

PREDICTING CUSTOMER CHURN IN BANKING: AN SQL BASED APPROACH FOR COHORTING CUSTOMERS AND MACHINE LEARNING ALGORITHMS FOR DATA VISUALIZATION

Sheebal M S

*RV Institute of Management
sheebalshinnu@gmail.com
Ph:- 9538183383*

M Praneet Kumar Reddy

*RV Institute of Management
praneethm.rvim23@gmail.com*

Pramod K L

*RV Institute of Management
pramodkl.rvim23@gmail.com
Ph:- 9481237850*

Dr. Jahnvi M

*Associate professor,
RV Institute of Management
Bangalore*

ABSTRACT

Customer churn, or consumers abandoning their engagement with a bank, is a key concern in the banking sector. This research attempts to provide an overview of customer turnover in the banking industry, investigating its origins, repercussions, and mitigation techniques. As customers discontinue their relationship with a bank, it can directly impact the bank's financial performance and overall business sustainability. Customer churn leads to reduced revenue and profitability for the bank. As customers leave for competing banks, the bank may lose its competitive edge and struggle to attract new customers.

By conducting a Postgre SQL-based study, banks can delve deep into their customer data, examining patterns and trends to identify the underlying causes of churn. This analysis enables banks to develop targeted strategies and initiatives to reduce churn rates and retain valuable customers.

In this study, we have conducted cohort analysis using Postgre SQL in order to identify the average number of products of churned customers, the average age of churned customers, the average balance of churned customers, and the average credit score of churned customers. Later, we explored the top 3, 5, or 10 customers based on certain characteristics and highlight trends/patterns. Further, we transferred the structured data from Postgre SQL to Python to visualize and understand the behaviour of the exited customers due to different factors, such as customer age, credit score, tenure, number of products, and geographical location.

number of products of churned customers, the average age of churned customers, the average balance of churned customers, and the average credit score of churned customers. Later, we explored the top 3, 5, or

10 customers based on certain characteristics and highlight trends/patterns. Further, we transferred the structured data from Postgre SQL to Python to visualize and understand the behaviour of the exited customers due to different factors, such as customer age, credit score, tenure, number of products, and geographical location.

Keywords:

Business Intelligence, Churn analysis, Cohort analysis, Data visualization, Postgre SQL, Python.

INTRODUCTION

Customer churn, also known as customer attrition, is the rate at which customers discontinue their relationship with a company or business. It represents the percentage of customers who stop using a company's products or services over a specific period. Churn can be voluntary (customer choice) or involuntary (external factors). Measuring churn helps businesses assess customer retention and identify strategies to reduce attrition, as retaining existing customers is typically more cost-effective than acquiring new ones. Reducing customer churn is crucial for sustaining profitability and fostering long-term customer loyalty. (Bilal, et al.)

PostgreSQL is an open-source database management system with versatile applications. In general, it is used for data storage, web applications, data analysis, geospatial data, and scientific research. In the banking industry, PostgreSQL securely manages customer data, aids in customer segmentation, churn analysis, and predictive analytics. It optimizes marketing strategies, enhances customer service, and supports mission-critical applications. (Kumar, et al.)

Python is a strong and well-liked programming language that has become very popular in data visualisation. It has a variety of libraries and features that make it a great option for producing attractive and educational data visualisations. Python is a well-liked choice for activities involving data visualisation, whether they include data exploration, analysis, or the dissemination of insights to stakeholders. Python's status as the go-to language for data scientists and analysts in numerous fields is further strengthened by its integration with libraries for data manipulation and analysis like NumPy, Pandas, and SciPy.

