Bank Customer Churn Analysis Project Report

Executive Summary

Project Overview

This analysis investigates customer churn patterns in banking data to identify key predictors that can help financial institutions reduce customer attrition. Using exploratory data analysis techniques on a dataset of 10,000 bank customers, this study reveals critical insights into customer behavior and retention factors.

Key Findings

- Customers who file complaints have a **90.1% churn rate** compared to **15.3%** for non-complainers.
- Customers with a **satisfaction score of 2** show the highest churn rate (**55%**), while those with a score of 5 churn only **8%**.
- Account balance levels reveal a **U-shaped relationship** with churn: low balance customers have a **34.7% churn rate**, while very high balance customers also show elevated churn compared to the mid-balance group.
- Customers with **poor credit scores (<500)** churn at **32%**, compared to **15%** for customers with excellent credit (>800).
- Supporting variables add nuance: younger customers (<30) churn more, and customers with only one product churn more than those with 2–3 products.

Business Impact

- Reducing complaint-driven churn by **50%** could lower overall churn from **24.4% to ~19%**, representing a savings of approximately **$5M annually** (assuming an average customer lifetime value of $1,000).
- Targeting low satisfaction customers with retention campaigns could save **1 in 4 at-risk customers**, potentially reducing attrition costs by an additional **$2M annually**.

Technologies Used

- Python (Pandas, NumPy, Matplotlib, Seaborn, Plotly)
- Jupyter Notebook for analysis
- Streamlit for dashboard development
- Statistical analysis and data visualization techniques

---

1. Introduction and Methodology

1.1 Business Problem

Customer churn represents a significant cost to financial institutions. This analysis aims to identify the primary factors that predict customer departure, enabling proactive retention strategies.

1.2 Dataset Description

- **Source**: Kaggle - Bank Customer Churn Dataset by Radheshyam Kollipara
- **Size**: 10,000 customers with 18 features
- **Target Variable**: Exited (0 = Retained, 1 = Churned)
- **Baseline Churn Rate**: 24.4%

1.3 Variables Analyzed

Primary Predictors (Initial Hypothesis):

- **Complain**: Customer complaint status
- **SatisfactionScore**: Customer satisfaction rating (1–5)
- **Balance**: Account balance levels

- **CreditScore**: Customer credit rating

Expanded Predictors (Iterative Analysis):

- **NumOfProducts**: Number of bank products held
- **IsActiveMember**: Account activity status
- **Age**: Customer age groups
- **Geography**: Customer location
- **Tenure**: Length of banking relationship
- **EstimatedSalary**: Income level

(Note: This iterative approach reflects my learning process — I began with 4 variables, then expanded to 6 after realizing the importance of supporting features.)

1.4 Analysis Approach

1. Data quality assessment (missing values, duplicates, data types).
2. Exploratory data analysis of key variables.
3. Hypothesis testing on primary predictors vs. baseline churn rate.
4. Comparative analysis across customer segments.
5. Development of business insights and retention recommendations.

---

2. Data Quality Assessment

2.1 Data Completeness

- **Missing values**: None detected.
- **Duplicates**: None detected.
- **Data types**: All features correctly typed (categorical and numerical).

2.2 Data Distribution

- Balanced representation of genders and geographies.
- Income distribution right-skewed, but consistent with real-world salaries.
- Customer age ranges from 18–92, median = 39.
- Most customers hold 1–2 products; fewer have 3+.

---

3. Analysis Results

3.1 Primary Hypothesis Validation

3.1.1 Complaint Analysis

Finding: 90.1% churn for complainers vs. 15.3% for non-complainers.

Business Interpretation: Complaints are the single strongest churn predictor. A lack of resolution almost guarantees attrition.

Statistical Significance: Well above baseline churn (24.4%).

3.1.2 Satisfaction Score Impact

Finding: Churn highest at Satisfaction = 2 (55%). Lowest at Satisfaction = 5 (8%).

Pattern Identified: Negative linear relationship between satisfaction and churn.

Risk Segmentation: Satisfaction ≤ 3 = high-risk; ≥ 4 = low-risk.

3.1.3 Balance Level Analysis

Finding: Low balance (<$5k) customers churn at **34.7%**, high balance (> $200k) also elevated at **30%**, mid-range balances lowest at **20%**.

Business Logic: Low balance customers may not find banking valuable; very high balance customers may leave for better investment options.

