MACHINE LEARNING 2– BUSINESS REPORT

SONA

PGPDSBA.O.OCT24.A

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Problem Statement - Thera bank - Guided Project

Business Context

The Thera bank recently saw a steep decline in the number of users of their credit card, credit cards are a good source of income for banks because of different kinds of fees charged by the banks like annual fees, balance transfer fees, and cash advance fees, late payment

fees, foreign transaction fees, and others. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances.

Customers' leaving credit card services would lead the bank to loss, so the bank wants to analyze the data of customers and identify the customers who will leave their credit card services and the reason for same – so that the bank could improve upon those areas

Objective

You as a Data Scientist at Thera Bank need to come up with a classification model that will help the bank improve its services so that customers do not renounce their credit cards

You need to identify the best possible model that will give the required performance

1. Explore and visualize the dataset.
2. Build a classification model to predict if the customer is going to churn or not
3. Optimize the model using appropriate techniques
4. Generate a set of insights and recommendations that will help the bank

Data Dictionary:

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

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

1. Data Sanity Checks

Import all the necessary libraries and load the dataset.

Shape of the Dataset

```
(10127, 20)
```

Table 1: Shape of the dataset

There are **10127 rows and 20 columns** in the dataset.

Displaying the last few rows of the dataset

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Rel
0	768805383	Existing Customer	45	M	3	High School	Married	$60K – 80K	Blue	39	
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K – 120K	Blue	36	
3	769911858	Existing Customer	40	F	4	High School	NaN	Less than $40K	Blue	34	
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K – 80K	Blue	21	

Table 2: Head – first 5 rows of dataset

Displaying the last few rows of the dataset

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Tota
10122	772366833	Existing Customer	50	M	2	Graduate	Single	$40K – 60K	Blue	40	
10123	710638233	Attrited Customer	41	M	2	NaN	Divorced	$40K – 60K	Blue	25	
10124	716506083	Attrited Customer	44	F	1	High School	Married	Less than $40K	Blue	36	
10125	717406983	Attrited Customer	30	M	2	Graduate	NaN	$40K – 60K	Blue	36	
10126	714337233	Attrited Customer	43	F	2	Graduate	Married	Less than $40K	Silver	25	

Table 3: Tail – last 5 rows of dataset

Checking the data types of the columns for the dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  object
 2   Customer_Age              10127 non-null  int64
 3   Gender                    10127 non-null  object
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           8608 non-null   object
 6   Marital_Status            9378 non-null   object
 7   Income_Category           10127 non-null  object
 8   Card_Category             10127 non-null  object
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(10), object(6)
memory usage: 1.6+ MB
```

Table 4: Data type

- Three different data types are present: 5 float, 10 int and 6 object data types.
- There are **null values present in the Education_Level and Marital_Status fields**.

Checking for duplicate values

```
0
```

There are **no duplicate values** in the dataset.

Checking for missing values

CLIENTNUM	0.000
Attrition_Flag	0.000
Customer_Age	0.000
Gender	0.000
Dependent_count	0.000
Education_Level	15.000
Marital_Status	7.400
Income_Category	0.000
Card_Category	0.000
Months_on_book	0.000
Total_Relationship_Count	0.000
Months_Inactive_12_mon	0.000
Contacts_Count_12_mon	0.000
Credit_Limit	0.000
Total_Revolving_Bal	0.000
Avg_Open_To_Buy	0.000
Total_Amt_Chng_Q4_Q1	0.000
Total_Trans_Amt	0.000
Total_Trans_Ct	0.000
Total_Ct_Chng_Q4_Q1	0.000
Avg_Utilization_Ratio	0.000
dtype: float64	

Table 5: Missing values

- The Education_Level column is missing 15% of its values from the total observations.
- The Marital_Status column is missing approximately 7% of its values from the total observations.

Checking the statistical summary

	count	mean	std	min	25%	50%	75%	max
CLIENTNUM	10127.000	739177606.334	36903783.450	708082083.000	713036770.500	717926358.000	773143533.000	828343083.000
Customer_Age	10127.000	46.326	8.017	26.000	41.000	46.000	52.000	73.000
Dependent_count	10127.000	2.346	1.299	0.000	1.000	2.000	3.000	5.000
Months_on_book	10127.000	35.928	7.986	13.000	31.000	36.000	40.000	56.000
Total_Relationship_Count	10127.000	3.813	1.554	1.000	3.000	4.000	5.000	6.000
Months_Inactive_12_mon	10127.000	2.341	1.011	0.000	2.000	2.000	3.000	6.000
Contacts_Count_12_mon	10127.000	2.455	1.106	0.000	2.000	2.000	3.000	6.000
Credit_Limit	10127.000	8631.954	9088.777	1438.300	2555.000	4549.000	11067.500	34516.000
Total_Revolving_Bal	10127.000	1162.814	814.987	0.000	359.000	1276.000	1784.000	2517.000
Avg_Open_To_Buy	10127.000	7469.140	9090.685	3.000	1324.500	3474.000	9859.000	34516.000
Total_Amt_Chng_Q4_Q1	10127.000	0.760	0.219	0.000	0.631	0.736	0.859	3.397
Total_Trans_Amt	10127.000	4404.086	3397.129	510.000	2155.500	3899.000	4741.000	18484.000
Total_Trans_Ct	10127.000	64.859	23.473	10.000	45.000	67.000	81.000	139.000
Total_Ct_Chng_Q4_Q1	10127.000	0.712	0.238	0.000	0.582	0.702	0.818	3.714
Avg_Utilization_Ratio	10127.000	0.275	0.276	0.000	0.023	0.176	0.503	0.999

Table 6: Statistical summary

- The maximum income is 666,666, significantly exceeding the mean, suggesting it might be an outlier.

- Age varies widely, ranging from 1893 to 1996.
- Some users have birth years before 1900, while the campaign year we're analyzing is 2016, making it highly improbable that these individuals are still alive. This could indicate a reporting error, which we will investigate further.

	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

