following subsections.

Phase 1: data preprocessing

In this phase, we apply the three data preprocessing techniques (see "Data preprocessing methods" section) to prepare the dataset for model training. Each preprocessing procedure corresponds to an operational step, as detailed below.

Step 1: FE preprocessing

Most churn models use features that are not readily available in the dataset. Moreover, the original feature set may not be sufficiently informative to predict customer churn. Therefore, creating new features through FE is critical to expanding the number of potential predictor candidates.

In the first step of the proposed framework, we create new features based on RFM, which summarize customers' purchasing behavior (Fader et al. 2005; Zhang et al. 2015) and are strongly related to churn behavior in the banking industry. Recency is the time of the most recent purchase; frequency corresponds to the number of prior purchases; and monetary value depicts the average purchase amount per transaction (Fader et al. 2005; Zhang et al. 2015; Heldt et al. 2021).

In our propositions, we considered the concept of the traditional RFM approach (Fader et al. 2005) and disaggregated per product category (RFM/P) (Heldt et al. 2021). We used credit cards, overall credits, and investments as categories. We proposed new features based on the original data and aligned them with the recency concept, which comprehends the overall recency, recency per product category, and recency per channel. We also proposed new features derived from frequency, overall, or per product category. We registered the number of periods with at least one transaction. In each period, we recorded the total number of transactions, the overall binary indicator of purchase incidence, the binary indicator of purchase incidence, and the binary indicator of using a specific channel. Finally, consistent with the monetary value concept, we proposed new features, such as the overall contribution margin, overall revenue, and overall value transacted per product category.

Step 2: IDT-over preprocessing

Aimed at data balancing, in the second step of the framework, we generated synthetic churned customers that act on the gaps in the decision boundary between the classes using the ADASYN algorithm. Such artificial churned customers must not change the data distribution, so we controlled significant differences between the original rare class distribution and the new data distribution using the Kolmogorov–Smirnov test (KS test). The KS test is a well-known nonparametric test that compares the distance between two cumulative distribution functions. Its null hypothesis is that both samples come from the same distribution.

Step 3: IDT-under preprocessing

After adding artificial churned customers, the third step removes noisy and redundant retained customers by applying the NEARMISS algorithm. We did it to equalize the remaining number of maintained and churned customers. Once again, we used the KS test to ensure that data removal did not modify data distribution.

The ADASYN algorithm inserts artificial instances of the rare class in regions where examples of this class are closer to the majority class, improving the distinction between classes. However, the NEARMISS algorithm removes instances from the majority class closer to the rare class (e.g., specimens hard to classify) to reduce noise and redundancies. The proposed framework first employs the ADASYN algorithm to reinforce the frontier between classes. After that, the NEARMISS algorithm provides more precise information on the best candidate instances of the majority class to remove in the subsequent step. We used the package "themis" in the R programming language (Hvitfeldt 2022) to implement NEARMISS and ADASYN.

Figure 2 illustrates the IDT preprocessing stages through a hypothetical example, with the original churned customers representing 20% of the data and the retained customers representing 80%. ADASYN added 5% of artificial churned customers, and NEARMISS removed 65% of the retained customers. The result is a balanced dataset comprising 50% churn and 50% nonchurn customers.

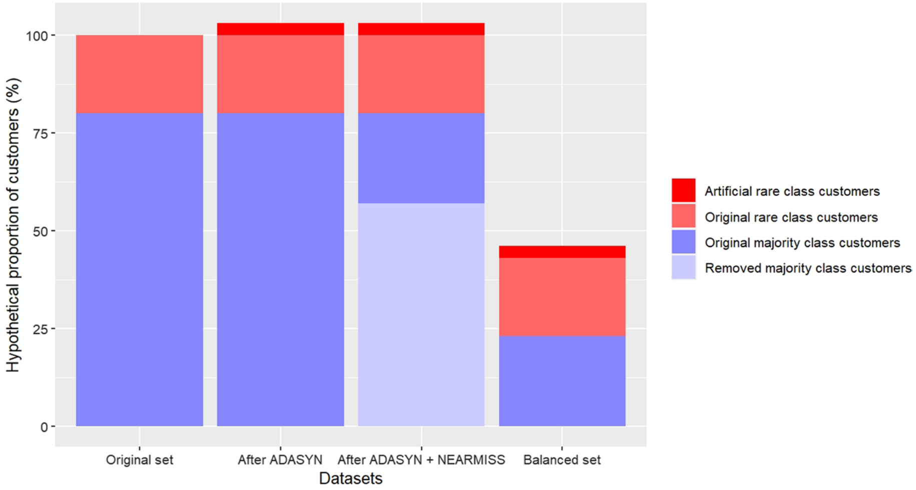

Fig. 2 Illustration of the IDT stages with a hypothetical example

Phase 2: model training and testing

In the final preprocessing phase, we carry out Step 4 using the R package "tune" for Bayesian hyperparameter optimization (Kuhn 2022) of the classification techniques. Using XGBoost, we optimized (1) the number of trees in the ensemble, (2) the minimum number of data points in a node required for further splitting, (3) the maximum depth of each tree, (4) the rate at which the boosting algorithm adapts from iteration to iteration, (5) the reduction in the loss function required for further splitting, and (6) the amount of data exposed to the fitting routine. Then, using the elastic net, we optimized alpha (mixture parameter between a pure ridge model and a pure lasso model) and lambda (regularization penalty). The optimization algorithm maximized the average PR-AUC on a tenfold cross-validation procedure.

In the fifth step, we estimated the model parameters for XGBoost and elastic net classification algorithms using the hyperparameters gaged by the Bayesian optimization in Step 4. We conducted the procedure on samples in the training set.

In the sixth step, we assessed the model's predictive performance using the following metrics: (1) accuracy, which is the proportion of churned and retained customers correctly classified; (2) specificity, which depicts the proportion of correctly classified retained customers against the total retained customers; (3) PR-AUC, which represents the area under the curve of a plot with recall and precision in the axes; and (4) recall, which denotes the model's capacity to correctly classify retained customers (Sofaer et al. 2019).

Data and empirical context

We used data from a major bank operating in Rio Grande do Sul State, Brazil. We selected 36 months of customer transaction history with the bank (December 2018 to November 2021), retrieving about 170 million transactions and more than 3 million customers. The class imbalance is high—2.25% of the customers belong to the minority class (churned customers), and the remaining customers belong to the majority class (retained customers). For more details, an exploratory data analysis is provided in "Appendix A".

In this study, we restricted the scope of analysis to business-to-consumer. The bank offers several financial services, including credit, investment, insurance, and private pension services. It operates under a contractual context, meaning that it experiences customer churn if a customer closes all accounts with the bank.

The relational database relies on 23 tables with customer transaction information, including sociodemographic features, full transaction logs, product categorization, and online channel usage. We present the 32 original features in Table 2.

Results and discussion

In this section, we present the framework validation considering different configurations of preprocessing stages and classification techniques.

