PROJECT 3: ADVANCED MACHINE LEARNING

--

— Credit Card User Churn Prediction—

Tiffany Rhodes

February 7, 2025

Contents / Agenda

- Executive Summary
- Objectives
- Business Problem Overview
- Solution Approach & Methodology
- Data Processing
- Model Performance Summary-Hyperparameter Tuning
- Conclusion
- Appendix

Executive Summary

- Thera Bank has recently experienced a decline in its credit card user base, posing a potential risk to revenue from various fees associated with credit card usage. To mitigate this challenge, the bank seeks to leverage advanced machine learning techniques to identify customers at risk of churning and implement targeted retention strategies.

- By analyzing customer data, the bank aims to uncover key factors influencing customer attrition and develop actionable insights to improve service offerings, enhance customer satisfaction, and ultimately reduce churn. This data-driven approach will enable Thera Bank to proactively address customer concerns, strengthen engagement, and safeguard its long-term financial performance.

Objectives

The primary goal of this project is to develop a predictive machine learning model that accurately forecasts customer churn. This will enable Thera Bank to take preventive measures, improve customer engagement, and optimize service offerings. Specifically, the project focuses on:

- ✓ Exploring and visualizing the dataset to understand customer behavior and churn patterns.
- ✓ Building classification models to predict customer churn with high accuracy.
- ✓ Optimizing models using hyperparameter tuning for improved performance.
- ✓ Deriving insights and recommendations to guide business strategies for customer retention.

Executive Summary

Key Insights:

Inactivity is a Strong Predictor of Churn

- Customers who have been inactive for **3+ months** are at a **high risk of leaving**.
- A significant number of churned customers had **low transaction activity** in the last 12 months.

Transaction Frequency and Engagement Impact Retention

- Customers with **higher transaction counts are less likely to churn**.
- A **decline in transaction volume between Q4 and Q1** strongly correlates with attrition.

Credit Utilization Patterns Affect Churn

- Customers with **low revolving balances and lower utilization ratios** tend to churn more.
- High-credit-limit customers show **lower engagement**, indicating the need for enhanced loyalty programs.

Demographic Factors Play a Role in Customer Churn

- Customers aged **40-55** are the most stable segment, while **younger customers have higher churn rates**.
- Customers with **fewer banking products (1-2) are more likely to leave** compared to those using multiple services.

Contact Frequency and Customer Support Engagement

- Customers who **contact the bank frequently** may indicate dissatisfaction and are at **greater risk of churn**.
- The majority of customers engage with the bank **only 2-3 times a year**, highlighting a need for **better engagement strategies**.

Data Dictionary

- **CLIENTNUM:** Client number. Unique identifier for the customer holding the account
- **Attrition_Flag:** Internal event (customer activity) variable - if the account is closed then "Attrited Customer" else "Existing Customer“
- **Customer_Age:** Age in Years
- **Gender:** Gender of the account holder
- **Dependent_count:** Number of dependents
- **Education_Level:** Educational Qualification of the account holder - Graduate, High School, Unknown, Uneducated, College(refers to a college student), Post-Graduate, Doctorate.
- **Marital_Status:** Marital Status of the account holder
- **Income_Category:** Annual Income Category of the account holder
- **Card_Category:** Type of Card
- **Months_on_book:** Period of relationship with the bank
- **Total_Relationship_Count:** Total no. of products held by the customer
- **Months_Inactive_12_mon:** No. of months inactive in the last 12 months

Data Dictionary

- **Contacts_Count_12_mon:** No. of Contacts between the customer and bank in the last 12 month
- **Credit_Limit:** Credit Limit on the Credit Card
- **Total_Revolving_Bal:** The balance that carries over from one month to the next is the revolving balance
- **Avg_Open_To_Buy:** Open to Buy refers to the amount left on the credit card to use (Average of last 12 months)
- **Total_Trans_Amt:** Total Transaction Amount (Last 12 months)
- **Total_Trans_Ct:** Total Transaction Count (Last 12 months)
- **Total_Ct_Chng_Q4_Q1:** Ratio of the total transaction count in 4th quarter and the total transaction count in 1st quarter
- **Total_Amt_Chng_Q4_Q1:** Ratio of the total transaction amount in 4th quarter and the total transaction amount in 1st quarter
- **Avg_Utilization_Ratio:** Represents how much of the available credit the customer spent

Business Problem Overview

Introduction

Thera Bank has been facing a decline in its credit card user base, which directly impacts its revenue from transaction fees, annual fees, interest charges, and other financial services. Customer churn in the credit card sector is a critical issue, as acquiring new customers is significantly more expensive than retaining existing ones. This project aims to **analyze customer behavior, predict churn, and develop strategies to enhance customer retention.**

Key Challenges

- **Revenue Loss Due to Customer Attrition**
  - Churned customers reduce the bank's revenue from credit card transactions, interest payments, and service fees.
  - Loss of long-term customers impacts the bank's profitability and growth potential.

- **Limited Customer Engagement & Usage**
  - Customers with **low transaction frequency and long inactivity periods** are more likely to churn.
  - The absence of targeted engagement strategies leads to **missed opportunities for customer retention**.

- **Understanding the Drivers of Churn**
  - Customer churn is influenced by multiple factors such as **credit utilization, inactivity, contact frequency, transaction behavior, and demographic attributes**.
  - The bank needs a **data-driven approach to accurately identify churn predictors** and mitigate risks.

- **Class Imbalance in Customer Data**
  - The dataset shows **more active customers than churned ones**, which can affect predictive model accuracy.
  - Proper balancing techniques, such as **oversampling or undersampling,** are required to improve model performance.

Business Problem Overview

What Is a Revolving Balance?

A **revolving balance** refers to the **unpaid portion of a credit card bill that carries over to the next month** if the full amount is not paid. This balance accrues interest and affects a customer's financial obligations.

Business Insight:

- Customers who **consistently carry a high revolving balance** may be more reliant on credit but also at risk of financial strain.
- **Churn Risk:** Customers with **low or zero revolving balances** may be using the credit card minimally, which could indicate **disengagement or the likelihood of switching to another financial institution**.

What is the Average Open to Buy?

The **"Open to Buy" (OTB)** refers to the **remaining credit available for spending**. The column **Avg_Open_To_Buy** represents the **12-month average of this available credit**.

Business Insight:

- Customers with **high open-to-buy limits** may have **low credit utilization**, suggesting they are not actively using the card.
- A **low open-to-buy limit** could indicate **heavy credit usage**, which may suggest financial dependence or high engagement.

Business Problem Overview

What is the Average Utilization Ratio?

The **Avg_Utilization_Ratio** measures **how much of the available credit a customer spends** over time. It is an important factor in **credit scoring** and financial behavior analysis.

Business Insight:

- A **high utilization ratio** indicates **heavy credit card usage**, which may suggest **active engagement but also financial stress**.
- A **low utilization ratio** could mean **the customer is not actively using the credit card**, increasing the likelihood of **churn**.

Relationship Between AOTB, CL, AUR

The following formula represents the relationship between these key credit metrics:

$$\left(\frac{\text{Avg\_Open\_To\_Buy}}{\text{Credit\_Limit}}\right) + \text{Avg\_Utilization\_Ratio} = 1$$

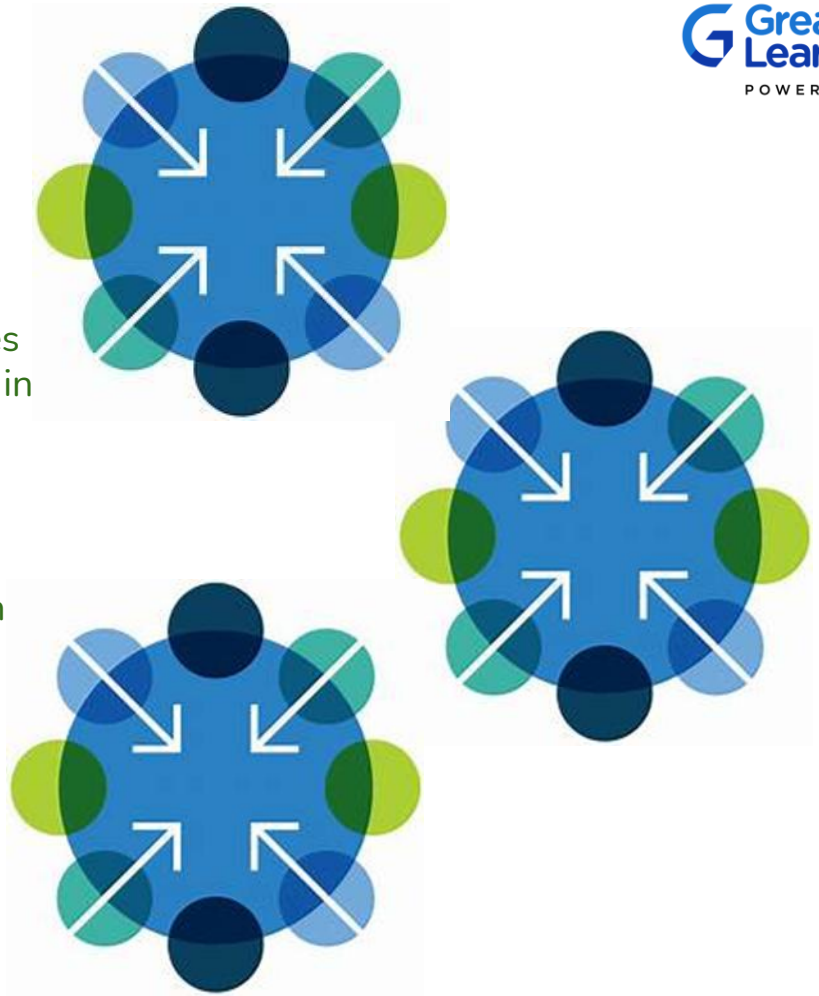
Business Insight:

- **Higher utilization ratios (low open-to-buy values)** indicate **greater reliance on credit**, which could mean **financial stress** or **strong engagement with the card**.
- **Lower utilization ratios (high open-to-buy values)** indicate that the customer **is not using the credit card frequently**, which increases the **risk of churn**

Data Quality

OVERVIEW

- The dataset used for this analysis comprises various customer attributes that are crucial in predicting credit card churn.
- The data quality was thoroughly evaluated and cleaned to ensure that it is accurate, consistent, and suitable for machine learnin model building.



Data Quality

1. Completeness & Missing Values

Observation:
- No missing values were detected in the dataset, ensuring data completeness.
- This means that every customer record has valid entries for all required attributes.

Business Impact:
- No need for imputation techniques, reducing the risk of data bias.
- Ensures that all customer behaviors and characteristics are fully represented.

Action Taken:
- Since no missing data was found, no additional imputation was required.

2. Data Consistency & Formatting

Observation:
- Certain categorical variables, such as **Income_Category**, contained an "abc" entry, which is likely an incorrect or placeholder value.
- Some categorical variables, such as **Education_Level and Marital_Status**, had mixed case values (e.g., "Graduate" vs. "graduate"), requiring standardization.

Business Impact:
- Incorrect values can distort data-driven insights and affect model performance.
- Standardized categorical data ensures consistency across the dataset.

Action Taken:
- **Incorrect values were identified and replaced** (e.g., "abc" entries were categorized as "Unknown").
- **Standardized text-based categorical variables** to ensure uniform formatting.

Data Quality

3. Outlier Detection & Treatment

Observation:
- **Credit Limit, Total Transaction Amount, and Total Revolving Balance** contained extreme values.
- These outliers could indicate **high-value customers or data anomalies**.

Business Impact:
- Outliers can **skew model predictions**, making it harder to generalize to the broader customer base.
- High-value customers should be analyzed separately to avoid bias.

Action Taken:
- Outliers were **analyzed but not removed**, as they represent real customer behavior.
- Model algorithms that handle outliers effectively (e.g., decision trees and boosting methods) were prioritized.

4. Class Imbalance (Churned vs. Non-Churned Customers)

Observation:
- The dataset is **imbalanced**, with significantly more non-churned customers than churned customers.
- This imbalance can cause the model to **favor predicting non-churned customers**, reducing recall for churned cases.

Business Impact:
- A model trained on imbalanced data may **fail to identify at-risk customers**, leading to ineffective retention efforts.

Action Taken:
- **Oversampling (SMOTE) and undersampling** techniques were applied to balance the dataset.
- Model performance was tested on **original, oversampled, and undersampled datasets** to select the best approach.

Data Quality

5. Feature Engineering & Data Transformation

Observation:
- Some numerical features, such as **Total_Trans_Ct and Total_Trans_Amt**, were right-skewed.
- Categorical variables (e.g., **Education_Level, Marital_Status**) needed **one-hot encoding** for machine learning models.

