Churn Prediction

Diandra MELO

SUMMARY

1. CONTEXT

Fictional bank company called **EuroBank**:

- Main profit: bank accounts
  - 20 % of client's salary if above the salaries average;
  - 15 % of client's salary otherwise.

Business Problem

IThe Analytics team wants to predict the **probability of a client to enter in churn**, based on previous churn clients profiles

Business Understanding

The **CEO wants to reduce the clients evasion** in order to maintain the company's margin of profit stable

2. CHALLENGES

Problem:

Churn probability definition for **avoiding clients evasion**

Solution:

Apply data analysis and **Machine Learning algorithms** for churn prediction

Outputs:

- Action plan of **incentives for clients in Churn**
- Report with **model performance and financial impacts**

3. Methodology

Data description, Hypothesis,
Exploratory Data Analysis, ML models

3.1. Data Description

Number of Columns: 14

Number of Rows: 10,000

row_number	int64
customer_id	int64
surname	object
credit_score	int64
geography	object
gender	object
age	int64
tenure	int64
balance	float64
num_of_products	int64
has_cr_card	int64
is_active_member	int64
estimated_salary	float64
exited	int64

Data available at [Kaggle]

3.2. Descriptive Statistics

Numerical Variables

	mean	median	std	min	max	range	skew	kurtosis
credit_score	650.53	652.00	96.65	350.00	850.00	500.00	-0.071607	-0.425726
age	38.92	37.00	10.49	18.00	92.00	74.00	1.011320	1.395347
tenure	5.01	5.00	2.89	0.00	10.00	10.00	0.010991	-1.165225
balance	76485.89	97198.54	62397.41	0.00	250898.09	250898.09	-0.141109	-1.489412
num_of_products	1.53	1.00	0.58	1.00	4.00	3.00	0.745568	0.582981
estimated_salary	100090.24	100193.92	57510.49	11.58	199992.48	199980.90	0.002085	-1.181518

- Highly positive-skewed and high kurtosis distribution of age;
- Most variables present low kurtosis, which may indicate absence of outliers.

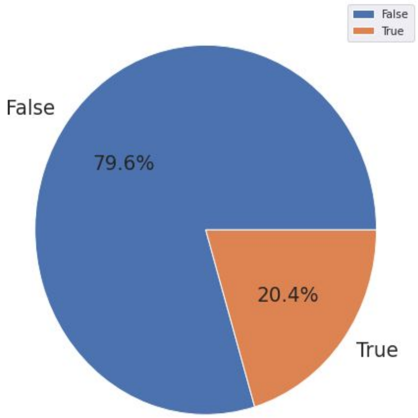
Some relevant average values:

