APPLICATION OF MACHINE LEARNING IN THE SEGMENTATION OF CUSTOMERS ADOPTING OPEN FINANCE IN BRAZIL

APLICAÇÃO DE MACHINE LEARNING NA SEGMENTAÇÃO DE CLIENTES ADERENTES AO OPEN FINANCE NO BRASIL

Felipe Lima de Holanda
Mestre em Governança, Tecnologia e Inovação pela Universidade Católica de Brasília (Brasília/Brasil).
E-mail: felipelhol@gmail.com

Paulo Fernando Marschner
Doutor em Administração pela Universidade Federal de Santa Maria (Santa Maria/BraisI).
Professor da Universidade Católica de Brasília (Brasília/Brasil).
E-mail: paulofernandomarschner@hotmail.com

Hércules Antônio do Prado
Doutorado em Computação pela Universidade Federal do Rio Grande do Sul (Porto Alegre/Brasi).
Professor da Universidade Católica de Brasília (Brasília/Brasil).
E-mail: hercules@p.ucb.br

Recebido em: 10 de junho de 2025
Aprovado em: 12 de agosto de 2025
Sistema de Avaliação: Double Blind Review
RGD | v. 22 | n. 2 | p. 27-51 | jul./dez. 2025
DOI: https://doi.org/10.25112/rgd.v22i2.4263

Revista Gestão e Desenvolvimento

e-ISSN: 2446-6875
p-ISSN: 1807-5436

ABSTRACT

This study aimed to apply machine learning techniques to identify the profile of customers most likely to adopt Open Finance within a major Brazilian financial institution classified as S1 by the Central Bank of Brazil, meaning it holds assets equal to or greater than 10% of the national GDP or has international relevance. The research employed real, individual-level data, enabling the development of robust predictive models with high practical applicability. Five techniques widely recognized in the literature were evaluated: Random Forest, Support Vector Machine (SVM), Decision Tree, Logistic Regression, and XGBoost. Among the models tested, XGBoost demonstrated the best performance, achieving an AUC of 0.90 and an Accuracy of 0.82. The most relevant predictor of Open Finance adoption was the customer's digital profile, followed by individuals with income up to R$2,000.00 and/or investments up to R$5,000.00, and customers aged over 64. These findings offer valuable insights for financial institutions and policymakers by highlighting the importance of segmented communication strategies and digital inclusion.

Keywords: Machine Learning, Open Finance, Financial Institution, Marketing.

RESUMO

Este estudo teve como objetivo aplicar técnicas de machine learning para identificar o perfil dos clientes mais propensos a aderir ao Open Finance em uma grande instituição financeira brasileira classificada como S1 pelo Banco Central do Brasil, ou seja, com ativos iguais ou superiores a 10% do PIB ou com relevância internacional. A pesquisa utilizou dados reais, em nível individual, o que permitiu construir modelos preditivos robustos e com alto potencial de aplicação prática. Foram avaliadas cinco técnicas amplamente reconhecidas na literatura: Random Forest, Support Vector Machine (SVM), Decision Tree, Logistic Regression e XGBoost. Entre os modelos testados, o XGBoost apresentou o melhor desempenho, com uma AUC de 0,90 e Acurácia de 0,82. O atributo mais relevante para prever a adesão ao Open Finance foi o perfil digital do cliente, seguida por indivíduos com renda de até R$2.000,00 e/ou investimentos de até R$5.000,00, e clientes com mais de 64 anos. Esses achados oferecem subsídios valiosos para instituições financeiras e formuladores de políticas, ao indicar a importância de estratégias segmentadas de comunicação e inclusão digital.

Palavras-chave: Aprendizado de máquina, Open Finance, Instituição Financeira, Marketing.

e-ISSN: 2446-6875
p-ISSN: 1807-5436

1 INTRODUCTION

In the context of rapid technological transformation and growing competition in the financial sector, institutions are seeking innovative strategies to add value and enhance the customer experience. In this regard, banks have advanced initiatives to strengthen digital channels, foster partnerships for the development of new ideas, and modernize internal resources through technologies such as cloud computing and artificial intelligence (CARAFFINI et al., 2023). Open Finance, or the open financial system, has emerged in this scenario as a regulatory initiative that promotes the sharing of financial data among banks, fintechs, and other service providers authorized by the Central Bank of Brazil (BANCO CENTRAL DO BRASIL, n.d.).

According to the 2024 Annual Report of Open Finance Brasil, the Brazilian ecosystem reached 27.7 million unique consents in 2023, totaling 42 million active consents and 946 participating institutions, approximately 15% of the banked population. By August 2024, this number had risen to around 32.98 million, reflecting a 19% increase compared to the end of 2023. In comparison, the United Kingdom (one of the pioneers in implementing the system) covers 13% of its banked population.

In Brazil, Open Banking laid the groundwork for Open Finance by introducing structured data sharing for checking and savings accounts. Open Finance expands this scope to include investment products, insurance, pensions, and other services (Banco Central do Brasil, 2023). Although Open Banking has been widely discussed in the literature (ADKE *et al.*, 2022; BARTELS, 2022; HJELKREM; LANGE, 2023; IMAN *et al.*, 2023), Open Finance still lacks in-depth studies, especially in emerging markets.

Voluntary adherence to Open Finance heightens the need for strategies that consider user behavior and profile. Identifying customers who are more likely to adopt the system enables the development of personalized actions, enhances customer retention, and optimizes resources. The literature emphasizes that retaining customers is more cost-effective than acquiring new ones (XIAO *et al.*, 2015; WEN *et al.,* 2019), underscoring the importance of understanding consumer behavior in the Open Finance environment.

However, despite valuable contributions such as Targher (2023), who analyzes the system from a regulatory perspective, national literature has yet to explore the identification of adopter customer profiles. Internationally, Grassi *et al.* (2022) discuss the value of data for banks and fintechs, focusing on organizational challenges, while Grassi (2024) analyzes data-sharing willingness in the insurance sector. Mishra *et al.* (2024) explore personalization through emerging technologies, without directly addressing Open Finance or customer segmentation.

Simultaneously, studies applying machine learning algorithms to analyze financial institutions' customer profiles have gained prominence. Palaniappan *et al.* (2017) and Dawood *et al.* (2019) used

e-ISSN: 2446-6875
p-ISSN: 1807-5436

algorithms such as Naïve Bayes, Random Forest, and Decision Tree for segmentation and marketing campaigns. Patil and Dharwadkar (2017) demonstrated the performance of Artificial Neural Networks (ANNs). Yang and Zhang (2018) and Niloy and Navid (2018) focused on default prediction using LightGBM and XGBoost. Despite these advances, the studies focus on other markets and do not consider the specificities of the Brazilian context.

Dawood *et al.* (2019) emphasize that banks must segment their vast datasets to understand customer behavior and formulate more effective strategies. Segmentation provides a rich description of customer profiles based on specific attributes and behaviors. In this regard, machine learning significantly contributes to forecasting future patterns and supporting strategic decisions in personalized financial products.

Given this context, the objective of this study is to apply machine learning techniques to identify the profile of customers most likely to adopt Open Finance in a major Brazilian financial institution classified as S1 by the Central Bank of Brazil (institutions with assets ≥10% of GDP or international relevance). The use of real, individual-level data enabled the development of a robust predictive model, avoiding overestimated forecasts and ensuring practical applicability.

The results showed that the XGBoost model was the most effective, with an AUC of 0.90 and Accuracy of 0.82, outperforming other classifiers. The customer's digital profile emerged as the most relevant attribute for predicting Open Finance adoption, followed by customers with an income of up to R$2,000.00 and/or investments of up to R$5,000.00, and clients aged over 64. One of the main contributions of this research is the use of real and representative data, which differentiates it from studies based on limited samples or synthetic data. Although the General Data Protection Law (LGPD) poses challenges to accessing high-quality data in the financial sector, as noted by Reddy (2020), this study overcame that obstacle by handling sensitive data responsibly, providing valuable insights for the industry.

