Python Project: Customer Churn Analysis

Hugo Collot, Juliette De Ketelaere
M2 Quantitative Economic Analysis

December 2022

Introduction

In this paper, we analyze the statistical relationship between banking, gender, geographic, and economic variables altogether and with a target variable, which is the exit churn rate of customers from a bank. Hence, processing a probit model analysis, we find little evidence that *Age*, *Balance*, $CreditScore$, $IsActiveMember$, and *Age* have a significant impact on the *Exited* target variable. Then, in the second part, we run a machine learning analysis implementing a classification KNN model. We find the optimal model regarding the number of neighbors and assess its prediction abilities.

1 Data

1.1 Description of the dataset

In this part, we analyze the summary statistics and correlation matrix of the customer churn rate dataset. First, let's observe that the dataset contains 10,000 observations (individuals), and it has 10 explanatory variables, and one target variable, which is *Exited.* We now focus on the main features of the descriptive statistics table. Customers from France are the most represented among customers from only 3 countries in Europe, with more than 5000 individuals. Males are over-represented with more than 54% of the observations. Furthermore, on average the individuals are 38 years old, and they have a €76,000 balance. They retain at least 1.5 bank products and have 0.71 credit cards. Then, we can deduce that they have mainly saving and insurance products than a current account with a credit card. Moreover, on average, bank customers earn more than €100,000 per year, and exit the bank in 20% of the cases.

Statistics	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
Count	10000.00	10000	10000	10000.00	10000.00	10000.00	10000.00	10000.00	10000.00	10000.00	10000.0
Unique	NaN	3	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Top	NaN	France	Male	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Frequency	NaN	5014	5457	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Mean	650.53	NaN	NaN	38.92	5.01	76485.89	1.53	0.71	0.52	100090.24	0.2
Std	96.65	NaN	NaN	10.49	2.89	62397.41	0.58	0.46	0.50	57510.49	0.4
Min	350.00	NaN	NaN	18.00	0.00	0.00	1.00	0.00	0.00	11.58	0.0
25%	584.00	NaN	NaN	32.00	3.00	0.00	1.00	0.00	0.00	51002.11	0.0
50%	652.00	NaN	NaN	37.00	5.00	97198.54	1.00	1.00	1.00	100193.92	0.0
75%	718.00	NaN	NaN	44.00	7.00	127644.24	2.00	1.00	1.00	149388.25	0.0
Max	850.00	NaN	NaN	92.00	10.00	250898.09	4.00	1.00	1.00	199992.48	1.0

We then analyze the correlation matrix between all the explanatory and target variables taken into account. Despite almost no correlation between most of the variables, we can highlight two main feature correlations between the $NumOfProducts$ and $Balance$, which are negatively correlated (-0.30). Furthermore, the *Exited* variable is positively correlated to *Age* variable (0.29). Hence, the higher the age, the higher the probability of a customer exit.

CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1.00	-0.00	0.00	0.01	0.01	-0.01	0.03	-0.00	-0.03
-0.00	1.00	-0.01	0.03	-0.03	-0.01	0.09	-0.01	0.29
0.00	-0.01	1.00	-0.01	0.01	0.02	-0.03	0.01	-0.01
0.01	0.03	-0.01	1.00	-0.30	-0.01	-0.01	0.01	0.12
0.01	-0.03	0.01	-0.30	1.00	0.00	0.01	0.01	-0.05
-0.01	-0.01	0.02	-0.01	0.00	1.00	-0.01	-0.01	-0.01
0.03	0.09	-0.03	-0.01	0.01	-0.01	1.00	-0.01	-0.16
-0.00	-0.01	0.01	0.01	0.01	-0.01	-0.01	1.00	0.01
-0.03	0.29	-0.01	0.12	-0.05	-0.01	-0.16	0.01	1.00

1.2 Variable Interactions

1.2.1 Gender

There is a majority of men in this dataset, for 45.43% of women. The financial characteristics between the two groups seem at first sight similar. The age distribution in the dataset is similar for men and women, with an average age of around 38 years, as we can see in the plots in appendix. The distribution between countries is also similar for the two genders.

