electronics

MDPI

Article

Improving Churn Detection in the Banking Sector: A Machine Learning Approach with Probability Calibration Techniques

Alin-Gabriel Văduva [1,2,*], Simona-Vasilica Oprea [1], Andreea-Mihaela Niculae [1,2], Adela Bâra [1] and Anca-Ioana Andreescu [1]

[1] Department of Economic Informatics and Cybernetics, Bucharest University of Economic Studies, 010374 Bucharest, Romania; simona.oprea@csie.ase.ro (S.-V.O.); andreea.niculae@csie.ase.ro (A.-M.N.); bara.adela@ie.ase.ro (A.B.); anca.andreescu@ie.ase.ro (A.-I.A.)

[2] Doctoral School of Economic Informatics, Bucharest University of Economic Studies, 010374 Bucharest, Romania

* Correspondence: alin.vaduva@csie.ase.ro

Abstract: Identifying and reducing customer churn have become a priority for financial institutions seeking to retain clients. Our research focuses on customer churn rate analysis using advanced machine learning (ML) techniques, leveraging a synthetic dataset sourced from the Kaggle platform. The dataset undergoes a preprocessing phase to select variables directly impacting customer churn behavior. SMOTETomek, a hybrid technique that combines oversampling of the minority class (churn) with SMOTE and the removal of noisy or borderline instances through Tomek links, is applied to balance the dataset and improve class separability. Two cutting-edge ML models are applied—random forest (RF) and the Light Gradient-Boosting Machine (LGBM) Classifier. To evaluate the effectiveness of these models, several key performance metrics are utilized, including precision, sensitivity, F1 score, accuracy, and Brier score, which helps assess the calibration of the predicted probabilities. A particular contribution of our research is on calibrating classification probabilities, as many ML models tend to produce uncalibrated probabilities due to the complexity of their internal mechanisms. Probability calibration techniques are employed to adjust the predicted probabilities, enhancing their reliability and interpretability. Furthermore, the Shapley Additive Explanations (SHAP) method, an explainable artificial intelligence (XAI) technique, is further implemented to increase the transparency and credibility of the model's decision-making process. SHAP provides insights into the importance of individual features in predicting churn, providing knowledge to banking institutions for the development of personalized customer retention strategies.

Keywords: churn; banking sector; classification; probability calibration; explainable AI (XAI)

check for updates

Citation: Văduva, A.-G.; Oprea, S.-V.; Niculae, A.-M.; Bâra, A.; Andreescu, A.-I. Improving Churn Detection in the Banking Sector: A Machine Learning Approach with Probability Calibration Techniques. *Electronics* **2024**, *13*, 4527. https://doi.org/10.3390/electronics13224527

Academic Editors: Muhammad Shahid Anwar, Maria Torres Vega and Muhammad Salman Pathan

Received: 21 October 2024
Revised: 11 November 2024
Accepted: 13 November 2024
Published: 18 November 2024

1. Introduction

The banking sector is marked by fierce competition, with customers having a wide range of choices, and the challenge of retaining customers and employees is one of the critical issues for such a business. A significant obstacle to this goal is the phenomenon of churn, expressed by the customer churn rate, an indicator aimed at measuring the proportion of customers who choose to terminate their relationship with the bank within a certain time frame [1]. Understanding and preventing churn is, on the one hand, a factor contributing to establishing a solid customer base but also to maximizing profits and improving the long-term reputation of financial institutions. Emerging technologies and data analysis offer opportunities for banks to address this challenge. ML algorithms have become essential tools in predicting customer behavior [2], offering the ability to identify trends that may be difficult to detect using classical methods. By using ML techniques, banks may develop internal models that anticipate churn potential based on a multitude of variables, from demographic data to transactional history and service usage behavior.

In this paper, we focus on calibrating the probabilities returned by the random forest (RF) and Light Gradient-Boosting Machine (LGBM) Classifier algorithms to obtain predictive models capable of detecting churn behavior for bank customers using synthetic data extracted from the Kaggle (https://www.kaggle.com/datasets/shubhammeshram579/bank-customer-churn-prediction/data, accessed on 7 November 2024) platform. RF is one of the state-of-the-art learning algorithms based on ensemble methods of decision trees. It is known for its accuracy and ability to adjust to the phenomenon of overfitting on the data it uses. It is also known for its ability to handle both categorical and continuous spectrum data properly. In the financial domain, it contributes to improving predictions of customer behavior and identifying risk factors [3], as well as identifying transactions with potentially high risk. The LGBM model is also an advanced classification method, notable for its efficiency in processing large datasets, having good runtime speeds and providing precise results. This gradient-based model is optimal for risk analysis [4], fraud detection, and marketing campaign optimization due to its ability to learn from complex and unevenly distributed data. It also has internal capabilities to handle null or infinite values, which facilitates the data preprocessing step. On the other hand, ensemble models are based on aggregating predictions from multiple decision trees to make the final prediction. The probabilities offered by the RF and LGBM models are generally the average probabilities predicted by each internal decision tree of the two algorithms. Metrics such as the Brier score [5] or logarithmic loss (log-loss) are relevant for models with correctly calibrated probabilities. For the Brier score, a value closer to 0 indicates a very good classification, while a value close to 1 highlights a poor classifier.

SHAP (Shapley Additive Explanations) analysis emerges as a transformative approach in enhancing the interpretability of complex ML models used in the banking sector. By decomposing the prediction of a model into the sum of effects of each feature, SHAP provides a detailed understanding of how each attribute influences a given prediction. This insight is important for refining the algorithms for churn prediction and for ensuring compliance with transparency and fairness requirements of the model. Moreover, SHAP analysis facilitates the identification of key factors contributing to customer retention, enabling targeted interventions that can more effectively mitigate churn. In our research methodology, we further employ SMOTETomek, a hybrid technique that combines oversampling of the minority class with SMOTE and the removal of noisy or borderline instances through Tomek links, to balance the dataset and improve class separability.

The contribution of our research is the development and calibration of predictive models for customer churn in the banking sector using cutting-edge ML algorithms, specifically RF and LGBM classifiers. We aim to create balanced models that predict customer churn based on open-source data, with a focus on calibrating the probabilities returned by these models to improve their accuracy. The use of SHAP analysis enhances the interpretability of the models, providing insights into how specific features influence predictions, which may further help banks target interventions to reduce churn. Additionally, our research offers a solution for preventing customer churn in the banking sector, improving transparency in the financial sector by ensuring model fairness and regulatory compliance.

Despite extensive studies on customer churn prediction using ML models, there remains a significant need for models that not only predict churn accurately but also provide calibrated, reliable probabilities that are interpretable for financial institutions. Many existing models lack probability calibration, leading to unreliable probability outputs, which may affect decision-making. Additionally, while model accuracy is prioritized, interpretability often remains overlooked, making it challenging for financial institutions to understand the drivers behind churn predictions and effectively utilize insights for targeted customer retention strategies. Our research aims to fill these research gaps by focusing on probability calibration and model interpretability through SHAP analysis, enhancing the usability and transparency of churn prediction models in the banking sector.

By addressing the following research questions (RQ), our research seeks to contribute an interpretable framework for churn prediction that enhances transparency, fairness and actionable insights.

RQ1: How can probability calibration techniques improve the reliability of predicted churn probabilities in ML models (specifically RF and LGBM) used in the banking sector?

RQ2: What are the most influential features affecting customer churn predictions, as identified by the SHAP method, and how can these insights support personalized retention strategies in banking?

RQ3: How does the Brier score, alongside other performance metrics, assess the calibration quality of predicted probabilities for churn, and how does this impact decision-making within financial institutions?

RQ4: What is the comparative performance of RF and LGBM classifiers in predicting customer churn within the banking industry, particularly in terms of accuracy, interpretability and probability calibration?

Our research paper is organized into several sections: an introduction section that highlights the general context of customer churn in the banking sector and the use of advanced ML classification algorithms; a literature review that examines the latest publications in the field; a research methodology outlining the proposed process to enhance the classification approach; a results section presenting the simulations and comparisons conducted; and discussion and conclusion sections summarizing the key findings of the study.

2. Literature Review

A multitude of researchers have aimed to construct models predicting customer churn phenomenon in the business environment, using relevant historical data. The issue firms generally seek to address is related to the feasibility of retaining an existing customer or acquiring a new one: which of these options is more costly? Clearly, retaining loyal customers is more relevant from this perspective. The authors of [6] propose several approaches to study this issue in the telecommunications domain, using stochastic gradient booster, RF, logistic regression (LR), and the nearest K-neighbors (K-nearest neighbors-KNN) algorithms, obtaining accuracy percentages of 83.9%, 82.6%, 82.9%, and 78.1%, respectively. In [7], the authors proposed a gravitational search method to find the most important variables for modeling, also applying cross-validation methods to reduce overfitting effects, and finally employing ensemble methods by aggregating predictions from multiple different algorithms. For AdaBoost and eXtreme Gradient Boosting (XGB) methods, accuracy percentages of 81.71% and 80.8% were obtained, with an AUC score of 84% for both models.

In another research [8], the authors utilized a dataset extracted from the Kaggle platform describing credit-card-holding bank customers. Apart from classical supervised learning methods, they also used a methodology based on customer segmentation, considering techniques from the unsupervised learning spectrum to achieve this segmentation, specifically the K-means clustering method. Based on these newly created segments, various methods such as LR, RF, and support vector machine (SVM) were applied, obtaining the highest accuracy of around 97.25%. However, most churn rate study methods have been limited to modeling the phenomenon of customer loss in the following time interval. The authors of [9] present a dynamic approach to predicting customer loss based on anticipated time, studied on a database of clients of a European private bank. The data format is panel-type, which also implies a time-series-based approach, providing dynamism to the instances in the training sets used. Furthermore, using a single algorithm independently is often not sufficient in studying customer churn rate. By integrating supervised learning techniques based on classical algorithms such as SVM and the specific kernel functions of this method with artificial neural networks (ANN), optimal results were obtained, thus creating a hybrid classification model [10] with accuracy values exceeding 97%. A synthesis of classification methods using ML and data mining in e-commerce is further presented in [11], where most of the difficulties in predicting customer churn rate were outlined alongside the evolution of techniques based on deep learning (DL) and those based on classical ML models.

In their exploration of customer churn prediction within the banking sector, the authors of [12] emphasized the utility of SHAP analysis in elucidating the influence of individual predictors in ML models. By integrating SHAP values, the authors significantly enhanced the interpretability of predictive models, allowing for more informed strategic decisions aimed at customer retention. Additionally, in [13], the authors investigated customer churn prediction in the banking sector using a dataset from a Brazilian financial institution. They conducted a comprehensive comparison of various supervised ML algorithms, including decision trees, KNN, elastic net, LR, SVM, and RF. The study finds that RF consistently outperformed other models across several metrics, confirming its suitability for churn prediction in financial settings. Additionally, the analysis identified key predictive features, such as the frequency of financial transactions and loan volume, which are essential for anticipating customer churn and devising proactive retention strategies.

Moreover, in their study on customer churn prediction, the authors of [14] focused on applying ML methods to classify customer portfolios in the banking sector. By evaluating a variety of techniques, such as LR, KNN, SVM, RF, and gradient-boosting machines (GBM), the authors emphasized the effectiveness of ensemble models. Their proposed ROS-voting (RF-GBM) approach outperformed traditional methods, achieving a notable accuracy rate of 95%. It highlighted the potential of hybrid techniques in addressing the challenges of class imbalance and improving the accuracy of churn predictions in the financial domain.

Hard and soft data fusion has also been used to model customer churn in the banking industry, as presented in [15]. This approach combined the use of hard data, which includes financial, behavioral and demographic information, with soft data, such as expert opinions, to enhance the accuracy of churn predictions. The authors employed a decision tree model for hard data and the Dempster–Shafer theory for soft data, ultimately fusing the two types to achieve a more dynamic analysis. The study demonstrated that the fusion of both hard and soft data leads to improved churn rate predictions, offering insights into customer behavior and informing retention strategies.

In the context of subprime auto loans, in [16], a detailed study conducted on predicting customer churn in the subprime auto loan market by comparing two models: a restricted model focusing solely on customer character and a full model incorporating the 4 Cs of lending: capacity, collateral, credit, and character. The study highlighted the importance of including borrower character variables, such as residential stability and employment duration, in predicting churn, especially for subprime borrowers. While the full model, which also considers factors like loan-to-value ratios and credit scores, performed well overall, the character-focused model provided important insights. For instance, customers with longer residential tenure and stable employment were less likely to churn, even when their credit scores were relatively low. This indicates that in subprime lending, assessing character-related factors may be as critical as traditional credit metrics. The findings suggest that financial institutions may improve churn predictions and retention strategies by balancing both quantitative (hard) data and qualitative (soft) customer characteristics. Authors of [17] further explored the challenges posed by heterogeneous data and class imbalance in the context of bank churn prediction. Their study focused on improving prediction accuracy by using the SMOTE technique for data balancing combined with ensemble-based methods. The results demonstrated that RF consistently outperformed other models, achieving an accuracy of 86% and a corresponding F1 score. Importantly, the model has been made interpretable through SHAP values and feature importance analysis, which reveal that customer age had the most significant impact on churn predictions, while other factors like credit card ownership contribute minimally. It is highlighted how balancing imbalanced data significantly improves the performance of ML models in banking churn prediction

A linear discriminant boosting (LD-Boosting) algorithm was also proposed as a novel approach to churn prediction in the banking sector [18]. This method included the strengths of linear discriminant analysis (LDA) combined with the boosting framework to improve predictive accuracy, especially in cases involving highly imbalanced datasets. By emphasizing the most discriminative features in each iteration, LD-Boosting avoids time-consuming numerical optimizations and significantly enhances computational efficiency. Applied to a real-world bank customer dataset, LD-Boosting outperforms traditional algorithms such as ANN, decision trees, and SVM, with improved prediction accuracy. The algorithm's design specifically targets the minority churn class, reducing error rates and achieving more precise predictions, demonstrating its potential to outperform existing methods in customer churn prediction. A hybrid DL-based approach for customer churn prediction, focusing on feature selection using the Archimedes Optimization Algorithm (AOA), was proposed in [19]. Their model, applied in the telecom industry, integrated convolutional neural networks with autoencoders (CNN-AE) for the churn prediction process and employs a thermal equilibrium optimization technique for hyperparameter tuning. By optimizing both the feature set and the DL model's parameters, the proposed method achieved a remarkable accuracy of 94.65%. The study highlighted that combining bio-inspired optimization algorithms with DL significantly improves churn prediction performance, making this technique highly effective for large-scale datasets in customer retention strategies.

Combinations of DL methods have been employed in studying the customer churn problem. The authors of [20] present a combined DL network model aimed at predicting customer churn in the banking sector. The proposed model operates on two levels: Level 0 consists of three DL neural networks (DNN), while Level 1 utilizes a LR model to combine the outputs of the base models. The ensemble approach is designed to harness the strengths of different DNN architectures, resulting in superior predictive performance. With this method, the model achieves an impressive accuracy of 96.6% along with precision, recall, and F1 scores exceeding 90%.

