Honor code

By submitting this assignment, I affirm the following:

1. If I used AI tools like ChatGPT, Co-Pilot, etc., I only sought guidance or clarification. Any generated content has been fully understood and appropriately modified to align with the assignment.

2. I understand the submitted code and can explain my work if asked.

We declare that we have read, understood, and agree to abide this honor code.

Group 25	
Student names	Student numbers
Jianxin Cui	664138jc
Vincent Decraene	605856vd
Samuel Vervier	738542sv
Kevin Vu	713837tv

Date: 26/11/2024

MSc. Business Information Management
Course: Big Data Management and Analytics
Course Code: BM04BIM
Group Assignment

Part 2: Data Analytics (20 Points)

1. Business Context

EuroBank International (EBI) is facing the challenge of customer churn, which means that its customers are leaving their service for various reasons. The bank seeks to understand the underlying causes of customer attrition and proactively identify customers who are at risk of leaving. EBI has engaged your group to conduct a comprehensive data analytics study. Your task is to deliver useful insights to the bank.

2. Data pre-processing

1. Categorical variable encoding

Gender and nationality were the two categorical variables in the dataset. LabelEncoder was used to convert these into numerical representations. The nation field was similarly classified into numeric categories, while the gender variable was binary (0 for female, 1 for male). Compatibility with machine learning models—which need numerical inputs—was guaranteed by this procedure.

2. Removal of Irrelevant Features

Since the customer_id variable is a unique identifier and has no bearing on forecasting customer attrition, it was eliminated from the dataset.

3. Handling Missing Values

The dataset's missing values were found and imputed using the corresponding column means. This method preserved the completeness of the data without appreciably increasing bias.

4. Outlier Detection and Removal

The Interquartile Range (IQR) approach was used to identify outliers in the numerical data. For every characteristic, values that fall outside of Q1−1.5×IQR and Q3+1.5×IQR were eliminated since they were deemed outliers. By lessening the effect of extreme values, this step enhanced the analysis's robustness.

5. Final Dataset

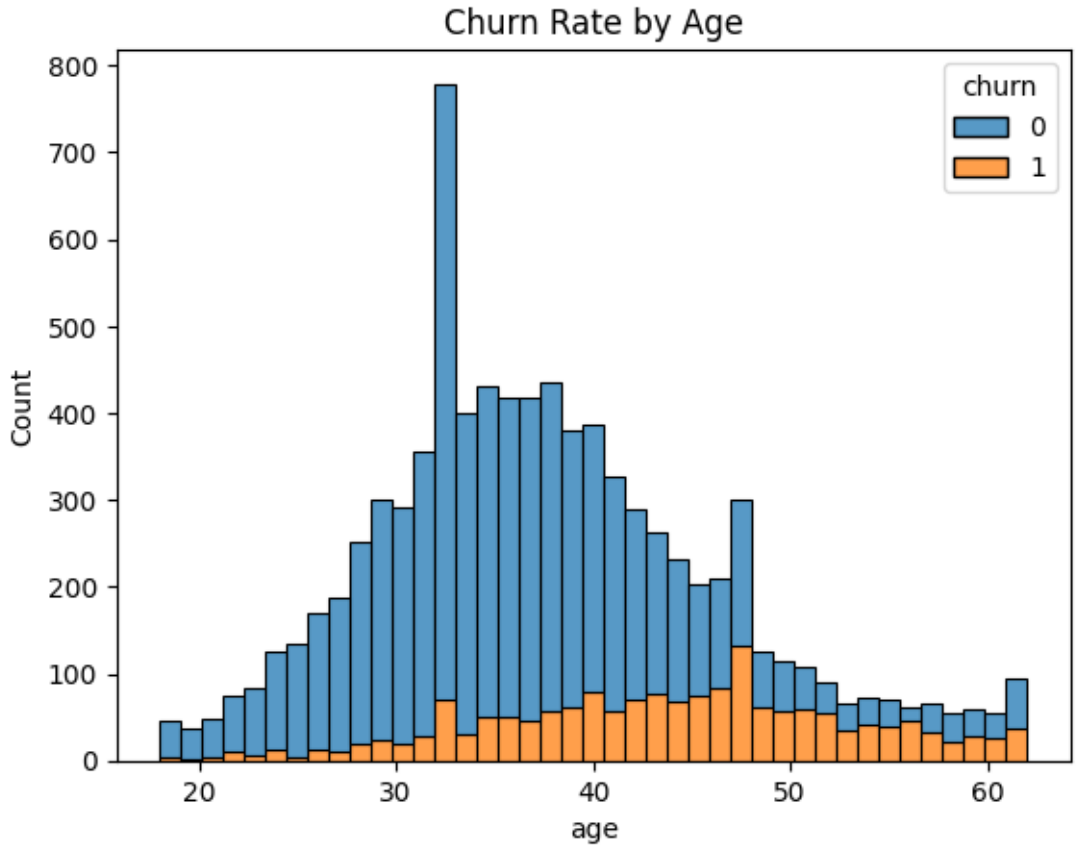
Finally, we get our final cleaned dataset with 8,612 rows and 12 columns.