IMPORTANCE OF POSTGRE SQL IN CHURN ANALYSIS FOR BANKS

Analyzing customer churn in the banking industry through a Postgre SQL is significant due

to its direct impact on a bank's financial performance and customer retention. Customer churn, the rate at which customers discontinue their relationship with a bank, can lead to financial losses, increased customer acquisition costs, and reduced market competitiveness. By conducting a Postgre SQL-based study, banks can delve deep into their customer data, examining patterns and trends to identify the underlying causes of churn. This analysis enables banks to develop targeted strategies and initiatives aimed at reducing churn rates and retaining valuable customers. (Singh, et al)

A Postgre SQL-based approach offers the advantage of efficiently handling vast datasets, characteristic of the banking industry. The study allows for seamless querying, filtering, and processing of data, facilitating comprehensive churn analysis and the generation of actionable insights. The insights gained from churn analysis guide banks in enhancing customer experiences and service quality. By understanding the reasons behind churn, banks can identify pain points and areas of improvement, leading to enhanced customer satisfaction and loyalty. This, in turn, strengthens the bank's reputation and fosters long-term customer relationships. (Kumar, et al.).

STUDY OBJECTIVES

- Identify the key drivers that significantly influence churn rates, helping the bank prioritize efforts to address those specific factors.
- Conducting cohort analysis using Postgre SQL in order to identify average number of products of churned customers, the average age of churned customers, the average balance of churned customers, and the average credit score of churned customers.
- To explore the top 3, 5, or 10 customers based on certain characteristics and highlight trends/patterns.
- Data visualization using Python, to understand the behaviour of the exited customers due to different factors, such as customer age, credit score, tenure, number of products, and geographical location.

LITERATURE REVIEW

1. Linares-Mustarós, R. P., García-Sánchez, J. A., & Segovia-Vargas, M. V. "Customer Churn Prediction in Retail Banking: A Data Mining Approach."

This research employs a data mining approach to predict customer churn in the retail banking sector. By analyzing historical customer data, the study aims to identify patterns and factors that contribute to customer attrition. The findings provide valuable insights for banks to develop effective customer retention strategies and enhance overall business performance, ensuring sustained profitability in a competitive market.

2. Abdulaziz, M., & Muhammad, A. N. "Customer Churn Prediction in Retail Banking: A Data Mining Approach Using R."

This study explores a data mining technique using the R programming language to predict customer churn in the retail banking industry. By analyzing customer behaviour and transaction data, the authors aim to develop accurate churn prediction models. The research assists banks in identifying customers at risk of churn and implementing targeted retention strategies, ultimately improving customer loyalty and satisfaction.

3. Mukherjee, B., Sarkar, S., & Sarkar, D. "Predicting Customer Churn in Banking Industry Using Ensemble Learning."

This research utilizes ensemble learning techniques to predict customer churn in the banking industry. By combining multiple algorithms, the authors aim to enhance churn prediction accuracy. The study enables banks to identify high-risk customers, develop personalized retention strategies, and minimize customer attrition.

4. Abbas, S., & Qi, S. "A Review of Customer Churn Prediction in Telecommunication Industry."

This review article explores various customer churn prediction approaches in the telecommunication industry. By summarizing

existing research, the study provides an overview of different methodologies and their effectiveness. The research assists telecommunication companies in understanding the state-of-the-art in churn prediction and implementing advanced techniques for customer retention.

5. Mathur, R., & Tayal, D. "Churn Prediction in Telecom Industry Using Advanced Data Mining Techniques."

This research applies advanced data mining techniques to predict customer churn in the telecom industry. By analyzing customer call records, usage patterns, and demographic data, the authors aim to build accurate churn prediction models. The findings help telecom companies identify churn-prone customers, enhance customer satisfaction, and reduce churn-related losses.

6. Khanam, A., & Zulkernine, F. H. "Churn Analysis and Prediction in Telecommunication Industry Using Machine Learning Techniques."

This study utilizes machine learning techniques to analyze and predict churn in the telecommunication industry. By considering various features such as call patterns, customer complaints, and subscription data, the research aims to develop effective churn prediction models. The findings enable telecommunication companies to optimize customer retention strategies and improve service quality.

