PIG E. BANK – Customer Retention Analysis

Data Analyst: Nadia Ordonez Roman

Date: 08.01.2024

Context:

To increase customer retention, the sales team wants to identify the leading indicators that a customer will leave the bank. Use client attributes such as age, estimated salary, etc, to identify the top risk factors that contribute to client loss and model them in a decision tree.

Excel processes:

To address the above situation, I first started by creating a copy to address data exploration and cleaning steps (Step 1). Once the dataset was cleaned, I compared basic statistics between both groups: customers who "Stayed" and those who "Exited", selecting relevant variables to create our decision tree (Step 2). Next, I evaluate and select the top 3 variables that would be included in the decision tree (Step 3). Finally, I created a decision tree using the selected variables (Step 4).

Step 1.- Cleaning steps and data exploration.

1.1 Cleaning steps

Data set

- **Raw data size**
  Total observations (rows) = 991 & Total variables (columns) = 14.
- **Value detail**
  "1" values equal to Yes.

Data cleaning

- **Deleted variables**

The variable "Last_Name" was deleted due to PII laws under the GDPR.

- **Duplicates**

No duplicates were found.

- **Missing values**

Variables	*Missing values*	*Treatment*
Row_Number	N	-
Customer_ID	N	-
Credit Score	3	Deleted
Country	N	-
Gender	1	Deleted
Age	1	Deleted
Tenure	N	-
Balance	N	-
NumOfProducts	N	-
HasCrCard?	N	-
IsActiveMember?	N	-
Estimated Salary	2	Deleted
ExistedFromBank	N	-

After consultation with the stakeholders, the 6 rows containing missing values were deleted. These observations represent 0.6% of our entire dataset. Thus, deleting them will not affect our main project goals.

- Inconsistencies

Variables	*Inconsistencies*	*Treatment*
Row_Number	N	-
Customer_ID	N	-
Credit Score	N	-
Country	Y. Same values are described in multiple ways.	Values were relabeled.
Gender	Y. Same values are described in multiple ways.	Values were relabeled.
Age	Y. Age out of expected range.	Age was imputed.
Tenure	N	-
Balance	N	-
NumOfProducts	N	-
HasCrCard?	N	-
IsActiveMember?	N	-
Estimated Salary	N	-
ExistedFromBank	N	-

Inconsistency 1	
Variable	Country
Value	22 'DE' values were changed to 'Germany'
Value	117 'ES' values were changed to 'Spain'
Value	244 'FR' values were changed to 'France'

Inconsistency	*2*
Variable	Gender
Value	49 'M' values were changed to 'Male'
Value	19 'F' values were changed to 'Female'

Inconsistency	*3*
Variable	Age
Value	11 cells with a '2' value were imputed*.

*** Imputation age**

After further investigation, 11 accounts linked to females from Spain were detected as customers with 2 years old age. This is a mistake since our stakeholders confirmed that minors can't own bank accounts in this dataset. Age was estimated taking into account the Gender, Country, and Salary range of alike customers.

Eg. Estimating the age for customers assigned 2 years old as age.

1.- Here, I filtered for Female, Spain, and Estimated Salary between 15 000 and 20 000.

2.- Assigned an average age, in this case = 38 yo. I also changed the format of the imputed cell to track this change to a grey color.

Country	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard?	IsActiveMember	Estimated Salary	ExitedFromBank?
Spain	Female	45	2	$122,311.21	1	1	1	$19,482.50	0
Spain	Female	36	6	$132,311.71	1	0	0	$15,462.84	0
Spain	Female	40	1	$146,502.07	1	1	0	$19,162.89	0
Spain	Female	30	2	$34,013.63	1	1	0	$19,570.63	0
Spain	Female	32	9	$0.00	2	1	1	$18,924.92	0
Spain	Female	2	6	$118,879.35	2	1	1	$19,131.71	0
Spain	Female	2	3	$86,605.50	3	1	0	$16,649.31	1

1.2 Data Exploration

Following the completion of the data cleaning procedures, our dataset now comprises 985 rows and 13 variables. Notably, there is a slight gender imbalance, with a marginally higher percentage of males (53%) compared to females. Geographically, the majority of our customers are from France, constituting 48.6% of the dataset, followed by clients from Germany at 26.0% and Spain at 25.4%. It is also important to highlight that 20.51% of our customers, totaling 202 observations, have exited the bank. These observations bear significance as the presence of imbalances or lower representation within specific groups can impact the accuracy of our decision tree model. Therefore, acknowledging these nuances is crucial for ensuring the robustness and reliability of our analytical outcomes.

Step 2. “Stayed” vs “Exited” comparison

In our dataset, various variables can be assessed to discern the factors influencing whether a customer chooses to exit or stay with our bank. Notably, "Row_Number" and "Customer ID" bear no impact on the customer's decision and are thus excluded from our analyses.

