PREDICTION OF CUSTOMER CHURN IN THE BANKING INDUSTRY

Created by:

Walter Edward

Introduce About the Company

TCB is one of the leading financial institutions in Vietnam, established in 1993. With more than 25 years of operation, TCB has built a strong position. in the banking industry by providing diverse products and services to meet the financial needs of individuals and businesses.

TCB stands out with its modern business model and innovation in technology, being one of the pioneer banks in applying digital technology to banking operations. This includes providing convenient, safe and efficient online banking, mobile and digital payment services.

In addition, TCB is always committed to bringing value and satisfaction to customers through quality products and services, while supporting customers in financial and investment activities. TCB also actively contributes to social and community development through charitable activities and social support programs.

With the vision of becoming one of the leading banks in Vietnam, TCB constantly improves service quality and innovation to meet the increasingly diverse needs of customers and promote the sustainable development of the economy. country economy.

TABLE OF CONTENTS

01

Preparation

02

Exploratory data analysis

03

PREDICTING

04

SUGGESTION

01 Preparation

Problem Statement

Customer retention is as crucial as customer acquisition when it comes to increasing revenue. Also we know, it is much more expensive to sign in a new client than keeping an existing one. So, churn rate needs to be minimized.
To do so it is necessary for banks to know what leads a client towards the decision to leave the bank. Also churn prediction allows companies to develop loyalty programs and retention campaigns to keep as many customers as possible

Objective

This project involves analyzing the customer churn rate in order to understand why customers leave and then building a classification model to predict whether a customer will churn out or not so that proper actions can be taken to prevent this churn.

Data Understand

Data contains:

- 14 Features
- 10000 rows
- There are no Null values

Column	Non-Null Value	Data Type
RowNumber	10000	Int64
CustomerID	10000	Int64
Surname	10000	object
Credit Score	10000	int64
Geography	10000	object
Gender	10000	object
Age	10000	int64
Tenure	10000	int64
Balance	10000	float64
NumOfProducts	10000	int64
HasCrCard	10000	int64
IsActiveMember	10000	int64
EstimatedSalary	10000	float64
Exited	10000	int64

Data Understand

Because we are studying customer characteristics to predict whether customers will decide to leave or not. So it seems like features like RowNumber, CustomerId, and Surname don't contribute much to the analysis, and make the model more cumbersome. Decided to drop them from the data frame, to facilitate analysis.

Dataset after processing contains:

- 11 Features
- 10000 rows
- There are no Null values

Column	Non-Null Value	Data Type
Credit Score	10000	int64
Geography	10000	object
Gender	10000	object
Age	10000	int64
Tenure	10000	int64
Balance	10000	float64
NumOfProducts	10000	int64
HasCrCard	10000	int64
IsActiveMember	10000	int64
EstimatedSalary	10000	float64
Exited	10000	int64

Data Processing

Feature Engineering Processing: Age_Group

- Students (18-22 years old): Need products such as student credit cards, savings accounts with preferential interest rates, fast money transfer services, and easy-to-use mobile applications.

- Young Professionals (23-30 years old): There is a higher demand for personal loans, credit cards with higher limits, investment products, and insurance.

- Mid-career Individuals (31-45 years old): Often interested in home loans, car loans, family insurance, and education plans for children.

- Pre-retirement Individuals (46-60 years old): May be more interested in retirement investment plans, health insurance, and wealth management.

- Early Retirement (60-75 years old): There is still demand for low-risk investments, premium health insurance, and savings accounts with good interest rates.

- Advanced Seniors (over 75 years old): Can focus on asset management, asset transfer, and banking products with more dedicated support and advice.

Values	Age
Students	Under 22
Young Professionals	From 22 to under 30
Mid-career Individuals	From 30 to under 45
Pre-retirement Individuals	From 45 to under 60
Early Retirement	From 60 to under 75
Advanced Seniors	Above 75

Data Processing

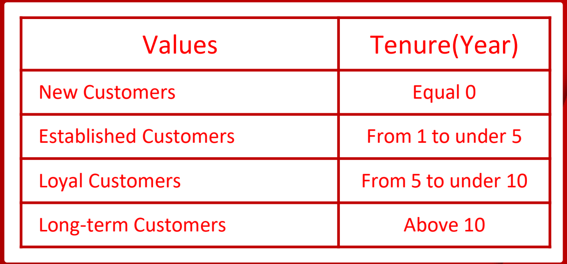
Feature Engineering Processing: Tenure_Seg

1. New Customers (New Customers)
- The first stage of the relationship when customers are still learning, discovering and using products and services.

2. Established Customers
- The stage when customers are familiar with the products/services and begin to feel comfortable and consider expanding relationships. For example, investing more in new products or services.

3. Loyal Customers (Loyal Customers)
- The stage where customers are very familiar and have a solid relationship with the bank. Customers begin to evaluate long-term relationships and the sustainability of services.

4. Long-term Customers (Long-term Customers)
- Customers have deep relationships and regularly interact with the bank in important services. A period when customers are extremely loyal and often enthusiastic advocates for the bank, possibly considering privilege or VIP programs.



Values	Tenure(Year)
New Customers	Equal 0
Established Customers	From 1 to under 5
Loyal Customers	From 5 to under 10
Long-term Customers	Above 10

Data Understand

Dataset after processing contains:

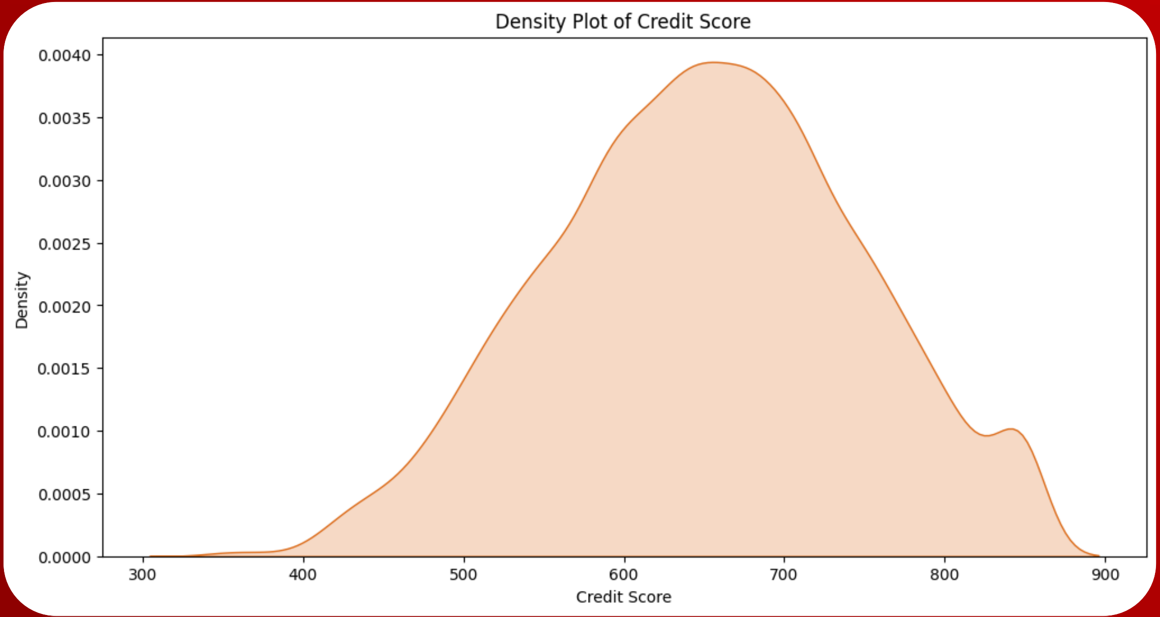
- 11 Features
- 10000 rows
- There are no Null values

