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

A PROJECT REPORT

Submitted by,

Mr. Nishchay B N – 20201ISE0086

Mr. Vikas H– 20201ISE0056

Mr. Sridar S –20201ISE0093

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY
BENGALURU
2024

ABSTRACT

Customer attrition in the banking industry occurs when consumers cease using the goods and services offered by the bank for a significant period and, eventually, terminate their relationship with the bank altogether. This phenomenon poses a substantial challenge for banks, making customer retention paramount in today's fiercely competitive banking market. Retaining customers not only stabilizes the bank's revenue stream but also plays a vital role in attracting new customers. A strong and loyal customer base fosters trust and generates positive word-of-mouth referrals, which are invaluable for acquiring new clientele. Consequently, reducing client attrition is a critical objective that banks must prioritize.

In our research, we undertake a comprehensive examination of bank data to predict which users are most likely to discontinue using the bank's services and subsequently become inactive or leave the bank. We employ a variety of sophisticated machine learning algorithms to analyze this data. Our approach includes performing a comparative analysis across different evaluation metrics to determine the most effective predictive models. This rigorous analysis enables us to identify patterns and trends that indicate potential customer attrition.

To facilitate this analysis and make it accessible for data science and management purposes, we have developed an advanced Data Visualization RShiny app. This application is designed specifically for customer churn analysis, providing an intuitive and interactive platform for exploring the data. By utilizing this app, bank managers and data scientists can visualize key trends and insights, making it easier to identify customers who are at risk of leaving.

Analyzing customer data through these sophisticated methods allows the bank to gain a deeper understanding of the factors contributing to attrition. With this knowledge, the bank can implement targeted strategies aimed at retaining customers on the verge of leaving. Proactive measures, informed by data-driven insights, can significantly enhance customer satisfaction and loyalty, ultimately reducing attrition rates and fostering a more stable and prosperous customer base.

TABLE OF CONTENT

Introduction

Customer attrition, also known as customer churn, is the phenome non 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 safe guard their reputation. The financial impact of customer attrition on bank scan 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 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 de mographics, 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 recom mendations with management, marketing teams, or other relevant parties.

The success of any business model relies on having a large customer base, which entails achieving two primary objectives: acquiring new customers and retaining 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. To address customer attrition, previous studies have discussed customer relationship management (CRM) systems and three approaches for retention. These approaches 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. 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. In a related study, 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 establish a starting point for conducting further research in this field.

Successful financial firms must provide useful customer assistance. By examining how consumers use goods, employing machine learning (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 or extend customer engagement, whether it is a customer who has forgotten about the service or 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. Practically, employing such a model to predict the likelihood of customer attrition allows banks to focus on this group of customers. 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.

Table 1:

Workflow of the Data Visualization RShiny App for Data Science and Management
Steps Functionalities

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.
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. Supports data-driven decision making.
10. Outputs	Supports data-driven decision making. Provides interactive visualizations and analysis results. Allows for comparative analysis outcomes. Creates predictive churn models Provides real-time monitoring updates. Outputs customized reports and exported data.

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.

Description of dataset

The dataset used in this study, accessible at Kaggle, comprises 10,000 rows of customer information. 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 onboarding, ensuring 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 includes Boolean measurements, such as 0 or 1, and other sections with two or more classes. These can be classified as follows: country, gender, has a credit card, being an active member, and churned. The final column, "exited," determines the current state of the customer, where 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)
Precision	True Positive/(True Positive+False Positive)	Specificity	TN/(TN+FP)
Recall	True Positive/(True Positive+False Negative)	SMOTE	Synthetic minority oversampling technique

Dataset Analysis

We preprocessed the dataset to effectively unify and visualize the data. Figure 1 provides a visual summary of the scientific contribution of the study.

Table 3: Variables, Null Count, and Unique Count of the Dataset

Variables	Variables	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

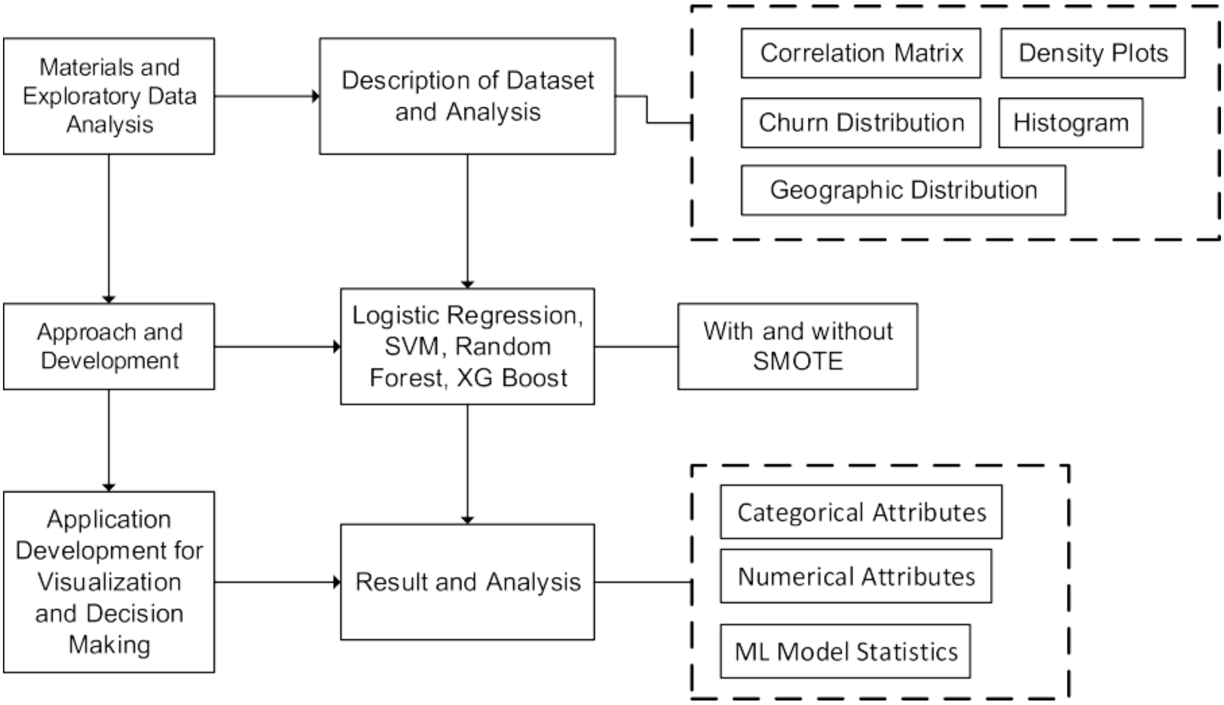


Fig. 1. Visual summary of the scientific contribution of the study.

(a) Customer Churn Distribution

The pie chart shows that 80% of customers did not churn, while 20% did.

This indicates that the dataset is highly imbalanced.

(b) Gender, Active Members, Credit Cards, and Country-Based Analysis

Gender:

25% of female customers churned (1,139 out of 4,543).

16% of male customers churned (898 out of 5,457).

Active vs. Inactive Customers:

Inactive customers have a churn rate of 27%.

Active customers have a churn rate of 14%.

Countries:

Germany has the highest churn rate at 40%.

France and Spain both have a churn rate of 16%.

(c) Balance, Owned Product Quantity, Credit Score, and Tenure-Based Analysis

Balance: Customers with a balance over 85,000 are more likely to churn.

Owned Products: Customers with more products at the bank are less likely to churn.

Credit Score:

Customers with a credit score below 400 are more likely to churn.

Other credit scores and tenure do not significantly impact churn rates.

