Discover Sustainability

Research

Ensemble-based customer churn prediction in banking: a voting classifier approach for improved client retention using demographic and behavioral data

Ruchika Bhuria[1] · Sheifali Gupta[1] · Upinder Kaur[2] · Salil Bharany[1] · Ateeq Ur Rehman[3] · Seada Hussen[6] · Ghanshyam G. Tejani[4] · Pradeep Jangir[5]

Received: 27 October 2024 / Accepted: 6 January 2025
Published online: 14 January 2025

Abstract
Customer turnover is a crucial issue in banking since maintained profitability depends on keeping clients. This work aims to categorize consumer turnover in banks by using a new ensemble approach combining many machine learning methods, hence enhancing churn prediction models. Using a comprehensive dataset including demographic, financial, and behavioral data—such as credit score, account balance, tenure, and activity levels—the study employs the goal variable revealing if a customer has left the bank. The study starts with univariate, bivariate, and multivariate feature exploration and subsequently uses the Interquartile Range (IQR) approach to identify outliers thereby improving the data quality. Five models—K-Nearest Neighbors, Support Vector Classifier, Random Forest, Decision Tree, and XGBoost—a Voting Classifier ensemble—are used to estimate project churn. Building upon all the strengths of each model, this approach improves the prediction of classification and provides a balanced and highly robust classification system. The applied approaches are K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest, Decision Tree, and XGBoost within a Voting Classifier configuration. The performance of the Voting Classifier without SMOTE yields the following results: Accuracy: 0.87, precision: 0.87, recall: 0.80, and F1-Score: 0.87. The proposed model that extend the base model using SMOTE (Synthetic Minority Over-sampling Technique), yields a higher prediction accuracy of 0.90, precision of 0.90, recall of 0.90 and F1-Score of 0.90. This enhancement is proving the efficiency of SMOTE to handle the class imbalance problem in order to render the churn prediction more balanced and reliable system. The proposed approach assures a reliable solution to the strategies to retain the customers in the banking organisations.

Keywords Machine learning models · Ensemble methods · Explainability · Challenges in statistical methods · Insights into customer behavior · Resilience · Inclusivity · Sustainability · Empowerment

✉ Ateeq Ur Rehman, 202411144@gachon.ac.kr; ✉ Seada Hussen, seada.hussen@aastu.edu.et; Ruchika Bhuria, ruchika.bhuria@chitkara.edu.in; Sheifali Gupta, sheifali.gupta@chitkara.edu.in; Upinder Kaur, upinderkaur45@gmail.com; Salil Bharany, salil.bharany@gmail.com; Ghanshyam G. Tejani, p.shyam23@gmail.com; Pradeep Jangir, pkjmtech@gmail.com | [1]Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India. [2]Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab 144411, India. [3]School of Computing, Gachon University, Seongnam-si 13120, Republic of Korea. [4]Jadara Research Center, Jadara University, Irbid 21110, Jordan. [5]Department of CSE, Graphic Era Deemed, To Be University, Dehradun, Uttarakhand 248002, India. [6]Department of Electrical Power, Adama Science and Technology University, 1888 Adama, Ethiopia.

1 Introduction

Customer retention is crucial for financial institutions, as acquiring new customers is significantly more costly than retaining existing ones. One of the challenges faced by banks and other financial institutions is predicting which customers are likely to leave (or "churn") over time. Bank churn prediction, which involves identifying customers who may terminate their relationship with the bank, has become an essential task in the financial sector. Accurate churn prediction models allow banks to proactively engage with at-risk customers through targeted interventions, personalized offers, or loyalty programs, thus reducing churn rates and enhancing customer satisfaction.

The banking industry has applied machine learning to enhance retention policies and more precisely project consumer disengagement than more conventional methods. ML models allow banks react early on by examining vast volumes of data, identifying trends, and estimating which customers are most likely to leave. Many studies have demonstrated how effectively ML performs in churn prediction—that is, in spotting variables generating customer dissatisfaction and improved projection of potential churners. Customer behavior, account activity, demographic data, and transaction patterns are the main factors ML models use to estimate turnover, so enabling banks to adapt retention efforts [1, 2].

Machine learning has transformed churn prediction systems from traditional statistical methods to more sophisticated, automated processes. While prior approaches needed significant manual feature engineering and battled with high-dimensional data, machine learning algorithms identify patterns automatically from data, hence improving prediction accuracy. Churn prediction has become somewhat well-known among the supervised learning methods—among models like decision trees, random forests, support vector machines (SVM), and gradient boosting. These models can run huge databases, create respectable outputs, and spot challenging trends. Moreover used to identify underlying trends in consumer data and variables generating turnover are unsupervised learning methods such clusterings and association rule mining [3, 4].

Especially for imbalanced datasets—common in churn prediction as non-churners often outnumber churners—ensemble techniques such gradient boosting and random forests blend numerous models to improve prediction accuracy. Combining forecasts from many models helps ensemble methods increase their accuracy by letting the model identify minority groups. Every method offers benefits based on the information and context. For example, since interpretability and transparency—qualities needed for regulatory compliance—make logistic regression, random forests, and decision trees in the banking industry rather popular [5–7].

Decision trees give a natural, understandable framework for analyzing consumer behavior in churn prediction. They also use the input variables to set decision points that the stakeholders can track how the forecasts are made. A number of decision trees in random forests reduce overfitting, and, therefore improve model predictions. GBM has good results in noisy or in high-dimensional data sets to work similarly; that is, to continuously correct mistakes of the preceding trees. In the field of churn prediction, another attractive algorithmic class applied in the banking industry, for instance, is the support vector machines (SVM) since they are efficient in dealing with high dimensions as well as nonlinearity due to their capability from handling interaction and outliers [8].

More so, in many fields such as banking where legal issues require transparency and hence explainability, churn prediction systems are basic. Although more complex models such as random forests and gradient boosting machines could be considered as black boxes, simpler models such decision trees and logistic regression are naturally interpretable. Techniques such as LIME and SHAP (Shapley Additive Explanations) have been developed to offer post-hoc explanations to aid to interpret the contributions of various characteristics to model predictions and so address this. Especially in the financial industry where decisions can affect client relationships and confidence, explainability guarantees that the predictions of the model match corporate aims and legal standards [9]. The body of current research on churn prediction in the banking industry mostly centers on the application of support vector machines, random forests, and decision trees—machine learning models. Particularly in imbalanced datasets where non-churners greatly outnumber churners, there is a knowledge vacuum on how best to combine several models to raise prediction accuracy. Moreover, although a lot of study has been done on model performance, little has been done to guarantee explainability and transparency in these models—qualities absolutely essential for following banking sector regulations. Focusing on its application to imbalanced datasets, this work attempts to close these gaps by investigating ensemble-based techniques for churn prediction and incorporating explainability methods to improve model transparency in a highly regulated sector [10].

Basically, churn prediction is a vital task for the banking industry since maintaining customer loyalty determines profitability. Machine learning techniques offer a good approach to project turnover since they allow one to look

at big data and spot complex trends.By using ensemble strategies and assuring model explainability, banks can at last change client retention strategies, satisfy regulatory criteria, and improve the accuracy of their forecasts. As competition from both established banks and new fintech firms increases, churn prediction models supported by machine learning provide banks with a practical tool to maintain ahead in the market.These three contribution will ensure that our research paper on banks customer turnover forecast is robust: -

- This work proposes a Voting Classifier using many applications of machine learning algorithms—including Random Forest,Decision Tree, Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and XGBoost. The technique strengthens not just prediction accuracy but also model generalizing power and resilience, therefore providing banks with a handy tool to correctly identify at-risk customers.
- The research does a thorough exploratory data analysis (EDA) using one-dimensional, two-dimensional, and multi-dimensional techniques to identify relationships between consumer attributes and attrition behavior.This complete EDA provides banks with tools to find the underlying trends and features of their client base by way of sharp examination of the variables producing customer turnover.
- This work emphasizes the need of data quality by means of a whole pretreatment pipeline encompassing fixing nonexistent values, feature measuring, one-hot conserving for categorization variables, and Interquartile Range (IQR) approach outlier identification. These methods ensure that the dataset is strong and suited for machine learning, therefore enhancing the model's capacity to identify significant trends and produce accurate forecasts.

The plan of this research divides it into five major sections, each of which is dedicated to important concerns of the investigation. Related Work, described in Sect. 2, presents an overview of the existing literature on customer churn prediction with an emphasis placed on the techniques that have been used in similar scenarios. Section 3, Proposed Methodology, presents the procedure adopted in this study starting with a description of the basic input dataset followed by use of Synthetic Minority Over-sampling Technique (SMOTE) for dealing with imbalance class. The data preprocessing step is explained in detail in Sect. 3.1 In Sect. 3.2 to Sect. 3.6, a one-dimensional, two-dimensional, and multi-dimensional analysis of the plotted dataset are presented to identify the structure of data and relationships among features. In Sect. 3.7 the Voting Classifier is introduced as the major model adopted in the study and it uses other machine learning classifiers to enhance on accuracy of the final prediction. Section 4, Results and Discussions, contain the analysis of the models by using quality criteria: accuracy, precision, recall, and F1-score. Last but not the least, Sect. 5, Conclusion and Future Work, concludes the study by putting forward the specific findings, the relevance of the work done and directions for the future research study in consideration of real-time predictive systems and other optimality fact.

2 Related work

The modelling of customer churn using an ML approach has quickly grown to be one of the most important research areas in the banking sector as organizations work to reduce churn and operating costs. This body of work also present several works that concern different methods that give different perspectives and new strategies to enhance churn prediction models. Vu [11] implemented a complex method that involved the use of a number of models in the forecasting of customer churn in the commercial banks, which emphasized the importance of system integration. The author also pointed out the use of classifiers such as Random Forest and Gradient Boosting allows the produce an ensemble model yielding to the different behaviors present in banking customers. Likewise, Huseyinov and Okocha [12] have dealt with churn issue with the assistance of the ML approach insisting on the necessity to enhance feature selection and hyperparameter for better prediction. Their work also pointed out that data preprocessing was a vital step in reaching the targeted model outcomes particularly in cases where there are noisy and sectional numerous pieces of data. Soni and Nelson [13] have described a churn prediction framework for profit motivated context in which, churn forecasting models are aligned to profit. In the second technique, the authors integrated profit metrics into churn prediction with the check that the resulting predictions are not only the customers at risk of churn but also the customers whose retention has the most significant impact on the organisation's bottom line.

Customer churn was investigated by Asfaw [14] using the Ethiopian case while using ML to study the banking industry and the possible social, economic and culture implications. This study also looked at how the models proposed are exportable when handling localized data but this does not seem to impact on them. Hui et al. [15] reported detailed study on customer churn risk situation in ABC Multistate Bank under Decision Trees, Support Vector Machines and Neural

Networks. They identified that these methods in four aspects of interpretability, speed and accuracy's and suggested the four methods be used but in an integrated manner. In conjunction with the current study, Lukita et al. [16] employed the data mining and ML for the purpose of the predictive analysis in conjunction with the EDA methodology with integration of model development in order to generate useful insights for the banking institutions. Khandelwal in [17] applied this principle component analysis to this area of churn prediction and explained that by applying this process of dimensionality reduction, the computational efficiency of the outcome is enhanced without necessarily having to compromise on the rate of correct prediction. This is particularly useful where datasets contain many variables due to issues that come with multicollinearity. Chen et al. [18] have conducted this study to European banking customers and integrating the aspect of behavior where the values of transactions, complaints, etc. can be used to construct customers models and insights. They observed that any scale up in the interaction level of a customer posed greater risk than a customer, who is a novice or an occasional shopper.

Interpretability is also a requirement where high risk industries such as banking are involved and this was well handled by Murindanyi et al. [19]. Explaining their approaches, the authors proved that it is possible to develop accurate algorithms in conjunction with the compliance with rules and regulations within the identified decision trees and SHAP (SHapley Additive exPlanations) values. Similarly, Simsek and Tas [20] used the approach of customer portfolio churn by classification techniques where the attributes of the portfolio enhance the predictive model. This approach is special because it points to the use of the personal analytic in churn management. Talwadia et al. [21] associated churn and credit card analysis and defined a single model to identify churn possibility of customers and classify them on the basis of their credit card usage. It is thus possible to bifurcate the error analysis for purposes of making right interferences. The problem of class imbalance was addressed as usual in churn prediction datasets by Hambali and Andrew [22] where Synthetic Minority Over-sampling TECHNIQUE (SMOTE) was used. They also compared imbalanced and balanced datasets that they got from the study and realised that hence approaches meant to balance distributions up improves the reliability of the result of ML models. Wang & Chen [23] provided an extensive discussion on the application of ML and DL techniques for credit card churn prediction; the accuracies, as well as computational efficiency were then compared. Along the same vein, Al-Sultan and Al-Baltah [24] has contributed to the Random Forest algorithms for the balanced and unbalanced data sets The authors impose an adjustable weighting mechanism to enhance the facial of algorithms stating and practicing the minority class with no extra complexity.

This is in line with Gkonis and Tsakalos [25] where issues of imbalanced learning were dealt with when invoking the use of grid search as well as auto tuning mechanisms. In addition, carrying out the analysis of churn up to banking and telecom arena, Bhaal et al. [26] attempts to compare the patterns of churn analysis in both the sectors. The obtained results might provide valuable insights on how to transfer and apply the ML models between various industries and companies. In [27], Kaya discusses on the commission rates in brokerage firms and banks as well as claiming the application of the use of the ML algorithms on the management of structured financial datasets. More practically in Random Forest for bank churn prediction, Zhang et al. [28] further provided a discussion on how to develop this framework. It also demonstrated how their algorithm could handle the scale and type of interaction found in these datasets. Gurung et al. [29] noted the maneuverability of models based on AI for churn prediction in the US market by applying for both neural networks and ensemble methods to focus on the aspects of the market. The others such as Baby et al. [30] called for an Artificial Neural Network on the basis of a case it therefore supports the view that the tactics of testing theories involves use of real life application of the conceptual models. In the same vain Raj et al. [31] examined the concerns of scalability and computational effectiveness of CMS to address he rising demand for real time customer churn prediction systems in banks. Other authors such as Galal et al., [32] used the works of others to introduce the hybrid meta-learners in the aspect of youth loyalty in digital banking, where meta-learners apply the ensemble techniques to solve the problem by combining the strong points of base models.

