Rounak Saha
14/04/2025

Customer Churn Analysis Report

1. Introduction

This report summarizes the exploratory data analysis, data cleaning and preprocessing, and key insights derived from our customer churn dataset. The goal is to provide a clear, concise foundation for predictive modeling and guide future analytical efforts to reduce churn.

2. Data Sources

Selected Datasets and Rationale The dataset consists of multiple sheets from an Excel file. The following datasets were selected based on their relevance to Customer Churn prediction. All datasets have their respective data with the Customer ID which helps uniquely identify a customer.

- **Customer Demographics:** Contains features like Age, Gender, Marital Status and Income Level. These features are critical since demographic features often influence customer behaviour. There are 1000 samples in the dataset, one for each customer ID.

- **Transaction History: Included features are:** Transaction ID, Transaction Date, Amount Spent and Product Category. The Transaction ID was a redundant column since it is used to identify each transaction which has no effect on customer churn. There are 5054 samples in the dataset. Transaction patterns reveal spending behaviour which may affect churn.

- **Customer Service:** Include features like Interaction ID, Interaction Date, Interaction Type and Resolution Status. Customer service Interactions like Unresolved complaints could be a crucial indicator for customer churn. Samples: 1001

- **Online Activity:** This dataset contains the various online service usage of customers. Includes features like Last Login Date, Login Frequency and Service Usage. The online service usage patterns could signal a customer likely to churn. A high user may not be as likely to leave as their counterpart. Samples: 1000

- **Churn Status:** The target variable ‘ChurnStatus’ indicates whether a customer has churned (1) or not (0). 1000 unique samples in the dataset.

3. Exploratory Data Analysis (EDA)

- **Age Distribution**: Age is broadly distributed; a slight concentration in the 25–45 range, with churn rates higher among younger customers.
- **Spending Patterns**: Most customers spend modest amounts; high spenders exhibit lower churn.
- **Login Frequency**: Higher login frequency correlates with retention; low-activity users show elevated churn.
- **Service Usage**: Certain service packages like Mobile App and Online banking has higher churn rate than website.
- **Demographic Segments**: Churn roughly equally distributed in gender, marital status, and income level.

4. Data Cleaning & Preprocessing

4.1. Feature Engineering

CustomerID was a feature present in every dataset which identified a unique customer. New features were created based on the aggregation data in the dataset**.**

- **Total_amount_spent:** From the transaction history dataset, the amount spent by each customer was aggregated to form this new feature.

- **DaysSinceLastInteraction:** From the customer service dataset, the Interaction Date for each customer was grouped and the number of days from the latest date in the group to the latest date in the dataset was found.

- **DaysSinceLastLogin:** From the online activity dataset, days since the last login of the customer was found.

4.2. Encoding Categorical Variables

One-hot encoded Variables

- MaritalStatus >> MaritalStatus_Married, MaritalStatus_Single, MaritalStatus_Widowed

- ServiceUsage >> Online Banking, Website

- ProductCategory >> Books, Clothing, Electronics, Furniture, Groceries

- InteractionType >> Complaint, Feedback, Inquiry
- ResolutionStatus >> Resolved, Unresolved

Label Encoded Variables

- IncomeLevel (low, medium, high) >> (0,1,2)
- Gender >> Male(0) , Female(1)

4.3. Handling Missing Values

The original datasets did not have any missing values. The features after feature engineering had missing values.

The fill value 0 for columns Complaint, Feedback, Inquiry, Resolved and Unresolved represent the real scenario that the customer has never had that specific interaction.