Table 2 Original features

01. Adjusted overdraft limit [real]
02. Balance value (inflow-outflow) of credit portability [real]
03. Balance value (inflow-outflow) of salary portability [real]
04. Category of the customer's bank agency [cat]
05. Checking account balance [real]
06. Cohort [int]
07. Credit card limit [real]
08. Credit score categorical [cat]
09. Credit score continuous [real]
10. Digital maturity [cat]
11. Education level [cat]
12. Gender [cat]
13. Marital status [cat]
14. Maximum viable overdraft limit [real]
15. Overall income [real]
16. Overall overdraft limit [real]
17. Overdraft limit [real]
18. Paycheck value [real]
19. Professional profile [cat]
20. Savings account balance [real]
21. Segment (Method A) [cat]
22. Segment (Method B) [cat]
23. The region where the customer's bank agency is located [cat]
24. Type of credit portability [cat]
25. Type of salary portability [cat]
26. Value customer takes the credit value in the banking system [real]
27. Value of inflow of credit portability [real]
28. Value of inflow of salary portability [real]
29. Value of outflow of credit portability [real]
30. Whether the customer has received a salary with the bank [bin]
31. Whether the customer is a civil servant [bin]
32. Whether the customer is an employee in the company [bin]

Binary features [bin] $\in \{0,1\}$; real features [real] $\in R$; integer non-negatives features [int] $\in Z^{+}$; and categorical features [cat]

Framework testing

Combining all three methods sampled for the preprocessing phase (see "Phase 1: data preprocessing" section), we derive eight combinations: FE (yes or no) × IDT-over (ADASYN or no) x IDT-under (NEARMISS or RU). We compare all the combinations against a ninth alternative taken as a baseline of not doing any preprocessing phase (no FE, no IDT-over, and no IDT-under) using only the 32 original features. Moreover, we combined all nine preprocessing choices with the two classification techniques (XGBoost or elastic net) examined, deriving 18 configurations for testing. They are presented in the first five columns of Table 5 as a framework.

ADASYN algorithm does not determine a priori the optimal number of churned customers to synthesize. We want to add a certain number of churned customers that reinforce the boundary between classes without compromising the sample's representativeness. We performed experiments by adding as many artificial churned customers as possible without identifying any significant difference between the empirical cumulative densities of the distribution of rare class—original versus ADASYN.

We also performed experiments to define the KNN parameter, as both ADASYN and NEARMISS rely on it. Numerical experiments suggest that $k = 5$.

The implementations used the R programming language version 4.2.1 (R Core Team 2022) package "themis" (Hvitfeldt 2022) for NEARMISS and ADASYN implementations, package "tune" for Bayesian hyperparameter optimization (Kuhn 2022), package "xgboost" (Chen et al. 2022) for XGBoost, and package "glmnet" (Friedman et al. 2010) for elastic net. Next, we describe the process of creating new features in the FE stage.

Table 3 Created features

Recency	Frequency	Monetary value
RFM aggregated		
33. Purchase recency [int]	34. Purchase frequency [int]	39. Total value transacted monthly [real]
	35. Relative purchase frequency [real]	40. Total contribution margin [real]
	36. Purchase incidence [bin]	
	37. Effective total amount transacted monthly [real]	
	38. Number of distinct monthly transactions [int]	
RFM disaggregated per product category		
41. Credit purchase recency [int]	44. Credit purchase incidence [bin]	56. Overall investment value [real]
42. Investment purchase recency [int]	45. Investment purchase incidence [bin]	57. Overall credit value [real]
43. Use of mobile channels recency [int]	46. Use of mobile channels incidence [bin]	58. The ratio of overall credit value over credit value taken by the customer with all banks in the market [real]
	47. Number of direct debits [int]	59. Number of investment products terminated [int]
	48. The number of interactions in mobile channels [int]	60. Number of credit contracts finished or terminated [int]
	49. The number of transactions and purchases in mobile channels [int]	
	50. Number of distinct products purchased [int]	
	51. Whether the customer has used a credit card [bin]	
	52. Whether the customer has done any debt renegotiation [bin]	
	53. Whether the customer has any credit taken [bin]	
	54. Whether the customer has any overdue credit [bin]	
	55. Whether the customer has a credit card [bin]	

Binary features $\left[\text{bin}\right] \in \{0, 1\}$; real features $\left[\text{real}\right] \in R$; integer non-negatives features $[\text{int}] \in Z^{+}$; and categorical features [cat]

New predictor features through FE

We use aggregated and disaggregated RFM per product category to create new predictor features. Table 3 presents the set of 28 features produced through the FE stage. The code for the new features starts at 33 and is in sequence with the original ones. We find that eight of these new features follow the traditional approach of RFM (Fader et al. 2005), aggregating variables related to RFM for each customer. The additional 20 created features follow the RFM per product approach (Heldt et al. 2021), disaggregating variables related to RFM per product category for each customer.

Purchase frequency counts the number of consecutive months in which a customer purchased any product (this counting process restarts after any month with no purchase). Purchase recency sums the number of months since the last purchase. We split the remaining features into two alternatives based on the feature type. We take the most recent value in the training time frame for traits at nominal and binary levels, converting it to dummy features using the one-hot encoding procedure. For the other continuous features, we take the median of the last six periods.

In the following subsection, we analyze the impact of preprocessing stages and classification models on churn prediction performance. We also examine the influence of the IDT-over and IDT-under algorithms on the majority and rare class distributions.

Impact of IDT-over and IDT-under techniques on data distributions

Table 4 presents the amount and proportions of customers in the IDT preprocessing stages. The final ratio between classes is 50% each.

After applying ADASYN and NEARMISS algorithms, we checked whether the resulting distributions of all features for both majority and rare classes are like their original distributions using the KS test. We found no statistically significant difference between all pairs of variables (before and after IDT-over and IDT-under), with *p values* ranging from 0.9 to 1. Figure 3 depicts the empirical cumulative distributions (original rare class × after ADASYN and original majority class × after NEARMISS) of two

Table 4 The proportion between classes in the IDT preprocessing stages

Stage	Total customers	Churned	Retained
Original set	3,283,332 (100.00%)	73,798 (2.25%)	3,209,534 (97.75%)
After IDT-Over ADASYN	3,298,107 (100.45%)	88,573 (2.70%)	3,209,534 (97.75%)
After IDT-Over ADASYN+ IDT-Under NEARMISS	177.146 (5.40%)	88,573 (2.70%)	88,573 (2.70%)

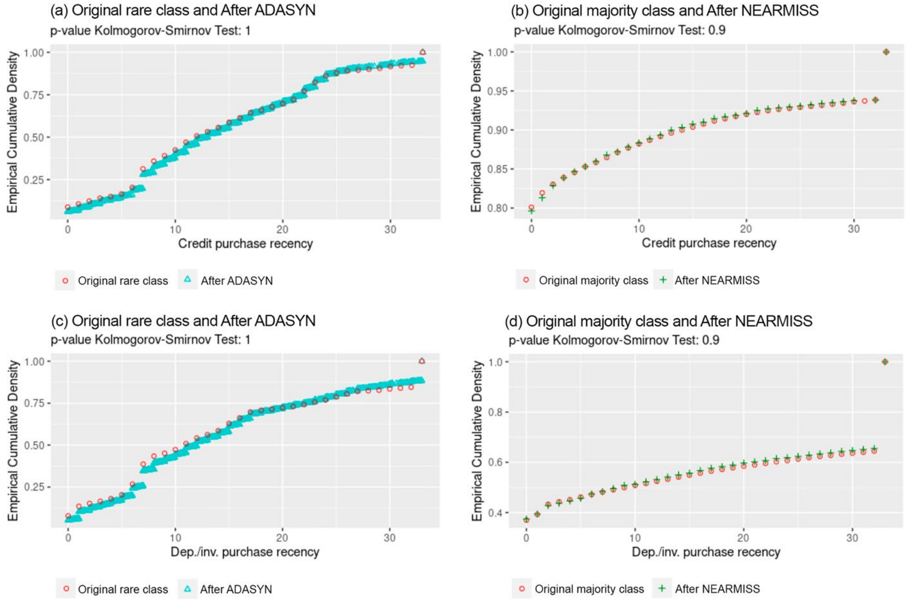

Fig. 3 Comparison among cumulative distributions from the rare and majority classes after applying IDT preprocessing stages

features—credit purchase recency and investment purchase recency. These features are among the most important for predicting churn behavior according to the importance features index (*Gain*) from XGBoost (discussed at the end of this section).

