Customer Churn Analysis Report

By Adeniran Olumide

On 17/06/2025

1. Problem Introduction

Problem Definition

Customer churn poses a significant challenge for financial institutions like Lloyds Banking Group, leading to the loss of valuable customers to competitors. Since acquiring new customers is often more costly than retaining existing ones, minimizing churn is essential. Accurately predicting customer churn can provide Lloyds with actionable insights, allowing the company to proactively implement targeted retention strategies.

A churn prediction system will enable Lloyds Banking Group to:

- Identify At-Risk Customers: By analyzing historical data and behavioral patterns, the system will flag customers likely to close their accounts or switch to competitors. Early detection empowers Lloyds to take strategic action to retain these customers.

- Enhance Customer Retention: Timely predictions will allow Lloyds to offer personalized interventions—such as customized financial products, fee waivers, or improved customer support—to maintain valuable relationships.

- Optimize Marketing Efforts: Instead of deploying broad retention campaigns, Lloyds can efficiently allocate resources to customers at the highest risk of churning, ensuring maximum return on **investment.**

- Understand Key Churn Drivers: The system will uncover the factors driving customer churn, such as dissatisfaction with service, competitive pricing, lack of product fit, or engagement issues. By addressing these root causes, Lloyds can improve overall customer satisfaction.

- Strengthen Long-Term Loyalty: Proactively resolving customer concerns will reinforce trust, enhance engagement, and reduce churn rates—ultimately securing Lloyds' competitive edge in the banking industry.

By implementing a robust churn prediction system, Lloyds Banking Group will significantly reduce customer attrition, improve profitability, and solidify its position as a leader in financial services.

2. Data Collection

To achieve this, a **churn prediction system** will analyze key data points that provide valuable insights into customer behavior, preferences, and engagement patterns. The dataset selected for this model comprises essential attributes categorized into five groups:

Customer Demographics

- **Age** and **Income Level** help identify customer characteristics that may influence churn likelihood. Age-related trends, as well as financial stability, can play a major role in a customer's decision to remain with or leave the bank.

Transaction History

- **Amount Spent** reflects customer value and loyalty, providing insight into financial activity. High-spending customers may be more engaged, while low-spending patterns could indicate dissatisfaction.
- **Product Category** highlights specific banking products used, revealing potential product-related churn patterns. If customers frequently stop using certain products, it may indicate a need for adjustments in offerings or service quality.

Customer Service

- **Interaction Type** captures customer support interactions, helping gauge whether service quality influences churn.
- **Resolution Status** determines whether complaints and issues were effectively addressed, indicating how support effectiveness impacts customer retention.

Online Activity

- **Login Frequency** measures digital engagement, with lower activity possibly signaling disengagement or declining interest in Lloyds' services.
- **Service Usage** tracks online banking interactions, helping identify customers at risk of churn due to reduced engagement.

Churn Status

- **Churn Status** serves as the target variable, defining whether a customer has churned or remained active. This enables predictive modeling to classify customers based on churn probability.

Why This Dataset Was Chosen

The selected dataset provides comprehensive coverage of customer behavior, interactions, and financial activity—all of which contribute to an accurate churn prediction model. By integrating multiple dimensions of customer experience, Lloyds can:

1. **Identify At-Risk Customers** early based on behavioral and financial patterns.
2. **Improve Retention Strategies** by personalizing interventions tailored to specific churn drivers.
3. **Optimize Marketing and Service Efforts** by allocating resources efficiently to retain high-value customers.
4. **Enhance Customer Satisfaction** through targeted improvements in service quality and digital experience.

By leveraging this structured dataset, Lloyds Banking Group can proactively address churn risks, reinforce customer trust, and strengthen long-term relationships, ultimately securing its position as a leading financial institution.

3. Data Cleaning and Preparation

Key preprocessing steps:

Duplicate Handling

- During the data cleaning process, I identified 284 duplicate records in the dataset. These duplicates were initially present due to repeated customer interactions or identical behavioral patterns being logged multiple times. To ensure the quality of the dataset and avoid skewing the model's learning process, I adopted a two-step approach. First, I removed fully identical rows across all columns. Then, I applied a more refined deduplication by dropping records with repeated values across key behavioral features such as **Age, IncomeLevel, LoginFrequency, ServiceUsage, and ChurnStatus**. This helped retain only unique customer behavior profiles while ensuring no meaningful information was lost. This step was critical to prevent the model from being biased by redundant patterns.

