MACHINE LEARNING – 2

GUIDED PROJECT

BANK CHURNERS

LIST OF CONTENTS

SL. NO.	DESCRIPTION	PAGE NO.
1	Problem Statement (Context)	7
2	Data Dictionary	7
3	Data Overview	9
4	Statistical Summary of the dataset	9
5	Categorical Summary of the dataset	12
6	Insights from unique values	14
7	EDA (Univariate Analysis)	16
8	Bivariate Analysis	58
9	Data Reprocessing	107
10	Train Test Split	110
11	Hyperparameter Tuning	121
12	Tuning AdaBoost using Under sampled data	124
13	Tuning Gradient Boosting using under sampled data	128
14	Tuning Gradient Boosting using original data	131
15	Tuning Gradient Boosting using oversampled data	132
16	Tuning XG Boost model with original data	135
17	Model comparison and final model selection	138
18	Training Performance comparison table	141
19	Feature Importance	144
20	Actionable Insights and Business Recommendations	147

LIST OF TABLES

SL. NO.	DESCRIPTION	PAGE NO.
1	Categorical Summary of the dataset	12
2	Various feature predicting customer attrition with numerical values	108
3	Table showing number of missing values	111
4	Model Building on Original Data	113
5	Model Building on Over sampled data	116
6	Model Building on Under sampled data	119
7	Confusion Matrix on Training set	122
8	Confusion Matrix on Validation set	123
9	Confusion Matrix on Training set	125
10	Confusion Matrix on Validation set	127
11	Confusion Matrix on Under sampled training set	129
12	Confusion Matrix on Validation set	130
13	Confusion Matrix on Over sampled Training set	133
14	Confusion Matrix on Validation set	134
15	Confusion Matrix on Original Training set	136
16	Confusion Matrix on Validation set	137
17	Training Performance Comparison Table	139, 141

LIST OF FIGURES

SL. NO.	DESCRIPTION	PAGE NO.
1	‘Customer_Age’ Distribution Plot	16
2	‘Months_on_Book’ Distribution Plot	18
3	‘Credit_Limit’ Distribution Plot	20
4	‘Total_Revolving_Bal’ Distribution Plot	22
5	‘Avg_Open_to_Buy’ Distribution Plot	24
6	‘Total_Trans_Ct’ Distribution Plot	26
7	‘Total_Amt_Chng_Q4_Q1 Distribution Plot	28
8	Total_Trans_Amt’ Distribution Plot	30
9	‘Total_Ct_Chng_Q4_Q1’ Distribution Plot	32
10	‘Avg_Utilization_Ratio’ Distribution Plot	34
11	‘Dependent_Count’ Bar Plot	35
12	‘Total_Relationship_Count’ Bar Plot	37
13	‘Months_Inactive_12_mon’ Bar Plot	39
14	‘Contacts_Count_12_mon’ Bar Plot	41
15	‘Gender’ Bar Plot	43
16	‘Education’ Bar Plot	45
17	‘Marital_Status’ Bar Plot	47
18	‘Income_Category’ Bar Plot	49
19	‘Card_Category’ Bar Plot	51
20	‘Attrition_Flag’ Bar Plot	53
21	Histogram of Distribution of Various Features	55
22	Heatmap of Correlation Check	59
23	Stacked Bar Plot of ‘Gender’	61
24	Stacked Bar Plot comparing ‘Marital_Status’ between ‘Attrited_Customer’ & ‘Existing_Customer’	63
25	Stacked Bar Plot comparing ‘Education_Level’ between ‘Attrited_Customer’ & ‘Existing_Customer’	65
26	Stacked Bar Plot comparing ‘Income_Category’ between ‘Attrited_Customer’ & ‘Existing_Customer’	67

27	Stacked Bar Plot comparing ‘Contacts_Count_12_mon’ between ‘Attrited_Customer’ & ‘Existing_Customer’	69
28	Stacked Bar Plot comparing ‘Months_Inactive_12_mon’ between ‘Attrited_Customer’ & ‘Existing_Customer’	71
29	Stacked Bar Plot comparing ‘Total_Relationship_Count’ between ‘Attrited_Customer’ & ‘Existing_Customer’	73
30	Stacked Bar Plot comparing ‘Dependent_Count’ between ‘Attrited_Customer’ & ‘Existing_Customer’	76
31	Distribution & Box Plot of ‘Total_Revolving_Bal’ w.r.t ‘Attrition_Flag’	78
32	Distribution & Box Plot of ‘Credit_Limit’ w.r.t ‘Attrition_Flag’	81
33	Distribution & Box Plot of ‘Customer_Age’ w.r.t ‘Attrition_Flag’	84
34	Distribution & Box Plot of ‘Total_Trans_Ct’ w.r.t ‘Attrition_Flag’	87
35	Distribution & Box Plot of ‘Total_Revolving_Bal’ w.r.t ‘Attrition_Flag’	90
36	Distribution & Box Plot of ‘Total_Ct_Chng_Q4_Q1’ w.r.t ‘Attrition_Flag’	93
37	Distribution & Box Plot of ‘Avg_Utilization_Ratio’ w.r.t ‘Attrition_Flag’	96
38	Distribution & Box Plot of ‘Months_on_Book’ w.r.t ‘Attrition_Flag’	99
39	Distribution & Box Plot of ‘Total_Revolving_Bal’ w.r.t ‘Attrition_Flag’	102
40	Distribution & Box Plot of ‘Avg_Open_to_Buy’ w.r.t ‘Attrition_Flag’	105
41	Creating new pipeline (AdaBoost Classifier)	122
42	Creating new pipeline (AdaBoost Classifier)	125
43	Creating new pipeline (Gradient Boosting)	129

44	Creating new pipeline (Gradient Boosting)	132
45	Creating new pipeline (XG Boost Classifier)	136
46	Feature Importance Plot	145

PROBLEM STATEMENT

❑ CONTEXT

Thera Bank has noticed that many customers are cancelling their credit cards, which is a problem because the credit card generates significant income through various fees. If more customers stop using their credit cards, the bank could lose a lot of money. To prevent this the bank wants to understand why customers are leaving and identify those who are likely to cancel their cards in future.

We have to create a model that can predict which customers are at a risk of leaving. This will help the bank improve its services and keep the customers from canceling their credit cards.

❑ DATA DICTIONARY

- CLINTNUM: Unique identifier for the customer holding the account
- Attrition_Flag: Customer activity status. Indicates “Attrited Customer” if the account is closed otherwise “Existing Customer”
- Customer_Age: Age of the account holder in years
- Gender: Gender of the account
- Dependent_Count: Number of dependents associated with the account holder

- Education_Level: Educational qualification of the account holder categorized as Graduate, High School, Unknown, Uneducated, College (refers to a college student), Post-Graduate, or Doctorate
- Marital_Status: Marital status of the account holder
- Income_Category: Annual income category of the account holder
- Card_Category: Type of credit card held by the account holder
- Months_on_book: Duration of the customer's relationship with the bank, in months
- Total_Relationship_Count: Total number of financial products held by the customer
- Months_Inactive_12_mon: Number of months the customer was inactive in last 12 months
- Contacts_Count_12_mon: Number of contacts between the customer and the bank in the last 12 months
- Credit_Limit: Credit limit on the customer's credit card
- Total_Revolving_Bal: Balance carried over from one month to next, also known as revolving balance
- Avg_Open_To_Buy: Average amount of available credit left on the card over 12 months
- Total_Trans_Amt: Total transaction amount over last 12 months
- Total_Trans_Ct: Total number of transactions over the last 12 months

- Total_Ct_chng_Q4_Q1: Ratio of total transaction count in fourth quarter to the first quarter
- Total_Amt_Chng_Q4_Q1: Ratio of total transaction amount in the fourth quarter to first quarter
- Avg_Utilization_Ratio: Average percentage of the available credit that the customer used

❑ DATA OVERVIEW

Here is an overview of the columns in the data set.

- The dataset consists of **10,127 entries (rows) and 21 columns.**
- The dataset contains 10 integers columns, 5 float columns and 6 objects (categorical) columns.
- There are some missing data in the dataset:

a) Education_Level: 1,519 missing values
b) Marital_Status: 749 missing values

❑ STATISTICAL SUMMARY OF THE DATASET

1. Customer Age:

- The average age of the customer is 46 years, with a range from 26 to 73 years.

- The IQR range (41 to 52 years) shows that most customers are middle- aged.

2. Months on Book (Tenure with the bank):

- The average customer has been with the bank for about 36 months, with some customers having a tenure as short as 13 months and others as long as 56 months.
- Most customers have been with the bank for 31 to 40 months.

3. Dependent Count:

- On an average, customer have about 2 to 3 dependents.
- Some customers have no dependents while the maximum number of dependents is 5.

4. Total Relationship Count:

- Customers have an average of nearly 4 different types of relationships with the bank. (E.g. Accounts, services)
- The most engaged customers have up to 6 relationships.

5. Month Inactive in last 12 months:

- On average customers were inactive for about 2 months in the past year.
- There are some customers who have not been inactive at all, while others were inactive for up to 6 months.

6. Contacts with the bank in last 12 months:

- The average number of contacts with the bank is about 2.5 times in the past year.
- Some customers have had no contact, while others contacted the bank up to 6 times.

7. Credit Limit:

- The average credit limit is approximately $8,632 with significant variation.
- The lowest credit limit is $1,438, while the highest is- $34,516.

8. Total Revolving Balance:

- The average revolving balance is $1,162 with a range from $0 to $2,517.
- This indicates that while some customers carry a high balance others may pay off their balance regularly.

9. Average Open to Buy (Remaining Credit):

- On average, customers have $7,469 of their credit available to spend.
- The available credit varies widely, similar to the credit limit.

10. Total Amount Change (Q4 to Q1):

- The average change in transaction amount from Q4 to Q1 is 0.76 (76%) with a maximum change of 339.7%
- This suggests that some customers noticeably increased their spending while others did not change their spending patterns.

11. Total Transaction Amount:

- The average transaction amount is $4,404 but this varies greatly with some customers spending as little as $510 and others as much as $18,484.

12. Total Transaction Count:

- Customers make an average of 65 transactions annually with some making as few as 10 and others making up to 139 transactions.

13. Change in Transaction Count:

- The average change in number of transactions from Q4 to Q1 is 0.71 (71%) suggesting that customer's activity level fluctuate throughout the year.

14. Avg. Utilization Ratio:

- The average credit utilization ratio is 27% meaning the customers typically use a little over a quarter of their available credit.
- Some customers use almost their entire credit limits (up to 99.9%) while others use none.

❑ CATEGORICAL SUMMARY OF THE DATASET

	count	unique	top	freq
Attrition_Flag	10127	2	Existing Customer	8500
Gender	10127	2	F	5358
Education_Level	8608	6	Graduate	3128
Marital_Status	9378	3	Married	4687
Income_Category	10127	6	Less than $40K	3561
Card_Category	10127	4	Blue	9436

TABLE 1

➢ Insights Based on the Table:

1. Attrition Flag:

- Existing Customers: Out of 10,127 customers 8,500 (83.92%) are existing customers, meaning they have not yet left the bank's credit card services.
- Attrited Customers: The remaining 1,627 (16.08%) have canceled their credit cards, suggesting potential areas of concern for the bank.

2. Gender:

- Female Customers: The majority of customers are female with 5,258 (52.89%) women using the bank's credit card services.
- Male Customers: Male customer accounts for 4,769 (47.11%) of the total customer base.

3. Education Level:

- Graduate- Level Education: Among the 8,608 customers who provided their education level, the most common level of education is ‘Graduate’ with 3,128 (36.34%) customers holding a graduate degree.
- Missing data: Education level data is missing for 1,519 customers.

4. Marital Status:

- Married Customers: Out of 9,378 customers who disclosed their marital status 4,687 (46.32%) are married, making it the most common status.
- Missing data: Marital status data is missing for 749 customers (7.39%)

5. Income Category:

- Lower Income bracket: The most common income category is ‘Less than 40k’ with 3,561 customers (35.16%) falling into this bracket.
- This indicates a significant portion of the customer base has a lower annual income which could influence their credit card usage and financial behavior.

6. Card Category:

- Blue Card: The majority of the customers 9,436 (93.18%) have a blue card indicating that this is the most popular or accessible card option offered by the bank.

➢ INSIGHTS FROM THE UNIQUE VALUES IN THE KEY CATEGORIES:

1. Attrition Flag:

- Existing Customer: The majority of the customers (8,500) are still using the bank's credit card services, while 1,627 have left, representing a significant portion (16.08%) that requires attention to reduce attrition.

2. Gender:

- Female Dominance: There are slightly more female customers (5,358) compared to the male customers. (4,769). This nearly equal distribution suggests that both the genders are similarly engaged with the bank's credit card services.

3. Education Level:

- Graduate- Education Level: The most common education level is 'Graduate' with 3,128 customers. A significant number of customers also have 'High School' (2,013) and 'Uneducated' (1,487) background.
- The distribution suggests that the bank serves a diverse customer base with varying levels of education attainment which could influence financial literacy and credit card usage.

4. Marital Status:

- Married Majority: Married customers form the largest group (4,687) followed by the single customer (3,943) and a smaller group of divorced customers (748). The high

number of married customers might indicate the importance of family related financial products.

5. Income Category:

- Lower Income Prevalence: The most common income category is ‘Less than $40K’, with 3,561 customers followed by ‘$40K-$60K’ (1,790) and ‘$80K-$120K’ (1,535).
- There is an unexpected category labelled ‘abc’ with 1,112 entries, indicating potential data entry issues or a need for data cleaning.

6. Card Category:

- Blue Card Popularity: The majority of the customers (9,436) hold a blue card which likely represents the bank standard or most accessible card option.
- A smaller number of customers have a higher tier card such as ‘Silver’ (555), ‘Gold’ (116) and ‘Platinum’ (20) suggesting that premium card offerings are less common among the customer base.

❑ EXPLORATORY DATA ANALYSIS (EDA)

UNIVARIATE ANALYSIS:

♦ ‘CUSTOMER AGE’ DISTRIBUTION PLOT:

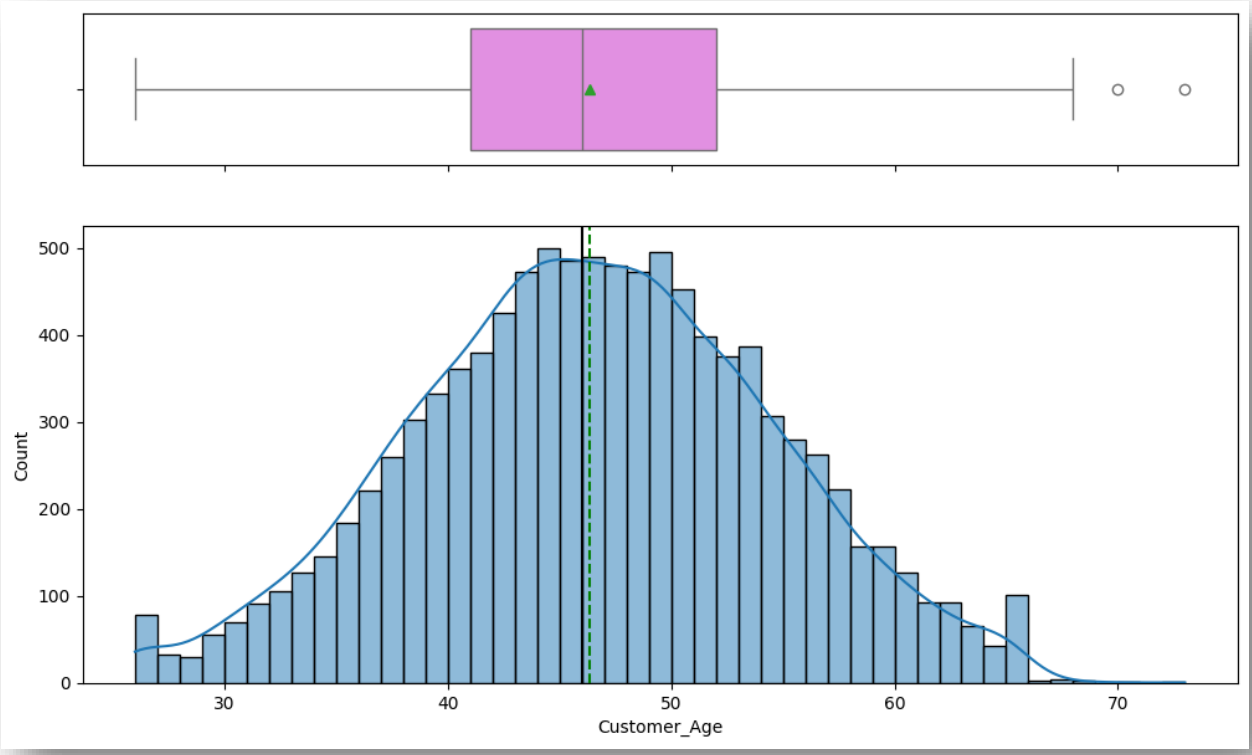

FIGURE 1

➢ Insights based on the Customer_Age plot:

1. Age Distribution:

- The histogram shows that the customer age distribution is approximately normal with a slight right skew. Most of the customers fall within the age range of 40 to 55 years.
- The peak of the distribution is around the age of 50, where the highest number of customers is concentrated.

2. Central Tendency:

- The mean age is around 46 years which aligns closely with the median, confirming the relatively symmetric distribution.
- The box plot above confirms the central tendency with the median age around the same mark.

3. Spread and Outliers:

- The IQR indicates that the middle 50% of the customers lie between 41-52 years old.
- There are few outliers on the higher end (ages above 65) indicating that while most customers are middle aged there are some old customers as well.

4. Skewness:

- The slight right skew in the histogram suggests that there are more younger customers slightly below the mean age, but a small number of older customers extend the tail to the right.
- **Overall, the plot shows that Thera bank credit card customers are predominantly middle aged.**

♦ 'MONTHS ON BOOK' DISTRIBUTION PLOT:

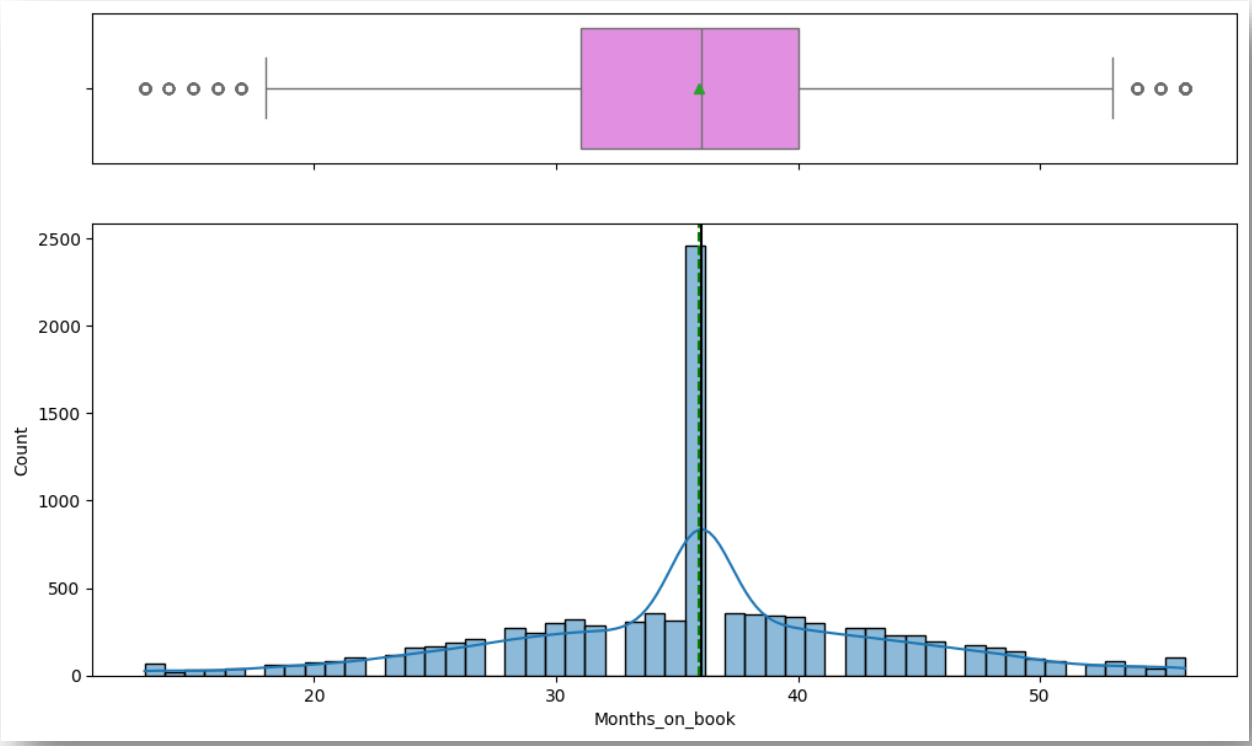

FIGURE 2

➢ Insights based on the 'Months_on_Book' Distribution Plot:

1. Distribution of Months on Book:

- The histogram shows that the most customers have been with the bank for around 36 months, with a significant spike. This suggests that the bank likely has a large group of customers who joined around the same time.

- The distribution is fairly symmetrical but with a notable peak at 36 months.

2. Central Tendency:

- The mean is around 36 months aligning closely with the median.
- The box plot shows that the median number of months on book is centered at 36 months, indicating that the majority of customers have been with the bank for approximately 3 years.

3. Spread and Outliers:

- The IQR in the box plot shows that the middle 50% of the customer have been with the bank for approximately 31-40 months.
- There are few outliers at both the lower ends and higher ends indicating some customers have either just started or have been with the bank for a long period.

4. Symmetry and Peaks:

- The sharp peak at 36 months could suggest a specific promotional period or a significant event 3 years ago that attracted many customers to the bank.
- The overall distribution is fairly symmetrical around this central value, with fewer customer at the extreme ends.
- **Overall, this plot suggests that Thera Bank has a large group of customers who have been with them for around 3 years, with fewer long term or very new customers.**

♦ ‘CREDIT LIMIT’ DISTRIBUTION PLOT:

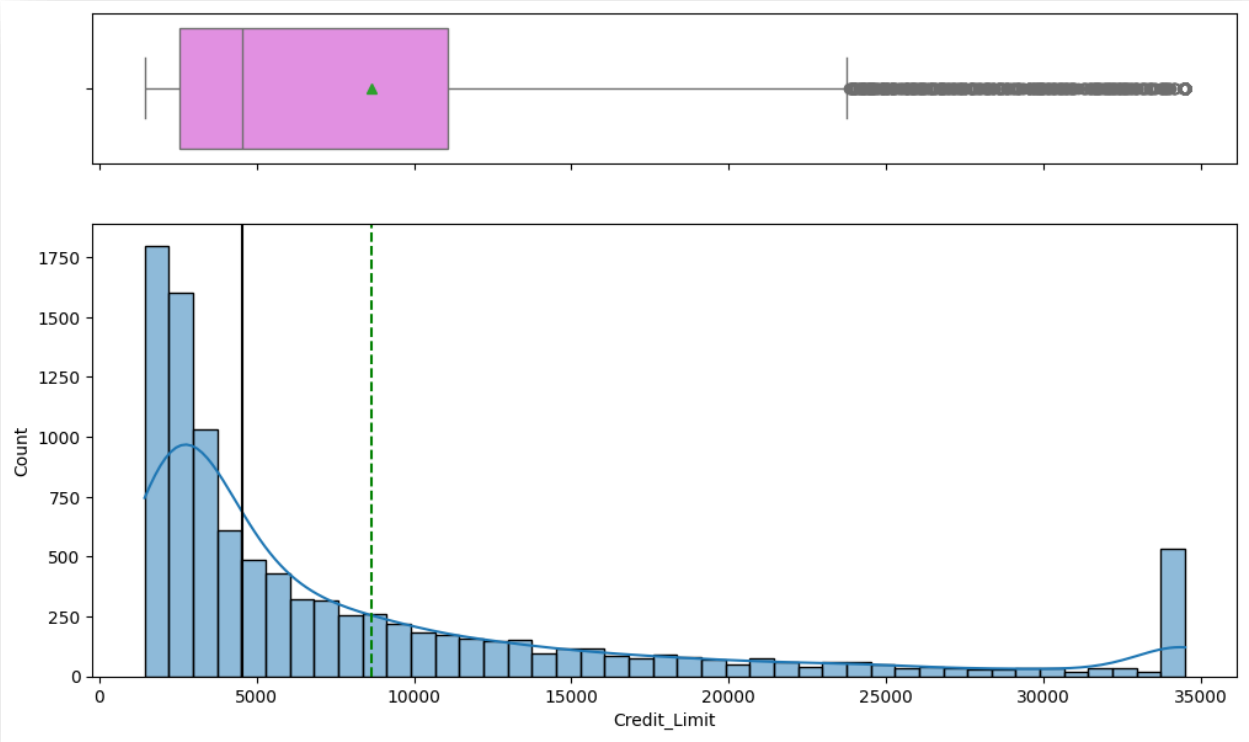

FIGURE 3

➢ Insights based on the Credit_Limit Distribution Plot:

1. Distribution of Credit Limit:

- The distribution is highly right skewed, indicating that majority of customers have relatively low credit limits, with a long tail extending towards higher credit limit.

- Most customers have credit limits below $10,000 with significant concentration between $1,500 and $5000.

2. Central Tendency:

- The mean credit limit is around $8,000; however, the median is lower indicating that the mean is pulled up by a few customers with very high credit limits.
- The box plot also shows that the median credit limit is much lower than the mean, reinforcing the skewness of data.

3. Spread and Outliers:

- The IQR indicates that the middle 50% of the customers have credit limit below $2,500 and $11,000.
- There is a significant number of outliers on the higher end, with credit limits extending up to $35,000. These outliers represent a small proportion of customers with exceptionally high credit limits.

4. Bimodal Distribution:

- There is a noticeable secondary peak around the $35,000, suggesting that small group of customers have been granted the maximum credit limit available. This could indicate a premium customer segment with special privileges.
- **Overall, the plot reveals that while most customers have relatively low credit limits, there is a small distinct group with very high limits, likely representing a more affluent or trusted customer base.**

♦ ‘TOTAL REVOLVING BAL’ DISTRIBUTION PLOT:

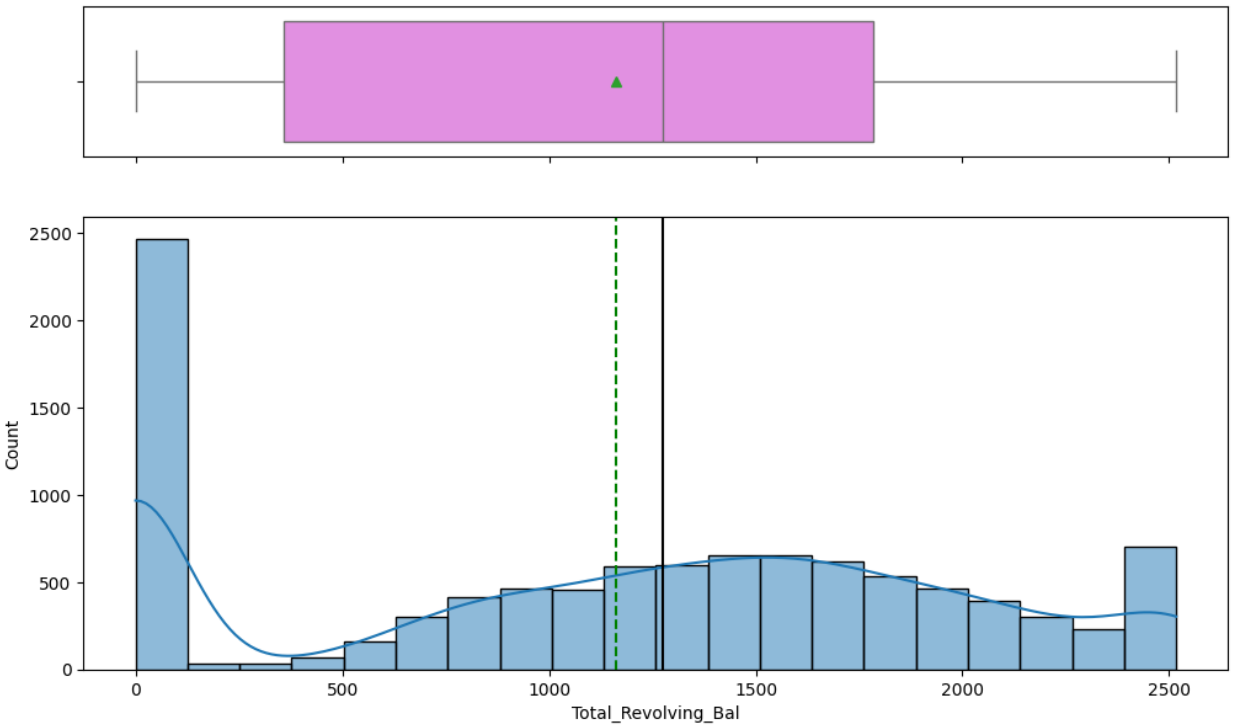

FIGURE 4

➢ Insights based on the ‘Total_Revolving_Bal’ Distribution Plot:

- The ‘Total revolving bal’ ranges approximately from 0 to 2,500.
- The IQR suggests that most of the data is concentrated in a relatively narrow range.

