IBM SPSS Modeler Project Report

Predictive Analytics in Bank Customer Churn

Name/Roll no. :

1) Piyush Kumar Sharma(AJU/231666) , 31

Class/Sec: B.Tech CSE (IBM) ‘E’

Index

Project Brief

- **Project Title:** Bank Customer Churn Using IBM SPSS Modeler

  This project aims to develop a predictive model to identify customers at high risk of churning and to propose actionable strategies and interventions to retain them, thereby reducing customer attrition and increasing the bank's lifetime customer value.

- **Project Goal**
  The primary goal is to **minimize customer churn** and improve customer retention rates by leveraging data-driven insights.
  Quantitative Goal: Achieve a **10% reduction** in the annual customer churn rate within one year of model deployment and strategy implementation.
  Qualitative Goal: Improve **customer satisfaction** and optimize marketing/retention campaign resource allocation.

---

2. Problem Statement
High customer churn erodes the bank's revenue, increases acquisition costs, and negatively impacts market share. The bank currently lacks a systematic, proactive method to reliably identify at-risk customers *before* they leave. Retention efforts are often generic or reactive, leading to inefficient resource use.

- **Scope**
  In-Scope Activities:
  Data Collection & Exploration (EDA): Gathering, cleaning, and analyzing historical customer data (e.g., demographics, account balance, transaction history, product usage, service interaction logs).
  Model Development: Building and optimizing machine learning models (e.g., Logistic Regression, Random Forest, Gradient Boosting) to predict the **probability of churn** for each customer.

- **Tools Used:**
  - IBM SPSS Modeler, focusing on stream construction and model palette for both regression and classification tasks

Introduction

In the highly competitive and increasingly digital financial services industry, customer churn—the rate at which customers cease their relationship with the bank—represents a critical threat to profitability and sustainable growth. Industry research consistently shows that the cost of acquiring a new customer is significantly higher (often 5 to 25 times more) than the cost of retaining an existing one. Every customer who closes an account represents not only a direct loss of revenue (fees, interest, product usage) but also a loss of future customer lifetime value and potential negative word-of-mouth.

For a bank to maintain its market position and maximize shareholder value, a fundamental shift from reactive customer service to proactive customer retention is essential.

Currently, customer retention efforts are often general, delayed, or poorly targeted, leading to inefficient resource allocation. The bank lacks a precise, forward-looking mechanism to identify the customers who are genuinely *at risk* of churning and, more importantly, *why* they are at risk.

This project will employ a **Data Science and Machine Learning** approach. We will leverage historical customer data to train classification models (such as Gradient Boosting, Random Forest, or Logistic Regression).

Feasibility Study

Data Source Flexibility (Ingestion)

This is the ability of the data pipeline to pull in and integrate information from diverse and potentially new systems.

- **Requirement:** Churn is rarely predicted by a single data source. It requires combining **hard data** (transactional records) and **soft data** (customer feedback, call center logs).
- **Application:** A flexible system can seamlessly integrate:
  - **Core Banking Data:** Account balances, tenure, number of products.
  - **Behavioral Data:** Website/app login frequency, feature usage.
  - **Interaction Data:** Call center wait times, Net Promoter Scores (NPS), social media sentiment.

Feature Flexibility (Engineering)

This is the ability to easily generate new predictive variables (**features**) from existing raw data. The most effective churn drivers are often not raw data points but metrics calculated from them.

- **Requirement:** A single metric can quickly lose predictive power as market dynamics change (e.g., the average balance of one year ago might be less relevant today).
- **Application:** Flexible feature engineering allows for the creation of *dynamic* features, such as:
  - **Recency, Frequency, Monetary (RFM) metrics** adapted for banking (e.g., *Recency* of last transaction, *Frequency* of branch visits).
  - **Trend Variables:** Calculating the **percentage change in average monthly balance** over the last three months (a strong indicator of potential churn).

Model Flexibility (Robustness)

This refers to the model's ability to maintain its predictive performance even when the underlying data distributions shift (a real-world concern in banking).

Project Details

The core objective of this project is to apply advanced analytics and machine learning to proactively identify customers at risk of leaving the bank, allowing for targeted and timely retention efforts.

Data Overview: Defining the Customer and Churn

The quality and breadth of the data are the foundation of a successful predictive model. This section details the key data categories and the primary target variable.

1. Target Variable Definition (The Churn Label)

The dependent variable for the prediction model is **Exited** or **Churn**.

Attribute	**Definition**	**Purpose**
Exited (Binary)	A customer is labeled **'1' (Churn)** if they have closed all their accounts with the bank within the observation period (e.g., the last 12 months). A customer is labeled **'0' (Retained)** if they remain an active customer.	To serve as the historical outcome for training the classification model.
Observation Period	12 to 24 months of historical customer activity.	To capture sufficient historical behavior leading up to the churn event.

Project Scope: Boundaries and Focus

The scope defines what the project will deliver and, crucially, what it will *not* cover, ensuring focused efforts and manageable expectations.

1. In-Scope Activities (The Deliverables)

- **Predictive Model Development:** Building and validating a Machine Learning model (e.g., using algorithms like XGBoost, Random Forest, or others) to predict the probability of a customer churning.
- **Churn Driver Analysis:** Identifying the top 5-10 features (e.g., "low balance in combination with high credit score") that most contribute to churn probability, ensuring the results are explainable.

- **Targeting Recommendations:** Providing the Marketing/CRM team with a ranked list of high-risk customers and recommended churn probability thresholds for targeted intervention campaigns.

- Key Constraints

Constraint	Description
Timeframe	The core model build and analysis phase is limited to **4 months**.
Data Privacy	All analysis must comply with bank data governance policies and relevant regulations (e.g., GDPR, CCPA). Customer PII (Personally Identifiable Information) must be anonymized or aggregated.
Model Type	The final model must be **interpretable** enough to explain the risk factors to business stakeholders, favoring models like Logistic Regression or Tree-based ensembles over deep "black-box" Neural Networks.

Conclusion and Summary

The project successfully developed and validated bank customer churn using IBM SPSS Modeler.

- **Shift to Predictive Targeting**

The bank should immediately utilize the churn risk scores to transition from a broad-based marketing approach to a **highly targeted retention strategy**. Customers in the **top 10-20%** risk segment should be prioritized for intervention, saving budget on customers who are unlikely to leave (False Positives) and focusing resources on those most likely to leave (True Churners).

- **Segmentation**

Retention offers must be tailored based on the customer's **Value** and their **Churn Driver**:

> **High-Value, High-Risk (The Priority):** These customers should receive high-touch, personalized interventions (e.g., proactive calls from relationship managers, premium product upgrades).
>
> **Low-Value, High-Risk:** Interventions should be automated and cost-effective (e.g., personalized digital messages, automated offers to try a new service).

- **Continuous Monitoring and Retraining**

Given that customer behavior and product usage evolve (**concept drift**), the predictive model is not a static asset. A flexible data pipeline must be established to **retrain the model automatically** (e.g., quarterly) on fresh data to ensure its continued accuracy and relevance.

In conclusion, this project delivers the necessary analytical capability to transform customer retention from a costly, uncertain process into a strategic, data-driven engine for sustainable growth.