This approach is similar to what has been done by Poudel et al. [33] whereby tabular ML models, especially that of churn prediction, incorporated integrated interpretability improvements such as LIME to enhance users' trust. Kim et al. [34] have done a study on an extent of large panel data in churn prediction and the analysis done for the same are as follows that explain how much granularity of the data counts in it. Thus, they contribute to augmenting the notion of flexible data preprocessing pipelines that must be good for any size of datasets. On similar note, Kharat and Rane [35] employed theories for churn prediction and in the process developed a user friendly web application that the banking practitioners could easily employ in the analysis. Manzoor et al. [36] have also provided a very good understanding of the ML techniques for churn prediction while giving a good code of conduct for business practitioners to select and use models efficiently. Another study by Peng et al. [37] outlined the issue of interpretability of churn predictive models – a topic that has received a lot of focus lately thanks to AI systems that make decisions with no indication of

the underlying process. As such, Vu [38] proposed a neural network based churn prediction model to outline the most advanced approach regarding model sophistication and performance. Individually, all these work offers a historical perspective with the churn prediction methodologies and recognized different and improved methods from statistic, machine learning, and deep learning methods. Current Areas of Interest are; improvement of interpretability and profit, establishing new models and significant aspects. It is found that there is immense scope for future work in the real-time churn prediction and understanding of the ethical aspects of using AI, and going across the domain to further enhance this important field.

3 Proposed methodology

For the banking industry, customer turnover—the phenomena whereby customers stop interacting with a company—poses a major problem. Maintaining profitability and market share depends on the identification of consumers likely to leave as the expense of gaining new ones typically exceeds that of keeping current ones. This work focuses on the use of machine learning methods based on a dataset including a broad spectrum of consumer characteristics to forecast bank customer attrition. This work intends to create a predictive model that enables banks to foresee which consumers are at risk of leaving and implement preventative measures by using advanced categorization techniques.Comprehensive information about every bank customer—including demographic traits (age, gender, location), financial metrics (credit score, account balance, estimated salary), and behavioral indicators—tenure with the bank, number of products owned, credit card ownership, and activity level—is included in the dataset used for this study. Dubbed "Exited," the goal variable shows whether or not a consumer left the bank. One-hot encoding for categorical variables, feature scaling, and missing data management were part of a comprehensive data preparation pipeline.

The analysis began with one-dimensional, two-dimensional, and multi-dimensional exploratory data analysis to uncover relationships across the features and the target variable. One-dimensional analysis involved investigating the distribution of each individual feature, such as examining the age distribution of churned versus retained customers. Two-dimensional analysis was used to explore relationships between pairs of features, such as how the interaction between credit score and balance affects churn. Multi-dimensional analysis extended this further by analyzing combinations of multiple features, providing deeper insights into complex patterns within the dataset. Establishing the model's resilience requires first a thorough awareness of outlays. Outliers were found using the Interquartile Range (IQR) technique and either handled or eliminated to stop them from distorting the forecasts of the model as seen in Fig. 1. This guaranteed the model learnt significant patterns free from effect of extreme, aberrant data points, hence improving the general quality of the data.

To predict customer churn, researchers applied an ensemble learning approach employing a voting classifier. Multiple machine learning techniques are combined in this method to raise general accuracy and resilience. Five classifiers in particular—XGBoost, Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC)—were employed. Combining these models helps the Voting Classifier to use the strengths of every method, thereby lowering the overfitting risk and enhancing generalizing on fresh data. Under this method, the class with the most votes is chosen as the final prediction by means of majority voting—where each classifier votes on the expected class.

3.1 Input dataset with smote

The dataset provides comprehensive information on bank customers and their churn status, therefore revealing whether they have left the bank. It is a useful tool for creating predictive models to identify customers who might be leaving as well as for looking at and assessing the factors affecting bank customer turnover. The dataset contains 10,000 rows of customer information for a bank churn prediction task and 13 features. Important traits are the client's Surname, a sequential Row Number for every entry, and an original Customer ID for identification. Apart that geography location showing the consumer, the dataset comprises the creditworthiness indicator, the CreditScore. Important demographic data include gender, age, and tenure—years spent with the bank; balance shows the potential account balance the client could be able to reach. Aside from that, HasCrCard tells whether a credit card is owned and NumOf things displays the banking item count. Although the Active Member tool displays the customer's activity level with the bank, estimated salary provides information about their financial status. The Exited feature indicates if the consumer has left the bank, thereby serving as the goal variable for predictive modelling. This dataset helps one understand consumer behaviour and improve methods of banking sector retention.

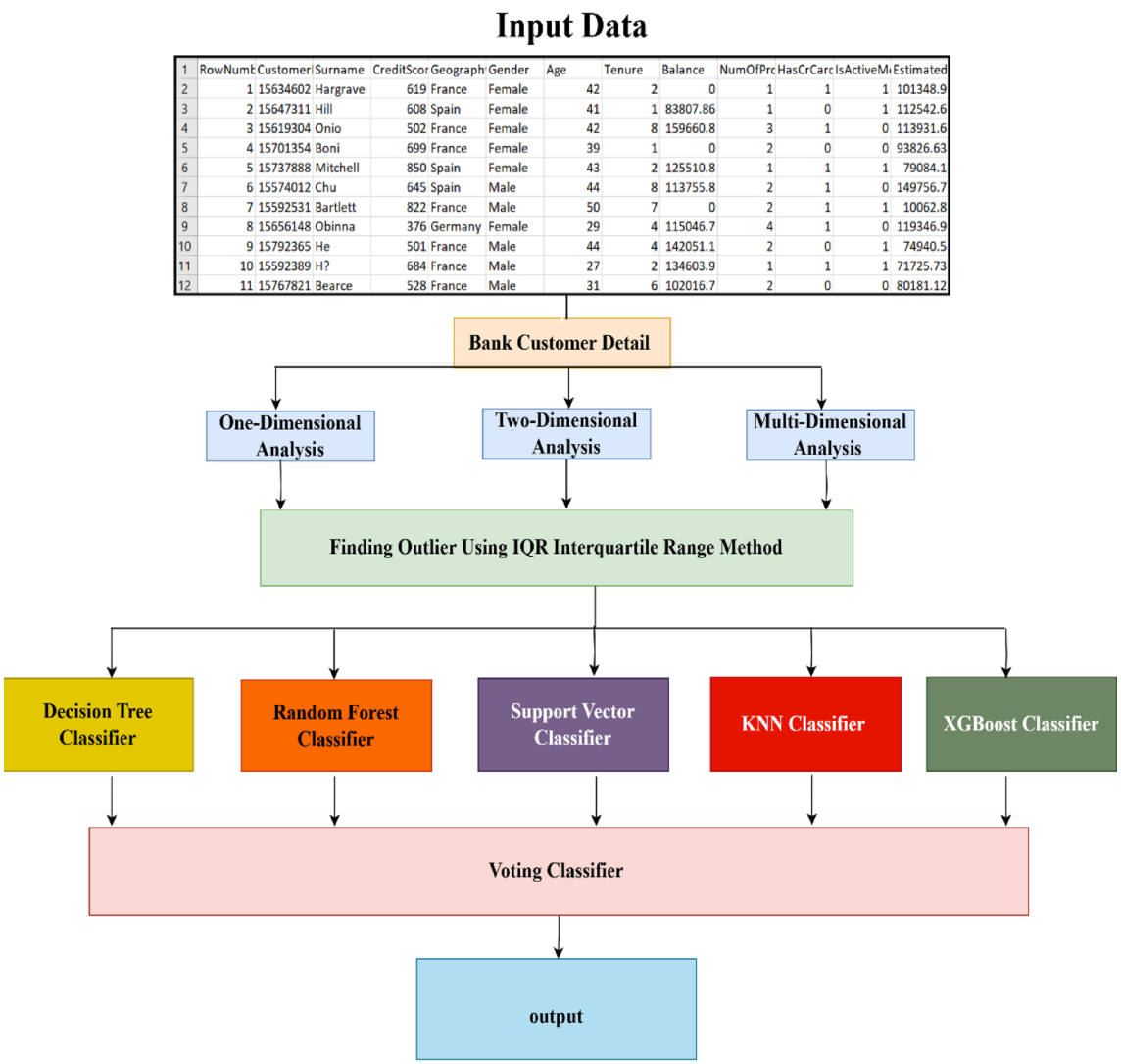


Fig. 1 The Flowchart for Bank churn customer classification using machine learning

There are 10,000 rows total in the dataset; 2,037 rows are churned customers (Exited = 1) and 7,963 rows are non-churned customers (Exited = 0). This mismatch in the dataset—where non-churned consumers far outnumber churned ones—may cause a model to be biassed toward estimating the majority class (non-churners). We implemented the Synthetic Minority Over-sampling Technique (SMote) to solve this class imbalance. Based on the feature space of the current churned records, SMote generates synthetic data points for the minority class—churned customers—so augmenting the number of churned customer instances.

By boosting the number of lost consumers, SMOTE will help to balance the dataset from 2,037 to 8,000. The dataset will have 7,963 non-churned and 8,000 churned consumers following SMote, so generating 15,963 rows total. This balanced dataset will allow the model to more effectively understand the patterns and attributes generating customer turnover and help avoid it from being biassed toward the non-churned class. The model will be able to accurately estimate both non-churn and churn scenarios if the two classes have more equal share. Figure 2 gives the detail of bank churn customer.

3.2 One-dimensional analysis

One-dimensional analysis is, the examination of individual traits within a dataset in isolation that helps one to understand data. This stage helps define the statistical characteristics and distributions of every variable therefore enabling analysts to obtain important insights before performing more complex research involving several variables. The major

Fig. 2 Bank customer and their churn status parameters

RowNumb	Customer	Surname	CreditScor	Geograph	Gender	Age	Tenure	Balance	NumOfPrc	HasCrCarc	IsActiveM	Estimated	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.9	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.6	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.6	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.8	1	1	1	79084.1	0
6	15574012	Chu	645	Spain	Male	44	8	113755.8	2	1	0	149756.7	1
7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.7	4	1	0	119346.9	1
9	15792365	He	501	France	Male	44	4	142051.1	2	0	1	74940.5	0
10	15592389	H?	684	France	Male	27	2	134603.9	1	1	1	71725.73	0
11	15767821	Bearce	528	France	Male	31	6	102016.7	2	0	0	80181.12	0
12	15737173	Andrews	497	Spain	Male	24	3	0	2	1	0	76390.01	0

objective of univariate analysis is a descriptive statistical summary of every variable. By examining the "Age" data, for example, analyzers can determine the mean, median, and mode—so estimating the average age of bank consumers. Among visual tools that highlight skewness or outliers and indicate these central tendencies are histograms and box graphs. Using a Data Frame including bank customer information, this study does univariate analysis. First it chooses all numerical columns and generates histograms for every one of them as shown in Fig. 3 (a) showing their distribution together with kernel density estimations to show the underlying trends of the data. It then finds categorical columns and creates count graphs for the top 10 most often occurring values in every category, therefore offering understanding of the distribution of categorical data. The distribution of Customer ID is illustrated in Fig. 3b, where the X-axis ranges from approximately 1560 to 1585 and the Y-axis represents counts from 0 to 700. The green bars indicate most counts around 600 per bin, with slight dips at the edges, and the black line provides a smooth overlay of the distribution. Similarlily Fig. 3c–k shows the histogram distribution of credit score, age, tenure, balance, number of products, credit card, active member, estimated salary and exited status respectively.

Understanding the features of the data prior to more research or modeling is made easier by this all-encompassing visualizing method. It especially shows the distribution of the Surname, Geography, and Gender columns, thereby offering information on the demographic makeup of the clientele. Every subplot shows counts of events for the most often occurring categories, therefore facilitating a simple comparison of consumer traits. Understanding the elements influencing consumer behavior and turnover depends on this study, which also guides next modeling projects. Figure 4a illustrates a horizontal bar chart titled "Count of Surname," which displays the number of occurrences for various surnames. The Y-axis lists the different surnames, while the X-axis represents their counts, ranging from 0 to 35. The color-coded bars indicate that Smith has the highest count, whereas Wright has the lowest. Under "Count of Geography," Fig. 4b shows a horizontal bar chart showing individual distribution throughout three areas. With almost 5000 people, France boasts the biggest representation on the table; Germany comes second with almost 2500; Spain comes third with almost 2000. This graphic depiction does a good job of stressing the three countries' demographic differences. With a horizontal bar chart labeled "Count of Gender," Fig. 4c displays the two gender boundaries. —"Male" and "Female"—their respective counts. Whereas the y-axis shows gender categories, the x-axis runs from 0 to 6000. Visually indicating a greater male count in the dataset, the light blue bar for "Male" stretches to almost 5500 while the bright orange bar for "Female" reaches roughly 4500.