- The median appears to be slightly below 1,000 suggesting that half of the data points have a balance below this value.
- There are no extreme outliers present as the whiskers extend almost to the maximum and minimum values.
- The distribution is right skewed with a long tail extending towards higher values of ‘total revolving balance’.
- A large concentration of data is near ‘0’ showing that many individuals have little or no revolving balance.
- The histogram shows a bimodal distribution, with a smaller peak around 2,500, suggesting that there might be 2 distinct groups: one with low or zero balances and another with higher balances.
- The right skewness suggests that a smaller portion of population carries significantly higher revolving balances which could be a risk factor for default or could indicate higher interest payments for these individuals.

♦ ‘AVERAGE OPEN TO BUY’ DISTRIBUTION PLOT:

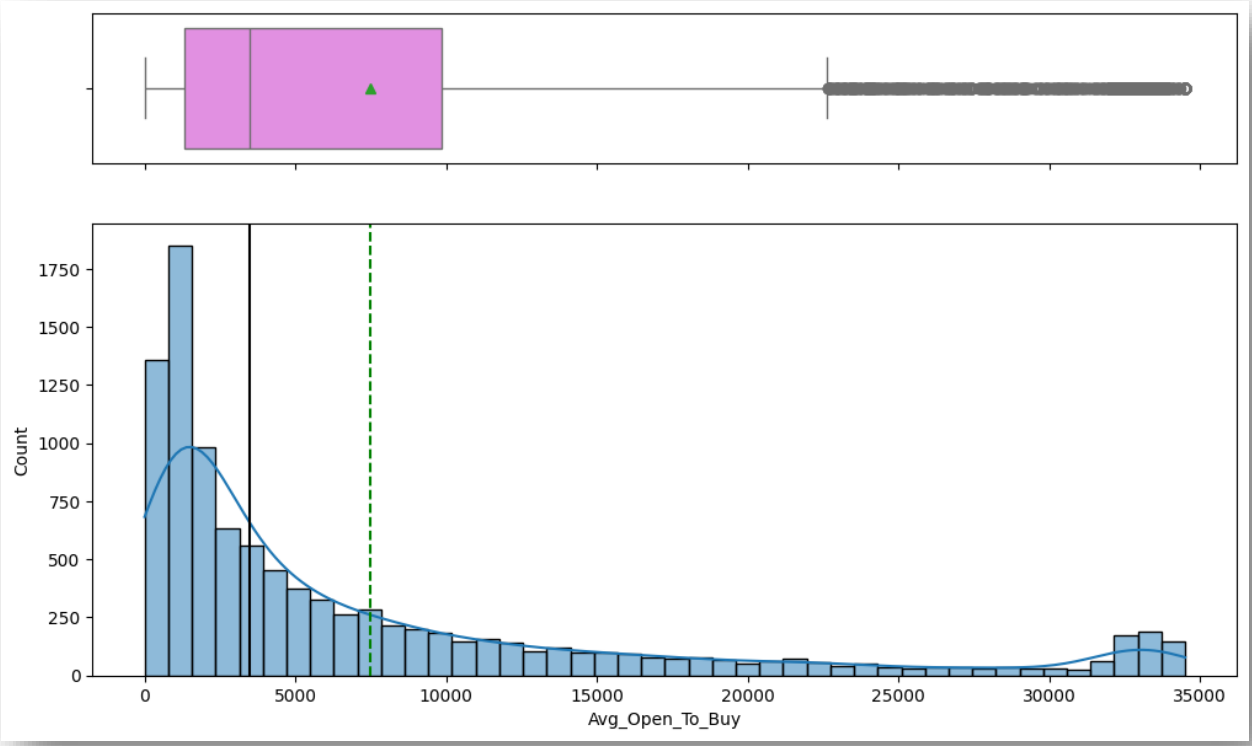

FIGURE 5

➢ Insights based on the 'Avg_Open_To_Buy' Distribution Plot:

- The 'average open to buy' variable ranges from 0 to 35,000.
- The IQR lies between approximately 2,000 and 10K.
- The median is around 6,000 indicating half of the data points have an 'average open to buy' balance below this value.
- There is a significant number of high value outliers as indicated by numerous points beyond the upper whisker.

This indicates that while most of the data is concentrated in a lower range, there are some individuals with significantly higher open-to-buy balances.

- Th distribution is heavily right skewed with a long tail extending towards higher values of ‘Average open to buy’. This indicates that while most individuals have a relatively low average open to buy balance, but there is a small group with very high balances.
- The distribution hints at the possibility of two distinct customer segments: one with lower available credit limit (likely more common) and another with significantly higher available credit. This could be useful for the risk assessment or targeted marketing.

♦ **‘TOTAL TRANS CT’ DISTRIBUTION PLOT:**

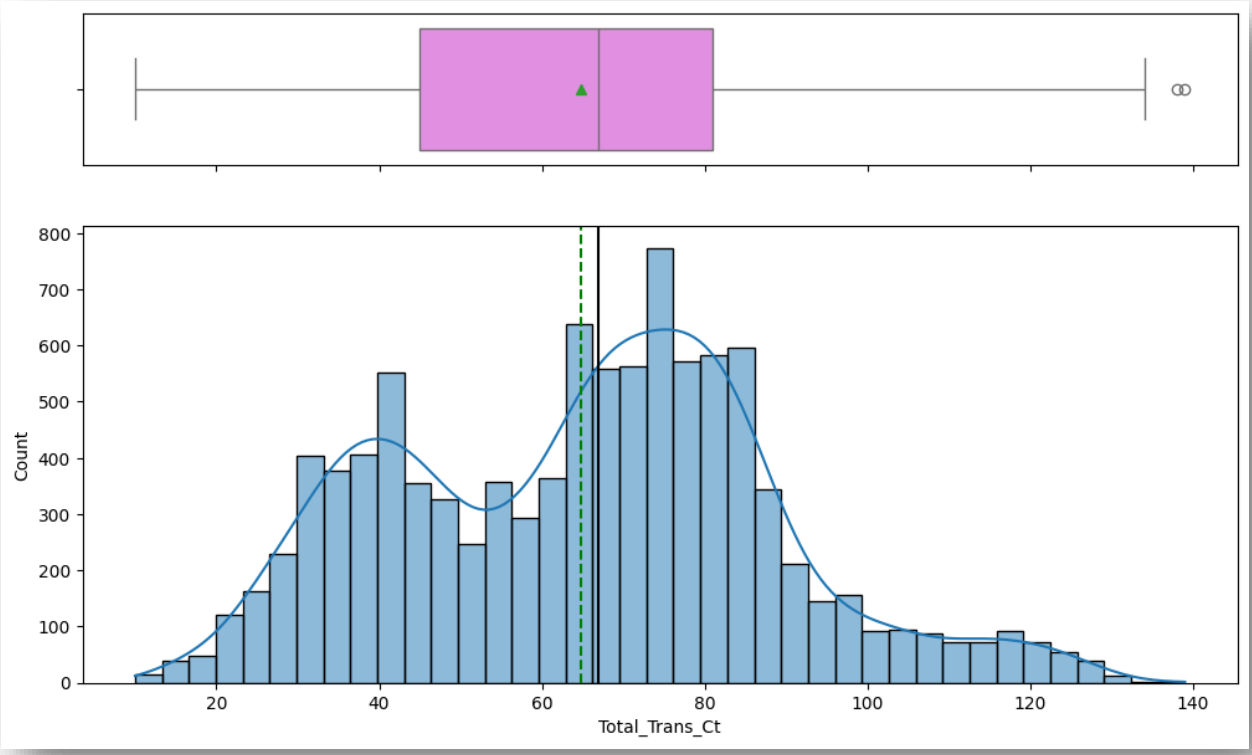

FIGURE 6

➢ **Insights based on the 'Total_trans_Ct' Distribution Plot:**

- The 'Total_Trans_Ct' spans from around 10 to about 40.
- The middle 50% of the data lies between approximately 40 to 80 transactions.
- The median is close to 60 transactions suggesting that half of the customers have fewer than 60 transactions.
- The distribution appears to be slightly right skewed with more customers having a moderate number of

transactions and fewer customers having a very high transaction count.

- The bimodal nature of histogram suggests 2 predominant customer segments: one group that makes fewer transactions (around 50) and another that make more (around 80). These groups could represent different levels of engagement or spending habits.
- A small number of customers have a very high transaction count, which could indicate heavy users or possibly business accounts.
- The distribution can help in identifying typical customer transaction pattern, which could be used for targeted marketing, identifying high- value customers or offering rewards to increase engagement among lower transaction customers.

♦ ‘TOTAL AMOUNT CHANGE Q4 Q1’ DISTRIBUTION PLOT:

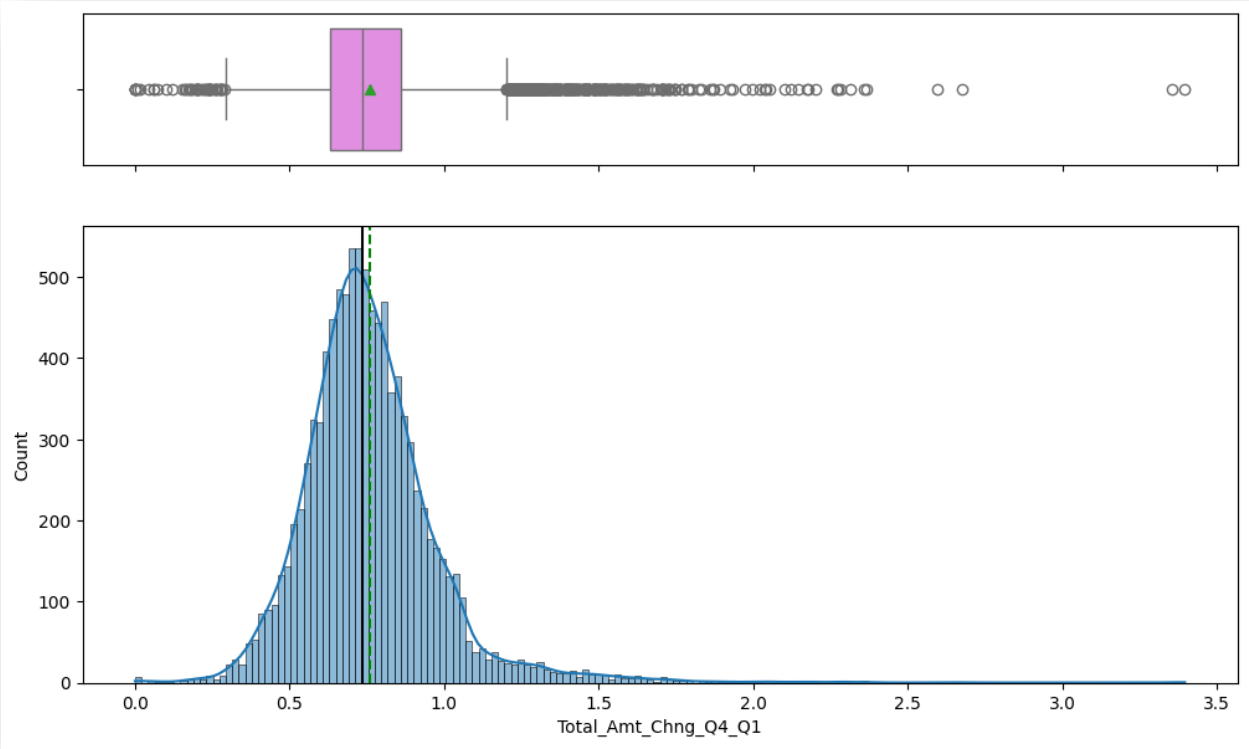

FIGURE 7

➢ Insights based on the 'Total_Amt_Chng_Q4_Q1' Distribution Plot:

- The 'total amt chng Q4 Q1' variable spans from approximately 0 to 3.5.
- The IQR lies between approximately 0.6 to 1.2.
- The median is around 1.0 suggesting that for half of the customers the change in spending between Q4 and Q1 is either stable or slightly increased.

- There are many outliers on both the ends of the distribution, especially on the higher end, suggesting that some customers have experienced significant changes in their spending behavior, either increased or decreased.
- The histogram shows a fairly normal distribution slightly skewed to the right. Most of the data is concentrated around the mean value, with a long tail extending towards higher values.
- The right- skewness indicates that while most customers experienced little change, a smaller number of customers significantly increased their spendings.
- The distribution could be useful for segmenting customer based on spending changes. For e.g., customers who significantly increased their spending might be good candidates for targeted or promotional offers, while those who reduced spendings might need retention efforts.

♦ ‘TOTAL TRANS AMT’ DISTRIBUTION PLOT:

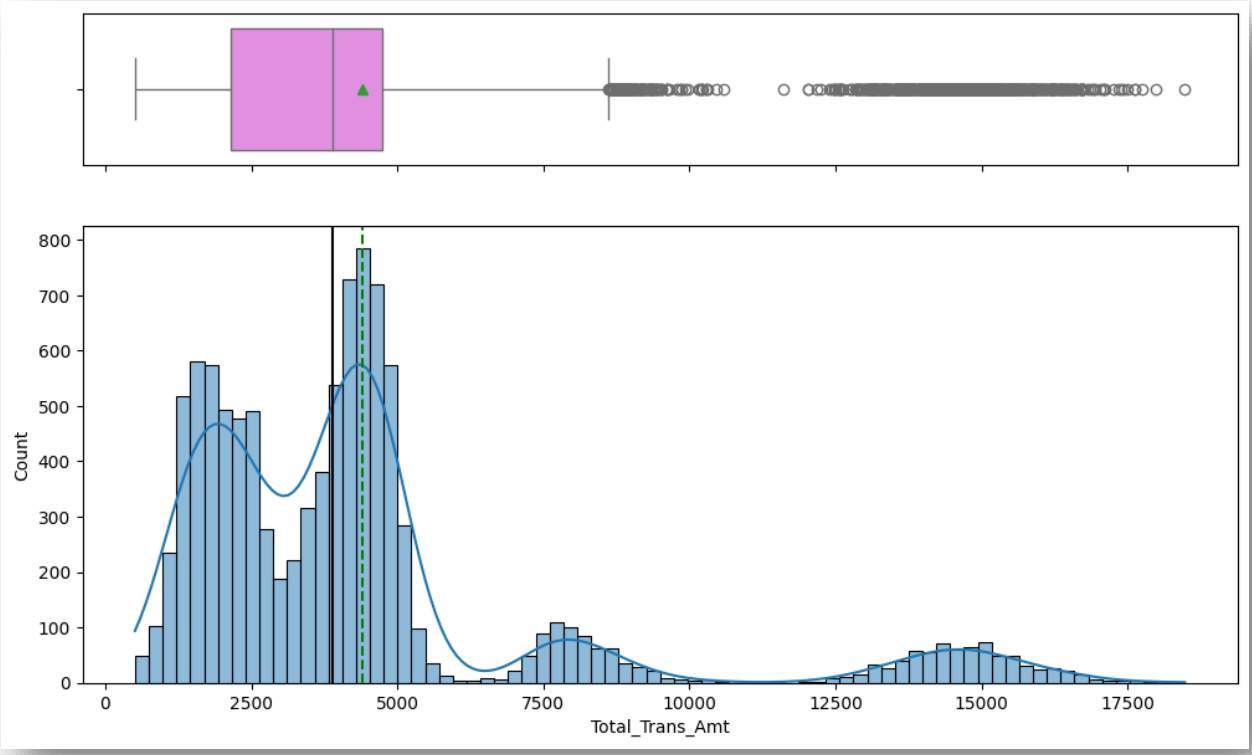

FIGURE 8

➢ Insights based on the 'Total_trans_Amt' Distribution Plot:

- The 'Total_Trans_Amt' ranges from approximately 0 to over 17,500.
- The IQR lies between 2,500 and 7,000.
- The median is close to 5,000 indicating that half of the customers have a total transaction amount below this value. There are many outliers on the higher end,

suggesting that some customer have significantly higher transaction amount compared to the majority.

- The distribution is right- skewed with a long tail extending towards higher transaction amounts.
- The histogram shows a multi modal distribution with several peaks. This suggests that there are 2 distinct group with varying transaction amounts. Some customers have moderate spending (around 5,000) while others have significantly higher spending, creating secondary peaks.
- The distribution can be used to segment customer into different spending tiers, helping in the development of personalized financial products or services.

♦ ‘TOTAL CT CHNG Q4 Q1’ DISTRIBUTION PLOT:

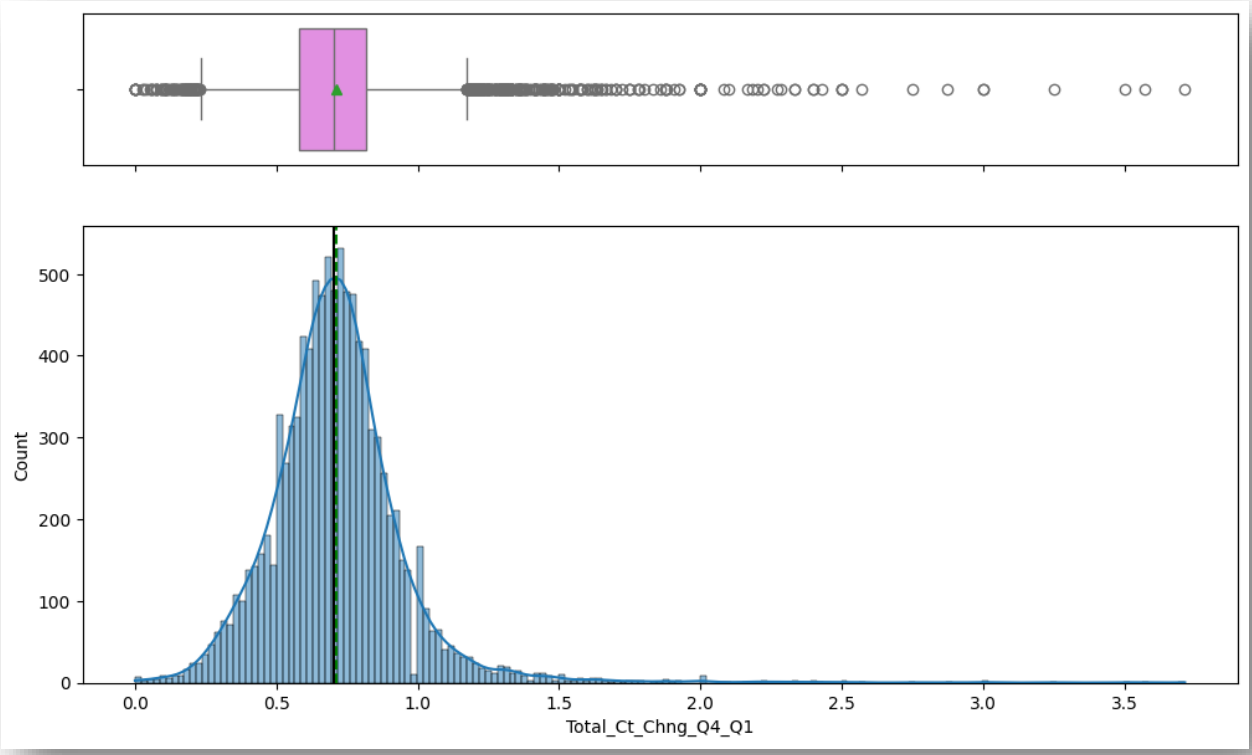

FIGURE 9

➢ Insights based on the 'Total_Ct_Chng_Q4_Q1' Distribution Plot:

- The 'Total_Ct_Chng_Q4_Q1' ranges from about 0 to approximately 3.5.
- The IQR lies between roughly 0.4 to 1.0.
- The median is close to 0.7, indicating that half of the customer have a total count change below this value.
- There are number of outliers on both the lower and higher ends of the distribution, indicating that some customers

experienced significantly larger or smaller changes compared to the majority.

- The histogram is approximately normal but slightly right skewed with most values centered around the median.
- Customers should be segmented based on their 'Total _Ct_Chng_Q4_Q1' values to understand their behavior better and make marketing strategies accordingly. Those with noticeable positive or negative changes might need further investigation to understand the causes behind these shifts.

♦ 'AVG UTILIZATION RATIO' DISTRIBUTION PLOT:

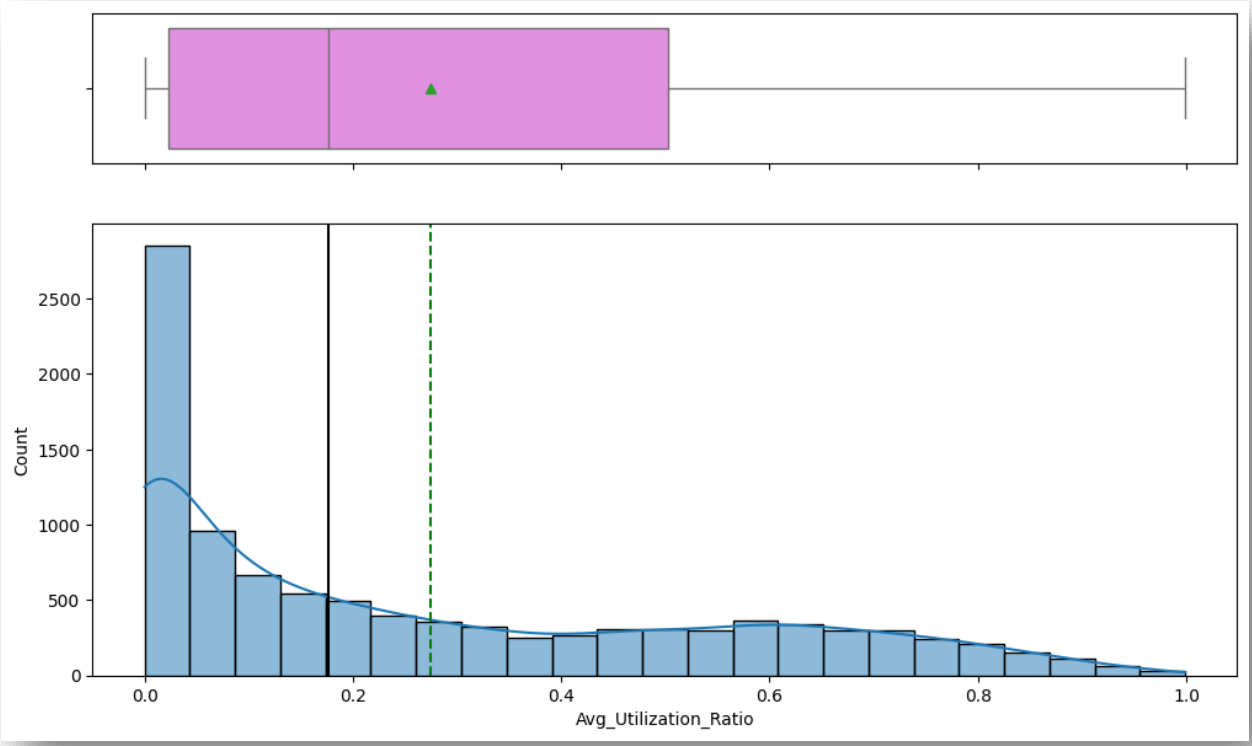

FIGURE 10

➢ Insights based on the 'Avg_Utilization_Ratio' Distribution Plot:

- The boxplot shows the median utilization ratio is quite low around 0.1.
- There are no significant outliers present, but the whiskers are extended, suggesting a wide range of values.
- The boxplot is heavily skewed to the left (lower values) indicating most of the data is clustered towards lower utilization ratios.

- The most frequent utilization ratio is very close to 0, indicating that many entities have low average utilization.
- It might be worthwhile to analyze why there is such a concentration of low utilization if there are barriers to higher utilization that could be addressed.

♦ BAR CHART OF 'DEPENDENT COUNT' PLOT:

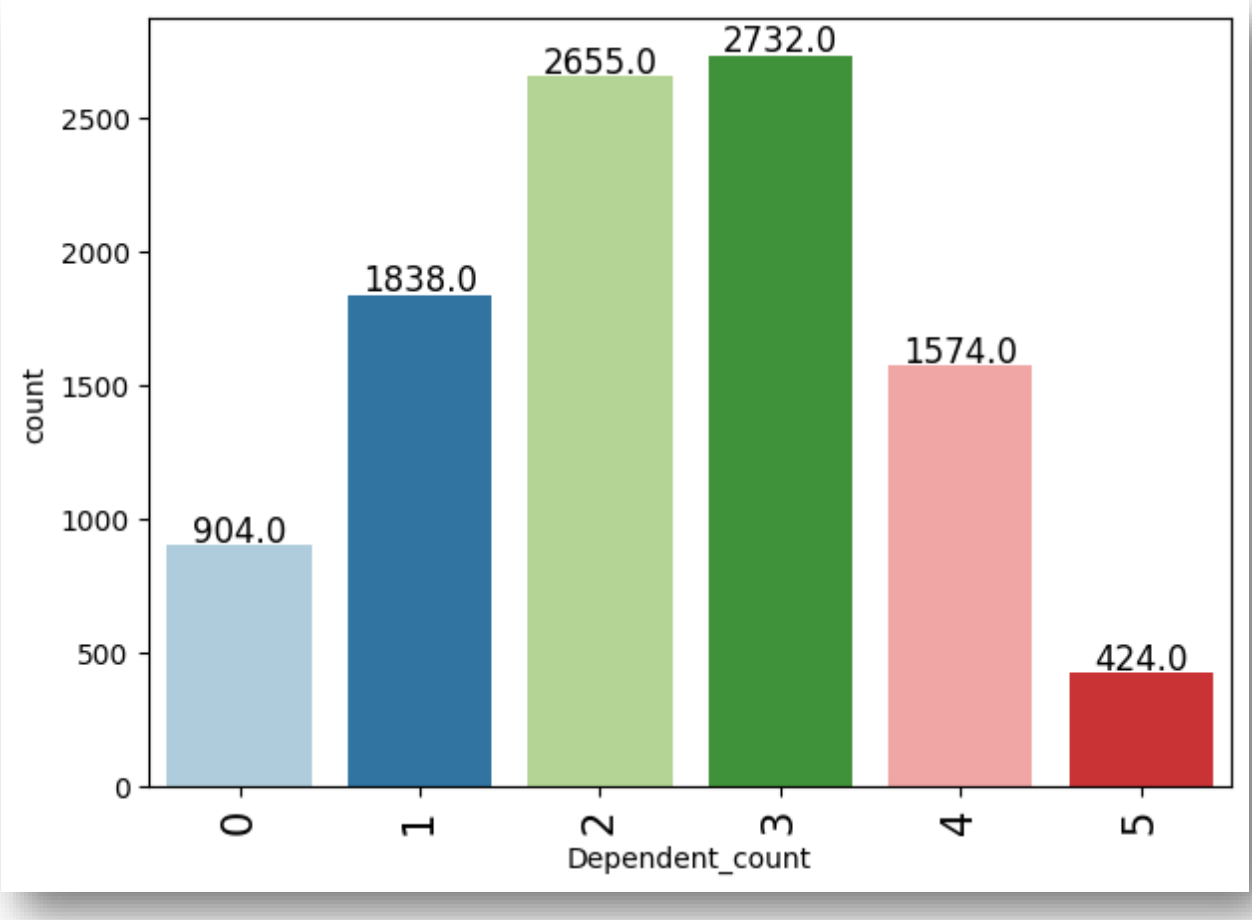

FIGURE 11

➢ Insights based on the 'Dependent_Count' Bar Plot:

- The most common dependent count is '3' with 2,732 occurrences. This indicates that the majority of entities in the dataset have 3 dependents.
- The distribution forms a symmetrical pattern around the dependent counts of '2', '3', which have the highest counts. (2,655 and 2,732 respectively)
- The number of dependents gradually decreases as the dependent count increases beyond 3.
- The least common dependent count is '5' with only 424 occurrences. This indicates larger dependent counts are relatively rare in the dataset.
- Notable trends:

a. There is a significant drop in frequency from 3 dependent to 4 dependents (from 2,732 to 1,574). This indicates that while 3 dependents are very common, families with 4 or more dependents are less frequent.

b. The frequency of having no dependents ('0') is quite low at 904, showing that most entities in the dataset have at least one dependent.