The contributions of this work are particularly relevant for financial institutions and policymakers, as they point to directions for personalized communication strategies and digital inclusion, essential for expanding Open Finance adoption. However, reliance on data from a single institution and the limitations of the study's scope suggest caution in generalizing the results. Future research may expand the database, include behavioral variables, and adopt longitudinal approaches to better understand the evolution of Open Finance adoption over time.

Revista Gestão e Desenvolvimento

e-ISSN: 2446-6875
p-ISSN: 1807-5436

2 LITERATURE REVIEW

Understanding customer behavior and characteristics is a central element in developing effective strategies across various sectors, particularly in banking. In recent years, the use of machine learning algorithms has emerged as a promising approach for analyzing and segmenting customer profiles. Recent studies have demonstrated the power of techniques such as neural networks, decision trees, random forests, among others, in identifying consumption and default patterns, as well as in predicting the adoption of financial services. In the context of Open Finance, which promotes greater transparency and interoperability in financial services, the use of these tools can provide deeper insights into customer needs and behaviors, enabling more personalized and efficient financial product offerings.

Palaniappan *et al.* (2017) developed a customer profiling model using data from a Portuguese retail bank collected between 2008 and 2013, aiming to enhance customer profile definition and identify those with a high probability of subscribing to long-term deposits. Three classification algorithms were applied: Naïve Bayes, Random Forest, and Decision Tree. The models' performance was evaluated based on metrics such as accuracy, precision, and recall rate. The results showed that classification techniques are effective in predicting customer profiles and increasing the efficiency of telemarketing campaigns, thereby improving fundraising strategies.

Patil and Dharwadkar (2017) proposed a prediction and classification model using Artificial Neural Networks (ANN) on two datasets of bank customers. The study used the ANN algorithm as the core technique, followed by a weighting of the results. The findings showed that the algorithm performed satisfactorily on both datasets, with accuracy rates of 72% for the first dataset and 98% for the second, indicating variations according to the characteristics of each dataset. These findings highlight ANNs as effective tools for predictive analysis in the banking sector.

Yang and Zhang (2018) proposed a classification model to predict credit card defaults using data from a bank in Taiwan. The study compared five data mining methods: Logistic Regression, Support Vector Machines (SVM), Neural Networks, XGBoost, and LightGBM. Model validation was conducted through cross-validation, focusing on estimating the area under the curve (AUC) and accuracy rate. LightGBM achieved the best results, with a precision rate of 89.34% based on the F1-score metric, followed by XGBoost. The results indicate that both algorithms perform well in predicting categorical variables, showing great potential for big data applications.

Niloy and Navid (2018) developed a classification model to identify defaulting customers based on a credit card default dataset from a bank in Taiwan. The study compared two machine learning algorithms: the Naïve Bayes Classifier and Decision Trees. Both algorithms were evaluated for their predictive capabilities in the context of credit risk management. The Naïve Bayes Classifier achieved higher accuracy

e-ISSN: 2446-6875
p-ISSN: 1807-5436

compared to Decision Trees. Both methods are widely used in supervised learning and data mining. The results indicate that the Naïve Bayes Classifier is more efficient in predicting customer reliability, with greater accuracy in distinguishing between defaulters and non-defaulters.

Dawood *et al.* (2019) analyzed the evolution of credit cards in the banking sector, emphasizing customer profiling as a key strategy to improve decisions related to offers and credit limits. Effective profiles enable better understanding of customers and increase profitability by targeting the most valuable ones. Unlike previous studies that used transactional or demographic data separately, this study combined both for greater accuracy and reduced risk. Techniques such as k-means, enhanced k-means, fuzzy c-means, and neural networks were applied to a labeled dataset. A distinctive feature was the creation of a new label for neural network classification, which reduced execution time and increased accuracy. A comparison of accuracy rates showed that the neural network was the most effective technique.

Carbo-Valverde *et al.* (2020) investigated the digitization process of banking customers, highlighting the importance of this transition for strategies aimed at attracting and retaining online users, as well as the increasing competition from BigTechs and FinTechs. Using machine learning, they applied random forests, conditional inference trees, and causal forests to identify predictive characteristics of digital banking service adoption. The results showed that random forests were the most accurate, with a prediction rate of 88.41% for adoption and use of online services. Adoption follows a sequence, beginning with informational services and progressing to transactional services. The diversification in online channel use is explained by awareness of services and perceived security. The study identified a complementarity between digital and non-banking channels and suggested that banks adopt a segmented approach, offering personalized services, while policymakers should promote awareness of digital financial services.

Despite the significant contributions of the reviewed studies, it is important to highlight some limitations, particularly regarding the generalization of results to different contexts. Many of the studies were conducted in specific markets, such as European and Asian ones, which may limit the applicability of the models to other economic and cultural settings. Additionally, most research focuses on historical data from traditional banking customers, without considering the emerging dynamics of Open Finance, which involves greater data sharing across institutions and new financial platforms. In this regard, understanding the profile of Open Finance customers in Brazil using machine learning is extremely relevant, as the Brazilian context presents unique challenges and opportunities in terms of economic diversity, access to financial services, and regulation. With the advancement of Open Finance, there is great potential for service personalization, but complex issues also arise regarding privacy, security, and financial inclusion. Therefore, it is essential for future research to focus on adapting predictive models to the Brazilian context, promoting a more accurate understanding of customer profiles and optimizing the financial experience of consumers within the new digital financial ecosystem.

e-ISSN: 2446-6875
p-ISSN: 1807-5436

3 METHODOLOGICAL PROCEDURES

To achieve the proposed objectives, this study adopted a quantitative and descriptive methodological approach, based on the collection of secondary data.

3.1 RESEARCH DESIGN

With respect to its nature, this study is classified as quantitative. Research of this type can be employed when the central aim is to model the relationship between variables, or to measure attitudes, perceptions, behaviors, feelings, and practices, among others (MANZATO; SANTOS, 2012). Quantitative research also follows a pre-established design, taking into account measures defined by the researcher so that the analysis of events is carried out in a precise and objective manner (PROETTI, 2018).

In terms of objectives, this study is classified as descriptive. Within the descriptive process, the goals may involve identifying, recording, and analyzing the attributes, motivations, or factors related to the phenomena under investigation, since descriptive research inherently contains an observational component. The main contribution of this type of research is that it enables different perspectives on what is already known (NUNES *et al.,* 2016). Regarding technical procedures, this study is classified as quantitative, with a focus on analyzing key machine learning models to understand the profile of customers most likely to adopt Open Finance. This approach allows the institution to develop tailored strategies to better serve these segments.

3.2 DATA AND RESEARCH SAMPLE

The data for this study were obtained from a large Brazilian financial institution participating in Open Finance Brazil. The dataset contains 200,000 anonymized records, ensuring banking confidentiality. A total of 100,000 customers who opted in and 100,000 who did not were selected, resulting in a balanced dataset (50/50). The data were collected in April 2024 and cover consents granted between May 2023 and April 2024. The dataset contains 29 attributes, comprising 28 descriptors and 1 class attribute, with all records referring to individual customers residing in Brazil during the analyzed period. Only 1,938 consents from legal entities were identified and excluded, as their volume and representativeness were negligible. A stratified random sampling was performed, based on whether customers had consented to share data via Open Finance with another institution. Thus, the findings are generalizable to the rest of the dataset, increasing external validity. Table 1 presents the description of all attributes used in the study, including field names, descriptions, and data types for each.

