Credit Card Users Churn Prediction

Project 3: Machine Learning (AI/ML course) by Sarah Choi

Date: 10/12/23

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary

Executive Summary - Conclusions

	Conclusion
Predictive Model	• Analyzed the dataset from Thera Bank for their credit card customers and applied four different type of ensemble algorithm to build several models, all models were also tuned to select the best Classification predictive model. • The bank can deploy the final model to i) predict and the customers who will leave their credit card services at the bank. ii) to find the drivers of attrition. iii) based on which the bank can take appropriate actions to improve service, so customers do not renounce their credit cards.
Model performance	• 15 models were built and validated with test dataset; 4 models were then fine tuned to get the best performance for final selection.
Feature Importance	• Total Transaction Amount, Total Transaction Counts and Total Revolving Balance are the most important variables in predicting customer attrition.
Outliers	• There are outliers in many variables (Total_Trans_Amt, Total_Amt_Chng_Q4_Q1, Avg_Open_To_Buy), some of which has nearly 26% of the data. Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results. This type of outlier can be a problem in regression analysis
Customer attrition Observations	• Our data analysis from the data collection shows the customers who have contacted the bank 6 times in the last 12 months are 100% attrite customers, followed by the customers who contacted 5 times, ~34% are attrite customers. • It seems the number of times the customers contacted the bank could contribute to the reason the customers become attrite, we should investigate further on the reason for these customer contact. • There are overall 16% attrite customers out of the total who have been inactive up to 6 months. • Total transaction amount and total transaction counts show significant difference between attriting and non-attriting customers. These scales can act as preliminary step to understand dissatisfaction of the customers. Lower the Total transaction amount and total transaction counts higher are the chances of attrition.

Executive Summary - Recommendations

	Recommendations
Most Significant Attributes	According to the model predication: a) If a customer credit card transaction amount, transaction count, and total revolving balance start to consistently drop over a period, then there's a very high chance the customer will discontinue their credit card service with the bank b) This is also reflected in the past 12 months data where the total transaction amount and transaction counts are significantly lower with the attrite customers as compared to the existing card customers. These features are the most important in predicting whether customer will leave the credit card service, the bank should monitor the customer card usage and provide action to prevent them from leaving.
Credit Card Marketing Strategy	• Banks are not in touch with their customers. They rarely talk with their customers and lack data-driven banking insights on what makes them happy or unhappy with the bank. • Banks rely on a one-size-fits-all approach. Because of the lack of personal touch, banks tend to use the same approach for all customers, regardless of their financial situation or personality. • The above two points reflect in the data analysis for Thera Bank where we found the current attrite customers contacted the bank many times in the past 12 months, their account activities became inactive, and the credit card usage eventually reduced. Poor service and poor financial advice emerged as top reasons why customers leave their banks. • Thera Bank's credit card strategy should include collecting customer experience data in real-time across all channels and touchpoints, identify key drivers and take action to improve customer satisfaction and loyalty.
Outliers	• It is observed that there are large quantities of outliers in many variables (Total_Trans_Amt, Total_Amt_Chng_Q4_Q1, Avg_Open_To_Buy). Total_Trans_Amt being the most important feature to predict customer attrition, has nearly 26% outlier. • Outliers are problematic for many statistical analyses because they can cause tests to either miss significant findings or distort real results. • Recommended approach: analyze how the initial machine learning or data science problem was framed, how the target population and sample were chosen, and how so many outliers made into the dataset. • The stakeholders (people who directly benefit or lose from the project) should be informed of the decision.

Business Problem Overview and Solution Approach

Problem Statement

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 objective of this analysis is to build a classification model that will help the bank improve its services so that customers do not renounce their credit cards

Solution Approach

Identity focus area

To extract actionable insights from the data that Thera Bank has collected to identify key areas of focus.

Perform Data Analysis & Build Predictive Model

- Perform EDA on Variables that influence the number of credit card users.
- Build a classification model to predict if the customer is going to churn or not.
- Optimize the model using appropriate techniques.

Summarize Recommendation

Summarize the insights and recommendations that will help the bank's business.

Data Overview

The data contains the different data related to a customer at Thera Bank. The detailed data dictionary is given below.

Variable	Description
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

Data Overview – con't

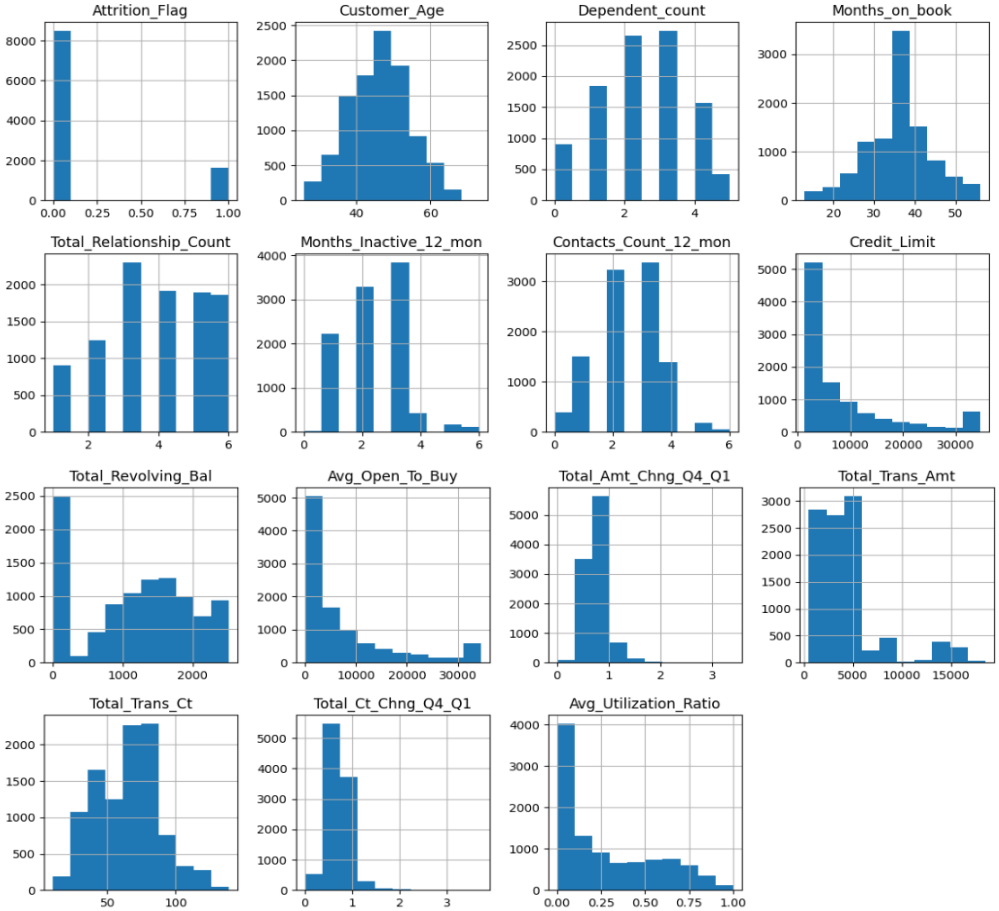
Continued with the rest of the data dictionary below.

Variable	Description
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

Summary:

- There is a total number of 10,127 unique records in the data frame with 21 columns for the data analysis.
- There are 6 object types, and the rest are numerical values.
- There is no duplicate row in the data.
- Education_Level column has 15% missing values out of the total observations.
- Marital_Status column has 7.4% missing values out of the total observations.

EDA Results Summary - Univariate Analysis



EDA Results Summary - Univariate Analysis

Below show the summary of the EDA findings, detailed charts can be found in the Appendix section

Univariate Analysis

- ***Customer_Age:*** Age has symmetrical and normal distribution, the number of customers who have accounts above the median age of 46 are slightly higher than those customers who are aged below 46
- ***Months_on_book***: The Period of relationship customers with the bank have a very wide range from 13 months to 56 months with most of the customer relationship average at 36 months. There are many outliers that are less than 13 months and more than 56 months
- ***Credit_Limit:*** Distribution of Credit limit is right-skewed with many outliers.
  From the boxplot, we can see many customers with credit limit that are more than ~$11000 that are consider outliers.
- ***Total_Revolving_Bal:*** The distribution of the total revolving balance is slightly skewed towards the right due to high number of the customer account 2,470 have zero balance. The average revolving balance is around ~$1160
- ***Avg_Open_To_Buy:*** This variable represent the amount left on customer's credit card to use. and this column represents the average of this value for the last 12 months. It has almost identical distribution as the variable "Credit Limit" with many outliers for the accounts that have ~9900 dollars left on their card. This pattern is consistent and seems to tell us customers will spend whatever they are allow to spend within the credit limit on the credit card.
- ***Total_Trans_Ct:*** The distribution of this variable is slightly left skewed with an average of 65 total transaction per customer over the last 12 months. There are a few outliers of around 130 counts that might need further investigation. All such records with a total transaction count of more than 130 belongs to existing customers have they all have contact with the bank at least once in the last 12 months, hence they seem to represent true pattern and we don't need to treat them as outliers.
- ***Total_Amt_Chng_Q4_Q1:*** The distribution of the Ratio of the total transaction amount in 4th quarter and the total transaction amount in 1st quarter is very evenly distributed, but there's large number of outliers with ration that are greater than 1.3 and smaller than 0.2.
- ***Total_Trans_Amt:*** The distribution of the Total transaction amount is skewed towards the right.* There are many outliers in this variable and the values above $4,700 are being represented as outliers by the boxplot. a total of 2637 rows. 26% of the total records. 296 are attrite customers, 2341 are existing customers.
- ***Total_Ct_Chng_Q4_Q1:*** The distribution of the Average utilization ratio, represents how much of the available credit the customer spent, is skewed towards the right due to majority of this ratio is zero.

EDA Results Summary – Univariate Analysis con't

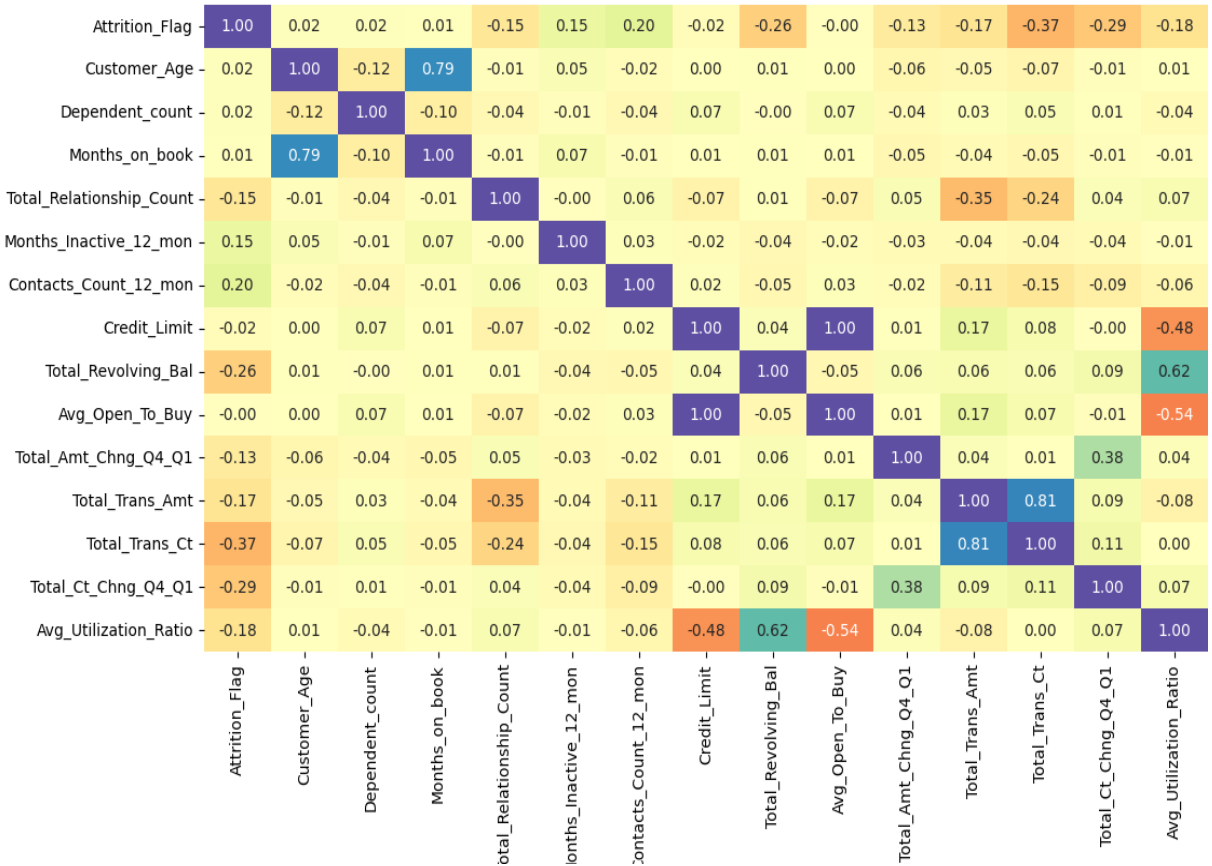
Below show the summary of the EDA findings, detailed charts can be found in the Appendix section

Univariate Analysis

- ***Avg_Utilization_Ratio***: The distribution of the Average utilization ratio, represents how much of the available credit the customer spent, is skewed towards the right due to majority of this ratio is zero.
- ***Dependent_count***: Half of the customers base (53%) have either 2 or 3 dependents
- ***Total_Relationship_Count:*** 22.8% of the customers have 3 products with the bank followed by customers who have 4 or 5 or products with ~18% respectively out of the total.
- ***Months_Inactive_12_mon:*** 38% of the customers are inactive up to 3 months, followed by 32.4% who are inactivies up to 2 months.
- ***Contacts_Count_12_mon:*** 33.4% of the customers contacted the bank 3 times, followed by 31.9% customers contacted the bank 2 times in the last 12 months.
- ***Gender:*** 52.9% of the customers are female and 47.1% are male.
- ***Education_Level:*** 30.9% of the customers have graduate degree followed by 19.9% customers have High school degree
- ***Marital_Status:*** 46.3% of the customers are married, followed by 38.9% customers are singles and 7.4% are divorced.
- ***Income_Category:*** 35.2% of customers have income less than $40K followed by customers who have income between 40K-60K. There is 11% that have anomalous data in the dataset which will be cleaned up
- ***Card_Category:*** 93.2% of the customers have Blue type of credit card followed by 5.5% of Silver card.
- ***Attrition_Flag:*** 83.9% are existing customer where 16.1% are attrite customers

EDA Results Summary – Bivariate Analysis

Below show the correlation matrix of the variables.