♦ BAR PLOT OF 'TOTAL RELATIONSHIP COUNT':

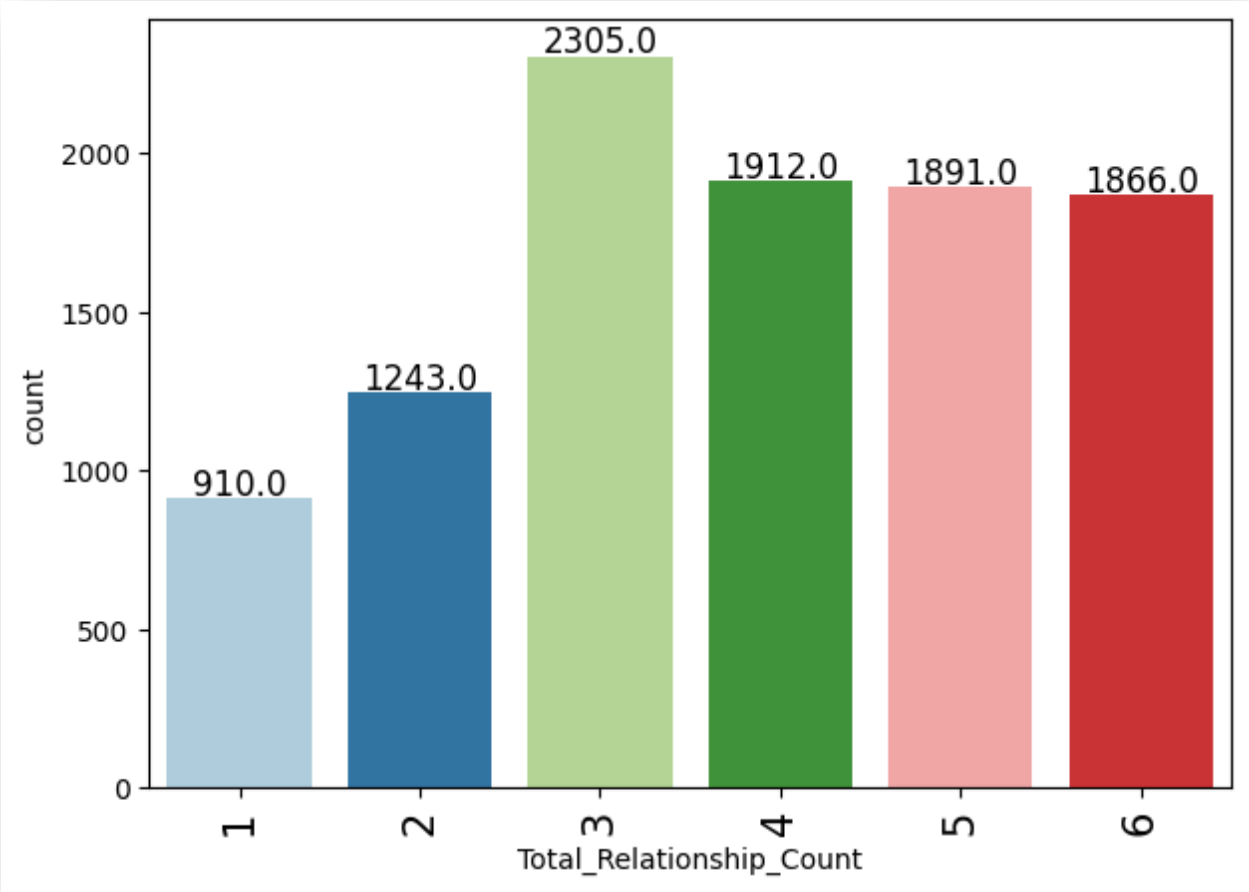

FIGURE 12

➢ Insights based on the 'Total_Relationship_Count' Bar Plot:

- The most common 'Total_Relationship_Count' is '3' with 2,305 occurrences. This suggests that a large portion of entities in the dataset have a total of 3 relationships.

- The distribution is bell- shaped with the frequency peaking at ‘3’ relationships and then gradually decreasing as the relationship count increases to ‘6’.
- There is a smaller count for 1 and 2 relationships with 910 and 1,243 occurrences respectively, suggesting that fewer entities have only one or two relationships.
- The counts for ‘4’, ‘5’ and ‘6’ relationships are relatively high and quite similar with 1,912, 1,891 and 1,866 occurrences, respectively. This indicates that the entities with more extensive relationship are fairly common in the dataset.
- There is a noticeable symmetry in the distribution with the count increasing up to 3 relationships and then decreasing beyond that, though the decline is not steep.

♦ BAR PLOT OF ‘MONTHS INACTIVE 12 MON’:

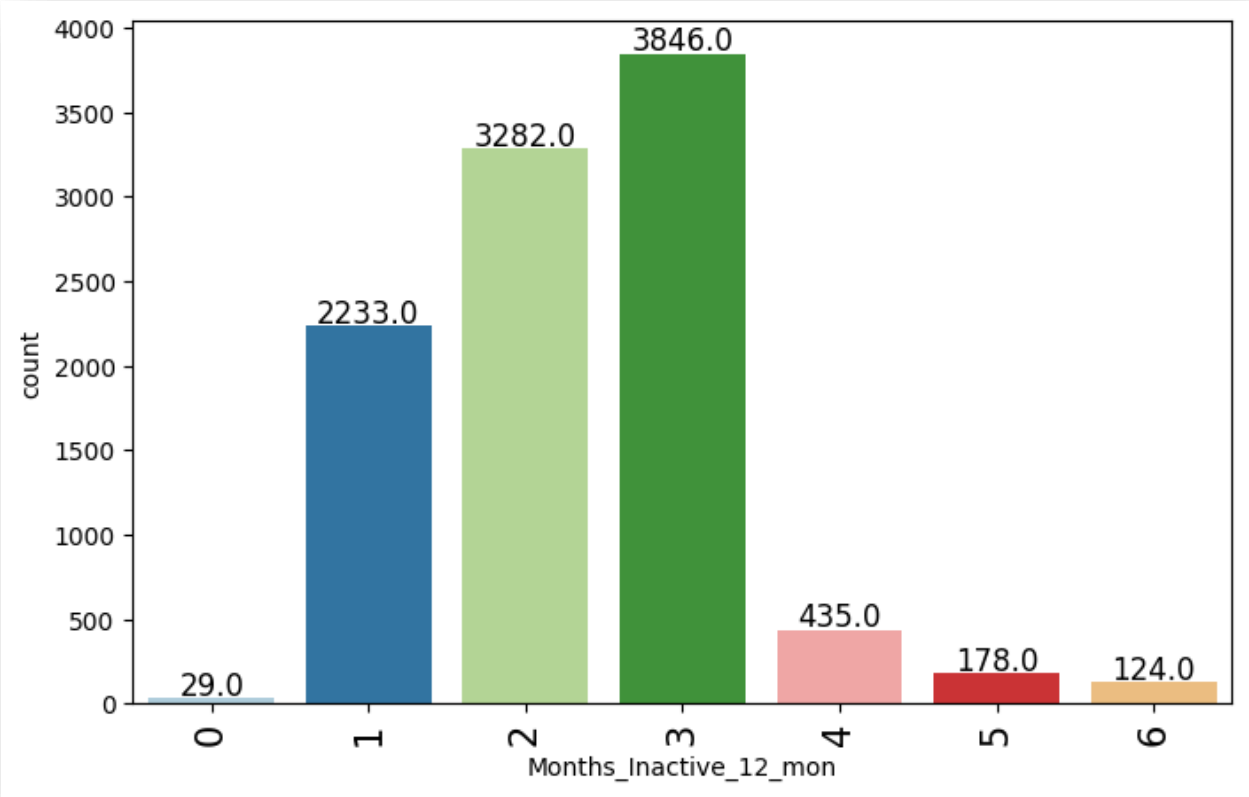

FIGURE 13

➢ Insights based on the 'Months_Inactive_12_mon' Bar Plot:

- The highest number of customers 3,846 have been inactive for 3 months, making it the most common inactivity period among customers.
- There is a gradual decrease in number of inactive customers as the number of inactive months increases beyond 3.

- After 3 months, the number of inactive customers drops sharply. For e.g., there are 3,282 customers inactive for 2 months and 2,233 customers inactive for 1 month.
- The number of customers inactive for 4 months (435), 5 months (178) and 6 months (124) is significantly lower compared to those inactive for 1-3 months.
- Only 29 customers have zero month of inactivity, suggesting that almost all the customers have been inactive at least once during the year.
- The fact that most customers have some months of inactivity, particularly 1-3 months could indicate periodic disengagement, which may be due to seasonal factors, change in product offers.
- The sharp drop in customer counts after 3 months of inactivity suggests a critical point where the chances of re-engagement might decrease, leading to long term inactivity.
- The low count for zero months of inactivity suggests that customer engagement might be an area that require attention.

♦ **BAR PLOT OF 'CONTACTS COUNT 12 MON':**

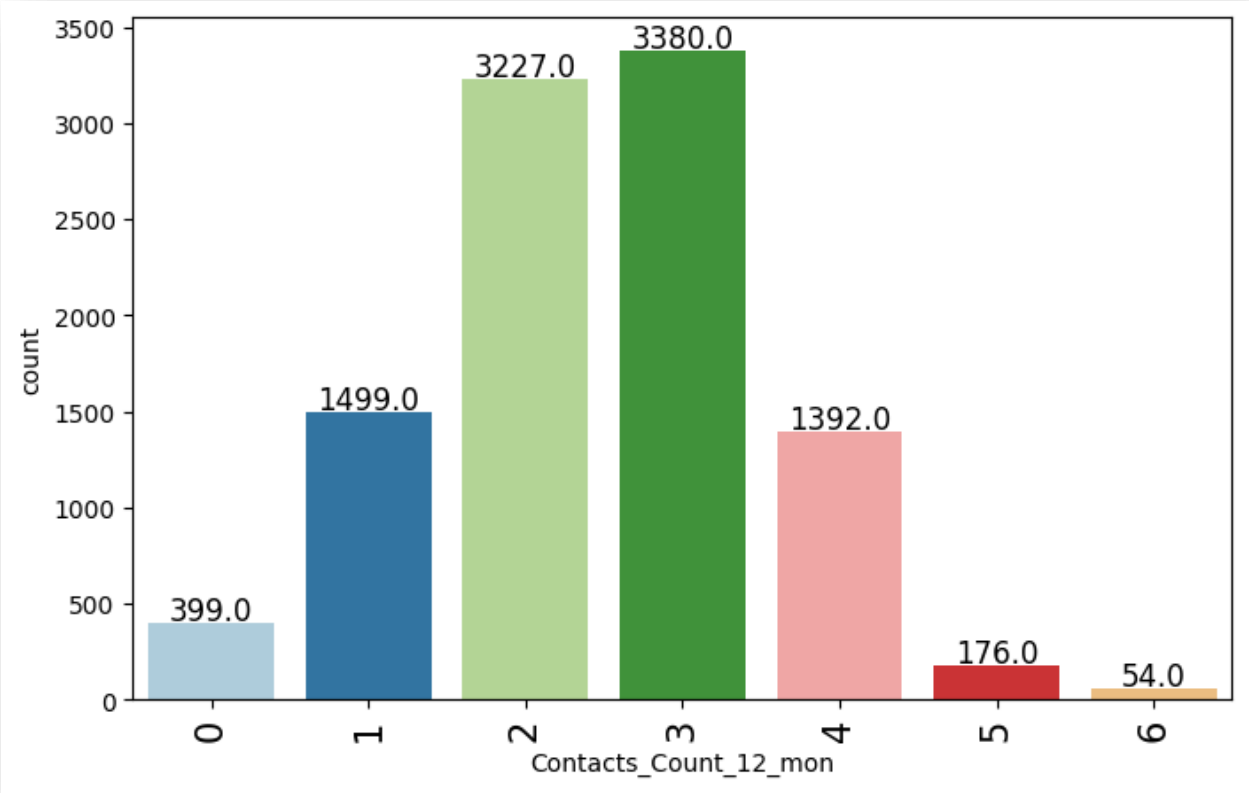

FIGURE 14

➢ Insights based on the 'Contacts_Count_12_mon' Bar Plot:

- The most common contact frequency is 3 times in last 12 months, with 3,380 customers falling into this category. This suggests that many customers engage with the bank approximately once every 4 months.
- A significant portion of customers have been contacted 2 (3,227 customers), 1 (1,499 customers) or 4 times (1,392

customers). The number indicates a consistent pattern where most customers are contacted between 2 to 4 times a year.

- There is a sharp decline in the number of customers contacted more than 4 times. Only 176 customers were contacted 5 times and just 54 customers had 6 contacts in a year. This indicates that frequent contacts are relatively rare.
- A notable number of customers (399) have had zero contacts with the bank over the past year, indicating a segment of customers who are potentially disengaged.

♦ **BAR PLOT OF 'GENDER':**

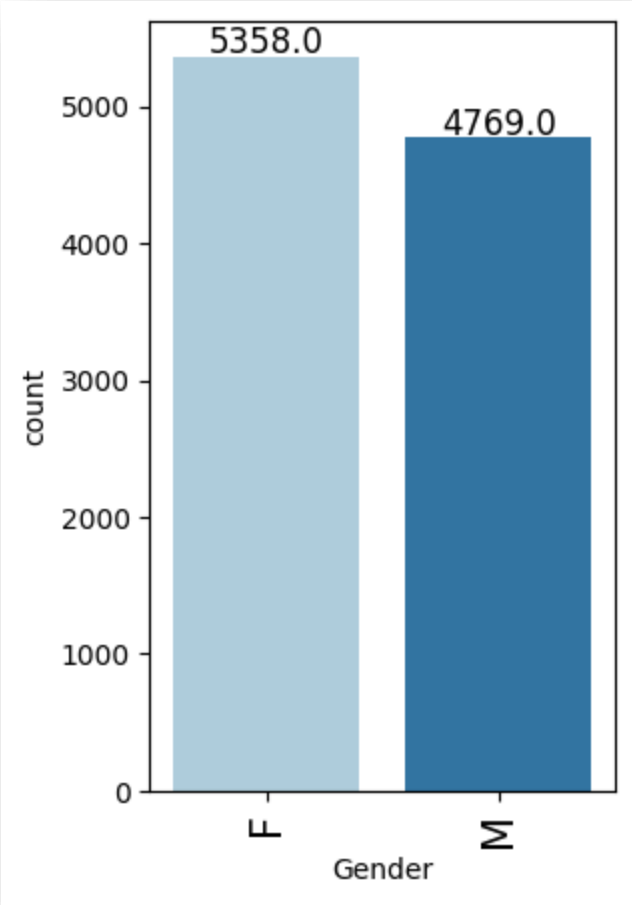

FIGURE 15

➢ Insights based on the 'Gender' Bar Plot:

- The bar plot shows that there are more female customers (5,358) compared to male customers (4,769). This suggests that female make up a slightly larger proportion of the customer base.

- Despite the difference, the gender distribution is relatively balanced with females representing approximately 53% of the customer base and males about 47%. This balance indicates that the bank's services appeal almost equally to both the genders.
- While the distribution is fairly balanced, the slight female majority might indicate the potential for gender specific marketing that cater to the needs and preferences of the female customers.

♦ BAR PLOT OF 'EDUCATION_LEVEL':

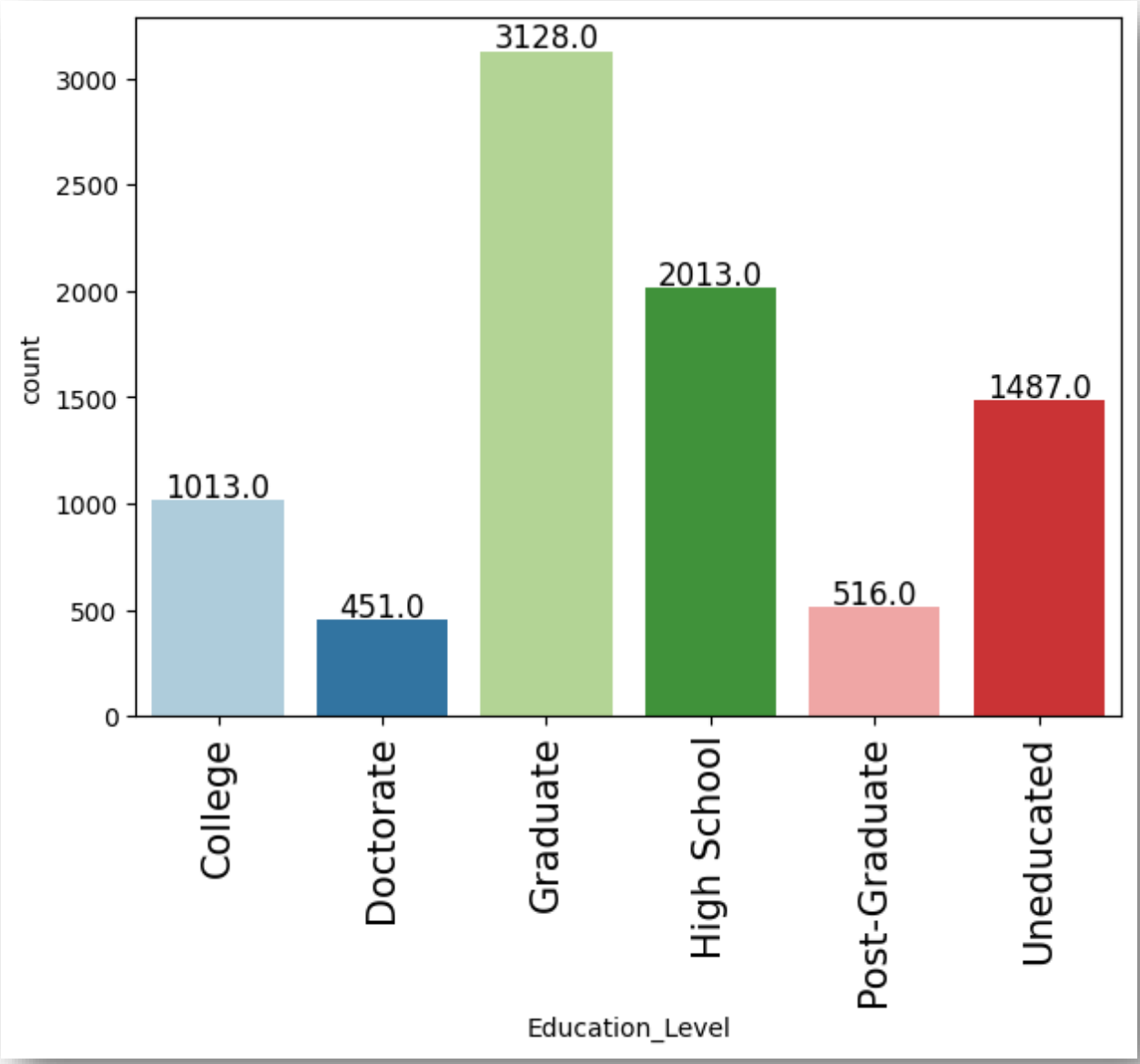

FIGURE 16

➢ **Insights based on the 'Education_Level' Bar Plot:**

- The largest group is those with a ‘graduate level’ of education with 3,128 individuals. This suggests that a significant portion of population has pursued education beyond high school likely influencing their financial stability and credit usage patterns.
- The second largest group comprises ‘high school graduates’ (2,013 individuals). This indicates that substantial number of customers might be starting their careers or opting for jobs that require a high school diploma, possibly affecting their financial behavior differently of those with higher education.
- There is a noticeable segment of 1,487 customers with no formal education. This group is larger than those with ‘Doctorate’ and ‘Post Graduate’ levels, suggesting a significant portion of customer base that may have lower access to financial literacy or opportunities which could impact their financial decisions and credit card usage.
- Both the ‘Post Graduate’ (516) and ‘Doctorate’ (451) groups represent a smaller portion of the customer base. This indicates a more niche segment of customers who have specialized financial needs or higher earning potential, potentially leading to different credit behavior.
- With 1,013 customers, those with a college education make up a moderate portion of base, likely balancing between those who pursued further education and those who did not.

♦ BAR PLOT OF ‘MARITAL STATUS’:

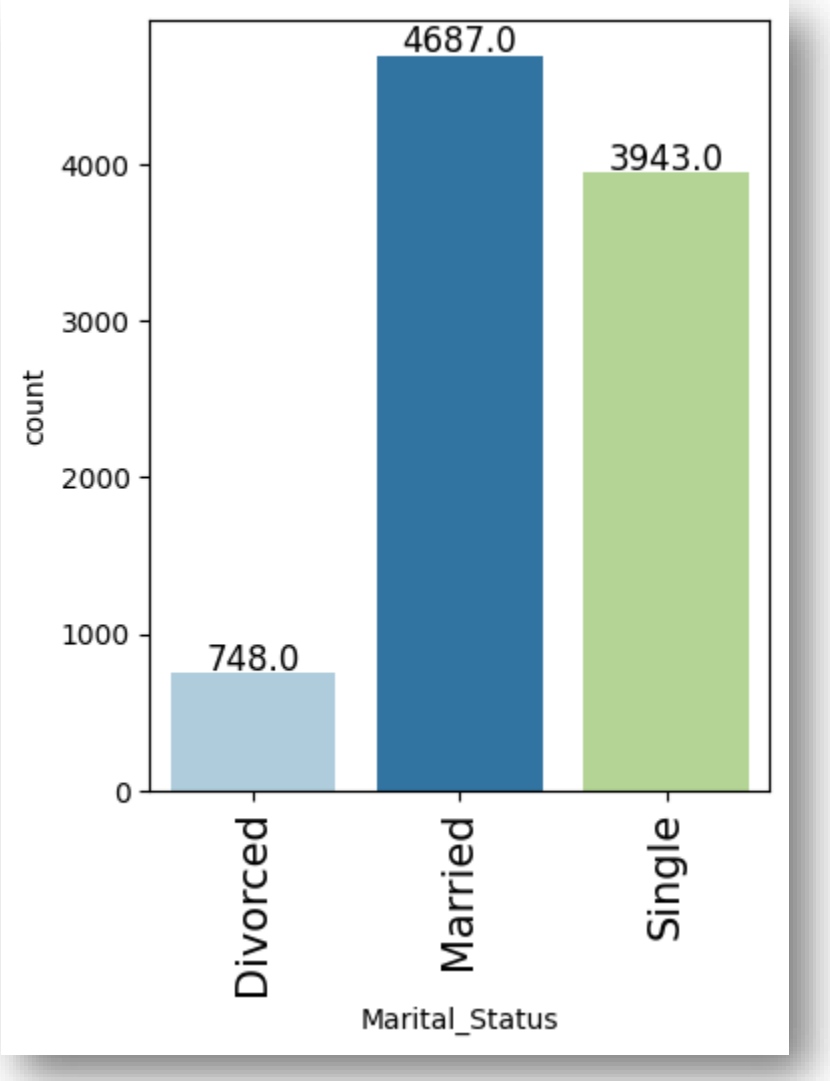

FIGURE 17

➢ **Insights based on the ‘Marital_Status’ Bar Plot:**

- The largest group of customers is those who are ‘Married’ with a count of 4,687. This indicates a significant portion of customer base is likely to be families or individuals in stable relationships which might influence their financial behavior and needs.
- The second largest group consists of ‘Single’ individuals with a count of 3,943. This shows a substantial portion of customer base is single which could focus on different financial products or services that cater to individual needs.
- The smallest group is ‘Divorced’ customers with a count of 748. This smaller segment might have unique financial needs.
- The large numbers in the married and single categories suggest these groups could be the primary focus for marketing efforts while the divorced group although smaller, might benefit from more personalized offerings to meet their specific financial circumstances.

♦ BAR PLOT OF ‘INCOME CATEGORY’:

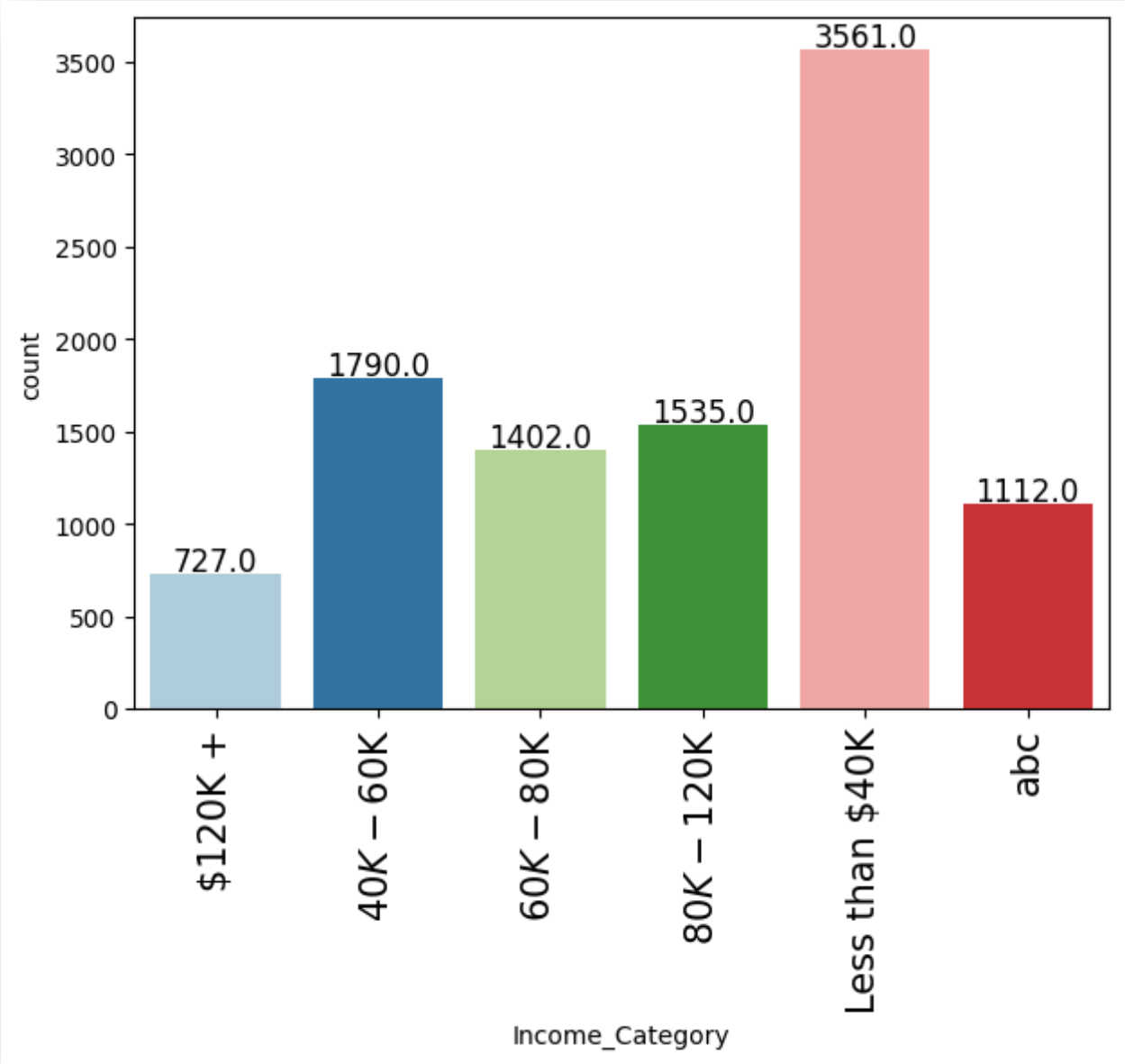

FIGURE 18

➢ **Insights based on the 'Income_Category' Bar Plot:**

- The largest group of customers fall under the “Less than $40K” income category, with a count of 3,561. This indicates that a significant portion of customer base earns below $40,000 annually, indicating potential opportunities for the products and services tailored to lower income individuals.
- The income categories “$40K-$60K”, “$60K-$80K” and “$80K-$120K” have counts of 1,790, 1,402 & 1,535 respectively. These mid- range income groups are fairly evenly distributed, indicating a diverse middle income customer base.
- The “$120K +” category has the smallest representation with 727 customers. This smaller high- income segment might be targeted with premium services.
- There is a category labelled “abc” with a count of 1,112. This appears to be an error suggesting that there might be an issue with data categorization. Addressing this issue is crucial to ensure accurate analysis.
- The concentration of customers in the lower-income bracket indicates that the strategies focusing on accessibility and affordability may resonate well. Whereas, the mid-range and high-income categories may require more diverse and tailored offerings to meet their specific financial needs.

