



# Hackaton Results

Esteban Braganza

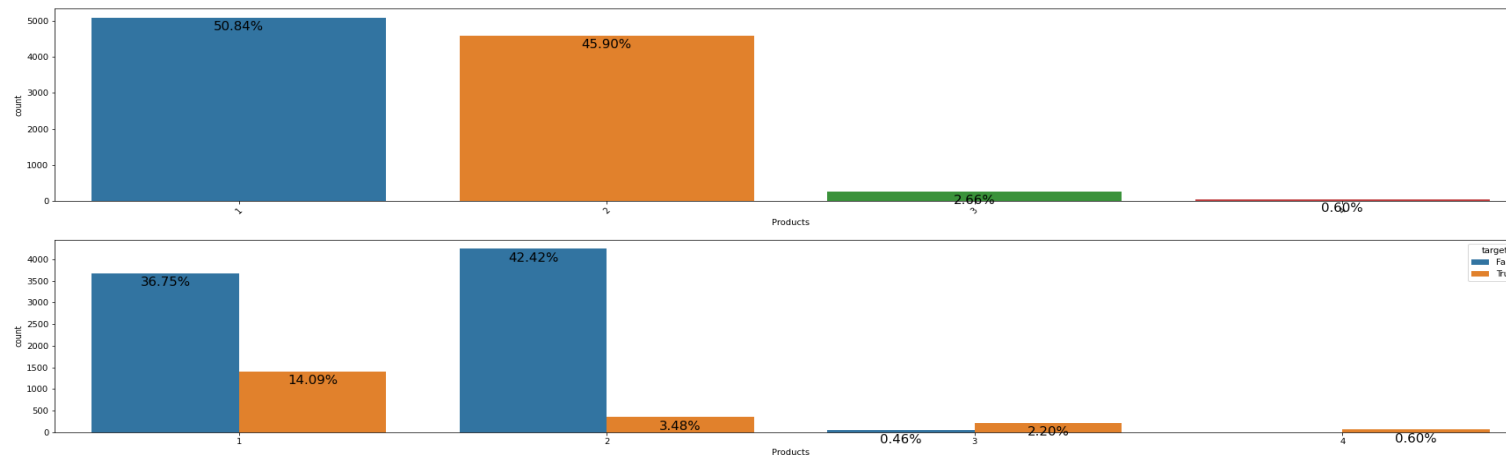
For the Finance and Risk Analyst position

# Desired Population

Filter Name	Action	Number of rows
Initial number of records	Check number of records	1545000
1 <sup>st</sup> filter	Erase clients with contracts before 2015	623242
2 <sup>nd</sup> filter	Erase Italian clients as they are not part of the company anymore	
3 <sup>rd</sup> filter	Erase clients with more than 75% nulls (count number of variables with nulls)	
4 <sup>th</sup> filter	Drop duplicated rows	
5 <sup>th</sup> filter	Erase clients with less than 2 years of information (those who entered until November 2017)	10000

# New Variables Created

	EstimatedSalary	age	account_balance	Score	Products
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	100090.239881	38.922000	76485.889288	650.528800	1.530200
std	57510.492818	10.488523	62397.405202	96.653299	0.581654
min	11.580000	18.000000	-0.000000	350.000000	1.000000
25%	51002.110000	32.000000	0.000000	584.000000	1.000000
50%	100193.915000	37.000000	97198.540000	652.000000	1.000000
75%	149388.247500	44.000000	127644.240000	718.000000	2.000000
max	199992.480000	92.000000	250898.090000	850.000000	4.000000