3. Exploratory data analysis

Overall Churn Rate:

The rate of churn is roughly 20.26%. This suggests that approximately one out of every five clients is quitting the bank.

How does the Churn rate of customers vary by demographic



Low Churn for Younger Customers (Ages 18–30):

Customers between the ages of 18 and 30 have churn rates that are continuously below 15%, with the lowest rates seen at ages 19 (4.17%) and 25 (2.96%).

Observation: Younger people are less likely to churn It might be linked tot the fact that they need simpler financial services

Moderate Churn for Middle-Aged Customers (Ages 31–40):

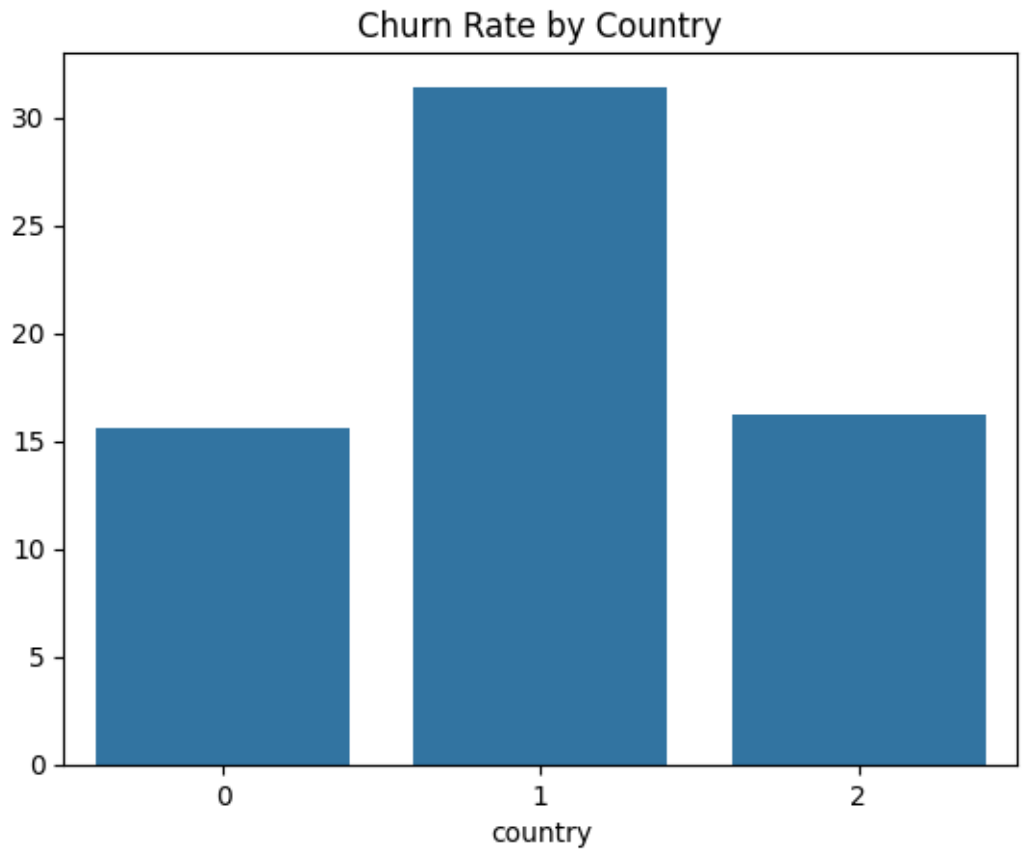
At age 40, churn rates increase gradually to 20.21%.

Observation: Clients in this age range may start looking into other options or become more financially complicated.

High Churn for Older Customers (Ages 41–62):

After age 40, churn rates increase significantly, reaching a peak of 59.35% for clients 52 years of age and older.