```
Unique values in Attrition_Flag are :
Attrition_Flag
Existing Customer    8500
Attrited Customer    1627
Name: count, dtype: int64
************************************************
Unique values in Gender are :
Gender
F    5358
M    4769
Name: count, dtype: int64
************************************************
Unique values in Education_Level are :
Education_Level
Graduate         3128
High School      2013
Uneducated       1487
College          1013
Post-Graduate     516
Doctorate         451
Name: count, dtype: int64
************************************************
Unique values in Marital_Status are :
Marital_Status
Married     4687
Single      3943
Divorced     748
Name: count, dtype: int64
************************************************
Unique values in Income_Category are :
Income_Category
Less than $40K    3561
$40K - $60K       1790
$80K - $120K      1535
$60K - $80K       1402
abc               1112
$120K +            727
Name: count, dtype: int64
************************************************
Unique values in Card_Category are :
Card_Category
Blue        9436
Silver       555
Gold         116
Platinum      20
Name: count, dtype: int64
************************************************
```

Table 7: Unique values

Dropped the CLIENTNUM as it is unique for each customer and will not add value to the model and encoding existing and attrited customers to 0 and 1 respectively, for analysis.

2. Exploratory Data Analysis

2.1 Univariate Analysis

Customer_Age

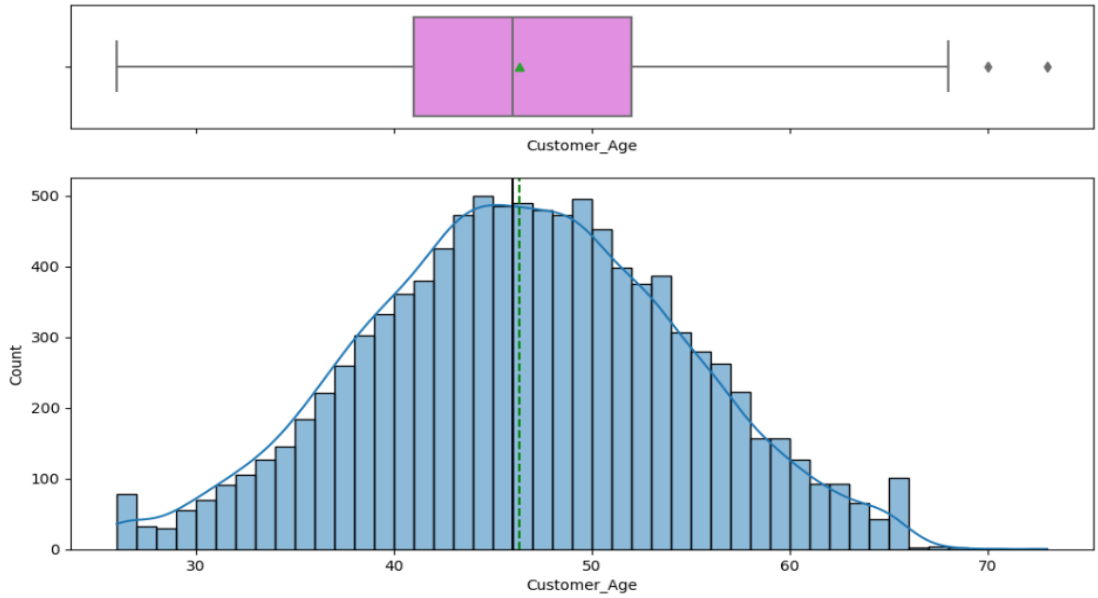

Fig 1. Customer_Age

- The age distribution is **normal**.
- The boxplot indicates the presence of **outliers on the right side**.
- We will not address these outliers, as they reflect the actual market trend.

Months_on_book

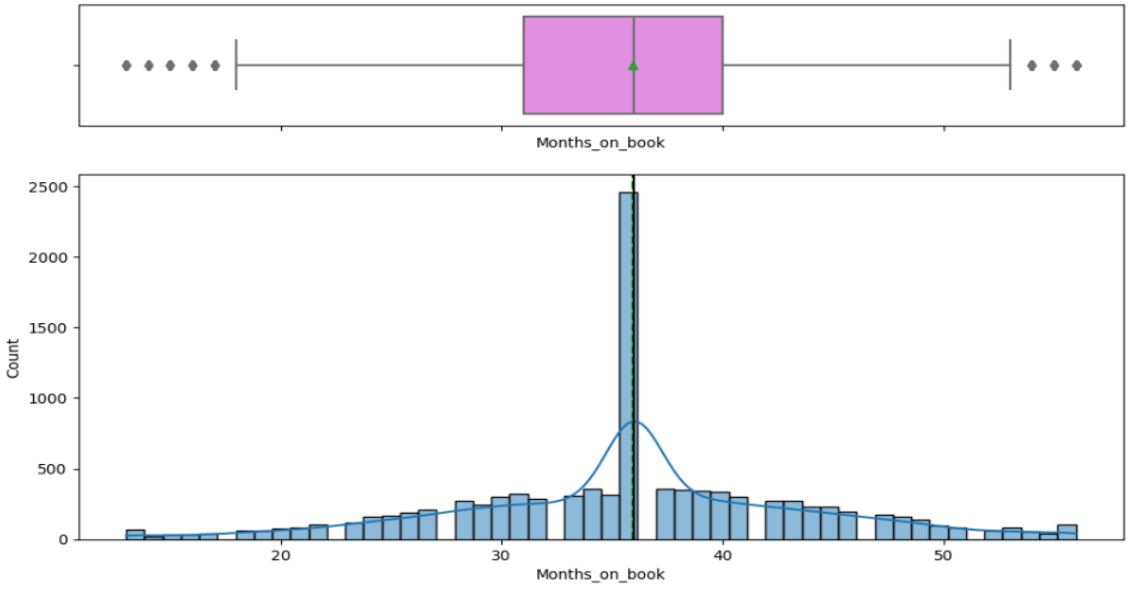

Fig 2. Months_on_book

- The distribution of the amount spent on fruits is **nearly normal**.
- The median for these customers at Thera Bank is approximately 37 months.
- While there are **a few outliers on both the right and left ends of the boxplot**, we will not address them, as some variation is to be expected in real-world scenarios involving spending amounts.

Credit_Limit

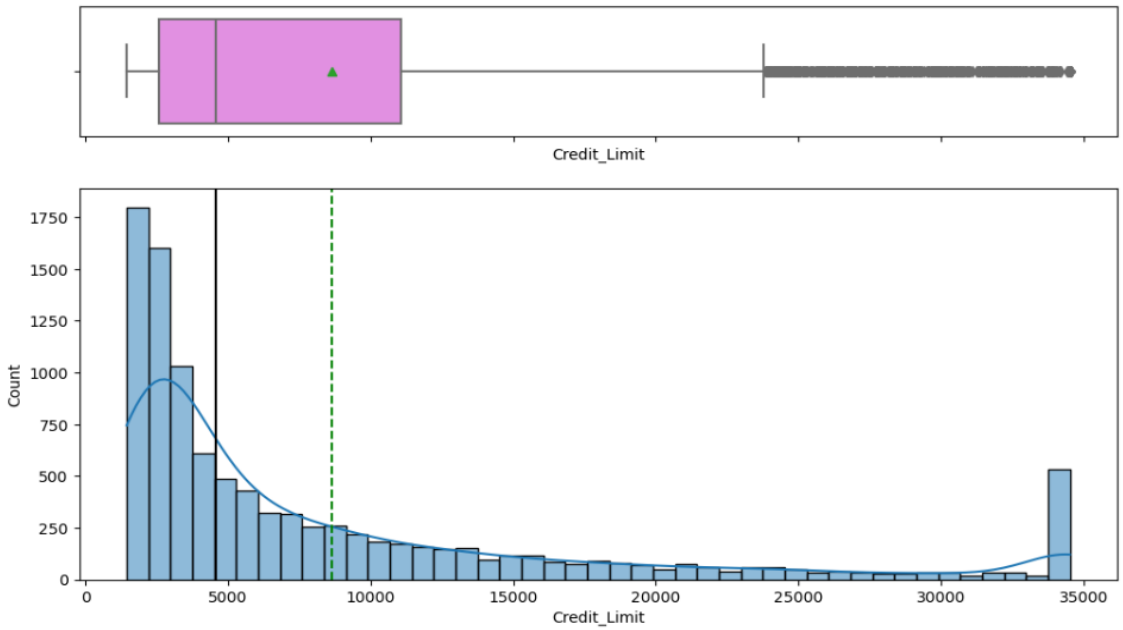

Fig 3. Credit_Limit

- The distribution of the credit amount is **right-skewed**.
- The boxplot indicates the presence **of outliers on the right side**.
- We will not address these outliers, as they reflect the actual market trend.

Total_Revolving_Bal

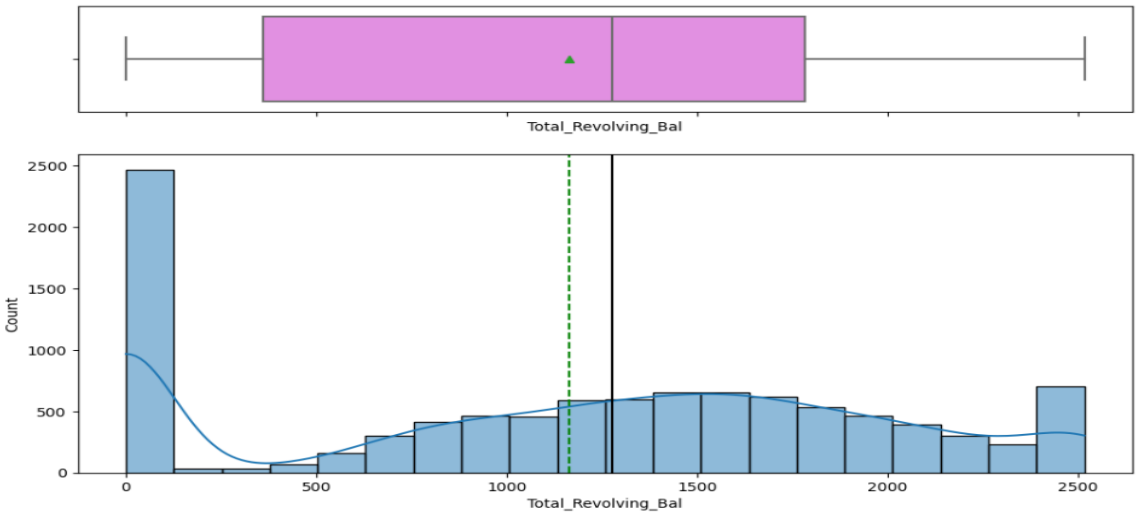

Fig 4. Total_Revolving_Bal

- The distribution of the balance that carries over from month to month, which reflects the revolving credit, is **approximately normal**.
- The boxplot reveals **outliers on the left side**.
- We will not address these outliers, as they represent the actual market trend.

Avg_Open_To_Buy

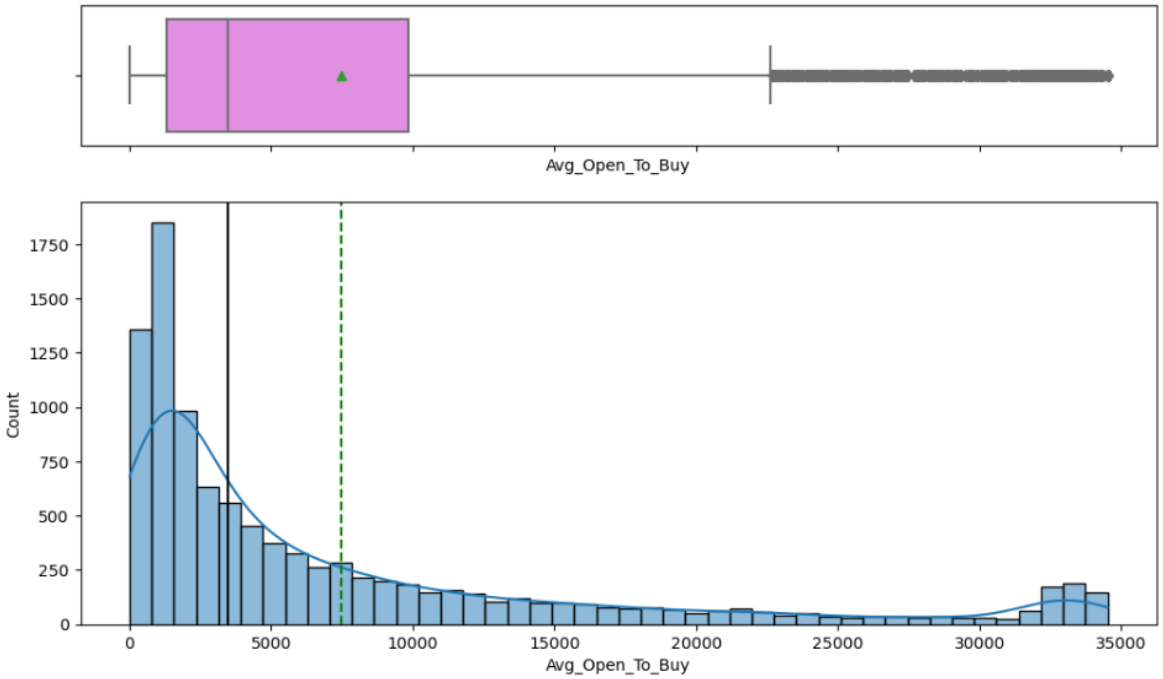

Fig 5. Avg_Open_To_Buy

- The distribution of the amount remaining on the credit card (averaged over the last 12 months) is **right-skewed**.
- There are **several observations at the far right** that may be considered outliers.
- We will not remove all of these data points, as they reflect genuine customer trends.

Total_Trans_Ct

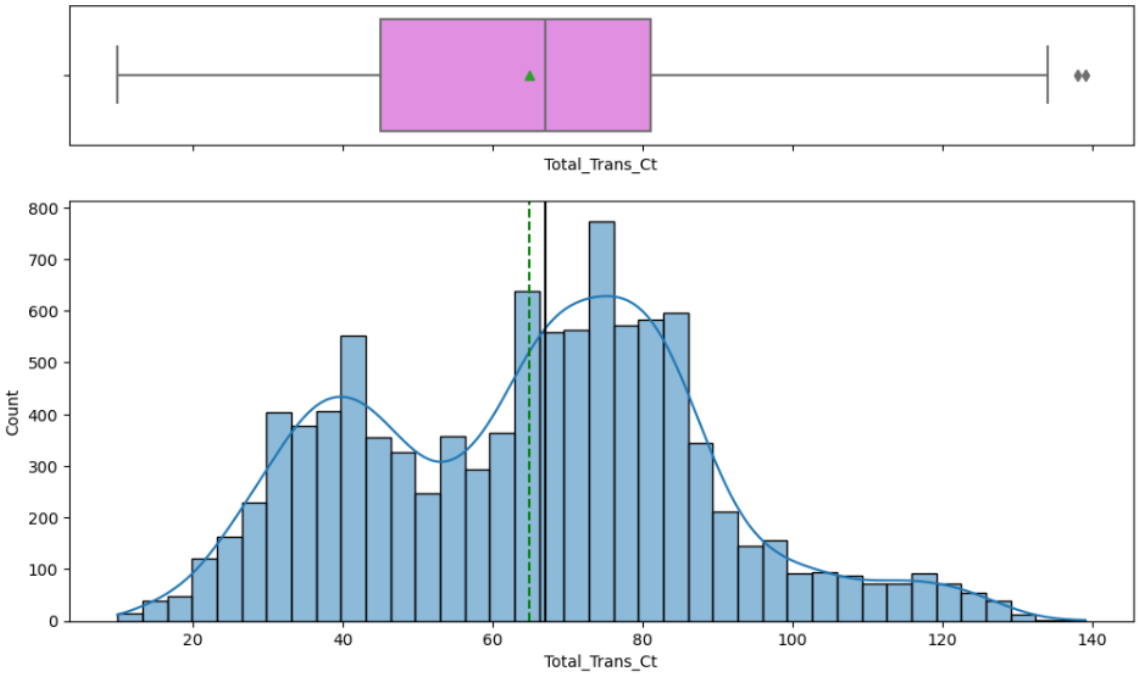

Fig 6. Total_Trans_Ct

- The majority of customers have **approximately 65 transactions in the last 12 months**.
- There are some **extreme values observed at the far right end**.

Total_Amt_Chng_Q4_Q1

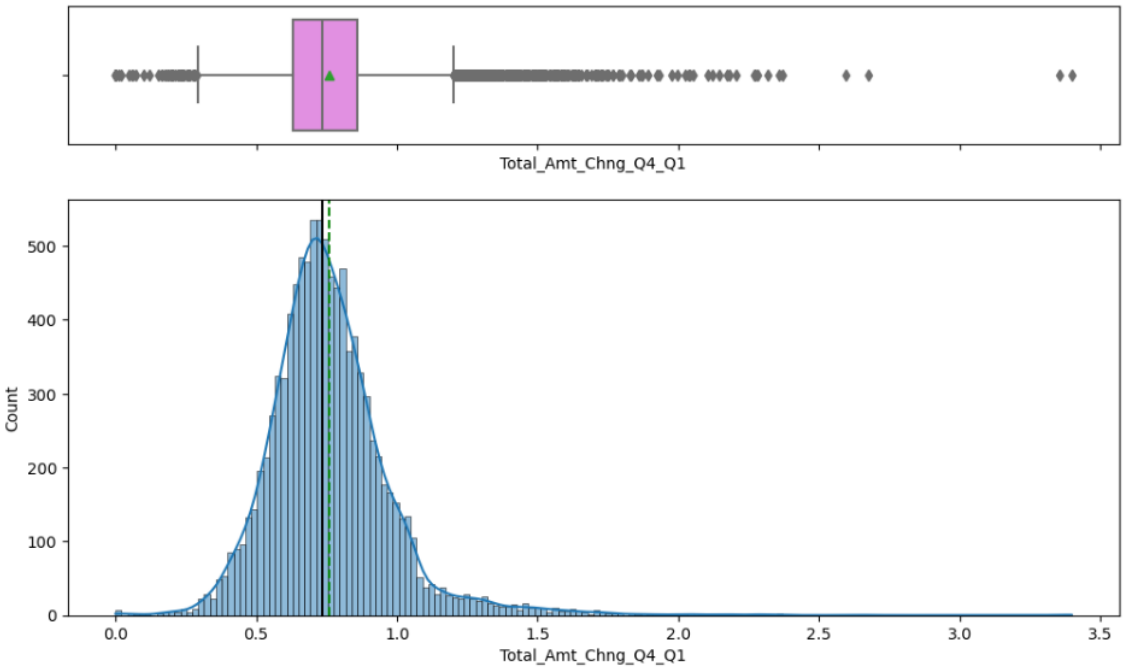

Fig 7. Total_Amt_Chng_Q4_Q1

- The distribution is **right-skewed**.
- The **median of the distribution is approximately 0.7** for the ratio of total transaction amounts in the fourth quarter to those in the first quarter.
- We observe **several extreme values** in this variable.

Let's see total transaction amount distributed

Total_Trans_Amt

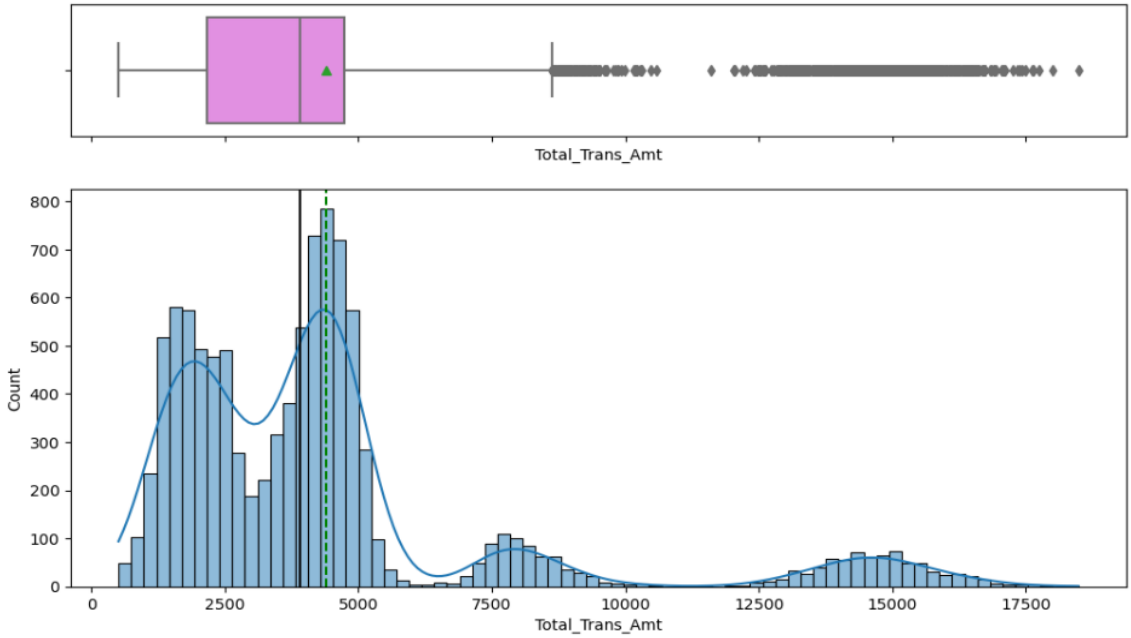

Fig 8. Total_Trans_Amt

- The distribution for the Total Transaction Amount (Last 12 months) **is right-skewed.**