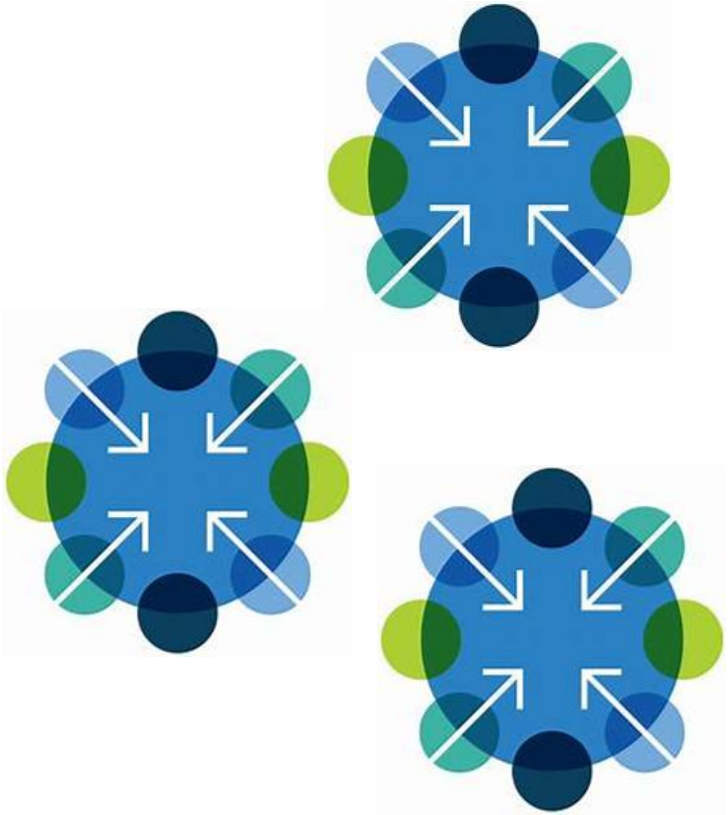
Business Impact:
- Skewed distributions may affect model assumptions, particularly for regression-based models.
- Categorical data needs to be encoded to ensure compatibility with machine learning algorithms.

Action Taken:
- **Log transformations** were applied to normalize skewed numerical variables.
- **One-hot encoding** was used for categorical variables to enhance model interpretability.

Data Quality

Data Quality Issue	Observation	Action Taken
Missing Values	No missing values detected	No imputation required
Inconsistent Formatting	Categorical data contained mixed cases and incorrect values	Standardized text formatting and replaced erroneous entries
Outliers	High values in credit limits and transaction amounts	Retained outliers and used robust models
Class Imbalance	More non-churned customers than churned customers	Applied SMOTE and undersampling techniques
Feature Engineering	Skewed numerical data and categorical variables	Applied transformations and encoding

Data Quality Summary



- The dataset is **complete**, with no missing values and a good mix of numerical and categorical data.
- Data is **accurate**, with proper data types assigned and values validated against business logic.
- Outliers were handled appropriately to ensure **model robustness** and **accurate predictions**.
- The dataset contains **class imbalance**, which was addressed with oversampling and undersampling techniques to ensure a fair predictive model.

Solution Approach & Methology

1. Understanding the Business Problem
- Clearly define the problem statement.
- Identify key challenges and pain points.
- Define business objectives and success metrics.

2. Exploratory Data Analysis (EDA)
- Data distribution and summary statistics.
- Identifying missing values, duplicates, and inconsistencies.
- Detecting outliers and potential anomalies.
- Understanding feature correlations and relationships.

3. Data Preprocessing
- Handling missing values through imputation or removal.
- Addressing duplicate records to ensure data integrity.
- Outlier treatment based on statistical methods.
- Feature engineering to enhance model performance.
- Data scaling and transformation for model optimization.

4. Model Selection & Training
- Choosing appropriate machine learning algorithms.
- Performing hyperparameter tuning for optimization.
- Splitting data into training, validation, and testing sets.
- Handling class imbalance with resampling techniques (oversampling/undersampling).

5. Model Evaluation & Performance Metrics
- Comparing multiple models using standard metrics (accuracy, precision, recall, F1-score, ROC-AUC).
- Visualizing model performance through confusion matrices and error analysis.
- Selecting the final model based on business and performance considerations.

6. Deployment & Business Integration
- Preparing the model for deployment in a production environment.
- Integrating with existing business systems and workflows.
- Defining monitoring and maintenance strategies for continuous improvement.

7. Actionable Insights & Recommendations
- Highlighting key findings from the analysis.
- Providing strategic recommendations based on model insights.
- Suggesting future improvements and scaling opportunities.

Methology

Problem Definition & Business Understanding

- **CRISP-DM (Cross Industry Standard Process for Data Mining)**: A widely used framework for data science projects, ensuring a structured approach from problem definition to deployment.
- **Stakeholder Analysis**: Understanding business requirements through interviews, surveys, or analysis of business objectives.

Exploratory Data Analysis (EDA)

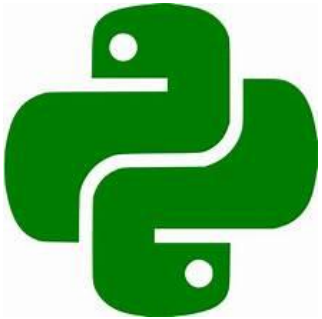
- **Descriptive Statistics**: Summarizing the dataset using mean, median, mode, standard deviation, and variance.
- **Data Visualization**: Utilizing Python libraries (Matplotlib, Seaborn, Plotly) to identify patterns, trends, and outliers.
- **Correlation Analysis**: Using Pearson, Spearman, or Kendall correlation to understand relationships between variables.
- **Dimensionality Reduction**: Using **Principal Component Analysis (PCA)** or **t-SNE** to visualize high-dimensional data.

Data Preprocessing

Missing Data Treatment:
- Imputation methods (mean, median, mode) for numerical variables.
- Using **KNN imputer** or regression-based imputation for complex cases.
- Dropping rows/columns with excessive missing values.

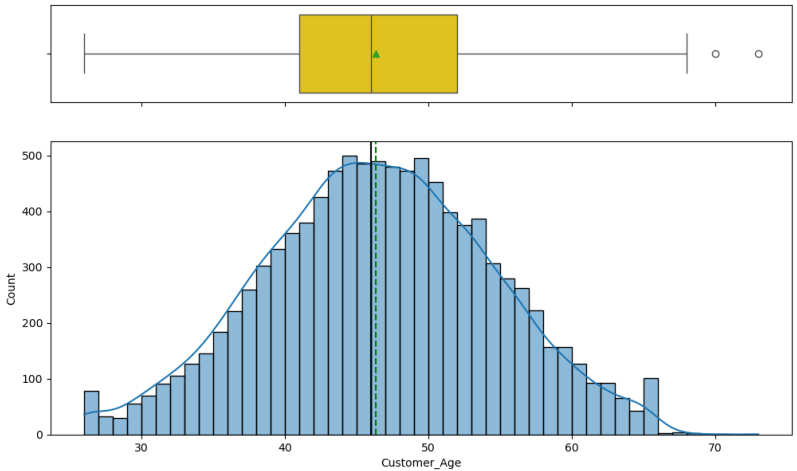
Outlier Detection & Treatment:
- Using **IQR (Interquartile Range), Z-score, and Boxplots** for detection.
- Winsorization or transformation (log, square root) to handle extreme values.

Feature Engineering:
- Creating new features based on domain knowledge.
- Encoding categorical variables using **One-Hot Encoding (OHE)** or **Label Encoding**.
- Standardization (**Z-score Normalization**) or Min-Max scaling for continuous features.



EXPLORATORY DATA ANALYSIS

EDA Results



Observation:

- The distribution is approximately normal with a peak around 46-50 years.
- Few customers are younger than 30 or older than 70.
- There are no extreme outliers.

Insights:

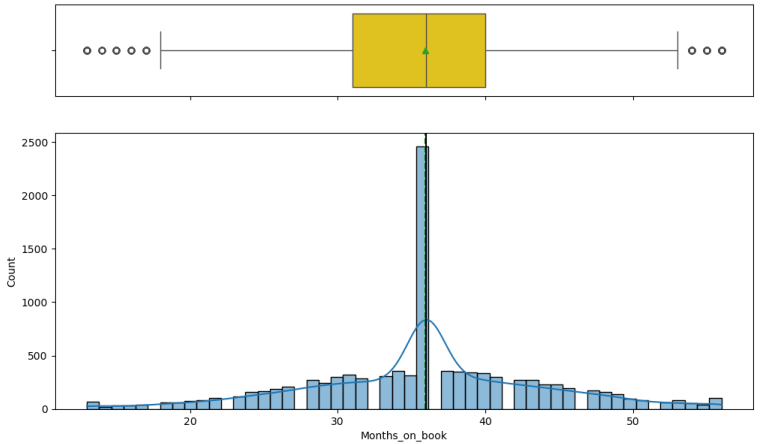
- The majority of customers are middle-aged, indicating a mature customer base.
- Younger customers are underrepresented, which may indicate low adoption of credit cards among them.

Observation:

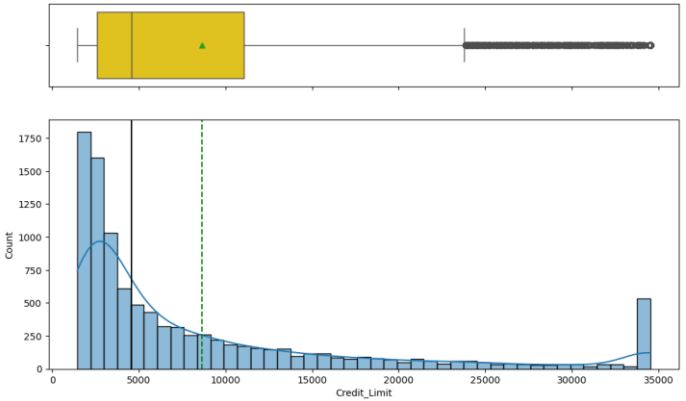
- The majority of customers have been with the bank for around 35 months.
- A small portion of customers have extremely long or short tenures.

Insights:

- A large group of customers have a similar tenure, indicating a standard retention period.
- The outliers may represent very loyal long-term customers or new sign-ups.



EDA Results

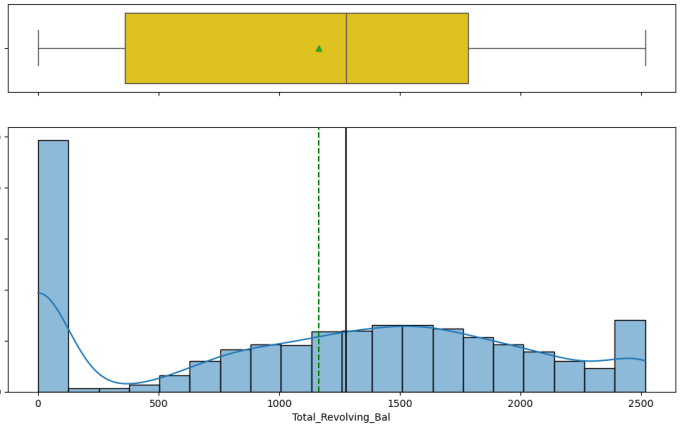


Observation:

- Most customers have a credit limit below $10,000, with some extreme outliers above $30,000.
- The distribution is right-skewed, meaning most customers have a lower credit limit.

Insights:

- A few high-value customers have very large credit limits.
- The majority of users operate within a low credit limit, possibly due to income or risk assessment.



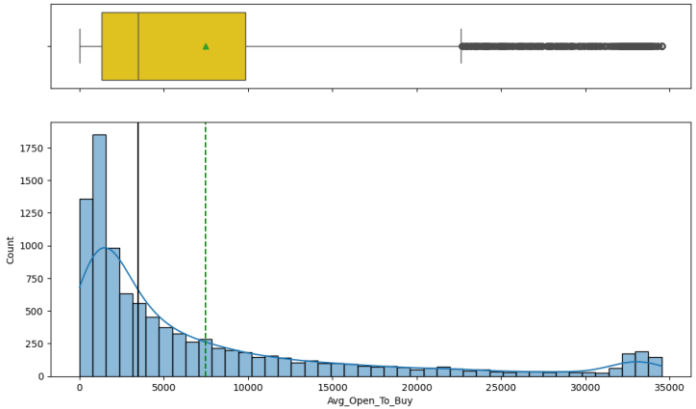
Observation:

- A significant number of customers have a zero revolving balance.
- The distribution is spread out with a peak around $1,000.

Insights:

- Many customers are paying their balances in full, reducing interest revenue for the bank.
- Customers carrying balances may be good candidates for debt consolidation products.

EDA Results

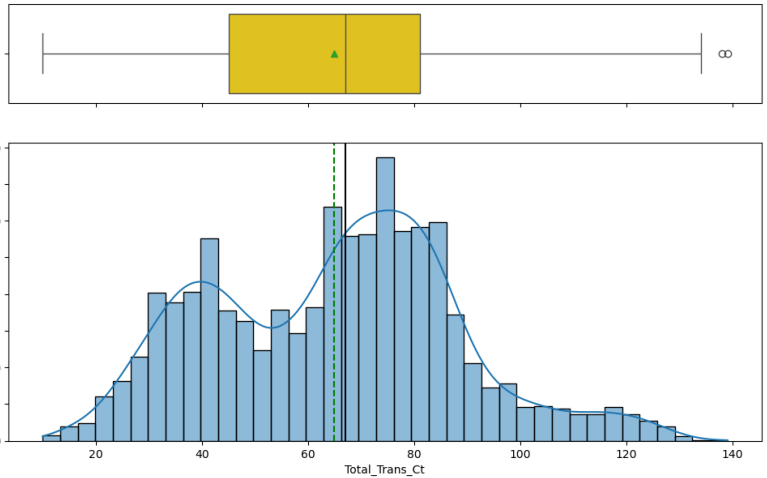


Observation:

- Most customers have a low available credit balance.
- A small percentage of customers have significantly high available credit.