3.3 Two-dimensional analysis

By concurrently investigating the link between two variables, two dimensional analysis expands the insights obtained by one dimensional analysis. Investigating possible relationships between individual consumer traits and the phenomena of attrition, or client turnover, benefits especially from this analytical method. For this study, for example, two dimensional analysis can be used to investigate the interaction between the target variable—customer turnover—and demographic elements including geography and gender. Count graphs and cross-tabulations help analysts see how churn rates vary between sexes and over geographical areas. Plotting churn rates versus geographical areas, for instance, can highlight notable differences suggesting that consumers in particular areas are more inclined to depart the bank. This all-encompassing visualizing technique allows one to grasp the characteristics of the data before extra investigation or modeling. It particularly provides details on the demographic composition of the clients by showing the distribution of the Surname, Geography, and Gender columns. Every subplot displays counts of events for the

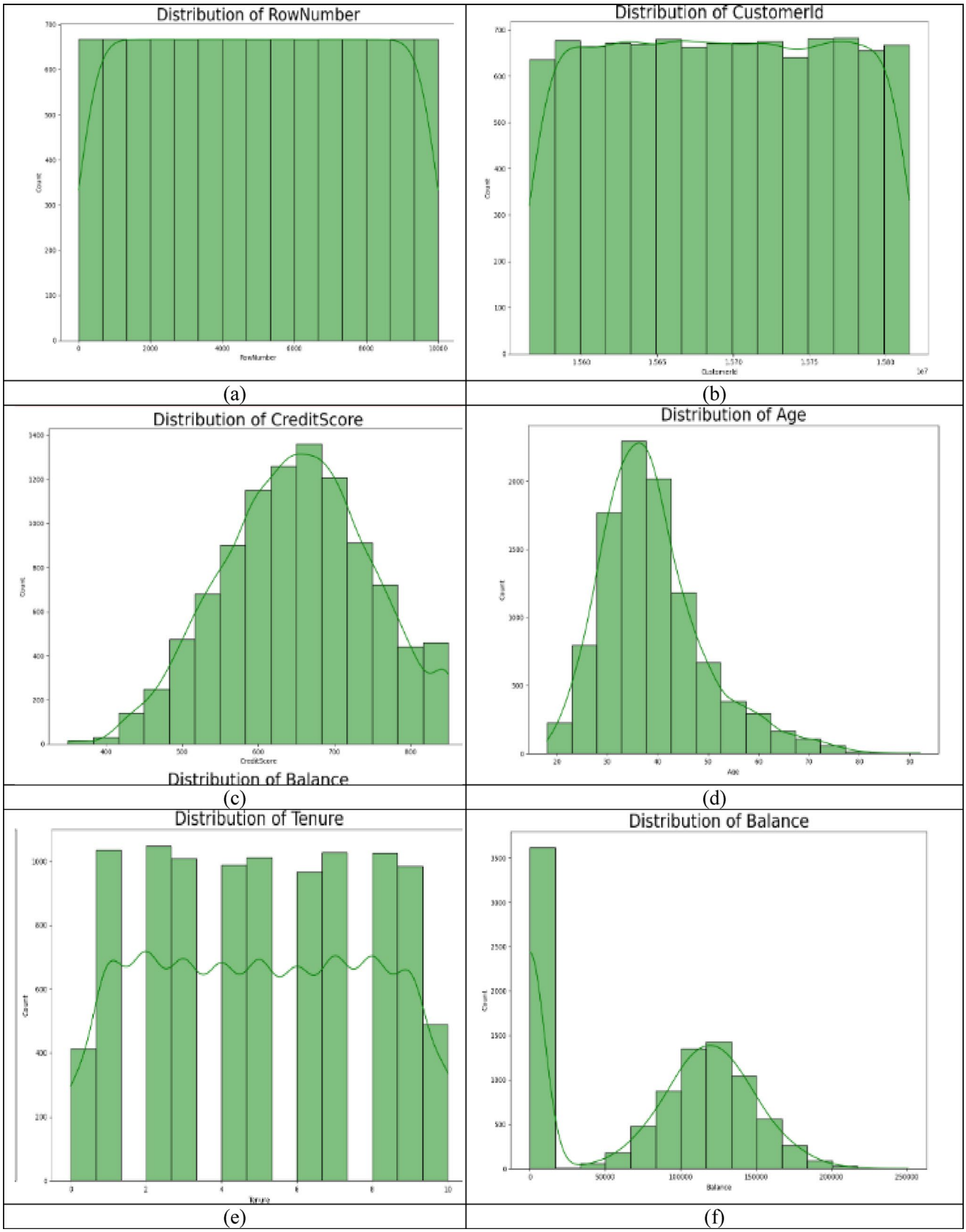


Fig. 3 Gives the information numerical and statistical data bank churn customer **a** Distribution of Row number, **b** Distribution of Custormer id, **c** Distribution of Credit score, **d** Distribution of Age, **e** Distribution of Tenure, **f** Distribution of Balance, **g** Distribution of Num of Products, **h** Distribution of Cr Card, **i** Distribution of active member, **j** Distribution of Estimated Salary, **k** Distribution of Exited

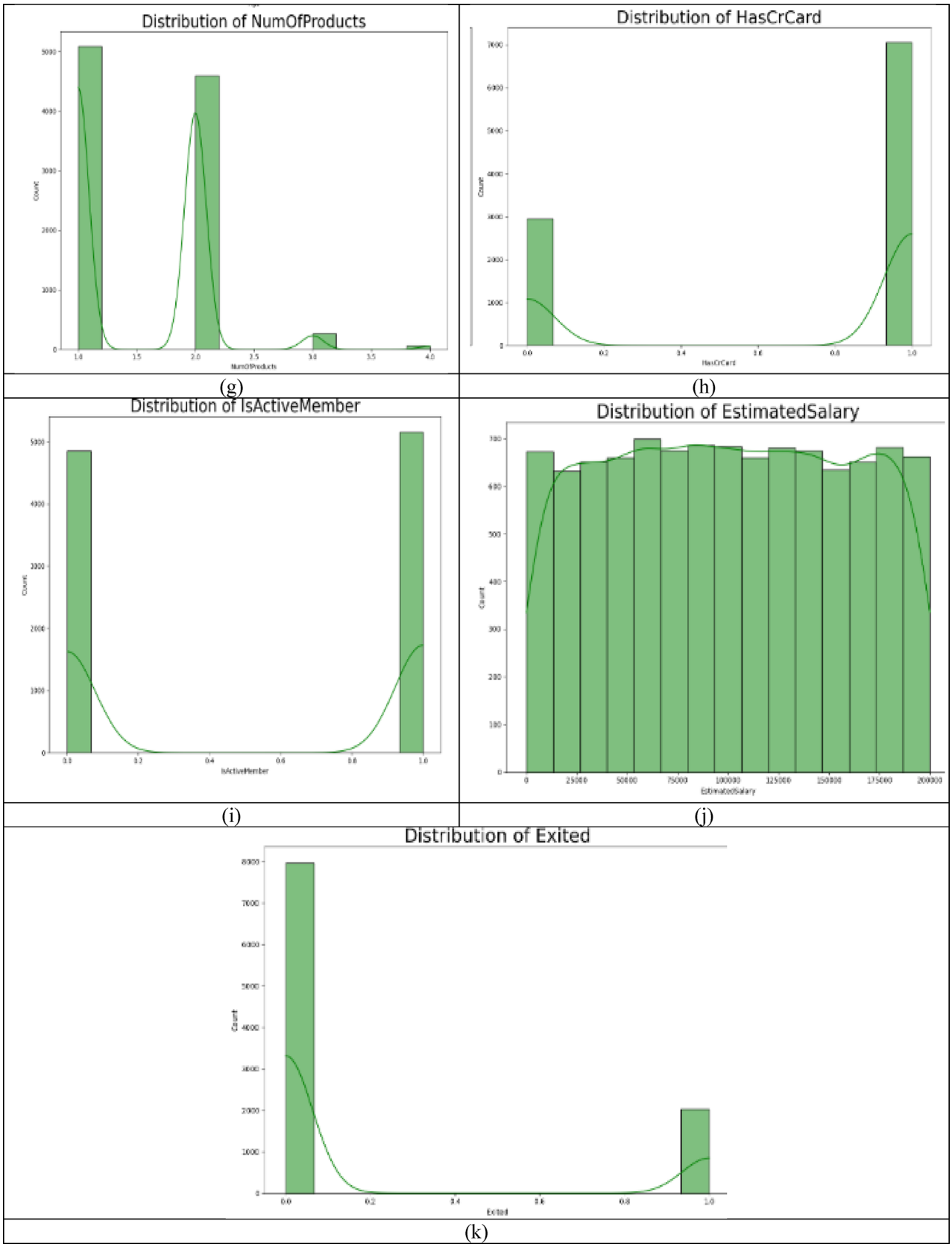


Fig. 3 (continued)

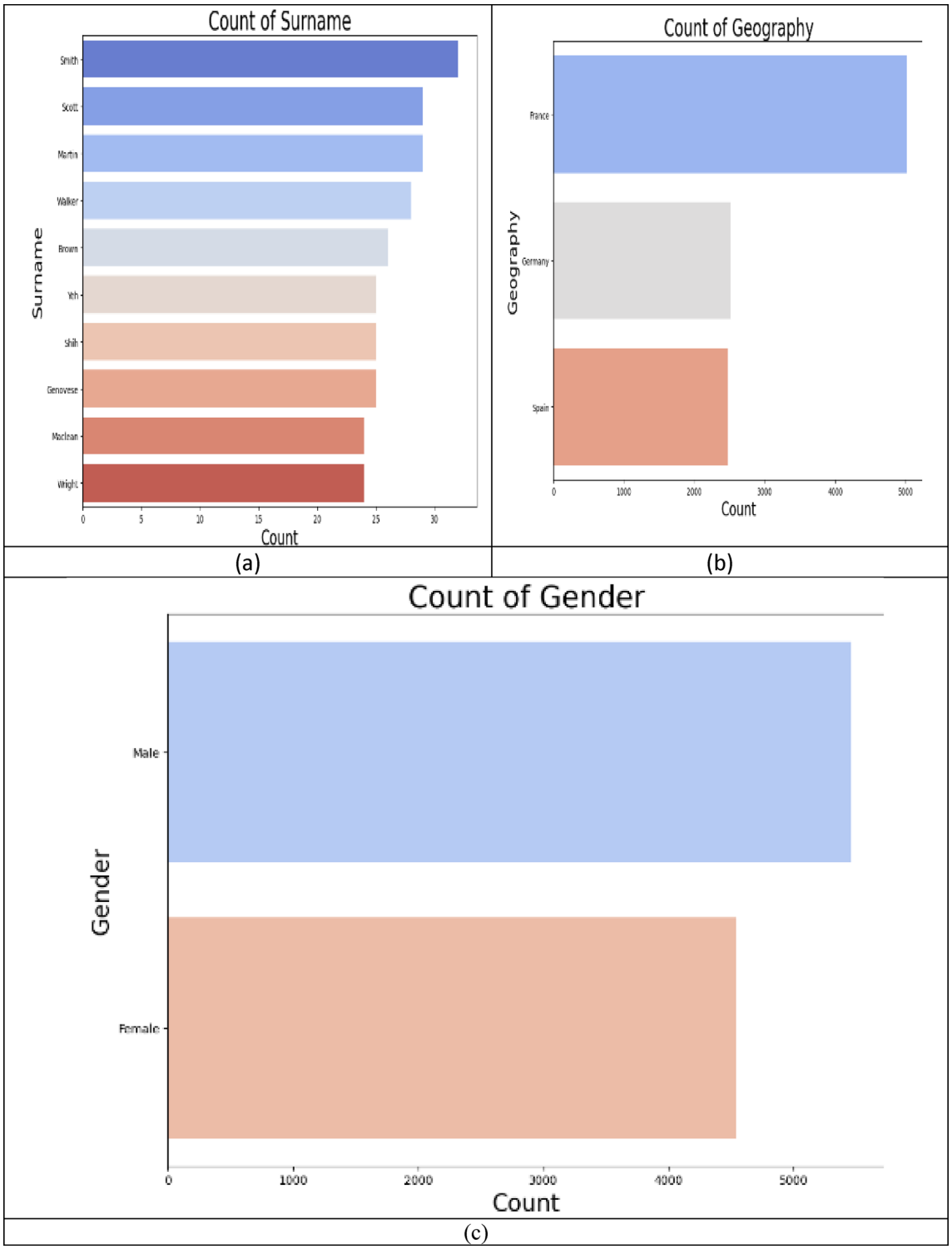


Fig. 4 the demographic distribution of the customer like **a** Count of Surname, **b** Count of geography, **c** Count of gender

most regularly occurring categories, therefore enabling a basic comparison of consumer characteristics. Figure 5a shows a box plot contrasting the two groups of the "Exited" variable's "RowNumber" distribution 0 and 1. The Y-axis shows "RowNumber," which can range from 0 to 10,000; the X-axis shows the "Exited" variable. For "Exited" = 0, box colors are green; for "Exited" = 1, orange. With identical interquartile ranges and whiskers, both groups show a median "RowNumber" close to 5000, hence stressing the spread and central tendency among these groups. Especially, this

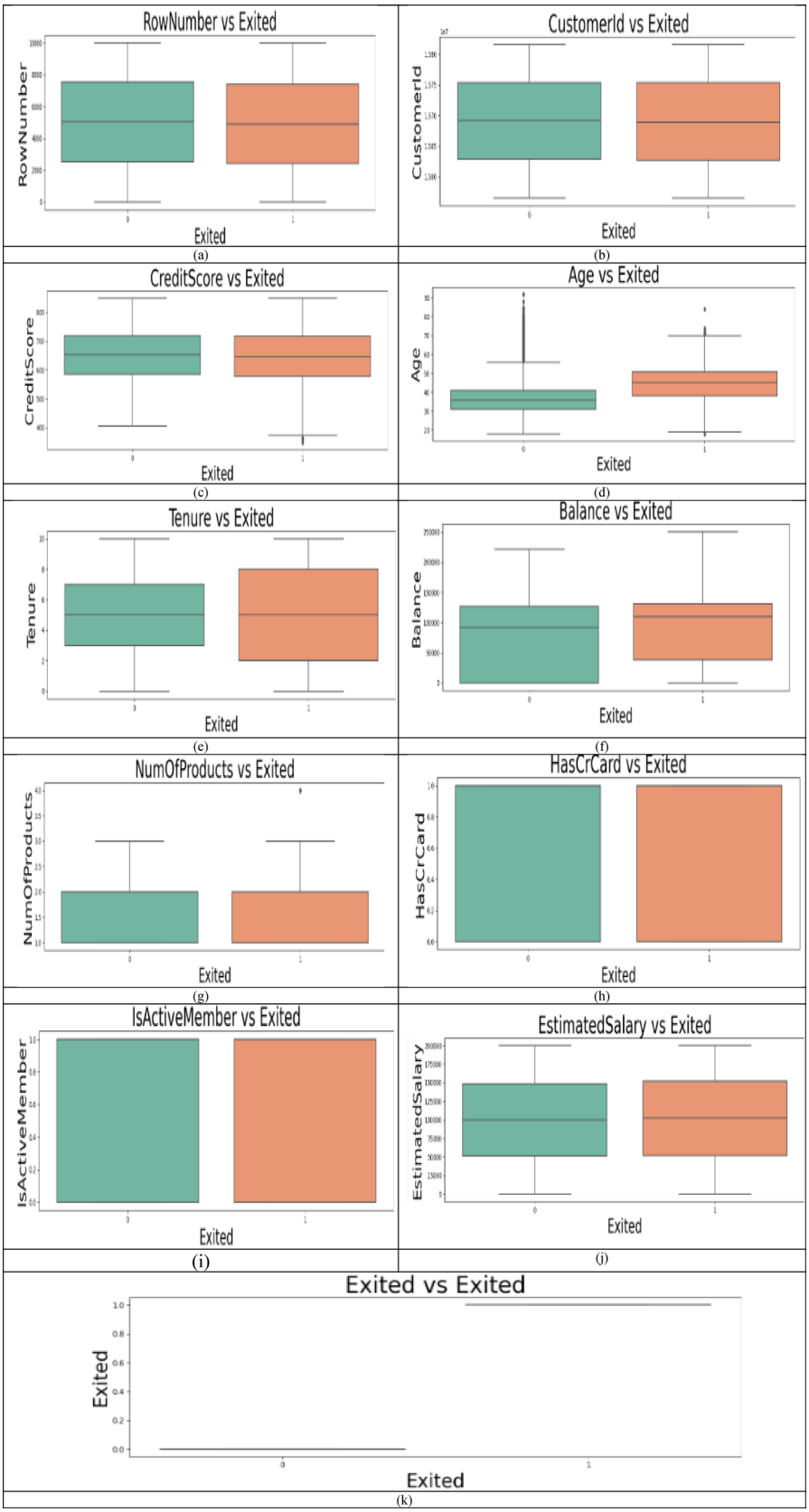


Fig. 5 Box plot comparing the **a** Row Number vs Exited, **b** Customer ID vs Exited, **c** Credit Score vs Exited, **d** Age vs Exited, **e** Tenure vs Exited, **f** Balance vs Exited, **g** NumofProducts vs Exited, **h** HasCrcard vs Exited, **i** IsActive Member vs Exited, **j** Estimatedsalary Vs Exited, **k** Exited vs exited

