Customer Churn Analysis and Prediction in Banking Sector

Software Design Document (SDD)

Volkan Çamlı

Table of Contents

1. INTRODUCTION

1.1. Purpose

This Software Design Document (SDD) provides a detailed description of the architecture and design of the “Customer Churn Analysis in the Banking Sector” project. It follows the IEEE 1016-2009 standard and aims to document all major components, interactions, and data structures used in the software system to ensure maintainability, scalability, and extensibility.

1.2. Scope

The project aims to predict and analyze customer churn using machine learning algorithms and provide actionable insights via a Power BI dashboard. The system includes a backend developed in Python using Jupyter Notebook, data preprocessing, model training and evaluation components, and a visualization layer using Power BI.

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

The rest of this document outlines the system's architectural design, component breakdown, data design, interface design, and dynamic behavior. Each component is discussed in terms of its functionality, interactions, and constraints.

2. SYSTEM OVERVIEW

2.1. System Context

The system is designed to analyze customer churn in the banking sector by integrating machine learning techniques, data preprocessing pipelines, and interactive dashboards. It operates as a data-driven decision support system for banking professionals who aim to understand, predict, and reduce customer attrition.

At a high level, the system is composed of the following layers:

1. Data Layer: Responsible for collecting, cleaning, and transforming data.

2. Modeling Layer: Hosts machine learning models for training and prediction.

3. Visualization Layer: Presents insights using Power BI dashboards.

4. Interaction Layer: Allows users (e.g., business analysts, data scientists, bank managers) to interact with the dashboard and derive actionable decisions.

2.2. System Architecture

The architecture follows a modular pipeline-based design that promotes reusability and maintainability. It consists of the following major modules:

a. Data Ingestion Module

- Functionality: Imports raw data from external sources (Kaggle dataset in .csv format).

- Tools: pandas, numpy

- Inputs: train.csv, test.csv

- Outputs: Structured DataFrames ready for preprocessing

b. Data Preprocessing & Feature Engineering Module

- Functionality: Cleans data, handles missing values, encodes categorical variables, and adds engineered features (e.g., risk classification, age groups, income levels).

- Outputs: Cleaned and enriched datasets
- Techniques:
    - Label encoding & one-hot encoding
    - Feature binning (age, income, credit score)
    - Clustering-based segmentation (K-means)

c. Model Training & Evaluation Module

- Functionality: Trains and evaluates several machine learning models.
- Models Implemented:
    - Logistic Regression
    - Decision Trees
    - Random Forest
    - LightGBM
    - Neural Networks
- Evaluation Metrics:
    - Accuracy
    - Precision / Recall
    - F1-Score
    - AUC-ROC
    - Confusion Matrix

d. Model Comparison & Selection Module

- Functionality: Compares models based on evaluation metrics and selects the best performing one.
- Technique: Cross-validation, hyperparameter tuning with grid/random search

e. Visualization Module (Power BI)

- Functionality: Generates an interactive dashboard for data exploration and decision-making.
- Components:
    - Churn distribution pie chart
    - Bar charts by gender, geography, age
    - Heatmaps for correlations
    - Customer segmentation maps
    - Churn prediction table with risk scores
- Platform: Microsoft Power BI Desktop

2.3. System Deployment

The system is executed locally in the development environment. The machine learning pipeline runs in Jupyter Notebook, and the resulting .csv files are imported into Power BI for dashboard creation. Deployment to production environments (e.g., via APIs or web services) is not in scope for this academic project but may be considered for future work.

2.4. System Interaction Diagram

Below is a textual representation of how system components interact:

[Raw Dataset]
→ (Data Ingestion)
→ (Preprocessing & Feature Engineering)
→ (Model Training & Evaluation)
→ (Model Selection)
→ [Predictions Output]
→ [Power BI Dashboard]

2.5. Assumptions and Dependencies

The dataset used is static and reliable (Bank Customer Churn Dataset from Kaggle).

All computations are done locally using Python (Jupyter Notebook).

Power BI is used only for visualization, not for model deployment or real-time inference.

No external APIs or cloud infrastructure is used.

Ethical usage of customer data is assumed, and no sensitive personal information is present in the dataset.