- Credit: € 650.53
- Age: 39 years old
- Tenure: 5 months

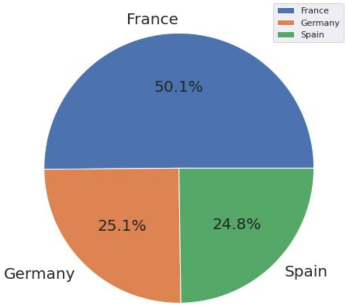
3.2. Descriptive Statistics

Categorical Variables

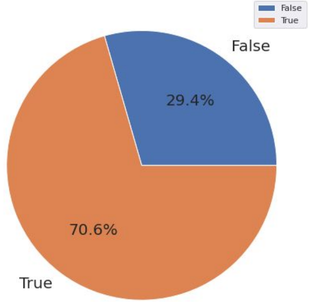
exited



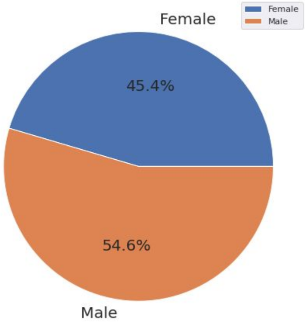
geography



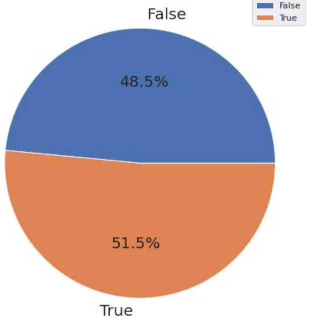
has_cr_card



gender

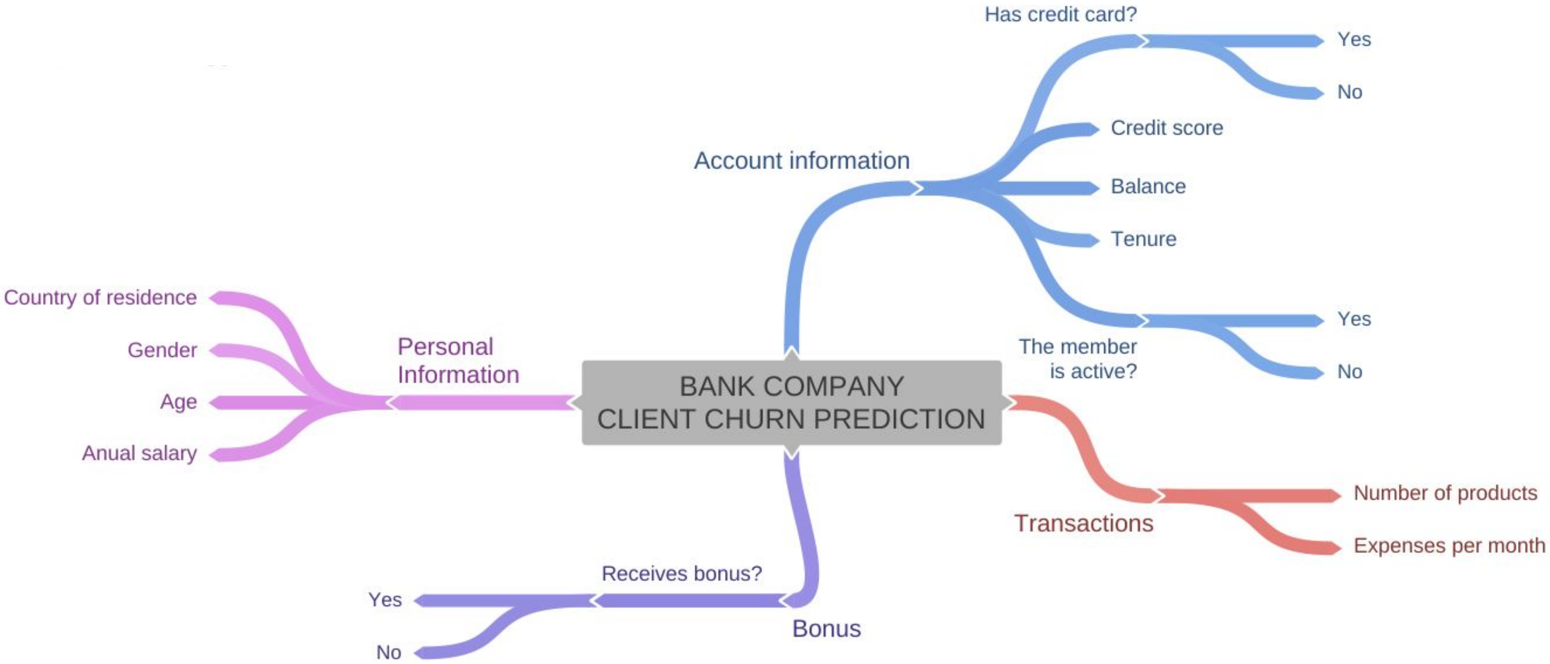


is_active_member



3.3. Hypothesis Creation

Hypothesis MindMap



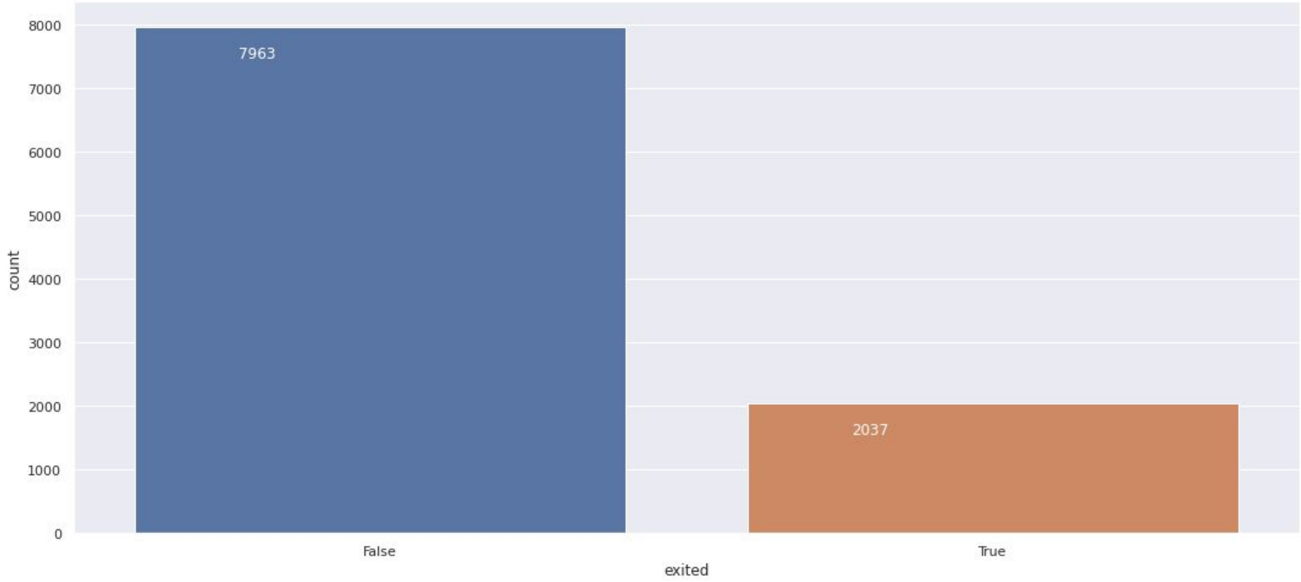
3.3. Hypothesis Creation

Hypothesis List

1. Younger clients are more likely to churn
2. Churn percentage does not change much between countries in Europe
3. Clients with higher estimated salary are less likely to churn
4. There are possibly no significant differences between gender in churning
5. Clients with no credit card are more likely to churn
6. Accounts with a balance of more than 50,000 € are less likely to indicate client churning
7. Accounts who are active for less than 2 years are more likely to close
8. Active members are less likely to churn
9. Clients who acquired more products are less likely to churn

