Customer Churn Analysis and Prediction in Banking Sector

Software Requirements Specification (SRS)

Volkan Çamlı

Table of Contents

1. INTRODUCTION

1.1. Purpose and Significance of the Research

Customer churn represents a major operational challenge for banking institutions. The departure of a customer not only reduces revenue but also increases acquisition costs for replacements, impacting long-term profitability. In a highly competitive and digitized banking environment, customer retention has become a key performance metric. The purpose of this project is to investigate the drivers of customer churn and to develop predictive machine learning models that can identify customers who are at risk of leaving. The research emphasizes not just prediction accuracy, but also interpretability and actionable insight, enabling banks to proactively implement retention strategies.

This research is significant in helping banks move from reactive churn management to a predictive and strategic approach, leveraging data analytics. Additionally, the use of interactive visualization tools like Power BI bridges the gap between complex machine learning outputs and business user accessibility.

1.2. Scope of the Research Project

The project involves several integrated components: preprocessing and analyzing customer data, applying feature engineering to enhance predictive signals, training various machine learning models, and visualizing results through an interactive dashboard. The dataset used is synthetic and publicly available, simulating a real banking environment with customer demographics, account activity, and churn labels. The models developed include logistic regression, decision trees, random forests, gradient boosting machines, and neural networks. These models are compared using standard evaluation metrics to determine the most effective approach.

The deliverables include:

- A fully functional data pipeline for churn analysis.
- A suite of trained and validated predictive models.
- A Power BI dashboard for visual insights.
- A written report detailing methodology, findings, and practical recommendations.

1.3. Definitions, Acronyms, and Abbreviations

Customer Churn (Attrition): The phenomenon where a customer stops doing business or ends their relationship with a bank**.**

Churn Prediction: The process of identifying customers likely to stop using a service or leave a business.

Feature Engineering: The process of transforming raw data into features that better represent the underlying problem to predictive models.

Clustering: An unsupervised machine learning technique used to group similar data points.

Correlation Matrix: A table showing correlation coefficients between variables to identify relationships.

Model Evaluation Metrics: Techniques used to assess the performance of machine learning models.

Train/Test Split: A technique for evaluating the performance of machine learning models by training on one set and testing on another.

KYC: Know Your Customer

GBM: Gradient Boosting Machines

ML: Machine Learning

AI: Artificial Intelligence

AUC-ROC: Area Under the Receiver Operating Characteristic Curve

DAX: Data Analysis Expressions (used in Power BI)

KPI: Key Performance Indicator

CRM: Customer Relationship Management

ReLU: Rectified Linear Unit (activation function in neural networks)

Adam: Adaptive Moment Estimation (optimization algorithm)

1.4. References

- Rangala Mahesh. Bank Customer Churn Dataset. Kaggle
- IEEE 830-1998. IEEE Recommended Practice for Software Requirements Specifications.
- KERAMATI, A., GHANEEI, H., & MIRMOHAMMADI, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. Financial Innovation, 2(1), 1-13.

- BRÂNDUŞOIU, I., TODEREAN, G., & BELEIU, H. (2016). Methods for churn prediction in the pre-paid mobile telecommunications industry. In 2016 International conference on communications (COMM) (pp. 97-100). IEEE.
- Guliyev H., Yerdelen Tatoğlu F. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning model. Journal of Applied Microeconometrics, cilt.1, sa.2, ss.85-99, 2021 (Hakemli Dergi)
- Preet Singh P., Islam Anik F., Senapati R., Sinha A., Sakib N., Hossain E. (2024). Data Science and Management, volume 7, issue 1, pages 7-16.

1.5. Overview of the Document Structure

This SRS is divided into four major sections. The introduction presents the background, scope, and significance of the research. The second section describes the research context, methodology, and deliverables. The third section outlines the system and functional requirements, while the final section includes diagrams, terminology, and appendices that support system implementation and understanding.

2. OVERALL PROJECT DESCRIPTION

2.1. Research Background and Context

Customer attrition poses a substantial risk to banks by reducing customer lifetime value and weakening the institution's reputation. With the growth of digital services, switching costs for customers have declined, heightening the need for banks to predict and understand churn behavior. Modern data science methods offer the ability to identify patterns and develop models capable of predicting customer churn with high accuracy. Through this project, data-driven decision-making in customer relationship management is not only possible but scalable and efficient.