Figure 3a, c depict that the artificially added churned customers did not modify the original distribution (*p value* 1.0), although ADASYN does not use information from the distribution. Figure 3b, d depict the same result for the NEARMISS algorithm—the distributions of retained clients and after-NEARMISS maintained customers are not significantly different (*p value* 0.9), although NEARMISS does not use information from the distribution.

Performance results

The predictive performance of the proposed framework (configurations C1 and C2 in Table 5) was tested against 16 alternative configurations by (1) using or not using FE, (2) using ADASYN for IDT-over or not using ADASYN, and (3) using NEARMISS or RU for IDT-under. We alternate the classification techniques (e.g., XGBoost and elastic net) in such configurations to assess their performance on churn prediction. We trained and tested each of these 18 configurations using 100 folds cross-validation. We used configurations C17 and C18 as references, not undergoing preprocessing stages. Table 5 presents the average PR-AUC, accuracy, recall, and specificity sorted by decreasing PR-AUC. "Appendix B" depicts a confusion matrix of the average of the 100-fold cross-validation procedure for the recommended configurations C1 and C2.

Table 5 reveals that, on average, configurations C1 and C2 (e.g., FE, ADASYN, and NEARMISS coupled with XGBoost and elastic net) yielded the highest predictive

Table 5 Average of the 100-fold cross validation predictive performance for the 18 assessed configurations

Configuration	FE	IDT-Over	IDT-Under	Classification model	PR-AUC	Accuracy	Recall	Specificity
C1	Yes	ADASYN	NEARMISS	XGBoost	0.9518	0.8860	0.8967	0.8884
C2	Yes	ADASYN	NEARMISS	Elastic Net	0.9157	0.8632	0.8654	0.8592
C3	Yes	ADASYN	RU	XGBoost	0.8857	0.8194	0.8401	0.8184
C4	Yes	None	NEARMISS	Elastic Net	0.8843	0.8208	0.8509	0.8225
C5	No	ADASYN	NEARMISS	Elastic Net	0.8614	0.7769	0.7802	0.7755
C6	Yes	None	RU	Elastic Net	0.8610	0.8045	0.8199	0.8041
C7	Yes	ADASYN	RU	Elastic Net	0.8564	0.7845	0.8300	0.7890
C8	No	None	NEARMISS	Elastic Net	0.8397	0.7435	0.7623	0.7429
C9	No	ADASYN	RU	Elastic Net	0.8110	0.6843	0.7655	0.6856
C10	No	None	RU	Elastic Net	0.8103	0.7128	0.7531	0.7141
C11	Yes	None	RU	XGBoost	0.8025	0.7444	0.7544	0.7468
C12	Yes	None	NEARMISS	XGBoost	0.7987	0.7399	0.7549	0.7445
C13	No	ADASYN	NEARMISS	XGBoost	0.7055	0.6337	0.6435	0.6320
C14	No	ADASYN	RU	XGBoost	0.6543	0.5671	0.6095	0.5630
C15	No	None	NEARMISS	XGBoost	0.5934	0.5263	0.5495	0.5283
C16	No	None	RU	XGBoost	0.5870	0.5295	0.5337	0.5301
C17	No	None	None	XGBoost	0.5000	0.9989	0.0000	1.0000
C18	No	None	None	Elastic Net	0.5000	0.9988	0.0000	1.0000

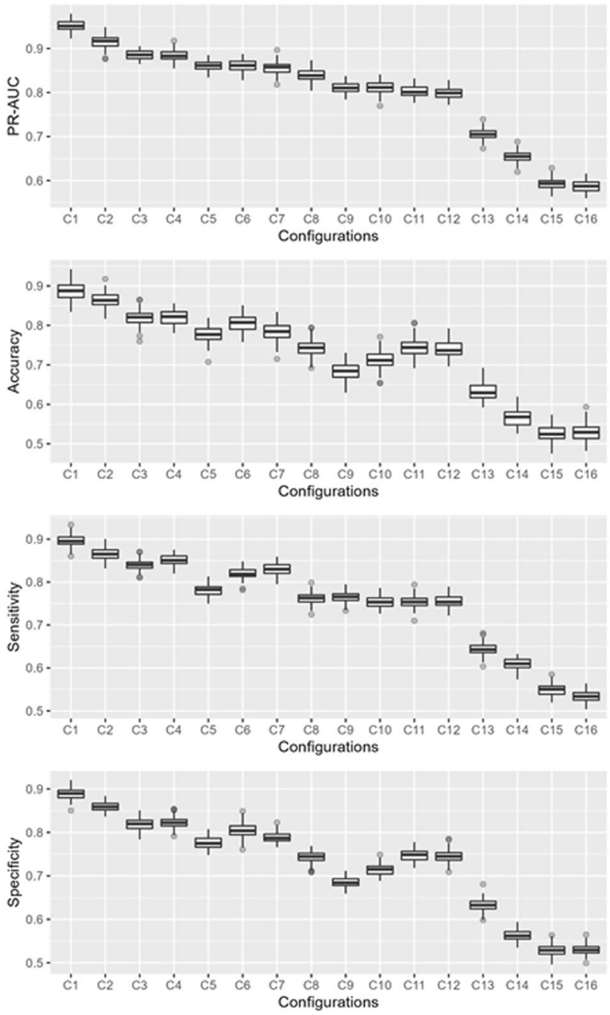

Fig. 4 Boxplot for 16 configurations and metrics

performances across all performance metrics. It highlights that the recommended ADASYN and NEARMISS strategies outperformed all the benchmark models tested, including configurations relying on RU—an IDT-under algorithm usually reported by the literature.

These results corroborate the hypothesis that inserting new features based on a solid conceptual background, such as RFM and RFM/P, increases the performance of churn predictions. The ADASYN that precisely inserted churned customers in the minority class reinforced the decision frontier between the majority and minority classes. Similarly, the NEARMISS applied to the majority class selected the less contributive customers to discard (instead of randomly choosing such customers), reinforcing the decision frontier and improving the classification performance. We find a superior performance of the XGBoost classification technique, which significantly outperformed elastic net. Finally, configurations C17 and C18 reveal that a strongly imbalanced dataset with no preprocessing drops the classifier performance for both XGBoost and elastic net. Both classifiers did not detect any pattern of churned customers (zero recall) and classified all

Table 6 MANOVA test

Stage	*df*	Pillai	Approximated F	Numerator df	Denominator df	Pr(> F)
FE	1	0.8115	1714.38	4	1592	0.0000
IDT-Over	1	0.2739	150.15	4	1592	0.0000
IDT-Under	1	0.2255	115.94	4	1592	0.0000
Classification model	1	0.5646	516.30	4	1592	0.0000
Residuals	1595					

customers as retained, which accounted for a perfect specificity. Although the accuracies of C17 and C18 are the highest among all configurations, they are misleading as they did not classify a single churned customer correctly. Figure 4 depicts the boxplots for all 16 configurations and four performance metrics (we omit configurations C17 and C18 due to the effects of data unbalance on assessed performance metrics). Such boxplots depict a similar variability of all performance metrics in the 100 replicates, suggesting high stability of the assessed metrics even when the performance levels are low.