3.4. Exploratory Data Analysis

Univariate Analysis

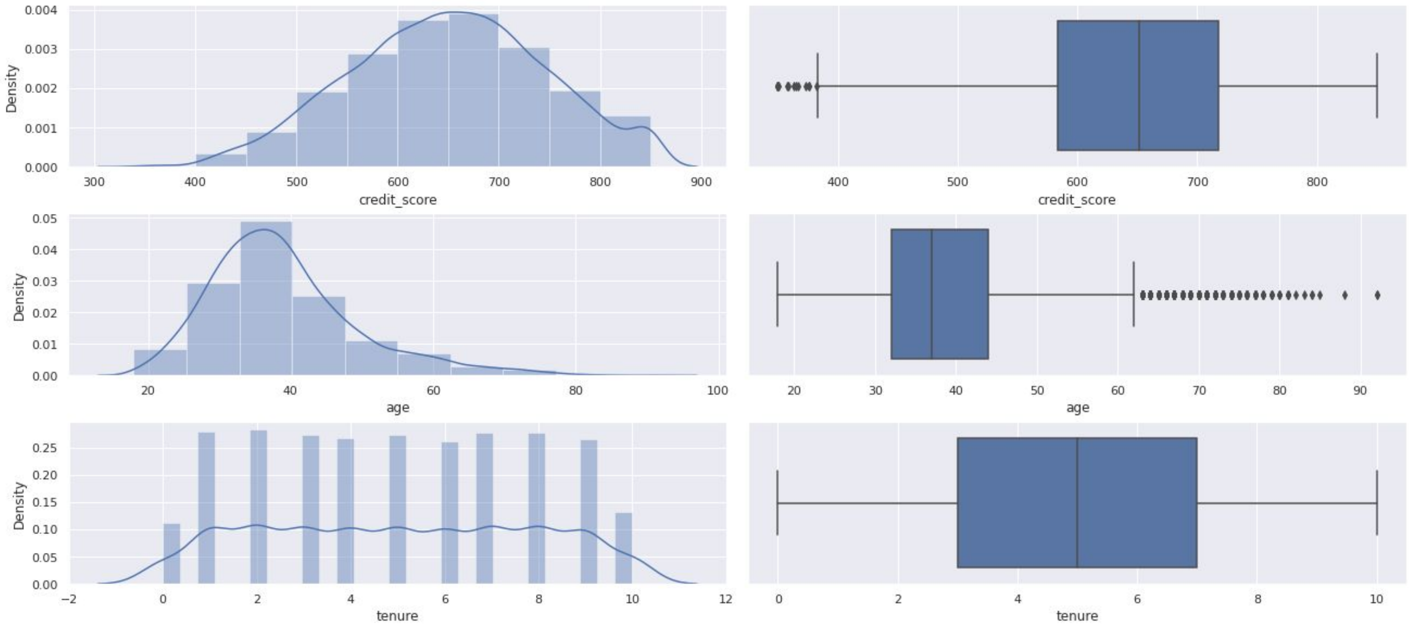


- Percentage of clients in churn: 20.37 %

3.4. Exploratory Data Analysis

Univariate Analysis

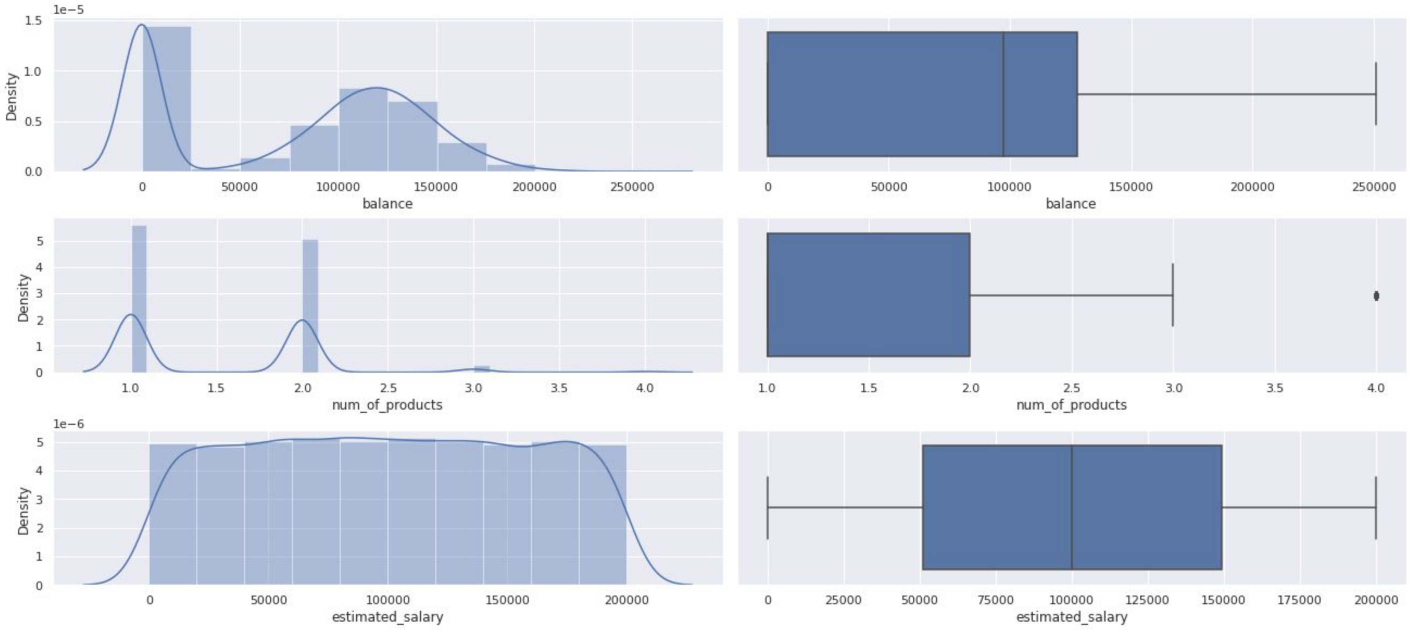
Numerical attributes: credit score / age / tenure