- There are **many outliers in the amount spent on above 12500**. We will not remove all such data points as they represent real customers.

Total_Ct_Chng_Q4_Q1

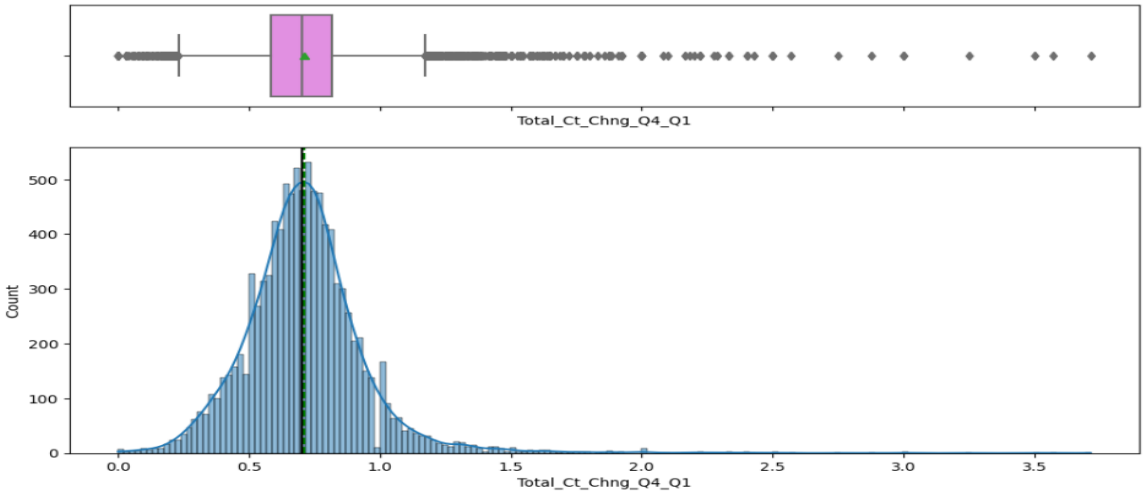

Fig 9. Total_Ct_Chng_Q4_Q1

- The **median of the distribution is around 0.6**, meaning that 50% of customers have a ratio of approximately 0.6 or less for the total transaction count in the fourth quarter compared to the first quarter.
- There are also **some extreme values** observed in this variable.

Avg_Utilization_Ratio

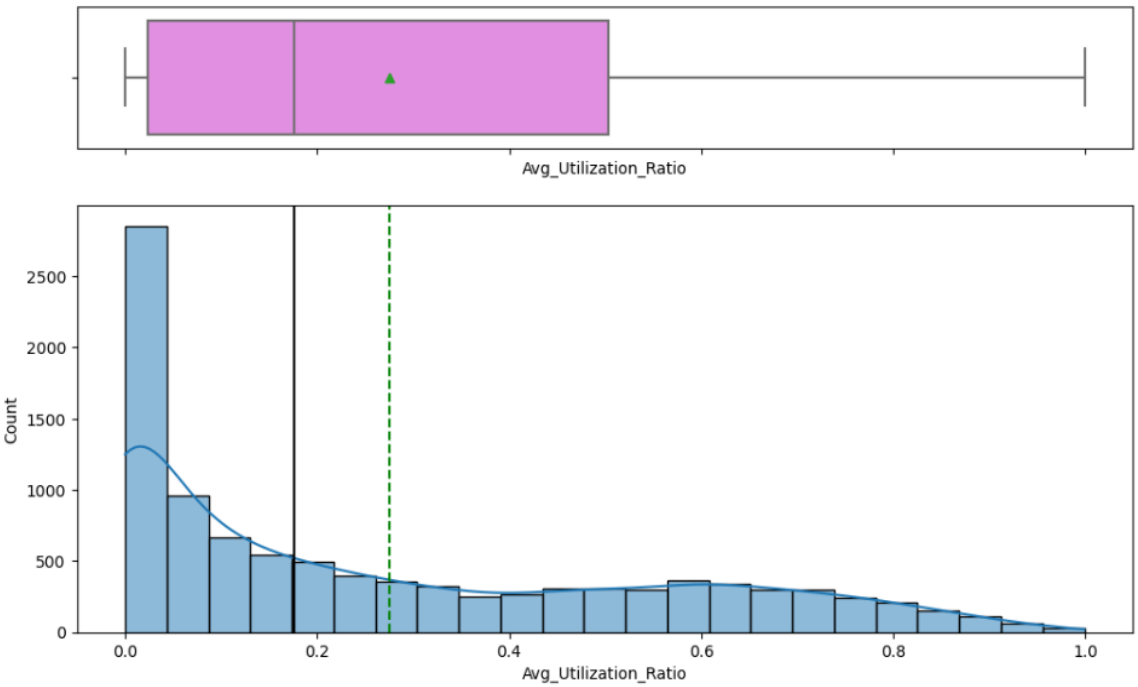

Fig 10. Avg_Utilization_Ratio

- The distribution is **right-skewed**.
- There are very few observations showing a representation greater than 0.8 of how much of the available credit the customer has utilized. **On average, customers use about 28%** of their total credit.

- There are **no outliers** in this variable.

Dependent_count

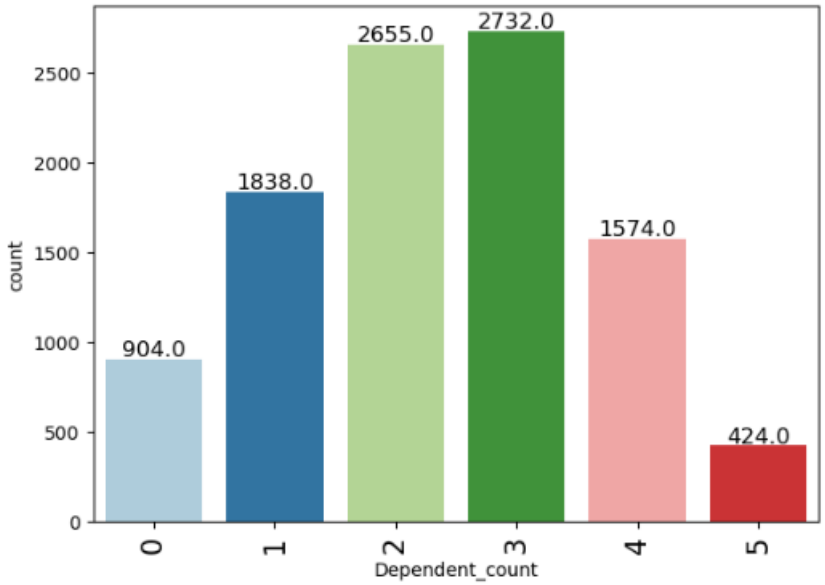

Fig 11. Dependent_count

Most customers have between **two and three dependents**.

Total_Relationship_Count

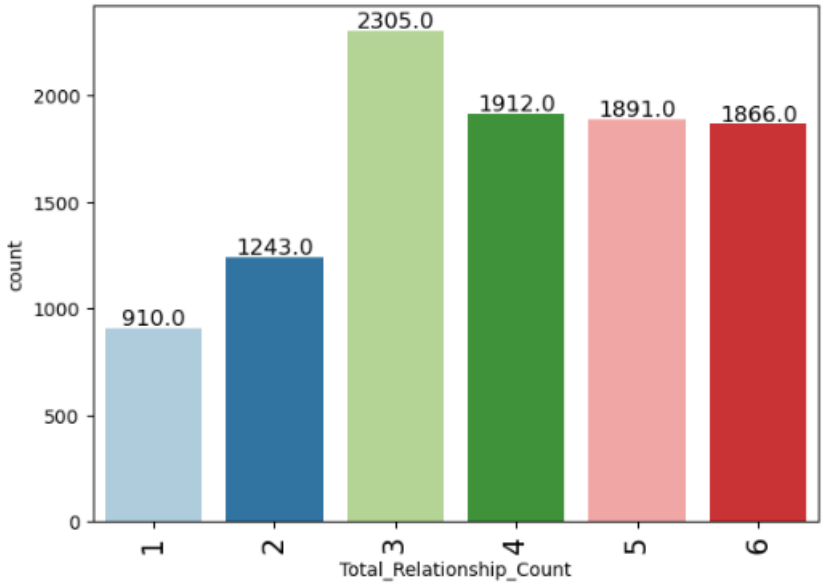

Fig 12. Total_Relationship_Count

- **Most customers hold three or four products** with the bank.
- **Attrited customers** typically have **fewer products** with the bank.

Months_Inactive_12_mon

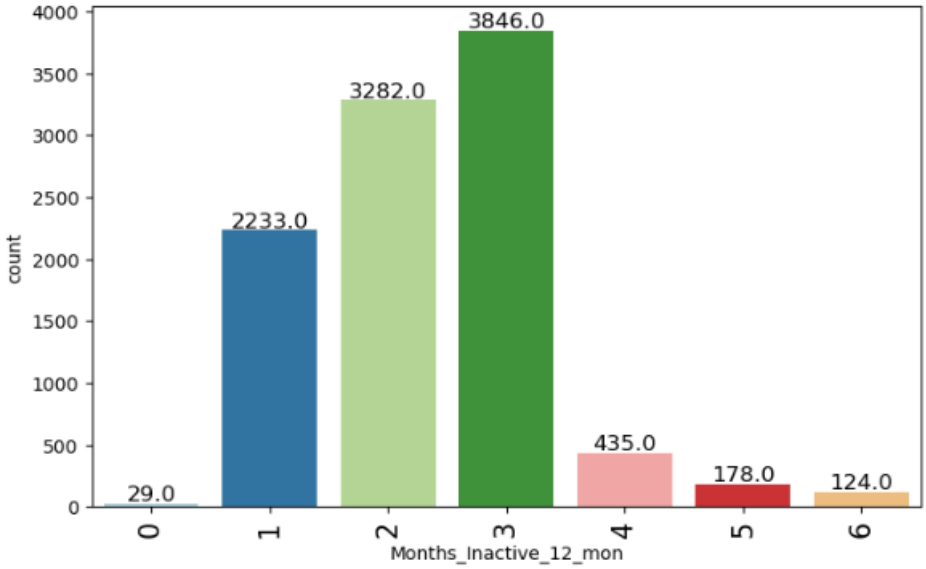

Fig 13. Months_Inactive_12_mon

- **Most customers have been inactive for 1 to 3 months** over the past 12 months, with the average duration slightly exceeding 2 months.
- **Attrited customers generally exhibit a longer period of inactivity** compared to existing customers.

Contacts_Count_12_mon

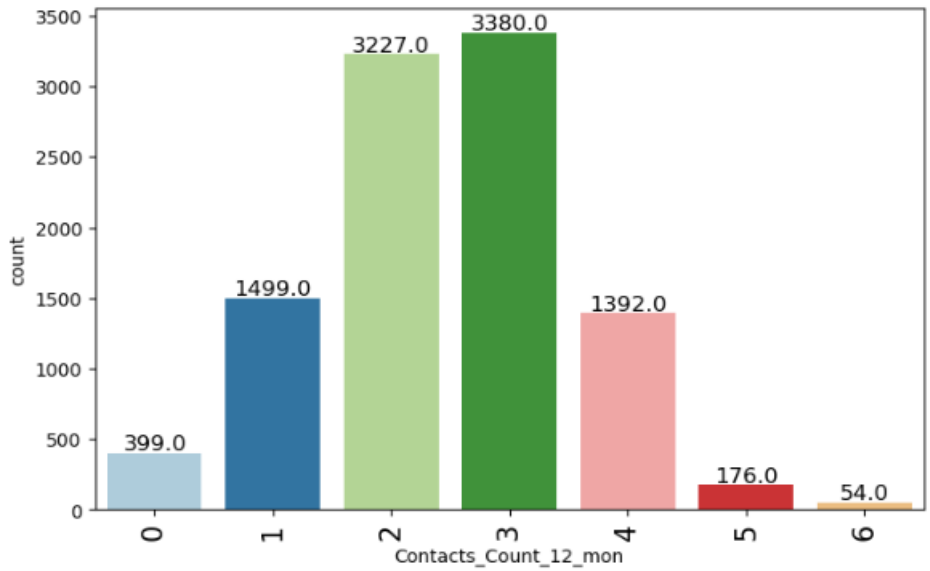

Fig 14. Contacts_Count_12_mon

- The plot indicates that **most customers have had three contacts** with the bank in the past 12 months.
- Approximately **51% of customers have had at least one contact**.

Gender

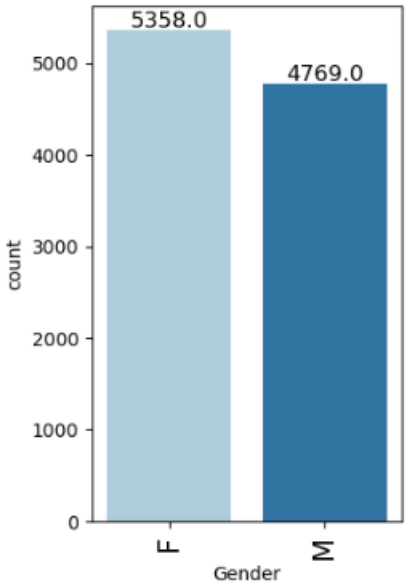

Fig 15. Gender

- **Female customers are obtaining more** credit cards than male customers.
- Approximately 47% of customers are male, while 53% are female.

Let's see the distribution of the level of education of customers

Education_Level

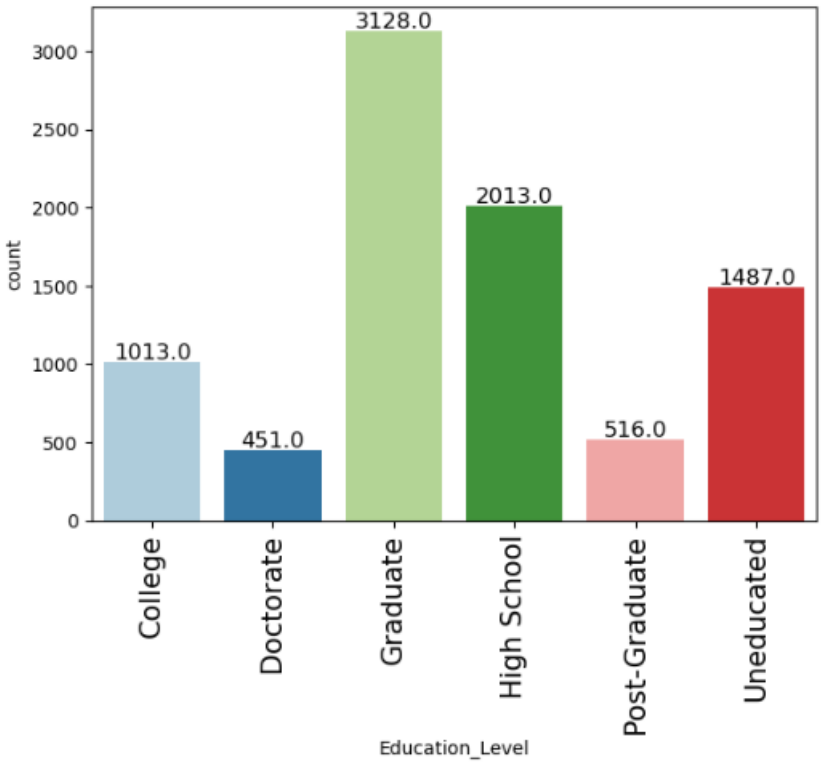

Fig 16. Education_Level

- The **majority of customers are about 31%**, who obtain credit cards hold a graduate degree.
- Approximately **19% are high school graduates**.
- Only 14% of customers have no formal education, which may affect their eligibility for a credit card.

Marital_Status

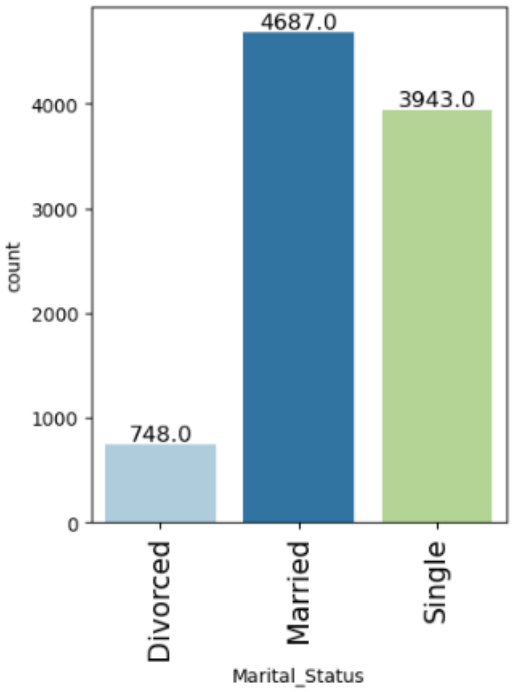

Fig 17. Marital_Status

- **Approximately 46% of customers are married**, which is understandable given the popularity of joint accounts.
- Around **39% of customers are single**.
- It's noteworthy that there is a low percentage of divorced customers.

Income_Category

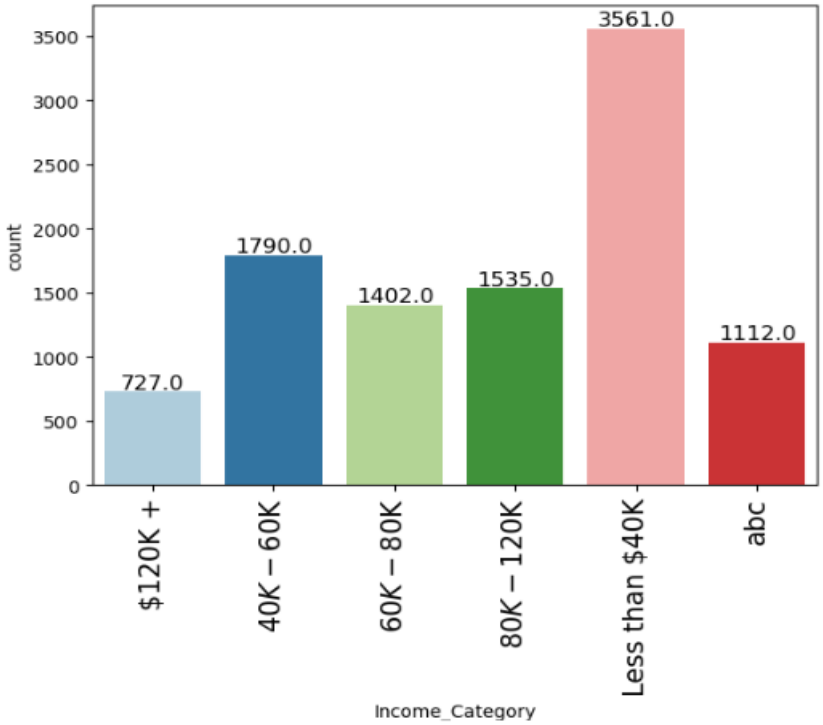

Fig 18. Income_Category

- **Most customers have an income of less than 40k**.
- A considerable number of customers have not specified their income category (Unknown). This is due to values labeled as 'abc' in the original dataset.

Card_Category

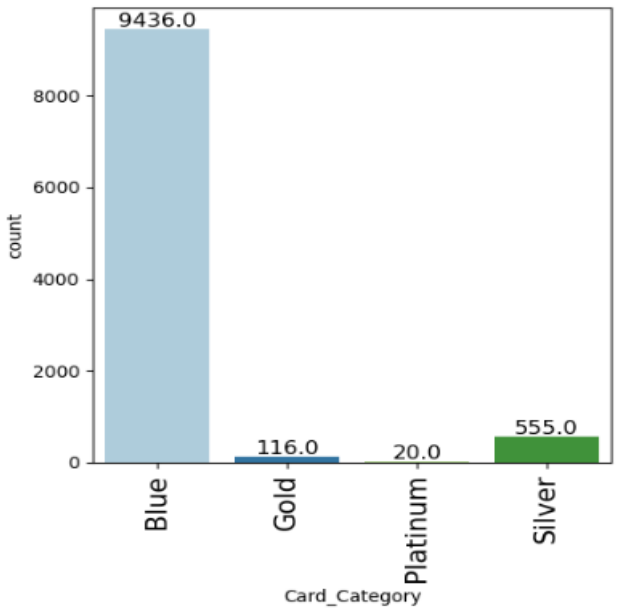

Fig 19. Card_Category

- The **majority of customers 94%,** belong to the Blue category.
- Only about **1% of customers are in the Gold category**, which is understandable as these individuals likely have high credit or income.
- There are very few observations, approximately 0.2%, in the Platinum category.

Attrition_Flag

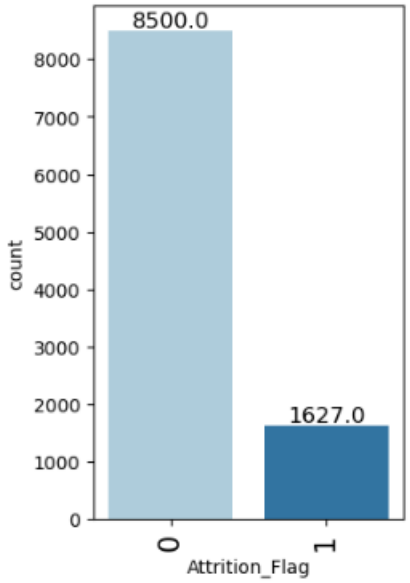

Fig 20. Attrition_Flag

- The **majority of customers are existing ones**.
- As noted earlier, the class distribution in the target variable is imbalanced.
- We have 83% of observations for existing customers and **16% for attritioners**.

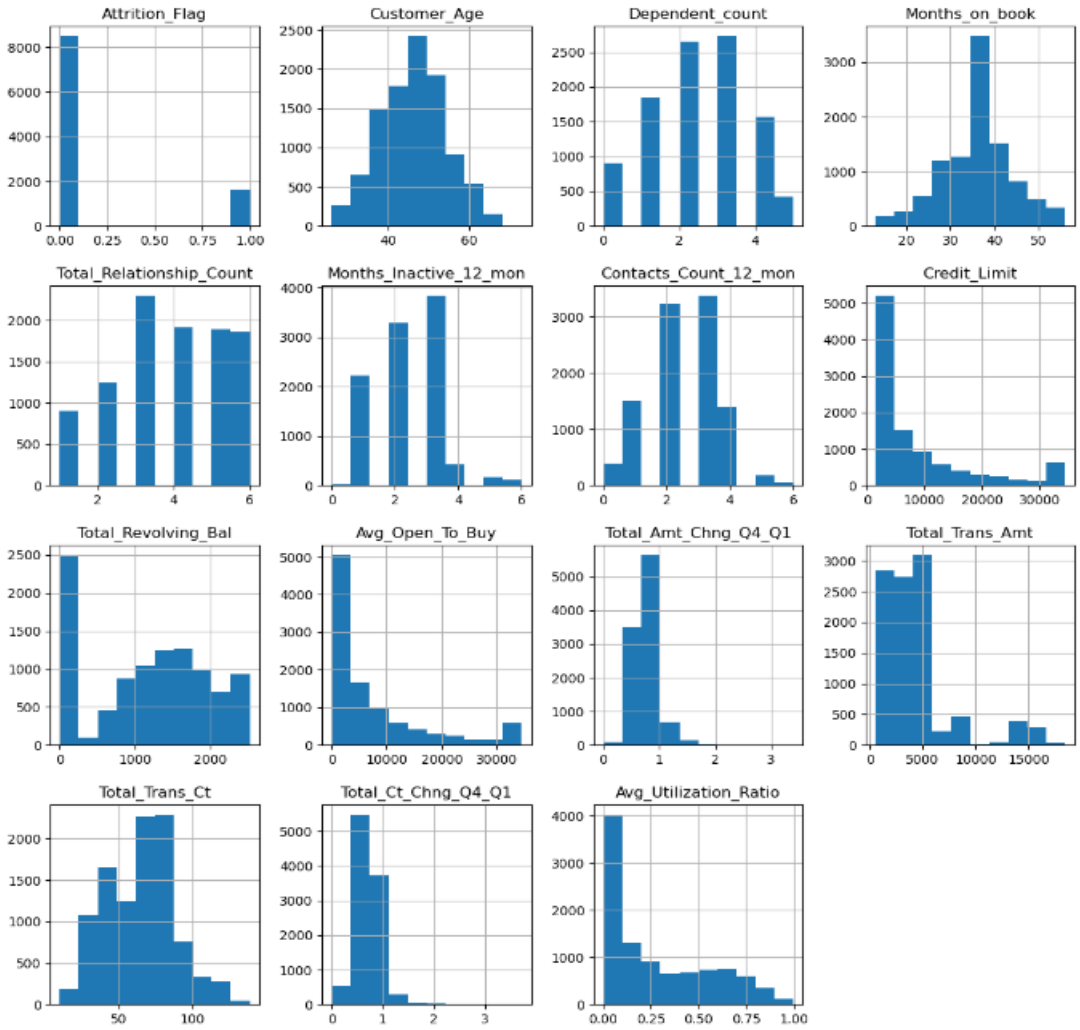

Fig 21. Histogram

Above illustration is the histogram of the dataset.

2.2 Bivariate Analysis

Let's see the attributes that have a strong correlation with each other

Correlation Check

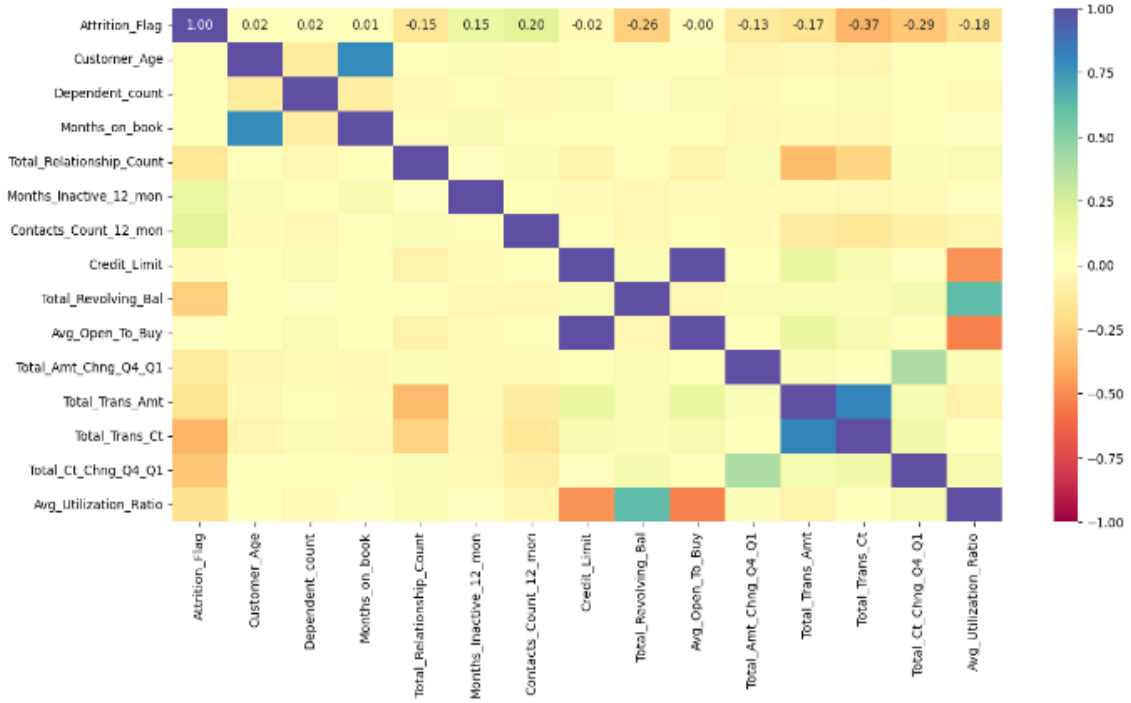

Fig 22. Heatmap

- **Customer age and months on book exhibit a strong positive correlation**, which is logical.