graph clearly shows the way "RowNumber" is distributed between those who left and those who did not. CustomerId values for two groups—those who have left (1) and those who have not (0) are shown in Fig. 5b as a box plot. Similarlily Fig. 5c–k credit score, age, tenure, balance, active member, estimated salary and exited respectively. Here Fig. 5(c), 5(d) and 5(g) are showing outliers in the boxplot of credit score, age and number of products respectively.

Using count graphs for the top ten values of every category feature in respect to the goal variable, Exited, the study effort performed a two dimensional analysis. This indicates customer turnover. The count plots for every categorical column show the most often occurring categories while separating consumers who left from those who did not via color coding. This method enables a comparison of the relationships between various category values and customer turnover, therefore offering information on which demographic or behavioral element might affect the probability of a client leaving the bank. The way trends and patterns are clearly shown by the visualization helps guide focused interventions and client retention policies. Figure 6. Shows the Count Graph to explore the relationship between catagorial feature and customer (exited). Figure 6a bar chart, titled "Surname vs Exited," is analyzing the relationship between different surnames and whether individuals have exited or not. "Smith" has the highest number of non-exited individuals, and "Scott" has the highest count of exited individuals. Figure 6b bar chart, titled "Geography vs Exited," is analyzing the relationship between different surnames and whether individuals have exited or not. "Germany" has the highest number of non-exited individuals, and "France" has the highest count of exited individuals. Figure 6c shows, by gender (Female and Male), a bar chart comparing the total number of people who left a service or organization. According to the chart, almost 3,000 women and almost 1,000 left; almost 4,500 men did not go compared to almost 500 who left.

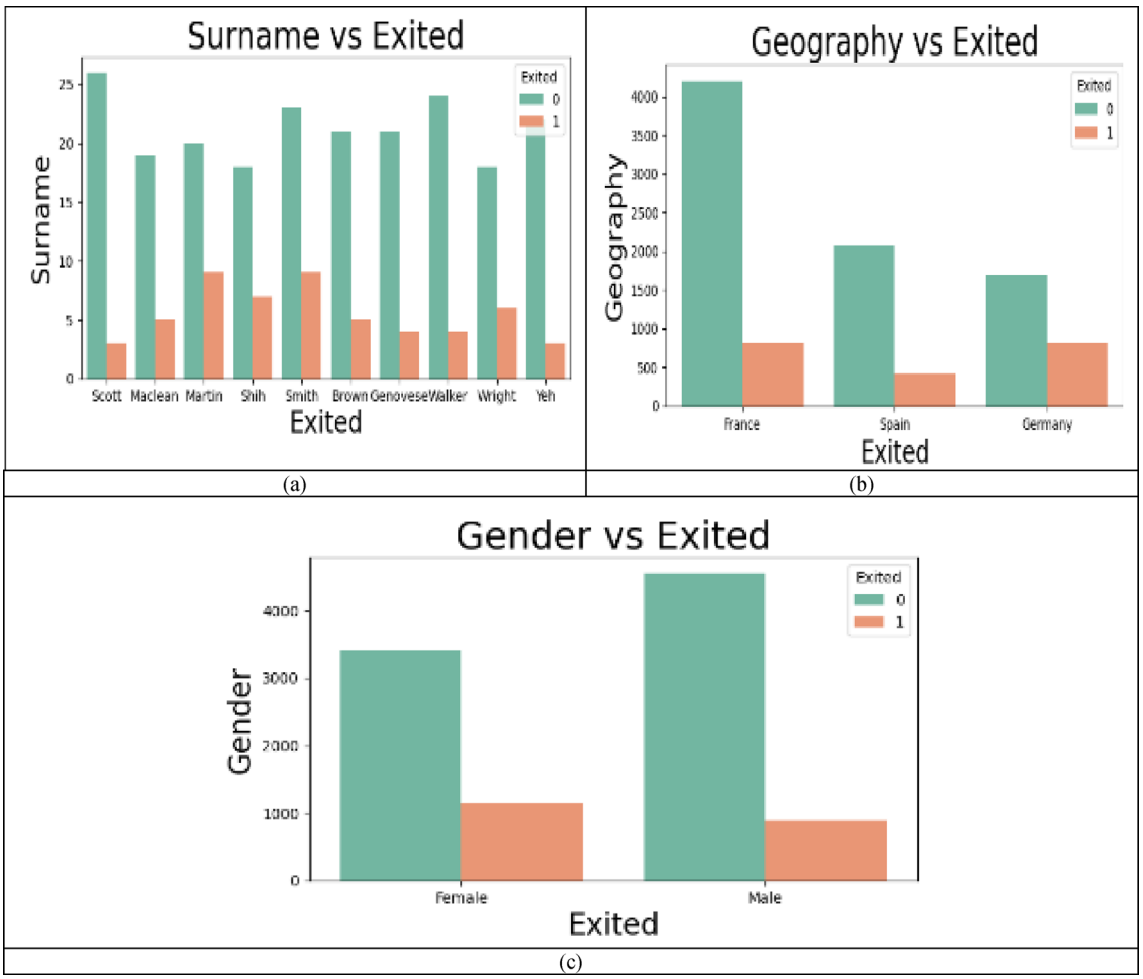


Fig. 6 Count Graph to explore the relationship between catagorial feature and customer graph (exited) **a** surname vs exited, **b** geography vs exited, **c** gender vs exited

3.4 Multi-dimensional analysis

Examining several variables concurrently in the next step of data analysis multi-dimensional analysis—allows one to probe intricate trends and linkages that might not be seen from univariate or bivariate analysis. This method helps one to have a more complete knowledge of the interactions among several elements causing client turnover. Utilizing scatter plots, analysts can visualize the associations between numerical variables, such as Credit Score against Balance, Age versus Estimated Salary, and Tenure versus the Number of Products. These scatter plots can be enhanced by color-coding the data points according to the Exited feature, which indicates whether a customer has churned. To examine the correlations between particular pairs of numerical features in respect to the goal variable, Exited, scatter plots are created indicating customer turnover. Three particular pairs—Credit Score vs. Balance, Age vs. Estimated Salary, Tenure vs. NumOf Products—have the emphasis of the study. Every scatter plot depicts how these variables interact with churn by color coding differentiating those who departed from those who stayed. This targeted analysis reveals patterns and potential correlations—such as whether age impacts predicted salary or whether better credit ratings are correlated with larger balances. These kinds of disclosures enable strategies to tackle factors generating turnover and increase client retention. Figure7 shows the scatter plots, analysts can visualize the associations between numerical variables, such as Credit Score against Balance, Age versus Estimated Salary, and Tenure versus the Number of Products. Figure 7(a) shows scatter plot labeled "CreditScore vs Balance." meant to show the relationship between credit score and balance for individuals, categorized by whether they exited (left) or didn't exit (stayed). Credit score is ranging from about 400 to 850 where as Balance is ranging from 0 to 250,000. Green dots represent those who stayed (Exited = 0) and orange dots for those who left (Exited = 1). The plot is dense with data points across the entire range of both axes, without a clear pattern. Overall, it suggests no obvious correlation between credit scores and balances in relation to the exit status. Figure7(b) shows a scatter plot, titled "Age vs Estimated Salary," maps individual data points with age on the x-axis and estimated salary on the y-axis. Notably, a dense cluster forms in the middle age range (around 30 to 60 years) and across a broad salary spectrum. Figure 7(c) shows, separated by gender and exit status, the link between tenure and the quantity of products. The number of products is shown horizontally; the tenure values run vertically. Different lines show gender groups; green and orange dots mark retention and exit status. Notably, about 1,000 people left; men had better retention rates than women.

The illustration demonstrates a scatter plot headed "Tenure vs. NumOf Products." "Exited," the x-axis labels; "Gender," the y-axis labels. With an extra classification by gender, the figure seems to indicate the association between the number of products kept by consumers (NumOfProducts) and their tenure. The x-axis spans 0 to 10; the y-axis values run from 1 to 4. The top left corner's legendary color coding for the "Exited" state is shown. Creating a correlation matrix for the dataset lets one assess the interactions of the numerical elements. Selecting columns of type float64 and int64 computes the correlation matrix, therefore determining the degree of linear relationship between pairs of features. The resultant heat map shows these interactions using colors suggesting their strength and direction and annotations for clarity denoting their values. This study aids in the identification of likely strongly connected characteristics both with one another and with regard to the aim variable, Exited. Development of predictive models targeted at addressing feature selection and customer attrition rely on an awareness of these interactions. Figure 8 displays the numerical feature correlation matrix for the dataset thereby enabling an extensive assessment of their interactions.

Developing a correlation matrix for the dataset enables one to assess the interactions of the numerical elements. Selecting columns of type float64 and int64 computes the correlation matrix, therefore determining the degree of linear relationship between pairs of features. Figure 8 shows the resultant heat map shows in these interactions using colors suggesting their strength and direction and annotations for clarity denoting their values. This study aids in the identification of likely strongly connected characteristics both with one another and with regard to the aim variable, Exited. Development of predictive models targeted at addressing feature selection and customer attrition rely on an awareness of these interactions. The bar chart in Fig. 9a shows the churn rate by age group, revealing how likely different age groups are to leave. The age groups range from 0–30 up to 80–90. Interestingly, those aged 50–60 seem the most likely to churn, with the highest bar. Contrastingly, the youngest (0–30) and the oldest (80–90) have the lowest churn rates. This insight can be vital for targeting customer retention strategies. Under three countries—France (about 0.16), Germany (about 0.32), and Spain (about 0.18)—Fig. 9b, "Churn Rate by Geography," shows the churn rates. Germany shows the highest employee turnover, implying notable regional variations. This graphic offers insightful analysis of retention of clients dynamics in several markets, which motivates more research on fundamental

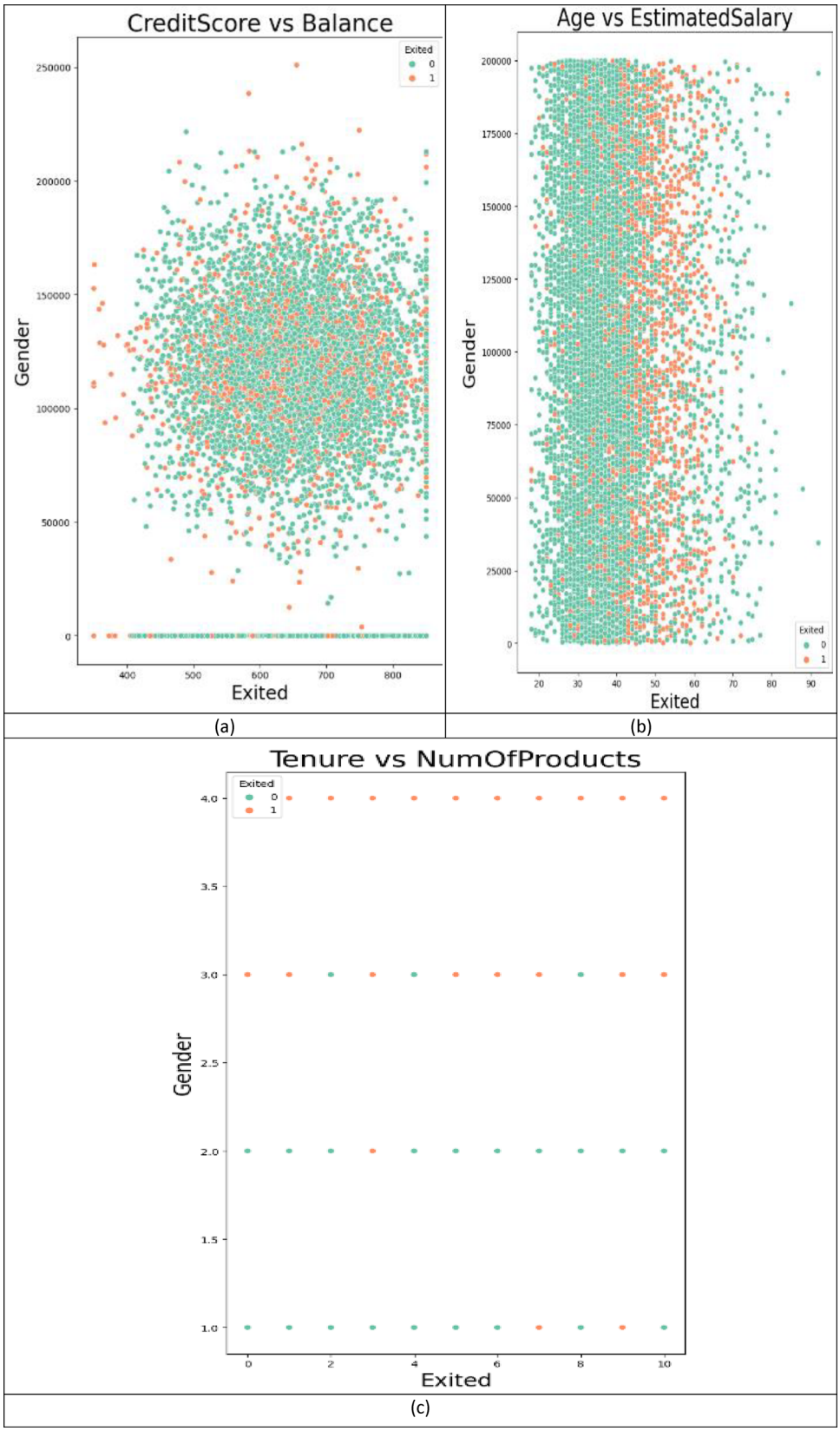


Fig. 7 Shows the scatter plots, analysts can visualize the associations between numerical variables, **a** Credit Score against Balance, **b** Age versus Estimated Salary, **c**Tenure versus the Number of Products

