Credit Card Users Churn Prediction

Ensemble Techniques and Model Tuning
Machine Learning and Artificial Intelligence PG

August 18th 2023

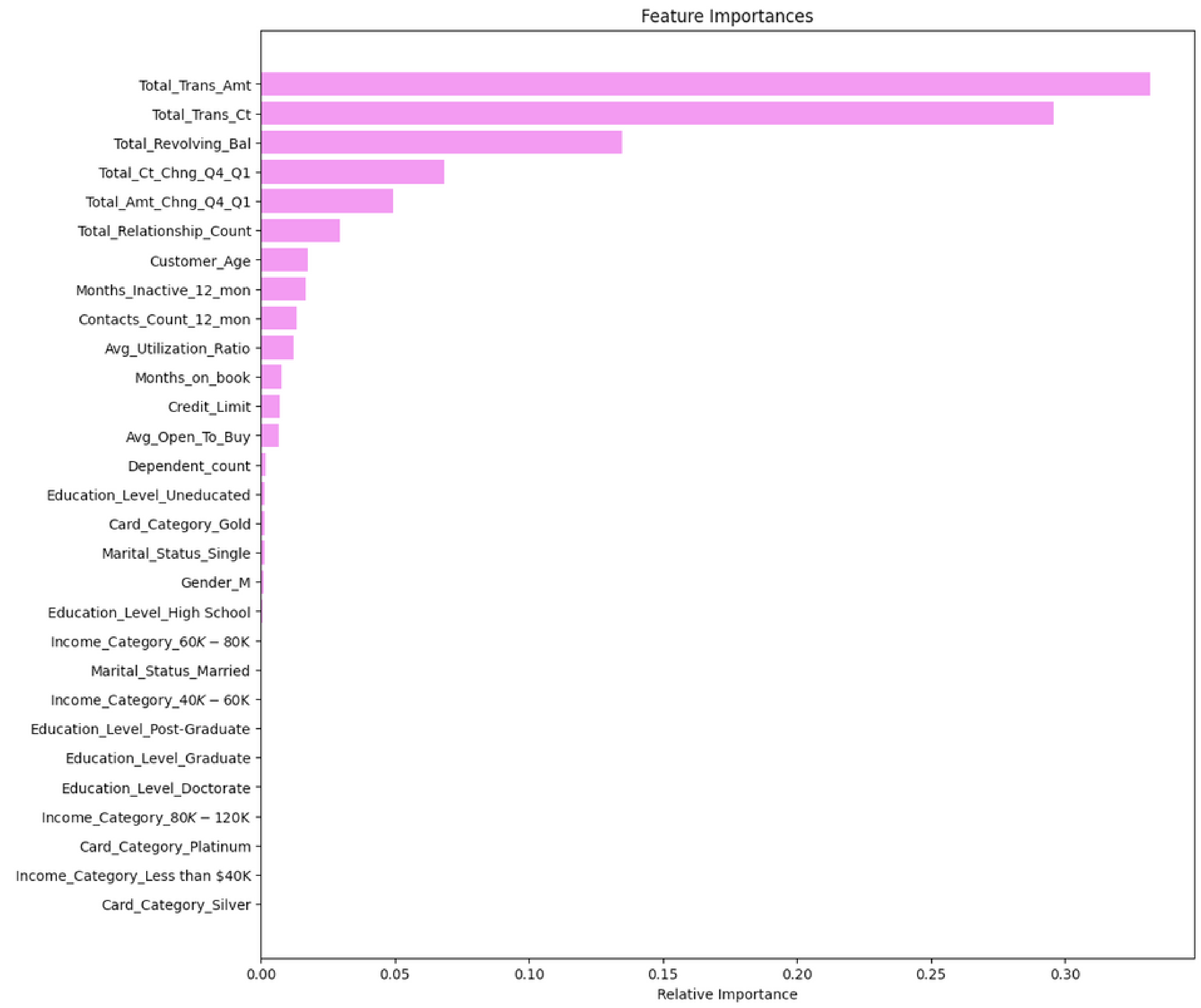
Prepared by Andrés Herrejón Maya.

Contents / Agenda

- **EXECUTIVE SUMMARY**
- **BUSINESS PROBLEM OVERVIEW AND SOLUTION APPROACH**
- **EDA RESULTS**
- **DATA PREPROCESSING**
- **MODEL PERFORMANCE SUMMARY FOR HYPERPARAMETER TUNING.**
- **APPENDIX**

Executive Summary

CHART FEATURE IMPORTANCE



After conducting an in-depth statistical analysis of the data and testing various models to predict customer attrition from the bank's services, the model that performed best in Recall Score was Gradient Boosting with undersampled data. This model produced a graph showing the importance of the dataset's features, with the total annual transaction amount and annual transaction count being the most important. These findings suggest that efforts to increase both the frequency and amounts of transactions are priorities for the bank.

The third most important factor is the total revolving balance of the card, which hinders customers from settling the card and leaving the service. This can be incentivized through interest-free months and various promotions, resulting in revolving balances without direct financial costs to the customer but retaining them within the customer base.

Executive Summary

CHART FEATURE IMPORTANCE

On the other hand, the seasonal use of the card or a decrease in both amount and transaction count throughout the year is also an essential variable to monitor. Strategies like gamification, where a streak of card usage unlocks benefits or rewards, can be engaging for some customers.

The number of products that customers have with the bank is also an essential feature. Even though it doesn't help distinguish between staying or leaving, it aids in better understanding the customers to sell them more products, thereby maintaining long-term loyalty. Products like savings accounts or investment accounts can foster greater loyalty.

Age is also a crucial characteristic, revealing that most of our customers are over 40. Various strategies can broaden the customer base to a younger market and develop products that create loyalty at an early age, as well as cater to existing clients with appropriate language and customer service design.

Prolonged activity by customers is vital for retention; regular card usage is crucial. Data also reveals that clients who left the bank had a higher number of contacts with it. Further investigation is needed to determine if these contacts resulted from unresolved issues or poor service, but it is an important feature. Developing applications or self-service options may guarantee fewer customer contacts, helping to retain them.

Another valuable insight is that credit limits may be a secondarily important variable since the average credit utilization has medium importance, but the total transaction amount is significant in the long term. However, risk must be kept low as the majority of customers have incomes under $40K.

Business Problem Overview and Solution Approach

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.

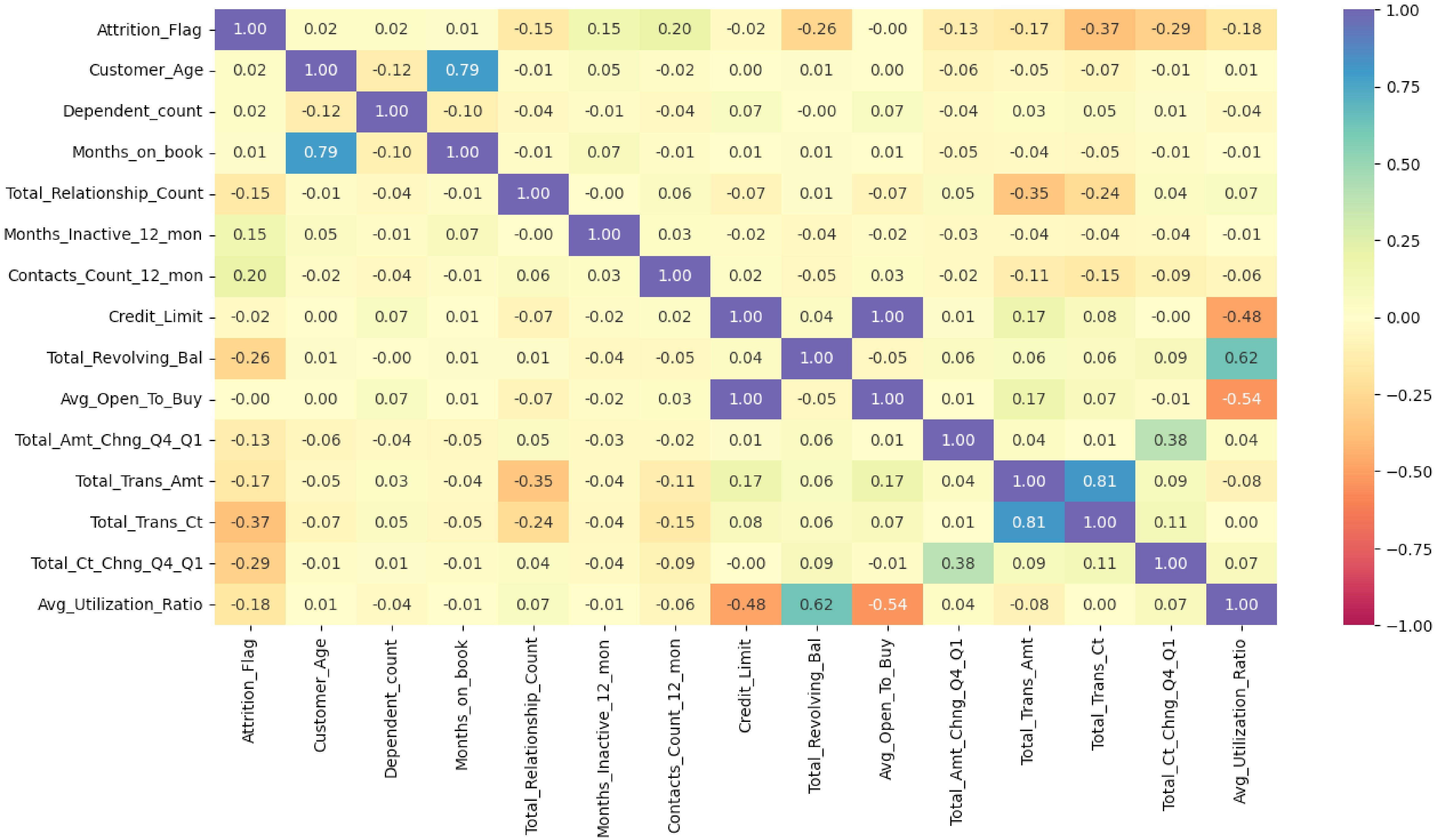
Customers' leaving credit cards services would lead bank to loss, so the bank wants to analyze the data of customers and identify the customers who will leave their credit card services and reason for same – so that bank could improve upon those areas

The solution approach is to come up with a classification model that will help the bank improve its services so that customers do not renounce their credit cards.

Review all important features (univariate analysis) in the data set and how they correlate (heatmap analysis) with other features and our target variable Attrition_Flag (bivariate analysis) insights and recommendations of how we can use this learnings for the Thera bank will be given in this color code format.

EDA Results

Heatmap Analysis



	Attrition_Flag	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
Attrition_Flag	1.00	0.02	0.02	0.01	-0.15	0.15	0.20	-0.02	-0.26	-0.00	-0.13	-0.17	-0.37	-0.29	-0.18
Customer_Age	0.02	1.00	-0.12	0.79	-0.01	0.05	-0.02	0.00	0.01	0.00	-0.06	-0.05	-0.07	-0.01	0.01
Dependent_count	0.02	-0.12	1.00	-0.10	-0.04	-0.01	-0.04	0.07	-0.00	0.07	-0.04	0.03	0.05	0.01	-0.04
Months_on_book	0.01	0.79	-0.10	1.00	-0.01	0.07	-0.01	0.01	0.01	0.01	-0.05	-0.04	-0.05	-0.01	-0.01
Total_Relationship_Count	-0.15	-0.01	-0.04	-0.01	1.00	-0.00	0.06	-0.07	0.01	-0.07	0.05	-0.35	-0.24	0.04	0.07
Months_Inactive_12_mon	0.15	0.05	-0.01	0.07	-0.00	1.00	0.03	-0.02	-0.04	-0.02	-0.03	-0.04	-0.04	-0.04	-0.01
Contacts_Count_12_mon	0.20	-0.02	-0.04	-0.01	0.06	0.03	1.00	0.02	-0.05	0.03	-0.02	-0.11	-0.15	-0.09	-0.06
Credit_Limit	-0.02	0.00	0.07	0.01	-0.07	-0.02	0.02	1.00	0.04	1.00	0.01	0.17	0.08	-0.00	-0.48
Total_Revolving_Bal	-0.26	0.01	-0.00	0.01	0.01	-0.04	-0.05	0.04	1.00	-0.05	0.06	0.06	0.06	0.09	0.62
Avg_Open_To_Buy	-0.00	0.00	0.07	0.01	-0.07	-0.02	0.03	1.00	-0.05	1.00	0.01	0.17	0.07	-0.01	-0.54
Total_Amt_Chng_Q4_Q1	-0.13	-0.06	-0.04	-0.05	0.05	-0.03	-0.02	0.01	0.06	0.01	1.00	0.04	0.01	0.38	0.04
Total_Trans_Amt	-0.17	-0.05	0.03	-0.04	-0.35	-0.04	-0.11	0.17	0.06	0.17	0.04	1.00	0.81	0.09	-0.08
Total_Trans_Ct	-0.37	-0.07	0.05	-0.05	-0.24	-0.04	-0.15	0.08	0.06	0.07	0.01	0.81	1.00	0.11	0.00
Total_Ct_Chng_Q4_Q1	-0.29	-0.01	0.01	-0.01	0.04	-0.04	-0.09	-0.00	0.09	-0.01	0.38	0.09	0.11	1.00	0.07
Avg_Utilization_Ratio	-0.18	0.01	-0.04	-0.01	0.07	-0.01	-0.06	-0.48	0.62	-0.54	0.04	-0.08	0.00	0.07	1.00

EDA Results

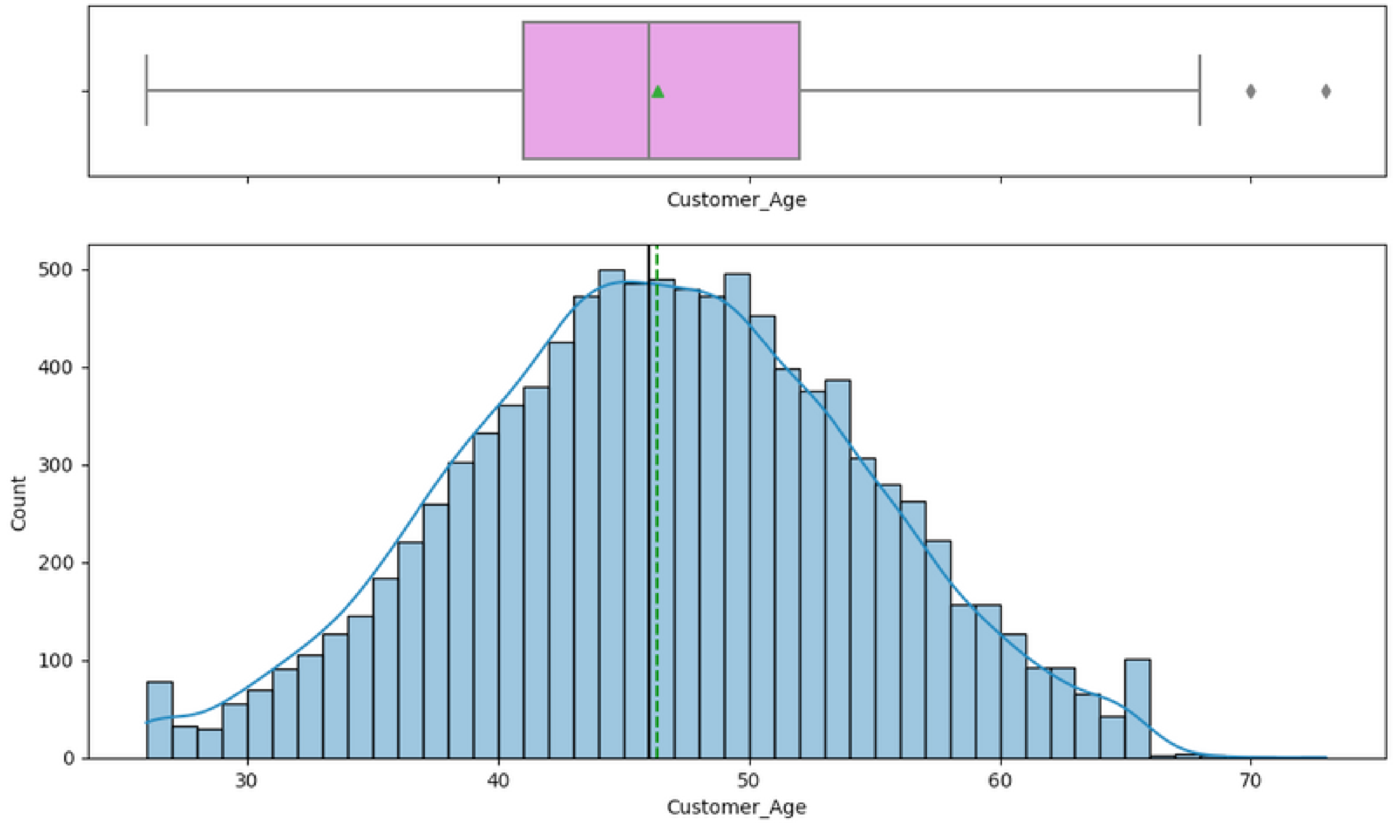
Heatmap Analysis

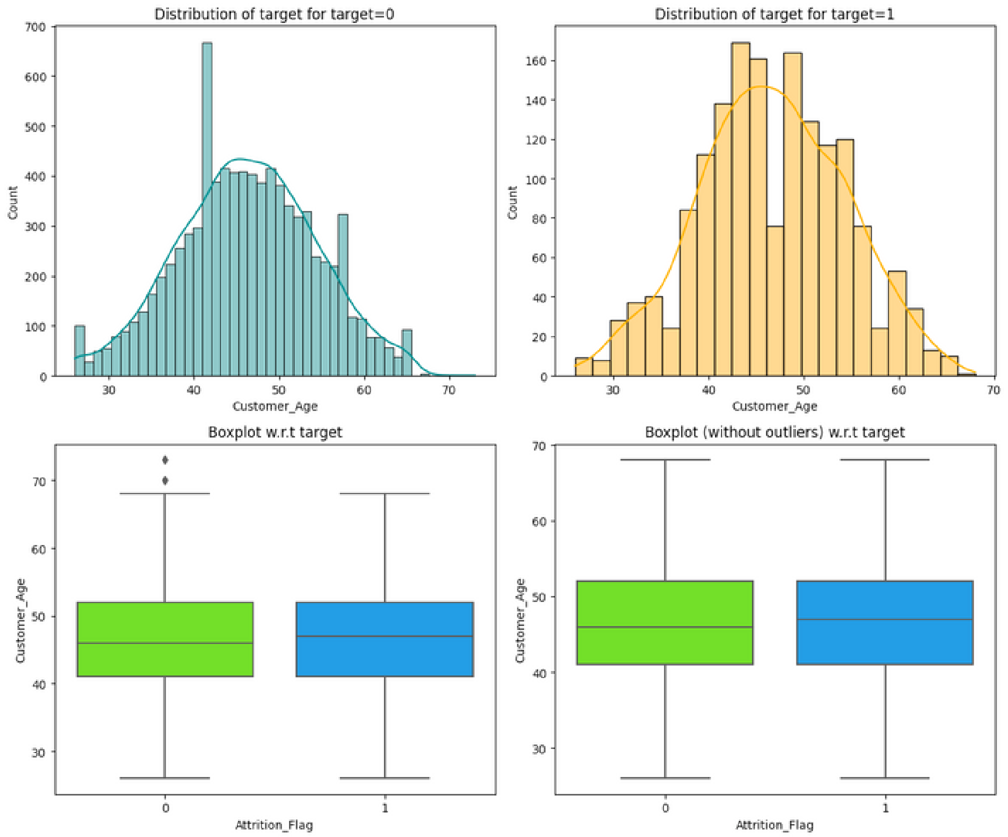
We can see important positive and negative correlations with our target variable Attrition_Flag:

- Total_Relationship_Count (-0.15): The less products the client has, the more prone to leave he is.
- Months_Inactive_12_mon (0.15): The more months inactive is more probable that the client leaves
- Contacts_Count_12_mon (0.20): The more contacts count the probability of abandonment increases as well.
- Total_Revolving_Bal (-0.26): Customers with higher revolving balances are less likely to attrite.
- Total_Amt_Chng_Q4_Q1 (-0.13): Customers who increase their transaction amounts over the year (from Q1 to Q4) are less likely to close their accounts or leave the bank's services.
- Total_Trans_Amt (-0.17): The less Total Transaction Amount is more probable that the client leaves
- Total_Trans_Ct (-0.37): The less Total Transaction Count the more probability of leaving the client has.
- Total_Ct_Chng_Q4_Q1 (-0.29): This shows that more than the amount, is the count that show more than double negative correlation with our target variable, so the customers who increase the number of transactions over the year (from Q1 to Q4) are less likely to close their accounts or leave the bank's services.
- Avg_Utilization_Ratio (-0.18): This shows that the more of the available credit the customer has spent the less likely it is for the customer to leave.

