Thera Bank

Project : Credit Card Users Churn Prediction
Course : Supervised Learning

Document Version : 1.0
Document Owner : Rahul Kulkarni
Document ID : Project 3 - SL - Thera Bank (Credit Card Users).pdf
Submission Date : 19th August 2023

Contents

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview & Analysis
- EDA - Univariate & Bivariate Analysis
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

Executive Summary – Business Context

- **Business Context:** Credit Cards are a good source of income for the banks because of different kinds of fees charged such as the Annual Fees, Balance Transfer Fees, Cash Advance Fees, Late Payment Fees, and Foreign Transaction Fees etc. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances.

- **The Problem Statement :** Thera Bank has seen a sharp decline in the number of their Credit Card users. Customer's leaving their Credit Cards services would mean a significant loss of revenue to the bank, and hence the bank wants to identify the customers who will leave their Credit Card services and understand the key reasons for such attrition. Based on the analysis, the Bank intends to improve its Credit Card Services, that will help retain Credit Card Users.

- **Solution Approach:** In order to resolve the above problem, we will undertake the following key tasks:

  - Perform a deep-dive on the previous Bank Churners dataset using libraries such as numpy and pandas for data manipulation, and seaborn and matplotlib for data visualisation

  - Perform exploratory data analysis on the dataset to deliver key findings and insights

  - Identify key customer attributes of the dataset that are most significant in driving attrition

  - Build a classification model that will be able to predict whether a customer will leave or not

  - Identify the key services that would need improvisation order to lower customer attrition

  - Recommend opportunities for improvement that will help the bank to boost customer retention and potential acquisition

Executive Summary – Model Evaluation Criteria & Approach

- **Model Evaluation Criteria :** The primary objective for building the model is to predict whether an existing customer will attrit or not and the key reasons for leaving the Credit Card Services. Using the confusion matrix as guiding principle, it is imperative to focus on reducing the False Negatives (FN) i.e., predicting that a customer will not attrite, but eventually attrites the Credit Card Services. Losing an existing customer would be a significant loss of revenue to the Bank. So, if FN is high, that means the attrition will be high. This implies that **reducing False Negatives** should be of utmost importance to the business
  - Key Criteria – Recall: The bank should therefore use Recall as the key model evaluation criteria – higher the Recall, greater are the chances of minimising False Negatives

- **Model Building Approach:** We have split the data into Training, Validation and Testing datasets, and built 5 base models.
  - Model 1 - Bagging
  - Model 2 - Random Forest
  - Model 3 - Gradient Boosting (GBM)
  - Model 4 - AdaBoost
  - Model 5 - Decision Tree (DTree).

  Using the original data, and Over-Sampling & Under-Sampling techniques, we have evaluated the performance (Recall) of the above models on Training & Validation datasets. Based on their **Recall** scores, we have **shortlisted** the **3 best performing** models (**1-GBM using Original dataset, 2-GBM using Under-Sampled dataset and 3-AdaBoost using Original dataset**) and tuned their hyper-parameters to achieve the best performance. We have then selected the **best model (GBM using Under-Sampled dataset)** and evaluated its performance on the Testing dataset

Executive Summary – Model Performance Summary

- **Model Performance without Hyper Parameter Tuning:** Following is the performance of all models for their Recall scores on Training & Validation datasets. The model performance was optimised and evaluated using original (no-sampling), oversampled and undersampled data. Based on the Recall score, Gradient Boosting (GBM) had the best performance followed by AdaBoost

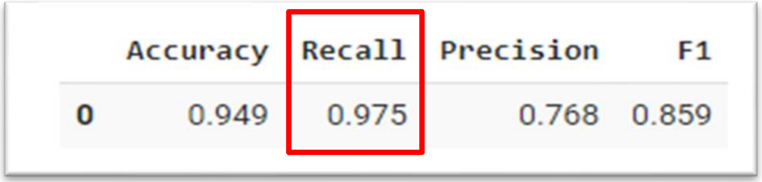
#	Model Type	Original Dataset		Oversampled Dataset		Undersampled Dataset	
		Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset
1	Model – 1: Bagging	0.9838	0.8148	0.9986	0.8765	0.9946	0.9382
2	Model – 2: Random Forest	1.0	0.7530	1.0	0.8888	1.0	0.9506
3	Model – 3: Gradient Boosting (GBM)	0.8840	0.9012	0.9810	0.9382	0.9823	0.9382
4	Model – 4: AdaBoost	0.8471	0.8641	0.9670	0.8888	0.9516	0.9506
5	Model – 5: Decision Tree (DTree)	1.0	0.8024	1.0	0.8024	1.0	0.9135

- **Model Performance - Hyper Parameter Tuned:** Optimised performance of GBM & AdaBoost on Training & Validation Datasets

#	Model Type - Hyper Tuned	Accuracy		Recall		Precision		F1	
		Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset
1	Model – 3: GBM using Original Data	0.988	0.984	0.948	0.926	0.978	0.974	0.963	0.949
2	Model – 3: GBM using Undersampled Data	0.996	0.955	0.998	0.975	0.994	0.790	0.996	0.873
3	Model – 4: AdaBoost using Original Data	0.993	0.982	0.975	0.914	0.982	0.974	0.978	0.943

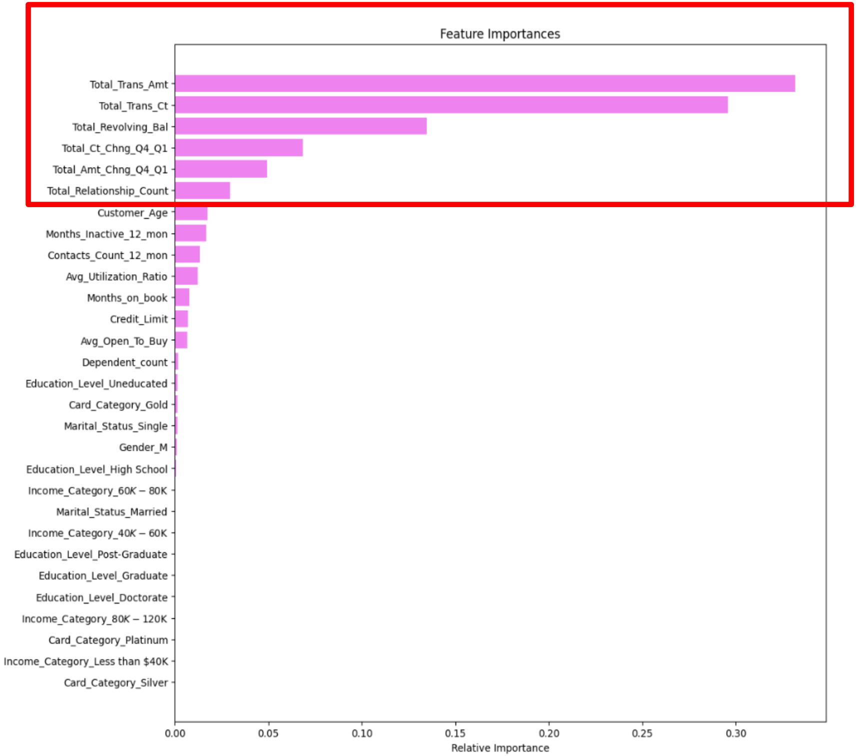
Executive Summary – Best Model Performance Summary

- **Best Model - Model – 3: GBM using Undersampled Data:** With a Recall score of 99.8% and 97.5% on Training & Validation datasets the Gradient Boosting Model using Undersampled dataset has generalised its performance and is the best model. **The model has delivered a score of 97.5% on the Testing dataset**



	Accuracy	Recall	Precision	F1
0	0.949	0.975	0.768	0.859

- **Most Important Features:** Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Total_Amt_Chng_Q4_Q1, and Total_Relationship_Count are the most important features



Executive Summary - Conclusion

- Of all the models, Model – 3: GBM using Undersampled Data (post hyper-tuning) seems to be the best fit model for Thera Bank.

- The model has high Recall score of 0.975 on the Testing dataset. This is in line with the Recall scores of 0.998 and 0.975 achieved on the Training & Validation datasets, respectively. With such high Recall scores, the model will minimise False Negatives, which is of utmost importance to the business.

- The model built can be used to predict whether a customer will attrit or not., which will help the bank to target the potential customers, who have a higher probability of attriting, and proactively incentivise them in order to maximise customer retention

- The Bank should focus on improving key services that are related to Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Total_Amt_Chng_Q4_Q1, and Total_Relationship_Count features

Executive Summary - Key Actionable Business Insights

- **Summarised Key Observations :**
  - There is a significant imbalance in dataset since there are high number of existing customers (~84%) than compared to 16% of attrited customers
  - High concentration of customers who attrited were observed with 1) A lower total transaction amount, 2) A lower total transaction count, 3) A lower utilization ratio, 4)A lower transaction count change Q4 to Q1 and 5) A significantly higher contacts with or by the bank
  - Majority of customers are in Blue card category and very few in Gold & Platinum category, which implies that customers are not using their credit card as much as possible and there are opportunities for service improvement
  - Average Open to Buy has lots of high-end outliers, which implies that there are customers who use only very small amount of their credit limit
  - Approx. 65% of customers have been contacted by the bank very often, which implies that that there have been issues where the customer needed support
  - The attrition levels of females is slightly higher than males
  - The attrition levels is high for single and divorced than compared to married customers
  - The attrition levels for advanced degrees such as Doctorate and Postgraduates is much higher than compared to other degree / non-degree educated customers

Executive Summary - Key Actionable Business Insights

- **Summarised Key Insights :**
  - Customer attributes such as Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Total_Amt_Chng_Q4_Q1, and Total_Relationship_Count are the most important features in predicting potential customers that will attrit. These features are negatively correlated with the Attrition_Flag, which implies that lower the values of these features, the higher the chances of a customer to attrite
  - Bank should increase the frequency of customer contact and provide them with various incentives and schemes to increase relationships of the customer with the bank
  - Bank should incentivise customers with cashback schemes on using credit cards, which might encourage customers on using their credit cards more often
  - Bank should also increase the credit limit for customers who regularly us their credit cards, which will lead to an increase in credit card spend / transaction amounts
  - The banks could potentially introduce an annual 0% interest free offer on large products, which will encourage customers to buy high value items using their credit cards. The payments towards this can be made on a monthly basis using a credit card. This would increase the total transaction amount, transaction counts and the revolving balance.

Executive Summary - Our Recommendation

Based on our key observations and insights, we recommend the following areas of improvement / opportunities that will drive business growth and lead to a better customer experience

- **Implement Customer Incentivisation Scheme:** Incentivising customers by offering them cashback schemes and discounts / vouchers on credit card purchases will encourage frequent spending and will drive customer gowth and increase revenue

- **Implement Tier based Rewards:** The Bank should introduce a Tier based Loyalty & Rewards Scheme for credit card purchases. Cumulative loyalty points above a certain threshold will promote the customer to a new tier, that will offer specific rewards such as First-Class Lounge access at Airports, Spa & Well-Being discounts etc

- **Implement Customer Satisfaction Survey:** The Bank should initiate a targeted Customer Satisfaction Survey to understand customer pain points and implement the findings to improve retention ratio of such customers

Business Problem Overview
& Solution Approach

Business Problem Overview and Solution Approach

- **Business Context:** Credit Cards are a good source of income for the banks because of different kinds of fees charged such as the Annual Fees, Balance Transfer Fees, Cash Advance Fees, Late Payment Fees, and Foreign Transaction Fees etc. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances.