Concerning the financial variables, the salaries between men and women look also similarly distributed as we can see in the appendix on the boxplot, the means, quartiles, and extreme values are roughly equal. Both genders present the same share of the products offered by bank, they use 1 or 2 products in majority and both have a majority of cards, around 70%. In terms of time as a bank customer, as we can see from the boxplot, women seem to have less time spend as client in majority (in terms of quartiles) which are rather concentrated around 2 and 7 years, compared to between 3 and 8 years for men. Men seem to be slightly in terms of recent transactions, 52% against 50% for women.

1.2.2 Geography

The database gathers banking information of customers from three countries: France, Spain, and Germany. In terms of social data, the share between men and women remains roughly the same, as well as age distribution, as we can see in the density curves which follow the same pattern among countries.
As for the financial data, we observe different distributions of the balance of payments between countries. While France and Spain have a high concentration of their balance around zero, most of Germany's density is around 11200 (although there is also a high concentration of this balance for the other two countries). Wages are similarly distributed across countries (see boxplots), as are the products used by costumers. As for the number of years as a client of the bank, it is more spread out for Germany, and France has a lower spread than the other countries. Member activity varies very slightly between countries, ranging from 49.7% for Germany to 52.3% for Spain.

2 How do the different variables affect churn?

2.1 Financial variables

There is no substantial difference between the two groups (exited and non-exited) in the following variables: Estimated Salary, Balance, and Credit Score. We can see this with the following boxplot: for each of the variables listed above, the exited and non-exited groups have the same median and quartiles, as well as extreme values. As for the distributions around the groups, the exited and non-exited have similar densities between the variables. However, the balances of the clients' bank accounts stand out slightly and show slightly different characteristics depending on whether they belong to exited or non-exited clients: exited clients show a higher proportion of balances around 0 (see density graph), and which pulls the mean and quartiles down (see boxplots). However, the number of product used seem correlated with the churn rate seem the proportion of using 1 product is slightly higher for exited, as well as using 3.

These results suggest that exited customers are part of the same population in terms of income, as financial factors do not seem to vary significantly between groups, except for the balance and number of products used. We need to look at other variables to find out which are the outflows of the bank's clients.

2.2 Social and geographic variables

The information on social and geographical factors concerns the gender of the clients, their country of origin, and their age. We can therefore highlight differences in these variables for exited and non-exited clients. These factors seem to be much more significantly correlated with customer exit. Indeed, we observe that among women, the proportion of exited clients is higher than among men. Similarly, the proportion of exited clients of French origin is much higher than those of Spanish origin. However, it is the age population that seems to present the most differences: the population of exited clients is on average older and is concentrated around 50 years old, compared to 30-40 years old for the clients who stay. These characteristics allow us to draw a picture of the exiters, who seem to be more concentrated among women, older, and more in Germany and France.

2.3 Probit model

Finally, we seek to establish to what extent and magnitude the different variables affect the output. The variable of interest (exited) is here a dummy variable (0 for non-exited, 1 for exited) (binary dependent variable). To address this problem, we build a probit model to establish how the variables affect the churn rate.

The results are presented in the table below, where $CustomChurn_i$ is a dummy variable taking the value 1 if the customer i left the bank and 0 otherwise. $CreditScore_i$ stands for the customer's credit score, Age_i his age, $Tenure_i$ the number of years the customer has been a client of the bank, $Balance_i$ his bank account's balance, $NumOfProducts_i$ the number of products contracted with the bank, $HasCrCard_i$, a dummy variable taking 1 if he has a credit card and zero otherwise, $IsActiveMember_i$a dummy variable taking 1 if he has been recently active, $Estimated_Salary_i$ his annual estimated salary and $Geography_i$ his geographic location.