Observations:

- Total Transaction counts have the strongest correlation with Total Transaction amount (0.81), indicates as total transaction counts go up, total transaction amount go up.
- Customer's age has high positive correlation with the period the customer has relationship with the bank (0.79), this indicates as age goes up, the longer the customers have the relationship with the bank.
- Total revolving balance is also positively correlated (0.62) to Average utilization ratio. As customer spend more on the amount left on the credit card each month, the revolving balance increases if they do not pay the balance in full.
- Total_Ct_Chng_Q4_Q1 is positively correlated (0.38) to the Total_Amt_Chng_Q4_Q1, as the ratio goes up on the total count, so does the ratio of the total amount.
- Based on the Reatlion bteween Avg_Open_To_Buy, Credit_Limit and Avg_Utilization_Ratio: (Avg_Open_To_Buy / Credit_Limit) + Avg_Utilization_Ratio = 1. The correlation matrix reflected accordingly, Avg_Open_To_Buy is negatively correlated to Avg_Utilization_Ratio (-0.54), the Avg_Utilization_Ratio represents how much of the available credit the customer spent, if this ratio goes up, the Avg_Open_To_Buy, the amount left on the credit card to use goes down.
- Avg_Open_To_Buy is strongly positively correlated to Credit_Limit (1.00)since customer can only spend on what is available with their credit limit.

EDA Results Summary – Bivariate Analysis

Below show the summary of the EDA findings, detailed charts can be found in the Appendix section

Attrition_Flag vs Gender :

- There are more Female attrite customers than Male attrite customers

Attrition_Flag vs Marital_Status:

- The ratio of attrite customers in the three marital status are closed to the same, marital status does not seem to have an impact on the target variable

Attrition_Flag vs Education_Level:

- Within the customers population who have Doctorate degree, it consists the most attrite customers (~21%).
- ~16% of the customers in all different type of education level are attrite customers.

Attrition_Flag vs Income_Category:

- There are ~17% of attrite customers with most of the income category except for the 60K-80K which is ~13%.

Attrition_Flag vs Contacts_Count_12_mon:

- For the customers who have contacted the bank 6 times in the last 12 months, they are 100% attrite customers, followed by the customers who contacted 5 times, ~34% are attrite customers.
- It seems the higher the number of times the customers contacted the bank could be one of the reason the customers become attrite, we should investigate further.

Attrition_Flag vs Months_Inactive_12_mon:

- There are overall 16% attrite customers out of the total who are have been inactive up to 6 months.
- With the customers who are the least inactive (1 month) also consist of the least number of attrite customers.

Attrition_Flag vs Total_Relationship_Count:

- Customers who have 4 to 6 products with the bank has the least number of attrite customers, where the customers who have 1 or 2 products have the highest number of attrite customers.

Attrition_Flag vs Dependent_count:

- There are ~17% of attrite customers across all dependent count category.
- The ratio of attrite customers in the Dependent count category are closed to the same, marital status does not seem to have an impact on the target variable.

EDA Results Summary – Bivariate Analysis con't

Below show the summary of the EDA findings, detailed charts can be found in the Appendix section

Attrition_Flag vs Total_Revolving_Bal:

- The distribution of the Total_Revolving_Bal is extremely skewed to the right for the group of customers, mainly due to the high number of customers who does not carry any revolving balance to next month which skewed the entire data population.
- Majority of the attrite customers do not have a revolving balance and carry a lot lower balance compared to the existing customers.

Attrition_Flag vs Credit_Limit:

- The distribution of the credit is extremely skewed to the right for the group of existing customers and the attrite customers.
- The distribution and the total amount of the credit limit are similar between the existing customers and the attrite customers indicates credit limit does not have much impact on the target variable.

Attrition_Flag vs Customer_Age:

- The distribution of the customer Age has very normal distribution between the existing customers and the attrite customers. There is not much difference in how data is dispersed between the two groups.

Total_Trans_Ct vs Attrition_Flag:

- There's a significant difference in the total transaction counts between the existing customers and the attrite customers.
- Attrite customers have much lower total transaction counts with average between 30 to 50 vs 58 to 82 for the existing customers.

Total_Trans_Amt vs Attrition_Flag:

- There's a similar difference in the total transaction amount between the existing customers and the attrite customers. Attrite customers have much lower total transaction amount with average between $1800 to $2800 vs $2500 to $4800 for the existing customers.

Total_Ct_Chng_Q4_Q1 vs Attrition_Flag:

- The distribution of Total_Ct_Chng_Q4_Q1 is quite normal between the existing customers and the attrite customers.
- The ratio is lower with the attrite customers although the range is bigger as compared to the existing customers.

Avg_Utilization_Ratio vs Attrition_Flag:

- The distribution of the Avg_Utilization_Ratio is extremely skewed to the right for the attrite customer, mainly due to the high number of customers who does not have any credit card utilization which skewed the entire data population.
- There is much higer Avg_Utilization_Ratio with the existing customers than the attrite customers.

EDA Results Summary – Bivariate Analysis con't

Below show the summary of the EDA findings, detailed charts can be found in the Appendix section

Attrition_Flag vs Months_on_book:

- 50% of the customers for both existing and attrite have almost the same number of months of relationship with the bank, between 28 to 40 months.

Attrition_Flag vs Total_Revolving_Bal:

- The distribution of the total revolving balance is skewed to the right for the two group of customers, mainly due to the high number of customers who does not have any credit card utilization which skewed the entire data population.
- The total revolving balance for the attrite customers is much lower than the existing with 60% range from 0 to 1200 dollars.
- The total revolving balance for the existing customers is higher with 50% range from 800 to 1800 dollars.

Attrition_Flag vs Avg_Open_To_Buy

- The distribution of the Average Open to buy is almost the same between the attrite customers and the existing customers and is very right skewed.

Data Preprocessing

- **Duplicate value check:** There's no duplicate row in the data frame.
- **Missing value treatment:** We imputed the missing values in the columns Education_Level column and Marital_Status column, and have validated there is no column has missing values in the train or test datasets.
- **Outlier check (treatment if needed):** Outliers were identified as follows, they were ignored as part of the EDA.

Columns	% of outlier
Attrition_Flag	16.07
Customer_Age	0.02
Months_on_book	3.81
Months_Inactive_12_mon	6.21
Contacts_Count_12_mon	9.72
Credit_Limit	9.51

Columns	% of outlier
Avg_Open_To_Buy	9.51
Total_Amt_Chng_Q4_Q1	3.91
Total_Trans_Amt	8.85
Total_Trans_Ct	0.02
Total_Ct_Chng_Q4_Q1	3.89

- **Data preprocessing for modeling:** checking for anomalous Values

➢ Replaced the anomalous values of "abc" with "NaN" in the column Income_Category.

Model Building

The Followings are the classification model building steps :

- EDA (Univariate and Bivariate) was performed on the data frame, detail is described in slides 6- 21 to get good evaluation of the data before building the model, also data preprocessing was performed to ensure all anomalies are handled.
- Before proceeding with building the model, the data was split the data into train data (80%) and test data (20%). Then the test dataset was further split into validation data (75%) and test data (25%) to be able to evaluate the model that was built on the train data, encode categorical features and scale numerical values.
- Total rows for train data is 8,101, for validation data is 507, and for test data is 1,519 after the two data split executions.
- Cross validation techniques to ensure the best model is selected to achieve a generalized model performance in production.
- Bagging, Random Forest, Gradient Boosting, AdaBoosting, and XG Boosting ensemble learning techniques were used to produce the best model for prediction.
- Hyperparameters tuning was done to tune the models after the models were fitted with oversampled data and undersampled data to get generalized performance

Model Evaluation Criterion

Model evaluation criterion

- The objective is to predict the customers who will leave their credit card services and reasons so that bank could improve upon those areas.

Model can make wrong predictions as:

1) Predicting a customer will attrite and the customer doesn't attrite

2) 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.

Model Performance Summary (original data)

- During **initial built of the model**, Bagging, Random Forest, Gradient Boosting, AdaBoosting, and DecisionTree ensemble learning techniques were used with the original dataset to train 5 classification models to check their performance.
- Model built using Random Forest and Decision Tree perfectly classify all the data points on the training set with 100% recall score. 0 errors on the training set, each sample has been classified correctly. However, since there's no restriction applied, the trees continue to grow with all possible patterns in the training set. The result on the performance of the validation set is 75% and 80% Recall. This huge disparity suggests that the models are overfitting.

Training Methods	Training Performance (Recall Score)	Validation Performance (Recall Score)
Bagging	0.9838709677419355	0.8148148148148148
Random Forest	1.0	0.7530864197530864
Gradient Boosting	0.8840245775729647	0.9012345679012346
AdaBoosting	0.84715821812596	0.8641975308641975
Decision Tree	1.0	0.8024691358024691

- Gradient Boosting has the best performance followed by AdaBoost model as per the validation performance.

Model Performance Summary (oversampled data)

- When combining both random sampling methods to train the models can occasionally result in overall improved performance in comparison to the methods being performed in isolation as seen in the result from training the models with the original data.
- We applied a modest amount of oversampling to the minority class, which improves the bias to the minority class examples, whilst we also perform a modest amount of under sampling on the majority class to reduce the bias on the majority class examples.
- The following are the performance metrics for training and validation with oversampled data by using Bagging, Random Forest, Gradient Boosting, AdaBoosting, and DecisionTree ensemble algorithm.

Training Methods	Training Performance (Recall Score)	Validation Performance (Recall Score)
Bagging	0.9986762759229298	0.8765432098765432
Random Forest	1.0	0.8888888888888888
Gradient Boosting	0.9810266215619944	0.9382716049382716
AdaBoosting	0.9670539785262539	0.8888888888888888
Decision Tree	1.0	0.8024691358024691

- Gradient Boosting has the best performance followed by AdaBoost model as per the validation performance.

Model Performance Summary (undersampled data)

- When combining both random sampling methods to train the models can occasionally result in overall improved performance in comparison to the methods being performed in isolation as seen in the result from training the models with the original data.
- We applied a modest amount of oversampling to the minority class, which improves the bias to the minority class examples, whilst we also perform a modest amount of under sampling on the majority class to reduce the bias on the majority class examples.
- The following are the performance metrics for training and validation with undersampled data by using Bagging, Random Forest, Gradient Boosting, AdaBoosting, and DecisionTree ensemble algorithm.

Training Methods	Training Performance (Recall Score)	Validation Performance (Recall Score)
Bagging	0.9946236559139785	0.9382716049382716
Random Forest	1.0	0.9506172839506173
Gradient Boosting	0.9823348694316436	0.9382716049382716
AdaBoosting	0.9516129032258065	0.9506172839506173
Decision Tree	1.0	0.9135802469135802

- Based on the validation result, performance on all models have improved compared to the models built from oversampled data.
- AdaBoosting has the best performance followed by Gradient Boosting model as per the validation performance.