The significance of variables was initially assessed through percentage and average comparisons between the two groups. Subsequently, variables were categorized into those deemed relevant and those considered irrelevant for further analysis.

Relevant	Irrelevant
Country	*Row_Number*
Gender	*Customer ID*
Age	*Credit score*
Tenure Balance	*HasCrCard?*
NumOfProducts	*Estimated Salary*
IsActiveMember?	

Step 3. Selecting the top 3 variables for a decision tree

For irrelevant variables Credit score, HasCrCard? and Estimated Salary, visuals, and statistical testing analyses were performed to further support our initial findings that they don’t play a significant role in predicting whether a customer would exit the bank. The performed t-tests with a 5% significance level showed that there is no significant difference between the credit score or estimated salary averages among those clients who stayed or exited the bank. Similarly, the percentages of clients with or without credit cards did not differ among the two groups.

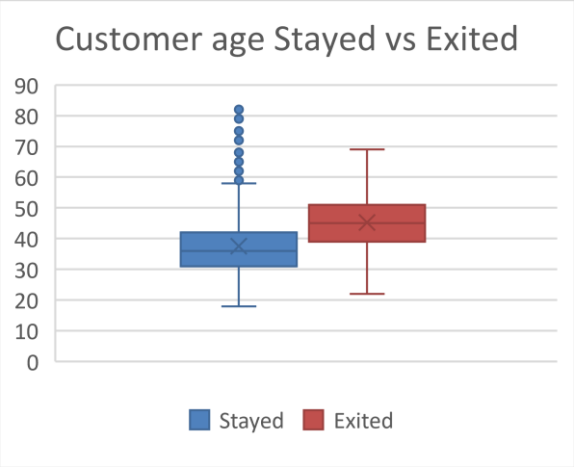
For selected relevant variables, I also created supporting visuals and ran statistical tests to highlight their relevance in predicting whether a customer would exit our bank. Based on their percentages among those who stayed or exited the bank, the age, country, and IsActiveMember variables were selected.

Top 3 variables driving the probability of a customer to exit the bank

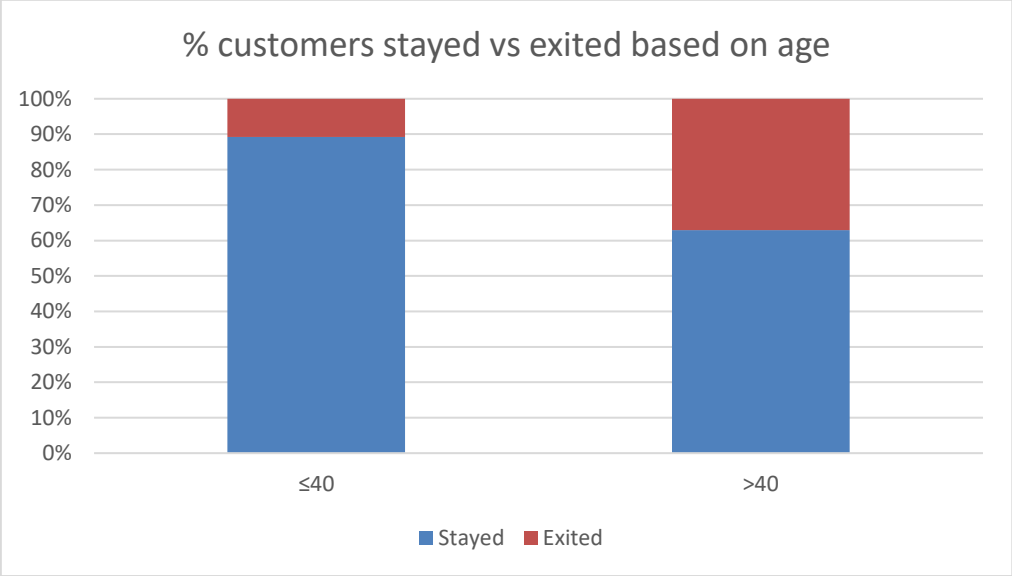
Variable	Condition	**Exited**	Stayed
Age	>40-year-olds	**37.09%**	62.91%
Country	Germany	**29.30%**	70.70%
IsActiveMember	0 = No	**29.25%**	70.75%

3.1 Age

The t-test with a 5% significance level showed that the average age between groups that stayed and exited was not similar. Older customers tend to exit our bank.