Age:

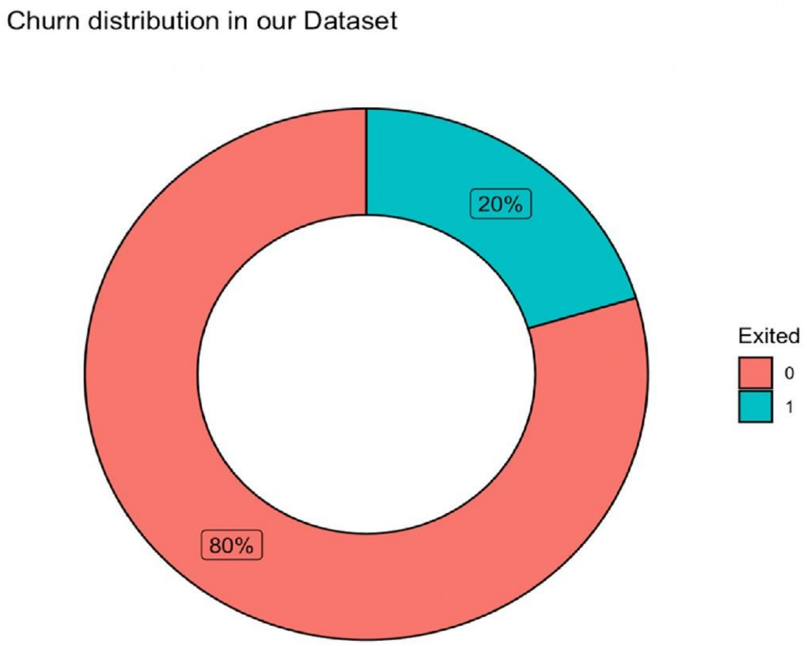
Older customers (above 40) are more likely to churn.

Customers aged 40-70 have a higher churn rate, possibly due to better plans for seniors at other banks.

(d) Correlation Matrix Analysis

No strong correlations between pairs of variables, indicating no multicollinearity.

The only notable correlation is between the number of products and balance.



Churn distribution in our Dataset

Theory, Approach, and Development

This section introduces different machine learning (ML) techniques applied to the dataset to predict customer churn in the banking sector. The focus is on core ML approaches, including logistic regression, support vector machine (SVM), random forest, and eXtreme Gradient Boosting (XGBoost). Additionally, a visualization tool for stakeholders is introduced to summarize various data aspects in a unified manner using PowerBI.

ML Techniques

We briefly discuss the theoretical underpinnings of the ML techniques used in this research:

Logistic Regression

A fundamental classification technique crucial for predicting customer churn. The logistic curve equation connects the dependent variable (probability of churn) and the independent variable.
The logistic regression model uses parameters to regulate how rapidly the probability changes with changes in the independent variable.

Random Forest

Creates a training dataset by sampling rows and features from the original data. An individual decision tree is created for each subset, and the majority vote of the trees determines the final classification.
Advantages include high accuracy and the ability to handle large datasets, though ensuring robustness in predicting unknown data can be challenging.

Support Vector Machine (SVM)

Provides the distance to the decision boundary and converts it to probability. Different techniques may perform better depending on the specific issue being addressed.

Gradient-Boosted Tree Algorithm (XGBoost)

A supervised learning approach based on function approximation by maximizing certain loss functions and employing regularization methods. XGBoost is known for its practical implementation and effectiveness in predictive analysis.

The objective functions (loss function and regularization) must be minimized during iterations to prevent overfitting and enhance model performance. Visualization Tool Using PowerBI

To make the insights actionable, we developed a visualization tool using PowerBI tailored explicitly for stakeholders to summarize different data aspects related to customer churn in the banking sector. PowerBI allows us to:

Visualize Customer Churn Distribution: Create pie charts and bar graphs to show the proportion of churned vs. non-churned customers, highlighting the imbalance in the dataset.

Analyze Gender and Demographics: Display the churn rates segmented by gender, age, and other demographic factors.

Evaluate Active vs. Inactive Members: Show comparisons between active and inactive customers, highlighting the differences in churn rates.

Assess Geographic Distribution: Map the churn rates across different regions or countries to identify high-risk areas.

Examine Financial Metrics: Use scatter plots and density plots to analyze the relationship between churn rates and financial metrics such as account balance, credit score, and tenure.

Identify Patterns and Trends: Utilize heatmaps and correlation matrices to detect patterns and relationships between various factors contributing to churn.
By integrating these ML techniques with PowerBI visualizations, stakeholders can better understand customer churn dynamics and make informed decisions to improve customer retention strategies in the banking sector.

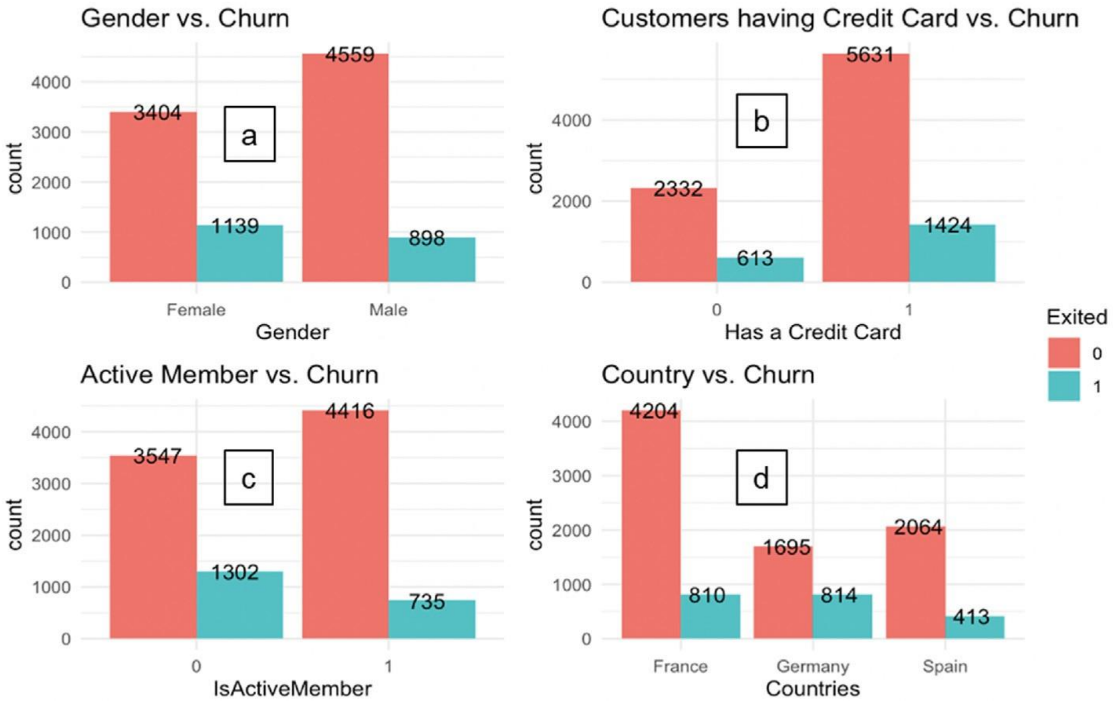


Fig. 3. Histograms of the dataset. (a) Gender vs. churn; (b) customers having credit card vs. churn; (c) active member vs. churn; (d) country vs. churn.