- **The Problem Statement :** Thera Bank has seen a sharp decline in the number of their Credit Card users. Customer's leaving their Credit Cards services would mean a significant loss of revenue to the bank, and hence the bank wants to identify the customers who will leave their Credit Card services and understand the key reasons for such attrition. Based on the analysis, the Bank intends to improve its Credit Card Services, that will help retain Credit Card Users.

- **Solution Approach:** In order to resolve the above problem, we will undertake the following key tasks:

  - Perform a deep-dive on the previous Bank Churners dataset using libraries such as numpy and pandas for data manipulation, and seaborn and matplotlib for data visualisation

  - Perform exploratory data analysis on the dataset to deliver key findings and insights

  - Identify key customer attributes of the dataset that are most significant in driving attrition

  - Build a classification model that will be able to predict whether a customer will leave or not

  - Identify the key services that would need improvisation order to lower customer attrition

  - Recommend opportunities for improvement that will help the bank to boost customer retention and potential acquisition

Data Overview & Analysis

Data Overview & Analysis

- The Bank Churners dataset has the following Data-Structure:

#	Columns	Data-type	Total Rows	Description
1	CLIENTNUM	Integer 64	10127	Client number. Unique identifier for the customer holding the account
2	Attrition_Flag	Integer 64	10127	Internal event (customer activity) variable - if the account is closed then "Attrited Customer" else "Existing Customer"
3	Customer_Age	Integer 64	10127	Age in Years
4	Gender	object	10127	Gender of the account holder
5	Dependent_count	Integer 64	10127	Number of dependents
6	Education_Level	object	8608	Educational Qualification of the account holder - Graduate, High School, Unknown, Uneducated, College(refers to a college student), Post-Graduate, Doctorate.
7	Marital_Status	object	9378	Marital Status of the account holder
8	Income_Category	object	9015	Annual Income Category of the account holder
9	Card_Category	object	10127	Type of Card
10	Months_on_book	Integer 64	10127	Period of relationship with the bank
11	Total_Relationship_Count	Integer 64	10127	Total no. of products held by the customer
12	Months_Inactive_12_mon	Integer 64	10127	No. of months inactive in the last 12 months
13	Contacts_Count_12_mon	Integer 64	10127	No. of Contacts between the customer and bank in the last 12 months
14	Credit_Limit	Float 64	10127	Credit Limit on the Credit Card
15	Total_Revolving_Bal	Integer 64	10127	The balance that carries over from one month to the next is the revolving balance

Data Overview & Analysis (Cont’d)

- The Bank Churners dataset has the following Data-Structure:

#	Columns	Data-type	Total Rows	Description
16	Avg_Open_To_Buy	Float 64	10127	Open to Buy refers to the amount left on the credit card to use (Average of last 12 months)
17	Total_Amt_Chng_Q4_Q1	Float 64	10127	Ratio of the total transaction amount in 4th quarter and the total transaction amount in 1st quarter
18	Total_Trans_Amt	Integer 64	10127	Total Transaction Amount (Last 12 months)
19	Total_Trans_Ct	Integer 64	10127	Total Transaction Count (Last 12 months)
20	Total_Ct_Chng_Q4_Q1	Float 64	10127	Ratio of the total transaction count in 4th quarter and the total transaction count in 1st quarte
21	Avg_Utilization_Ratio	Float 64	10127	Represents how much of the available credit the customer spent

- Shape of the Dataset: Total No. Of Columns - 21 | Total No. Of Rows - 10,127

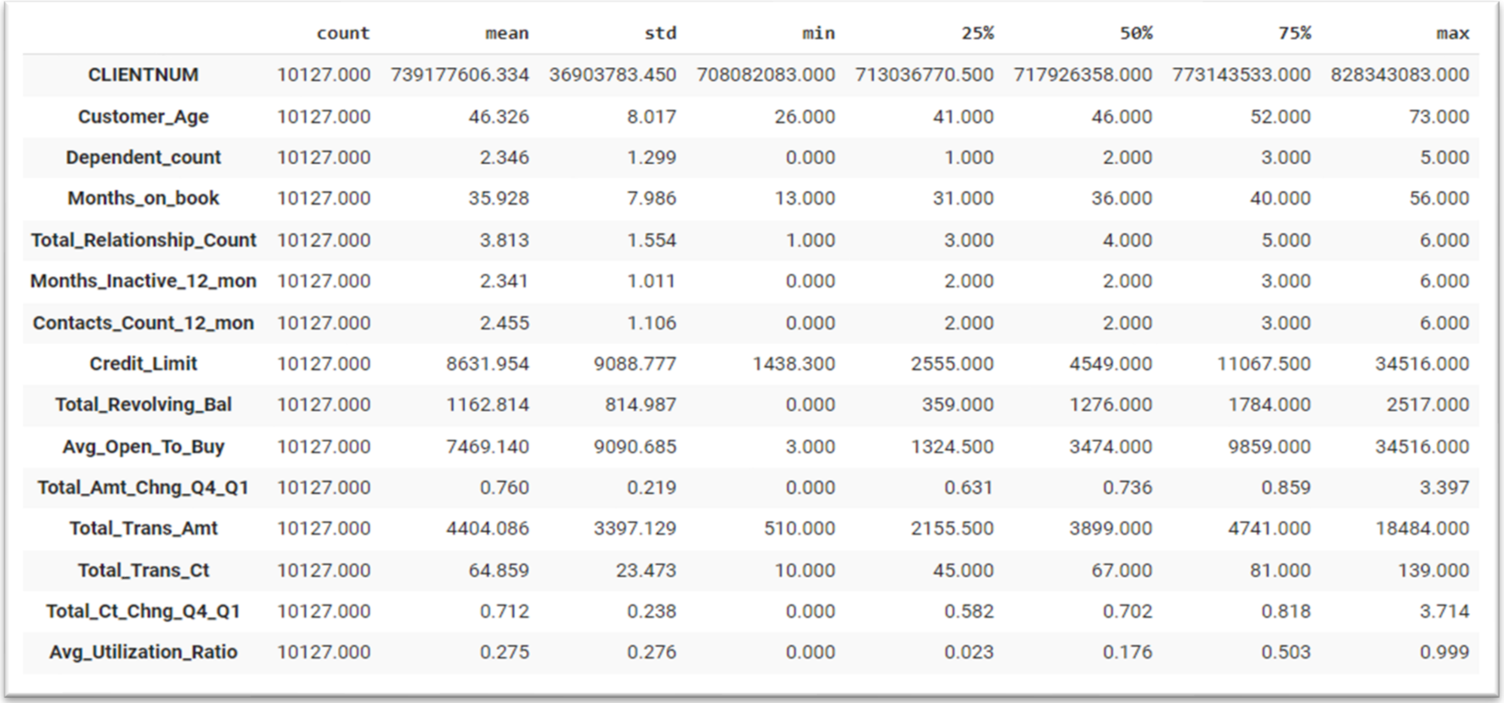
- Column Data-types: 15 of 21 columns are of numerical data types(Float64(5) & Integer64(10)) , and 6 are Object data types

- Missing & Junk Values: All columns, except 2 (Education_Level and Marital_Status) have 10,127 non-null values. Education_Level and Marital_Status have 8,608 and 9,378 records respectively. There are 1,519 and 749 N/A and NAN missing values in Education_Level and Marital_Status respectively. Income_Category has 1,112 junk values such as “abc”

- Duplicate Values: There are no duplicate values in the dataset

Data Overview & Analysis (Cont'd)

- Statistical Summary: Following is the statistical summary of the dataset



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

Data Overview & Analysis – Key Observations & Insights

- **Key Observations:**
  - There are missing and junk values in the dataset. Education_Level & Marital_Status has circa 15% and 7.4% missing values out of the total observations, respectively. Income_Category has 11% junk values out of the total observations
  - There are no duplicate values in the dataset
  - The minimum and maximum Customer Age is 26 years and 73 years respectively, whereas the mean Age is 46.32 years
  - The minimum and maximum number of Dependents is 0 and 5 respectively, whereas the mean size is approx. 2.34
  - The minimum and maximum banking relationship (months on book) is 13 and 56 months respectively, whereas the mean period is approx. 35.92 months
  - The minimum and maximum number of products held by the customer (total relationship count) is 1 and 6 respectively, whereas the mean number of products held by the customer is approx. 3.81
  - The minimum and maximum inactivity period (Months_Inactive_12_mon) is 0 and 6 months respectively, whereas the mean inactivity period is approx. 2.34
  - The minimum and maximum contacts made with the customer over the last 12 months is 0 and 6 times respectively, whereas the mean number of contacts is approx. 2.45 times

Data Overview & Analysis – Key Observations & Insights

- **Key Observations (Cont'd):**
  - The minimum and maximum credit limit is USD $1,438.3 and USD $34,516.0 respectively, whereas the mean is approx. USD $8,631.95
  - The minimum and maximum total revolving balance is USD $0 and USD $2,517.0 respectively, whereas the mean is approx. USD $1,162.81
  - The minimum and maximum open to buy credit line is USD $3 and USD $ 34,516.0 respectively, whereas the mean is approx. USD $7,469.14
  - The minimum and maximum total transaction amount is USD $510.0 and USD $ 18,484.0 respectively, whereas the mean is approx. USD $4,404.08
  - The minimum and maximum total number of transactions is 10 and 139 respectively, whereas the mean number of transactions is approx. 139
  - The minimum and maximum average utilization ratio is 0 and 0.99 respectively, whereas the mean is approx. 0.27

Data Overview & Analysis – Key Observations & Insights

- **Key Insights:**
  - Out of 10,127 records, 399 customers haven't been contacted at all. 1499 customer have been contacted only once, 3,227 have been contacted twice, 3380 have been contacted thrice, 1,392 have been contacted 4 times, 176 have been contacted 5 times and 54 have been contacted 6 times
  - Circa. 83.9% of the customers are existing customers (8500 of 10,127) than compared to 16.1% (16270 of 10,127) of attrited customers
  - Circa. 52.9% of the customers (5358 of 10,127) are females compared to 47% (4769 of 10,127) of male customers
  - There are higher number of customers with a Graduate degree that compared to those with other qualifications. 36.3% (3128 of 10,127) of customers are Graduates, followed by 23.4% who have attended only High School. 17.3% of customers are Uneducated and 11.8% of customers attended only college. 6% of customers had a Post-Graduate degree whereas 5.2% of customers held Doctorate degrees
  - There are higher number of customers who are married than compared to single and divorced. 50% (4687 of 10,127) customers are married than compared to 42.04% (3943 of 10,127) of customers who are single and 8% (748 of 10,127) of customers who are divorced