7. Chris, "A PostgreSQL-based study on banking churn analysis."

This study investigates banking churn analysis using PostgreSQL as the database management system. By utilizing SQL queries and data processing capabilities of PostgreSQL, the research aims to analyze customer behaviour and churn patterns in the banking industry. The study provides insights to help banks implement effective retention strategies and improve customer satisfaction and loyalty.

8. Kaur, & Kaur. "Customer Churn in the Banking Industry: A Literature Review."

This literature review delves into the topic of customer churn in the banking industry. By summarizing and analyzing existing research, the study provides a comprehensive overview of churn-related factors, challenges, and strategies in the banking sector. The review helps researchers and practitioners gain insights into the current state of knowledge on customer churn and informs the development of future studies and practical interventions to address churn in banks.

9. Bilal, et al. "Customer Churn Analysis in Banking Sector: Evidence from Explainable Machine Learning Models."

This study presents an analysis of customer churn in the banking sector using explainable machine learning models. By interpreting model outputs, the research aims to provide evidence-based insights into churn patterns and contributing factors. The findings assist banks in understanding the reasons behind customer attrition and devising targeted retention measures for increased customer satisfaction and long-term loyalty.

10. Singh, et al. "Churn Prediction in the Retail Industry Using PostgreSQL."

This research explores churn prediction in the retail industry, utilizing PostgreSQL as the database management system. By employing SQL queries and PostgreSQL's capabilities, the study aims to predict customer churn in the retail sector. The insights derived from the analysis help retailers develop effective customer retention strategies, optimize marketing efforts, and enhance overall business profitability.

11. Kumar, et al. "Churn Analysis in the Banking Sector Using PostgreSQL."

This study conducts churn analysis in the banking sector, leveraging PostgreSQL as the database management system. By analyzing customer data and behaviour using SQL queries and PostgreSQL features, the research aims to identify churn patterns and risk factors in banks. The findings enable banks to address churn challenges proactively, retain valuable banking

industry as they will aid in developing effective churn reduction strategies. Banks can use this knowledge to improve customer customers, and foster sustainable growth in the competitive financial landscape.

12. Chen, L., & Li, W. "A Novel Hybrid Approach for Customer Churn Prediction in Retail Banking."

This study proposes a novel hybrid approach for predicting customer churn in retail banking. By integrating different methodologies and features, the research aims to improve churn prediction accuracy. The findings provide valuable insights to help retail banks proactively address churn, personalize customer interactions, and optimize customer retention strategies.

RESEARCH GAP

This research paper offers a thorough and promising approach for forecasting customer churn in the banking sector, but there is still a significant research gap with regard to the incorporation of real-time data and the assessment of the model's long-term performance.

The research focuses specifically on using historical data to cohort customers and makes use of different machine learning methods for data visualisation, which unquestionably offers useful insights into customer attrition tendencies. However, the inability to capture dynamic shifts in consumer behaviour and churn-related issues results from the lack of real-time data integration. The dynamic nature of the banking industry necessitates the development of a model that can quickly adjust to shifting consumer preferences and macroeconomic situations.

METHODOLOGY

The study will encompass an analysis of historical customer data from the bank's database, focusing on transactional information, customer demographics, and behaviour patterns. The analysis will cover a specific time frame and may

satisfaction, enhance customer engagement, and optimize marketing efforts, ultimately leading to improved financial performance.

DATA COLLECTION AND UNDERSTANDING

SECONDARY DATA

Variables	Description
CustomerId	Unique ID issued by the bank to the account holder
Surname	Name of the account holder
CreditScore	Credit score of the account holder
Geography	Geographic location of the account holder
Gender	Gender of the account holder
Age	Age of the account holder
Tenure	The total period of time the customer associates with bank
Balance	Current balance of the account holders
NumOfProducts	Total number of products the customer had availed
HasCrCard	Credit card owned by customers
IsActiveMember	Current status of the customer
EstimatedSalary	Estimated salary of account holders
Exited	Churned customers from the bank

From the bank's data warehouse, using SAS Base, a sample of 10,000 customers from a European bank from the period of January to June 2018 was taken. These customers would churn during the period of July to December 2018. The investigation of this project will primarily focus on consumer behaviour before churning, as was already noted. Binary variables will be used to track customer behaviour and highlight the different actions and inactions that may indicate the likelihood of churn. The value '1' indicates that the customer has churned while the value '0' indicates no churn.