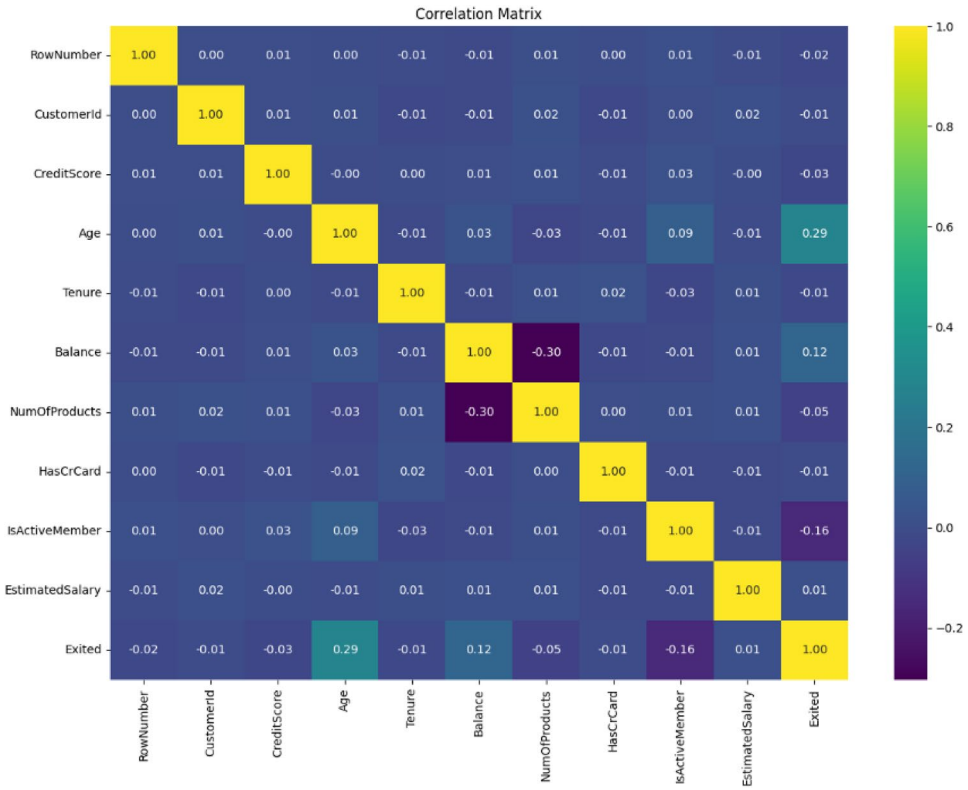


Fig. 8 Correlation Matrix for bank churn customer

causes and trends affecting these rates. Figure 9c shows the "Churn Rate by Gender" bar chart, which shows the rate of turnover among many sexes. The data shows that, based on the bar height of 0.25, about 25% of women had left; compared to about 15% for men at 0.15, which suggests This draws attention to a clear variance in churn rates, which motivates more research into the fundamental causes of this variation. Figure 9d presents the bar chart titled "Churn Rate by Number of Products," highlighting the correlation between customer churn rates and product ownership. The churn rate is approximately 25% for 1 product, decreases to 5% for 2 products, but surges to 80% for 3 products and 100% for 4 products, suggesting that increased product ownership significantly raises churn likelihood.

The boxplot shown in Fig. 10a is titled "Churn Rate by Balance." It looks at customer balances to determine if they have exited or not. Balance on y-axis is ranging from 0 to 250,000. For customers who haven't exited (0), the median balance is about 100,000, with most balances between roughly 25,000 and 150,000. For customers who have exited (1), the median balance is higher, around 150,000, with balances mostly between about 50,000 and 200,000. Seems like higher balances are more common among those who have exited. The box plot "Churn Rate by Estimated Salary" illustrates similar median salaries for retained and exited customers, suggesting that estimated salary may not have a major impact on banking turnover among consumers shown in 10(b).

3.5 Finding outliers

Apart from the objective variable, the function produces box graphs for every numerical feature in the dataset. Figure 11 shows box graphs help to give understanding of the distribution of the data by precisely showing important statistical characteristics including the median, quartiles, and probable outliers for every feature. This study provides information that could guide subsequent modeling and feature selection processes and helps identify any anomalies or trends influencing customer churn. The configuration is aimed to allow the numerical columns, thereby ensuring a nice and orderly presentation of the results. This boxplot shown in Fig. 11a "Boxplot of Row Number," lays out the distribution of row numbers in a dataset. Box Spans from about 2500 to 7500, capturing the middle 50% of the data. Median Line sits at around 5000, showing the central value. Whiskers stretch from 0 to 10,000, representing the full range of data. Similarliy Fig. 11b–j shows the boxplots of credit score, age, tenure, balance, number of products, credit card, active member, estimated salary, credit for row number respectively. Out of these Fig. 11c, d, g have outliers in credit score, age, number of products which will be removed in next phase. For examble A boxplot titled "Boxplot of Age" is shown in Fig. 11d where

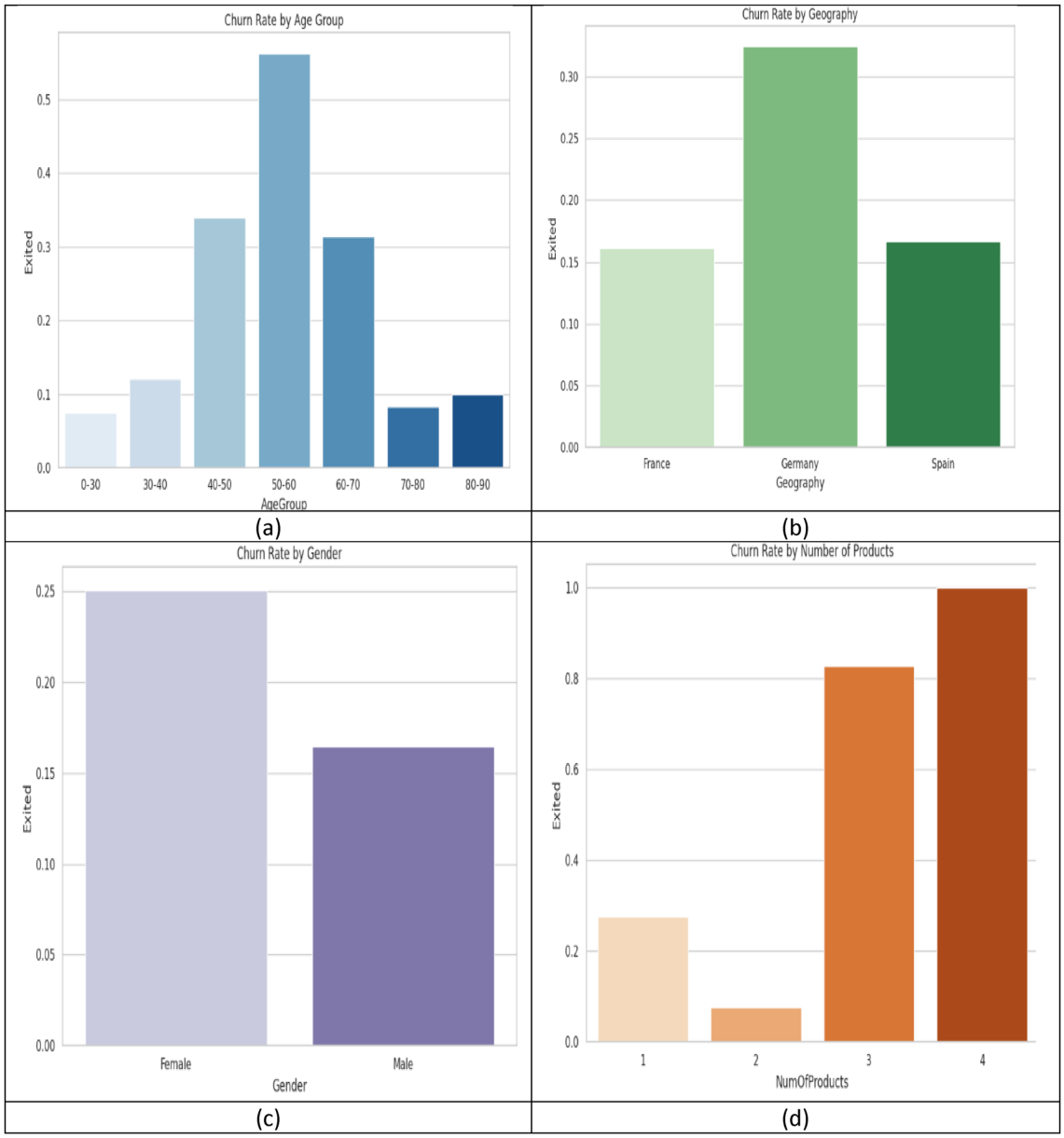


Fig. 9 Bar Chart **a** Churn rate by age, **b** Churn rate by geography, **c** Churn rate by gender, **d** Churn rate by Num of Products

the box captures the interquartile range (IQR) of the data, stretching from around 30 to 45. The median age sits around 37, shown by the line inside the box. Whiskers extend from about 20 to 60, pointing to the main range of ages. There are outliers above 60, with the highest touching 90.

3.6 Handling outliers by interquartile range (IQR) method

Apart from the objective variable, the function generates box graphs for every numerical aspect in the data. By exactly displaying significant statistical features including the median, quartiles, and likely outliers for every feature, the box graphs aid to provide comprehension of the distribution of the data. This study provides data that could guide next modeling and feature selection activities and helps identify any anomalies or trends influencing customer churn. The design is aimed to allow the numerical columns, so providing a good and orderly presentation of the findings. In statistics, the Interquartile Range (IQR) approach is a method for identifying dataset outliers. Apply the IQR approach first after determining the first quartile (Q1) and the third quartile (Q3) of the data, which consequently show the 25th and 75th

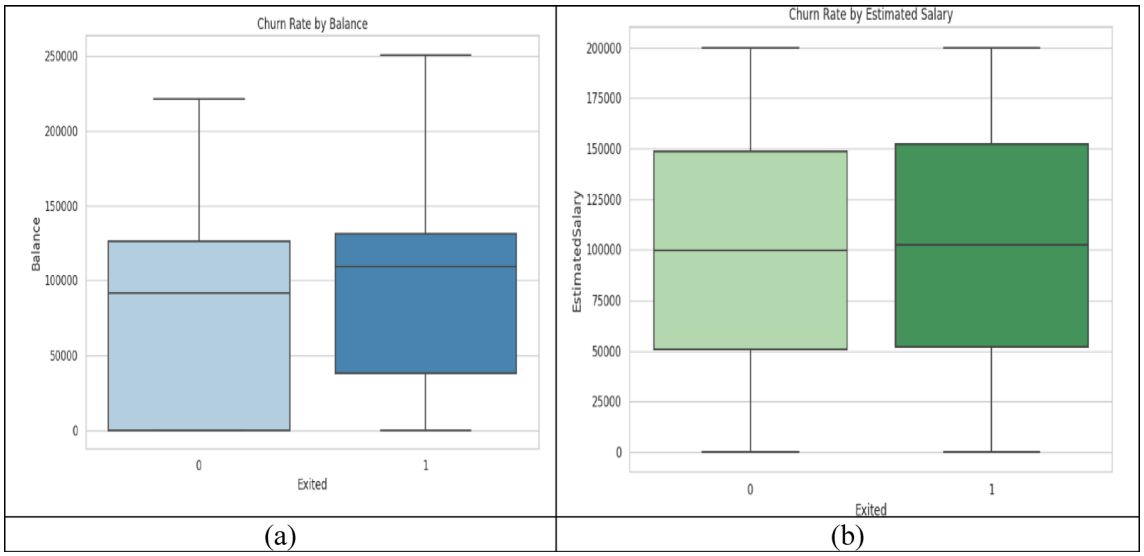


Fig. 10 Churn Rate **a** by Balance, **b** by Estimated Salary

percentiles, respectively. Subtracting Q1 from Q3 then helps one find the IQR. Considered outliers are any data points either below the lower bound or over the upper bound. This approach facilitates the identification of dataset abnormalities, therefore enabling improved data analysis and modeling. Figure 12a, c, e, shows the boxplot of credit score, age, num of products respectively before applying interquartile range (IQR) and Fig. 12b, d, f shows their boxplots after applying interquartile range in which outliers are removed.

3.7 Voting classifier

Here six different classifiers i.e. random forest, decision tree, k-nearest neighbors,, support vector XGBoost and voting classifier are used in machine learning to identify bank churn clients depending on important criteria. Highly interpretable models called decision trees divide data according on the most important factors, therefore facilitating the understanding of consumer behavior patterns resulting in attrition. nevertheless they can be prone to overfitting—especially in complicated data. Perfect for high-dimensional data, Support Vector Classifier is a strong method aimed to maximize the distance between the two classes, hence separating churners from non-churners. Even if it can struggle with big datasets and noisy data, K-nearby Neighbors is a basic, non-parametric strategy that organizes consumers depending on the behavior of their surrounding neighbors. For churn prediction activities with a combination of categorical and numerical information, advanced models like as XGBoost (Extreme Gradient Boosting) provide better accuracy by forming an ensemble of decision trees, where each tree corrects the errors of the preceding one, thus greatly efficient. Ultimately, Voting Classifier averages the predictions or uses majority voting to combine the strengths of several models decision trees, SVC, KNN, XGBoost therefore offering a more balanced and strong method of churn detection. Using performance criteria including accuracy, precision, recall, and F1-score, banks may compare and tune these algorithms to more precisely find possible churners and create focused retention plans.

