Python Project: Customer Churn Analysis

Sofia Ruiz Romanos, Sarah Radoui

Master 2-Quantitative Economics and Analysis

November 8, 2024

How can banks identify and predict key factors driving customer churn, and what strategies can be implemented to reduce the churn based on customer profiles and behavior patterns?

Every result mentioned in this document is illustrated by either a chart or a table available in the attached python notebook. We choose to show only specific figures, to respect the number of pages restriction given for this project.

1. Context

A European bank wants to identify the main factors contributing to customer churn (i.e., customers leaving the bank and closing their accounts). By doing so, the bank can target such customers with incentives, or use this knowledge to propose new products that are better suited to their needs.

To try to identify the main factors contributing to customer churn we will first make a general review of the variables we are working with, their distributions and trends. This will give us more perspective over the dataset. In a second part of the paper we will try to identify and predict these key factors driving customers churn by interacting our variables of interest with the **Exited variable (Target).**

2. Correlation Analysis

We first construct a correlation matrix and look at what might affect churning.

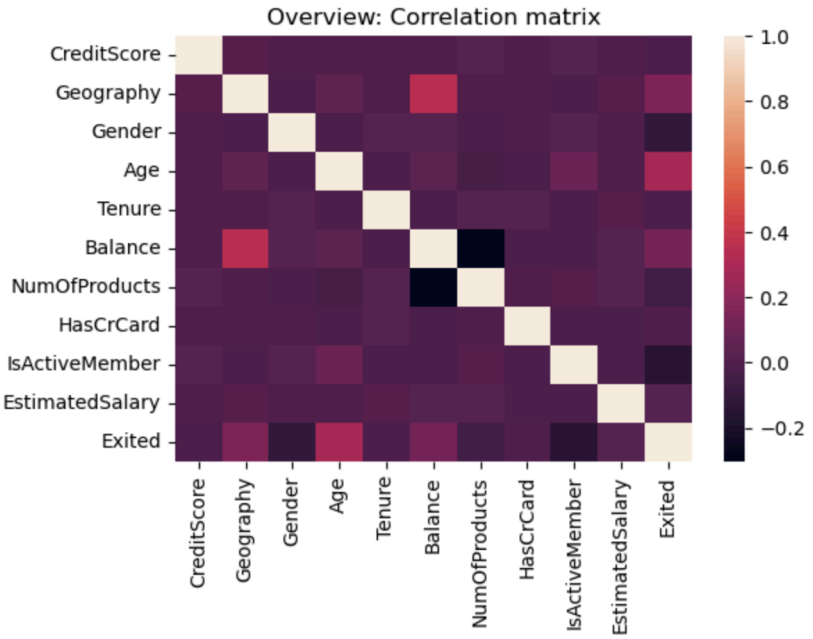


Fig 1. Correlation Matrix

Fig 1 shows that the strongest correlations to customer exit are with *Geography, Age and Balance.* While not directly correlated to churn, some other variables present interesting patterns to investigate further. *Balance*, for instance, appears to be interestingly correlated with *Geography*, potentially indicating that customers with a certain level of balance are concentrated in a specific country. Additionally, *Age* also holds a minimal correlation with customer's *Activity*.

3. General Analysis

A) INDIVIDUAL DISTRIBUTION OF VARIABLES

We have divided variables into three categories: those indicative of a customer's **identity**, those pertaining to each customers' **financial information**, and finally those relating to a customer's **"activity" or behavior** with respect to the bank and its products. The variables *CustomerId* and *Surname* act as indexes and references to every client but are not relevant to our analysis.

CUSTOMER PROFILE

The bank's customer base comprises 45.43% women (4,543 clients) and 54.57% men (5,457 clients) **(Fig 2)**, with ages ranging from 18 to 92 years (median age of 37, standard deviation of 10.49) **(Fig 3)**.

Most clients are based in France (50.14%), followed by Germany (25.09%) and Spain (24.77%). We speculate that the bank might be facing some competition in the German and Spanish markets **(Fig 4**), or that it might have originated in France before expanding to Germany and Spain, which would explain the smaller samples in these two countries with respect to France.