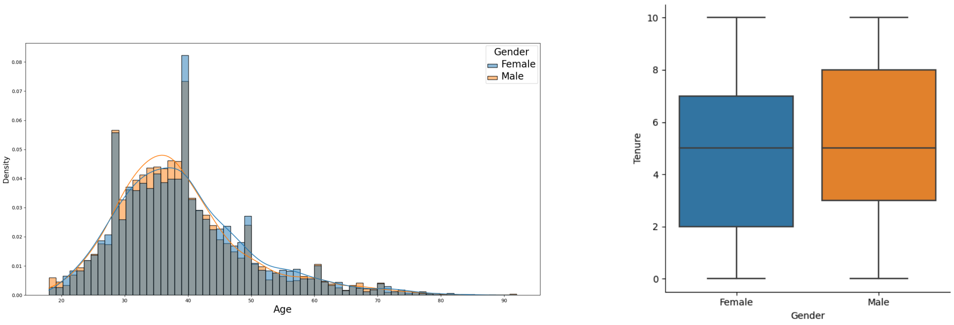


Figure 1: Age density (left) and Tenure for each gender group (right)

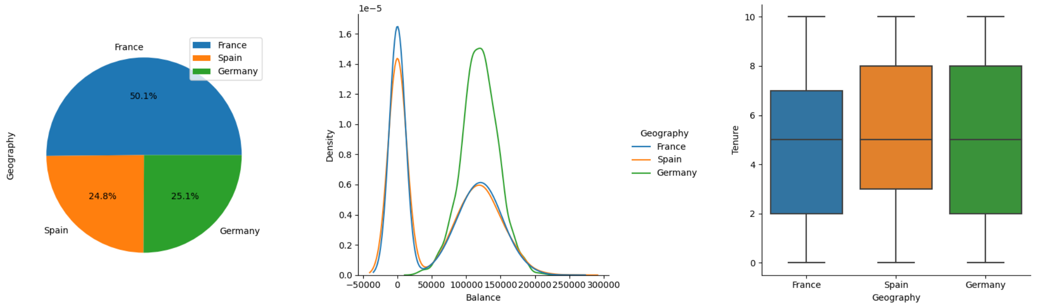


Figure 2: Country repartition (left), Balance density among countries (middle), Tenure among countries (right)