TOOLS USED FOR THIS STUDY

- **PostgreSQL** - A powerful database management system, was utilized to perform cohort analysis in this study on customer churn in the banking industry, allowing for in-depth exploration of customer behaviour over time. PostgreSQL is a potent and well-liked open-source relational database management system, and pg Admin 4 is a free and open-source web-based administration and management application for it. Database administrators, programmers, and other users can easily interact with PostgreSQL databases thanks to its

user-friendly interface.

- **Python** - A versatile programming language, was employed for data visualization in this study on customer churn in the banking industry, enabling the creation of insightful charts and graphs to present key findings and patterns in the dataset.

DATA PREPARATION

The following tasks are performed for data preparation:

- **Data Import** - A new PostgreSQL database and table are created to store the dataset. The bank dataset is then imported into the newly created table using PostgreSQL's COPY command or any other suitable method.
- **Check for missing values** - Identifying and handling any missing data points in the dataset, either by imputing missing values or removing the affected rows/columns.
- **Handle duplicates** - Duplicated and repeated values are eliminated.
- **Data type conversion** - Variables changed to their appropriate data format (e.g., numerical, categorical, date, etc.) for analysis.
- **Handling Categorical Variables** - Converted categorical variables and transformed categorical variables into numerical representations using techniques like label encoding or one-hot encoding, depending on the model's requirements.
- **Scaling and Normalization** - Normalize numerical features, like Credit Score, Age, and estimated Salary, to bring them to a similar scale to avoid domination by features with large ranges.
- **Outlier Detection** - Identify outliers with any extreme values that might influence the analysis and decision-making process.
- **Data Cleaning** - Perform data cleaning steps in PostgreSQL to handle missing values and duplicates

DATA ANALYSIS AND INTERPRETATION COHORT ANALYSIS

The following results were obtained from Cohorting the customers from the available sample:

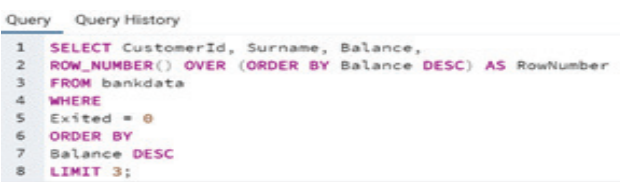
Average Number of Products of Churned Customers	1.5
Average Age of Churned Customers	44.8
Average Balance of Churned Customers	91108.5
Average Credit Score of Churned Customers	645.3

It is obvious that the bank needs to concentrate on enhancing its product offerings given the low average number of items held by churned clients. To better fulfill the demands of clients, this can entail offering new products or improving the functionality of existing ones.

The high average balance and the average age of churned customers suggest that they may have been long-term clients with sizable deposits in the bank. To keep these crucial consumers, it is crucial to concentrate on improving the customer experience. This may entail enhancing customer service, giving individualized services, or introducing loyalty programs.

The fact that churned customers have lower average credit scores than the overall population suggests that creditworthiness is a contributing factor. The bank should concentrate on addressing these issues by giving credit education and counseling to customers in order to help them raise their credit ratings as well as by providing substitute products or services to those who might not fulfill standard credit requirements.

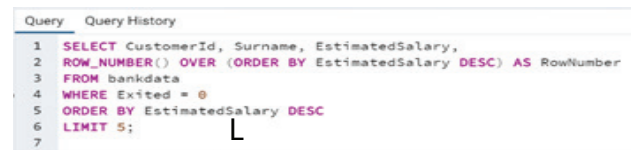
1] TOP 3 CUSTOMERS WITH THE HIGHEST BALANCE:



customerid	integer	surname	character varying (255)	balance	double precision	rownumber	bigint
1	15571958	McIntosh		221532.8		1	
2	15599131	Dilke		214346.96		2	
3	15769818	Moore		212778.2		3	

From the above output, we can easily identify that McIntosh, Dilke, and Moore are the top 3 customers with the highest balance. This way, the bank can utilize post SQL to identify the customers with the highest balance.