Voting classifier is an ensemble machine learning method intended to increase general model accuracy and performance by aggregating the predictions of several independent classifiers. Support Vector Classifiers, Random Forests, and Decision Trees taken together create a more robust prediction model. A voting classifier aggregates predictions using two major approaches: soft and hard voting. The class with most votes from the individual classifiers is the final prediction in hard voting. This method is effective when the classifiers exhibit comparable performance standards. Conversely, soft voting evaluates the predicted probability for every class to select the one with the best average among all the classifiers. Combining the forecasts from several classifiers, a voting classifier generates a last prediction. Two ways of operation are hard voting and soft voting.. In hard voting, every classifier produces a prediction—a class label; the class with the most votes is selected as the final prediction. Hard voting is modeled by Eq. 1; yˆis the expected label. Though the classifiers generate probabilities for every class, under soft voting the class with the highest average probability is selected as the final prediction. The soft

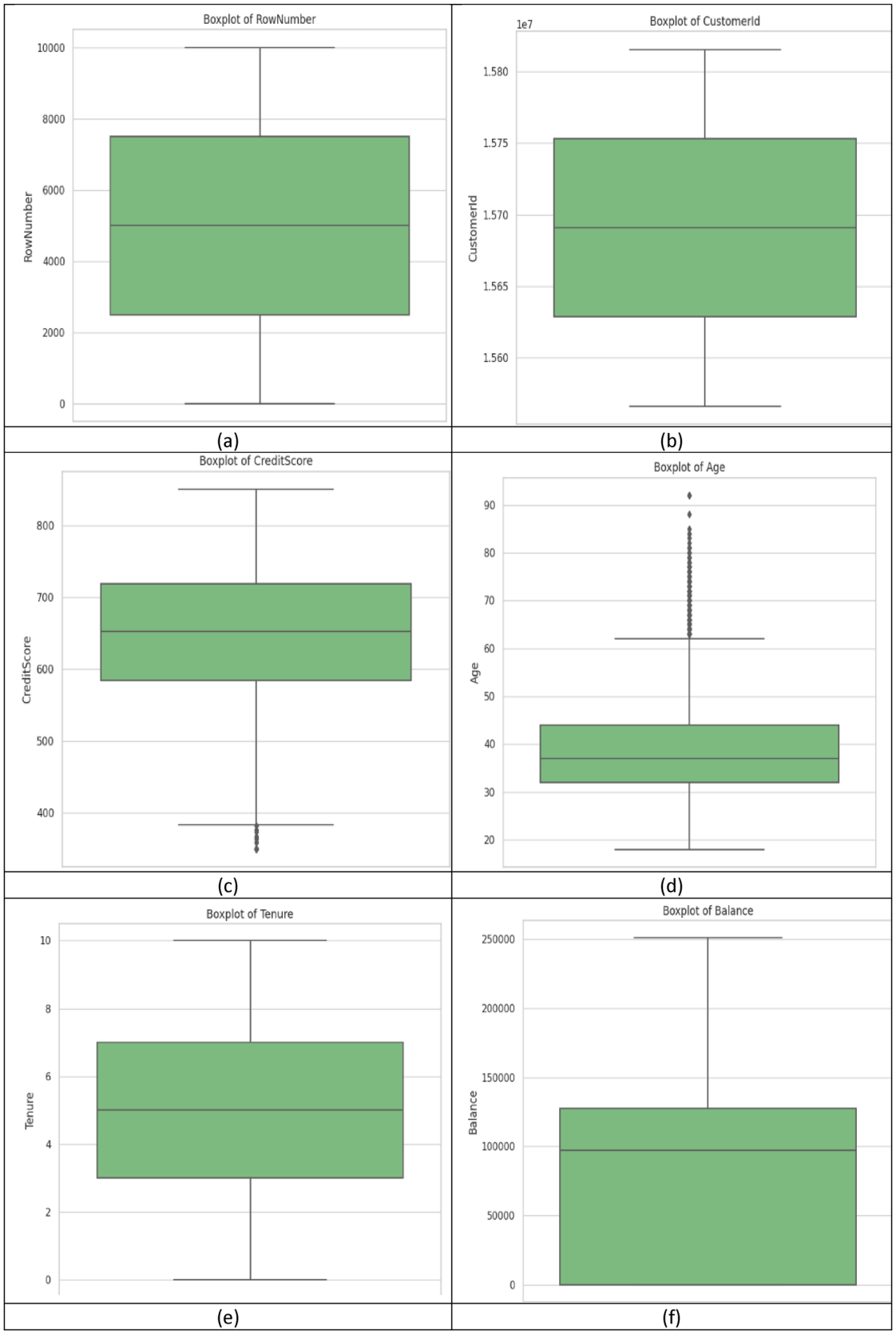


Fig. 11 Gives the information about Boxplot of **a** Row number, **b** Customer-id, **c** Credit score, **d** Age, **e** Tenure, **f** Balance, **g** Num of Products, **h** Has cr Card, **i** Active Member, **j** Estimated salary

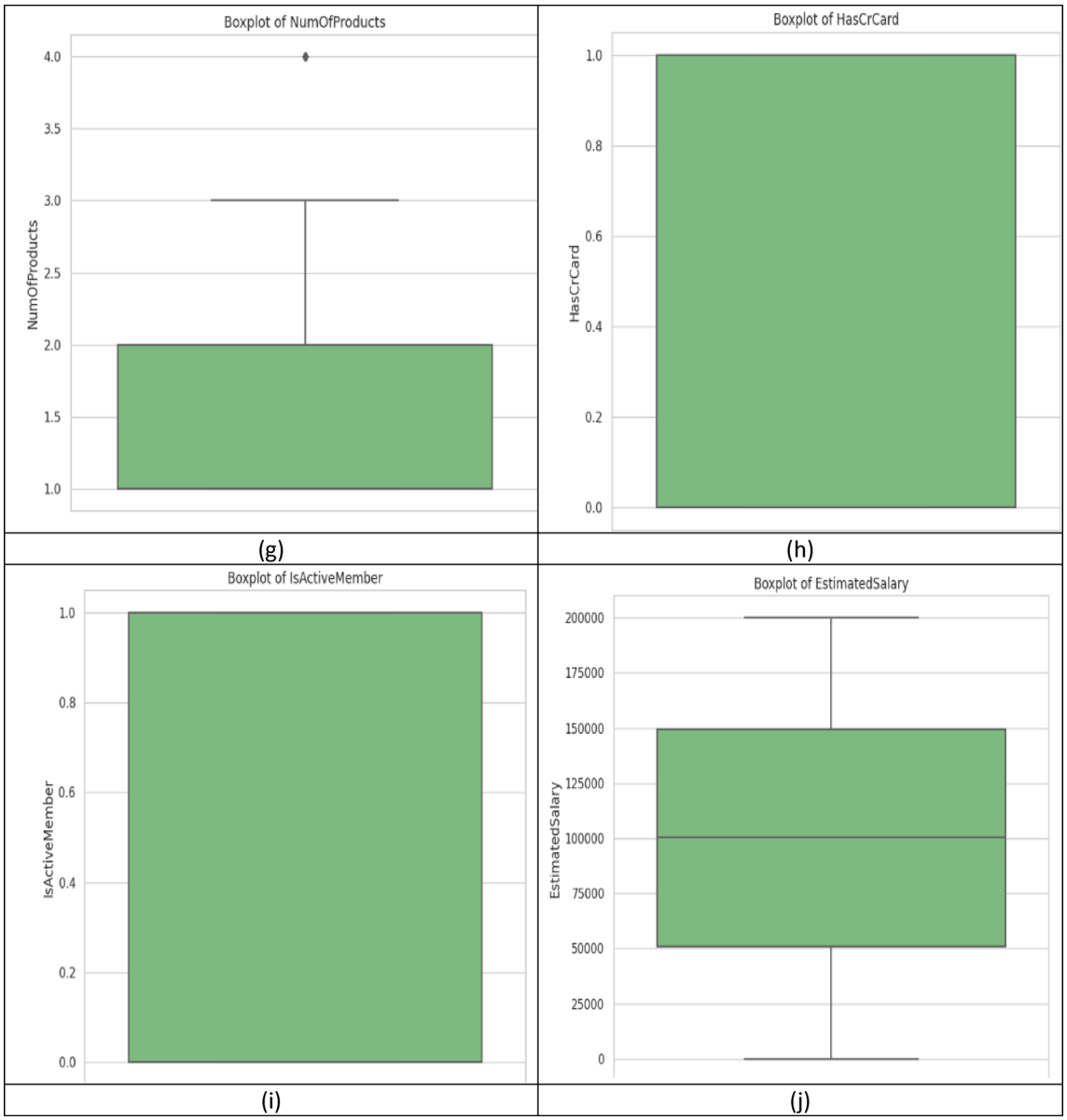


Fig. 11 (continued)

voting equation is Eq. 2. Pi (c) is the anticipated probability of class using m classifiers; argmaxc decides on the class with the best average probability.

$$\hat{y} = (\hat{y}_1 \hat{y}_2 \ldots \hat{y}_n) \tag{1}$$

$$\hat{y} = argmax_c\left(\frac{1}{m}\sum_{i-1}^{m} P_i(c)\right) \tag{2}$$

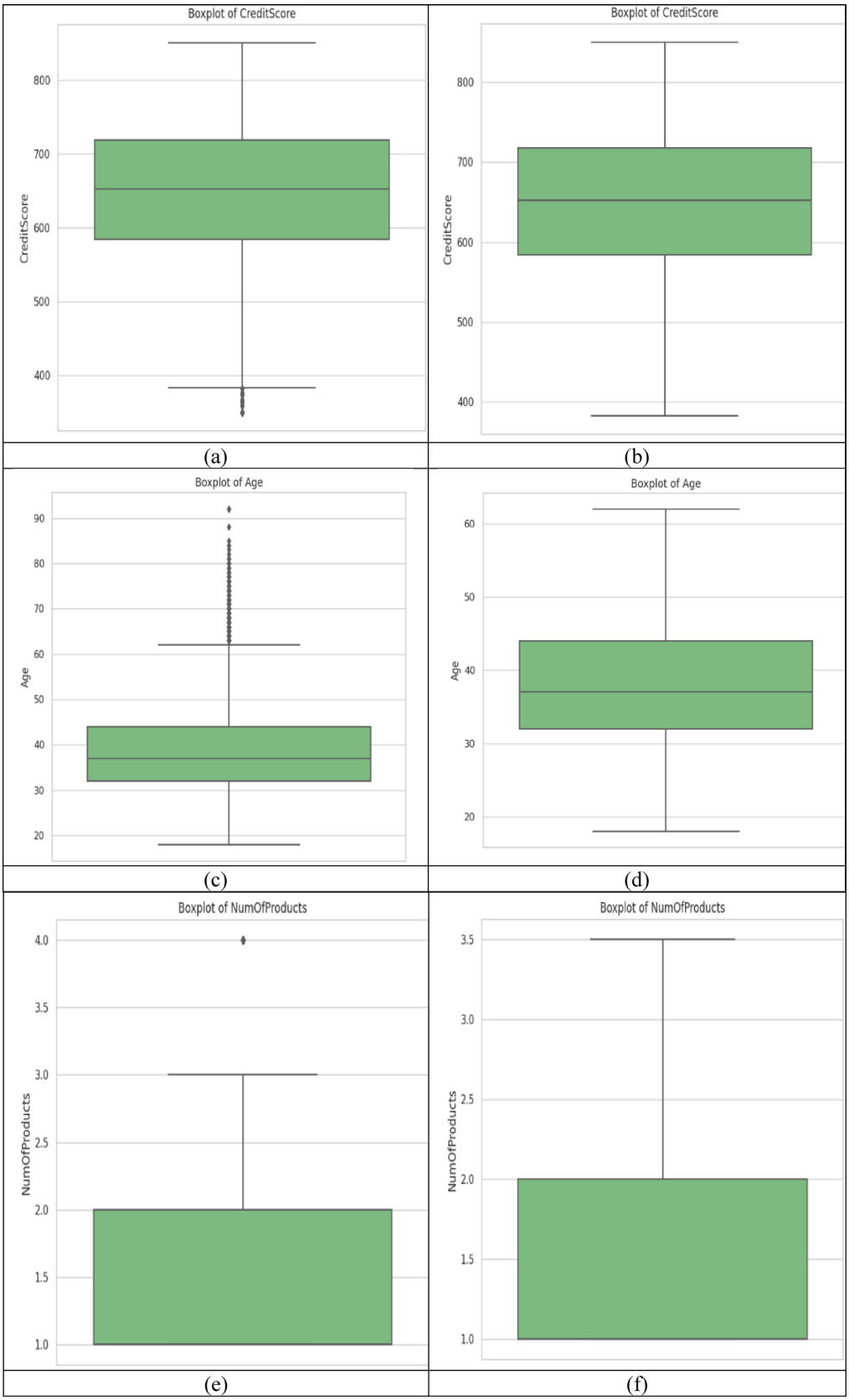


Fig.12 Boxplot **a** Credit score before IQR, **b** Credit score after IQR, **c** Age before IQR, **d** Age after IQR, **e** Num of preducts before IQR, **f** Num of products after IQR

4 Results and discussions

In this study, we implemented five individual classifiers—Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), and XGBoost—alongside an ensemble Voting Classifier to predict bank customer churn. Each classifier was trained and evaluated separately on the dataset to understand their individual contributions to the churn prediction task. The Voting Classifier was then applied to combine the predictions from these five models, leveraging their collective strengths to enhance overall performance. By aggregating the outputs of diverse algorithms, the Voting Classifier provided a more balanced and accurate prediction, mitigating the weaknesses of any single model. This approach allowed us to achieve significant improvements in classification metrics such as accuracy, precision, recall, and F1-score, ensuring a robust solution for identifying at-risk customers.