It would be interesting to see if the age profile of customers in each country provides insight into our theory. For example, a younger age distribution in Germany and Spain than in France could indicate a higher penetration of the bank's services in the two branches amongst young customers who do not have an established bank account yet. However, age distribution analysis reveals no significant differences across countries, suggesting a similar customer age profile in all three markets **(Fig 5)**.

Fig 2. Gender Distribution

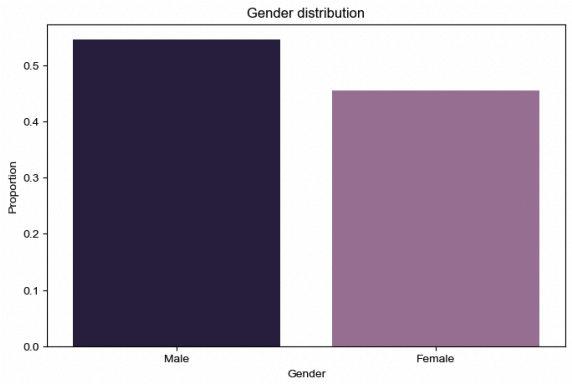
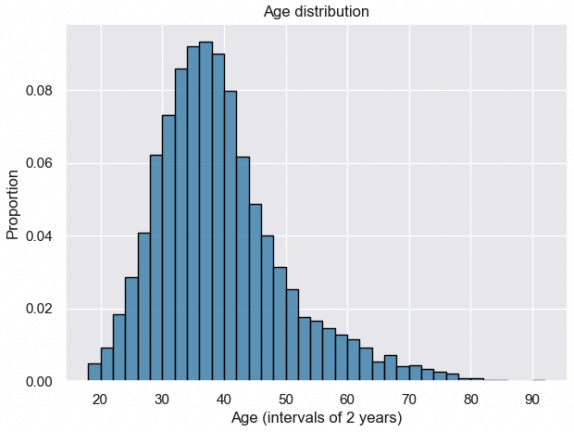


Fig 3. Age distribution



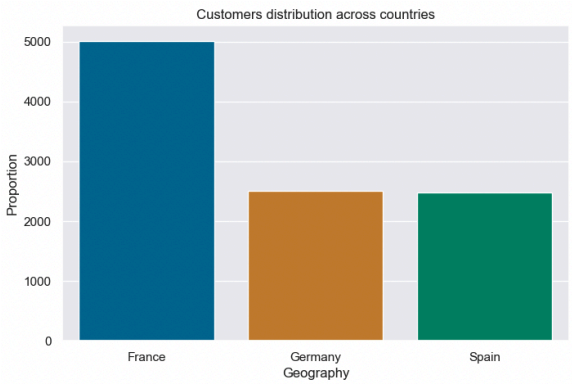


Fig 4. Customers distribution across countries

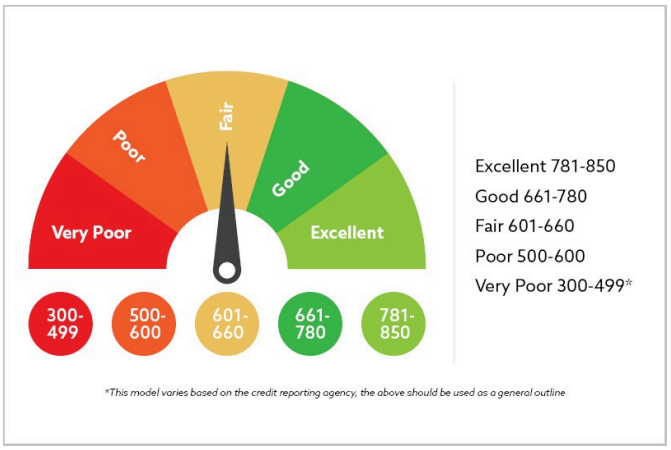
Fig 5. Age distribution by country

FINANCIAL ANALYSIS

Our financial variables are credit score, balance, and estimated salary. The financial profile of the bank's customers reveals three main insights:

1. Credit scores are generally centered around a "good" level

Credit score is a number from 300 to 850 that rates a consumer's creditworthiness. It represents a customer's credit history and your ability to pay your loans and financial obligations. In particular, the FICO score divides credit score into different intervals to simplify the interpretation of the score according to each customer.