2] Top 5 customers with the highest estimated salary:



customerid	integer	surname	character varying (255)	estimatedsalary	double precision	rownumber	bigint
1	15662021	Lucciano		199992.48		1	
2	15634359	Dyer		199970.74		2	
3	15697270	Gannon		199953.33		3	
4	15762331	Moss		199929.17		4	
5	15709136	Adams		199909.32		5	

From the above output, we can easily identify that Lucciano, Dyer, Gannon, Moss, and Adams are the top 5 customers with the highest estimated salary. This way, the bank can use Postgre SQL in order to identify the customers with the highest estimated salary.

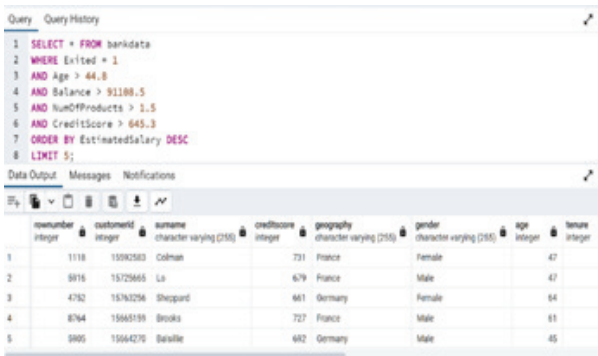
3] TOP 10 CUSTOMERS WITH THE MOST PRODUCTS:



customerid	integer	surname	character varying (255)	numofproducts	integer	rownumber	bigint
1	15744881	Tsai		3		1	
2	15661545	Nicolay		3		2	
3	15671269	Holden		3		3	
4	15616028	Tao		3		4	
5	15732299	Boniwell		3		5	
6	15600651	Ijendu		3		6	
7	15628860	Lee		3		7	
8	15697574	Stewart		3		8	
9	15649182	Johnston		3		9	
10	15711718	Mackenzie		3		10	

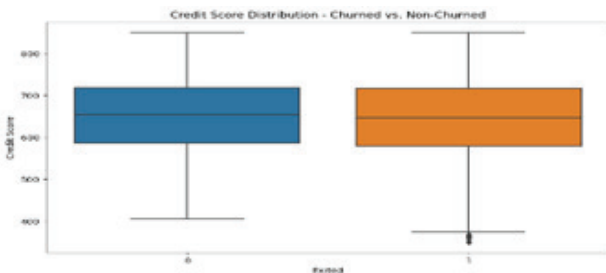
From the above output, we can easily identify that Tsai, Nicolay, Holden, Tao, Boniwell, Ijendu, Lee, Stewart, Johnston, and Mackenzie are the Top 10 customers with the most products. This way, the bank can utilize Postgre SQL in order to identify the customers with most products.

14] TOP 5 CUSTOMERS TO BE OFFERED A REDUCED INTEREST RATE

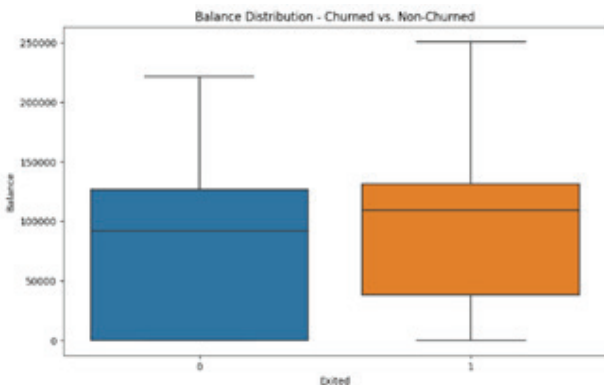


From the above output, we can easily identify that Colman, Lo, Sheppard, Brooks, and Balsillie are the top 5 customers to be offered a reduced interest rate. This way, the bank can utilize the PostgreSQL to identify the top customers to whom the reduced interest rate must be offered.