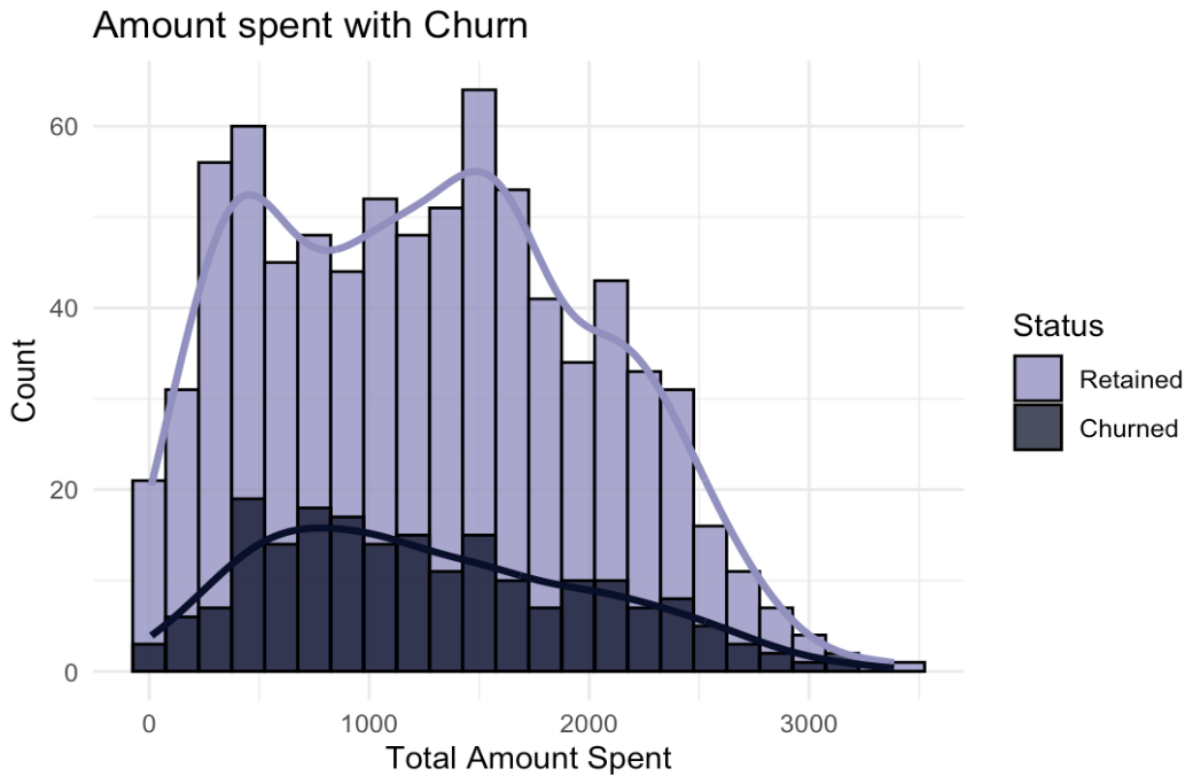
For the column DaysSinceLastInteraction, null values are filled using the median of the column.

