Data Science and Management 7 (2024) 7–16

Contents lists available at ScienceDirect

Data Science and Management

journal homepage: www.keaipublishing.com/en/journals/data-science-and-management

Research article

Investigating customer churn in banking: a machine learning approach and visualization app for data science and management

Pahul Preet Singh [a], Fahim Islam Anik [b], Rahul Senapati [a], Arnav Sinha [a], Nazmus Sakib [c,*], Eklas Hossain [d]

[a] Institute for Artificial Intelligence and Data Science, University at Buffalo, The State University of New York, Amherst, 14068, United States
[b] Department of Mechanical Engineering, Khulna University of Engineering & Technology, Khulna, 9203, Bangladesh
[c] Department of Information Technology, Kennesaw State University, Marietta, 30067, United States
[d] Department of Electrical and Computer Engineering, Boise State University, Boise, 83725, United States

ARTICLE INFO

Keywords:
Bank customer attrition
Churn prediction
Machine learning
XGboost
Random forest

ABSTRACT

Customer attrition in the banking industry occurs when consumers quit using the goods and services offered by the bank for some time and, after that, end their connection with the bank. Therefore, customer retention is essential in today's extremely competitive banking market. Additionally, having a solid customer base helps attract new consumers by fostering confidence and a referral from a current clientele. These factors make reducing client attrition a crucial step that banks must pursue. In our research, we aim to examine bank data and forecast which users will most likely discontinue using the bank's services and become paying customers. We use various machine learning algorithms to analyze the data and show comparative analysis on different evaluation metrics. In addition, we developed a Data Visualization RShiny app for data science and management regarding customer churn analysis. Analyzing this data will help the bank indicate the trend and then try to retain customers on the verge of attrition.

1. Introduction

Customer attrition, also known as customer churn, is the phenomenon where customers terminate their relationship with a business or organization. In the context of banking, customer attrition occurs when customers close their accounts or discontinue utilizing services of a particular bank. Effectively understanding and managing customer attrition are crucial for banks to maintain financial stability and safeguard their reputation. The financial impact of customer attrition on banks can be significant, resulting in potential revenue loss across various banking services. Consequently, establishing and nurturing long-term customer relationships is highly valuable for banks. By gaining insight into attrition patterns, banks can identify customers at risk of leaving and implement strategies to retain them. This approach enhances overall customer lifetime value and bolsters bank profitability.

Moreover, customer attrition has repercussions on a bank's reputation and brand perception. High churn rates often indicate underlying issues, such as poor customer experience, inefficient processes, or a lack of competitive products and features. Therefore, understanding and managing customer attrition are crucial for banks to address these challenges and enhance their overall customer experience. Within the competitive banking industry, monitoring and managing customer attrition can provide banks valuable insights into customer preferences, needs, and pain points. This knowledge can help banks develop targeted strategies to differentiate themselves from their competitors and enhance customer retention.

The Data Visualization RShiny app, a web application framework developed using the R programming language, plays an important role in analyzing customer churn. It empowers users to interact with churn-related data through interactive visualizations and dashboards, fostering a deeper understanding of the data and enabling the identification of patterns and trends related to customer attrition. The app offers real-time monitoring capabilities by connecting live data sources, allowing banks to track attrition rates, customer behavior, and other relevant indicators in

Peer review under responsibility of Xi'an Jiaotong University.
* Corresponding author.
E-mail address: nsakib1@kennesaw.edu (N. Sakib).

https://doi.org/10.1016/j.dsm.2023.09.002
Received 16 January 2023; Received in revised form 12 September 2023; Accepted 18 September 2023
Available online 28 September 2023

real time. This feature enables prompt action and decision making. Additionally, the app supports comparative analysis, facilitating the comparison of different customer segments based on demographics, product usage, or behavior. This functionality provides valuable insights into which segments are more susceptible to churn and guides the development of targeted retention strategies. Predictive modeling is another crucial aspect of an application. It integrates machine-learning (ML) algorithms or statistical models to forecast customer churn. By generating churn predictions and visualizing the probability of churn for individual customers, the app assists banks in identifying high-risk customers and taking proactive measures to prevent churn. Furthermore, the app facilitates reporting and communication by allowing the creation of customized reports and presentations. This feature enables stakeholders to easily share churn insights and recommendations with management, marketing teams, or other relevant parties. The workflow of the Data Visualization RShiny app for data science and management involves the functions and processes listed in Table 1.

The success of any business model relies on having a large customer base, which entails achieving two primary objectives: acquiring new customers and retain existing ones. Winning new customers involves designing products and advertising them to the appropriate demographics. The second challenge, retaining customers, is essential for any business model to thrive, as lost customers are highly unlikely to return. Our problem statement primarily addresses the concern about maintaining customers and predicting their patterns, which eventually contributes to solving the customer attrition problem (De Caigny et al., 2020; Lee and Shin, 2020; Shirazi and Mohammadi, 2019). To address customer attrition, previous studies have discussed customer relationship management (CRM) systems and three approaches for retention (De Caigny et al., 2020; De Lima Lemos et al., 2022; Rahman and Kumar, 2020). These articles include post-purchase evaluations, periodic satisfaction surveys, and continuous satisfaction tracking. They provide an excellent foundation for exploring the reasons for customer dissatisfaction.

This study aims to extend the scope of the aforementioned CRM systems, with a primary focus on identifying and predicting the likelihood of customer attrition (Amuda and Adeyemo, 2019; Domingos et al., 2021; Ho et al., 2019). The findings of this study can be applied in real-world scenarios to assist banks in determining customer defection and taking preventive measures to retain such customers (Geiler et al., 2022; Machado and Karray, 2022). In a related study (Lemmens and Gupta, 2020), the authors discussed managing churn to maximize profits and investigated the profit-loss ratio concerning when customers stop using products. Apart from practical applications for predicting bank customer attrition, this study helps establishing a starting point for conducting further research in this field (Al-Mashraie et al., 2020; Baghla and Gupta, 2022; Schaeffer and Sanchez, 2020).

Table 1
Workflow of the data visualization RShiny app for data science and management.

Steps	Functionalities
1. Data input	Allows users to input relevant data sources. Supports various formats (CSV, Excel, and databases).
2. Data preprocessing	Cleans and prepares the data for analysis. Handles missing values, normalization, and other features.
3. Interactive filtering	Enables users to filter and select variables. Focuses on specific subsets of the data.
4. Visualizations	Generates a variety of visualizations. Includes scatter plots, bar charts, and line graphs.
5. Comparative analysis	Presents insights on customer churn patterns. Allows comparison of metrics and customer segments.
6. Predictive modeling	Analyzes churn rates, customer behavior, and other parameters. Integrates ML for churn prediction.
7. Real-time monitoring	Visualizes probability of churn for individual customers. Connects to live data sources.
8. Reporting and exporting	Updates visualizations and metrics in real-time. Generates customized reports.
9. Decision support	Exports visualizations and analysis results. Provides interactive dashboards and visuals.
10. Outputs	Supports data-driven decision making. Provides interactive visualizations and analysis results. Allows for comparative analysis outcomes. Creates predictive churn models and provides real-time monitoring updates. Outputs customized reports and exported data.