Column	Non-Null Value	Data Type
Credit Score	10000	int64
Geography	10000	object
Gender	10000	object
Age_Groups	10000	object
Tenure_Segs	10000	object
Balance	10000	float64
NumOfProducts	10000	int64
HasCrCard	10000	int64
IsActiveMember	10000	int64
EstimatedSalary	10000	float64
Exited	10000	int64

02

EDA

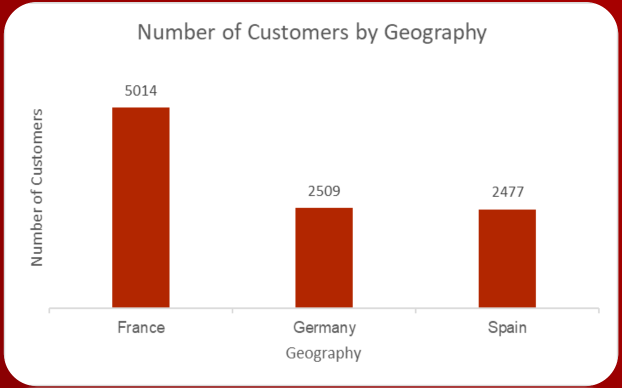
Credit Score



Insights:

- The average credit score of customers in the data sample is 650, with the 25% and 75% percentiles being 584 and 718 respectively. The median is 652, close to the mean(650), showing the symmetrical, and even, distribution of credit scores in the data sample.
- The standard deviation of the credit score is 96. Indicates large fluctuations in the credit scores of customers in the data sample.
- There is a significant number of customers with credit scores lower than the 25% percentile (2534 customers). Banks may face higher risks in repaying debt, as these customers have poor repayment ability and therefore may be more likely to churn.
- There is also a significant portion of customers with credit scores higher than the 75% percentile (2501 customers). These can be considered elite customers with good debt repayment ability, banks may need to focus on maintaining relationships and providing good services to retain them, as this can be a stable source of income. set for the bank.

Geography



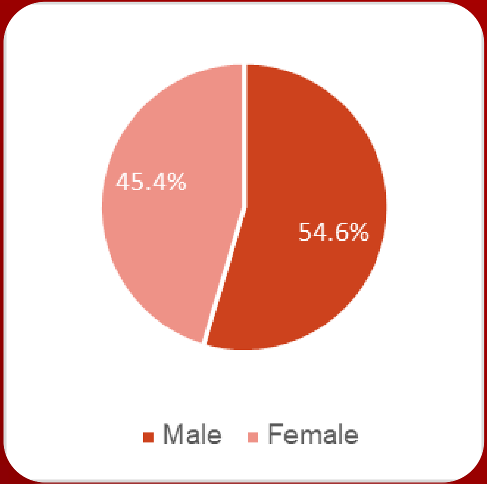
- France is the country with the highest percentage of customers (5014 customers), half of the total number of customers.

- The number of customers from Spain (24.8%) and Germany (25.1%) are relatively close to each other, but Germany seems to be a little larger than Spain.

- Due to the large number of customers coming from France, this country can play an important role in the bank's marketing and customer retention strategies.

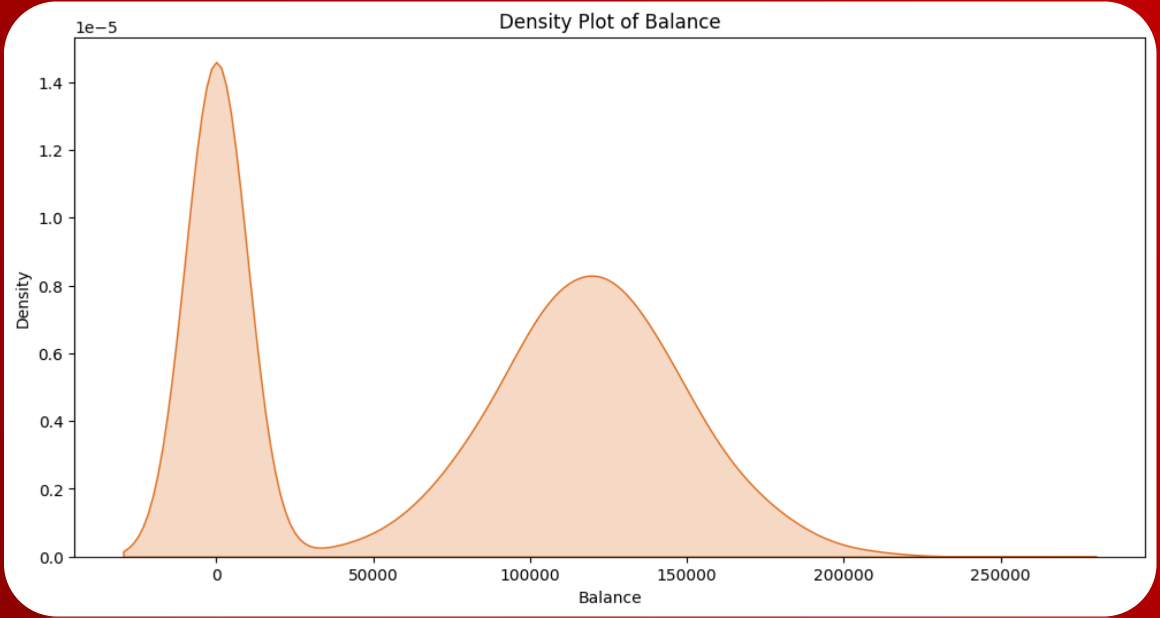
- However, customers from Spain and Germany should not be overlooked, as they still represent a significant proportion of the total and can offer important market potential.