3.4. Exploratory Data Analysis

Univariate Analysis

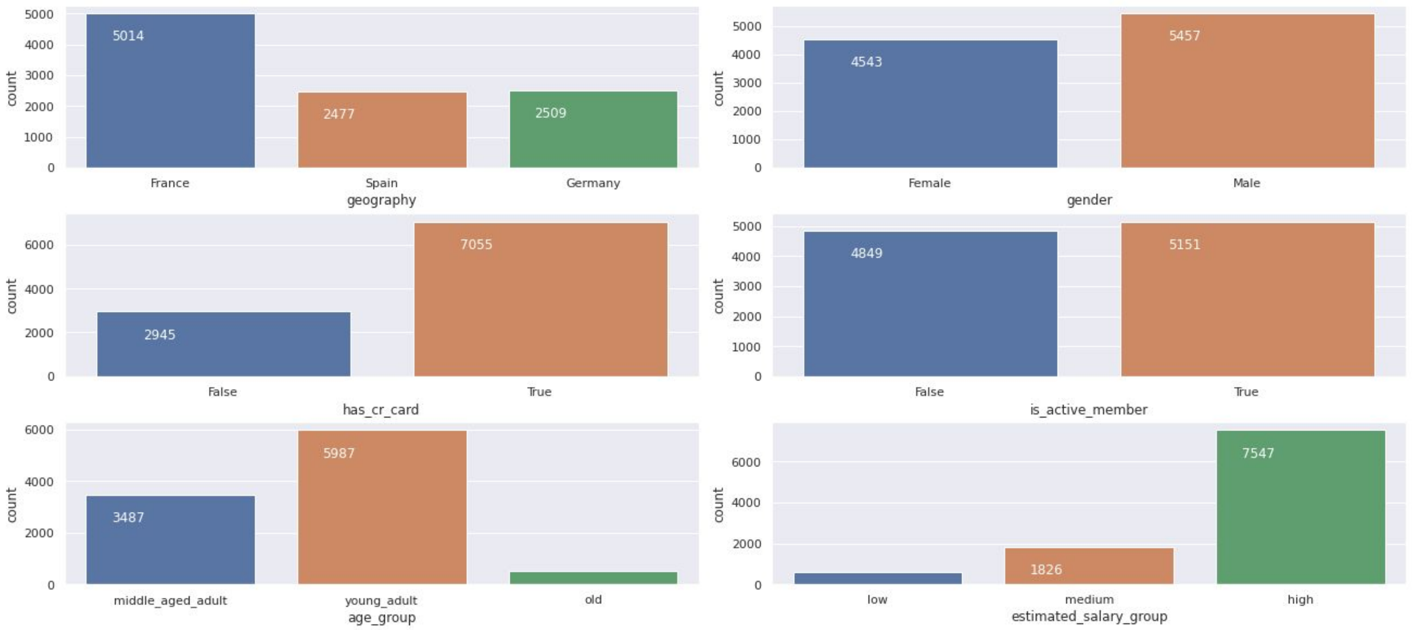
Numerical attributes: balance / number of products / estimated salary