3.1.4 Credit Score Relationship

Finding: Poor credit (<500) → 32% churn; Excellent credit (>800) → 15%.

Risk Assessment: Lower creditworthiness correlates with higher churn, but not as strongly as complaints/satisfaction.

3.2 Supporting Variable Insights

- **NumOfProducts**: 1 product → 36% churn; 2 products → 15%; 3+ products → 22%.
- **IsActiveMember**: Active → 14% churn; Inactive → 28%.
- **Age**: Younger (<30) churn more than middle-aged customers.
- **Geography**: Regional differences exist; churn higher in some geographies.

3.3 Variable Ranking by Predictive Power

1. Complaints (90% churn rate)
2. Satisfaction Score (linear relationship)
3. IsActiveMember
4. NumOfProducts
5. Balance
6. Credit Score
7. Age
8. Geography

---

4. Business Insights and Recommendations

4.1 Immediate Action Items

High Priority (30 days):

- Implement rapid complaint-resolution protocol.
- Launch retention campaigns for customers with satisfaction ≤ 3.

Medium Priority (90 days):

- Strengthen engagement campaigns for inactive members.
- Promote cross-selling to single-product customers.
- Develop “premium loyalty offers” for very high-balance customers.
-

4.2 Customer Segmentation Strategy

- **High-Risk**: Complainers, Satisfaction ≤ 3, Inactive members.
- **Medium-Risk**: Low balance customers, Poor credit customers.
- **Low-Risk**: Multi-product customers, High satisfaction, Active members.

4.3 ROI Projections

- **Annual churn cost (baseline)**: 2,440 customers × $1,000 = **$2.44M**.
- **Complaint resolution impact**: Save ~500 customers/year = **$500k** savings.
- **Satisfaction improvement**: Save ~250 customers/year = **$250k** savings.
- **Total potential savings**: ~$750k annually.

---

5. Technical Implementation

5.1 Environment Setup

The project was conducted in Google Colab, which provided a cloud-based environment with direct access to Kaggle datasets

---

5.2 Analysis Pipeline

1. **Data Loading & Validation →** The dataset was accessed via KaggleHub and imported into pandas. Data quality checks (missing values, duplicates, dtypes) confirmed consistency.
2. **Exploratory Data Analysis (EDA) →** Conducted using matplotlib, seaborn, and plotly to identify trends, distributions, and correlations.
3. **Hypothesis Testing & Comparisons →** Groupby operations and visual breakdowns validated the impact of key churn drivers.
4. **Dashboard Development →**
    - **Notebook Dashboard:** Built with ipywidgets + Plotly for interactive filtering and visualization inside Colab.
    - **Streamlit App:** External interactive dashboard built for production-like deployment, offering a user-friendly interface to explore churn trends dynamically.

5.3 Technologies Used

- **Python:** Pandas, NumPy, Matplotlib, Seaborn, Plotly
- **Jupyter/Colab:** For analysis, EDA, and notebook-based dashboarding
- **ipywidgets:** For interactive in-notebook exploration
- **Streamlit:** For external dashboard deployment
- **KaggleHub:** For dataset integration

5.4 Key Technical Achievements

- Robust binning for balance and credit score.
- Clear segmentation analysis with visualizations.
- Interactive dashboard to filter churn risks.
- Scalable pipeline for future predictive modeling.

---

6. Conclusions and Future Work

6.1 Key Takeaways

- Complaints and satisfaction are the most powerful churn predictors.
- Active engagement and cross-selling reduce churn significantly.
- Dashboard enables managers to identify and act on high-risk segments.

6.2 Model Limitations

- Based only on EDA; no predictive model built yet.
- No time-series element (churn measured at one snapshot).

6.3 Future Research Directions

- Build machine learning models (logistic regression, random forest).
- Develop churn scoring system for real-time monitoring.
- Test retention campaigns with A/B experiments.

---

7. Appendices

Appendix A: Variable Definitions

Kaggle Dataset: Bank Customer Churn

Appendix B: Code Repository

- **GitHub Repository:** https://github.com/DynamicDataMindset/Bank-Churn-Streamlit-Dashboard-Analysis
- **Streamlit Dashboard**: Bank Customer Churn Analysis · Streamlit

---

Author: Boniface Ramushu
Date: 09/28/2025
Institution: Data Science Internship – Hex Softwares
Contact: bonifaceramushu28@gmail.com