♦ BAR PLOT OF DISTRIBUTION OF ‘CARD CATEGORY’:

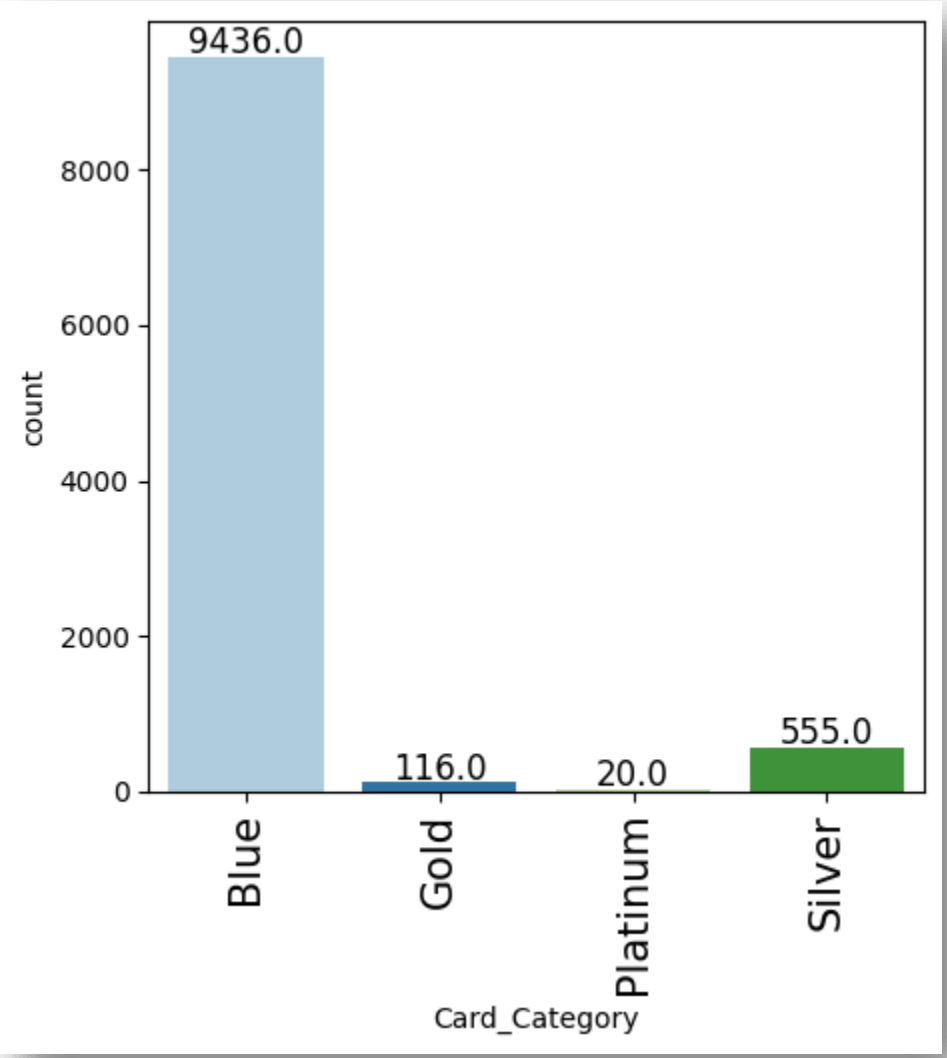

FIGURE 19

➢ Insights based on the 'Card_Category' Bar Plot:

- The majority of the customers (9,436) are using the 'Blue' card category. This suggests that the 'Blue' card is the

most popular or widely accessible option among customers. It also shows that this card offers basic features or benefits that appeal to the largest segments of the customer base.

- The ‘Silver’ card category has 555 customers, making it second most popular option. This shows a moderate interest in a card that possibly offers more benefits than the ‘Blue’ card but it is not as premium as higher tiers.
- The ‘Gold’ card is used by 116 customers indicating limited interest in this mid-tier card which might offer more perks than the ‘Silver’ card but is not widely used.
- The ‘Platinum’ card category has the smallest customer base with only 20 users. This indicates that the ‘Platinum’ card is likely the most premium option is the least popular, possibly due to higher fees or features that appeal to a very niche group.
- The overwhelming performance for the ‘Blue’ card might be an opportunity to educate customers about the benefits of upgrading to higher-tier cards.

♦ **BAR PLOT ON DISTRIBUTION OF ‘ATTRITION FLAG’:**

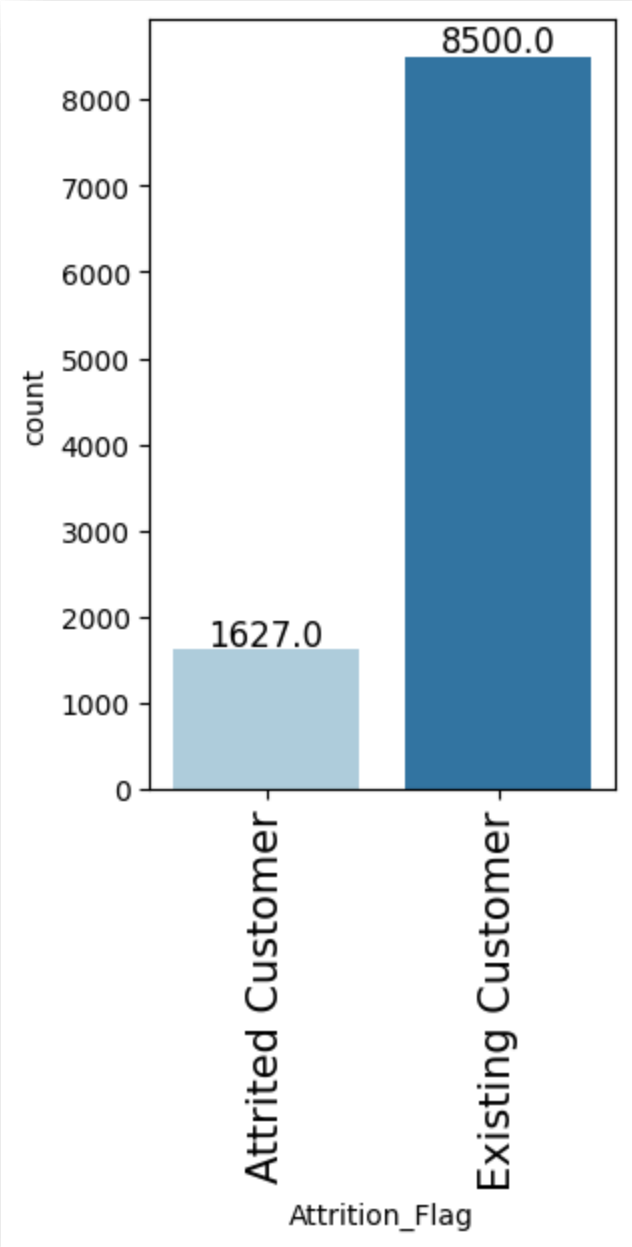

FIGURE 20

- **Insights based on the 'Attrition_Flag' Bar Plot:**

- There is a significant imbalance between the 2 classes. The majority class (‘0’) has 8,500 instances whereas, minority class (‘1’) has only 1,627 instances. This suggests that the occurrences of one category if far more frequent than the other.
- The imbalance might lead to challenge in model training if the data is used for predictive modelling such as bias towards the majority class.
- It also suggests that a much smaller proportion of customers are leaving compared to those staying. This indicates a relatively stable customer base.

♦ HISTOGRAM SHOWING THE DISTRIBUTION OF VARIOUS FEATURES IN THE DATASET:

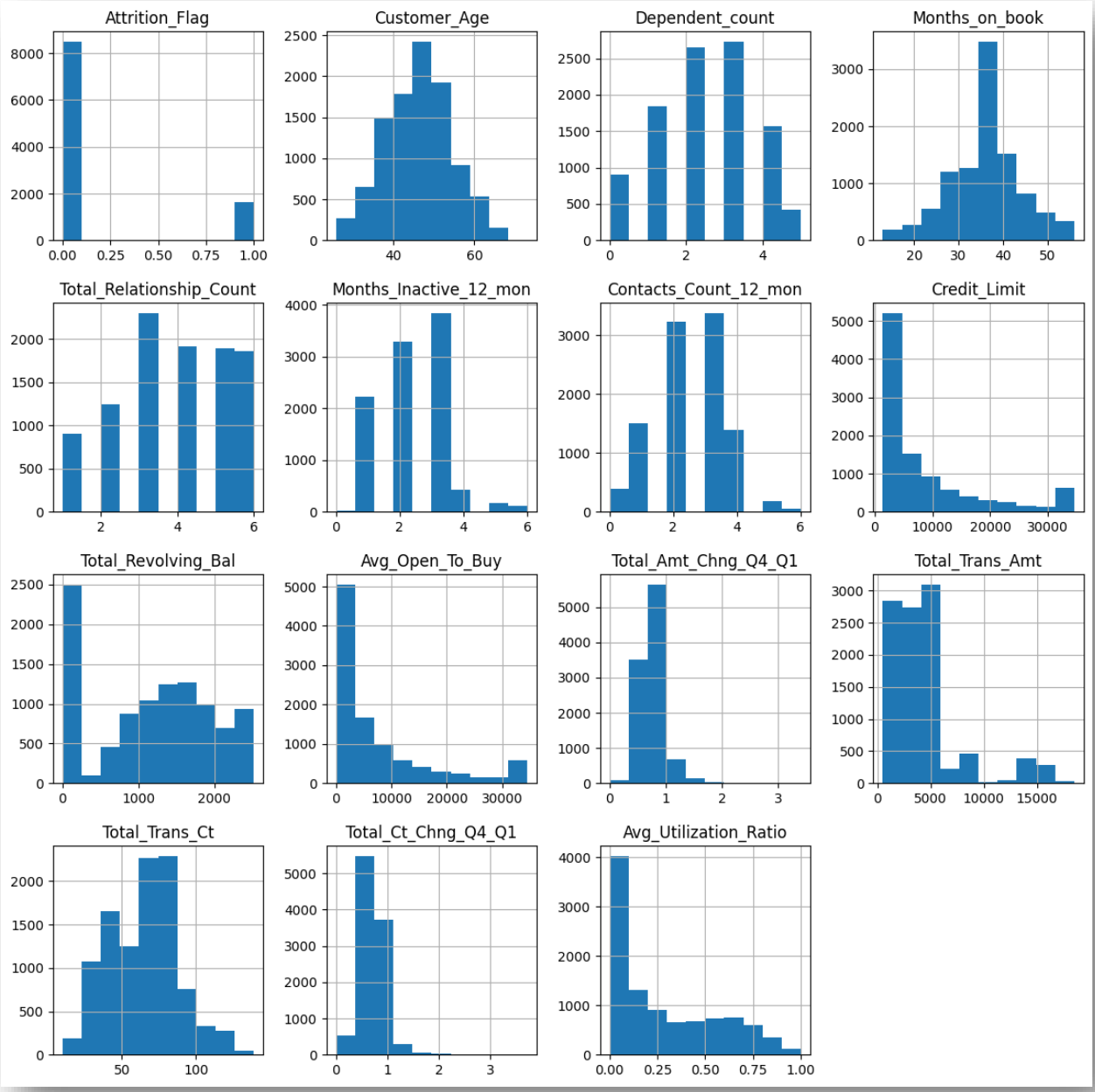

FIGURE 21

➢ Insights based on the histogram of various features:

1. Attrition_Flag:

- The class imbalance is evident here with a majority of data points labelled as ‘0’ (No Attrition) and a smaller fraction labelled as ‘1’ (Attrition).

2. Customer_Age:

- The age distribution appears to be roughly normally distributed, with most customers aged between 40-60 years, peak around 50 years.

3. Dependent Count:

- The majority of the customers have 1-3 dependents with fewer customers having 4 or more dependents.

4. Months on Book:

- This feature seems to have a normal distribution centered around 30 to 40 months, indicating the tenure of most customers fall within this range.

5. Total Relationship Count:

- The distribution suggests a reasonable spread across different relationship counts, with peaks at 2,3 and 5. this suggests customers typically have multiple relationships with the bank.

6. Months Inactive 12 mon:

- Most customers have been inactive for 2 to 3 months over the past year, with very few having no inactivity or high inactivity (4-6 months).

7. Contacts Count 12 mon:

- The distribution suggests that the most customers have been in contact with the bank between 2-3 times in the past year.

8. Credit Limit:

- The distribution is highly skewed to the right suggesting that most customers have a relatively low credit limit, but there are few with very high limits (up to 30,000)

9. Total Revolving Balance:

- This feature is also skewed with most customers having a revolving balance around 0 to 2,000.

10. Average Open to Buy:

- The distribution is heavily skewed to right showing that many customers have a high available credit after the most recent payment, indicating a cautious use of credit.

11. Total Amt Chng Q4 Q1:

- The distribution is slightly right-skewed with most customers showing a change ratio between 0.5 to 2. this suggests that many customers increase their transaction amounts between Q4 and Q1.

12. Total Trans Amt:

- This distribution is right-skewed with a peak around 5,000. Only a few customers have very high transaction amounts up to 15,000.

13. Total Trans Ct:

- This distribution is somewhat normally distributed peaking around 60 to 80 transactions indicating that most customers fall within this transaction range annually.

14. Total Ct Chng Q4 Q1:

- The distribution shows that most customers experience a slight to moderate change in the number of transactions between Q4 and Q1, with a peak around 0.5 to 1.5.

15. Avg. Utilization Ratio:

- The utilization ratio is right-skewed with most customers having a ratio below 0.5 indicating that many customers do not utilize their available credit fully.

❑ BIVARIATE ANALYSIS

Now, let's check the attributes that have a strong correlation with each other.

CORRELATION CHECK:

- **HEATMAP OF CORRELATION CHECK:**

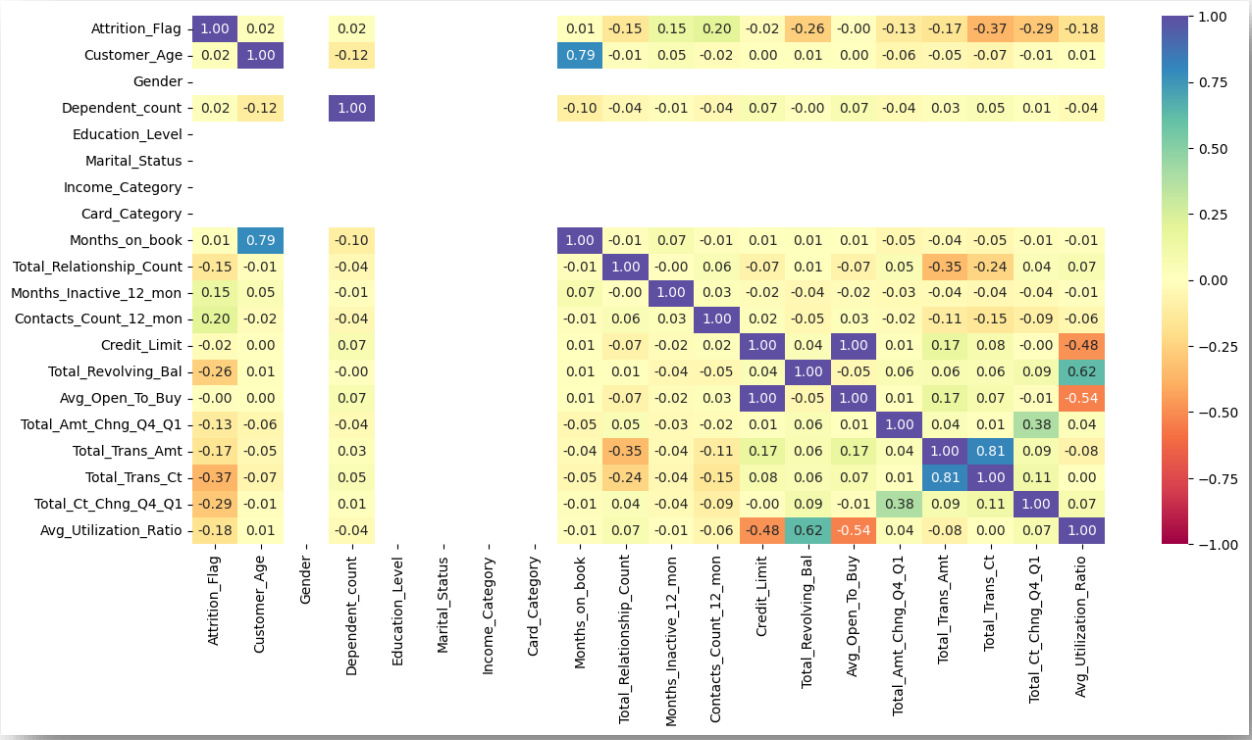

FIGURE 22

➢ Insights based on the Heatmap:

1. Correlation with Attrition Flag:

a) **Contacts_Count_12_mon & Months_Inactive_12_mon** have the highest positive correlation with attrition flag, suggesting that customers with higher contact counts or more inactive months in the last 12 months are more likely to leave.

b) **Total_Trans_Ct & Total_Ct_Chng_Q4_Q1** have the strongest negative correlations with attrition flag,

indicating the customers with higher transaction count or significant increase in transaction counts between quarters are less likely to leave.

2. Correlation among other features:

a) **Credit_Limit & Avg_Open_to_Buy** have a very high positive correlation which is expected since available credit after a recent payment is closely related to the credit limit.

b) **Total_Trans_Ct & Total_Trans_Amt** are also strongly correlated meaning that customers who perform more transactions tend to spend more overall.

c) **Total_Revolving_Balance & Avg_Utilization_Ratio** shows moderate positive correlation with each other, indicating that customers with high revolving balance tend to utilize more of their available credit.

3. Other Notable Correlations:

a) Months_on_Book & Customer_Age shows a strong positive correlation, indicating that older customers generally have longer tenure with the bank.

b) **Total_Relationship_Count, Total_Trans_Ct & Total_Ct_Chng_Q4_Q1** implying that customers with more bank relationship tend to have fewer transactions and less change in transaction counts between quarters.

★ Contacts_Count_12_mon, Months_Inactive_12_mon, Total_Trans_Ct and Total_Ct_Chng_Q4_Q1 stand out to be potentially significant predictors for customer attrition.

- **STACKED BAR PLOT COMPARES THE ‘GENDER’ BETWEEN ‘ATTRITION CUSTOMER’ & ‘EXISTING CUSTOMER’:**

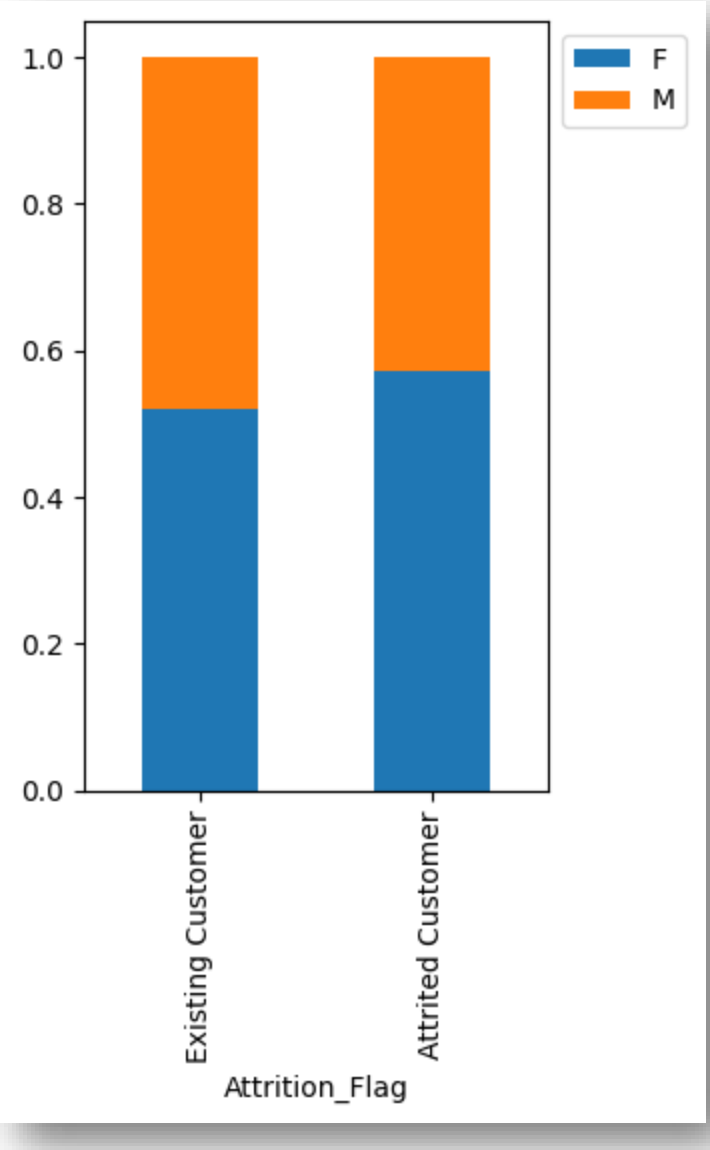

FIGURE 23

➢ Insights based on the Stacked Bar plot of ‘Gender’:

- The gender distribution between existing customer and attrited customer is very similar. Both categories show a nearly equal proportion of male and female customers.
- Since the proportion of male and female are almost the same in both existing and attrited categories, it suggests that customer attrition is not strongly influenced by gender.
- This implies that gender may not be a key differentiator in predicting customer attrition.

♦ STACKED BAR PLOT COMPARES THE ‘MARITAL STATUS’ BETWEEN ‘ATTRITION CUSTOMER’ & ‘EXISTING CUSTOMER’:

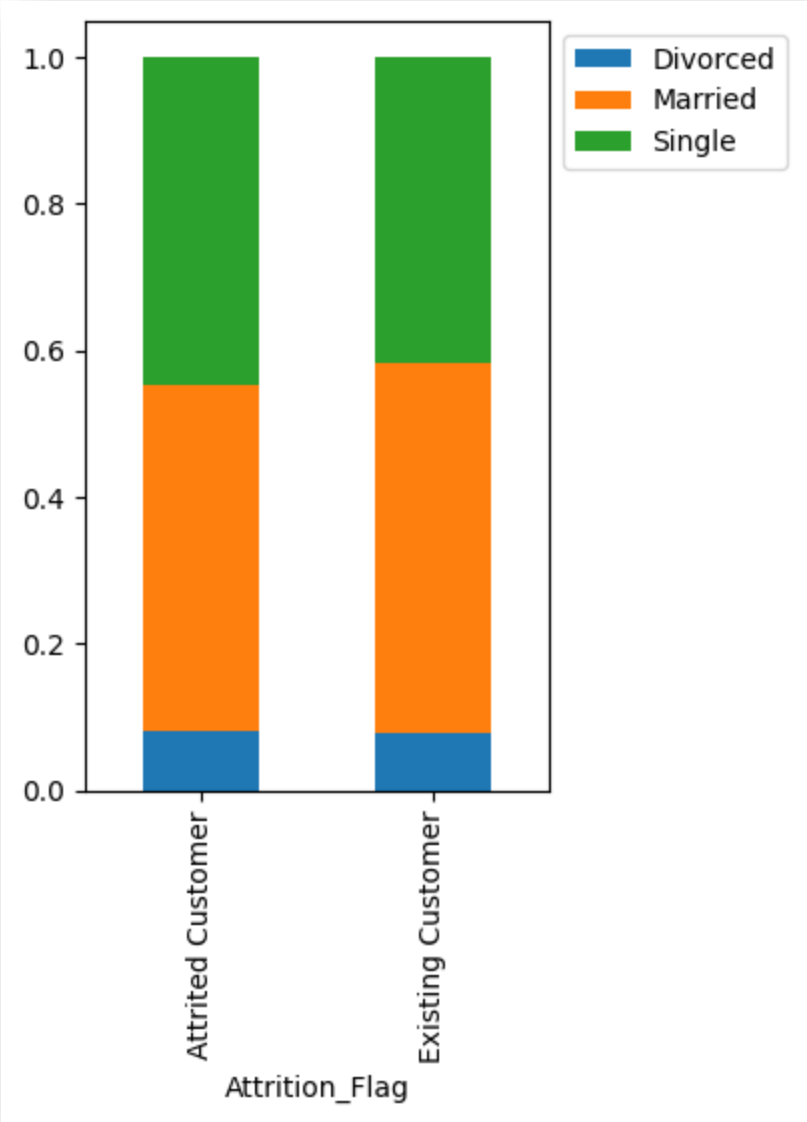

FIGURE 24

- **Insights based on the Stacked Bar plot of 'Marital Status':**

- The proportion of single customers is higher among attrited customers compared to the existing customers. This could indicate that single customers are more likely to leave.
- The proportion of married customers is slightly lower in the attrited group compared to the existing group suggesting that married individuals may be more likely to remain customers.
- The proportion of divorced customers is relatively small and appears in similar groups, indicating that being divorced may not significantly impact customer attrition.
- The marital status seems to play a role in customer attrition, with single customers more likely to attrite compared to the married customers.

♦ STACKED BAR PLOT COMPARES THE 'EDUCATION LEVEL' BETWEEN 'ATTRITION CUSTOMER' & 'EXISTING CUSTOMER':

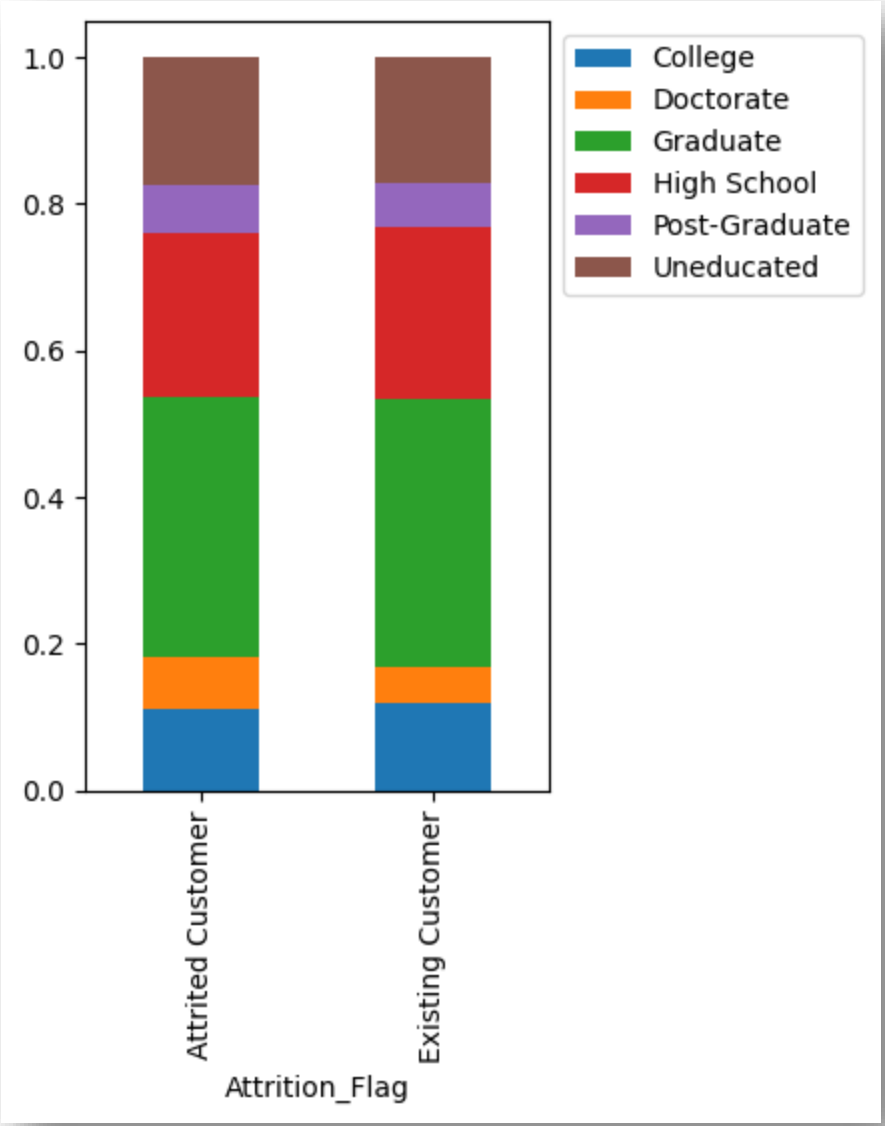

FIGURE 25

- **Insights based on the Stacked Bar plot of 'Education Level':**

- The distribution of education levels between attrited and existing customers is very similar. This suggests that education level does not significantly influence the customer attrition.
- The largest segment in both the groups are customers with a ‘High School’ and ‘Graduate’ education. This suggests that most customers fall within these education levels, and there is no significant difference in attrition across these groups.
- The ‘College’ and ‘Post Graduation’ groups are slightly more represented among the existing customer than attrited customer, though the difference is minimal.
- The ‘Uneducated’ group is slightly more represented in attrited customers, but again the difference is very small.
- Education level appears to have a minimal impact on customer attrition, as the proportions across all education categories are nearly identical between existing and attrited customers. This suggests that other factors beyond education may be more critical in determining the customer attrition.

