LLOYDS
BANKING GROUP

CUSTOMER RETENTION ENHANCEMENT THROUGH PREDICTIVE ANALYTICS

Task 1: Customer churn - Exploratory data analysis

Submitted for the forage job simulation

Contents

Executive Summary

This report presents a thorough analysis of customer churn data of SmartBank, a subsidiary of Lloyds Bank. The analysis includes an exploratory data analysis, data preprocessing, and preparation for predictive modelling.

The analysis integrates data from five distinct sources within the database infrastructure:

1. **Demographics Data**: Customer age, gender, marital status, and income level.
2. **Service Usage Data**: Login frequency, service platform preferences, and engagement patterns.
3. **Transaction Data**: Customer spending behaviour.
4. **Churn Data**: Customer churn status.
5. **Customer Service Interactions**: Inquiry, feedback, and complaint resolution tracking

This multi-source integration provides a complete view of customer behaviour, combining static demographic attributes with dynamic interaction patterns to create a comprehensive foundation for churn prediction modelling. The unified dataset of 1,000 customers with 22 features captures both the "who" (demographic profile) and "how" (behavioural patterns) of customer relationships.

1. Dataset Overview

1.1.Variable Selection

The Final Dataset consist of 1,000 records corresponding to number of customers and 19 variables, specifically selected for customer churn prediction modelling.

1.1.1. Rationale for variable inclusion.

Demographic variables:

- **Age**: Age is a critical demographic factor that can influence customer behaviour, preferences, and loyalty. Different age groups often have distinct financial needs and engagement patterns. For instance, younger customers may prefer digital and mobile services, while older customers may value personalized service or branch-based interactions. Identifying age-related churn trends allows the company to design age-specific retention strategies and targeted marketing campaigns.
- **Gender:** Gender differences can provide insights into variations in customer preferences and service usage. Men and women may prioritize different financial products or respond differently to customer service experiences. Understanding whether churn tendencies differ by gender helps tailor communication styles, product offerings, and engagement strategies for each segment.
- **Marital Status**: Marital status often correlates with life stage, financial responsibilities, and risk preferences. For example, married customers might be more interested in joint accounts, mortgage products, or family insurance, whereas single customers may focus more on personal loans or investment opportunities. Including marital status helps uncover whether family or relationship-related factors affect churn likelihood, enabling personalized financial service recommendations.
- **Income Level**: Income level directly affects customers' spending capacity, product usage, and sensitivity to service costs or benefits. Higher-income customers may expect premium services, while lower-income customers might be more price-sensitive. This variable helps in identifying whether dissatisfaction or churn risk is linked to unmet financial expectations, poor product fit, or lack of perceived value across income brackets.

Service Usage Variables

- **Login Frequency**: Login frequency indicates how actively a customer engages with the company's online or mobile platforms. A declining login trend may signal decreasing interest or dissatisfaction, which is often a leading indicator of churn. Conversely, highly active users tend to have stronger brand engagement and lower churn risk. Monitoring this metric enables early intervention through re-engagement campaigns.

- **Service Usage**: The types and number of services used (e.g., loans, savings, credit cards) provide valuable insights into customer dependency on the company's offerings. Customers who use multiple services tend to have higher switching costs and are less likely to churn, while those with minimal engagement may be more vulnerable. Understanding service usage patterns helps identify which services drive loyalty and which may need improvement.
- **Customer Total Amount Spent**: The total amount a customer spends reflects both loyalty and engagement level. High-spending customers might expect premium support, while low-spending ones may feel neglected or undervalued. Sudden drops in spending can signal dissatisfaction or a shift to competitors. Tracking spending behaviour helps identify at-risk customers and develop personalized retention offers.

