PREDICTIVE MODELING AND RISK SCORING FOR BANK CUSTOMER CHURN

Author: S A HAR GANESH

DATA SCIENCE INTERNSHIP (3 MONTHS)

EMAIL – hariganeshsa@gmail.com

1. Introduction and Business Context

Customer churn is a critical challenge in the banking industry. Acquiring a new customer is significantly more expensive than retaining an existing one, and frequent customer exits directly affect revenue stability, customer lifetime value, and long-term competitiveness. Despite having access to rich customer-level data, many banks struggle to identify which customers are likely to leave, how risky each customer is, and why customers decide to exit.

This project, **Predictive Modeling and Risk Scoring for Bank Customer Churn**, addresses these challenges by building a data-driven churn intelligence system. The goal is not only to predict whether a customer will churn but also to assign a churn risk score and provide explainable insights that can support proactive retention strategies. The entire analysis has been designed to align with real-world banking use cases and to be interpretable by both technical and non-technical stakeholders.

---

2. Problem Statement

Despite having comprehensive customer data, banks often face three major gaps:

1. **Lack of accurate churn prediction models** – Traditional approaches identify churn only after it has occurred, making retention reactive.
2. **Absence of quantitative churn risk scores** – Without risk scoring, all customers are treated equally, leading to inefficient and costly retention campaigns.
3. **Limited explainability of churn drivers** – Business teams often do not understand *why* customers leave, which prevents targeted and meaningful interventions.

This project aims to solve these gaps by developing an early-warning churn prediction system that delivers accurate predictions, probability-based risk scores, and clear behavioral insights.

---

3. Dataset Overview and Domain Understanding

The dataset used in this project represents customer-level banking information, including demographic, financial, and behavioral attributes. Each row corresponds to a single customer, and the target variable indicates whether the customer has exited the bank.

Key categories of features include:

- **Demographic attributes**: Age, Gender, Geography
- **Financial indicators**: Credit Score, Account Balance, Estimated Salary
- **Relationship and engagement indicators**: Tenure, Number of Products, Credit Card ownership, Active membership status
- **Target variable**: Exited (1 = Customer churned, 0 = Customer retained)

From a business perspective, engagement and relationship strength are expected to play a more important role in churn behavior than static demographics alone. This assumption guided the subsequent exploratory analysis and feature engineering steps.

---

4. Data Preprocessing and Cleaning

The initial phase of the analysis focused on ensuring data quality and relevance.

4.1 Data Inspection

The dataset was examined for:

- Data types and structure
- Missing values
- Summary statistics and distributions

The inspection confirmed that the dataset was well-structured, with no critical missing values requiring imputation.

4.2 Removal of Non-Informative Features

Certain columns such as customer identifiers and surnames were removed. These fields uniquely identify customers but do not contribute to churn behavior and can introduce noise into predictive models.

4.3 Encoding and Scaling

- Categorical variables (such as Geography and Gender) were encoded using one-hot encoding.
- Numerical variables were scaled where required to ensure fair contribution to model training, particularly for distance- and gradient-based models.

This preprocessing ensured that the dataset was suitable for machine learning while preserving interpretability.

---

5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand customer behavior patterns and identify early signals of churn.

5.1 Univariate Analysis

- The churn variable showed class imbalance, with retained customers forming the majority. This highlighted the importance of recall and ROC-based evaluation metrics.
- Age distribution analysis indicated that churn rates were higher among older customers.
- Balance analysis revealed that high account balances alone do not guarantee customer loyalty.

5.2 Bivariate Analysis

- **Engagement vs Churn**: Inactive customers exhibited significantly higher churn rates.
- **Product Holding vs Churn**: Customers with fewer bank products were more likely to exit.
- **Tenure vs Churn**: Shorter relationships with the bank correlated with higher churn risk.

These findings reinforced the hypothesis that churn is driven primarily by engagement and relationship strength rather than purely demographic factors.

---

6. Feature Engineering

To enhance the predictive power of the models and better capture customer behavior, several derived features were created:

- **Balance-to-Salary Ratio**: Measures financial engagement relative to earning capacity.
- **Product Density**: Captures the intensity of product usage over the customer's tenure.
- **Engagement–Product Interaction**: Combines activity status with product ownership.
- **Age–Tenure Interaction**: Reflects lifecycle and relationship maturity effects.

These engineered features translated raw data into business-meaningful indicators and played a crucial role in improving model performance.