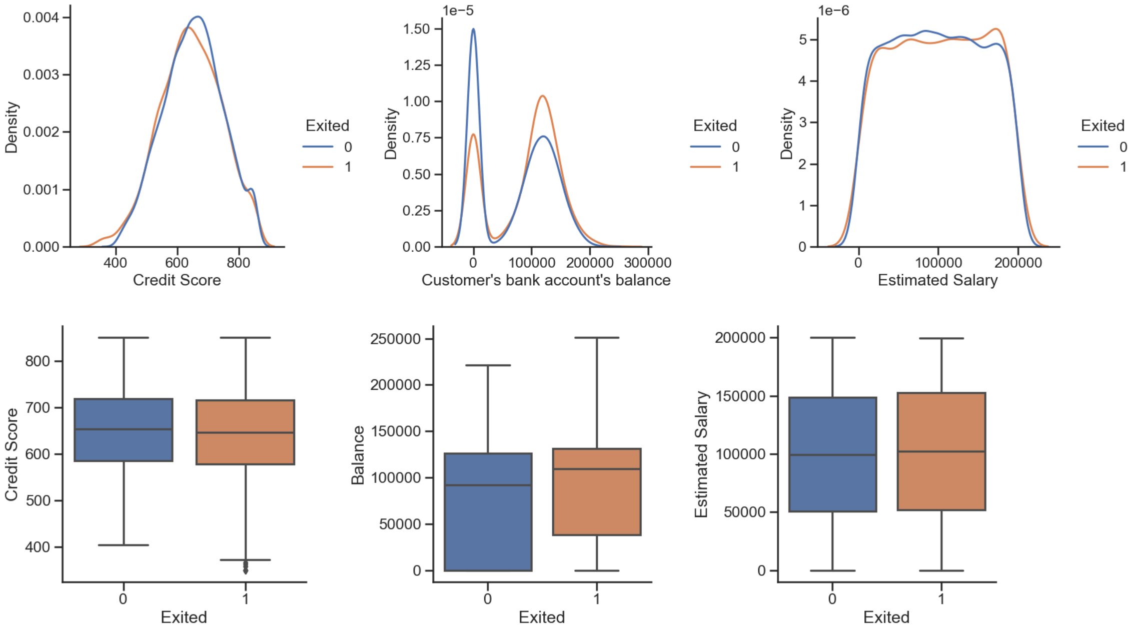


Figure 3: Financial variables and Exited customers

Dep. Variable:	Exited	**No. Observations:**	10000
Model:	Probit	**Df Residuals:**	9988
Method:	MLE	**Df Model:**	11
Pseudo R-squ.:	0.1527	**Log-Likelihood:**	-4282.9
converged:	True	**LL-Null:**	-5054.9
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
CreditScore	-0.0004	0.000	-2.524	0.012	-0.001	-8.93e-05
Age	0.0417	0.001	28.940	0.000	0.039	0.045
Tenure	-0.0086	0.005	-1.616	0.106	-0.019	0.002
Balance	1.521e-06	2.84e-07	5.358	0.000	9.64e-07	2.08e-06
NumOfProducts	-0.0532	0.025	-2.115	0.034	-0.103	-0.004
HasCrCard	-0.0277	0.034	-0.823	0.411	-0.094	0.038
IsActiveMember	-0.5795	0.032	-18.279	0.000	-0.642	-0.517
EstimatedSalary	2.785e-07	2.68e-07	1.039	0.299	-2.47e-07	8.04e-07
Geography_France	-0.9470	nan	nan	nan	nan	nan
Geography_Germany	-0.5031	nan	nan	nan	nan	nan
Geography_Spain	-0.9239	nan	nan	nan	nan	nan
Gender_Female	-1.0350	nan	nan	nan	nan	nan
Gender_Male	-1.3390	nan	nan	nan	nan	nan

Figure 4: Probit Regression on the variable Exited

Among the variables tested, 5 significantly influence the churn rate: $CreditScore$, Age, $Balance$, and $NumberofProducts$ and $IsActiveMember$. The credit score seems to negatively affect the probability of the churn rate, which indicates that people with a higher score are less likely to stay in the bank. This result is predictable since it is predictable that a customer who has an easier time borrowing (higher score) will be more likely to stay and take advantage of the bank's services.
The balance positively affects the churn rate, suggesting that clients with a higher bank account will have a higher probability to leave. Finally, the age of the customers is significant, which seems to confirm that older customers will tend to leave the bank more easily. Also, in terms of financial variables, the coefficient of wether the client was active recently, $IsActiveMember$, have a significant negative coefficient, suggesting that clients who where active recently have a lower probability to leave the bank.
On the other hand, the geography and gender variables do not seem to affect the churn rate: the p-value is too high to reject the null hypothesis.

Hence, the exit rate seem de be led with higher probability by financial data : Balance, Credit Score, and Number of Products, but also by the age factor, since older clients seem to have a higher probability to leave the bank.

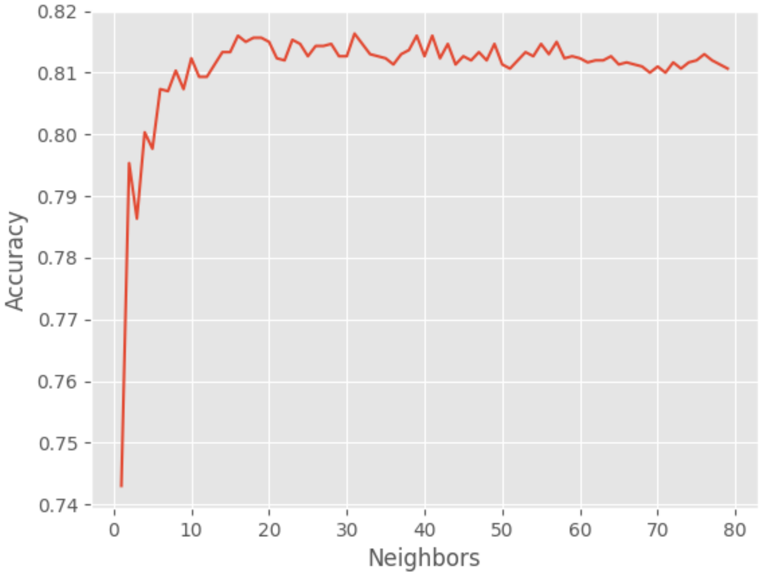
3 Machine Learning Model

In this part, we build and implement to the data a simple machine-learning model. Hence, using a supervised-learning KNN algorithm, we search for the model that provides the highest accuracy rate by tuning the KNeighborsClassifier parameter. Furthermore, we use a KNN algorithm to classify each individual whether she has exited (*Exited* variable is equal to 1) or not (conversely, *Exited* variable is equal to 0), considering the 'Exited' variable as the target variable, and the other variables as the explanatory variables. This classification is made based on the properties identified of the k closest neighbors of each given individual.