Gender



- The total number of male customers accounts for the majority of the data sample (5460 customers), with a rate of 54.6%, the remaining 45.4% belongs to women.
- Based on the data, it may be necessary to focus on special marketing strategies to retain male customers, while attracting more female customers, to balance the ratio between male and female customers, as both can be an important potential market.

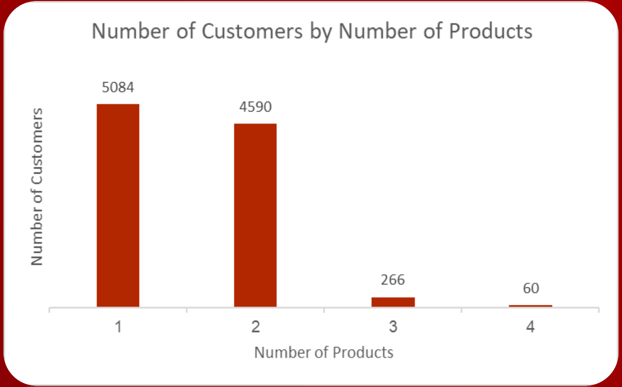
Balance


Density Plot of Balance
1e−5
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Density
0
50000
100000
150000
200000
250000
Balance

Insights:

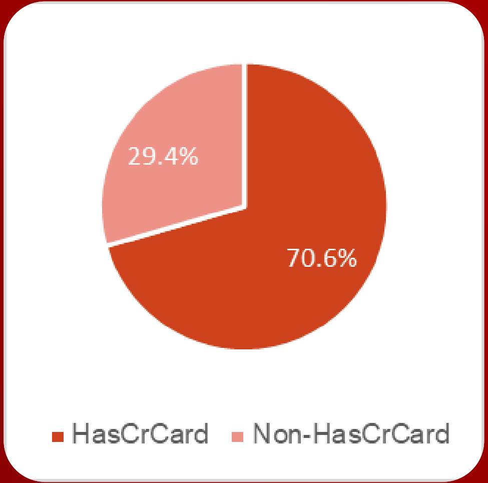
- The average customer balance in the data sample is $76,485.
- 25% of customer balances in the sample are 0. This indicates a large number of customers with no balance (3617 customers, 36% of the total). This affects the bank's revenue, because these customers do not generate profit from interest, but the bank still has to pay operating costs to maintain these accounts such as customer support costs. , infrastructure costs, especially when this number of customers is large, the cost will be significant for the bank.
- Median is $97198 and 75% of the customer balance in the data sample is $127644.
- The standard deviation of the balance is $62,387. This indicates large fluctuations in customer account balances in the data sample.
- The distribution of account balances is uneven, with a large portion of customers having zero balances and a few having higher balances.
- You should focus on attracting and retaining customer groups with high balances. At the same time, pay attention to providing appropriate services and products for customer groups with low balance or no balance.

Number of Products



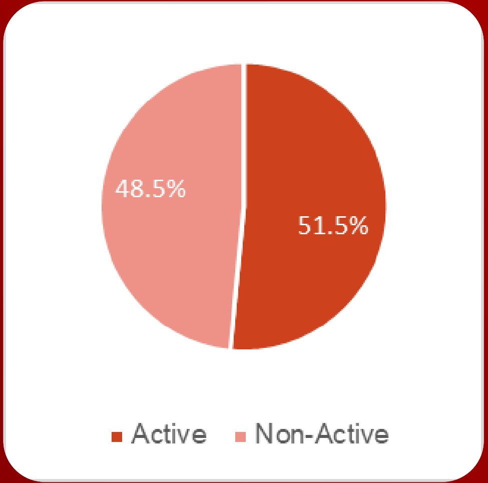
- The majority of customers (96.7%) only use 1 (5084 customers) or 2 (4590 customers) products.
- The number of customers using 3 or 4 products is quite small, accounting for only 3.3% of the total number of customers in the data sample.
- Can focus on developing and promoting new product packages for customers to use more products. Optimize service and support to enhance customer experience and strengthen relationships.

Has Credit Card



- The proportion of customers using credit cards is quite high, accounting for nearly 71% of the total number of customers in the data sample.
- Can focus on developing, improving quality and promoting credit card-related services to optimize customers' card usage. Marketing strategies can focus on attracting new customers and strengthening relationships with existing customers through incentives and support services such as customer care, card customer support 1 effective and safe way.
- On the banking side, there needs to be effective risk management policies and procedures related to the issuance and use of credit cards.

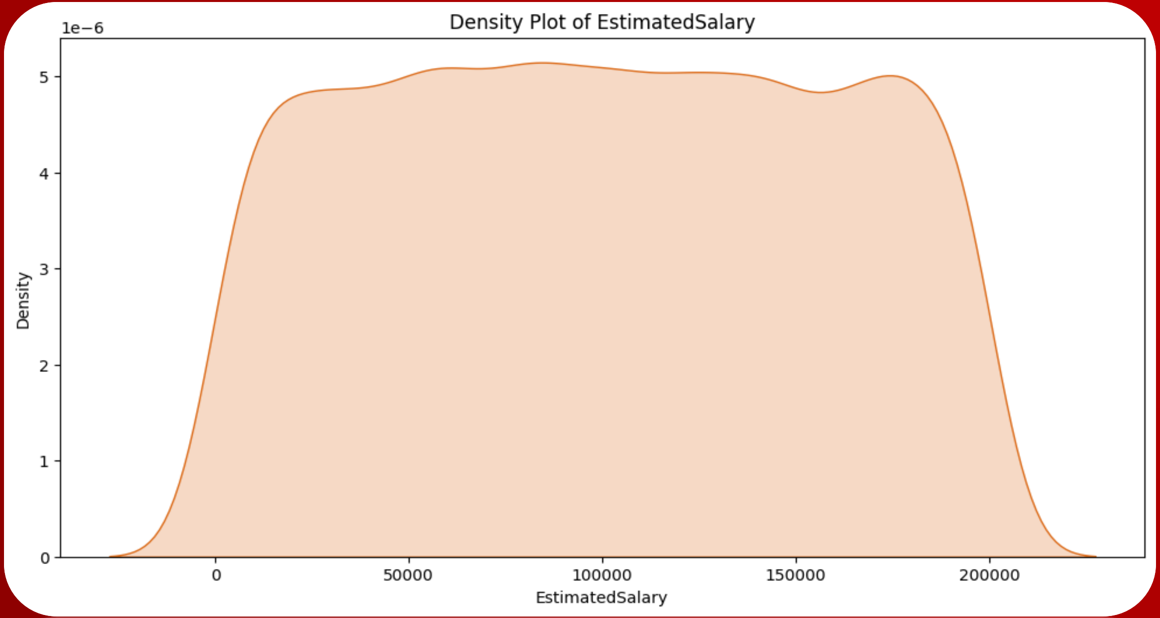
Is Active Member |



- The proportion of active customers is 51.5%, accounting for more than half of the total number of customers in the data sample.