# Distributions and data separability

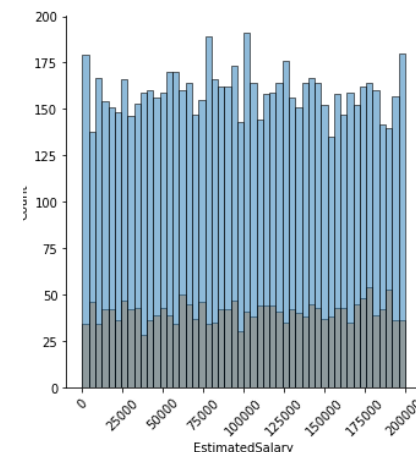
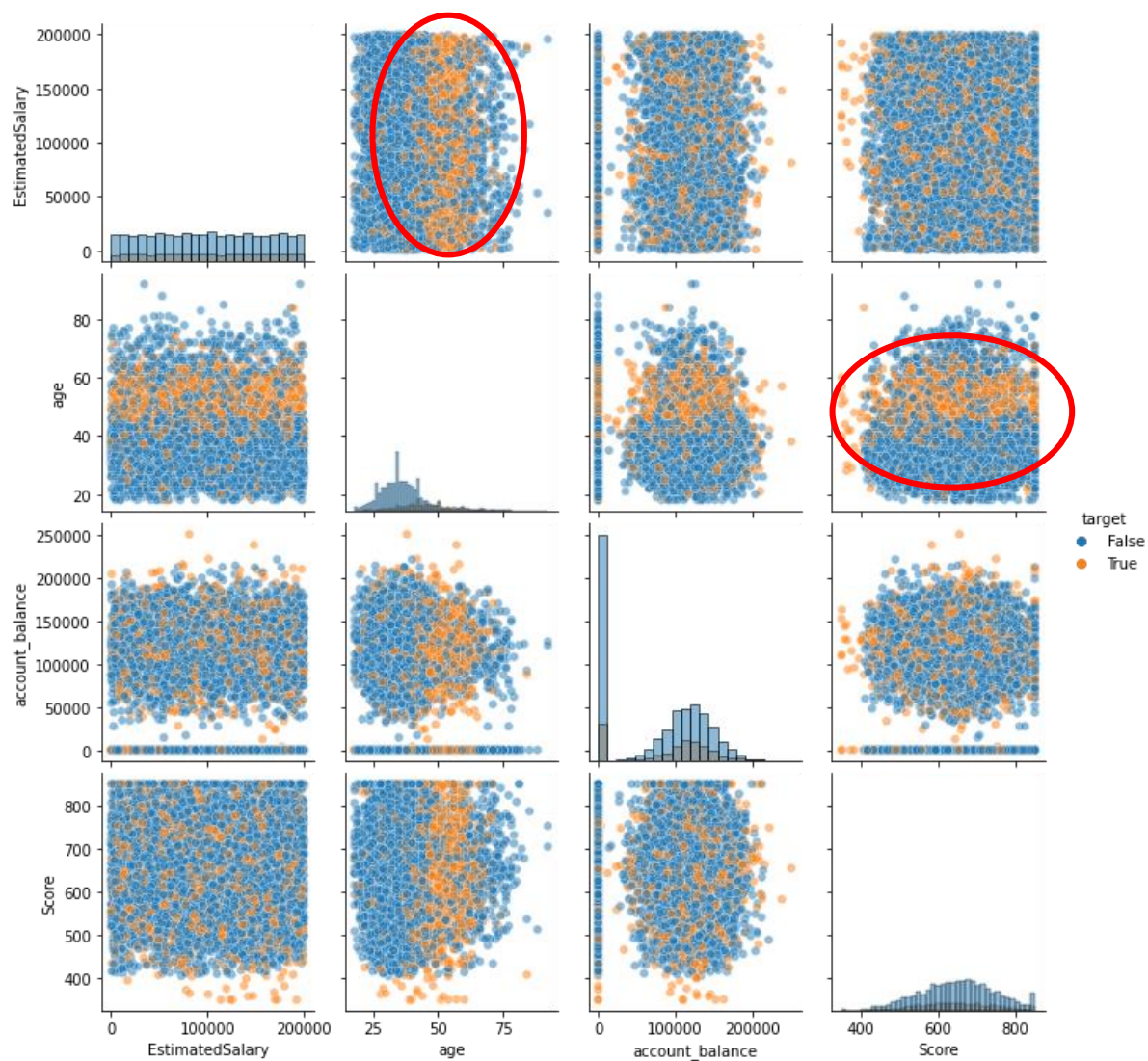