- **Credit limit and utilization ratio** show a **moderate negative correlation**.
- Total revolving balance and utilization ratio are strongly positively correlated.
- There are also other intuitive correlations present.

Attrition_Flag vs Gender

Attrition_Flag	0	1	All
Gender			
All	8500	1627	10127
F	4428	930	5358
M	4072	697	4769

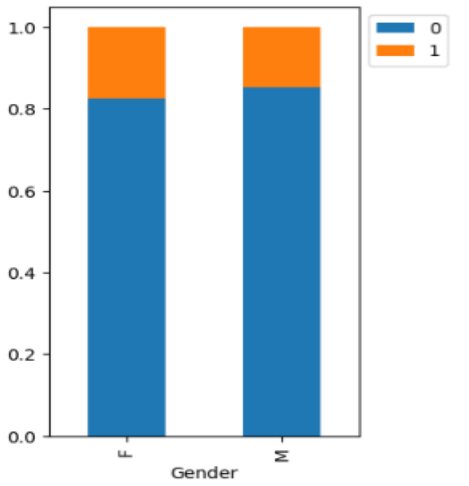

Fig 23. Attrition_Flag vs Gender

- The dataset has a **slightly greater number of female customers** compared to male customers.
- This trend is also seen in the slightly higher number of attrited female customers than attrited male customers.

Attrition_Flag vs Marital_Status

Attrition_Flag	0	1	All
Marital_Status			
All	7880	1498	9378
Married	3978	709	4687
Single	3275	668	3943
Divorced	627	121	748

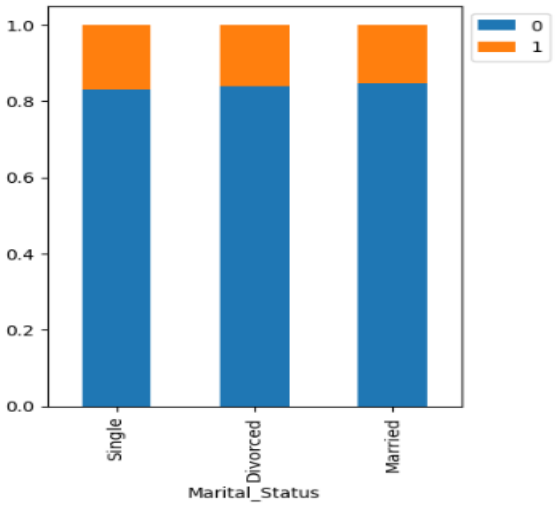

Fig 24. Attrition_Flag vs Marital_Status

- The **majority of customers are married**.
- Single customers make up the next largest group.

Attrition_Flag vs Education_Level

Attrition_Flag	0	1	All
Education_Level			
All	7237	1371	8608
Graduate	2641	487	3128
High School	1707	306	2013
Uneducated	1250	237	1487
College	859	154	1013
Doctorate	356	95	451
Post-Graduate	424	92	516

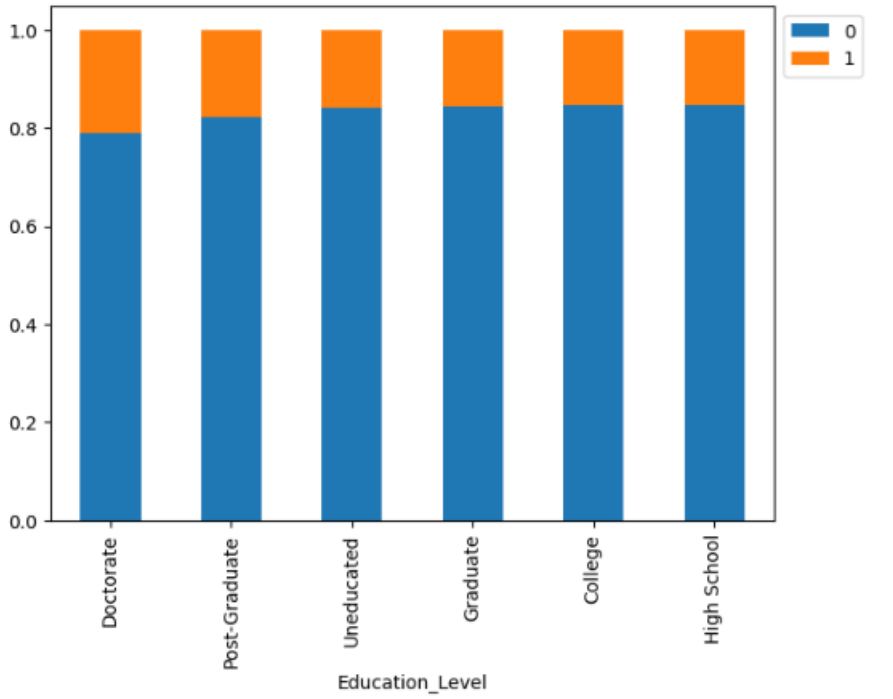

Fig 25. Attrition_Flag vs Education_Level

- The distribution of values in the 'Education_Level' column reveals that the **majority of customers are graduates.**
- This is followed by customers with a high school education, and then those classified as uneducated.

Attrition_Flag vs Income_Category

Attrition_Flag	0	1	All
Income_Category			
All	8500	1627	10127
Less than $40K	2949	612	3561
$40K - $60K	1519	271	1790
$80K - $120K	1293	242	1535
$60K - $80K	1213	189	1402
abc	925	187	1112
$120K +	601	126	727

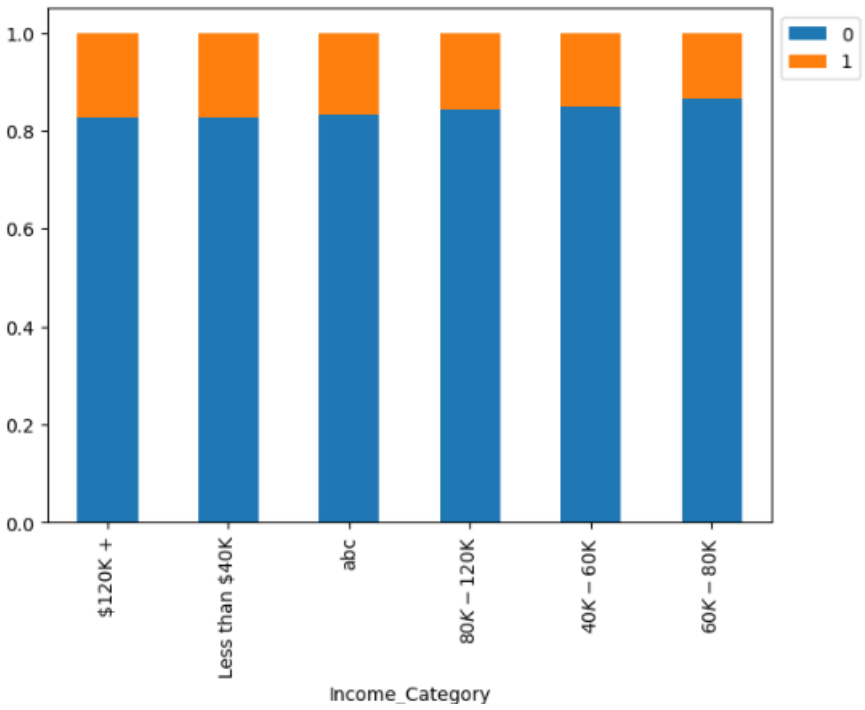

Fig 26. Attrition_Flag vs Income_Category

- Customers from the income groups earning **less than 40K** exhibit the highest attrition rates.
- This suggests that both low and high-income earners are more likely to attrite.

Attrition_Flag vs Contacts_Count_12_mon

Attrition_Flag	0	1	All
Contacts_Count_12_mon			
All	8500	1627	10127
3	2699	681	3380
2	2824	403	3227
4	1077	315	1392
1	1391	108	1499
5	117	59	176
6	0	54	54
0	392	7	399

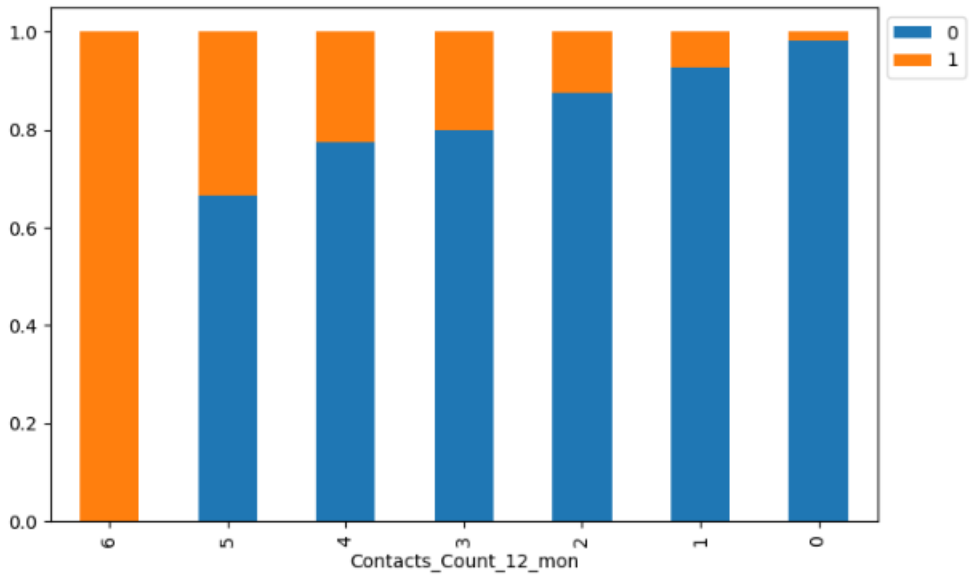

Fig 27. Attrition_Flag vs Contacts_Count_12_mon

- **Most customers have been contacted 2 to 3 times** in the past 12 months.
- **Attrited customers** seem to have **received more contact**.

Let's see the number of months a customer was inactive in the last 12 months (Months_Inactive_12_mon) vary by the customer's account status (Attrition_Flag)

Attrition_Flag vs Months_Inactive_12_mon

Attrition_Flag	0	1	All
Months_Inactive_12_mon			
All	8500	1627	10127
3	3020	826	3846
2	2777	505	3282
4	305	130	435
1	2133	100	2233
5	146	32	178
6	105	19	124
0	14	15	29

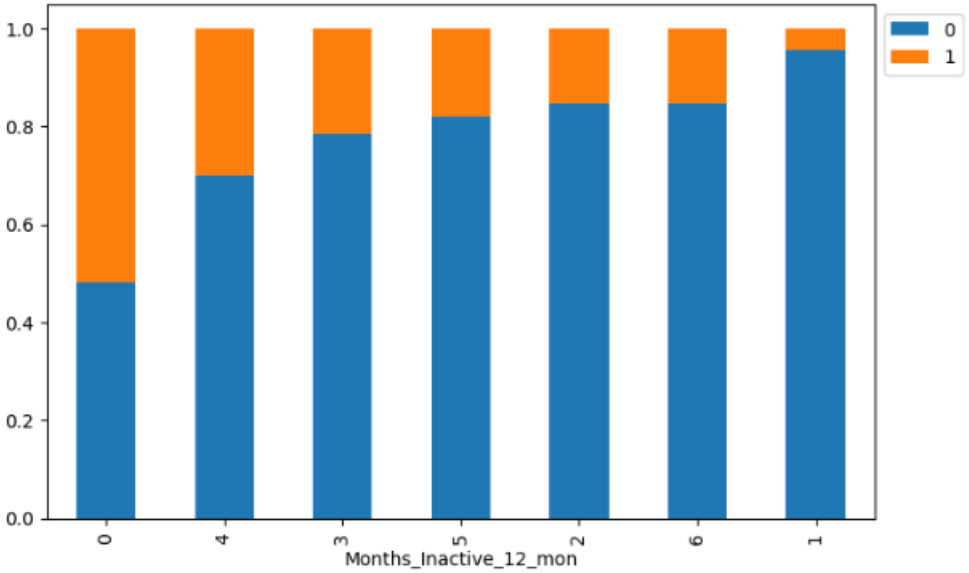

Fig 28. Attrition_Flag vs Months_Inactive_12_mon

- **Most customers have been inactive for 1 to 3 months** over the past year, with the average duration slightly exceeding 2 months.
- Attrited customers generally exhibit a longer period of inactivity compared to existing customers.

Attrition_Flag vs Total_Relationship_Count

Attrition_Flag	0	1	All
Total_Relationship_Count			
All	8500	1627	10127
3	1905	400	2305
2	897	346	1243
1	677	233	910
5	1664	227	1891
4	1687	225	1912
6	1670	196	1866

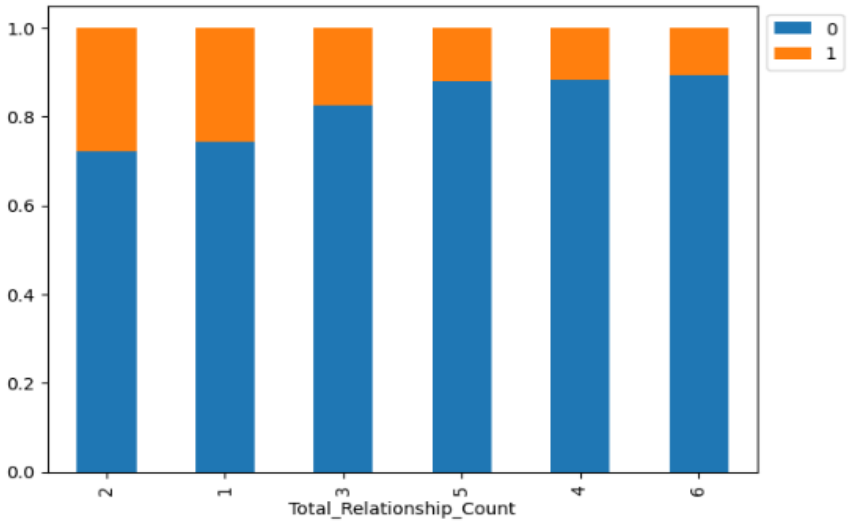

Fig 29. Attrition_Flag vs Total_Relationship_Count

- **Most customers hold three or four products** with the bank.
- Attrited customers typically have fewer products with the bank.

Attrition_Flag vs Dependent_count

Attrition_Flag	0	1	All
Dependent_count			
All	8500	1627	10127
3	2250	482	2732
2	2238	417	2655
1	1569	269	1838
4	1314	260	1574
0	769	135	904
5	360	64	424

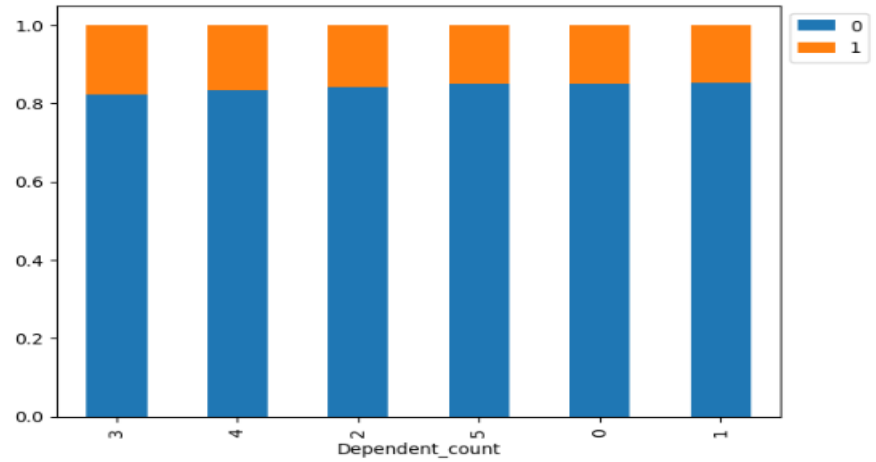

Fig 30. Attrition_Flag vs Dependent_count

- Higher numbers of dependents are associated with increased attrition, as greater responsibilities may lead to financial instability for these customers.