- Banks should focus on maintaining and strengthening relationships with active customers to retain them and increase sales. At the same time, pay attention to stimulating activity for inactive customers through promotions or customer care services.

Estimated Salary



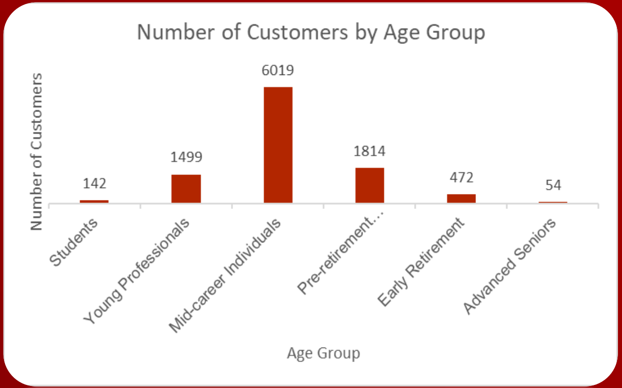
Insights:

- The median income of customers in the sample is $100,090, with the 25% percentile being $51,002 and the 75% percentile being $149,388. The median is $100,193, which is close to the median ($100,090), indicating that income in the data sample is symmetrically distributed.

- The standard deviation of income is $57,510. This indicates large fluctuations in the income levels of customers in the data sample.

- Can focus on strengthening relationships with high-income customers, to optimize sales and profits for the bank.

Age Group

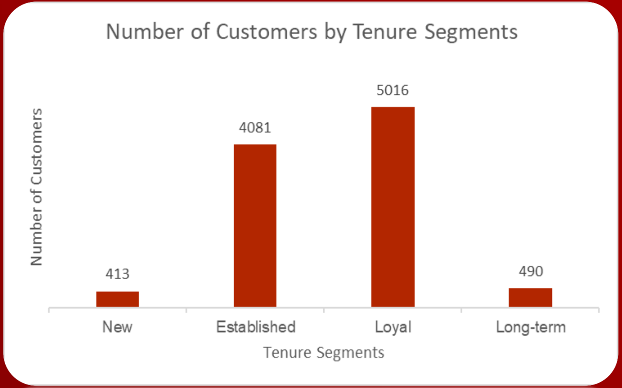


- There is a large number of customers in the age group from 23 to 60, accounting for about 93% of the total number of customers in the data sample.

- The remaining age groups have a relatively small number of customers, but still contribute to the diversity of the data sample.

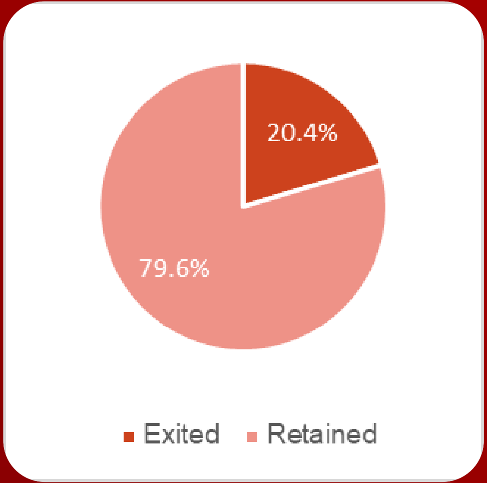
- Marketing strategies tailored to each age group can be developed to optimize customer experience and strengthen relationships. It is necessary to conduct market research to better understand the needs and priorities of each age group, thereby providing a basis for developing appropriate products and services.

Tenure Segments



- The majority of customers in the data sample belong to the 'Established' and 'Loyal' groups, accounting for more than 90% of the total number of customers.
- The number of customers in the 'New' and 'Long-term' groups is relatively small compared to the total number of customers in the sample (about 9%).
- There is a need to focus on maintaining and strengthening relationships with ‘Loyal’ and ‘Established’ customers, possibly through promotions and customized services. At the same time, pay attention to attracting and retaining 'New' customers, and maintaining relationships with 'Long-term' customers.

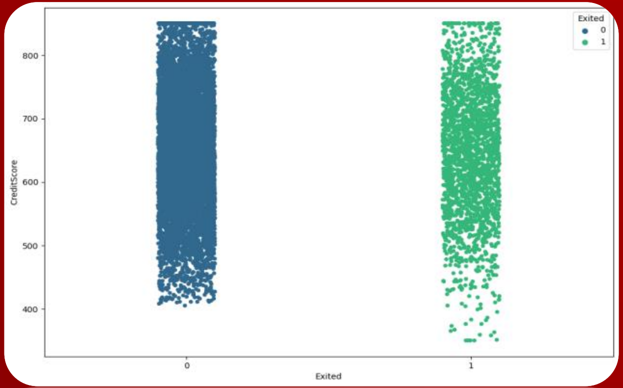
Exited



- The proportion of retained customers accounts for the majority of the data sample (nearly 80%), and the group of customers who have left accounts for 20.4% of the total number of customers.

- Understanding the "exited" target variable helps banks predict and prevent customer loss. Therefore, it is necessary to focus on developing customer retention strategies, as well as improving customer experience to minimize churn rates. Customer care is also an important factor to create a good environment for customer retention. Besides, banks also need to attract new customers.

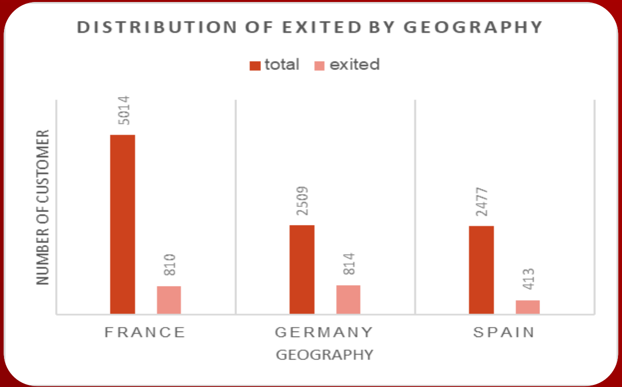
Credit Score and Exited



- Customers who stay (exited = 0): have the densest data density (80% of customers stay), credit score is evenly distributed, and close to each other. It shows that the customer's credit score is quite stable and fluctuates little from 400 to 900. However, credit score tends to concentrate from 500 to 800 points.