The average credit score is around 650.5288, with a standard deviation of 96.65 points. We observe an interesting peak in the proportion of people who have a credit score around 850. Most clients' credit scores in this bank are between 'poor' and 'fair', with 'good' being the most common and very few scoring 'excellent' or 'very poor' credit **(Fig 6)**.

2. Balance distribution is highly skewed

The average balance is of 76485.89€ and the median is of 97198.54€. This is because the proportion of people whose balance is at 0 is considerable: 36.17 % **(Fig 7)**.

Fig 6. Customers distribution by Credit Score

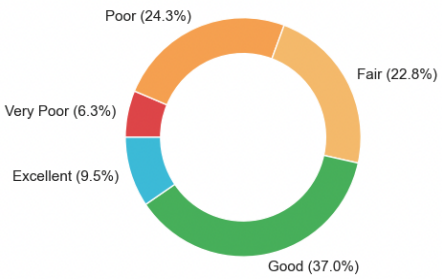
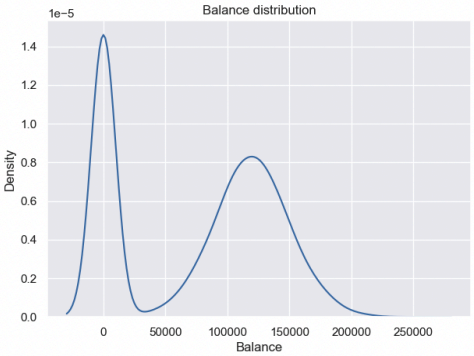


Fig 7.Balance distribution



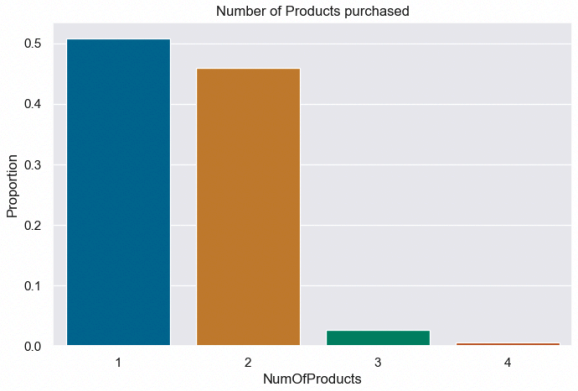
3. Estimated salary is evenly distributed across the customer base.

The average estimated salary is of 100,090.24€ and the median is of 100,193.92€ **(see. Python notebook)**. It is interesting to note that we observe a very uniform distribution across all values, with a slight peak at the end of the distribution. People all over the distribution of salary seem to be clients of this bank.

However, it is very interesting to note the disparity between this result and the very skewed distribution in balance. If we group people having a very low balance, around zero, and those having a higher balance, we do not actually see any difference in the estimated salary distributions of the two groups. This implies that there seems to be an important element beyond estimated salary that will likely affect balance and which we fail to observe.

ACTIVITY DESCRIPTION

Fig 8. Number of products purchased by the customers



The analysis of the number of products bought by each customer shows that all clients have at least one product, likely a basic bank account, and nearly half hold two products, while only a small fraction (under 5%) have three or four products **(Fig 8)**.

Credit card ownership is common (70% of clients), though not all customers with two products have a credit card, indicating it's not always the second product purchased. Notably, active and inactive clients are almost evenly split, and customer tenure is well distributed, though clients with 10-year tenure form a smaller group. The lower customer count in Germany and Spain may indicate higher competition in these markets as seen previously**. (see. Python notebook)**

B) HOW DO THE DIFFERENT VARIABLES INTERACT WITH EACH OTHER?