♦ STACKED BAR PLOT COMPARES THE ‘INCOME CATEGORIES’ BETWEEN ‘ATTRITION CUSTOMER’ & ‘EXISTING CUSTOMER’:

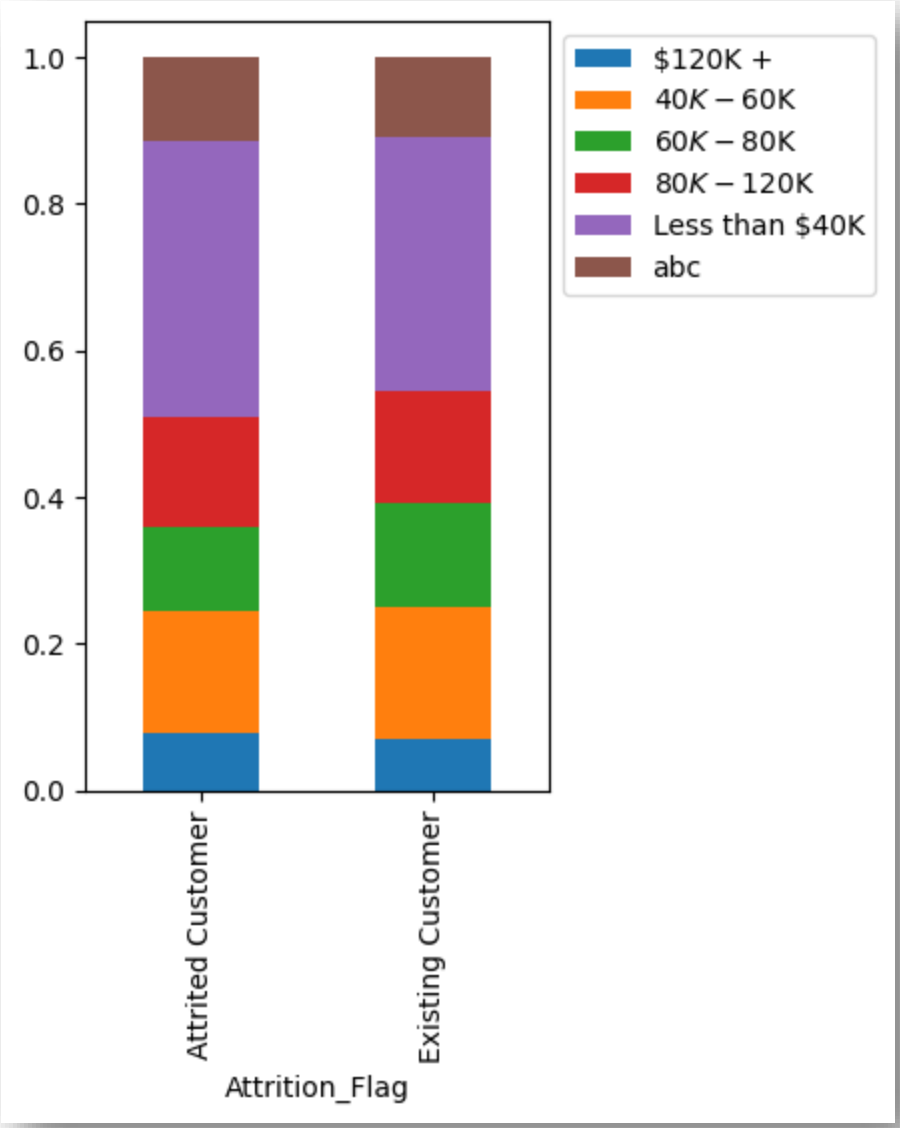

FIGURE 26

- **Insights based on the Stacked Bar plot of 'Income Categories':**

- Both attrited customer and existing customer groups have similar distribution across the income categories. This suggests that income might not be a significant differentiator between customers who leave and those who stay.
- The “Less than $40K” income category has the largest proportion in both customer groups, suggesting that a significant portion of customers belong to this lower income bracket.
- The “$120K +” category has the smallest representation in both the groups, indicating that higher-income individuals are less common among these customers.
- The “abc” category appears in the legend, which does not correspond to a valid income category. This might be an error in the data labelling and it should be investigated and corrected if necessary.
- Overall, the similarity in income distribution across both customer groups indicates that income alone might not be a strong predictor of customer attrition.

♦ STACKED BAR PLOT COMPARES THE ‘CONTACTS COUNT 12 MON’ BETWEEN ‘ATTRITION CUSTOMER’ & ‘EXISTING CUSTOMER’:

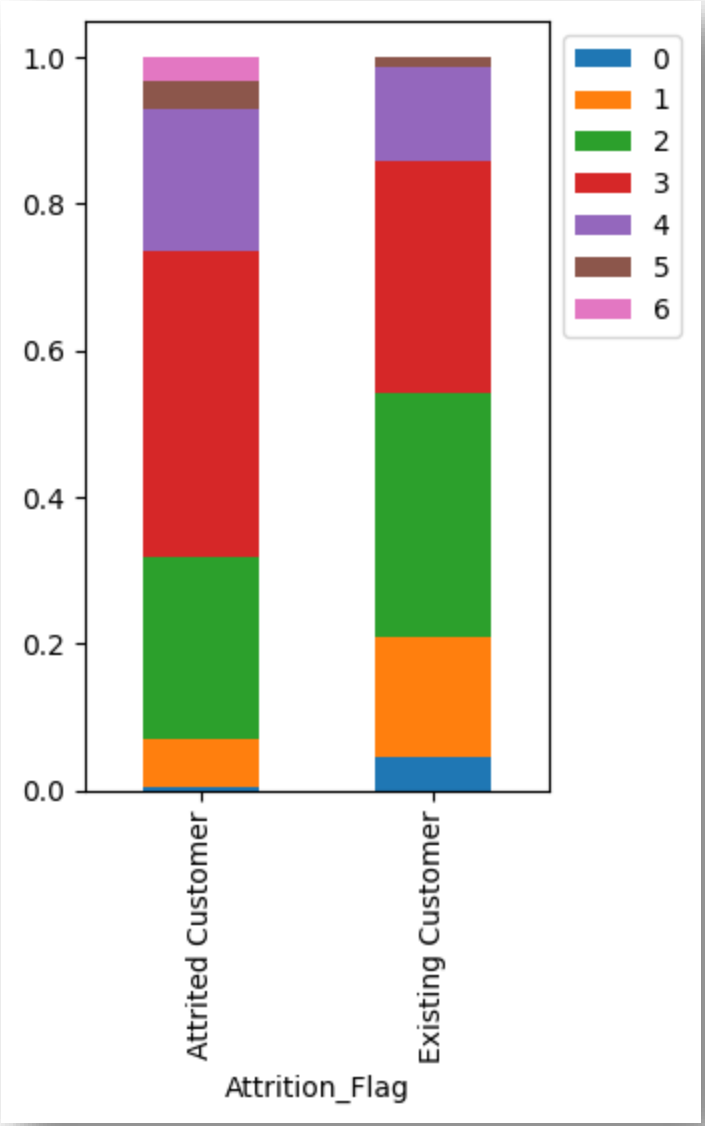

FIGURE 27

- **Insights based on the Stacked Bar plot of 'Contacts Count 12 mon':**

- The attrited customer group has more diverse distribution across the categories compared to the existing customer group, where certain categories dominate. For e.g. Categories 2,3 and 4 have higher proportions in both the groups, but category 4 is notably more prevalent among attrited customer.
- Category 4 makes up a significant portion of attrited customer group indicating that this category might be associated with a higher likelihood of attrition.
- The existing customer group has a higher proportion in the lower categories (1,2) whereas these categories are less represented in the attrition customer group. This indicates that customers in these lower categories are more likely to remain with the bank.
- The attrited customer group include a small but notable proportion of categories 5 and 6, which are nearly absent in the existing customer group. This shows that these categories might be linked to specific factors that contribute to customer attrition.
- The legends include categories 5 and 6 which might require verification to ensure that they are correctly labeled and interpreted.
- This analysis indicates that certain categories are more prevalent among customers who have left.

- **STACKED BAR PLOT COMPARES THE ‘MONTHS INACTIVE 12 MON’ BETWEEN ‘ATTRITION CUSTOMER’ & ‘EXISTING CUSTOMER’:**

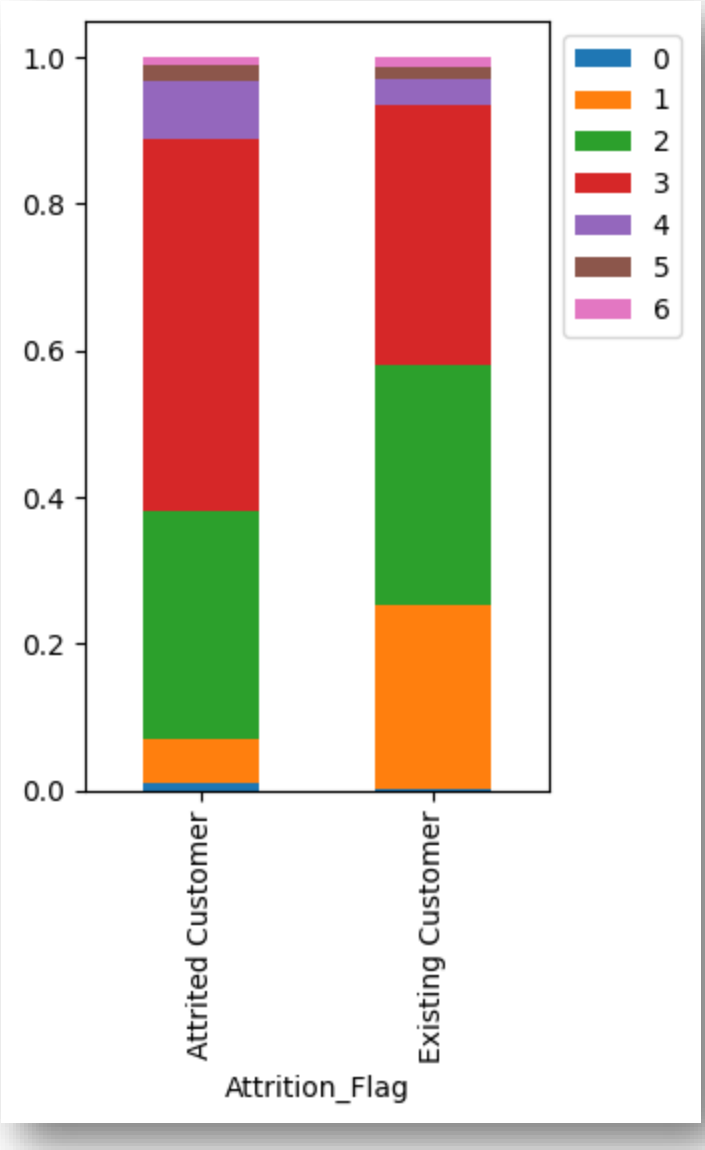

FIGURE 28

➢ Insights based on the Stacked Bar plot of 'Months Inactive 12 mon':

- Categories 2 and 3 dominate in both the customer groups. However, category 3 appears slightly more prominent among attrited customer which could suggest that individuals in this category might be at higher risk of attrition.
- Category 1 has a smaller representation in the attrited customer group compared to the existing customer group. This shows that customers in the category 1 are more likely to remain with the bank.
- The higher categories (4,5,6) are present in both the groups, with a slightly larger proportion in the attrited customer group. This might suggest that as the customer move into these higher categories, their likelihood of attrition increases though the differences are not very pronounced.
- Category 0 is barely present in either group showing that this category is not important factor in customer segmentation for attrition analysis.
- Despite some differences, the overall distribution across the categories is quite similar for both attrited and existing customer. This suggests that factors outside of these categories might also be influencing customer attrition.

- **STACKED BAR PLOT COMPARES THE 'TOTAL RELATIONSHIP COUNT' BETWEEN 'ATTRITION CUSTOMER' & 'EXISTING CUSTOMER':**

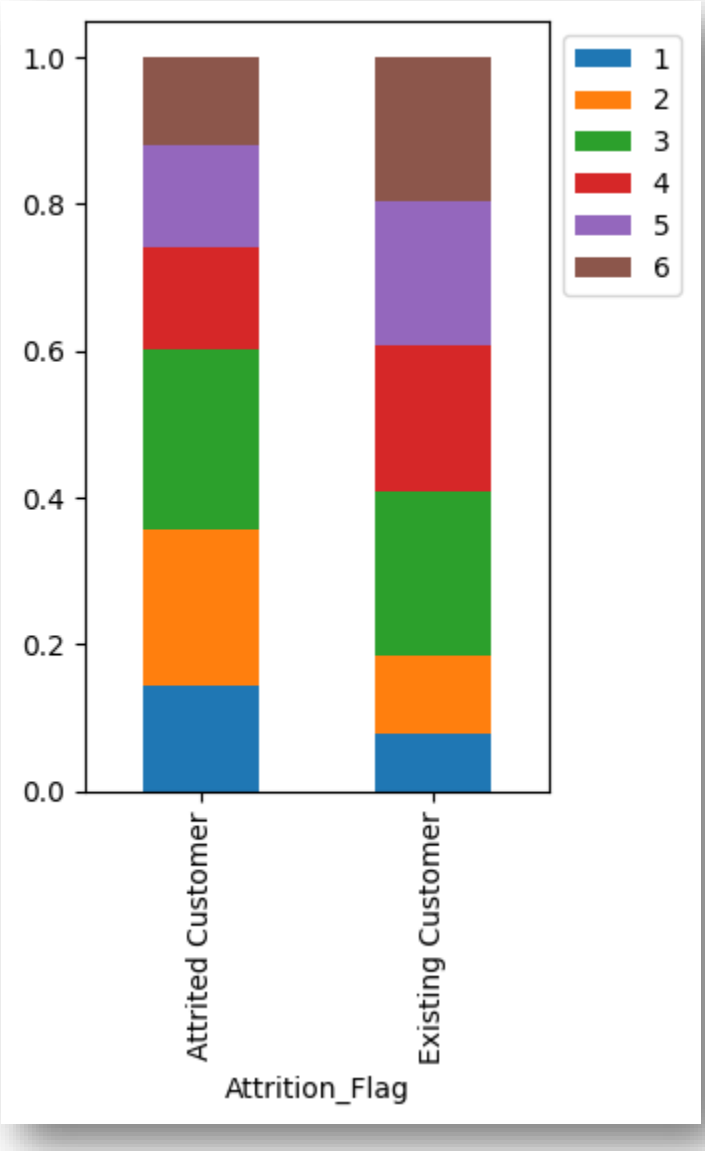

FIGURE 29

➢ Insights based on the Stacked Bar plot of 'Total Relationship Count':

- The distribution of categories is very similar between attrited customer and existing customer groups. This shows that the categorical variable represented in this plot may not be a strong differentiator between the customers who leave and those who stays.
- In both the customer groups, category 6 has the highest representation. This suggests that the majority of customers fall into this category regardless of whether they stay or leave.
- Categories 3 and 4 occupy an important portion of the distribution in both the groups, slightly less than category 6, but more than categories 1,2 and 5. This indicates that these middle categories are also important but not as dominant as category 6.
- Categories 1 and 2 have the smallest representation in both the groups, suggesting that fewer customers fall into these categories. Their consistent proportions across both the groups suggests that they are equally likely to be retained or attrited.
- Category 5 has a noticeable presence but is not as prominent as categories 3,4,6. This indicates it represents a smaller but still significant portion of the customer base.

- Overall, the similarity in distribution across categories suggests that this categorical variable alone may not provide strong predictive power for customer attrition. Other factors need to be considered to understand what drives customer behavior more effectively.

♦ **STACKED BAR PLOT COMPARES THE ‘DEPENDENT COUNT’ BETWEEN ‘ATTRITION CUSTOMER’ & ‘EXISTING CUSTOMER’:**

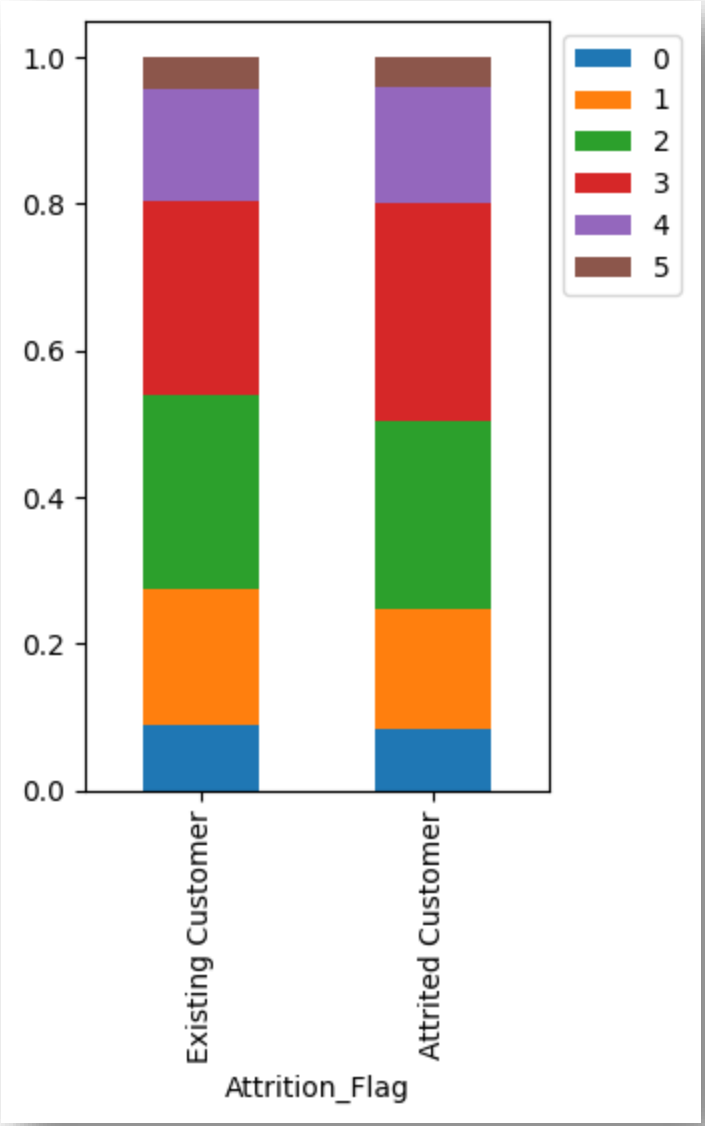

FIGURE 30

- **Insights based on the Stacked Bar plot of 'Dependent Count':**

- The distribution of categories across existing customer and attrition customer groups is nearly identical. This shows that the categorical value shown here does not significantly differentiate between the customers who stay and those who leave.
- Categories 2, 3 and 4 make up the majority of the distribution for both the groups. This suggests that these categories represent the bulk of the customer base, whether they are retained or attrited.
- Categories 0 and 1 have the smallest presence in both the customer groups. Their minimal representation suggests that fewer customer belong to these categories and they have a similar likelihood of staying or leaving.
- Category 5 has a noticeable but small presence in both the groups. Its consistent representation across both groups shows it does not strongly correlate with customer attrition.
- The uniform distribution of these categories between the 2 groups highlight that this categorical variable alone may not be a strong predictor of whether a customer will leave or remain.
- Overall, the plot shows that there is little or no difference in the categorical distribution between the existing and attrited customers, implying that other variables may play more critical role in understanding customer attrition.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'TOTAL REVOLVING BAL' w.r.t 'ATTRITION FLAG':

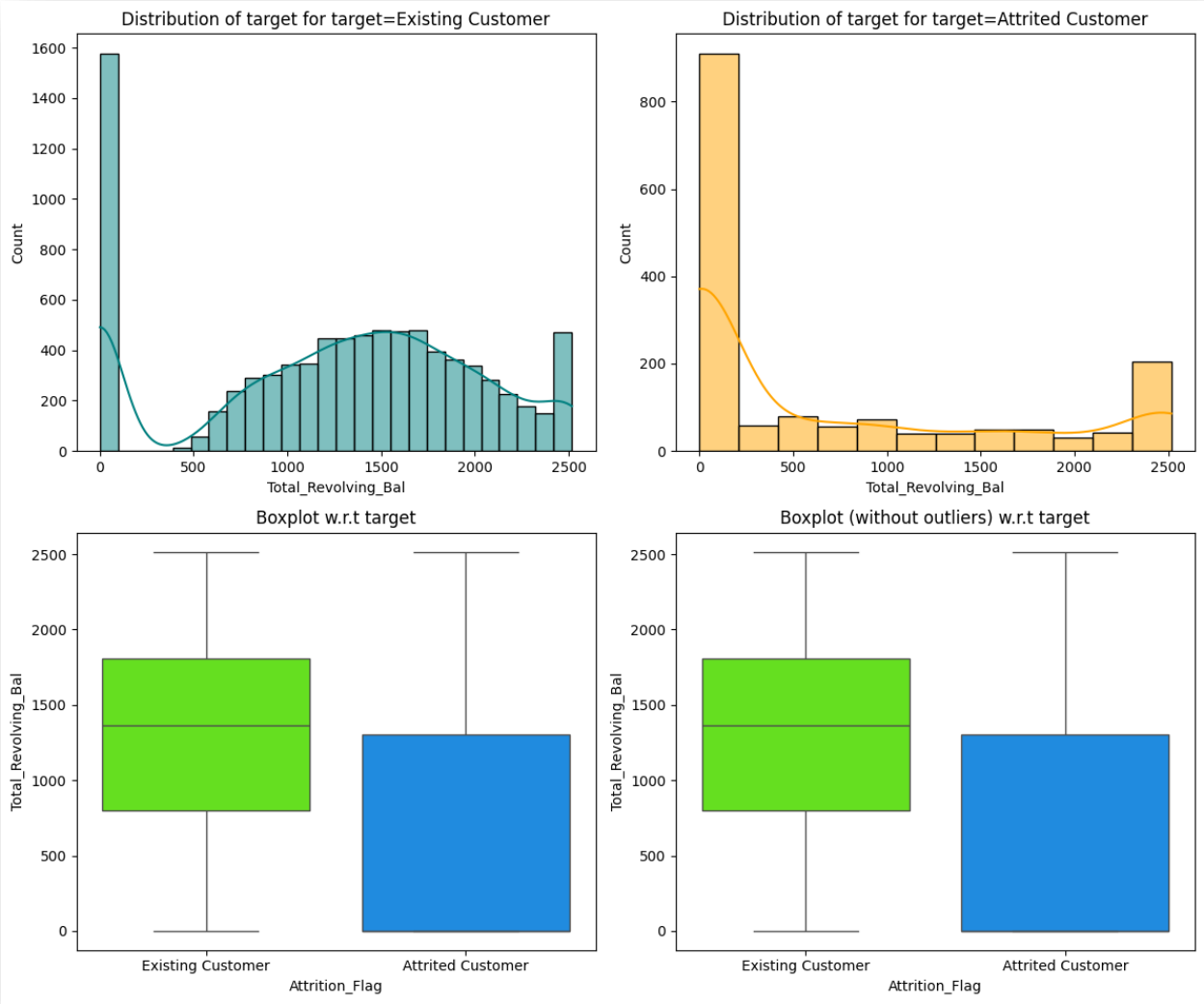

FIGURE 31

➢ Insights based on the Distribution and Bar plot of 'Total_Revolving_Bal':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- The distribution of 'Total Revolving Bal' for existing customer is more spread out and relatively uniform between 500 and 2,000 with noticeable peaks at 0 and 2,500.
- A significant number of existing customers have a revolving balance of 0, indicating either no revolving debt or full payment of balances regularly.
- There is also a smaller peak at upper limit (2,500), suggesting some existing customers max out their revolving balance.

2. Attrited Customer:

- The distribution for attrited customers is heavily skewed towards lower revolving balances, with a large peak at 0.
- The majority of attrited customers have very low revolving balances, with few having high balances.
- The distribution drops off quickly after the initial peak, indicating that customers with higher balances are less likely to leave.

→ BOX PLOTS:

1. Box plot with outliers:

- The box plot shows the median ‘Total Revolving Balance’ is higher for existing customers compared to the attributed customers.
- The IQR is also larger for existing customers indicating more variability in their revolving balances.
- The lower quartile for attrited customer is at or near zero reinforcing the observation that many leaving customers have very low balances.

2. Box plot without outliers:

- Similar trends are observed without outliers, with existing customers having higher median balances and a wider IQR compared to attrited customers.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR ‘CREDIT LIMIT’ w.r.t ‘ATTRITION FLAG’:

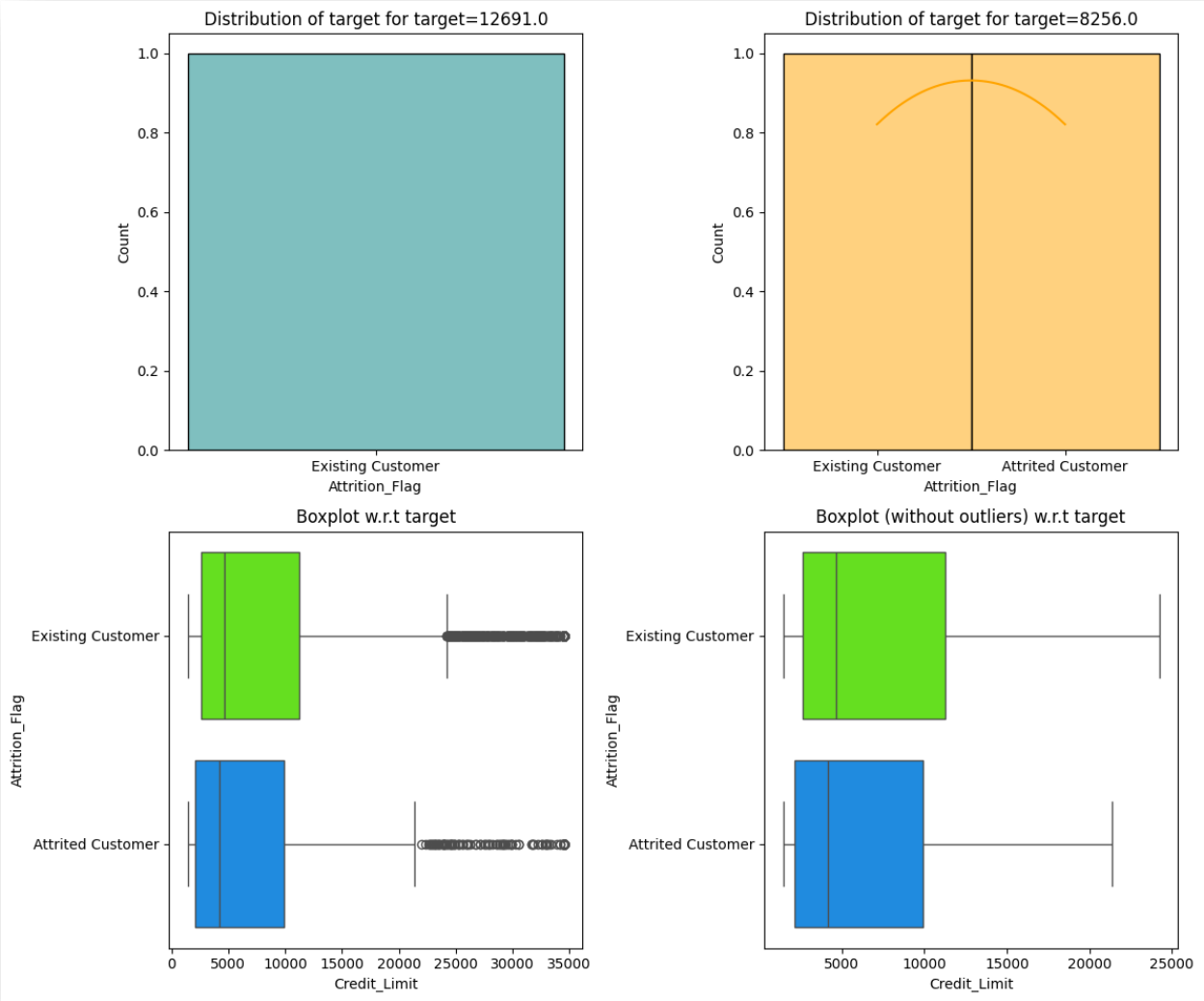

FIGURE 32

➢ Insights based on the Distribution and Bar plot of 'Credit_Limit':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- The distribution for existing customer with a specific ‘Credit Limit’ value of 12,691. This suggests a uniform distribution for this specific limit within the existing customer group, possibly due to specific credit or product offerings.

2. Attrited Customer:

- For attrited customer with a ‘Credit Limit’ value of 8,256, the distribution shows a slight curve, indicating that the credit limit might be more common or preferred among attrited customers.
- The distribution is uniform, indicating no significant variation among attrited customers for their credit limit.

→ BOX PLOTS:

1. Box plots with outliers:

- The box plot shows the median ‘Credit Limit’ for existing customer is higher than that of attrited customer.
- There is a wider range of credit limits for existing customer with more outliers on the higher end, indicating some on the higher end, indicating some customers have exceptionally high credit limits.

- Attrited customers have a smaller IQR and fewer high limit outliers, showing that they generally have lower credit limits.