- **Attrition rates are relatively low** for customers with 0 or 1 dependent.

Total_Revolving_Bal vs Attrition_Flag

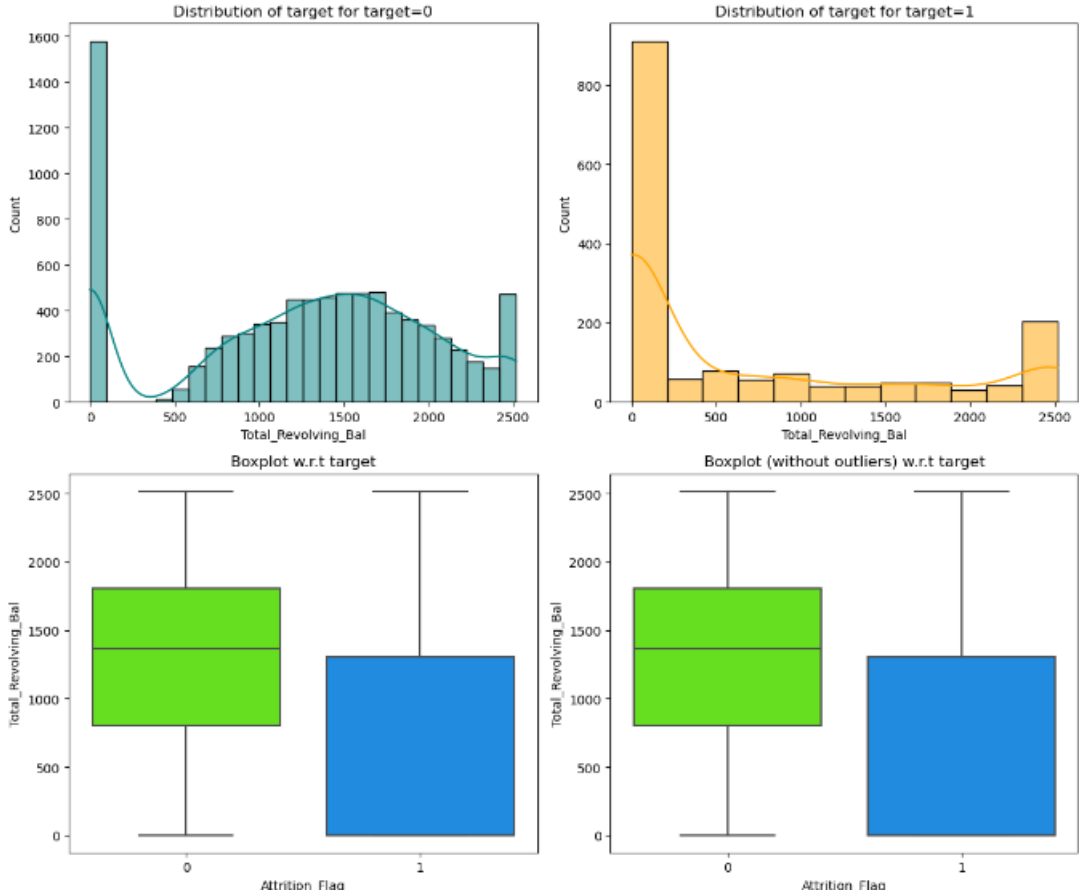

Fig 31. Distribution plot on Total_Revolving_Bal vs Attrition_Flag

- Customers with **lower total revolving balances tend to attrite**.
- These customers **likely paid off their dues** and **opted out of the credit card service**.

Attrition_Flag vs Credit_Limit

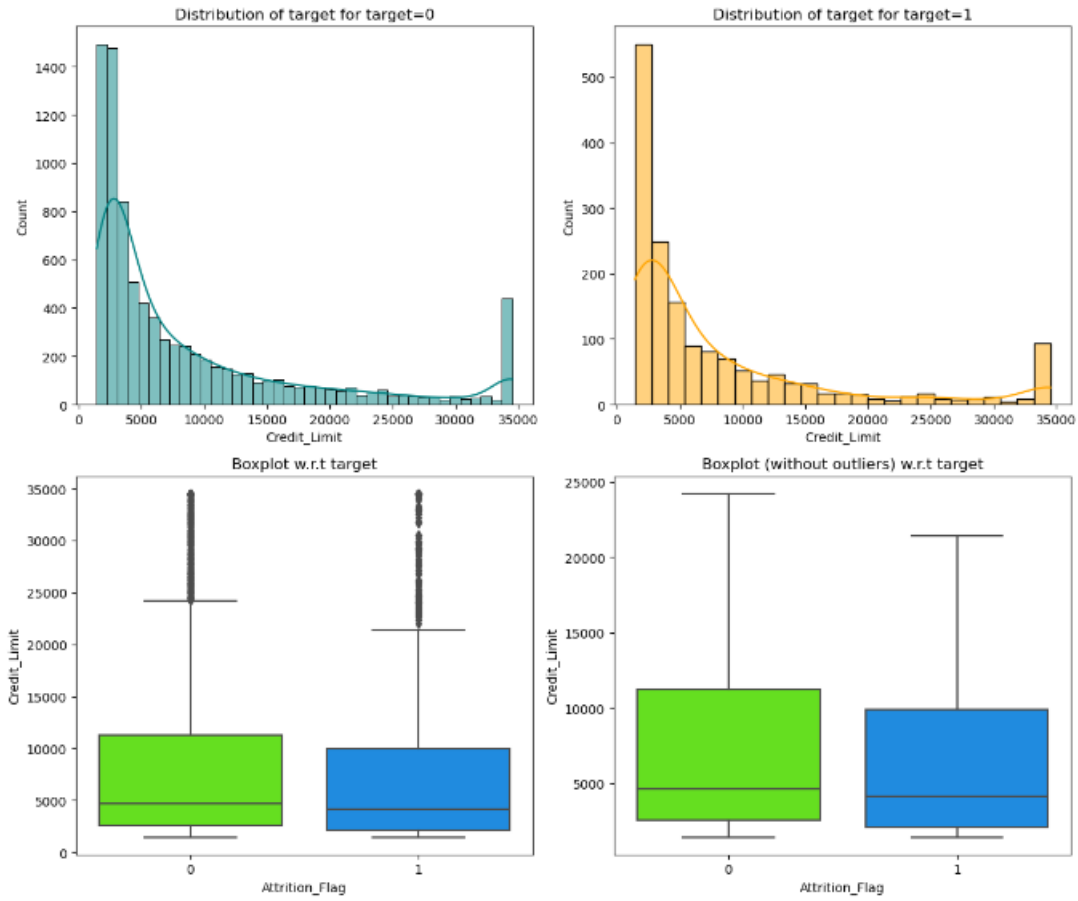

Fig 32. Distribution plot on Attrition_Flag vs Credit_Limit

- **Most customers have** a credit limit **below $10,000**.
- The data shows a **significant right skew**.

Attrition_Flag vs Customer_Age

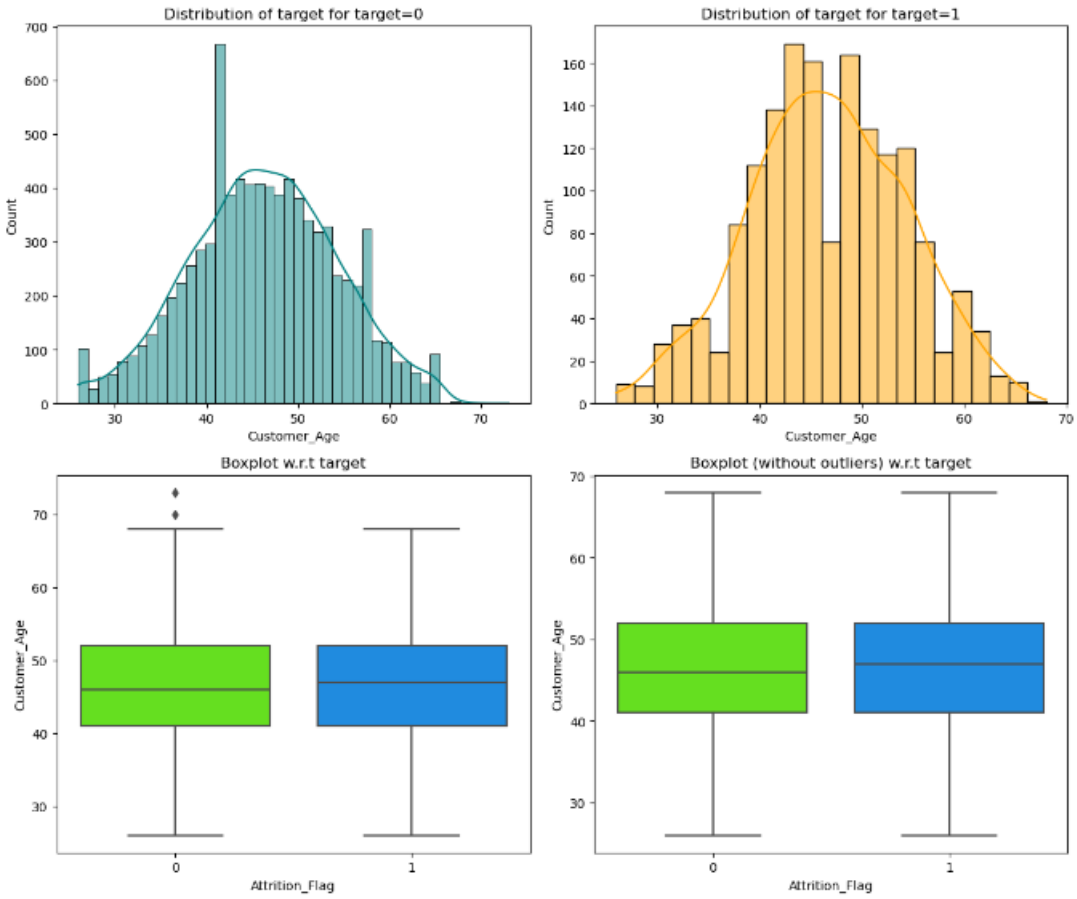

Fig 33. Distribution plot on Attrition_Flag vs Customer_Age

- The **distribution is approximately normal**, with most customers being in their 40s and 50s.
- The boxplot also shows a **few outliers** on the higher age range.

Total_Trans_Ct vs Attrition_Flag

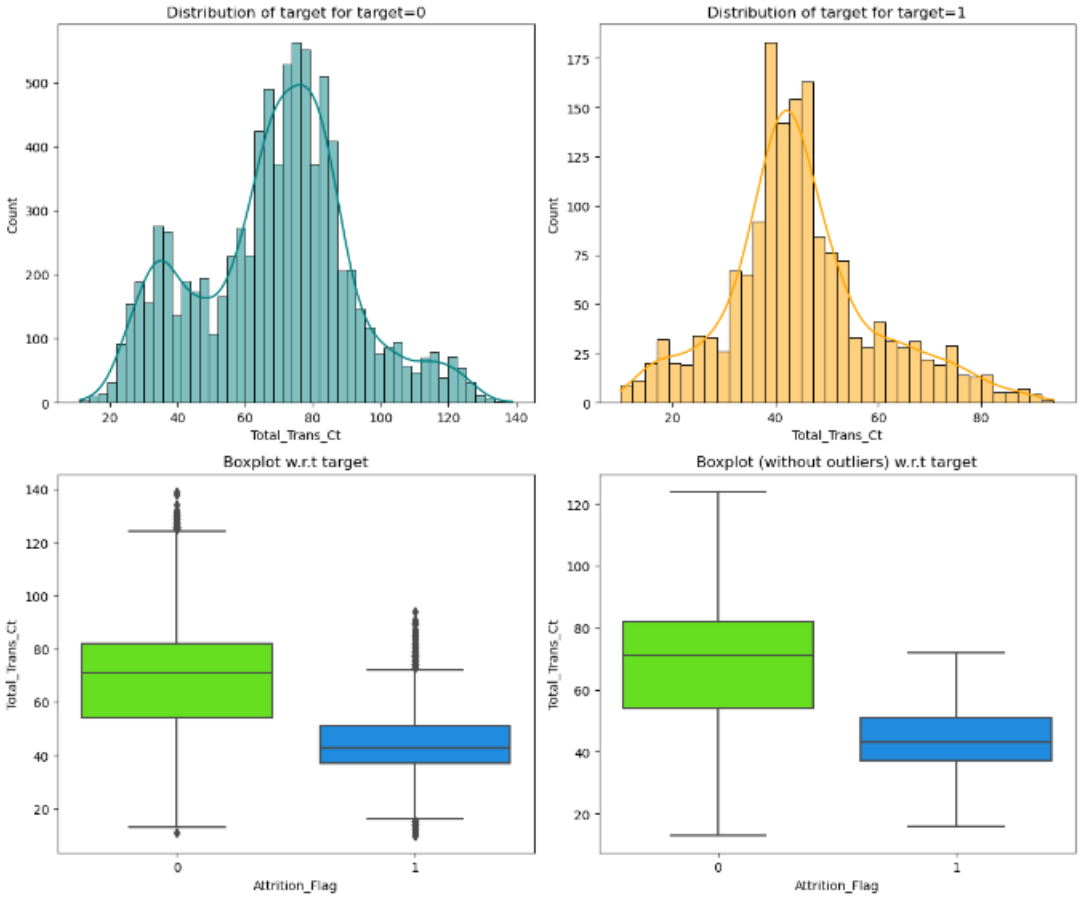

Fig 34. Distribution plot on Total_Trans_Ct vs Attrition_Flag

- **Most customers** have a total transaction count **between 60 and 100**.
- **Attrited customers** generally have a **lower total transaction** count than those who are still active.

Total_Trans_Amt vs Attrition_Flag

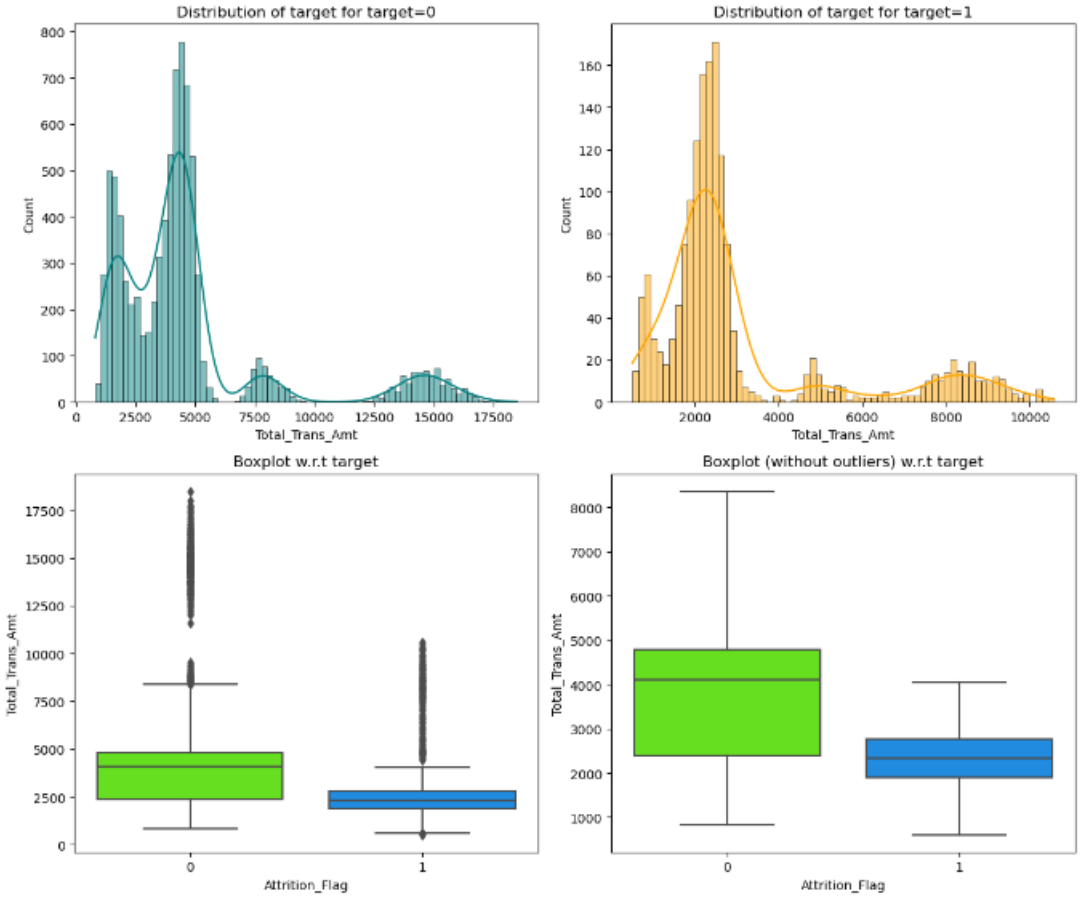

Fig 35. Distribution plot on Total_Trans_Amt vs Attrition_Flag

- The **majority of customers** have a total transaction amount of **under $5,000.**
- **Attrited customers typically have a lower** total transaction amount than active customers.

Let's see the change in transaction amount between Q4 and Q1 (total_ct_change_Q4_Q1) vary by the customer's account status (Attrition_Flag)

Total_Ct_Chng_Q4_Q1 vs Attrition_Flag

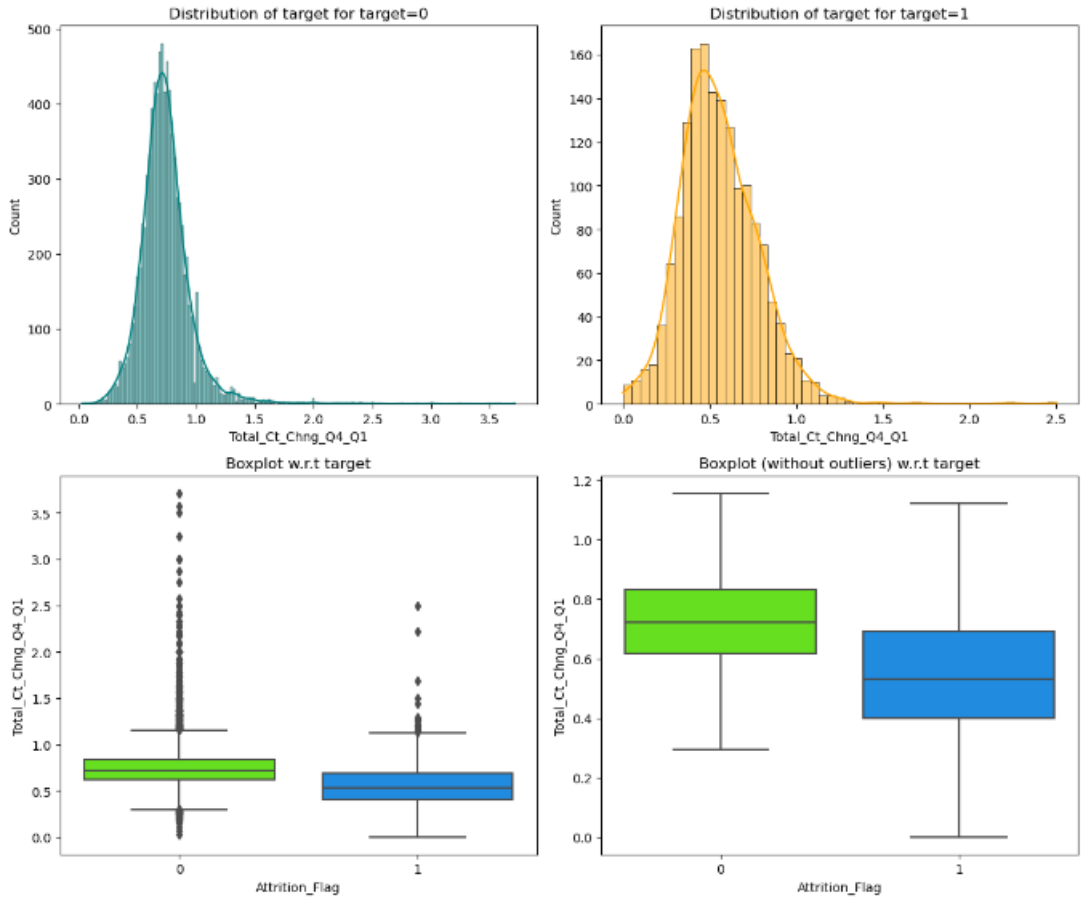

Fig 36. Distribution plot on Total_Ct_Chng_Q4_Q1 vs Attrition_Flag

- Most customers have a ratio of **approximately 0.7**.
- **Attrited** customers usually have a **lower ratio**.

Avg_Utilization_Ratio vs Attrition_Flag

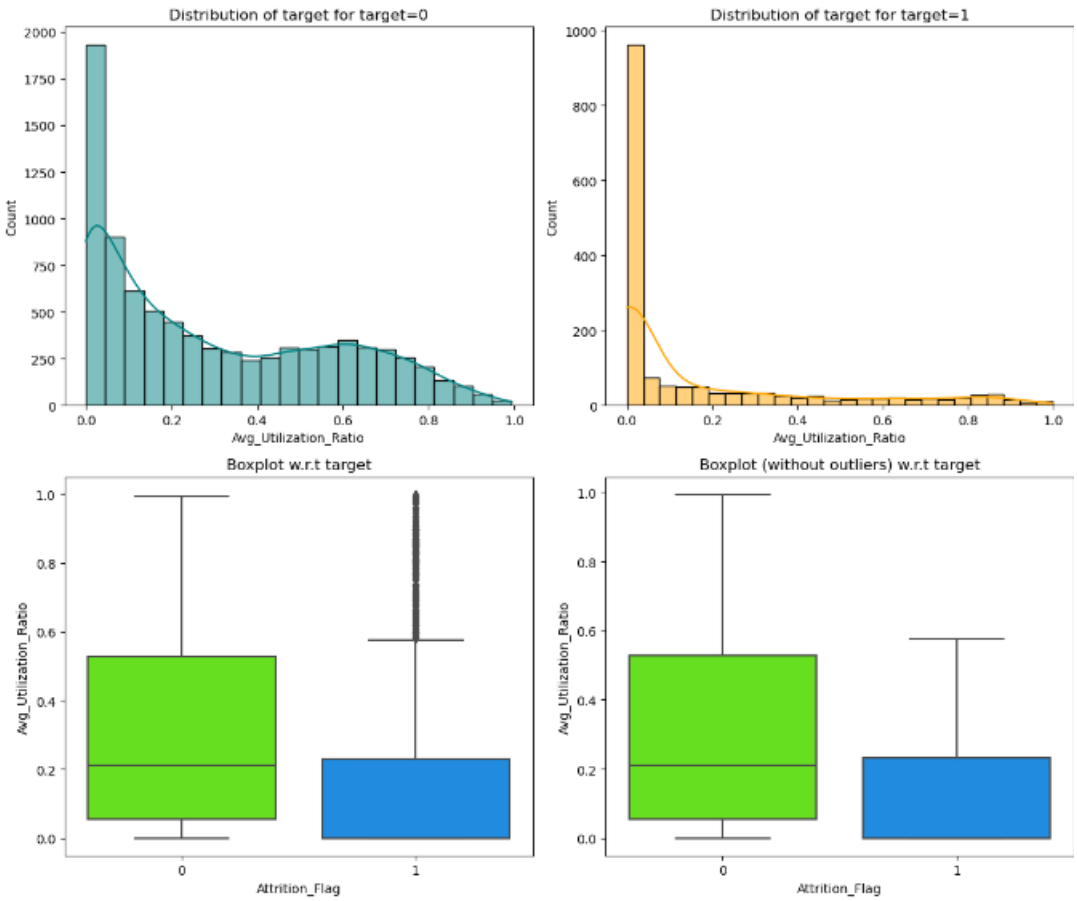

Fig 37. Distribution plot on Avg_Utilization_Ratio vs Attrition_Flag

- **Many customers maintain a low utilization ratio**, with a notable portion having a ratio close to 0.
- **Attrited customers have a higher percentage** with a near-zero utilization ratio compared to those who remain active.

Attrition_Flag vs Months_on_book

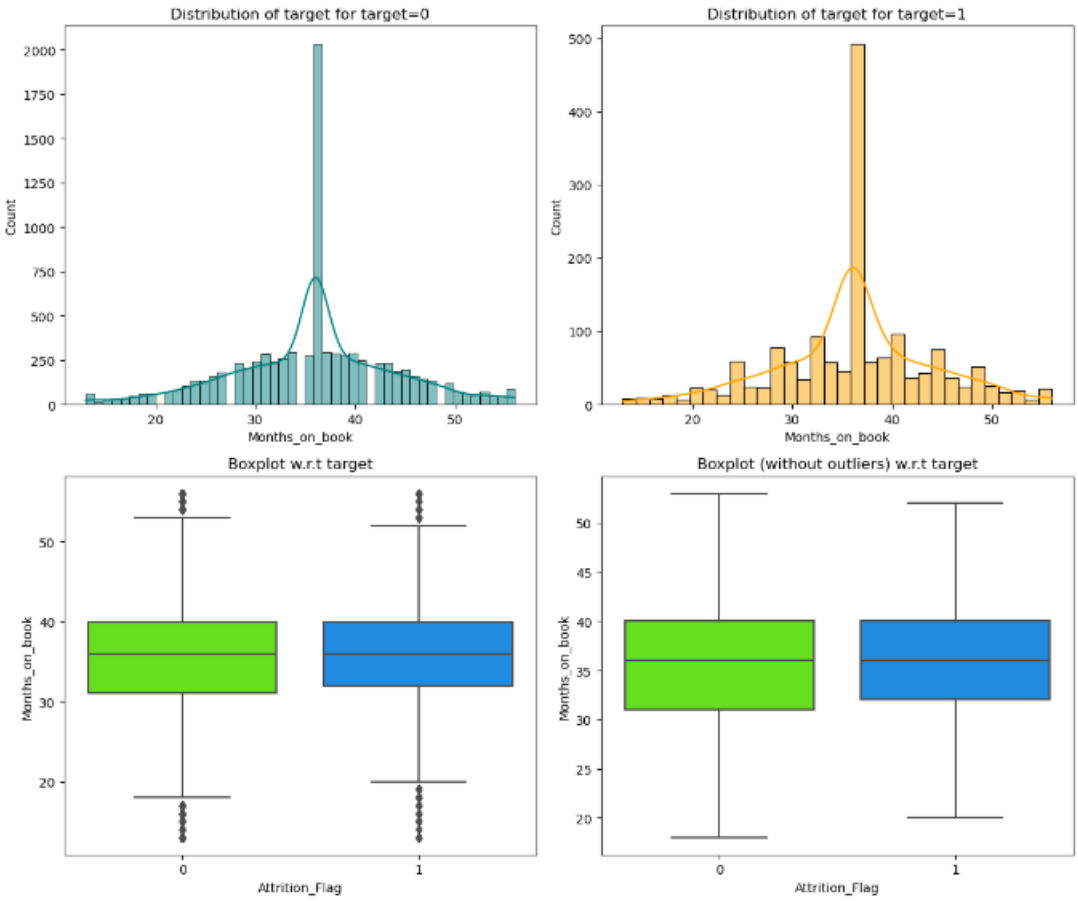

Fig 38. Distribution plot on Attrition_Flag vs Months_on_book

- Most customers have been with the bank for **30 to 40 months**, peaking around 36 months.
- The **distribution is roughly normal** but has a **slight rightward skew**.
- There is no notable difference in the time on book between attrited and active customers.

Attrition_Flag vs Total_Revolving_Bal

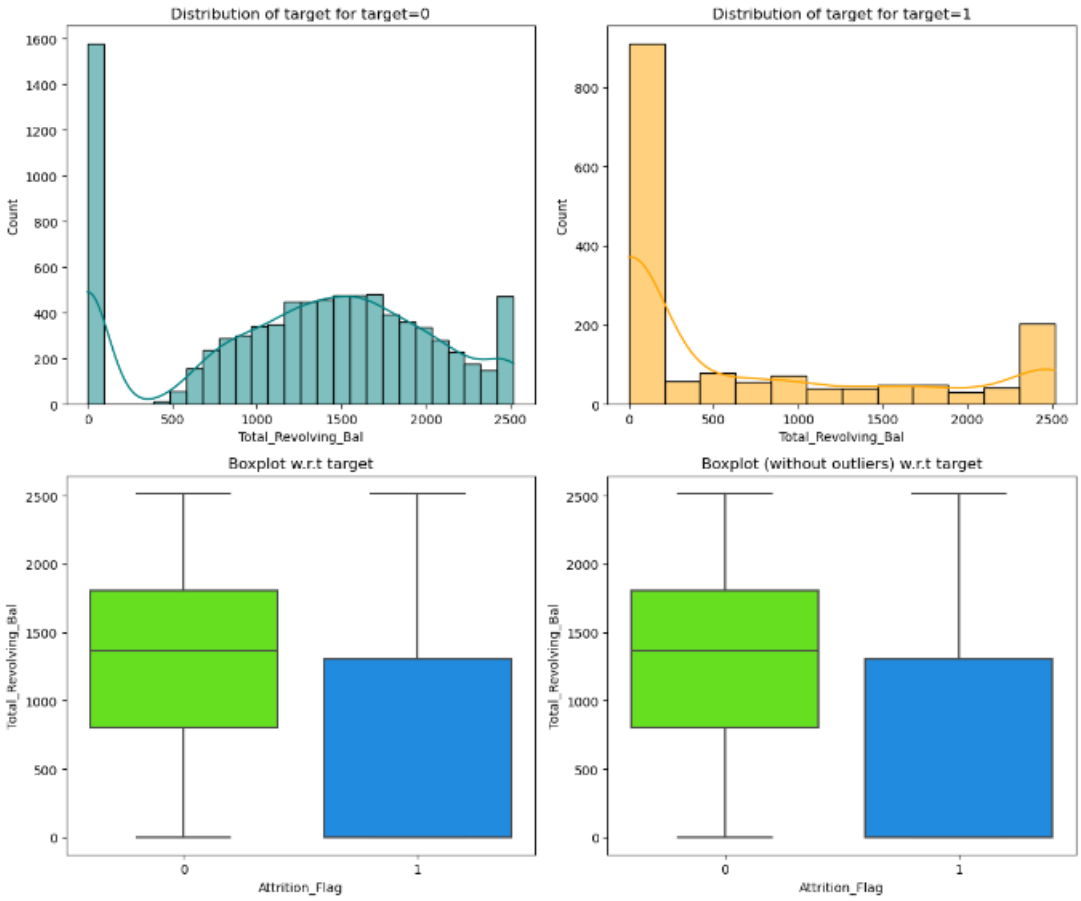

Fig 39. Distribution plot on Attrition_Flag vs Total_Revolving_Bal

- **Many customers** maintain a **low revolving balance**.
- **Attrited customers** generally have even **lower revolving balances**.

Attrition_Flag vs Avg_Open_To_Buy

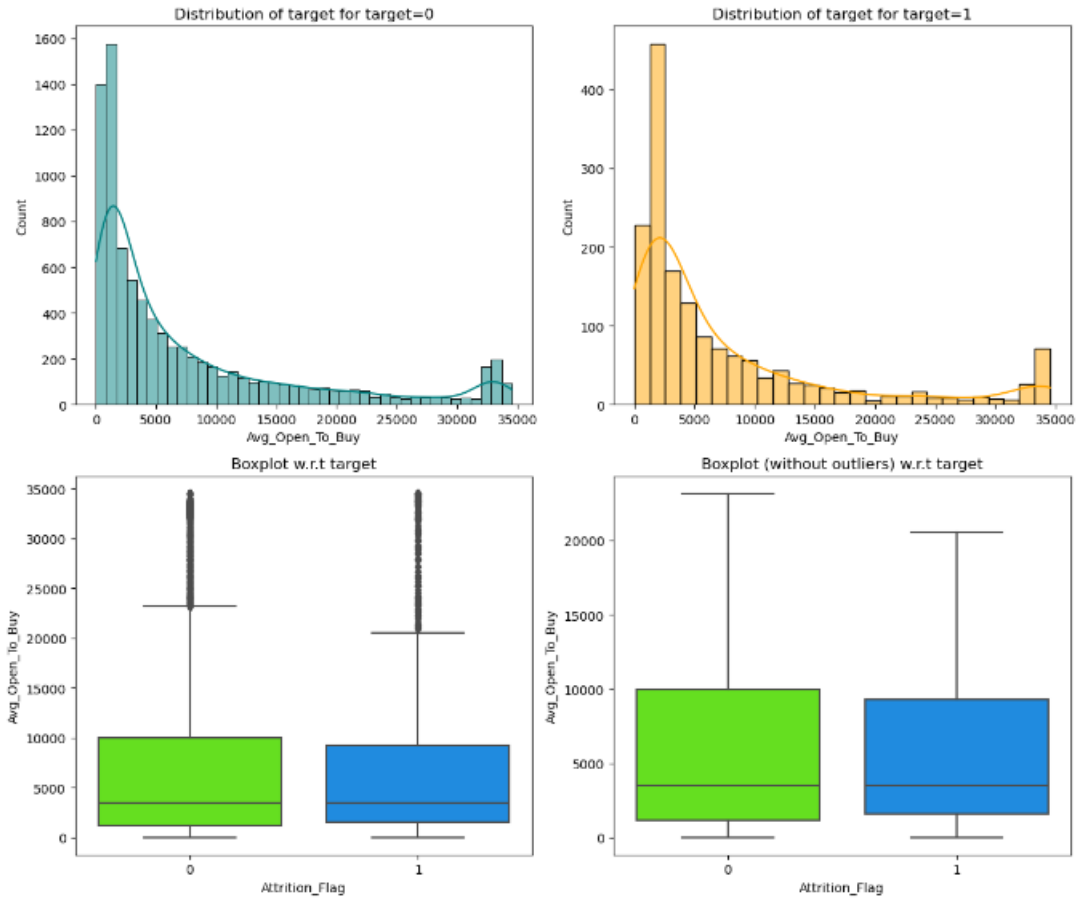

Fig 40. Distribution plot on Attrition_Flag vs Avg_Open_To_Buy

- **Most customers have less than $10,000** in available credit.
- The **highest** concentration is **around $2,000**.

3. Data Preprocessing

3.1 Outlier Detection

Attrition_Flag	16.066
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
dtype: float64	

Table 8. Outlier Detection

None of these values appear to be incorrect. Income levels vary significantly among individuals, and some may reach out to the bank more often than others.

3.2 Train-Test Split

Attrition_Flag	0
Customer_Age	0
Gender	0
Dependent_count	0
Education_Level	1519
Marital_Status	749
Income_Category	1112
Card_Category	0
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
Total_Ct_Chng_Q4_Q1	0
Avg_Utilization_Ratio	0
dtype: int64	

Table 9. Data info

The anomalous **values have been successfully replaced** with new ones to ensure accuracy and relevance in the data.

3.3 Training and validation set