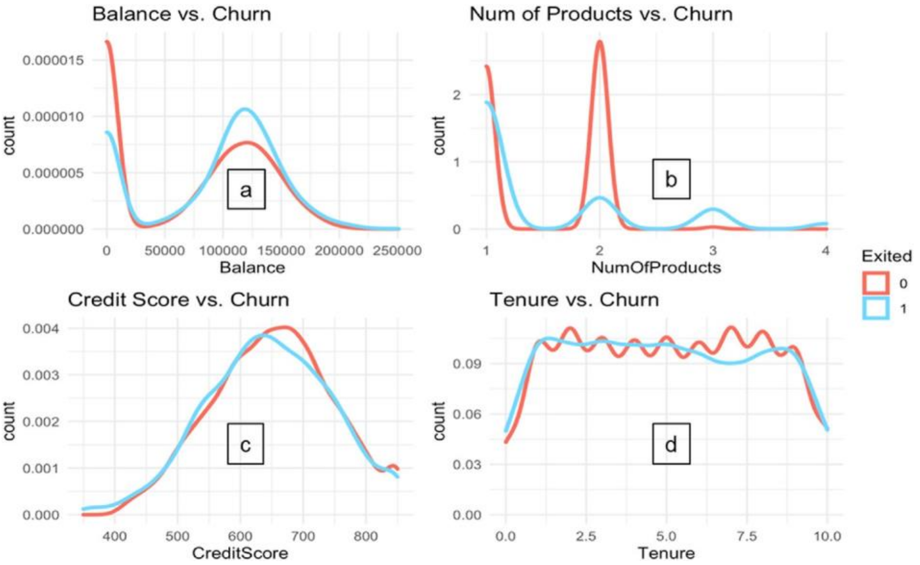
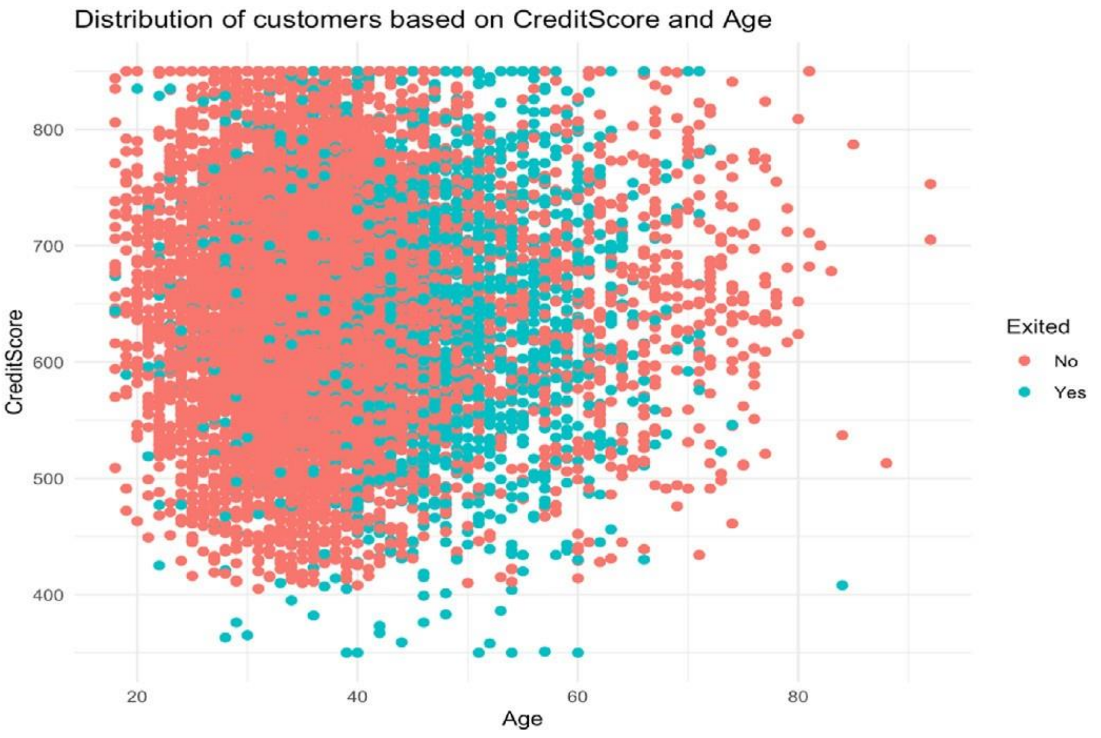


Fig. 4. Density plots of (a) observed balance, (b) owned product quantity, (c) credit score, and (d) tenure.



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 using a two-step workflow pipeline:

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

Grid search parameters were configured for each model. Four grid-search cross-validation functions were set up, using the pipeline and parameters as inputs. All grids were then fitted to the training dataset.

Logistic Regression Model

Achieved an accuracy of 85%, a sensitivity of 38%, a specificity of 97%, and an AUC of 0.83 on validation data.

The large discrepancy between sensitivity and specificity was due to data imbalance, causing model bias.

To address the imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was used, which creates synthetic data points based on original data points to augment the dataset without losing information.

Handling Data Imbalance with SMOTE

SMOTE was applied to the dataset before and after addressing the imbalance.
Methods like undersampling were avoided to prevent loss of information.

Customer churn prediction involved fitting and evaluating multiple models after using SMOTE.

Feature Engineering and Data Cleaning

Irrelevant rows, such as "Row Number" and "Surname," were removed.
Three new variables were created: "TenureByAge," "BalanceSalaryRatio," and "CreditScoreGivenAge," to enhance model performance.

This addressed the research question on selecting pertinent attributes and removing outliers.

Model Evaluation and Selection

After thorough evaluation, the best model was chosen based on a combination of metrics, ensuring adaptability to dynamic data patterns.
Privacy was maintained as the data did not contain any information traceable to individuals.

By following this approach, we aimed to predict customer churn accurately for an imbalanced dataset and identify the most reliable model. This also involved selecting relevant attributes and removing outliers to improve model performance.

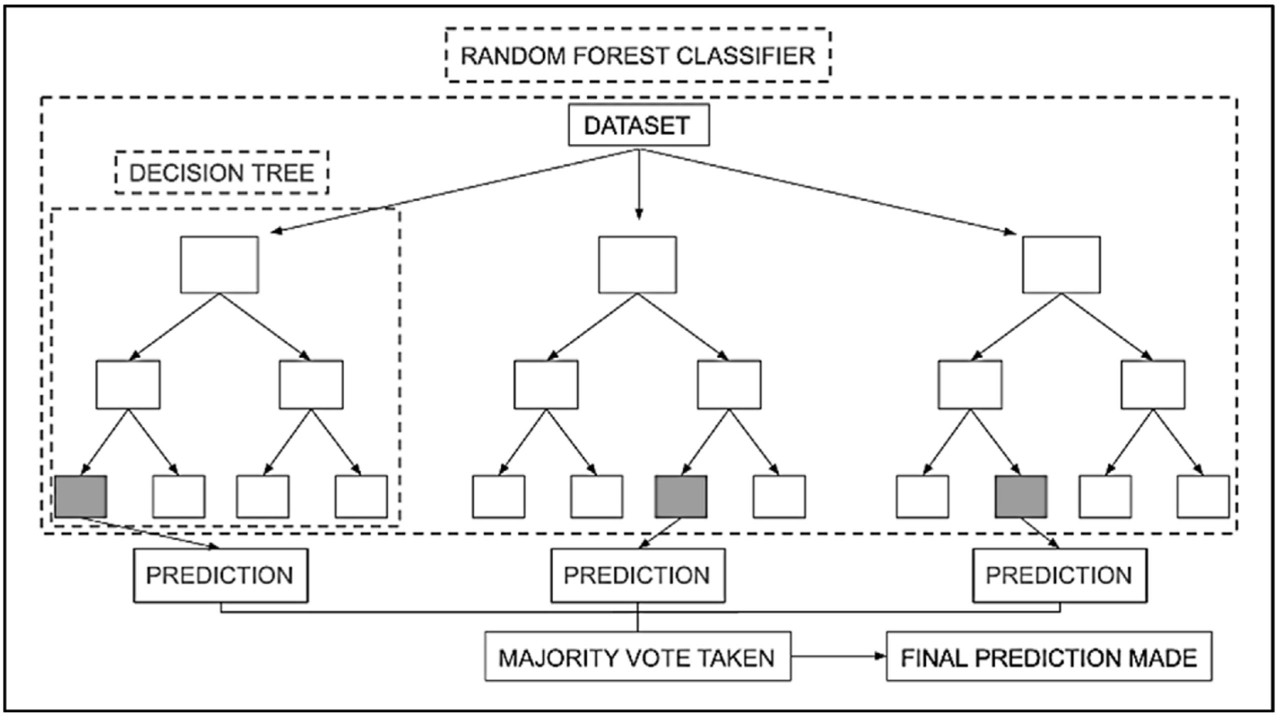


Fig. 8. Random forest classifier with dataset.