Successful financial firms must provide useful customer assistance. By examining how consumers use goods, employing ML in the financial sector enables businesses to provide customized offers and services that cater to customer demands. The primary role of ML in customer retention is to monitor and forecast customer turnover by tracking behavioral changes. Research has revealed that acquiring new customers costs significantly more than retaining existing ones. ML enables firms to recognize clients who are on the verge of leaving and take prompt action to retain them. Additionally, it can help boost client trust and maintain/extend customer engagement, whether it is the customer who has forgotten about the service or the one who has had a bad experience. Having a model that provides insights to banks about customers likely to leave will assist them in taking the necessary steps and specifically targeting these customers rather than expending resources tracking all customers (De Lima Lemos et al., 2022; Guliyev and Tatoğlu, 2021; He et al., 2014; Karvana et al., 2019; Patil and Dharwadkar, 2017). Practically, employing such a model to predict the likelihood of customer attrition allows banks to focus on this group of customers (Dias et al., 2020; Vo et al., 2021). However, most studies do not adequately compare various ML techniques to help banks make informed decisions based on a comparison of the results and domain knowledge. To bring this idea to fruition, the suitable adaptation of a particular application for this purpose and its representation have not been well demonstrated in the literature.

This study holds potential implications for stakeholders in the banking industry. Stronger customer retention methods will lead to personalized offers, improved customer service, and customized banking solutions, all of which will enhance customer experiences. By deploying resources and training programs targeted at enhancing customer service, employees may benefit from better work environments and greater job satisfaction. As customer churn decreases, and customer lifetime value and profitability increase, shareholders should anticipate improved financial performance. Additionally, implementing research findings can boost a bank's image and brand impression, attract new clients, and encourage long-term company growth. Furthermore, the study emphasizes the significance of data-driven decision making, enabling stakeholders to make informed decisions based on churn analysis insights and encouraging an industry-wide culture of evidence-based decision making. These outcomes will support the overall expansion and achievements of banking institutions.

This study makes several important contributions to the body of knowledge regarding customer churn analysis and ML in the banking sector. The first part of the article presents a comprehensive preprocessing method that guarantees data correctness and consistency, addressing the critical issue of data preparation unique to customer churn analysis in banking. Second, the study thoroughly examines different ML methods and evaluates their effectiveness in anticipating customer attrition. This comparative analysis offers valuable insights into the performance of various churn prediction algorithms in the banking industry. Furthermore, by offering a user-friendly tool for displaying churn-related insights, the development of the Data Visualization RShiny app enhances the practical application of churn analysis. Finally, this study yields useful implications for banks, emphasizing the importance of understanding customer attrition and providing practical recommendations to improve client retention.

Our unique contributions of this study are summarized as follows:

(1) It presents a comprehensive preprocessing approach that effectively unifies diverse data in a consistent format.

(2) It conducts a thorough investigation of different ML algorithms for the specific purpose of predicting bank customer attrition.
(3) An application is developed to provide stakeholders with extensive visualizations, empowering them to make informed decisions.

The remainder of the article includes four sections: materials and exploratory data analysis; theory and approach; results and discussion; and conclusions. Fig. 1 summarizes the key contributions of the study.

2. Materials and exploratory data analysis

This section provides an overview of the dataset used, as well as the different analyses and summaries, to better understand the different parameters contributing to the prediction modeling. Table 2 lists the notations and descriptions used in this study.

2.1. Description of dataset

The dataset (https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction/data) used in this study comprises 10,000 rows of customers. Typically, each bank has an elaborate process, with a know your customer (KYC) assessment conducted for every new customer. Several critical processes or steps are involved during the onboarding, which ensures that the bank data obtained can be considered complete and reliable. Consequently, all the necessary customer information acquired is accessible and legitimate. Each customer is differentiated by a unique customer ID and associated surname. The dataset includes customer details, such as credit scores, age, tenure, balance, number of products, and estimated salary. The data include Boolean measurements, such as 0 or 1, and other sections, with two or more classes. These can be classified as follows: county, gender, has a credit card, being an active member, and churned. The final column, "exited" determines the current state of the customer, and 1 implies that customer attrition occurred. We aimed to feed the bank data into a model and determine the outcome, the exit column, if it becomes 1. The captured data varied according to the customer's location, economic status, and gender. The number of products a user uses is proportional to how loyal and profitable the customer is to the bank. A mix of such wide-ranging data helps to draw factual and statistically accurate inferences. Categorical data were converted into a numerical form to prevent information loss during modeling. "RowNumbers", "CustomerID", and "Surname" were removed from our dataset, as they are not pertinent to our analysis. Table 3 summarizes the data.

Table 2
Notations and descriptions.

Notation	Description	Notation	Description
CRM	Customer relationship management	ROC	Receiver operating characteristic curve
KYC	Know your customer	TP	True positives
SVM	Support vector machine	FP	False positives
XGBoost	eXtreme gradient boosting	TN	True negatives
GLM	Generalized linear model	FN	False negatives
CV	Cross-validation	Accuracy	(TP + TN)/(TP + TN + FP + FN)
AUC	Area under the ROC curve	Sensitivity	TP/(TP + FN)
Precision	True positive/(true positive + false positive)	Specificity	TN/(TN + FP)
Recall	True positive/(true positive + false negative)	SMOTE	Synthetic minority oversampling technique

2.2. Dataset analysis

We preprocessed the dataset to effectively unify and visualize the diverse input data parameters in a consistent format.

(a) Customer church distribution. The pie chart in Fig. 2 depicts the distribution of our dependent variable (churned) in the dataset. 80% of the records are for "not churned" customers, and 20% are "churned". Thus, every 5th customer was churned, and our dataset is highly imbalanced.

Table 3
Variables, null count, and unique count of the dataset.

Variables	Null count	Unique count
RowNumber	0	10,000
CustomerID	0	10,000
Surname	0	2,932
CreditScore	0	460
Geography	0	3
Gender	0	2
Age	0	70
Tenure	0	11
Balance	0	6,382
NumOfProducts	0	4
HasCrCard	0	2
IsActiveMember	0	2
EstimatedSalary	0	9,999
Exited	0	2

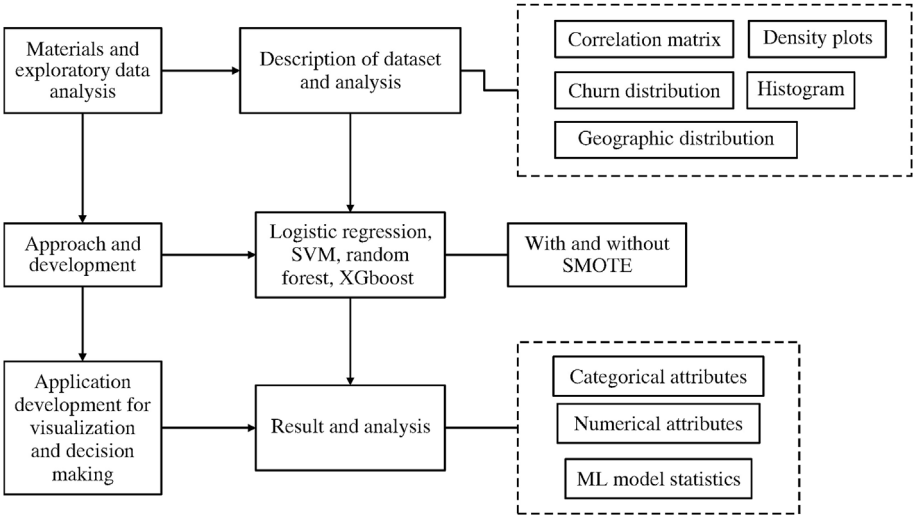

Fig. 1. Visual summary of the scientific contribution of the study.