3.4. Exploratory Data Analysis

Univariate Analysis

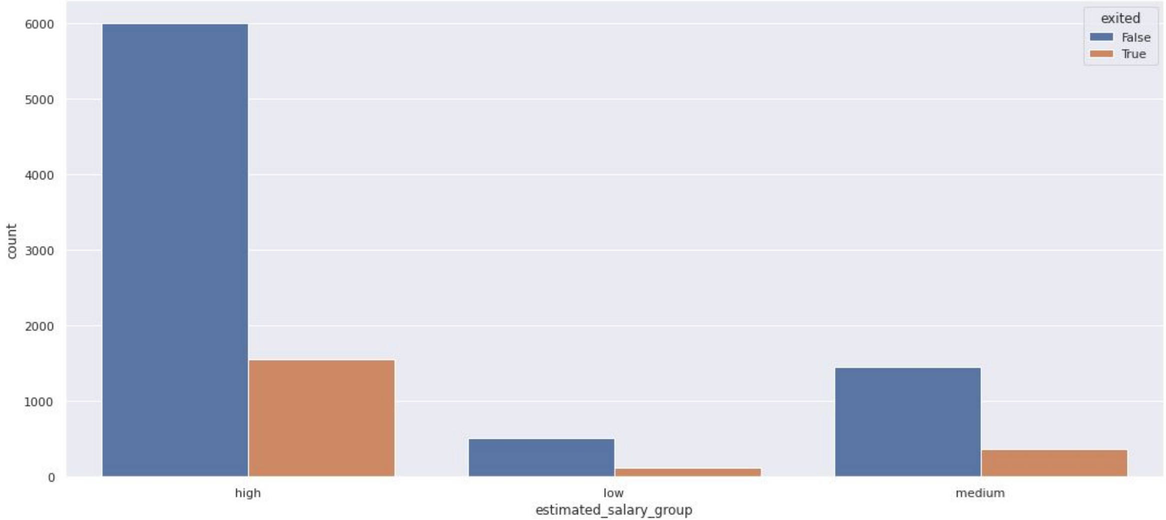
Categorical attributes



3.4. Exploratory Data Analysis

Hypothesis 03

Clients with higher estimated salary are less likely to churn



This is a **FALSE** hypothesis

Churn percentage for low salary range: 18.98 %

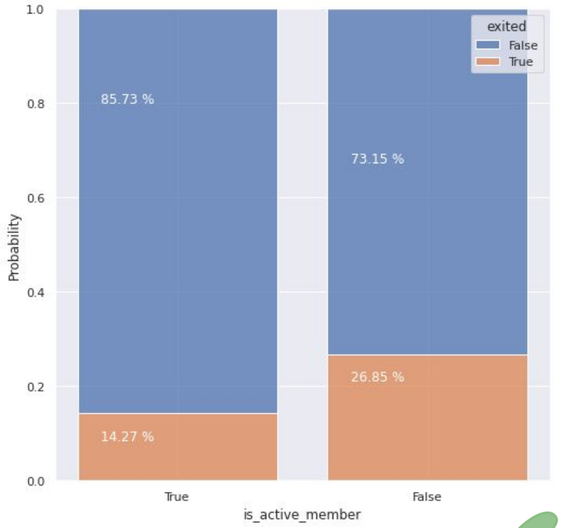
Churn percentage for medium salary range: 20.26 %

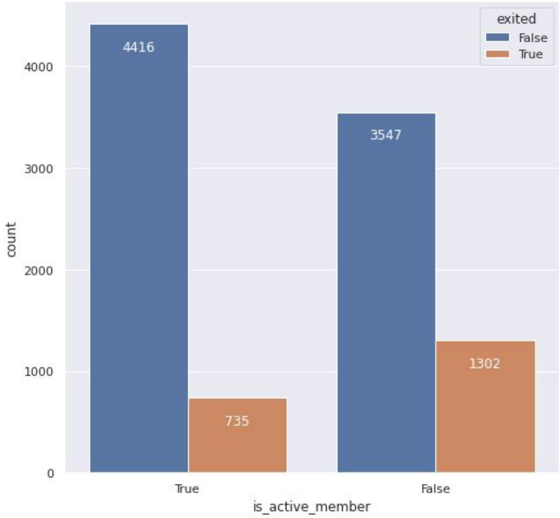
Churn percentage for high salary range: 20.51 %

3.4. Exploratory Data Analysis

Hypothesis 08

Active members are less likely to churn





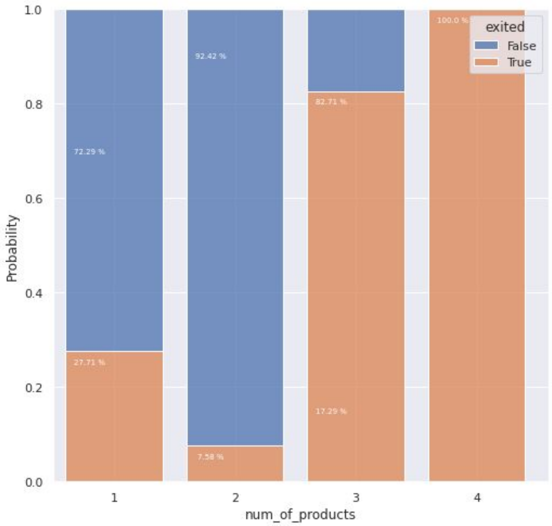
This is a **TRUE** hypothesis

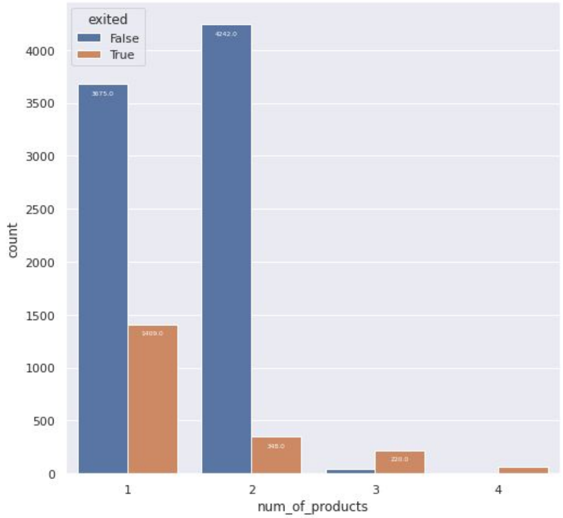
The percentage difference is of about 13 %