EDA Results

Age



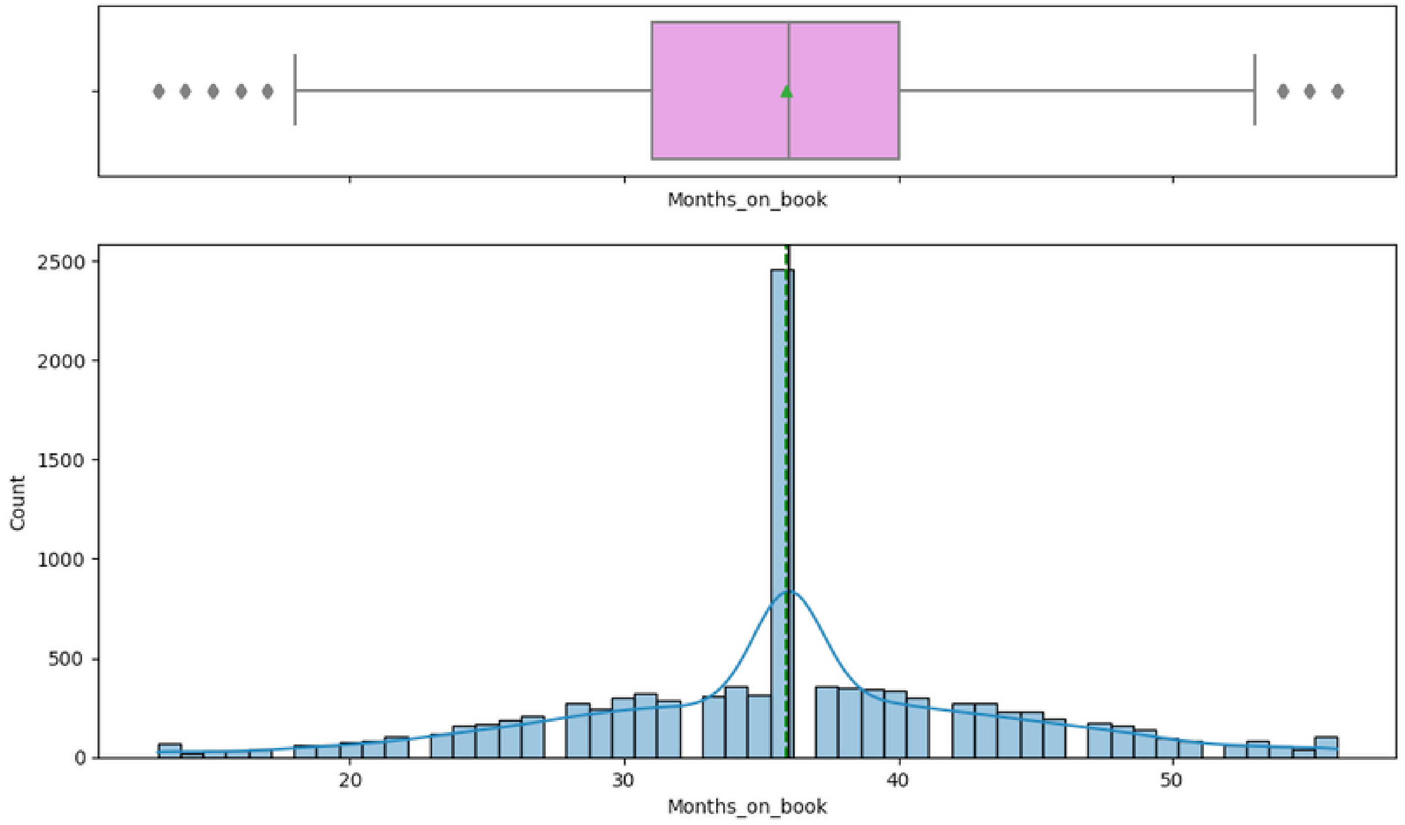


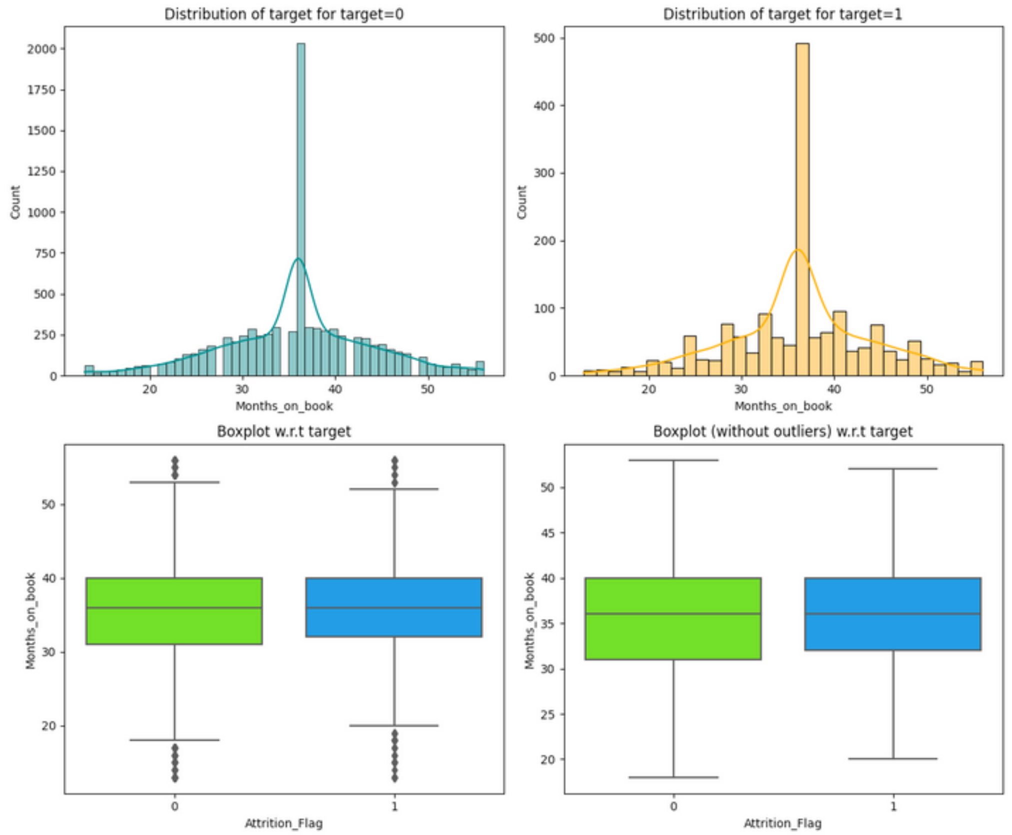
The average age of customers is around 46 years, with most customers ranging between 41 and 52 years. Normal distribution of ages observed on customers that kept and left their credit card services.
The only strong correlation found was with Months_on_book, which is logical that the older the customer the longer relationship with the bank has. Some negative correlation with Dependant_count is found as the older the customer, the less people depend on them.

Targeting strategies can be tailored to the prevalent age group (41-52) and tailoring products to different age groups can attract diverse customers.

EDA Results

Months_on_book





The average duration of a customer's relationship with the bank is approximately 3 years (35.9 months), with a range spanning from 31 to 40 months. There is a noticeable variance in this duration among different customers. The correlation analysis does not reveal any significant relationship between the number of months as a customer (Months_on_book) and the target variable (Attrition_Flag). The distribution of this feature is generally normal, but there is a noticeable peak that might indicate a large influx of customers at a particular time. This peak could be the result of a successful marketing campaign that attracted many customers simultaneously, and this pattern is observed across both attrition flags.

EDA Results

Months_on_book

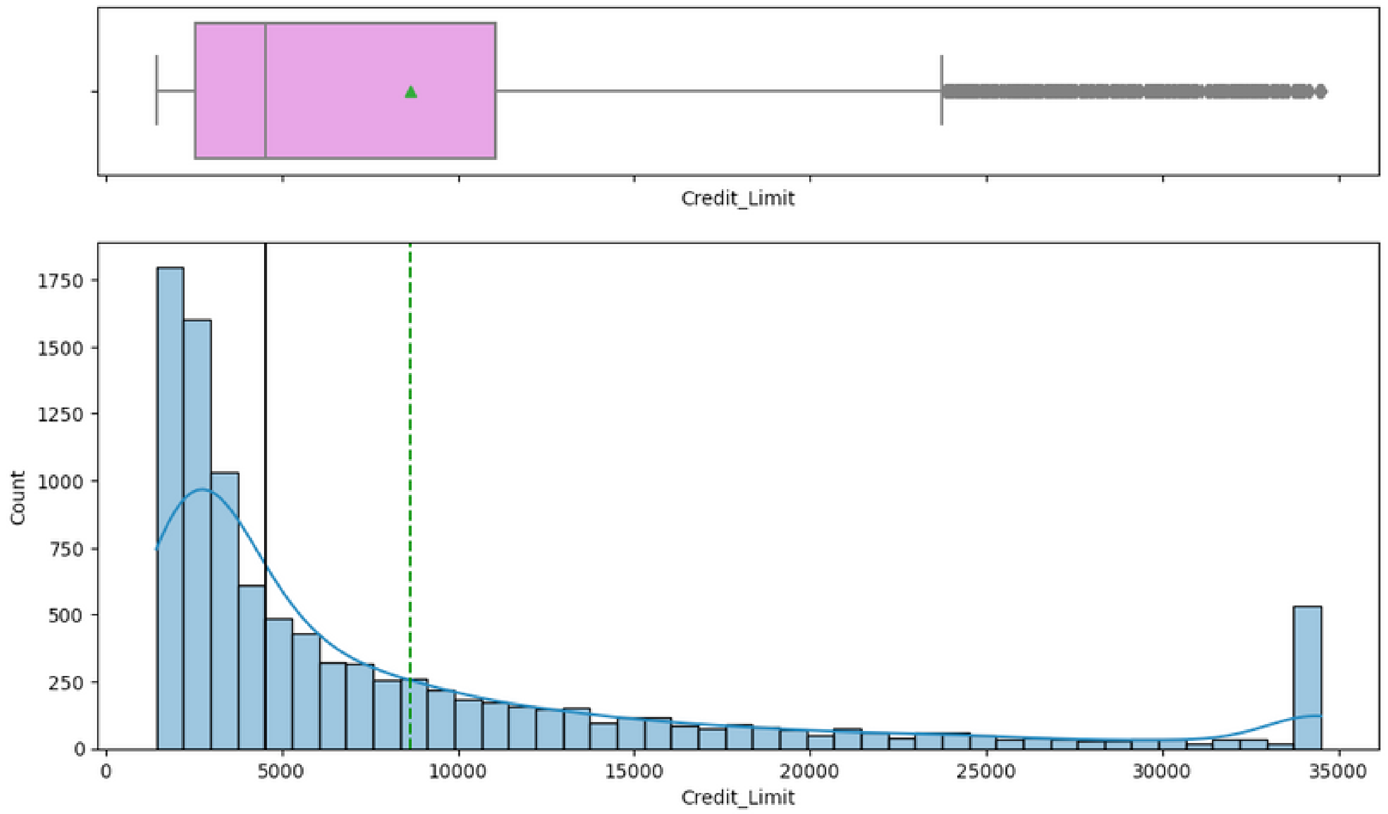
The duration of a customer's relationship with the bank, even if it extends to 3 years, does not appear to be a determining factor in retaining them. Merely having a presence in the customer's wallet for an extended period may not be sufficient to ensure their loyalty or prevent them from leaving.

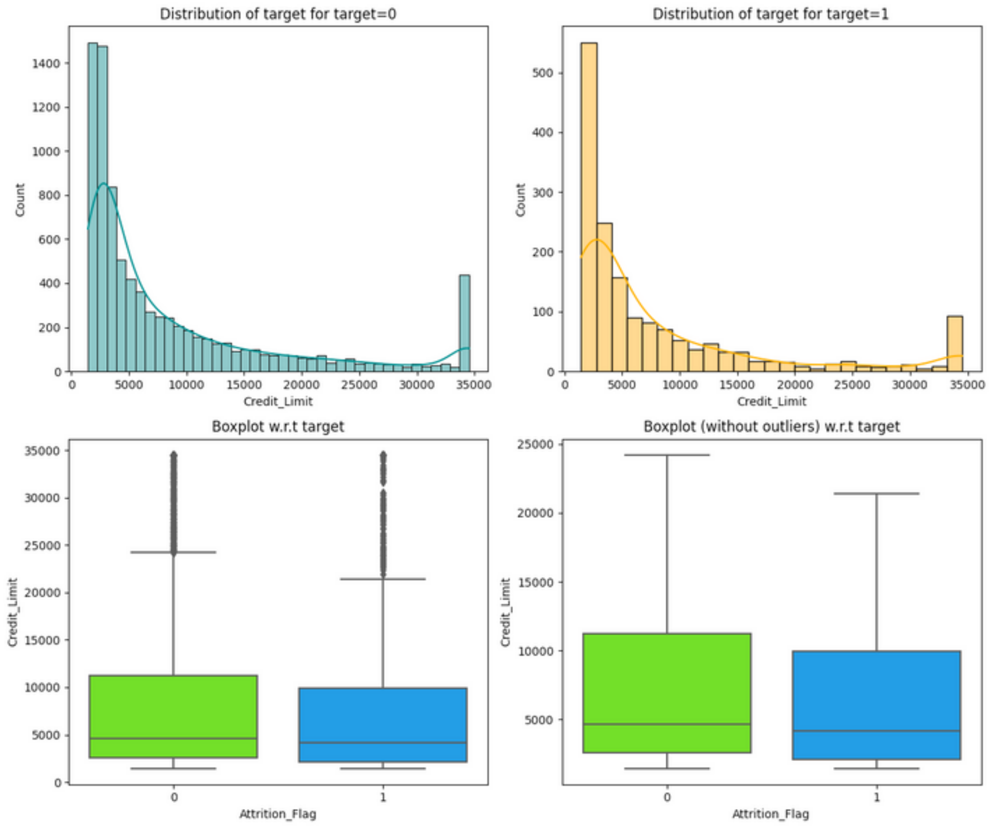
The bank need to incentivize the constant usage of the products.

Getting insights like income category can give insights on what kind of products could match. Benefits on basic necessity products for lower income market could be attractive.

EDA Results

Credit_Limit





There is a pronounced right-skewed distribution in the data, with 50% of customers having a credit limit below $4,550. This is significantly lower than the average credit limit of $8,631, explaining the high standard deviation and the presence of outliers.

EDA Results

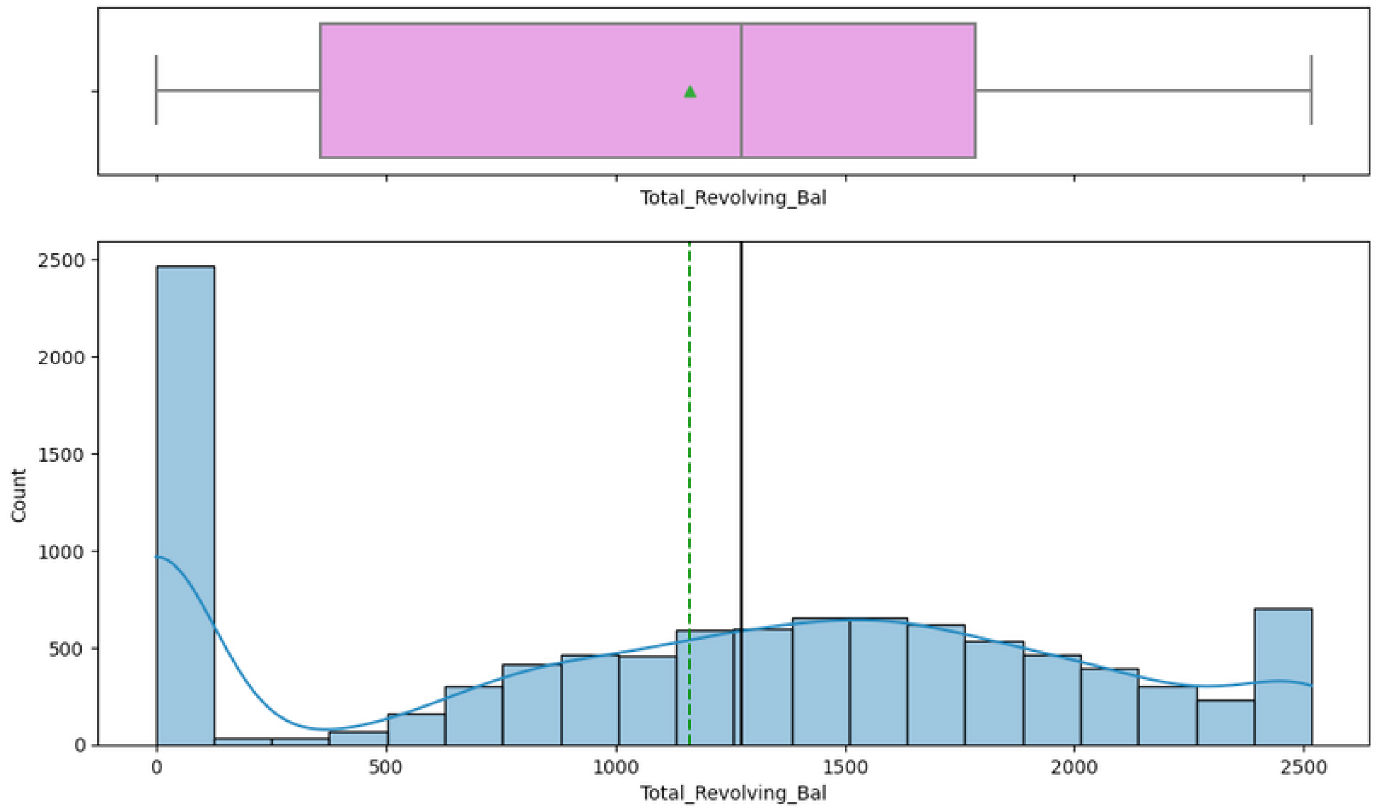
Credit_Limit

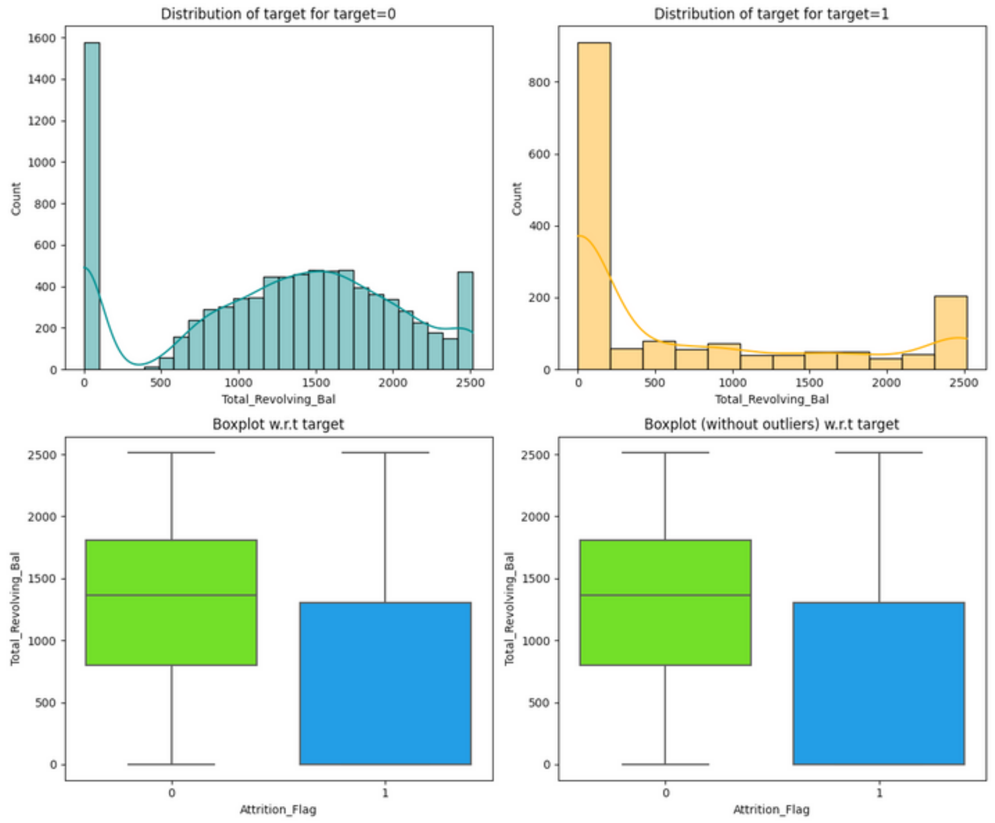
The negative correlation between Credit_Limit and Avg_Utilization_Ratio suggests that customers are not fully utilizing their credit cards, and this is not due to a lack of available credit limit or a low Avg_Open_To_Buy. Logically, Credit_Limit and Avg_Open_To_Buy show a perfect positive correlation. There is a slight positive correlation with Total_Transaction_Amt, indicating that a higher credit limit correlates with a higher transaction amount, but not necessarily with a higher total transaction count (Total_Transaction_Count).
There is no significant correlation between Credit_Limit and the target variable Attrition_Flag.