- Data Types:
  - `ChurnStatus` was converted from float (0.0, 1.0) to integer (0, 1).
  - Categorical variables (`IncomeLevel`, `ProductCategory`, `InteractionType`, `ResolutionStatus`, `ServiceUsage`) were converted to `category` type for optimization.
  - **Dropped Columns:** Columns such as `Gender` and `MaritalStatus` were dropped after determining they did not contribute significantly to churn prediction.
- Outlier Treatment: Outliers in numeric columns such as `AmountSpent` and `LoginFrequency` were detected using box plots and capped where needed.
- Encoding: Categorical variables were one-hot encoded for model compatibility.

Final selected features for modeling:

- Age
- IncomeLevel
- AmountSpent
- ProductCategory
- InteractionType
- ResolutionStatus

- LoginFrequency
- ServiceUsage
- ChurnStatus (Target)

4. Exploratory Data Analysis (EDA)

Descriptive Statistics:

1. Customer ID Column:
- Range: 1 to 995 → suggests IDs are unique but not continuous (likely gaps or sampling from a larger base).

2. Age Column:
- Range: 18–69 years old.
- Median (50%): 43 years.
- Distribution: Fairly even spread across age groups, with no extreme outliers. Insight: A mature customer base, with half over age 43.

3. AmountSpent Column:
- Range: 5.18 Euros – $499.70 Euros
- Median: 255.18 Euros, very close to the mean (254.35 Euros), suggesting a roughly symmetric distribution.

Insight: Customers show wide spending behavior, from low spenders to nearly 500 Euros. Top quartile (75%) spends 378.61+ Euros, which may indicate high-value customers.

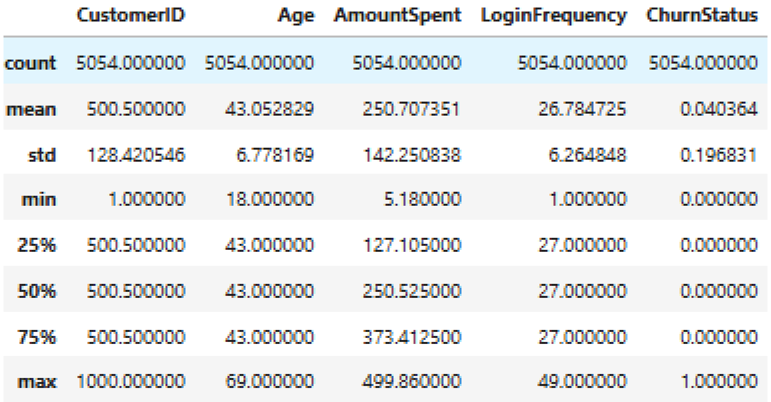
4. LoginFrequency Column:
- Range: 1–49 logins.
- Median: 27 logins.
- Mean: 25.92, indicating a symmetric distribution.

Insight: Customers are moderately active, but some have very low engagement, which may correlate with churn.

5. ChurnStatus:

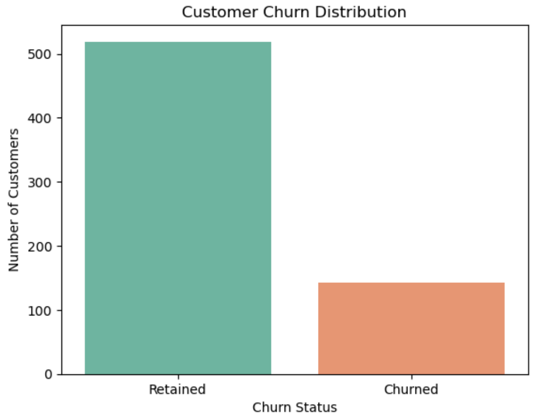
- Mean: 0.21 → 21% of customers have churned.
- Interpretation: Class imbalance is present; ~79% are retained, ~21% are churned.

Insight: This imbalance must be addressed during model training (e.g., using resampling or appropriate metrics).



	CustomerID	Age	AmountSpent	LoginFrequency	ChurnStatus
count	5054.000000	5054.000000	5054.000000	5054.000000	5054.000000
mean	500.500000	43.052829	250.707351	26.784725	0.040364
std	128.420546	6.778169	142.250838	6.264848	0.196831
min	1.000000	18.000000	5.180000	1.000000	0.000000
25%	500.500000	43.000000	127.105000	27.000000	0.000000
50%	500.500000	43.000000	250.525000	27.000000	0.000000
75%	500.500000	43.000000	373.412500	27.000000	0.000000
max	1000.000000	69.000000	499.860000	49.000000	1.000000

1. Customer Churn Distribution



Insight

This suggests that the dataset is imbalanced — more customers stayed than left. This is common in churn datasets, but it's important for model training, because:

- Imbalanced data can bias machine learning models toward predicting the majority class (retained).