A framework based on ANN for predicting customer churn in the banking sector, built on customer churn data containing variables such as credit scores, age, and account balances was also proposed in [21]. The ANN model, consisting of one hidden layer and a sigmoid-activated output layer, achieved an accuracy of 87%, with precision, recall, and F1 scores of 87%, 98%, and 92%, respectively. By implementing a straightforward ANN architecture, the authors demonstrated how DL can outperform traditional ML models such as KNN, XGB, and decision trees, which were used in previous studies on the same dataset.

Probability calibration techniques have been increasingly applied to enhance the reliability of ML models in churn prediction. Ref. [22] introduced a method to calibrate RF for probability estimation, addressing the challenge of providing accurate predictions across different centers or time periods. They propose two strategies for updating RF: the first was based on Elkan's approach, applicable to any ML method that yields consistent probability estimates; the second is an LR-based approach specifically tailored for RF. Through simulation studies, they demonstrated that while both methods improve probability estimates, the LR-based recalibration is preferable when covariate distributions differ between datasets. Their method ensured that the RF model produced calibrated probabilities, thus offering a more precise prediction tool that can be effectively used in dynamic environments such as customer churn prediction in the banking sector.

A comprehensive comparison of various probability calibration techniques for ML models is provided in [23]. In total, 10 calibration methods were assessed—including logistic calibration, beta calibration, Platt scaling, and isotonic regression—across multiple ML algorithms, such as elastic net, gradient boosting, RF, and SVM. Using both Monte Carlo simulations and real-world datasets, the authors demonstrated that regression-based approaches, specifically logistic and beta calibrations, consistently yielded the most reliable probability estimates. The research also highlights the advantages of beta calibration due to its flexibility, as it estimates separate parameters for both logit-transformed probabilities. The issue of calibration drift in ML models used for clinical decision support, specifically focusing on acute kidney injury prediction is presented in [24]. The authors compared the performance of various ML methods, including RF and ANN, with traditional regression models. Over a nine-year validation period, they found that ML models, particularly RF, maintain calibration better than regression-based approaches. Calibration drift, or the tendency of models to overpredict risk over time, was more pronounced in LR models. It emphasized the need for efficient recalibration protocols to maintain model accuracy and user confidence in dynamic settings such as banking churn prediction, where probabilities must be consistently reliable for decision-making.

In the context of churn modeling, a multi-level stacking model tailored to the banking sector has shown promise in enhancing predictive accuracy. In [25], the authors applied this approach by integrating four ML algorithms: K-nearest neighbor (KNN), XGBoost, RF, and support vector machine (SVM), at the first level, followed by higher-level models such as logistic regression, recurrent neural networks (RNN), and deep neural networks (DNN). This layered structure achieved impressive results, including an accuracy of 91.08% and a ROC-AUC score of 98%. Additionally, in [26], an experimental analysis was conducted to understand the impact of different hyperparameters on DNN for churn prediction within the banking sector. The research explored various configurations, including activation functions, batch sizes and training algorithms, to optimize model performance. Findings revealed that using a rectifier function in the hidden layers alongside a sigmoid function in the output layer yielded the best results for DNN. Furthermore, the research highlighted the importance of smaller batch sizes and the RMSProp algorithm in enhancing accuracy, providing valuable heuristic knowledge for practitioners tuning hyperparameters for churn modeling in banking.

Recognizing the importance of effective churn prediction, customer behavior was analyzed in [27] using a combination of ML algorithms to enhance predictive accuracy in the banking sector. The study employed ensemble methods—RF, XGBoost, and LGBM—to assess customer data and pinpoint stages where churn risks are heightened. By focusing on key features and reducing the dimensionality of the data, the approach achieved an accuracy of up to 89%, highlighting the advantage of ensemble techniques in refining predictions and addressing customer retention challenges.

Targeting the youth segment in retail banking, authors of [28] developed an ensemble model to predict churn by focusing on critical factors specific to young customers. Using ML techniques, including the ExtraTreesClassifier, the research identified drivers like mobile banking accessibility, zero-interest loans, and ATM service quality as significant contributors to churn. The final model demonstrated good performance, achieving 92% accuracy and an AUC of 91.88%. This research emphasizes the importance of catering to the expectations of young customers to mitigate churn risk.

Regarding the research on explainable AI, comprehensive research on predicting bank creditworthiness through ML is presented in [29]. It employed gradient-boosting and RF models, achieving notable performance metrics, with accuracy scores reaching 82.49% for gradient-boosting and 82.39% for RF. These models were further enhanced using SHAP values to determine feature importance, which highlighted repayment status and credit limits as key predictors of credit risk. By visualizing these feature contributions, the authors provided decision-makers with clear, interpretable insights into model predictions, aligning the approach with regulatory standards for responsible AI.

In an effort to improve customer churn prediction within the banking sector, ref. [14] examined a comprehensive classification approach leveraging both traditional and ensemble machine learning models. The study tested algorithms such as logistic regression, KNN, SVM, and decision trees, with RF and gradient-boosting as core components of a robust ensemble model. To address the class imbalance common in churn datasets, the researchers employed random oversampling (ROS), which improved prediction metrics. The proposed ROS-Voting ensemble model, combining RF and LGBM, achieved an accuracy of 95%, underscoring the value of ensemble techniques in enhancing predictive power and accuracy for identifying high-risk customer groups.

Fuzzy rule-based approaches for modeling customer churn are employed in [30]. Authors applied this approach within the telecommunications industry, focusing on identifying at-risk customers to improve retention. Using real-world data from a large mobile operator in Poland, the study leveraged Mamdani and Sugeno fuzzy inference models to address data uncertainty and improve interpretability. Key usage and payment features were extracted through fuzzy clustering and decision trees, leading to a concise rule set with high prediction accuracy. This work demonstrates the effectiveness of customized churn models in handling complex customer data, an approach that can be adapted to the banking sector to enhance predictive accuracy and inform retention strategies in similar customer-centric settings. Additionally, Ref. [31] addresses the issue of class imbalance in customer churn prediction by developing a focal-loss-enhanced LightGBM model. Their approach specifically targets challenging cases in churn prediction by adjusting the loss function to focus more on the minority class (i.e., churned customers). Using a Kaggle dataset of 10,127 credit card users, they compared the performance of their FocalLoss_LightGBM model to other popular machine learning models, such as SVM, RF, XGBoost, and the standard LGBM. The results demonstrated that the FocalLoss_LightGBM model provided improved stability and accuracy, particularly in identifying churned customers, thus offering a promising solution for customer retention efforts in banking where class imbalance is often a challenge.

Ensemble methods, alongside XAI techniques, provide a great framework for increasing credibility in the customer churn area. In this context, Ref. [32] explores customer churn prediction in the telecommunications industry through ensemble learning models, utilizing algorithms such as decision trees, boosted trees, and RF. The study highlights the value of XAI techniques, specifically LIME and SHAP, to enhance model interpretability, making predictive insights more accessible for decision-makers. With an RF model achieving 91.66% accuracy, the research emphasizes the strategic advantage of churn prediction in retaining customers.

Table 1 compares the various previous studies on customer churn prediction, summarizing the objective, methods, and main results.

Table 1. Comparative analysis of the previous research.

Ref.	Objective	Methods	Main Results
[6]	Predicting customer churn in the telecommunications domain	Stochastic gradient booster, RF, LR, KNN	Accuracy: 83.9% (SGD), 82.6% (RF), 82.9% (LR), 78.1% (KNN)
[7]	Improve churn prediction using variable selection and ensemble methods	Gravitational search method, AdaBoost, XGB, cross-validation	Accuracy: 81.71% (AdaBoost), 80.8% (XGB), AUC: 84% for both models
[8]	Segment bank customers for churn prediction	K-means clustering, LR, RF, SVM	Accuracy: 97.25% (LR, RF, SVM)
[9]	Dynamic churn prediction based on time-series data in private banking	Supervised learning: SVM, ANN	Hybrid model achieved accuracy > 97%
[10]	Improve churn prediction with hybrid ML models	SVM with ANN	Optimal results with accuracy > 97%
[11]	Synthesis of ML methods for churn prediction in e-commerce	Various ML techniques, including DL and classical models	Outlines difficulties in predicting churn and evolution of techniques
[12]	Interpretability of ML models in banking churn prediction using SHAP values	SHAP analysis with ML models	Enhanced interpretability, aiding customer retention strategies
[13]	Comparison of supervised ML algorithms for churn prediction in banking	Decision trees, KNN, elastic net, LR, SVM, RF	RF outperformed other models; identified key features like financial transaction frequency and loan volume
[14]	Churn prediction using ensemble models in banking	LR, KNN, SVM, RF, gradient-boosting machines (GBM), ROS-voting (RF-GBM)	ROS-voting approach achieved accuracy of 95%
[15]	Use of hard and soft data fusion in banking churn prediction	Decision tree, Dempster–Shafer theory	Fusion of data types improved churn predictions
[16]	Churn prediction in subprime auto loans	Restricted model (customer character) vs full model (4 Cs of lending)	Both models useful; character-focused model provided critical insights on customer retention

Table 1. *Cont.*

Ref.	Objective	Methods	Main Results
[17]	Addressing class imbalance in churn prediction using SMOTE	RF, ensemble methods, SMOTE technique	RF achieved 86% accuracy– customer age was the most significant factor
[18]	Novel LD-Boosting algorithm for churn prediction in banking	Linear discriminant analysis (LDA) with boosting	LD-Boosting outperformed traditional algorithms, reducing error rates in churn prediction
[19]	DL-based churn prediction with bio-inspired optimization	Convolutional neural networks with autoencoders (CNN-AE), Archimedes Optimization Algorithm (AOA)	Achieved 94.65% accuracy in telecom churn prediction
[20]	Combined DL network model for churn prediction in banking	Level 0 (3 DNNs), Level 1 (LR model)	Accuracy of 96.6%, precision, recall, and F1 > 90%
[21]	ANN model for churn prediction in banking	ANN with 1 hidden layer, sigmoid-activated output layer	Accuracy: 87%, precision: 87%, recall: 98%, F1: 92%
[22]	Probability calibration for RF in churn prediction	Elkan's approach, LR-based recalibration	LR-based recalibration improved probability estimates in dynamic settings
[23]	Comprehensive comparison of probability calibration methods for churn prediction	Logistic calibration, beta calibration, Platt scaling, isotonic regression	Beta calibration yielded most reliable estimates
[24]	Calibration drift in ML models over time	RF, ANN, regression-based models	RF maintained calibration better over time than regression models

Compared to previous studies, our research addresses several identified gaps and adds contributions in customer churn prediction, specifically within the banking sector, focusing on: (1) probability calibration for improved predictive reliability—unlike much previous work that primarily emphasizes model accuracy, our research highlights the importance of probability calibration to enhance the reliability of churn predictions. By calibrating the probabilities returned by RF and LGBM models, we improve the interpretability and trustworthiness of predictions; (2) enhanced model interpretability through SHAP analysis—though some previous works use SHAP for feature interpretation, our research applies SHAP to uncover the influence of individual predictors on churn probabilities, thereby enhancing model transparency and supporting compliance with regulatory requirements; (3) combination of RF and LGBM for better performance—using the two advanced ensemble-based algorithms, we provide a comparative analysis of their performance in the context of churn prediction, specifically focusing on accuracy, probability calibration, and interpretability; (4) detailed analysis of key performance metrics—in addition to standard metrics like accuracy and F1 score, our research further evaluates calibration-specific metrics such as the Brier score, which measures the accuracy of predicted probabilities.

3. Research Methodology

In this section, we describe the methodology for creating classification models of existing customers in the dataset, which describes 10,000 users of banking services from Spain, France, and Germany. The dataset is sourced from the Kaggle platform and is frequently referenced in customer churn research. Despite its broad usage, the dataset lacks verified documentation from any specific bank, limiting the ability to attribute it to a single institution. However, it is widely accepted in both academic and industry research as a suitable basis for evaluating churn prediction methodologies.

Due to stringent privacy regulations and compliance requirements within the banking industry, access to real-world, institution-specific customer data is difficult to obtain, making it challenging for researchers to utilize proprietary datasets for academic purposes. Consequently, publicly available datasets, like the one employed in our study, offer a foundation for developing and validating methodologies. While not directly representative of proprietary banking data, this dataset enables exploratory analysis and demonstrates methodological approaches that could later be adapted to more context-specific datasets. The features and modeling techniques used in this study, such as financial ratios and demographic characteristics, align closely with real-world applications of churn prediction in banking. Furthermore, while this dataset lacks time-series data, the approach demonstrated here could be enhanced with a panel data structure for longitudinal analysis, enabling a deeper examination of temporal patterns in customer behavior.

Publicly available datasets may contain biases due to their construction for broad accessibility and usability. To address this possibility, we conduct an in-depth exploratory data analysis (EDA), inspecting feature distributions, detecting class imbalances, and identifying patterns that could suggest data augmentation. Our analysis does not reveal any anomalies that might indicate synthetic bias characteristics. Additionally, we employ SMOTETomek to balance classes rigorously, enhancing the robustness and generalizability of our findings.

The methodology is graphically represented in Figure 1.