Interactions variable

- **Inquiry Interaction**: Frequent inquiries may suggest that the customer is uncertain, facing issues, or needs additional support and that may suggest ambiguity in company services or instructions. High inquiry volume can indicate potential frustration or confusion, which could lead to churn if not managed effectively.
- **Inquiry Status Resolved**: A high count of resolved inquiries reflects effective customer support, which can strengthen trust and reduce churn likelihood.
- **Inquiry Status Unresolved**: Unresolved inquiries are a strong predictor of churn, as they indicate dissatisfaction and unaddressed customer needs.
- **Feedback Interaction**: Frequent feedback submissions show that the customer is engaged and cares about improving services. This can be both a positive signal (engaged user) or an early warning of dissatisfaction depending on sentiment.
- **Feedback Status Resolved**: Resolved feedback demonstrates that the company values customer input, leading to increased satisfaction and retention.
- **Feedback Status Unresolved**: Unresolved feedback can cause frustration and signal neglect, increasing the risk of churn. This variable helps identify gaps in follow-up processes.
- **Complaint Interaction**: Complaint frequency is a direct measure of dissatisfaction. While occasional complaints are normal, repeated complaints indicate recurring issues or unmet expectations.
- **Complaint Status Resolved**: Effective complaint resolution is key to customer retention. This variable measures the responsiveness and efficiency of customer service, which can mitigate churn risk.
- **Complaint Status Unresolved**: Unresolved complaints are one of the strongest churn predictors, as they signify poor service recovery and persistent dissatisfaction.

Target Variable:

Churn Status: The *ChurnStatus* variable will serve as the dependent (target) variable in the prediction model. It will take binary values (*1 = churned*, *0 = retained*), indicating whether a customer has discontinued using the company's services. It will enable the model to identify patterns and relationships between the independent variables and the likelihood of churn, forming the foundation for predictive analytics and retention modelling.

1.2. Data Transformation

The data transformation process was conducted in Microsoft Excel. After selecting the relevant variables of interest, a new worksheet was created to serve as the consolidated dataset. Using the VLOOKUP function, demographic data such as age, gender, marital status, and income level were imported into this new sheet from the original source tables.

For the transaction data, the total amount spent by each customer was calculated by summing all transaction values associated with that customer. This approach provided an efficient representation of individual spending behaviour. The resulting Total Amount Spent column was then referenced and added to the new dataset using VLOOKUP.

Next, service usage data were incorporated into the same sheet, again using the VLOOKUP function to ensure accurate matching between customer IDs and their corresponding service information. For the service interaction data, the interaction records were broken down into six distinct variables to capture more detailed behavioural insights. Using the COUNT function, we calculated the number of occurrences for each interaction type (inquiries, feedback, and complaints) and further classified them based on their resolution status, counting how many were resolved and how many remained unresolved and that with respect to each customer.

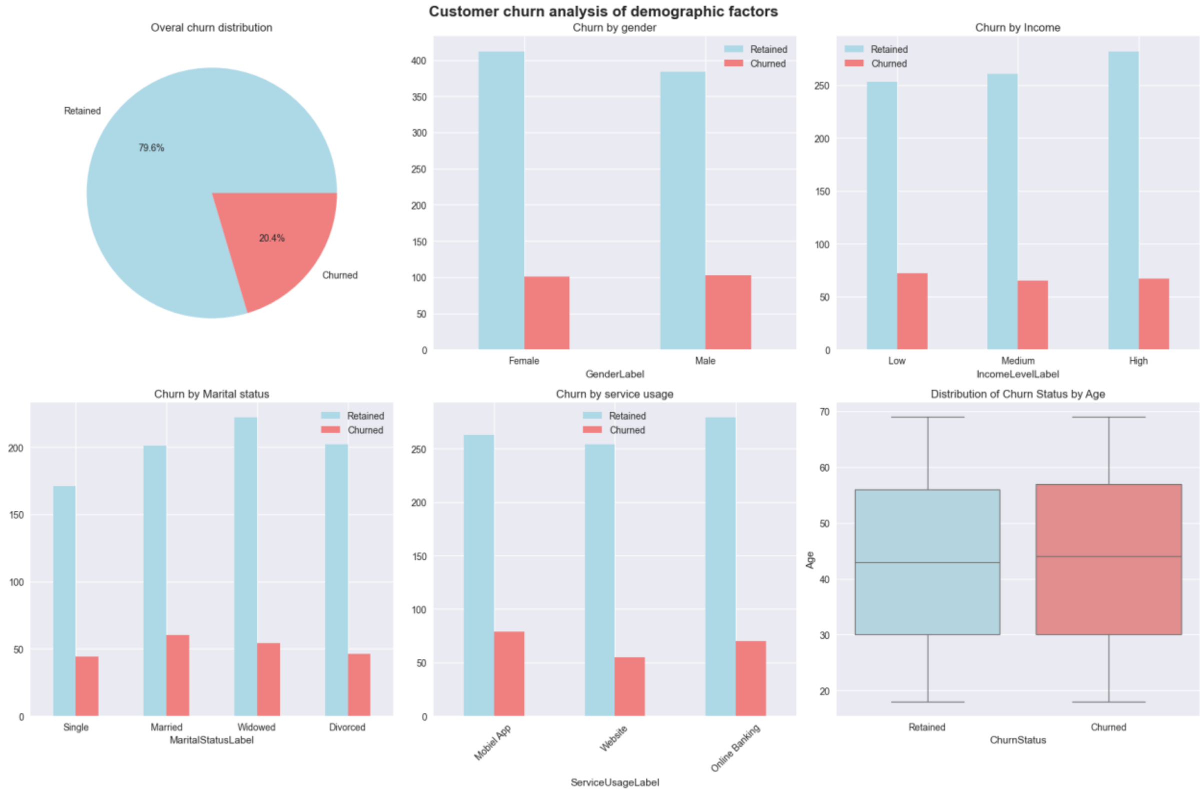
Finally, the ChurnStatus variable was added to the dataset using the VLOOKUP function, linking each customer's churn status from the original churn data table to the consolidated sheet.