4.1 Evaluation metrics

In this research, six performance parameter have been used for the evaluation of different models and proposed model for the bank churn customer prediction. The different parameters used are precision, recall, F1_score, accuracy, macro average, weighted _average whose definition are given below.

Precision: It is the ratio of precisely predicted positive observations to the overall predicted positives. It shows the actual correctness count among the predicted positives as shown in Eq. 3.

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP} \tag{3}$$

Recall: It is the ratio of all the precisely predicted positive observations to actual positive observations as shown in Eq. 4. It shows the model's positive instance classification strength.

$$\text{Recall} = \text{TP} / \text{TP} + \text{FN} \tag{4}$$

F1_Score: Combing the two measures, it is the harmonic mean of Precision and Recall as shown in Eq. 5. Dealing with imbalanced datasets is beneficial since it aggregates the false positives and false negatives.

$$\text{F1} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \tag{5}$$

Accuracy: It is the proportion of total correct predictions out of all predictions as shown in Eq. 6.

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{TN} + \text{FP} + \text{FN} \tag{6}$$

- Macro average: The average of the Precision, Recall, and F1-Score across all classes, treating all classes equally regardless of their support.
- Weighted average: It is the average of the Precision, Recall, and F1-Score across all classes weighted by the support of every class. It emphasizes even more the performance of the classes with more support.

4.2 Confusion matrices for different classifiers and proposed voting classifier

For customer prediction in the case of bank churn, a variety of classifiers which form the core of the machine learning algorithms are used in order to learn customer churn patterns. Beside the benefit of every classifier, there will also be drawback. The Decision Tree classifier is an easy to interpret model that finds behaviors that lead to churn using a tree of decisions but there is an overfitting problem. The decision tree model shown in Fig. 13a shows an accuracy of 77.40%, correctly classifying most cases. However, it misclassifies 264 false positives and 188 false negatives, indicating room for improvement. The Random Forest is much more robust as it is an ensemble method that uses multiple, iteratively constructed decision trees in order to make improved, generalized decisions that are also more accurate. The confusion matrix shown in Fig. 13b shows that the Random Forest classifier achieves an accuracy of 83.65%, with 1,443 true negatives, 230 true positives, 164 false positives, and 163 false negatives. This indicates the model is effective at predicting non-churn cases but has room for improvement in identifying churn customers. On the other hand, the Support Vector Machine (SVM) method also can enhance customers separation between churn and non-churn, through the construction

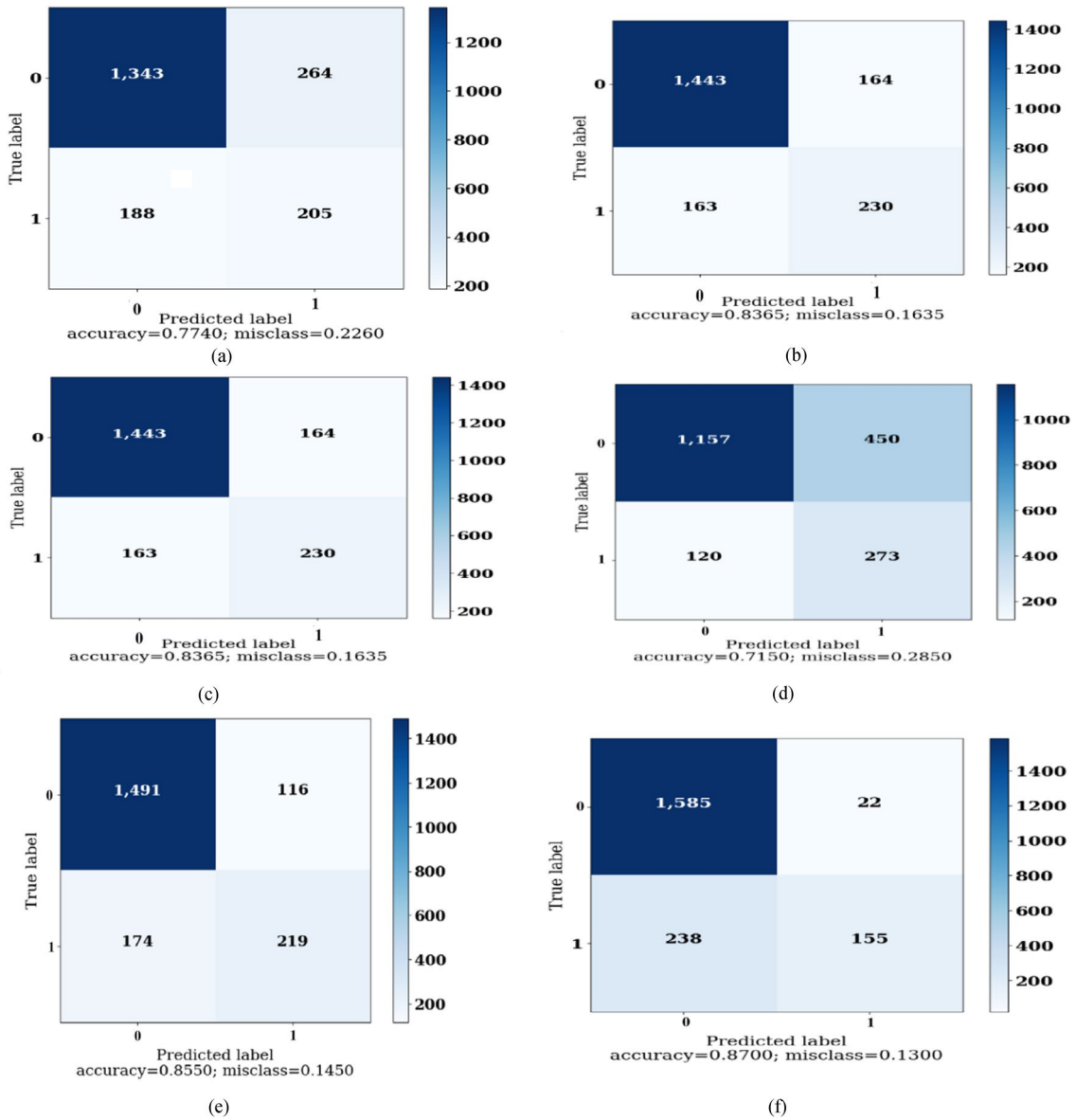


Fig. 13 Confusion Matrix For Different Classifier **a** Decision Tree, **b** Random Forest, **c** Support Vector Machine, **d** K-Nearest neighbor, **e** XGBoost, **f** Voting Classifier

of hyper-plane in features space, but may encountered an issue when exploring big data due to scalability problem. The confusion matrix shown in Fig. 13c for the SVM classifier shows an accuracy of 83.65%, with 1,443 true negatives, 230 true positives, 164 false positives, and 163 false negatives. This indicates that the model performs well in predicting non-churn customers but could improve in detecting churn cases. The same is true of the K-Nearest Neighbors (KNN) classifier, which decides on churn based on proximity in the feature space; this method also degrades as the number of customers increases. The confusion matrix shown in Fig. 13d for the KNN model has an accuracy of 71.50%, indicating it correctly classifies most cases, but there's room for improvement in reducing the false positives (450) and false negatives (120). Enhancing the model could boost performance significantly.Therefore, XGBoost is a Gradient Boosting algorithm widely used by applying multiple passes of learning in a single-stage tree model to refine the results of an actual prediction model to boost the results of an imprecise method used in bank churn datasets which face problems like missing values and imbalanced data. The confusion matrix shown in Fig. 13e for the XGBoost model has an accuracy of 71.50%, indicating it correctly classifies most cases. However, there's potential for improvement, particularly in reducing false positives (450) and false negatives (120), to boost overall performance. However, it's seen that the highest accuracy belongs to Voting Classifier where the model choice is made by voting together all models which can be Random Forest, XGBoost

Classifier, SVM etc. The confusion matrix for the Voting Classifier shows higher precision as well as recall values for both churn and non churn classes. The confusion matrix shown in Fig. 13f for the voting classifier achieves the best performance with an accuracy of 88.00%, effectively minimizing false positives (20) and false negatives (214). This highlights its superiority in accurately classifying most cases compared to other models. This has the advantage of overpowering various weaknesses of the individual models, but at the same time suggesting how banks can apply effective customer churn retention strategies through formulating a model from the overall results of the individual models.

4.3 Proposed voting classifier with SMOTE

To address class imbalance in the banking churn dataset, we apply SMOTE (Synthetic Minority Over-sampling Technique) to increase the minority class, "churned" customers, from 2,037 to 8,000. SMOTE generates synthetic samples by analyzing and interpolating between existing minority class data points, which helps prevent model bias. The research work started by splitting the data into features (X) and the target variable (y), then applying SMOTE only to the training set to avoid data leakage. This balanced dataset enables the model to better detect churned customers, improving accuracy and supporting more effective customer retention strategies in business decision-making.

Confusion matrix as seen in Fig. 14 provides important new perspectives on model performance in customer churn prediction. With a 90% accuracy the model shows general dependability by correctly recognizing 345 churn cases and 1453 non-churn cases. But it also produces 124 false positives, meaning non-churn consumers are mistakenly categorized as churn and could result in unwarranted retention policies. There are also 48 false negatives, suggesting some real churn cases go missing and could cause income loss. The precision and recall measures for Class 1, or churn, show that although the algorithm correctly points out those consumers most likely to leave others are nevertheless missed. Improving recall to reduce false negatives would help the model be more effective for customer retention strategies since it will ensure less missed situations of possible turnover. Reducing false positives and false negatives will increase the utility of the model in a company environment by harmonizing effective resource allocation with effective churn detection. By further improving predictions and maybe reducing false negatives, changing model thresholds or adding additional features could assist to improve the general influence of the model on customer retention initiatives.

Accuracy measures, recall, F1-score, over precision a well-balanced model performance show on the classification report. With a 0.88 for Class 0—non-churn—the model correctly projects 88% of projected non-churn customers. The model rightly notes 92% of actual non-churn events with a 0.92 recall. While recall is 0.89, so capturing 89% of the genuine churn cases for Class 1 consumers, precision is better at 0.92, so showing 92% accuracy in recognizing churn cases. Reflecting a solid mix between accuracy and recall, both classes have an F1-score of 0.90; the overall accuracy on the 3,504 instance dataset is 90%. Independent of class balance, consistent performance stands emphasized with all the macro and weighted averages for accuracy, recall, and F1-score—0.9. While the high recall for non-churn (Class 0) shows effectiveness in identifying consumers most likely to stay, the great accuracy for churn (Class 1) reduces false positives and promotes targeted measures. Using minimum misclassification, the

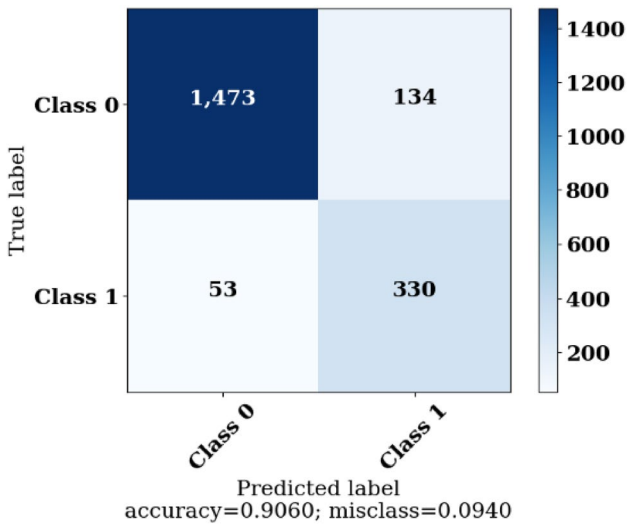


Fig. 14 Confusion matrix for voting Classifier with SMOTE

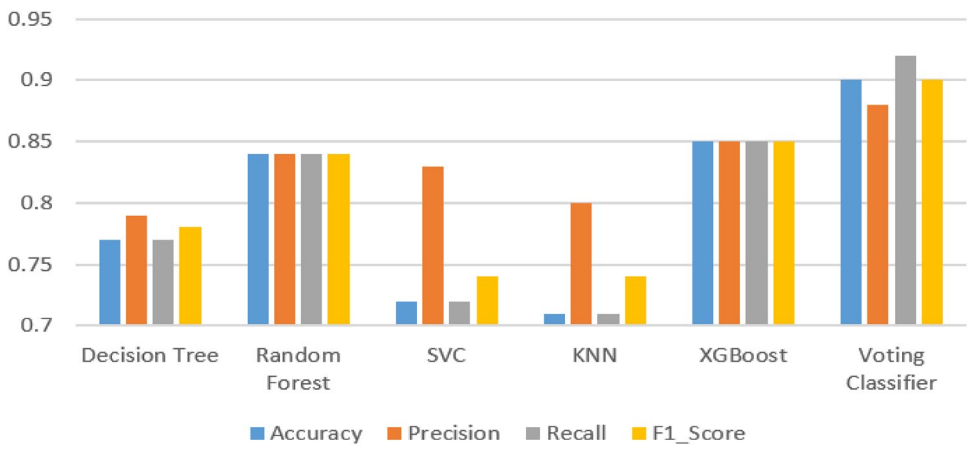
Fig. 15 Performance matrix for voting Classifier with SMOTE

	precision	recall	f1-score
0	0.88	0.92	0.90
1	0.92	0.89	0.90
accuracy			0.90
macro avg	0.90	0.90	0.90
weighted avg	0.90	0.90	0.90

Table 1 Performance comparison of different classifiers with proposed model

	Precision	Recall	F1_score	Accuracy
Decision tree	0.79	0.77	0.78	0.77
Random forest	0.84	0.84	0.84	0.84
Support vector machine	0.83	0.72	0.73	0.72
K-nearest neighbors	0.80	0.71	0.74	0.71
Xgboost	0.85	0.85	0.85	0.85
Voting classifier without smote	0.87	0.87	0.80	0.87
Proposed model (voting classifier with smote)	0.90	0.90	0.90	0.90

Fig. 16 Graphical analysis of different classifiers with proposed model



study provides a consistent model with excellent performance for churn prediction that could thus help initiatives aimed at client retention. Figure 15 shows the performance matrix.