Hyperparameter Tuning

- After building 15 models, it was observed that both the Gradient Boosting and Adaboost models, trained on an undersampled dataset, as well as trained on an oversampled dataset, exhibited strong performance on both the training and validation datasets.
- Sometimes models might overfit after undersampling and oversampling, so it's better to tune the models to get a generalized performance
- We have tuned these 4 models using the same data (undersampled or oversampled) as we trained them on before using hyperparemer tuning technique and the result is as followings:

Tuning AdaBoostClassifier with Undersampled data

Model performance on training data

	Accuracy	Recall	Precision	F1
0	0.999	1.000	0.998	0.999

Model performance on validation data

	Accuracy	Recall	Precision	F1
0	0.947	0.951	0.770	0.851

Tuning Gradient Boosting with Undersampled data

Model performance on training data

	Accuracy	Recall	Precision	F1
0	0.996	0.998	0.994	0.996

Model performance on validation data

	Accuracy	Recall	Precision	F1
0	0.955	0.975	0.790	0.873

Tuning Gradient Boosting with Oversampled data

Model performance on training data

	Accuracy	Recall	Precision	F1
0	0.985	0.987	0.983	0.985

Model performance on validation data

	Accuracy	Recall	Precision	F1
0	0.984	0.963	0.940	0.951

Tuning XGBoost with Original data

Model performance on training data

	Accuracy	Recall	Precision	F1
0	0.938	0.998	0.724	0.839

Model performance on validation data

	Accuracy	Recall	Precision	F1
0	0.925	0.963	0.690	0.804

Model Comparison and Final Model Selection

Model performance on training data in the final model selection

Training performance comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data	XBBoost trained with orginal data
Accuracy	0.996	0.985	0.999	0.938
Recall	0.998	0.987	1.000	0.998
Precision	0.994	0.983	0.998	0.724
F1	0.996	0.985	0.999	0.839

Model performance on validation data in the final model selection

Validation performance comparison:

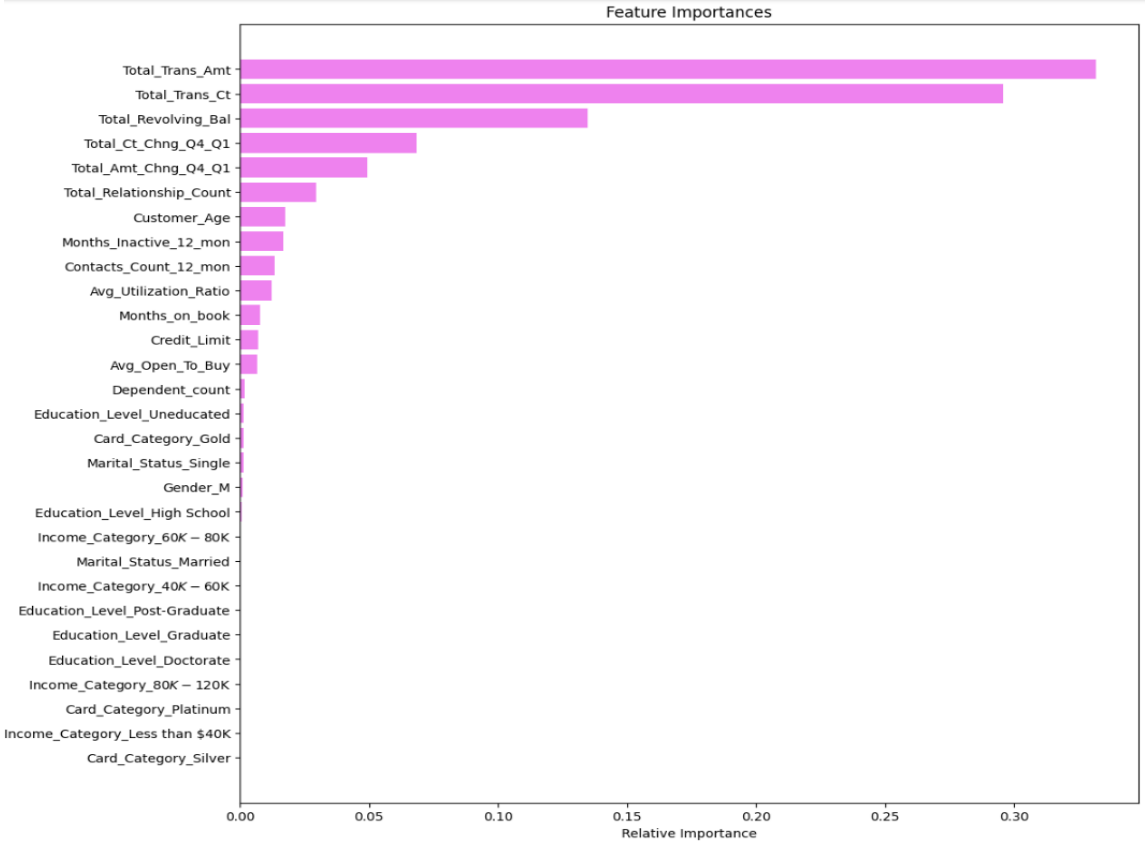
	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data	XBBoost trained with orginal data
Accuracy	0.955	0.984	0.947	0.925
Recall	0.975	0.963	0.951	0.963
Precision	0.790	0.940	0.770	0.690
F1	0.873	0.951	0.851	0.804

- Gradient Boosting model trained with undersampled data has the generalized performance and best recall score, therefore it's considered to be the best and final model. Model performance is then checked on unseen test data.

	Accuracy	Recall	Precision	F1
0	0.949	0.975	0.768	0.859

- The Gradient Boosting model trained on undersampled data has given 97.5% recall on the test set
- This performance is in line with what we achieved with this model on the train and validation sets and so this is a generalized model.

Feature Importance from Final Model

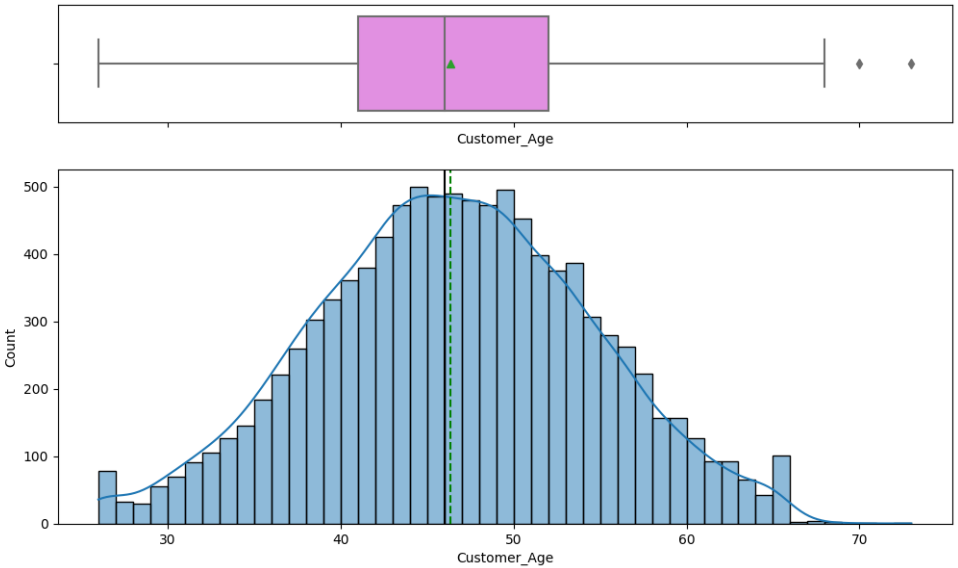


- Total_Trans_Amt, Total_Trans_Ct and Total_Trans_Ct are the Top 3 most important features for making predictions.

APPENDIX

EDA - Univariate Data Analysis

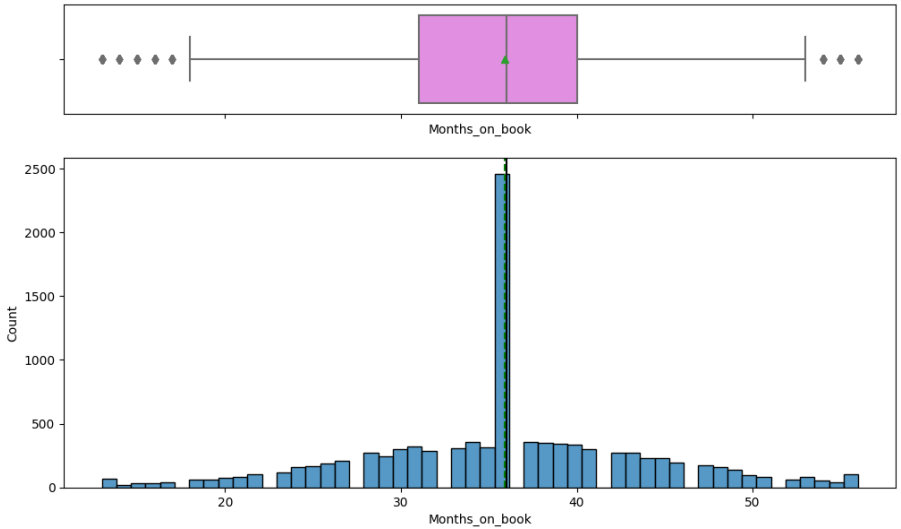
Customer's Age



Observations:

- Age has symmetrical and normal distribution, the number of customers who have accounts above the median age of 46 are slightly higher than those customers who are aged below 46

Month on Book

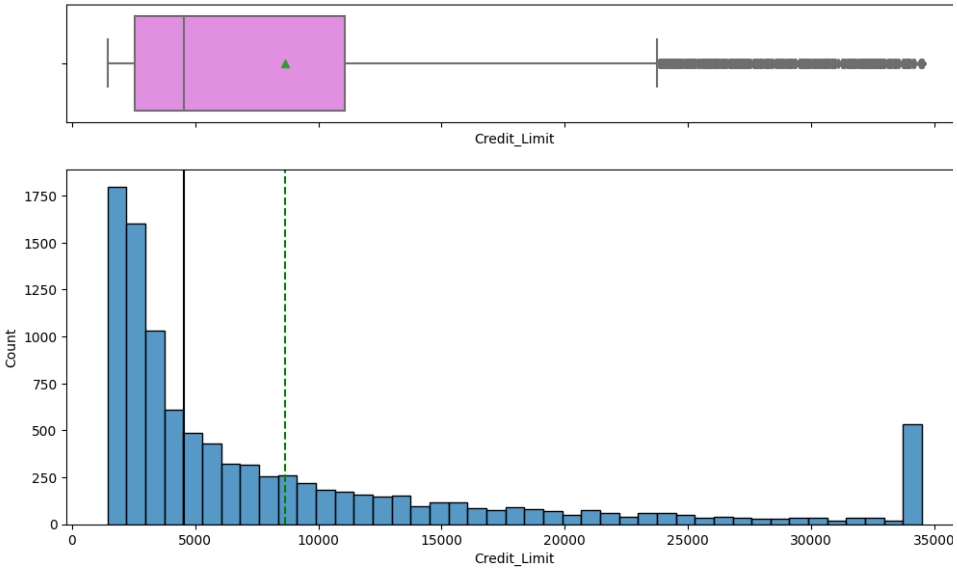


Observations:

- The Period of relationship customers with the bank have a very wide range from 13 months to 56 months with most of the customer relationship average at 36 months. There are many outliers that are less than 13 months and more than 56 months

EDA - Univariate Data Analysis

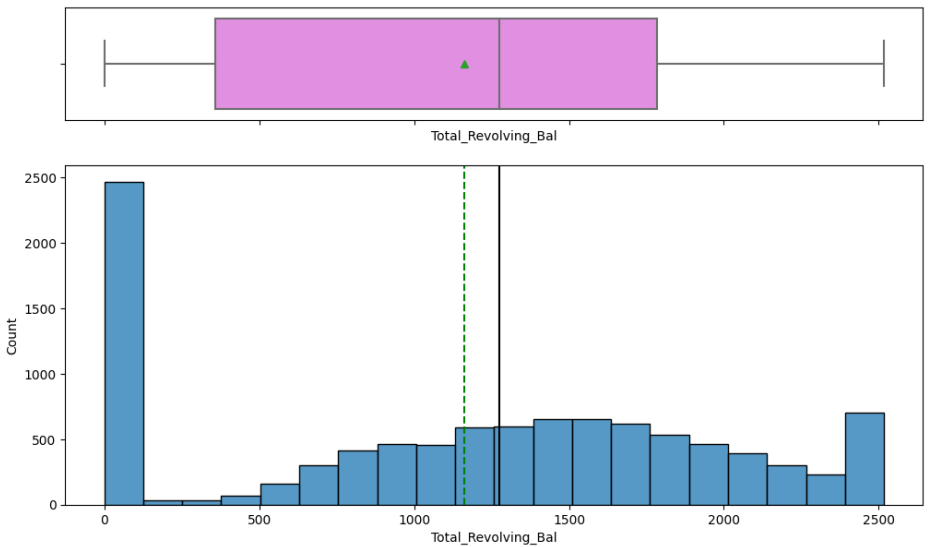
Credit Limit



Observations:

- Distribution of Credit limit is right-skewed with many outliers.
- From the boxplot, we can see many customers with credit limit that are more than ~$11000 that are consider outliers.