GENDER PATTERNS

Gender does not appear to be a predictor of activity, tenure, credit card ownership, or recent activity levels. While women are slightly more likely than men to hold more than two products, the low

numbers of clients with 3 or 4 products make meaningful comparison challenging. However, the average estimated salary is lower for women, with fewer women at the high end of the salary distribution (175k+), whereas men are more represented in that range **(Fig 9)**. Finally, we find no gender differences in balance or credit score, suggesting no link between salary differences and these financial variables**.**

Fig 9. Distribution of Salaries by gender

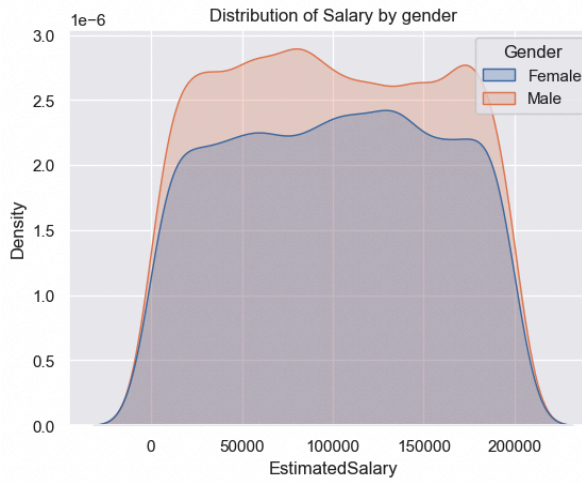
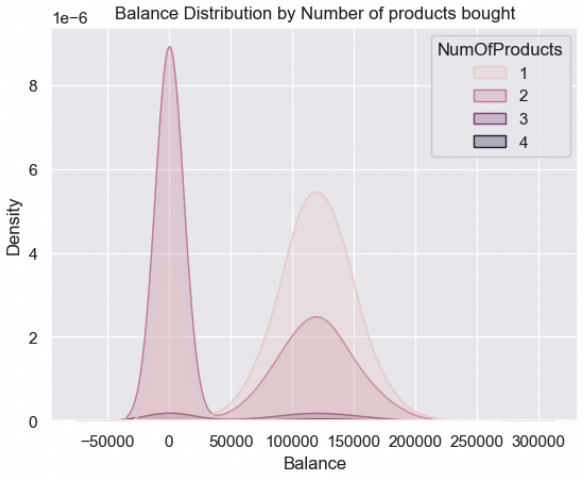


Fig 10. Balance Distribution by Number of products purchased



BALANCE PATTERNS

Balance does not show any notable correlation with activity variables either, contrary to what we might expect. Typically, we could anticipate that customers with higher balances would display higher engagement or retention rates, perhaps due to greater financial involvement with the bank. Additionally, higher balance holders might be more likely to own multiple products as part of their wealth management strategy. However, in this dataset, we observe no significant link between balance and the number of products purchased **(Fig 10)**, nor any clear trends indicating that balance affects customer activity overall.

SALARY PATTERNS

Estimated salary does not appear to predict activity patterns, as no clear trend emerges between salary levels and customer engagement. Additionally, given the small sample sizes for clients with 3 or 4 products, examining activity by product count is as mentioned before, not particularly insightful. Overall, estimated salary shows no correlation with individual activity patterns.

CREDIT SCORE PATTERNS

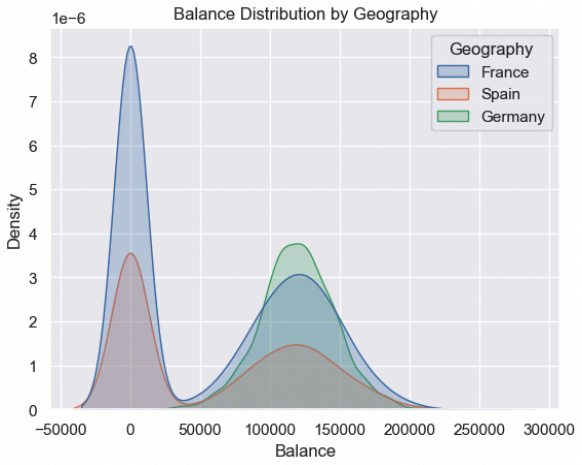
Regarding the credit score, we might assume that customers with higher credit scores would be more engaged or likely to hold multiple products, as their financial stability could encourage deeper involvement with the bank's offerings. Additionally, a higher credit score might be associated with lower churn rates, reflecting a more stable financial profile. However, in this dataset, we find no significant relationship between credit score and any activity variable, suggesting that

creditworthiness alone is not a key driver (or not enough) of customer engagement or retention patterns here.