While higher credit limits are associated with higher individual transaction amounts, they are not translating into more frequent usage or a lower attrition rate. The bank must consider alternative incentives that focus on encouraging regular use of the credit card, rather than simply increasing spending limits. This approach aligns with the goal of increasing customer engagement and loyalty

EDA Results

Total_Revolving_Bal





On average, customers carry a revolving balance of $1162.81. However, the distribution is notably right-skewed, with the majority of customers having lower revolving balances and only a few having higher balances. The KDE (Kernel Density Estimation) line in the distribution plot reveals two distinct peaks, indicating potential clusters or groupings within the data. Additionally, 25% of customers have a revolving balance below $359. The data also shows considerable variability, with a standard deviation of $814.99.

EDA Results

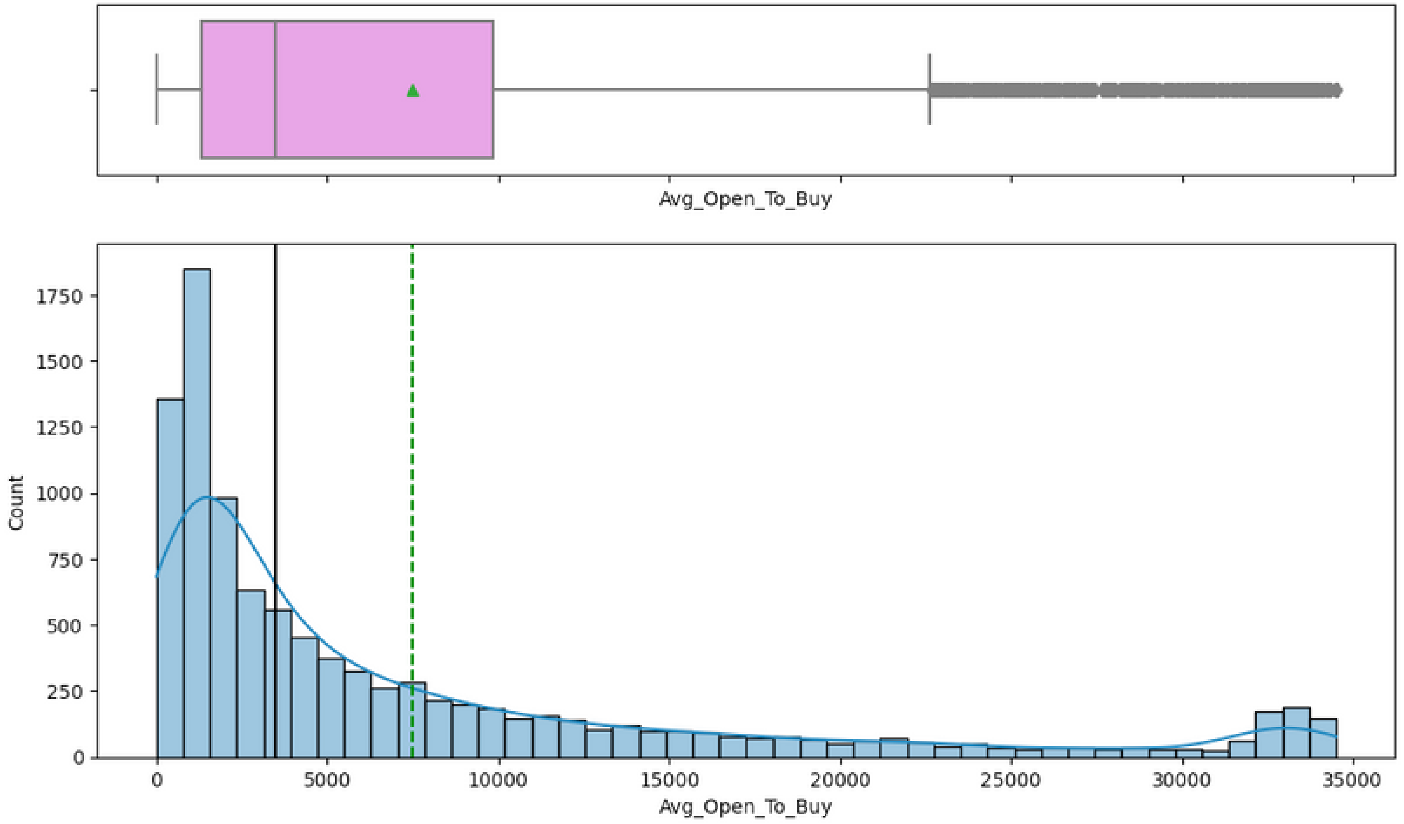
Total_Revolving_Bal

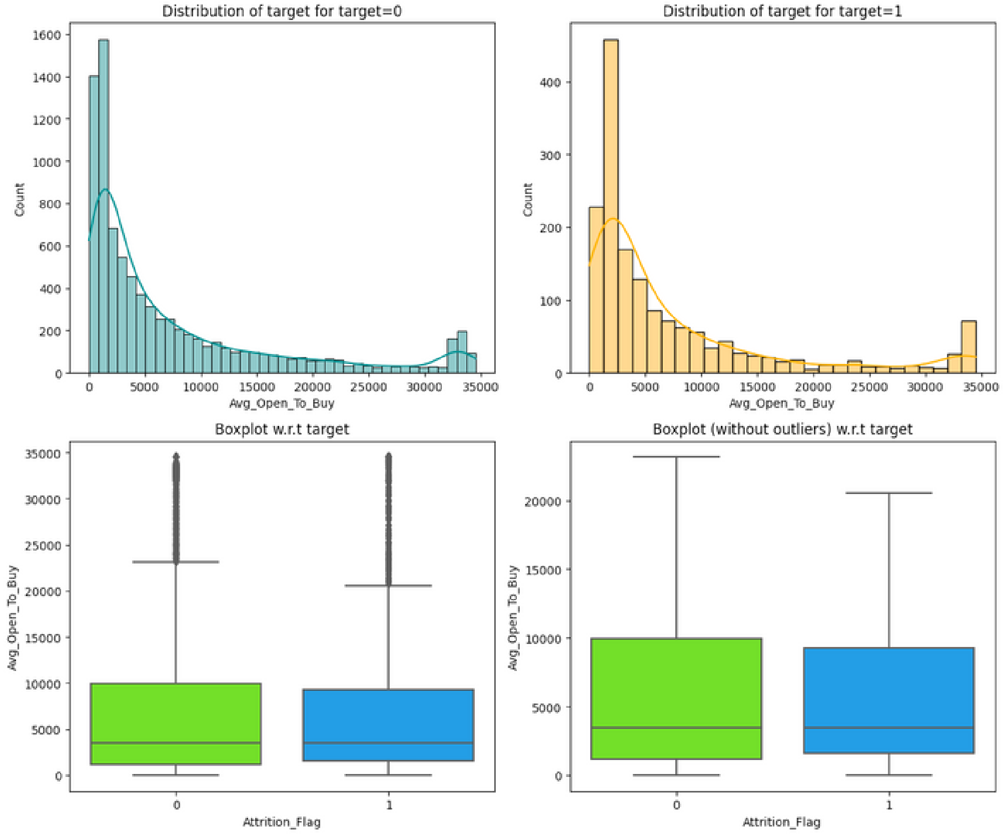
There is a negative correlation of -0.26 with our target variable, signifying that customers with higher revolving balances are less likely to leave the bank's services (attrite). This correlation might be explained by the fact that a higher balance would be more challenging to pay off, and therefore, customers are less likely to close their accounts.
In the bivariate analysis, we can observe that it's primarily the customers with lower balances who tend to leave the bank's services, as evidenced by the right skewness in the distribution. On the other hand, customers who do not attrite maintain the possible clusters of customers as indicated by the KDE line.

The insights reflect prudent risk management by the bank, as customers appear to pay off their credit card balances month-to-month. A balance transfer strategy from other banks, offering better interest rates to customers with good risk management, could be an appealing approach. Additionally, the bank might incentivize usage through interest-free installment plans. This could make carrying a balance more attractive even to customers who typically manage their credit well. Such strategies could align with the bank's goals of increasing customer engagement and maximizing the value provided by their credit products.

EDA Results

Avg_Open_To_Buy





Most customers exhibit a low 'Average Open to Buy' value, primarily because the majority of them have a low credit limit, as mentioned earlier. Since there is an overlap in the distributions of 'Average Open to Buy,' where both instances of attrition (existing and attrited customers) show a right-skewed pattern, this feature alone may not be a strong differentiator between the two groups.

The heatmap analysis also logically indicates that the lower the 'Open to Buy,' the higher the 'Average Utilization Ratio.' This is because customers have used more of their credit limit, reducing the available credit for further purchases. In other words, as the 'Open to Buy' decreases, the proportion of the credit limit that has been utilized increases, reflecting higher credit engagement.

EDA Results

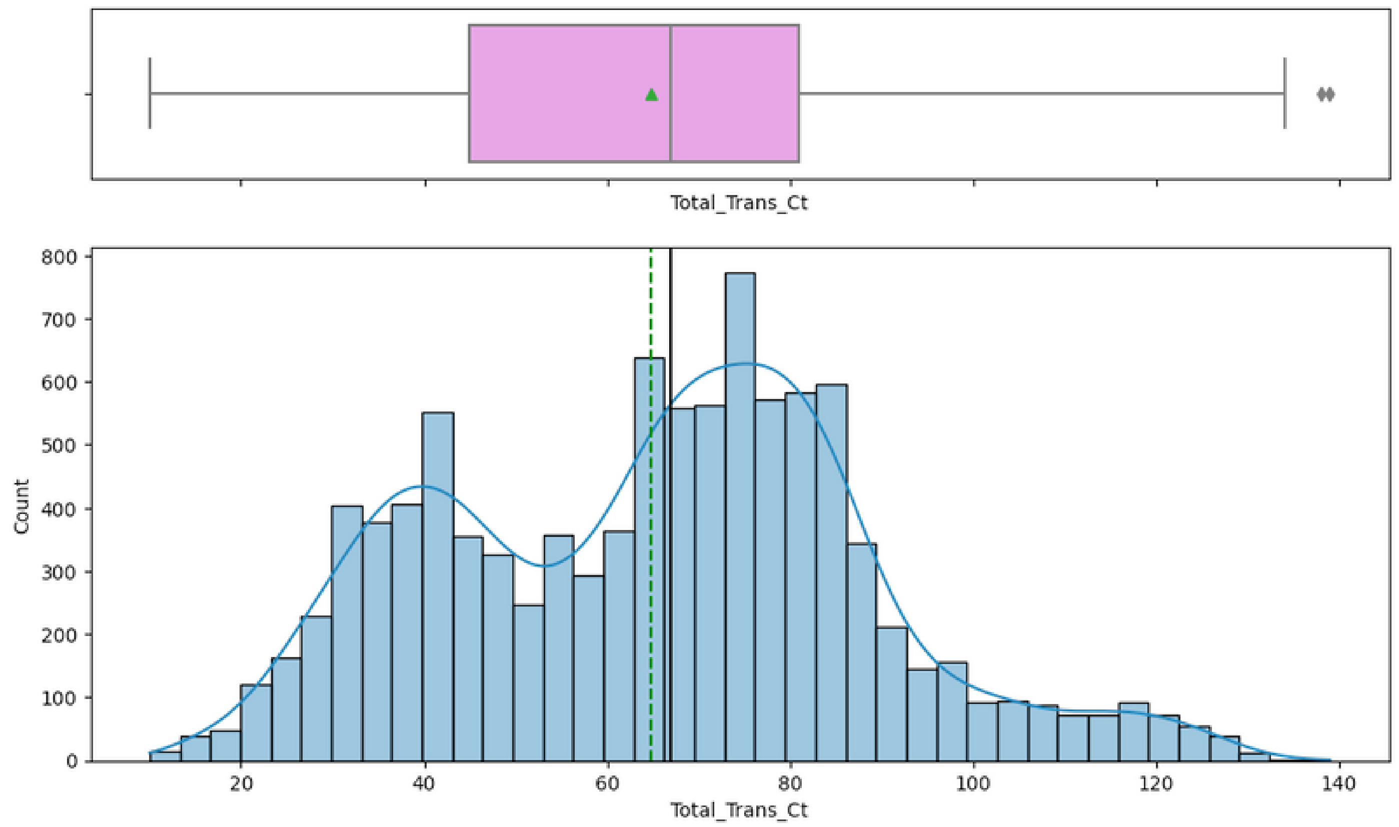
Avg_Open_To_Buy

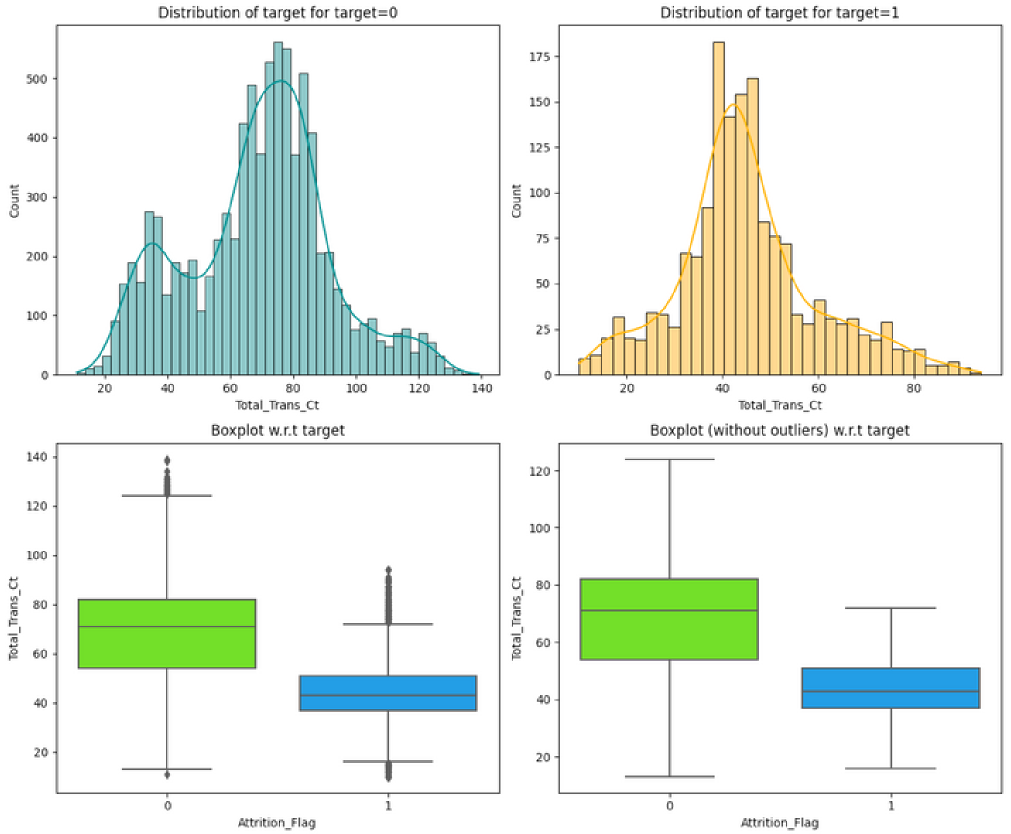
This information gives insights on the kind of products that we can market to extend the number of products that the customer has from the bank.

For example, saving accounts or investments could give benefits that increase the amount of Credit_Limit and give some space to increase the Avg_Open_To_Buy without adding risk to the practices.

EDA Results

Total_Trans_Ct





This feature shows the highest negative correlation of -0.37, indicating that constant use of the card to some extent helps retain the customer. On average, customers use the card approximately 65 times over 12 months. The standard deviation is 23.4, with 25% of customers making fewer than 45 annual transactions, 50% making fewer than 67 transactions, and 75% making fewer than 81 transactions. The observed transactions range from a minimum of 10 to a maximum of 139. Outliers are present in both groups of customers (current and attrited), but there are many more outliers among those who decided to cease being customers. This constant use can be interpreted as atypical behavior for a customer deciding to leave the credit card service.

EDA Results

Total_Trans_Ct

Based on the previous insights, it is recommended to encourage constant use of the card, which may eventually lead to a significant debt amount and possibly a balance carried over to subsequent billing periods. These incentives can be offered in the form of points and discounts where people make purchases. Even if the amounts are not large, keeping the card in constant use justifies the fees that the bank charges, regardless of whether the card is used or not, such as annual fees, etc.