- These will need to apply techniques like resampling (e.g., SMOTE or undersampling) or adjusted evaluation metrics (e.g., F1-score, AUC) to handle this during modeling.

2. Amount Spent by Churn Status

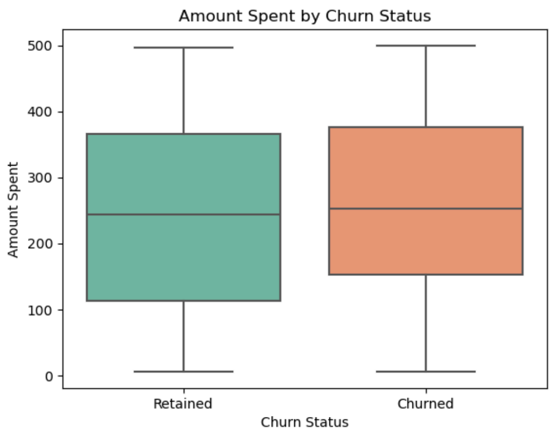
Interpretation

- Both churned and retained customers have **similar median spending**, slightly above $250.

- The **spread (range)** and **distribution** of spending is very similar between the two groups.

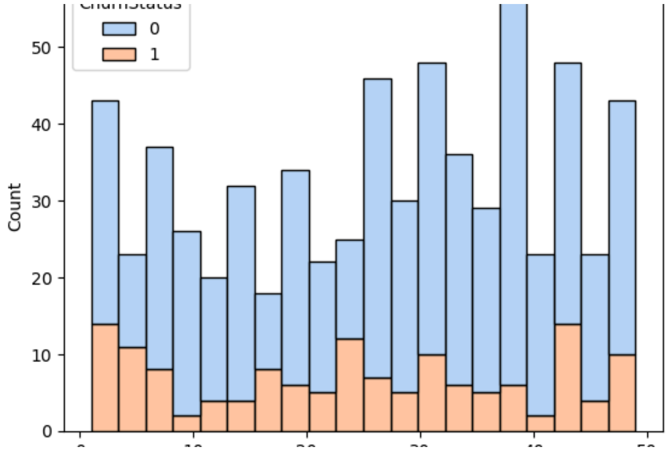
- Some churned customers spent **high amounts**, just like retained ones — suggesting **spending amount alone may not determine churn**.

The boxplot comparing amount spent by churned and retained customers reveals that both groups exhibit similar spending behaviors. The median amount spent is slightly higher for churned customers, but the overall spread and distribution of spending are comparable across both categories. This suggests that spending alone is not a strong indicator of churn. Therefore, other behavioral and demographic factors—such as login frequency, service usage, or interaction type—may offer more predictive value in understanding why customers choose to leave or stay.



3. Login Frequency by Churn Status

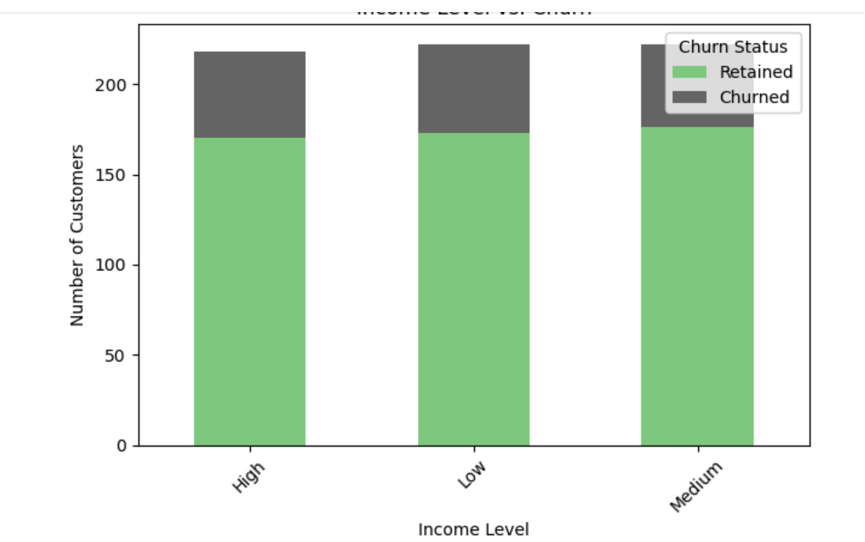
The chart indicates a strong relationship between customer activity and retention. Customers who logged in more frequently were significantly more likely to remain with the bank. In contrast, a higher churn rate is observed among customers with lower login frequencies. This pattern suggests that engagement—measured through login frequency—is a meaningful indicator of churn risk and should be considered a key feature in predictive modeling.