HOW IS GERMANY AN EXCEPTION?

While we see no geographic trends for credit card ownership, product count, or credit scores, German customers display a different balance distribution compared to other countries **(Fig 11)**. In Germany, balances are tightly centered around a peak of 125,000€, with no values near zero, unlike in France and Spain. This prompts us to further investigate the relationship between German customers and churn rates. One possible explanation can be a different financial engagement level among German clients.

Fig 11. Balance Distribution by Number of products purchased



Having explored the general customer profile, we now turn to analyzing the key factors driving customer churn. Starting with a review of summary statistics, we aim to understand how churning impacts other variables, providing insight into potential predictors of customers leaving the bank.

4. Key Factors driving customer churn

Table 1. Differences in Summary Statistics

variable	count	mean	std	min	25%	50%	75%	max
CreditScore	5926	6.5017	−4.66767	55	7	7	2	0
Age	5926	−7.42961	0.363801	0	−7	−9	−10	8
Tenure	5926	0.100535	−0.0554485	0	1	0	−1	0
Balance	5926	−18363.2	4487.25	0	−38340	−17276.6	−5023.05	−29365.3
EstimatedSalary	5926	−1727.29	−506.831	78.49	−1124.23	−2815.8	−3812.95	184.38

Table 1. provides an overview of key statistical metrics, including mean, maximum, minimum, and standard deviation values, offering a foundational analysis of the dataset's distribution and variability with respect to the **Exited target variable**. We can for example see that those who exit are on average 7 years younger and seem to have a smaller average balance (of almost 1900) compared to those who stay.

Around 20% of the overall customers have exited the bank.

A) CUSTOMER PROFILE

We see that non-exited customers are generally younger, with a more concentrated age range around 30-40 years old, which could come from the fact that, as we showed previously, the dataset is composed mainly of younger customers **(Fig 12)**. However, amongst exited customers there seems to be a pattern of older customers that appear more likely to exit the bank. We indeed see that the churn rate is the highest for the age group between 50 and 60 years old. This suggests that the bank may be more successful in retaining/attracting younger clients.

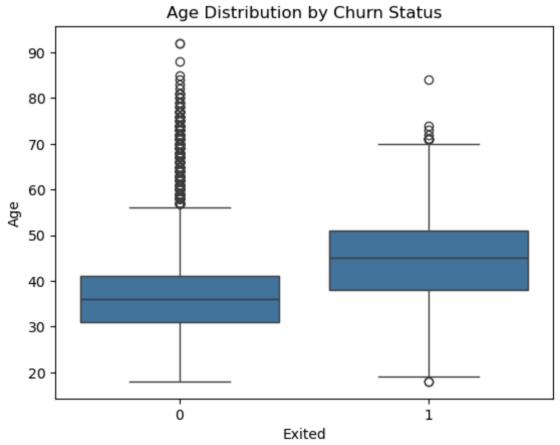
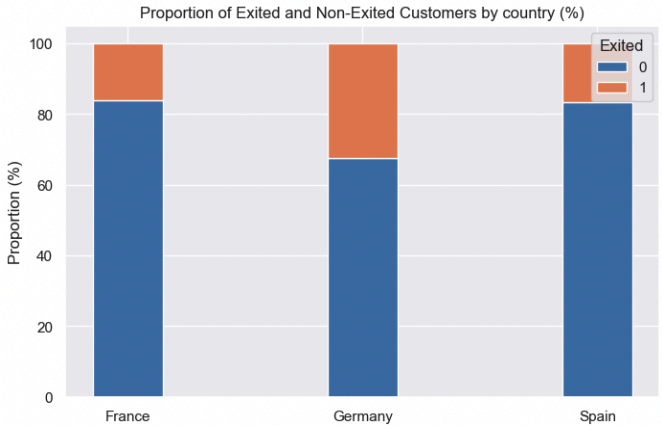


Fig 12. Age Distribution by Churn Status

Regarding the effect of a customer's location on exit probability we look at the interaction with the Geography variable (**Fig 13).** Within Germany the proportion of customers exiting is higher relative to the total population.