This systematic transformation process ensured that all relevant demographic, behavioural, and interaction data were accurately merged into a single, analysis-ready dataset for churn prediction modelling.

2. Exploratory analysis

The dataset reveals that a total of 796 customers (79.6%) were retained, while 204 customers (20.4%) churned, resulting in an overall churn rate of 20.4%. This indicates that while most customers remain loyal, a notable minority have discontinued their engagement. The average customer age is 43 years (ranging from 18 to 69, with a standard deviation of 15.24), suggesting a broad age distribution across the customer base. In terms of engagement, customers log in an average of 25.91 times (range: 1–49) reflecting diverse usage behaviour. The average total amount spent by customers is 1,267.07 with a standard deviation of 738.59, indicating substantial variation in spending patterns. Regarding categorical features, the gender distribution is nearly balanced with 51.3% female and 48.7% male, while married customers form the largest group across the four marital status categories. The income levels are relatively evenly spread across low, medium, and high brackets, and customers use a mix of website, mobile app, and online banking services. Finally, customer interaction metrics show an average of 0.31 inquiries, 0.36 feedback submissions, and 0.34 complaints per customer, with resolution rates varying by interaction type, suggesting differences in customer service responsiveness across categories.

2.1.Demographic Analysis of Churn

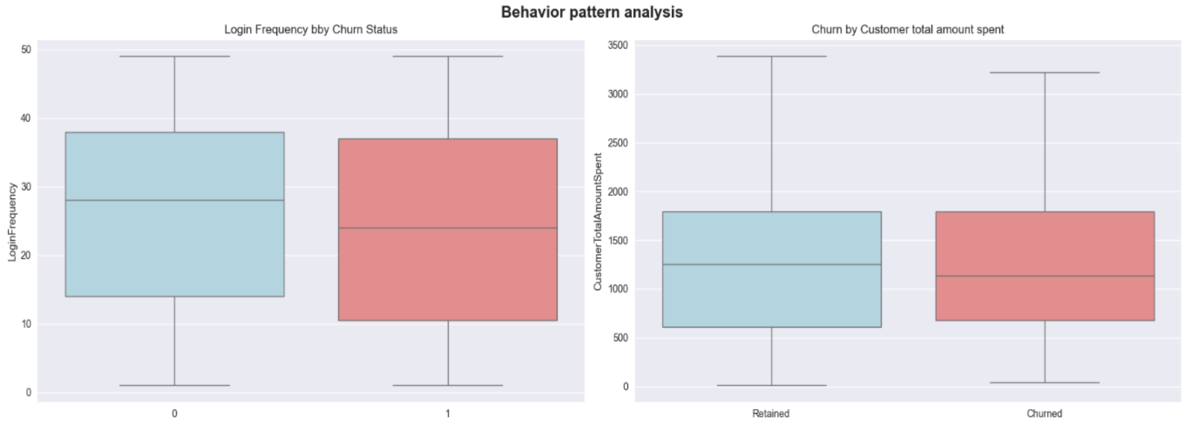


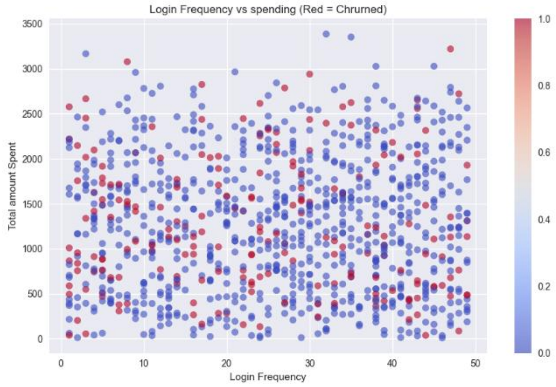
- **Churn by Gender**
  Both male and female customers show similar churn patterns, with no major difference in churn rate between genders. This suggests that gender may not be a strong predictor of churn in this dataset.
- **Churn by Income Level**
  Customers across low, medium, and high-income levels display similar churn proportions. However, there is a slight trend where lower-income customers churn slightly more, possibly indicating sensitivity to pricing or service costs.
- **Churn by Marital Status**
  Marital status appears to have **some influence on churn**. Married and widowed customers show slightly higher churn rates than single or divorced individuals. This may reflect differing financial responsibilities or product needs across life stages.
- **Churn by Service Usage**
  Service usage type (Mobile App, Website, or Online Banking) shows relatively balanced churn distributions. However, website users show a marginally higher churn

rate, which could imply lower satisfaction with that platform's user experience compared to mobile or online banking users.