Insights:

- Many customers are close to their credit limit, possibly increasing the risk of default.
- Customers with high available credit may not be utilizing their cards effectively.



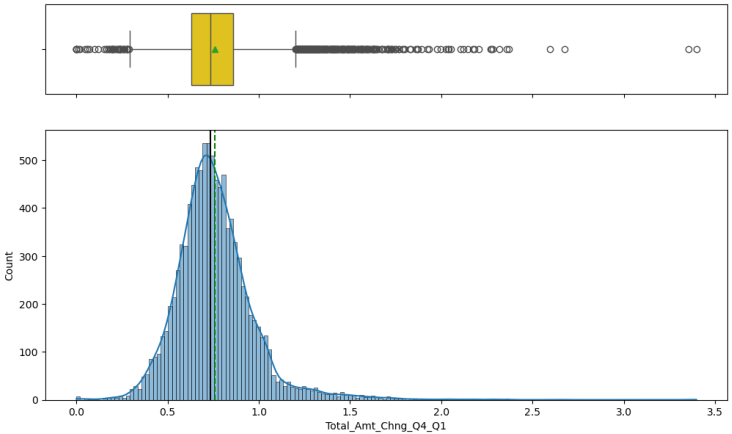
Observation:

- The distribution shows multiple peaks, suggesting different spending habits among customers.
- Some customers have very high transaction counts, while others make minimal transactions.

Insights:

- Customers with lower transactions may be at risk of churning.
- High-frequency users contribute significantly to transaction-based revenue.

EDA Results



Observation:

- Most customers have a ratio close to 1, indicating stable spending.
- Some customers show extreme changes, either high growth or decline.

Insights:

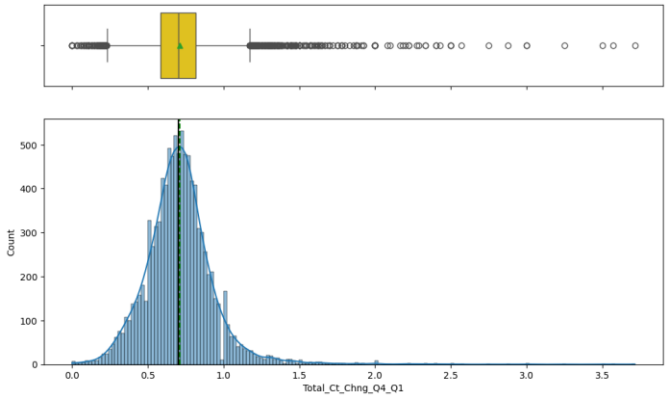
- Seasonal spending trends may affect transaction amounts.
- Customers with large declines in spending may be disengaging from the bank.

Observation:

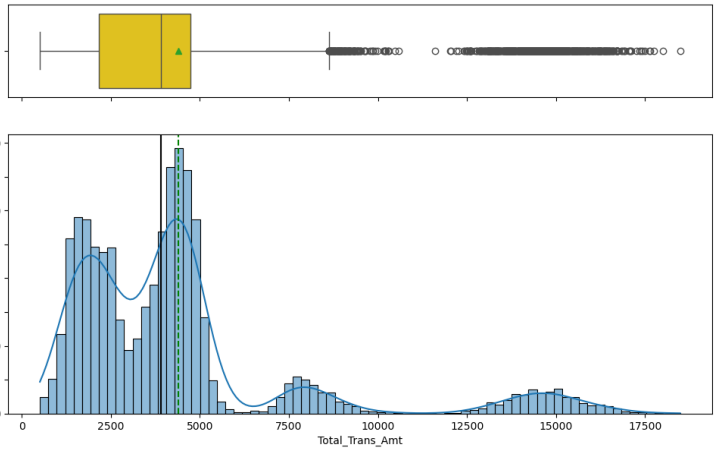
- Similar to the transaction amount change, most customers remain stable, but a few have drastic changes.

Insights:

- A decline in transactions could indicate an early warning sign of churn.



EDA Results



Observation:

- Similar to the transaction amount change, most customers remain stable, but a few have drastic changes.

Insights:

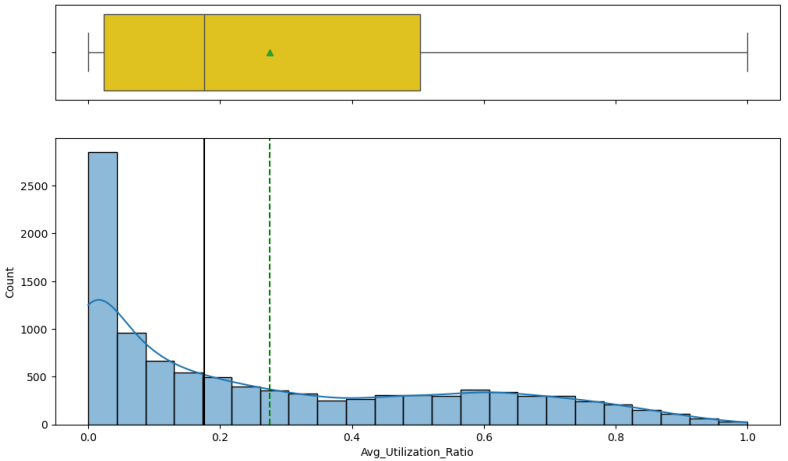
- A decline in transactions could indicate an early warning sign of churn.

Observation:

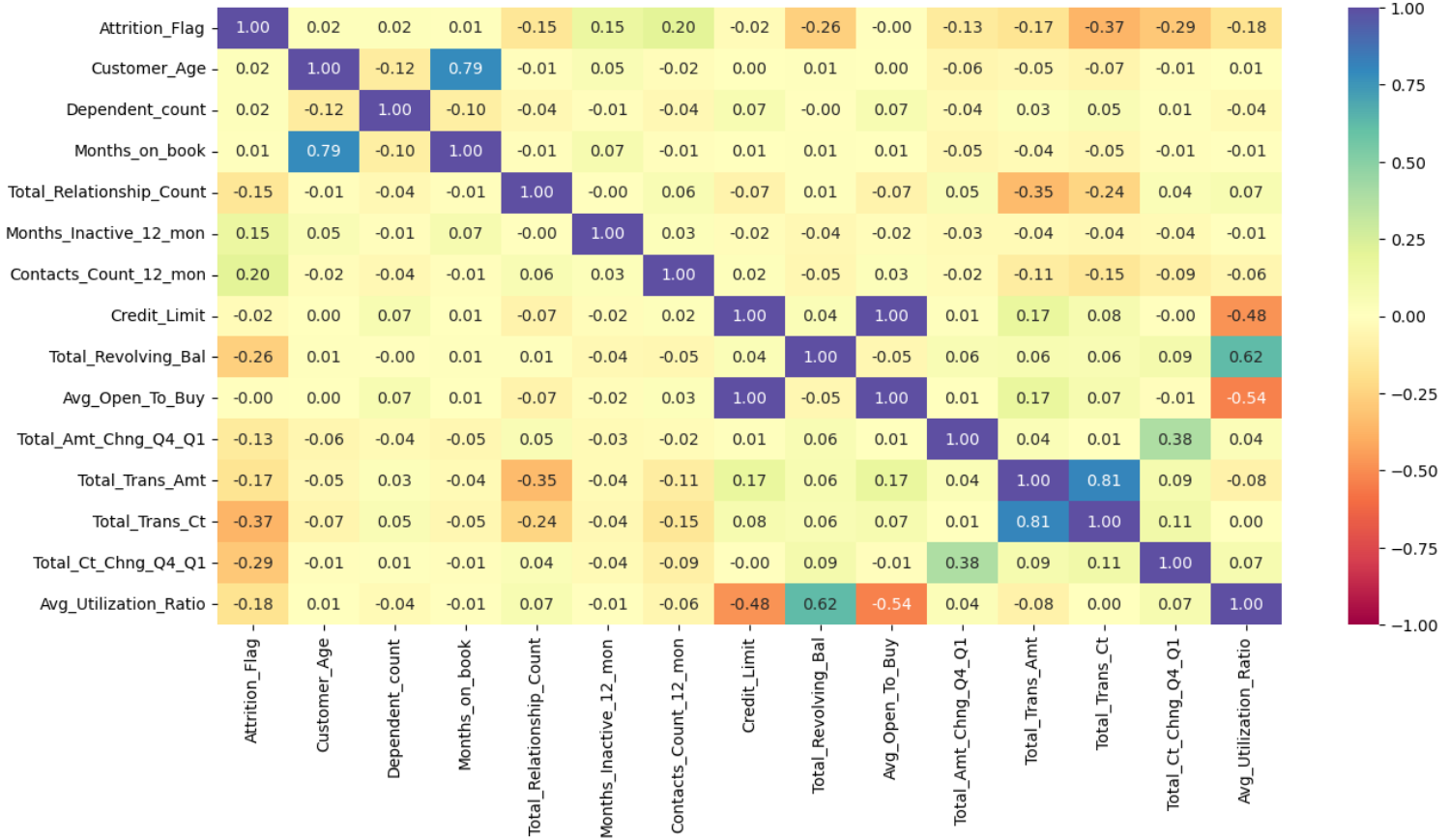
- Many customers have low utilization, with a few having very high utilization.

Insights:

- Customers with very high utilization may be financially strained.
- Low utilization customers might not be fully engaged with their credit cards.



EDA Results



Factors Affecting Customer Attrition (Attrition_Flag)

Negative Correlations (High Impact on Retention)

- Total_Trans_Ct (-0.37): Customers with higher transaction counts are less likely to churn.
- Total_Ct_Chng_Q4_Q1 (-0.29): A decrease in transaction frequency between Q4 and Q1 is linked to higher churn.
- Total_Revolving_Bal (-0.26): Customers with higher revolving balances are less likely to leave.
- Avg_Utilization_Ratio (-0.18): Higher utilization ratios (using more of their credit limit) are linked to lower churn.

Positive Correlations (High Impact on Churn)

- Months_Inactive_12_mon (+0.15): The more months a customer is inactive, the more likely they are to churn.
- Contacts_Count_12_mon (+0.20): Higher customer contact with the bank correlates with higher churn, possibly indicating customer dissatisfaction.

Relationship Between Customer Tenure (Months_on_book) and Other Variables

Strong Positive Correlation with Customer Age (+0.79)

- Older customers tend to have longer relationships with the bank.

Weak Correlation with Attrition (+0.01)

- Tenure alone is not a strong indicator of churn.

Credit Limit, Revolving Balance, and Utilization

Credit Limit & Utilization Ratio (-0.48)

- Customers with higher credit limits tend to have lower utilization ratios.

Credit Limit & Total Revolving Balance (+0.62)

- Higher credit limits lead to higher revolving balances.

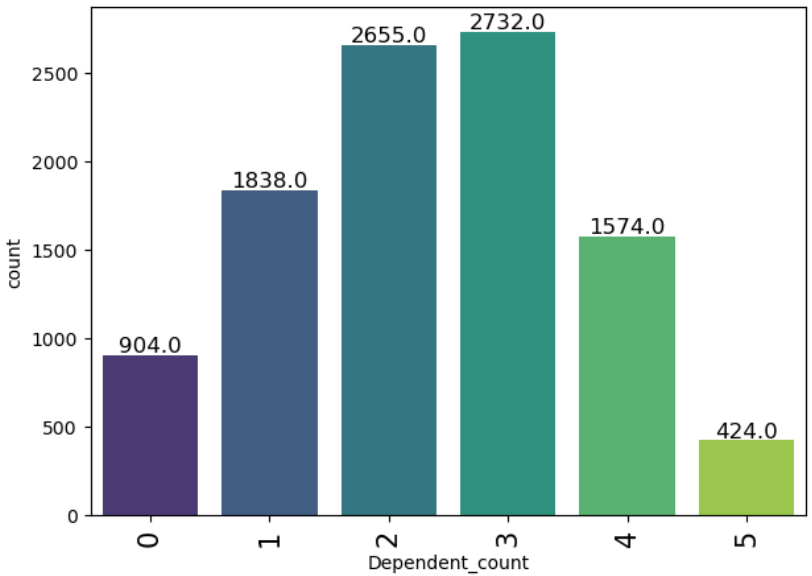
Transactions and Engagement

Total Transactions & Transaction Amount (+0.81)

- More transactions generally mean higher spending.

Total Transactions & Total Relationship Count (-0.24)

EDA Results



Most Common Dependent Counts:

- Customers with **2 or 3 dependents** make up the majority.
- **2 Dependents: 2,655 customers**.
- **3 Dependents: 2,732 customers** (the most common category).