Revista Gestão e Desenvolvimento

e-ISSN: 2446-6875
p-ISSN: 1807-5436

Table 1 – List of Attributes Used in the Study

Attribute	Description	Data Type
cSegmento	Customer segment. Classification performed by the financial institution based on criteria such as income and investment volume.	Categorical
iRelacionamento	Length of the customer's relationship with the financial institution	Numerical
iIdade	Customer's age	Numerical
bDigital	Indicates whether the customer used a digital channel for transactions in the last 12 months	Binary
bCorrente	Indicates whether the customer has a checking account	Binary
bPoup	Indicates whether the customer has a savings account	Binary
bCartCred	Indicates whether the customer has an active credit card	Binary
bCredCom	Indicates whether the customer has an active personal loan	Binary
bCredHab	Indicates whether the customer has an active mortgage loan	Binary
bProtecao	Indicates whether the customer holds a protection product (life, property)	Binary
bInad	Indicates whether the customer is delinquent on any credit operation	Binary
cGenero	Customer's gender	Categorical
vUF	Indicates the customer's state of residence	Categorical
mRenda	Customer's income	Numerical
bBenef	Indicates whether the customer receives any social benefits	Binary
bCredSal	Indicates whether the customer receives salary credit in the bank	Binary
bInvest	Indicates whether the customer is an investor	Binary
bServPub	Indicates whether the customer is a public servant	Binary
mVolCred	Indicates the volume of credit operations the customer has with the bank	Numerical
mVolInvest	Indicates the volume of the customer's investments with the bank	Numerical
mVolEndivid	Indicates the customer's total debt volume in the market	Numerical
iIFMercado	Indicates the number of financial institutions where the customer holds an account	Numerical
iTotalProd	Total number of products the customer holds with the bank	Numerical
iPoupMovQtdeUlt12Meses	Total number of savings account transactions in the last 12 months	Numerical
iCorrMovQtdeUlt12Meses	Total number of checking account transactions in the last 12 months	Numerical
dCorrentePrimeira	Opening date of the first checking account (dd/mm/yyyy)	Date
dPoupPrimeira	Opening date of the first savings account (dd/mm/yyyy)	Date
bClienteEncart	Indicates whether the customer is managed by a relationship manager	Binary
bCompartilha_Opf	Indicates whether the customer has shared data via Open Finance	Binary

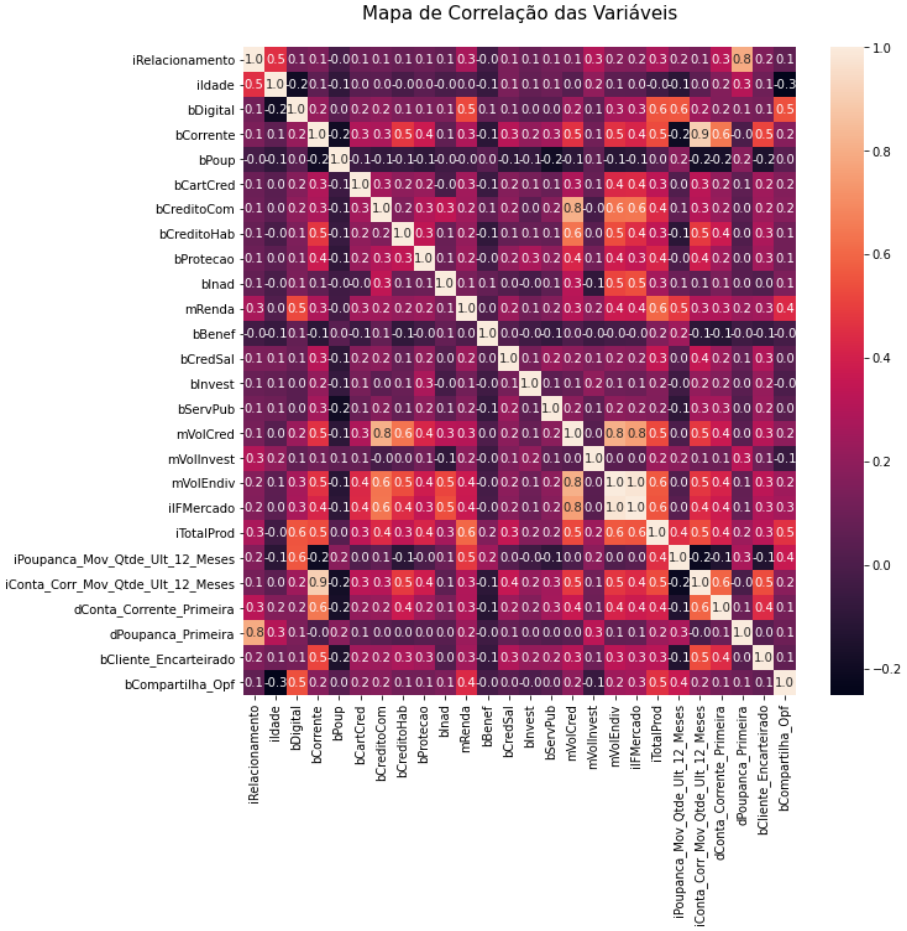
Source: Prepared by the authors.

Revista Gestão e Desenvolvimento

e-ISSN: 2446-6875
p-ISSN: 1807-5436

The table presents data on demographic aspects, banking behavior, relationship with the institution, and customers' financial characteristics. The variables analyzed (categorical, numerical, and binary) include the customer segment, number of active products, relationship history, and occurrence of delinquency. These data are essential for analyzing and segmenting customer profiles. An exploratory analysis was conducted using a dataset of 200,000 records and 29 attributes, resulting in a correlation map. Figure 1 illustrates the correlations between variables, using colors to represent the strength and direction of the associations: values close to 1 indicate a strong positive correlation, values close to -1 indicate a negative correlation, and values close to 0 indicate weak or no association.

Figure 1 – Correlation Matrix of the Variables



Source: Prepared by the authors.

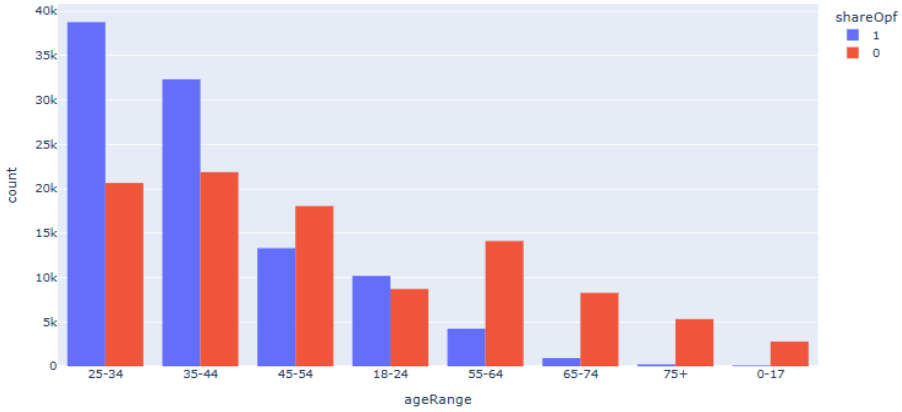
The correlation map identified significant patterns, with some variables showing strong positive correlations, indicating interdependencies that influence the modeling process. The variable *mVolCred*, for

Revista Gestão e Desenvolvimento
e-ISSN: 2446-6875
p-ISSN: 1807-5436

instance, showed a high correlation with *bCredCom* and *mVolEndivi* (ρ = 0.8), suggesting a relationship of similarity. Other variables exhibited low correlations (ρ close to 0), indicating independence. This analysis was crucial for identifying collinearity that could affect statistical models and machine learning algorithms. The use of Spearman's correlation ensured robustness against skewness and outliers, helping to exclude redundant variables. The main results are presented below.

