**Customer Churn Predictive Analysis:**

**Phase 1 Report**

**Author: Costas Pinto**

**Date:** August 2, 2025 **Author:** Data Science Team **Status:** Phase 1 Complete

**1.0 Executive Summary**

* **Objective:** This report details the foundational phase of a project aimed at developing a predictive model to identify customers at high risk of churning from SmartBank. The primary goal of this phase was to gather, analyze, and prepare the necessary data for machine learning.
* **Methodology:** Data from five distinct sources (Demographics, Transaction History, Customer Service, Online Activity, and Churn Status) were integrated. To create a rich analytical dataset, several behavioral features were engineered, including transaction recency, customer interaction frequency, and complaint ratios. A comprehensive Exploratory Data Analysis (EDA) was then conducted to uncover initial insights into the drivers of customer churn.
* **Key Findings:**
  + The dataset exhibits a significant **class imbalance**, with approximately 20.4% of customers having churned. This must be addressed in the modeling phase to ensure the model can accurately identify the minority class.
  + Customer **age** and **service interaction history** show a visible correlation with churn. Initial visualizations suggest that specific age brackets and customers with a higher proportion of complaints are more likely to leave.
  + The engineered features, such as DaysSinceLastTransaction and ComplaintRatio, appear to be valuable additions, showing variance that can be leveraged by a predictive model.
* **Next Steps:** The data preparation phase is now complete. The final, cleaned dataset has been generated and is ready for the development, training, and evaluation of predictive models in Phase 2.

**2.0 Data Gathering and Preparation**

* **2.1 Data Sources:** The analysis was based on the Customer\_Churn\_Data\_Large.xlsx file. All five sheets were utilized to create a comprehensive 360-degree view of each customer.
* **2.2 Feature Engineering:** To move beyond static data, the following key features were engineered to capture customer behavior:
  + **Transactional Aggregates:** TotalTransactions, TotalAmountSpent, AverageTransactionAmount.
  + **Behavioral Recency:** DaysSinceLastTransaction, DaysSinceLastLogin.
  + **Service Interaction Ratios:** ComplaintRatio (Complaints / Total Interactions).
* **2.3 Data Cleaning and Preprocessing:**
  + **Missing Values:** Handled by imputing with 0, representing an absence of activity for customers with no transaction or service history.
  + **Categorical Encoding:** Non-numerical columns like Gender and MaritalStatus were converted into a machine-readable format using one-hot encoding.
  + **Final Dataset:** All identifiers and original date columns were dropped to produce the final, model-ready dataset.

**3.0 Exploratory Data Analysis (EDA) Findings**

The following section presents the key findings from the EDA.

* **3.1 Summary Statistics:** The table below provides a statistical overview of the engineered numerical features.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Statistic** | **Age** | **Total Transactions** | **Total Amount Spent** | **Avg. Transaction Amt.** | **Days Since Last Transaction** | **Total Interactions** | **Complaint Count** | **Complaint Ratio** | **Login Frequency** | **Churn** | **Days Since Last Login** |
| **count** | 1000.00 | 1000.00 | 1000.00 | 1000.00 | 1000.00 | 668.00 | 668.00 | 668.00 | 1000.00 | 1000.00 | 1000.00 |
| **mean** | 43.27 | 5.05 | 1267.07 | 248.81 | 79.05 | 1.50 | 0.50 | 0.33 | 25.91 | 0.20 | -185.90 |
| **std** | 15.24 | 2.60 | 738.59 | 79.37 | 78.72 | 0.50 | 0.60 | 0.41 | 14.06 | 0.40 | 104.90 |
| **min** | 18.00 | 1.00 | 9.80 | 9.80 | 1.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 | -364.00 |
| **25%** | 30.00 | 3.00 | 626.69 | 203.73 | 21.00 | 1.00 | 0.00 | 0.00 | 13.75 | 0.00 | -273.25 |
| **50%** | 43.00 | 5.00 | 1232.88 | 250.12 | 53.00 | 1.50 | 0.00 | 0.00 | 27.00 | 0.00 | -190.50 |
| **75%** | 56.00 | 7.00 | 1791.90 | 295.02 | 107.00 | 2.00 | 1.00 | 0.50 | 38.00 | 0.00 | -97.00 |
| **max** | 69.00 | 9.00 | 3386.04 | 496.99 | 356.00 | 2.00 | 2.00 | 1.00 | 49.00 | 1.00 | 0.00 |

* **3.2 Churn Distribution:** The plot below illustrates the distribution of churned vs. retained customers in the dataset.
* A blue rectangular column with white text

  AI-generated content may be incorrect.**3.3 Feature Correlation:** The heatmap shows the correlation between different numerical features. This is crucial for identifying multicollinearity and understanding inter-feature relationships.
* A screenshot of a graph

  AI-generated content may be incorrect.**3.4 Age Distribution and Churn:** This visualization highlights how churn rates vary across different customer age groups.

A graph with blue and orange bars

AI-generated content may be incorrect.

**4.0 Deliverable and Next Steps**

The primary deliverable of this phase is the cleaned and preprocessed dataset, which has been saved to the following location:

* **File Path:** outputs/cleaned\_churn\_dataset\_for\_modeling.csv

This dataset forms the foundation for the next phase of the project, which will focus on building and evaluating a suite of machine learning models to accurately predict customer churn.