Observation: Dissatisfaction, a lack of individualized services, or shifting budgetary priorities could cause older clients to leave.



Country 0: Churn rate is 15.58%.

Country 1: Churn rate is 31.42%.

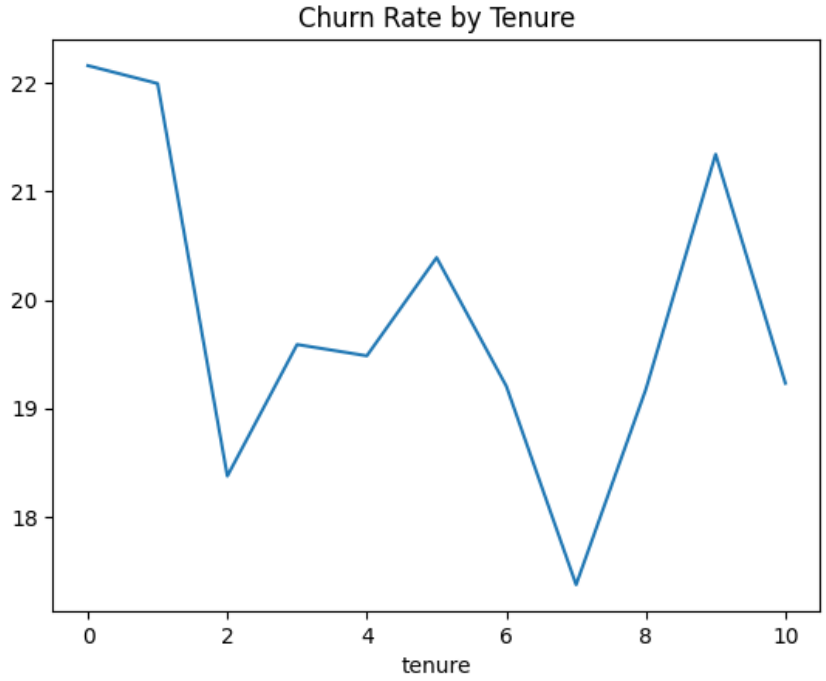
Country 2: Churn rate is 16.25%.

Compared to the other countries, Country 1 exhibits noticeably higher churn rates, suggesting a regional problem that may require more research.

Churn Rate by Gender:

- **Female customers (gender = 0):** The churn rate is **24.25%**.
- **Male customers (gender = 1):** The churn rate is **15.97%**.
- This suggests that female customers are more likely to churn compared to male customers.

Relationship Between Tenure and Churn:



First Tenure: 0–2 years

In the first year, the churn rate is high (over 22%), but by the second year, it has drastically decreased.

Observation: New clients are more likely to depart, perhaps as a result of unfulfilled expectations or discontent with the services they received initially.

Mid Tenure 3-6 years

For mid-tenure clients, the churn rate is around 18% to 20%.

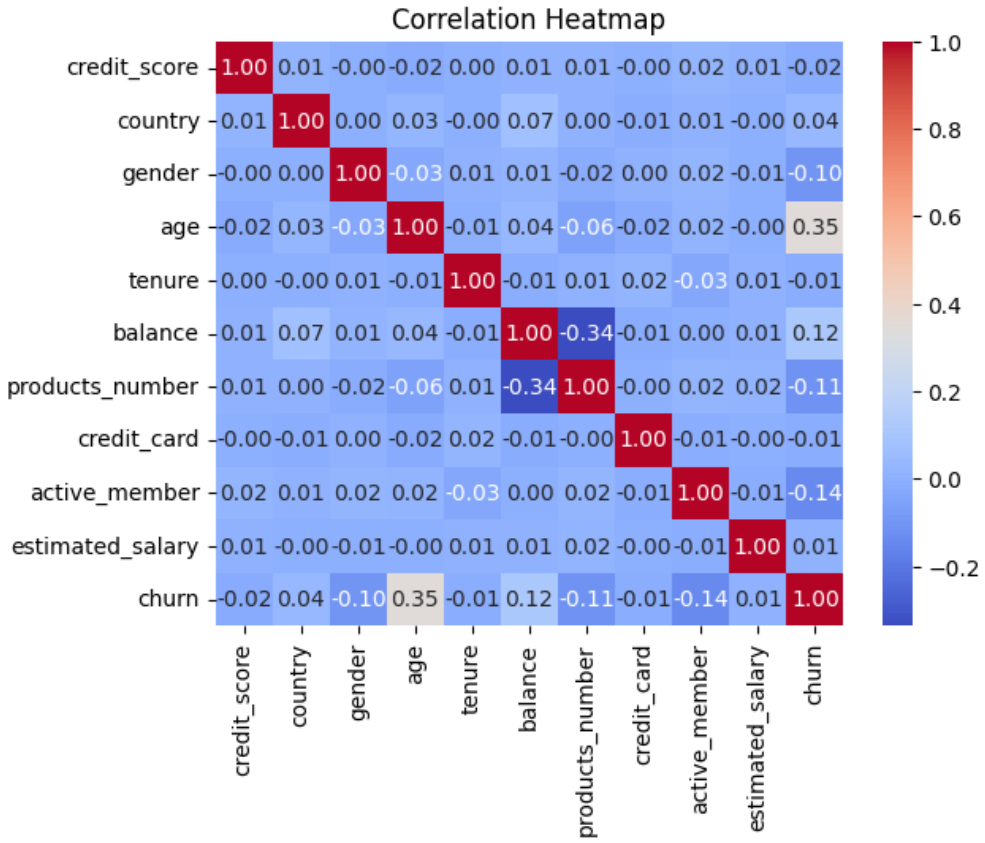
Observation: Clients that stick around after the first few months might regard the bank's services to be adequate, which would lower attrition.