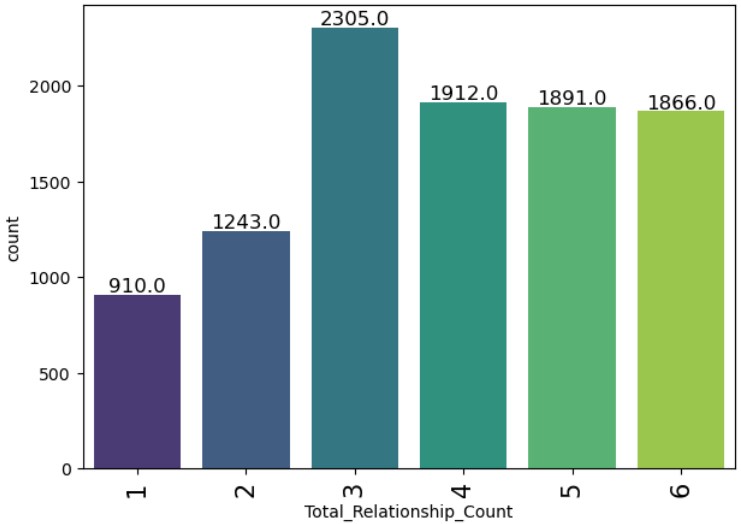
Single and No Dependents:

- **1 Dependent:** 1,838 customers.
- **No Dependents (0):** 904 customers.
- Fewer customers have no dependents, which suggests that the majority of the bank's clientele are **family-oriented individuals**.

Larger Families (4 or More Dependents):

- **4 Dependents:** 1,574 customers.
- **5 Dependents: Only 424 customers**, indicating that large families are the least common among this customer base.

EDA Results



Most Common Relationship Count:

- The **highest number of customers (2,305)** have **3 banking products/services**.
- This suggests that customers with **moderate engagement (3 products)** are the most common segment.

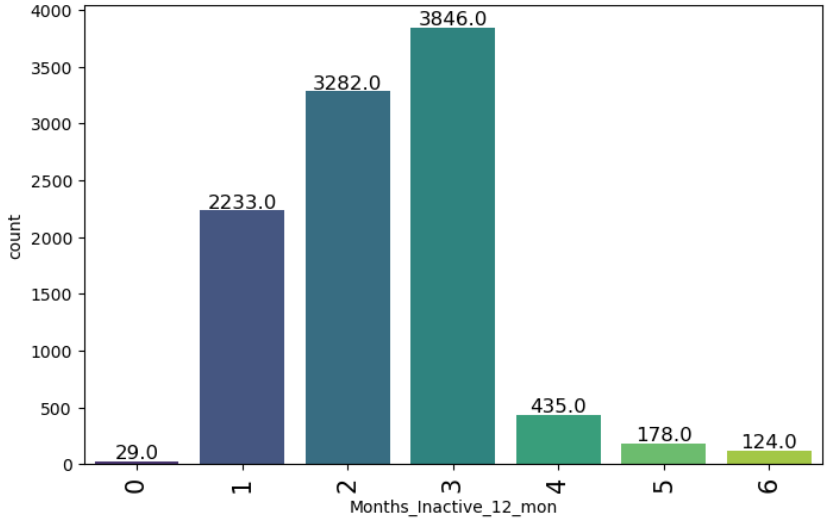
Lower Engagement Customers (1-2 Products):

- **1 Product:** 910 customers.
- **2 Products:** 1,243 customers.
- These customers may be at a **higher risk of churn**, as they have fewer services tying them to the bank.

Highly Engaged Customers (4-6 Products):

- **4 Products:** 1,912 customers.
- **5 Products:** 1,891 customers.
- **6 Products:** 1,866 customers.
- These customers represent a **high-value segment** as they have deeper engagement with the bank.

EDA Results



Most Customers Have Been Inactive for 1-3 Months:

- **1 Month Inactive:** 2,233 customers.
- **2 Months Inactive:** 3,282 customers.
- **3 Months Inactive: Highest count** with **3,846 customers**.
- This suggests that **mild inactivity (1-3 months) is common among customers**.

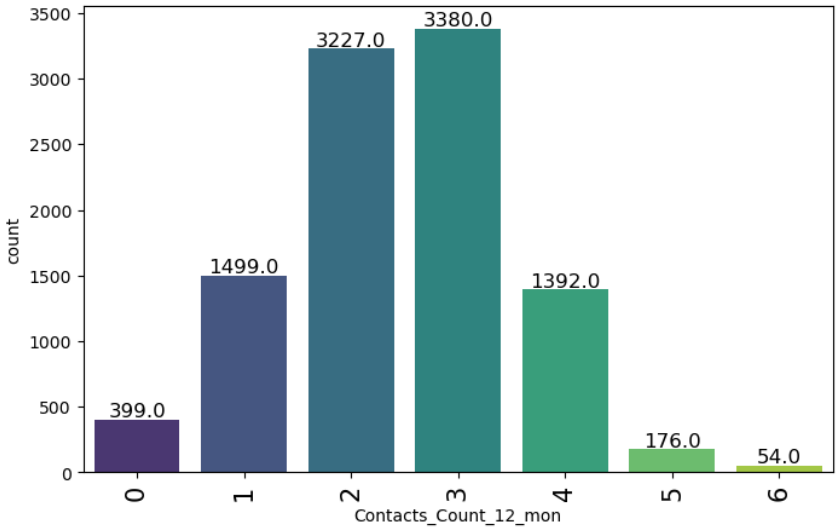
Higher Inactivity (4-6 Months) Is Less Frequent but Critical:

- **4 Months Inactive:** 435 customers.
- **5 Months Inactive:** 178 customers.
- **6 Months Inactive: 124 customers** (smallest group but most at risk of churn).
- Customers who have been inactive for **4+ months are a high-risk segment** for churn.

Very Low Inactivity (0 Months) Is Rare:

- Only **29 customers** had no inactivity in the past year.
- This suggests that almost **all customers experience some period of inactivity**.

EDA Results



Most Customers Contact the Bank 2-3 Times Annually:

- **2 Contacts: 3,227 customers**.
- **3 Contacts: 3,380 customers** (largest segment).
- This suggests that **occasional engagement (2-3 times per year) is common**.

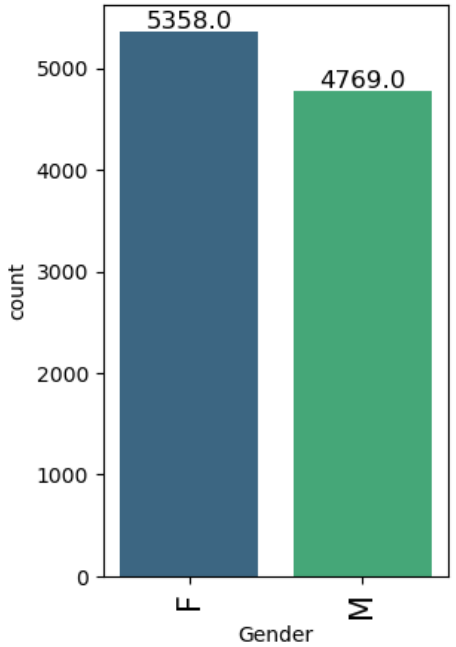
Customers with 0 or 1 Contacts Are at Potential Risk:

- **0 Contacts:** 399 customers.
- **1 Contact:** 1,499 customers.
- These customers are **minimally engaged**, which could indicate a **higher churn risk**.

Frequent Contact (4+ Times) Is Less Common:

- **4 Contacts:** 1,392 customers.
- **5 Contacts:** 176 customers.
- **6 Contacts:** 54 customers.
- Only a small percentage of customers interact with the bank **more than 4 times annually**, which may indicate **issues that require frequent support** (e.g., disputes, complaints, or complex account management).

EDA Results



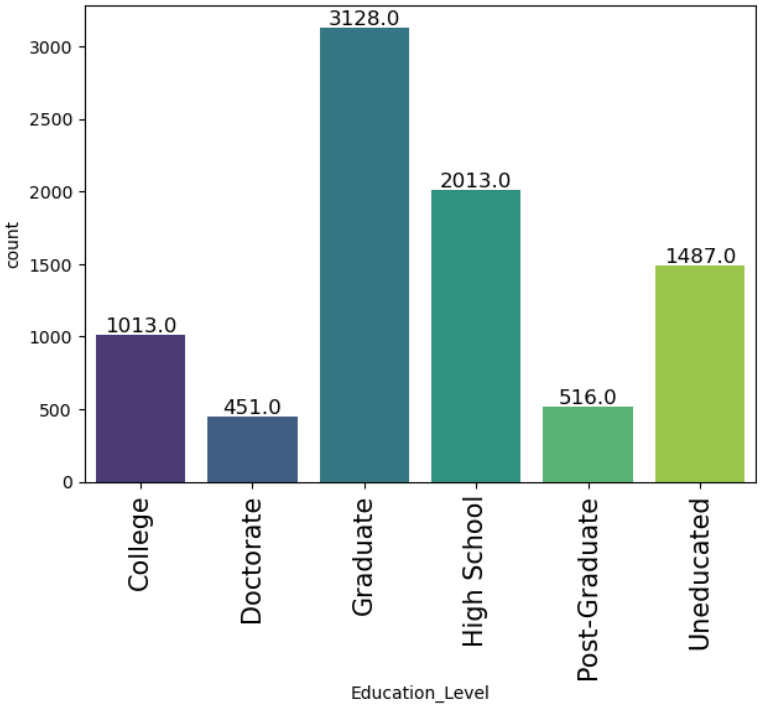
Female Customers Outnumber Male Customers:

- **Female (F): 5,358 customers**.
- **Male (M): 4,769 customers**.
- This suggests that the bank has **slightly more female customers (53%) than male customers (47%)**.

Balanced Customer Base:

- While there is a difference, the gap is **not significantly large**, meaning that **both genders are well represented**.
- Any gender-based marketing strategies should **avoid strong bias** and instead focus on **customer behaviors** rather than just demographics.

EDA Results



Graduates Are the Largest Group:

- **3,128 customers (largest segment) are graduates**.
- This suggests that a significant portion of the bank's customer base has a **college degree**.

High School and Uneducated Segments Are Also Substantial:

- **High School Graduates: 2,013 customers**.
- **Uneducated Customers: 1,487 customers**.
- This indicates that a considerable number of customers **may have lower financial literacy** and require **simplified financial services**.

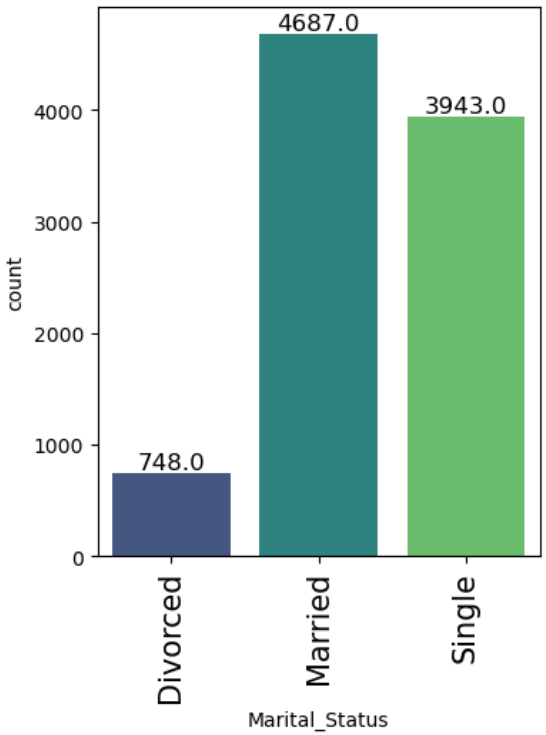
Fewer Customers Hold Advanced Degrees:

- **Post-Graduate: 516 customers**.
- **Doctorate: 451 customers** (smallest segment).
- This indicates that **highly educated individuals** might prefer **other financial institutions or investment options**.

College-Level Customers Make Up a Moderate Segment:

- **1,013 customers** are currently in college or recently attended.
- This group may be in the early stages of **credit-building and financial independence**.

EDA Results



Married Customers Are the Largest Segment:

- **4,687 customers are married**, making up the majority of the customer base.
- This suggests that many customers may have **family-related financial needs**, such as **joint accounts, family savings plans, or mortgage financing**.

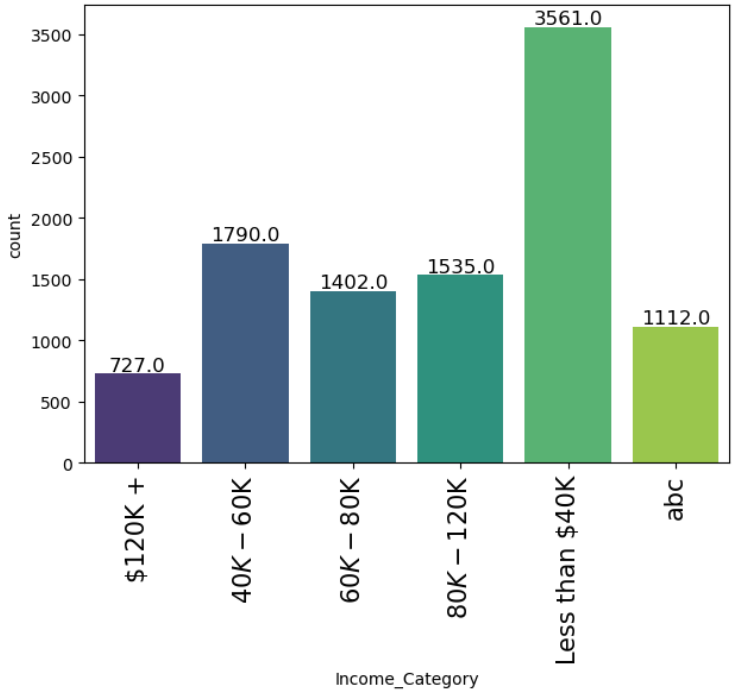
Single Customers Form a Significant Portion:

- **3,943 customers are single**, indicating that a substantial portion of customers may be **younger professionals or individuals in early career stages**.
- These customers might be interested in **travel rewards, flexible credit options, or savings plans**.