Next, we statistically assessed how the different preprocessing steps and classification techniques impacted the performance metrics. We first employed the multivariate analysis of variance (MANOVA) to test whether the three stages (FE, data balance, and classification model) are significant in the four predictive performance metrics. To apply the MANOVA test, we initially tested for the normality of the performance metrics using the Shapiro–Wilk test. Only three out of the 64 performance averages did not pass the test ($\alpha < 0.05$). Next, we carried out the MANOVA test. Table 6 presents the results.

The results indicate that all groups within the stages are different regarding the performance metrics. The preprocessing steps are informative for the predictive performance across all the metrics tested. The Pillai results for each stage indicate that FE had the highest impact on predictive performance (Pillai = 0.8115), followed by the classification model (Pillai = 0.5646), IDT-over (Pillai = 0.2739), and IDT-under method (Pillai = 0.2255).

Based on the MANOVA results, we conducted individual univariate ANOVA tests for each performance metric, as presented in Table 7. This was conducted to check the significance of each factor in each performance metric.

The univariate ANOVAs confirmed the significant differences when considering each performance metric separately. They also supported the higher impact on all performance metrics obtained from using FE relative to the other stages studied. Additionally, we assessed the statistical differences of the best-ranked configurations using C1 configuration (e.g., FE, ADASYN, NEARMISS, and XGBoost) as a reference. A post hoc analysis using pairwise t-tests revealed that the pairs C1–C2, C1–C3, C1–C4, and C1–C5 are all significantly different ($\alpha < 0.05$) for all performance metrics.

Given the impact of the FE stage in the framework proposed, we extracted the feature importance (*Gain*) using Eq. 1 for C1 configuration to compare the contribution of the most relevant original and created features. Table 8 presents the 10+ features, comprising 80%[1] of the accumulated feature importance.

[1] Using the classic Pareto rule, Appendix C exhibits the complete list of remaining features resulting from XGBoost.

Table 7 Univariate ANOVAs

Metric	Stage	*df*	Sum Sq	Mean Sq	F value	Pr(> F)
PR-AUC	FE	1	7.4743	7.4743	2678.21	0.0000
	IDT-Over	1	1.3519	1.3519	484.40	0.0000
	IDT-Under	1	0.4983	0.4983	178.53	0.0000
	Classification Model	1	4.6317	4.6317	1659.63	0.0000
	Residuals	1595	4.4513	0.0028		
Accuracy	FE	1	10.3780	10.3780	4037.89	0.0000
	IDT-Over	1	0.9675	0.9675	376.43	0.0000
	IDT-Under	1	0.7384	0.7384	287.30	0.0000
	Classification Model	1	3.4610	3.4610	1346.59	0.0000
	Residuals	1595	4.0994	0.0026		
Recall	FE	1	9.2224	9.2224	4260.11	0.0000
	IDT-Over	1	1.2777	1.2777	590.20	0.0000
	IDT-Under	1	0.2430	0.2430	112.24	0.0000
	Classification Model	1	4.4635	4.4635	2061.84	0.00000
	Residuals	1595	3.4529	0.0022		
Specificity	FE	1	10.58	10.5839	4627.61	0.0000
	IDT-Over	1	0.8925	0.8925	390.21	0.0000
	IDT-Under	1	0.7320	0.7320	320.05	0.0000
	Classification Model	1	3.4350	3.4350	1501.89	0.0000
	Residuals	1595	3.6480	0.0023		

Table 8 10 most important features

Feature	Original/from FE	Gain
41. Credit purchase recency	From FE	0.4623
43. Use of mobile channels recency	From FE	0.0503
15. Overall income	Original	0.0491
42. Investment purchase recency	From FE	0.0489
50. Number of distinct products purchased	From FE	0.0420
39. Total value transacted monthly	From FE	0.0378
38. Number of distinct monthly transactions	From FE	0.0375
18. Paycheck value	Original	0.0358
06. Cohort	Original	0.0280
30. Whether the customer has received a salary with the bank	Original	0.0270

Credit purchase recency is the topmost important feature, accounting for 46.23% of feature importance when predicting customer churn. This finding reinforces the conceptual foundations of customer behavior in the banking industry and the method used to support the proposition of this new feature. When acquiring credit products, clients sign contracts that usually last several months, so it is not surprising that a customer remains active for a while. Once the credit contract has ceased and recency starts to increase, the likelihood of defection also increases significantly, most likely because of low switching costs.

According to the traditional RFM concept, recency is more influential than frequency in defining the likelihood of a customer being active (Fader et al. 2005). Regarding the derived RFM/P concept, the same rationale is valid, with the only difference being the disaggregation per product category. In the present study, the relevance of credit purchase recency for customer retention in the banking industry confirms the findings of the existing literature.

Besides credit purchase recency, usage of mobile channels and investment purchase recency are also relevant for churn prediction using the recommended C1 configuration. It confirms the importance of features based on the RFM/P concept, which is consistent with the rationale that recency is critical to defining whether a customer is likely to remain active or not in the future.

Apart from the credit-related features, deposit and investment-related features were also influential for churn prediction. Customers' withdrawals from checking and investment accounts indicate possible churn events in the future. As recency measures the time between withdrawals, smaller recency suggests that a customer is likely to churn. The recency of customer interactions through mobile channels also contributes to that sense, helping to identify customers that might be making transactions with a competitor.

Overall customer income is also deemed one of the most relevant features by the proposed framework as it contributes to increasing the retention rate. Low-income customers are less likely to purchase investment products and more likely to default. Additionally, due to their low credit limits, such customers often acquire credit from different financial institutions.

The number of distinct products purchased was also relevant to predicting customer churn. This feature correlates with the concept of purchase frequency as it indicates that a customer has purchased at least one financial service in a period. However, it also encompasses information on how many different financial services a customer has purchased. Therefore, it reflects the success of cross-selling efforts made by sales teams. The relevancy of this feature to predict churn indicates that the higher the number of financial services a customer purchases, the less likely she is to churn.

We analyzed the impact of customer churn on a firm's financial performance in terms of the profit loss that would be avoided by employing the proposed framework. We observed that 97.25% of profitability could be maintained if successful action was taken to reverse the potential churn. When no action is taken, a firm would incur costs to acquire new customers in order to recover the loss in profitability.

Furthermore, customers that are most likely to defect are those currently interacting with other banks, with a median of 62% of businesses in other financial institutions ("Appendix D", Fig. 5). Based on the Mann–Whitney test, a significant difference (*p value* 0.00) was found in the share of wallets between the churn group and the group of retained customers. The median share of wallet for the churn group ($\widehat{\mu}_{median} = 0.38$) was substantially lower than the nonchurn group ($\widehat{\mu}_{median} = 0.78$). This suggests that the churn group has business with other financial institutions, and their defection represents a missed opportunity to exploit their business opportunities. By retaining customers, a financial institution can capitalize on its business potential and generate long-term revenue streams.

Regarding a bank's products, the profitability loss appears to come mainly from credit. For example, rural credit was the most affected product, accounting for 35% of the total lost margin and losing 8% of its margin during the sample period. This information suggests the need for adjustments in offering that product.