(b) Gender, active members, credit cards, and country-based analysis. Fig. 3 presents valuable insights regarding gender, active members, credit cards, and country-based analyses. We observed that out of 4,543 female customers, 1,139 were churned (~25%), whereas for the male population, 898 were churned out of 5,457 in total (~16%). In addition, 1/5th of all customers were churned irrespective of whether they owned credit cards. The likelihood of attrition of inactive customers is almost double that of active customers (27% inactive vs. 14% active). Germany has the highest customer churn rate (40%), followed by France (16%) and Spain (16%).

(c) Balance, owned product quantity, credit score, and tenure-based analysis. Fig. 4 illustrates a density plot to observe the balance, owned product quantity, credit score, and tenure-based analysis. We found that customers who maintain a balance of more than 85,000 are more likely to churn. Premium accounts and higher savings interest at other banks could be the root causes of this. Customers owning these two products at the bank were significantly less likely to leave. Numerical factors, such as credit scores and tenure, do not impact the customer attrition rate. However, customers with a poor credit history (that is, a credit score below 400) will undoubtedly leave the bank, which is visible in the scatterplot below (Fig. 5). In this figure, there are only blue dots (churned customers) below the credit score of 400, which could be due to the weak economic status of the customer.

It is evident from the box plot presented in Fig. 6 that older customers (above the age of 40) are more likely to churn than younger customers. The bee-swarm plot (Fig. 5) supports this assumption, as more pink dots (churned customers) are present in the age range of 40–70. This could be due to better plans offered for seniors at other banks.

(d) Correlation matrix analysis of the dataset. It appears from Fig. 7 that no pair of variables is strongly correlated, thus helping to satisfy the fundamental assumption (absence of multicollinearity) of modeling (Alin, 2010; Kim, 2019; Mansfield and Helms, 2012). The only notable correlation observed was between the number of products and balance.

Now that we have a comprehensive understanding of the different significant parameters of the dataset, the next section describes the theoretical methodological approach for predicting bank customer attrition.

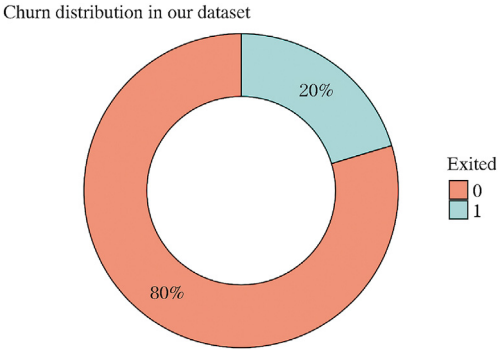

Fig. 2. Customer churn distribution.

3. Theory, approach, and development

This section introduces different ML techniques and applies them to the dataset. We focused on core ML approaches, including logistic regression, support vector machine (SVM), random forest, and eXtreme Gradient Boosting (XGBoost). In this section, we then introduce a visualization tool for stakeholders tailored explicitly to summarize different data in a single mode.

3.1. ML techniques

We briefly discuss the theoretical underpinnings of the ML techniques used in this research. First, logistic regression is a fundamental classifi-

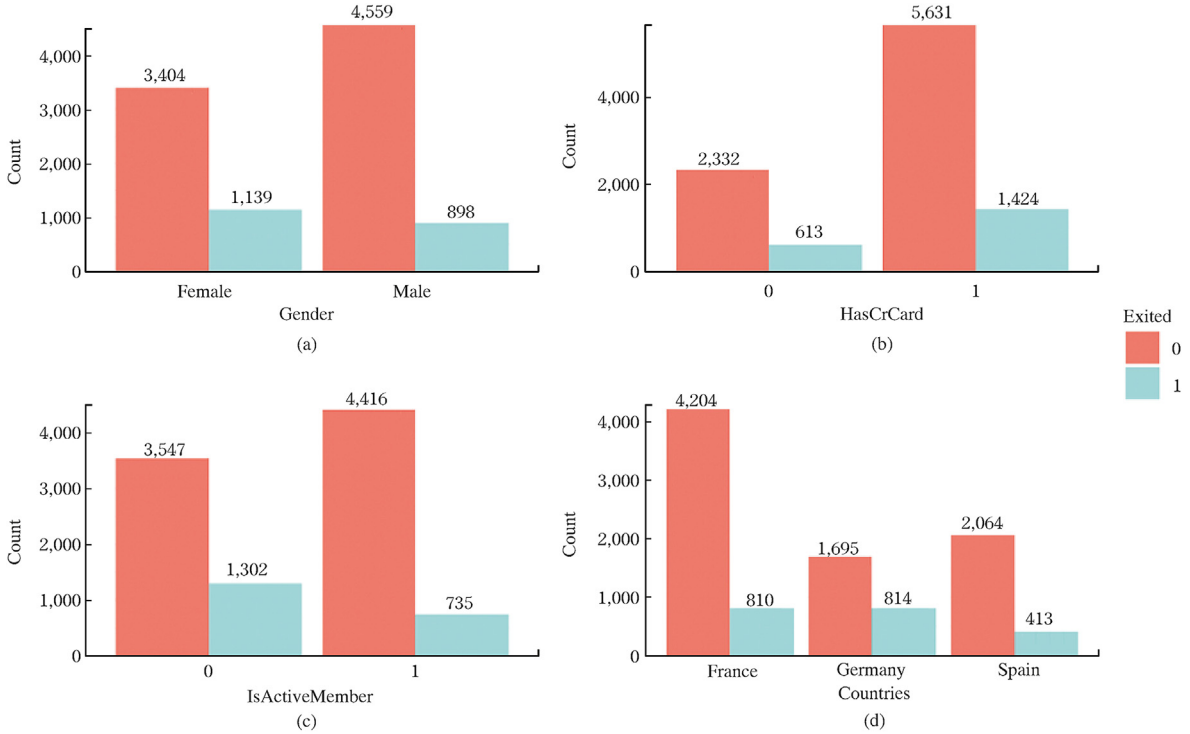

Fig. 3. Histograms of the dataset. (a) Gender vs. churn; (b) customers having credit card vs. churn; (c) active member vs. churn; (d) country vs. churn.