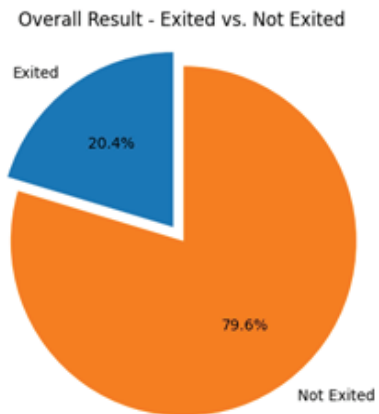
5] DATA VISUALIZATION USING PYTHON



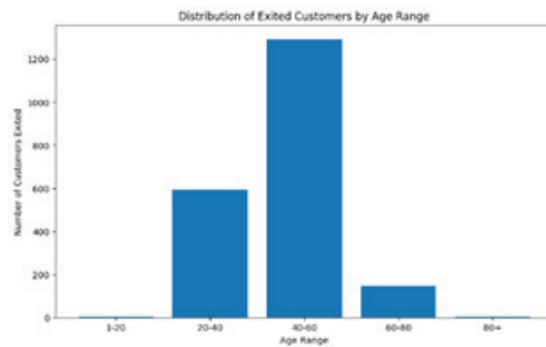
From the above output, we can interpret that most of the customers churn because their credit score falls in the range of 590 to 710, and also the customers who did not churn also fall in the range of 590 to 710. Hence, we can say that there is no contribution of credit score towards customer churn.



From the above output, we can interpret that most of the customers churn because of their balance which falls in the range of 1,00,000 – 1,50,000, and the customers who did not churn fall in the range of 50,000-1,00,000. Therefore, we can easily conclude that the customers who had slightly higher balance exited the bank when compared to those who did not.

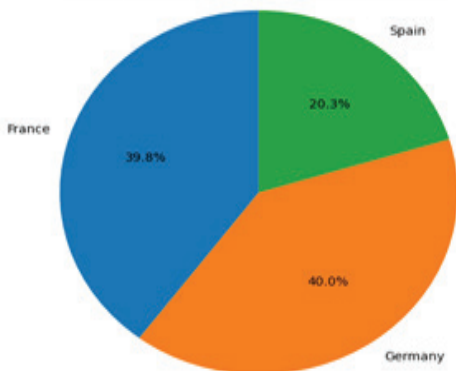


From the above output, we can interpret that nearly 20.4% of the customers existed the bank and 79.6% of the customers did not exit the bank. Therefore, we can conclude that the product and services provided by the bank is good but as to improve in order to retain the existing customers.



From the above output, we can interpret that more than 1200 customers exiting the bank fall in the age group of 40 to 60 years, nearly 600 customers existing the bank fall in the age group of 20 to 40 years and nearly 200 customers exiting the bank belong to the age group of 60 to 80 years.

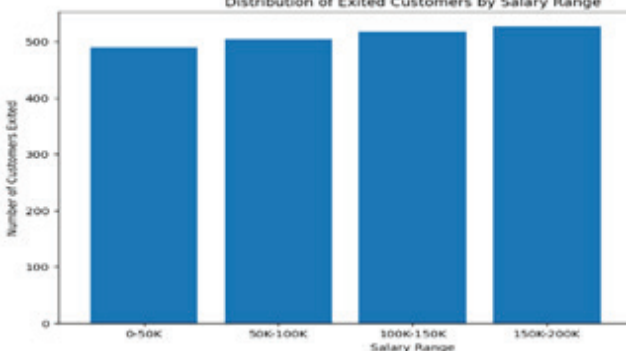
Distribution of Exited Customers by Geography



From the above output, we can interpret that female belonging to the age group of 40 to 60 years as exited the bank more, when compare to male belonging to the age group of 40 to 60 years. From this we can conclude that the bank to take steps in order to retain its female and male customers belonging to the age group of 40 to 60 years.