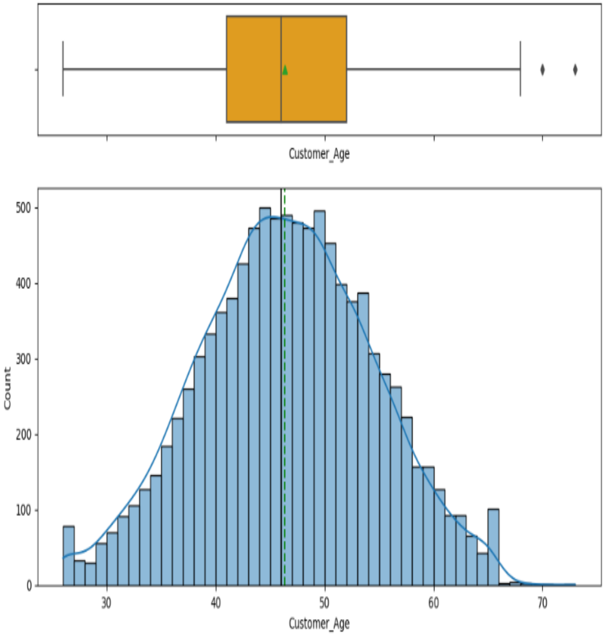
Data Overview & Analysis – Key Observations & Insights

- **Key Insights (Cont'd):**
  - There are higher number of customers with an income of less than $40K. 35.16% (3561 of 10,127) customers have income less than $40K, followed by 17.67% (1790 of 10,127) have income between $40K-$60K. 13.84% (1402 of 10,127) customers have an income between $60K-$80K, whereas 15.15% (1535 of 10,127) have an income between $80K-$120K. Only 7.17% (727 of 10,127) customers have an income of above $120K.
  - There are significantly larger number of customers in the Blue category than compared to other categories. 93.17% (9436 of 10,127) of customers lie in the Blue category. 5.48% (555 of 10,127) of customers lie in the Silver category, followed by 1.14% (116 of 10,127) of customers in the Gold and 0.19% (20 of 10,127) of customers in the Platinum category

EDA - Univariate Analysis

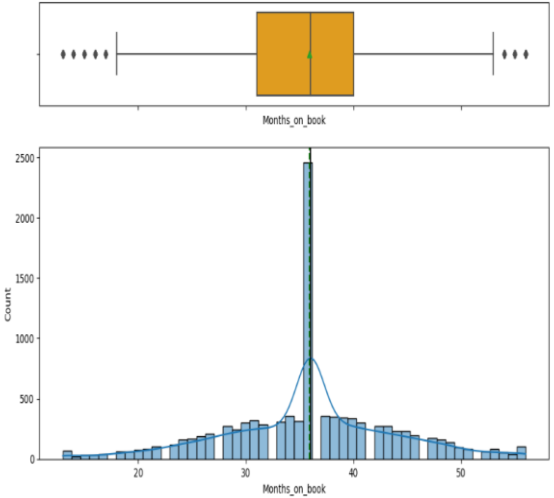
EDA - Univariate Analysis – Customer Age

- **Age:**
  - Highest Age Group: There are circa 500 customers 44 years of age
  - Minimum: The minimum age is 26 years
  - Q1: 25% of the population are less 41 years of age
  - Q3: 75% of the population are less 52 years of age
  - Maximum: The maximum age is 68 years
  - Median & Mean: The median and the mean age is approx. 46 years of age
  - Outliers: There are 2 outliers of 70 and 73 years
  - Skewness: From the plot, it can be observed that the graph has a normal distribution



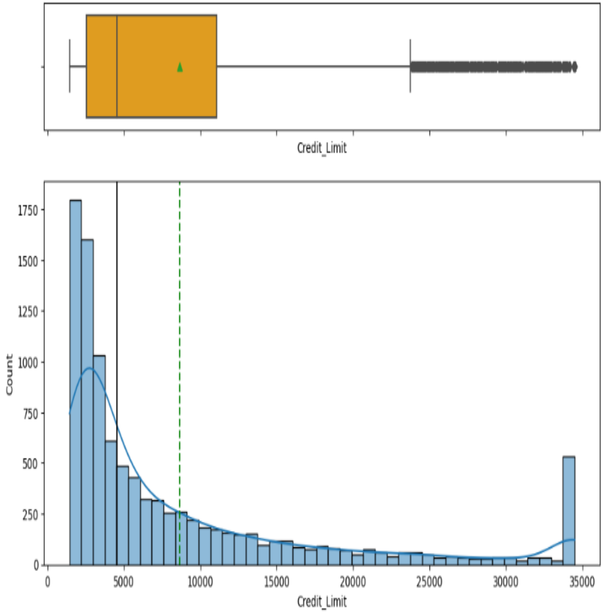
EDA - Univariate Analysis – Months On Book

- **Months_on_book:**
  - Highest: 2,463 customers have 36 months on book relationship with the bank
  - Minimum: The minimum months on book is 13 months
  - Lower Fence: The lower fence months on book is 18 months
  - Q1: 25% of the customers have less than 31 months on book
  - Q3: 75% of the customers have less than 40 months on book
  - Upper Fence: The lower fence months on book is 53 months
  - Maximum: The maximum months on book is 56 months
  - Median & Mean: The median and mean months on book is crica 36 months
  - Outliers: There are 8 outliers – 13, 14, 15, 17, 54, 55 and 56 months on book
  - Skewness: From the plot, it can be observed that the graph has a near to normal distribution



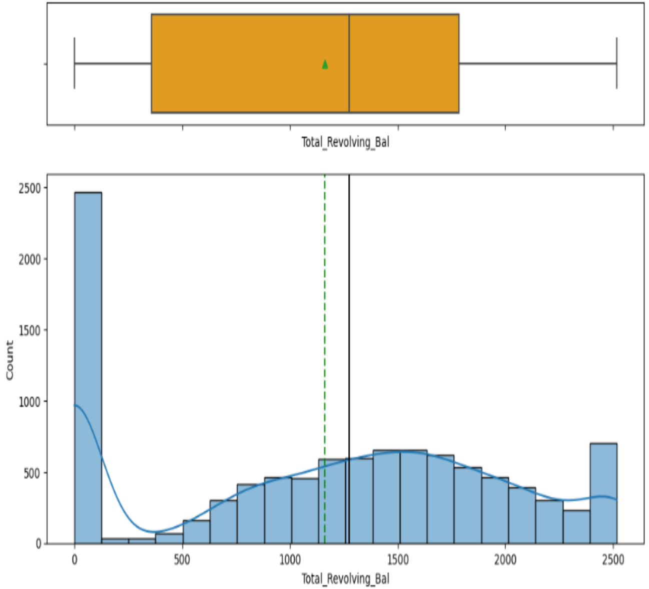
EDA - Univariate Analysis – Credit Limit

- **Credit_Limit:**
  - Highest: Approx. 1,020 customers have their credit limit between $2500 and $2999
  - Minimum: The minimum credit limit is $14383.0
  - Q1: 25% of the customers have a credit limit of less than $2,555
  - Q3: 75% of the customers have a credit limit of less $11K
  - Maximum: The upper fence and maximum credit limit is circa $23K and $34K respectively
  - Median: The median credit limit is $4,549
  - Mean: The mean credit limit is $8,632K
  - Outliers: There are several outliers ranging between $23.8K-$34.51K
  - Skewness: From the plot, it can be observed that the graph highly right skewed



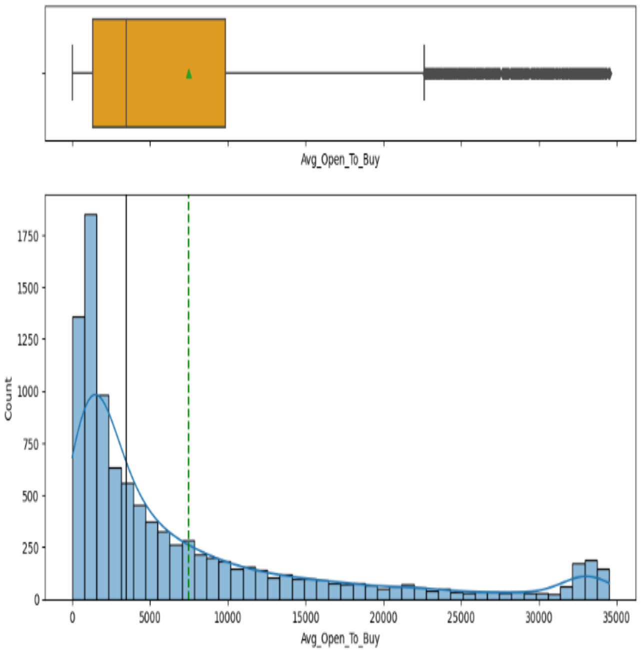
EDA - Univariate Analysis – Total Revolving Balance

- **Total_Revolving_Bal**
  - Highest Balance Group: Approx. 2,470 customers have a total revolving balance between $0K- $40
  - Minimum: The minimum total revolving balance is $0
  - Q1: 25% of the customers have a total revolving balance of less than $359
  - Q3: 75% of the customers have a total revolving balance of less than $1,784
  - Maximum: The maximum total revolving balance is $2,517
  - Median: The median total revolving balance $1,276
  - Mean: The mean total revolving balance $1,162
  - Outliers: There are no outliers
  - Skewness: From the plot, it can be observed that the graph has a near to normal distribution



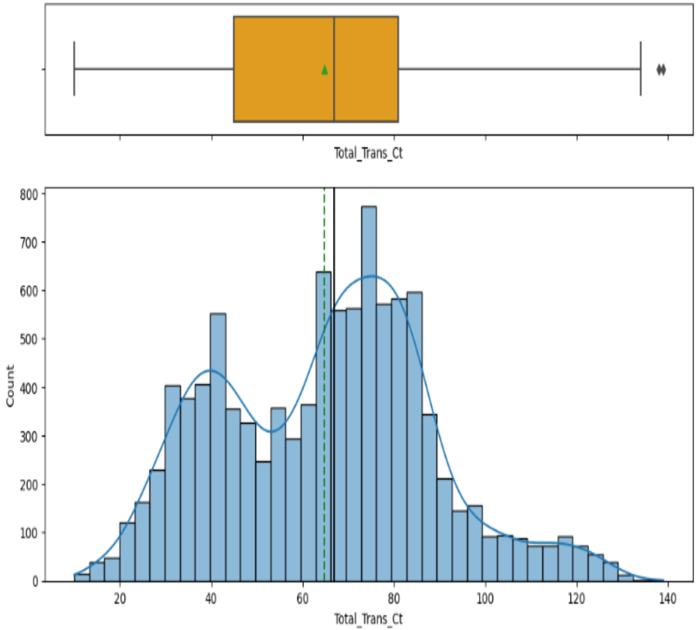
EDA - Univariate Analysis – Average Open To Buy

- **Avg_Open_To_Buy:**
  - Highest: Approx. 1,216 customers have an average open to buy between $750-$1250
  - Minimum: The minimum average open to buy is $3
  - Q1: 25% of the customers have an average open to buy of less than $1,324.5
  - Q3: 75% of the customers have an average open to buy of less than $9,859
  - Maximum: The maximum of average open to buy is $34,516
  - Median: The median average open to buy is $3,474
  - Mean: The mean average open to buy is $7,469
  - Outliers: There are several outliers ranging between $22.66K-$34.51K
  - Skewness: From the plot, it can be observed that the graph highly right skewed