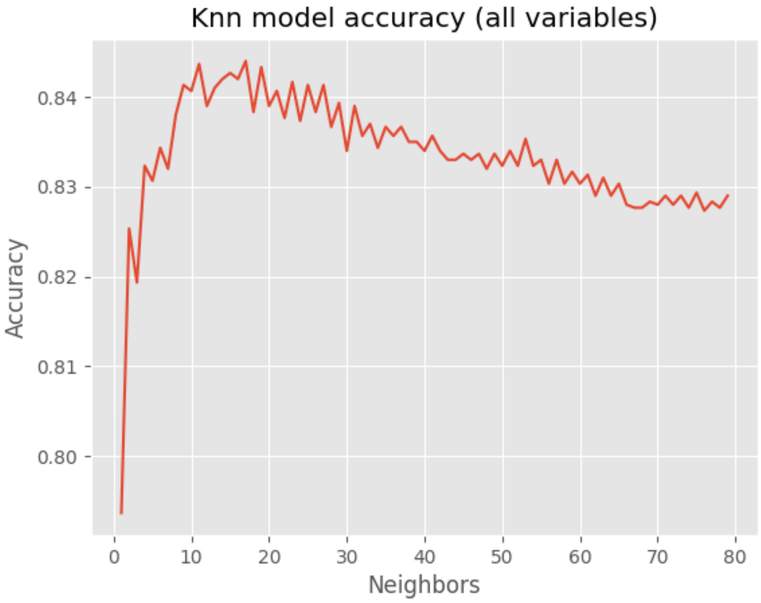
We divide our model implementation part as follows: we first preprocess the data, to drop non-relevant explanatory variables such as *CustomerId* and *Surname*, which could lead to overfitting issues. Second, using the *categorical_encoder* module, we transform categorical variables *Gender* and *Geography* into order numerical ones, affecting a given number to a gender/country-specific individual. Indeed, the KNN model used in this analysis does not treat categorical variables. Finally, we drop the target-variable column, and assign it to the output column of the model (Y), while the rest of the data frame is treated as the input part (X).

Moreover, to implement the KNN model, we split the customer's dataset to avoid any overfitting bias. We divide the data frame into a 70% training set, and a 30% test set. We first run and train the model on the training set, and then we predict the results using the trained model on the test set. We finally get the accuracy of the model on the test set, which is the proportion of true positive answers among all the test sample size.

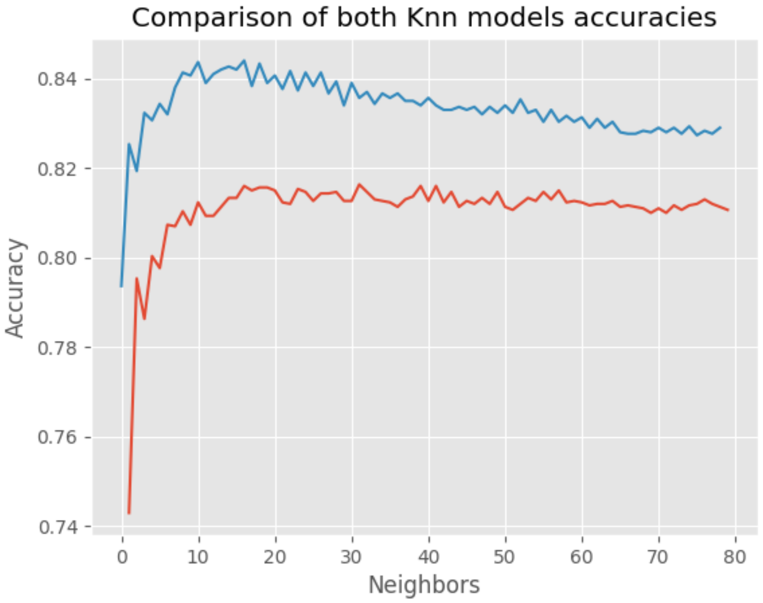
We iterate the algorithm by varying the number of neighbors taken into account for each individual and normalizing the distances of all the coordinates using *pipe*. The objective is to find the number of neighbors maximizing the accuracy of the model. First of all, we initiate the model parameter-varying loop with a limited scale of variables taken into account in the input (X). We keep only a few variables such as *CreditScore*, *Geography*, *Gender*, *Age*, *EstimatedSalary*. We draw out a graph featuring the accuracy of each model iteration varying with respect to the value of the neighbor parameter $KNeighborsClassifier$. In the figure below, we can observe that the accuracy is steeply increasing with the number of neighbors for small values i.e. between 1 and 15, and then reaching an up-and-down plateau. Eventually, we find that the model yielding the highest accuracy is reached for 30 neighbors. The maximum value of the accuracy is then 81.63%.