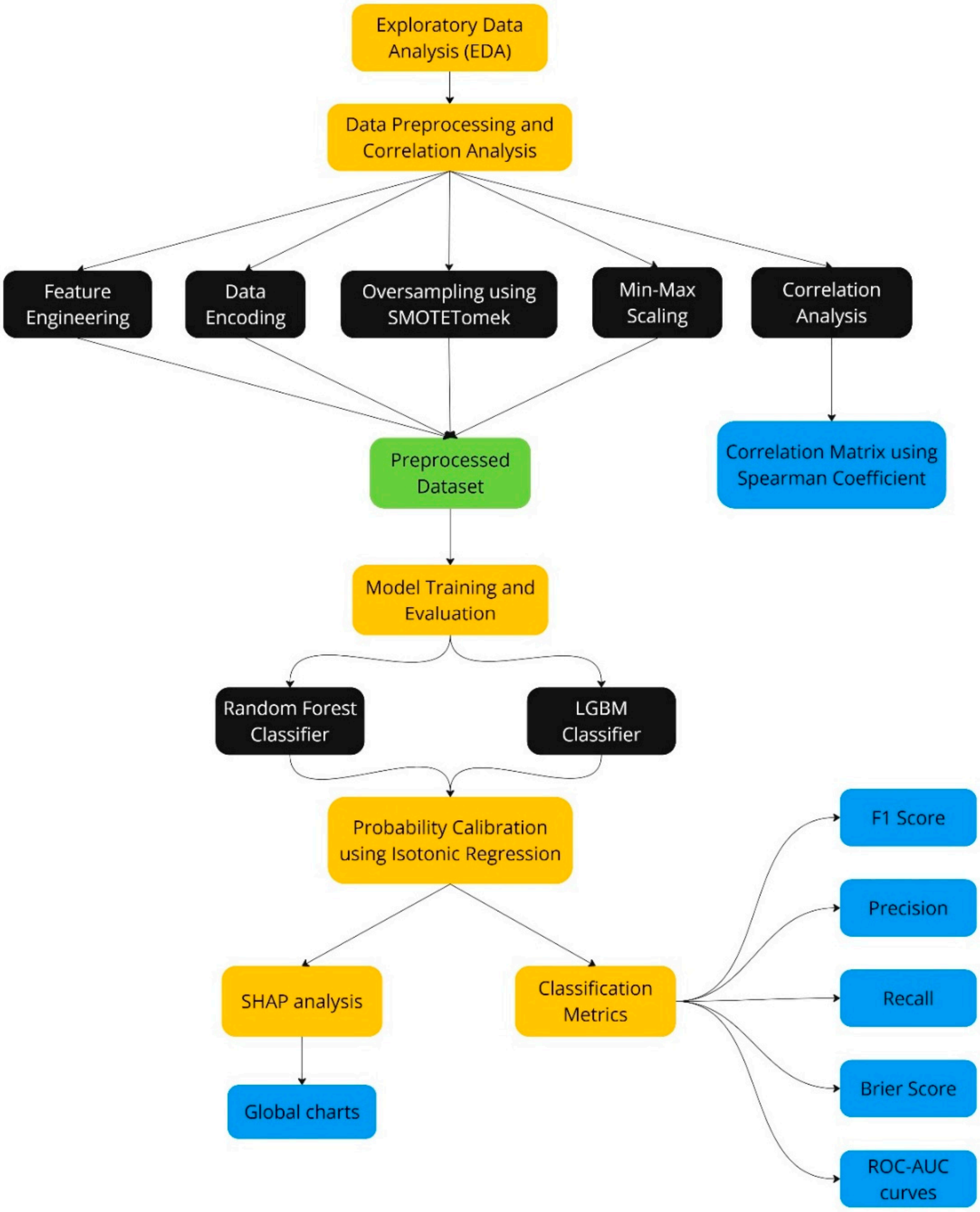

Exploratory Data Analysis (EDA)
Data Preprocessing and Correlation Analysis
Feature Engineering
Data Encoding
Oversampling using SMOTETomek
Min-Max Scaling
Correlation Analysis
Preprocessed Dataset
Correlation Matrix using Spearman Coefficient
Model Training and Evaluation
Random Forest Classifier
LGBM Classifier
Probability Calibration using Isotonic Regression
SHAP analysis
Classification Metrics
Global charts
F1 Score
Precision
Recall
Brier Score
ROC-AUC curves

Figure 1. Flowchart of the research methodology framework.

3.1. Exploratory Data Analysis (EDA)

The process begins with EDA to extract the most important information from raw data. Through this stage, we analyze the distributions of variables, whether our studied problem is balanced or not, by highlighting the balance of values of the dependent variable from the dataset. Additionally, we analyze the distribution of lost customers both geographically at the level of the three presented countries and by the gender of individuals.

3.2. Data Preprocessing and Correlation Analysis

In this section, we present the key steps involved in preparing the data for predictive modeling. These steps are essential to ensure that the dataset is properly structured, balanced and transformed to maximize the performance of the ML models. The preprocessing workflow includes multiple phases, beginning with feature engineering, followed by data encoding techniques and the application of oversampling methods to address class imbalance, and finally, a detailed correlation analysis to evaluate the relationships between variables.

3.2.1. Feature Engineering

Feature engineering is an important step in predictive modeling, as it involves the creation of new variables that capture latent patterns in the dataset. These variables are derived through mathematical transformations of the existing features, which are expected to improve the model's capacity to predict customer churn. Three derived features—BalanceToSalaryRatio, BalanceToProductRatio, and TenureToAge—are created.

BalanceToSalaryRatio measures the proportion of a customer's balance relative to their annual salary. This feature reflects the financial capacity of a customer to absorb economic shocks and their potential dependency on the bank's financial services. The formula for this ratio is defined as:

$$BalanceToSalary = \frac{Balance}{Salary} \tag{1}$$

where *Balance* represents the total amount of funds held by the customer in their account(s) and *Salary* denotes the annual income of the customer.

A higher value of this ratio may indicate a stronger financial dependence on the bank, which could influence the likelihood of churn, particularly in response to economic changes or alterations in banking services.

The second derived feature, *BalanceToProductRatio*, is designed to assess the relationship between a customer's total balance and the number of banking products they utilize. This ratio provides insights into the degree of a customer's engagement with the bank's offerings, reflecting both financial trust and product diversity. Mathematically, it is expressed as:

$$BalanceToProductRatio = \frac{Balance}{Number\ of\ products} \tag{2}$$

where *Number of products* refers to the total number of banking services or products the customer uses, such as savings accounts, loans, or credit cards.

A higher value in this ratio could signal concentrated financial engagement in fewer products, which may be associated with a greater risk of churn if the customer's reliance on these products changes.

Lastly, the *TenureToAge* feature quantifies the customer's tenure with the bank relative to their age. This ratio serves as an index of loyalty, with the assumption that customers who have had a longer relationship with the bank, relative to their age, are less likely to churn. The formula for *TenureToAge* is:

$$TenureToAge = \frac{Tenure}{Age} \tag{3}$$

where *Tenure* denotes the number of years the customer has been a client of the bank and *Age* is the customer's current age. A higher value suggests an earlier engagement with the bank and a longer, potentially more stable relationship, which could correlate with a reduced likelihood of customer churn.

3.2.2. Data Encoding

To ensure the categorical variables are suitable for ML models, appropriate encoding techniques were applied. Two methods—binary encoding and label encoding—were employed to transform the categorical data into numerical formats.

For the gender variable, binary encoding was implemented, in which male customers were assigned a value of 1 and female customers a value of 0. The Geography variable, consisting of three distinct categories (Spain, France, and Germany), was transformed using label encoding. Label encoding assigns a unique integer to each category, transforming the nominal data into numeric form. In this case, Spain, France, and Germany were encoded as 0, 1, and 2, respectively.

Furthermore, binary variables such as HasCrCard and IsActiveMember were also transformed. To enhance the model's ability to detect inverse relationships, the value 0 was replaced with -1, allowing the models to recognize when certain conditions (such as not having a credit card or not being an active member) might inversely affect the likelihood of churn.

3.2.3. Oversampling Using SMOTETomek

Given the inherent imbalance in the target variable (churn) where the minority class (customers who churn) is underrepresented, the oversampling technique known as SMOTETomek is employed. It is a hybrid technique that combines Synthetic Minority Oversampling Technique (SMOTE) with Tomek links to both generate synthetic samples for the minority class and remove overlapping instances between classes.

The SMOTE algorithm synthesizes new samples by interpolating between existing minority class examples and their nearest neighbors. Mathematically, this process is represented by:

$$x_{new} = x_{minority} + \lambda \cdot \left(x_{neighbor} - x_{minority}\right) \tag{4}$$

where $x_{minority}$ is a minority class instance, $x_{neighbor}$ is one of its k-nearest neighbors and λ is a random value in the range [0, 1].

The new synthetic instance, x_{new}, lies along the line segment connecting $x_{minority}$ and $x_{neighbor}$ increasing the representation of the minority class.

In addition, Tomek links are identified and removed to enhance class separability. A pair of instances $\left(x_i, x_j\right)$, where x_i belongs to the majority class and x_j belongs to the minority class, forms a Tomek link if:

$$\forall k, d(x_i, x_k) > d\left(x_i, x_j\right) \tag{5}$$

where $d\left(x_i, x_j\right)$ is the distance between x_i and x_j. The removal of these links ensures that noisy or overlapping instances between the two classes are eliminated, leading to clearer decision boundaries for the classifiers.

By applying SMOTETomek to both the training and testing sets, the model's capacity to learn from the minority class is enhanced, improving its generalizability and robustness in detecting churn behavior.

3.2.4. Correlation Analysis

To gain more insights about the relation between the variables used in our study, the Spearman correlation coefficient is used to build the correlation matrix, which measures the strength and direction of a monotonic relationship between two variables. The Spearman correlation is particularly suitable for this analysis because it does not assume linearity between the variables, but rather focuses on their rank-order relationships. The Spearman correlation coefficient is given by:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{6}$$

where d_i is the difference between the ranks of corresponding values in the two variables and n is the number of observations (instances). This coefficient ranges from -1 (perfect negative monotonic relationship) to +1 (perfect positive monotonic relationship), with values close to 0 indicating weak or no monotonic relationship.

3.2.5. Data Scaling

The final step in the preprocessing pipeline is the application of sin–max scaling to transform numerical variables into a uniform range, typically [0, 1]. Scaling ensures that variables with different units or magnitudes do not disproportionately influence the model's predictions. The min–max scaler transforms each feature according to the following formula:

$$fx' = \frac{fx - min(fx)}{max(fx) - min(fx)} \tag{7}$$

where fx is the original feature value, $min(fx)$ is the minimum value of the feature in the dataset, $max(fx)$ is the maximum value of the feature and fx' is the scaled value, rescaled to the range [0, 1].

3.3. Model Training and Evaluation

In this section, we describe the methodology used for training and evaluating the ML models. This process includes splitting the dataset into training and test sets, formalizing the mathematical basis for the RF and LGBM models, implementing probability calibration using isotonic regression, and concluding with post-model explainability using SHAP.

3.3.1. Train–Test Split Methodology

After the preprocessing phase, the dataset is split into training and test sets. For this research, we employed a standard 80–20% split, with 80% of the data used for model training and 20% reserved for testing. Given that the dataset becomes balanced after the application of SMOTETomek, the decision threshold for classification is set to the default value of 50%. Specifically, instances with predicted probabilities greater than or equal to 50% are classified as belonging to class 1 (churn), while those below this threshold are classified as belonging to class 0 (retention). To evaluate the models' performance, we compute a range of metrics, including accuracy, precision, sensitivity (recall), F1 score, and Brier score. In addition, we apply the calibration method through isotonic regression and compare the results before and after calibration. The receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) score are also analyzed to assess the discriminative power of the models.

3.3.2. Random Forest Classifier

RF is a widely used ensemble learning method that combines multiple decision trees to improve classification accuracy. The model constructs multiple decision trees during training, and each tree is built using a random subset of the features and a bootstrap sample of the training data.

Mathematically, the prediction of a RF for a given input X is the average of the predictions from each decision tree in the ensemble. Formally, the RF prediction is given by:

$$\hat{y} = \frac{1}{nT}\sum_{l=1}^{nT} T_l(X) \tag{8}$$

where nT is the number of decision trees in the forest, $T_l(X)$ is the prediction from the l-th decision tree for the input X and $\hat{y}$ is the final prediction, which is typically the class label with the highest average probability across the trees.

Each decision tree in the RF is trained independently on a bootstrap sample of the data, and at each split in the tree, a random subset of features is selected to reduce correlation between trees. This randomness ensures that the trees are decorrelated, thus improving the ensemble's robustness and reducing overfitting.

3.3.3. Light Gradient-Boosting Machine Classifier

LGBM is a highly efficient implementation of gradient boosting, which builds models sequentially by optimizing a differentiable loss function. Unlike RF, which averages the predictions of independent trees, LGBM builds each tree to correct the errors of the previous trees. The objective of LGBM is to minimize the following loss function:

$$L(y, \hat{y}) = \sum_{i=1}^{n} l(y_i, F(x_i)) \tag{9}$$

where $l(y_i, F(x_i))$ is the loss function, in our case the log-loss function, y_i is the true label of instance i and $F(x_i)$ is the predicted value for instance i from the current model.

In each iteration, LGBM fits a new decision tree to the negative gradient of the loss function, updating the model as follows:

$$F_m(X) = F_{m-1}(X) + \eta \cdot h_m(X) \tag{10}$$

where $F_{m-1}(X)$ is the model from the previous iteration, η is the learning rate, and $h_m(X)$ is the new decision tree fitted to the residuals (errors) from the previous model.

The LGBM model is highly optimized for efficiency, using techniques such as histogram-based algorithms for faster computation, leaf-wise growth strategy, and exclusive feature bundling to reduce memory usage and increase speed.

3.3.4. Probability Calibration Using Isotonic Regression

Many ML models, including RF and LGBM, tend to produce uncalibrated probabilities, which can affect decision-making when probabilities are needed for further business actions. To address this, we apply isotonic regression as a post-processing step to calibrate the predicted probabilities.

Isotonic regression is a non-parametric method that fits a monotonic function to the predicted probabilities, ensuring that the probabilities increase as the predicted class becomes more certain. The objective of isotonic regression is to minimize the following loss function:

$$\underset{f}{min} \sum_{i=1}^{n} (f(\hat{p}_i) - y_i)^2 \tag{11}$$

where $f(\hat{p}_i)$ is the isotonic function applied to the predicted probability $\hat{p}_i$, y_i is the true label for instance i.

The isotonic function used for probability calibration is a piecewise constant, non-decreasing function. While there is no specific closed-form equation for the isotonic function (since it depends on the data points it is fitted to), its form is derived from solving a constrained optimization problem where the function is required to be monotonic (non-decreasing).

The calibration process adjusts the predicted probabilities to better align with the true probabilities, improving the interpretability and reliability of the model's predictions.

3.3.5. Explainable AI with SHAP

To enhance the transparency and interpretability of our ML models, we implement the Shapley Additive Explanations (SHAP) method. SHAP values are based on cooperative game theory and provide a consistent and fair measure of the contribution of each feature to the model's prediction. The SHAP value for a feature *fx* is calculated as the weighted average of the marginal contributions of that feature across all possible subsets of features. The mathematical formulation of the SHAP value ϕ_{fx} for feature *fx* is:

$$\phi_{fx}(f,x) = \sum_{S \subseteq Nx \backslash \{fx\}} \frac{|S|!(|Nx| - |S| - 1)!}{|Nx|!}(f(S \cup \{fx\}) - f(S)) \quad (12)$$

where Nx is the set of all features, S is the subset of features excluding *fx*, $f(S)$ is the model's prediction using only the features in subset S, $f(S \cup \{fx\})$ is the prediction when feature *fx* is added to the subset S, and $\phi_{fx}(f,x)$ is the SHAP value, which represents the contribution of feature *fx* to the prediction for instance x_i.

SHAP values provide a global and local interpretability framework, helping to explain how each feature contributes to individual predictions and how important each feature is in the overall model. This method is particularly useful in explaining complex ML models, such as RF and LGBM, which are often treated as "black boxes".

3.3.6. Classification Metrics

In the case of classification problems, several statistical indicators are used to describe the generalization capabilities of a model. A very useful tool in describing the classification power of models is the confusion matrix. In the confusion matrix, we can identify true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). Based on these results identified from the confusion matrix, the main classification metrics are defined.