Fig 13. Churn proportion by country



It is interesting to keep in mind that in absolute terms we see the same number of customers exiting the bank in France and in Germany, due to the different sample sizes between the two countries shown in **Fig 4**. The relatively strong correlation we found in the correlation matrix between *Geography* and *Exited* is explained by the significant difference in percentage of exit in Germany.

Furthermore, The percentage of women leaving the bank seems higher overall (around 25%) than for men (around 75%). If we split by country we see the same results: in all the three countries the

percentage of female churning over the full sample of churned customers per country, is always higher **(see. table below)**

Table 2. Churning percentage by gender and country

Geography	Nbr Female Churning	Nbr Male Churning	Female Churning (%)	Male Churning (%)
France	460	350	56.790123	43.209877
Germany	448	366	55.036855	44.963145
Spain	231	182	55.932203	44.067797

B) FINANCIAL ANALYSIS

We then focus our analysis on the interaction of the *Exited* variable with the financial variables *Balance*, *Estimated Salary* and *Credit Score*.

We first see that what we see at country level is similar to what we concluded at aggregate level: having a small or big balance doesn't seem to predict possible exits amongst clients **(Fig 14)**, nor does the level of salary a customer holds, as we observe a quite uniform distribution of estimated salary by country **(Fig 15).**

Fig 14. Balance distribution by Exited status and country

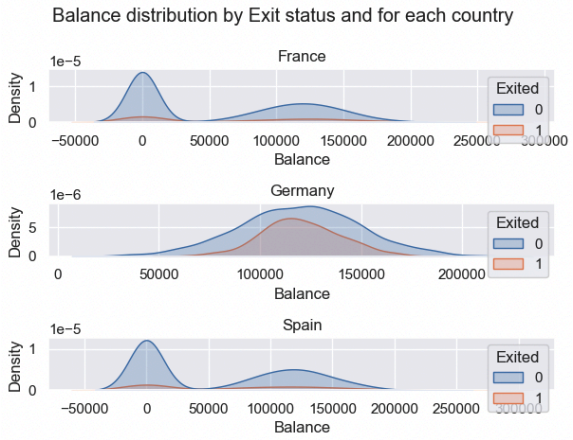


Fig 15. Salary distribution by Exited Status and country

Credit score could be an interesting parameter to use to understand a customer's financial behavior and reliability, because of factors such as payment history, outstanding debts, and credit history length. By examining these scores, we aim to identify potential trends or correlations between credit reliability and customers leaving the bank. These may serve as valuable indicators for assessing whether specific financial profiles are more likely to leave the bank.

However, in our dataset we can only see that a slightly bigger percentage of poor and very poor credit score customers leave the bank **(Fig 16),** but it's not significant enough to confirm our previous assumptions.

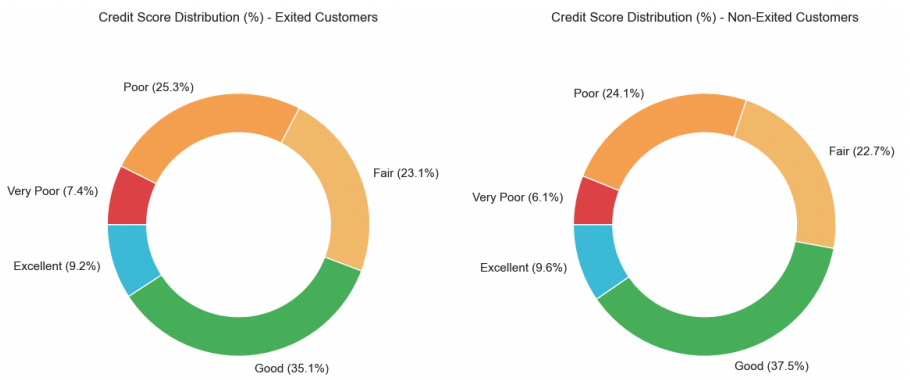


Fig 16. Credit Score Distribution (%) among Exited customers [left chart]and Non-Exited customers [right chart]

C) ACTIVITY ANALYSIS

Given our previous results, we now look at the effects of spending more or less years in the bank, purchasing more or less products, owning a credit card and the overall activity of the customers on churning.