Application Development for Visualization and Decision Making

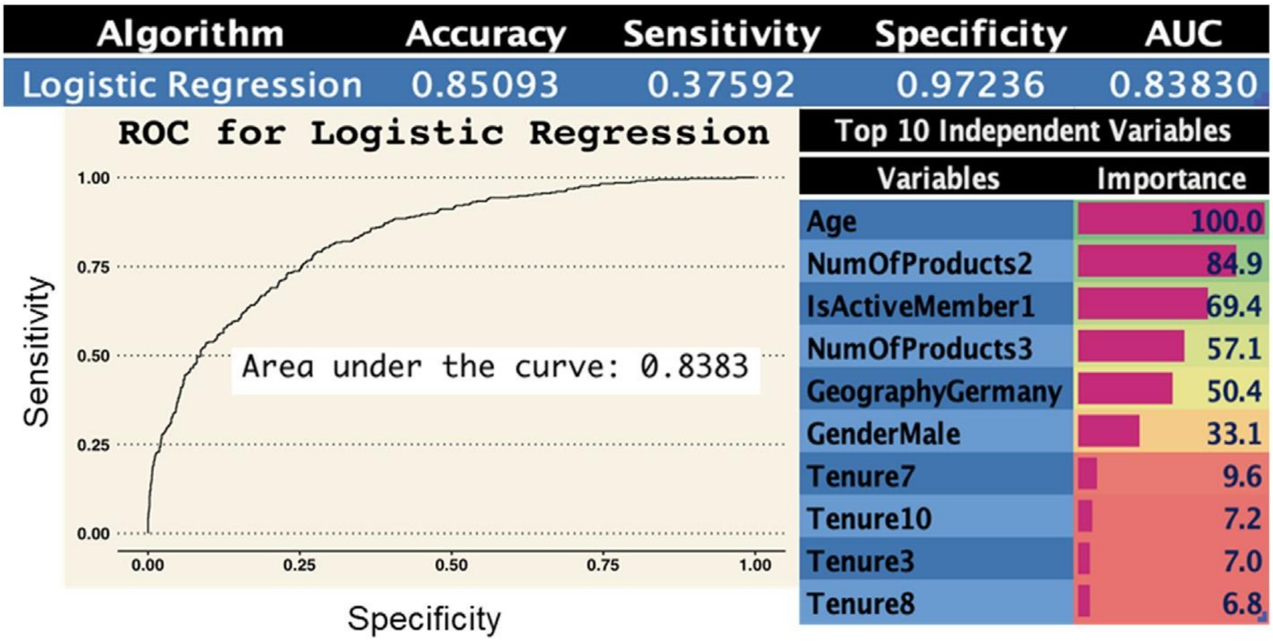
To provide a practical implementation, we developed an application using Plotly and Python to demonstrate our model calculations and customer analysis outcomes. The application interface consists of three primary tabs: data analysis, model analysis, and prediction.

Data Analysis Tab

Displays information on both categorical and numerical attributes.
Users can select attributes to view.
Categorical data is shown using donut charts (indicating the percentage presence or absence of an attribute) and bar charts (showing how presence or absence of an attribute varies with churn).
Numerical data is displayed using density plots, scatter plots with age, and box plots of the selected attributes.

Model Analysis Tab

Shows evaluation metrics including accuracy, sensitivity, specificity, AUC, F1 score, test-train data split percentage, and feature importance for all models.
Provides a comprehensive view of model performance.



Algorithm	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.85093	0.37592	0.97236	0.83830

Top 10 Independent Variables

Variables	Importance
Age	100.0
NumOfProducts2	84.9
IsActiveMember1	69.4
NumOfProducts3	57.1
GeographyGermany	50.4
GenderMale	33.1
Tenure7	9.6
Tenure10	7.2
Tenure3	7.0
Tenure8	6.8

Fig. 9. Logistic regression model statistics.

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

Prediction Tab

Allows users to set the independent variables to their chosen values to obtain the predicted dependent variable (churn).
Helps in determining the churn probabilities for all models.
Enables users to make informed decisions based on model predictions.

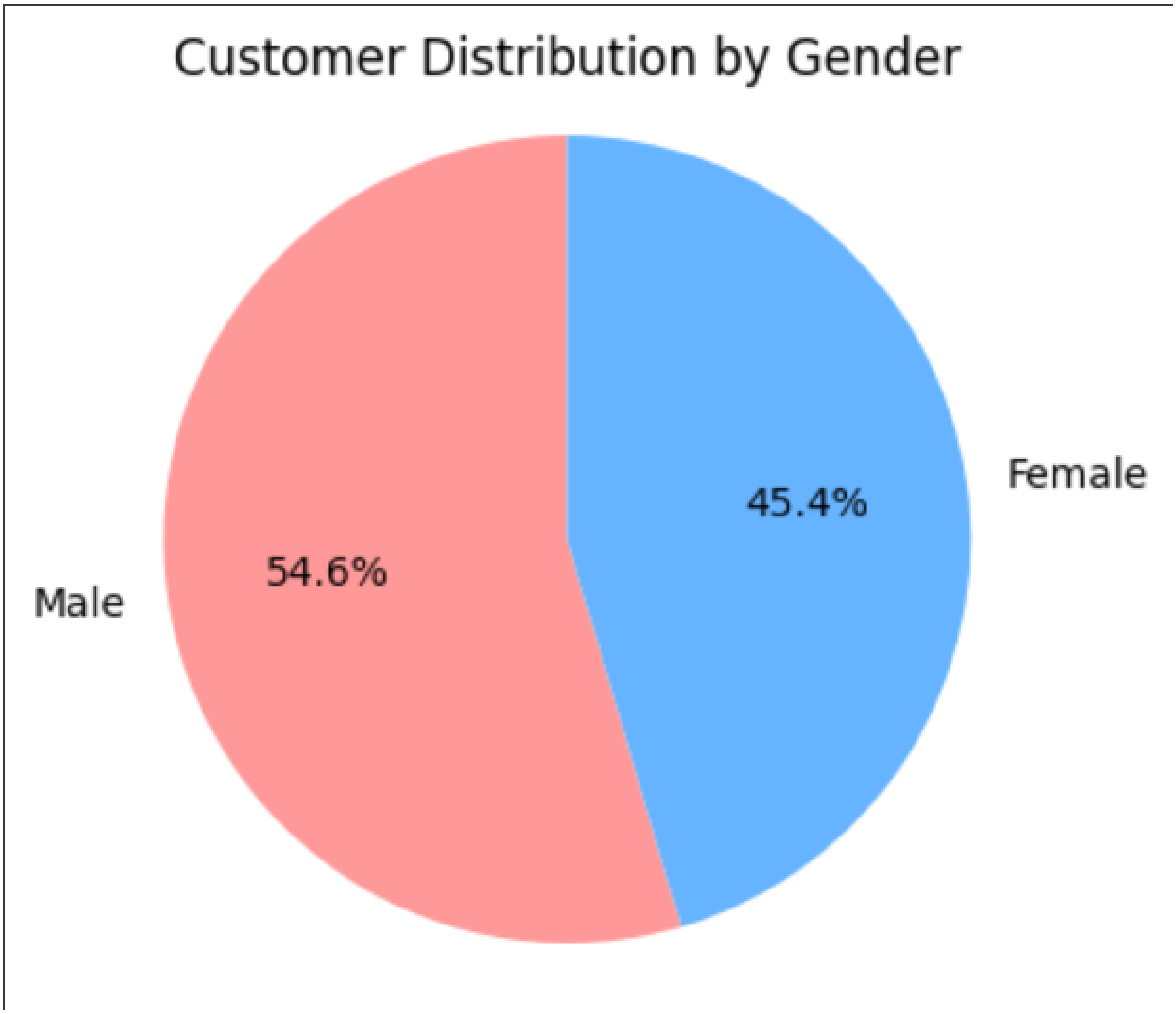
Results and Discussion

This section discusses the findings of the ML models using evaluation metrics such as accuracy, sensitivity, specificity, AUC, and F1 score.
Model Performance
The sensitivity of all models was initially insufficient, with values ranging from 17.3% to 44%.
After applying SMOTE for data balancing, XGBoost performed the best in terms of accuracy (83%), followed by random forest (78.3%).
XGBoost also led in F1 score (0.613), specificity (90.3%), and AUC (0.847). Random forest followed closely in these metrics.
Logistic regression exhibited the highest sensitivity (71.4%), followed by random forest (69.3%).