Figure size 2160x720 with 0 Axes

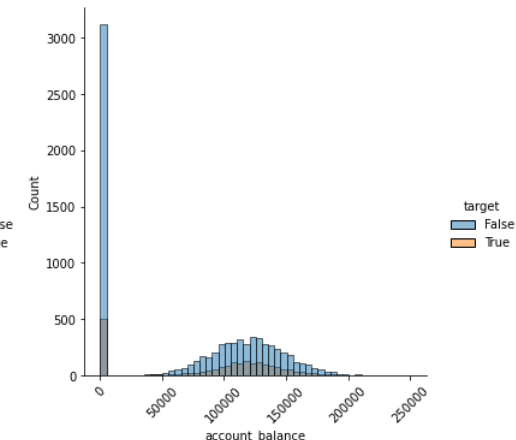
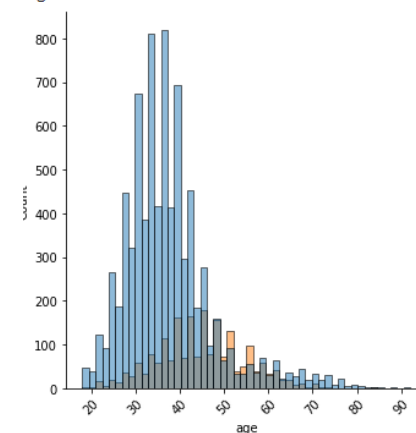
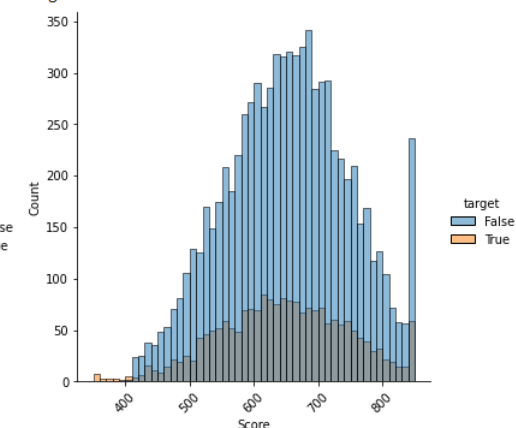
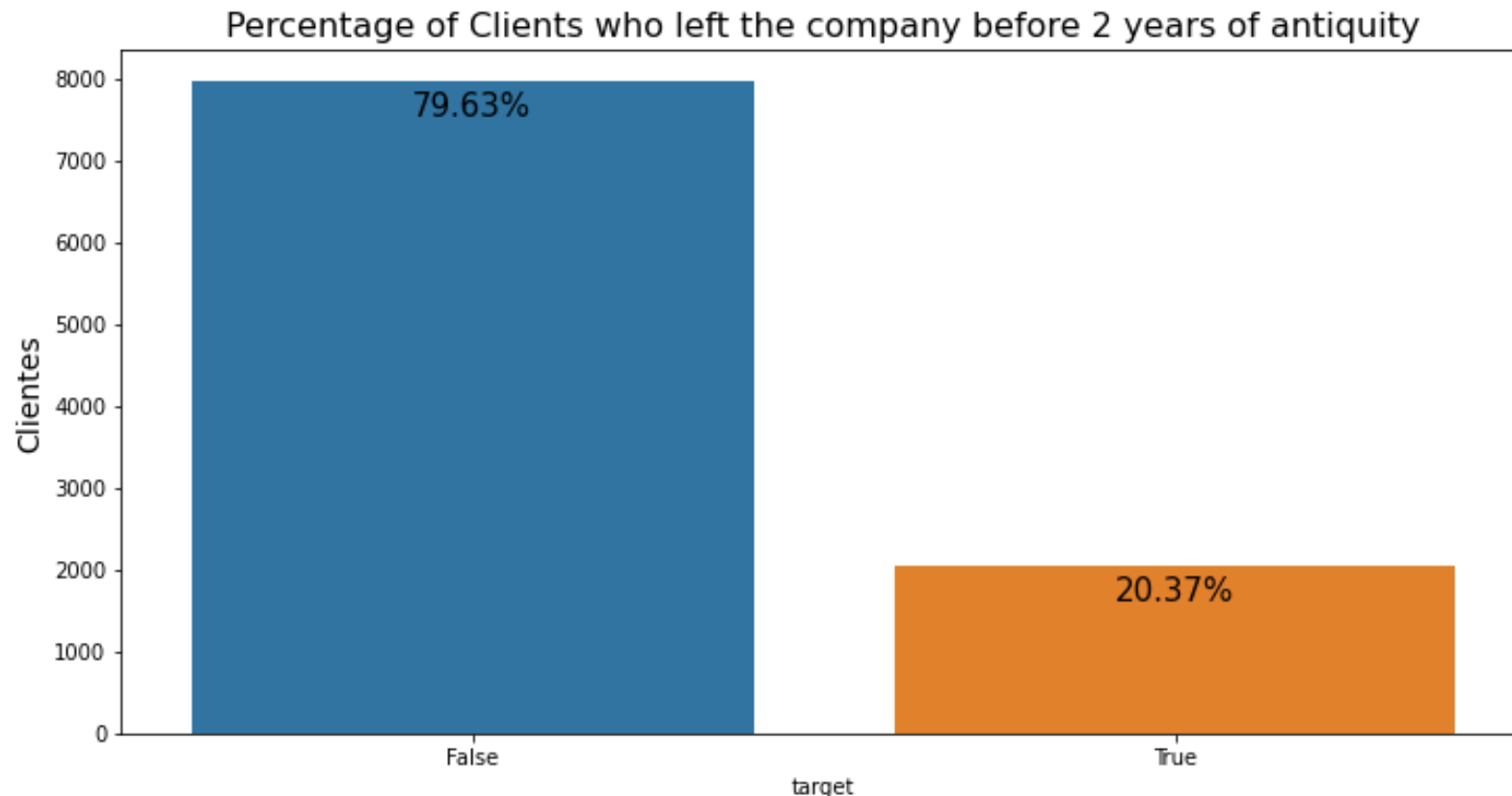