Managerial implications

To deal with customer churn, marketing managers try to identify in advance the decrease in customer engagement with their company and plan marketing efforts to prevent customers from defecting (Benoit and Poel 2012; Gordini and Veglio 2017). Regarding the retail banking sector, adequate customer retention management has become increasingly relevant due to the growing competition. The entry of new players relying on digital and scalable innovative solutions tends to replace traditional services offered by incumbents and reduce customers' switching costs (Lazari and Machado 2021).

Managers benefit from the framework proposed in this study in many ways. They can anticipate potentially churning customers three months before, even with highly imbalanced data. The framework not only allows managers to know the most likely churning customers but also gives them a reasonable amount of time to plan marketing efforts proactively to prevent such churn from happening.

In addition, based on the FE step of the framework (Step 1) and using the concept of RFM, we sought to generate features related to churn behavior in the retail banking industry (see Tables 2, 3). This provides a tailored list of adequate predictors that managers can use to harness their existing databases, allowing classification algorithms to predict customer churn more precisely and thus increase the effectiveness of the customer management team to monitor and manage potential churners.

Finally, based on these proposed features and after estimating the classification algorithms, we also identify features with higher gain (Table 8). Analyzing such metrics contributes to understanding features that are the most relevant churn predictors in the retail banking context. For instance, credit purchase recency is the most pertinent feature. Although customers might transfer a credit contract to another banking institution anytime they want, this finding suggests that the existence of credit contracts significantly increases the likelihood of retaining a customer. Another key feature is the use of mobile channel recency. As the competition is becoming increasingly dependent on building digital relationships with customers, engaging them through the frequent use of mobile channels is critical for their loyalty to a company. In summary, knowing features that impact churn behavior is relevant for managers to drive well-designed marketing efforts and avoid customer defection. Thus, adopting the proposed framework has important managerial implications, providing managers with resources to enhance customer retention management and thrive in a digital market environment with growing competition and lower switching costs.

Conclusions

Our study makes a valuable contribution to the research on decision support systems by proposing a framework to model customer churn behavior in the context of highly imbalanced classes. We emphasize the importance of conducting data preprocessing strategies (i.e., FE, IDT-over, and IDT-under processes) to improve model performance. The FE process is fundamental for building new and informative features to better characterize customers' behavior, directly impacting model effectiveness. Regarding the IDT-over and IDT-under strategies, they handle the class imbalance issue through the ADASYN and NEARMISS algorithms, respectively, providing the machine learning technique with a suitable number of instances at the modeling stage. In the IDT-over stage, the ADASYN algorithm reinforced the decision boundary toward the minority class (churned customers). In the IDT-under step, the NEARMISS provided an efficient way to undersampling the retained customers of the majority class such that the maintained customers have reduced noise and redundancies, making it easier to identify patterns with the XGBoost and elastic net models.

Compared with alternative configurations, we demonstrated that the algorithms used in our proposed framework perform well in PR-AUC, accuracy, sensibility, and specificity. Additionally, as we tested two classification models (XGBoost and elastic net), we demonstrated that adequate data preprocessing procedures improve the predictive power in classification models relying on different mathematical fundamentals.

The proposed framework also provided a methodological contribution to the literature on churn prediction by adequately dealing with highly imbalanced datasets. The thorough work on data preprocessing in the FE step and rebalancing classes, reinforcing the decision boundary between them, and reducing data redundancy and noise, combined with state-of-the-art classification techniques, led to a higher predictive performance of customer churn. Thus, our study substantially enhances customer retention practices, fostering the adoption of more effective marketing efforts to prevent churn. Preventing churn increases customer portfolio profitability.

These results confirm the effectiveness of conducting a detailed data preprocessing before proceeding to the model training. Given the lack of extant research on this topic, we encourage additional studies to investigate different data preprocessing methods. They should not only test diverse classification models to achieve gains in predictive performance but also test different combinations of data preprocessing techniques because it has a significant potential to provide even more accurate predictions.

Finally, we highlight some limitations in the proposed study that can become the subject of future research. One restriction is about using a single dataset, mainly due to the challenging task of obtaining real datasets from the retail banking industry. We encourage future studies to extend the proposed framework to other cases in this industry or even apply it to address binary classification problems other than churn prediction. Another limitation relates to the classification techniques tested in our experiments (XGBoost and elastic net); different algorithms, such as artificial neural networks and support vector machines, can also be tested. The decision to use only two types of algorithms was due to the focus of our study in the data preprocessing phase. Another limitation of the study is the absence of features regarding customer transactions with other companies. Data indicating customers' share-of-wallet change over time would probably increase predictive performance. For instance, a decrease in the number of transactions of a given customer with a focal company may increase the number of transactions of this customer with competitors. The only available information regarding such behavior was the binary feature on whether a customer took credit with other companies in the banking system. However, we found only a weak association of this binary feature with churn. Therefore, it does not appear as a good churn predictor. Given the lack of such features in our dataset, we had to apply the framework utilizing data mainly related to the strict relationship between the focal company and a customer. Future studies can benefit from using features regarding customer transactions with other companies.

Appendix A: Exploratory data analysis

In this appendix, we present exploratory data analysis to better describe the type of features used, how they are distributed, and how they are correlated to each other. In Table 9, we present the number of unique classes in each categorical feature as well as the amount of customers in the four most frequent classes of each categorical feature.

Table 10 shows the number of customers in each binary features class, as well as the average between both classes, to show how they are distributed. Among these features, we highlight that the Purchase incidence feature has an average close to 1 (Average = 0.98). It shows how most customers are highly active. However, this does not necessarily mean that they will be retained, even though the churn event is rare. It indicates the challenge of modeling churn in this context.

Table 9 Descriptive statistics of categorical features

Feature	Number of classes	Count*
04. Category of customer's bank agency	7	B: 888,415; C: 631,418; A: 613,032; D: 517,727
08. Categorical credit score	13	99: 1,569,874; 12: 461,279; 2: 404,631; 1: 206,043
10. Digital maturity	4	1: 2,130,840; 3: 805,861; 2: 267,148; 4: 79,483
11. Education level	15	13: 833,247; 9: 695,946; 7: 484,516; 21: 446,435
12. Gender	2	F: 1,815,657; M: 1,467,675
13. Marital status	8	6: 1,461,414; 1: 1,044,718; 4: 298,700; 5: 160,176
19. Professional profile	51	101: 659,322; 202: 546,260; 201: 458,209; 999: 260,877
21. Segment (Method A)	3	A: 2,483,565; B: 627,096; C: 172,671
22. Segment (Method B)	4	A: 2,696,468; B: 466,652; C: 94,184; D: 26,028
23. Region where customer's bank agency is located	17	14: 539,811; 4: 414,037; 6: 381,175; 37: 370,482
24. Type of credit portability	4	N: 3,250,665; E: 19,125; S: 13,154; ES: 388
25. Type of salary portability	2	N: 3,101,121; E: 182,211