- Customers leaving (Exited = 1): the credit scores of customers leaving are evenly distributed and close to each other, they are concentrated between 500 and 800 points. Some customers have high credit scores (over 800 points), but still leave, and customers with low credit scores (under 400), often tend to leave.

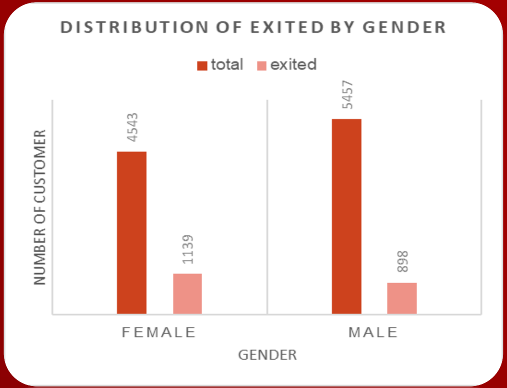
Geography and Exited



- The number of exited customers is mostly concentrated in France (810) and Germany (814), while in Spain it is less. However, when looking at each country, France has up to 5014 customers, 810 people leaving is only about 16% of the total in this country. But in Germany, there were 814 customers leaving, while there were only 2,509 customers, accounting for nearly 1/3 of the total number of customers here.

- Thereby, Germany has a very high customer churn rate. Further analysis may be needed to better understand the causes and factors that influence customers' decisions to leave, thereby taking measures to reduce churn rates in different countries.

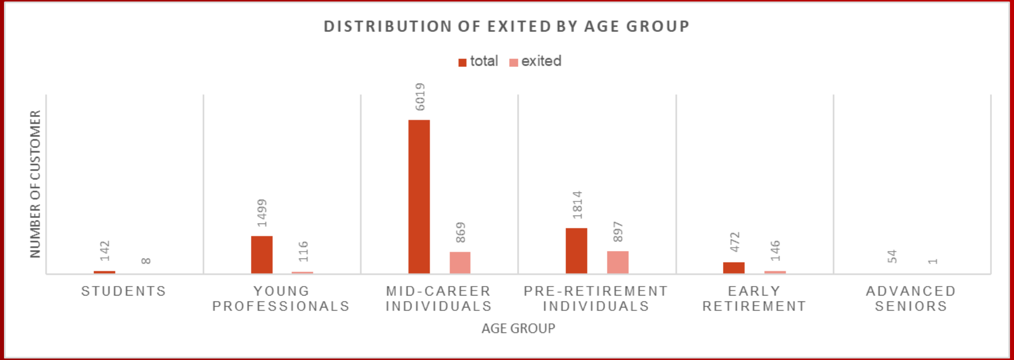
Gender and Exited



- From the above data, it can be seen that there is a certain difference between the number of male and female customers. Although Nam has a larger number of customers, the churn rate is lower. Females are a group of customers with a higher tendency to leave bank services.

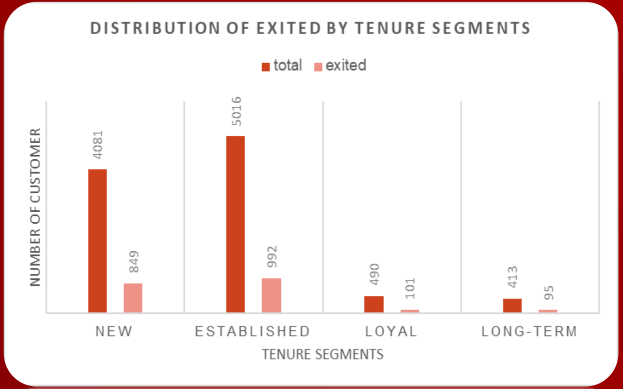
- Further analysis may be needed to better understand the causes and influencing factors to develop customized marketing and customer care strategies appropriate for each gender.

Age Group and Exited



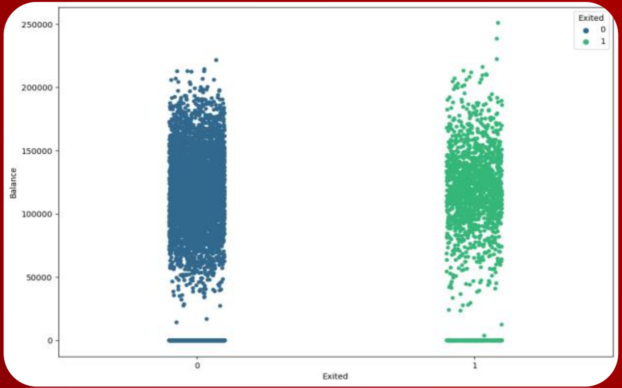
- The rate of exited customers is especially high in the Pre-retirement Individuals group (46-60 years old), with a rate of 50% out of a total of 1814 customers, and Early Retirement (60-75 years old), about 1 part 3 out of a total of 472 customers.
- This may reflect changing trends in customer needs and priorities with age.
- To reduce customer churn, banks can focus on developing products and services that suit the needs and priorities of each age group, especially from 46 to 75 years old.

Tenure Segments and Exited



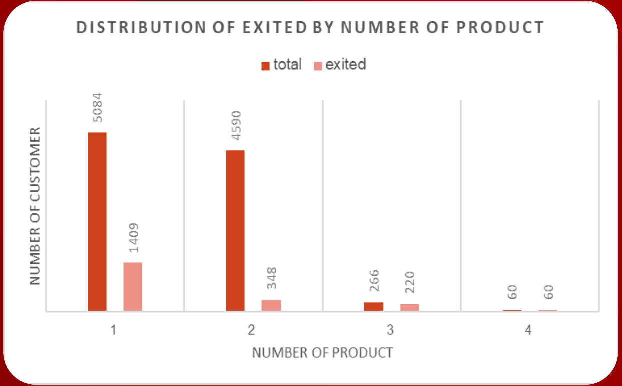
- The percentage of customers who have left is about 20% of the total 4081 'Established' customers, 20% of the total 5016 'Loyal' customers, 20% of the total 490 'Long-term' customers, and about 23% out of a total of 413 'New' customers.
- From the above data, we see that the rate of customers who have left is higher in the 'New' customer group when compared to the remaining groups. This may reflect customer stability and loyalty over time.
- To reduce customer churn, banks can focus on providing customized incentives and services to retain customers, especially in the beginning when customers are new to using the bank's services. row.
- Banks also need to establish customized marketing and customer care strategies for each service period to optimize customer experience and reduce churn rates.