Later Tenure 7–10 years

The churn rate first declines even further, reaching a low after six years, then rises once more at nine years before gradually dropping again.

Observation: Long-term clients may depart because of shifts in their financial requirements, better deals from rival banks, or a lack of interaction from the bank.

Interesting Patterns in the Dataset:



Churn and Active Membership:

Active_member and churn have a -0.14 connection. This suggests that, despite the weak relationship, active members are less likely to leave. Increasing consumer participation and engagement may be advantageous to banks.

Churn and Balance:

Customers with both big and small balances may go for non-financial reasons, as suggested by the nearly zero correlation between balance and turnover, which shows no obvious linear relationship.

Purchased Goods and Churn:

Customers that have more bank products are less likely to churn, according to the -0.11 connection between products_number and churn. This supports the notion that cross-selling enhances retention.

Churn and Country:

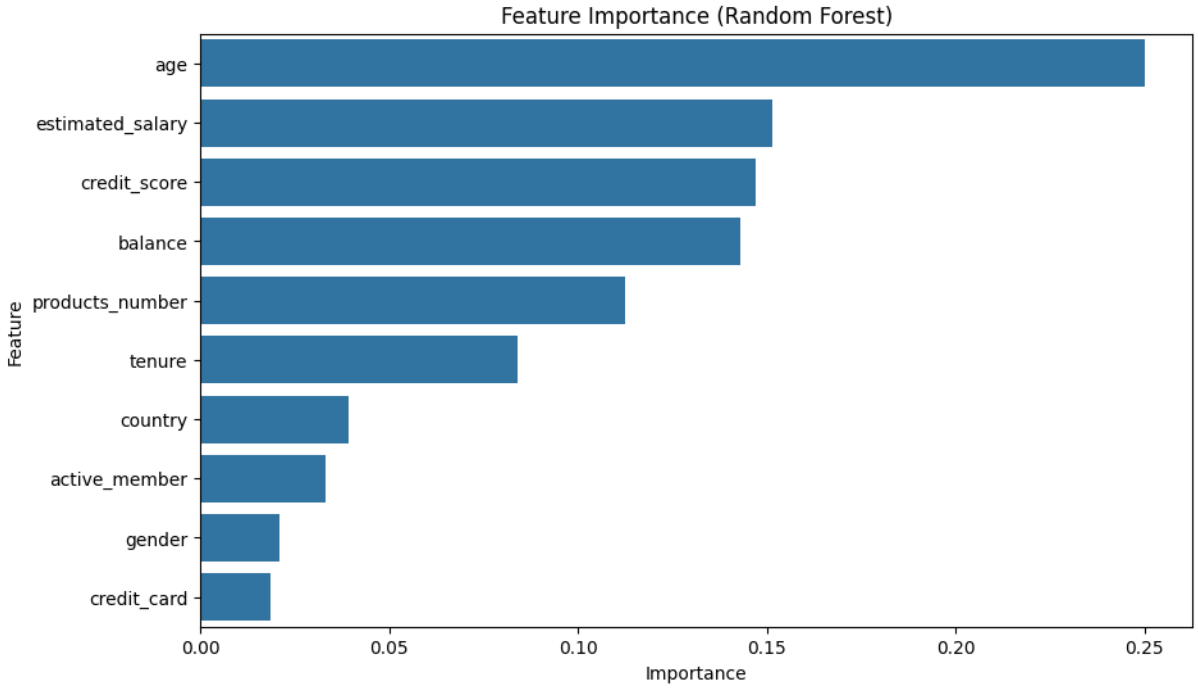
Although the previous churn rate research revealed significant variance by nation, the link between country and churn is minimal (0.10). This implies that non-linear interactions or outside influences may be the driving force behind regional churn patterns.