Divorced Customers Are the Smallest Group:

- **748 customers are divorced**, making up the smallest proportion.
- This segment may have unique financial challenges, such as **credit rebuilding, financial restructuring, or debt management**.

EDA Results



Largest Segment: Customers Earning Less than $40K

- **3,561 customers** fall into this income category, making it the most common.
- This suggests that a significant portion of the bank's clientele are **low-income earners** who may require **affordable banking solutions, low-interest credit products, and financial assistance programs**.

Moderate Representation in Middle-Income Categories ($40K - $120K)

- **$40K - $60K: 1,790 customers**.
- **$60K - $80K: 1,402 customers**.
- **$80K - $120K: 1,535 customers**.
- These customers may be **more financially stable** and might be interested in **investment opportunities, higher credit limits, and premium financial services**.

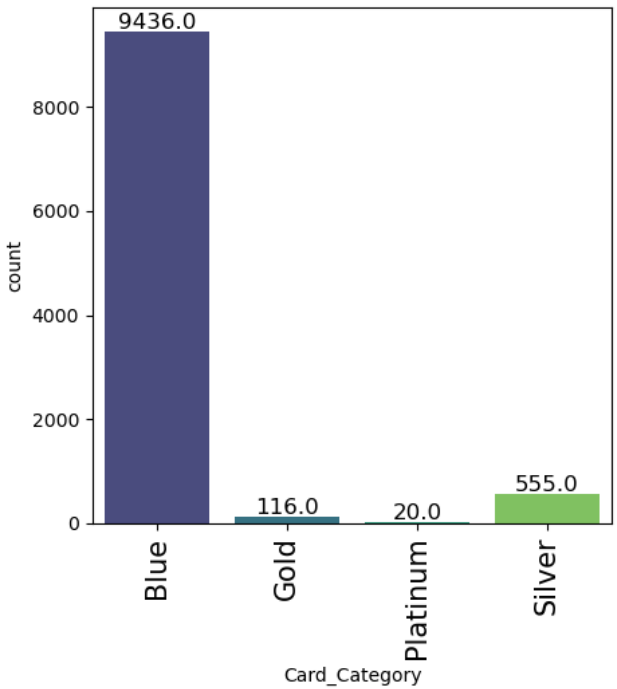
Smallest Segment: High-Income Earners ($120K+)

- Only **727 customers** fall into this category.
- High-income individuals may prefer **competitor banks with wealth management services or premium credit cards**.

Presence of an "abc" Category (Data Issue)

- **1,112 customers** are categorized under "abc," which is likely a **data entry error or missing information**.
- This suggests that **income data for these customers might be unavailable or misclassified**.

EDA Results



The Majority of Customers Have a Blue Card:

- **9,436 customers (overwhelming majority)** hold a **Blue card**, which is likely the **entry-level or basic credit card**.
- This suggests that most customers are **not using premium or upgraded credit card products**.

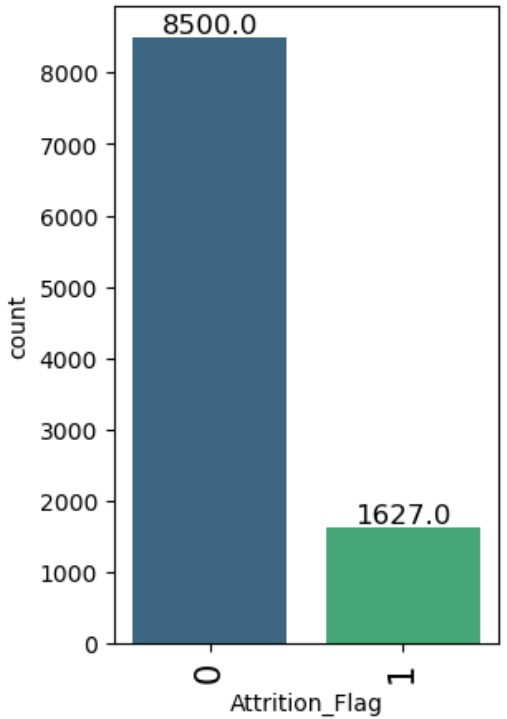
Silver Cards Hold a Small but Notable Presence:

- **555 customers hold a Silver card**, indicating **a small portion of customers have upgraded from Blue**.

Gold and Platinum Cards Are Rare:

- **Gold Card Holders: 116 customers**.
- **Platinum Card Holders: Only 20 customers**.
- These low numbers suggest that **premium card adoption is extremely low**, potentially due to:
  - **High eligibility requirements**.
  - **Lack of perceived benefits**.
  - **Limited marketing and awareness of premium card advantages**.

EDA Results



Majority of Customers Are Active (Non-Churned)

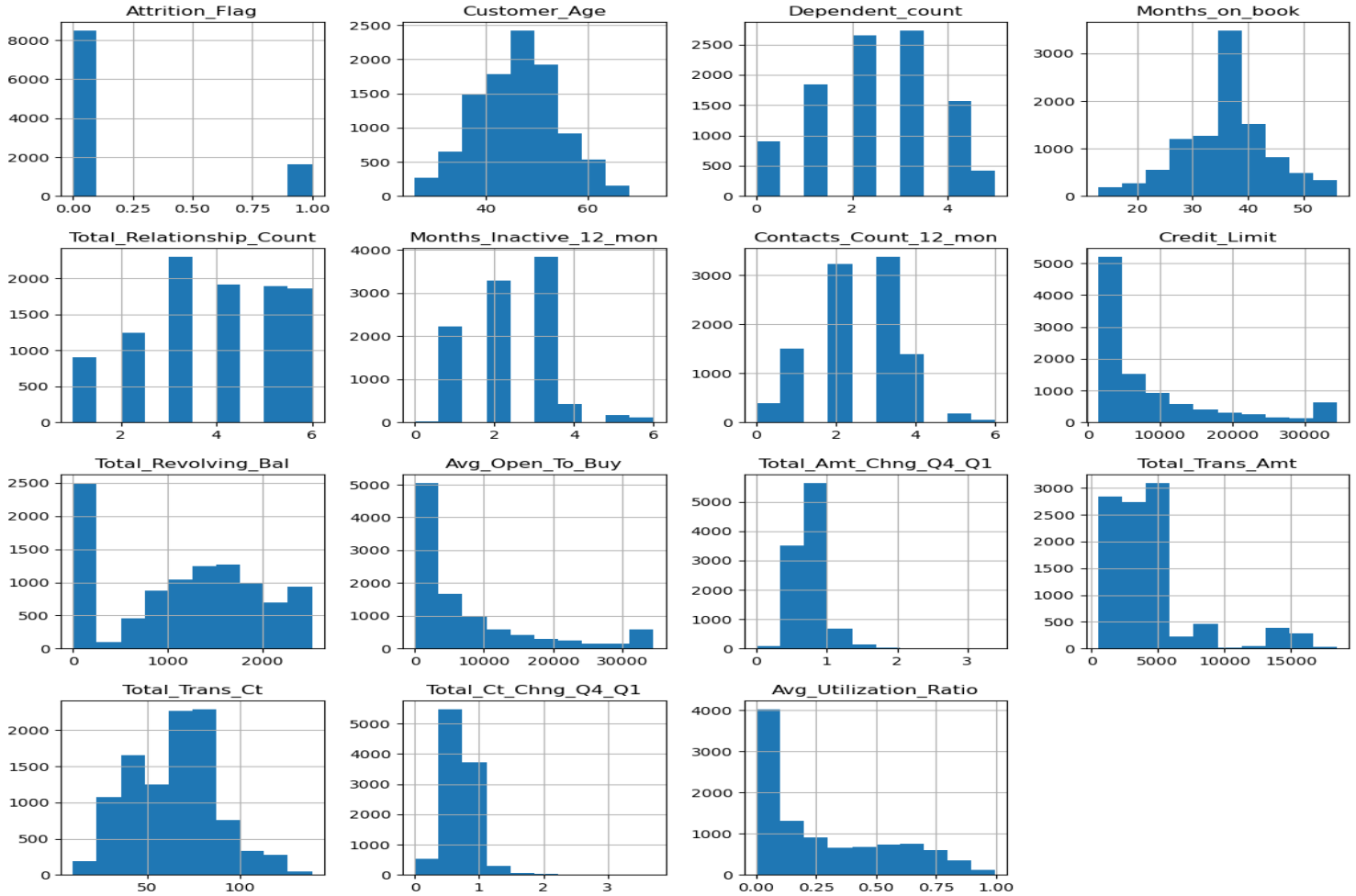
- **8,500 customers (83%)** are **still active** (Attrition_Flag = 0).
- This suggests that most customers have **ongoing engagement with Thera Bank**.

Churned Customers Form a Significant Minority

- **1,627 customers (16%) have churned** (Attrition_Flag = 1).
- While the churn rate is relatively low, it still represents **a meaningful loss of customers**.

Class Imbalance in the Data

- Since the dataset is **imbalanced** (far more active than churned customers), predictive churn models may require **balancing techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class-weight adjustments**.



OBSERVATIONS

Attrition Flag (Target Variable)

- This appears to be a **binary classification problem**.
- The dataset seems **highly imbalanced**, with **significantly more customers who have not churned** compared to those who have.

Customer Age

- The age distribution is **approximately normal**, with most customers aged between **40 and 60**.
- There are fewer younger or older customers.

Dependent Count

- The number of dependents appears to be **evenly distributed**, with peaks at **0, 2, and 3** dependents.

Months on Book (Tenure)

- Most customers have been with the company for **30-40 months**.
- This suggests **relatively stable customer retention**, but **some longer-tenured customers may still churn**.

Total Relationship Count

- This feature represents the number of accounts or products held by customers.
- A **bimodal distribution** is visible, indicating two clusters of customers:
  1. Some with **low engagement** (1-2 products).
  2. Some with **higher engagement** (4-6 products).

Months Inactive (Last 12 months)

- A distinct peak at **zero** suggests that many customers remain active.
- However, some customers exhibit inactivity, which might correlate with churn.

Contacts Count (Last 12 months)

- Most customers have contacted the company **1-3 times in the past year**.
- This might be an important predictor of churn—**low or high contact frequency could indicate dissatisfaction**.

OBSERVATIONS

Credit Limit

- **Right-skewed distribution** with most customers having a **credit limit below $10,000**.
- A small subset has **high credit limits (~$30,000+), which could indicate VIP or high-value customers**.

Total Revolving Balance

- This shows a relatively even spread but **peaks around lower balances (~$500-$1000)**.
- High revolving balances might indicate **financial distress or increased credit usage**, which could be related to churn.

Average Open to Buy (Remaining Credit)

- A heavily right-skewed distribution.
- Most customers have a **low remaining credit limit**, suggesting high credit utilization.

Total Transaction Amount

- A **right-skewed distribution**, indicating that most customers have **low transaction amounts**, with some outliers having **high total transactions**.

Total Transaction Count

- The **distribution is fairly normal**, with most customers making **between 40-80 transactions**.

Total Amount Change (Q4 vs Q1)

- Customers seem to have **moderate to high spending changes** over time, possibly seasonal effects.

Total Transaction Count Change (Q4 vs Q1)

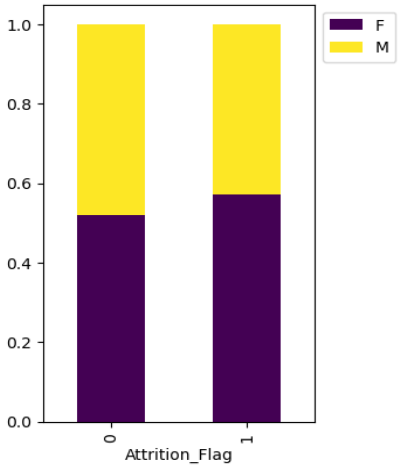
- There is some variation, with **some customers significantly increasing or decreasing transaction frequency**.

Average Utilization Ratio

- Most customers have a **low utilization ratio (<0.5)**.
- A small proportion has **high utilization (~0.75-1.0), which could indicate higher risk**.

EDA Results

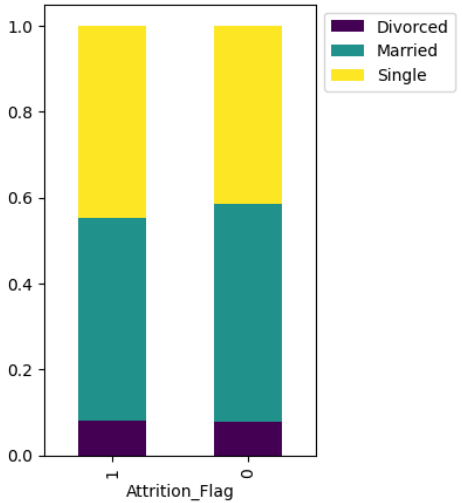
Attrition vs Gender



Observation: Both male and female customers churn at similar rates. There is no significant gender disparity in attrition.