2. Box plot without outliers:

- When outliers are removed, the trends remain similar: existing customers have a higher median credit limit and a wider range of limits.
- Attrited customers continue to show a lower median credit limit, indicating that higher credit limits may correlate with customer retention.
- Higher credit limit appears to be associated with existing customers while lower credit limits are more common among attrited customers. This suggests that customers with higher credit limits are more likely to stay, potentially due to better financial understanding.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'CUSTOMER AGE' w.r.t 'ATTRITION FLAG':

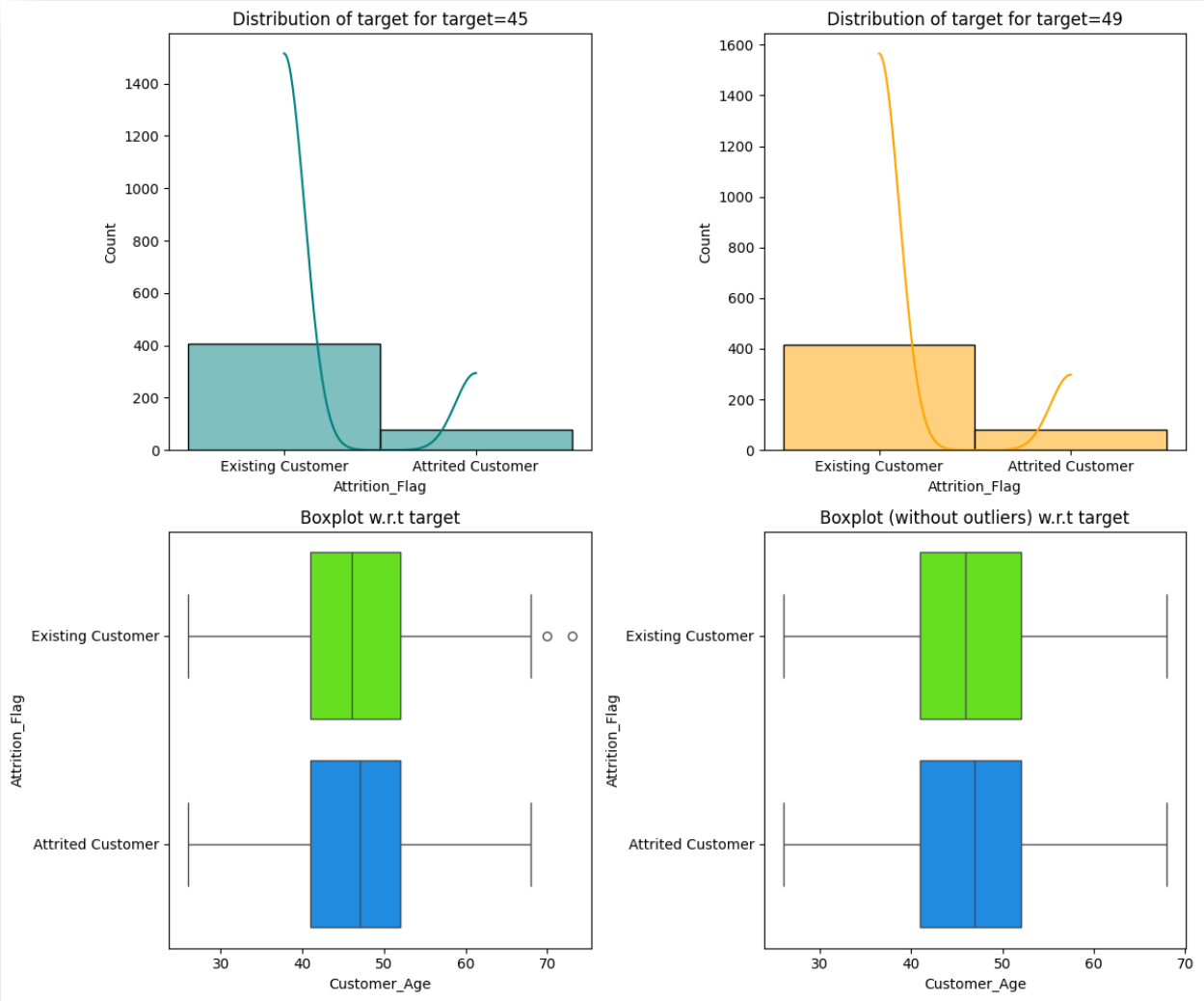

FIGURE 33

➢ Insights based on the Distribution and Bar plot of 'Customer_Age':

→ DISTRIBUTION PLOTS:

1. Age 45:

- Most existing customer fall at the age of 45 with very few attrited customer at this age. This suggests that age 45 is a point where customers are generally more satisfied or stable, leading to lower attrition rates.
- There is a sharp decline in attrition as we observe customer aged 45, indicating that this age group has a strong retention rate.

2. Age 49:

- Similar to age 45, most customers who are 49 years old tend to be existing customers. There is also a sharp decline in the number of attrited customer at this age.
- The pattern is consistent with the age 45 group, where the majority of customers at this age are retained.

→ BOX PLOTS:

1. Box plot with outliers:

- The median customer age for both existing and attrited customers is around 47 – 50 years.
- The IQR for both existing and attrited customers are similar, indicating that age alone may not be a strong differentiator for attrition.

- There are few outliers in the existing customers group at older ages (around 65 – 70 years) suggesting that some older customers remain loyal.

2. Box plot without outliers:

- The box plot without outliers mirrors the insights from the previous plot, where the age distribution for existing and attrited customer is quite similar.
- The absence of significant differences in the IQR and medians between the two groups further shows that age might not be a critical factor in determining whether a customer will stay or leave.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'TOTAL TRANS CT' w.r.t 'ATTRITION FLAG':

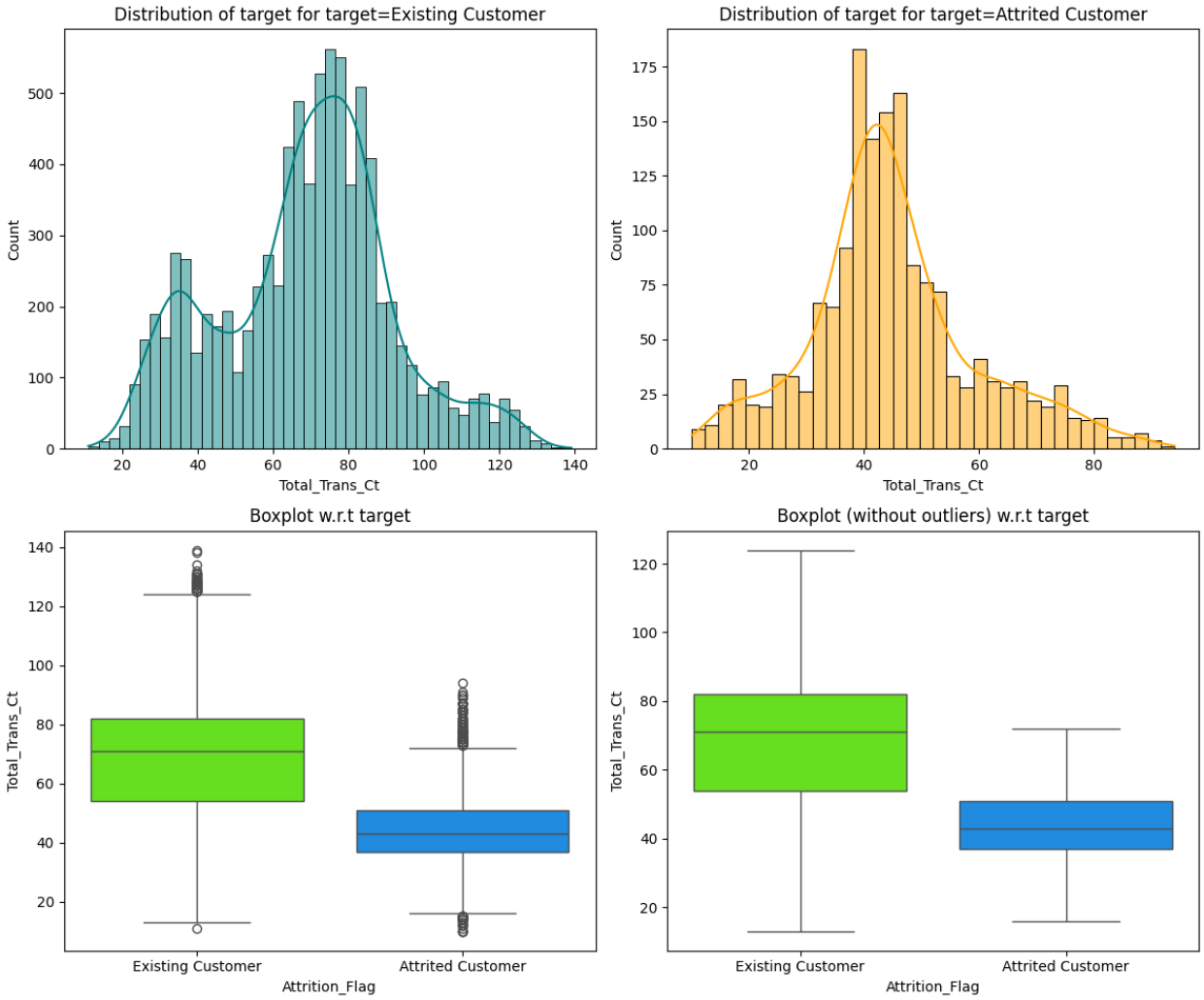

FIGURE 34

➢ Insights based on the Distribution and Bar plot of 'Total_Trans_Ct':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- Existing customer tend to have a higher total transaction count, with a distribution peak around 60 – 80 transactions.
- The distribution is relatively symmetric, indicating the existing customer generally maintain a consistent level of transactions.
- A noticeable number of customers have transaction counts as high as 120 – 140, suggesting high engagement among some existing customer.

2. Attrited Customer:

- Attrited customer have a low transaction count overall, with a distribution peaking around 40- 60 transactions.
- The distribution is also relatively symmetric but skewed towards lower transaction counts compared to existing customers.
- There are very few attrited customers with high transaction counts (over 80), indicating that lower engagement in terms of transaction counts might be associated with a higher likelihood of attrition.

→ BOX PLOTS:

1. Box plot with outliers:

- The median transaction count for existing customer is higher (around 65 – 70) compared to attrited customer (around 50).
- The IQR for existing customer is also wider, suggesting a broader spread of transaction activity among retained customers.
- There are several outliers in the existing customer group with very high transaction counts.

2. Box plot without outliers:

- Even without outliers, the trend remains the same: existing customer have higher transaction counts and a wider IQR compared to the attrited customer.
- The absence of significant outliers in the attrited customer group suggests that customers who leave tend to have consistently lower transaction activity.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'TOTAL TRANS AMT' w.r.t 'ATTRITION FLAG':

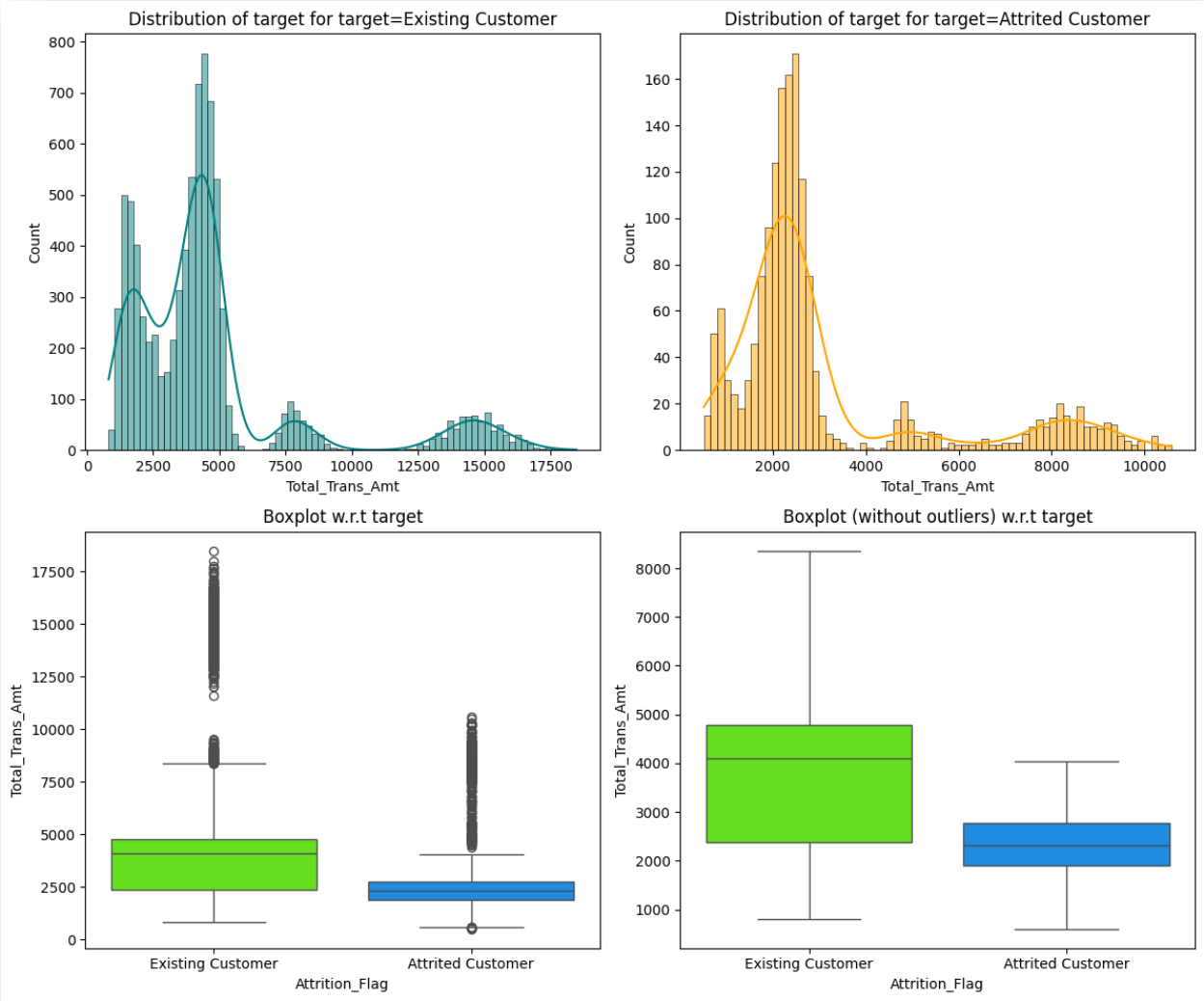

FIGURE 35

➢ Insights based on the Distribution and Box plot of 'Total_Trans_Amt':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- The distribution is skewed towards lower transaction amount with significant peak around 2,000 – 4,000 range.
- There are several smaller peaks observed at higher transaction amounts indicating a subset of existing customers who transact in larger sums.
- The density plot suggests multiple transaction behaviors with existing customer group.

2. Attrition Customer:

- The distribution is also skewed towards lower transaction amounts, but it is more concentrated around 1,000 – 3,000 range.
- There are few customers with high transaction amounts compared to the existing customers.

→ BOX PLOTS:

1. Box plot with outliers:

- Existing customers tend to have a higher median 'Total Trans Amt' than attrited customers.
- The spread of transaction amounts is broader for existing customer suggesting more variability in their spending behaviors.

- A significant number of outliers are present in both the groups, with existing customer having outliers extending to very high transaction amounts.

2. Box plot without outliers:

- The box plot without outliers reaffirm that existing customer generally transact higher amounts than attrited customers.
- The IQR is wider for existing customers showing more diverse transaction behavior.
- Attrited customer have more compact IQR indicating more consistency in their transaction amounts.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'TOTAL CT CHNG Q4 Q1' w.r.t 'ATTRITION FLAG':

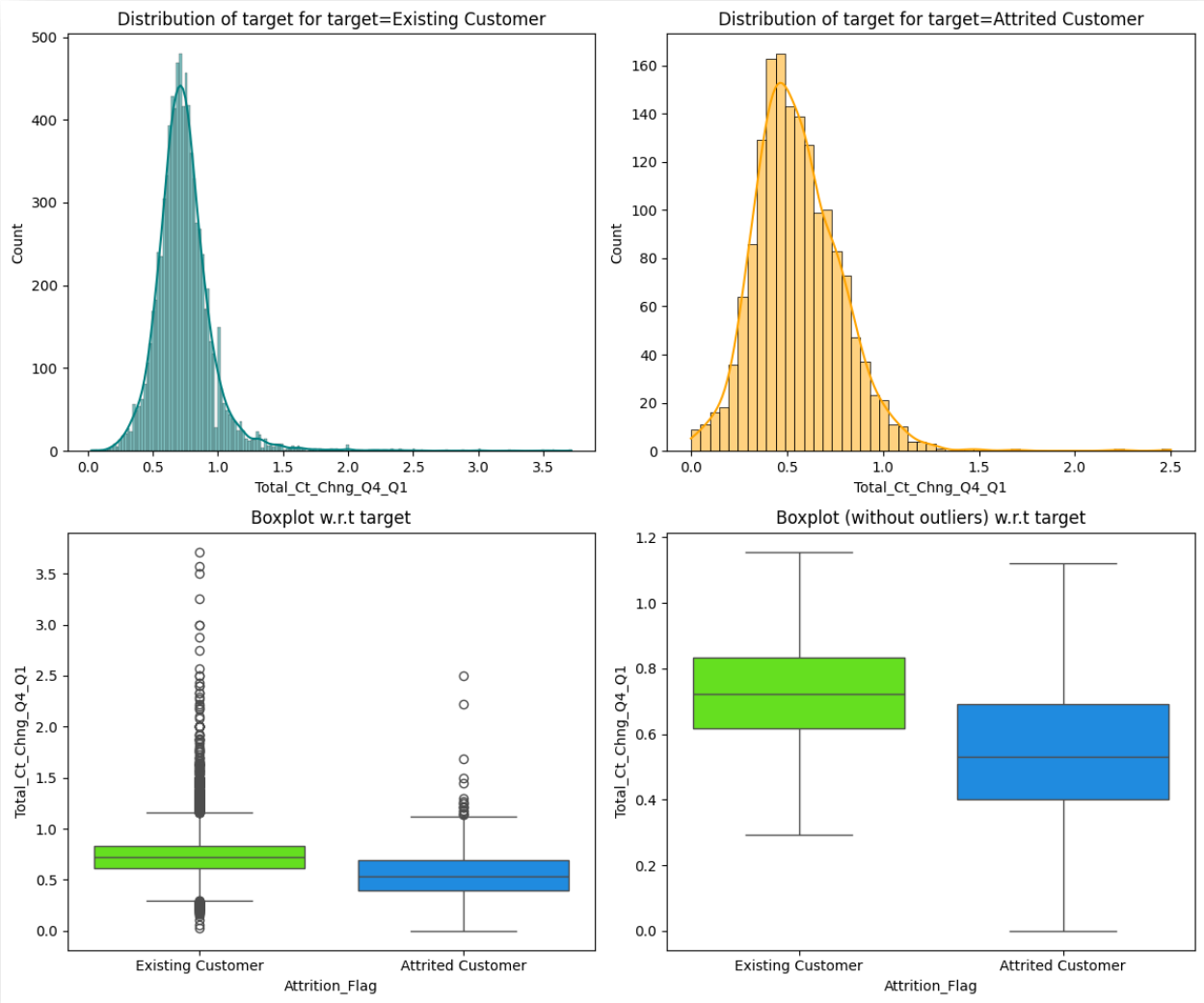

FIGURE 36

➢ Insights based on the Distribution and Bar plot of 'Total_Ct_Chng_Q4_Q1':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- The distribution is relatively narrow and centered around 0.9 to 1.0, suggesting that more existing customer show little or no change in number of transactions between Q4 and Q1.
- A small proportion of existing customer exhibit a significant increase in the transaction count, with the distribution tailing off to the right.

2. Attrition Customer:

- The distribution is slightly skewed to the left with most attrition customer having a 'total_ct_chng_Q4_Q1' value around 0.5 to 0.8.
- This indicates that attrition customers tend to have a reduction in transaction count between Q1 and Q4 with fewer customers experiencing an increase.

→ BOX PLOTS:

1. Box plot with outliers:

- Existing customers generally have a higher median change in transaction counts compared to attrited customers.
- There are more outliers among existing customers who show a significant positive change in transaction count, suggesting some existing customers have a considerably increased their transaction activity.

- Attrition customer have fewer extreme outliers and most of them are on the lower end, indicating a decline in transaction activity.

2. Box plot without outliers:

- The box plot shows that IQR for existing customers is higher than that for the attrited customers.
- The median 'Total_Ct_Chng_Q4_Q1' for existing customer is around 0.8 which is higher than the median for attrition customer (around 0.6).
- This suggests that a decrease in transaction count is more common among attrited customer, while existing customer tend to maintain or slightly increase their transaction activity.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'AVG. UTILIZATION RATIO' w.r.t 'ATTRITION FLAG':

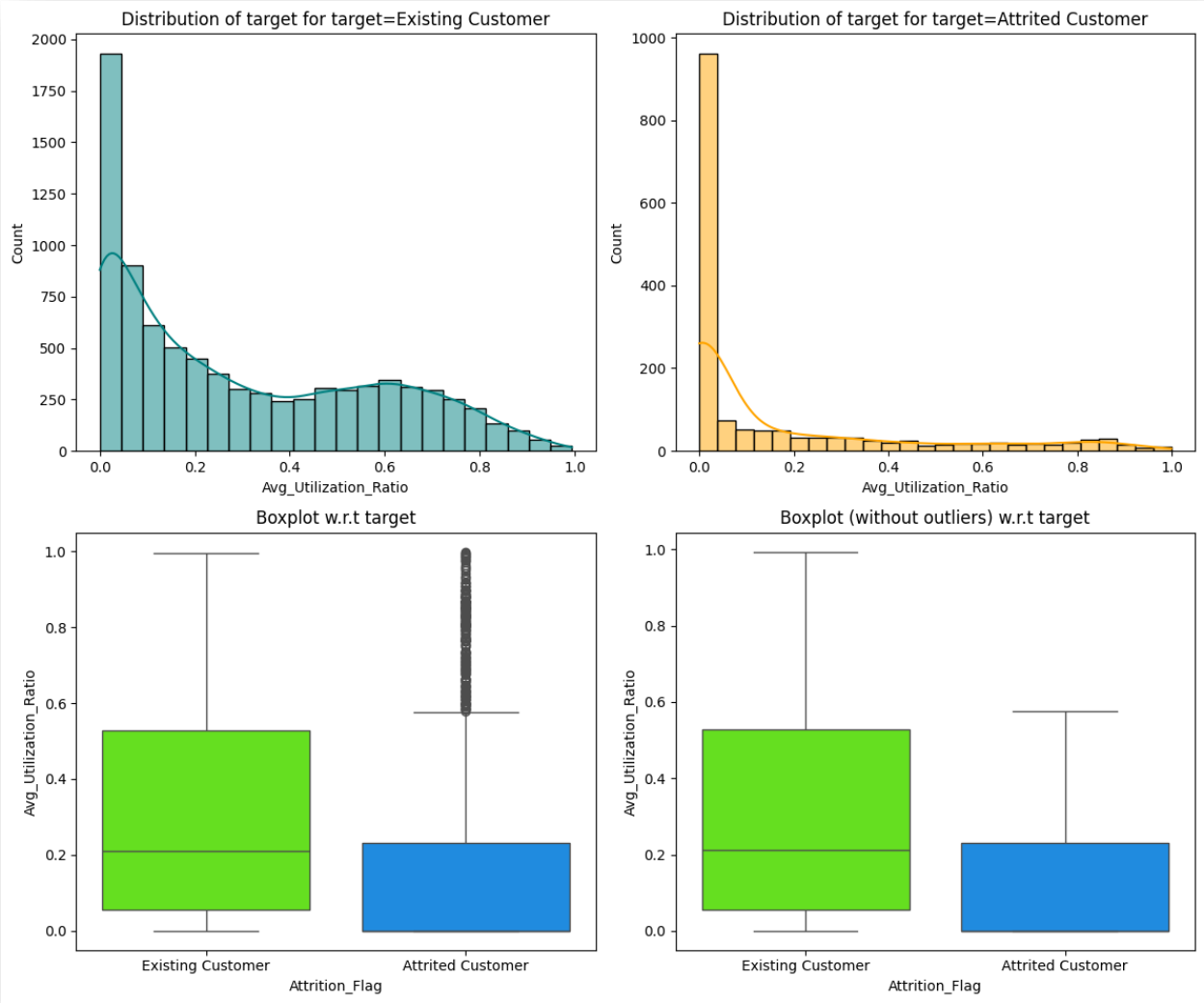

FIGURE 37

➢ Insights based on the Distribution and Bar plot of 'Avg_Utilization_Ratio':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- The distribution is right skewed with majority of existing customer having a low utilization ratio, particularly around 0.0 to 0.1.
- The distribution gradually decreases as the utilization ratio increases with fewer customers having a high utilization ratio.
- There is a notable spread across a wide range of utilization ratios, indicating variability in how existing customers use their available credit.

2. Attrited Customer:

- The distribution is also right- skewed but it is more sharply concentrated at very low utilization ratios, especially around 0.0 to 0.1.
- There are few attrited customers with high utilization ratios indicating that most attrited customers used a very small portion of their available credit.
- The tail of the distribution is shorter compared to existing customer meaning fewer attrition customer exhibit high utilization behavior.

→ BOX PLOTS:

1. Box plot with outliers:

- The median utilization ratio for existing customer is higher than that of attrited customer indicating that on average, existing customer utilize more of their available credit.
- Existing customer shows a broader range of utilization ratio with a significant number of outliers, suggesting that some existing customers are using a very high portion of their available credit.
- The tail of the distribution is shorter as compared to existing customer, meaning fewer attrition customer exhibit high utilization behavior.

2. Box plot without outliers:

- The box plot shows that the IQR for existing customer is wider, meaning there is more variation in utilization behavior among them.
- The median for existing customer is around 0.2, while for attrition customer it is lower around 0.1.
- This suggests that attrition customer generally maintain lower utilization of their credit, which could be linked to their decision to leave.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'MONTHS ON BOOK' w.r.t 'ATTRITION FLAG':

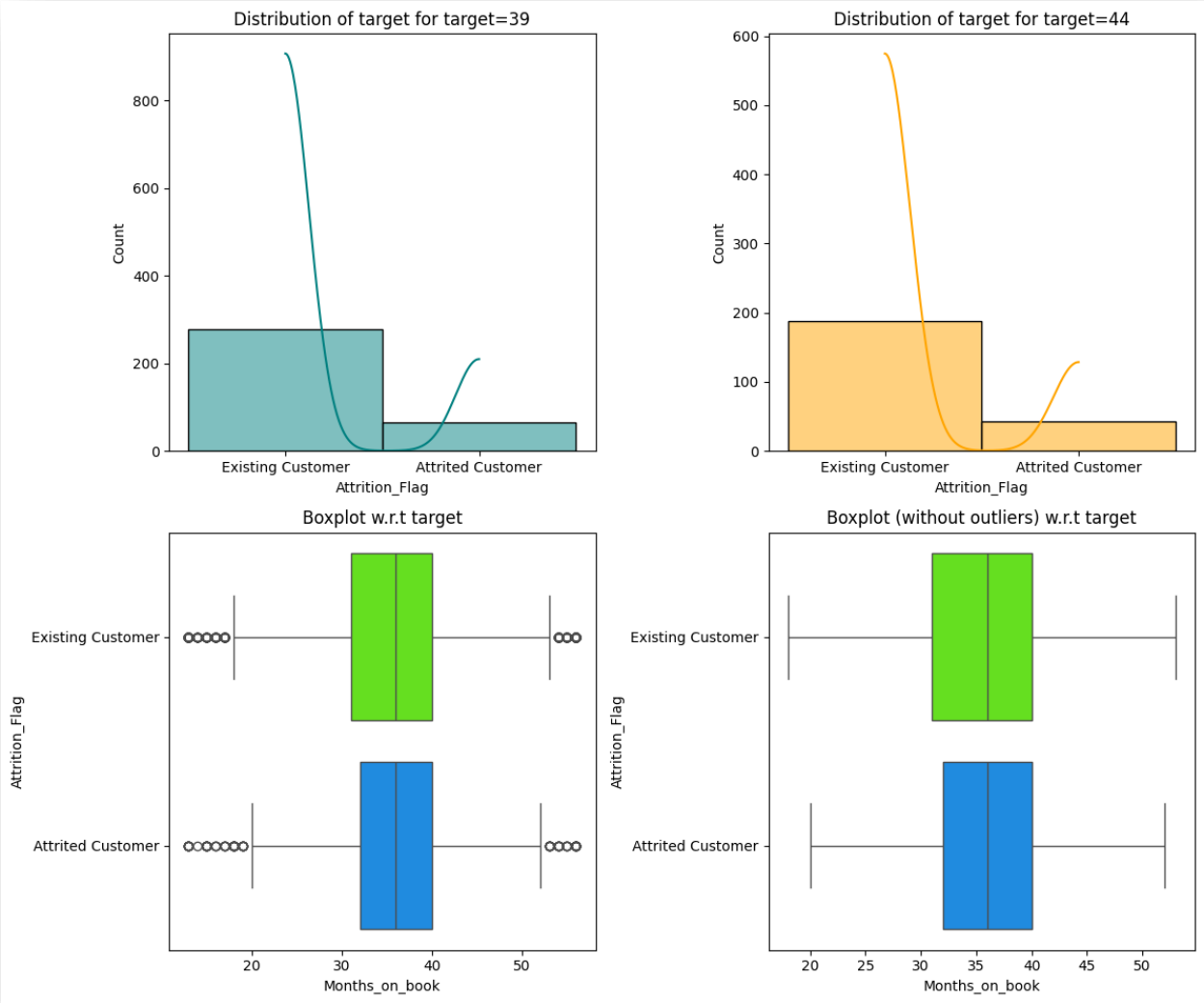

FIGURE 38

➢ Insights based on the Distribution and Box plot of 'Months_on_Book':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- For customers who have been with the bank for 39 months, the number of existing customers is higher than that of the attrition customer.
- The distribution has a sharp decline, suggesting that the majority of the customers at this tenure remain with the bank, with only few leaving.