The bank could incentivize the card usage by:

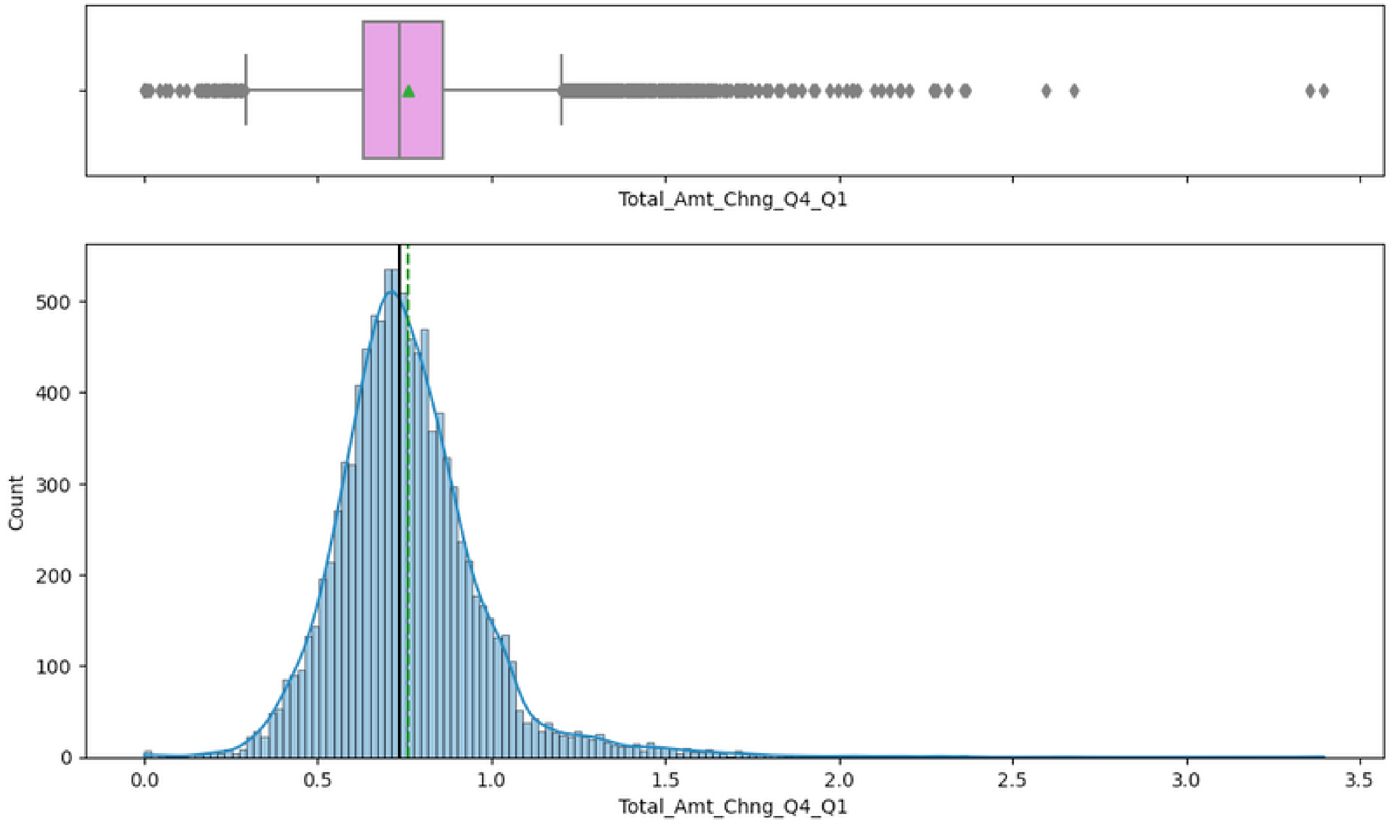
Implementing a loyalty program that rewards customers for frequent use of the card, even for small purchases.

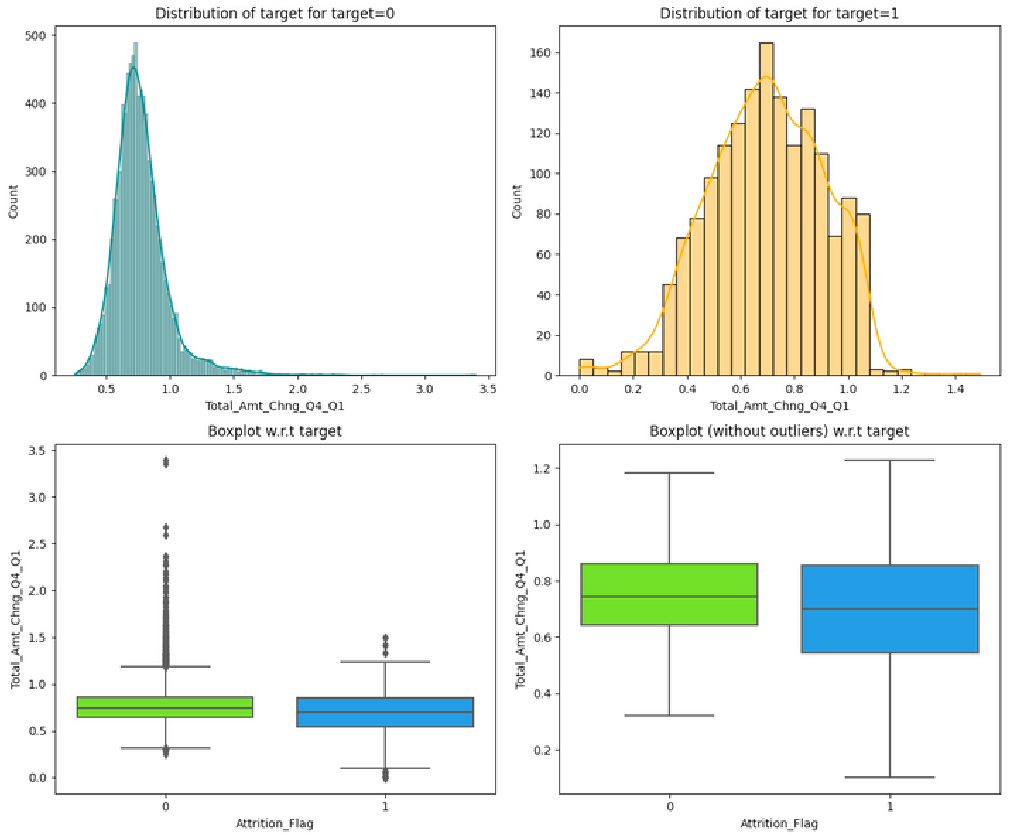
Creating targeted offers and discounts for specific customer segments to encourage more transactions.

Educating customers about existing and maybe develop new benefits of regular card usage and communicating the value-added services that come with the card.

EDA Results

Total_Amt_Chng_Q4_Q1





This feature represents whether the customer has increased the amount of credit card usage from the first quarter to the last quarter of the year. The statistical analysis shows that, on average, customers decreased the total amount on their card by 24%. The standard deviation of this amount change is 0.21, with a minimum of zero (indicating no change) and a maximum of 3.3 (indicating a tripling of the amount and being a clear outlier). The distribution shows that 25% of customers reduced their amount by approximately 37%, 50% by approximately 26%, and 75% by approximately 14%. Further analysis could determine whether this is seasonal or a trend.

EDA Results

Total_Amt_Chng_Q4_Q1

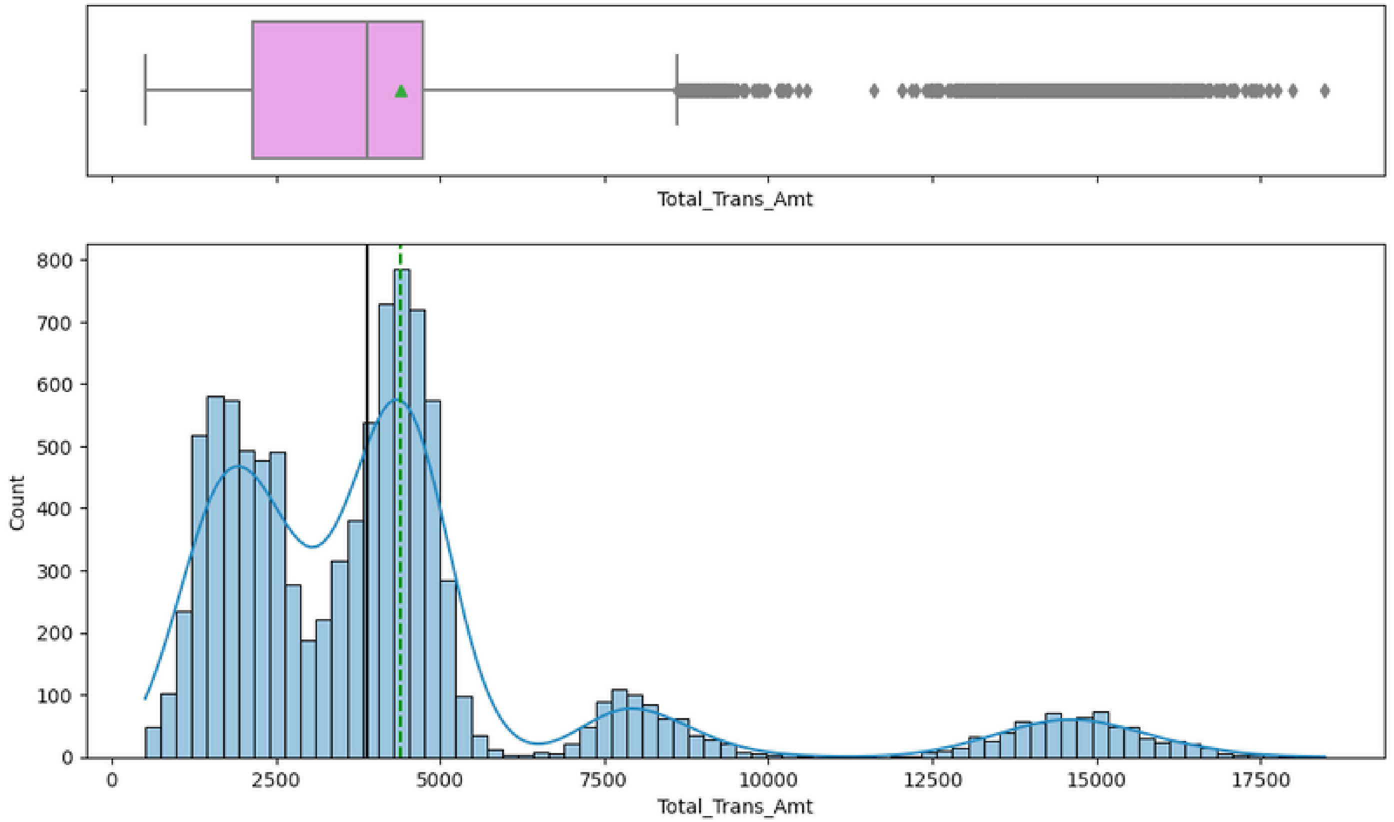
Generally, customers decrease their purchase amounts as the year progresses, which has a correlation of -0.13 with service abandonment.
The plot shows a normal distribution with the presence of some outliers on both the small amount side and the high amount side. Most customers maintained a decrease in their credit card amount in Q4 compared to Q1.
Since both customers who decided to leave the credit card services and those who decided to stay show a decrease, this feature alone may not be sufficient to determine whether a customer will remain or leave the bank.

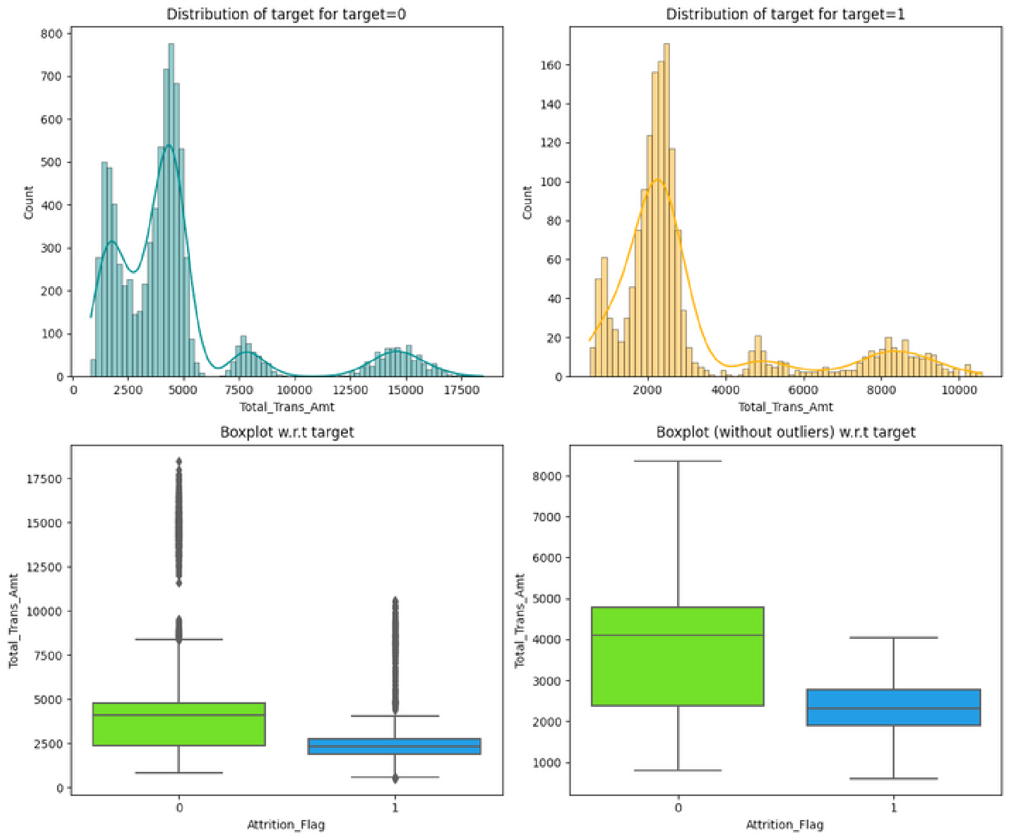
The bank should encourage consistent use of the credit card throughout the 12 months, promoting both a cumulative amount and a higher transaction count.

If the decrease in spending is found to be seasonal, consider introducing seasonal promotions or incentives to maintain or increase spending throughout the year.

EDA Results

Total_Trans_Amt





The "Total_Trans_Amt" represents the total annual transaction amount. On average, the amount is $4404, with a standard deviation of $3397, indicating a high variation. The minimum observed is $510, while the maximum is $18,484. 25% of customers accumulate less than $2155, 50% at least $3899, and 75% less than $4742 annually. The graph shows right skewness, and the KDE line reveals four clusters: the first around $2500, the second and largest around $4800, another around $7800, and the last around $15,000.

This feature may have limited ability to discern whether the customer will leave or remain with the bank, as it shows a similar distribution in both groups.

EDA Results

Total_Trans_Amt

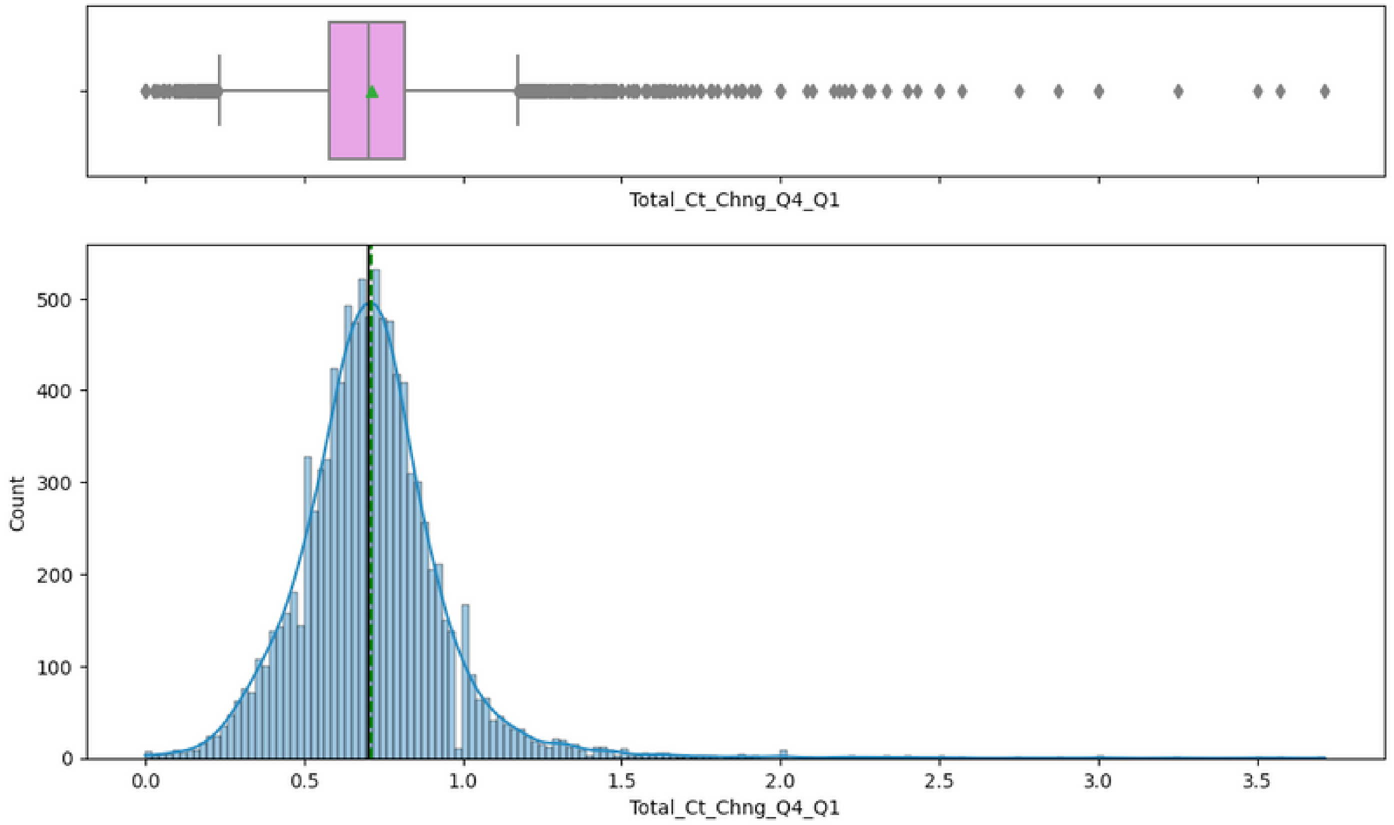
The graph indicates that higher amounts start to become outliers earlier among customers who decide to leave the bank (approximately from $5000 onwards for annual total debt).
A negative correlation with the Attrition variable indicates that lower amounts increase the likelihood of the customer leaving the bank. A high negative correlation with Total_Relationship_Count suggests that the more products a customer has with the bank, the lower the amount they will have on their card. Positive correlations exist with Credit_Limit, logically, and with Avg_Open_To_Buy. There's also a strong positive correlation with the total transaction count.

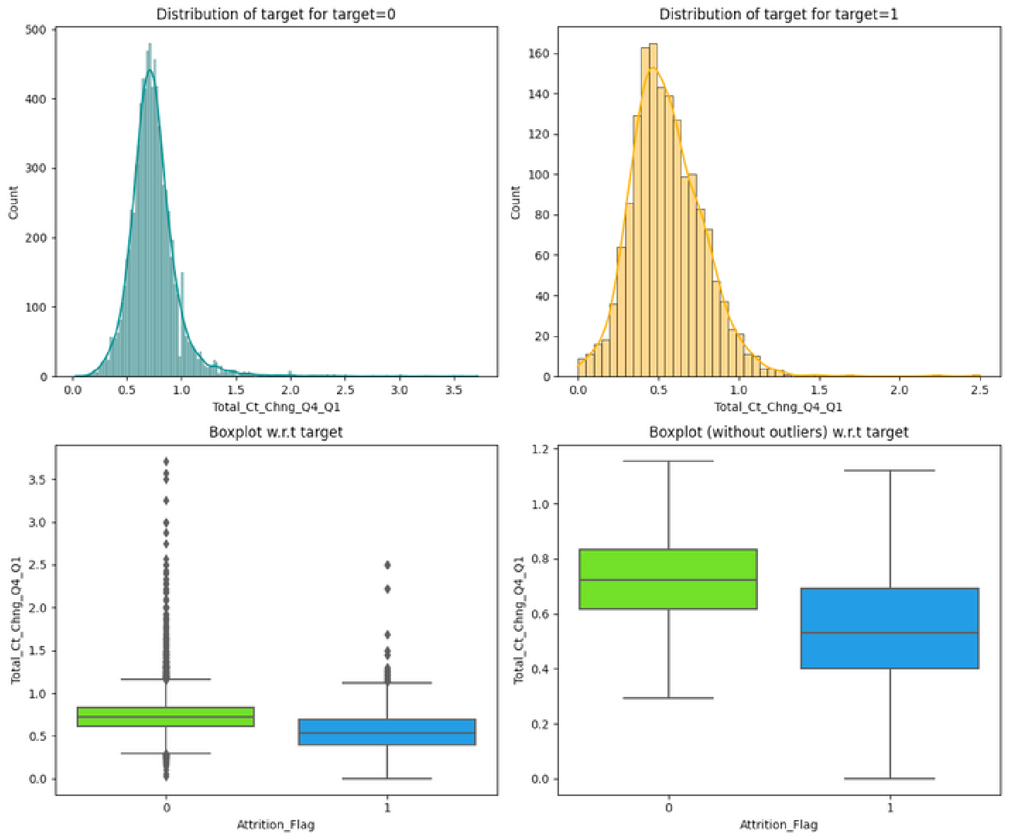
Consider customer segmentation based on the four identified clusters. Develop targeted offers and incentives that align with different spending levels and behaviors. Then, introduce loyalty and reward programs that incentivize spending within specific ranges, encouraging customers to move from one cluster to another by increasing spending.