Clients aged 25 to 44 lead the sharing of information, with the 25-34 age group standing out, representing nearly 40% of the total (Figure 2). This may be associated with higher digital activity and the adoption of new technologies that facilitate sharing. Furthermore, this stage of life typically involves career and financial development, requiring more interaction with financial institutions.

Figure 2 – Number of Clients by Age Group.



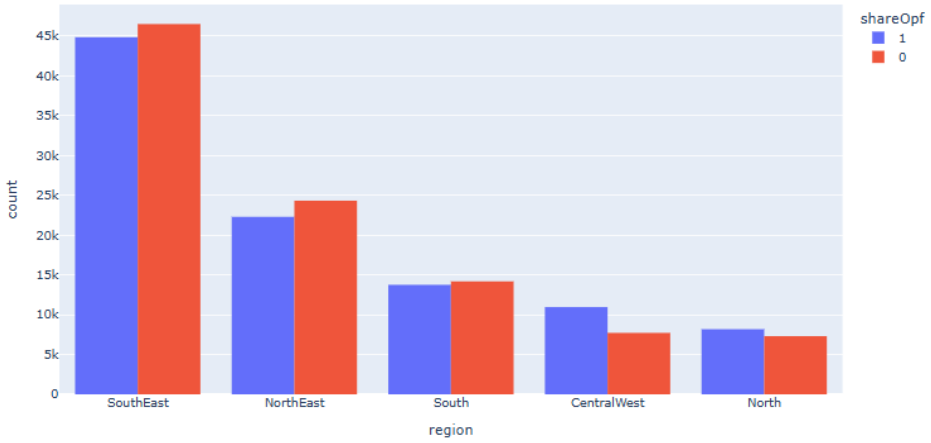
Note: The blue bars represent clients who share their information, while the red bars correspond to clients who choose not to share it.

Source: Prepared by the authors.

The Southeast region leads in information sharing, followed by the Northeast, which accounts for just over 20% of the total (Figure 3). This pattern may be associated with higher population density and more advanced economic development in the Southeast, which provides greater access to financial services and technology. Meanwhile, the Northeast, though also significant, shows behavior aligned with its recent economic growth and the increasing adoption of digital technologies.

Revista Gestão e Desenvolvimento
e-ISSN: 2446-6875
p-ISSN: 1807-5436

Figure 3 – Number of Clients by Region.

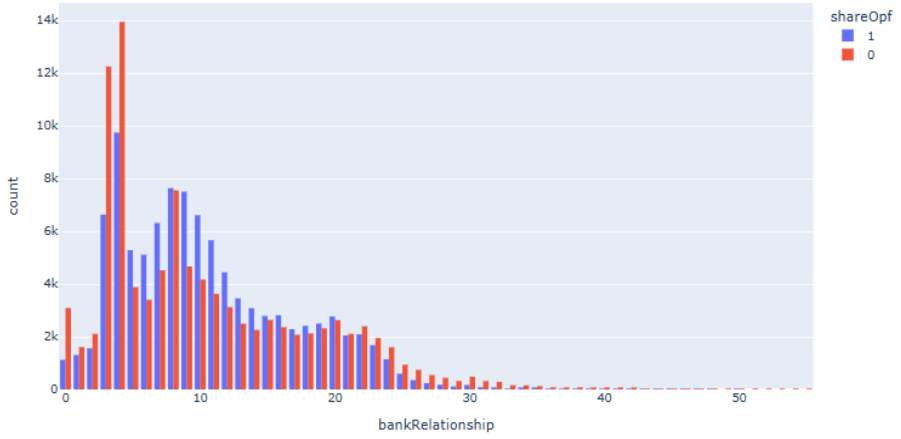


Note: The blue bars represent clients who share their information, while the red bars correspond to clients who choose not to share it.

Source: Prepared by the authors.

Clients with up to ten years of relationship with the financial institution are the ones who share data the most with other institutions (Figure 4). The reduction in sharing after this period suggests greater loyalty or trust in the primary institution, reducing the need to interact with others. This pattern reflects a strengthening of the bond, indicating a lower likelihood of churn and higher satisfaction with the services provided.

Figure 4 – Number of Clients by Length of Relationship.



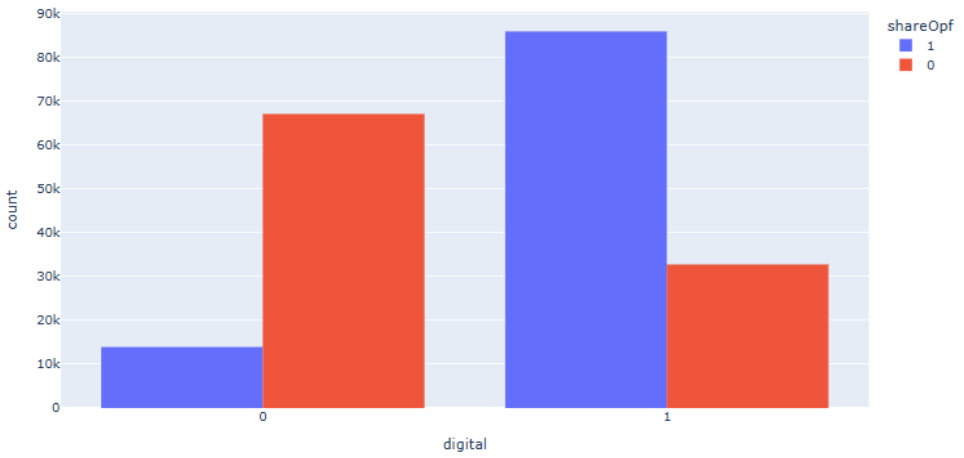
Note: The blue bars represent clients who share their information, while the red bars correspond to clients who choose not to share it.

Source: Prepared by the authors.

Revista Gestão e Desenvolvimento
e-ISSN: 2446-6875
p-ISSN: 1807-5436

Clients who share data the most have a digital profile, possibly due to greater familiarity and comfort with digital platforms (Figure 5). This may indicate that these clients seek convenience and better offers through constant comparisons. Financial institutions wishing to retain these clients should invest in technology and personalized digital offerings to keep them increasingly engaged and satisfied.

Figure 5 – Number of Clients with a Digital Profile.



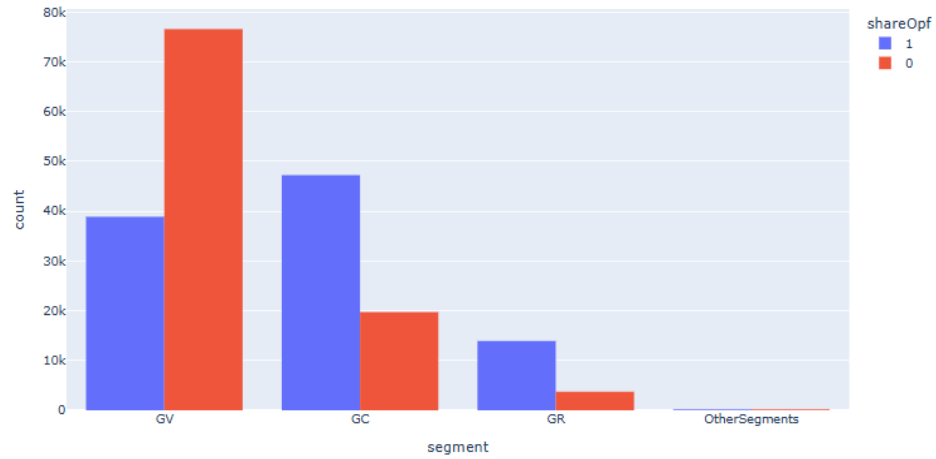
Note: The blue bars represent clients who share their information, while the red bars correspond to clients who choose not to share it.

Source: Prepared by the authors.