Total Revolving Balance

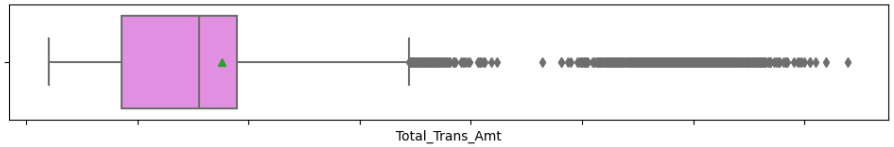


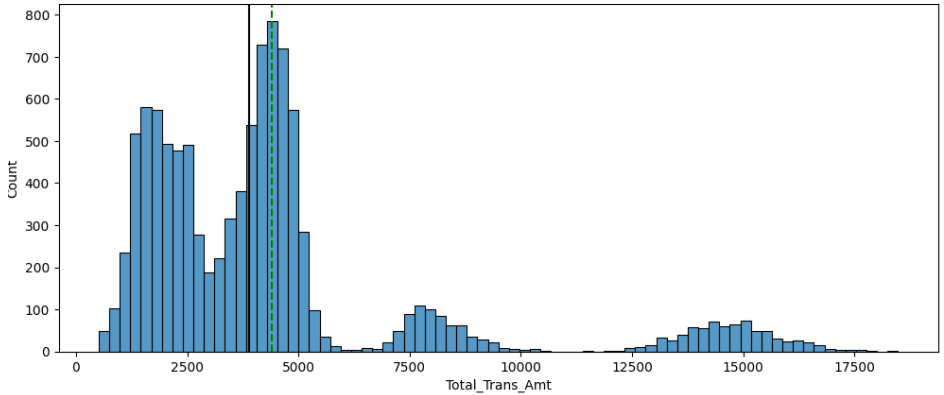
Observations:

- The distribution of the total revolving balance is slightly skewed towards the right due to high number of the customer account 2,470 have zero balance. The average revolving balance is around ~$1160

EDA - Univariate Data Analysis

Total Transaction Amount

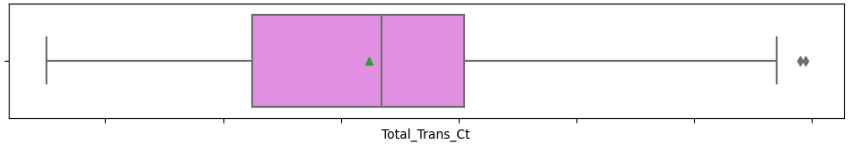


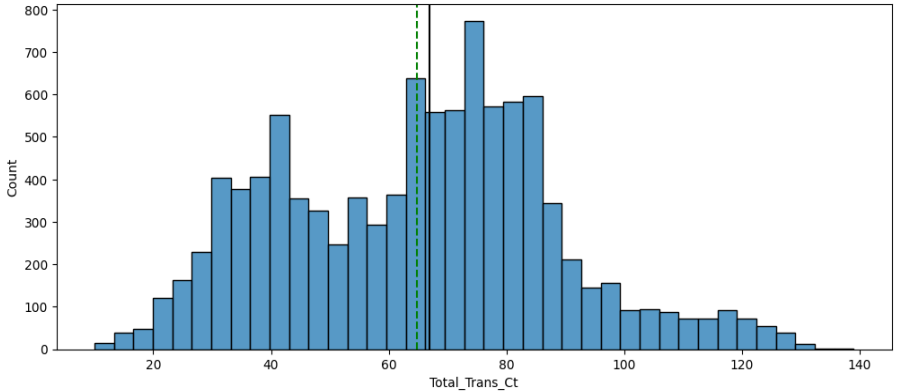


Observations:

- The distribution of the Total transaction amount is skewed towards the right.* There are many outliers in this variable and the values above $4,700 are being represented as outliers by the boxplot. a total of 2637 rows. 26% of the total records, 296 are attrite customers, 2341 are existing customers.

Total Transaction Counts



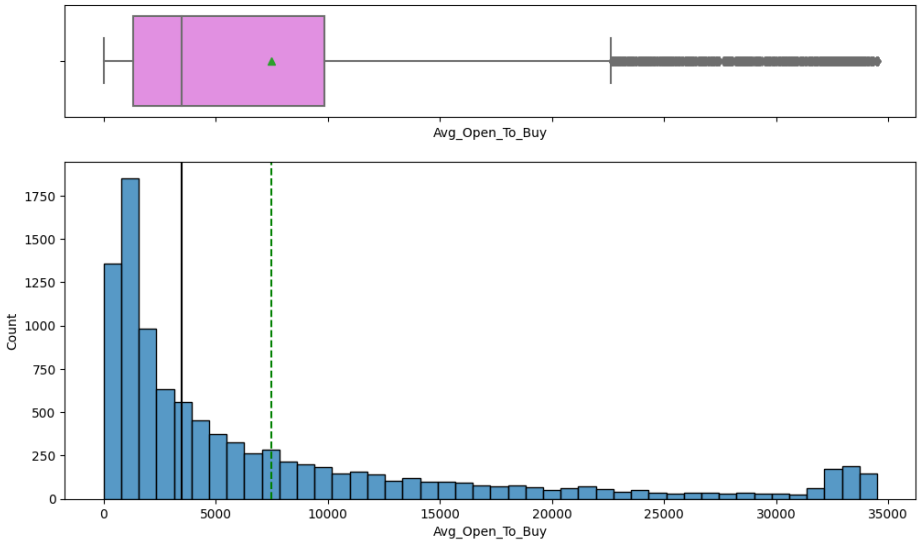


Observations:

- The distribution of this variable is slightly left skewed with an average of 65 total transaction per customer over the last 12 months. There are a few outliers of around 130 counts that might need further investigation.
- All such records with a total transaction count of more than 130 belongs to existing customers have they all have contact with the bank at least once in the last 12 months, hence they seem to represent true pattern and we don't need to treat them as outliers

EDA - Univariate Data Analysis

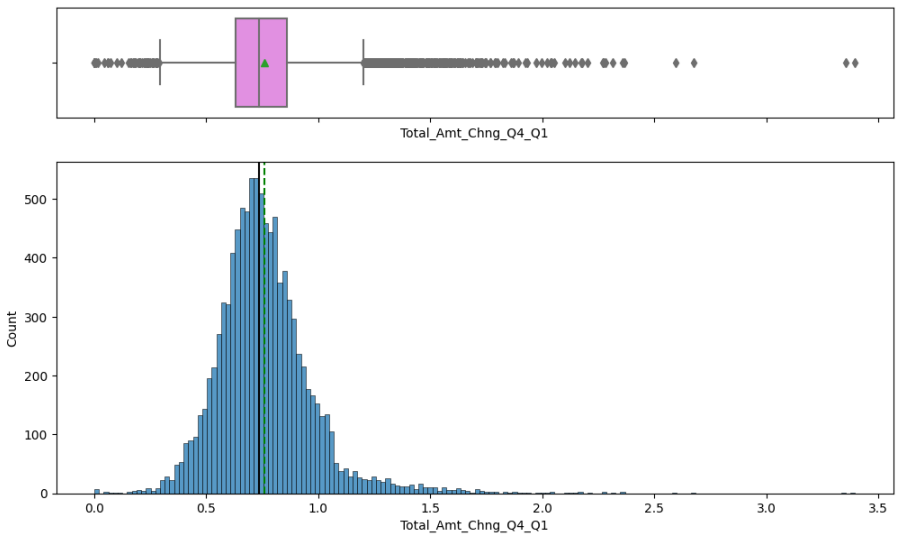
Average Open to Buy



Observations:

- This variable represent the amount left on customer's credit card to use. and this column represents the average of this value for the last 12 months. It has almost identical distribution as the variable "Credit Limit" with many outliers for the accounts that have ~9900 dollars left on their card. This pattern is consistent and seems to tell us customers will spend whatever they are allow to spend within the credit limit on the credit card.

Total Revolving Balance

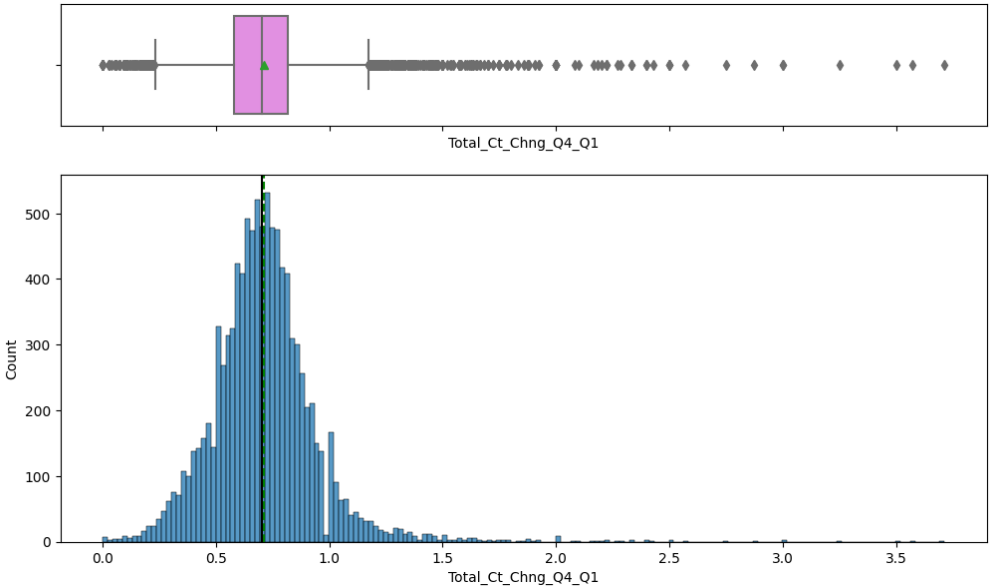


Observations:

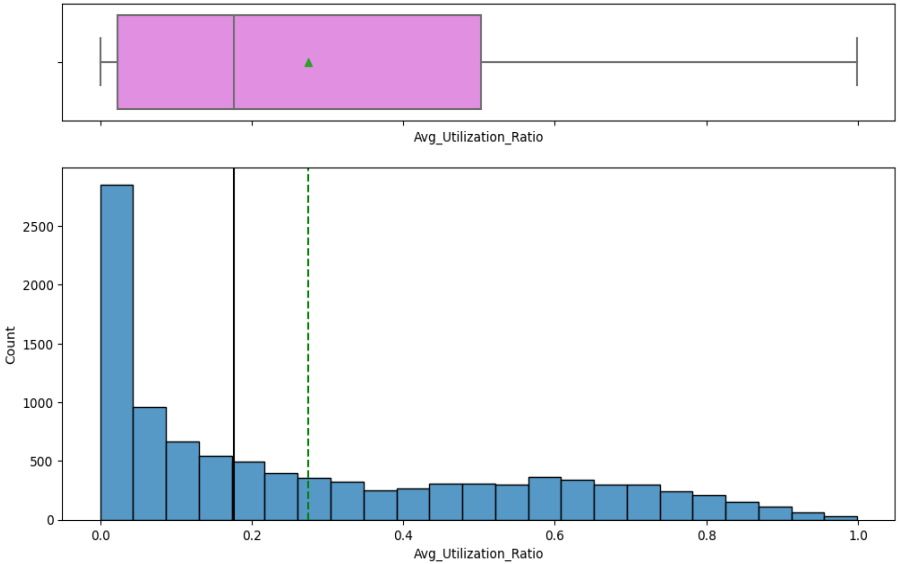
- The distribution of the total revolving balance is slightly skewed towards the right due to high number of the customer account 2,470 have zero balance.
- The average revolving balance is around ~$1160

EDA - Univariate Data Analysis

Total Count change Q4_Q1



Average Utilization Ratio



Observations:

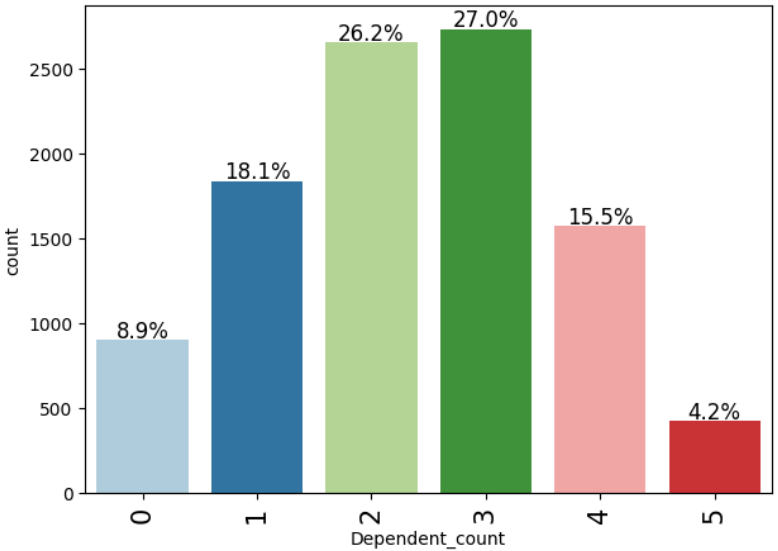
- The distribution of the Ratio of the total transaction count in 4th quarter and the total transaction amount in 1st quarter is very evenly distributed, with large number of outliers with ratio that are greater than 1.3 and smaller than 0.2. this pattern is very close to the distribution of the Total_Amt_Chng_Q4_Q1 which is a consistent reflection.

Observations:

- The distribution of the Average utilization ratio, represents how much of the available credit the customer spent, is skewed towards the right due to majority of this ratio is zero.