Accuracy is a general measure of a model's performance and represents the proportion of correct predictions (both positive and negative) out of the total cases. It is calculated using the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (13)$$

Precision measures the accuracy of positive predictions. It reflects the proportion of cases labeled as positive by the model that are actually positive. The calculation formula is:

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

Recall measures the model's ability to correctly identify all relevant (positive) cases. This is important in scenarios where missing positive cases are critical. It is calculated as:

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

F1 score represents the harmonic mean between precision and recall. It provides a balance between precision and recall and is useful when there is an uneven class distribution. The formula for the *F1 score* is:

$$F1Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (16)$$

The Brier score is a measure of the performance of a prediction model, especially in contexts where predictions are probabilities. The Brier score is defined as the mean of the squared differences between the predicted probabilities and the observed outcomes, where outcomes are coded as 0 or 1. This score is useful for evaluating the accuracy of predicted probabilities in binary classification cases and is often used as a calibration metric.

It is helpful for comparing the performances of different prediction models as it provides a single number summarizing the model's performance. Models with a lower Brier score are considered superior. In the context of probability calibration, the Brier score is efficient as it measures how close the predicted probabilities are to reality. If a model's probabilities are perfectly calibrated, then the Brier score reflects only the inherent errors in the data and the modeled event. Mathematically, the Brier score is defined as:

$$Brier\ score = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2 \tag{17}$$

where p_i is the predicted probability of the positive class for instance i, y_i is the actual binary outcome for instance i (1 if the event occurs, 0 otherwise).

The receiver operating characteristic (ROC) curve is a graphical representation used to evaluate the performance of binary classifiers by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold levels. The ROC curve provides insights into the trade-off between sensitivity (recall) and specificity for different classification thresholds. To summarize the performance of the classifier using the ROC curve, the area under the curve (AUC) is computed. The AUC is a scalar value that represents the entire two-dimensional area beneath the ROC curve, providing an aggregate measure of classifier performance across all classification thresholds.

The steps of the methodology are combined into a single algorithm (Algorithm 1) expressed as pseudocode:

Algorithm 1. Algorithm—Customer churn prediction.

```
# Step 1: Feature Engineering
FOR each customer IN dataset X:
    BalanceToSalaryRatio ← Balance/Salary                    // Equation (1)
    BalanceToProductRatio ← Balance/NumProducts              // Equation (2)
    TenureToAge ← Tenure/Age                                 // Equation (3)
ADD [BalanceToSalaryRatio, BalanceToProductRatio, TenureToAge] to X
# Step 2: Data Encoding
# Convert categorical variables into numerical format.
If Gender == 'Male':
        Gender ← 1
Else:
        Gender ← 0
Geography ← LabelEncoding(Geography)
FOR binary feature IN [HasCrCard, IsActiveMember]:
    REPLACE 0 with −1
# Step 3: Oversampling with SMOTETomek
# Address class imbalance by generating synthetic samples and removing Tomek links.
FOR each $x_{minority}$ IN minority_class:
        $x_{neighbor}$ ← k-nearest neighbors ($x_{minority}$)
        $x_{new} \leftarrow x_{minority} + \lambda \cdot (x_{neighbor} - x_{minority})$    // Equation (4)
        ADD $x_{new}$ TO minority_class
FOR each majority-minority pair ($x_i$, $x_j$):
        IF distance ($x_i$, $x_j$) IS minimal for both i and j:
        REMOVE pair ($x_i$, $x_j$) FROM X                    // Equation (5)
# Step 4: Correlation analysis
INITIALIZE correlation_matrix
FOR each pair ($x_i$, $y_i$) of variables (X, Y) IN dataset:
     $d_i$ ← COMPUTE RANK ($x_i$ , $y_i$)
     $\rho \leftarrow 1 - \frac{6\sum d_i^2}{n(n^2-1)}$      // Equation (6)
ADD $\rho$ TO correlation_matrix
# Step 5: Data Scaling
FOR each feature $fx$ IN dataset X:
$fx' \leftarrow (fx - min(fx))/(max(fx) - min(fx))$         // Equation (7)
     REPLACE $fx$ WITH $fx'$ in the dataset X
# Step 6: Model Training and Evaluation
X_train, Y_train, X_test, Y_test ← SPLIT (X, Y)
# Train Random Forest (RF) and Light Gradient Boosting Machine (LGBM) models.
// Random Forest Classifier
INITIALIZE aggregate_prediction
FOR each tree IN number_of_trees:
    bootstrap_sample ← DRAW bootstrap FROM X_train
    tree ← TRAIN DecisionTree(bootstrap_sample, random_feature_subset)
```

```
        predict_tree ← tree (X_train)
        aggregate_prediction ← aggregate_prediction + predict_tree
prediction_RF ← aggregate_trees/ number_of_trees                // Equation (8)
// LightGBM Classifier
INITIALIZE model_with_mean_prediction
FOR each iteration IN num_iterations:
        gradient ← COMPUTE gradient(errors) BASED ON loss_function        // Equation (9)
        tree ← FIT model (X_train, gradient)
        UPDATE LGBM_model WITH learning_rate × tree                        // Equation (10)
prediction_LGBM ← LGBM_model (X_train)
# Step 7: Probability Calibration with Isotonic Regression
# Calibrate model predictions to improve interpretability.            // Equation (11)
calibrated_probabilities ← FIT IsotonicRegression TO model_probabilities
# Explainability with SHAP
FOR each feature fx IN X:                                  // Equation (12)
        SHAP_value ← COMPUTE weighted marginal contributions ($fx$)
# Step 8: Evaluation Metrics
# Compute metrics to assess model performance.
COMPUTE confusion_matrix(predictions, actual_labels)
Accuracy ← (TP + TN)/(TP + TN + FP + FN)                                  // Equation (13)
Precision ← TP/(TP + FP                                                   // Equation (14)
Recall ← TP/(TP + FN)                                                     // Equation (15)
F1_score ← 2 × (Precision × Recall)/(Precision + Recall)              // Equation (16)
Brier_score ← $MEAN(predicted\_probability - actual\_label)^2$              // Equation (17)
PLOT ROC_curve AND COMPUTE AUC
END Algorithm
```

If using a real dataset from a specific bank, adjustments to the current methodology would improve model relevance and accuracy. Domain-specific feature engineering may help tailor engineered features to reflect unique bank characteristics, such as loan types, specific customer segments or engagement patterns. Additionally, incorporating bank-specific transaction behaviors (e.g., ATM withdrawals, loan repayment history) will provide further insights into the customer's banking profile. An expanded EDA would also be beneficial. This step would analyze additional variables unique to the bank, such as regional banking habits, preferred banking channels (online vs. branch visits) or specialized product usage. Recent churn trends and any identifiable reasons for customer churn, such as dissatisfaction or competitive offerings, should also be reviewed.

Therefore, churn drivers identified from customer feedback or survey data would provide valuable qualitative insights into churn motivation. These insights could be incorporated into feature engineering or used in post-model interpretability stages. In a real-world environment, customer behavior and market conditions change over time, affecting model accuracy. Implementing regular model drift checks and recalibrating probabilities using new customer data helps ensure the model remains accurate. Real-world churn behavior is often influenced by broader economic conditions, so it may be beneficial to integrate macroeconomic indicators, such as interest rates inflation rates, or unemployment statistics, to capture external factors that could affect customer retention.

4. Results

4.1. Exploratory Data Analysis

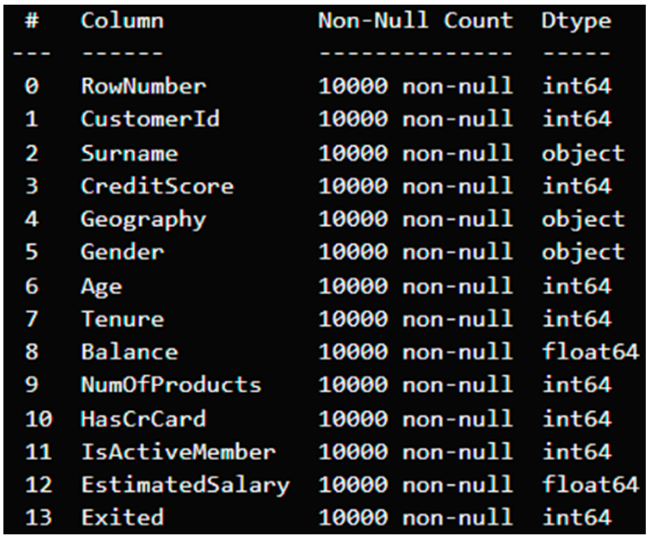
The explored dataset initially consisted of 14 variables, of which 11 are relevant to our problem. The dataset employed in this study has been widely utilized in previous research on customer churn prediction, appearing in multiple studies [20,33–42]. The input variables are presented in Table 2.

The first three variables are not predictive, so we will exclude them from our analysis. The next step involves checking for the presence of null values in the dataset.

From Figure 2, we observe that all variables contain non-null values, so it is not necessary to address specific methodologies for handling them.

Table 2. The variables and their descriptions.

Content	Details
RowNumber	Row index—insignificant
CustomerId	Unique customer identifier
Surname	Surname
CreditScore	Customer's credit score, calculated based on their credit history
Geography	Customer's country or region of residence, indicating the bank's geographic market
Gender	Customer's gender, which can be male or female
Age	Customer's age
Tenure	Number of years or months since the customer became a client of the bank
Balance	Amount of money in the customer's accounts at the bank at a given time
NumOfProducts	Number of products used by the customer
HasCrCard	Whether the customer holds a credit card (1) or not (0)
IsActiveMember	Whether the customer is active (1) or not (0)
EstimatedSalary	Estimated annual salary of the customer
Exited	Target variable, whether the customer has left the bank (1) or not (0)



```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
```

Figure 2. Variables in the dataset, their non-null values and each one's type.

The distributions of the variables CreditScore, Age and Balance are highlighted to illustrate some of the most important characteristics for an instance in the dataset, in Figure 3. The distribution of credit score has a shape that suggests an approximation of a normal distribution, with a central peak and symmetrical tails. However, some positive skewness is observed, where a small number of customers have very high credit scores compared to the central area. This distribution indicates that the majority of bank customers have moderate to high credit scores.

The age distribution plot indicates a bell-shaped distribution, with a pronounced peak around the median value. We observe the presence of positive skewness, reflecting a higher proportion of younger customers compared to older ones. The data points, in the right tail of the distribution, suggest that a smaller segment of the population has ages above the average threshold. Regarding the balance distribution, the plot shows a peak at zero, followed by a bimodal distribution. The presence of a significant number of customers with a zero balance may indicate customers who are not actively using the bank account for deposits or who have closed their accounts. The two modes observed in the distribution of positive balances may reflect different saving or investment behaviors of customers.

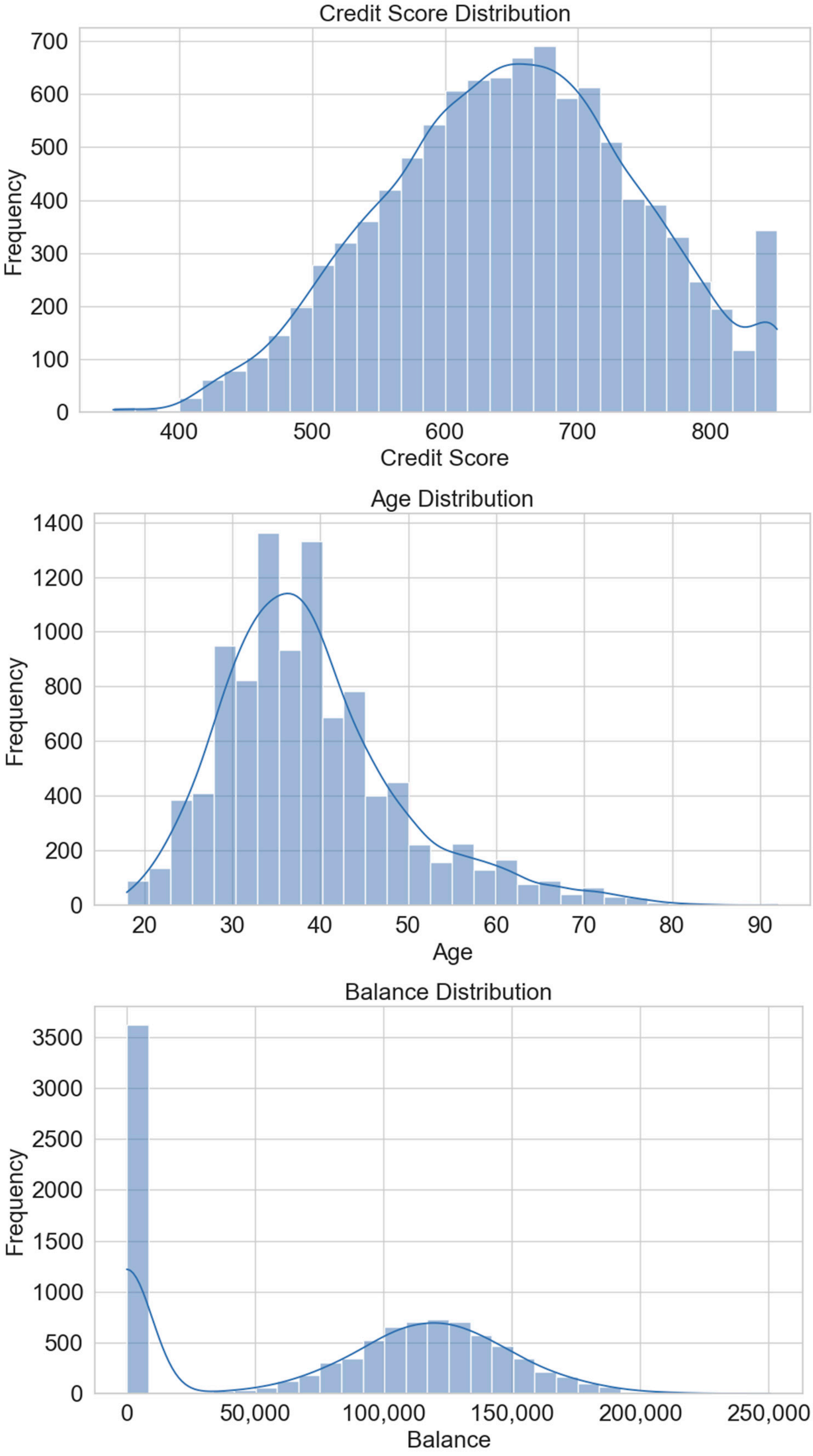


Figure 3. Distributions of CreditScore, Age, and Balance variables.