Figure size 2160x720 with 0 Axes



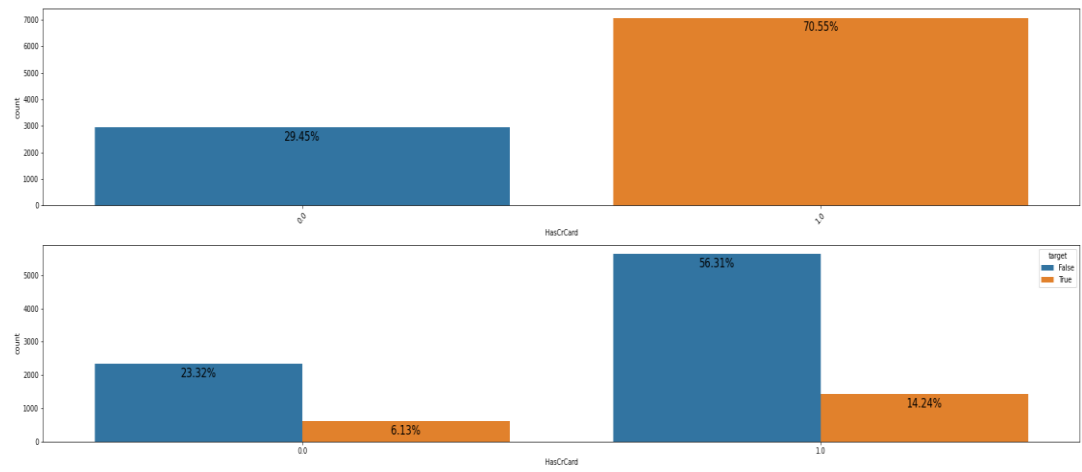
# Preprocessing: Target

- For the target variable we create a category for clients who left the company before 2 years of antiquity.
- We have a highly imbalanced data set.