Balance and Exited



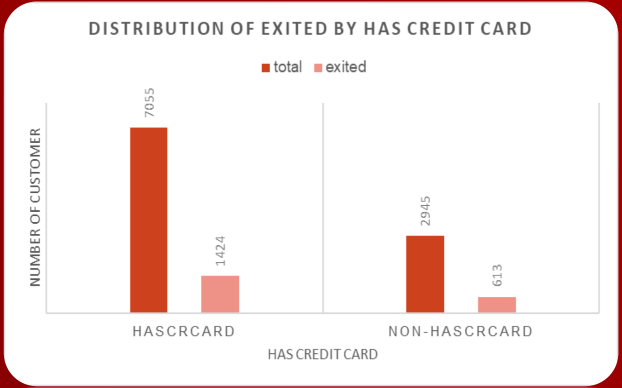
- Customers who stay (exited = 0): have the densest data density (most customers stay), the balance is evenly distributed, and close to each other. It shows that the customer's balance is quite stable and fluctuates little from $35,000 to $200,000. However, the balance tends to concentrate between $50,000 and $200,000.
- Exited customers (Exited = 1): the balance of leaving customers tends to be concentrated, stable and less volatile in the range of 75,000 to 175,000 USD. However, some customers with high balance (over $200,000) still leave, and customers with low balance (under $50,000) often tend to leave.

Number of Products and Exited



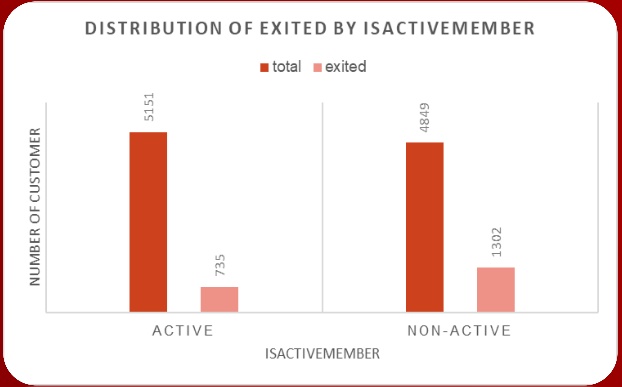
- About 97% of customers only use 1 to 2 bank products.
- Specifically, for customers who use 1 product, their churn rate is 27% out of a total of 5084 customers. And for those using 2 products, the churn rate was 7.5% across a total of 4,590 samples.
- From the above data, it can be seen that the greater the number of products, the lower the customer churn rate. This may indicate that customers who use multiple products are more likely to retain, or have a strong attachment to, the bank.
- To reduce customer churn, banks can focus on increasing promotion of complementary products or strengthening relationships with existing customers to increase the number of products used.

Has Credit Card and Exited



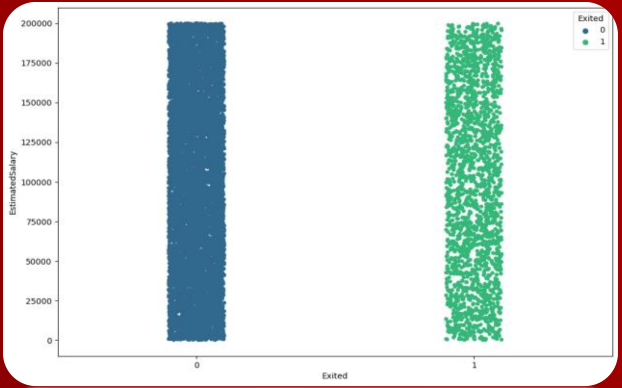
- The rate of customer departure is 20% of the total 2945 customers without credit cards, and 20% of the total 7055 customers with credit cards.
- From the above data, we see that the percentage of customers who have left is the same in both the group without a credit card and the group with a credit card (both 20%).
- This may indicate that the presence of a credit card is not an important factor in customer retention, but may encourage customers to use additional services such as credit cards to maximize revenue for the bank.

Is Active Member and Exited



- The percentage of customers who have churned is higher in the inactive group than in the active group.
- This indicates that customer activism can be an important factor in customer retention.
- To reduce customer churn, banks can focus on increasing customer engagement and operations, as well as providing customized offers and services to the inactive group.
- Customized customer care and marketing strategies need to be established for each active and inactive group to optimize customer experience and reduce churn.

Estimated Salary and Exited



Customers staying (exited = 0) and Customers leaving (exited = 1) both have Estimated Salary evenly distributed, and close to each other.

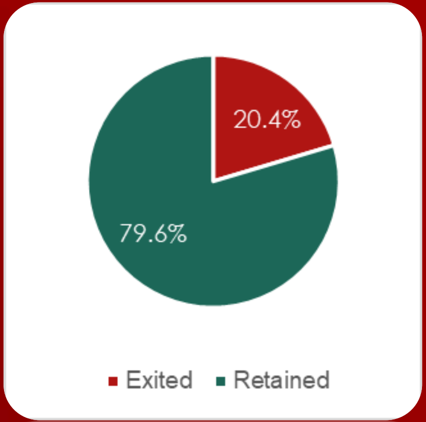
It shows that no matter what a customer's estimated salary is, from $0 to $200,000, they are likely to leave.

EstimatedSalary of the staying customer group (exited = 0), has denser data, just because the number of staying customers is greater.

03 Predicting

Data Validation and Preprocessing

Check Balance

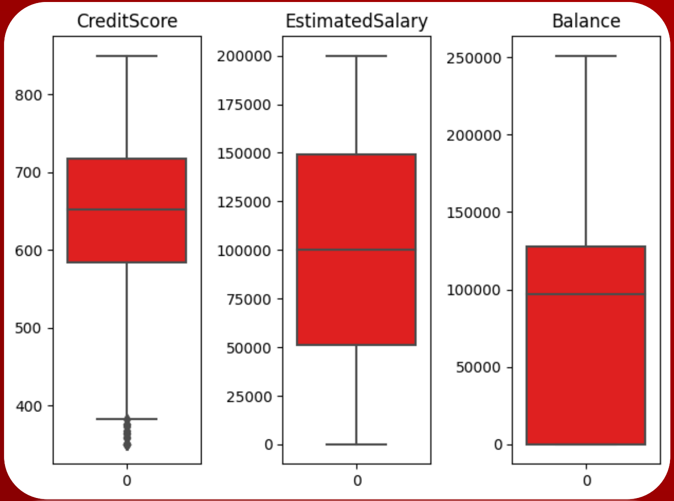


Data includes:

- 80 percent of the data belongs to the stay layer.
- There is an imbalance in the data. Not leaving is more than leaving

Data Validation and Preprocessing

Check Outliers

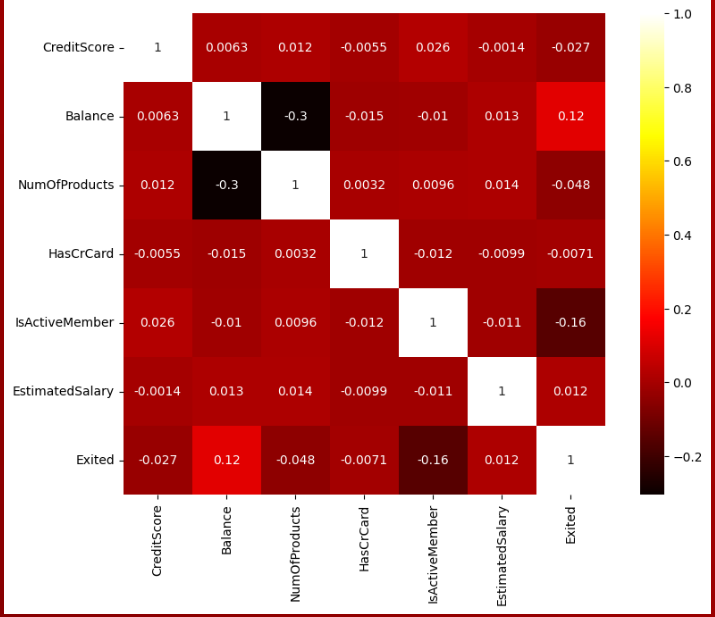


The data has very few outliers

=> Does not have much effect on model training.

Data Validation and Preprocessing

Check Correlation



	CreditScore	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
CreditScore	1	0.0063	0.012	-0.0055	0.026	-0.0014	-0.027
Balance	0.0063	1	-0.3	-0.015	-0.01	0.013	0.12
NumOfProducts	0.012	-0.3	1	0.0032	0.0096	0.014	-0.048
HasCrCard	-0.0055	-0.015	0.0032	1	-0.012	-0.0099	-0.0071
IsActiveMember	0.026	-0.01	0.0096	-0.012	1	-0.011	-0.16
EstimatedSalary	-0.0014	0.013	0.014	-0.0099	-0.011	1	0.012
Exited	-0.027	0.12	-0.048	-0.0071	-0.16	0.012	1

Through the process of examining the correlation of attributes, there are two factors that have the most impact on our Exited target variable in a positive and negative direction.

Among them, ranked first is isActiveMember (-16%), and second is Balance (12%).

We will focus more on these two factors when performing the next work.

Label Encoding for Each Object Data Type

Before Label Encoding

Column	Value
Geography	['France’, 'Germany’, 'Spain’]
Gender	['Female’, 'Male']
Age Group	[‘Students’, ‘Young Professionals’, ‘Mid-career Individuals’, ‘Pre-retirement Individuals’, ‘Early Retirement’, ‘Advanced Seniors’]
Tenure Seg	['New’, ‘Established’, ‘Loyal’, ‘Long-term’]

After Label Encoding

Column	Value
Geography	[0, 1, 2]
Gender	[0, 1]
Age Group	[0, 1, 2, 3, 4, 5]
Tenure Seg	[0, 1, 2, 3]

Use the sklearn.preprocessing.LabelEncoder() library to encode labels whose data type is object in the data.

Train Test Split

Column	Shape
Training feature set size	(12740, 10)
Test feature set size	(3186, 10)
Training variable set size	(12740,)
Test variable set size	(3186,)

Because our target variable y has imbalanced data (retained accounts for 80%, and exited accounts for 20%), we use smote to handle the imbalanced data. Then use the train_test_plits library to divide the data into the train and test set at a ratio of 80:20, random_state=42, and stratify according to variable y.

Model Logistic Regression

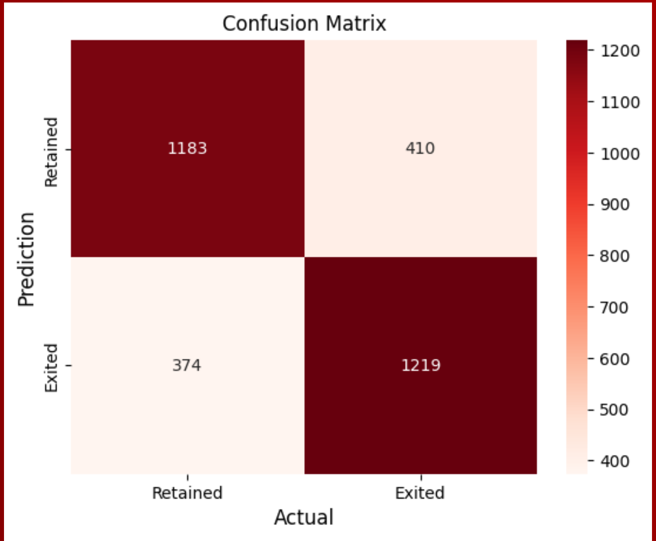
Hyperparameter	Best value
C	0.01
solver	newton-cg
fit_intercept	True
penalty	l2
tol	0.0001

Use the GridsearchCV library to test and find the most suitable hyperparameters for the model.

Then use these hyperparameters to train the model and make predictions.

Model Logistic Regression

Evaluating



	Accuracy	Precision	Recall	F1-Score
Values	0.75	0.75	0.77	0.76

- Accuracy = 0.75: the model predicts correctly about 75% of all cases.
- Precision = 0.75: 75% of customers predicted to churn actually churn. The rate of correctly predicting customer churn is quite good.
- Recall = 0.77: the model can capture about 77% of the total number of customers who have left.
- F1-Score = 0.76: the model has fairly consistent performance between accurately predicting churn cases and catching a high proportion of these cases.