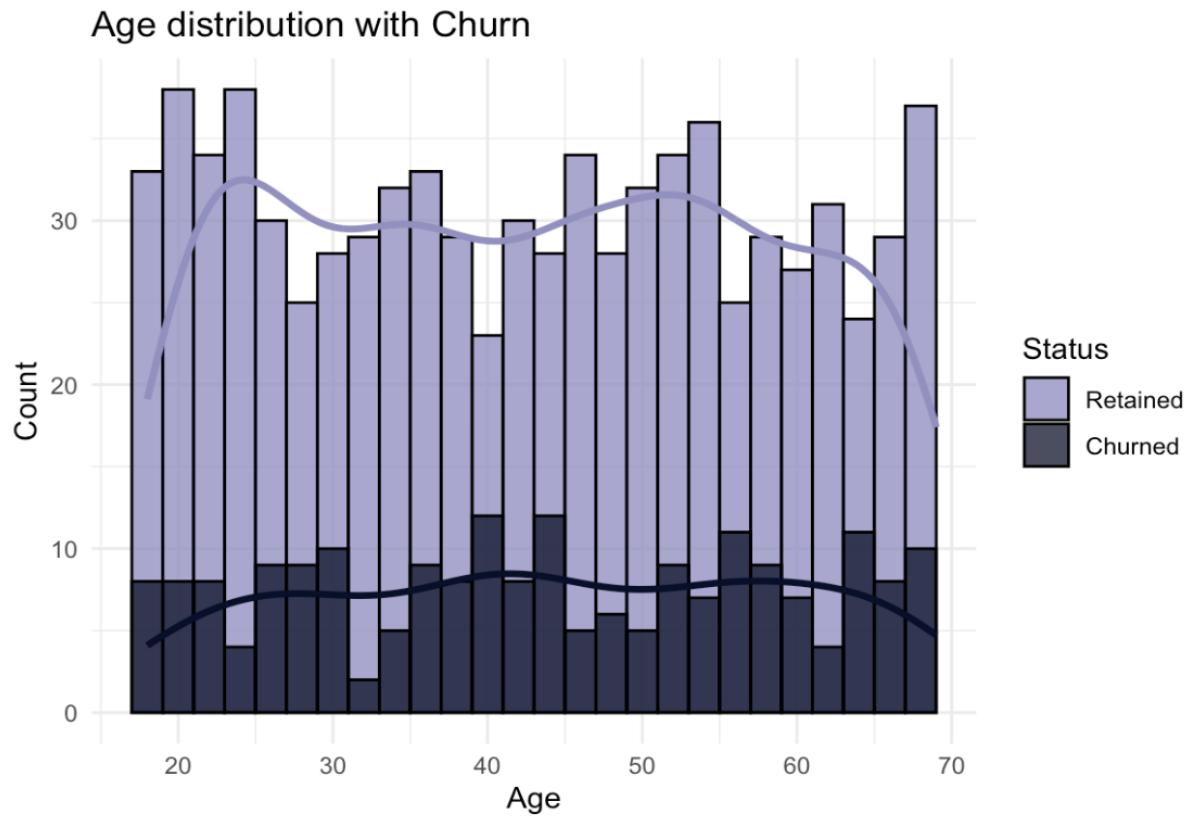
4.4. Standardization

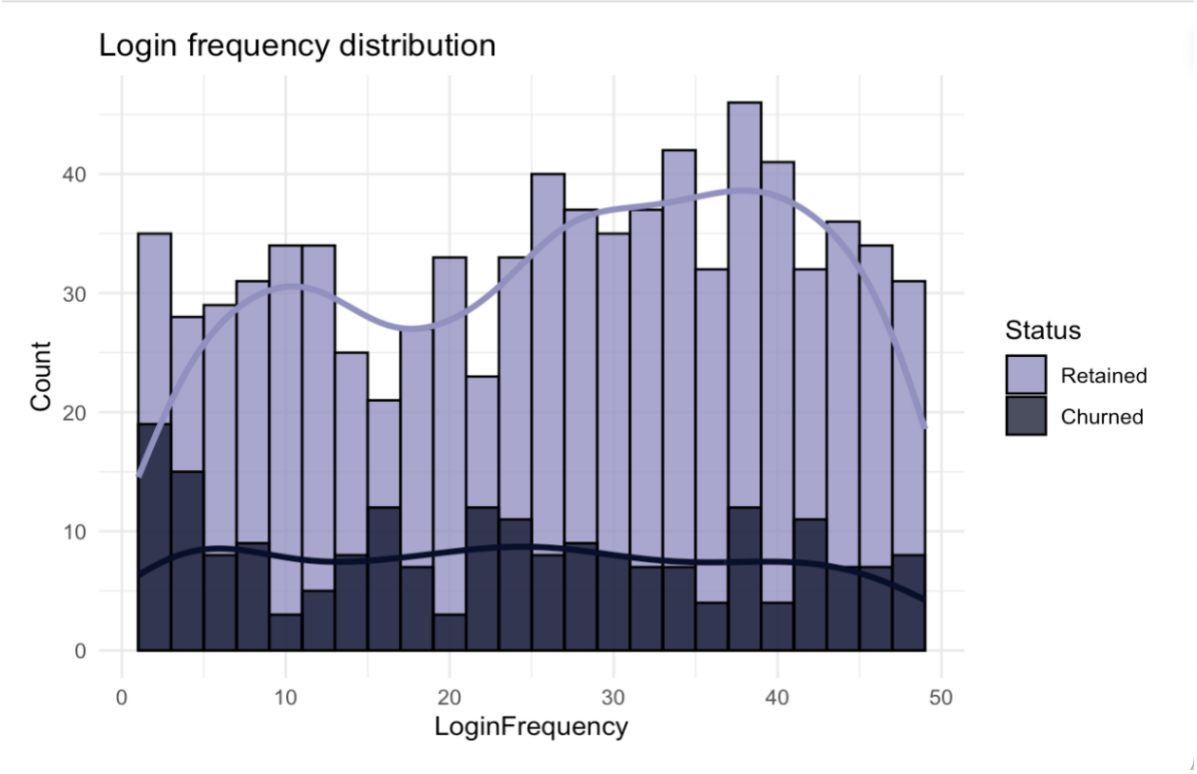
The data in the dataset is not scaled, we need to scale the data before training the model. Standardization is sufficient for logistic regression, so we are not normalizing the data.