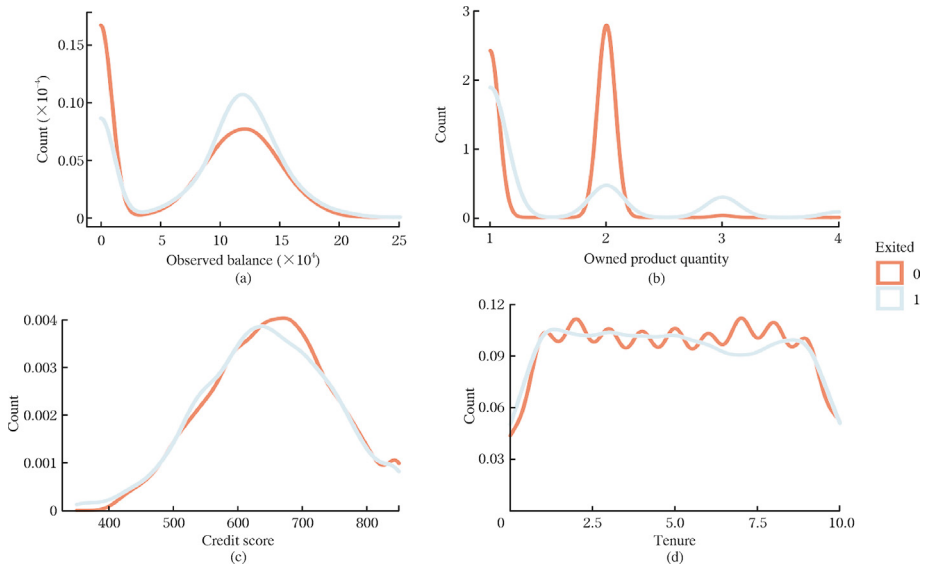

Fig. 4. Density plots of (a) observed balance, (b) owned product quantity, (c) credit score, and (d) tenure.

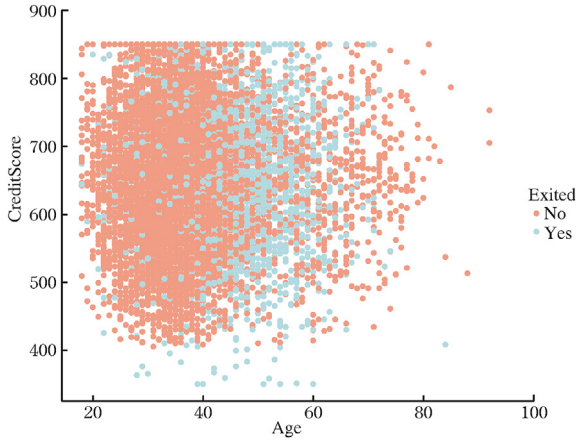

Fig. 5. Distribution of customers based on credit score and age.

cation technique that is crucial for predicting customer churn. The logistic curve equation is as follows:

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \tag{1}$$

The rolling mean of the *DV*, $P(\bar{Y})$, and independent variable, *X*, are connected by the logistic curve. In Eq. (1), *e* is the base of the natural logarithm (approximately 2.718), *a* and *b* are the model parameters, and *P* is the probability. When *X* is zero, the value of *a* produces *P*, and the value of *b* regulates how rapidly the probability changes when *X* is changed by a single unit (we can have standardized and unstandardized *b* weights in logistic regression, as in ordinary linear regression). *B* is not as easily interpreted in this model as in a typical linear regression, because the relationship between *X* and *P* is not linear.

In the case of random forest, a subset of the original data must first be made with row sampling and feature sampling to create a training dataset. Subsequently, an individual decision tree is created for each subset. Finally, considering the output of each decision tree, the majority vote is counted, as shown in Fig. 8. Random forest allows individuals to belong to a class for classification. The advantage of using random forest for classification is its high accuracy. However, ensuring the robustness of the model (or its generalizability) in predicting unknown data remains challenging.

An SVM provides the distance to the border, and several steps must be undertaken to convert it to probability (Cervantes et al., 2020). When applied to specific issues, one technique may outperform another (Anton et al., 2019; Sothe et al., 2020).

The gradient-boosted tree algorithm, which is a supervised learning approach, is based on function approximation by maximizing certain loss functions and employing a number of regularization approaches. XGBoost is one of the best-known and most practical implementations of this algorithm. To obtain a predictive analysis, the following objective functions (loss function and regularization) must be minimized at iteration *t*, as shown in Eq. (2) (Chen and Guestrin, 2016):

$$\zeta^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(X_i)\right) + \Omega(f_t) \tag{2}$$

Here, $\zeta^{(t)}$ represents the t-*th* iteration of the ensemble model. A summation over all the data points in the dataset, where *n* is the number of data points, is on the right hand side of the equation.

$(y_i, \hat{y}_i^{(t-1)} + f_t(X_i))$ is the loss function, which measures the difference between the true target value and the predicted target value. $\Omega(f_t)$, representing the regularization or complexity penalty, is applied to the t-*th* model prevent overfitting.

3.2. Approach toward churn prediction

Four pipelines were constructed for each model: logistic regression, random forest, SVM, and XGBoost. The models were then fitted to a training dataset. We orchestrated model training by constructing a two-step workflow pipeline.

- Normalize the features to bring them on the same scale;
- Instantiate the model to fit on.

We then configured the grid search parameters for each model. We set up four grid-search CV functions that used the pipeline and parameters as inputs. Finally, all grids were fitted to the training dataset. For the classification, we initially used a logistic regression model (Pearce and Ferrier, 2000). In order to compute the significant features, we used the "glm" function (Manning, 2007). We achieved an accuracy of 85%, a sensitivity of 38%, a specificity of 97%, and 0.83 as AUC when we applied the model to our validation data (Statinfer, 2017), as shown in Fig. 9. The large discrepancy between sensitivity and specificity was most

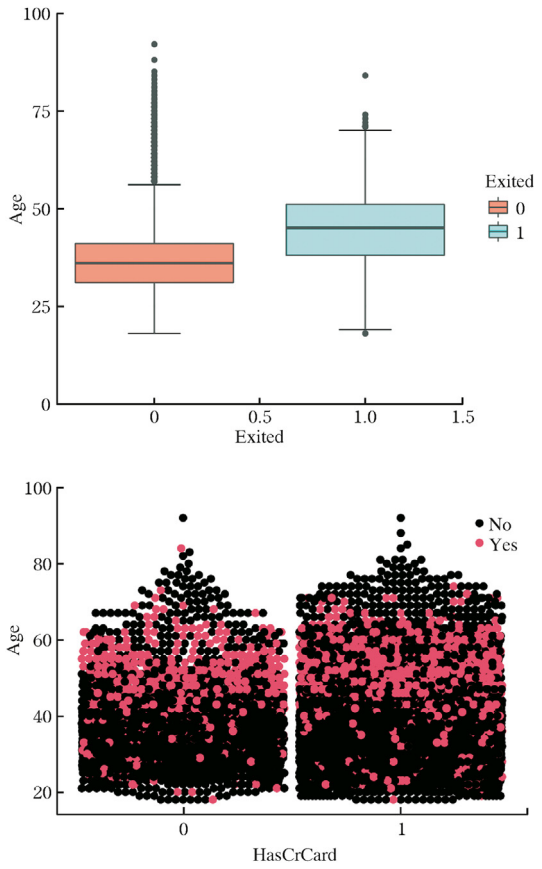

Fig. 6. Distribution of age of customers.

likely due to a large imbalance in the data, which caused bias in the model (He and Garcia, 2009; Mazumder, 2021). As the data were not balanced (Kaur et al., 2019; Sun et al., 2009), we used the synthetic minority oversampling technique (SMOTE) (Torgo et al., 2013).