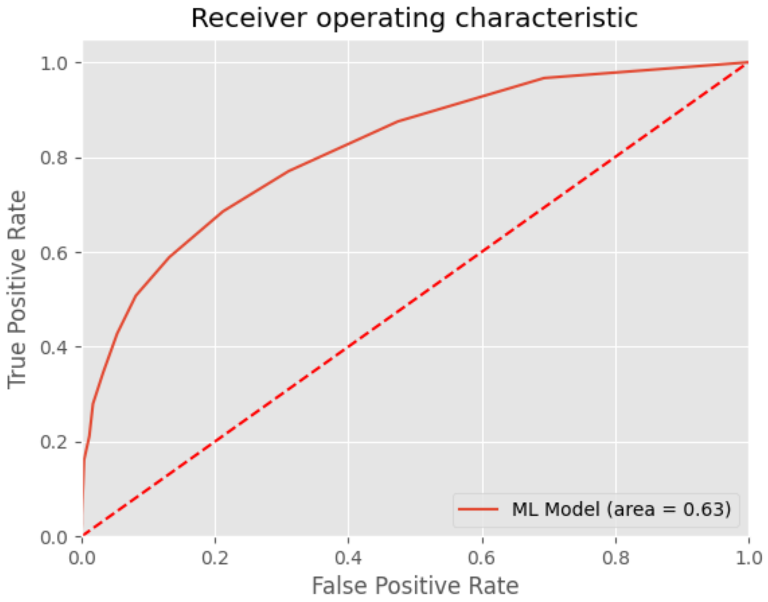
Furthermore, we iterate the model with all the variables contained in the input data frame and plot the corresponding graph of accuracy below. We find a high at 84.40% of accuracy for a neighbor parameter value of 16. Hence, the larger the set of input variables taken into account by the model, the lower the number of required neighbors by the optimal model (the most accurate one). However, compared to the previous case in which we only consider a limited set of input variables, after the highest point is reached and passed, the accuracy of the model is slightly decreasing with the number of neighbors.



In the next graph, we compare the accuracy for each category of the model, and we show that even slightly decreasing, the accuracy of the complete input variable set model is always higher than the one with a limited set of variables taken into account.



As a concluding remark for this section, we have plotted the graph of the receiver operating characteristics (below), which indicates the trade-off between sensitivity (through the TPR) and specificity of the model (through the 1 - FPR) of the model. Hence, a too-specific model would systematically overfit to data but would feature a very low sensitivity to data composition, whereas a very general model would feature high sensitivity. In this case, we study the best model found with the complete set of input variables and find that the area under the curve (AUC) is about 0.63, which tells us that the best model found is not too overfitted but has certain predictive abilities.



Conclusion

The aim of this project was to analyse what variable influenced the churn rate of a bank, in other words, what influenced the client to leave. We found that mostly financial variables increased the probability to leave the bank, but also the age of the client. We hence implemented a machine learning model in order to predict the churn rate of the bank.