```
(6075, 19) (2026, 19) (2026, 19)
```

Table 10. Train and test data shape

After splitting the data into **train test in the ratio 80:20** and **train test in the ratio 75:25**, the shape of Train, Validation and test data are (6075, 19), (2026, 19) and (2026, 19) respectively.

3.4 Missing value imputation

```
Customer_Age              0
Gender                    0
Dependent_count           0
Education_Level           0
Marital_Status            0
Income_Category           0
Card_Category             0
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
Total_Ct_Chng_Q4_Q1       0
Avg_Utilization_Ratio     0
dtype: int64
------------------------------
Customer_Age              0
Gender                    0
Dependent_count           0
Education_Level           0
Marital_Status            0
Income_Category           0
Card_Category             0
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
Total_Ct_Chng_Q4_Q1       0
Avg_Utilization_Ratio     0
dtype: int64
------------------------------
Customer_Age              0
Gender                    0
Dependent_count           0
Education_Level           0
Marital_Status            0
Income_Category           0
Card_Category             0
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
Total_Ct_Chng_Q4_Q1       0
Avg_Utilization_Ratio     0
dtype: int64
```

Table 11. Check missing values

All missing values have been addressed and handled appropriately in test and train data. And checked that no column has missing values.

```
Gender
F    3193
M    2882
Name: count, dtype: int64
******************************
Education_Level
Graduate         2782
High School      1228
Uneducated        881
College           618
Post-Graduate     312
Doctorate         254
Name: count, dtype: int64
******************************
Marital_Status
Married     3276
Single      2369
Divorced     430
Name: count, dtype: int64
******************************
Income_Category
Less than $40K    2783
$40K - $60K       1059
$80K - $120K       953
$60K - $80K        831
$120K +            449
Name: count, dtype: int64
******************************
Card_Category
Blue        5655
Silver       339
Gold          69
Platinum      12
Name: count, dtype: int64
******************************
```

Table 12. Train data