*Limited to the top four

Table 10 Descriptive statistics of binary features

Feature	Average	0 (FALSE)	1 (TRUE)
30. Whether the customer has received a salary with the bank	0.42	1,892,600	1,390,732
31. Whether the customer is a civil servant	0.16	2,748,934	534,398
32. Whether the customer has a dedicated account manager	0.15	2,790,652	492,680
44. Credit purchase incidence	0.83	570,809	2,712,523
45. Investment purchase incidence	0.37	2,080,999	1,202,333
36. Purchase incidence	0.98	58,702	3,224,630
46. Use of mobile channels incidence	0.61	1,290,043	1,993,289
55. Whether the customer has a credit card	0.20	2,635,776	647,556
53. Whether the customer has any credit taken	0.41	1,922,302	1,361,030
54. Whether the customer has any overdue credit	0.08	3,012,753	270,579
52. Whether the customer has done any debt renegotiation	0.01	3,252,890	30,442
51. Whether the customer has used a credit card	0.14	2,816,557	466,775

Table 11 Descriptive statistics of real and integer features

Feature	Average	Standard deviation	Minimum	Median	Maximum
01. Adjusted overdraft limit	256.40	630.09	0.00	0.00	9,250.00
02. Balance value (inflow-outflow) of credit portability	18.58	16,280.40	− 338,316.03	0.00	414,720.72
03. Balance value (inflow-outflow) of salary portability	306.95	1,266.32	− 5,531.10	0.00	0.00
05. Checking account balance	298.07	2,077,43	0.00	4.74	53,932.53
06. Cohort	57.31	18.43	6.00	68.00	68.00
07. Credit card limit	466.05	1,216.43	0.00	0.00	14,875.00
09. Credit score continuous	0.10	0.11	0.00	0.06	0.87
14. Maximum viable overdraft limit	370.53	894.32	0.00	0.00	13,900.00
15. Overall income	3,468.91	5,707.86	0.00	1,915.50	125,626.80
16. Overall overdraft limit	256.40	630.09	0.00	0.00	9,250.00
17. Overdraft limit	1,213.84	3,925.43	0.00	0.00	60,000.00
18. Paycheck value	1,678.21	2,449.37	0.00	1,100.00	32,995.03
20. Savings account balance	2,703.46	15,142.11	0.00	0.00	304,495.32
26. Value customer takes the credit value in the banking system	19,694.76	61,405.29	0.00	1,076.91	1,481,108.16
27. Value of inflow of credit portability	678.95	10,654.78	0.00	0.00	338,316.03
28. Value of inflow of salary portability	306.95	1,266.32	0.00	0.00	5,531.10
29. Value of outflow of credit portability	697.39	12,605.24	0.00	0.00	417,276.72
33. Purchase recency	0.08	0.89	0	0.00	33
34. Purchase frequency	30.24	7.49	0	33.00	33
35. Relative purchase frequency	0.97	0.17	0.00	1.00	1.00
37. Effective total amount transacted monthly	16,358.43	55,182.67	0.00	669.87	1,248,443.91
38. Number of distinct monthly transactions	12.92	15.79	0.00	7.00	1,160.00
39. Total value transacted monthly	19,282.47	58,117.48	0.00	2,265.93	1,297,626.13
40. Total contribution margin	75.11	232.77	− 79.47	3.40	28,222.49
41. Credit purchase recency	2.45	7.09	0	0.00	33
42. Investment purchase recency	14.86	14.81	0	9.00	33
43. Use of mobile channels recency	7.82	12.63	0	0.00	33
47. Number of direct debits	0.48	0.86	0	0.00	11
48. The number of interactions in mobile channels	16.41	33.39	0	0.00	2,200
49. The number of transactions and purchases in mobile channels	2.49	5.73	0	0.00	553
50. Number of distinct products purchased	3.43	2.79	0	3.00	20
56. Overall investment value	4,987.83	21,306.99	0.00	0.00	370,407.32
57. Overall credit value	10,182.29	47,298.56	0.00	95.14	1,062,054.80
58. The ratio of overall credit value over credit value taken by the customer with all banks in the market	0.27	0.40	0.00	0.00	1.00
59. Number of investment products terminated	1.44	4.47	0	0.00	32
60. Number of credit contracts finished or terminated	0.69	1.69	0	0.00	32

Table 11 shows the following descriptive statistics of each real and integer feature: (1) average, (2) standard deviation, (3) minimum value, (4) median, and (5) maximum value. We highlight how the distributions of the purchasing recency per product differ significantly from the distribution of the overall purchase recency feature (specifically, credit purchase recency and investment purchase recency). It indicates how the feature engineering process based on the recency, frequency, and monetary value concept (RFM) was important to generate features that uncover a relevant level of variability which was not observable by only using the overall purchase recency feature. It provides a richer set of predictor features for the learning algorithms increasing the likelihood of predicting the churn more precisely.

Appendix B: Confusion matrix C1 and C2

Table 12 shows the average confusion matrix average of the 100-fold cross-validation procedure for the recommended configurations C1 and C2.

Table 12 Average confusion matrix for the 100-fold cross-validation experiment

	Reference	
	Churn	**Retained**
C1		
Prediction		
Churn	6,631.10	36,133.96
Retained	763.90	287,647.04
C2		
Prediction		
Churn	6,399.63	45,588.36
Retained	995.37	278,192.64

Appendix C: Remaining features resulting from XGBoost

Table 13 shows the complete list remaining of original or feature engineering features resulting from XGBoost, sorted by descending Gain.

Table 13 Complete list of remaining features resulting from XGBoost

Feature	Original/from FE	Gain
41. Credit purchase recency	From FE	0.4623
43. Use of mobile channels recency	From FE	0.0503
15. Overall income	Original	0.0491
42. Investment purchase recency	From FE	0.0489
50. Number of distinct products purchased	From FE	0.0420
39. Total value transacted monthly	From FE	0.0378
38. Number of distinct monthly transactions	From FE	0.0375
18. Paycheck value	Original	0.0358
06. Cohort	Original	0.0280
30. Whether the customer has received a salary with the bank	Original	0.0270
60. Number of credit contracts finished or terminated	From FE	0.0224

Table 13 (continued)

Feature	Original/from FE	Gain
37. Effective total amount transacted monthly	From FE	0.0166
19. Professional profile	Original	0.0146
34. Purchase frequency	From FE	0.0142
40. Total contribution margin	From FE	0.0137
13. Marital status	Original	0.0122
57. Overall credit value	From FE	0.0117
59. Number of investment products terminated	From FE	0.0090
11. Education level	Original	0.0086
05. Checking account balance	Original	0.0082
56. Overall investment value	From FE	0.0073
45. Investment purchase incidence	From FE	0.0071
20. Savings account balance	Original	0.0058
48. The number of interactions in mobile channels	From FE	0.0053
46. Use of mobile channels incidence	From FE	0.0042
12. Gender	Original	0.0039
22. Segment (Method B)	Original	0.0036
33. Purchase recency	From FE	0.0031
54. Whether the customer has any overdue credit	From FE	0.0029
17. Overdraft limit	Original	0.0027
58. The ratio of overall credit value over credit value taken by the customer with all banks in the market	From FE	0.0026
23. The region where the customer's bank agency is located	Original	0.0016

Appendix D: Share of wallet between churned and retained customers (Fig. 5)