This process was performed prior to and after addressing the data imbalance. The imbalance was treated using SMOTE (Chawla et al., 2002; Fernández et al., 2018; Han et al., 2005) which is an improved method for managing imbalanced data in classification problems that performs data augmentation by creating synthetic data points based on original data points. This choice was made because methods such as undersampling could cause a potential loss of information. Customer churn was predicted by fitting and evaluating multiple models after addressing the imbalance using SMOTE. This answered two of the research questions posed, namely, the prediction of customer churn for an imbalanced dataset and the examination of multiple models to obtain the most reliable one. Some of the rows, such as "Row Number" and "Surname" were irrelevant; as a result, we removed them. Along with these changes, three new variables were created, namely "TenureByAge", "BalanceSalaryRatio," and "CreditScoreGivenAge." This was done to enhance the performance of the ML models. Removing irrelevant data and feature engineering also addressed a research question, namely, selecting pertinent attributes for evaluating the model and removing outliers. After a thorough examination, the best model was chosen based on a particular combination of metrics that contributed to the implementation of an adaptable model capable of addressing the dynamic nature of data patterns. In addition, because the data do not contain any information that could be publicly traced back to a person, privacy has been maintained.

3.3. Application development for visualization and decision making

Since our research aimed to provide a practical implementation, we

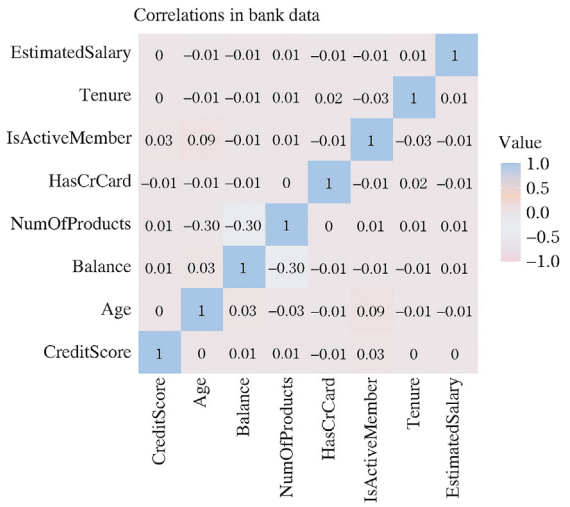

Fig. 7. Correlation matrix of the dataset.

deployed an application using Plotly (Van Der Donckt et al., 2022) and Python to demonstrate our model calculations and the possible outcomes of customer analysis. The application interface functions interactively comprise three primary tabs: data analysis, model analysis, and prediction. The data analysis tab shows information on both categorical and numerical attributes. The user can then select attributes. The categorical representations, as shown in Fig. 10, include a donut chart indicating the percentage of presence/lack of presence of a particular attribute and a bar chart indicating how the presence/absence of a particular distribution varies with churn. The numerical representation in Fig. 11 comprises a density plot, a scatter plot with age, and a box plot of the selected attributes. An image of the data analysis tab for "Number of Products" (categorical) and "CreditScore" (numerical) is given below.

The model analysis tab in Fig. 12 shows all evaluation metrics, including accuracy, sensitivity, specificity, AUC, F1 score, test-train data split percentage, and feature importance for all models. An image of the model analysis tab for prediction is provided below. Finally, the prediction tab shown in Fig. 13 lets the user set the independent variables to the values of their choice to obtain the dependent variable (which, in this case, is churn). Using this feature, we can determine the churns for all models mentioned above. An image of this tab for arbitrarily chosen data is shown below as a demonstration. These features help the application user draw inferences and make knowledgeable decisions.

4. Results and discussion

In this section, we discuss the findings of the aforementioned ML models. The evaluation metrics used include the accuracy, sensitivity, specificity, AUC, and F1 score (Sokolova et al., 2006). Table 4 presents the model performance of the evaluation metrics. Because the main objective is to predict churn, the combination of sensitivity and accuracy is a far more sensible choice than assigning equal weightage to all the other metrics. It is evident from Table 4 that the sensitivity of all models was not sufficient. The maximum and minimum values are 44% and 17.3%, respectively. Since accuracy and sensitivity are the most relevant metrics, data treatment was required in order to draw meaningful inferences. After SMOTE was performed, as shown in Table 3, we observed that XGBoost performed the best in terms of accuracy at 83%, followed closely by random forest at 78.3%. A similar trend was observed for the F1 score (XGBoost leads at 0.613, followed by random forest at 0.577), specificity (XGBoost leads at 90.3%, followed by random forest at 80.7%), and AUC (XGBoost leads at 0.847, followed by random forest at 0.831). The highest sensitivity was observed for logistic regression (71.4%), followed by random forest (69.3%). Since sensitivity and accuracy are the most relevant metrics, it is best to choose random forest

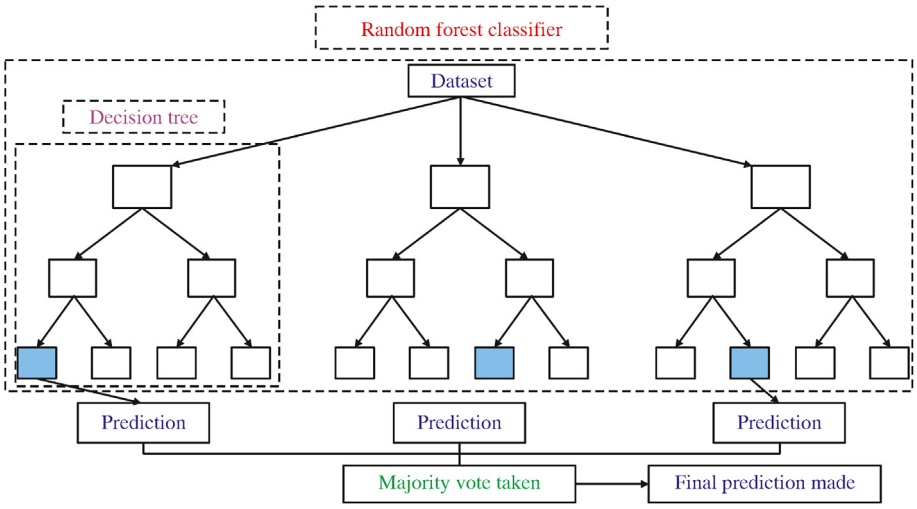

Fig. 8. Random forest classifier with dataset.

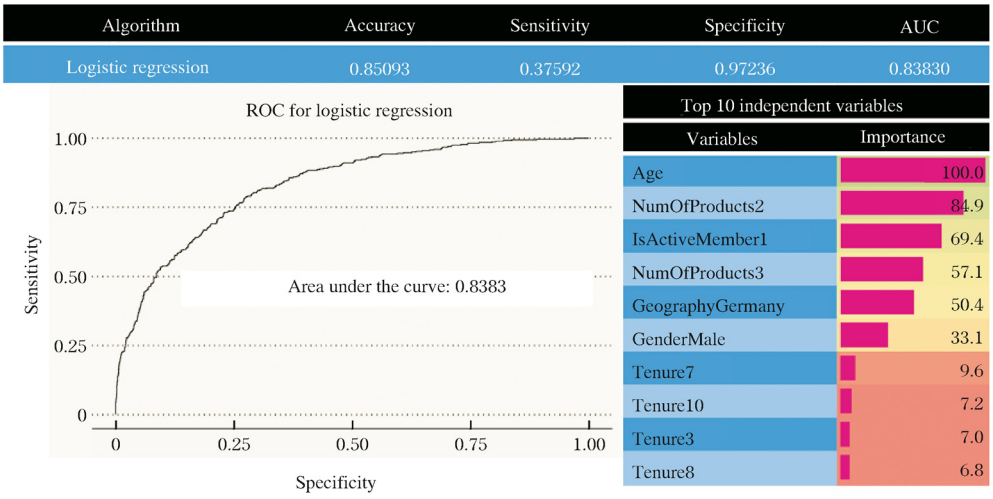

Fig. 9. Logistic regression model statistics.