Clients in the GV segment (income up to 2,000 and/or investments up to 5,000) share less data compared to clients in the GC (income up to 7,000 and/or investments up to 100,000) and GR (income above 7,000 and/or investments above 100,000) segments (Figure 6). This suggests that higher-income and higher-investment clients have more interactions with various financial institutions, possibly seeking better offers or personalized services, while lower-income clients may have more limited access to multiple institutions or may be more conservative in managing their data.

Revista Gestão e Desenvolvimento
e-ISSN: 2446-6875
p-ISSN: 1807-5436

Figure 6 – Number of Clients by Income Segment.

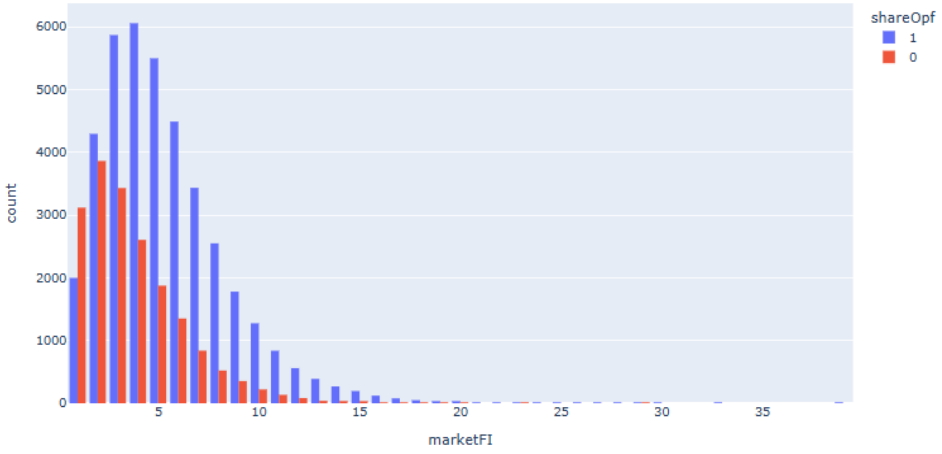


Note: The blue bars represent clients who share their information, while the red bars correspond to clients who choose not to share it.

Source: Prepared by the authors.

Finally, the number of data shares is higher among clients with relationships with four to six financial institutions, compared to other quantities (Figure 7). This range may represent a balance where clients seek to diversify their financial services without overly dispersing their relationships, making it easier to manage their personal finances and optimize the benefits received from each institution.

Figure 7 – Number of Clients by Total Number of Financial Institutions.



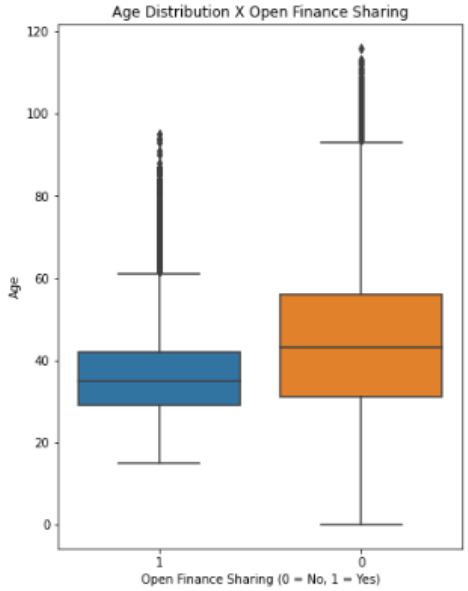
Note: The blue bars represent clients who share their information, while the red bars correspond to clients who choose not to share it.

Source: Prepared by the authors.

Revista Gestão e Desenvolvimento
e-ISSN: 2446-6875
p-ISSN: 1807-5436

The analysis of the boxplot graphs for the attributes mRenda, iIdade, and iTotalProdutos indicates the presence of outliers. For the mRenda variable, three values above 800,000 are observed, significantly discrepant from the others (Figure 8 on the right). In the case of iIdade, there is a higher concentration of elevated values, with many records above 100 years old among clients who did not share their data (Figure 9). As for the iTotalProdutos attribute, the mean is higher, showing greater values for clients who shared their data, along with an isolated value of 50 (Figure 10).

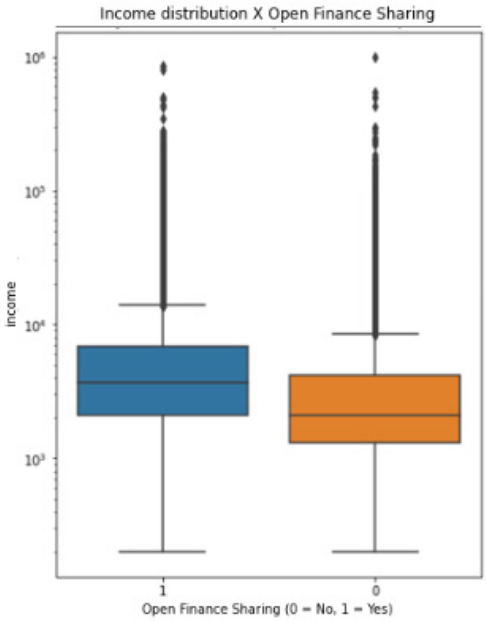
Figure 8 – Boxplot of mRenda.



Source: Prepared by the authors

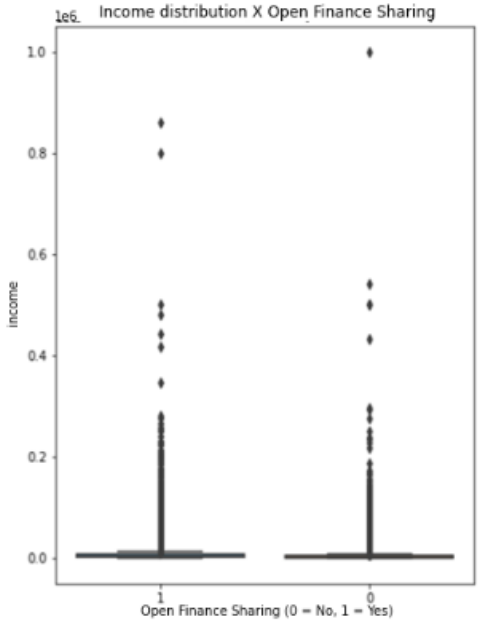
Revista Gestão e Desenvolvimento
e-ISSN: 2446-6875
p-ISSN: 1807-5436

Figure 9 – Boxplot of Age.



Source: Prepared by the authors

Figure 10 – Boxplot of Total Number of Products Contracted.



Source: Prepared by the authors.

Revista Gestão e Desenvolvimento
e-ISSN: 2446-6875
p-ISSN: 1807-5436

Upon completing the exploratory analysis, data preprocessing was performed to prepare the dataset for modeling. The steps included adjusting variables, converting dates into numeric values, transforming categorical variables into dummies, handling missing data, removing duplicates, and organizing the distribution of variables, with a focus on categorizing attributes such as age and region. Details of these steps can be found in Appendix 1, which presents tables with information on missing data, age distribution, and other transformations applied.

3.3 MACHINE LEARNING METHODOLOGY

In this stage, the data were analyzed using machine learning techniques, understood as computational methods capable of learning from accumulated experiences and improving their performance on specific tasks (SILVA; ZHAO, 2016). This field is described by Mitchell (1997) as lying at the intersection of computer science and statistics, with an emphasis on inductive learning, which generalizes patterns from observations and examples. According to Henrique, Sobreiro, and Kimura (2019), its applications include tasks such as classification, regression, and the extraction of association rules.

This study adopts a supervised learning approach, suitable for situations in which a labeled target variable is available - in this case, the binary variable bCompartilha_Opf, which indicates whether or not the customer has agreed to share data via Open Finance. The goal is to predict customers' propensity to join the system based on observable characteristics, using models that operate in three stages: (i) construction, (ii) training, and (iii) validation. It is important to note that the data used for validation were not reused during training, in accordance with best methodological practices.