3.4. Exploratory Data Analysis

Hypothesis 09

Clients who acquired more bank services are less likely to churn





This is a **FALSE** hypothesis

The percentage of clients that acquired more than 3 bank products that exited is **way higher** than the ones that did not exit.

3.4. Exploratory Data Analysis

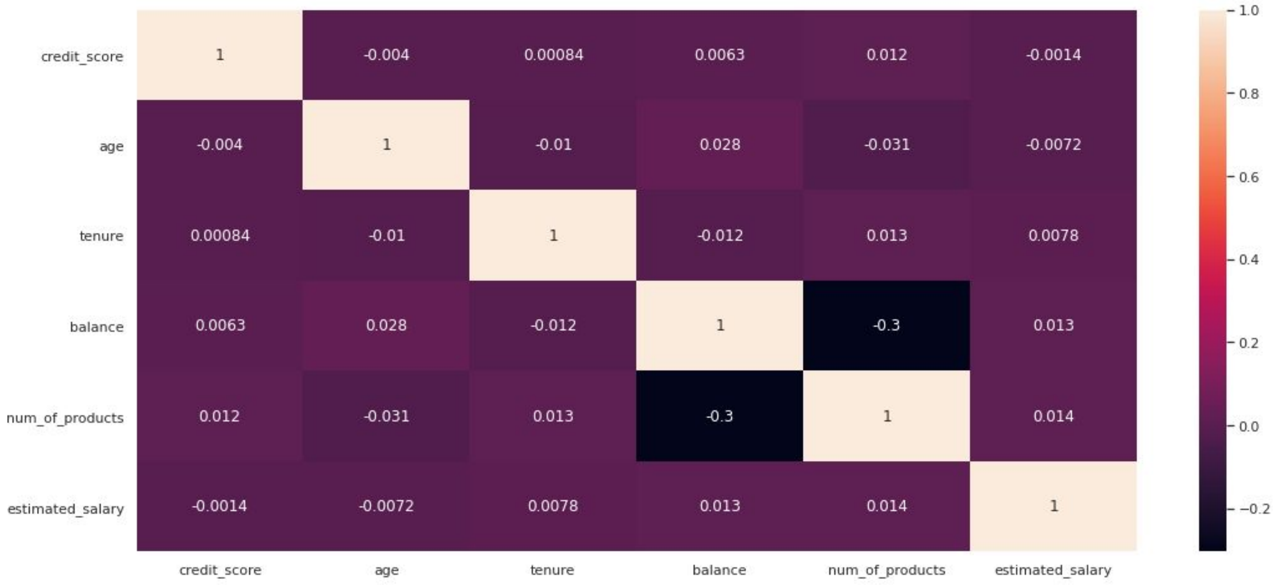
Hypothesis Final Table

Hypothesis	Conclusion	Relevance
H1 - Younger clients are more likely to churn	FALSE	High
H2 - Churn percentage does not change much between countries in Europe	FALSE	High
H3 - Clients with higher estimated salary are less likely to churn	FALSE	Low
H4 - There are possibly no significant differences between gender in churning	FALSE	Medium
H5 - Clients with no credit card are more likely to churn	TRUE	Low
H6 - Accounts with a balance of more than 50,000 € are less likely to indicate client churning	FALSE	Medium
H7 - Accounts who are active for less than 2 years are more likely to close	TRUE	Low
H8 - Active members are less likely to churn	TRUE	High
H9 - Clients who acquired more products are less likely to churn	FALSE	High