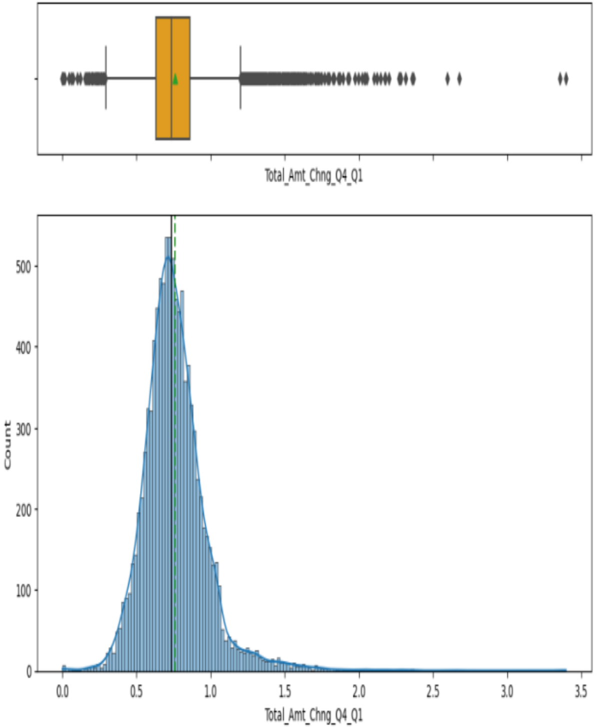
EDA - Univariate Analysis – Total Transaction Count

- **Total_Trans_Ct:**
  - Highest: Approx. 396 customers have a total transaction count between 70-71
  - Minimum: The minimum total transaction count is 10
  - Q1: 25% of the customers have a total transaction count of less than 45
  - Q3: 75% of the customers have a total transaction count of less than 81
  - Maximum: The maximum total transaction count is 139
  - Median: The median total transaction count is 67
  - Mean: The mean total transaction count is approx. 65
  - Outliers: There are 2 outliers -138 & 139
  - Skewness: From the plot, it can be observed that it has a near to normal distribution



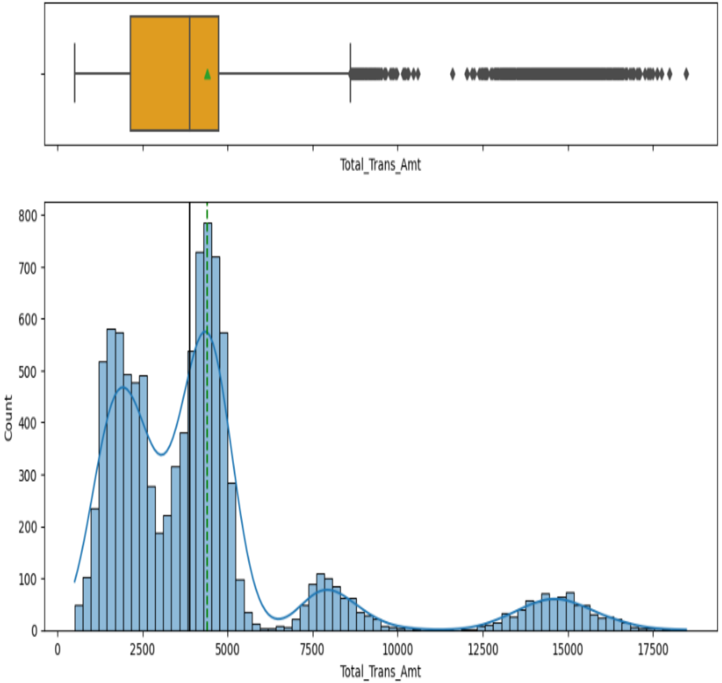
EDA - Univariate Analysis – Total Amount Change Q4-Q1

- **Total_Amt_Chng_Q4_Q1:**
  - Highest: Approx. 552 customers had a total amount change (q4 over q1) between 0.75 and 0.769
  - Minimum: The minimum total amount change (q4 over q1) is 0
  - Q1: 25% of the customers have a total amount change (q4 over q1) of less than 0.63
  - Q3: 75% of the customers have a total amount change (q4 over q1) of less than 0.85
  - Maximum: The maximum total amount change (q4 over q1) is 3.39
  - Median: The median total amount change (q4 over q1) is 0.73
  - Mean: The mean total amount change (q4 over q1) is 0.76
  - Outliers: There are several outliers ranging between 0 to 3.71
  - Skewness: From the plot, it can be observed that the graph normally distributed



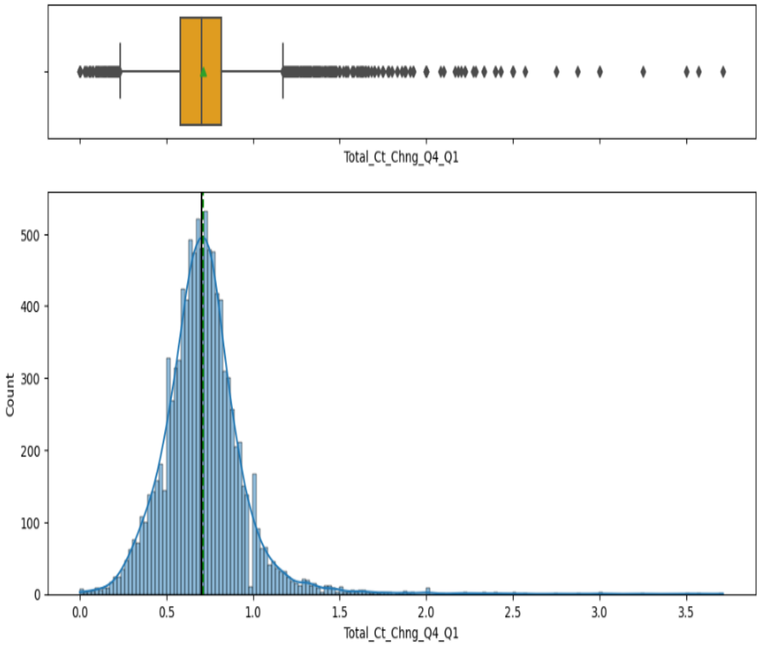
EDA - Univariate Analysis – Total Transaction Amount

- **Total_Trans_Amt:**
  - Highest: Approx. 687 customers have a total transaction amount between $4,400 and $4,599
  - Minimum: The minimum total transaction amount is $510
  - Q1: 25% of the customers have a total transaction amount of less than $2,155
  - Q3: 75% of the customers have a total transaction amount of less than $4,741
  - Upper Fence & Max: The maximum total transaction amount is $18,484 and the upper fence is $8,6189
  - Median: The median total transaction amount is $3,899
  - Mean: The mean total transaction amount is $4,404
  - Outliers: There are several outliers ranging from $8,620 - $18,480
  - Skewness: From the plot, it can be observed that it has a near to normal distribution



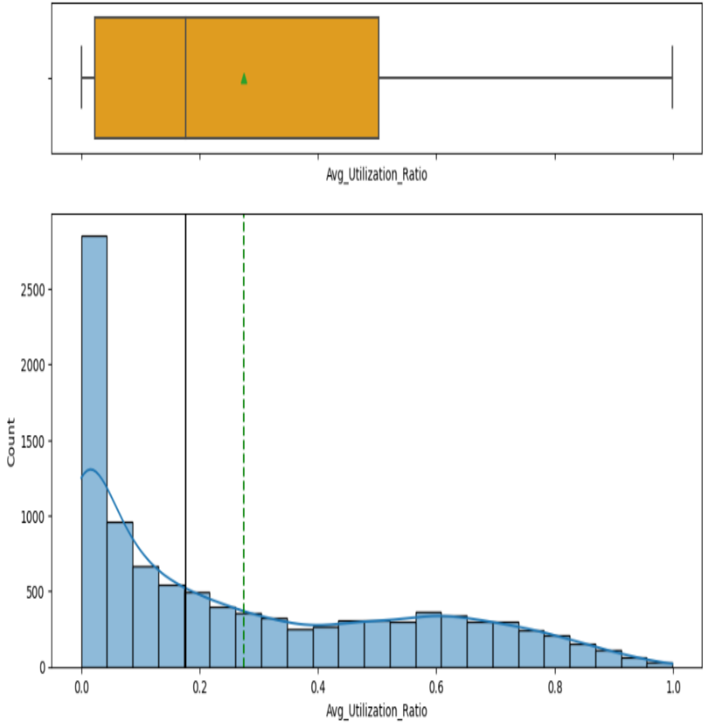
EDA - Univariate Analysis – Total Count Change Q4-Q1

- **Total_Ct_Chng_Q4_Q1:**
  - Highest: Approx. 552 customers had a total count change (q4 over q1) between 0.75 and 0.769
  - Minimum: The minimum total count change (q4 over q1) is 0
  - Q1: 25% of the customers have a total count change over q1) of less than 0.58
  - Q3: 75% of the customers have a total count change (q4 over q1) of less than 0.81
  - Maximum: The maximum total count change (q4 over q1) is 3.71
  - Median: The median total count change (q4 over q1) 0.7
  - Mean: The mean total count change (q4 over q1) is 0.71
  - Outliers: There are several outliers ranging between 0 and 3.71
  - Skewness: From the plot, it can be observed that the graph normally distributed



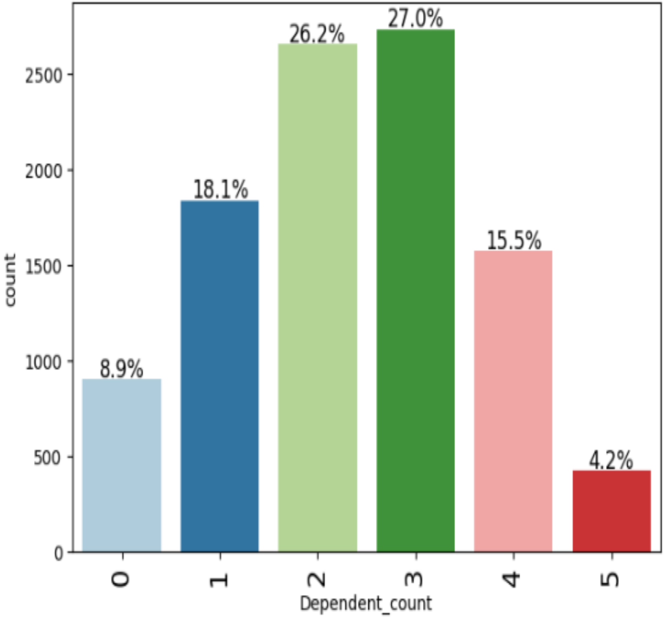
EDA - Univariate Analysis – Average Utilisation Ratio

- **Avg_Utilization_Ratio:**
  - Highest: Approx. 2,479 customers have an average utilisation ratio of 0
  - Minimum: The minimum average utilisation ratio is 0
  - Q1: 25% of the customers have an average utilisation ratio of less than 0.02
  - Q3: 75% of the customers have an average utilisation ratio of less than 0.50
  - Maximum: The maximum average utilisation ratio is 0.99
  - Median: The median average utilisation ratio is 0.17
  - Mean: The mean average open to buy is 0.27
  - Outliers: There are no outliers
  - Skewness: From the plot, it can be observed that the graph slightly right skewed