Avoid over-selling credit card products as they seem to accumulate less balance and is easier to drop the card with a small amount by distributing the debt between many cards.

EDA Results

Total_Ct_Chng_Q4_Q1





The feature represents the change in the transaction count from the first quarter to the fourth quarter. All numbers less than one indicate a decrease in the transaction count throughout the year, while numbers above one indicate an increase. On average, there is a 30% decrease, with a standard deviation of 0.23. The minimum observed is no change, and the maximum shows more than a threefold change in the count. 25% of customers have decreased by less than 42%, 50% by approximately 30%, and 75% by an estimated 19%.

A vast number of outliers are present showing that clients are most likely to decrease the card usage throughout the year and the distribution show that most clients that dropped the service decreased the count.

EDA Results

Total_Ct_Chng_Q4_Q1

In terms of correlations:

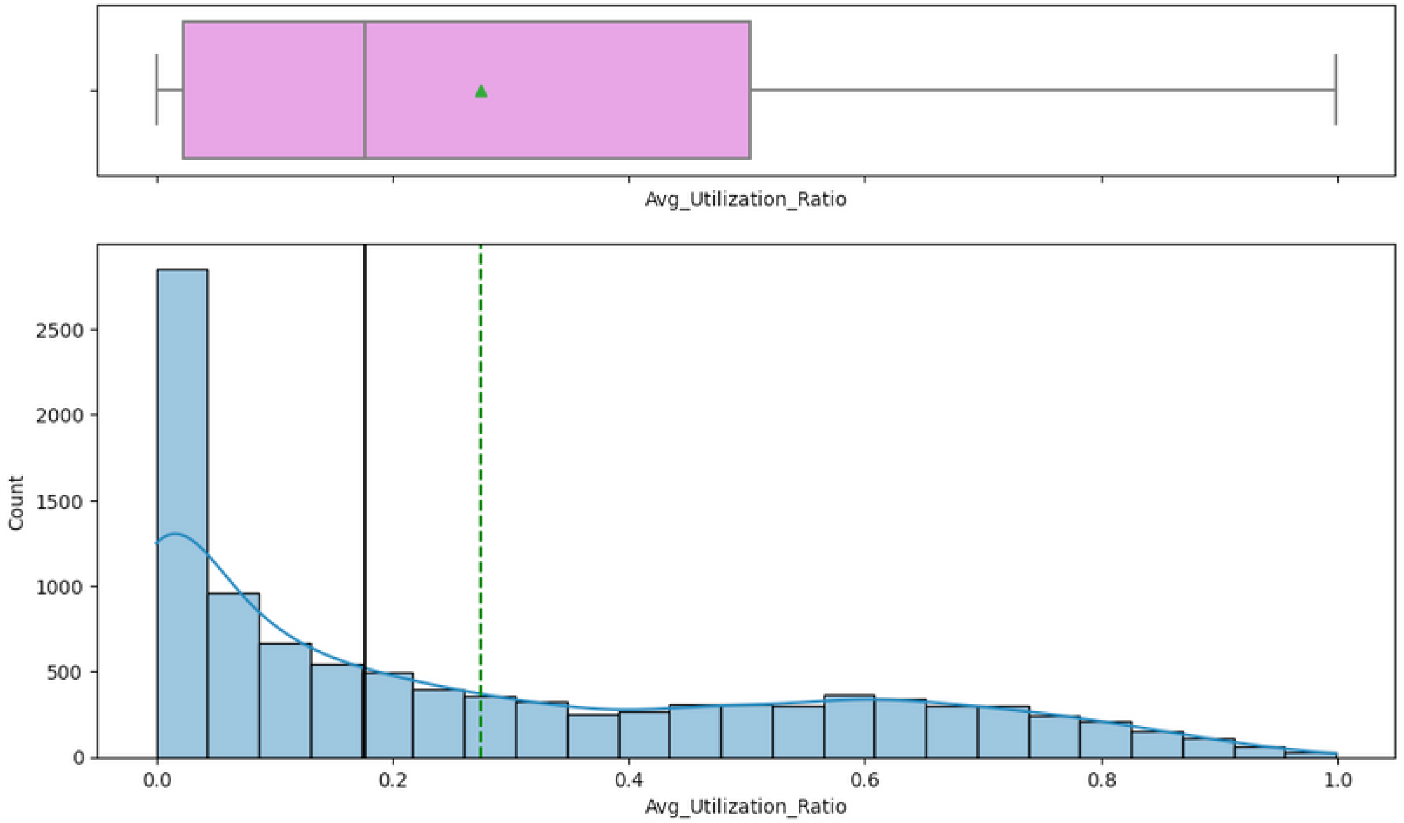
A negative correlation of -0.29 with the Attrition_Flag feature indicates that if the count decreases throughout the year, the likelihood of the customer leaving the services increases.
A positive correlation of 0.38 with Total_Amt_Chng_Q4_Q1 logically signifies that if the transaction count increases, it is likely that the amount will also increase throughout the year.
Another positive correlation with the total number of transactions is also quite logical.

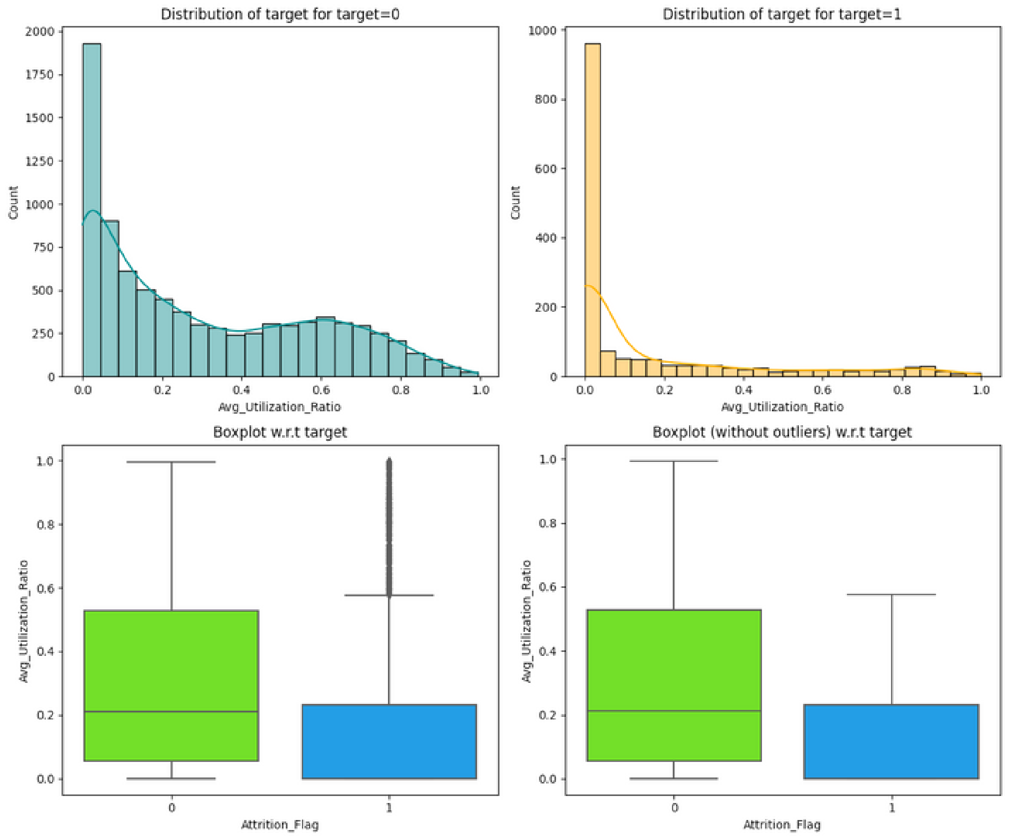
Encourage customers to use their credit cards throughout the year, even for small purchases, to avoid a decrease in the number of transactions over the year. This can be achieved through targeted promotions, loyalty points, or cashback offers that reward transaction frequency.

Analyze the seasonality of the transaction count and introduce offers that align with customer spending patterns in different quarters. This can help maintain a steady transaction count throughout the year.

EDA Results

Avg_Utilization_Ratio





This feature represents the proportion of available credit that the customer has used. On average, customers have used 27.5% of their available credit. The minimum utilization is 0%, and the maximum is nearly 100%, meaning they have used almost their entire credit line. Specifically:

- 25% of customers have used only 2.3%.
- 50% have used less than 17.6%.
- 75% have used less than 50% of their available credit line.

EDA Results

Avg_Utilization_Ratio

The distribution is skewed to the right, indicating that most people use a low portion of their available credit.

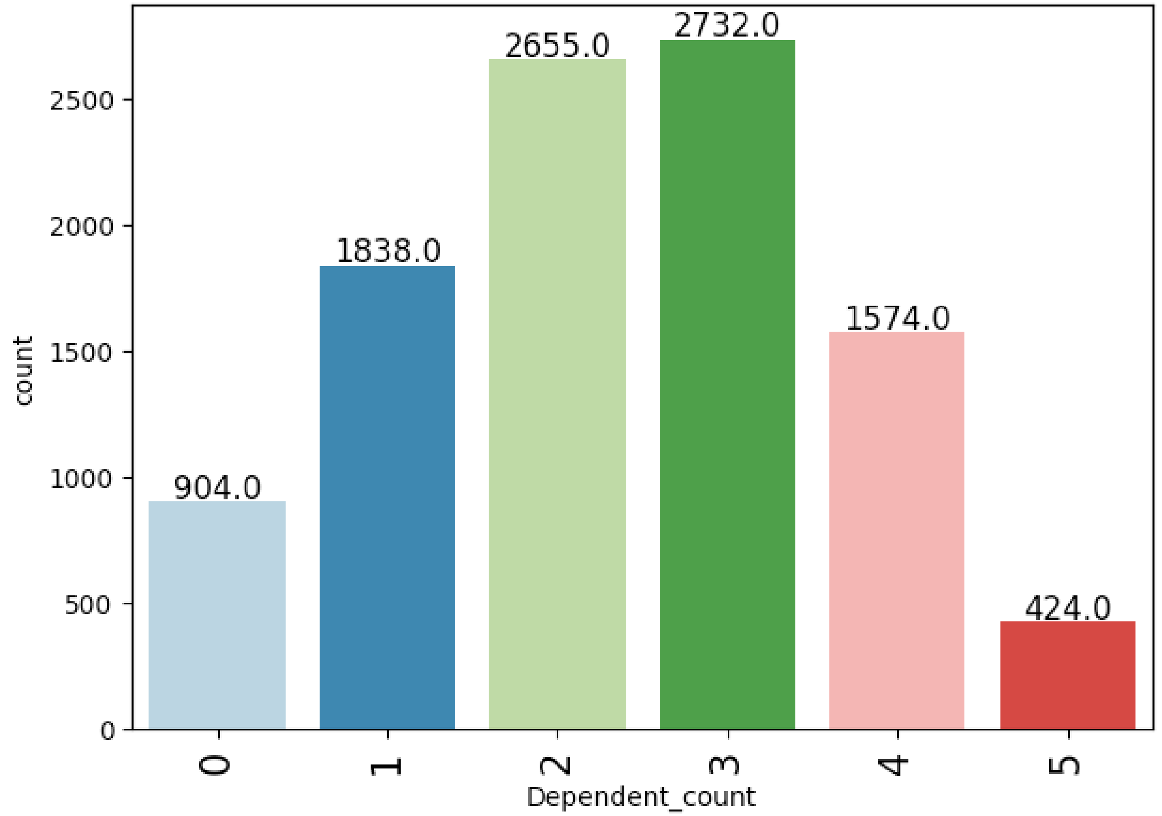
In terms of correlations:
A negative correlation of -0.18 with Attrition_Flag suggests that lower utilization of the credit line is associated with a higher likelihood of service abandonment.
A negative correlation of -0.48 with Credit_Limit is logical since a high credit limit represents a low portion of utilization.
A positive correlation of 0.62 with Total_Revolving_Bal also makes sense as higher utilization of the credit line leads to a higher revolving balance.
A negative correlation of -0.54 with Avg_Open_To_Buy is also clear, as the greater the utilization portion, the less credit is available for purchase.

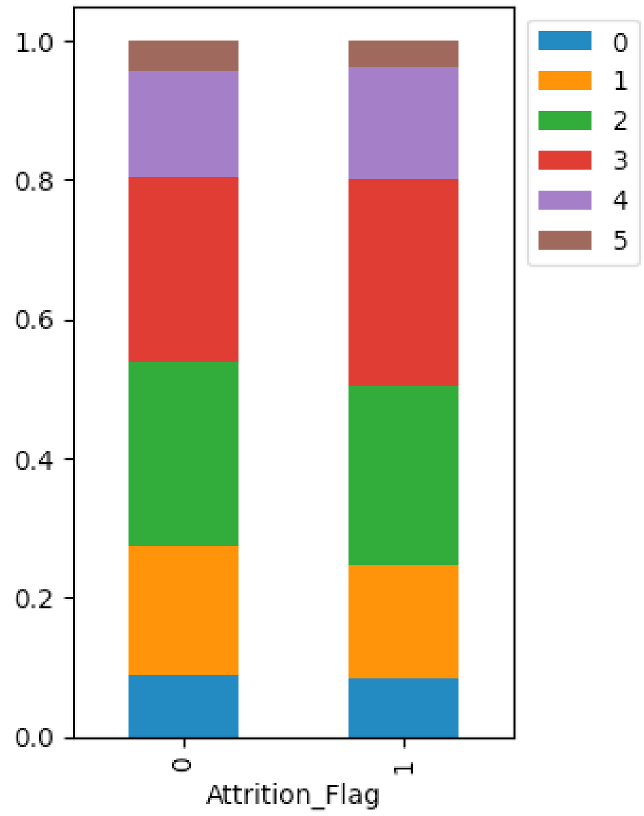
The graph shows a much lower credit utilization among customers who abandon the service compared to those who remain as customers.

For customers using a very low percentage of their credit, consider targeted offers that encourage responsible spending. This could include cashback offers, rewards, or specialized interest rates for specific spending categories.

EDA Results

Dependent_count





The feature "Dependent_count" represents the number of dependents that a customer has. On average, customers have 2.3 dependents, with a standard deviation of 1.3, a minimum of zero dependents, and a maximum of five dependents observed. Looking at the distribution, 25% of customers have one or fewer dependents, 50% of customers have two or fewer dependents, and 75% of customers have three or fewer dependents.

The graph reveals the largest group of customers to have three dependents, followed by two dependents, then one, four, zero, and lastly, five dependents.

EDA Results

Dependent_count

Very similar proportions are shown between customers who remain and customers who leave the credit card service regarding the number of dependents. Therefore, this feature is unlikely to serve as a distinguishing factor between customers who are going to abandon and customers who are going to remain.

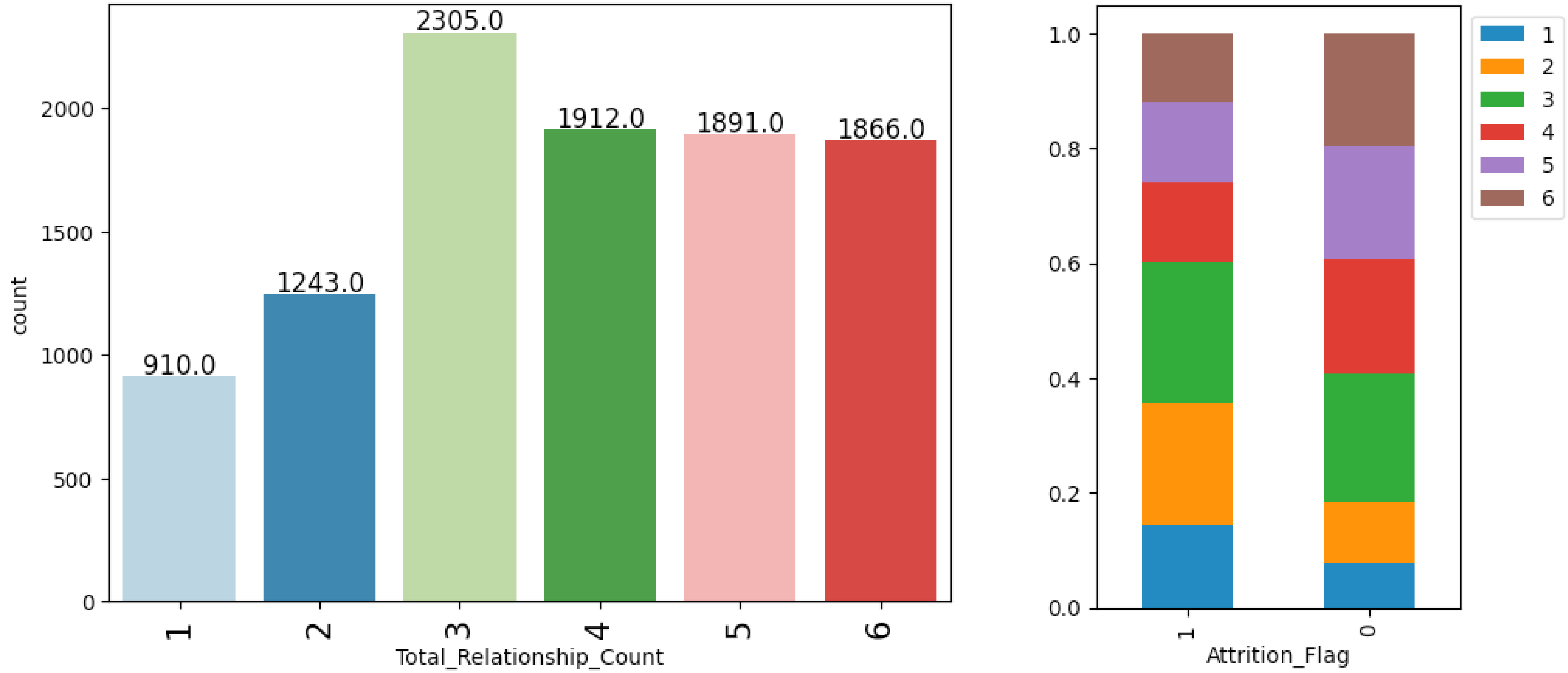
While not directly linked to attrition, the "Dependent_count" feature can still provide insights into the customer's lifestyle and needs, which can be leveraged to enhance customer satisfaction and loyalty.

Families with more dependents may have different spending patterns or credit needs compared to those with fewer or no dependents. Understanding these needs can help in tailoring specific products, offers, or services that resonate with different customer segments.

The bank might consider offering family-oriented benefits or rewards that cater to customers with multiple dependents. This could include discounts on family dining, entertainment, or travel packages.