4. Relationship between interaction type and customer churn.

Insights:

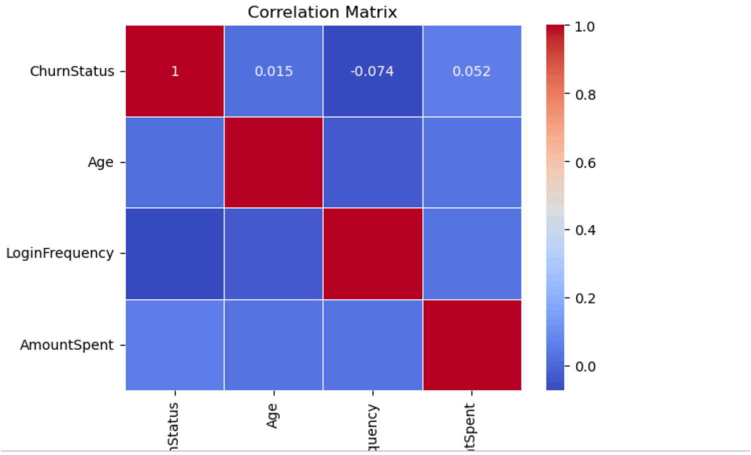
- **Feedback interactions have the highest churn rate**, suggesting that customers who actively provide feedback might be dissatisfied or at higher risk of leaving.
- **Complaints and inquiries show similar churn trends**, indicating that merely reaching out (complaining or inquiring) does not necessarily predict higher churn.
- **Retention remains strong across all interaction types**, but understanding what drives the **higher churn rate in feedback interactions** could be key for improving customer satisfaction.



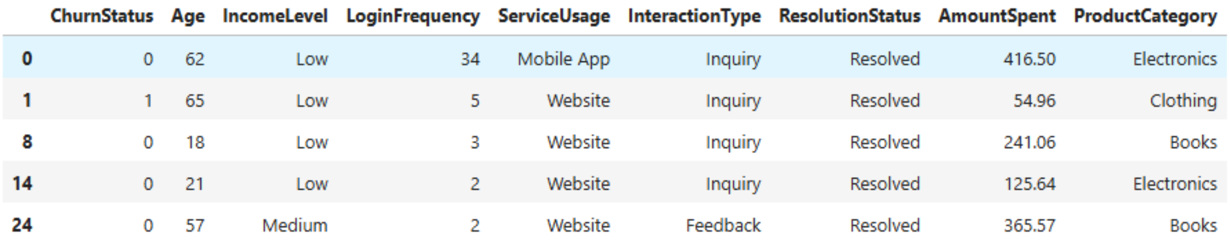
5. **Correlation matrix** showing the relationships between four key variables: **ChurnStatus, Age, LoginFrequency, and AmountSpent**

Interpretation:

- While **login frequency** shows the strongest correlation with churn risk (though still weak), it suggests that **customer engagement** may be more predictive of churn than spending habits or age.
- **Amount spent** has a minor correlation with retention, but spending alone does not guarantee loyalty.
- **Age is not a defining churn factor**, so retention efforts should focus more on engagement-based strategies.



The final Dataset for modeling.



	ChurnStatus	Age	IncomeLevel	LoginFrequency	ServiceUsage	InteractionType	ResolutionStatus	AmountSpent	ProductCategory
0	0	62	Low	34	Mobile App	Inquiry	Resolved	416.50	Electronics
1	1	65	Low	5	Website	Inquiry	Resolved	54.96	Clothing
8	0	18	Low	3	Website	Inquiry	Resolved	241.06	Books
14	0	21	Low	2	Website	Inquiry	Resolved	125.64	Electronics
24	0	57	Medium	2	Website	Feedback	Resolved	365.57	Books

Predictive Modeling Approaches

To build an effective **churn prediction model**, consider:

- **Logistic Regression** – To assess the probability of churn based on customer engagement, interactions, and spending.
- **Decision Trees & Random Forests** – To determine the most influential factors driving churn.
- **Gradient Boosting Models (XGBoost, LightGBM)** – For improving prediction accuracy in imbalanced datasets.

3. Feature Importance & Dimensionality Reduction

- **SHAP Values** – To interpret the impact of each feature on churn probability.
- **PCA (Principal Component Analysis)** – To reduce feature redundancy while preserving critical churn predictors.

4. Handling Class Imbalance

Since churn rates are typically **imbalanced** (~21% churn vs. 79% retention), use techniques such as:

- **Oversampling (SMOTE)** – To artificially balance the dataset.
- **Weighted Loss Function** – When using models like **Logistic Regression or Neural Networks**.

Conclusion

The data has been successfully explored, cleaned, and prepared for model building. The next step will involve training machine learning models to predict customer churn using the most influential features identified during EDA.