5. Visualization Highlights

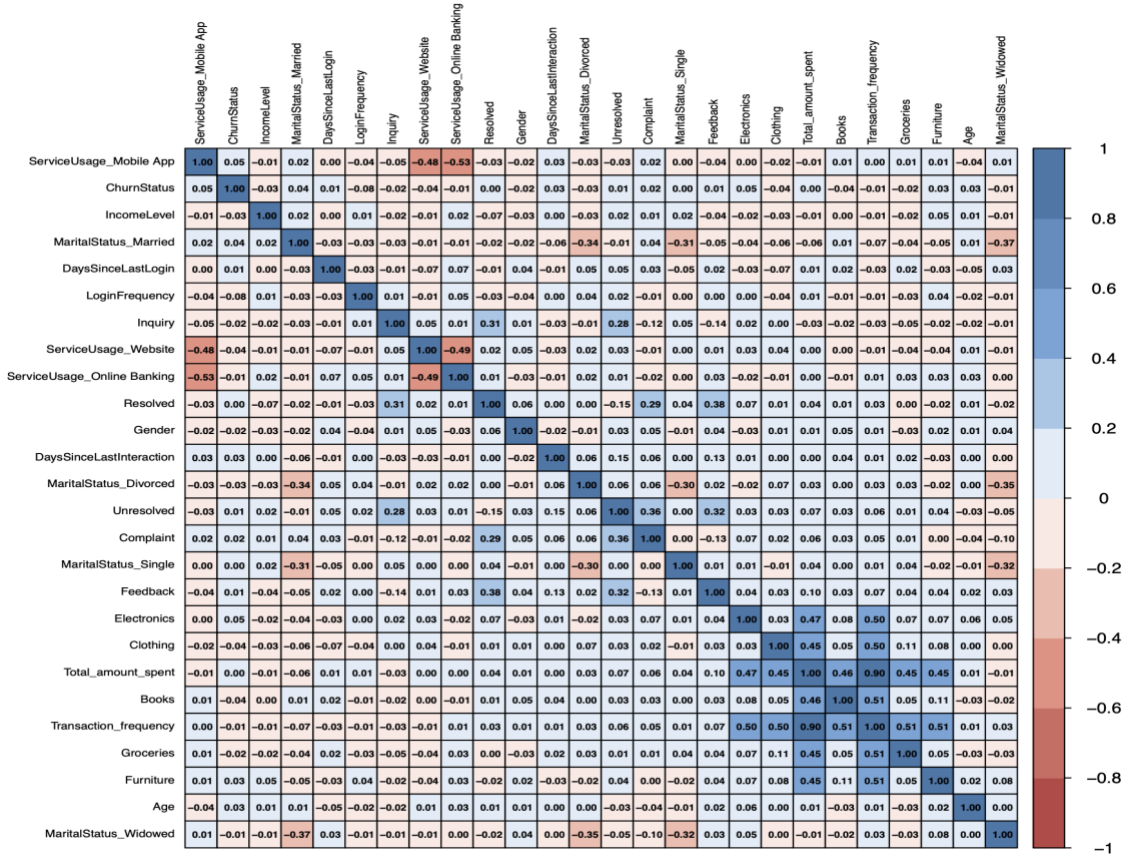
- **Histograms**: Illustrated distributions of age, spending, and login frequency by churn status.