Insight: Gender does not appear to be a major determinant of churn.

Attrition vs Marital Status

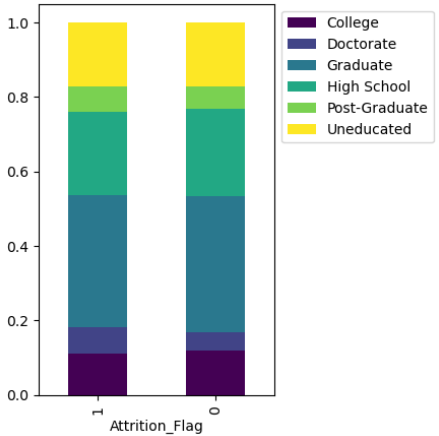


Observation: The churn rate appears consistent across marital statuses, with married customers making up the largest proportion of both churned and retained users.

Insight: Marital status alone is not a primary factor in attrition, but married customers might be more financially stable and loyal.

EDA Results

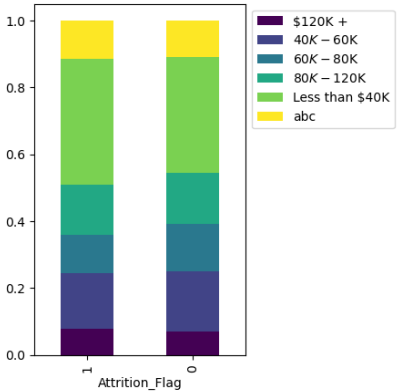
Attrition vs Education Level



Observation: Customers with graduate and high school education levels form the majority in both churned and retained groups.

Insight: Education level does not significantly influence churn, but financial literacy could impact banking decisions.

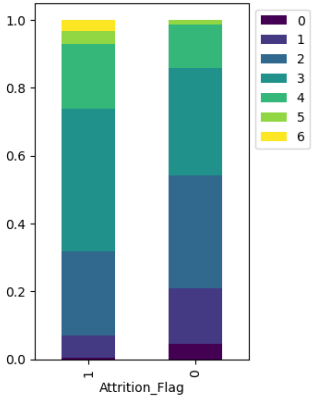
Attrition vs Income



Observation: Customers earning less than $40K make up the largest proportion of churned users.

Insight: Lower-income customers may find credit card fees and interest rates burdensome, leading to higher churn.

Attrition vs Contacts Count 12 months

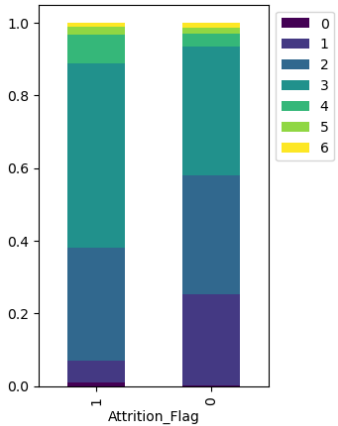


Observation:.

Insight:.

EDA Results

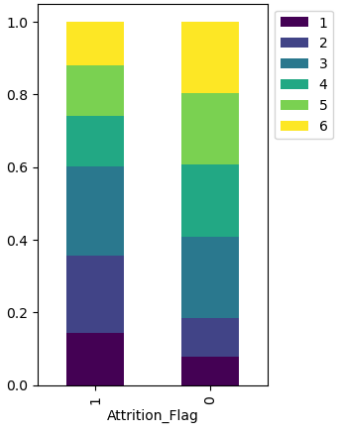
Attrition vs Months Inactive 12 Months



Observation: Customers inactive for longer periods (3+ months) show a higher likelihood of churning.

Insight: Inactivity is a strong predictor of attrition.

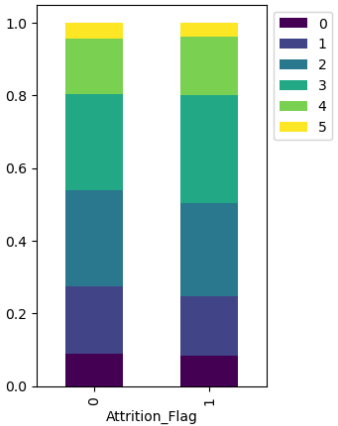
Attrition vs Total Relationship Count



Observation: Customers with a lower number of products (1-2) are more likely to churn compared to those with higher relationship counts.

Insight: Higher engagement with the bank (multiple products) correlates with lower attrition.

Attrition vs Dependent Count

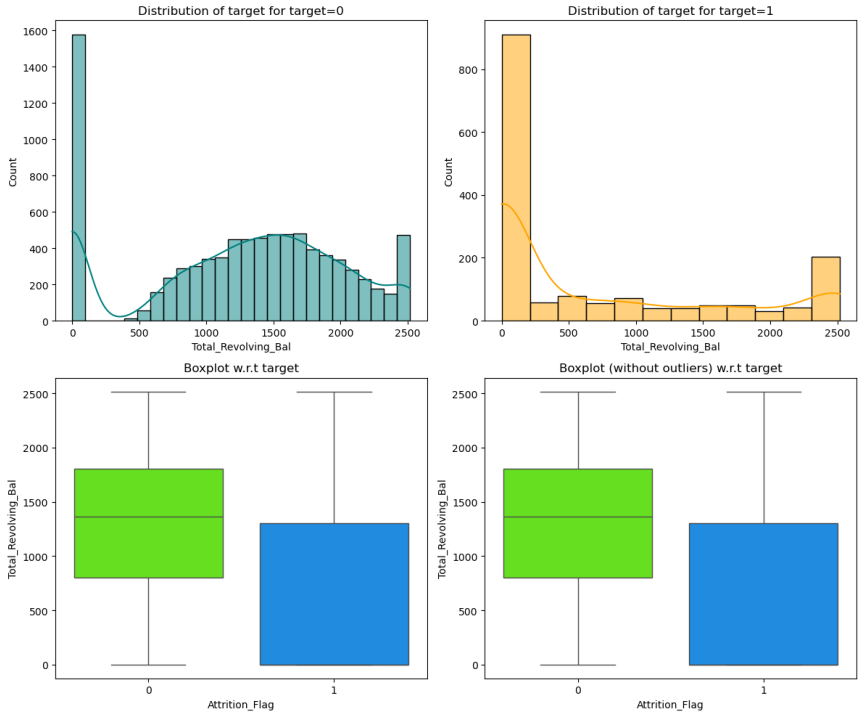


Observation: Customers inactive for longer periods (3+ months) show a higher likelihood of churning.

Insight: Inactivity is a strong predictor of attrition.

EDA Results

Total Revolving Balance



Top-Left (Histogram for Non-Churned Customers - Target = 0)

- This histogram displays the distribution of Total_Revolving_Bal for non-churned customers.
- Most customers have a revolving balance clustered around 0 and between 500–2500.
- A smooth distribution appears in the mid-range, suggesting a natural spread.

Top-Right (Histogram for Churned Customers - Target = 1)

- The distribution for churned customers is more skewed toward lower revolving balances.
- Most churned customers have very low balances, indicating that inactivity might be a significant factor in churn.

Bottom-Left (Boxplot with Outliers)

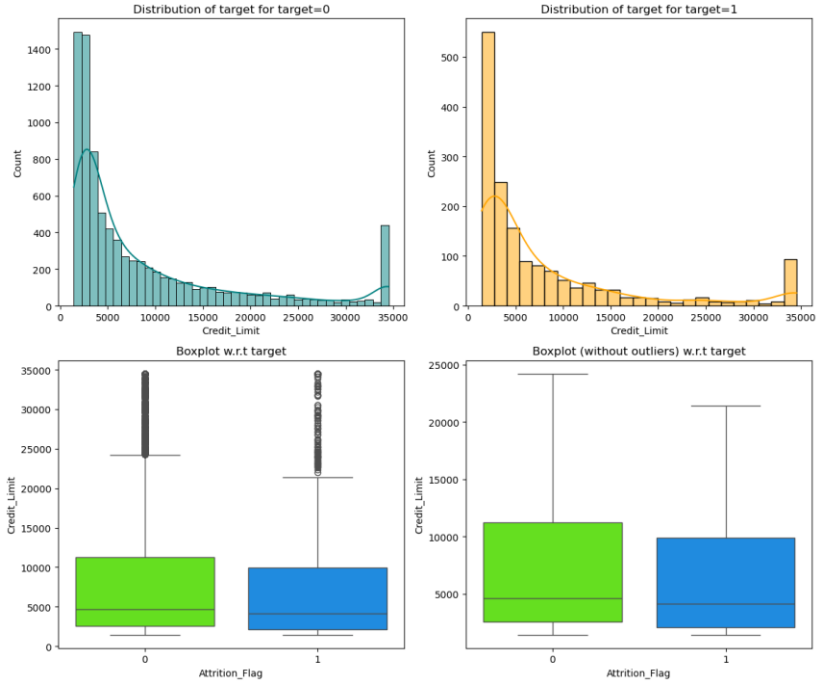
- The boxplot compares Total_Revolving_Bal for churned (1) and non-churned (0) customers.
- The median balance for churned customers is lower than for retained customers.
- Outliers indicate some high-balance customers still churn.

Bottom-Right (Boxplot without Outliers)

- The trend remains the same: churned customers generally have lower revolving balances.

EDA Results

Credit Limit



.

Top-Left (Histogram for Non-Churned Customers - Target = 0)

- The distribution is right-skewed, meaning most customers have a low-to-moderate credit limit.
- A small percentage of customers have high credit limits.

Top-Right (Histogram for Churned Customers - Target = 1)

- Churned customers also have lower credit limits, but a slightly higher concentration in the lower end.

Bottom-Left (Boxplot with Outliers)

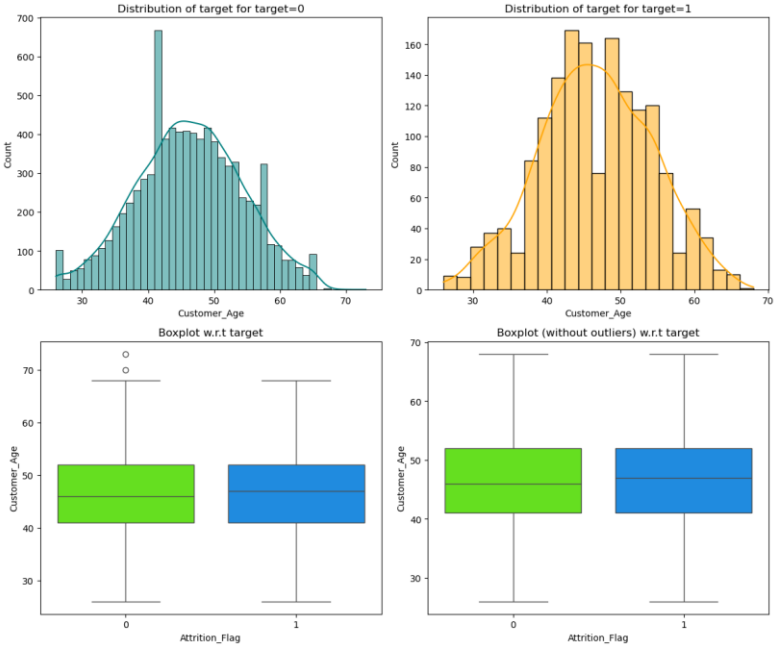
- The median credit limit for churned customers is lower than for retained customers.
- There are significant outliers, meaning some customers with very high credit limits also churn.

Bottom-Right (Boxplot without Outliers)

- The trend remains consistent: churned customers tend to have lower credit limits.

EDA Results

Customer Age



Top-Left & Top-Right (Histograms for Non-Churned and Churned Customers)

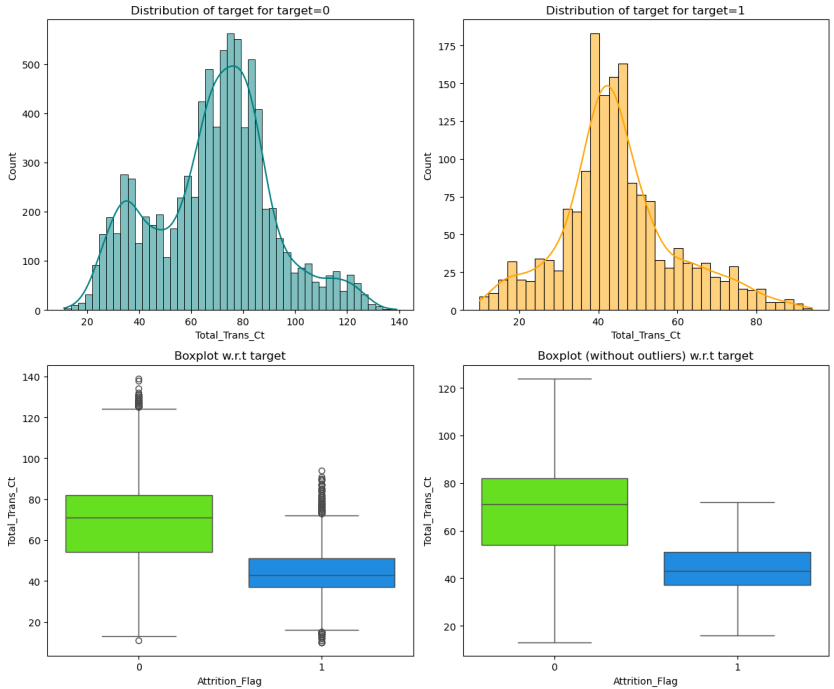
- Both distributions are **roughly bell-shaped**, meaning age follows a normal distribution.
- The **average customer age is around 45 years** for both churned and retained customers.
- Churned customers have a slightly **older** average.