Key insights:

Age is the most important factor impacting turnover, as mentioned above and shown in the heatmap. Active membership and product count come in second and third, respectively. Features like credit card ownership, balance, and expected salary have little effect on churn, indicating that other behavioural or demographic factors might be more important.

4. Model building

Strongest Predictors of Customer Churn:

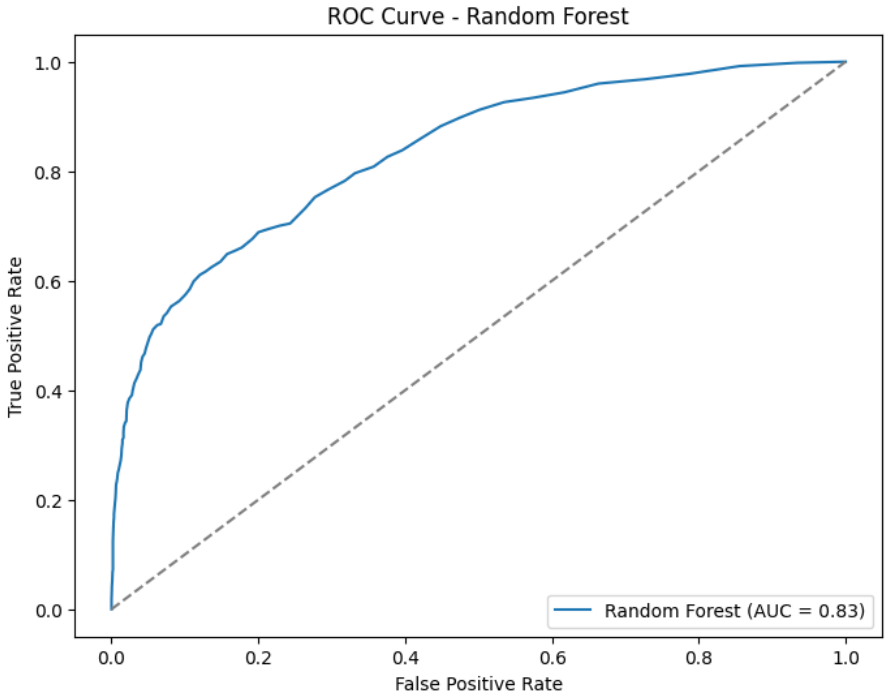


According to the graph, age has the greatest importance score (~0.23%) and is the most significant variable in predicting customer attrition. After 4 features are also important these are the Estimated Salary (~12.3%), the balance (~11.9%), the credit score (~11.8%) and the amount of product (~11.1%).

These are the strongest predictors of customer churn.

To identify them we used the Random Forest Feature relevance: For each split involving a feature, the Random Forest model analyses the drop in impurity (such as the Gini Index) to determine feature relevance. Features with higher significance scores are those that have a greater impact on lowering prediction error.

Evaluation Metrics Comparison Across Models



The Random Forest model's ability to predict customer attrition is shown by the ROC curve. The model's capacity to differentiate between churn and non-churn consumers is indicated by its Area Under the Curve (AUC) value of 0.83.

Metric	Logistic Regression	Decision Tree	Random Forest
Accuracy	83.12%	79.68%	**85.95%**
Precision (Churn)	66.17%	47.60%	**73.63%**
Recall (Churn)	26.55%	*47.51%*	**42.91%**
F1-Score (weighted)	80%	80%	**84%**

Due to its great accuracy, precision, and recall, the Random Forest model is the most effective. Because it strikes a balance between reducing false positives and identifying actual churn cases, it is advised for churn prediction in this dataset.