4.4 Comparison of different classifiers based on performance parameters

The analysis from the Table 1 and Fig. 16 demonstrates that the addition of SMOTE to the Voting Classifier proposed in this paper provides for a higher precision, recall, F1-score and accuracy (90%) as compared to all the other models considered in this paper. This shows how class imbalance has been well handled using SMOTE. Other individual classifiers include XGBoost which was slightly close to the first-best performers having a mean-score of 0.85 across all scores. Compared to other algorithms, SVM and KNN has considerably average recall and accuracy which suggests that they are not so efficient in dealing with characteristics of this particular dataset. The Basic Voting Classifier has reasonable accuracy but is less accurate than the SMOTE model reiterating the significance of balance of datasets.

Table 2 State of Art Comparison on Banking Customer Churn Prediction Dataset with 10,000 Customers Data

Reference Number	Year	Proposed Methodology	Key Finding	Accuracy
[15]	2023	Machine learning algorithms (e.g., random forest, logistic regression)	Utilized multiple machine learning algorithms to predict customer churn	84.76%
[39]	2021	Comparison of classification models Decision Trees, SVM, Random Forest)	Compared the performance of different classification models for churn prediction	79.60%
[40]	2023	Various machine learning techniques	Focused on improving churn prediction accuracy in the banking industry using machine learning	81%
[6]	2022	Ensemble-based methods, data balancing	Improved churn prediction by incorporating data balancing and explainable machine learning models	86%
[41]	2023	Decision tree with genetic algorithm	Proposed a hybrid approach combining decision trees with genetic algorithms for better churn prediction	75.95%
[12]	2022	Machine learning algorithms, logistic regression, SVM	Applied various machine learning models to predict customer churn with focus on model interpretability	85%
Proposed model		Voting classifier	Combines the strengths of multiple classifiers to predict churn more accurately	90%

4.5 State of art comparison

The present state-of-the-art methods shown in Table 2 compares the efficiency of proposed ensemble Voting Classifier with other state of art models, which reaches the maximum accuracy of 90%. All the authors shown in Table 2 have worked on banking customer churn prediction dataset with 10,000 customer's data. The insights presented in this paper compare the methodologies utilised in customer churn prediction with an emphasis on the recent improvements within the presented approach and those employed in the field of ensemble methods. Another traditional method include Random Forest and Logistic Regression, which has gained significant performance consistently with an accuracy of 84.76% as shown by [15]. In the same way, while, Decision Trees and SVM models are interpretable models, they have proven to perform poorly, with [39] achieving 79.60% of accuracy. There has been progress in the methods, including ensemble-based methods and data balancing, as discussed in [6], which successfully deployed both to obtain 86 per cent accuracy. That is why it is necessary to consider the results related to the solutions of class imbalance problem and potential of using the combined approach.

More complex methodologies still present a curious output, by adopting for instance in [41] the Decision Tree with Genetic Algorithm, an 75.95% of accuracy is obtained, therefore there is not always that the application of additional stages will give better results. The desire for models to be more explainable and interpretable like in [12] ensures that between performance and practicality the accuracy is seen as reaching a level of 85%.

The Voting Classifier model proposed is more effective compared to these benchmarks, provided enhanced 90% by various classifiers with SMOTE balance the data set. Thus, one can note that the integration of the features that predict the choice of one or another travel product together with the elimination of the problem of data imbalance boosts performance. This paper clearly establishes a new benchmark especially for the unbalanced data in churn prediction problem while being both highly accurate and robust. It makes a contribution to the advancement of tools for predicting customers' behavior based on signals and improving the best techniques as well as introducing an ensemble point of view into the churn analysis.

5 Conclusion and future work

This research investigates consumer turnover in the banking sector using machine learning techniques—especially ensemble learning using a Voting Classifier. By leveraging a whole dataset including demographic, financial, and behavioral features, the computer was able to accurately identify customers at risk of leaving the bank. The study consisted in a powerful data pretreatment pipeline including one-dimensional, two-dimensional, and multi-dimensional explorations together with outlier detection utilizing the Interquartile Range (IQR) technique. The results obtained suggest that the Voting Classifier, if augmented with SMOTE for dealing with class imbalance, presents a substantial improvement over the baseline. The model without SMOTE showed accuracies of 87% and recall/ precision of 80%/87%. But integrating SMOTE seemed to enhance the performance significantly and the model achieved 90% of accuracy and balanced precision, recall, F1-Score being 90%. This has greatly improved showing how SMOTE enhances handling of class imbalance and increases the models likelihood to accurately predict customers that will churn.

From the results presented in this work, it can be seen that the presented ensemble approach, as well as using techniques like SMOTE, indicates a very accurate and stable solution for customer retention in the banking industry. The model's balance between accuracy and recall supports its practical application even more since it minimizes false positives and false negatives, which is crucial in a corporate environment where misclassification could result in either pointless intervention or lost chances for customer retention.

By more precisely identifying at-risk clients using the information from this approach, banks may apply focused retention plans to lower turnover and increase customer loyalty. Understanding the main elements influencing customer turnover—such as account balance, tenure, and activity level—allows banks to make data-driven decisions to solve problems and keep important customers. Ultimately, this study emphasizes the importance of machine learning—especially ensemble techniques—in tackling challenging corporate problems including client turnover. Future research could investigate other feature engineering approaches, the integration of more intricate models including deep learning, or an emphasis on interpretability to further improve the model's usability in real-world banking applications. Building on the results of this study, next research can investigate numerous directions to

improve the customer attrition prediction. First, more sophisticated methods to capture complicated, non-linear correlations in the data could be applied like deep learning models or neural networks.

Author contributions Ruchika Bhuria: Conceptualization, data collection, preprocessing, and analysis, manuscript drafting, and editing. Sheifali Gupta: Data analysis, feature engineering, visualization, and manuscript review. Upinder Kaur: Methodology development, validation of models, and interpretation of results. Salil Bharany: Project supervision, methodology design, model evaluation, and review of manuscript drafts. Ateeq Ur Rehman: Software implementation, model optimization, and statistical analysis; contributed to manuscript revision. Seada Hussen: Overall project administration, and final manuscript review; also responsible for addressing reviewer comments. Ghanshyam G. Tejani: Software implementation,model optimization, and statistical analysis Pradeep Jangir: Software implementation, validation of models, and interpretation of result.

Data availability The dataset used in this study is the "Bank Customer Churn Prediction Dataset," which is publicly available on Kaggle. You can access it through the following link: https://www.kaggle.com/datasets/saurabhbadole/bank-customer-churn-prediction-dataset.

Declarations

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Guliyev H, Yerdelen Tatoğlu F. Customer churn analysis in banking sector: Evidence from explainable machine learning model. J Appl Microeconometr. 2021;1(2):85–99.
2. Jain H, Yadav G, Manoov R. Churn prediction and retention in banking, telecom and IT sectors using machine learning techniques. In: Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019 Singapore: Springer Singapore, 2020, pp. 137–156).
3. Dalmia H, Nikil CV, Kumar S. Churning of bank customers using supervised learning. In: Innovations in Electronics and Communication Engineering: Proceedings of the 8th ICIECE Springer Singapore 2019, pp. 681–691
4. Tran H, Le N, Nguyen VH. Customer churn prediction in the banking sector using machine learning-based classification models. IJIKM. 2023. https://doi.org/10.28945/5086.
5. Rahman M, Kumar V. Machine learning based customer churn prediction in banking. In: 2020 4th international conference on electronics, communication and aerospace technology (ICECA) IEEE 2020, pp. 1196–1201.
6. Tékouabou SC, Gherghina ȘC, Toulni H, Mata PN, Martins JM. Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. Mathematics. 2022;10(14):2379.
7. Shukla A. Application of machine learning and statistics in banking customer churn prediction. In: 2021 8th International Conference on Smart Computing and Communications (ICSCC), IEEE. 2021, pp. 37–41.
8. Singh PP, Anik FI, Senapati R, Sinha A, Sakib N, Hossain E. Investigating customer churn in banking: a machine learning approach and visualization app for data science and management. Data Sci Manage. 2024;7(1):7–16.
9. Sagala NTM, Permai SD. Enhanced churn prediction model with boosted trees algorithms in the banking sector. In: 2021 International Conference on Data Science and Its Applications (ICoDSA) IEEE, 2021, pp. 240–245.
10. Masrom S, Septiyanti R, Ahmad A, Rahman RA, Sulaiman N. Analysis of machine learning in classifying bank profitability with corruption factor. J Adv Res Appl Sci Eng Technol. 2024;40(2):13–21.
11. Vu VH. An efficient customer churn prediction technique using combined machine learning in commercial banks. In: Operations research forum (Vol. 5, No. 3, p. 66). Cham: Springer International Publishing. 2024.
12. Huseyinov I, Okocha O. A machine learning approach to the prediction of bank customer churn problem. In: 2022 3rd International Informatics and Software Engineering Conference (IISEC), IEEE, 2022, pp. 1–5.
13. Soni PK, Nelson L. PCP: profit-driven churn prediction using machine learning techniques in banking sector. Int J Perform Eng. 2023;19(5):303.
14. Asfaw T. Customer churn prediction using machine-learning techniques in the case of commercial bank of Ethiopia. Sci Temper. 2023;14(03):618–24.

15. Hui SH, Khai WK, XinYing C, Wai PW. Prediction of customer churn for ABC Multistate Bank using machine learning algorithms/Hui Shan Hon... [et al.]. Malaysian Journal of Computing (MJoC), 2023;8(2):1602–1619.
16. Lukita C, Bakti LD, Rusilowati U, Sutarman A, Rahardja U. Predictive and analytics using data mining and machine learning for customer churn prediction. J Appl Data Sci. 2023;4(4):454–65.
17. Khandelwal, V., 2023, June. Customer churn prediction in banking sector using PCA with machine learning algorithms. In AIP Conference Proceedings (Vol. 2782, No. 1). AIP Publishing.
18. Chen P, Liu N, Wang B. Evaluation of customer behaviour with machine learning for churn prediction: The case of bank customer churn in europe. In: Proceedings of the International Conference on Financial Innovation, FinTech and Information Technology, FFIT 2022, October 28–30, 2022, Shenzhen, China. 2023.
19. Murindanyi S, Mugalu BW, Nakatumba-Nabende J, Marvin G. Interpretable machine learning for predicting customer churn in retail banking. In: 2023 7th International conference on trends in electronics and informatics (ICOEI) (). 2023, pp. 967–974, IEEE.
20. Simsek M, Tas IC. A classification application for using learning methods in bank costumer's portfolio churn. J Forecast. 2024;43(2):391–401.
21. Talwadia V, Jain SK, Chanchawat L, Pepeti SK. An integrated bank customer and credit card holder churn/no churn analysis system using machine learning. Int Res J Innovat Eng Technol. 2023;7(5):114.
22. Hambali MA, Andrew I. Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis. Qeios, 2024
23. Wang S, Chen B. Credit card attrition: an overview of machine learning and deep learning techniques. Информатика Экономика Управление/Informatics Economics Management. 2023;2(4):0134–44.
24. Al-Sultan SY, Al-Baltah IA. An improved random forest algorithm (ERFA) utilizing an unbalanced and balanced dataset to predict customer churn in the banking sector. IEEE Access. 2024. https://doi.org/10.1109/ACCESS.2024.3395542.
25. Gkonis V, Tsakalos I. Deep dive into churn prediction in the banking sector: the challenge off hyperparameter selection and imbalanced learning. J Forecast. https://doi.org/10.1002/for.3194
26. Bhaal N, Adarsh A, Awasthi P, Usha G. A comparative framework for Churn analysis in banking and telecom sector. In: AIP Conference Proceedings (Vol. 3075, No. 1). AIP Publishing. 2024.
27. Kaya H. Using Machine Learning Algorithms to Analyze Customer Churn with Commissions Rate for Stocks in Brokerage Firms and Banks. Bitlis Eren Üniversitesi Fen Bilimleri Dergisi. 2024;13(1):335–45.
28. Zhang N, Zheng Y, Duan C. Bank customer churn prediction based on random forest algorithm. In: Proceedings of the 5th International Conference on Computer Information and Big Data Applications 2024, pp. 1031–1035.
29. Gurung N, Hasan MR, Gazi MS, Chowdhury FR. AI-based customer churn prediction model for business markets in the usa: exploring the use of ai and machine learning technologies in preventing customer churn. J Comp Sci Technol Stud. 2024;6(2):19–29.
30. Baby B, Dawod Z, Sharif MS, Elmedani W. Customer churn prediction model using artificial neural network: a case study in banking. In: 2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) 2023, IEEE, pp. 154–161.
31. Raj R, Gupta S, Mishra R, Malik M. Efficacy of customer churn prediction system. In: 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT), IEEE, 2024, pp. 546–553.
32. Galal M, Rady S, Aref M. Enhancing machine learning engineering for predicting youth loyalty in digital banking using a hybrid meta-learners. Int J Intell Comput Info Sci. 2024;24(2):28–40.
33. Poudel SS, Pokharel S, Timilsina M. Explaining customer churn prediction in telecom industry using tabular machine learning models. Mach Learn Appl. 2024;17: 100567.
34. Kim WJ, Ahn JJ, Oh KJ. Prediction of bank outstanding customers churn by panel data size. Quantitative Bio-Science. 2023;42(2):65–73.
35. Kharat SS, Rane CV. web application for banking churn prediction using ANN. In: XVIII International Conference on Data Science and Intelligent Analysis of Information. Cham: Springer Nature Switzerland 2023. pp. 621–629.
36. Manzoor A, Qureshi MA, Kidney E, Longo L. A review on machine learning methods for customer churn prediction and recommendations for business practitioners. IEEE Access. 2024. https://doi.org/10.1109/ACCESS.2024.3402092.
37. Peng K, Peng Y, Li W. Research on customer churn prediction and model interpretability analysis. PLoS ONE. 2023;18(12): e0289724.
38. Vu VH. Predict customer churn using combination deep learning networks model. Neural Comput Appl. 2024;36(9):4867–83.
39. Gavielidou, C., 2021. Big data analytics in banks: Comparison of classification models in predicting customers churn.
40. Charandabi SE. Prediction of customer churn in banking industry. 2023. arXiv preprint arXiv:2301.13099.
41. Li LJ, Junn KY. Decision tree with genetic algorithm for bank customer churn prediction. In: 2023 IEEE 21st Student Conference on Research and Development (SCOReD), IEEE 2023, pp. 426-431.