EDA Results

Total_Relationship_Count



The feature "Total_Relationship_Count" represents the number of products that a customer has with the bank. On average, customers have 3.8 products, with a standard deviation of 1.5, a minimum of one, and a maximum of six products. In terms of distribution, 25% of customers have three or fewer products, 50% have four or fewer products, and 75% have five or fewer products.

The plot is left skewed showing that most clients have more than 2 products.

EDA Results

Total_Relationship_Count

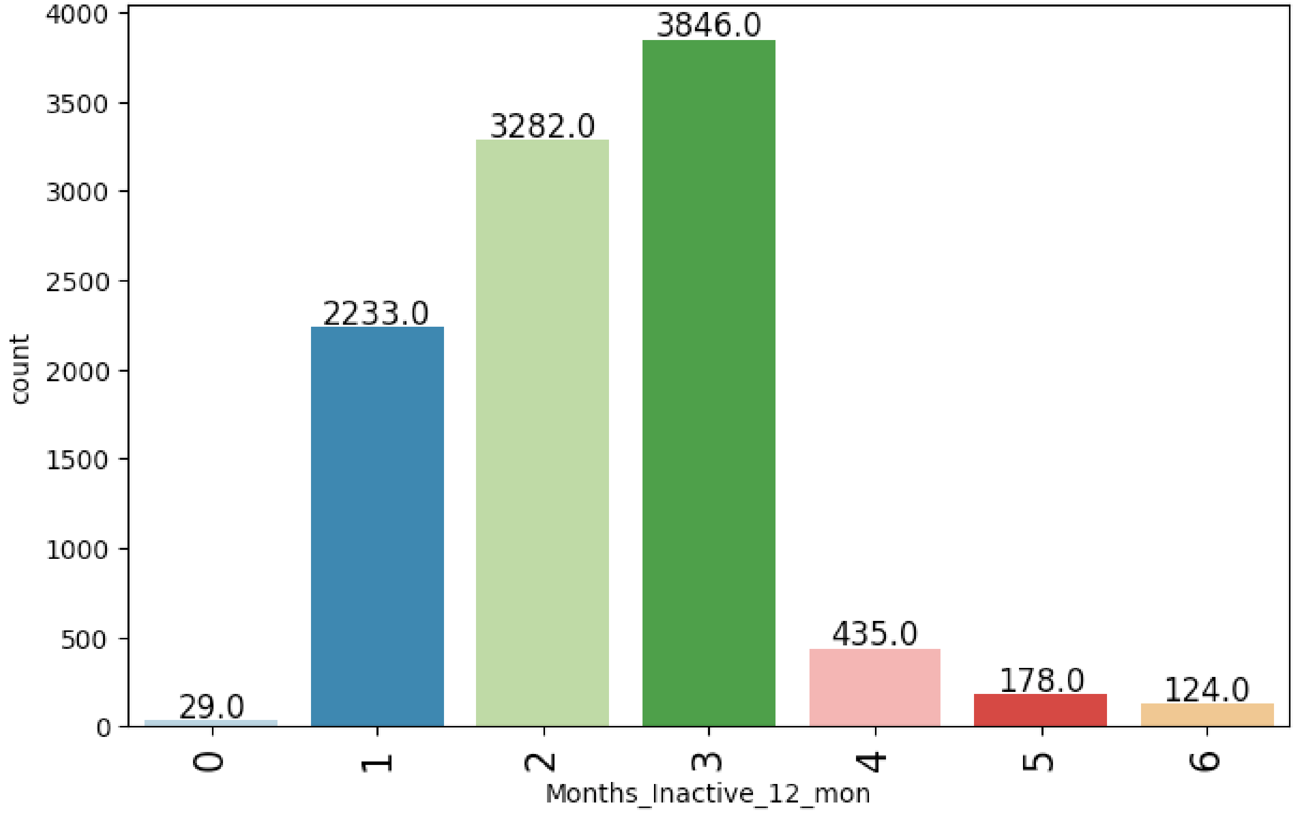
The "Total_Relationship_Count" has a slight negative correlation of -0.15 with the Attrition_Flag, indicating that to some extent, the fewer products a customer has, the higher the probability of abandonment. It also exhibits a negative correlation of -0.35 with the Total_Trans_Amt, meaning that the fewer cards a customer has, the higher the amount they spend on each card. Similarly, it correlates negatively at -0.24 with Total_Trans_Ct, suggesting that the fewer cards a customer has, the higher the transaction count.

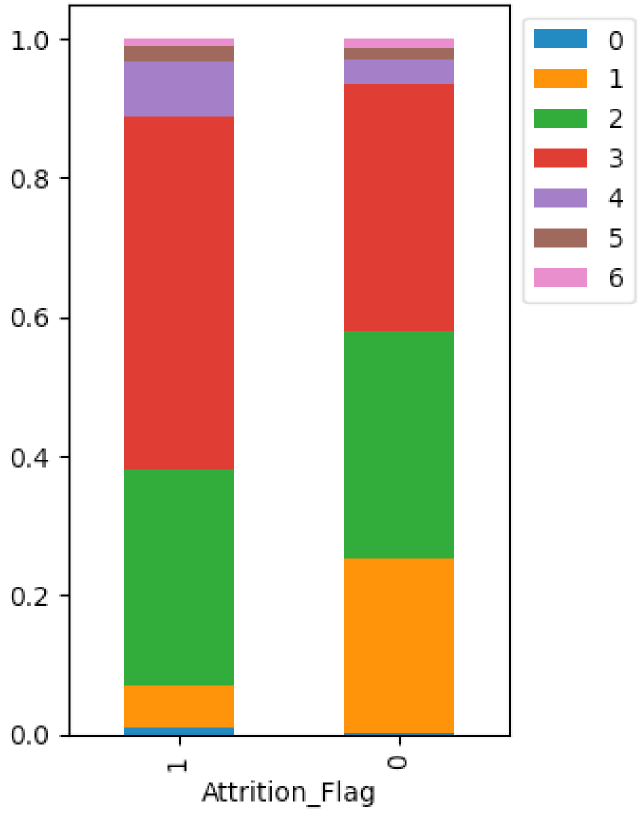
The bank should work on understanding why customers with fewer products are more likely to leave. Is it due to dissatisfaction with the existing products, lack of awareness of other offerings, or some other reason? This insight can guide targeted interventions to increase retention.

The negative correlation between the number of products and both transaction amounts and transaction counts suggests an opportunity for cross-selling and up-selling. By offering tailored products that align with a customer's lifestyle, age, number of dependents, and financial needs, the bank might encourage customers to increase their engagement across multiple products.

EDA Results

Months_Inactive_12_mon





The feature "Months_Inactive_12_mon" represents the number of months a customer's card was inactive in the last 12 months. The data shows that on average, customers leave their card inactive for around 2.3 months. There's a standard deviation of one month, with a minimum inactivity period of zero months and a maximum of six months. Half of the customers leave their card inactive for two months or less, and 75% of customers leave their card inactive for three months or less.

EDA Results

Months_Inactive_12_mon

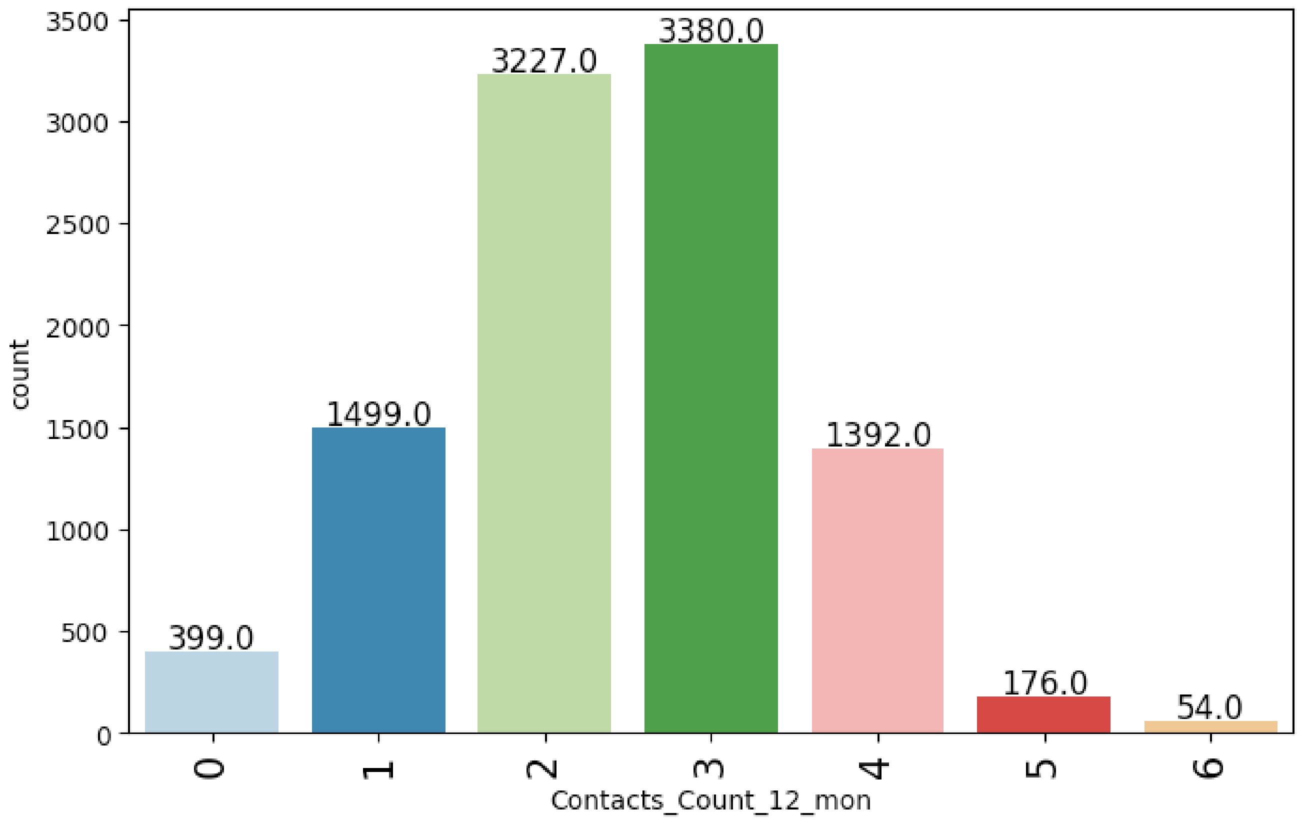
The distribution reveals that the vast majority of customers leave their card inactive for a maximum of three months, followed by a maximum of two months, and finally one month of inactivity. Notably, the segment of customers that abandoned the bank demonstrates a higher number of inactive months compared to those who remained with the bank.
There is a positive correlation of 0.15 with Attrition_Flag, indicating that the greater the number of inactive months, the higher the likelihood of account abandonment.

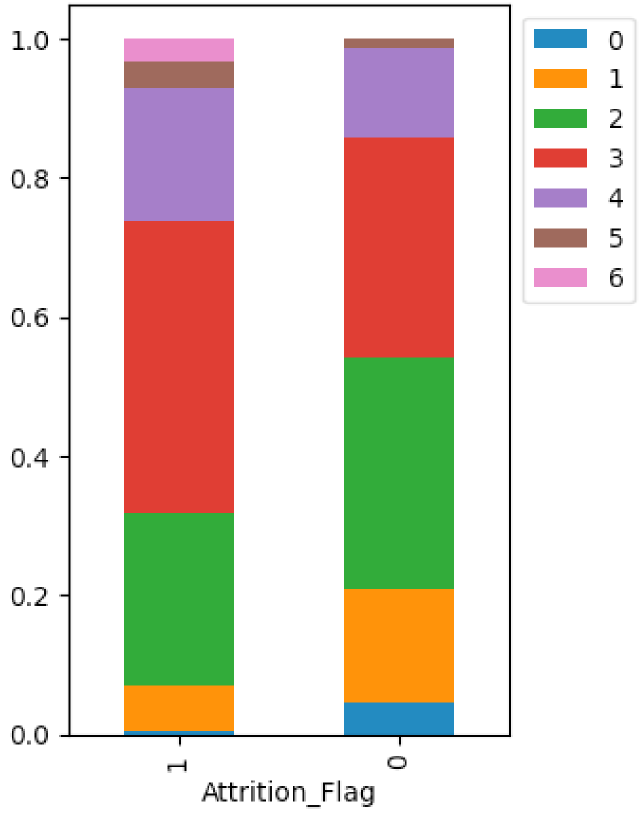
The bank should explore why customers are leaving their cards inactive.

The bank could consider implementing a gamification strategy where customers are encouraged to maintain a streak of card usage to win certain benefits. These could be financial incentives or partnerships with affiliated businesses. This approach could make card usage more engaging and rewarding, thereby reducing inactivity.

EDA Results

Contacts_Count_12_mon





The feature "Contacts_Count_12_mon" indicates the number of times a customer has been in contact with the bank over the past 12 months. On average, there have been 2.4 contacts, with a standard deviation of 1.1 contacts. The minimum number of contacts is zero, and the maximum is six. Half of the customers had two or fewer contacts, and 75% of the customers had three or fewer contacts.

EDA Results

Contacts_Count_12_mon

There is a positive correlation of 0.2 with Attrition_Flag, indicating that the more contacts a customer has had with the bank, the greater the likelihood that they will discontinue their services. There is also a negative correlation with transaction amounts and transaction counts, indicating that the more contacts a customer has, the lower the amounts and transactions they carry out.
The graphs show that customers who left the bank generally had a higher number of contacts.

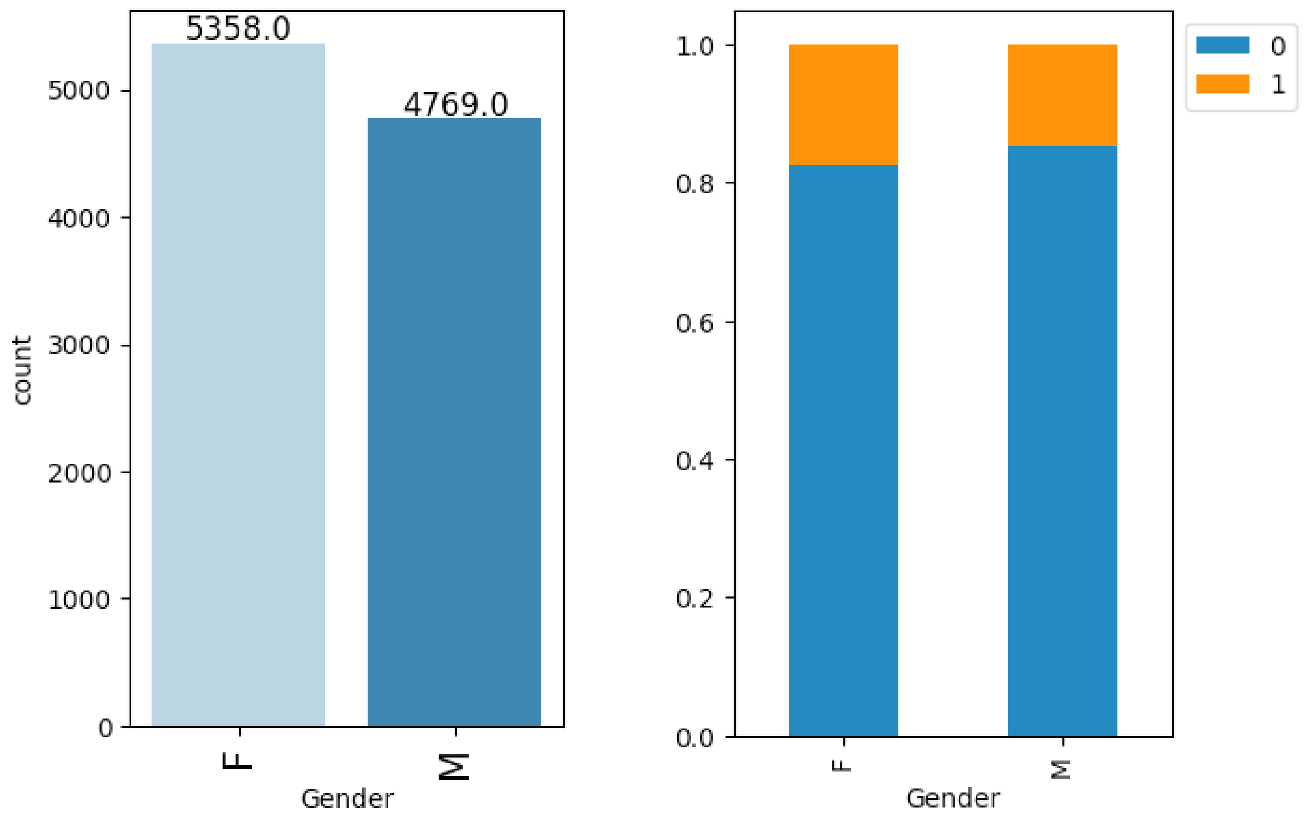
The bank should investigate the reasons for customer contacts. Are they reaching out due to issues, inquiries, complaints, or other reasons? Understanding the nature of these contacts can guide more effective customer support.

The correlation with Attrition_Flag suggests that more contacts might indicate dissatisfaction or unresolved issues. The bank should conduct surveys or direct interviews to gauge customer satisfaction and pinpoint specific areas that may need improvement. Whether it's slow response times, unmet needs, or inadequate resolutions, the bank needs to assess and enhance its customer support processes.

Leveraging chatbots, FAQs, or an App or self-service portals might reduce the need for direct contacts. By providing customers with quick and easy access to information and solutions, the bank can enhance convenience and reduce reliance on traditional customer support.

EDA Results

Gender

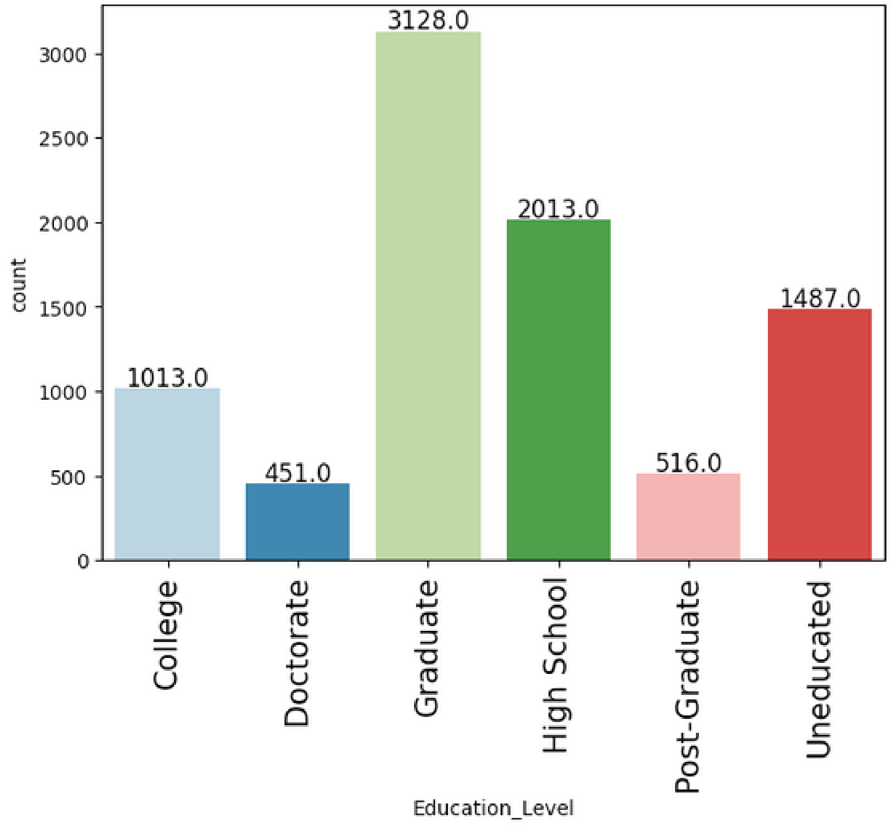


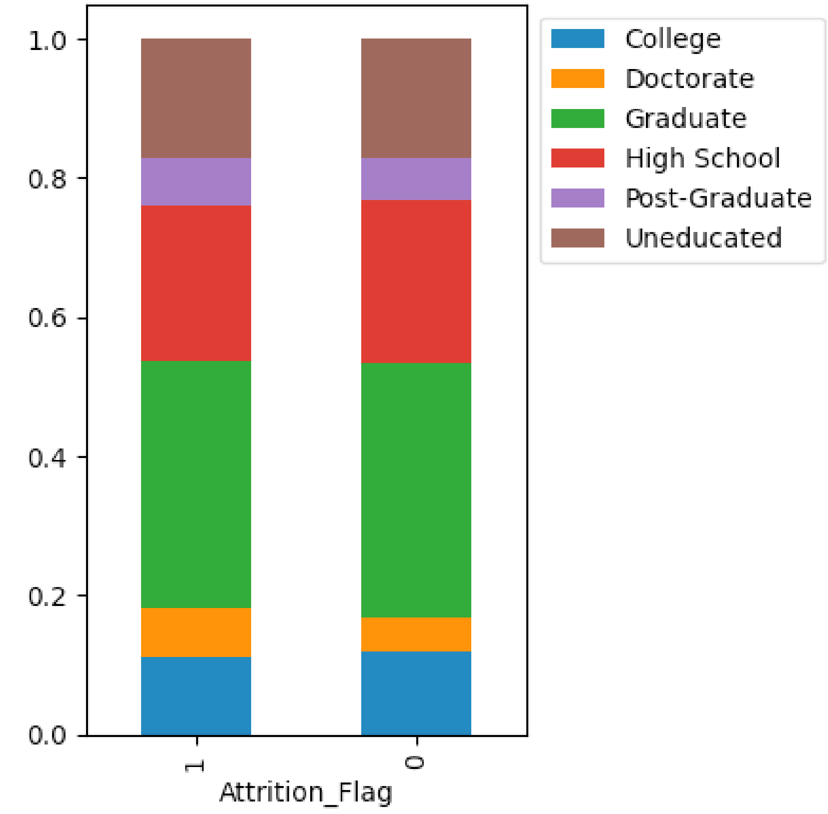
This feature show that nearly 53% of the customer base is female and 47% male. And this feature doesn't help in discriminating weather a client is going to drop the bank's services.