```
Gender
F    1095
M     931
Name: count, dtype: int64
******************************
Education_Level
Graduate         917
High School      404
Uneducated       306
College          199
Post-Graduate    101
Doctorate         99
Name: count, dtype: int64
******************************
Marital_Status
Married     1100
Single       770
Divorced     156
Name: count, dtype: int64
******************************
Income_Category
Less than $40K    957
$40K - $60K       361
$80K - $120K      293
$60K - $80K       279
$120K +           136
Name: count, dtype: int64
******************************
Card_Category
Blue        1905
Silver        97
Gold          21
Platinum       3
Name: count, dtype: int64
******************************
```

Table 13. Validation data

```
Gender
F    3193
M    2882
Name: count, dtype: int64
******************************
Education_Level
Graduate         2782
High School      1228
Uneducated        881
College           618
Post-Graduate     312
Doctorate         254
Name: count, dtype: int64
******************************
Marital_Status
Married     3276
Single      2369
Divorced     430
Name: count, dtype: int64
******************************
Income_Category
Less than $40K    2783
$40K - $60K       1059
$80K - $120K       953
$60K - $80K        831
$120K +            449
Name: count, dtype: int64
******************************
Card_Category
Blue        5655
Silver       339
Gold          69
Platinum      12
Name: count, dtype: int64
******************************
```

Table 14. Test data

3.5 Encoding categorical variables

```
(6075, 29) (2026, 29) (2026, 29)
```

Table 15. Impute missing values

	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal
800	40	2	21	6	4	3	20056.000	1602
498	44	1	34	6	2	0	2885.000	1895
4356	48	4	36	5	1	2	6798.000	2517
407	41	2	36	6	2	0	27000.000	0
8728	46	4	36	2	2	3	15034.000	1356

Table 16. Top 5 rows after encoding

After encoding there are **29 columns**. And displayed the first 5 rows of dataset.

4. Model Building - Original Data

Model evaluation criterion

Model can make wrong predictions as:

- **Predicting a customer will attrite and the customer doesn't attrite**
- **Predicting a customer will not attrite and the customer attrites**

Which case is more important?

- **Predicting that customer will not attrite but he attrites i.e. losing on a valuable customer or asset.**

How to reduce this loss i.e need to reduce False Negatives??

-Bank would want Recall to be maximized, greater the Recall higher the chances of minimizing false negatives. Hence, the focus should be on increasing Recall or minimizing the false negatives or in other words identifying the true positives(i.e. Class 1) so that the bank can retain their valuable customers by identifying the customers who are at risk of attrition.

Let's define a function to output different metrics (including recall) on the train and test set and a function to show confusion matrix so that we do not have to use the same code repetitively while evaluating models.