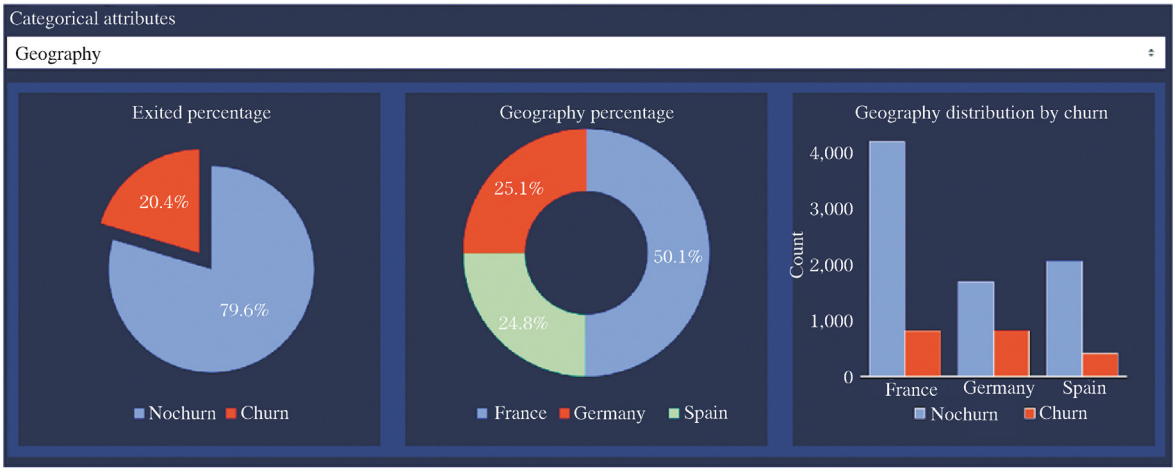

Fig. 10. Categorical attributes analysis.

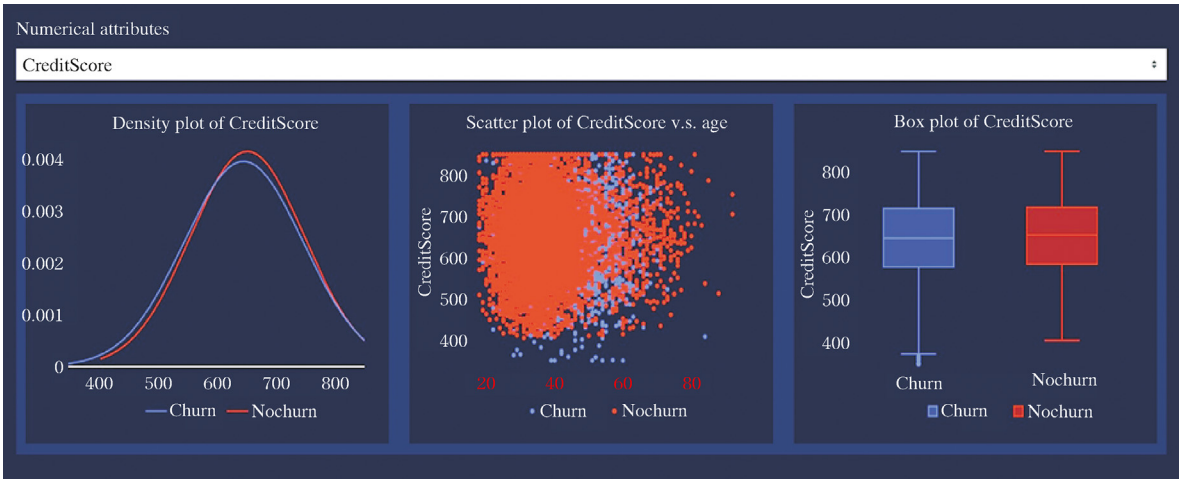

Fig. 11. Numerical attributes analysis.

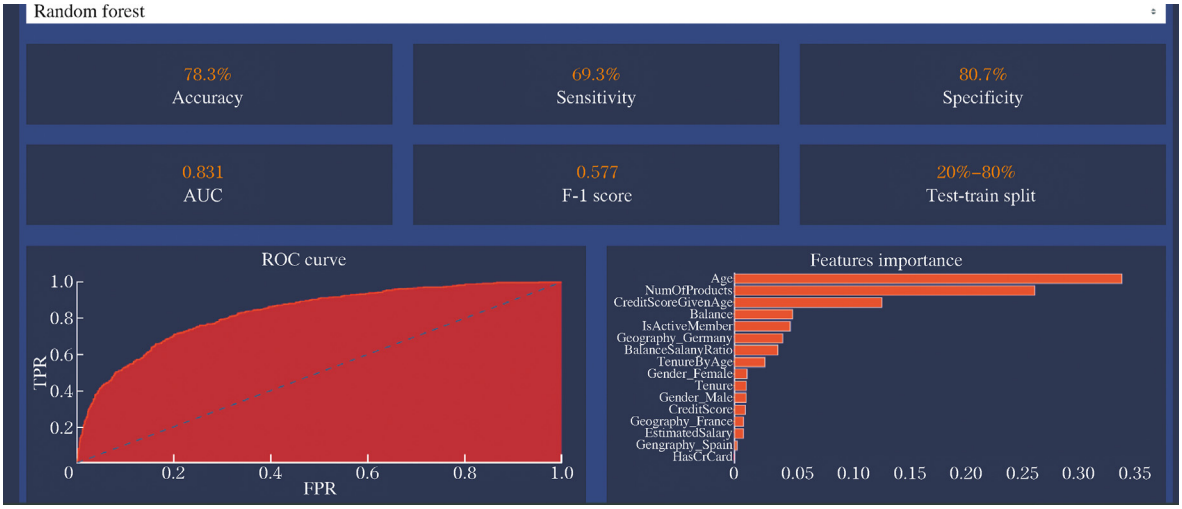

Fig. 12. Machine learning (ML) model statistics.