3.4. Exploratory Data Analysis

Multivariate Analysis

Numerical attributes

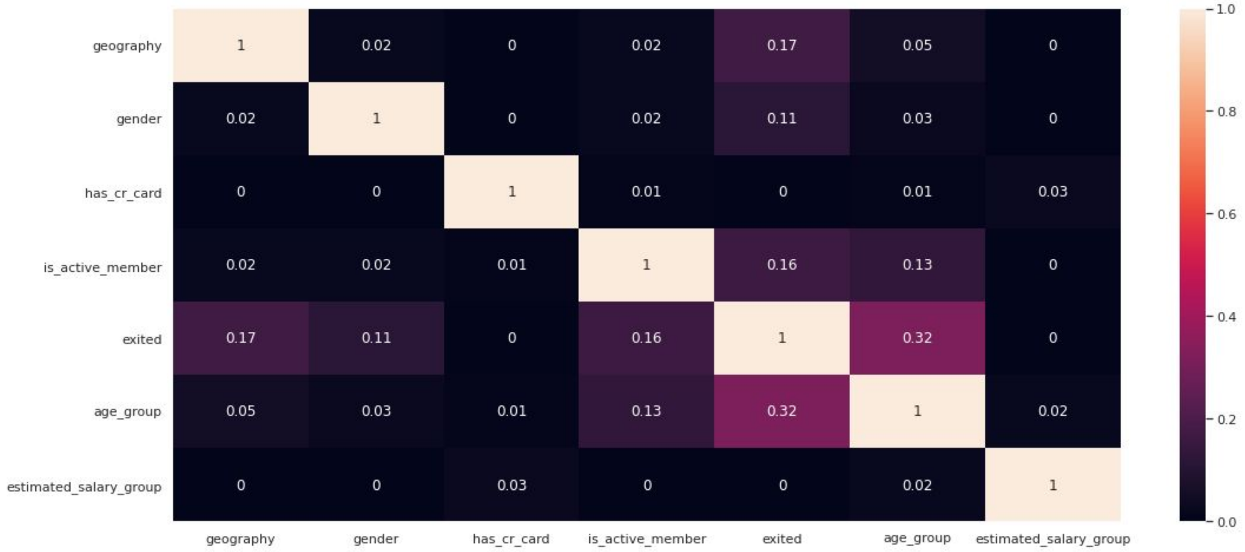


- Relevant correlation: balance & number of products

3.4. Exploratory Data Analysis

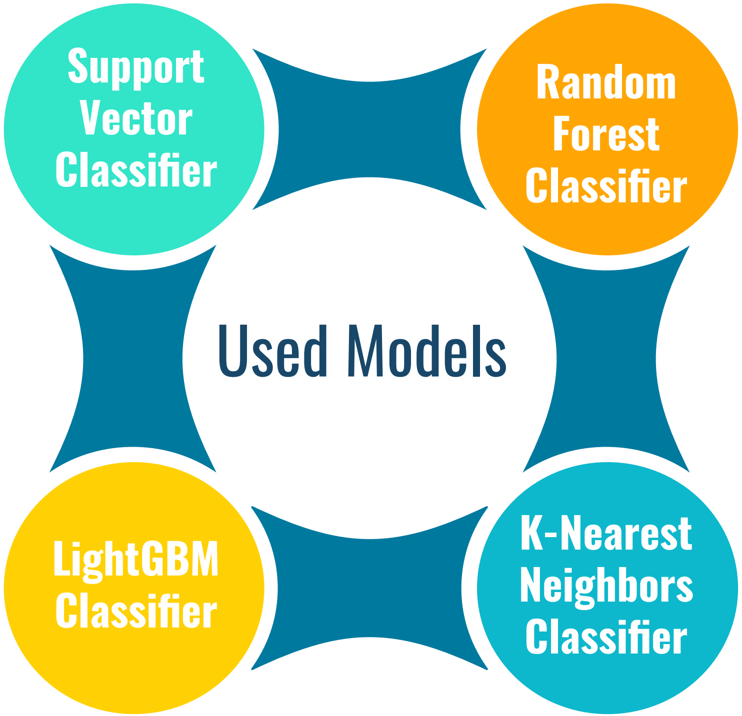
Multivariate Analysis

Categorical attributes



- Relevant correlations:
  - Exited (target variable) & age group;
  - Exited (target variable) & geography;
  - Exited (target variable) & is_active_member;

3.5. Machine Learning modelling



Model's Performance

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
LightGBM	0.8594 +/- 0.0096	0.7461 +/- 0.0406	0.4817 +/- 0.0402	0.5841 +/- 0.0322	0.8517 +/- 0.0147
RF	0.8610 +/- 0.0077	0.7698 +/- 0.0414	0.4647 +/- 0.0307	0.5784 +/- 0.0247	0.8472 +/- 0.0092
KNN	0.8264 +/- 0.0082	0.6368 +/- 0.0362	0.3619 +/- 0.0321	0.4608 +/- 0.0321	0.7485 +/- 0.0238
SVM	0.8387 +/- 0.0100	0.8166 +/- 0.0607	0.2791 +/- 0.0464	0.4138 +/- 0.0539	0.7966 +/- 0.0230

