EuroBank International

From Data Chaos to Insight: Predicting and Preventing Customer Atrition

Big Data Management and Analytics
Group Assignment
MSc BIM, 2025-2026
RSM, EUR

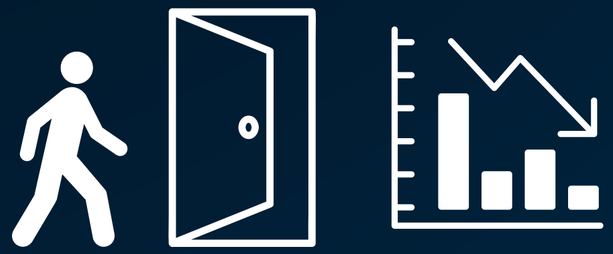
Group 25

Alisa Krajenbrink - 595886ak
Andreea Bula - 787876ab
Georgios Trikkas Britt - 774384gt
Natalia Poulakida - 752490np
Paola D'Incecco - 775206pd

The Challenge: Rising Customer Churn at EBI

Business impact of rising churn

- Higher acquisition costs
- Lower revenue
- Reduced customer loyalty



What EBI needs

- Clear insights on churn drivers
- Predictive model for flagging potential churners
- Actionable recommendations

Data Pre-Processing

1. No missing values or duplicates & unnecessary variable removed (“customer ID”)
2. Capped 16 outliers from “Credit Score” variable
3. Processed categorical variables: gender (binary) and country (one hot encoding)

Exploratory Analysis: Demographic Factors

Gender

Males exhibit a lower churn rate (**16.49%**) compared to females (**24.98%**)

Age

Churn rates for the 45–54 and 55–65 age groups are nearly at the **50%** mark.

Region

Germany shows the highest churn rate at **33.33%**, while France and Spain are both around **16%**.

Exploratory Analysis:
Tenure, Financial, & Engagement

Tenure

Churn rates remain consistent regardless of the time a customer was at the bank.

Financial

Higher account balance leads to higher churning behavior.
There is minimal impact with salary and credit score.

Engagement

Customer with 2 bank products are most loyal. Owning 3 or 4 products increases the churning risk.
Active customers have lower churning behavior.

Our predictive models

Logistic regression

Decision tree

Random forest

Clean data

Data pre-processing

80-20 split

Performance metrics (accuracy, sensitivity, precision, specificity)

Benchmark under same conditions

Train models

5-Fold cross validation

Optimization of models using ROC

Robustness

No overfitting

Evaluate

Performance metrics

Strongest predictors

Best model

Strongest predictors of churn

Model	Age	Balance	Number of products	Active member	Country Germany
Logistic regression	Younger customers more likely to churn	Customers with higher balance churn less	-	Inactive members churn more	Higher churn risk
Decision tree	Important split	Secondary split	Important split	Important split	Secondary split
Random forest	High importance	High importance	High importance	High importance	High importance

Conclusion: These variables appear statistically significant, visually (tree) and algorithmically (RF) as the most influential.

Comparing model performance

Model	Accuracy	Precision	Sensibility	Specificity	AUC	5-Fold Cross-Validation ROC
Logistic Regression	0.19	0.175	0.803	0.033	0.749	0.764
Decision Tree	0.855	0.786	0.393	0.972	0.75	0.76
Random Forest	0.857	0.774	0.445	0.9623	0.852	0.855

Conclusion: **Random Forest** offers the best overall balance between sensibility, precision and specificity. It has the highest accuracy, AUC and generalisation. Thus, it **is the strongest and most reliable model for predicting churn**.

Churn Reduction Recommandations

Retain High-Risk Age Segments(45-64)

Offer: loyalty programs, proactive check-ins, dedicated advisors

Boost Engagement of Inactive Customers

Actions: activation campaigns, personalized nudges, engagement rewards

Address Germany-Specific Issues

Localize offers, communication, and resolve market-specific pain points

Who to Target? Profit-Based Outreach Strategy

$$V \times P(churn) - €1 > 0$$

Scenario 1: Value = €5

Threshold: P(churn) > 20%

977 customers should be contacted

Scenario 2: Value = €10

Threshold: P(churn) > 10%

1000 customers (all) should be contacted