To assess the performance of different classification algorithms, a horserace was conducted, in which multiple models were tested under the same empirical framework. This strategy ensures comparability of the results by using metrics derived from data not previously exposed to the algorithms. Logistic regression was used as the baseline model, followed by the application of five widely recognized techniques for tabular data: Random Forest, Support Vector Machine (SVM), Decision Tree, Logistic Regression, and XGBoost. Table 2 presents the algorithms evaluated in the study, along with their respective approaches.

Table 2 – Evaluated Algorithms

Algorithm	**Approach**
Logistic **	Functions
Decision Tree	Tree
Random Forest	Tree
SVM	Functions
XGBoost	Meta-model

Note: The Logistic algorithm was used as a baseline due to its widespread use in the financial sector.

Source: Prepared by the authors

e-ISSN: 2446-6875
p-ISSN: 1807-5436

Training time was considered with a focus on the feasibility of model implementation, prioritizing those with faster processing. Slower models were only valued when the others exhibited unsatisfactory performance. The main metrics used were derived from the confusion matrix, which compares predicted and actual values across two classes: P (positive) and N (negative). The relevant intersections are: true positive (P correctly classified), true negative (N correctly classified), false positive (N classified as P), and false negative (P classified as N). Table 3 presents a schematic example. In this study, the positive class represents customers who consented to Open Finance, while the negative class refers to those who did not.

Table 3 – Confusion Matrix

Classes	Predicted Positive Class	Predicted Negative Class
Actual Positive Class	VP	FP
Actual Negative Class	FN	VN

Source: Prepared by the authors.

In the following subsections, the main classification algorithms used in this study are presented and described in greater depth, highlighting their characteristics, theoretical foundations, and practical applications.

3.3.1 Logistic Regression

Logistic regression is a statistical technique used to model the probability of binary events based on independent variables. It is widely applied in binary classification problems (HASTIE; TIBSHIRANI; FRIEDMAN, 2009), using the logistic function to map inputs to probabilities between 0 and 1. Recognized for its robustness and efficiency, even with many attributes (JAMES *et al.,* 2013), it is commonly used in fields such as biomedical analysis, credit prediction, and medical diagnosis. Its simplicity and interpretability make it accessible to users with limited mathematical background. Additionally, it provides insights into the relevance of independent variables through regression coefficients, which indicate the direction and magnitude of their impact on the event's probability (BISHOP, 2006).

3.3.2 Support Vector Machine

SVM is a machine learning technique used for classification and regression, notable for identifying the hyperplane that best separates the classes in the feature space. By maximizing the margin between data points from different classes, it demonstrates robustness against overfitting and effectiveness in high-dimensional data (CORTES; VAPNIK, 1995). For linearly separable data, the optimal decision function

e-ISSN: 2446-6875
p-ISSN: 1807-5436

is the one that maximizes this distance, represented by a maximum-margin hyperplane (CHERKASSKY; MA, 2004). SVMs are versatile, adapting to different kernels to model nonlinear relationships, which makes them suitable for contexts such as pattern recognition, bioinformatics, and finance (SCHÖLKOPF; SMOLA, 2002). Another advantage is their efficiency with moderately to large-sized datasets, as only the support vectors (a subset of the training data) are used in model formulation (CHANG; LIN, 2011).

3.3.3 Random Forest

Random Forest is a non-parametric technique developed by Breiman (2001) as an extension of the CART (Classification and Regression Trees) program, designed to improve predictive performance. The method combines multiple predictive trees (a forest), generated from randomly selected vectors drawn independently and with the same distribution for all trees. Within each tree, subdivisions are made from random subsets of predictor variables, selected based on the total number of available predictors. The final result of the Random Forest is obtained by averaging the outputs of all trees (BREIMAN, 2001).

3.3.4 Decision Tree

Decision trees are a popular technique that repeatedly splits data into more homogeneous subsets based on feature values, selecting at each split the feature that best separates the data according to a specific criterion (QUINLAN, 1986). One of their main advantages is interpretability, as each node represents a decision based on a single feature, making the decision-making process easy to understand and visualize. Additionally, decision trees handle both categorical and numerical data without the need for extensive preprocessing, making them suitable for various data types (BREIMAN et al., 1984). Although simple and interpretable, decision trees may suffer from overfitting when applied to complex datasets. Strategies such as pruning, limiting depth, and ensemble learning (such as Random Forest) help mitigate this issue and improve performance across different datasets (BREIMAN, 2001).

3.3.5 XGBoost

XGBoost is an ensemble method that combines bagging and boosting in its construction. Initially, a tree is sequentially improved through boosting, while new trees are created based on the bagging method, following a logic similar to that of Random Forests. Bagging generates different versions of decision trees from bootstrap samples of the original dataset (HASTIE *et al.*, 2009), combining them to make predictions. To increase diversity, only a portion of the predictor variables is used to build each tree, resulting in the model known as Random Forest, where the final decision is made by aggregating multiple trees (CHEN; GUESTRIN, 2016). In the case of boosting, the process begins with a tree characterized by high bias and

e-ISSN: 2446-6875
p-ISSN: 1807-5436

low variance, fitted to the training data. New trees are built sequentially to improve upon the previous model, reducing overfitting at each iteration (CHEN; GUESTRIN, 2016).

4. RESULTS AND DISCUSSION

Initially, model selection was performed using hold-out validation, in which 70% of the data was used for training and 30% for testing. This approach ensured that the evaluation was conducted on a dataset not used during the training process. As a baseline, the logistic regression model was first applied, followed by the evaluation of other popular algorithms for tabular data, such as Decision Tree, Random Forest, SVM, and XGBoost. Table 4 presents the comparative results among the evaluated models, including the performance metrics of Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Among the tested models, XGBoost stood out as the best classifier, achieving an AUC of 0.90 and an Accuracy of 0.82. These results indicate that the XGBoost model was the most effective in distinguishing between customers who adopted Open Finance and those who did not.

Table 4 - Model Results.

Classificador	TP	FP	TN	FN	Accuracy	Precision	Recall	F1-Score	AUC-ROC
modelXGB	42.11%	10.21%	40.02%	7.65%	0.82	0.80	0.85	0.83	0.90
Random Forest	42.19%	11.06%	39.18%	7.57%	0.81	0.79	0.85	0.82	0.89
Decision Tree	36.56%	12.83%	37.40%	13.21%	0.74	0.74	0.73	0.74	0.73
SVM	25.27%	9.74%	40.49%	24.49%	0.66	0.72	0.51	0.60	0.77
Logistic Regression	31.92%	12.87%	37.36%	17.84%	0.69	0.71	0.64	0.68	0.73

Where: True Positive (TP) refers to the number of actual positives correctly classified by the model; False Positive (FP) is the number of actual negatives incorrectly classified as positives; True Negative (TN) refers to the number of actual negatives correctly classified by the model; False Negative (FN) is the number of actual positives incorrectly classified as negatives. Accuracy is the proportion of correct predictions (positive and negative) relative to the total number of cases analyzed. Precision is the proportion of positive predictions that are actually correct. Recall is the proportion of actual positive cases that were correctly identified. F1-Score is the harmonic mean between Precision and Recall, used to balance these two metrics. AUC-ROC (Area Under the Receiver Operating Characteristic Curve) measures the model's ability to distinguish between positive and negative classes.

Source: Prepared by the authors.