The distributions for the number of banking products used (NumOfProducts) and the churn rate (Exited) are particularly relevant, given their direct impact on the financial stability and growth of the bank. In Figure 4, the distribution of the number of banking products illustrates a categorical distribution where most customers possess one or two banking products. A significantly smaller proportion of customers use three or four products, which may indicate a tendency for customers not to overly diversify the banking services used. The distribution of the churn variable (Exited) demonstrates a clear imbalance between the number of customers who have churned and those who have remained. Most customers, according to this variable, have not churned from the bank, suggesting a positive retention rate, but the proportion of those who have churned is not negligible, requiring further analysis to identify the factors associated with this decision.

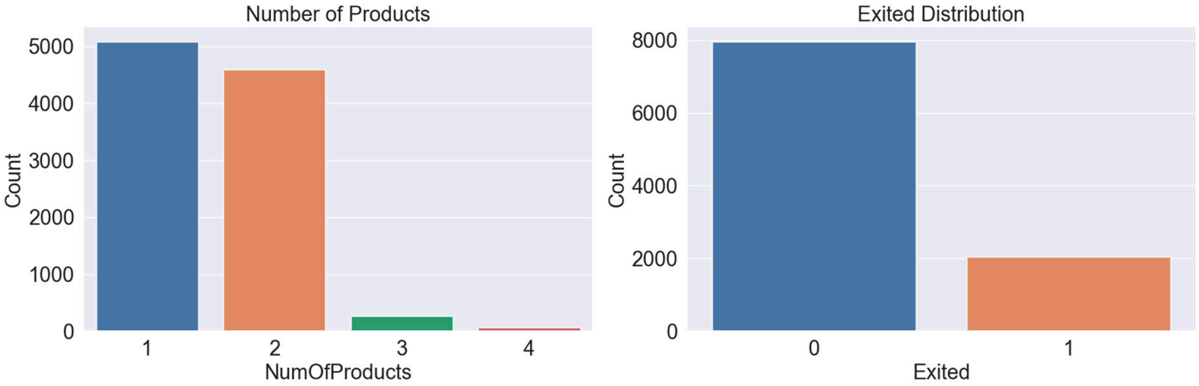


Figure 4. Distributions of the NumOfProducts and Exited variables.

We further analyze the distribution of the number of clients who retained the use of the banking services, marked with 0, compared to those who churned from the bank, highlighted with 1, based on geographical location expressed by the variable Geography, for the countries France, Spain, and Germany. It is observed that France presents the highest number of loyal customers, while the number of those who churned from the bank is highest in Germany. This trend indicates that regional or specific business policy factors may influence customers' decisions to continue or terminate their relationship with the bank.

Regarding the distribution of service churn based on gender (Gender variable), the two bars representing the number of female and male clients are distinguished (Figure 5). The graph highlights a gender difference in churn rate, with a higher number of women churning from the bank compared to men. This underscores the need for deeper analysis to understand the underlying reasons for these differences and to identify possible disparities in customer experience or satisfaction based on gender.

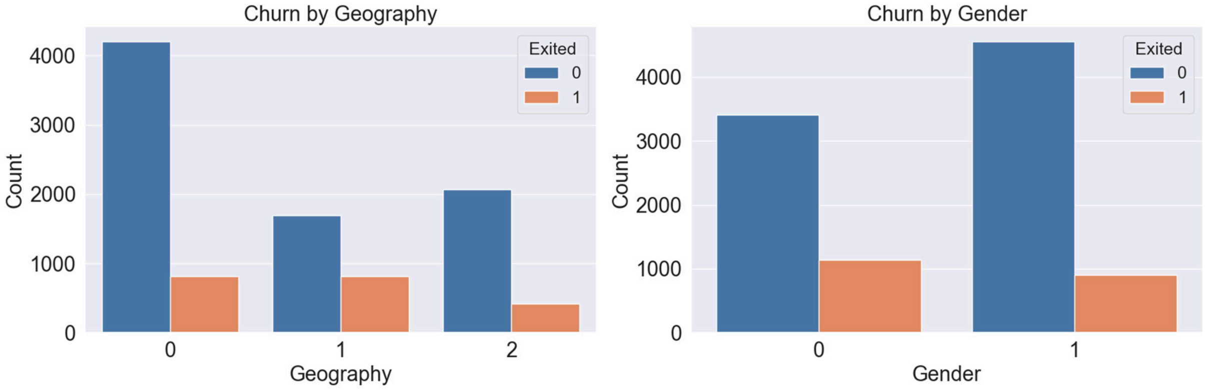
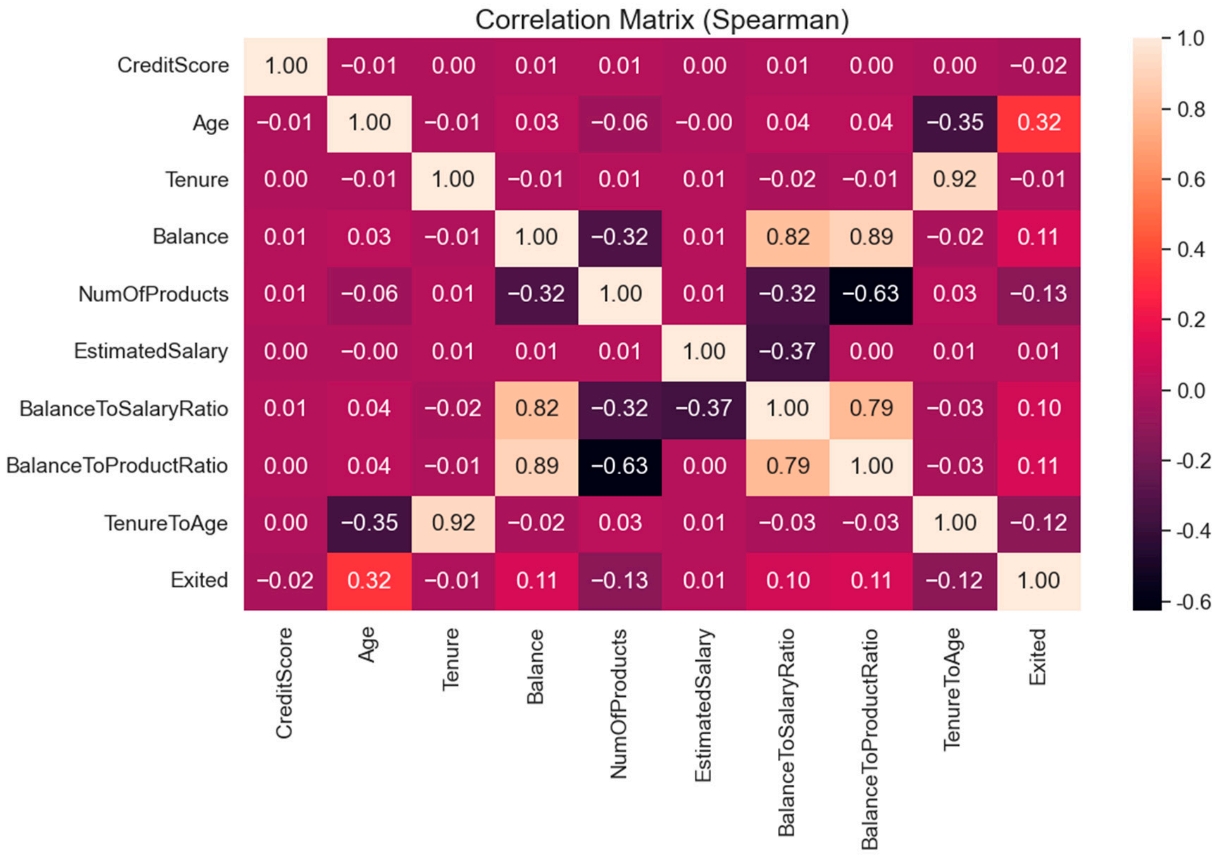


Figure 5. Distribution of the number of customers by categories based on country and gender.

For calculating the correlation matrix in Figure 6, we choose the Spearman correlation coefficient to measure monotonic associations between variables. This coefficient does not assume linearity between variables but rather the ability of one variable to predict the magnitude order of the other.



Correlation Matrix (Spearman)

	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	BalanceToSalaryRatio	BalanceToProductRatio	TenureToAge	Exited
CreditScore	1.00	−0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.00	−0.02
Age	−0.01	1.00	−0.01	0.03	−0.06	−0.00	0.04	0.04	−0.35	0.32
Tenure	0.00	−0.01	1.00	−0.01	0.01	0.01	−0.02	−0.01	0.92	−0.01
Balance	0.01	0.03	−0.01	1.00	−0.32	0.01	0.82	0.89	−0.02	0.11
NumOfProducts	0.01	−0.06	0.01	−0.32	1.00	0.01	−0.32	−0.63	0.03	−0.13
EstimatedSalary	0.00	−0.00	0.01	0.01	0.01	1.00	−0.37	0.00	0.01	0.01
BalanceToSalaryRatio	0.01	0.04	−0.02	0.82	−0.32	−0.37	1.00	0.79	−0.03	0.10
BalanceToProductRatio	0.00	0.04	−0.01	0.89	−0.63	0.00	0.79	1.00	−0.03	0.11
TenureToAge	0.00	−0.35	0.92	−0.02	0.03	0.01	−0.03	−0.03	1.00	−0.12
Exited	−0.02	0.32	−0.01	0.11	−0.13	0.01	0.10	0.11	−0.12	1.00

Figure 6. Correlation matrix between the numerical variables and the target Exited variable.

The moderate positive correlation between Age and Exited (0.32) suggests that older customers are more prone to churn, indicating age as a significant predictor for customer attrition. Conversely, the correlation between Age and Tenure is −0.35, revealing that younger customers often have longer tenure compared to older customers. This may reflect that newer, older customers are at higher risk of churn, potentially due to a lack of strong engagement or loyalty to the bank. The correlation matrix also highlights that Balance exhibits moderate correlations with other financial metrics, such as BalanceToSalaryRatio (0.82) and BalanceToProductRatio (0.89). These strong relationships indicate that customers' financial behavior, as reflected in their account balances and ratios, is interconnected. The variable NumOfProducts shows a slight negative correlation with Exited (−0.13), indicating that customers with more products are slightly less likely to churn. This could imply that deeper product engagement might act as a retention factor. There is also a notable negative correlation between NumOfProducts and BalanceToProductRatio (−0.63), indicating that as customers use more products, the ratio of their balance to the number of products decreases, possibly due to the distribution of balances across multiple products. The correlation matrix shows a negative correlation of −0.35 between Age and Tenure, indicating that younger customers tend to have longer tenure periods relative to their age, or conversely, older customers may have shorter tenure. This could suggest that older customers who have joined the bank more recently might be at higher risk of churn, as their shorter tenure may reflect less loyalty or engagement with the bank.

4.2. RF Architecture

To configure the RF model, we use the RandomForestClassifier method from the scikit-learn package. We use 100 decision tree estimators, the split criterion is gini, and a minimum of 2 instances are required in a node to split an internal decision tree.

The confusion matrix generated for the uncalibrated RF model (Figure 7—left side) highlights the distribution of predictions in relation to the actual values. The model correctly classified 1427 cases as belonging to the negative class (actual label 0), indicating a good ability of the model to identify customers who did not churn. In contrast, the model identified 1233 cases correctly as belonging to the positive class (actual label 1), representing customers who churned. However, the model exhibited a number of 341 false negative instances, cases where customers who actually churned were predicted to be retained (predicted label 0). This aspect underlines a limitation of the model in capturing all the dynamics associated with the churn decision. Additionally, there were 147 false positive cases, where customers who stayed with the bank were erroneously predicted to churn (predicted label 1).

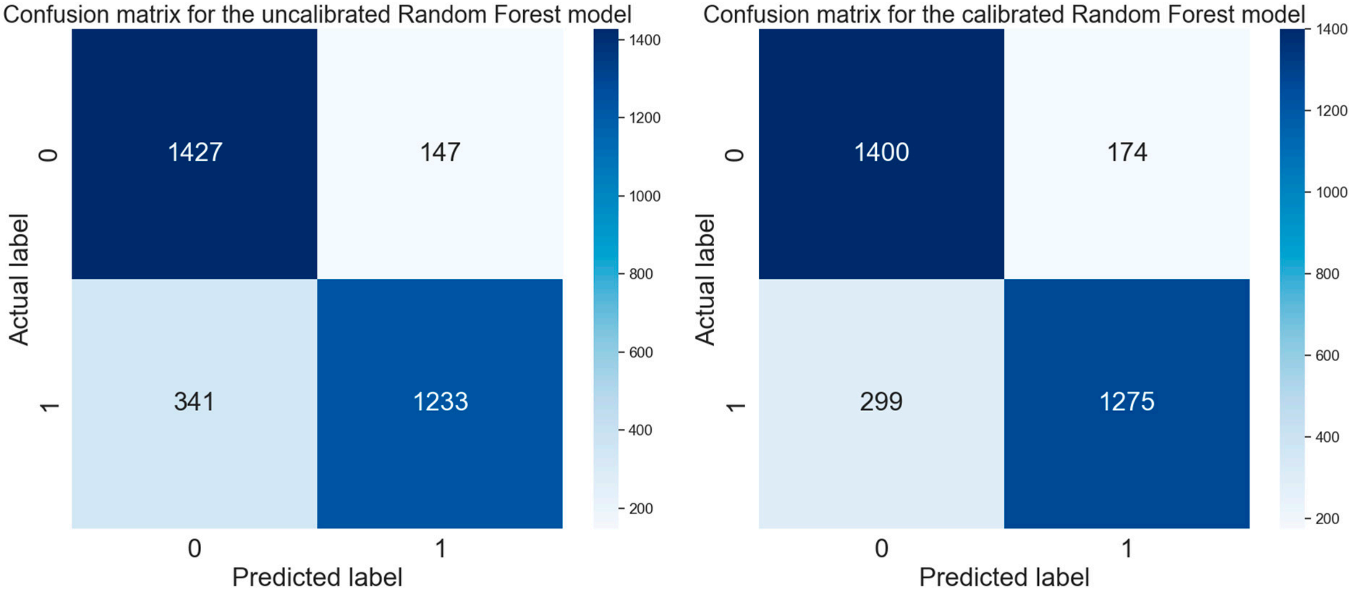


Figure 7. Confusion matrices for the uncalibrated and calibrated RF model.

The confusion matrix of the calibrated RF model (Figure 7—right side) indicates an improvement in identifying customers who churn, with an increase in true positive cases to 1275 and a decrease in false negative cases to 299. However, there is a slight increase in false positive classifications. These results suggest that the model calibration process has adjusted the classification threshold, leading to more precise detection of customers likely to churn, with an acceptable compromise regarding the increase in false positive errors.

In Figure 8, we illustrated the ranking of the most important variables that contributed to the RF model predictions. We observe that Age is the most influential feature, indicating that age differences among customers significantly contribute to the predictability model of bank churn behavior. Other significant features include the number of bank products used (NumOfProducts) and the tenure-to-age ratio (TenureToAge), which may reflect customer loyalty and engagement with the bank. Estimated salary (EstimatedSalary) and credit score (CreditScore) emphasize the importance of economic factors in the decision to churn from the bank.