2. Attrition Customer:

- For Customers who have been with the bank for 44 months, a similar pattern is observed, but the difference between existing customer and attrition customer is smaller compared to the 39 months (about 3 and a half years) group.
- The curve suggests a slightly higher attrition rate, though existing customers still dominate.

→ BOX PLOTS:

1. Box plot with outliers:

- The Box plot shows the distribution 'Months_on_Book' variable for existing and attrited customers including outliers.

- It appears that the median number of months on book is higher for existing customers compared to attrition customer.
- The distribution for existing customers is slightly more spread out, while attrition customer has a tighter distribution around the lower end.

2. Box plot without outliers:

- The box plot shows that existing customer generally have a tighter median tenure (Months on book) compared to attrited customer.
- The IQR is also higher for the existing customer, indicating that those who stay longer are likely to continue staying whereas, customers who leave tend to have a shorter tenure.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'TOTAL REVOLVING BAL' w.r.t 'ATTRITION FLAG':

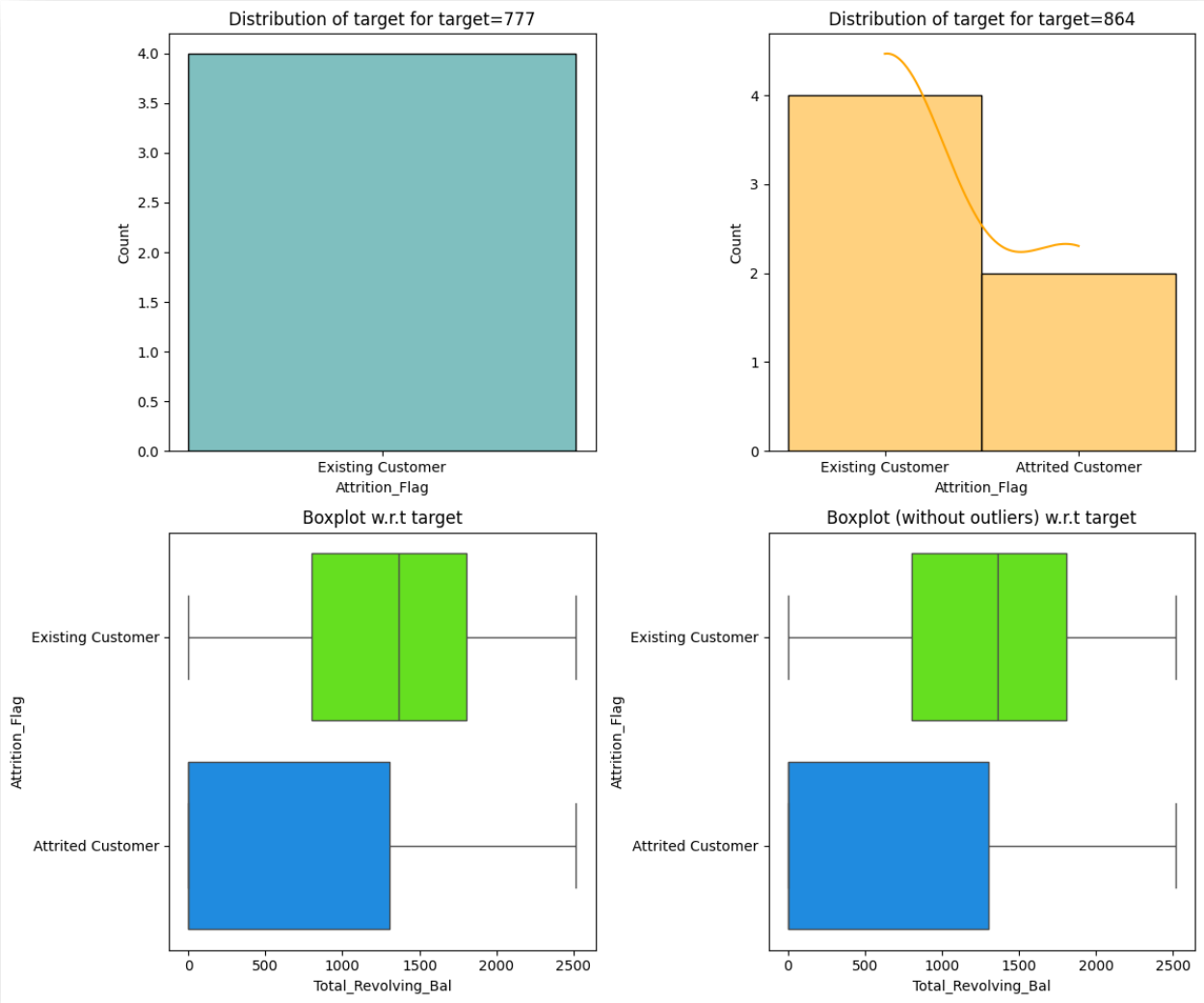

FIGURE 39

➢ Insights based on the Distribution and Bar plot of 'Total_Revolving_Bal':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- For customers with ‘Total_Revolving_Balance’ of 777, the entire group falls under existing customer with no attrition customer.
- This suggests that at this specific balance level, customers are likely to remain with the bank.

2. Attrition Customer:

- For customers with a total revolving balance of 864, both existing and attrition customers are present.
- However, there are more existing customer than attrited ones, though the difference is not as stark.
- This indicates that while customers at this balance level are still more likely to stay, there is a noticeable attrition rate

→ BOX PLOTS:

1. Box plot with outliers:

- Box plot shows that the distribution of ‘Total_Revolving_Balance’ for both existing and attrited customers.
- Existing customers have a higher total revolving balance with median around 1,500, whereas attrited customers have a lower median balance.

- The spread is wide for both the groups indicating a high variance in revolving balances.

2. Box plot without outliers:

- The box plot shows that existing customers tend to have higher revolving balances compared to the attrited customers.
- The median balance for existing customers is higher and the overall distribution is more concentrated around the middle, whereas attrited customer have a slightly lower and more dispersed distribution.

♦ DISTRIBUTION PLOTS & BOX PLOTS FOR 'AVG OPEN TO BUY' w.r.t 'ATTRITION FLAG':

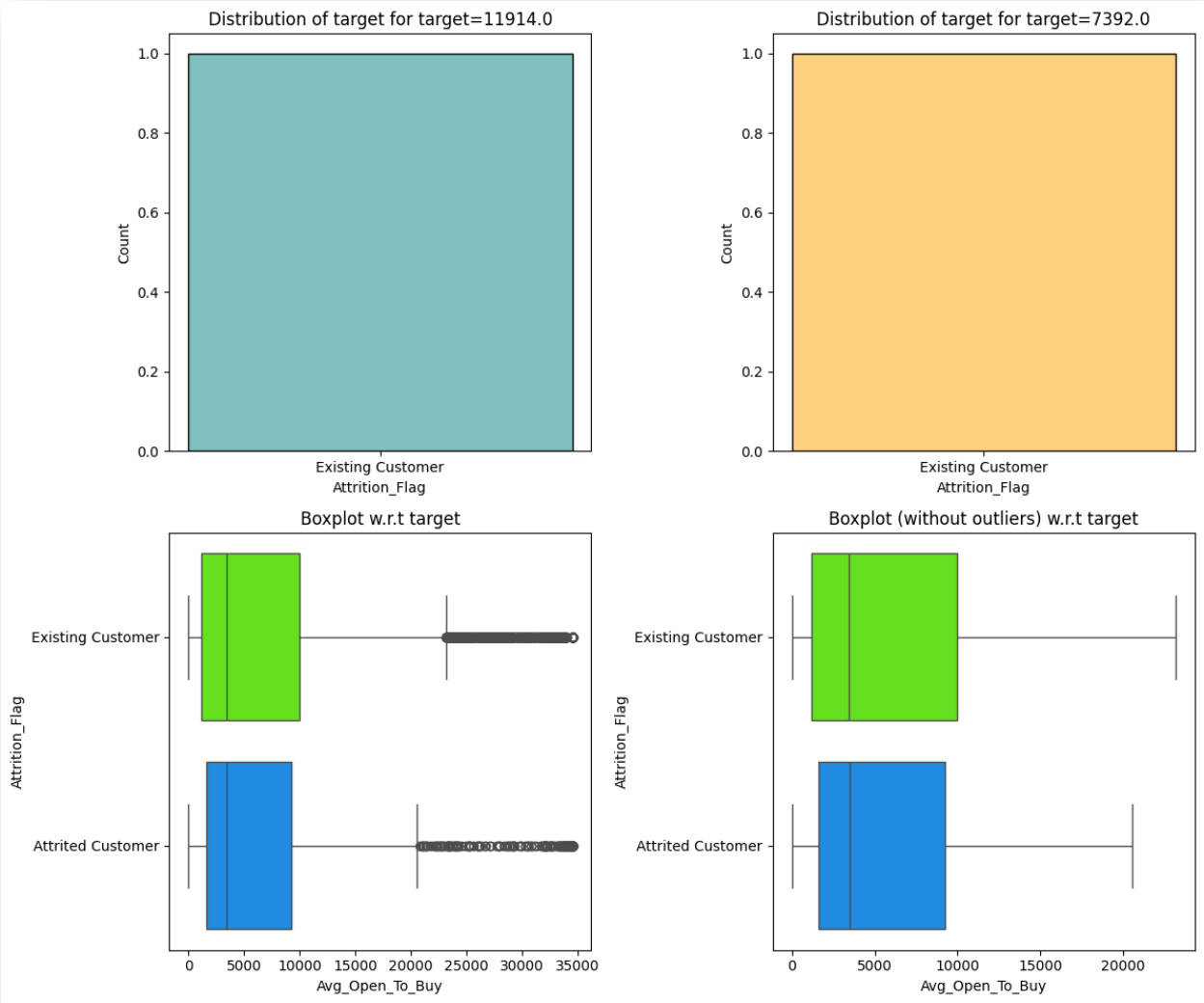

FIGURE 40

➢ Insights based on the Distribution and Bar plot of 'Avg_Open_to_Buy':

→ DISTRIBUTION PLOTS:

1. Existing Customer:

- For customers with an average open to buy amount of 11,914, all of them are existing customers.
- There is no attrition customer in this group, suggesting that customers with this level of available credit are more likely to stay.

2. Attrition Customer:

- For customer with an average open to buy amount of 7,392, all of them are also existing customers.
- This again suggests a strong retention rate among customers with available credit at this level.

→ BOX PLOTS:

1. Box plot with outliers:

- The box plot shows the distribution of 'Average_Open_to_Buy' for existing and attrited customers.
- Existing customers have a higher median and a wider range of available credit while attrited customer have a slightly lower median.
- The plot also shows many outliers among existing customers with a very high available credit (above 20,000)

which suggests that some customers with high credit availability are particularly likely to stay.

2. Box plot without outliers:

- The box plot shows that existing customer have a higher median ‘Avg_Open_to_Buy’ compared to attrited customers.
- The distribution for both the groups becomes more concentrated, but the difference in median remains, reinforcing the notion that higher available credit is associated with customer retention.

❑ DATA PREPROCESSING

Outlier Detection

♦ VARIOUS FEATURES IN PREDICTING CUSTOMER ATTRITION, WITH NUMERICAL VALUES REPRESENTING FEATURE IMPORTANCE SCORES:

	0
CLIENTNUM	0.000
Customer_Age	0.020
Dependent_count	0.000
Months_on_book	3.812
Total_Relationship_Count	0.000
Months_Inactive_12_mon	3.268
Contacts_Count_12_mon	6.211
Credit_Limit	9.717
Total_Revolving_Bal	0.000
Avg_Open_To_Buy	9.509
Total_Amt_Chng_Q4_Q1	3.910
Total_Trans_Amt	8.848
Total_Trans_Ct	0.020
Total_Ct_Chng_Q4_Q1	3.891
Avg_Utilization_Ratio	0.000

TABLE 2

➢ Insights based on the above table:

1. Key influencers:

- Contacts_Count_12_mon:

a) The number of contacts made in the past 12 months is the most significant factor among the given features. This suggests that higher customer engagement or frequent communication is crucial in understanding customer behavior and possibly predicting attrition.

- Credit_Limit:

a) Both the ‘total credit’ and ‘avg open to buy’ amounts are also highly influential. Customer with higher credit availability is more likely to engaged or valued, which could reduce the risk for attrition.

- Total_Trans_Amount:

a) The total transaction amount is another significant factor indicating that higher spending customers are more likely to be retained, due to the value they bring to the bank.

2. Moderate Influencers:

- Total_Amt_Chng_Q4_Q1 and Total_Ct_Chng_Q4_Q1:

a) Changes in transaction amounts and counts from one quarter to another are moderately important. Significant changes could signal shifts in customer behavior, such as reduced spending which correlates with higher attrition risk.

- Months_On_Book:

a) The length of time a customer has been with the bank also plays a moderate role, with longer- tenured customers potentially being more loyal.

- Months_Inactive_12_mon:

a) The number of inactive months in the past year is moderately influential, which is expected as inactivity often precedes attrition.

3. Less Significant Factors:

- Customer_Age and Total_Trans_Ct:

a) These features have minimal importance indicating that the age of the customer and total number of transactions might not be strong indicators of attrition on their own.

- Total_Relationship_Count, Total_Revolving_Balance, Avg_Utilization_Ratio, Dependent_Count:

a) These features have zero importance suggesting that they may not contribute significantly to predicting customer attrition in this specific context.

❑ TRAIN TEST SPLIT

♦ TABLE SHOWING NUMBER OF MISSING VALUES FOR EACH FEATURES:

	0
Attrition_Flag	0
Customer_Age	0
Gender	10127
Dependent_count	0
Education_Level	10127
Marital_Status	10127
Income_Category	10127
Card_Category	10127
Months_on_book	0
Total_Relationship_Count	0
Months_Inactive_12_mon	0
Contacts_Count_12_mon	0
Credit_Limit	0
Total_Revolving_Bal	0
Avg_Open_To_Buy	0
Total_Amt_Chng_Q4_Q1	0
Total_Trans_Amt	0
Total_Trans_Ct	0

TABLE 3

- **Insights based on the above table of missing values for different features:**

1. Data Completeness:

- Most features have no missing values. This indicates a high level of data completeness for these features which is advantageous for modelling and analysis since it reduces the need for imputation or data cleaning strategies.
- Missing data:

a) Education level has 1,519 missing values.

b) Marital Status has 749 missing values.

2. Impact of missing data:

- Missing values in education level might be problematic, especially if the variable is considered important for predicting customer behavior.
- Missing data in marital status is also important but to lesser extent. Missing values in this feature could also affect the model's performance if marital status is an important predictor.

❑ MODEL BUILDING

♦ MODEL BUILDING ON ORIGINAL DATA

```
Training Performance:

Bagging Classifier: 0.9753846153846154
Random Forest: 1.0
Gradient Boosting: 0.8923076923076924
AdaBoost: 0.8684615384615385
XGBoost: 1.0

Validation Performance:

Bagging Classifier: 0.8513513513513513
Random Forest: 0.8108108108108109
Gradient Boosting: 0.8513513513513513
AdaBoost: 0.8378378378378378
XGBoost: 0.9054054054054054
```

TABLE 4

- **Insights based on the Training and Validation Performances:**

1. Overfitting Observations:

- Random Forest and XG Boost both achieved perfect score (1.0) on the training set. However, their validation performances dropped to 0.8108 and 0.9054 respectively. This suggests that these models are overfitting the training data – learning the training data too well but failing to generalize as effectively to the unseen data.

2. Generalization Performance:

- XG Boost shows the highest validation performances (0.9054) suggesting it generalizes better to unseen data compared to other models. Despite high training score, its relatively strong validation score suggests that it has a good balance between fitting the training data and generalizing to new data.
- Bagging Classifier and Gradient Boosting both have the same validation performance (0.8514), but bagging shows higher training score (0.9754 vs 0.8923). The smaller gap between training and validation scores in gradient have the same validation performance (0.8514), but bagging shows higher training score (0.9754 vs 0.8923). The smaller gap between training and validation scores in Gradient Boosting suggests it is more resilient to overfitting compared to Bagging.

3. Under Fitting Concerns:

- AdaBoost has the lowest training performance (0.8685) and slightly better validation performance (0.8378). This indicates that AdaBoost is not only overfitting, but it also might not be complex enough to fully capture the patterns in the training data, leading to under fitting.

4. Model Selection Considerations:

- XG Boost appears to be the best model given its highest validation accuracy.
- Bagging Classifier and Gradient Boosting offers a balance between training and validation performance making it

more viable option if looking for more reliable generalization.

- Random Forest's high overfitting suggests that tuning might be necessary to improve its generalization performance.
- AdaBoost requires more complex base learners or additional tuning to improve both training and validation performance.

♦ MODEL BUILDING ON OVER SAMPLED DATA

- Before oversampling, counts of label 'Yes': 1300
- Before oversampling, counts of label 'No': 6801
- After oversampling, counts of label 'Yes': 6801
- After oversampling, counts of label 'No': 6801
- After oversampling, shape of train_X: (13602, 14)
- After oversampling, shape of train_y: (13602)

♦ MODEL BUILDING – OVERSAMPLED DATA:

```
Training Performance:

Bagging: 0.9979414791942361
Random forest: 1.0
Gradient Boosting: 0.9810322011468902
AdaBoost: 0.9716218203205411
XGBoost: 1.0

Validation Performance:

Bagging: 0.8783783783783784
Random forest: 0.8783783783783784
Gradient Boosting: 0.8783783783783784
AdaBoost: 0.8648648648648649
XGBoost: 0.918918918918919
```

TABLE 5

➢ Insights based on the Training and Validation Performances:

1. Over fitting Concerns:

- Random Forest and XG Boost both achieved perfect training scores (1.0). The perfect training scores indicate that these models are highly overfitting to the training data. However, XG Boost manages to achieve the highest validation score, indicating it might be overfitting less than Random Forest, but some level of overfitting is still present.

2. Validation Performance Comparison:

- XG Boost has the highest validation performance (0.9189). This suggests that among all the models, XG Boost is best at generalizing to unseen data. This makes XG Boost a strong candidate for the final model despite its perfect training score. Especially because it outperforms the other models on the validation set.
- Bagging, Random Forest and Gradient Boosting have identical validation performances (0.8784). This uniform performance suggests that these models are capturing similar patterns in the data. However, given their high training performance especially Random Forest and Bagging, these models are also overfitting to some degrees.

3. AdaBoost Performance:

- AdaBoost shows the lowest validation performance (0.8649), and a slightly lower training performance (0.9716) compared to other models. AdaBoost's lower training score indicates that it is less prone to overfitting, its lower validation performance also shows that it might not be as effective in capturing the complexity of data compared to other models.

4. Overfitting Vs Under Fitting Balance:

- Bagging and Gradient Boosting achieve very high training performance (0.9979 and 0.9810 respectively) and identical validation performance (0.8784). This shows that

while they are learning well from the training data, they are not generalizing as well as XG Boost.

- The gap between training and validation performance in AdaBoost is smaller suggesting it is better balanced between overfitting and under fitting.

5. Model Selection Considerations:

- XG Boost is the best performing model on the validation set, making it the most promising choice. However, attention is needed to reduce overfitting through tuning on cross – validation to ensure robust performance on unseen data.

♦ MODEL BUILDING ON UNDER SAMPLED DATA

- Before under sampling, counts of label 'Yes': 1300
- Before under sampling, counts of label 'No': 6801
- After under sampling, counts of label 'Yes': 1300
- After under sampling, counts of label 'No': 1300
- After under sampling, shape of train_X: (2600, 19)
- After under sampling, shape of train_y: (2600)

```
Training Performance:

Bagging: 0.9915384615384616
Random forest: 1.0
Gradient Boosting: 0.9792307692307692
AdaBoost: 0.953076923076923
XGBoost: 1.0

Validation Performance:

Bagging: 0.918918918918919
Random forest: 0.9594594594594594
Gradient Boosting: 0.9594594594594594
AdaBoost: 0.9054054054054054
XGBoost: 0.972972972972973
```

TABLE 6

➢ Insights based on the Training and Validation Performances:

1. Strong Generalization by XG Boost:

- XG Boost not only achieves a perfect training score (1.0) but also the highest validation performance (0.9730). This suggests that XG Boost is effectively capturing the patterns in the training data while also generalizing well to unseen data. Its strong validation performance indicates that it is handling the complexity of data better than other models.

2. High Validation Performance by Random Forest and Gradient Boosting:

- Both Random Forest and Gradient Boosting shows identical validation performance (0.9595) which is slightly lower than XG Boost. While Random Forest achieves perfect training score, Gradient Boosting shows slightly lower training performance (0.9792) indicating that Gradient Boosting is less prone to overfitting compared to Random Forest. However, both the model shows strong generalization capabilities.

3. Bagging Balance between Training and Validation:

- Bagging achieves a high training performance (0.9915) and a strong validation performance (0.9189). The small gap between training and validation scores suggests a good balance between fitting the training data and generalizing to new data.

4. AdaBoost Performance:

- AdaBoost shows the lowest training performance (0.9531) and second lowest validation performance (0.9054) among the models. This suggests that AdaBoost might be under fitting, failing to capture the underlying patterns in the training data. However, AdaBoost is less prone to overfitting.

5. Model Selection Consideration:

- XG Boost is the best – performing model on the validation set, making it most suitable.
- Random Forest and Gradient Boosting also performs well and could be considered.

- Bagging and AdaBoost is not as strong but can be useful in situation where reducing the overfitting or complexity is a priority.

❑ HYPERPARAMETER TUNING

♦ TUNING ADABOOST USING ORIGINAL DATA:

Best parameters are {'n_estimators': 100, 'learning_rate': 0.1, 'base_estimator': DecisionTreeClassifier(max_depth=3, random_state=1)} with CV score=0.6802208588957056

➢ Insights:

- The model's best parameter suggests a careful and methodical approach to learning, aiming for stability and generalization rather than aggressive fitting.
- The cross-validation score of 0.6802 indicates moderate performance, with room for further refinement through additional tuning or feature enhancement.
- The model appears well suited for applications where avoiding overfitting and ensuring consistent predictions are key concern.

♦ CREATING NEW PIPELINE WITH BEST PARAMETERS:

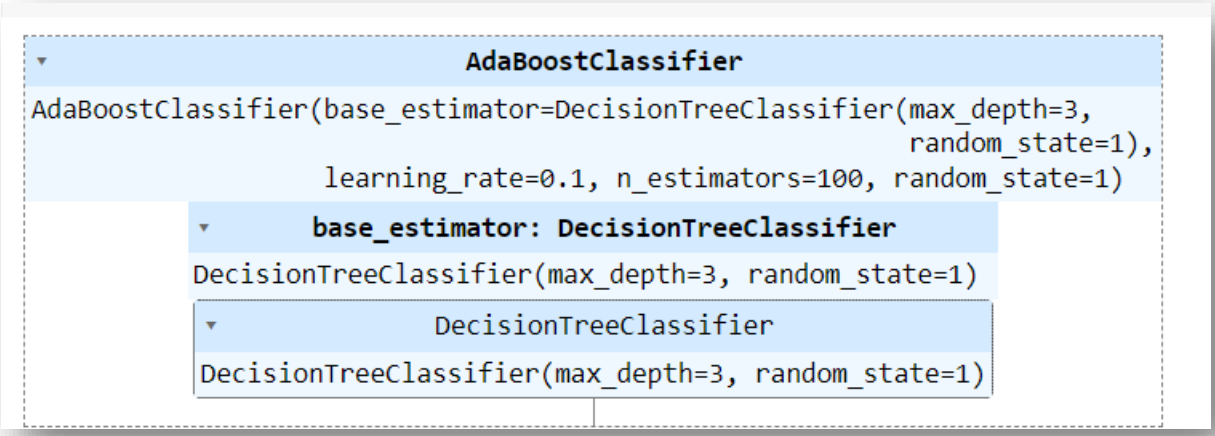

FIGURE 41

♦ CONFUSION MATRIX ON PERFORMANCE ON TRAINING SET:

	Accuracy	Recall	Precision	F1
0	0.981	0.916	0.963	0.939

TABLE 7

➢ Insights based on the performance on the training set:

- Accuracy: The model correctly predicted 98.1% of the training instances. The high accuracy suggests that the

model has learned the patterns in the training data very well.

- Recall: Recall of 91.6% means the model successfully identified 91.6% of all the positive instances in the training set. The high recall shows that the model is effective at capturing most of the true positive cases.
- Precision: With a precision of 96.3% when the model predicted a positive instance it was correct 96.3% of the time. The high precision suggests that the model is very reliable in its positive predictions, with few false positives.
- F1 score: This suggests that the model has achieved an effective compromise between recall and precision on training set.
- The model shows excellent performance on the training set, with high accuracy, recall, precision and F1. While this indicates that the model has learned the training data well there is a potential risk of overfitting.

♦ CONFUSION MATRIX ON PERFORMANCE ON VALIDATION SET:

	Accuracy	Recall	Precision	F1
0	0.972	0.892	0.917	0.904

TABLE 8

➢ **Insights based on the performance on validation set:**

- Accuracy: The model correctly classified 97.2% of the instances in the validation set. The high accuracy suggests that the model is performing well on the unseen data, maintaining a strong ability to correctly predict the majority of cases.
- Recall: A recall of 89.2% suggests that the model successfully identified 89.2% of the actual positive cases. While this is a strong recall, it is lower than the training recall suggesting that the model may be missing some positive cases in the validation set.
- Precision: With a precision of 91.7% the model correctly predicts 91.7% of positive case. The high precision implies that the model is effective at minimizing false positives.
- F1 score: The F1 score of 0.904, which balances recall and precision shows that the model maintains a good balance between recall and precision on the validation set.
- The model shows a strong generalization capabilities with a high accuracy of 97.2% on the validation set. It shows a slight decrease in recall and precision compared to the training set indicating that while it performs very well on unseen data there is room for some improvement.
- The balanced F1 score and high precision confirms model's reliability, making it a strong candidate for deployment.

❑ TUNING ADABOOST USING UNDERSAMPLED DATA

♦ CREATING NEW PIPELINE WITH BEST PARAMETERS:

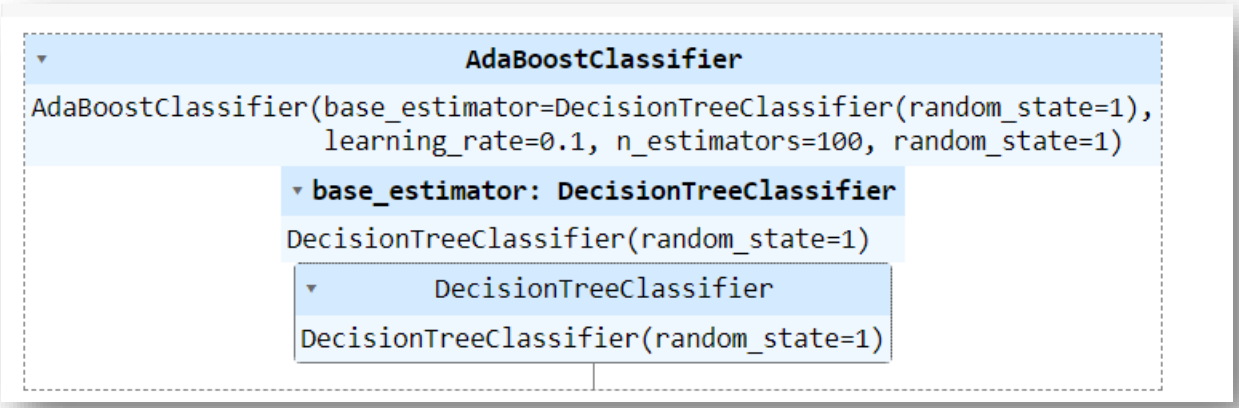

FIGURE 42

♦ CONFUSION MATRIX ON PERFORMANCE ON TRAINING SET:

	Accuracy	Recall	Precision	F1
0	0.981	0.916	0.963	0.939

TABLE 9

➢ **Insights based on the performance on the training set:**

- Accuracy: The model correctly classified 98.1% of the instances in the training set. The high accuracy indicates that the model has learned the patterns in the training data very effectively.
- Recall: With a recall of 91.6% the model successfully classified 91.6% of the true positive cases in the training data. The high recall suggests that the model is good at capturing most of the positive instances, minimizing false negatives.
- Precision: The model achieved the precision of 96.3% meaning that when it predicted a positive instance, it was correct 96.3% of the times. The high precision indicates that the model is effective at minimizing false positives.
- F1 score: The F1 score, which balances precision and recall is 0.939. This indicates that the model has achieved a good balance between identifying positive cases and ensuring that its positive predictions are accurate.
- The model exhibits excellent performance on the training set, with high accuracy, recall, precision and F1 scores. These metrics indicate that the model has learned the training data well and is effective at both identifying the positive instances and making accurate positive predictions.