EDA - Univariate Data Analysis

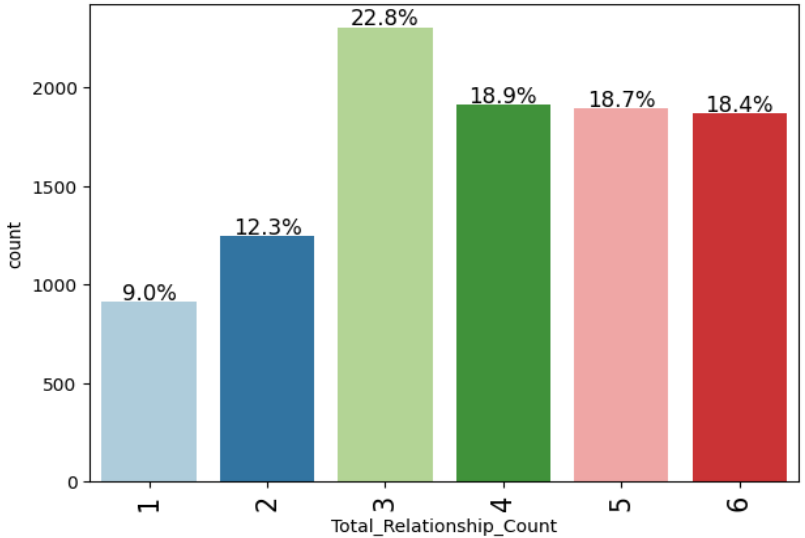
Dependent Count



Observations:

- Half of the customers base (53%) have either 2 or 3 dependents

Total Relationship Count

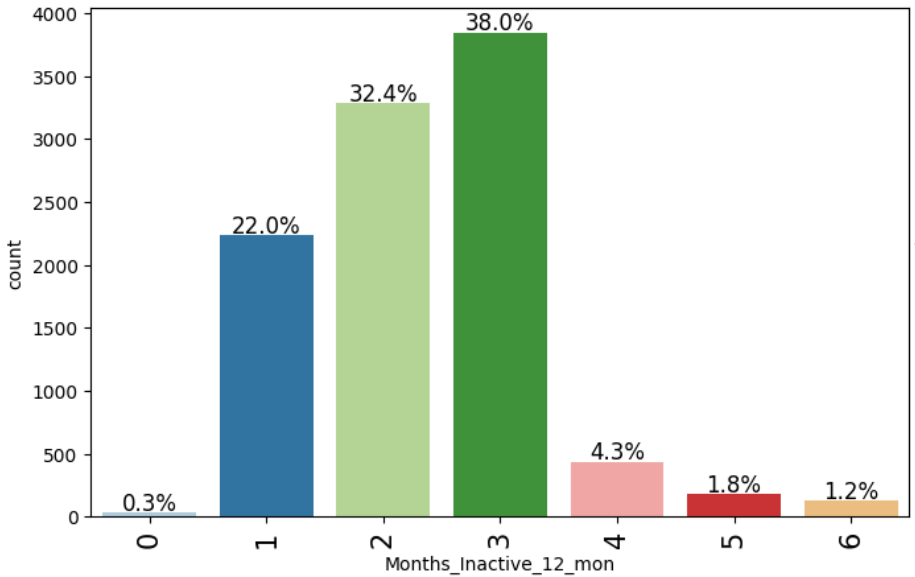


Observations:

- 22.8% of the customers have 3 products with the bank followed by customers who have 4 or 5 or products with ~18% respectively out of the total.

EDA - Univariate Data Analysis

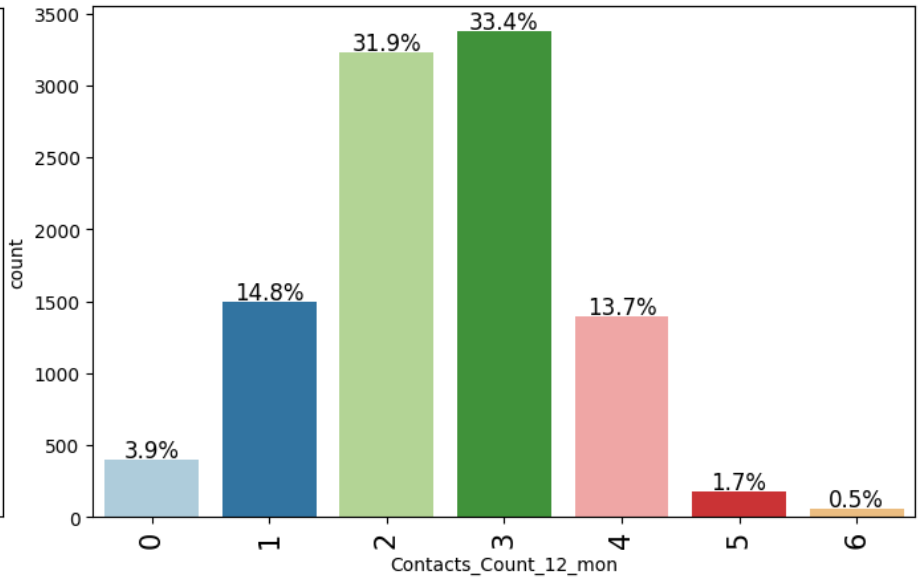
Months inactive in last 12 months



Observations:

- 38% of the customers are inactive up to 3 months, followed by 32.4% who are inactivated up to 2 months.

Contact Count in last 12 months



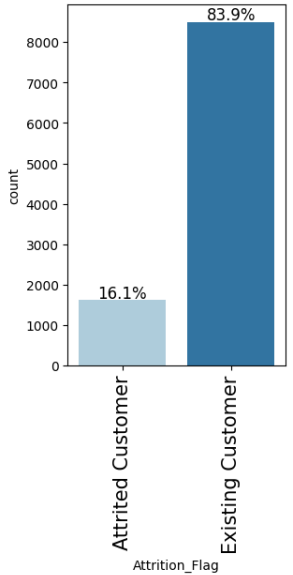
Observations:

- 33.4% of the customers contacted the bank 3 times, followed by 31.9% customers contacted the bank 2 times in the last 12 months.

EDA - Univariate Data Analysis

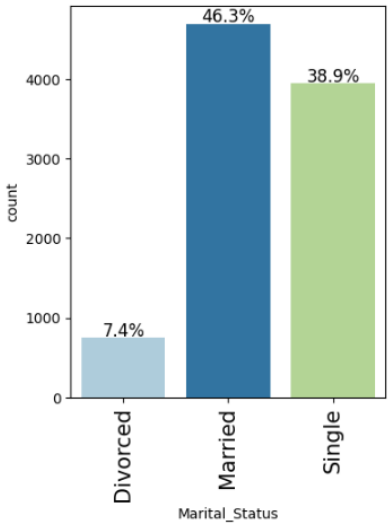
Great Learning
POWER AHEAD

Attrition Flag



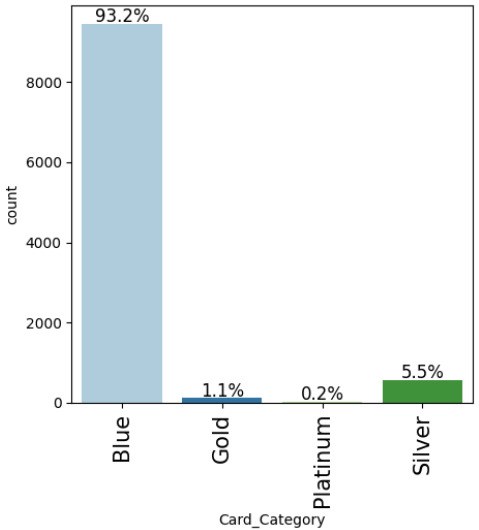
Observations:
- 83.9% are existing customer where 16.1% are attrite customers

Marital Status



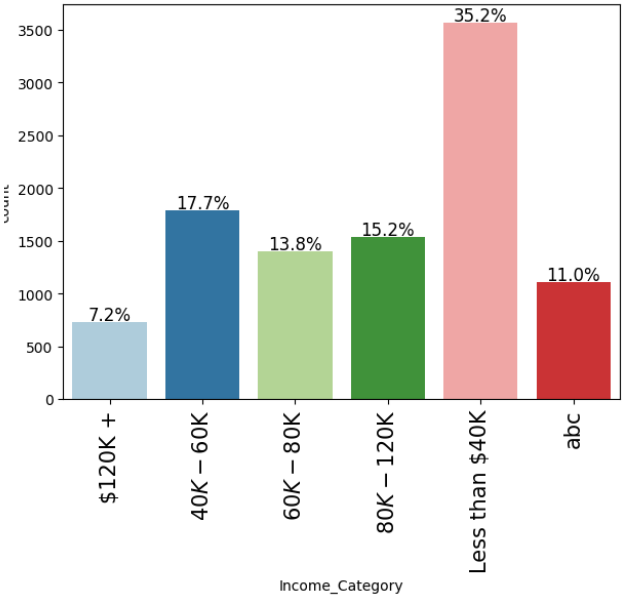
Observations:
- 46.3% of the customers are married, followed by 38.9% customers are singles and 7.4% are divorced.

Card Category



Observations:
- 93.2% of the customers have Blue type of credit card followed by 5.5% of Silver card.

Income Category

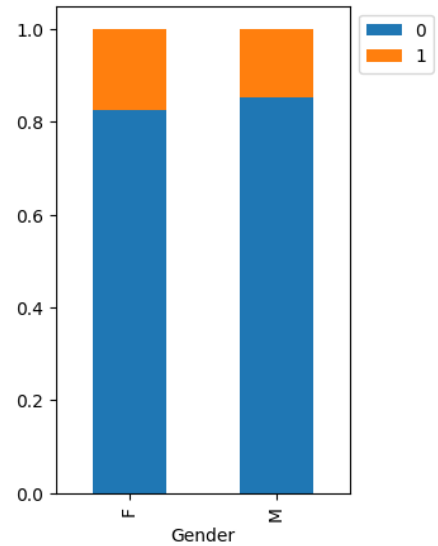


Observations:
- 35.2% of customers have income less than $40K followed by customers who have income between 40K-60K. There is 11% that have anomalous data in the dataset which will be cleaned up

EDA - Bivariate Analysis

Attrition_Flag vs Gender

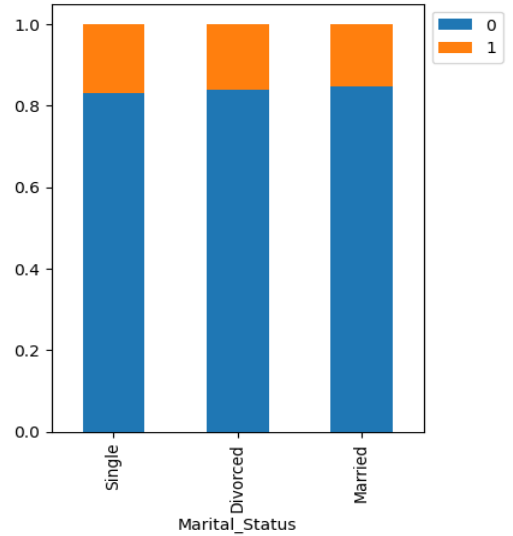
Attrition_Flag	0	1	All
Gender			
All	8500	1627	10127
F	4428	930	5358
M	4072	697	4769



Observations: 57% attrite customers are female , 43% are male

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

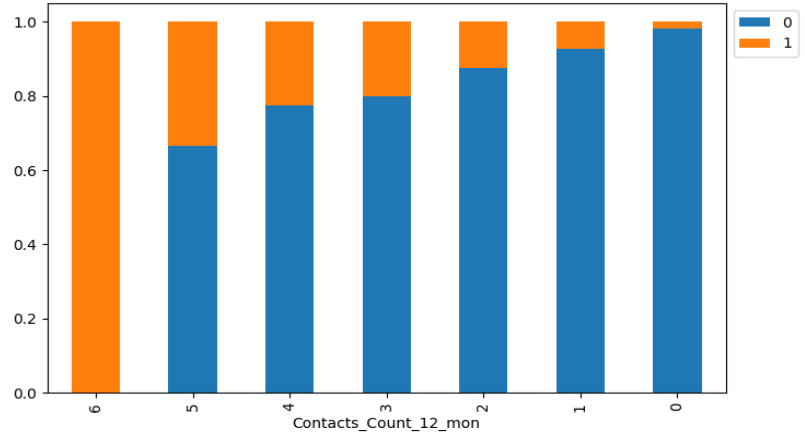


- **Observations:** The ratio of attrite customers in the three marital status are closed to the same, marital status does not seem to have an impact on the target variable

EDA - Bivariate Analysis

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

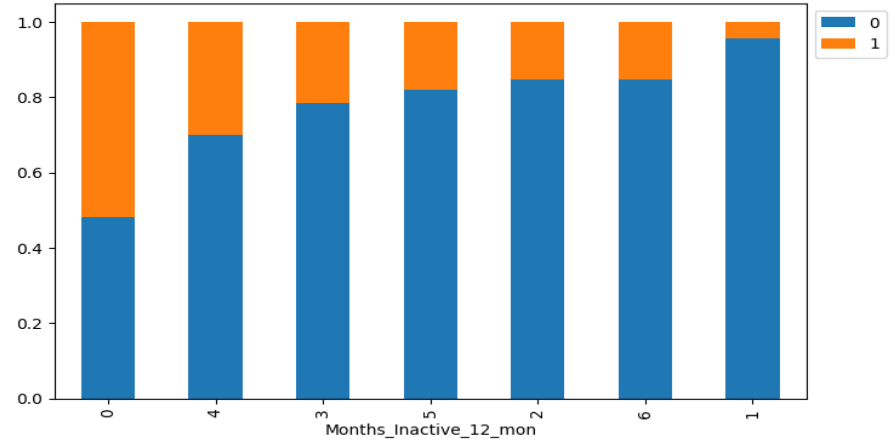


Observations:

- For the customers who have contacted the bank 6 times in the last 12 months, they are 100% attrite customers, followed by the customers who contacted 5 times, ~34% are attrite customers.
- It seems the higher the number of times the customers contacted the bank could be one of the reason the customers become attrite, we should investigate further.