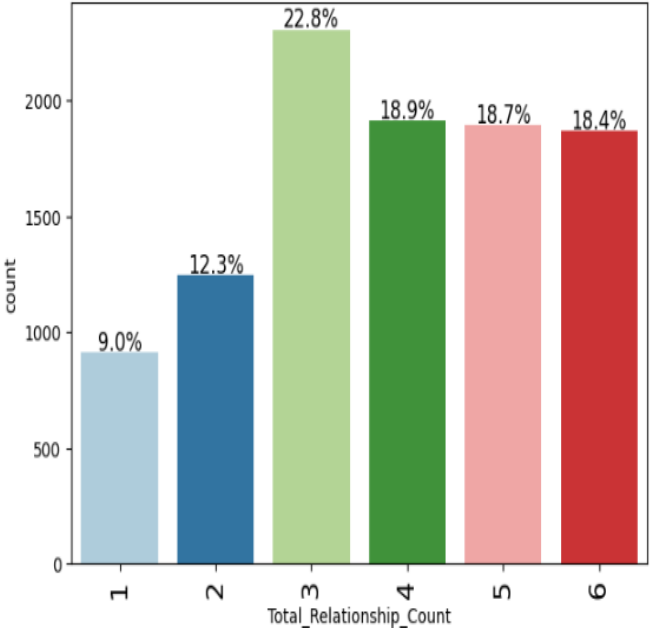
EDA - Univariate Analysis – Dependent Count

- **Dependent_count:**
  - Majority of customers - approx. 27% (2,732) have only 3 dependents in their family, followed by 26.2% (2,655) and 18.1% (1,838) who have 2 & 1 dependents in their families, respectively
  - 15.5% (1,574) have 4 dependents, whereas 4.2% (424) have 5 dependents in their families
  - 8.5% (904) have no dependents



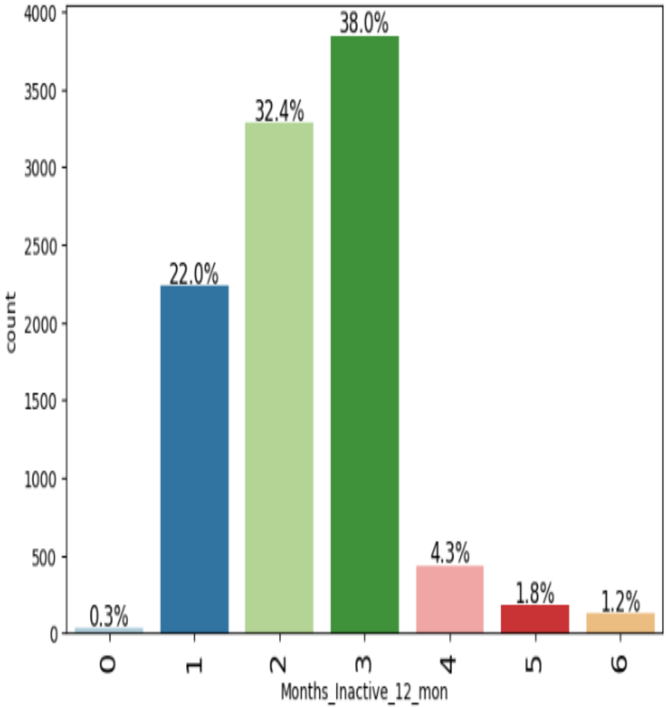
EDA - Univariate Analysis – Total Relationship Count

- **Total_Relationship_Count:**
  - Majority of customers - approx. 22.8% (2,305) have 3 products / relationship, followed by 18.9% (1,912) that have 4 products / relationship with the bank
  - 18.7% (1,891) and 18.4% (1,866) of customers have 5 & 6 products / relationship with the bank
  - 12.3% (1,243) of customers have 2 products / relationship with the bank, whereas 9% (910) has only 1 products / relationship



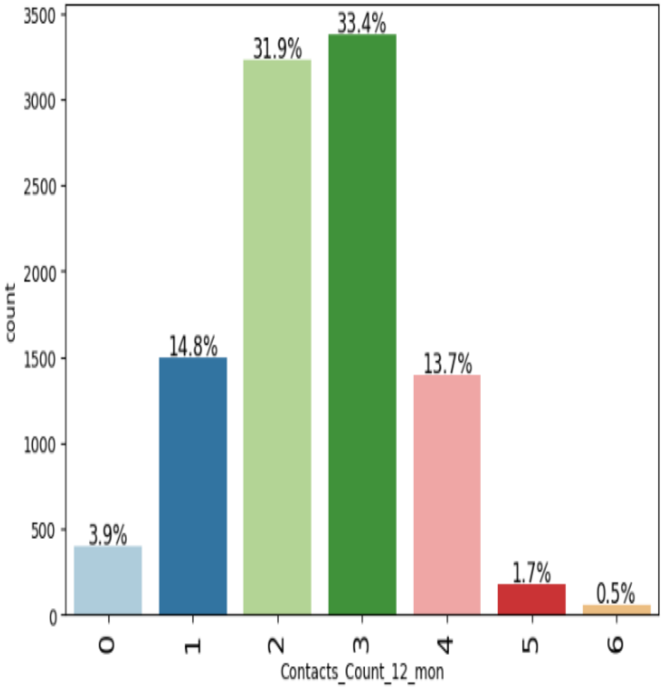
EDA - Univariate Analysis – Months Inactive – 12 Months

- **Months_Inactive_12_mon:**
  - Approx. 38% (3,846) customers have been inactive for 3 months, followed by 32.4% (3,282) that been inactive for 2 months, followed by 22% (2,233) that been inactive for 1 month
  - Approx. 4.3% (435) of customers have been inactive for 4 months, followed by 1.8% (178) that been inactive for 5 months.
  - Approx. 1.2% (124) of customers have been inactive for 6 months



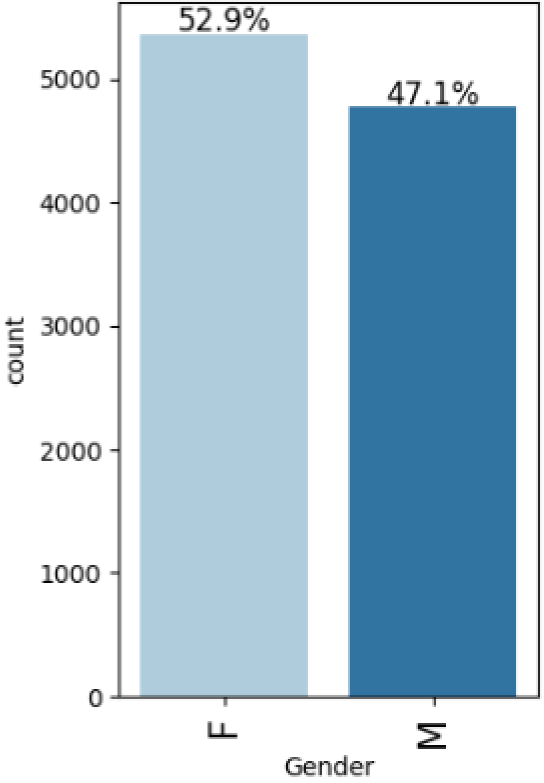
EDA - Univariate Analysis – Contacts Counts – 12 Months

- **Contacts_Count_12_mon:**
  - Approx. 33.4% (3,380) of customers have been contacted 3 times, and 31.9% (3,227) of customers have been contacted 2 times
  - Approx. 14.8% (1,499) of customers have been contacted once, and 13.7% (1,392) of customers have been contacted 4 times
  - Approx. 1.7% (176) of customers have been contacted 5 times, whereas 0.5% (54) customers have been contacted 6 times
  - Approx. 3.9% (399) customers haven’t been contacted at all



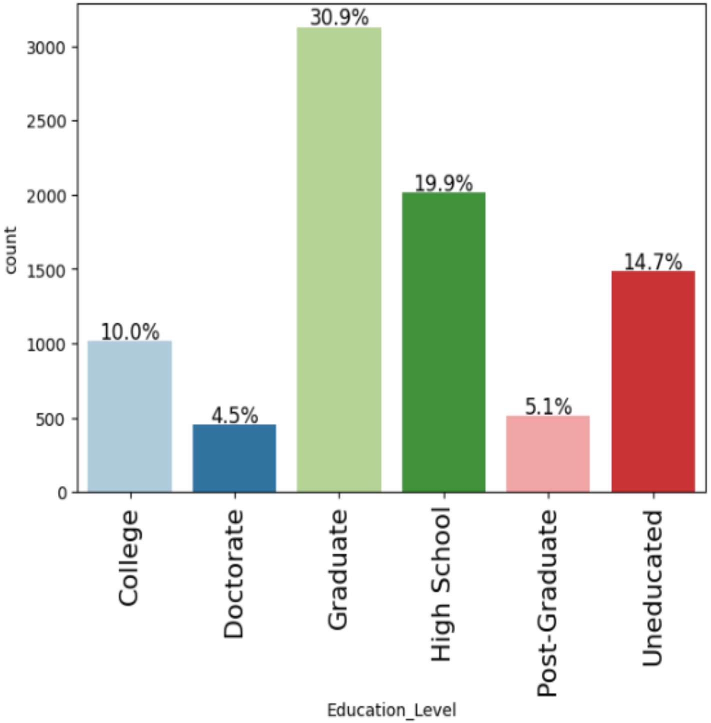
EDA - Univariate Analysis – Gender

- **Gender:**
  - The ratio of Female to Male customers is similar
  - Approx. 52.9% (5,358) of customers are Women, whereas 47.1% (4,769) of customers are Men



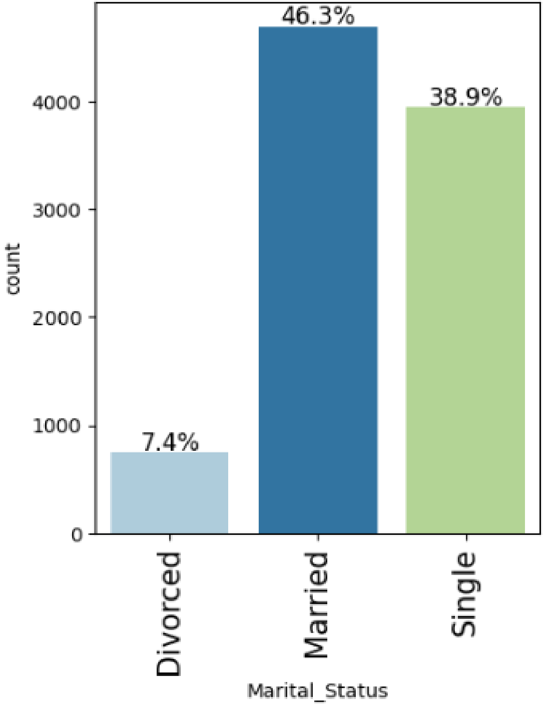
EDA - Univariate Analysis – Education Level

- **Education_Level:**
  - There are higher number of graduates than compared to other degree and non-degree holders.
  - Approx. 30.9% (3,128) of customers are graduates, followed by 19.9% (2,013) that educated till high-school.
  - Approx. 14.7% (1,487) of customers are uneducated, followed by 10% (1,013) that college educated.
  - Approx. 5.1% (516) of customers have a post graduate degree, and 4.5% (451) hold a doctorate degree.