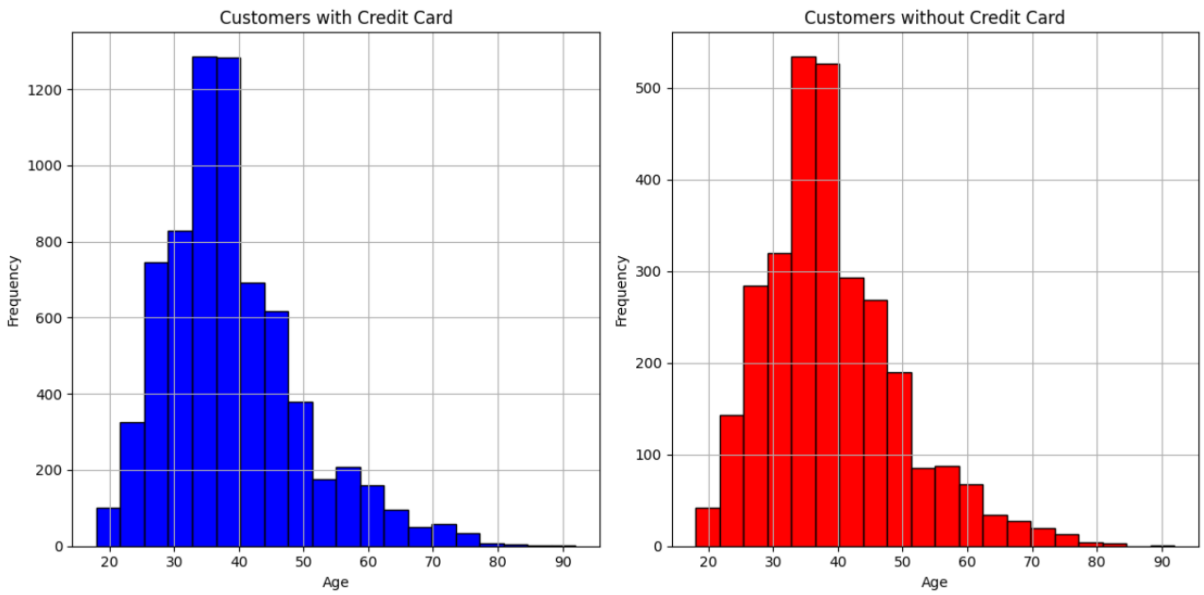
Choosing the Best Model

Since sensitivity and accuracy are the most relevant metrics, random forest is recommended

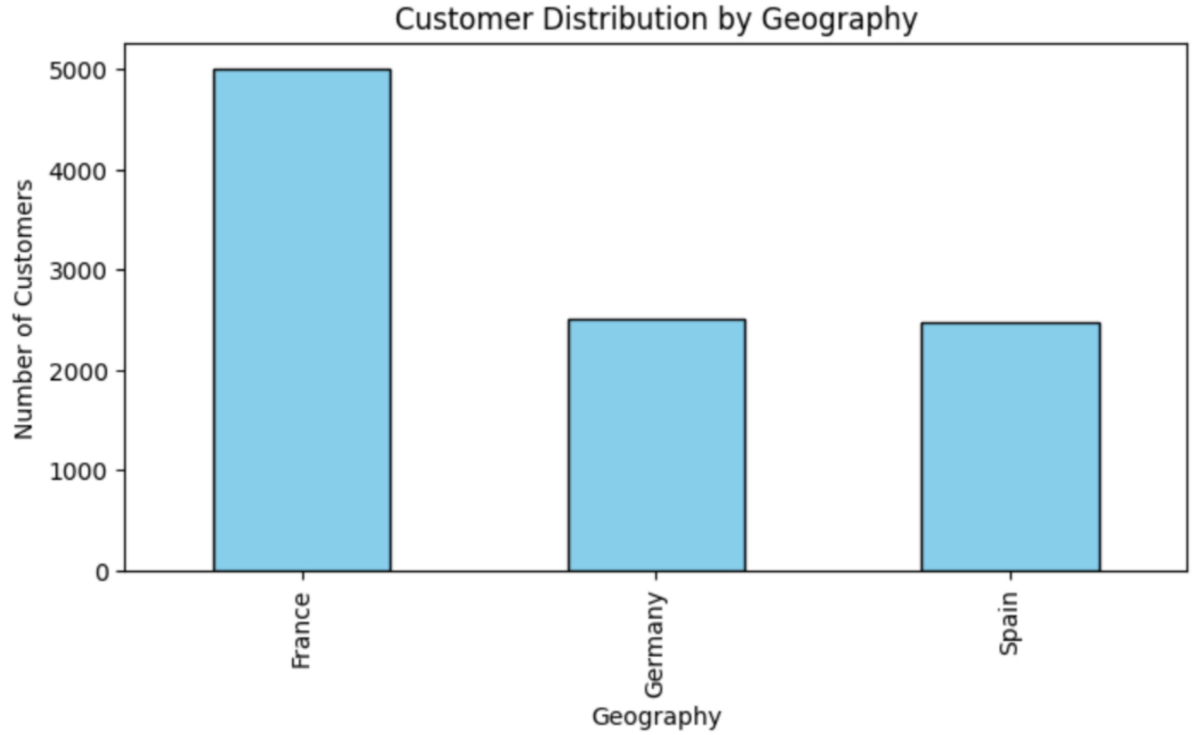
due to its balanced performance in both metrics.
Random forest can handle large datasets effectively and is robust against outliers, unlike linear regression, which is sensitive to outliers.
By implementing this structured approach, we provided a comprehensive tool for visualizing and predicting customer churn, enabling stakeholders to make data-driven decisions.