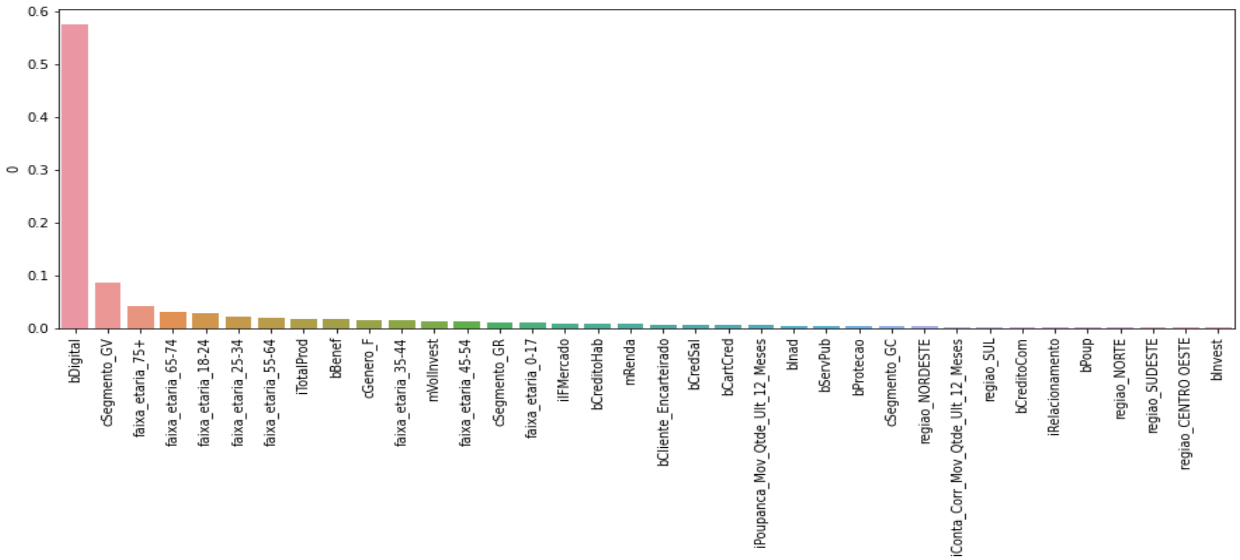
The comparison of model metrics indicates that, while logistic regression achieved an AUC of 0.73, other models such as Random Forest and Decision Tree also performed well, with AUCs of 0.89 and 0.73,

Revista Gestão e Desenvolvimento
e-ISSN: 2446-6875
p-ISSN: 1807-5436

respectively. On the other hand, the SVM model had the worst performance, with an AUC of 0.77, indicating that it was not suitable for this type of binary classification task. These results highlight the robustness of the XGBoost model, which combines strong predictive power with a highly satisfactory AUC. According to Yang and Zhang (2018), advanced techniques such as XGBoost have proven extremely effective in complex forecasting contexts, which was also observed in this study. The model not only demonstrated the best prediction accuracy, but also exhibited excellent capability in differentiating between customers who chose to share their data and those who did not.

In the subsequent stage, analysis of the attributes used by the models revealed relevant factors for Open Finance adoption. The customer's digital profile stood out as the most significant attribute, being strongly associated with the likelihood of consenting to data sharing. This finding is consistent with the conclusions of Carbo-Valverde et al. (2020), who emphasized the importance of digital channels for customer engagement in online financial services. Customers more familiar with technology are more likely to adopt Open Finance. This finding suggests that financial institutions should invest in digital channels and personalized offerings, as digitally savvy customers tend to adopt Open Finance more readily. Figure 11 illustrates this relationship.

Figure 11 – Ranking of Attributes with the Highest Predictive Power.



Source: Prepared by the authors.

Furthermore, the GV segment, composed of clients with an income of up to R$2,000.00 and/or investments up to R$5,000.00, also proved to be relevant. Although this group shows a lower adoption rate of Open Finance, it represents a significant portion of the population with growth potential. This

e-ISSN: 2446-6875
p-ISSN: 1807-5436

aligns with Dawood et al. (2019), who emphasized the importance of personalized strategies for specific customer segments, such as low-income individuals. Financial inclusion strategies could be implemented to encourage the use of financial products among these clients over time.

Another relevant group consists of clients over the age of 64, who, although less digitally engaged, tend to have a more stable income base, such as retirement benefits. For this group, tailored strategies involving financial products focused on healthcare plans and specific needs of elderly individuals could enhance the adoption of Open Finance. The total number of contracted products stands out as an indicator of more engaged clients who are familiar with financial services, making them more likely to recognize the benefits of Open Finance.

Additionally, some low-contribution attributes suggest that the analysis can be simplified by using fewer variables without compromising the model's efficiency.

In summary, the results of this study corroborate the findings of several authors who highlight the effectiveness of algorithms such as XGBoost and the importance of a segmented approach to predicting Open Finance adoption. They also emphasize the need for personalized strategies targeting specific groups, such as low-income and elderly clients. The analysis of digital profiles is also confirmed as essential, especially for enhancing the acceptance of financial technologies and increasing financial inclusion.

5. CONCLUSION

The main objective of this study was to analyze the factors influencing customer adoption of Open Finance, based on real data from a financial institution. To achieve this, a quantitative approach was employed using machine learning techniques. The methodology involved splitting the data into training and testing sets (hold-out validation), with the application of different classification algorithms, such as Logistic Regression, Decision Tree, Random Forest, SVM, and XGBoost. Model performance was evaluated using metrics including Accuracy, Precision, Recall, F1-Score, and AUC-ROC.

The results showed that the XGBoost model was the most effective, achieving an AUC of 0.90 and an Accuracy of 0.82, outperforming the other classifiers. These findings can inform more effective marketing strategies by highlighting profiles with a higher propensity for adoption, such as customers with high digital literacy, individuals with lower income and/or smaller investment volumes, and consumers aged over 64. By better understanding these segments, financial institutions can design more targeted campaigns, personalize communication strategies, and enhance digital journeys, thereby fostering broader inclusion and sustainable engagement with Open Finance.

e-ISSN: 2446-6875
p-ISSN: 1807-5436

These findings underscore the importance of segmented strategies for different customer profiles, especially those with lower income or older age, who tend to have lower digital familiarity but represent segments with significant potential. The contributions of this study are relevant for financial institutions and policymakers by pointing to pathways for personalized communication strategies and digital inclusion aimed at expanding Open Finance adoption. However, limitations such as the reliance on internal data and the scope restricted to a single financial institution warrant caution in generalizing the results. Future research may expand the data base, incorporate behavioral variables, and explore longitudinal approaches to track changes in adoption patterns over time.

REFERENCES

ADKE, V.; BAKHSHI, P.; ASKARI, M. Impact of disruptive technologies on customer experience management in ASEAN: A review. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTING – ICOCO, 2022, Kota Kinabalu, Malaysia. **Anais. Kota Kinabalu: IEEE**, 2022. p. 364–368. Disponível em: https://doi.org/10.1109/ICOCO56118.2022.10031882. Acesso em: 11 jun. 2024.

BANCO CENTRAL DO BRASIL. **Open Finance**. 2023. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/openfinance. Acesso em: 12 abr. 2024.

BARTELS, C. Cluster analysis for customer segmentation with open banking data. In: ACM INTERNATIONAL CONFERENCE, 2022. **Anais do ACM**, 2022. p. 87–94. Disponível em: https://doi.org/10.1145/3523181.3523194. Acesso em: 10 maio. 2025.

BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer Science & Business Media, 2006. 738 p.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018: Lei Geral de Proteção de Dados Pessoais (LGPD)**. Diário Oficial da União, Brasília, DF, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 20 abr. 2024.