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



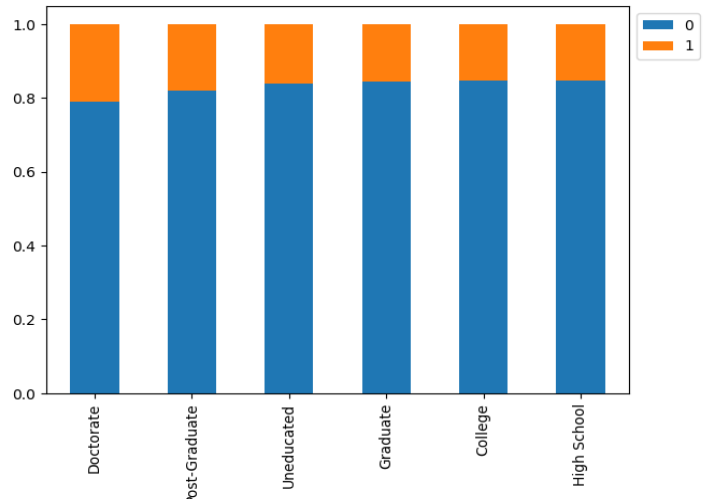
Observations:

- There are overall 16% attrite customers out of the total who are have been inactive up to 6 months.
- With the customers who are the least inactive (1 month) also consist of the least number of attrite customers.

EDA - Bivariate Analysis

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

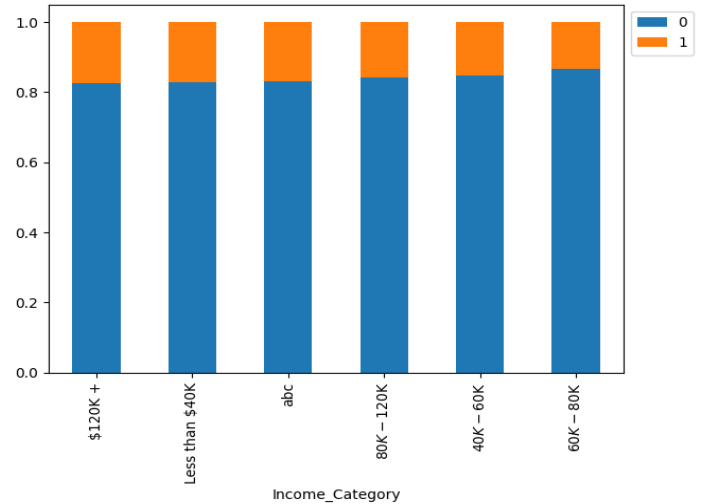


Observations:

- Within the customers population who have Doctorate degree, it consists the most attrite customers (~21%).
- ~16% of the customers in all different type of education level are attrite customers..

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



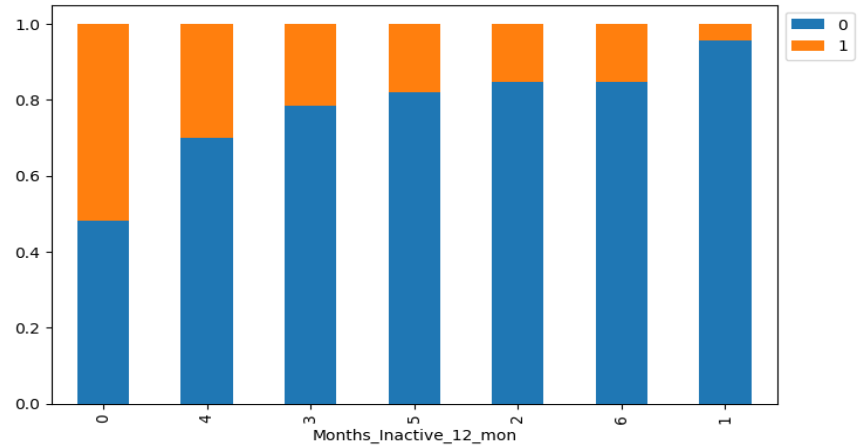
Observations:

- There are ~17% of attrite customers for most of the income category except for the 60K-80K which is ~13%.

EDA - Bivariate Analysis

Attrition_Flag vs Dependent_count

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

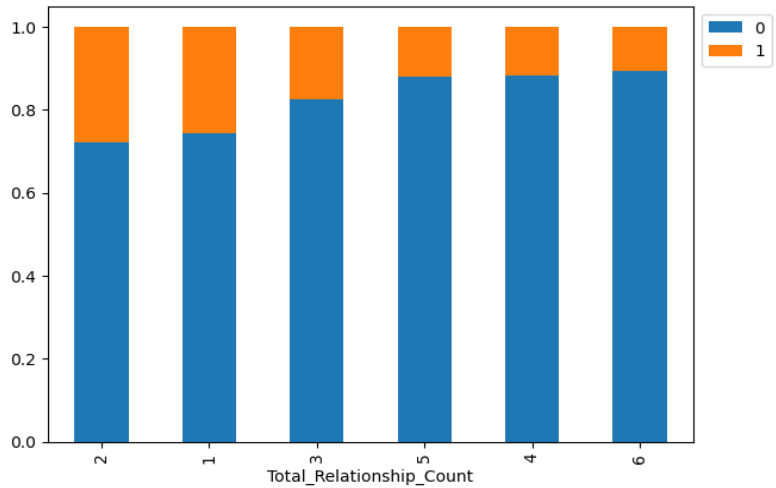


Observations:

- There are ~17% of attrite customers across all dependent count category.
- The ratio of attrite customers in the Dependent count category are closed to the same, marital status does not seem to have an impact on the target variable.

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

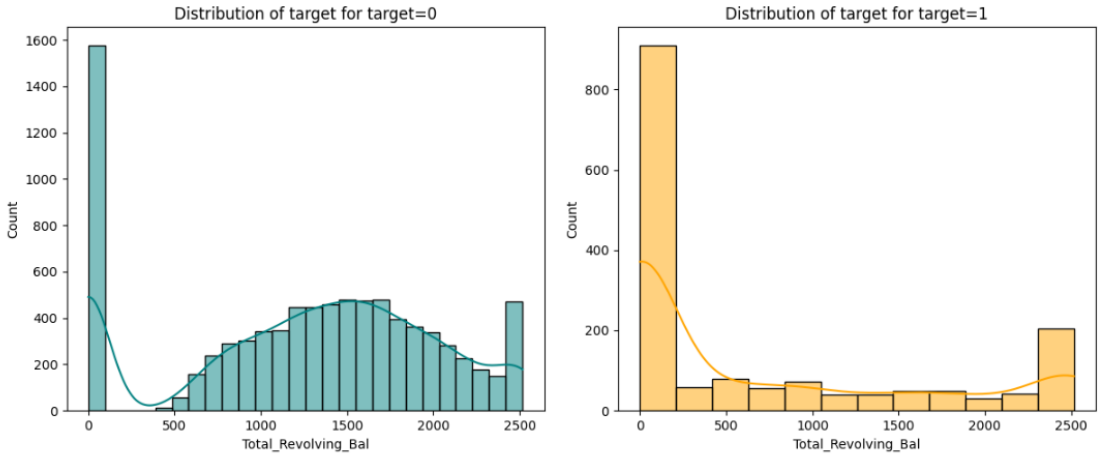


Observation:

- Customers who have 4 to 6 products with the bank has the least number of attrite customers, where the customers who have 1 or 2 products have the highest number of attrite customers.

EDA - Bivariate Analysis

Total_Revolving_Bal vs Attrition_Flag

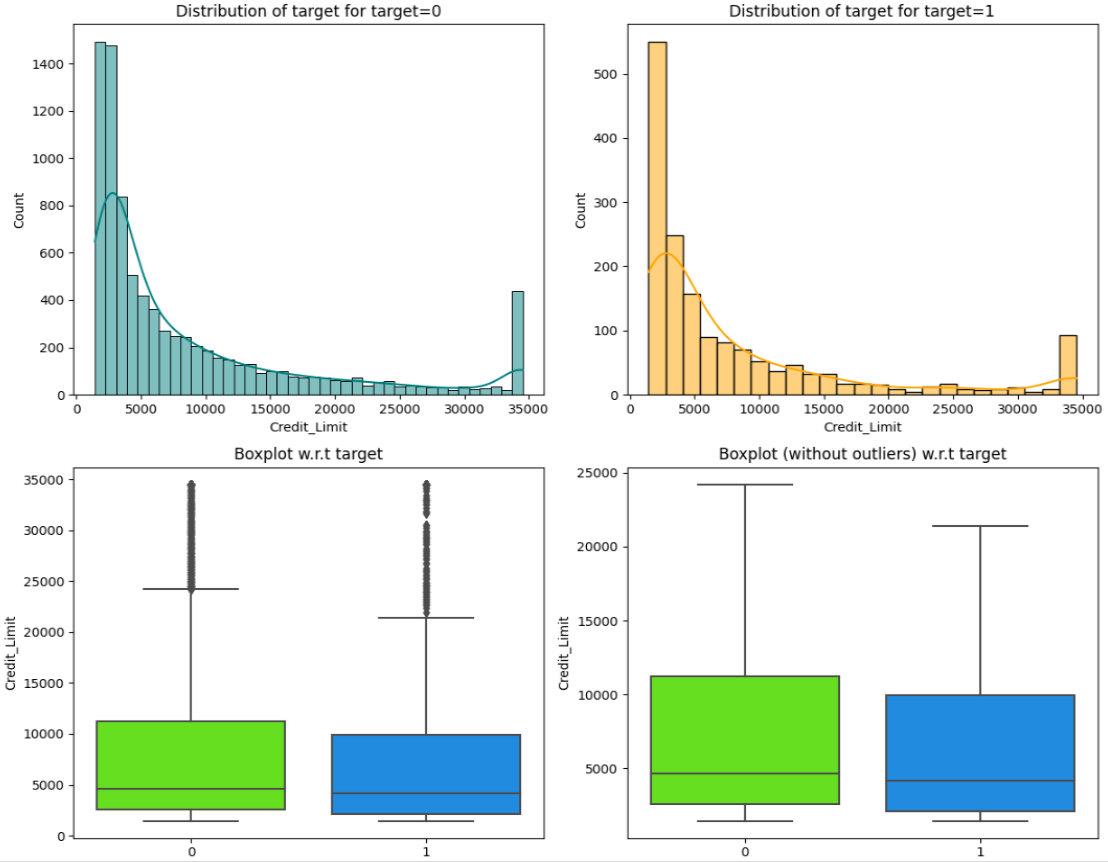


Observations:

- The distribution of the Total_Revolving_Bal is extremely skewed to the right for the group of customers, mainly due to the high number of customers who does not carry any revolving balance to next month which skewed the entire data population.
- Majority of the attrite customers do not have a revolving balance and carry a lot lower balance compared to the existing customers.