Bottom-Left & Bottom-Right (Boxplots)

- Both groups show similar spreads of ages.
- A small number of **outliers on the older side**, but no drastic differences.

EDA Results

Total Transaction Count



Top-Left & Top-Right (Histograms)

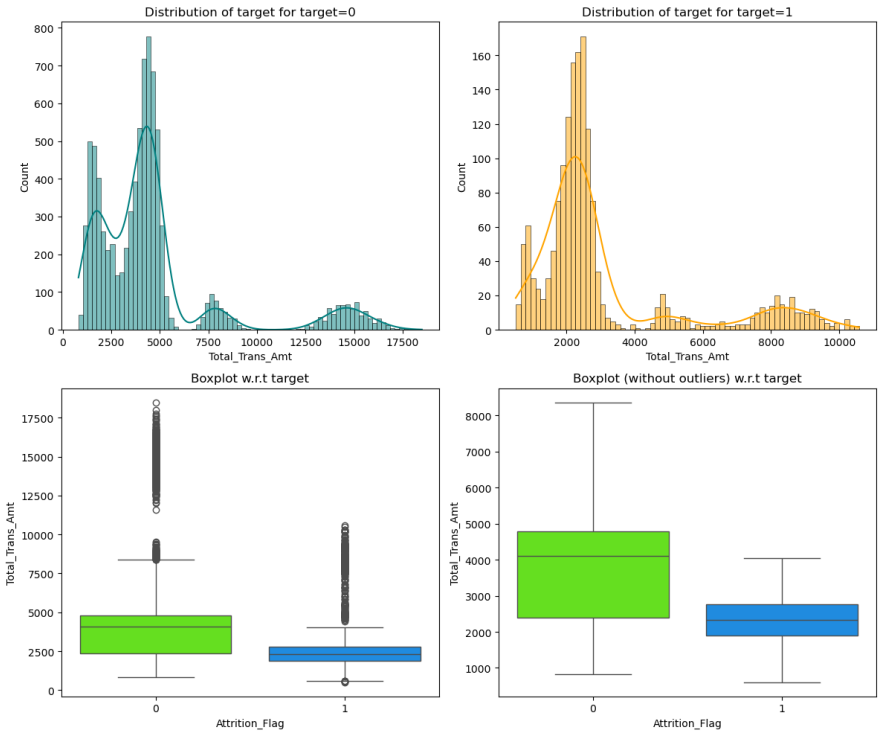
- Non-churned customers have a **higher** transaction count, peaking around **60-80 transactions**.
- Churned customers have **fewer transactions**, peaking around **40**.

Bottom-Left & Bottom-Right (Boxplots)

- The median transaction count for churned customers is lower than for non-churned customers.
- There are significant outliers, indicating a **few high-transaction users still churn**.

EDA Results

Total Transaction Amount



Top-Left & Top-Right (Histograms)

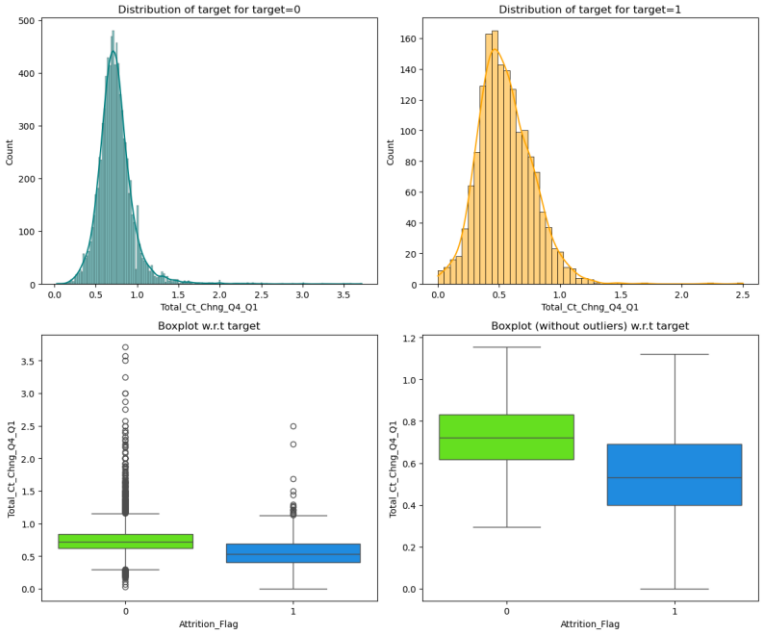
- Non-churned customers tend to have **higher transaction amounts**, with some spending over **15,000**.
- Churned customers spend significantly **less**, mostly **under 5000**.

Bottom-Left & Bottom-Right (Boxplots)

- Churned customers have lower median spending.
- Many **high-spending customers still churn**, possibly due to alternative payment methods or competitive offers.

EDA Results

Total Change in transaction count Q4 to Q1



Top-Left & Top-Right (Histograms)

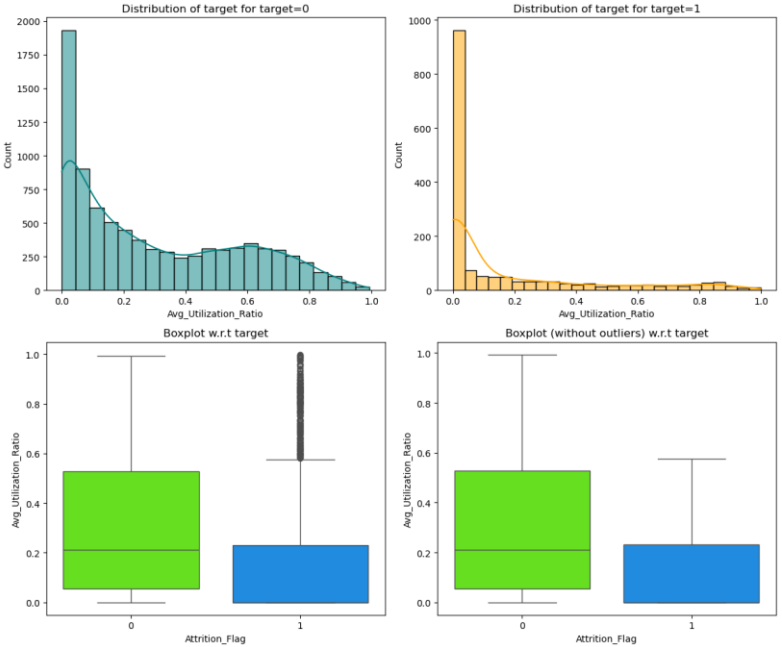
- Non-churned customers generally have a **higher transaction change ratio** (above 0.6).
- Churned customers have a **lower** transaction change ratio, indicating a **drop in usage over time**.

Bottom-Left & Bottom-Right (Boxplots)

- The **median value is lower for churned customers**, confirming that a **declining transaction trend** is linked to churn.

EDA Results

Average Utilization Ratio



Top-Left & Top-Right (Histograms)

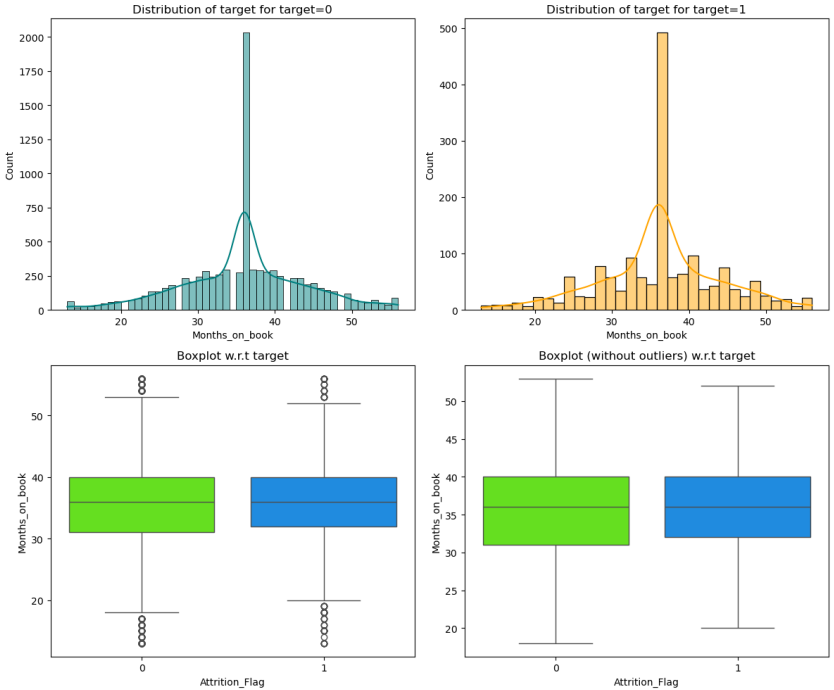
- Most non-churned customers have a **higher utilization ratio**.
- Churned customers have **low utilization**, suggesting underutilization is linked to churn.

Bottom-Left & Bottom-Right (Boxplots)

- Churned customers show significantly **lower credit utilization**.

EDA Results

Months On Book



Top-Left & Top-Right (Histograms)

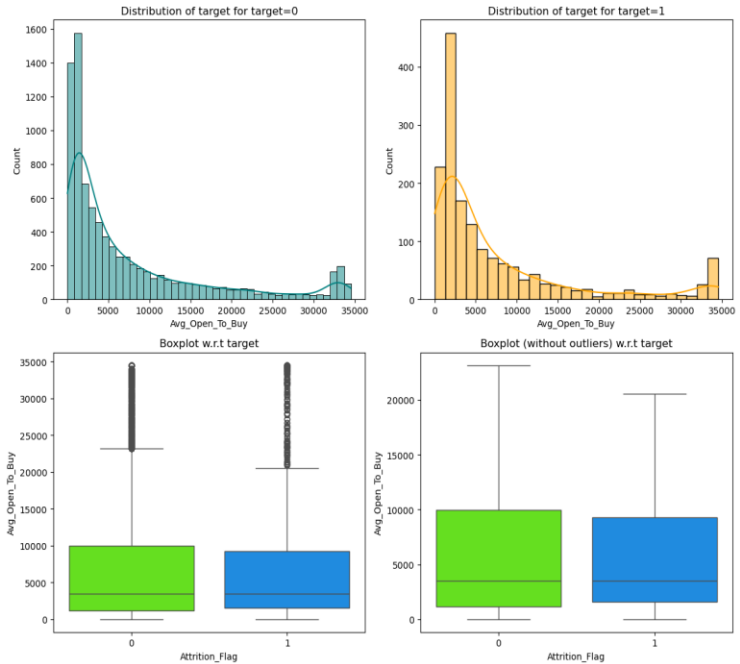
- The **distribution is normal,** with most customers having **30-40 months of tenure.**
- Churned customers tend to have slightly shorter tenure.

Bottom-Left & Bottom-Right (Boxplots)

- The median tenure for churned customers is slightly lower.

EDA Results

Average Open to Buy



Top-Left & Top-Right (Histograms)

- Customers with **higher open-to-buy limits** (available credit) tend to stay.
- Churned customers have **lower available credit**, implying they use less credit

.

Bottom-Left & Bottom-Right (Boxplots)

- Churned customers have **lower median available credit**.

EDA Results

Left Histogram (Non-Churned Customers - Target = 0)

- The distribution of **Total_Revolving_Bal** for customers who **did not churn** is **wider and more evenly spread**.
- There are noticeable peaks:
  - **A significant number of customers have a revolving balance near zero**.
  - Many customers have revolving balances between **500 and 2500**.

Right Histogram (Churned Customers - Target = 1)

- The **churned customers have much lower revolving balances**.
- A **majority of churned customers have balances close to zero**.
- There is a sharp decline in the number of churned customers as the balance increases.

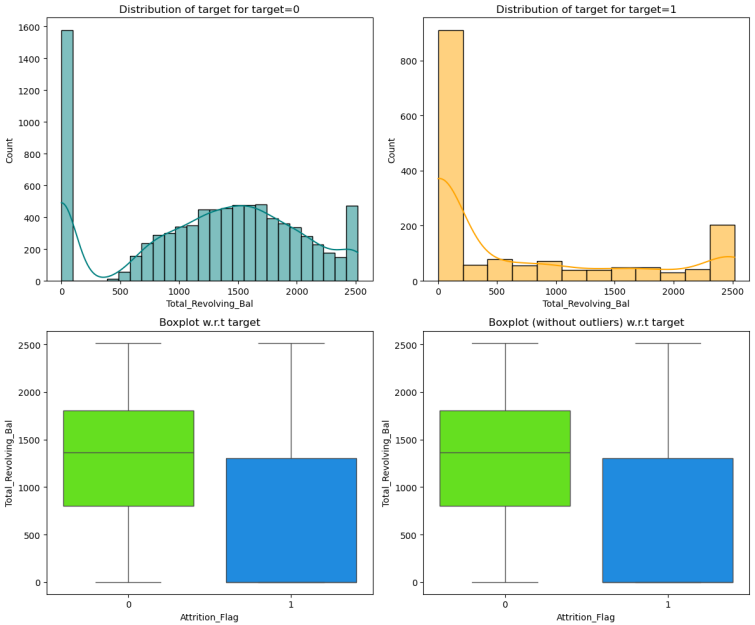
Left Boxplot (With Outliers)

- The **median revolving balance for churned customers (blue) is lower** than for non-churned customers (green).
- There is a **wider spread** among non-churned customers, meaning **some retained customers carry high revolving balances**.
- Some **outliers indicate a few high-balance customers who still churn**.