3. ARCHITECTURAL DESIGN

3.1. Design Approach

The architectural design follows a modular, layered, and pipeline-oriented structure. Each component is designed to be reusable and loosely coupled, allowing independent development and testing. The architecture is implemented entirely in Python (Jupyter Notebook) and integrated with Power BI for the visualization layer. The system adheres to the separation of concerns principle, dividing functionality across preprocessing, modeling, evaluation, and visualization stages.

3.2. Main Components

Data Preprocessing Module

- Responsible for data cleaning, encoding, and feature engineering. Includes classification of customers by credit score, age group, and income level. Irrelevant variables like CustomerID are dropped, and categorical features are converted into numerical formats.

Modeling Module

- Several machine learning algorithms (Logistic Regression, Decision Trees, Random Forest, LightGBM, Neural Networks) are trained and evaluated. Cross-validation and hyperparameter tuning are applied to improve performance.

Evaluation Module

- Models are compared using metrics such as Accuracy, F1-Score, AUC, and Confusion Matrix. The best-performing model is selected for final predictions.

Visualization Module (Power BI)

- A Power BI dashboard visualizes churn insights using charts, heatmaps, and segmentation filters. It helps stakeholders monitor churn trends and identify at-risk customer segments.

3.3. Interaction Flow

[Raw Dataset (.csv)]

↓

[Preprocessing & Feature Engineering]

↓

[Model Training & Evaluation]

↓

[Prediction Output (.csv)]

↓

[Power BI Dashboard]

4. DATA DESIGN

4.1. Dataset Overview

The system uses a publicly available dataset titled “Bank Customer Churn Dataset” sourced from Kaggle. The dataset includes both demographic and financial attributes for each customer, along with the target variable Exited, which indicates whether the customer has churned (1) or not (0).

Training Set: 165,000 records

Test Set: 110,000 records

Target Variable: Exited (binary classification)

4.2. Key Data Attributes

Feature Name	Description
CustomerId	Unique customer identifier (dropped during modeling)
Surname	Customer’s last name (not used in modeling)
CreditScore	Numeric credit score indicating financial reliability
Geography	Country (France, Germany, Spain)
Gender	Male / Female
Age	Customer's age
Tenure	Years the customer has been with the bank
Balance	Account balance
NumOfProducts	Number of products owned (e.g., credit card, mortgage)
HasCrCard	Boolean: whether the customer has a credit card
IsActiveMember	Boolean: activity level with the bank
EstimatedSalary	Approximate salary of the customer
Exited	Target variable (1 = churned, 0 = retained)

4.3. Feature Engineering

To enhance model performance and enable deeper insights, the following new features were derived:

- Credit Score Group: Categorized into “High Risk”, “Moderate Risk”, “Good”, “Excellent”
- Age Group: “Young”, “Middle-Aged”, “Senior”
- Income Group: “Low Income”, “Middle Income”, “Wealthy”, “Very Wealthy”
- Customer Segments: Created via K-Means clustering for behavior-based segmentation

All categorical features were encoded (using label or one-hot encoding), and all irrelevant columns (CustomerId, Surname) were dropped.

4.4. Data Integrity

No missing values were present in the dataset.

All features were numerically or categorically clean.

Class imbalance was noted: ~21% churn rate vs. ~79% retention.

5. CONCLUSION

This Software Design Document (SDD) presents a structured and modular design framework for the project “Customer Churn Analysis in the Banking Sector”, developed in alignment with the IEEE 1016-2009 standard.

The system is composed of a well-defined sequence of data analysis stages: data preprocessing, feature engineering, model training and evaluation, and result visualization. By using Python-based tools and machine learning libraries, the system ensures flexibility, scalability, and reproducibility in experimentation. The integration with Power BI provides a professional and interactive interface for data-driven decision-making.

The architectural choices—such as using a pipeline design, model comparability, and enriched feature creation—support both performance and interpretability. The decision to rely on multiple classification algorithms allows the system to adapt to data-specific challenges like class imbalance and nonlinear relationships.

This design does not include deployment to a real-time system, focusing instead on an offline, notebook-based predictive pipeline. However, the structure is extendable to future web-based or API-driven systems should the project evolve beyond academic use.

In conclusion, the design presented in this document offers a reliable foundation for churn prediction in the banking domain. It combines strong data engineering practices, proven machine learning techniques, and clear visualization strategies, resulting in a system that is both effective and accessible for business users and technical stakeholders alike.