Bank Customer Churn Prediction System

Technical Documentation

Developer: Anushka Patil

Date: January 21, 2026

A professional-grade machine learning system designed to predict and prevent customer churn in banking.

Table of Contents

1. Executive Summary

This project presents a comprehensive machine learning solution for predicting customer churn in banking. The system combines advanced data science techniques with production-ready deployment, enabling financial institutions to proactively identify at-risk customers and implement targeted retention strategies. The final model achieves 70% accuracy with strong recall (59%), ensuring most potential churners are identified.

2. Problem Statement

Business Challenge: Customer churn is a critical issue in banking, costing institutions significant revenue. Predicting which customers are likely to leave allows banks to implement proactive retention strategies and optimize resource allocation.

Technical Objectives:
• Build predictive models with high recall (catch most churners)
• Identify key drivers of churn through explainability analysis
• Deploy an interactive system for real-time predictions
• Provide actionable insights for business teams

3. Dataset Description

Metric	Value
Total Records	10,000 customers
Features	19 (demographic, financial, behavioral)
Target Variable	Exited (0=Retained, 1=Churned)
Churn Rate	37.05%
Missing Values	None
Time Period	Synthetic customer data

4. Data Preprocessing Pipeline

Step 1: Data Cleaning
Removed non-predictive features (RowNumber, CustomerId, Surname). No missing values were found in the dataset.

Step 2: Feature Engineering
Created three new features to enhance model performance:
• **AgeGroup:** Categorized continuous age into discrete groups
• **HighBalance:** Binary flag for customers with balance above training median
• **BalanceProductInteraction:** Interaction between balance and number of products

Step 3: Categorical Encoding
• One-Hot Encoding: Geography (3 countries → 3 binary features)
• Label Encoding: Gender, AgeGroup

Step 4: Feature Scaling
Applied StandardScaler to normalize numerical features (mean=0, std=1) for algorithms sensitive to feature magnitude.

Step 5: Class Imbalance Handling
Applied SMOTE (Synthetic Minority Over-sampling Technique) to training data only, increasing minority class from 2,964 to 5,036 samples.

5. Model Development

Model 1: Logistic Regression
Linear model serving as the baseline. Fast to train and interpretable coefficients provide clear feature importance.

Model 2: Random Forest
Ensemble method capturing non-linear relationships through 100 decision trees. Handles interactions between features naturally.

Model 3: XGBoost
Gradient boosting framework optimizing for classification loss. Leverages sequential error correction for superior performance.

6. Model Evaluation Results

Metric	Log. Reg.	Random Forest	XGBoost
Accuracy	64.95%	69.75%	67.90%
Precision	52.44%	59.24%	56.96%
Recall	57.89%	58.84%	54.66%
F1-Score	55.04%	59.04%	55.79%
ROC-AUC	70.47%	73.27%	73.80%
Specificity	69.10%	76.17%	75.69%

Key Findings:

✓ Random Forest achieved best accuracy (69.75%) and strong recall (58.84%)
✓ XGBoost achieved best ROC-AUC (73.80%), indicating superior ranking ability
✓ All models show high specificity (69-76%), minimizing false positives
✓ Recall > 54% ensures majority of churners are identified

7. Feature Importance & Explainability

Top 5 Churn Drivers (Aggregate):
1. **IsActiveMember** (64%) - Active engagement is the strongest churn predictor
2. **Gender** (35%) - Gender-based patterns exist in churn behavior
3. **Geography_France** (30%) - France shows distinct churn risk
4. **Geography_Germany** (26%) - Germany demonstrates higher churn tendency
5. **HasCrCard** (21%) - Credit card ownership correlates with retention

SHAP Analysis:
Computed SHAP (SHapley Additive exPlanations) values for model-agnostic feature attribution. This enables individual prediction explanations showing exactly which features pushed a specific prediction toward churn or retention.

8. Deployment Architecture

Web Application: Streamlit-based interactive platform enabling business teams to:
• Make single predictions for new customers
• Batch predict on customer CSV files
• View model evaluation visualizations
• Explore feature importance and SHAP explanations
• Access comprehensive project documentation

Model Serialization: All models and the preprocessing pipeline are pickled for reproducibility and production deployment.

9. Business Impact & Recommendations

Recommended Actions:
1. Deploy Random Forest model (70% accuracy) as primary production model
2. Use ROC-AUC metric to rank customers by churn probability for prioritized outreach
3. Focus retention programs on inactive members (strongest churn driver)
4. Implement geographic-specific strategies for Germany and France
5. Monitor model performance monthly and retrain quarterly with new data
6. A/B test retention interventions on high-probability churn customers

Expected ROI:
Assuming 10,000 customers and 37% churn rate: 3,700 potential churners. Catching 59% with our model = 2,183 identifiable churners. If retention efforts save 20% = 437 retained customers. At $5,000 customer lifetime value savings = **$2,185,000 potential impact.**