Appendix

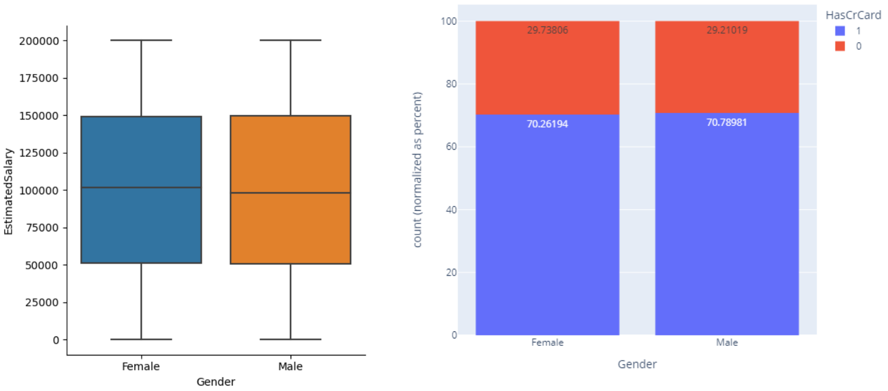


Figure 5: Gender variable : Estimated Salary (left) and Has Credit card (right)

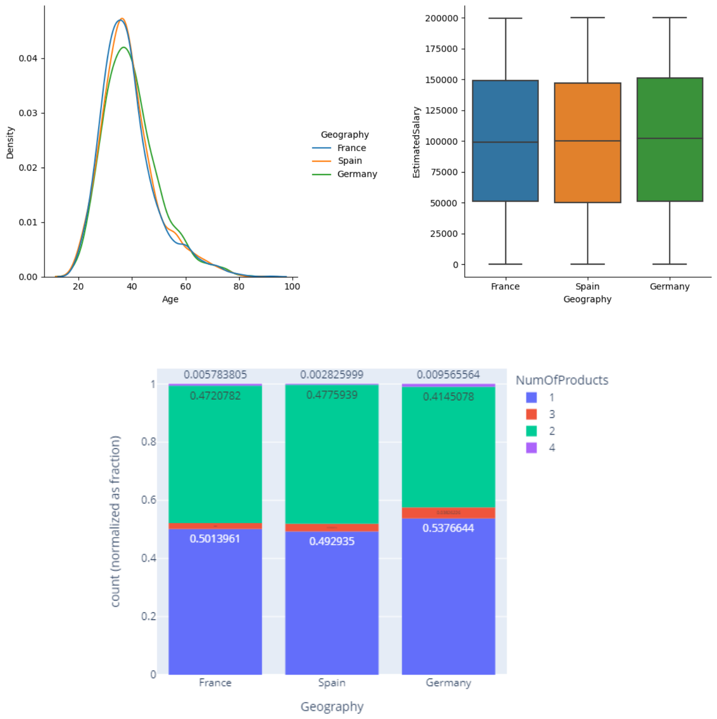


Figure 6: Geographical variable : Age (1st figure), Estimated Salary (second), and Number of products (3rd)

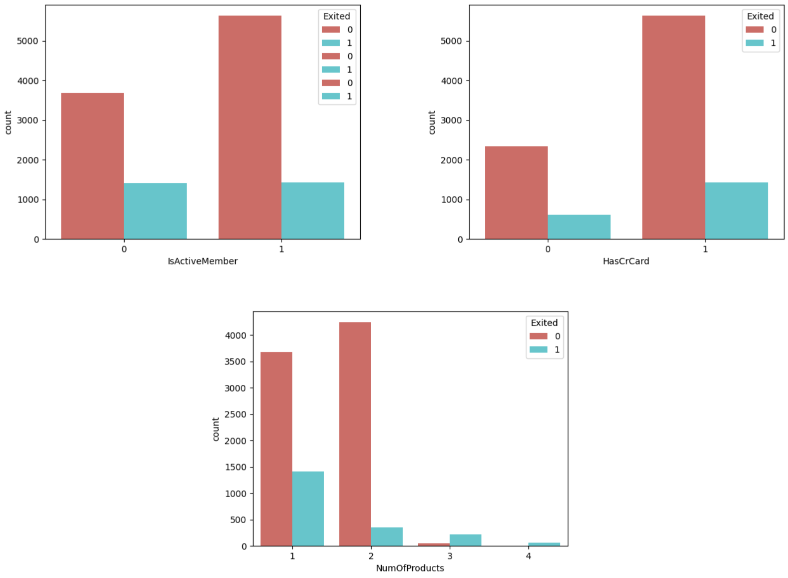


Figure 7: Financial variables among exited and non exited clinets : Is Active Member, Has a Credit Card, Number of products