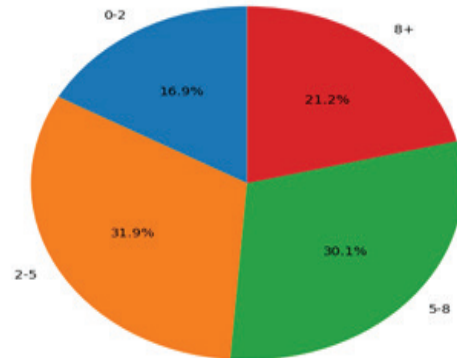
From the above output, we can interpret that 20.3% of the customers who existed the bank belong to Spain, nearly 39.8% of people of customers who exited the bank belong to France and majority of the customers who exited which is nearly 40% belong to Germany.

Distribution of Exited Customers by Salary Range



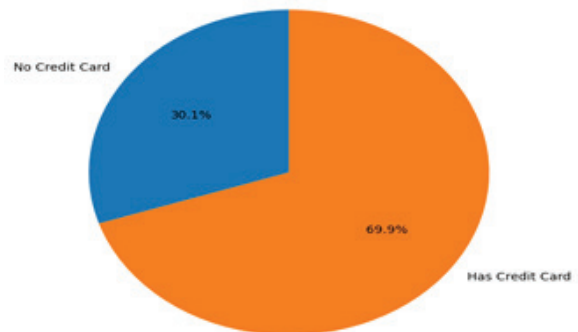
From the above output, we can interpret that most of the customers who's having a salary range of 0K to 50 K and 150 to 200 K have exited the bank, whereas customers salary range of 150K to 200K has exited more when compared to other salary ranges.

Distribution of Exited Customers by Tenure Range



From the above output, we can interpret that nearly 31.9% of customers who had a tenure range of 2 to 5 exited the bank more, whereas the 5 to 8 tenure range has a 30.1% of customers exiting, 0 to 2 tenure range has 16.9% of customers exiting and finally, eight plus tenure range has 21.2% of customers exiting the bank.

Distribution of Exited Customers by Credit Card



From the above output, we can interpret that 69.9% of customers who exited the bank had credit card and 30.1% of the customers who exited the bank did not have credit cards, So we can conclude that the majority of the customers who exited the bank had credit cards.

FINDINGS, LIMITATIONS, AND SUGGESTIONS

FINDINGS

- The bank should focus on improving its product offerings as churned clients tend to have a low average number of items held. This can involve introducing new products or enhancing the functionality of existing ones to better meet customer needs.
- The high average balance and the average age of churned customers suggest that To retain long-term clients with substantial deposits, the bank must focus on improving the customer experience, including enhanced customer service, personalized services, and loyalty programs.
- The fact that churned customers have lower average credit scores, the bank should offer credit education and counselling to customers, helping them improve their credit scores. Additionally, providing alternative products or services for those not meeting standard credit requirements can mitigate churn risks.
- Credit score of exited falls in the range of 590 to 710 which is similar to not exited. So, the credit score doesn't significantly impact churn.

SUGGESTIONS

- The bank needs to Improve Product Offerings by Offering personalized products, by Offering competitive rates and fees, focusing on convenience and by Provide financial education.
- Enhance Customer Experience by Simplify the processes, Personalized communication, by Using customer feedback, and by Providing proactive customer service.
- Addressing the Creditworthiness Concerns by Verifying income and Employment, by Monitoring credit behavior, using credit insurance, collaterals or guarantees.
- Nearly 79.6% of the customers did not exit the bank, but the bank needs to enhance its Product/service quality to retain churned customers by 20.4%.
- More than 1200 plus customers, who exited the bank belonged to the age group of 40 -60 years,

then followed by customers with the age group 20 - 40 years and 60-80 years.