```
Training Performance:

Logistic regression: 0.44672131147540983
Bagging: 0.985655737704918
Random forest: 1.0
GBM: 0.875
Adaboost: 0.826844262295082
Xgboost: 1.0
dtree: 1.0

Validation Performance:

Logistic regression: 0.5030674846625767
Bagging: 0.8098159509202454
Random forest: 0.7975460122699386
GBM: 0.8588957055214724
Adaboost: 0.852760736196319
Xgboost: 0.901840490797546
dtree: 0.8098159509202454
```

Table 17. Model Building - original data

After building and appending the models,

- The **Random Forest and Decision Tree models are overfitting the data**.
- The **Gradient Boosting** model scored **low in both** training and validation, which could indicate underfitting.
- **Logistic Regression** showed **moderate performance**.
- **XGBoost, Bagging, and AdaBoost** performed reasonably **well and present good opportunities for enhancement**.

5. Model Building - Oversampled Data

```
Before Oversampling, counts of label 'Yes': 976
Before Oversampling, counts of label 'No': 5099

After Oversampling, counts of label 'Yes': 5099
After Oversampling, counts of label 'No': 5099

After Oversampling, the shape of train_X: (10198, 29)
After Oversampling, the shape of train_y: (10198,)
```

Table 18. Model Building – Oversampled data 1

```
Training Performance:

Logistic regression: 0.803098646793489
Bagging: 0.9976465973720338
Random forest: 1.0
GBM: 0.9792116101196313
Adaboost: 0.964698960580506
Xgboost: 1.0
dtree: 1.0

Validation Performance:

Logistic regression: 0.7760736196319018
Bagging: 0.8619631901840491
Random forest: 0.8680981595092024
GBM: 0.9049079754601227
Adaboost: 0.901840490797546
Xgboost: 0.9294478527607362
dtree: 0.8650306748466258
```

Table 19. Model Building – Oversampled data 2

After building models with oversampled data,

- **The Random Forest and Decision Tree** models **continue to overfit the data**.
- **Logistic Regression** achieved the **best performance**.
- **XGBoost, Bagging, AdaBoost, and Gradient Boosting Machine** (GBM) all performed well.

6. Model building - Undersampled data

```
Before Under Sampling, counts of label 'Yes': 976
Before Under Sampling, counts of label 'No': 5099

After Under Sampling, counts of label 'Yes': 976
After Under Sampling, counts of label 'No': 976

After Under Sampling, the shape of train_X: (1952, 29)
After Under Sampling, the shape of train_y: (1952,)
```

Table 20. Model Building – Undersampled data 1

```
Training Performance:

Logistic regression: 0.8217213114754098
Bagging: 0.9907786885245902
Random forest: 1.0
GBM: 0.9805327868852459
Adaboost: 0.9528688524590164
Xgboost: 1.0
dtree: 1.0

Validation Performance:

Logistic regression: 0.8251533742331288
Bagging: 0.9294478527607362
Random forest: 0.9355828220858896
GBM: 0.9570552147239264
Adaboost: 0.9601226993865031
Xgboost: 0.9693251533742331
dtree: 0.9202453987730062
```

Table 21. Model Building – Undersampled data 2

After building models on undersampled data,

- Overfitting in the **Random Forest and Decision Tree models** has decreased, although they **still exhibit some overfitting**.
- **XGBoost and AdaBoost** appear to be delivering the **best performance**.
- Logistic Regression continues to perform well.

7. Model Performance Improvement using Hyperparameter Tuning

7.1 Tuning AdaBoost using original data

Created new pipeline with best parameters obtained from tuning and **fit the model on original data**.

```
Best parameters are {'n_estimators': 100, 'learning_rate': 0.05, 'base_estimator': DecisionTreeClassifier(max_depth=3, random_state=1)} with CV score=0.
9467346938775512:
CPU times: user 631 ms, sys: 124 ms, total: 755 ms
Wall time: 7.09 s
```

Table 22. Adaboost – original data 1

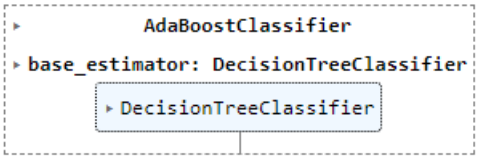

Table 23. Adaboost – original data 2

	Accuracy	Recall	Precision	F1
0	1.000	1.000	1.000	1.000

Table 24. Adaboost – Training set

	Accuracy	Recall	Precision	F1
0	0.934	0.960	0.723	0.825

Table 25. Adaboost – Validation set

Checked the performance on both Training set and Validation set.

7.2 Tuning AdaBoost using undersampled data

Created new pipeline with best parameters obtained from tuning and **fit the model on undersampled data**.

```
Best parameters are {'n_estimators': 90, 'learning_rate': 1, 'base_estimator': DecisionTreeClassifier(max_depth=2, random_state=1)} with CV score=0.8
647409733124019:
CPU times: user 1.26 s, sys: 294 ms, total: 1.55 s
Wall time: 22.1 s
```

Table 26. Adaboost – Undersampled data 1

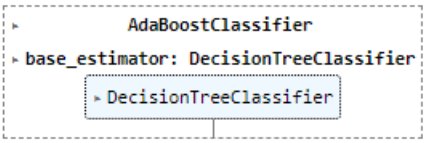

Table 27. Adaboost – Undersampled data 2

	Accuracy	Recall	Precision	F1
0	1.000	1.000	1.000	1.000

Table 28. Undersampled – Training set

	Accuracy	Recall	Precision	F1
0	0.968	0.877	0.920	0.898

Table 29. Undersampled – validation set

Checked the performance on both Training set and Validation set.

7.3 Tuning Gradient Boosting using undersampled data

Created new pipeline with best parameters obtained from tuning Gradient Boosting and **fit the model on undersampled data**.

```
Best parameters are {'subsample': 0.9, 'n_estimators': 75, 'max_features': 0.7, 'learning_rate': 0.1, 'init': AdaBoostClassifier(random_state=1)} with C
V score=0.9508267922553637:
CPU times: user 752 ms, sys: 285 ms, total: 1.04 s
Wall time: 9.6 s
```

Table 30. Gradient Boosting - Undersampled data 1

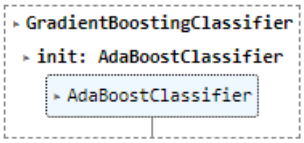

Table 31. Gradient Boosting - Undersampled data 2

	Accuracy	Recall	Precision	F1
0	0.970	0.977	0.964	0.970

Table 32. Gradient Boosting – Training set

	Accuracy	Recall	Precision	F1
0	0.938	0.957	0.738	0.833

Table 33. Gradient Boosting – validation set

Checked the performance on both the undersampled Training set and undersampled Validation set.

7.4 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=0.8104395604395604:
CPU times: user 1.26 s, sys: 376 ms, total: 1.64 s
Wall time: 20.2 s
```

Table 34. Gradient Boosting – original data 1

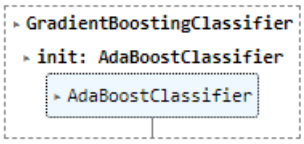

Table 35. Gradient Boosting – original data 2

Created new pipeline with best parameters obtained from tuning Gradient Boosting and **fit the model on original data**.

7.5 Tuning Gradient Boosting using over sampled data

	Accuracy	Recall	Precision	F1
0	0.943	0.977	0.745	0.846

Table 36. Gradient Boosting – Oversampled train set

	Accuracy	Recall	Precision	F1
0	0.938	0.957	0.738	0.833

Table 37. Gradient Boosting – Oversampled validation set

Checked the performance on both oversampled train set and validation set.

7.6 Tuning XGBoost Model with Original data

Created new pipeline with best parameters obtained from tuning XGBoost and **fit the model on original data**.

```
Best parameters are {'subsample': 0.7, 'scale_pos_weight': 5, 'n_estimators': 50, 'learning_rate': 0.01, 'gamma': 3} with CV score=0.9979591836734695:
CPU times: user 615 ms, sys: 238 ms, total: 853 ms
Wall time: 4.51 s
```

Table 38. XGBoost – Original data 1

```
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, device=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='logloss',
              feature_types=None, gamma=1, grow_policy=None,
              importance_type=None, interaction_constraints=None,
              learning_rate=0.01, max_bin=None, max_cat_threshold=None,
              max_cat_to_onehot=None, max_delta_step=None, max_depth=None,
              max_leaves=None, min_child_weight=None, missing=nan,
              monotone_constraints=None, multi_strategy=None, n_estimators=50,
              n_jobs=None, num_parallel_tree=None, random_state=1, ...)
```

Table 39. XGBoost – Original data 2

	Accuracy	Recall	Precision	F1
0	0.577	1.000	0.275	0.432

Table 40. XGBoost - train set

	Accuracy	Recall	Precision	F1
0	0.584	1.000	0.279	0.436

Table 41. XGBoost - validation set

Checked the performance on both original train set and validation set.

8. Model Performance Comparison and Final Model Selection

Training performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Undersampled data
Accuracy	0.996	0.963	0.992
Recall	0.997	0.997	0.967
Precision	0.995	0.813	0.982
F1	0.996	0.896	0.975

Table 42. Training performance comparison

Validation performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Undersampled data
Accuracy	0.996	0.963	0.992
Recall	0.997	0.997	0.967
Precision	0.995	0.813	0.982
F1	0.996	0.896	0.975

Table 43. Validation performance comparison

Now we have our final model, so let's find out how our final model is performing on unseen test data.

The **Gradient Boosting model trained on the original data** demonstrates **strong generalization performance**, so we will consider it the best model.

	Accuracy	Recall	Precision	F1
0	0.975	0.914	0.931	0.922

Table 44. Performance on test set

Checked the performance of best model on test data.

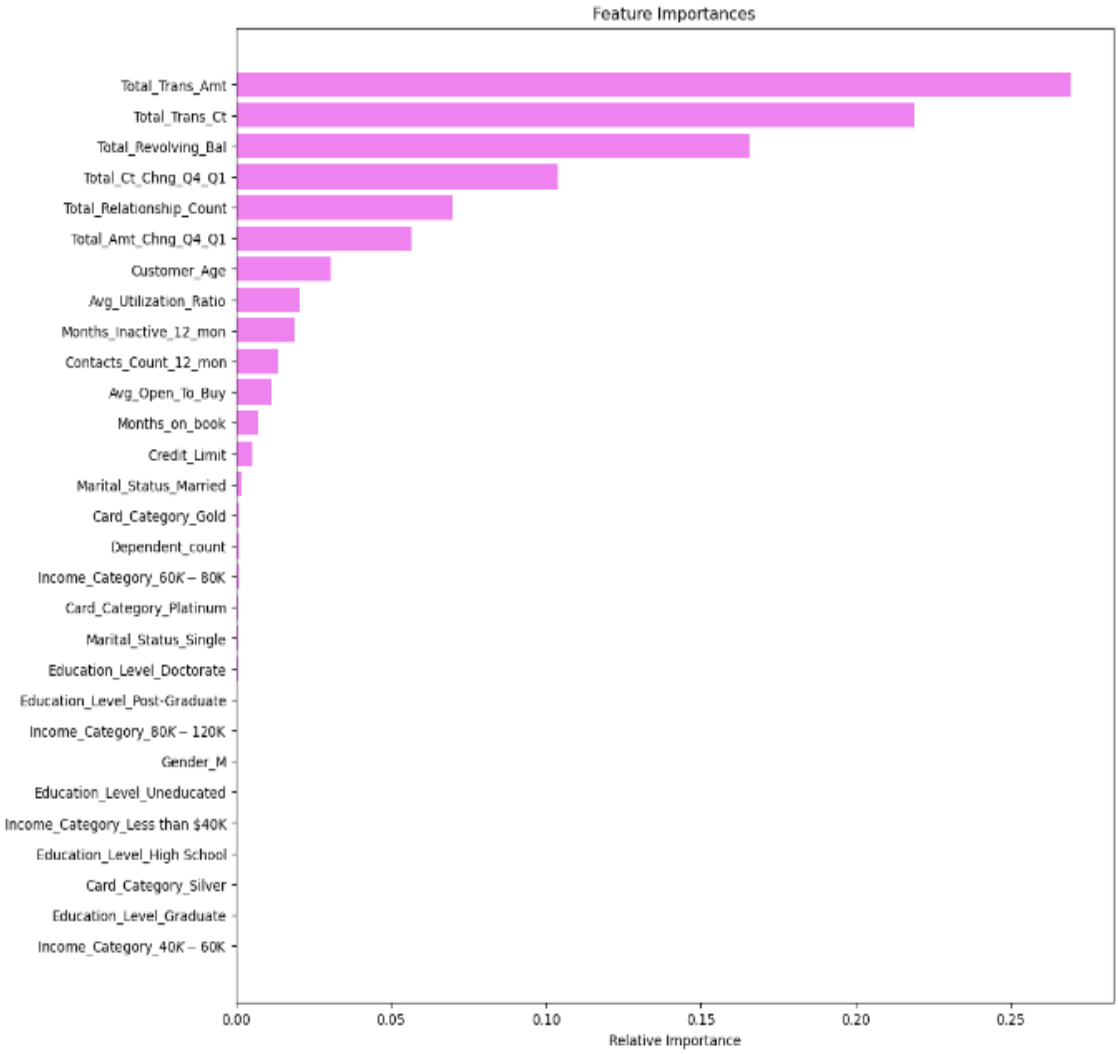

Fig 41. Feature importances

The **total count of transactions ranks as the most important** feature, **followed by the total revolving monthly balance** and the total number of products a customer has with the bank, when predicting whether a customer will attrite.

9. Actionable Insights & Recommendations

Actionable Insights

1. **Key Factors Contributing to Attrition:** Total_Trans_Ct, Total_Revolving_Bal, Total_Trans_Amt, and Total_Relationship_Count. We have successfully developed a predictive model,

- the bank can implement to identify customers at risk of attrition.
- it can help the bank uncover the key factors contributing to attrition.
- the bank can base appropriate actions to enhance customer retention strategies.

2. **Transaction Amount**: A lower number of transactions can result in decreased total transaction amounts, which may lead to customer attrition.

3. **Transaction Count**: A reduced number of transactions per year can contribute to customer attrition. To encourage card usage, the bank could implement incentives such as cashback offers or special discounts, motivating customers to use their cards more frequently.

4. **Months Inactive**: Increased inactivity correlates with higher attrition rates, particularly when customers are inactive for 2 to 4 months. The bank can send automated messages to re-engage customers, sharing updates on their monthly activity, new offers, and available services.

5. **Relationship Count**: Attrition is notably higher among customers who utilize only one or two products from the bank, accounting for approximately 55% of total attrition.

6. **Revolving Balance:** Customers with lower total revolving balances are more likely to attrite, likely having paid off their dues and opted out of credit card services. After customers settle their dues, the bank could solicit feedback on their experiences to better understand the reasons for attrition.

Recommendations

1. **Key Factors Contributing to Attrition:** Total_Trans_Ct, Total_Revolving_Bal, Total_Trans_Amt, and Total_Relationship_Count.

2. The **highest attrition rates** are observed among customers who **frequently interacted with the bank**. This suggests that the bank may be failing to address the issues these customers face, leading to their attrition.

3. Implementing a **feedback collection system** could help **assess customer satisfaction** with the resolutions provided; if customers are dissatisfied, the bank should take appropriate actions to address their concerns.

4. The bank should investigate the issues these customers face with their products; enhancing customer support and transparency could improve retention.

5. To **encourage card usage**, the bank **could implement incentives** such as cashback offers or special discounts, motivating customers to use their cards more frequently.

6. **Female customers should be the focus** of any marketing campaigns, as they tend to utilize their credit more and make higher-value transactions. However, since their credit limits are lower, increasing these limits could be beneficial for the bank.