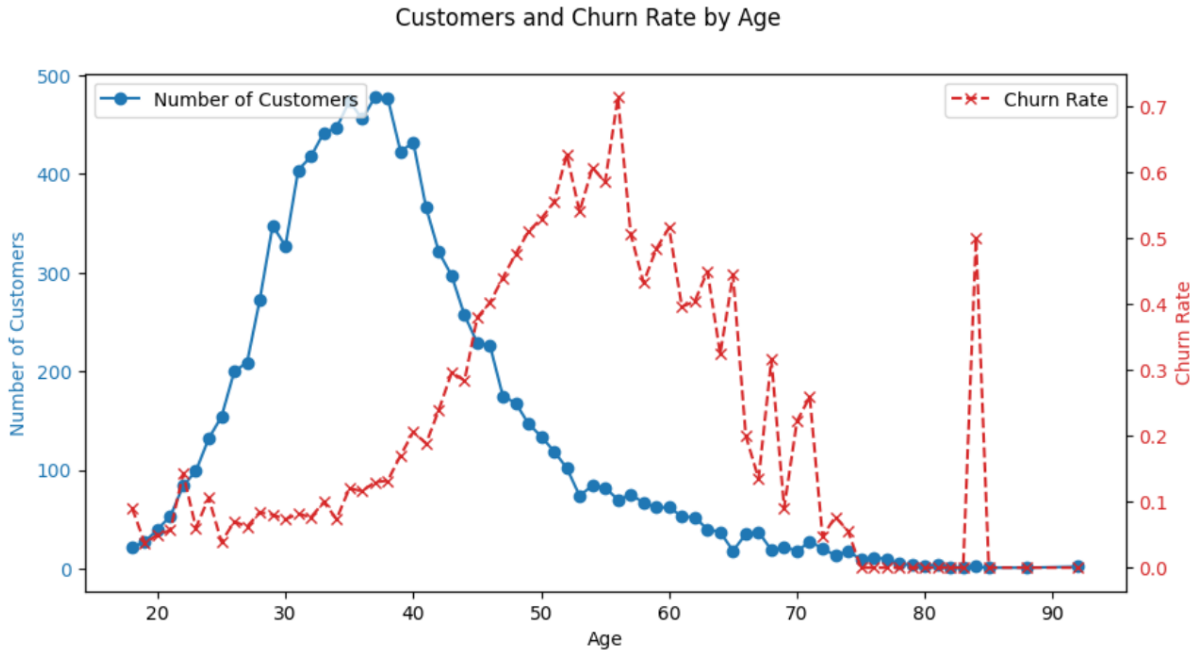
PIE Chart



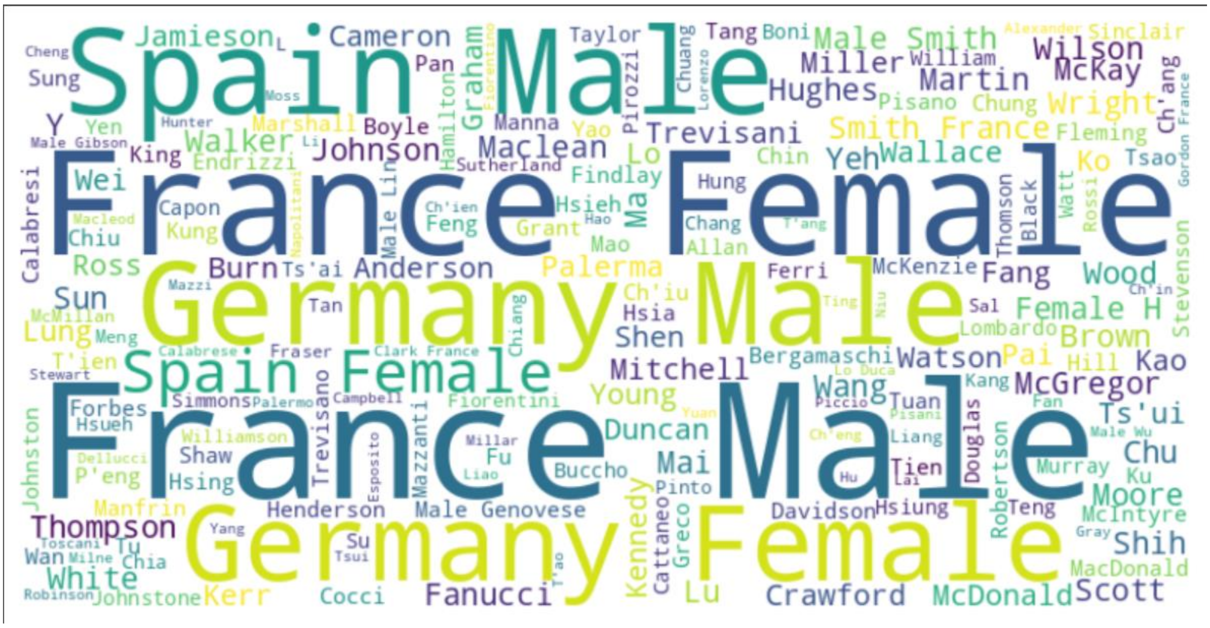
Histogram



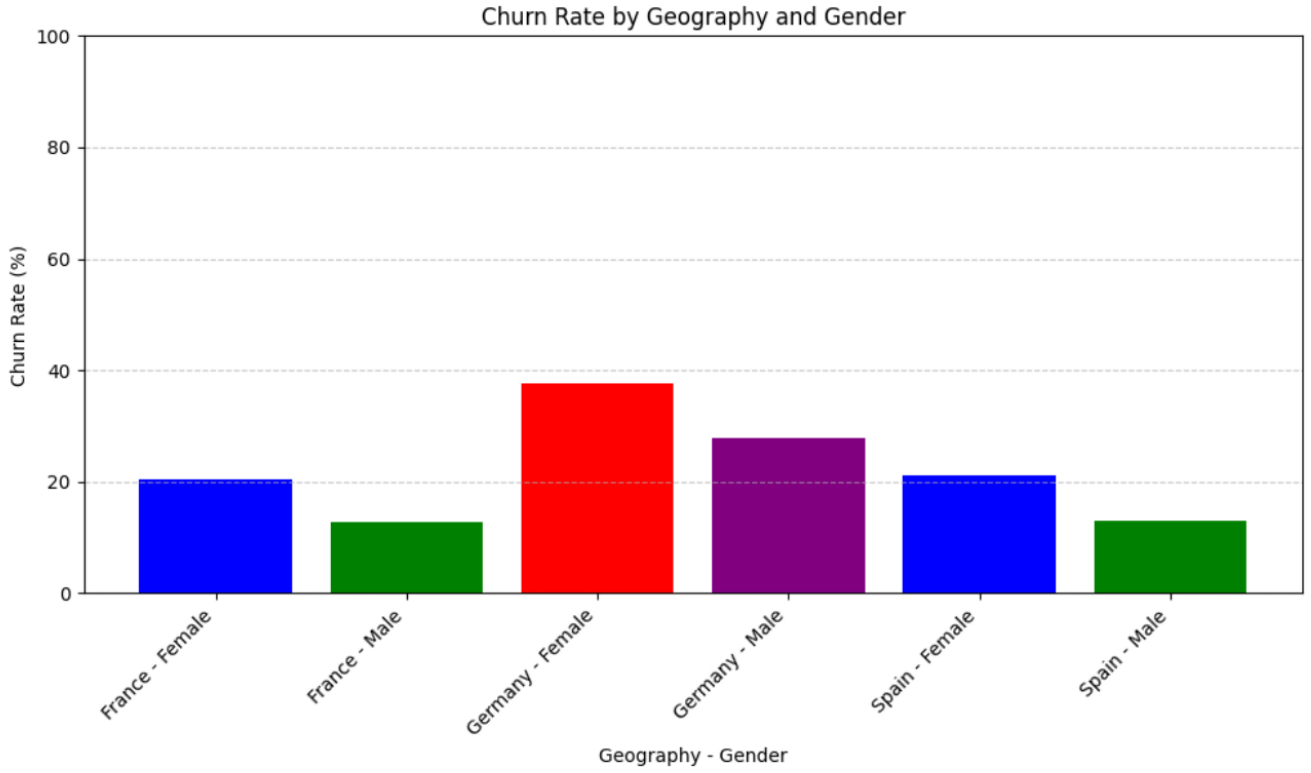
Bar Chart



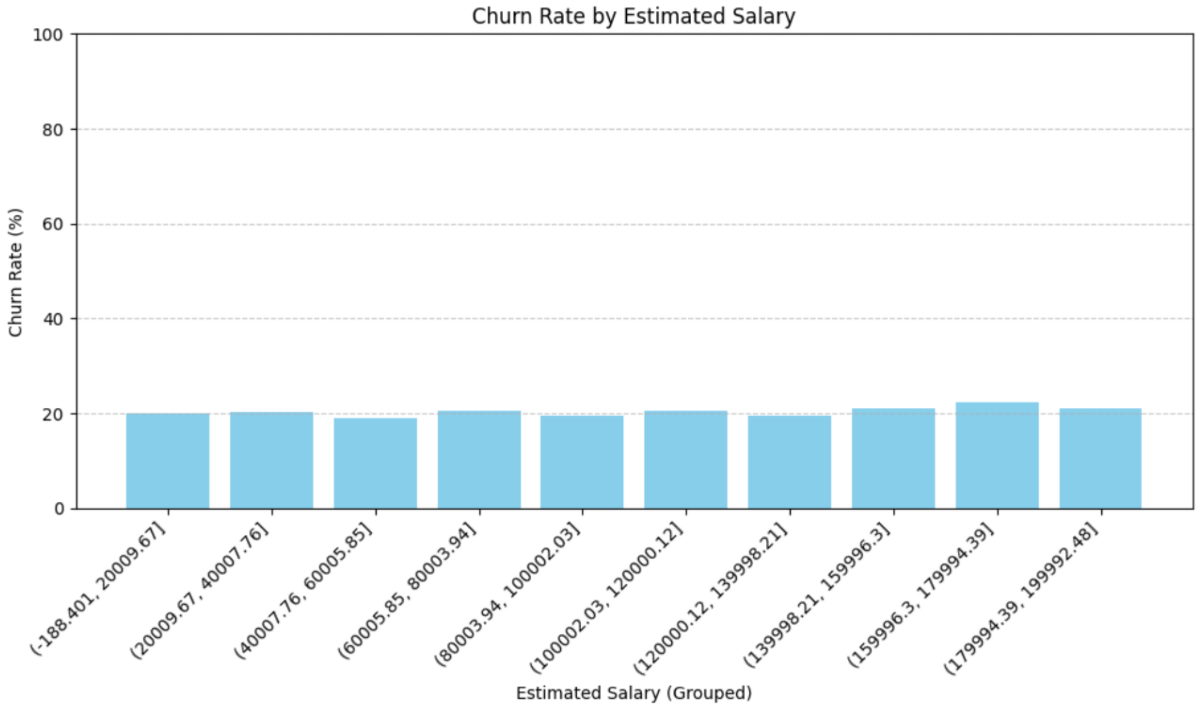
Line plot



Word cloud

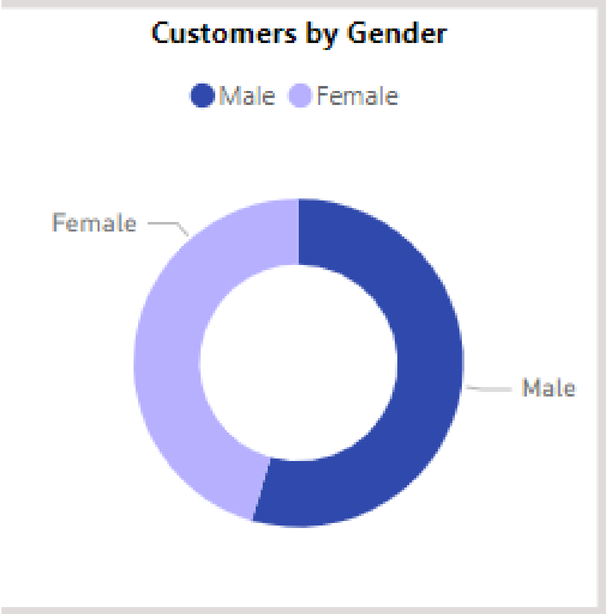


Bar graph

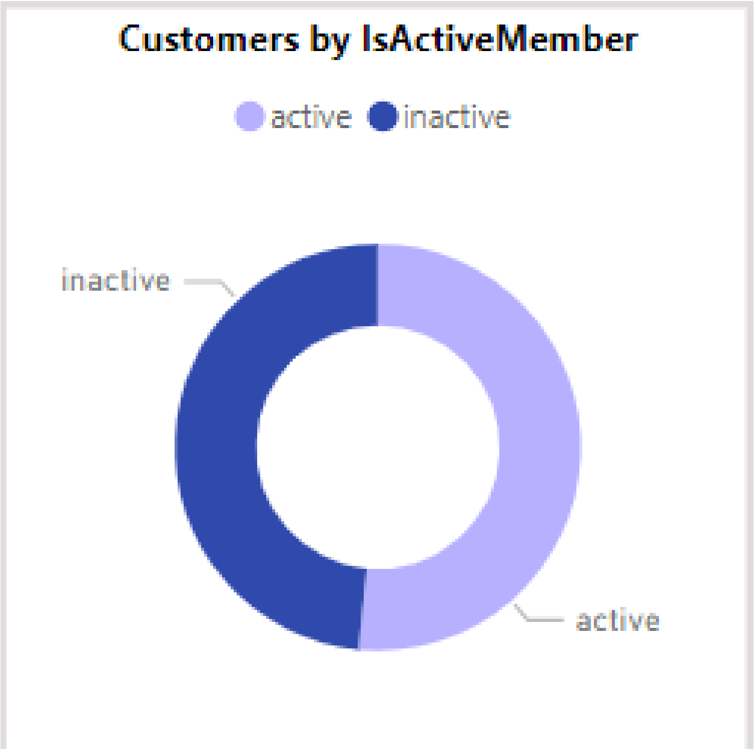


Bar graph

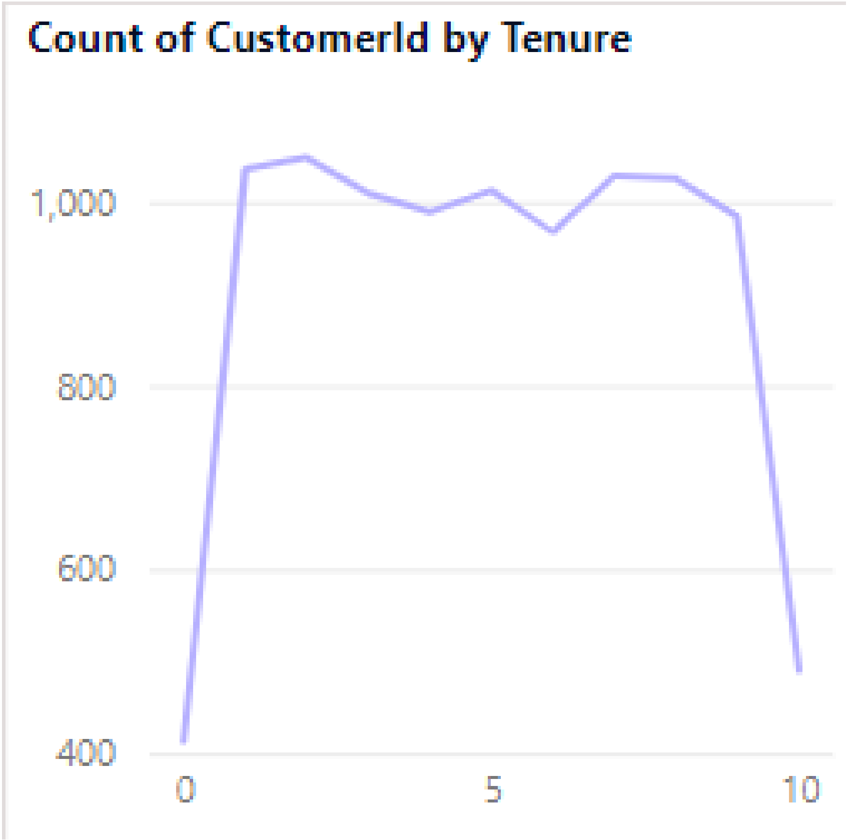
Using Power BI

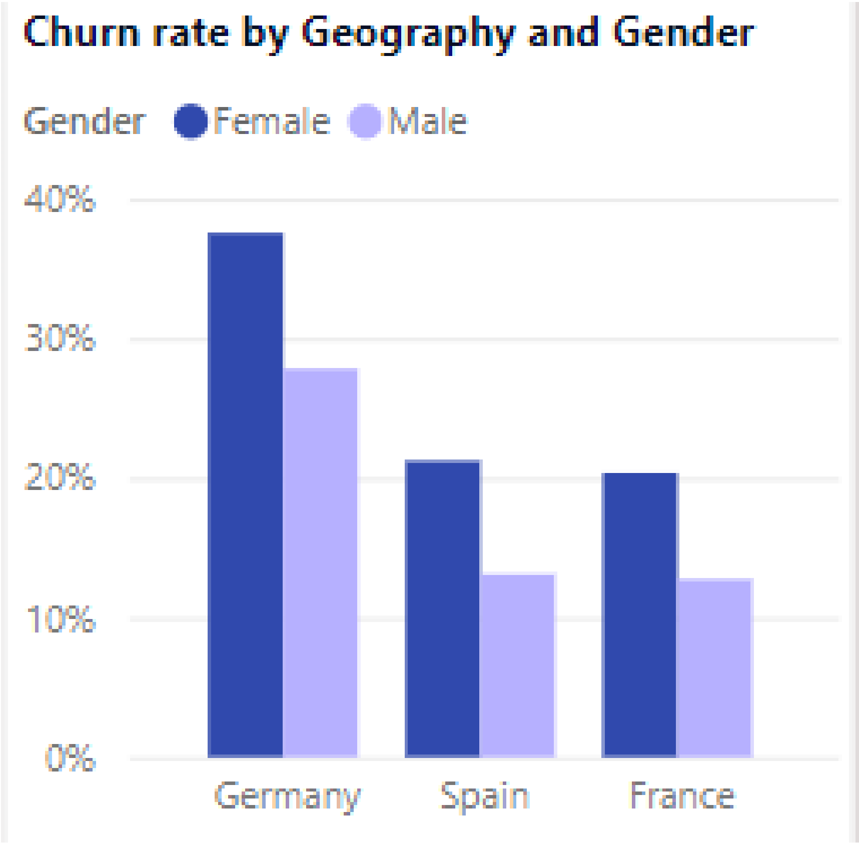


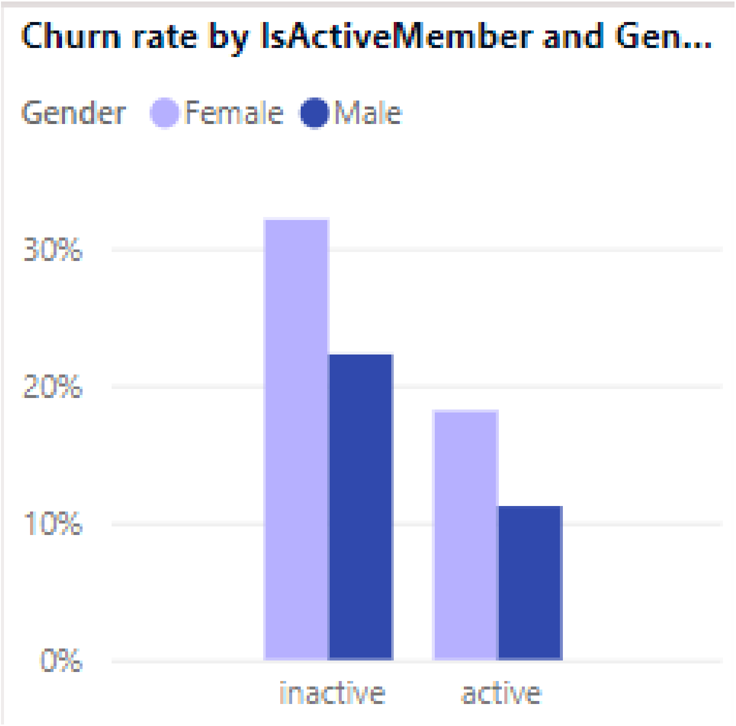
Donut

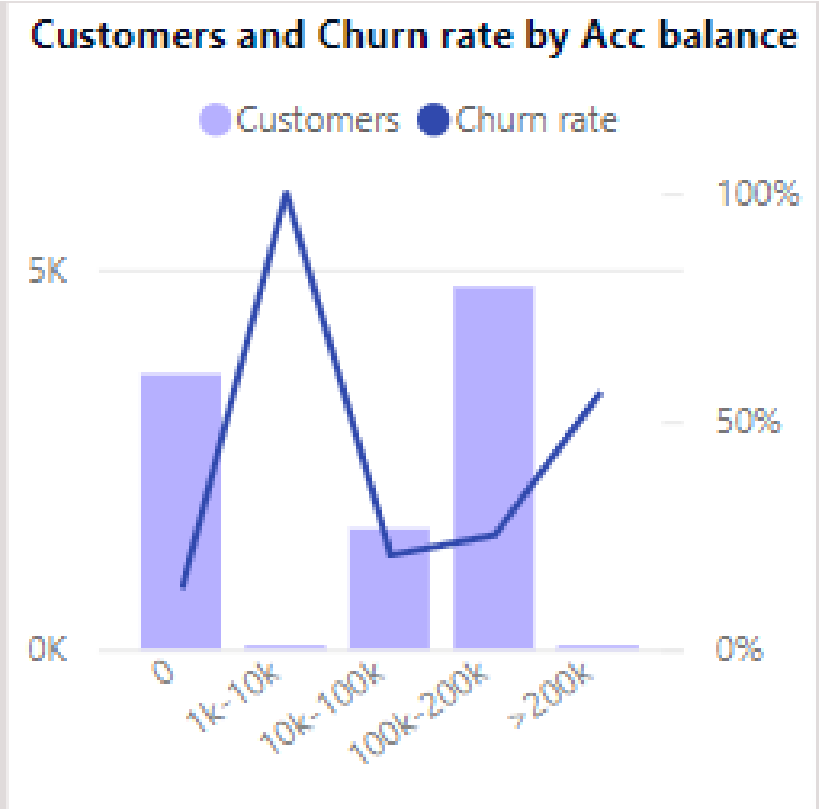


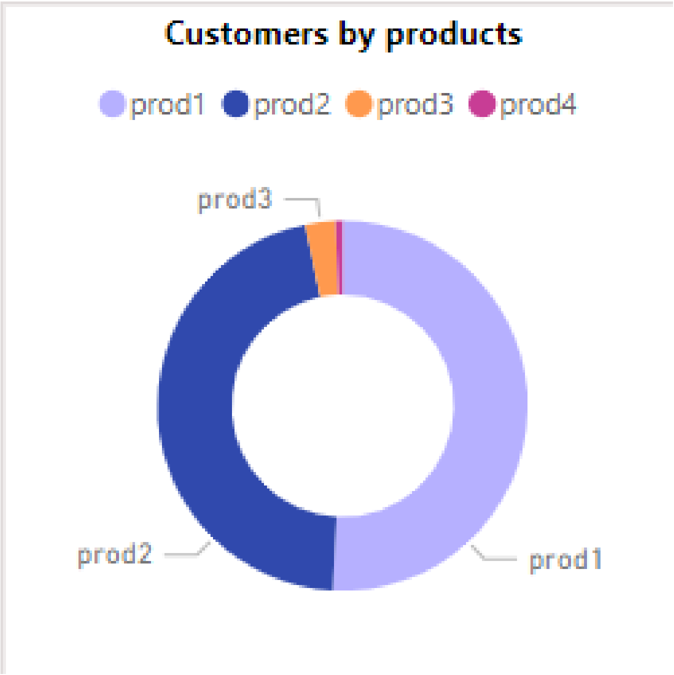