BREIMAN, L. **Random forests**. Machine Learning, Dordrecht, v. 45, p. 5–32, 2001. DOI: m: https://doi.org/10.1023/A:1010933404324.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. **Classification and regression trees.** Boca Raton: CRC Press, 1984. 368 p.

e-ISSN: 2446-6875
p-ISSN: 1807-5436

CARAFFINI, J. P. T. S.; SOUZA, R. B. de L.; BEHR, A.; VENTURINI, L. D. B. Transformação digital e desempenho no setor bancário: uma abordagem com análise envoltória de dados. **Revista Gestão e Desenvolvimento**, v. 20, n. 2, p. 54–79, 2023. DOI: https://doi.org/10.25112/rgd.v20i2.3262.

CHANG, C. C.; LIN, C. J. LIBSVM: A library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, v. 2, n. 3, p. 1–27, 2011. DOI: https://doi.org/10.1145/1961189.1961199.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. **Anais. ACM**, New York: Association for Computing Machinery, 2016. p. 785–794. DOI: https://doi.org/10.1145/2939672.2939785.

CHERKASSKY, V.; MA, Y. Practical selection of SVM parameters and noise estimation for SVM regression. **Neural Networks**, Amsterdam, v. 17, n. 1, p. 113–126, 2004. DOI: https://doi.org/10.1016/S0893-6080(03)00169-2.

CARBO-VALVERDE, S.; CUADROS-SOLAS, P.; RODRÍGUEZ-FERNÁNDEZ, F. A machine learning approach to the digitalization of bank customers: Evidence from random and causal forests. **PLoS ONE**, v. 15, n. 10, e0240362, 2020. DOI: https://doi.org/10.1371/journal.pone.0240362.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995. DOI: https://doi.org/10.1007/BF00994018.

DAWOOD, E. A. E.; ELFAKHRANY, E.; MAGHRABY, F. A. Improve profiling bank customer's behavior using machine learning. **IEEE Access**, v. 7, p. 109320–109327, 2019. DOI: https://doi.org/10.1109/ACCESS.2019.2934644.

GRASSI, L. In a world of Open Finance, are customers willing to share data? An analysis of the data-driven insurance business. **Eurasian Business Review**, v. 14, n. 3, p. 727–753, 2024. DOI: https://doi.org/10.1007/s40821-024-00263-w.

GRASSI, L.; FIGINI, N.; FEDELI, L. How does a data strategy enable customer value? The case of FinTechs and traditional banks under the open finance framework. Financial Innovation, v. 8, n. 75, p. 1–34, 2022. DOI: https://doi.org/10.1186/s40854-022-00378-x.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: Data mining, inference, and prediction. 2. ed. New York: Springer, 2009. 745 p.

e-ISSN: 2446-6875
p-ISSN: 1807-5436

HENRIQUE, B. M.; SOBREIRO, V. A.; KIMURA, H. Literature review: Machine learning techniques applied to financial market prediction. **Expert Systems with Applications**, v. 124, p. 226–251, 2019. DOI: https://doi.org/10.1016/j.eswa.2019.01.012.

HJELKREM, L. O.; LANGE, P. E. D. Explaining deep learning models for credit scoring with SHAP: A case study using open banking data. **Journal of Risk and Financial Management**, v. 16, n. 4, p. 1–19, 2023. DOI: https://doi.org/10.3390/jrfm16040221.

IMAN, N.; NUGROHO, S. S.; JUNARSIN, E.; PELAWI, R. Y. Is technology truly improving the customer experience? Analysing the intention to use open banking in Indonesia. **International Journal of Bank Marketing**, v. 41, n. 7, p. 1521–1549, 2023. DOI: https://doi.org/10.1108/IJBM-09-2022-0427.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. New York: Springer, 2013. 426 p.

MANZATO, A. J.; SANTOS, A. B. **A elaboração de questionários na pesquisa quantitativa**. Florianópolis: UFSC, 2012.

MITCHELL, T. M. **Machine learning**. New York: McGraw-Hill, 1997. 414 p.

MISHRA, A. K.; TYAGI, A. K.; AROWOLO, M. O. Future trends and opportunities in machine learning and artificial intelligence for banking and finance. In: IRFAN, M. et al. (Org.). **Applications of blockchain technology and artificial intelligence**: Financial mathematics and fintech. Cham: Springer, 2024. p. 211–238. DOI: https://doi.org/10.1007/978-3-031-47324-1_13.

NILOY, N. H.; NAVID, M. A. I. Naïve Bayesian classifier and classification trees for the predictive accuracy of probability of default credit card clients. **American Journal of Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 1, 2018. DOI: https://doi.org/10.11648/j.ajdmkd.20180301.11.

NUNES, G. C.; NASCIMENTO, M. C. D.; ALENCAR, M. A. C. de. Pesquisa científica: conceitos básicos. **Id On Line Revista de Psicologia**, v. 10, n. 1, p. 144, 28 fev. 2016.

OPEN FINANCE BRASIL. **Relatório anual 2023 do Open Finance Brasil**. 2024. Disponível em: https://ob-wp-media-files.s3.amazonaws.com/wp-content/uploads/2024/05/07141329/2023_Relatorio-Anual-OFB.pdf. Acesso em: 12 abr. 2024.

PALANIAPPAN, S.; MUSTAPHA, A.; FOOZY, C. F. M.; ATAN, R. Customer profiling using classification approach for bank telemarketing. International **Journal of Informatics Visualization**, v. 1, n. 2, p. 214–217, 2017. DOI: https://doi.org/10.30630/joiv.1.4-2.68.

e-ISSN: 2446-6875
p-ISSN: 1807-5436

PATIL, P. S.; DHARWADKAR, N. V. Analysis of banking data using machine learning. In: INTERNATIONAL CONFERENCE ON IoT IN SOCIAL, MOBILE, ANALYTICS AND CLOUD (I-SMAC), 2017. **Anais do Piscataway**: IEEE, 2017. p. 876–881. DOI: https://doi.org/10.1109/I-SMAC.2017.8058305.

PROETTI, S. As pesquisas qualitativa e quantitativa como métodos de investigação científica: um estudo comparativo e objetivo. **Revista Lumen**, v. 2, n. 4, p. 1-23, 1 jun. 2018.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, n. 1, p. 81–106, 1986. DOI: https://doi.org/10.1007/BF00116251.

REDDY, S. M.; MIRIYALA, S. Security and privacy preserving deep learning. **arXiv**, Ithaca, v. preprint arXiv:2006.12698, 2020. DOI: https://doi.org/10.48550/arXiv.2006.12698.

SILVA, T. C.; ZHAO, L. **Machine learning in complex networks**. Cham: Springer, 2016.

SCHÖLKOPF, B.; SMOLA, A. J. **Learning with kernels**: Support vector machines, regularization, optimization, and beyond. Cambridge: MIT Press, 2002. 644 p.

TARGHER, R. M. **Open finance no Brasil**: Levantamento de desafios e lições aprendidas. 2023. 135 f. Dissertação (Mestrado em Administração de Empresas) – Fundação Getúlio Vargas, Escola de Administração de Empresas de São Paulo, São Paulo, SP, 2023.

WEN, Z.; YAN, J.; ZHOU, L.; LIU, Y.; ZHU, K.; GUO, A.; ZHANG, F. Customer churn warning with machine learning. In: KRÖMER, P. et al. (Org.). **Proceedings of the Fifth Euro-China Conference on Intelligent Data Analysis and Applications**. ECC 2018. Cham: Springer, 2019. p. 629–638. DOI: https://doi.org/10.1007/978-3-030-03766-6_39.

XIAO, J.; XIAO, Y.; HUANG, A.; LIU, D.; WANG, S. Feature-selection-based dynamic transfer ensemble model for customer churn prediction. **Knowledge and Information Systems**, v. 43, p. 29–51, 2015.

YANG, S.; ZHANG, H. Comparison of several data mining methods in credit card default prediction. **Intelligent Information Management**, v. 10, n. 5, p. 115, 2018. DOI: https://doi.org/10.4236/iim.2018.105010.