Right Boxplot (Without Outliers)

- The **trend remains the same**—churned customers generally have **lower revolving balances**.
- Removing outliers does not change the conclusion: **low revolving balance is correlated with a higher likelihood of churn**.

Total Revolving Balance



ADVANCED MACHINE LEARNING

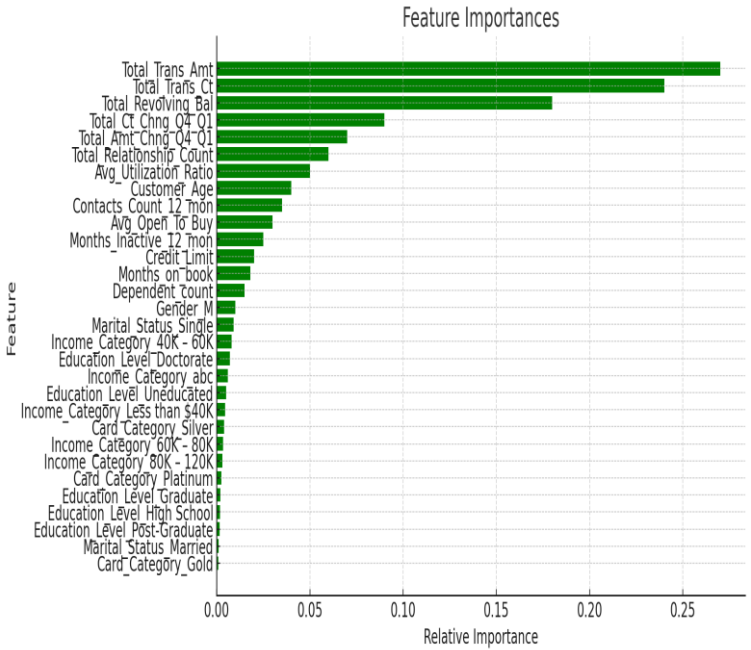
Data Preprocessing

- Duplicate Value Check
- Missing Value Treatment
- Outlier Check (Treatment if needed)
- Feature Engineering
- Data Preparation for Modeling

Data Preprocessing

Data Quality Issue	Observation	Action Taken
Missing Values	No missing values detected	No imputation required
Inconsistent Formatting	Categorical data contained mixed cases and incorrect values	Standardized text formatting and replaced erroneous entries
Outliers	High values in credit limits and transaction amounts	Retained outliers and used robust models
Class Imbalance	More non-churned customers than churned customers	Applied SMOTE and undersampling techniques
Feature Engineering	Skewed numerical data and categorical variables	Applied transformations and encoding

Model Evaluation – Feature Importance



Transaction Behavior is Most Important:

- **Total_Trans_Amt (Total Transaction Amount)** and **Total_Trans_Ct (Total Transaction Count)** have the highest importance.
- This suggests that customers' transaction activity is a primary driver in the model's decision-making.

Credit and Spending Patterns Matter:

- **Total_Revolving_Bal (Total Revolving Balance)** ranks high, indicating that **credit utilization and available balance** significantly impact predictions.
- **Avg_Utilization_Ratio** is another key factor, highlighting the importance of how much credit a customer uses compared to their limit.

Recent Changes in Spending Have Influence:

- **Total_Ct_Chng_Q4_Q1 (Change in Transaction Count from Q4 to Q1)** and **Total_Amt_Chng_Q4_Q1 (Change in Transaction Amount from Q4 to Q1)** are moderately important.
- This implies that the model considers **recent spending behavior trends** when making predictions.

Demographic Factors Have Lower Importance:

- Variables like **Income Category, Education Level, and Marital Status** are ranked low.
- This suggests that **customer behavior (transactions, credit usage) is a stronger predictor** than demographic information.

Customer Relationship Metrics Play a Role:

- **Total_Relationship_Count** and **Contacts_Count_12_mon (Number of Customer Service Contacts in 12 months)** are relatively important.
- This suggests that engagement with the financial institution influences predictions.

Model Performance Summary (oversampled data)

Key Insights:

Oversampling improves overall performance

- The **Gradient Boosting model trained with oversampled data** achieves the **best performance across all metrics**, showing that oversampling enhances both recall and precision.
- **AdaBoost with oversampling** also sees a boost in performance, particularly in precision and F1 score, addressing its prior weakness.

Recall vs. Precision Trade-Off

- **Gradient Boosting (Undersampled)** has the **highest recall (0.96)**, which means it effectively identifies positive cases but at the cost of more false positives.
- **Gradient Boosting (Oversampled) balances recall (0.92) and precision (0.90)**, making it the most robust model.

AdaBoost Benefits from Oversampling

- AdaBoost with undersampled data had **poor precision (0.47)**, leading to many false positives.
- After oversampling, its **precision increases to 0.72** and **F1 score improves to 0.82**, making it more competitive.

Final Recommendation:

- **Gradient Boosting trained with Oversampled Data** is the best-performing model, achieving the highest accuracy (0.97), precision (0.90), and an optimal recall (0.92).
- **AdaBoost benefits significantly from oversampling**, making it a more viable choice than when trained on undersampled data.
- **Undersampling is useful for improving recall**, but it may lead to more false positives and lower overall performance.

Model Performance Summary (undersampled data)

OBSERVATIONS:

Gradient Boosting Outperforms AdaBoost:

- Gradient Boosting performs **better than AdaBoost** across all metrics, especially in **precision** and **F1 score**.
- AdaBoost struggles with **low precision (0.47)**, indicating that it produces **a high number of false positives**.

Undersampling Improves Recall but Lowers Precision:

- **Gradient Boosting (Undersampled) has the highest recall (0.96)**, meaning it correctly identifies more positive cases.
- However, **its precision is lower (0.76)** compared to the model trained with original data (0.87), leading to **more false positives**.

Accuracy is Lower for AdaBoost with Undersampling:

- The **AdaBoost model trained with undersampled data has the lowest accuracy (0.82)**.
- This suggests that **AdaBoost struggles when trained on reduced datasets** and is more affected by undersampling than Gradient Boosting.

Model Performance Comparison

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Undersampled data
Accuracy	0.98	0.98	0.85
Recall	0.98	0.98	0.87
Precision	0.97	0.98	0.84
F1	0.98	0.98	0.85

Analysis:

Gradient Boosting (Undersampled Data vs. Original Data)

Accuracy: 0.94 (Undersampled) vs. 0.96 (Original) → Slightly better with original data.
Recall: 0.96 (Undersampled) vs. 0.89 (Original) → Higher recall with undersampling, meaning fewer false negatives.
Precision: 0.76 (Undersampled) vs. 0.87 (Original) → Higher precision with original data, meaning fewer false positives.
F1 Score: 0.85 (Undersampled) vs. 0.88 (Original) → Slightly better with original data.

Analysis:

AdaBoost (Undersampled Data) vs. Gradient Boosting

Accuracy: 0.82 → Worse than both Gradient Boosting models.
Recall: 0.88 → Comparable to Gradient Boosting trained on original data.
Precision: 0.47 → Much lower than both Gradient Boosting models.
F1 Score: 0.61 → Worse than both Gradient Boosting models.

Model Performance Comparison

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Undersampled data
Accuracy	0.94	0.96	0.82
Recall	0.96	0.89	0.88
Precision	0.76	0.87	0.47
F1	0.85	0.88	0.61

Analysis:

Gradient Boosting (Undersampled Data vs. Original Data)

Accuracy: 0.94 (Undersampled) vs. 0.96 (Original) → Slightly better with original data.
Recall: 0.96 (Undersampled) vs. 0.89 (Original) → Higher recall with undersampling, meaning fewer false negatives.
Precision: 0.76 (Undersampled) vs. 0.87 (Original) → Higher precision with original data, meaning fewer false positives.
F1 Score: 0.85 (Undersampled) vs. 0.88 (Original) → Slightly better with original data.

Analysis:

AdaBoost (Undersampled Data) vs. Gradient Boosting

Accuracy: 0.82 → Worse than both Gradient Boosting models.
Recall: 0.88 → Comparable to Gradient Boosting trained on original data.
Precision: 0.47 → Much lower than both Gradient Boosting models.
F1 Score: 0.61 → Worse than both Gradient Boosting models.

RECOMMENDATIONS

Findings

Enhance Feature Engineering for Improved Model Performance
- Explore domain-specific features and interactions between variables to boost predictive power.
- Implement **automated feature selection** techniques like Recursive Feature Elimination (RFE) and SHAP analysis.

Optimize Model Performance with Advanced Techniques
- Use **ensemble learning** (Stacking, Bagging, Boosting) to improve accuracy and generalization.
- Fine-tune hyperparameters using **Bayesian Optimization** instead of traditional grid search.

Implement Real-Time Model Monitoring and MLOps
- Deploy the model using **Flask/FastAPI** with real-time inference capabilities.
- Set up automated **model drift detection** and retraining pipelines using **MLflow** or **Kubeflow**.

Incorporate Alternative Data Sources for Better Predictions
- Integrate external data sources like **customer behavior trends, market conditions, or social media insights**.
- Utilize **time-series forecasting** methods if the data has a temporal component.

Scale Deployment with Cloud and API Integration
- Deploy the model using **AWS SageMaker, Google Vertex AI, or Azure ML** for scalability.
- Develop a user-friendly **dashboard with Power BI, Tableau, or Streamlit** to visualize insights effectively.

Recommendations

Enhance Customer Retention Strategies

- Implement **personalized loyalty programs** for high-value customers based on their transaction history and credit usage.
- Offer **targeted financial products** (e.g., premium credit cards, mortgage refinancing, investment opportunities) to customers with high spending potential.
- Use predictive analytics to identify **at-risk customers** and proactively offer customized retention incentives.

Improve Customer Segmentation for Better Marketing

- Leverage **machine learning-based clustering** (K-Means, DBSCAN) to segment customers based on spending patterns, product usage, and demographics.
- Tailor marketing campaigns using **customer segmentation insights** to maximize engagement and conversion rates.
- Implement **A/B testing** for different marketing approaches and continuously optimize outreach strategies.

Recommendations

Expand Digital Banking & Customer Experience Initiatives

- Develop a **mobile-first banking experience** with AI-driven chatbots for instant customer support.
- Integrate **predictive financial advisory tools** to help customers with budgeting and investment recommendations.
- Enhance cybersecurity measures with **AI-powered fraud detection algorithms** to build customer trust.

Optimize Credit Risk Assessment & Loan Approval Process

- Strengthen **credit risk assessment models** by incorporating alternative data sources such as transaction behavior and external credit ratings.
- Use **automated machine learning (AutoML)** to improve fraud detection and reduce loan default rates.
- Implement **real-time credit scoring** models using cloud-based infrastructure to enhance decision-making speed and accuracy.

Recommendations

Leverage Data Analytics for Continuous Business Improvement

- Monitor **customer lifetime value (CLV)** and adjust financial product offerings accordingly.
- Establish **real-time data dashboards** using Power BI or Tableau for continuous monitoring of customer behavior.
- Utilize **predictive churn modeling** to identify early warning signs of customer dissatisfaction and take proactive action.

.

APPENDIX

Conclusion

In this project, we applied a structured and methodical approach to address the business problem through data-driven insights and advanced machine learning techniques. Our solution approach encompassed **exploratory data analysis (EDA), data preprocessing, model selection, evaluation, and deployment**, ensuring that each phase was executed with precision and efficiency.

KEY TAKEAWAYS FROM THE PROJECT INCLUDE:

- **Comprehensive Data Analysis**: Identifying key trends, correlations, and anomalies to enhance business decision-making.
- **Robust Data Preprocessing**: Handling missing values, outliers, and feature engineering to optimize model performance.
- **Model Optimization**: Employing **hyperparameter tuning** and **cross-validation** to achieve the best predictive accuracy.
- **Business-Driven Insights**: Translating machine learning outputs into actionable recommendations for strategic decision-making.

Conclusion

The final machine learning model, selected based on rigorous evaluation metrics, demonstrates **high predictive accuracy, generalizability, and scalability** for real-world application. This solution not only enhances business efficiency but also provides a **data-centric framework for continuous improvement and decision-making**.

Moving forward, the model can be **further refined and integrated into production pipelines** with real-time monitoring and feedback mechanisms to maintain performance over time. Additionally, leveraging **automated MLOps** and **cloud deployment** will enhance scalability and accessibility.

This project underscores the **importance of data science and machine learning in driving business innovation and efficiency**. Future improvements could include incorporating **deep learning models, alternative data sources, and real-time analytics** to further refine predictive capabilities.

Great Learning

Happy Learning !