Donut

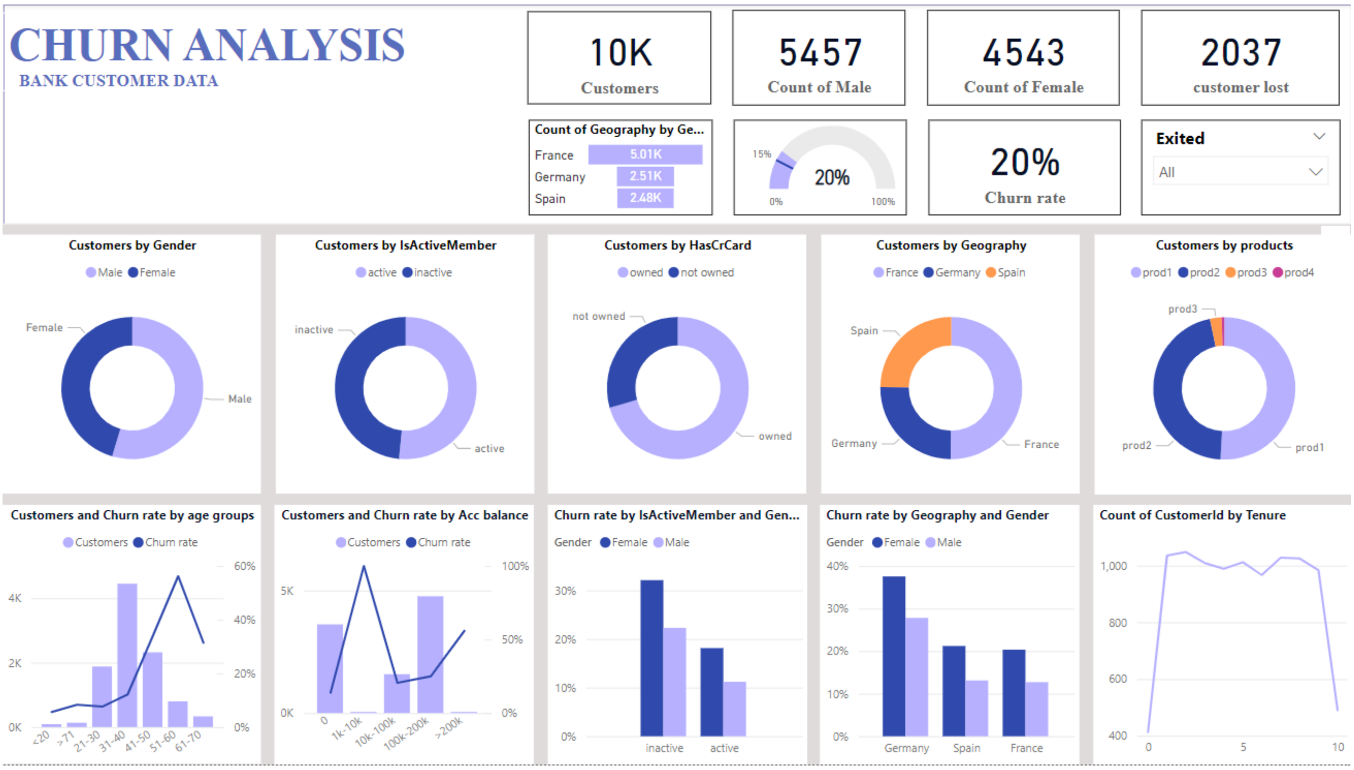

Count of CustomerId by Tenure
1,000
800
600
400
0
5
10


Churn rate by Geography and Gender
Gender
● Female
● Male
40%
30%
20%
10%
0%
Germany
Spain
France


Churn rate by IsActiveMember and Gen...
Gender ● Female ● Male
30%
20%
10%
0%
inactive
active


Customers and Churn rate by Acc balance
● Customers ● Churn rate
5K
0K
100%
50%
0%
0
1k-10k
10k-100k
100k-200k
>200k





Dashboard

1. Conclusion

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 pre- dictions. 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 devel- oping 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 general-usability 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.

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.

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 over-
sampling 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., et al., 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.

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.

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, 04.

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.