The Random Forest model's classification performance is emphasized by its ROC curve. With an AUC of **0.83**, the model shows a high degree of ability to differentiate between customers who are likely to churn and those who are not. The curve demonstrates a successful trade-off between the false positive rate and the actual positive rate (sensitivity), confirming the robustness of the model. The Random Forest model's status as the most dependable churn prediction model in this dataset is further confirmed by the AUC value, which supplements the assessment criteria (such as accuracy, precision, and recall).

5. Recommendations

Recommendations to Manage Customer Churn

Put in place targeted retention initiatives for older citizens.

Rationale: Older customers have greater churn rates, making age the best indicator of attrition. These clients may have unmet demands or feel cut off from contemporary financial services.
Action: Create specialized retention tactics to appeal to this group, such as loyalty plans, individualized services, or rewards.

Increase Product Engagement

Rationale: The strong relationship between products_number and turnover indicates that customers with fewer banking products are more likely to leave.

Action: By providing bundled services or savings for using multiple products, cross-selling campaigns can encourage users to adopt more banking products.??

Focus on Customers in High-Churn Regions

Rationale: Because of local rivalry or unfulfilled service expectations, customers from some countries (such as Country 1) have far higher turnover rates…

Action: To pinpoint particular causes (such as competition or service gaps), conduct a thorough geographical analysis. Then, employ customized marketing campaigns, improved services, or competitive pricing to address these issues.??

Recommendations Based on Profitability Analysis

Retention Value = €5:

Threshold: 0.6

Number of Customers to Target: 92

Total Expected Profit: €368

Explanation: At this threshold, the bank will target customers who have a churn probability greater than 60%. The expected profit is calculated as the number of high-risk customers (92) multiplied by the profit per customer (€4).

Threshold: 0.7

Number of Customers to Target: 69

Total Expected Profit: €276

Explanation: Raising the threshold to 70% reduces the number of customers targeted (69). However, these customers are more likely to churn, leading to a slightly lower total expected profit compared to the 0.6 threshold.

Threshold: 0.8

Number of Customers to Target: 46

Total Expected Profit: €184

Explanation: At a threshold of 80%, only the most at-risk customers (46) are targeted. This results in a further decrease in expected profit due to targeting fewer customers.

Retention Value = €10:

Threshold: 0.6

Number of Customers to Target: 92

Total Expected Profit: €828

Explanation: With a retention value of €10, the bank would target the same number of customers (92) at the 60% threshold, but the expected profit increases significantly. This is because the retention value has increased, so the expected profit per customer is now €9 instead of €4, resulting in a higher total profit.

Threshold: 0.7

Number of Customers to Target: 69

Total Expected Profit: €621

Explanation: The number of targeted customers drops to 69 at the 70% threshold, but due to the higher retention value (€10), the total expected profit is higher compared to the €5 retention scenario at the same threshold.

Threshold: 0.8

Number of Customers to Target: 46

Total Expected Profit: €414

Explanation: With the 80% threshold, fewer customers (46) are targeted, but the higher retention value results in a total expected profit of €414, which is still higher than the €5 scenario at this threshold.

Comparison Between €5 and €10 Retention Values:

Threshold	Retention Value = €5	Number of Customers to Target (5)	Total Expected Profit (5)	Retention Value = €10	Number of Customers to Target (10)	Total Expected Profit (10)
0.6	€4	92	€368	€	92	€828
0.7	€4	69	€276	€	69	€621
0.8	€4	46	€184	€	46	€414

Higher Retention Value: Increasing the retention value from €5 to €10 significantly boosts the total expected profit. At every threshold, the expected profit is higher when the retention value is €10.

Threshold Impact: As the threshold increases (from 0.6 to 0.8), the bank targets fewer customers, and the expected profit decreases. However, the quality of customers (those with a higher predicted churn probability) improves as the threshold increases, so the bank is focusing on the most at-risk customers.

Recommendations:

Low Threshold (0.6): At this threshold, the bank targets a larger group of customers. The total expected profit is moderate, but the bank may retain a wider range of at-risk customers. This strategy is more suitable for a €10 retention value, as it generates a significantly higher total expected profit compared to the €5 scenario.

Moderate Threshold (0.7): The bank will target fewer customers but with a higher likelihood of churning. The expected profit at this threshold is higher than at the 0.6 threshold, and still reasonable in both the €5 and €10 scenarios.

High Threshold (0.8): At this threshold, the bank targets the highest-risk customers, but the total expected profit is lower due to the smaller number of customers targeted. However, for the €10 retention value, this threshold still yields a decent expected profit, making it a good option for focusing on only the most likely churners.