The number of years people have been clients of the bank for does not appear to be a meaningful predictor of churn, as its distribution remains consistent across those who have exited and those who remain. Similarly, the number of products purchased and credit card ownership do not show significant differences in relation to exited clients.

However, a customer's *Activity Status* presents a more compelling trend: across all countries, a higher proportion of exited clients are classified as non-active compared to retained clients **(Table 3).** For example, in France, 63.21% of exited members were not recently active, and similar patterns can be observed in Germany and Spain. This suggests that inactivity could be a key factor associated with churn. However, as our data lacks a time-series element, it is difficult to determine whether inactivity directly contributed to churn or was simply a result of clients already preparing to leave the bank.

Table 3. Churning percentage by gender and country

		Active Members (%)	Non-Active Members (%)
Geography	**Exited**		
France	0	54.543292	45.456708
	1	36.790123	63.209877
Germany	0	56.165192	43.834808
	1	36.363636	63.636364
Spain	0	56.734496	43.265504
	1	34.140436	65.859564

Overall, we see some striking trends such as, Germany being the country with the highest rate of churning, and with no customers with a balance around 0. Furthermore, the lower the balance the

higher the probability to have purchased two products. We also see slight trends for Age in our activity variables, as well as a small difference in salary distribution for each gender. Finally, we wonder whether not being recently active could be a predictor of customer churn.

However, we fail to see any clear indicative variable that might be able to predict customer churn.

5. A machine learning application: Can we predict the exit of customers ?

We build a decision tree to try to predict the customers who will exit the bank. The most important problem we have run into while building this tree is the small sample of the dataset, which reduced the training sample for the tree. We stratify the tree to obtain representative samples in train and test. We then use accuracy score as the maximizing metric to build a *CVGridSearch* to allow us to pick the best parameters for this model, and we analyze its performance.

The first column of table **Table 4** presents our results. Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. Our accuracy score means that the model correctly predicted the churn status of the customer 85.23% of the instances in our test set. Precision (also called positive predictive value) is the fraction of customers that correctly exited amongst the predicted exited customers (*True Exits / Predicted Exits*). Our precision implies that 75% of the people who were predicted by the algorithm to exit were correctly predicted. Recall (also known as sensitivity) is the fraction of the total number of relevant exit instances that were actually retrieved (*True Exits / Total Exits*). Only around 41% of the people who exited in our database were correctly predicted by the model.

Table 4. Tree Performance

Maximising metric:	Accuracy	Recall
Accuracy	0.8523	0.8496
Precision	0.7530	0.7030
Recall	0.4091	0.4533

The general accuracy and precision of our model are not bad, especially given the sample size. However, our model fails particularly in its recall. It is not good at identifying all the people who exited, which poses a problem because this is what we want this algorithm's main task to be for future customer datasets.

To try and improve the recall of our model, we run another grid search for our best parameters using recall as the maximizing metric (see column 2 of table **Table 4**). The resulting tree improves recall score only by 5% (going from 40% to 45%) while reducing accuracy and precision slightly (around 1% and 5% approximately), which is not a significant difference.

The confusion matrix **(Fig 17)** clearly shows us the problem with the model: **amongst our true exited, more than half are incorrectly predicted as stayed.** We are able to think of 2 reasons to explain the poor recall. First, that there are no clear variables or combinations of variables in our database that are good predictors of churn. This would explain the absence of clear trends during our previous analysis, and we can also see this in how complicated the tree nodes' decisions are. Secondly, that a bigger data sample could maybe have brought more clarity and trends into our variables to explain customer churn.

Figure 17. Confusion Matrix