♦ CONFUSION MATRIX ON PERFORMANCE ON VALIDATION SET:

	Accuracy	Recall	Precision	F1
0	0.972	0.892	0.917	0.904

TABLE 10

➢ Insights based on the performance on validation set:

- Accuracy: The model correctly classified 97.2% of the instances, which is a strong indicator of overall performance.
- Recall: Recall score suggests that the model correctly identified 89.2% of the relevant true positives. This is slightly lower than precision suggesting that while the model is good at identifying positives, it may miss some true positives, showing slight tendency towards false negatives.
- Precision: With precision of 91.7% the model effectively predicted positive instances. The high precision suggests that when the model predicts a positive, it is likely correct.
- F1 score: The F1 score of 0.904 reflects the model's ability to maintain a good trade-off between precision and recall. It indicates the model is generally well calibrated.

- The model demonstrates high accuracy and precision making it reliable for tasks where positive identification is crucial.
- The slight lower recall indicates that the model misses some relevant instances, but high precision compensates for this.
- The F1 score consolidates the model's effectiveness, indicating a good balance between recall and precision.

♦ TUNING GRADIENT BOOSTING USING UNDERSAMPLED DATA:

Best parameters are {'subsample': 0.9, 'n_estimators': 100, 'max_features': 0.5, 'learning_rate': 0.1, 'init': AdaBoostClassifier(random_state=1)} with CV score=0.6353921661160925

➢ Insights:

- The best parameter identified for the model suggests an optimal configuration that balances the complexity and performance.
- With a cross-validation score of 0.635, the model has a moderate predictive power.

- These parameters suggest a well-regularized model that is cautious in its learning approach which is crucial for handling complex data, with potential noise or overfitting risk.

♦ CREATING NEW PIPELINE WITH BEST PARAMETERS:

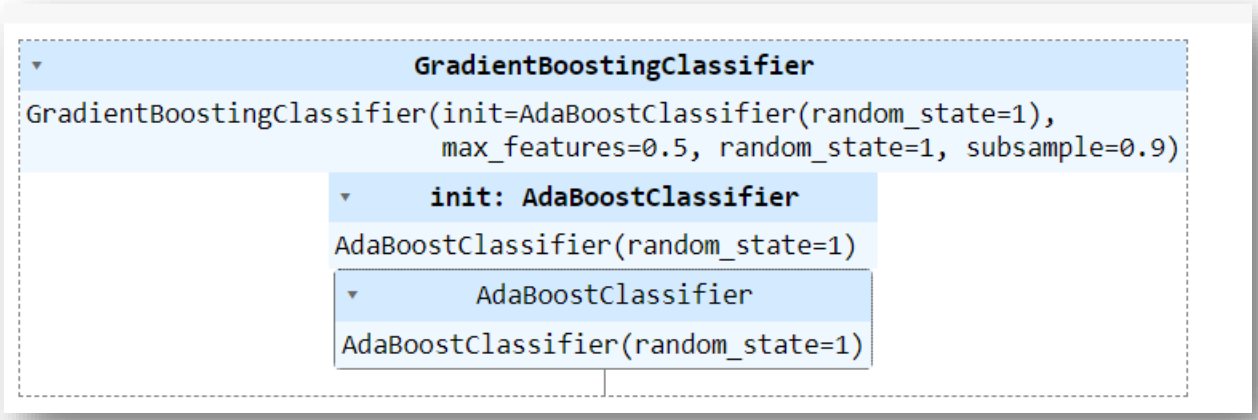

FIGURE 43

♦ CONFUSION MATRIX ON PERFORMANCE ON UNDERSAMPLED TRAINING SET:

	Accuracy	Recall	Precision	F1
0	0.974	0.979	0.969	0.974

TABLE 11

➢ Insights based on the performance on the Under Sampled training set:

- Accuracy: The model correctly classified 97.4% of the instances, showing high overall effectiveness.
- Recall: With a recall of 97.9%, the model effectively identifies almost all the true positive cases, meaning it is highly sensitive and has low false negative rate.
- Precision: A precision of 96.9% shows the model makes very few false positive predictions.
- F1 score: F1 score is 0.974, indicating an excellent balance between precision and recall, reflecting model's ability to make accurate predictions.
- The model performs exceptionally well on the under sampled training set, achieving both high recall and precision.
- The high recall is noteworthy, as it suggests that the model is well suited for scenarios where missing true positives is costly.

♦ CONFUSION MATRIX ON PERFORMANCE ON VALIDATION SET:

	Accuracy	Recall	Precision	F1

0	0.949	0.946	0.761	0.843

TABLE 12

➢ Insights based on the performance on validation set:

- Accuracy: With an accuracy of 94.9%, the model correctly classifies the majority of instances.
- Recall: The model has a high recall of 94.6%, showing that it successfully identifies most of the true positive cases. This suggests that the model is effective at detecting relevant instances with a low false negative rate.
- Precision: The precision is considerably lower at 76.1% suggesting that a significant proportion of the positive predictions are false positives. This suggests that the model is over predicting the positive class, leading to more false alarms.
- F1 score: The F1 score of 0.843, which balances recall and precision reflects the model's ability to make accurate predictions.
- The model is highly sensitive as indicated by the high recall, making it reliable for detecting true positives.
- Though the model's performance suggests model is effective at identifying relevant instances, there is a room to improve its precision.

♦ TUNING GRADIENT BOOSTING USING ORIGINAL DATA:

Best parameters are {'subsample': 0.9, 'n_estimators': 100, 'max_features': 0.5, 'learning_rate': 0.1, 'init': AdaBoostClassifier(random_state=1)} with CV score=0.6353921661160925

➢ Insights:

- The identified best parameters suggest a well-tuned model configuration aimed at balancing performance and over fitting.
- With a CV score of 0.635, the model shows moderate predictive power. While the score is not exceptionally high it reflects a model that is likely robust and well regularized suited for handling complex data.

♦ CREATING NEW PIPELINE WITH BEST PARAMETERS:

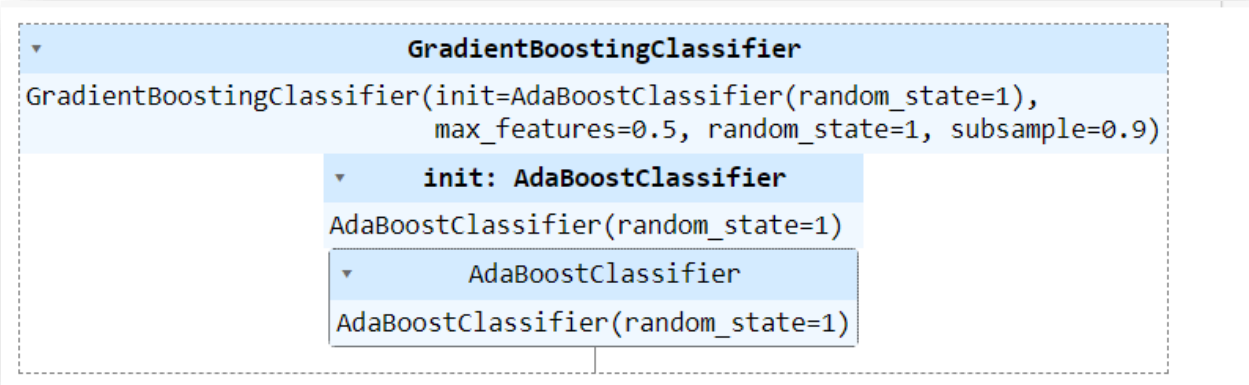

FIGURE 44

- **TUNING GRADIENT BOOSTING USING OVER SAMPLED DATA:**

- **CONFUSION MATRIX ON PERFORMANCE ON OVER SAMPLED TRAINING SET:**

	Accuracy	Recall	Precision	F1
0	0.975	0.881	0.958	0.918

TABLE 13

- **Insights based on the performance on over sampled training set:**

- Accuracy: The model achieved a high accuracy of 97.5% indicating it correctly classified most of the instances. This

suggests a strong overall performance on the over sampled data.

- Recall: The recall of 88.1% shows that the model correctly predicts most of the true positives but still misses some.
- Precision: With a precision of 95.8%, the model is excellent at minimizing false positives, meaning most positive predictions are correct.
- F1 score: F1 score is 91.8%, which shows a well-balanced performance, though the lower recall pulls it down slightly.
- The over sampling technique appears to have helped boost the overall performance, especially in accuracy and precision, but further tuning is needed.
- While recall is slightly lower, it remains robust, suggesting the model is well suited for application where both minimizing false positives and maintaining a high detection rate of true positives are important.

♦ CONFUSION MATRIX ON PERFORMANCE ON THE VALIDATION SET:

	Accuracy	Recall	Precision	F1
0	0.949	0.946	0.761	0.843

TABLE 14

➢ **Insights based on the performance on the validation set:**

- Accuracy: The model has a high accuracy, of 94.9% meaning it correctly classifies the majority of the cases.
- Recall: The recall of 94.6% shows that the model is highly effective at identifying true positive cases, with a low rate of false negatives. This is crucial in cases where missing a positive instance is more harmful than the false positives.
- Precision: The precision is relatively lower at 76.1% indicating that a significant number of positive predictions are incorrect
- F1 score: The F1 score is 84.3%, reflecting the tradeoff between the model's ability to detect positives and accuracy of those predictions.

♦ **TUNING XG BOOST MODEL WITH ORIGINAL DATA:**

Best parameters are {'subsample': 0.9, 'scale_pos_weight': 5, 'n_estimators': 100, 'learning_rate': 0.01, 'gamma': 1} with CV score=0.7521113732892873

➢ **Insights:**

- The best parameter indicates a well-tuned model with a focus on handling class imbalance and controlled learning.
- With a cross-validation score of 0.752, the model shows good generalization ability. The parameters reflect a deliberate approach to handling class imbalance while ensuring the model remains well regularized and stable during training.

♦ CREATING NEW PIPELINE WITH BEST PARAMETERS:

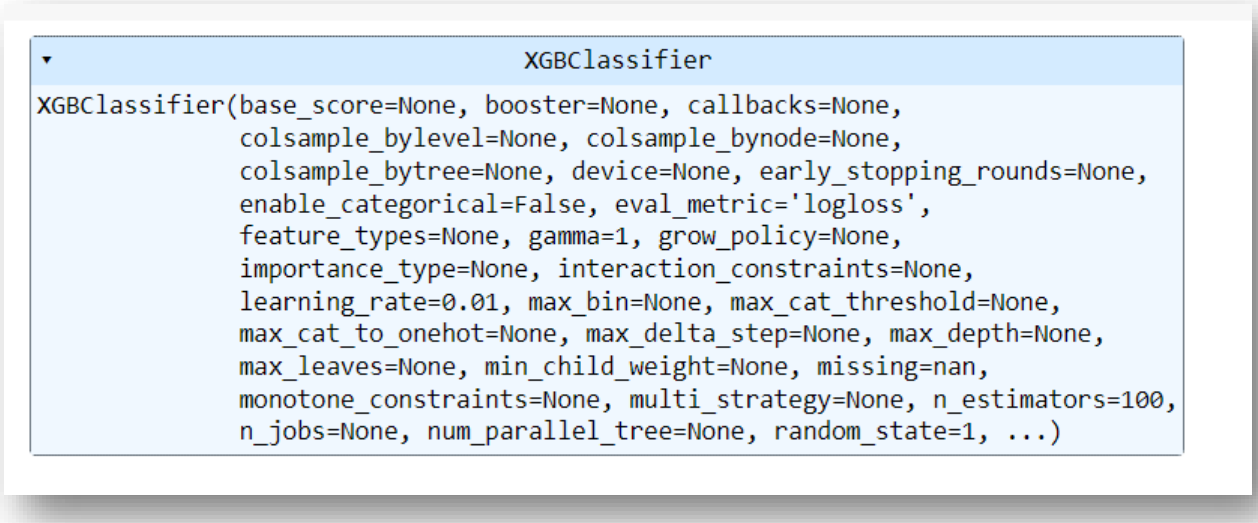

FIGURE 45

♦ CONFUSION MATRIX ON PERFORMANCE ON THE ORIGINAL TRAINING SET:

	Accuracy	Recall	Precision	F1
0	0.949	0.965	0.773	0.858

TABLE 15

- **Insights based on the performance on the original training set:**

- Accuracy: The model achieved a high accuracy of 94.9% indicating that it correctly classifies the majority of cases.
- Recall: With a recall of 96.5%, the model is highly effective at identifying true positives. This is valuable in scenarios where detecting all positive cases is crucial.
- Precision: The precision is relatively lower at 77.3% suggesting that while the model is good at identifying positives, it also generates a fair number of false positives. This could be a concern in applications where incorrect positive predictions carry significant costs.
- F1 score: The F1 score of 85.8% which balances precision and recall, indicates a reasonably good performance overall. However, the lower precision slightly diminishes the model's overall effectiveness.

- **CONFUSION MATRIX ON THE PERFORMANCE ON VALIDATION SET:**

	Accuracy	Recall	Precision	F1
0	0.935	0.946	0.707	0.809

TABLE 16

- **Insights based on the performance on the validation set:**

- Accuracy: The model correctly classifies 93.5% of the cases, indicating solid performance overall.
- Recall: Recall of 94.6% shows that the model is highly effective at identifying true positives, with low rate of false negatives, making it reliable for detecting true positives, which is crucial in applications where missing positive cases is costly.
- Precision: Precision is relatively lower at 70.7% suggesting that a significant proportion of positive predictions are false positives. This indicates the model is less effective at distinguishing between true positives and false positives, which could lead to unnecessary errors.
- F1 score: F1 score of 80.9% reflects the tradeoff between precision and recall. It shows that the model performs reasonably well overall, the lower precision pulls down the F1 score indicating that the model could benefit from

improvement in making more accurate positive predictions.

❑ MODEL COMPARISON AND FINAL MODEL SELECTION:

♦ TRAINING PERFORMANCE COMPARISON TABLE:

The table compares the performance of different models trained with various data sampling techniques:

Training performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Undersampled data	XGBoost trained with Original data
Accuracy	0.974	0.975	0.981	0.949
Recall	0.979	0.881	0.916	0.965
Precision	0.969	0.958	0.963	0.773
F1	0.974	0.918	0.939	0.858

TABLE 17

➢ Insights based on the table:

1. Gradient Boosting with under sampled data:

a) Accuracy (0.974): High accuracy, indicating the model performs well even with reduced data.

b) Recall (0.979): Excellent recall, suggesting that the model is very good at identifying true positives.
c) Precision (0.969): The precision is high, indicating only few false positives.
d) F1 score (0.974): Balanced performance with high recall and precision, making this approach very effective.
e) Gradient Boosting with under sampled data performs exceptionally well, particularly in maintaining a high F1 score, making it a robust choice when data is limited.

2. Gradient Boosting with Original Data:

a) Accuracy (0.975): Slightly higher accuracy than the under sampled version showing robust performance.
b) Recall (0.881): The recall is lower compared to the under sampled model indicating that it misses truer positives.
c) Precision (0.958): High precision but slightly lower than the under sampled version.
d) F1 score (0.918): The drop in F1 reflects the tradeoff between precision and recall with the original data leading to more missed positives.
e) Gradient Boosting with original data has a good overall performance but falls behind the under sampled version, indicating that under sampling helped to improve its recall significantly.

3. AdaBoost with Under sampled data:

a) Accuracy (0.981): The highest accuracy among the models, indicating strong overall performance.

b) Recall (0.916): Good recall, though not as high as Gradient Boosting with under sampled data.
c) Precision (0.963): Very high precision, suggesting the model is excellent at minimizing false positives.
d) F1 score (0.939): High F1 score, showing a well-balanced model, slightly better than Gradient Boosting with the original data.
e) AdaBoost with under sampled data achieves the best overall performance, with highest accuracy, and a strong balance between recall and precision.

4. XG Boost with Original data:

a) Accuracy (0.949): Lower accuracy compared to the other models, suggesting it struggles more with the data.
b) Recall (0.965): High recall, indicating strong sensitivity to true positives.
c) Precision (0.773): Significantly lower precision, meaning the model produces a considerable number of false positives.
d) F1 score (0.858): The lowest F1 score among the model's reflecting the imbalanced tradeoff between high recall and low precision.
e) AdaBoost with under sampled data shows high recall but at the cost of precision, leading to more false positives. The model may require further tuning or balancing to improve its precision.

♦ TRAINING PERFORMANCE COMPARISON TABLE:

The table compares the validation performance of different models and data sampling techniques:

	Gradient boosting validated with Undersampled data	Gradient boosting validated with Original data	AdaBoost validated with Undersampled data	XGBoost validated with Original data
Accuracy	0.949	0.949	0.972	0.935
Recall	0.946	0.946	0.892	0.946
Precision	0.761	0.761	0.917	0.707
F1	0.843	0.843	0.904	0.809

TABLE 18

➢ Insights based on the table:

1. **Gradient Boosting Validated with Under sampled data and Original Data:**

a) Accuracy (0.949): Both the models achieved same accuracy of 94.9%, suggesting that the overall performance in classifying cases in consistent regardless of the sampling technique.

b) Recall (0.946): Both the models also have the same recall of 94.6% showing strong sensitivity and effectiveness in capturing the true positives. The choice between under sampling and using original data did not affect the recall.

c) Precision (0.761): Precision is the same for both the models at 76.1% indicating a moderate rate of false positives. The models may produce some incorrect positive predictions, but the effect is balanced across both the sampling methods.
d) F1 score (0.843): Both the models share the same F1 score of 84.3% reflecting a consistent balance between precision and recall. These results suggests that under sampling does not significantly alter the model's validation performance compared to using the original data.

2. AdaBoost Validated with Under sampled data:

a) Accuracy (0.972): The highest accuracy among the models, indicating strong generalization.
b) Recall (0.892): Slightly lower recall compared to Gradient Boosting, suggesting it misses more true positives.
c) Precision (0.917): The precision among the models, meaning it is particularly good at minimizing false positives.
d) F1 score (0.904): The highest F1 score showing a well-balanced model with a strong ability to generalize.
e) AdaBoost validated with under sampled data performs the best overall with highest accuracy, precision, F1 score indicating that it balances sensitivity and specificity well.

3. XG Boost Validated with Original Data:

a) Accuracy (0.935): Slightly lower accuracy compared to other models, indicating some challenges in generalization.

b) Recall (0.946): High recall similar to gradient boosting, showing strong sensitivity to true positives.
c) Precision (0.707): The lowest precision, indicating a higher number of false positives, which reduces the model's specificity.
d) F1 score (0.809): The lowest F1 score among the models, reflecting the tradeoff between high recall and low precision.
e) XG Boost validated with original data has a strong recall but struggles with precision, leading to a lower F1 score. It is sensitive to true positives but generates falser positives than the other models.

★ In summary, AdaBoost under sampled data is the most effective model in terms of generalization, while gradient boosting maintains a consistent performance across different data sampling strategies. XG Boost, despite its strong recall, needs further tuning to improve precision and overall balance.

❑ FEATURE IMPORTANCE

The feature importance plot provides valuable insights into which variables have the most influence on the model's predictions:

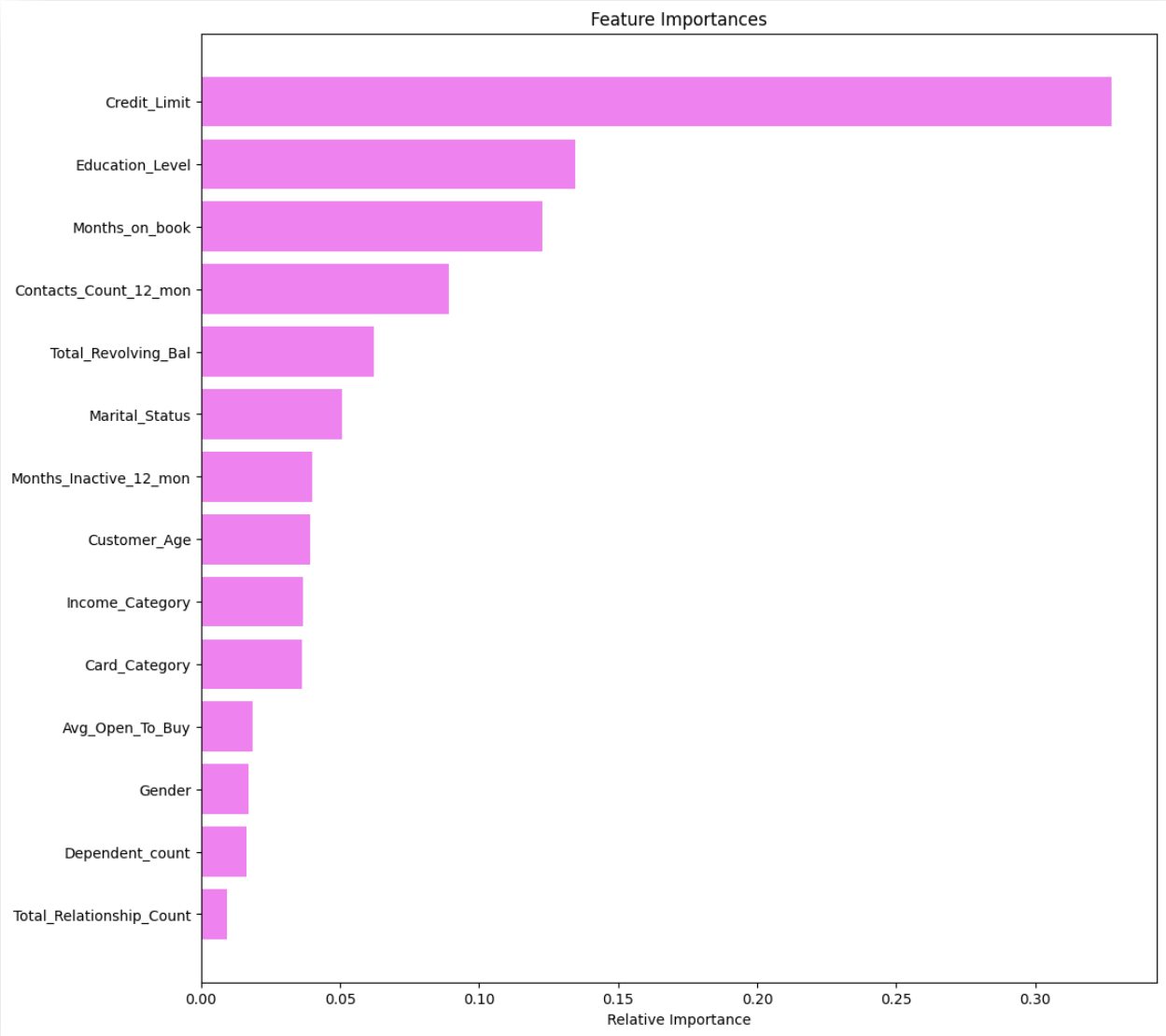

FIGURE 46

➢ Insights based on the plot:

1. Credit Limit:

- Dominant Feature: The ‘Credit_Limit’ feature stands out as the most important variable by a significant margin. This suggests that the amount of credit available to a customer is the strongest predictor in the model. This could be because credit limit often correlates with the financial behavior and risk profiles.

2. **Education Level:**

- Second most important: ‘Education_Level’ is the second most important feature indicating, that the level of education a customer has attained plays a significant role in predicting outcomes. This could relate to how education level impacts earning potential and financial literacy.

3. **Months on Book:**

- Stability indicator: ‘Months_on_Book’ is also a crucial feature, reflecting the length of time a customer has their account. A longer relationship with the bank might indicate stability and customer loyalty, which are important factors in predicting customer behavior.

4. **Contacts Count in last 12 Months:**

- Engagement indicator: The number of times a customer has contacted the bank in last 12 months is also influential. High contact frequency might indicate issues or dissatisfaction.

5. **Total Revolving Balance:**

- Credit Utilization: ‘Total_Revolving_Bal’ represents the balance on revolving credit accounts such as credit card.

This is a key metric in credit scoring models as it indicates how much of the available credit is being used.

6. **Marital Status:**

- It has a moderate influence suggesting that a customer's marital status may correlate with financial behavior, perhaps due to combined income or financial responsibilities.

7. **Months Inactive in the last 12 months:**

- Activity Level: 'Months_Inactive_12_mon' reflects how often a customer is inactive over the year. This could be a signal of declining engagement or satisfaction.

8. **Customer Age:**

- Age Factor: 'Customer_Age' shows that the age is moderately important factor. Different age groups have varying financial behaviors and risk profiles.

9. **Other Factors:**

- Income category, card category and average open to buy: These features also contribute to the model but to a lesser extent. They likely offer additional nuance to the predictions by capturing income, spending patterns and available credit.
- Gender, dependent count and total relationship count: These features have the least impact, indicating that while they might add some value, they are not as predictive as other variables.

- ✓ The model heavily relies on the financial metrics such as credit limit, revolving balance and months on book to make predictions.
- ✓ Demographic factors like education level and marital status plays a significant but secondary role.
- ✓ The engagement of customer, as measured by contact frequency and account activity is also important suggesting that customer behavior is a key indicator of outcomes.
- ✓ Feature related to age and income provide additional context but are less critical to the model's predictions.

❑ ACTIONABLE INSIGHTS & BUSINESS RECOMMENDATIONS

♦ ACTIONABLE INSIGHTS:

1. Strengthen Customer Relationship:

- Use predictive analytics to identify customers with fewer products and a history of inactivity. Launch personalized offer for credit increases, new savings accounts or investment services.

2. Enhance Customer Engagement:

- Implement automated systems to flag and contact customer who have been inactive for a specific period. Schedule regular check-ins or offer loyalty rewards to encourage activity.

3. Improve Credit Utilization:

- Provide educational content through emails or the bank's app. Offer tools that help customer monitor their credit usage and suggest ways to increase it for credit score benefits.

4. Address High Contact Rates:

- Implement a feedback loop for high contact customers to identify common pain points. Use this information to improve customer service and reduce the need for frequent interactions.

5. Retention Strategy for High-Risk Customers:

- Create a loyalty program that rewards the long-term customers with exclusive benefits, such as lower fees, higher credit limits, or personalized financial planning services.

♦ BUSSINESS RECOMMENDATIONS:

1. Deepen Customer Relationships:

- Increase the number of products per customers to improve the customer retention.

- Identify the customer with fewer than 3 products (Total relationship count) and promote bundled offerings (e.g. pairing a savings account with a credit card or offering insurance products).
- Use personalized marketing strategies leveraging data analytics to target offers based on customer profiles, such as income level or existing products. It will strengthen customer loyalty and increase lifetime value by encouraging customers to engage more deeply with the bank.

2. Enhance Customer Retention for At- Risk Segments:

- Reduce the attrition rate by addressing key factors leading to customer churn.
- Develop an early warning system to detect customers showing signs of disengagement, such are reduced transaction activity or increased inactivity.
- Proactively reach out to these customers with personalized offer, such as fee waivers, credit limit increases, or loyalty rewards.
- Implement a targeted retention program focusing on customer who have been with the bank for over 36 months but show declining activity.
- This will lower attrition rates and improved customer satisfaction by addressing issues before they lead to churn.

3. Optimize Credit Utilization:

- Encourage more active use of credit lines to enhance profitability and customer engagement.
- Educate customer with low utilization ratios on the benefits of responsible credit usage.
- Implement targeted campaigns to offer higher credit limits to customer with a track of responsible credit management, encouraging them utilize more of their available credit.
- This will increase the revenue from interest and fees due to higher utilization rates and a more engaged customer base.

4. **Improve Customer Experience through Proactive Support:**

- Reduce the frequency of customers contacts by addressing underlying issues that lead to dissatisfaction.
- Analyze the reasons for high contact rates among attrited customers to identify common service issues or gaps in customer support.
- Enhance customer service training and implement a customer feedback loop to address and resolve issues more effectively on the first contact.
- Offer self-service tools and FAQs on the bank's website or app to reduce the need for frequent customer service interactions.

- This will lead to enhanced customer satisfaction and reduced operational costs by decreasing the volume of customer service enquiries.

5. **Leverage data-driven Marketing for Targeted offers:**

- Increase transaction volumes and frequency by tailoring offers to specific customer segments.
- Use transaction data to identify customers with declining transaction amounts or frequencies. Offer cashback rewards, discounts or points to incentivize increased usage.
- Develop seasonal or event-driven promotions targeted at customers with historically low transaction activity during certain periods.
- This will help in higher transaction volumes and increased customer engagement.

By focusing on deepening relationship, enhancing retention, optimizing credit usage, improving customer experience and leveraging data-driven marketing, the Thera Bank can significantly improve customer satisfaction, reduce attrition and drive profitability.