EDA - Bivariate Analysis

Attrition_Flag vs Credit_Limit

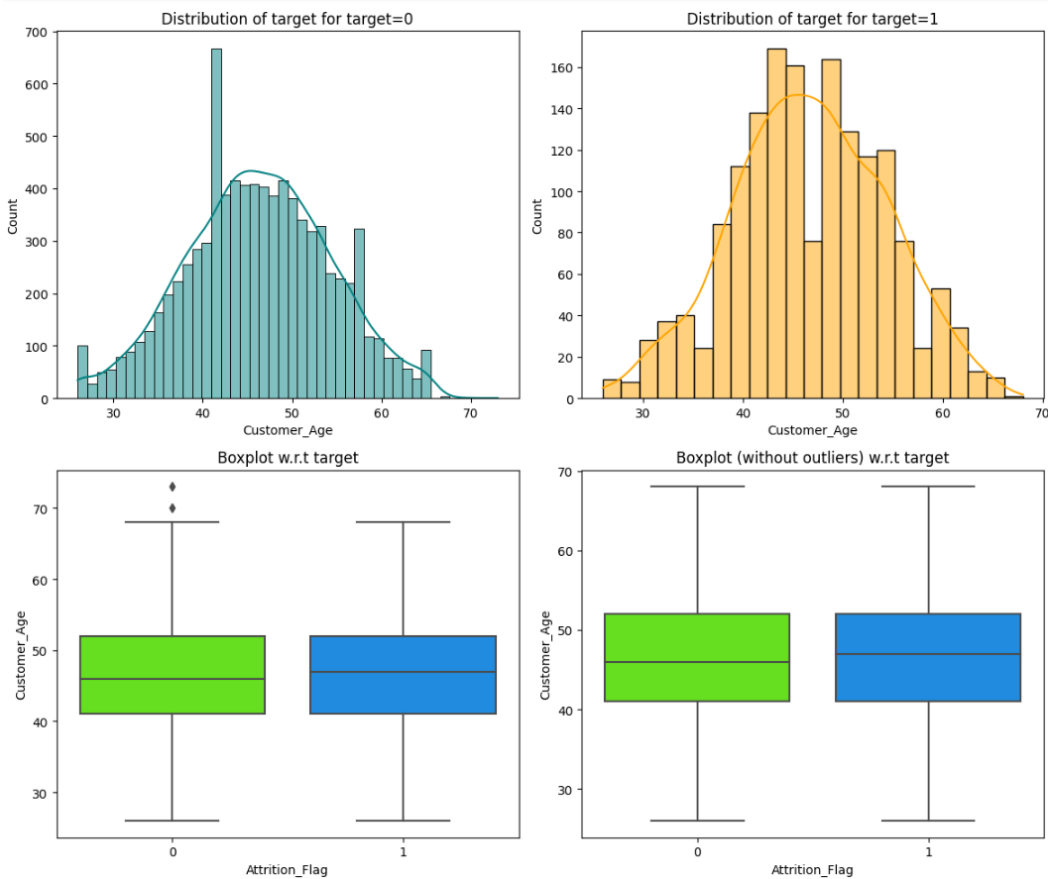


Observations:

- The distribution of the credit is extremely skewed to the right for the group of existing customers and the attrite customers.
- The distribution and the total amount of the credit limit are similar between the existing customers and the attrite customers indicates credit limit does not have much impact on the target variable.

EDA - Bivariate Analysis

Attrition_Flag vs Customer_Age

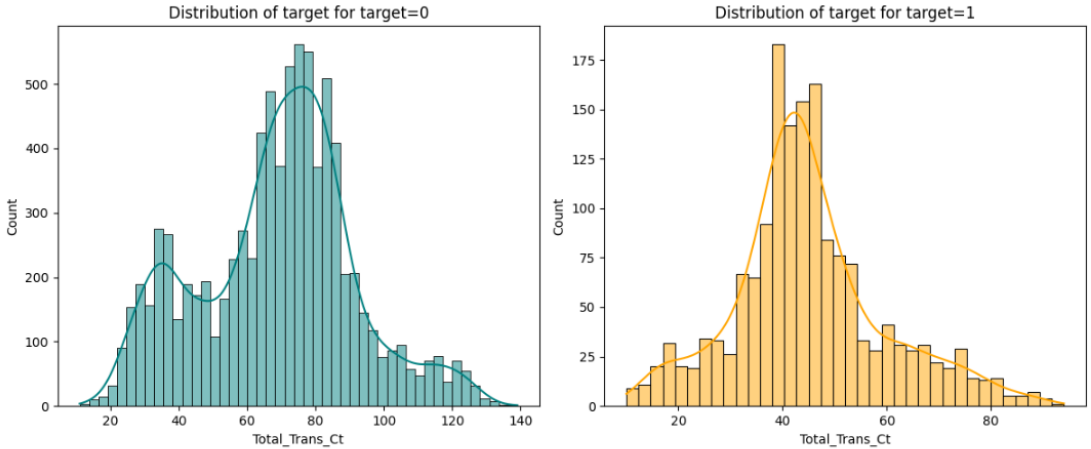


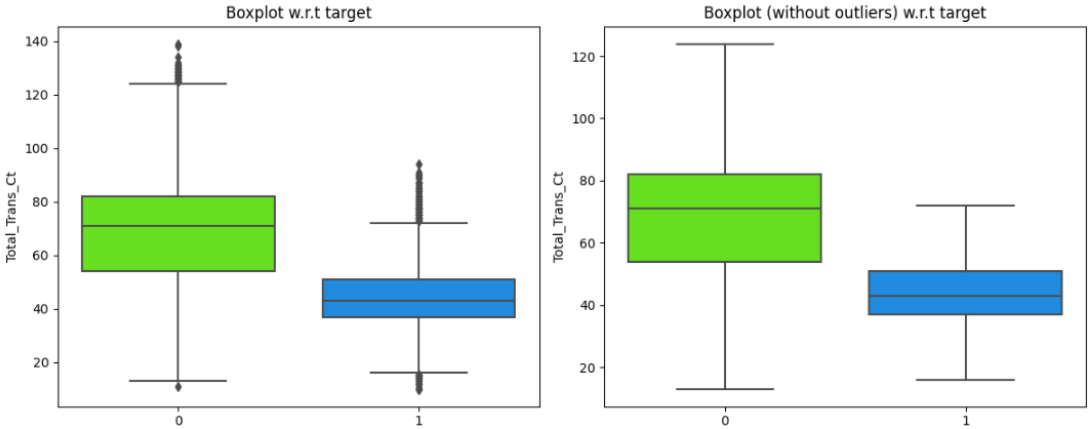
Observations:

- The distribution of the customer Age has very normal distribution between the existing customers and the attrite customers. There is not much difference in how data is dispersed between the two groups.

EDA - Bivariate Analysis

Total_Trans_Ct vs Attrition_Flag



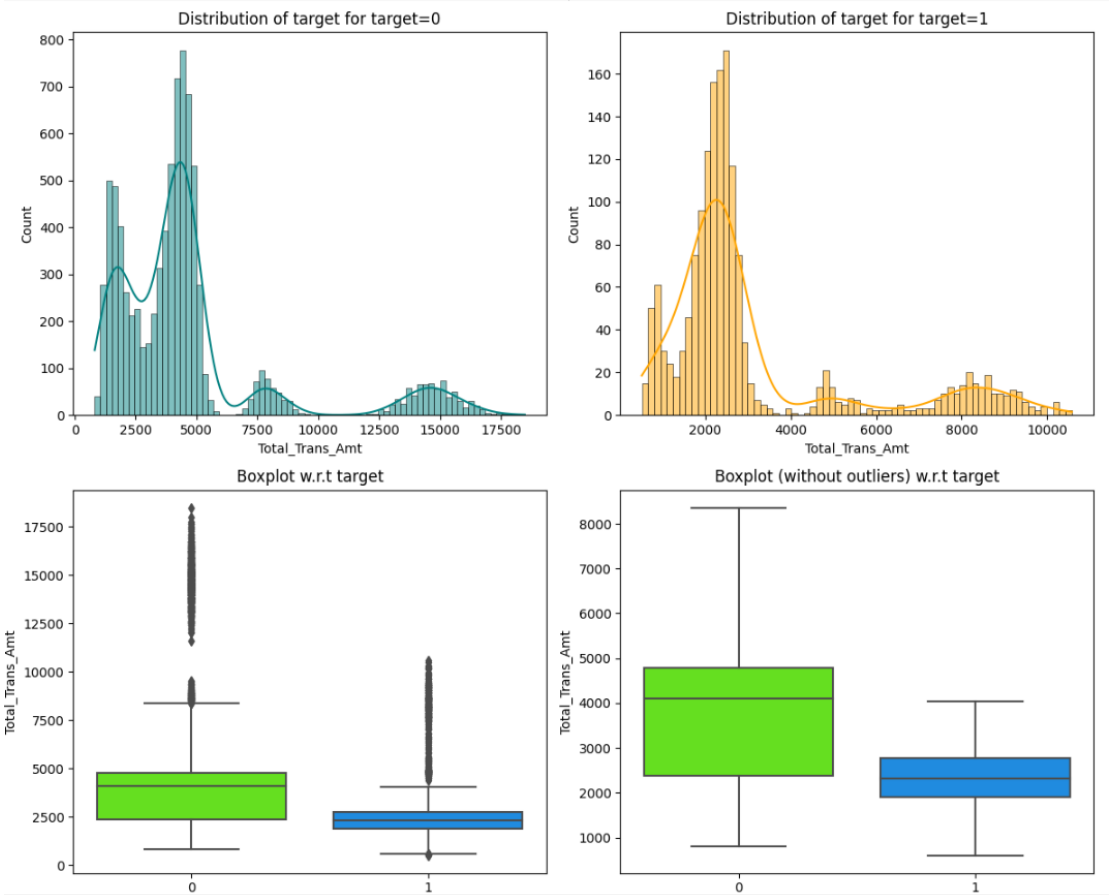


Observations:

- There's a significant difference in the total transaction counts between the existing customers and the attrite customers.
- Attrite customers have much lower total transaction counts with average between 30 to 50 vs 58 to 82 for the existing customers.

EDA - Bivariate Analysis

Total_Trans_Amt vs Attrition_Flag

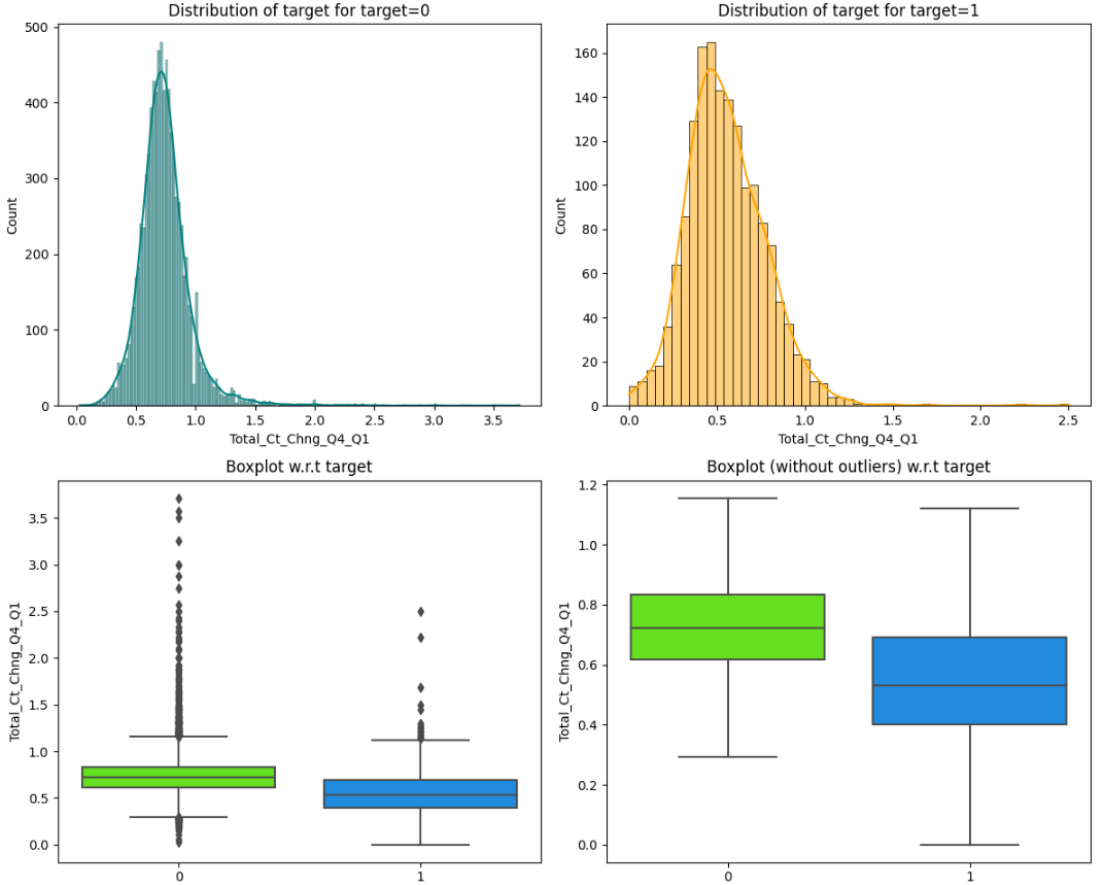


Observations:

- There's a similar difference in the total transaction amount between the existing customers and the attrite customers. Attrite customers have much lower total transaction amount with average between $1800 to $2800 vs $2500 to $4800 for the existing customers.