Amount spent with Churn
Count
Total Amount Spent
0
20
40
60
0
1000
2000
3000
Status
Retained
Churned

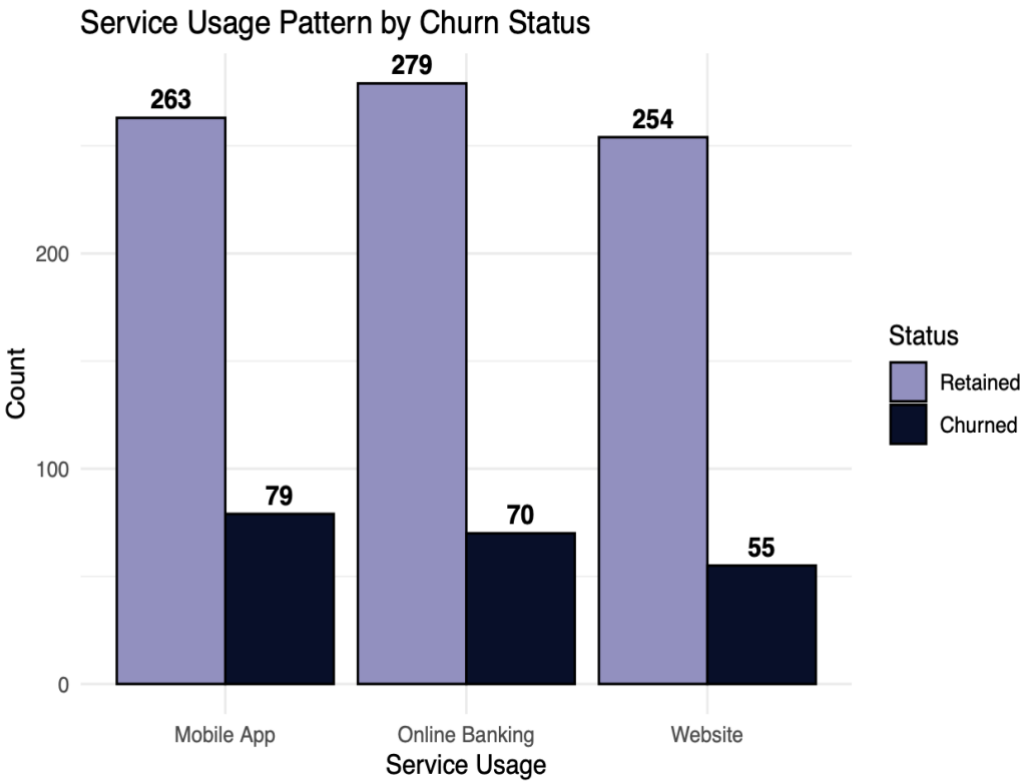

Age distribution with Churn
Count
Age
0
10
20
30
20
30
40
50
60
70
Status
Retained
Churned