While gender may not directly impact attrition, understanding the customer base is crucial to offer products according to different lifestyles and needs.

EDA Results

Education_Level





The education feature reveals that the majority of the customer base falls under the "Graduate" classification, followed by "High School," "Uneducated," "College," "Post Graduate," and lastly "Doctorate." Unfortunately, this feature displays an almost identical distribution between the customers who left and those who stayed, so it doesn't have utility for distinguishing between customers who will stay and those who will leave.

EDA Results

Education_Level

Despite the apparent lack of discrimination power, this understanding of the educational background of the customer base can still offer some insights and guide strategies:

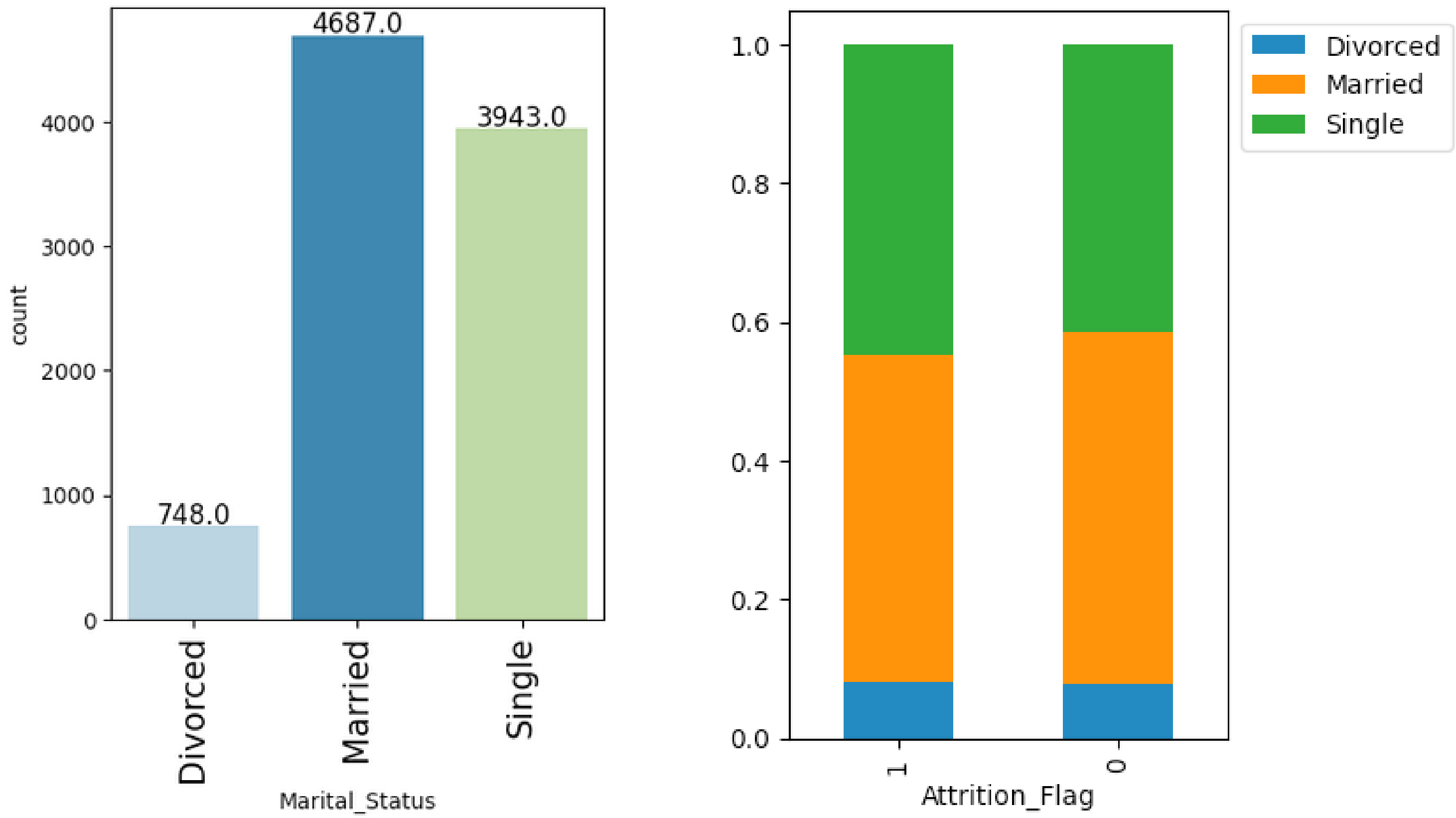
Knowing the education levels of customers allows the bank to tailor its marketing and communication efforts. For instance, financial products can be explained and advertised in ways that resonate with different education levels.

Education levels may correlate with financial needs and preferences. The bank can design and offer products or services that align with the specific life stages and financial goals of different education groups. For example, the bank could offer financial literacy programs or investment workshops tailored to various education levels.

The fact that education does not distinguish between staying and leaving customers might be an essential insight in itself. It suggests that education is not a critical factor in customer retention for this particular bank, and efforts might be better focused on other features and attributes that show a stronger correlation with attrition.

EDA Results

Marital_Status



This feature indicates that 4687 people are married, 3943 are single, and 748 are divorced. Similarly, this feature does not assist in discriminating between customers who will leave the bank's services and those who will remain.

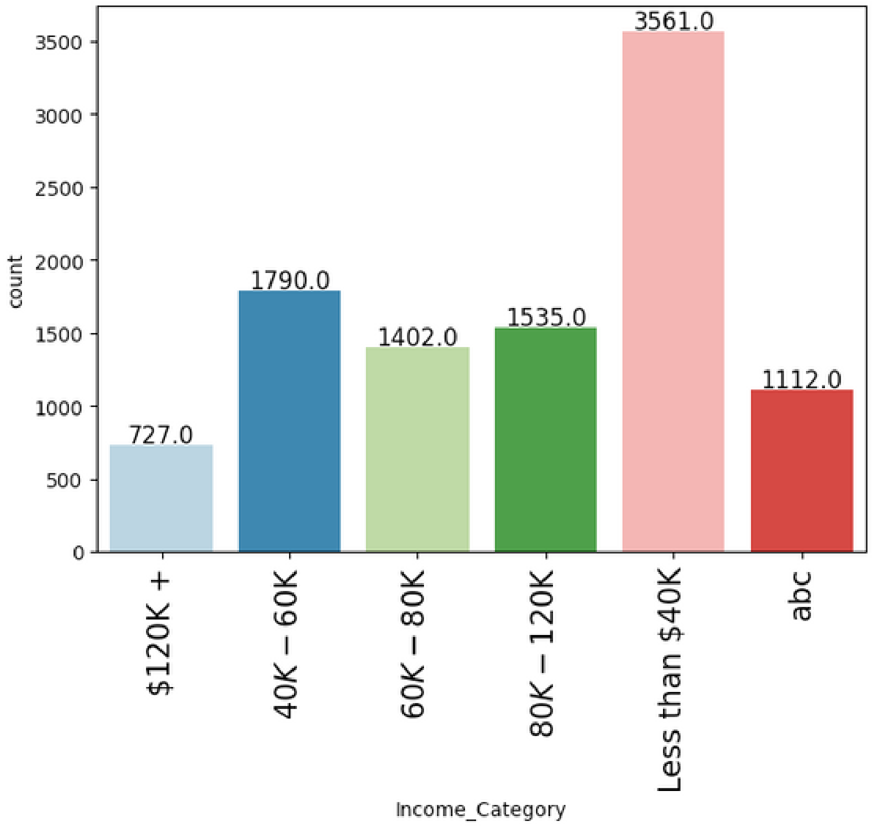
EDA Results

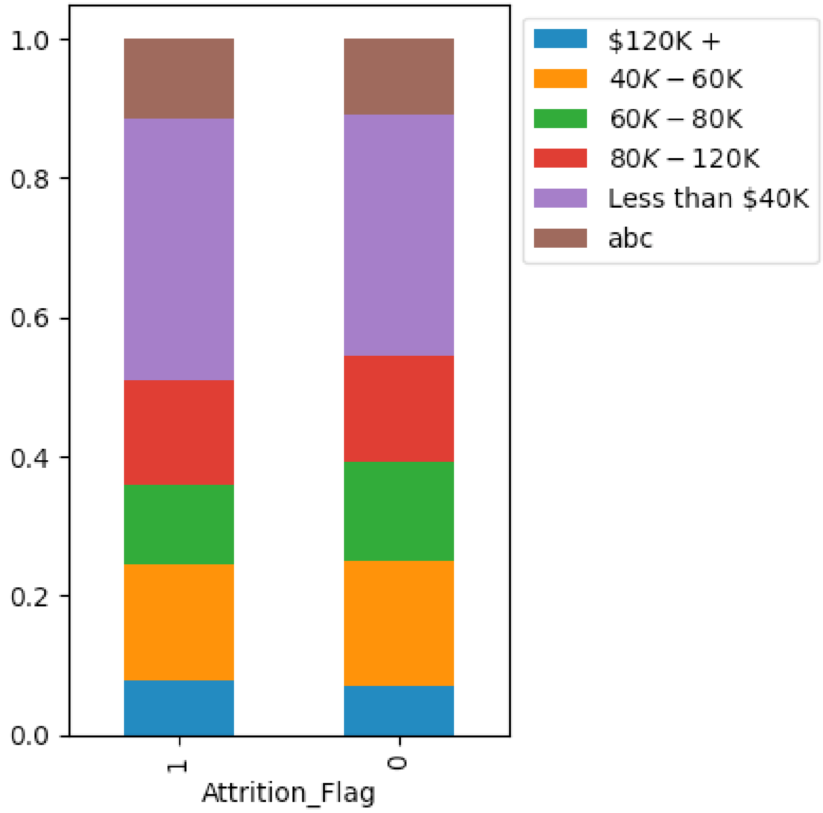
Marital_Status

While marital status may not directly impact attrition, understanding the marital composition of the customer base can guide in tailoring products and services. For instance, married couples might be more interested in joint accounts, mortgages, or family insurance plans, while single individuals might have different financial needs and goals. The bank could consider personalized offerings and targeted marketing based on marital status, even if it doesn't directly correlate with customer retention it does affect on the Total_Relationship_Count and that variable has a positive impact on customer retention.

EDA Results

Income_Category





This category shows that the group earning less than $40K is the largest, and the groups between $40K - $60K, $60K - $80K, and $80K - $120K are more or less even, along with a group called "abc" that is likely an error. The smallest group earns $120,000 and above.
Both the customers who remain and those who leave show almost identical proportions in the income clusters.

This might explain the low Credit_Limits, the biggest base are from clients that fall on the income category of less than $40K

EDA Results

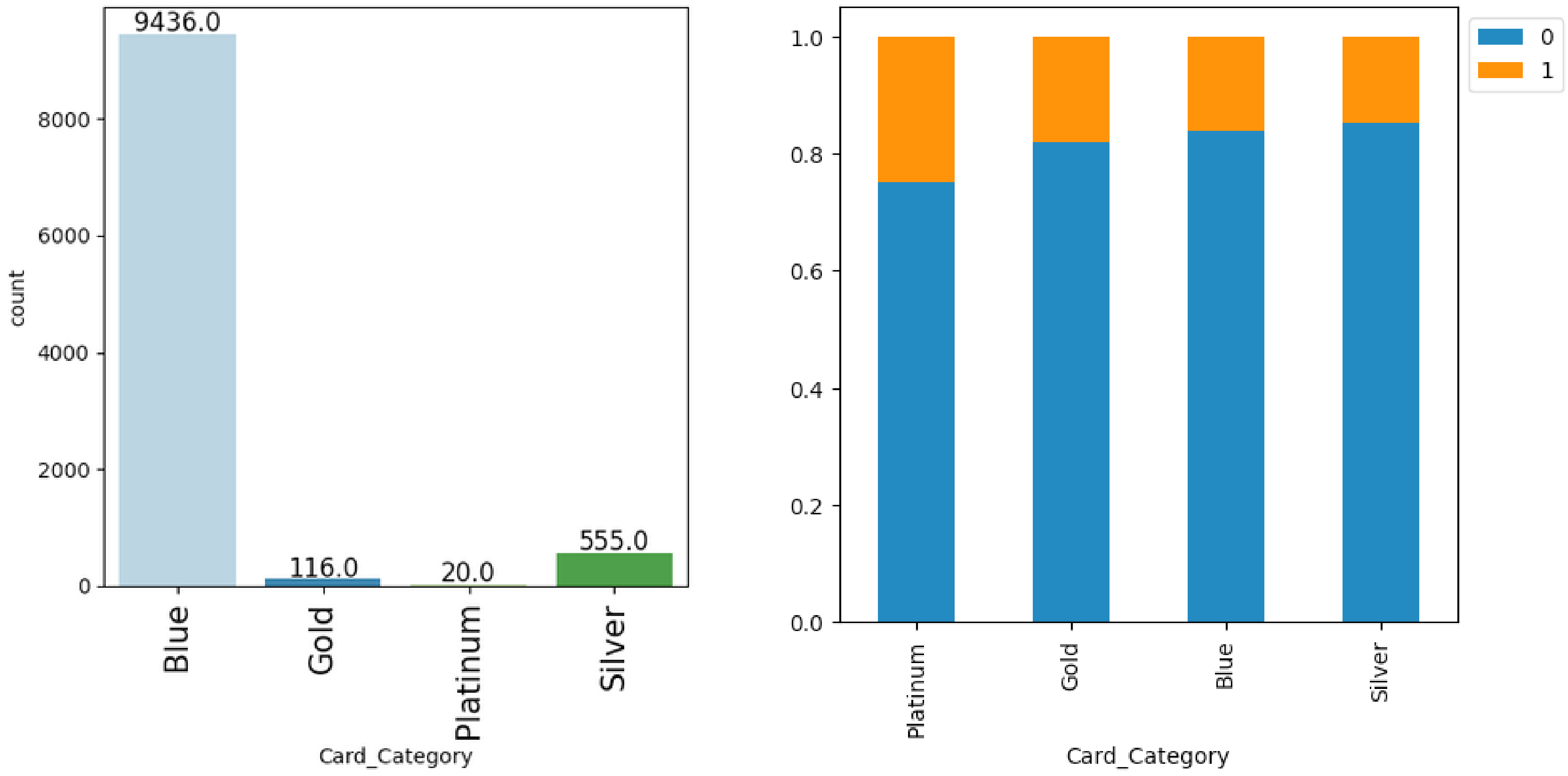
Income_Category

The distribution reveals that the vast majority of customers leave their card inactive for a maximum of three months, followed by a maximum of two months, and finally one month of inactivity. Notably, the segment of customers that abandoned the bank demonstrates a higher number of inactive months compared to those who remained with the bank.
There is a positive correlation of 0.15 with Attrition_Flag, indicating that the greater the number of inactive months, the higher the likelihood of account abandonment.

Different credit offers, investment opportunities, and savings plans can be tailored to various income levels that can lead to the increase of products adopted by the customer.

EDA Results

Card_Category



This category shows the type of credit card the customer has, with "Blue" being the predominant one with 9436 users, followed by "Silver" with 555 users, "Gold" with 116 users, and "Platinum" with only 20 users. The product that shows a slight increase in the proportion of customers leaving the services is the Platinum card, but overall the distribution looks quite similar among the different credit cards, so this feature hardly helps us discriminate between customers who leave and those who stay.

EDA Results

Card_Category

The distribution reveals that the vast majority of customers leave their card inactive for a maximum of three months, followed by a maximum of two months, and finally one month of inactivity. Notably, the segment of customers that abandoned the bank demonstrates a higher number of inactive months compared to those who remained with the bank.
There is a positive correlation of 0.15 with Attrition_Flag, indicating that the greater the number of inactive months, the higher the likelihood of account abandonment.

The bank might investigate why Platinum cardholders show a slightly higher attrition rate. Is there dissatisfaction with the benefits or services associated with this card? Could the bank offer additional features or incentives to retain these customers? Understanding card categories can lead to better-tailored card offers, enhanced customer satisfaction, and possibly, improved retention strategies.

Data Preprocessing

- **DUPLICATE VALUE CHECK**

No duplicates were found. Unique CLIENTNUM observed.

- **MISSING VALUE TREATMENT**

Education_Level showed 1519 missing values, Marital_Status 749, and Income_Category 1112 after replacing "abc" for NaN value. A simple imputer with "Most Frequent" strategy is used after splitting the data.
No missing values observed after this action.

- **OUTLIER CHECK (TREATMENT IF NEEDED)**

Attrition_Flag 16% of data is considered an outlier. 9.7% for Credit_Limit, Avg_Open_To_Buy shows 9.5%, Total_Trans_Amt 8.8%, Contacts_Count_12_mon 6.2%. Beyond that a little percentage is considered an outlier on each feature. No treatment needed as these are actual customers.