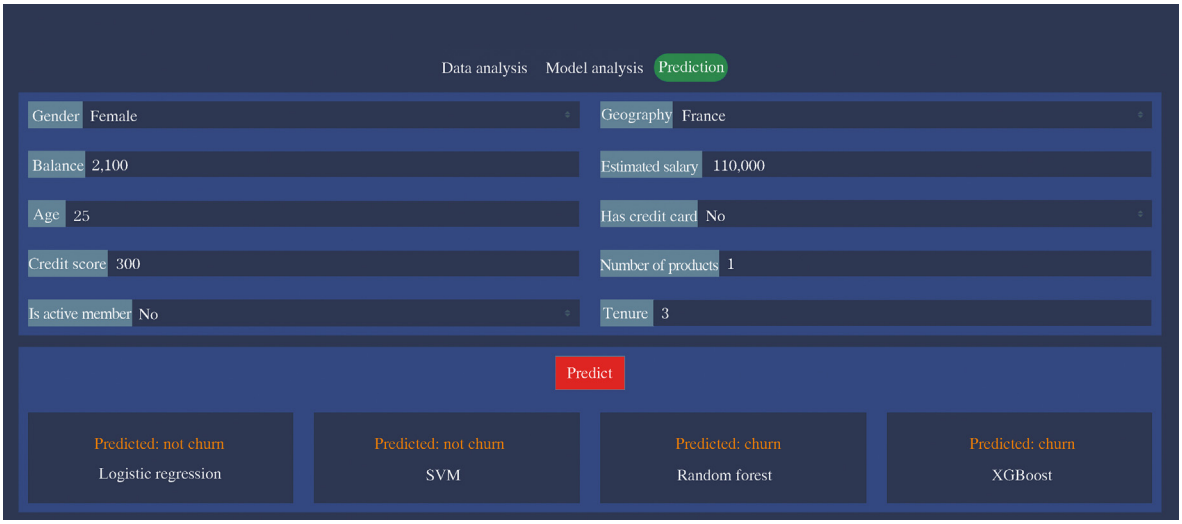

Fig. 13. Prediction for a given set of inputs.

Table 4
Evaluation metrics (without SMOTE) comparison for different approaches.

Metrics	Logistic regression	Support vector machine	Random forest	XGBoost
Accuracy	0.793	0.802	0.844	0.852
Specificity	0.961	0.987	0.977	0.963
Sensitivity	0.173	0.119	0.353	0.440
AUC	0.763	0.750	0.831	0.842
F1 score	0.263	0.205	0.491	0.559

Note: synthetic minority oversampling technique (SMOTE); area under the receiver operating characteristic (ROC) curve (AUC).

Table 5
Evaluation metrics (with SMOTE) comparison for different approaches.

Metrics	Logistic regression	Support vector machine	Random forest	XGBoost
Accuracy	0.691	0.719	0.783	0.839
Specificity	0.685	0.731	0.807	0.903
Sensitivity	0.714	0.672	0.693	0.601
AUC	0.767	0.765	0.831	0.847
F1 score	0.497	0.505	0.577	0.613

Note: synthetic minority oversampling technique (SMOTE); area under the receiver operating characteristic (ROC) curve (AUC).

because it exhibits reasonably good performance in both metrics. Another noteworthy advantage is that the random forest technique can handle large datasets owing to its ability to work with many variables. In addition, linear regression is very sensitive to outliers, as opposed to random forest, which helps us justify the latter's choice. We can draw valuable insights from the exploratory data analysis. The chance of churning increases by 0.93 when the customer is German as compared to French and Spanish customers (Table 5).

Additionally, the odds of churning are reduced by 0.5 when a male customer joins the bank compared to a female customer. However, the likelihood of a customer leaving the bank decreases by 1.5 if the new customer owns two products, whereas the probability of leaving increases by 2.5 if the new customer owns three bank products. Additionally, customers who maintain a balance of more than 85,000 are more likely to churn. Premium accounts and higher savings interest at other banks could be root causes for this. Similarly, customers with two bank products are more likely to stay. Crucial inferences can be drawn from these results. XGBoost achieved the best accuracy, 83.9%. The highest observed sensitivity in logistic regression was 71.4%. Random forest exhibited the best overall performance, with an accuracy of 78.3% and a sensitivity of 69.3%. Random forest also has the advantage of handling large datasets and accommodating many features.

The key findings of this research indicate that the random forest model performed best in terms of accuracy (78.3%) and sensitivity (69.3%), followed by XGBoost, with an accuracy of 83.9%. The analysis revealed critical factors influencing customer churn, such as nationality, gender, number of products owned, and account balance. For example, German customers were found to have a higher likelihood of churning than French or Spanish customers. Male customers have a lower probability of churning than female customers. Additionally, customers with two bank products are more likely to stay, whereas those with three products are more likely to churn. Customers maintaining a balance above 85,000 are also more likely to churn, potentially because of attractive offerings from other banks.

Banks can leverage this information to improve customer retention through the implementation of targeted strategies. For instance, they can focus on retaining German customers through personalized or tailored services. They can also offer incentives or rewards to customers who own multiple bank products to increase loyalty. Moreover, banks can enhance their efforts to understand and cater to female customers' specific needs and preferences to reduce churn. By identifying customers with high account balances, banks can proactively engage in offering premium account benefits or personalized financial solutions to mitigate the risk of attrition. The other alternative ways banks might utilize the findings of the present study include the following:

- Proactive customer retention strategies**.** By using ML algorithms and the insights derived from this research, banks can identify customers who are most likely to churn. This will enable them to implement proactive retention strategies such as personalized offers, targeted communication, or enhanced customer support for those at risk of attrition.
- Enhanced customer experience**.** Understanding the key factors contributing to customer attrition can help banks address pain points and improve the overall customer experience. By focusing on areas that drive dissatisfaction or disengagement, banks can make necessary improvements and increase customer satisfaction, thereby reducing churn.
- Tailored marketing and product offerings. The findings can guide banks in tailoring their marketing campaigns and product offerings. By identifying the patterns or characteristics associated with customer attrition, banks can develop targeted marketing messages and introduce new products or features that cater to specific customer needs, thereby increasing their value propositions and reducing the likelihood of churn.
- Effective decision making**.** The Data Visualization RShiny app developed in this study provides stakeholders with a comprehensive visualization, enabling them to make informed decisions. By utilizing the app and insights derived from this research, banks can gain a clearer understanding of churn trends, customer behavior, and the effectiveness of retention strategies. This will empower them to make data-driven decisions and effectively allocate resources to improve customer retention.

Furthermore, the findings of this study affect numerous other industries in addition to the banking sector. The examination of customer attrition and comprehension of the underlying variables are applicable to a variety of businesses, including insurance, telecommunications, subscription-based services, and e-commerce. Adopting a thorough preparation strategy and combining various pieces of data guarantee the accuracy and consistency of data analyses across sectors. Similarly, using ML algorithms to forecast customer churn or other important business outcomes enables the optimization of proactive client retention and marketing strategies. Decisions in areas such as CRM, marketing initiatives, resource allocation, and product development are supported by the creation of the Data Visualization RShiny app. The research conclusions thus have useful ramifications that may be applied to a variety of sectors, directing data-driven decision making and improving client retention methods.

5. Conclusion

This study helps to predict churn among bank customers with relative success. However, there is scope for improvement in the future. Due to the sensitive nature of banking data, access to large datasets is restricted. Access to more data points would enhance the generalizability of predictions. Additionally, having access to more granular data would contribute to improved forecasts. The current attributes are more specific to a customer's profile than their recency (the metrics that record behavior immediately before churning). Prospective researchers can derive these metrics, which will help track a shift in customer behavior just before they churn, and thus better identify churn patterns. There is also an opportunity for prospective researchers to improve our app by automating the model-training process, incorporating new features and data points, and generate updated models. This would help build a feedback loop in the models, ensuring greater veracity of model prediction by changing patterns and increasing the dataset. In addition, they

can go further by incorporating additional prediction algorithms that can be integrated into the visualization app for comparative analysis and better churn management. This would enable the deployment of this app in multiple businesses, where it can be used as a centralized churn management system. Another potential research topic involves developing localized prediction models that predict churn for only a subset of customers. For example, the given datasets could have different models for customers from different countries. The hypothesis is that individual models would drive higher accuracy and cumulatively greater generalizability than a model that is supposed to fit all situations. Additionally, the research indicates that different prediction algorithms are more suitable for different customer buckets, leading to more impact predictions.