Model Random Forest Classification

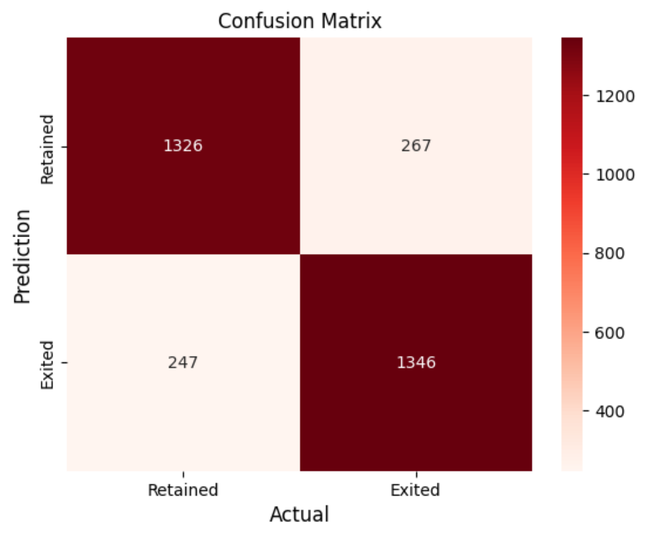
Hyperparameter	Best value
n_estimators	500
criterion	entropy
max_depth	20
min_samples_split	2
min_samples_leaf	1
max_features	sqrt
bootstrap	True
class_weight	balanced
random_state	42

Use the GridsearchCV library to test and find the most suitable hyperparameters for the model.

Then use these hyperparameters to train the model and make predictions.

Model Random Forest Classification

Evaluating



	Accuracy	Precision	Recall	F1-Score
Values	0.84	0.83	0.84	0.84

- Accuracy = 0.84: the model predicts correctly about 84% of all cases.
- Precision = 0.83: 83% of customers predicted to churn actually churn. The rate of correctly predicting customer churn is quite good.
- Recall = 0.84: the model can capture about 84% of the total number of customers who have left.
- F1-Score = 0.84: the model has fairly consistent performance between accurately predicting churn cases and catching a high proportion of these cases.

Model Extreme Gradient Boost (XGBoost) Classification

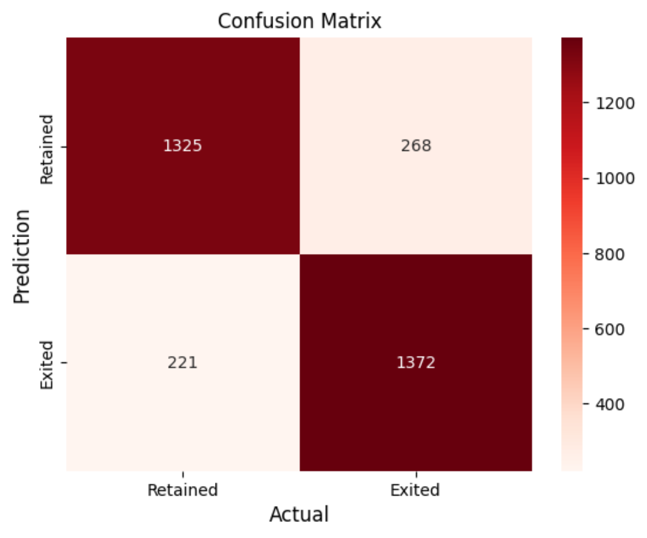
Hyperparameter	Best value
learning_rate	0.2
n_estimators	500
max_depth	10
min_child_weight	1
gamma	0.01
reg_alpha	0
reg_lambda	0.1

Use the GridsearchCV library to test and find the most suitable hyperparameters for the model.

Then use these hyperparameters to train the model and make predictions.

Model Extreme Gradient Boost (XGBoost) Classification

Evaluating



	Accuracy	Precision	Recall	F1-Score
Values	0.85	0.84	0.86	0.85

- Accuracy = 0.85: the model predicts correctly about 85% of all cases.
- Precision = 0.84: 84% of customers predicted to churn actually churn. The rate of correctly predicting customer churn is quite good.
- Recall = 0.86: the model can capture about 86% of the total number of customers who have left.
- F1-Score = 0.85: the model has fairly consistent performance between accurately predicting churn cases and catching a high proportion of these cases.

Evaluate the model's prediction results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.75	0.75	0.77	0.76
Random Forest Classification	0.84	0.83	0.84	0.84
Extreme Gradient Boost	0.85	0.84	0.86	0.85

- accuracy: random forest and xgboost models both have higher accuracy than logistic regression. The higher the accuracy of the model, the better the model proves.
- Precision: the random forest and xgboost models both have higher precision than logistic regression, showing their ability to correctly predict customer churn.
- Recall: xgboost has the highest recall, showing that it is capable of finding more customer churn than the other two models.
- F1-score: xgboost model has the highest f1-score, achieving the best balance between precision and recall

=> based on these indices, the extreme gradient boost (XGBoost) model shows the best performance in predicting customer churn in the banking sector, with high accuracy, good precision and recall, along with f1- stable score. Random forest also has good performance, but XGBoost is the top choice based on current metrics. We can continue to refine and use the XGBoost model to predict customer churn with the best performance.

04 Suggestion

Suggestion for BOD

Customers with low credit scores (below 400) often tend to leave, banks should give recommendations so customers can improve their credit scores, including making specific suggestions such as spending reasonable, repay on time and maintain a positive credit history.

Germany is the country with the highest churn rate (33%), further research is needed to better understand the causes of high customer churn rates in Germany, due to service not meeting expectations, spending High fees, fierce competition, or poor customer support,... then take appropriate measures to reduce the churn rate.

Women are a group of customers with a higher tendency to leave banks than men. Further research may be needed to better understand the causes, thereby developing customized customer care products and services suitable for each gender, especially female customers.

Customers with high balance (over 200k) and low balance (under 50k) often tend to leave. Maybe the products and services are not suitable for these two customer groups. Banks should increase communication and personalized consulting to better understand each customer's needs and financial goals, helping the bank provide more suitable financial solutions and products. For example, for customers with high balance, banks can recommend investment packages or high-yield assets. For customers with low balance, flexible savings products or suitable credit packages can be provided.

Customers who use more products have higher retention rates. To reduce the customer churn rate, banks can focus on enhancing customer experience, promoting consulting and providing information about products, offering combined product packages, and promotional programs. attractive to increase the number of products used.

Active customer activity is an important factor that helps banks retain customers. Banks need to create interactive activities, provide attractive incentive information, and design flexible and convenient products and services. Regularly contact and interact with inactive customers to create close relationships, this can be via email, phone, text messages, etc. Creating a personal connection can be an important factor. so that customers feel excited, and come back to use your services.

Customers aged 46 to 75 years old have a high churn rate (45%), which may reflect the changing trend in customer needs and priorities with age. Banks can focus on developing products and services that suit the needs and priorities of each age group. Especially from 46 to 75 years old.

New customers are a group of customers with a higher churn rate when compared to other groups. Banks need to create a welcoming and supportive environment, and offer attractive product and service packages, especially in the early stages when customers are new to using the bank's services. This helps customers want to stick with the bank for a long time, and increases customer loyalty to the bank.

THANKS

Do you have any questions?

nhudaitran1510@gmail.com
+84-9 6862 93 64