EDA - Bivariate Analysis

Total_Ct_Chng_Q4_Q1 vs Attrition_Flag

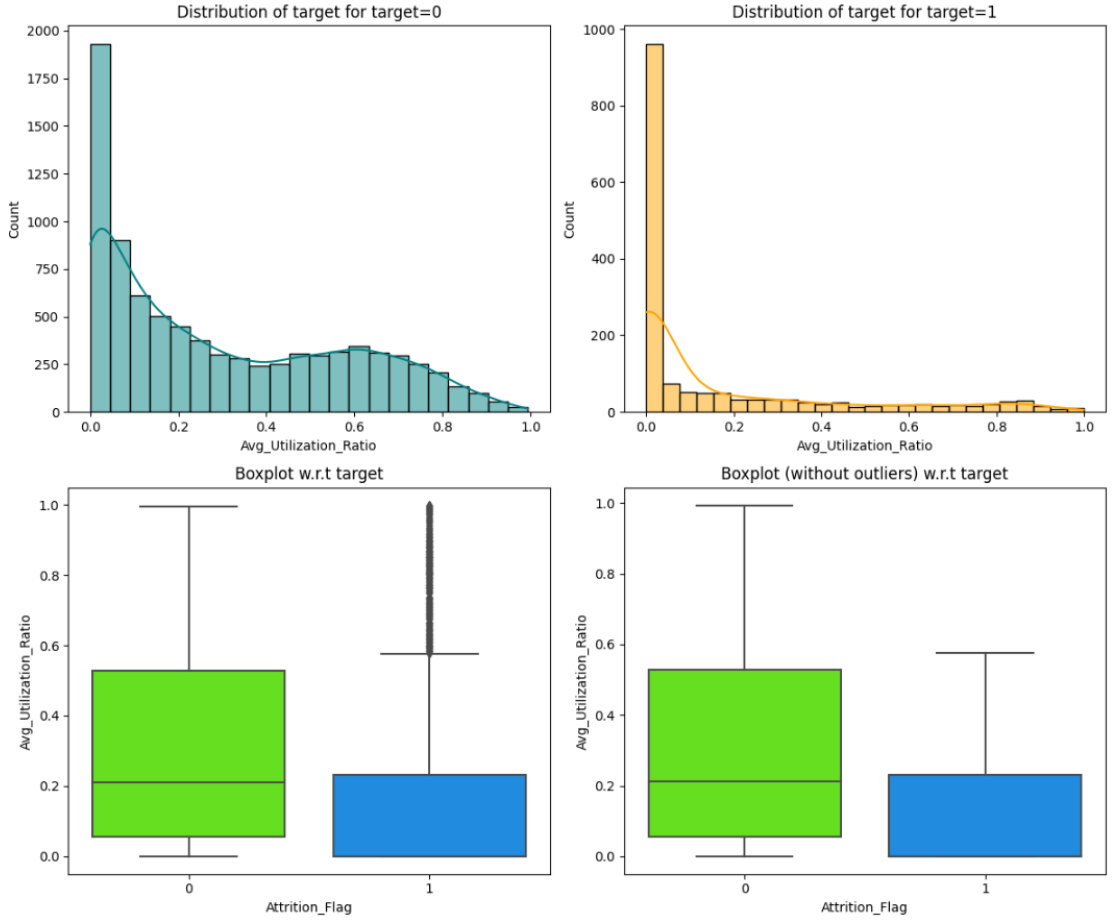


Observations:

- The distribution of Total_Ct_Chng_Q4_Q1 is quite normal between the existing customers and the attrite customers.
- The ratio is lower with the attrite customers although the range is bigger as compared to the existing customers.

EDA - Bivariate Analysis

Avg_Utilization_Ratio vs Attrition_Flag

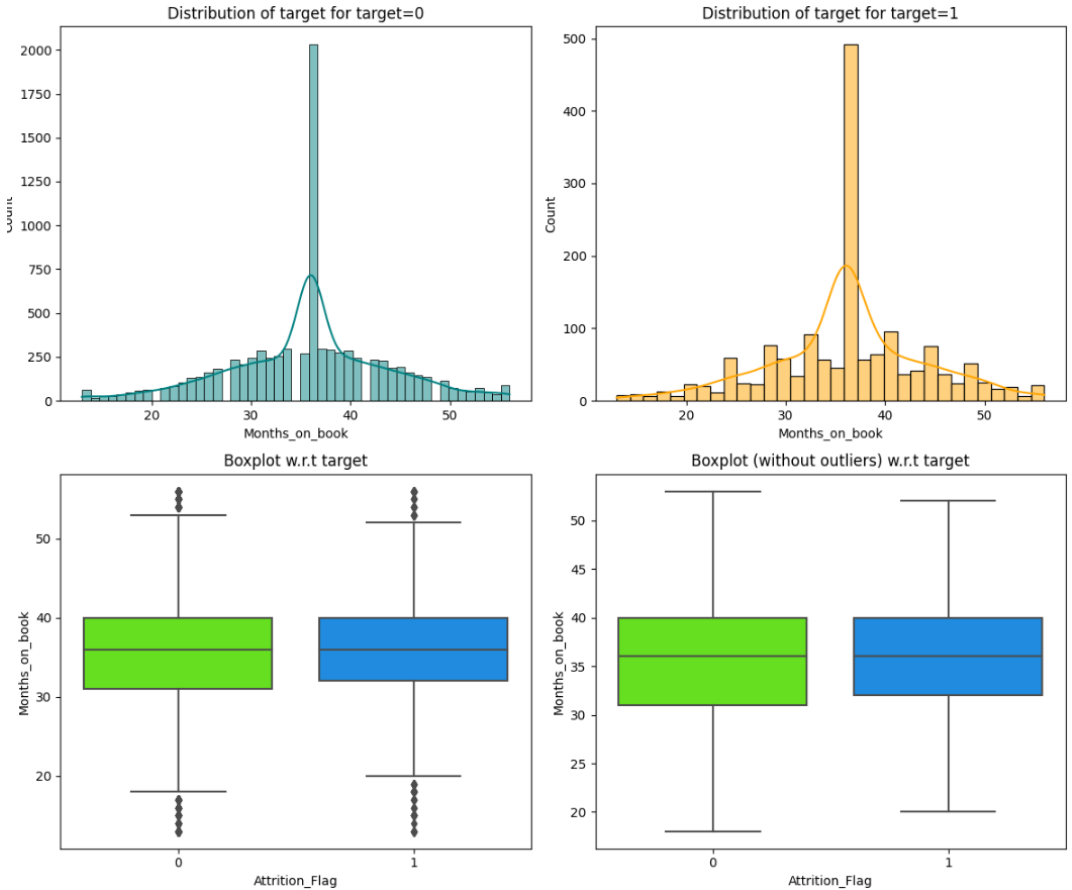


Observations:

- The distribution of the Avg_Utilization_Ratio is extremely skewed to the right for the attrite customer, mainly due to the high number of customers who does not have any credit card utilization which skewed the entire data population.
- There is much higer Avg_Utilization_Ratio with the existing customers than the attrite customers.

EDA - Bivariate Analysis

Attrition_Flag vs Months_on_book

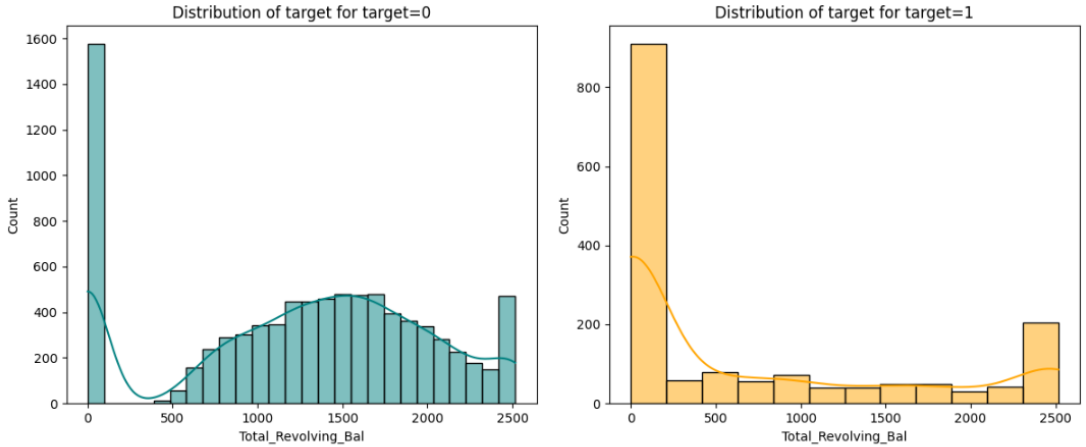


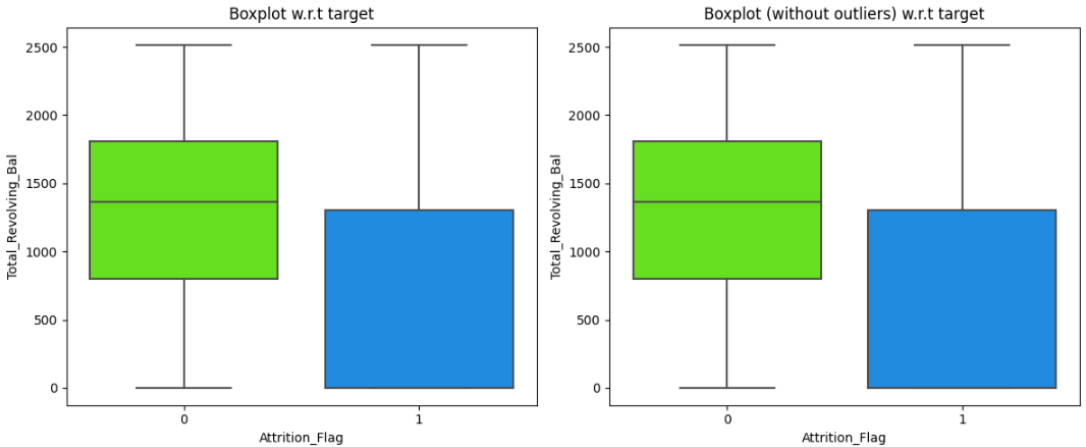
Observations:

- 50% of the customers for both existing and attrite have almost the same number of months of relationship with the bank, between 28 to 40 months.

EDA - Bivariate Analysis

Attrition_Flag vs Total_Revolving_Bal



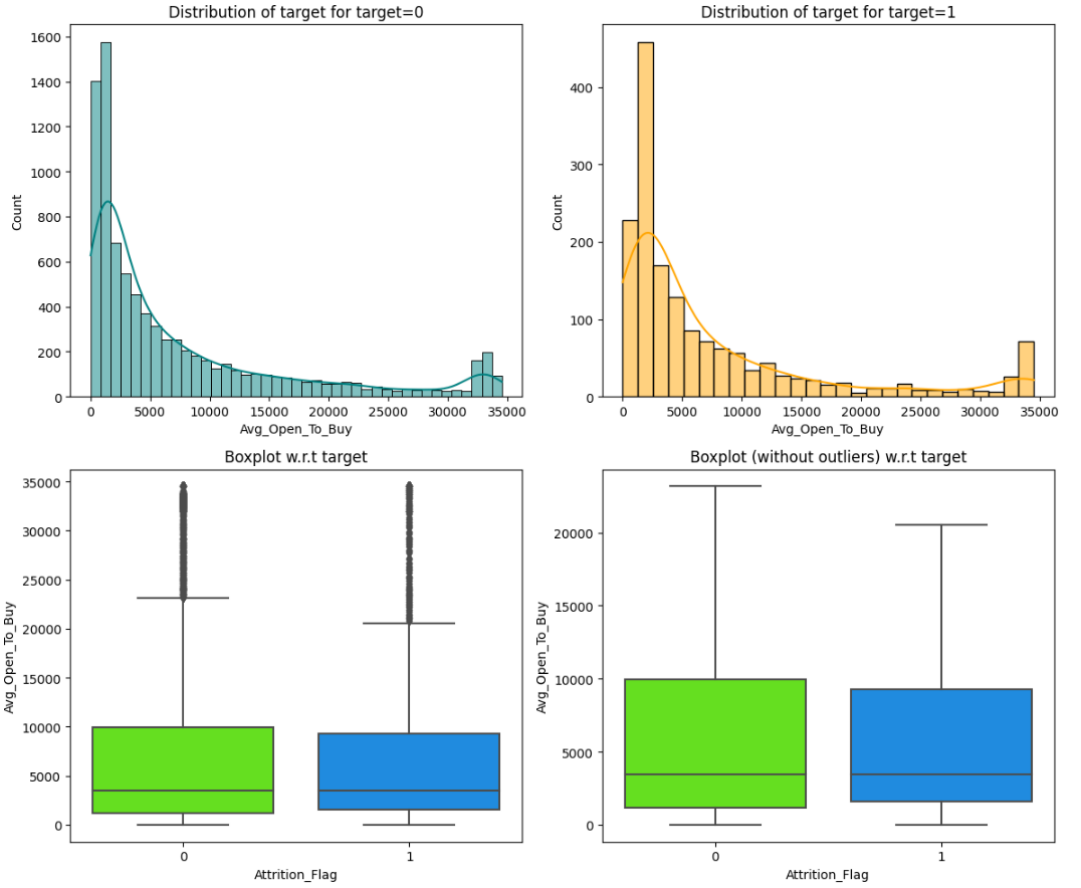


Observations:

- The distribution of the total revolving balance is skewed to the right for the two group of customers, mainly due to the high number of customers who does not have any credit card utilization which skewed the entire data population.
- The total revolving balance for the attrite customers is much lower than the existing with 60% range from 0 to 1200 dollars.
- The total revolving balance for the existing customers is higher with 50% range from 800 to 1800 dollars.

EDA - Bivariate Analysis

Attrition_Flag vs Avg_Open_To_Buy



Observations:

- The distribution of the Average Open to buy is almost the same between the attrite customers and the existing customers and is very right skewed.