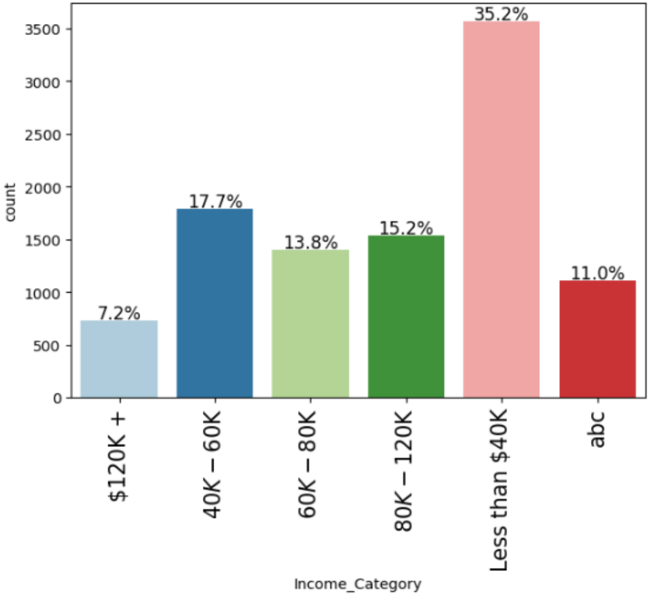
EDA - Univariate Analysis – Marital Status

- **Marital_Status:**
  - Most of the customers are married or single, whereas a significant minority of customers are divorced
  - Approx. 46.3% (4,687) of customers are married, whereas 38.9% (3,943) customer are single
  - Approx. 7.4% (748) of customers are divorced



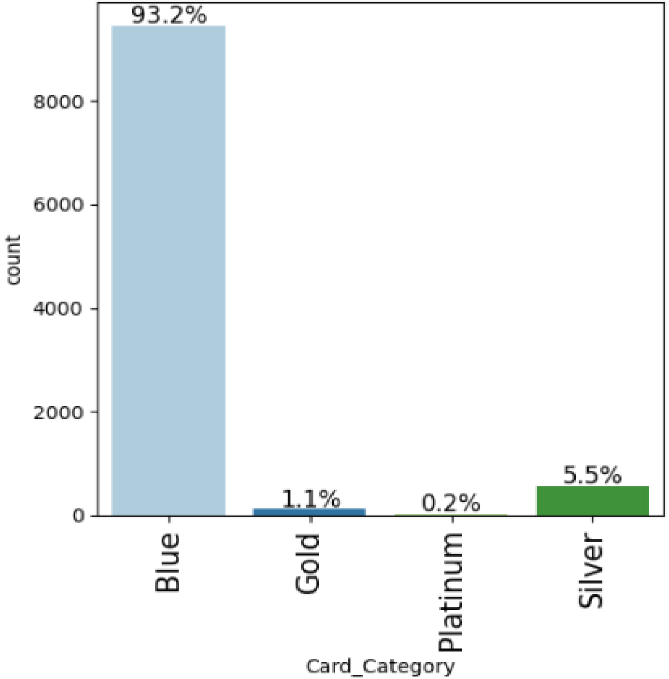
EDA - Univariate Analysis – Income Category

- **Income_Category:**
  - Majority of customer have an income of less than $40K, whereas a minority of customers have an income of above $120K
  - Approx. 35.2% (3,561) of customers have an income of less than $40K, whereas only 7.2% (727) customers have an income of above $120K
  - Approx. 17.7% (1,790) of customers have an income between $40K-$60K, 13.8% (1,402) of customers have an income between $60K-$80K, and 15.2% (1,535) of customers have an income between $80K-$120K
  - Approx. 11% (1,112) of records have anomalous data in the dataset that will need to be rectified



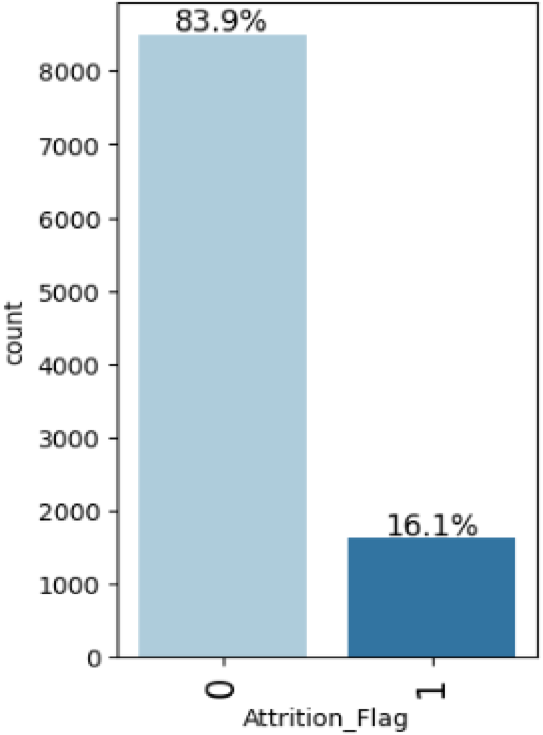
EDA - Univariate Analysis – Card Category

- **Card_Category:**
  - Majority of customers hold Blue card category whereas a significant minority of customers hold Gold & Platinum cards
  - Approx. 93.2% (9,436) of customers belong to the Blue card category, whereas only 1.1% (116) and 0.2% (20) customers belong to the Gold and Platinum card category
  - Approx. 5.5% (555) of customers belong to the Silver card category

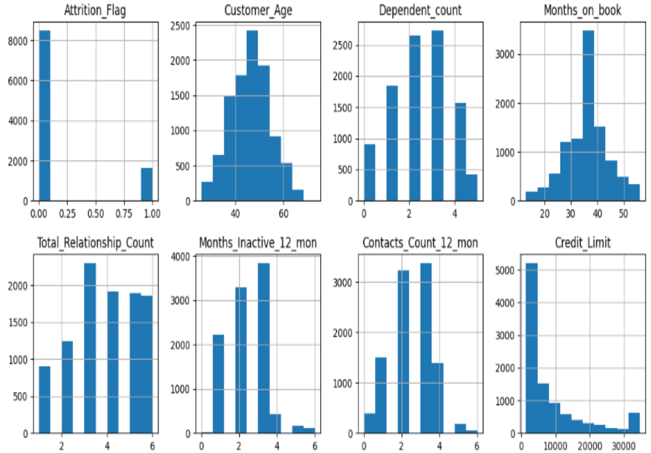


EDA - Univariate Analysis – Attrition Flag

- **Attrition_Flag:**
  - There is a significant imbalance in data since there are high number of existing customers than compared to those that have attrited
  - Approx. 83.9% (8,500) are existing that compared to 16.1% (1,627) customers that have left the credit card services
  - The high imbalance in data would need to be fixed prior to model building

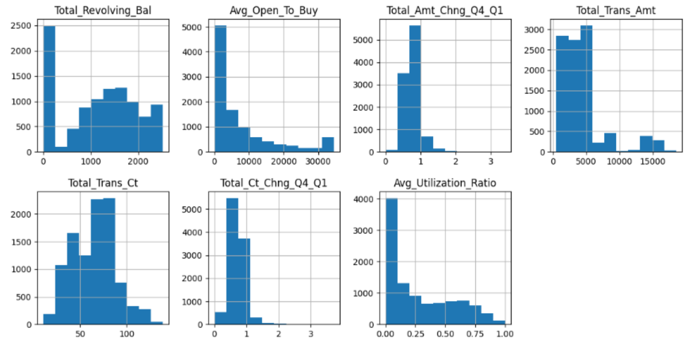


EDA - Univariate Analysis – Key Observations & Insights



- **Key Observations:**
  - 83.9% are existing customer than compared to 16.1% that have left the credit card services
  - 93.2% of customers belong to the Blue card category, whereas only 1.1% and 0.2% customers belong to the Gold and Platinum cards
  - 35.2% of customers have an income of less than $40K, whereas only 7.2% customers have an income of above $120K
  - 46.3% of customers are married, whereas 38.9% customer are single, and 7.4% of customers are divorced
  - 30.9% of customers are graduates, whereas only 5.1% are post-graduates and 4.5% hold a doctorate degree
  - 52.9% of customers are Women, whereas 47.1% of customers are Men
  - 65% of customers have been contacted 2 or 3 times, whereas a significant minority have been contacted very rarely
  - 70% of customers have been inactive for a period of 2-3 months
  - Approx. 2,479 customers have an average utilisation ratio of 0
  - Approx. 2,470 customers have a total revolving balance between $0K- $40
  - Average Open to Buy has lots of higher end outliers, which means there are customers who uses only very small amount of their credit limit

EDA - Univariate Analysis – Key Observations & Insights

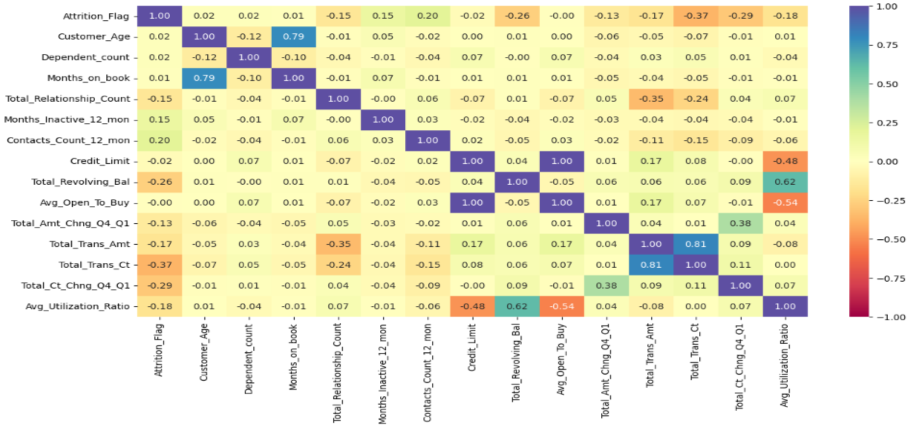


- **Key Insights:**
  - There is a significant imbalance in dataset since there are high number of existing customers than compared to those that attrited
  - Majority of customers are in Blue card category and very few in Gold & Platinum category, which implies that customers are not using their credit card as much as possible and there are opportunities for service improvement
  - 25% of customers have a total revolving balance of 0, which could imply that these customer don't often use their cards
  - Average Open to Buy has lots of high-end outliers, which means that there are customers who uses only very small amount of their credit limit
  - Approx. 65% of customers have been contacted by the bank very often, which implies that that there have been issues where the customer needed support
  - Majority of customers have been inactive for a period of 2-3 months, which could imply that there could be facing some challenges using the credit card services
  - 30% of customers have a nil average utilsation ratio
  - The ratio of Female to Male customers is similar
  - Majority of customer have an income of less than $40K, whereas a minority of customers have an income of above $120K

EDA - Bivariate Analysis

EDA - Bivariate Analysis – Correlation Check

- **Correlation Amongst Variables :**
  - There is a significant positive correlation of 0.79 between Customer_Age and Months_On_Book, which implies that customers have been with the bank for a signifcant period of time.
  - There is a significant positive correlation of 0.81 between Total_Trans_Amt and Total_Trans_Ct, which implies that higher the number of transactions greater is the generated revenue
  - There is a strong positive correlation of 0.62 between Total_Revolving_Bal and Avg_Utilisation_Ratio
  - There is a weak positive correlation of 0.38 between Total_Amt_Chng_Q4_Q1 and Total_Ct_Chng_Q4_Q1
  - There is an insignificant positive correlation of 0.20 between Contacts_Count_12_mon and Attrition_Flag, and 0.15 between Months_Inactive_12_mon and Attrition_Flag
  - There is a strong negative correlation of -0.54 and weak negative correlation of -0.48 between Avg_Utilisation_Ratio, and Avg_Open_To_Buy, and Credit_Limit respectively. This implies that less usage of the card gives more available credit to be spent