- **FEATURE ENGINEERING**

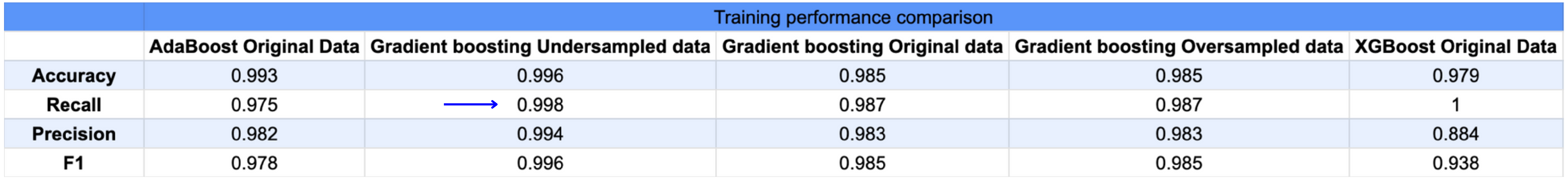
After getting dummies for categorical variables, we ended up with 29 columns

- **DATA PREPARATION FOR MODELING**

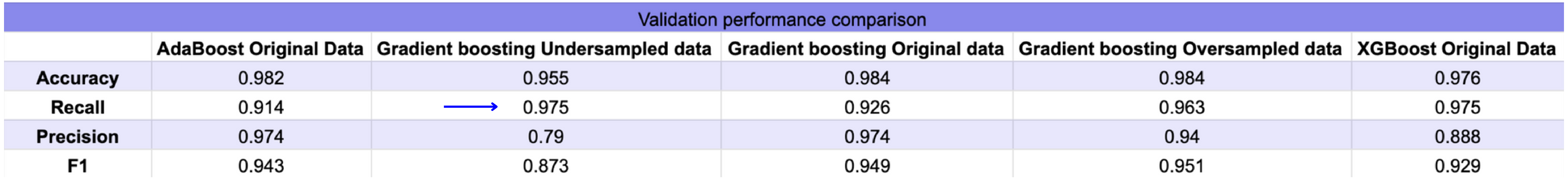
Defined Attrition_Flag as 'y' variable.
Data split into 80% for training, 5% for validation and 15% for testing.

Model Performance Summary

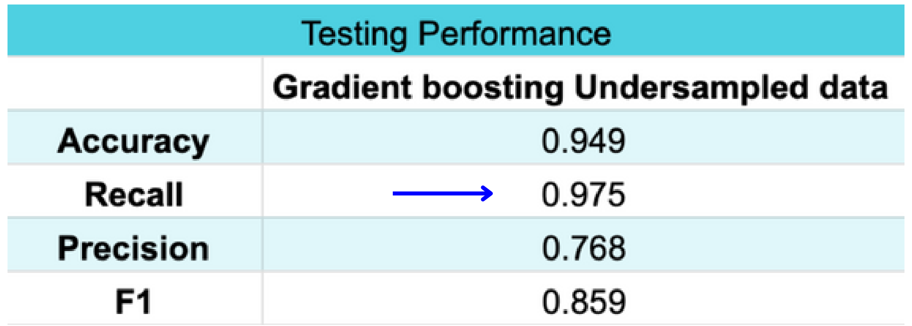
Bank would want Recall to be maximized, greater the Recall higher the chances of minimizing false negatives. Hence, the focus should be on increasing Recall or minimizing the false negatives or in other words identifying the true positives (i.e. Class 1) so that the bank can retain their valuable customers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



	Training performance comparison				
	AdaBoost Original Data	Gradient boosting Undersampled data	Gradient boosting Original data	Gradient boosting Oversampled data	XGBoost Original Data
Accuracy	0.993	0.996	0.985	0.985	0.979
Recall	0.975	→ 0.998	0.987	0.987	1
Precision	0.982	0.994	0.983	0.983	0.884
F1	0.978	0.996	0.985	0.985	0.938



	Validation performance comparison				
	AdaBoost Original Data	Gradient boosting Undersampled data	Gradient boosting Original data	Gradient boosting Oversampled data	XGBoost Original Data
Accuracy	0.982	0.955	0.984	0.984	0.976
Recall	0.914	→ 0.975	0.926	0.963	0.975
Precision	0.974	0.79	0.974	0.94	0.888
F1	0.943	0.873	0.949	0.951	0.929



	Testing Performance
	Gradient boosting Undersampled data
Accuracy	0.949
Recall	→ 0.975
Precision	0.768
F1	0.859

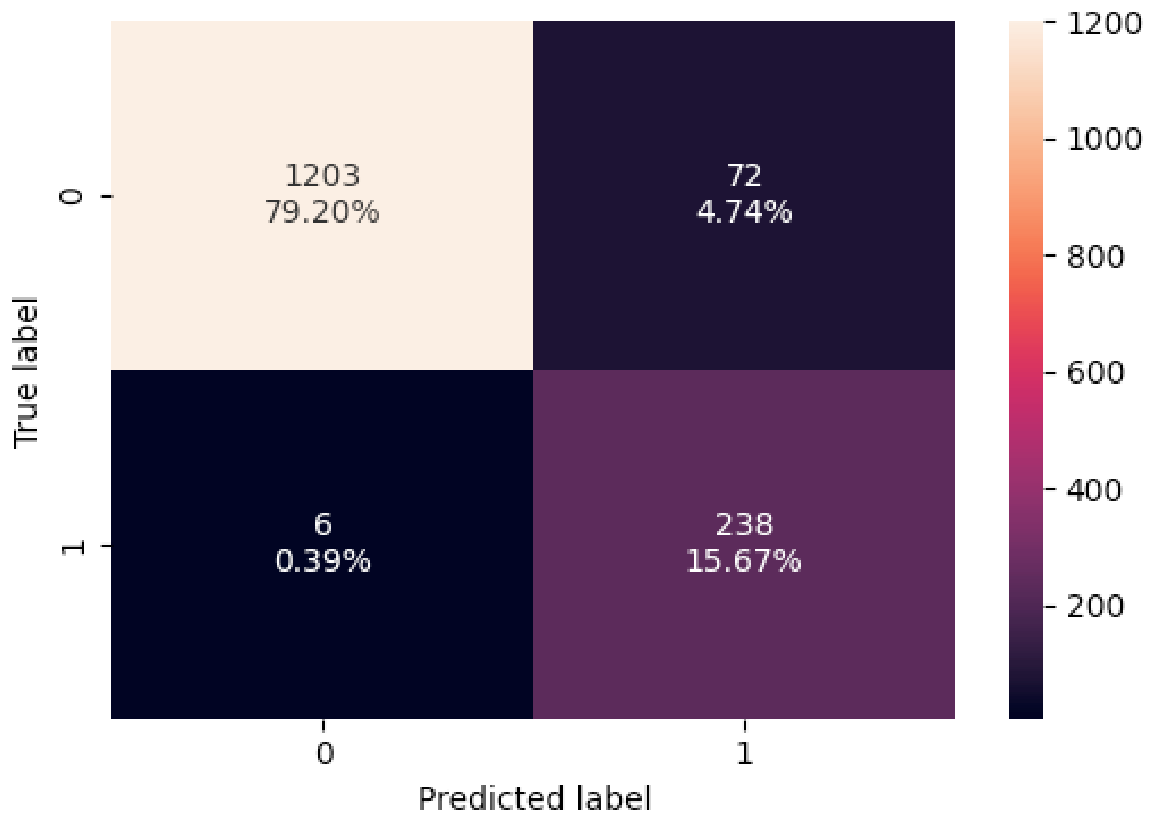
We observed the best performance on Recall score with Gradient boosting with undersampled data model. Although XGBoost with original data showed the same performance on training, the computational time comes with no benefit on performance.

LINK TO APPENDIX SLIDE ON MODEL ASSUMPTIONS

APPENDIX

Model Performance Summary

Gradient Boosting Classifier Best parameters
max_features: 0.7
Initial Model: Ada Boost Classifier
Random state = 1
learning_rate: 0.2
n_estimators: 125
subsample: 0.7
CV score=0.95239021514

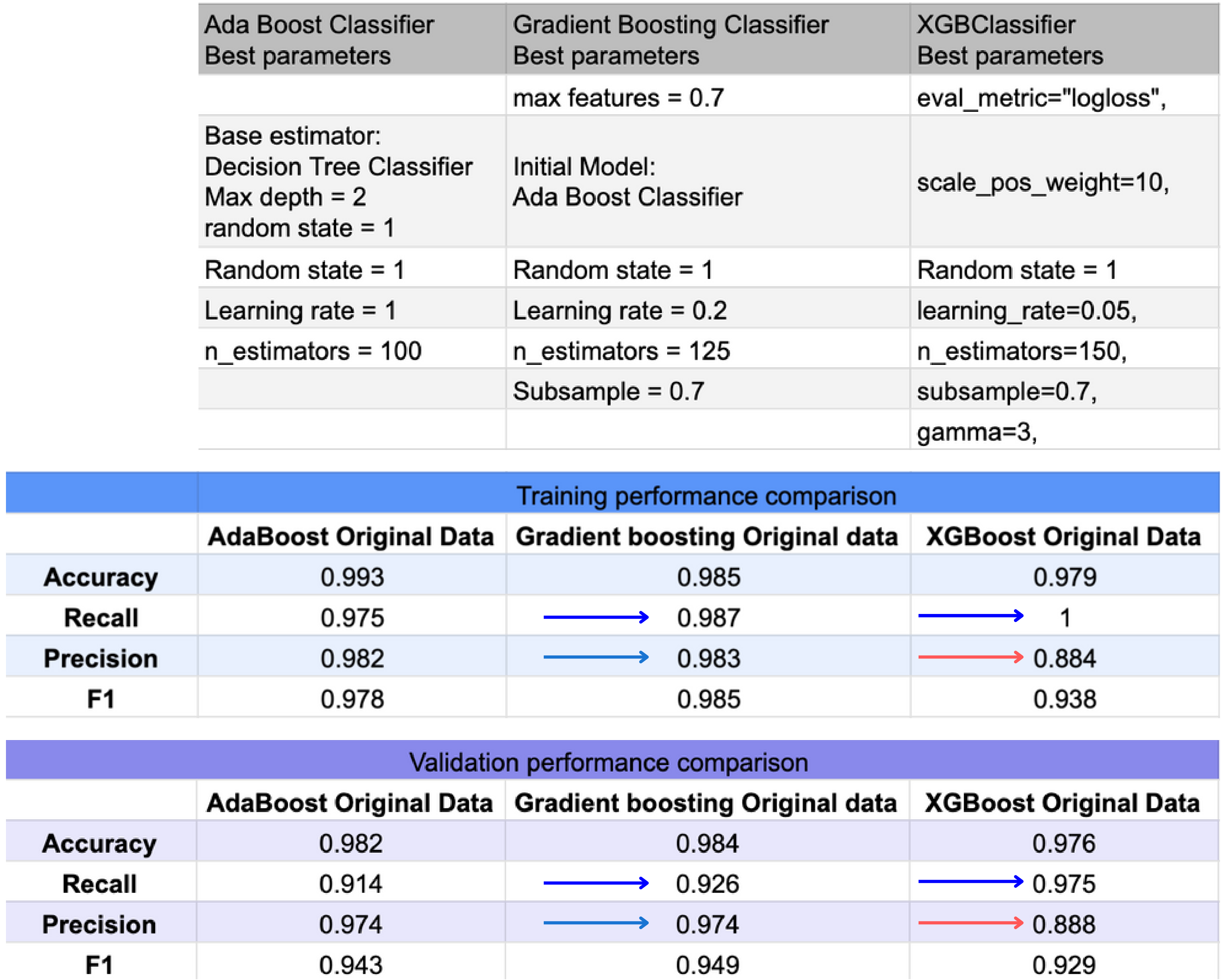


Testing Performance	
	Gradient boosting Undersampled data
Accuracy	0.949
Recall	→ 0.975
Precision	0.768
F1	0.859

A high value was obtained on Recall score as the bank objective appointed.

Precision performed low which means that some efforts maybe taken on clients that weren't thinking on leaving the bank but still can result on long term loyalty.

Model Performance Summary (original data)



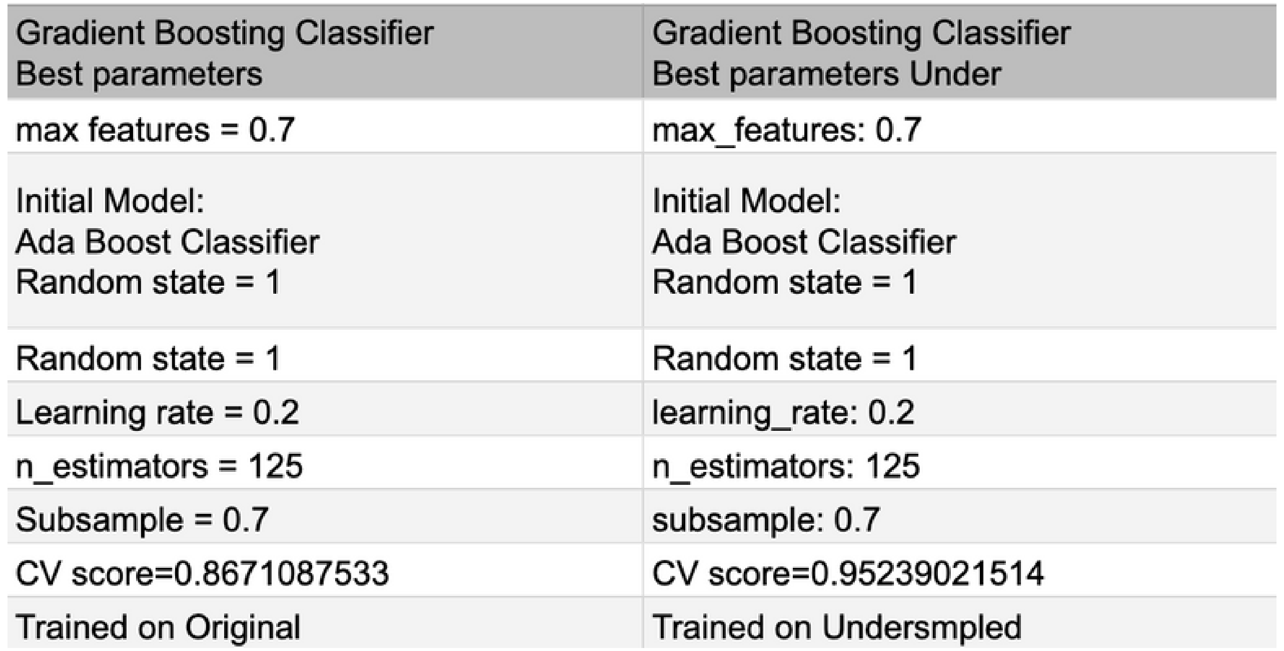
Ada Boost Classifier Best parameters	Gradient Boosting Classifier Best parameters	XGBClassifier Best parameters
	max features = 0.7	eval_metric="logloss",
Base estimator: Decision Tree Classifier Max depth = 2 random state = 1	Initial Model: Ada Boost Classifier	scale_pos_weight=10,
Random state = 1	Random state = 1	Random state = 1
Learning rate = 1	Learning rate = 0.2	learning_rate=0.05,
n_estimators = 100	n_estimators = 125	n_estimators=150,
	Subsample = 0.7	subsample=0.7,
		gamma=3,

	Training performance comparison		
	AdaBoost Original Data	**Gradient boosting Original data**	**XGBoost Original Data**
Accuracy	0.993	0.985	0.979
Recall	0.975	→ 0.987	→ 1
Precision	0.982	→ 0.983	→ 0.884
F1	0.978	0.985	0.938

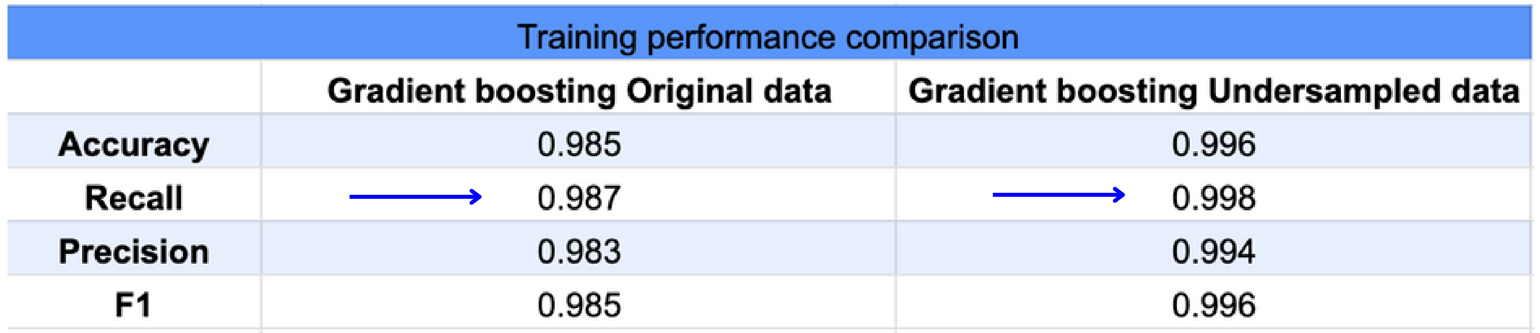
	Validation performance comparison		
	AdaBoost Original Data	**Gradient boosting Original data**	**XGBoost Original Data**
Accuracy	0.982	0.984	0.976
Recall	0.914	→ 0.926	→ 0.975
Precision	0.974	→ 0.974	→ 0.888
F1	0.943	0.949	0.929

Some overfitting is observed on training data using the XGBoost model with original data, which is expected. On validation performance it gives the best performance on Recall Score.

Model Performance Summary (undersampled data)

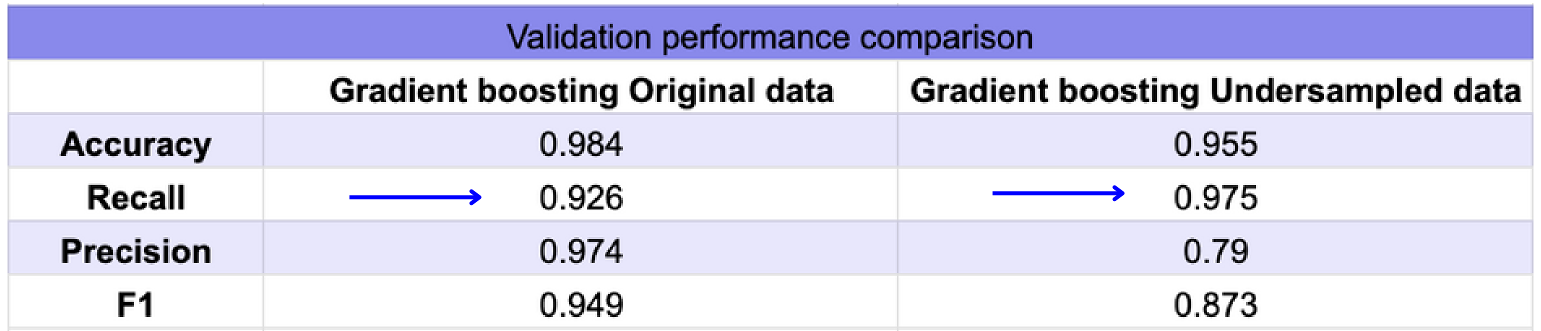


Gradient Boosting Classifier Best parameters	Gradient Boosting Classifier Best parameters Under
max features = 0.7	max_features: 0.7
Initial Model: Ada Boost Classifier Random state = 1	Initial Model: Ada Boost Classifier Random state = 1
Random state = 1	Random state = 1
Learning rate = 0.2	learning_rate: 0.2
n_estimators = 125	n_estimators: 125
Subsample = 0.7	subsample: 0.7
CV score=0.8671087533	CV score=0.95239021514
Trained on Original	Trained on Undersmpled



Training performance comparison

	Gradient boosting Original data	Gradient boosting Undersampled data
Accuracy	0.985	0.996
Recall	→ 0.987	→ 0.998
Precision	0.983	0.994
F1	0.985	0.996



Validation performance comparison

	Gradient boosting Original data	Gradient boosting Undersampled data
Accuracy	0.984	0.955
Recall	→ 0.926	→ 0.975
Precision	0.974	0.79
F1	0.949	0.873

A better performance on the score of interest is observed when undersampled training data is used. A recall Score of 0.998 on training and 0.975 on validation vs 0.987 and 0.926 for training with original data.

Precision may be a trade off as mentioned, but we are after the best recall score.

Great Learning

Happy Learning !