The balance-to-product ratio (BalanceToProductRatio), balance (Balance) and tenure (Tenure) have moderate importance, suggesting that customer financial management and the duration of the relationship with the bank are also relevant, but to a lesser extent.

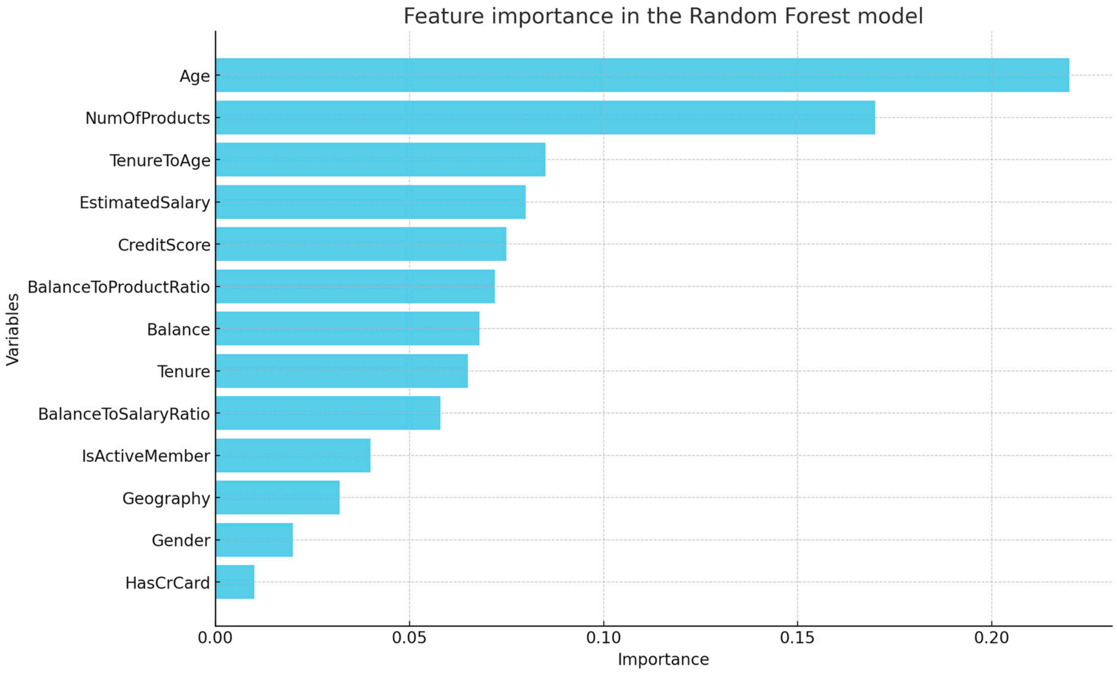


Figure 8. Ranking of variable importance in the RF model.

Features with the lowest importance in the model are BalanceToSalaryRatio, which compares the customer's financial resources to their income, whether they are an active member (IsActiveMember), demographic factors such as geography (Geography), gender (Gender), and possession of a credit card (HasCrCard). Their low importance suggests that although they play a role in the model, they are not the main predictors of churn behavior within the analyzed dataset.

The Brier score associated with the RF model is approximately equal to 0.115, indicating that the model performs well for the probability threshold for predictions equal to 0.5. For the calibrated model, the Brier score is equal to 0.1503. A higher Brier score for the calibrated model could reflect a closer adherence of predicted probabilities to the complex realities of the data, even if these probabilities are less confident and thus lead to a slightly weaker Brier score.

In assessing the performance of the RF model, the classification report shows changes between the uncalibrated and calibrated versions of the model. For the uncalibrated model, the precision for the positive class (class 1, indicating customers who churn) was 0.89, while the recall was 0.78, resulting in an F1 score of 0.83. This suggests that although the model was quite precise in its predictions, it missed a considerable proportion of actual churned customers.

After calibration, an improvement in recall for the positive class to 0.81 was observed, with a slight decrease in precision to 0.88, suggesting that the calibrated model was more efficient in identifying customers likely to churn. The improvement in the F1 score to 0.84 for the positive class indicates better harmony between precision and recall after calibration. For the negative class (class 0, indicating customers who remain with the bank), there was a slight improvement in precision from 0.81 to 0.82 and a decrease in recall from 0.91 to 0.89 following calibration. These changes suggest that the calibrated model became slightly more selective, sacrificing some of its ability to recognize all loyal customers in favor of identifying those who churn. Overall, the calibration process balanced the performance between the positive and negative classes, leading to a general increase in model accuracy from 0.84 to 0.85. This overall improvement in precision and recall across the entire dataset is reflected in the F1 score. Thus, calibration brought an advantage to the model in the context of balancing the costs associated with different types of classification errors.

4.3. LBGM Model

For the LGBM model, we select a hyperparameter configuration with a maximum depth of the internal classification estimators set to 6, a maximum number of leaves for estimators set to 31, a learning rate around 0.1, and, similar to the RF model, a total number of internal decision trees set to 100. The boosting method used is the default gradient-boosting decision tree (GBDT).

In Figure 9, we highlight the confusion matrices for both the uncalibrated and calibrated approaches. Comparing the confusion matrices for the uncalibrated and calibrated LGBM models, we observe significant differences in classification performance compared to the RF model. In the uncalibrated model, the number of true negative cases is 1464, and the number of true positive classifications is 1386, indicating a robust ability of the model to correctly classify both the customers who stay and those who churn. However, there are 188 false negative instances and 110 false positive instances, suggesting opportunities for improvement in accurately identifying all churn cases. After calibration, the confusion matrix shows a slight increase in the number of true negative instances to 1489 and a slight decrease in true positive instances to 1366. False negative cases increase to 208, while false positives decrease to 85. This indicates that while the calibrated model is better at avoiding misclassifying loyal customers as at risk of churn (fewer false positives), it is slightly less efficient in identifying actual churners (increased false negatives).

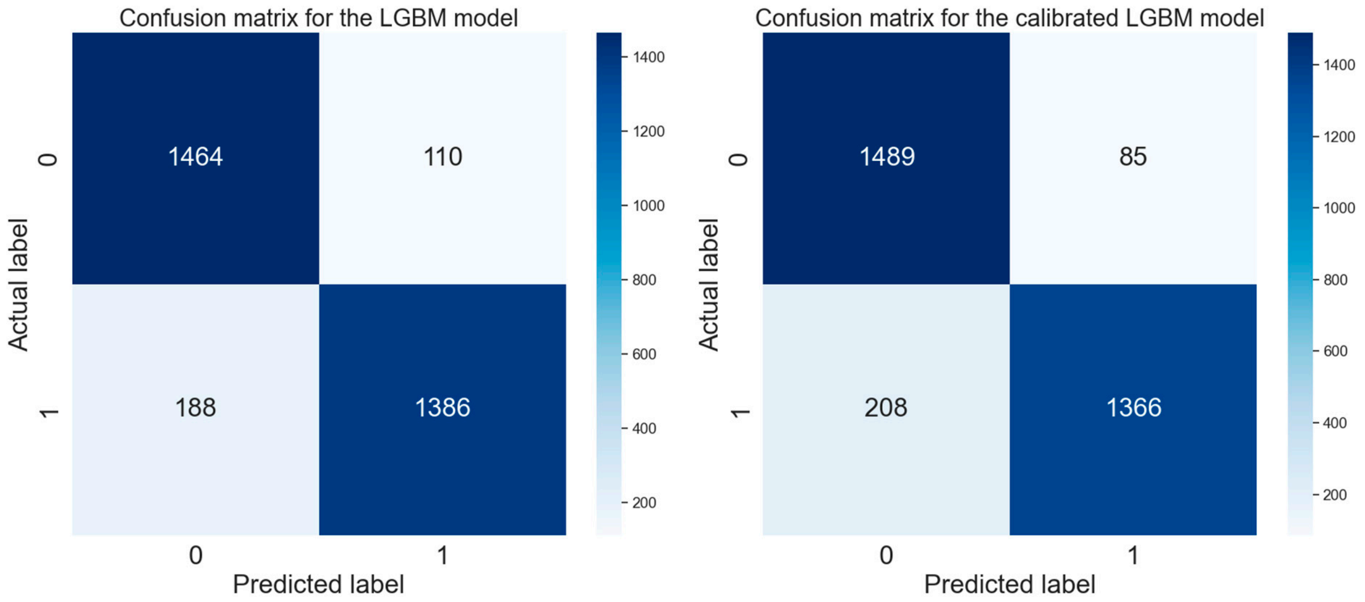


Figure 9. Confusion matrices for the uncalibrated and calibrated LGBM.

From the variable importance plot (Figure 10) for the LGBM model, we can infer that Age is also the most significant feature for the model's prediction, indicating a major influence of customers' age on their decision to churn or stay with the bank. Tenure (the duration of the relationship with the bank) and EstimatedSalary follow, suggesting that the time spent as a bank customer and the estimated income level are also important factors.

The variables Balance (account balance) and CreditScore present moderate importance, which may indicate their role in modeling customer satisfaction and financial behavior. NumOfProducts (number of bank products used) and derived ratios like TenureToAge (the ratio of tenure to age) and BalanceToSalaryRatio (the ratio of account balance to estimated salary), although less significant than the top variables, reflect other aspects of customer commitment and financial stability. Geography, IsActiveMember (whether the customer is an active member), Gender, and HasCrCard (whether the customer holds a bank-issued credit card) are at the bottom of the ranking, indicating relatively low influence in the current model.

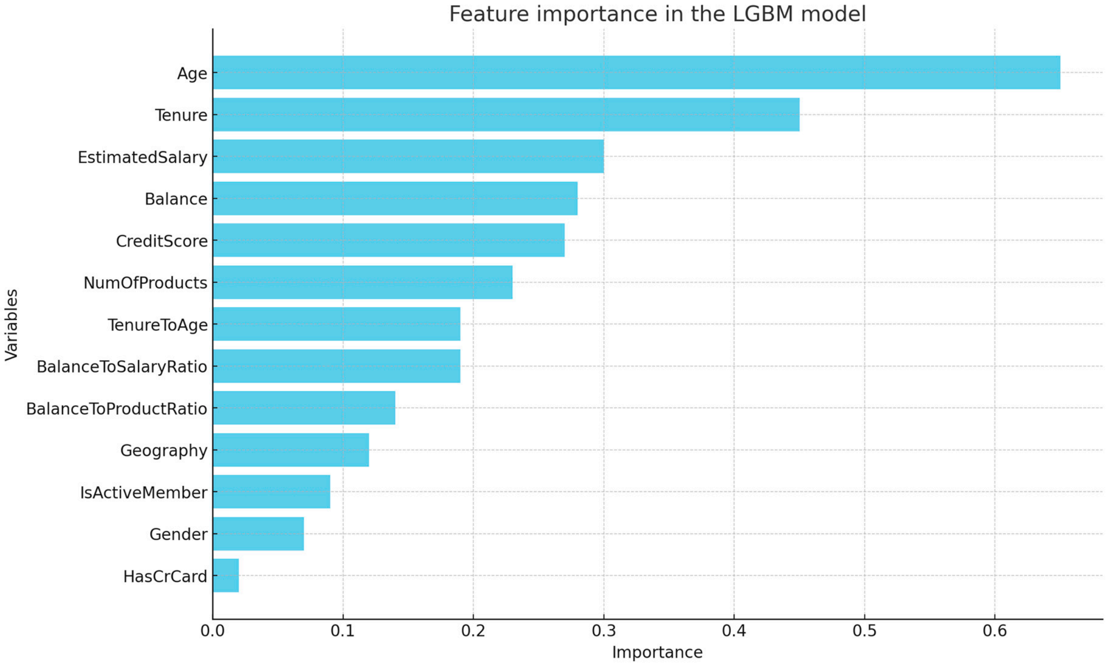


Figure 10. Feature importance in the LGBM model.

The comparative classification report for the LGBM model shows similar prediction performance between the uncalibrated and calibrated versions, with minor variations between metrics. For the uncalibrated model, we have a precision of 0.89 for class 0 and of 0.93 for class 1, suggesting an excellent ability to correctly identify both customers who do not churn and those who do churn. The recall for class 0 is 0.93, indicating that most loyal customers are correctly recognized, and that for class 1 is 0.88, reflecting a good ability to identify customers who churn. The F1 score, which balances precision and recall, is 0.91 for class 0 and 0.90 for class 1, demonstrating a solid overall performance of the model.

After calibrating the model, precision decreases slightly for class 0 to 0.88 while increasing to 0.94 for class 1. Recall increases for class 0 to 0.95, reflecting an improvement in recognizing loyal customers, and decreases for class 1 to 0.87. The F1 score remains constant at 0.91 for class 0 and decreases marginally to 0.90 for class 1. The overall accuracy remains at 0.91, indicating that overall, the calibrated model does not provide a general improvement over the uncalibrated model. Individual changes in metrics suggest that the calibrated model may be slightly more proficient in identifying loyal customers but with minimal cost in its ability to detect customers who will churn. These differences may be relevant depending on the specific context and business objectives, especially if prioritizing the reduction of false negatives or false positives.

The Brier score for the uncalibrated model is 0.0732, and for the calibrated one, the score is around 0.0697. A lower Brier score for the calibrated model indicates a marginal improvement in the accuracy of predicted probabilities after calibration. The reduction from 0.0732 to 0.0697 represents a positive, albeit modest, adjustment of the LGBM model regarding the correspondence between predicted probabilities and observed realities in the test data. This suggests that the calibration process helps to align the predicted probabilities closer to the actual label distribution in the dataset, possibly by correcting biases or adjusting probability estimates that were initially either too pessimistic or too optimistic. In practice, a model with a lower Brier score is preferred as it provides a more solid basis for risk-based decision-making.

As a final comparison criterion between the two models, we present the ROC curves and the associated AUC scores for these curves to highlight each model's performance at different probability thresholds. In Figure 11, the LGBM model, represented by the orange curve, shows slightly superior performance with an AUC of 0.96 compared to the RF model, represented by the blue curve, which has an AUC of 0.93. AUC values close to 1 indicate excellent model capacity to discriminate between positive and negative classes. The higher AUC value of the LGBM model suggests that it has an overall better performance in correctly classifying positive and negative cases. The dashed line represents random classification, with an AUC of 0.5, meaning that any model with an ROC curve above this line has better performance than a random classifier. Both models demonstrate better capability than a random model, but the LGBM curve is closer to the upper-left corner of the graph, indicating a higher rate of true positives and a lower rate of false positives for most decision thresholds used. The ROC curve is invariant to calibrated/uncalibrated model distinction as it is based on the ranking of predictions rather than the absolute probability values.