Model's choice: **Random Forest Classifier**

- Better model accuracy, with good F1-Score in comparison to others

4. Conclusion

Business performance, Model performance

4.1. Business Performance

19.65 %

Current churn rate

7,463,041.16 €

Profit loss if all clients in Churn leave

3,532,126.35 €

Expected ROI with the model

4.1. Business Performance

€ 10,000 investment

Business plan - reducing churn

01 Top 100 customers with highest churn probability

- Give € 100 to each possible client in churn
  ROI: 16,773.52 %

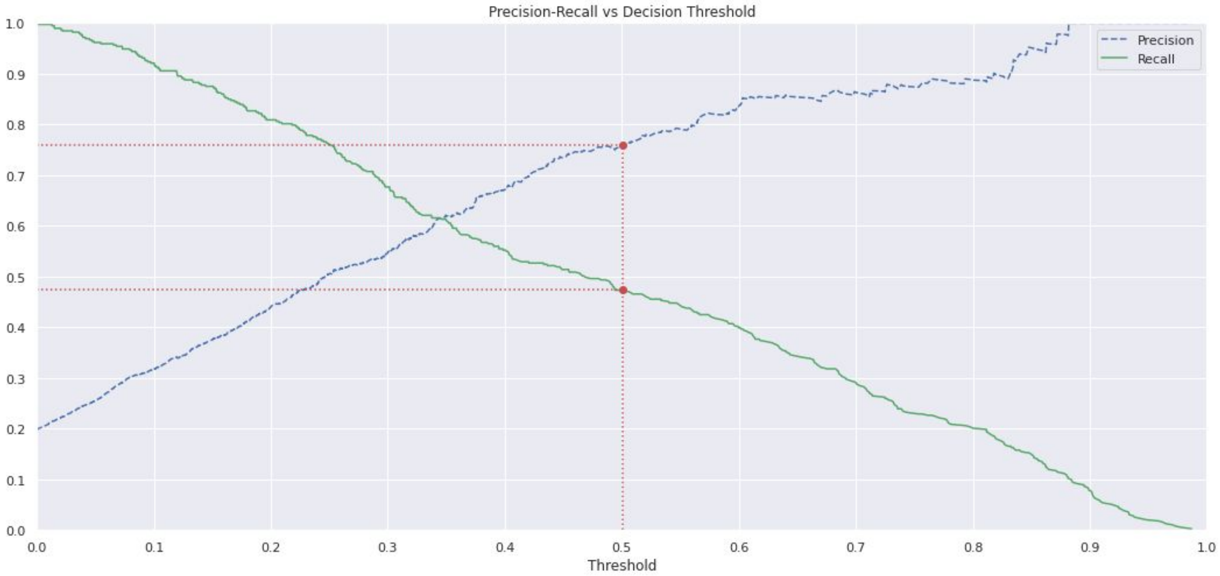
02 Maximum returned value with "0-1 Knapsack-Problem" Approach

- Select the optimal combination of clients that maximize the ROI
- ROI: 25,471.09 %

03 Realistic Approach MRV with "0-1 Knapsack Problem" approach

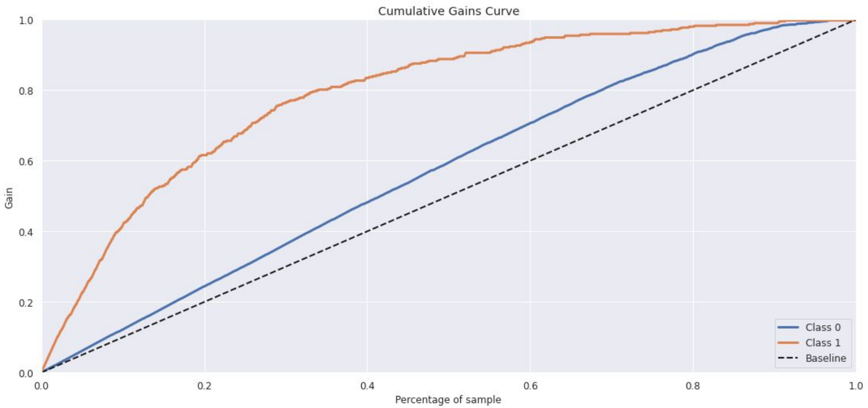
- Consider real probability constraints
- ROI: 33,379.67 %

4.2. ML Model Performance



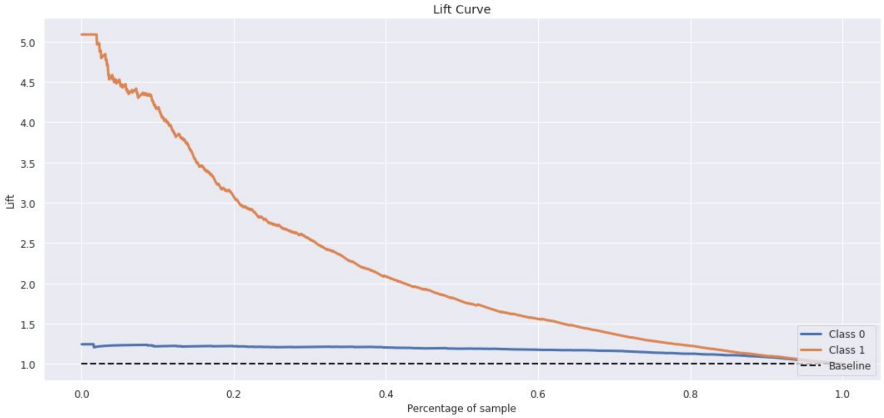
- The threshold between precision and recall is 50.1 %
  - which means, this is the predicted probability of an observation belonging to the positive class

4.2. ML Model Performance



The **cumulative gain** for the model considering 20% of the sample is around 58%, compared to the baseline model

From the **lift curve,** it is possible to observe that the model allows addressing 3x more targets for the 20% group, compared with random guessing



5. Next Steps

- Work on the model deployment for further predictions through an API

Thanks!

Any questions?

You can find me at:

- LinkedIn: Diandra Melo
- GitHub: @diandramelo
- Gmail: dcxsmelo@gmail.com