CRediT author statement

Pahul Preet Singh: Conceptualization, Methodology, Software. **Fahim Islam Anik**: Writing-Original draft preparation. **Rahul Senapati**: Visualization, Investigation. **Arnav Sinha**: Software, Validation. **Nazmus Sakib**: Supervision, Correspondence, Writing-Reviewing and Editing. **Eklas Hossain**: Writing-Reviewing and Editing.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

References

Al-Mashraie, M., Chung, S.H., Jeon, H.W., 2020. Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: a machine learning approach. Comput. Ind. Eng. 144 (Jun.), 106476.

Alin, A., 2010. Multicollinearity. Wiley Interdiscip. Rev. Comput. Stat. 2 (3), 370–374.

Amuda, K.A., Adeyemo, A.B., 2019. Customers churn prediction in financial institution using artificial neural network. Available at: https://arxiv.org/abs/1912.11346.

Anton, S.D.D., Sinha, S., Schotten, H.D., 2019. Anomaly-based intrusion detection in industrial data with SVM and random forests. In: 2019 International Conference on Software, Telecommunications and Computer Networks. IEEE, pp. 1–6.

Baghla, S., Gupta, G., 2022. Performance evaluation of various classification techniques for customer churn prediction in E-commerce. Microprocess. Microsyst. 94 (Oct.), 104680.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, et al., 2020. A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing 408 (Sep.), 189–215.

Chawla, N.V., Bowyer, K.W., Hall, L.O., et al., 2002. SMOTE: synthetic minority oversampling technique. J. Artif. Intell. 16 (Jun.), 321–357.

Chen, T., Guestrin, C., 2016. XGboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

De Caigny, A., Coussement, K., De Bock, K.W., et al., 2020. Incorporating textual information in customer churn prediction models based on a convolutional neural network. Int. J. Forecast. 36 (4), 1563–1578.

De Lima Lemos, R.A., Silva, T.C., Tabak, B.M., 2022. Propension to customer churn in a financial institution: a machine learning approach. Neural Comput. Appl. 34 (14), 11751–11768.

Dias, J., Godinho, P., Torres, P., 2020. Machine learning for customer churn prediction in retail banking. In: International Conference on Computational Science and its Applications. Springer, Berlin, pp. 576–589.

Domingos, E., Ojeme, B., Daramola, O., 2021. Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. Comput. Times 9 (3), 34.

Fernández, A., Garcia, S., Herrera, et al., 2018. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J. Artif. Intell. 61 (Apr.), 863–905.

Geiler, L., Affeldt, S., Nadif, M., 2022. An effective strategy for churn prediction and customer profiling. Data Knowl. Eng. 142 (Nov.), 102100.

Guliyev, H., Tatoğlu, F.Y., 2021. Customer churn analysis in banking sector: evidence from explainable machine learning models. J. Appl. Mic. Econ. 1 (2), 85–99.

Han, H., Wang, W.Y., Mao, B.H., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing. Springer, Berlin, pp. 878–887.

He, B., Shi, Y., Wan, Q., et al., 2014. Prediction of customer attrition of commercial banks based on SVM model. Procedia Comput. Sci. 31 (Jan.), 423–430.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21 (9), 1263–1284.

Ho, S.C., Wong, K.C., Yau, Y.K., et al., 2019. A machine learning approach for predicting bank customer behavior in the banking industry. In: Machine Learning and Cognitive Science Applications in Cyber Security. IGI Global, pp. 57–83.

Karvana, K.G.M., Yazid, S., Syalim, A., et al., 2019. Customer churn analysis and prediction using data mining models in banking industry. In: 2019 International Workshop on Big Data and Information Security. IEEE, pp. 33–38.

Kaur, H., Pannu, H.S., Malhi, A.K., 2019. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput. Surv. 52 (4), 1–36.

Kim, J.H., 2019. Multicollinearity and misleading statistical results. Korean J. Anesthesiol. 72 (6), 558–569.

Lee, I., Shin, Y.J., 2020. Machine learning for enterprises: applications, algorithm selection, and challenges. Bus. Horiz. 63 (2), 157–170.

Lemmens, A., Gupta, S., 2020. Managing churn to maximize profits. Market. Sci. 39 (5), 956–973.

Machado, M.R., Karray, S., 2022. Applying hybrid machine learning algorithms to assess customer risk-adjusted revenue in the financial industry. Electron. Commer. Res. Appl. 56 (Nov.), 101202.

Manning, C., 2007. Generalized linear mixed models. Available at: https://nlp.stanford.edu/~manning/courses/ling289/GLMM.pdf.

Mansfield, E.R., Helms, B.P., 2012. Detecting multicollinearity. The American Statistician 36 (3a), 158–160.

Mazumder, S., 2021. 5 Techniques to handle imbalanced data for a classification problem. Available at: https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/.

Patil, P.S., Dharwadkar, N.V., 2017. Analysis of banking data using machine learning. In: 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud). IEEE, pp. 876–881.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133 (3), 225–245.

Rahman, M., Kumar, V., 2020. Machine learning based customer churn prediction in banking. In: 2020 4th International Conference on Electronics, Communication and Aerospace Technology. IEEE, pp. 1196–1201.

Schaeffer, S.E., Sanchez, S.V.R., 2020. Forecasting client retention—a machine-learning approach. J. Retailing Consum. Serv. 52 (Jan.), 101918.

Shirazi, F., Mohammadi, M., 2019. A big data analytics model for customer churn prediction in the retiree segment. Int. J. Inf. Manag. 48 (Oct.), 238–253.

Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In: Australasian Joint Conference on Artificial Intelligence. Springer, Berlin, pp. 1015–1021.

Sothe, C., De Almeida, C.M., Schimalski, et al., 2020. Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data. Gisci. Remote Sens. 57 (3), 369–394.

Statinfer, 2017. Calculating sensitivity and specificity in R. Available at: https://statinfer.com/203-4-2-calculating-sensitivity-and-specificity-in-r/.

Sun, Y., Wong, A.K., Kamel, M.S., 2009. Classification of imbalanced data: a review. Int. J. Pattern Recognit. Artif. 23 (4), 687–719.

Torgo, L., Ribeiro, R.P., Pfahringer, B., et al., 2013. SMOTE for regression. In: Portuguese Conference on Artificial Intelligence. Springer, Berlin Heidelberg, pp. 378–389.

Van Der Donckt, J., Van der Donckt, J., Deprost, E., et al., 2022. Plotly-resampler: effective visual analytics for large time series. In: 2022 IEEE Visualization and Visual Analytics (VIS). IEEE, pp. 21–25.

Vo, N.N., Liu, S., Li, X., et al., 2021. Leveraging unstructured call log data for customer churn prediction. Knowl. Base Syst. 212 (Jan.), 106586.