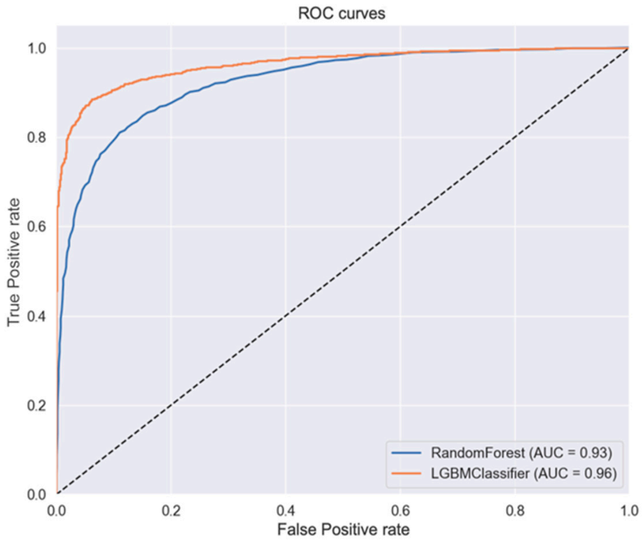


Figure 11. ROC curves associated with the two models.

Table 3 presents the hyperparameters used for configuring the RF and LGBM models, detailing the key settings considered when building the two methods.

In Table 4, we compare the performance metrics for class 1 across different models, including baseline and calibrated versions. The baseline models refer to those built on undersampled data without using SMOTETomek, and the predictions for these models are uncalibrated.

Table 3. Models' hyperparameters.

Model	Hyperparameter	Value
RF	Number of estimators	100
	Criterion	Gini
	Minimum samples required to split	2
	Maximum depth	None
	Bootstrap	True
	Random state	42
	Max features	'sqrt'
LGBM	Number of estimators	100
	Learning rate	0.1
	Maximum depth	6
	Maximum number of leaves	31
	Boosting method	Gradient Boosting Decision Tree (GBDT)
	Random state	42

Table 4. Metrics comparison of the RF and LGBM classifiers.

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1 Score (Class 1)	Brier Score
RF Baseline	0.87	0.78	0.45	0.57	0.1033
LGBM Baseline	0.86	0.74	0.47	0.58	0.1003
RF	0.84	0.89	0.78	0.83	0.115
RF Calibrated	0.85	0.88	0.81	0.84	0.1503
LGBM	0.91	0.93	0.88	0.90	0.0732
LGBM Calibrated	0.91	0.94	0.87	0.90	0.0697

4.4. SHAP Analysis

The last step in our analysis is the employment of the SHAP analysis to explain the internal decision-making process of both RF and LGBM. We applied SHAP scores on the uncalibrated versions of the two models in order to understand better the reasoning behind the generated predictions.

Summary plots describe the overall influence of the most important variables in a so-called black box model. In Figure 12, we observe a comprehensive visualization of feature contributions across all classifications generated by the LGBM model. The plot is arranged with features listed in order of importance from top to bottom, determined by the sum of the absolute SHAP values across all observations. Each dot on the plot represents an individual SHAP value, indicating the impact of a feature value on the model output relative to a baseline value, typically the mean of the feature across the dataset. The color of the dots varies from blue to pink, representing the low to high spectrum of feature values, respectively. Notably, the plot's horizontal dispersion of dots for each feature shows the variability of that feature's impact on the model's predictions. A wider dispersion suggests a more significant variation in how the feature affects the model output across different instances. This could imply interactions with other features or a non-linear relationship within the model. The variable Age stands out as the most influential feature, indicating a significant impact on churn prediction, where older ages tend to increase the likelihood of churn. Additionally, NumOfProducts also shows substantial influence, suggesting that the number of products a customer uses is imperious in predicting their retention or churn, with fewer products generally leading to higher churn rates.

Figure 13 displays the results for the SHAP analysis considering the RF classifier. The variables Age and NumOfProducts stand out as the most significant, with Age showing a wide dispersion of effects, indicating its complex relationship with churn likelihood. This variability suggests that older customers may be more prone to churn under certain conditions. Conversely, NumOfProducts negatively impacts churn, implying that customers with fewer products are more likely to disengage. Other notable features include Tenure, IsActiveMember and Balance, where being an active member notably decreases the likelihood of churn, highlighting the importance of customer engagement in retention strategies. Lesser impactful but still significant features like Geography and Gender also demonstrate moderate effects, suggesting demographic factors play a role in churn dynamics. This analysis illuminates the multifactorial nature of customer behavior and retention in the studied context.

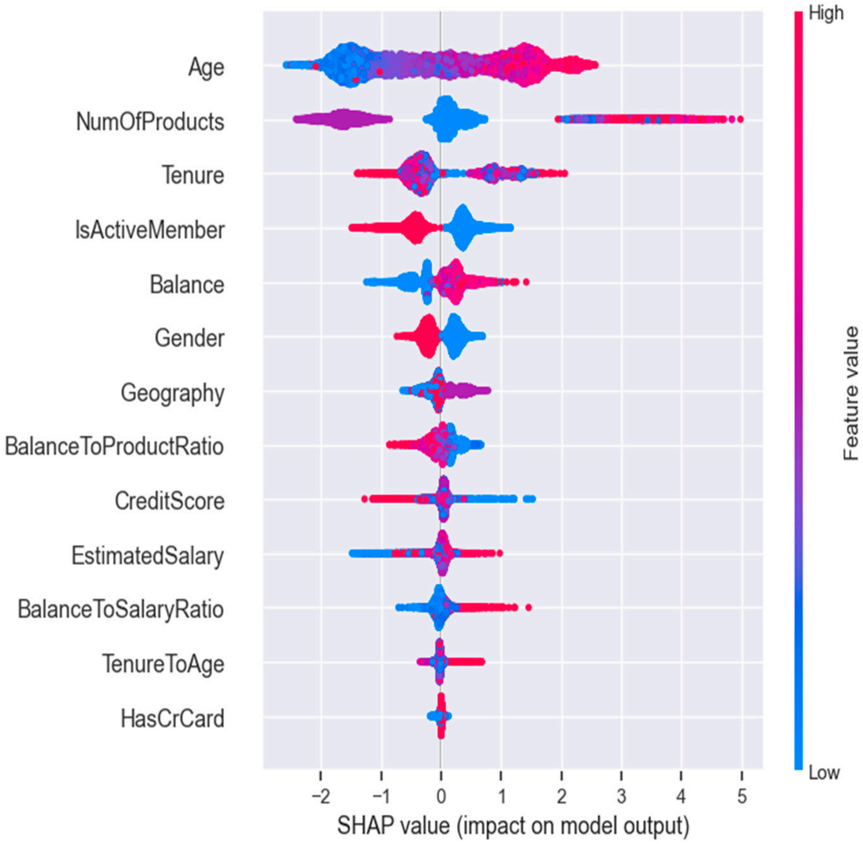


Figure 12. SHAP summary plot for the LGBM classifier.

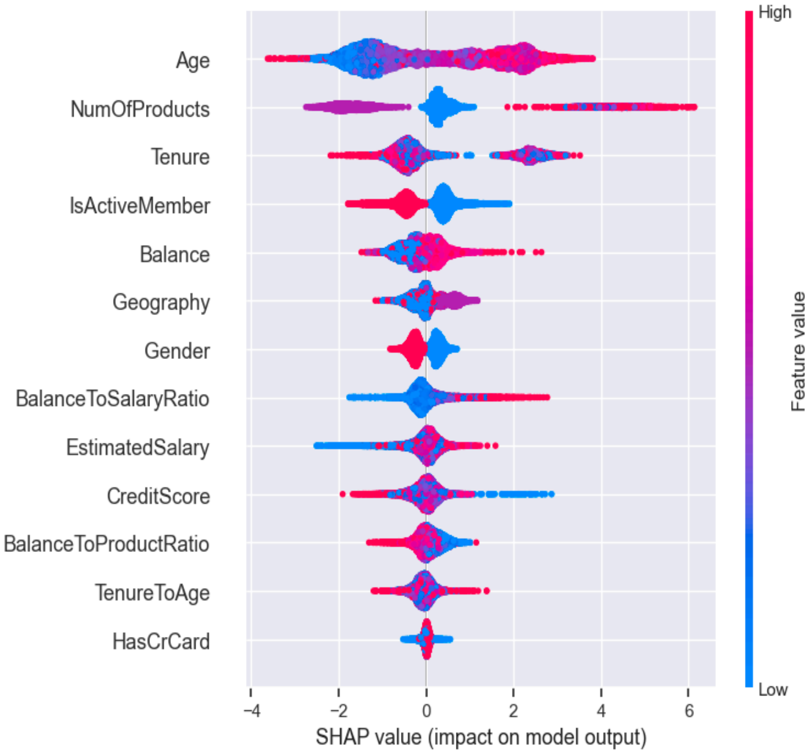


Figure 13. SHAP summary plot for the RF classifier.

5. Discussion

The results obtained from our analysis provide insights into the factors influencing customer churn in the banking sector, the effectiveness of RF and LGBM models, and the roles of model calibration and interpretability techniques like SHAP in enhancing predictive performance and transparency.

The exploratory analysis reveals key patterns in customer demographics and financial behaviors that correlate with churn. Credit scores show a slightly positive skew, with most customers maintaining moderate to high scores. This indicates that while most customers may have good financial standing, there is a subset with particularly high scores, potentially representing loyal, creditworthy clients. The bimodal distribution of balances, with a notable peak at zero, suggests a large segment of inactive customers or those who may have recently closed accounts. The dual peaks in positive balances could reflect varying saving and spending habits, with one group maintaining minimal balance and another group utilizing the bank as their primary financial institution. Customers mostly possess one or two products, with fewer customers holding three or four products. This limited use of bank products may indicate a shallow engagement with the bank, potentially increasing the churn likelihood. Additionally, the imbalance in the churn rate variable confirms that while a majority remain with the bank, a notable minority do leave, requiring focused retention strategies.

The analysis of customer distributions by geography and gender suggests that churn patterns vary by region and gender, with Germany showing higher churn rates and women churning more frequently than men. These demographic variations point toward cultural or market-specific factors and potentially different experiences or expectations across customer segments.

The correlation matrix and SHAP importance plots emphasize several factors as significant predictors of churn:

(1) Age and Churn—age exhibits a moderate positive correlation with churn (Spearman coefficient of 0.32), indicating that older customers are more likely to leave the bank. This is confirmed by SHAP analysis, where age consistently ranks as the most influential feature in both the RF and LGBM models. This relationship could suggest that older customers might feel less connected to digital banking offerings or are more likely to reevaluate their financial relationships.

(2) NumOfProducts and Retention—the number of products negatively correlates with churn (-0.13), suggesting that customers with more products tend to remain with the bank. This implies that a higher product engagement fosters loyalty, possibly due to increased familiarity with bank offerings and a greater sense of reliance on the bank's services.

(3) Tenure and Balance—longer tenure and higher balances correlate with lower churn, although their impact is less pronounced than age or product engagement. These factors imply that customers with longer banking relationships and substantial balances are more stable clients, potentially because they have deeper financial roots in the institution.

In addition to financial implications, customer churn in banking entails various operational, reputational and regulatory risks. Financial risks are immediate, as banks lose revenue from departing customers, but operational costs rise as resources are diverted to acquire new clients. Reputational risks also impact customer perception, particularly if high churn rates signal dissatisfaction, which can deter potential new clients. Regulatory risks are further heightened, as high churn could attract compliance scrutiny of customer service practices. By examining influential factors such as age and product usage through SHAP analysis, our model identifies specific areas for targeted interventions, which could mitigate these risks. For instance, strategies focused on retaining older customers or those with limited product engagement could help reduce the overall risk exposure associated with churn.

Interestingly, geographic location and gender have lower feature importance scores, although they present some churn-related tendencies in exploratory analysis. This discrepancy suggests that while demographic characteristics contribute to churn, behavioral and financial variables (e.g., age, product usage, balance) have a more direct influence on churn prediction.

The RF and LGBM models demonstrated strong predictive performance, but calibration played a significant role in balancing false positives and false negatives, as illustrated in the confusion matrices and performance metrics. Calibration improved the recall of churn cases in the RF model, reducing false negatives, which is valuable for identifying at-risk customers. It slightly increased false positives, indicating a trade-off where the model prioritized detecting churners at the expense of some loyal customers. This shift may be preferable in situations where capturing churners early on has greater financial implications than occasionally targeting loyal customers. The calibrated LGBM model showed higher precision and recall for churn predictions, achieving an F1 score of 0.90 and an accuracy of 0.91. Its lower Brier score (0.0697) after calibration indicates improved reliability in its probability estimates, making it a robust choice for scenarios requiring confident probability-based decision-making. Overall, the LGBM model outperformed RF in predictive accuracy and calibration, suggesting its suitability for churn prediction in banking. The higher AUC score (0.96) of LGBM over RF (0.93) reinforces this, indicating that it discriminates better between churners and loyal customers across thresholds.

SHAP analysis provides valuable insights into how individual features influence churn predictions, enhancing transparency. Age emerges as the most impactful feature across both models. This aligns with banking industry trends, where older customers might face barriers to digital adoption or exhibit lower engagement. Additionally, SHAP plots reveal that older customers consistently have higher SHAP values for churn, underscoring the need to address their specific needs and concerns.

SHAP values for NumOfProducts show that customers with fewer products are more likely to churn, suggesting that promoting additional products could enhance retention. The bank could offer tailored promotions or introduce multi-product bundles to encourage product adoption among low-engagement customers.

Features like EstimatedSalary, Balance, and CreditScore also show moderate importance, reflecting that customers' financial stability influences churn behavior. This suggests that customers with greater financial resources are generally more loyal, possibly due to stronger financial commitments or satisfaction with the bank's services.

SHAP analysis also highlights the interplay between Tenure and Age in customer behavior, where newer, older customers exhibit a higher likelihood of churn. This may suggest that older customers with shorter tenure feel less loyalty or may have recently switched banks due to dissatisfaction with prior institutions. Targeted engagement strategies tailored for newer, older customers could help mitigate this risk.

The insights from model performance and SHAP analysis reveal several actionable strategies for improving customer retention in the banking sector. The bank could focus on retention initiatives for older customers and those with shorter tenure, as these groups show higher churn risks. Personalized engagement, enhanced services, or digital assistance for older clients could improve their experience and loyalty. Encouraging customers to diversify their product portfolio may strengthen their attachment to the bank. Offering rewards or discounts for customers who adopt additional products (e.g., savings accounts, credit cards or loans) may help improve retention. Customers with high balances and stable credit scores show lower churn rates, suggesting that targeting financially stable clients with exclusive perks or tailored financial products could further reinforce loyalty.

6. Conclusions

Our research provides an in-depth exploration of developing and calibrating predictive models for customer churn in the banking sector using advanced machine learning techniques, namely RF and LGBM classifiers. The LGBM model demonstrated a slight edge in predictive performance, with higher AUC values in ROC analysis, showcasing its ability to capture the nuances in customer behavior and identify churn with greater accuracy. This indicates the model's robustness in distinguishing between customers who are likely to leave and those who will stay, even within complex data distributions.