- Female and male customers who churned more belong to the age group 40-60 age group. the bank needs to take steps in order to retain its female and male customers of that age group.
- Majority of the customers who exited belong to Germany which was nearly 40%, followed by France which is 39.8% and then by Spain 20.3%.
- Customers salary range of 150K to 200K has exited more when compared to other salary ranges.
- Tenure range of 2 to 5 exited the bank more i.e., nearly 31.9% of customers. followed by 5 to 8 tenure range with 30.1% churn.
- 69.9% of churned customers had credit cards. 30.1% of the churned customers did not have credit cards.
- credit score range of 550 to 650 and 650 to 750 exited the bank irrespective of whether they were credit cardholders or non-credit cardholders.

LIMITATIONS

Granular data, or specific information about specific consumers and their behaviors, is necessary for cohort analysis. Banks may contain a tonne of data, and as the data becomes more detailed, the database size increases and query performance decreases. This may require optimization methods and slow down the analysis performance.

Cleansing and Integrity of Data for cohort analysis depends on high-quality data. Results can be misled by incomplete or inconsistent data. It can be difficult to ensure data purity and integrity, especially when working with data from several sources inside the bank. For churn analysis to detect long-term trends,

historical data is frequently needed. The amount of historical data that is currently accessible may be constrained based on the bank's data retention policy, making it challenging to do substantial long-term analysis.

As sensitive consumer information is handled by banks, it is crucial to protect data privacy and security. Data management must be done carefully while conducting cohort analysis in order to avoid security breaches or unauthorized access.

All pertinent contextual information that could affect churn rates may not be included in cohort analysis. There may be omissions of elements like customer encounters, support requests, or consumer feedback, which limits the depth of insights.

CONCLUSION

Customer churn, also known as customer attrition, is the rate at which customers discontinue their relationship with a company or business. customer churn leads to reduced revenue and profitability for the bank. Lost customers mean lost potential income from fees, interest, and other banking services. This can be particularly impactful when high-value customers, such as those with substantial deposits or loans, churn. As customers leave for competing banks, the bank may lose its competitive edge and struggle to attract new customers.

Finally, customer churn can damage the bank's reputation and trust among customers. Frequent churn signals dissatisfaction and can deter potential customers from choosing the bank

So, through this study, we were able to focus on customer churn in the banking industry. We performed cohort analysis using Postgre SQL to identify the average number of products of churned customers, the average age of churned

customers, the average balance of churned customers, and the average credit score of churned customers. Later from the results we concluded that the bank should focus on improving its product offerings as churned clients, the bank must focus on improving the customer experience, and also should offer credit education and counselling to customers.

Finally, we transferred the structured data from Postgre SQL to Python for data visualization and to understand the behaviour of the exited customers due to different factors, such as customer age, credit score, tenure, number of products, and geographical location. In the end, we have recommended a few suggestions like Offering personalized products, providing proactive customer service, using customer feedback, monitoring credit behavior, etc to retain its customers.

REFERENCES

- [1] Linares-Mustarós, R. P., García-Sánchez, J. A., & Segovia-Vargas, M. V. "Customer Churn Prediction in Retail Banking: A Data Mining Approach."
- [2] Abdulaziz, M., & Muhammad, A. N. "Customer Churn Prediction in Retail Banking: A Data Mining Approach Using R."
- [3] Mukherjee, B., Sarkar, S., & Sarkar, D. "Predicting Customer Churn in Banking Industry Using Ensemble Learning."
- [4] Abbas, S., & Qi, S. "A Review of Customer Churn Prediction in Telecommunication Industry."
- [5] Mathur, R., & Tayal, D. "Churn Prediction in Telecom Industry Using Advanced Data Mining Techniques."
- [6] Khanam, A., & Zulkernine, F. H. "Churn Analysis and Prediction in Telecommunication Industry Using Machine Learning Techniques."
- [7] Chris820629. "A PostgreSQL-based study on banking churn analysis."
- [8] Kaur, & Kaur. "Customer Churn in the Banking Industry: A Literature Review."