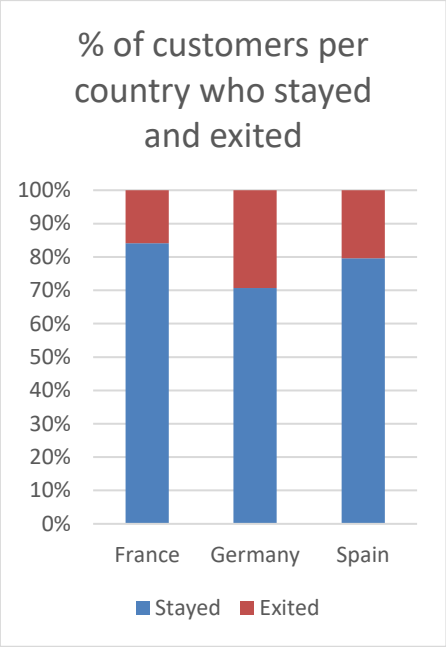


For a simple decision tree schema and based on the average age between groups, I grouped age into those who are 40 or younger and older than 40 years old. In this case, 37.09% of customers who were older than 40 years old exited the bank. While only, 10.79% of those who are 40 and younger exited the bank.



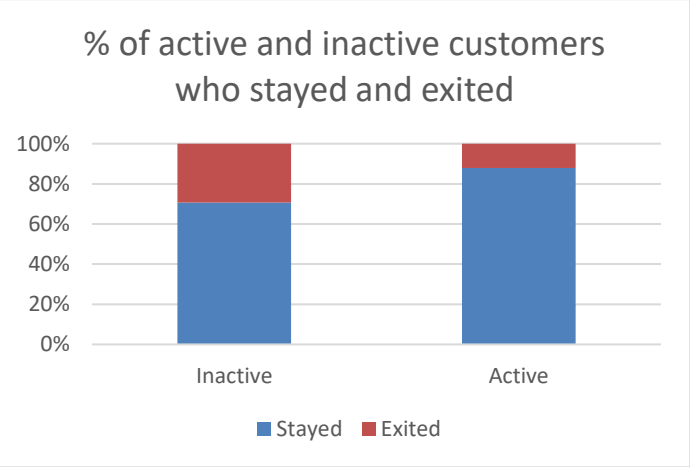
3.2 Country

German clients exited the bank at higher rates followed by customers from Spain. Of our 256 initial German customers, 29.3% exited the bank. In contrast, French customers show a tendency to remain with the bank. Only 15.87% of our French customers exited the bank.



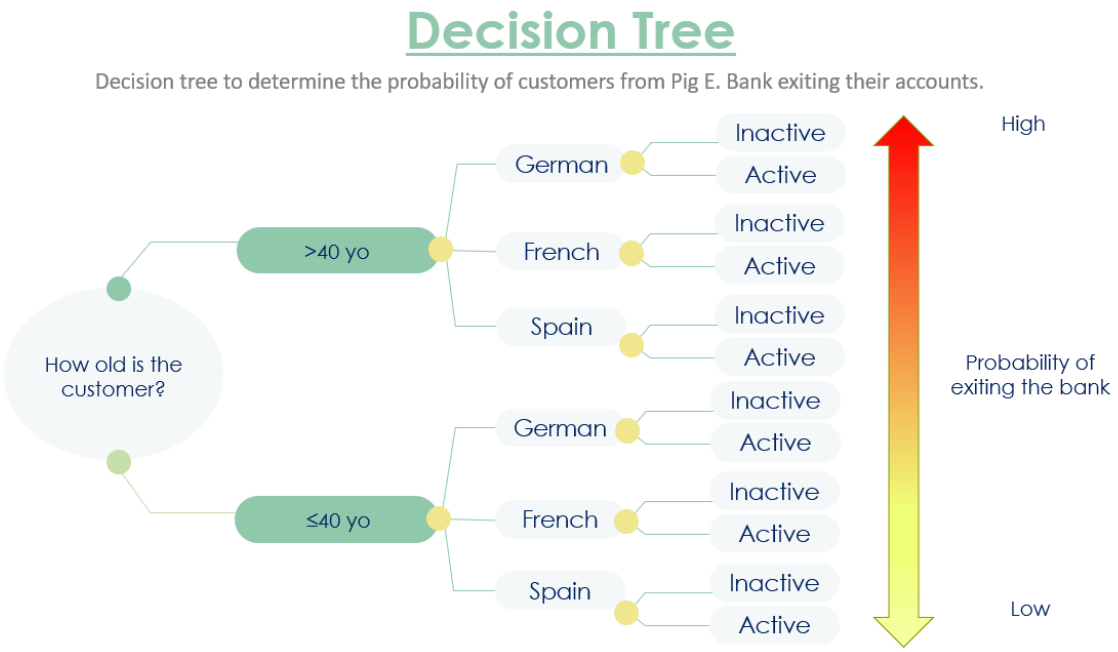
3.3 IsActiveMember?

Clients who exited the bank were usually inactive customers. In our dataset from 482 inactive customers, 29.25% of them exited the back. Active customers usually stay with the bank, with only 12.13% of them exiting the bank.



Step 4. Decision tree

Once the top 3 variables were chosen, I built a decision tree with age as the top variable driving the probability of a customer to exit the bank, followed by country of origin and their active status.



When applying our decision tree to our current dataset, 63.3% of the filtered customers exited the bank. To improve the accuracy of our model, more variables can be included.