Calibration of the models resulted in noticeable performance improvements, particularly for the RF model. The calibrated RF model saw a significant increase in recall for class 1 (churn cases), rising from 0.78 to 0.81, indicating a better capacity to identify customers likely to churn. However, this improvement in recall came with a slight drop in precision, from 0.89 to 0.88, reflecting a trade-off where the model became more sensitive to detecting churn at the expense of a higher number of false positives. The enhancement in the F1 score from 0.83 to 0.84 suggests that calibration helped achieve a better balance between precision and recall. Despite these benefits, the Brier score increased from 0.115 to 0.1503, showing that while the calibrated model better differentiated between churn and non-churn customers, its probability estimates became less confident.

The impact of calibration on the LGBM model was more modest. Precision for class 1 improved slightly from 0.93 to 0.94, while recall saw a minor decline from 0.88 to 0.87, resulting in a consistent F1 score of 0.90. The decrease in the Brier score from 0.0732 to 0.0697 after calibration indicated a positive adjustment, suggesting a closer alignment of predicted probabilities with actual outcomes. This showed that calibration had a more substantial impact on the RF model, highlighting the importance of refining models that initially exhibit less accurate probability estimations.

Data balancing also played a crucial role in improving model performance. A comparison of baseline models with those incorporating undersampling and SMOTETomek techniques demonstrated substantial gains, particularly in recall. For instance, the RF baseline model achieved a recall of only 0.45, while the calibrated RF model reached 0.81, showing that balancing the data helped the model identify a significantly higher number of actual churn cases, thereby reducing the number of false negatives. Similarly, the LGBM baseline model's recall increased from 0.47 to 0.88 in the uncalibrated version, further emphasizing the positive impact of addressing class imbalance.

To further enhance model transparency and interpretability, SHAP analysis was employed. SHAP allowed for a clear understanding of how individual features influenced model predictions, revealing age, number of banking products and tenure as the most significant drivers of churn. This interpretability is introduced for regulatory compliance and building trust, as it provides a data-driven rationale for each prediction and highlights key areas for targeted customer retention strategies.

The study also revealed important patterns during the data exploration phase, such as the distribution of customer age, balance and number of banking products. The obtained insights helped contextualize the churn dynamics, showing that certain demographic and financial factors may have substantial impacts on customer decisions to stay or leave. Overall, the calibrated models provided more balanced predictions, improving the recall of churn cases while maintaining precision, thus offering banks more reliable tools for managing customer attrition.

Based on the insights gained from feature importance analysis, banks can prioritize retention efforts around the factors most strongly associated with churn. For example, customer age emerged as a significant predictor, suggesting that personalized strategies that address the needs of different age groups could be effective. Younger customers may be more responsive to digital banking solutions and rewards programs, while older customers might value personalized financial advice or better interest rates on savings accounts. Financial factors such as the number of bank products used, credit score, and estimated salary also showed considerable influence on churn behavior. Banks could implement strategies such as offering incentives to increase product usage, providing credit score improvement programs, or offering salary-based tailored financial services to retain these customers. Furthermore, for customers with lower tenure (shorter relationship duration with the bank), onboarding programs that engage new customers with exclusive benefits or loyalty programs may help build a stronger long-term relationship.

This study has several limitations that should be acknowledged. First, the models relied on a static dataset, meaning that the analysis was based on historical data snapshots rather than real-time or longitudinal data. Consequently, the predictions may not account for rapidly changing customer behaviors. Second, while data balancing techniques like SMOTETomek improved recall, they may also introduce synthetic noise, potentially affecting model generalization. Additionally, only random forest and LightGBM models were considered, leaving out other potentially valuable algorithms like neural networks or support vector machines. Finally, the models used standard hyperparameter settings, and further fine-tuning could potentially yield better performance.

Future research could extend this study by exploring additional machine learning techniques or model ensembles that may further improve churn prediction. Incorporating more advanced probability calibration techniques or optimizing hyperparameters specific to each model could also enhance the alignment between predicted probabilities and actual outcomes. Additionally, the integration of more customer behavioral data or external economic indicators might reveal further insights into churn patterns. Another avenue for future work involves investigating real-time churn prediction, where models are updated continuously with streaming data, enabling timely interventions. A comparative analysis of different data balancing techniques beyond SMOTETomek, such as ADASYN or generative adversarial networks (GANs), could be conducted to understand their specific effects on model performance.

Author Contributions: Conceptualization, S.-V.O., A.-M.N., A.B. and A.-I.A.; data curation, A.-G.V. and A.-M.N.; investigation, A.-G.V., A.-M.N. and A.B.; methodology, A.-G.V. and S.-V.O.; supervision, S.-V.O., A.B. and A.-I.A.; validation, S.-V.O.; visualization, A.-G.V. and A.-I.A.; writing—original draft, A.-G.V., S.-V.O. and A.-M.N.; writing—review and editing, S.-V.O., A.-M.N., A.B. and A.-I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI—UEFISCDI, project number COFUND-DUT-OPEN4CEC-1, within PNCDI IV.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

References

1. Saran Kumar, A.; Chandrakala, D. A Survey on Customer Churn Prediction Using Machine Learning Techniques. *Int. J. Comput. Appl.* **2016**, *154*, 13–16. [CrossRef]
2. Gür Ali, Ö.; Aritürk, U. Dynamic Churn Prediction Framework with More Effective Use of Rare Event Data: The Case of Private Banking. *Expert. Syst. Appl.* **2014**, *41*, 7889–7903. [CrossRef]
3. Chen, T.H. Do You Know Your Customer? Bank Risk Assessment Based on Machine Learning. *Appl. Soft Comput.* **2020**, *86*, 105779. [CrossRef]
4. Hemalatha, P.; Amalanathan, G.M. A Hybrid Classification Approach for Customer Churn Prediction Using Supervised Learning Methods: Banking Sector. In Proceedings of the International Conference on Vision Towards Emerging Trends in Communication and Networking, ViTECoN, Vellore, India, 30–31 March 2019. [CrossRef]
5. Karvana, K.G.M.; Yazid, S.; Syalim, A.; Mursanto, P. Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. In Proceedings of the 2019 International Workshop on Big Data and Information Security, IWBIS, Bali, Indonesia, 11 October 2019; pp. 33–38. [CrossRef]
6. Lalwani, P.; Mishra, M.K.; Chadha, J.S.; Sethi, P. Customer Churn Prediction System: A Machine Learning Approach. *Computing* **2022**, *104*, 271–294. [CrossRef]
7. Naik, K.S. Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach. *arXiv* **2021**, arXiv:2110.02206.
8. Petkovic, D.; Altman, R.; Wong, M.; Vigil, A. Improving the Explainability of Random Forest Classifier—User Centered Approach. *Pac. Symp. Biocomput.* **2018**, *23*, 204–215. [CrossRef]
9. Prabadevi, B.; Shalini, R.; Kavitha, B.R. Customer Churning Analysis Using Machine Learning Algorithms. *Int. J. Intell. Netw.* **2023**, *4*, 145–154. [CrossRef]
10. Rufibach, K. Use of Brier Score to Assess Binary Predictions. *J. Clin. Epidemiol.* **2010**, *63*, 938–939. [CrossRef]

11. Singh, P.P.; Anik, F.I.; Senapati, R.; Sinha, A.; Sakib, N.; Hossain, E. Investigating Customer Churn in Banking: A Machine Learning Approach and Visualization App for Data Science and Management. *Data Sci. Manag.* **2024**, *7*, 7–16. [CrossRef]
12. Guliyev, H.; Tatoğlu, F.Y. Customer Churn Analysis in Banking Sector: Evidence from Explainable Machine Learning Models. *J. Appl. Microeconometrics* **2021**, *1*, 85–99. [CrossRef]
13. De Lima Lemos, R.A.; Silva, T.C.; Tabak, B.M. Propension to Customer Churn in a Financial Institution: A Machine Learning Approach. *Neural Comput. Appl.* **2022**, *34*, 11751–11768. [CrossRef] [PubMed]
14. Simsek, M.; Tas, I.C. A Classification Application for Using Learning Methods in Bank Costumer's Portfolio Churn. *J. Forecast.* **2024**, *43*, 391–401. [CrossRef]
15. Alizadeh, M.; Zadeh, D.S.; Moshiri, B.; Montazeri, A. Development of a Customer Churn Model for Banking Industry Based on Hard and Soft Data Fusion. *IEEE Access* **2023**, *11*, 29759–29768. [CrossRef]
16. Valluri, C.; Raju, S.; Patil, V.H. Customer Determinants of Used Auto Loan Churn: Comparing Predictive Performance Using Machine Learning Techniques. *J. Mark. Anal.* **2022**, *10*, 279–296. [CrossRef]
17. Tékouabou, S.C.K.; Gherghina, Ș.C.; Toulni, H.; Mata, P.N.; Martins, J.M. Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods. *Mathematics* **2022**, *10*, 2379. [CrossRef]
18. Xie, Y.; Li, X. Churn Prediction with Linear Discriminant Boosting Algorithm. In Proceedings of the 7th International Conference on Machine Learning and Cybernetics, ICMLC, Kunming, China, 12–15 July 2008; Volume 1, pp. 228–233. [CrossRef]
19. Mengash, H.A.; Alruwais, N.; Kouki, F.; Singla, C.; Abd Elhameed, E.S.; Mahmud, A. Archimedes Optimization Algorithm-Based Feature Selection with Hybrid Deep-Learning-Based Churn Prediction in Telecom Industries. *Biomimetics* **2023**, *9*, 1. [CrossRef] [PubMed]
20. Vu, V.H. Predict Customer Churn Using Combination Deep Learning Networks Model. *Neural Comput. Appl.* **2024**, *36*, 4867–4883. [CrossRef]
21. Zaky, A.; Ouf, S.; Roushdy, M. Predicting Banking Customer Churn Based on Artificial Neural Network. In Proceedings of the 5th International Conference on Computing and Informatics, ICCI, Cairo, Egypt, 9–10 March 2022; pp. 132–139. [CrossRef]
22. Dankowski, T.; Ziegler, A. Calibrating Random Forests for Probability Estimation. *Stat. Med.* **2016**, *35*, 3949. [CrossRef]
23. Ojeda, F.M.; Jansen, M.L.; Thiéry, A.; Blankenberg, S.; Weimar, C.; Schmid, M.; Ziegler, A. Calibrating Machine Learning Approaches for Probability Estimation: A Comprehensive Comparison. *Stat. Med.* **2023**, *42*, 5451–5478. [CrossRef]
24. Davis, S.E.; Lasko, T.A.; Chen, G.; Siew, E.D.; Matheny, M.E. Calibration Drift in Regression and Machine Learning Models for Acute Kidney Injury. *J. Am. Med. Inform. Assoc.* **2017**, *24*, 1052–1061. [CrossRef]
25. Ngo, V.-B.; Vu, V.-H. Multi-Level Machine Learning Model to Improve the Effectiveness of Predicting Customers Churn Banks. *Cybern. Inf. Technol.* **2024**, *24*, 3–20. [CrossRef]
26. Domingos, E.; Ojeme, B.; Daramola, O. Experimental Analysis of Hyperparameters for Deep Learning-Based Churn Prediction in the Banking Sector. *Computation* **2021**, *9*, 34. [CrossRef]
27. Elyusufi, Y.; Kbir, M.A. Churn Prediction Analysis by Combining Machine Learning Algorithms and Best Features Exploration. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 615–622. [CrossRef]
28. Vijayakumar Bharathi, S.; Pramod, D.; Raman, R. An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers. *Data* **2022**, *7*, 61. [CrossRef]
29. Chang, V.; Xu, Q.A.; Akinloye, S.H.; Benson, V.; Hall, K. Prediction of Bank Credit Worthiness through Credit Risk Analysis: An Explainable Machine Learning Study. *Ann. Oper. Res.* **2024**, 1–25. [CrossRef]
30. Zdziebko, T.; Sulikowski, P.; Sałabun, W.; Przybyła-Kasperek, M.; Bąk, I. Optimizing Customer Retention in the Telecom Industry: A Fuzzy-Based Churn Modeling with Usage Data. *Electronics* **2024**, *13*, 469. [CrossRef]
31. Li, J.; Bai, X.; Xu, Q.; Yang, D. Identification of Customer Churn Considering Difficult Case Mining. *Systems* **2023**, *11*, 325. [CrossRef]
32. Chang, V.; Hall, K.; Xu, Q.A.; Amao, F.O.; Ganatra, M.A.; Benson, V. Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. *Algorithms* **2024**, *17*, 231. [CrossRef]
33. Kavyarshitha, Y.; Sandhya, V.; Deepika, M. Churn Prediction in Banking Using ML with ANN. In Proceedings of the 2022 6th International Conference on Intelligent Computing and Control Systems, ICICCS, Madurai, India, 25–27 May 2022; pp. 1191–1198. [CrossRef]
34. Soni, A.; Mishra, J.; Dixit, M. Comparative Study of Bank Customers Churn Prediction Using AI/ML. In Proceedings of the 2024 IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT), Jabalpur, India, 6–7 April 2024; pp. 1359–1365. [CrossRef]
35. Hui, S.H.; Khai, W.K.; XinYing, C.; Wai, P.W. Prediction of Customer Churn for ABC Multistate Bank Using Machine Learning Algorithms / Hui Shan Hon. . . [et al.]. *Malays. J. Comput. (MJoC)* **2023**, *8*, 1602–1619.
36. Rahman, M.; Kumar, V. Machine Learning Based Customer Churn Prediction In Banking. In Proceedings of the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 5–7 November 2020; pp. 1196–1201. [CrossRef]
37. Li, L.J.; Junn, K.Y. Decision Tree with Genetic Algorithm for Bank Customer Churn Prediction. In Proceedings of the 2023 IEEE 21st Student Conference on Research and Development, SCOReD, Kuala Lumpur, Malaysia, 13–14 December 2023; pp. 426–431. [CrossRef]
38. Charandabi, S.E. Prediction of Customer Churn in Banking Industry. *arXiv* **2023**, arXiv:2301.13099. [CrossRef]

39. Han, S. Machine Learning Based Customer Churn Prediction in Banking Sector. *Highlights Bus. Econ. Manag.* **2024**, *40*, 378–384. [CrossRef]
40. Saxena, A.; Singh, A.; Govindaraj, M. Analyzing Customer Churn in Banking: A Data Mining Framework. *Multidiscip. Sci. J.* **2023**, *5*, 2023ss0310. [CrossRef]
41. Yang, C. Machine Learning Algorithms Based Prediction for Customer Churn in Banks. *Highlights Bus. Econ. Manag.* **2024**, *40*, 352–358. [CrossRef]
42. Khine, S.T.; Myo, W.W. Mining Customer Churns for Banking Industry Using K-Means and Multi-Layer Perceptron. In Proceedings of the IEEE International Conference on Control and Automation, ICCA, Yangon, Myanmar, 27–28 February 2023; pp. 220–225. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.