- **Distribution of Churn Status by Age**
  The boxplot shows that churned customers tend to be slightly older on average than retained ones. The median age for churned users is higher, suggesting that older customers may be less engaged with newer digital services or have different service expectations that aren't being met.

2.2. Behavioural Pattern Analysis





- **Login Frequency by Churn Status**
  The boxplot indicates that churned customers generally log in less frequently than retained ones. This suggests a strong relationship between low engagement (as reflected by login activity) and churn behavior.
- **Customer Total Amount Spent by Churn Status**
  Churned customers have a slightly lower median spending compared to retained ones. This could imply that lower-spending customers are less invested in the company's services, making them more likely to leave.
- **Login Frequency vs. Spending**
  The scatterplot shows that most churned customers cluster in the lower-to-mid range of both login frequency and total spending. This reinforces the idea that lower engagement and lower financial activity correlate with a higher likelihood of churn**.**

3. Data Cleaning and Preprocessing

3.1. Missing Data

No Missing data detected

3.2. Data Encoding

Categorical variables were pre-processed through label encoding to convert qualitative data into numerical format suitable for machine learning models. The encoding scheme applied was as follows:

- GenderLabel: 0 = Female, 1 = Male
- MaritalStatusLabel: 1 = Single, 2 = Married, 3 = Widowed, 4 = Divorced
- IncomeLevelLabel: 1 = Low, 2 = Medium, 3 = High
- ServiceUsageLabel: 1 = Mobile App, 2 = Website, 3 = Online Banking

This standardized encoding ensures that categorical attributes were represented numerically while maintaining interpretability. All features were thus transformed into a consistent, machine-readable structure, enabling reliable input for the churn prediction model.

3.3. Data scaling Standardization

Standardization was performed on Customer Total Amount Spent, as it exhibited the widest range and highest standard deviation. Age and Login Frequency were also standardized due to their continuous nature and differing numeric scales. Additionally, count-based variables such as Inquiry Interaction, Feedback Interaction, Complaint Interaction and the other variables related to them, were standardized to maintain consistency across numeric features. This was done in Python using *sklearn.preprocessing.StandardScaler*, which transforms numeric features to have a mean of 0 and a standard deviation of 1. In contrast, categorical, binary, and identifier variables were left unstandardized, as standardization is not appropriate for their data types.