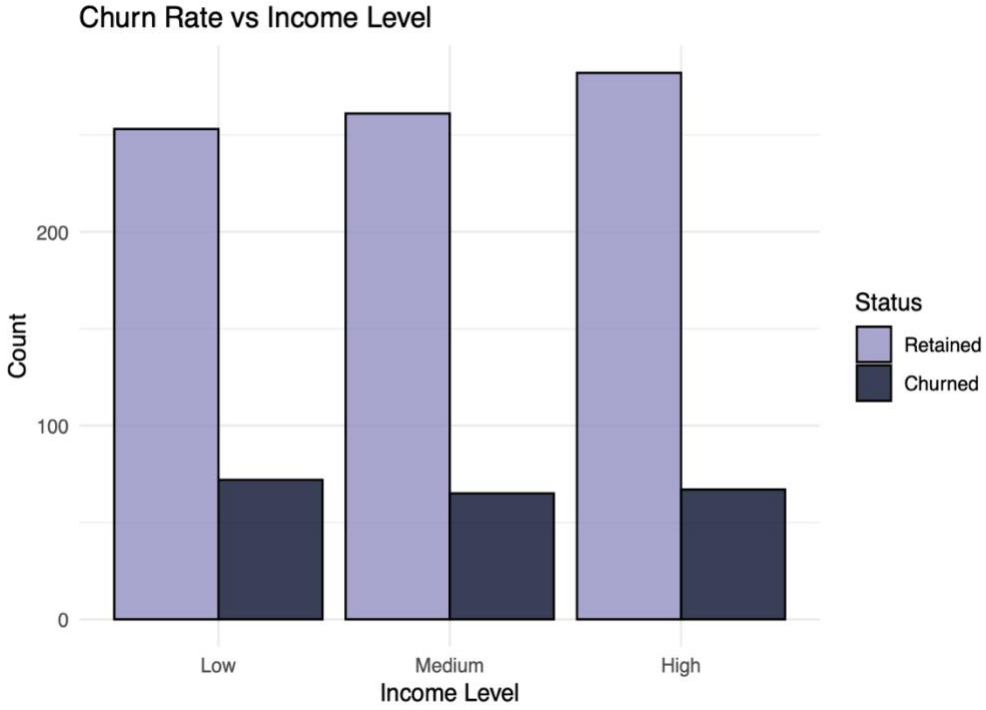


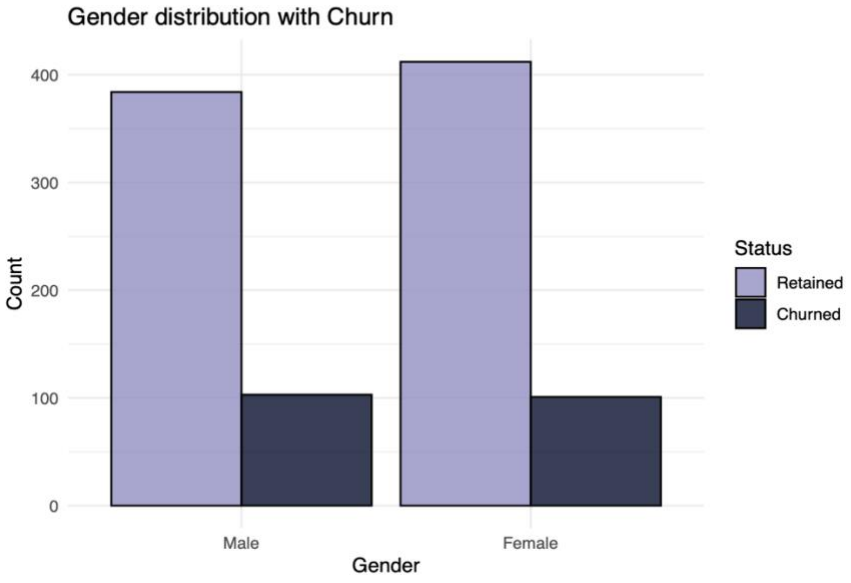
- **Heat map**: Showed the relationship between Age and AmountSpent, highlighting churn clusters.



- **Countplots**: Compared churn counts across gender, service usage, income levels.







6. Conclusions & Recommendations

Key Insights:

- Younger, low- login frequency, and low-spending customers churn at higher rates.
- service usage category significantly influence retention.

Actionable Recommendations:

- **Targeted Engagement**: Design loyalty programs for younger and low-frequency users.
- **Personalized Offers**: Develop incentives for high-risk segments based on service usage.
- **Monitoring & Alerts**: Implement real-time churn-risk scoring to trigger retention actions.

7. Next Steps

- **Model Development**: Train and validate classification models (e.g., logistic regression, Decision Tree, random forest, XGBoost).
- **Evaluation**: Use precision, recall, AUC, and business KPIs to select the optimal model.
- **Deployment**: Integrate churn prediction into CRM workflows for proactive outreach.
- **Continuous Improvement**: Monitor model performance, retrain regularly, and incorporate new data (e.g., customer feedback).