---

7. Train–Test Strategy

The dataset was split into training and testing sets using a **stratified split**, ensuring that the churn proportion remained consistent across both sets. This approach reflects real-world deployment conditions, where models must generalize to unseen customers.

The test set was reserved exclusively for evaluation to avoid data leakage and to provide an unbiased estimate of model performance.

8. Model Development

Multiple models were developed and compared to balance predictive performance, stability, and interpretability.

8.1 Logistic Regression (Baseline Model)

Logistic Regression served as an interpretable benchmark. While easy to explain, it showed limited ability to capture complex, non-linear churn patterns.

8.2 Decision Tree

The Decision Tree model improved performance over the baseline by capturing non-linear relationships but showed signs of instability and overfitting.

8.3 Random Forest

Random Forest significantly improved predictive accuracy and robustness by aggregating multiple decision trees. It handled interactions well and reduced overfitting.

8.4 Gradient Boosting

Gradient Boosting further enhanced performance by sequentially correcting errors from previous models. It delivered strong discrimination between churned and retained customers.

8.5 XGBoost

XGBoost, an optimized boosting algorithm, achieved performance comparable to Gradient Boosting, particularly in identifying high-risk churners.

9. Model Evaluation and ROC Analysis

Model performance was evaluated using multiple metrics, including Accuracy, Precision, Recall, F1-score, and ROC-AUC. Given the business objective, ROC curves were especially important.

For each model, ROC curves were plotted using a One-vs-Rest (OvR) approach for the binary target:

- **Class 0**: Retained customers
- **Class 1**: Churned customers

Key Findings from ROC Curves:

- Logistic Regression curves were closer to the diagonal, indicating weaker discrimination.
- Decision Trees showed moderate separation but lacked stability.
- Random Forest, Gradient Boosting, and XGBoost produced curves close to the top-left corner, indicating strong predictive power.

Gradient Boosting and XGBoost emerged as the top-performing models, with Gradient Boosting selected as the final model due to its balance of performance, stability, and interpretability.

---

10. Churn Risk Scoring and Validation

Beyond binary predictions, the selected model generated churn probabilities for each customer in the test set. These probabilities were converted into churn risk scores (0–100%) and grouped into Low, Medium, and High-risk categories.

Validation showed that:

- Low-risk customers had very low actual churn rates.
- Medium-risk customers showed moderate churn behavior.
- High-risk customers exhibited significantly higher churn rates.

This confirmed that the risk score meaningfully prioritizes customers based on their likelihood of leaving.

---

11. Key Insights: Why Customers Are Leaving

The analysis clearly indicates that customer churn is driven by behavioral and relationship factors rather than demographics alone. The most influential drivers include:

- Low engagement and inactive membership
- Limited product ownership
- Weak or short-term relationships with the bank
- Misalignment between customer lifecycle stage and engagement

Demographic attributes such as gender and geography played a comparatively minor role.

---

12. Business Recommendations

Based on the findings, the following actions are recommended:

1. Deploy Gradient Boosting as the primary churn prediction model.
2. Use churn risk scores to prioritize retention efforts.
3. Focus on improving customer engagement rather than offering generic discounts.
4. Strengthen early-stage customer relationships through onboarding and cross-sell strategies.
5. Integrate the model into a decision-support dashboard for continuous monitoring.
6. Improve customer engagement through personalized offers, rewards, and regular communication.
7. Identify inactive customers early using churn prediction and trigger retention campaigns.
8. Increase product adoption through cross-selling (credit cards, loans, insurance bundles).
9. Strengthen onboarding experience for new customers during the first 3–6 months.
10. Implement segment-based strategies for older customers with low engagement.
11. Monitor churn rate through a real-time dashboard and take proactive action.

---

13. Overall Project Summary

This project successfully transformed raw customer data into a practical churn intelligence system. By combining robust data preprocessing, insightful exploratory analysis, advanced machine learning models, and probability-based risk scoring, the project addresses the core challenges faced by banks in churn management.

The final solution enables early identification of at-risk customers, explains the underlying drivers of churn, and supports proactive, targeted, and cost-effective retention strategies. Overall, the project demonstrates how data science can directly support strategic business decision-making in the banking domain.

Streamlit Dashboard Link: https://europeanbankanalysis.streamlit.app/

GitHub Link: https://github.com/HariGanesh2003/Predictive-Modeling-and-Risk-Scoring-for-Bank-Customer-Churn/tree/main