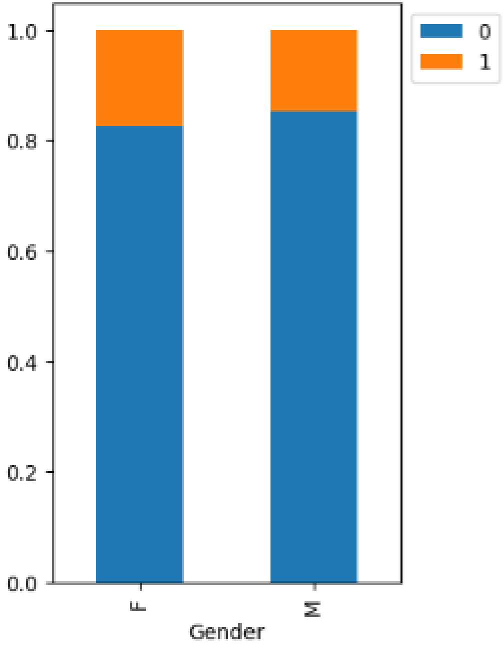


EDA - Bivariate Analysis – Correlation Check

- **Correlation Amongst Variables (Cont'd) :**
  - There is an insignificantly positive correlation of 0.17 and 01.7 between Total_Trans_Amt, and Credit_Limit & Avg_Open_To_Buy respectively
  - There is a weak negative correlation of -0.35 and -0.24 between Total_Relationship_Count, and Total_Trans_Amt & Total_Trans_Ct respectively, which implies that owing a single product (instead of multiple products) such as a credit card increases the probability of using it often thereby increasing revenue
  - There is a weak negative correlation of -0.37, -0.29 and -0.26 between Attrition_Flag, and Total_Trans_Ct, Total_Ct_Chng_Q4_Q1 and Total_Revolving_Bal
  - There is an insignificantly negative correlation of -0.18, -0.17, -0.15 and -0.13 between Attrition_Flag, and Avg_Utilization_Ratio, Total_Trans_Amt, Total_Relationship_Count and Total_Amt_Chng_Q4_Q1
  - There is an insignificantly negative correlation of -0.15 and -0.11 between Contacts_Count_12_mon, and Total_Trans_Ct & Total_Trans_Amt, respectively
  - Customer attributes such as Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Total_Amt_Chng_Q4_Q1, and Total_Relationship_Count are the most important features in predicting potential customers that will attrit. These features are negatively correlated with the Attrition_Flag, which implies that lower the values of these features, the higher the chances of a customer to attrite

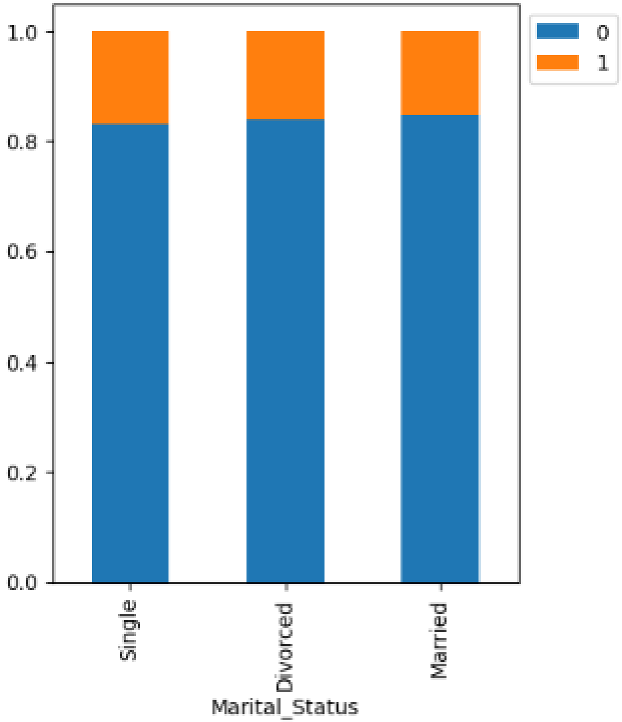
EDA - Bivariate Analysis – Attrition vs Gender

- **Attrition_Flag vs Gender:**
  - The attrition levels of females is slightly higher than males.
  - 17.4% (930 of 5,358) of females have attrited than compared to 14.6% (697 of 4,769) of males that have attrited



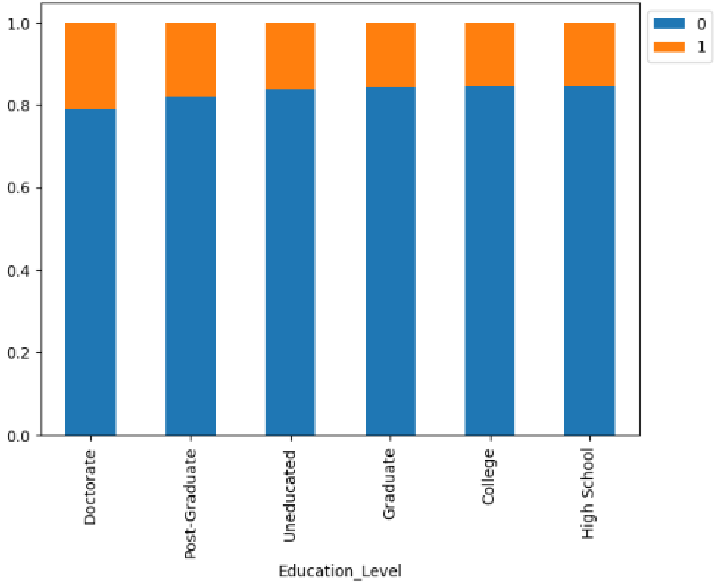
EDA - Bivariate Analysis – Attrition vs Marital Status

- **Attrition_Flag vs Marital_Status:**
  - The attrition levels is high for single and divorced than compared to married customers.
  - 16.9% (668 of 3,943) of singles and 16.2% (121 of 748) of divorced have attrited than compared to 15.1% (709 of 4,687) of married customers who have attrited



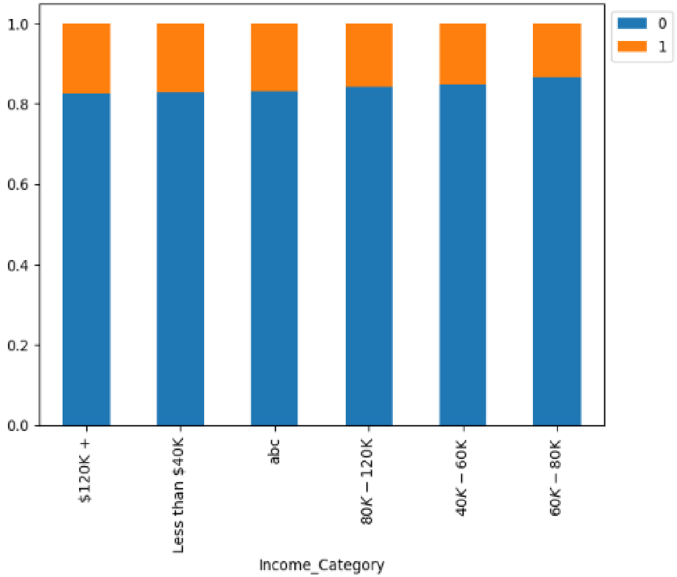
EDA - Bivariate Analysis – Attrition vs Education Level

- **Attrition_Flag vs Education_Level:**
  - The attrition levels for advanced degrees such as Doctorate and Postgraduates is much higher than compared to other degree / non-degree educated customers
  - 21.1% (95 of 451) of Doctorates and 17.8% (92 of 516) of Postgraduates have attrited
  - Graduates, College & High School educated, and uneducated customers have a similar attrition rate
  - 15.9% (237 of 1,487) of Uneducated, 15.6% (487 of 3,128) of Graduates, 15.2% (1547 of 1,013) of College, and 15.2% (306 of 2,013) of High School have attrited



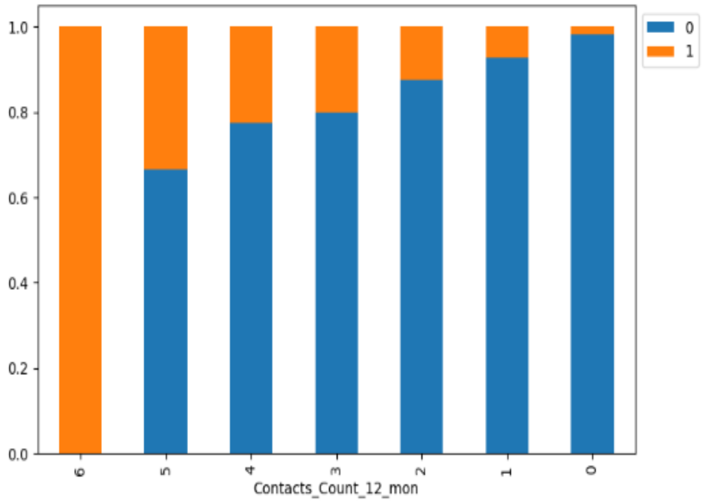
EDA - Bivariate Analysis – Attrition vs Income Category

- **Attrition_Flag vs Income_Category:**
    - Customers with an income above $120K and those with an income below $40K have higher rates of attrition than compared to customers who have an income within the range of $40K-$120K
    - 17.3% (126 of 727) of customers with an income above $120K, and 17.2% (612 of 3,561) of customers with an income below $40K have attrited
    - 15.8% (242 of 1,535) of customers with an income between $80K-$120K, and 15.1% (271 of 1,790) of customers with an income between $40K-$60K have attrited
    - 13.5% (189 of 1,402) of customers with an income between $60K-$80K have attrited



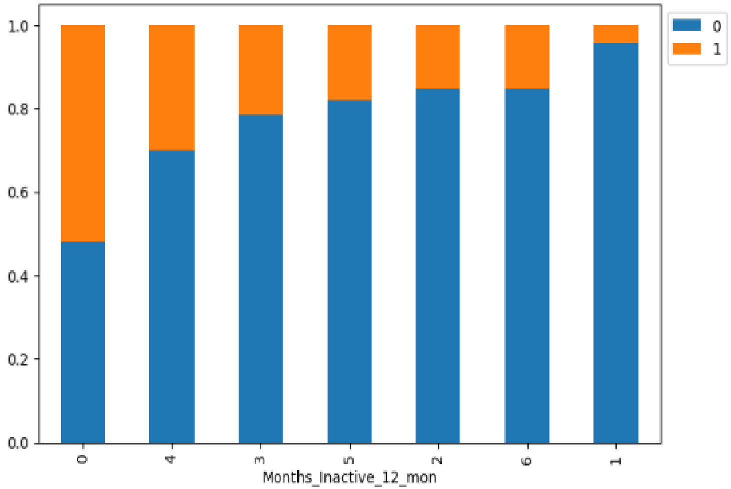
EDA - Bivariate Analysis – Attrition vs Contacts Counts

- **Attrition_Flag vs Contacts_Count_12_mon:**
  - All customers with a contact count of 6 have attrited than compared to other customers with a relatively low contact count
  - 33.5% (59 of 176) of customers with contact count of 5 have attrited, followed by 22.6% (315 of 1,392) of customers with contact count of 4 have attrited.
  - 20.1% (681 of 3,380) of customers with contact count of 3 have attrited, whereas 12.5% (403 of 3,227) of customers with contact count of 2 have attrited
  - 7.2% (108 of 1,499) of customers with contact count of 1 have attrited, whereas 1.8% (7 of 399) of customers with contact count of 0 have attrited