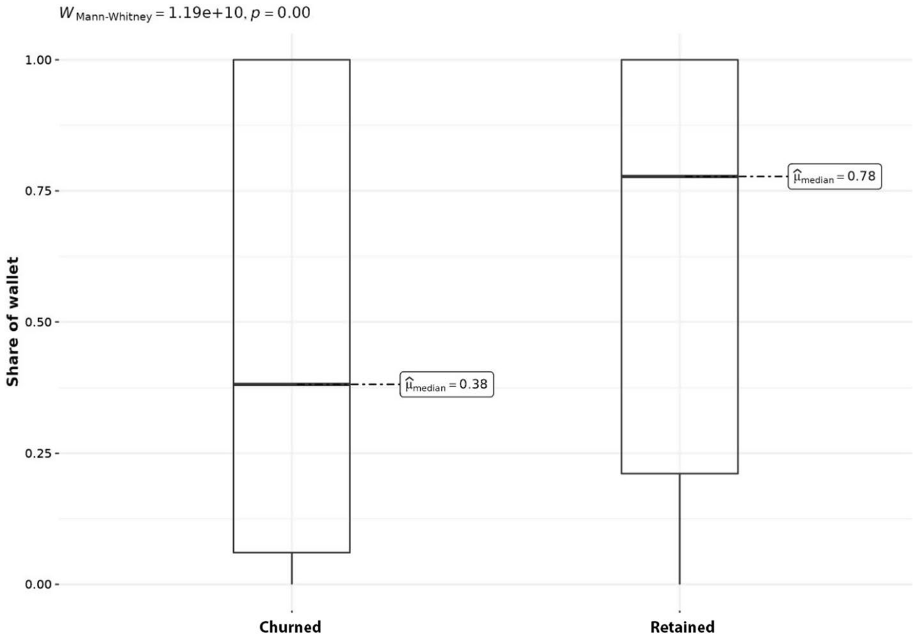

Fig. 5 Mann–Whitney test to compare the share of wallet between churned and retained customers

Abbreviations

C1	Configuration 1
C2	Configuration 2
C3	Configuration 3
C4	Configuration 4
C5	Configuration 5
C6	Configuration 6
C7	Configuration 7
C8	Configuration 8
C9	Configuration 9
C10	Configuration 10
C11	Configuration 11
C12	Configuration 12
C13	Configuration 13
C14	Configuration 14
C15	Configuration 15
C16	Configuration 16
C17	Configuration 17
C18	Configuration 18
CCP	Customer churn prediction
FE	Feature engineering
FS	Feature selection
IDT	Imbalanced dataset treatment
IDT-over	Imbalanced dataset treatment oversampling
IDT-under	Imbalanced dataset treatment undersampling
KNN	K-Nearest neighbor
KS-test	Kolmogorov–Smirnov test
MVI	Missing values imputation
PR-AUC	Precision-recall area under curve
OT	Outliers treatment
RFM	Recency, frequency, and monetary value.
RFM/P	Recency, frequency, and monetary value/product
RU	Random undersampling
SMOTE	Synthetic minority oversampling technique

Acknowledgements

Not applicable.

Author contributions

JBGB designed the model, and the computational framework analyzed the data and carried out the implementation. GBB, RH, JLB, and MJA verified the analytical methods and contributed to writing the manuscript with input from all authors. CSS, FBL, and MJA supervised the findings of this work and contributed to the final version of the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are not publicly available due to a data privacy agreement signed with the financial institution but are available from the corresponding author on reasonable request.

Declarations

Competing interests

The authors declare no competing interests.

Received: 16 September 2022 Accepted: 20 September 2023
Published online: 13 January 2024

References

Ascarza E (2018) Retention futility: targeting high-risk customers might be ineffective. J Mark Res 55:80–98. https://doi.org/10.2139/ssrn.2759170

Ascarza E, Hardie BGS (2013) A joint model of usage and churn in contractual settings. Mark Sci 32:570–590. https://doi.org/10.1287/mksc.2013.0786

Ascarza E, Neslin SA, Netzer O et al (2018) In pursuit of enhanced customer retention management: review, key issues, and future directions. Cust Need Solut 5:65–81. https://doi.org/10.1007/s40547-017-0080-0

Bafna R, Jain R, Malhotra R (2023) A comparative study of classification techniques and imbalanced data treatment for prediction of software faults. Res Sq. https://doi.org/10.21203/rs.3.rs-2809140/v1

Benoit DF, den Poel DV (2012) Improving customer retention in financial services using kinship network information. Expert Syst Appl 39:11435–11442. https://doi.org/10.1016/j.eswa.2012.04.016

Broby D (2021) Financial technology and the future of banking. Financ Innov 7:47. https://doi.org/10.1186/s40854-021-00264-y

Broby D (2022) The use of predictive analytics in finance. J Finance Data Sci 8:145–161. https://doi.org/10.1016/j.jfds.2022.05.003

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, New York, pp 785–794

Chen T, He T, Benesty M, et al (2022) xgboost: Extreme gradient boosting. CRAN R package version 1.6.0.1: https://CRAN.R-project.org/package=xgboost

Datta H, Foubert B, Van Heerde HJ (2015) The challenge of retaining customers acquired with free trials. J Mark Res 52:217–234. https://doi.org/10.1509/jmr.12.0160

Dey I, Pratap V (2023) A comparative study of SMOTE, borderline-SMOTE, and ADASYN oversampling techniques using different classifiers. In: 2023 3rd international conference on smart data intelligence (ICSMDI), pp 294–302

Fader PS, Hardie BGS, Lee KL (2005) "Counting your customers" the easy way: an alternative to the pareto/NBD model. Mark Sci 24:275–284. https://doi.org/10.1287/mksc.1040.0098

Farquad MAH, Ravi V, Raju SB (2014) Churn prediction using comprehensible support vector machine: an analytical CRM application. Appl Soft Comput 19:31–40. https://doi.org/10.1016/j.asoc.2014.01.031

Fernandez A, Garcia S, Herrera F, Chawla NV (2018) SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 61:863–905. https://doi.org/10.1613/jair.1.11192

Feyen E, Frost J, Gambacorta L et al (2021) Fintech and the digital transformation of financial services: implications for market structure and public policy. BIS Papers 117. https://www.bis.org/publ/bppdf/bispap117.htm

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22

Galar M, Fernandez A, Barrenechea E et al (2012) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C (appl Rev) 42:463–484. https://doi.org/10.1109/TSMCC.2011.2161285

García S, Luengo J, Herrera F (2014) Data preprocessing in data mining. Springer, Berlin

Geiler L, Affeldt S, Nadif M (2022) A survey on machine learning methods for churn prediction. Int J Data Sci Anal. https://doi.org/10.1007/s41060-022-00312-5

Gordini N, Veglio V (2017) Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Ind Mark Manag 62:100–107. https://doi.org/10.1016/j.indmarman.2016.08.003

Gür Ali Ö, Arıtürk U (2014) Dynamic churn prediction framework with more effective use of rare event data: the case of private banking. Expert Syst Appl 41:7889–7903. https://doi.org/10.1016/j.eswa.2014.06.018

He H, Ma Y (2013) Imbalanced learning: foundations, algorithms, and applications. Wiley, New York

He H, Bai Y, Garcia EA, Li S (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. Hong Kong

He B, Shi Y, Wan Q, Zhao X (2014) Prediction of customer attrition of commercial banks based on SVM model. Procedia Comput Sci 31:423–430. https://doi.org/10.1016/j.procs.2014.05.286

Heldt R, Silveira CS, Luce FB (2021) Predicting customer value per product: from RFM to RFM/P. J Bus Res 127:444–453. https://doi.org/10.1016/j.jbusres.2019.05.001

Huang B, Kechadi MT, Buckley B (2012) Customer churn prediction in telecommunications. Expert Syst Appl 39:1414–1425. https://doi.org/10.1016/j.eswa.2011.08.024