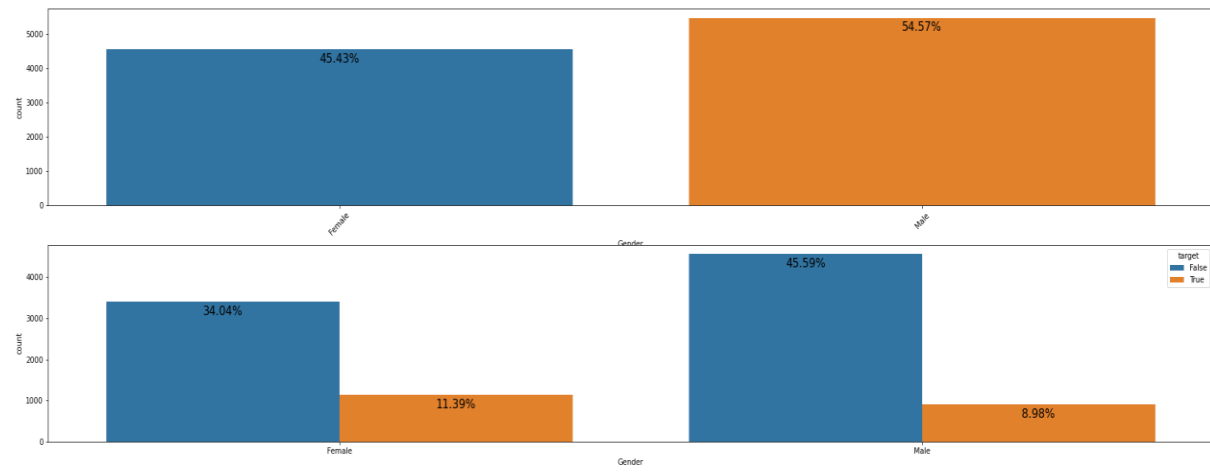


# Some relations of other variables

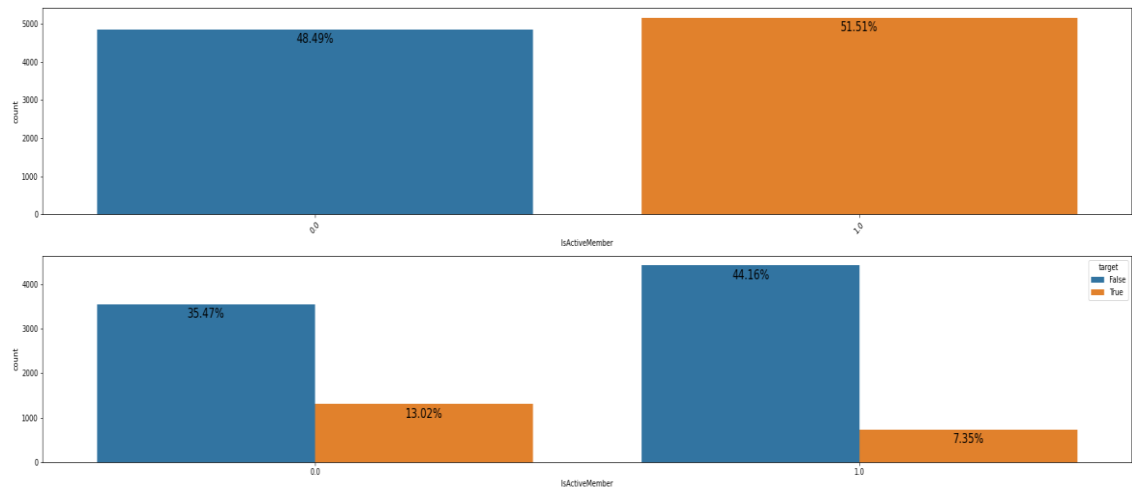
HasCrCard



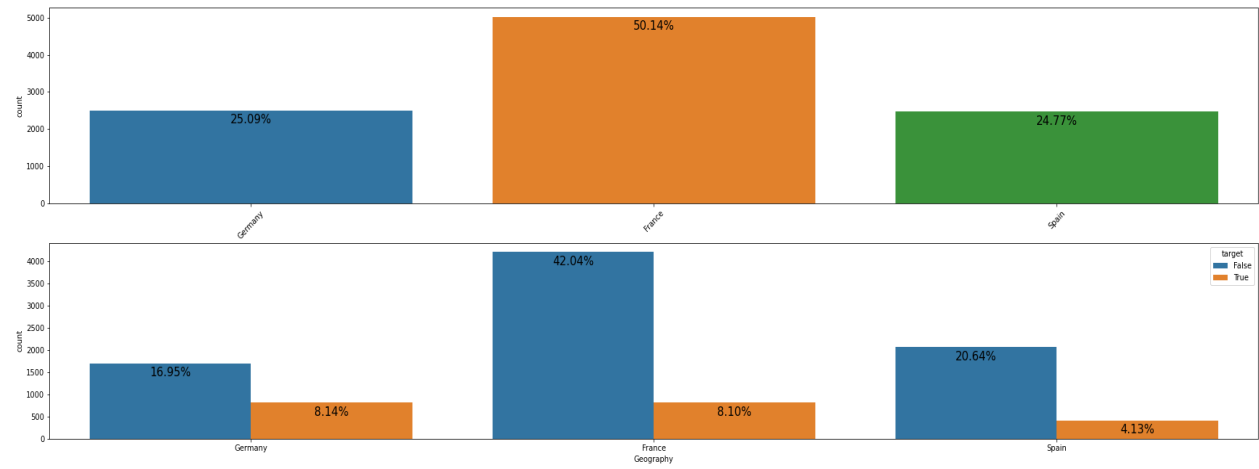
Gender



IsActiveMember



Geography

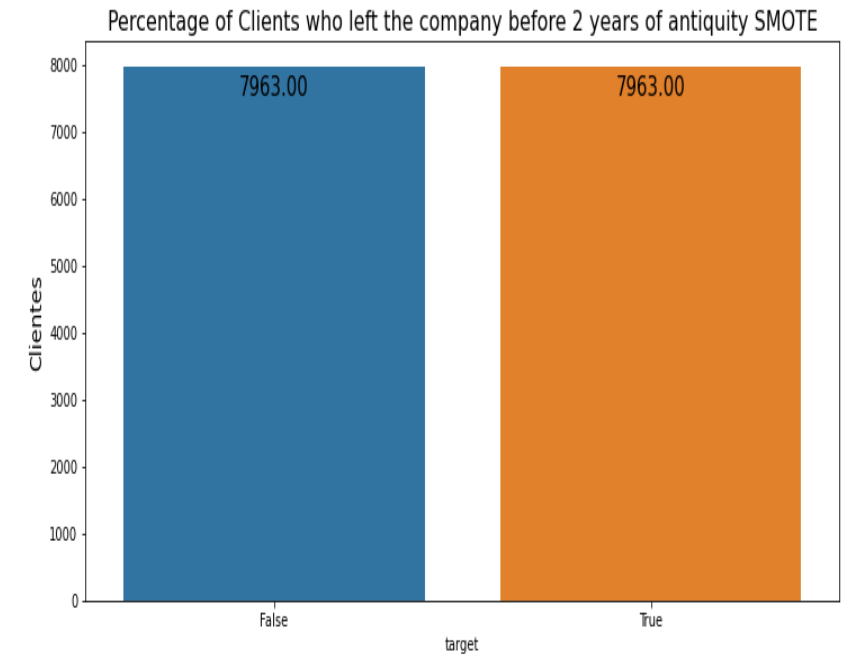


# Dummy Creation and Smote

- For preprocessing we looked for more interesting relations by creating dummy variables from the numerical variables we had at first.
- Performed a SMOTE approach to solve imbalance in data.

	Gender	HasCrCard	IsActiveMember	target	only_one_Product	more_than_one_product	Geography_France	Geography_Germany	Geography_Spain	EstimatedSalary_ (-188.401, 20009.67]	EstimatedSalary_ (20009.67, 40007.76]	EstimatedSalary_ (40007.76, 60005.85]
CustomerId												
15610711	1	0.0	0.0	False	1	0	0	1	0	0	0	0
15764170	0	1.0	0.0	True	1	0	0	1	0	0	1	0
15594720	1	1.0	1.0	False	0	1	0	1	0	0	0	0
15706552	0	0.0	1.0	False	1	0	1	0	0	0	0	0
15785358	0	1.0	1.0	True	1	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
15689152	0	0.0	0.0	False	1	0	0	0	1	1	0	0
15787204	1	0.0	0.0	False	1	0	0	0	1	0	0	0
15606887	1	1.0	0.0	True	1	0	1	0	0	0	0	0
15778320	1	0.0	0.0	False	1	0	0	1	0	0	0	0
15601172	0	1.0	0.0	False	1	0	1	0	0	0	0	0

40000 rows x 13 columns



# Models Results and Feature Importance

## Baseline: Logistic Regression

Accuracy Score: 0.8509102322661645  
F1 Score: 0.8503724745071615  
ROC AUC Score: 0.8504776392326246

## Random Forest Classifier

Accuracy Score: 0.8562460765850597  
F1 Score Macro: 0.8562297033136803  
ROC AUC Score: 0.856390198149345