2.2. Problem Statement and Research Questions

Despite the availability of customer data, many banks lack the infrastructure or expertise to make meaningful use of it for churn prediction. This project addresses that gap through the following research questions:

- What customer characteristics are most indicative of churn?
- Which machine learning models are most effective for predicting churn in banking?
- How can the insights from churn prediction models be visualized to inform management decisions?

By answering these questions, the project aims to contribute to both academic understanding and practical application in the field of banking analytics.

2.3. High-Level Research Methodology

The project follows a structured methodology:

Data Collection: Use of a public dataset simulating bank customer data.

Preprocessing: Handling missing values, encoding categorical variables, and normalizing numerical features.

Feature Engineering: Creating new features such as credit risk levels, age groups, income brackets, and clusters using K-means.

Model Development: Training multiple models including logistic regression, decision trees, random forests, GBMs (e.g., XGBoost), and feedforward neural networks.

Model Evaluation: Comparing models using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.

Visualization: Designing a Power BI dashboard that provides descriptive and predictive insights.

2.4. Project Constraints and Assumptions

The dataset is synthetic and may not fully replicate real-world complexity.

The system is developed and tested in Python and Power BI, with limited external deployment considerations.

Interpretability is prioritized alongside predictive power to ensure business usability.

Real-time prediction and deployment into production systems are out of scope for this academic study.

2.5. Expected Contributions or Deliverables

- A research report detailing methodology, results, and implications.
- Python-based Jupyter Notebook codebase for preprocessing, modeling, and evaluation.
- Power BI dashboard enabling interactive exploration of churn predictions.
- Recommendations for customer retention strategies based on empirical findings.

3. FUNCTIONAL AND SYSTEM REQUIREMENTS

3.1. Description of Tools and Modules

The system includes the following components:

Data Preprocessing Module: Loads, cleans, and transforms input data.

Feature Engineering Module: Generates domain-specific features.

Machine Learning Module: Trains and compares multiple models.

Evaluation Module: Assesses models using standard classification metrics.

Visualization Module: Generates exportable outputs and feeds into Power BI.

All components are developed in Python, leveraging libraries such as pandas, numpy, scikit-learn, matplotlib, seaborn, xgboost, and keras.

3.2. Functional Requirements

The system must be able to load and process large CSV datasets (>100,000 rows).

The preprocessing pipeline must support missing value handling, one-hot encoding, and scaling.

The system must support the training of multiple classifiers with hyperparameter tuning.

Evaluation results must be stored and visualized (confusion matrix, AUC curve, etc.).

Outputs must be compatible with Power BI (CSV or Excel formats).

3.3. System Architecture and Workflows

The workflow begins with data ingestion and is followed by feature transformation. The processed data is then used to train models in parallel, the results of which are saved for evaluation and dashboarding.

Workflow Steps:

Import raw data →

Clean and transform features →

Train models →

Evaluate performance →

Export predictions →

Load into Power BI dashboard

3.4. Interface Requirements

User Interface: The Power BI dashboard provides a visual layer for exploring model outputs, filtering by demographics, geography, and account activity.

Data Interface: The system reads CSV inputs and produces cleaned datasets, prediction outputs, and evaluation summaries for export.

API Considerations: While out of scope for this phase, the modular codebase is designed for future API integration.

3.5. Non-functional Requirements

Performance: All models should train in under 5 minutes on datasets of ≤165,000 rows.

Scalability: Code should support expansion to datasets of 1 million rows with minor adjustments.

Reproducibility: Experiments and outputs should be reproducible using saved seeds and version-controlled code.

Usability: All features and findings must be accessible to a business audience via the Power BI interface.

Security & Privacy: The dataset is synthetic and contains no personally identifiable information.

4. FUNCTIONAL AND SYSTEM REQUIREMENTS

4.1. Diagrams

System Workflow Diagram: Visual depiction of the sequential steps from data ingestion to dashboard deployment.

Model Architecture: Illustrations for decision trees and neural networks used in the system.

Dashboard Design Layout: Sketches or screenshots of Power BI components including filters, maps, KPI indicators, and prediction tables.

4.2. Glossary of Terms

Attrition Rate: Percentage of customers who leave within a specified period.

Precision: True positives divided by the sum of true and false positives.

Recall: True positives divided by the sum of true positives and false negatives.

Feature Importance: Numerical ranking of how influential a feature is in a predictive model.