Hvitfeldt E (2022) themis: Extra recipes steps for dealing with unbalanced data. CRAN R package version 1.0.0: https://CRAN.R-project.org/package=themis

Jassim MA, Abdulwahid SN (2021) Data mining preparation: process, techniques and major issues in data analysis. IOP Conf Ser: Mater Sci Eng 1090:012053. https://doi.org/10.1088/1757-899X/1090/1/012053

Keramati A, Ghaneei H, Mirmohammadi SM (2016) Developing a prediction model for customer churn from electronic banking services using data mining. Financ Innov 2:10. https://doi.org/10.1186/s40854-016-0029-6

Khoh WH, Pang YH, Ooi SY et al (2023) Predictive churn modeling for sustainable business in the telecommunication industry: optimized weighted ensemble machine learning. Sustainability 15:8631. https://doi.org/10.3390/su15118631

Kou G, Olgu Akdeniz Ö, Dinçer H, Yüksel S (2021a) Fintech investments in European banks: a hybrid IT2 fuzzy multidimensional decision-making approach. Financ Innov 7:39. https://doi.org/10.1186/s40854-021-00256-y

Kou G, Xu Y, Peng Y et al (2021b) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis Support Syst 140:113429. https://doi.org/10.1016/j.dss.2020.113429

Kuhn M (2022) tune: Tidy tuning tools. CRAN R package version 1.0.1. https://CRAN.R-project.org/package=tune

Kuhn M, Johnson K (2019) Feature engineering and selection: a practical approach for predictive models. CRC Press, Boca Raton

Lähteenmäki I, Nätti S (2013) Obstacles to upgrading customer value-in-use in retail banking. Int J Bank Mark 31:334–347. https://doi.org/10.1108/IJBM-11-2012-0109

Lahmiri S, Bekiros S, Giakoumelou A, Bezzina F (2020) Performance assessment of ensemble learning systems in financial data classification. Int J Intell Syst Account Finance Manag 27:3–9. https://doi.org/10.1002/isaf.1460

Lazari N, Machado G (2021) The future of banking: growing digitalization of Brazil's financial system will foster efficiency and intensify competition. S&P Global Ratings

Lemmens A, Croux C (2006) Bagging and boosting classification trees to predict churn. J Mark Res 43:276–286. https://doi.org/10.1509/jmkr.43.2.276

Lemmens A, Gupta S (2020) Managing churn to maximize profits. Mark Sci 39:956–973. https://doi.org/10.1287/mksc.2020.1229

Li T, Kou G, Peng Y, Yu PS (2022) An integrated cluster detection, optimization, and interpretation approach for financial data. IEEE Trans Cybern 52:13848–13861. https://doi.org/10.1109/TCYB.2021.3109066

Lin W-C, Tsai C-F, Hu Y-H, Jhang J-S (2017) Clustering-based undersampling in class-imbalanced data. Inf Sci 409–410:17–26. https://doi.org/10.1016/j.ins.2017.05.008

Livne G, Simpson A, Talmor E (2011) Do customer acquisition cost, retention and usage matter to firm performance and valuation? J Bus Financ Acc 38:334–363. https://doi.org/10.1111/j.1468-5957.2010.02229.x

Megahed FM, Chen Y-J, Megahed A et al (2021) The class imbalance problem. Nat Methods 18:1270–1272. https://doi.org/10.1038/s41592-021-01302-4

Murinde V, Rizopoulos E, Zachariadis M (2022) The impact of the FinTech revolution on the future of banking: opportunities and risks. Int Rev Financ Anal 81:102103. https://doi.org/10.1016/j.irfa.2022.102103

Mutanen T, Nousiainen S, Ahola J (2010) Customer churn prediction—a case study in retail banking. In: Data mining for business applications, pp 77–83. https://doi.org/10.3233/978-1-60750-633-1-77

Pousttchi K, Dehnert M (2018) Exploring the digitalization impact on consumer decision-making in retail banking. Electron Markets 28:265–286. https://doi.org/10.1007/s12525-017-0283-0

Pyle D (1999) Data preparation for data mining (The Morgan Kaufmann series in data management systems), Book&CD-ROM 1st. Morgan Kaufmann, Burlington

R Core Team (2022) R: a language and environment for statistical computing. R Project. https://www.R-project.org/

Reichheld FF, Sasser WE (1990) Zero defections: quality comes to services. Harvard business review. https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Sammut C, Webb GI (eds) (2010) Data preprocessing. In: Encyclopedia of machine learning. Springer, Boston, MA, pp 260–260

Snoek J, Larochelle H, Adams RP (2012) Practical bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems 25 (NIPS 2012). NeurIPS proceedings

Sofaer HR, Hoeting JA, Jarnevich CS (2019) The area under the precision-recall curve as a performance metric for rare binary events. Methods Ecol Evol 10:565–577. https://doi.org/10.1111/2041-210X.13140

Sun Y, Wong AKC, Kamel MS (2009) Classification of imbalanced data: a review. Int J Patt Recogn Artif Intell 23:687–719. https://doi.org/10.1142/S0218001409007326

Tékouabou SCK, Gherghina ŞC, Toulni H, Neves Mata P, Mata MN, Martins JM (2022) A Machine Learning Framework towards Bank Telemarketing Prediction. J Risk Financ Manag 15:269. https://doi.org/10.3390/jrfm15060269

Triguero I, Derrac J, Garcia S, Herrera F (2012) A taxonomy and experimental study on prototype generation for nearest neighbor classification. IEEE Trans Syst, Man, Cybern C 42:86–100. https://doi.org/10.1109/TSMCC.2010.2103939

Victoria AH, Maragatham G (2021) Automatic tuning of hyperparameters using Bayesian optimization. Evol Syst 12:217–223. https://doi.org/10.1007/s12530-020-09345-2

Weiss GM (2004) Mining with rarity: a unifying framework. SIGKDD Explor Newsl 6:7–19. https://doi.org/10.1145/1007730.1007734

Xie Y, Li X, Ngai EWT, Ying W (2009) Customer churn prediction using improved balanced random forests. Expert Syst Appl 36:5445–5449. https://doi.org/10.1016/j.eswa.2008.06.121

Zhang J, Mani I (2003) KNN Approach to Unbalanced Data Distributions: a case study involving information extraction. In: Proceeding of international conference on machine learning. ICML United States, Washington DC

Zhang Y, Bradlow ET, Small DS (2015) Predicting customer value using clumpiness: from RFM to RFMC. Mark Sci 34:195–208. https://doi.org/10.1287/mksc.2014.0873

Zhao J, Dang X-H (2008) Bank Customer churn prediction based on support vector machine: taking a commercial bank's VIP customer churn as the example. In: 2008 4th international conference on wireless communications, networking and mobile computing. IEEE, Dalian, China, pp 1–4

Zhao H, Zuo X, Xie Y (2022) Customer churn prediction by classification models in machine learning. In: 2022 9th international conference on electrical and electronics engineering (ICEEE). pp 399–407

Zheng A, Casari A (2018) Feature engineering for machine learning: principles and techniques for data scientists. O'Reilly Media, Inc

Zhu B, Baesens B, Backiel A, vanden Broucke SKLM (2018) Benchmarking sampling techniques for imbalance learning in churn prediction. J Oper Res Soc 69:49–65. https://doi.org/10.1057/s41274-016-0176-1

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B (stat Methodol) 67:301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Publisher's Note