## Support Vector Classifier

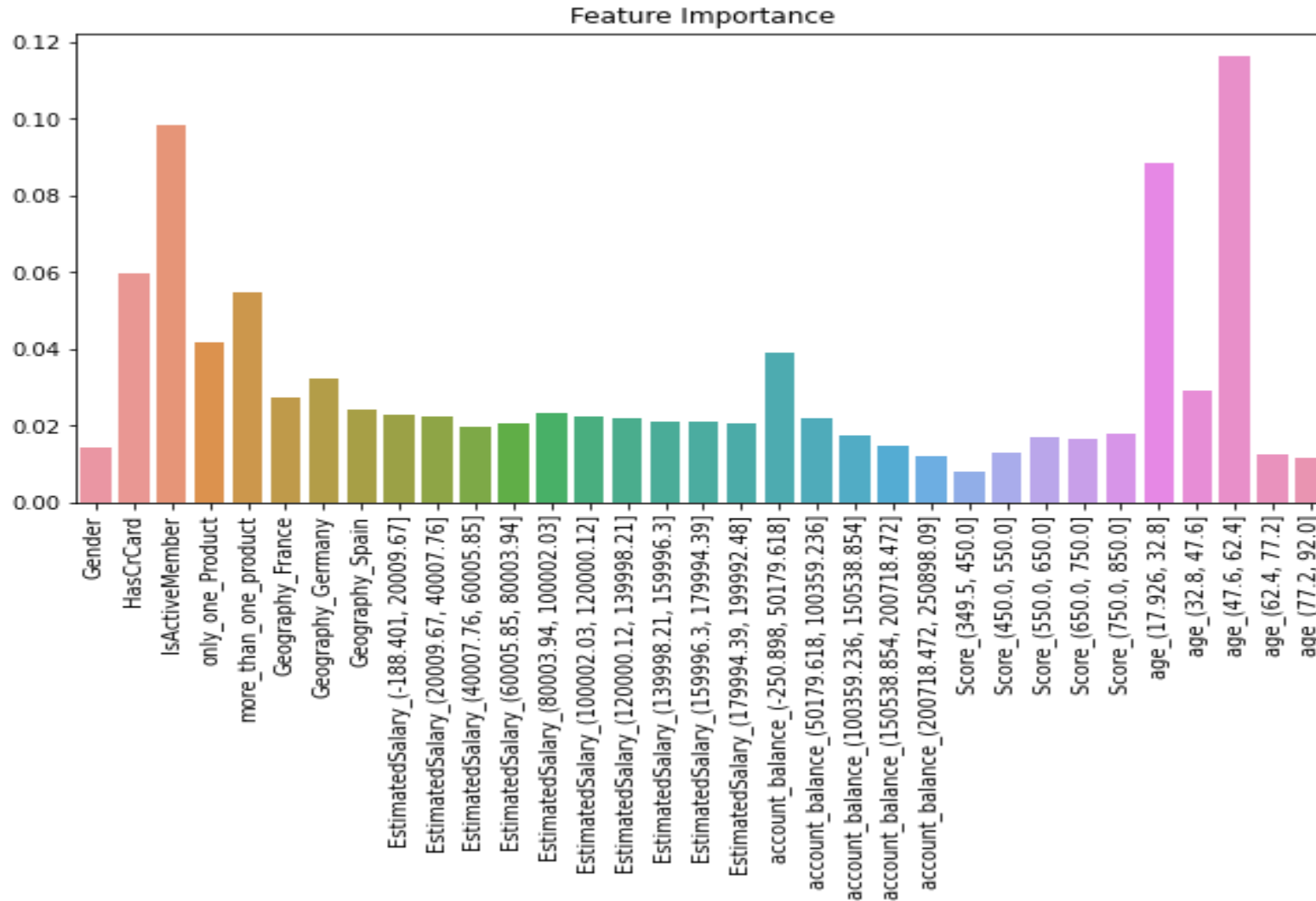
Accuracy Score: 0.879472693032015  
F1 Score: 0.8788846438957633  
ROC AUC Score: 0.8789625928086132

## XGBoost Classifier

Accuracy Score: 0.8788449466415568  
ROC AUC Score: 0.8785602251052225  
F1 Score Macro: 0.8786302483854158



# Feature Importance in XGBoost



## Churn Client Profile

Between 18 and 33 years  
And between 45 and 65 years

Active member with Credit Card  
Has bought more than one product.

Is in the lowest interval of account  
Balance.

More likely are from Germany