EDA - Bivariate Analysis – Attrition vs Months Inactive

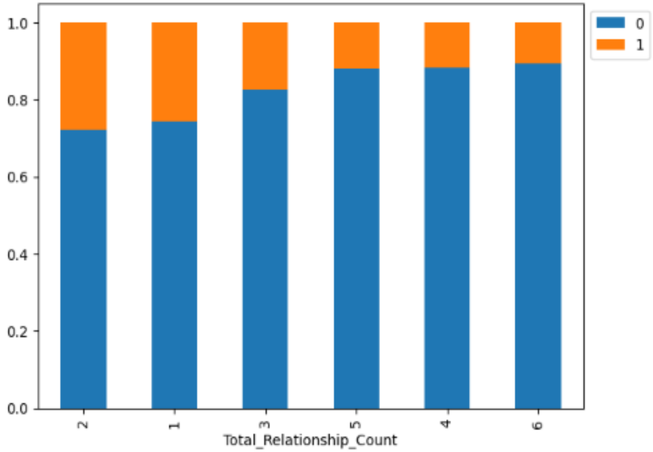
- **Attrition_Flag vs Months_Inactive_12_mon:**
  - Nearly half of the customers that have been fully active over a 12-month period have attrited than compared to those customers who have been inactive for longer periods
  - 51.7% (15 of 29) of customers that have been fully active have attrited than compared to 15.3% (19 of 124) of customers that have been only active for only 6 months
  - 29.9% (130 of 435) of customers who have been inactive for 4 months, and 21.5% (826 of 3,846) of customers who have been inactive for 3 months have attrited
  - 18% (32 of 178) of customers who have been inactive for 5 months, and 15.4% (505 of 3,282) of customers who have been inactive for 2 months have attrited
  - Only 4.5% (100 of 2,233) of customers who have been inactive for 1 month have attrited



Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

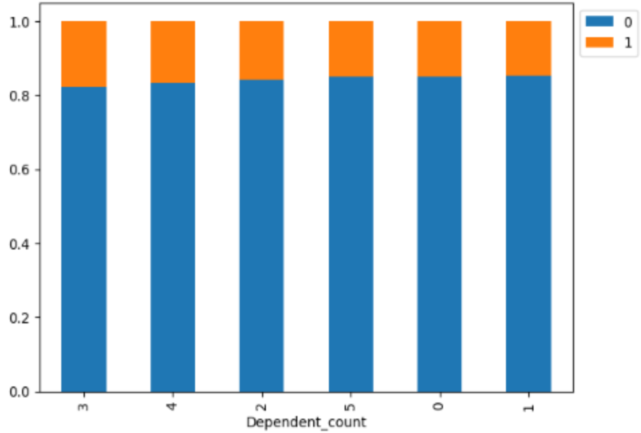
EDA - Bivariate Analysis – Attrition vs Total Relationship Count

- **Attrition_Flag vs Total_Relationship_Count:**
  - Customers having 1 or 2 relationship counts / products have a higher rate of attrition that compared to those who have a relationship count of 3 and above
  - 27.8% (346 of 1,243) and 25.6% (233 of 910) of customers with a relationship counts / products of 2 & 1 respectively have attrited
  - 17.4% (400 of 2,305) of customers with a relationship count of 3, 12% (227 of 1,891) of customers with a relationship count of 5, 11.8% (225 of 1,912) of customers with a relationship count of 4, and 10.5% (196 of 1,866) of customers with a relationship count of 6 have attrited



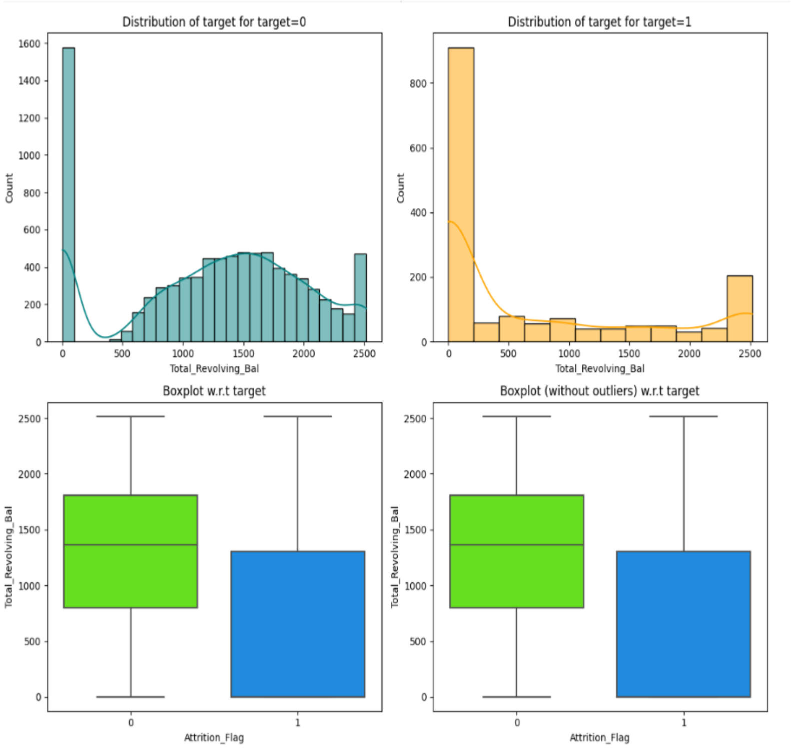
EDA - Bivariate Analysis – Attrition vs Dependent Count

- **Attrition_Flag vs Dependent_Count:**
  - Customers having 3 or more dependents have a higher rate of attrition that compared to those who having less than 3 dependents
  - 17.6% (482 of 2,732) and 16.5% (260 of 1,574) of customers with dependents of 3 & 4 respectively have attrited
  - 15.7% (417 of 2,655) of customers with a dependent count of 2, 15.1% (64 of 424) of customers with a dependent count of 5, 14.9% (135 of 904) of customers with a dependent count of 0, and 14.6% (269 of 1,838) of customers with a dependent count of 1 have attrited



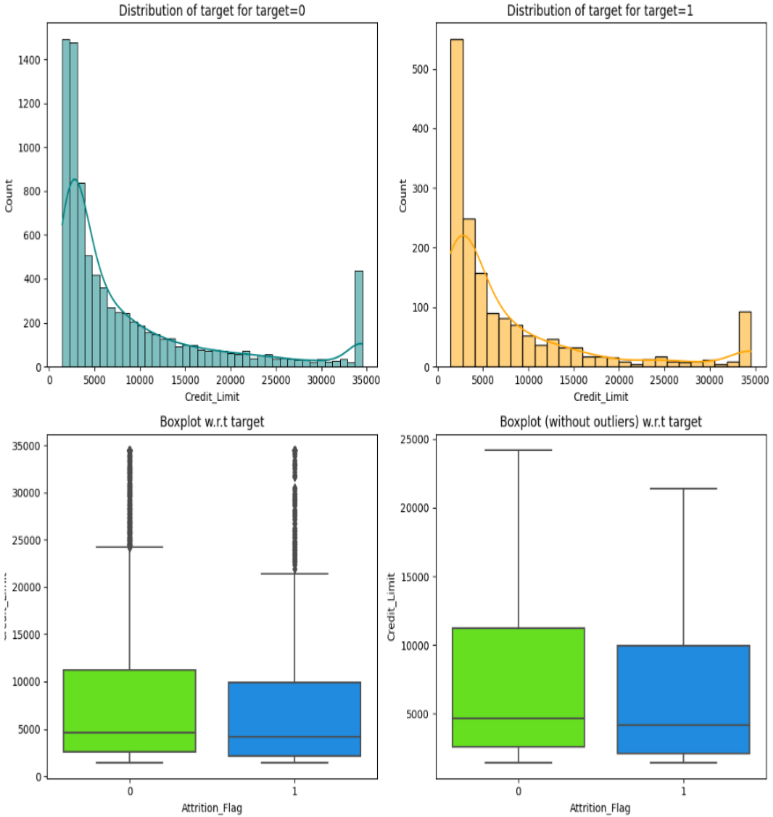
EDA - Bivariate Analysis – Attrition vs Total Revolving Bal

- **Attrition_Flag vs Total_Revolving_Bal :**
  - Minimum: Customers that are likely to attrit have 0 total revolving balance
  - Maximum: The maximum and the upper fence was $2,517
  - Q1: 25% of the customers that attrited had a total revolving balance of less than $0
  - Q3: 75% of the of the customers that attrited had a total revolving balance of less than $1,303
  - The mean total revolving balance for attrited customer was $672
  - The median total revolving balance for attrited customer was $0



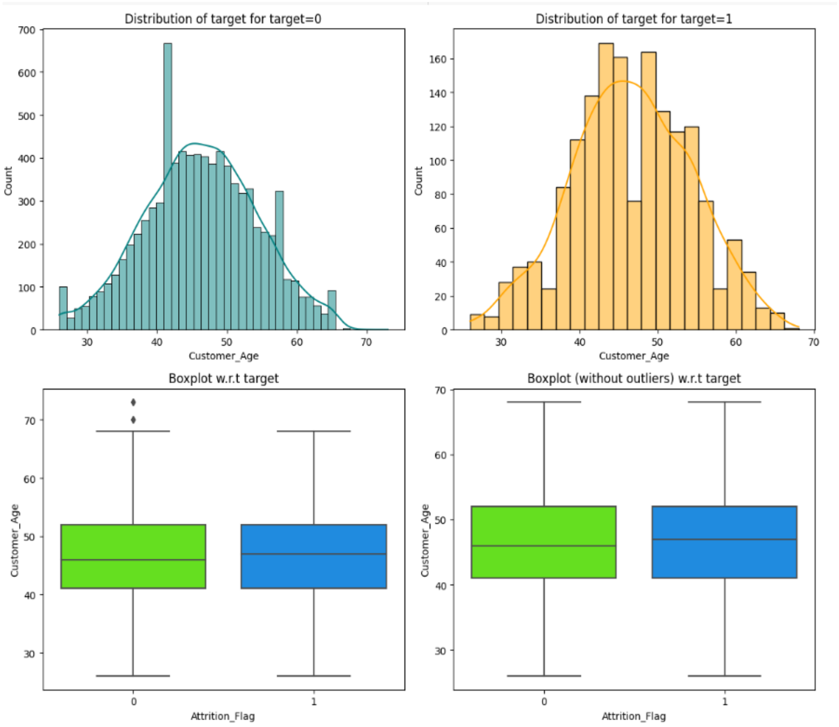
EDA - Bivariate Analysis – Attrition vs Credit Limit

- **Attrition_Flag vs Credit_Limit:**
    - Minimum: Customers that attired had a minimum credit limit of $1,438
    - Maximum: Attrited customers had an upper fence and maximum credit limit of $21.3K and 34.5K respectively
    - Q1: 25% of the customers that attrited had a credit limit of less than $2,114
    - Q3: 75% of the of the customers that attrited had credit limit of less than $9.93K
    - The mean credit limit for attrited customer was $8,136
    - The mediancredit limit for attrited customer was $4,178



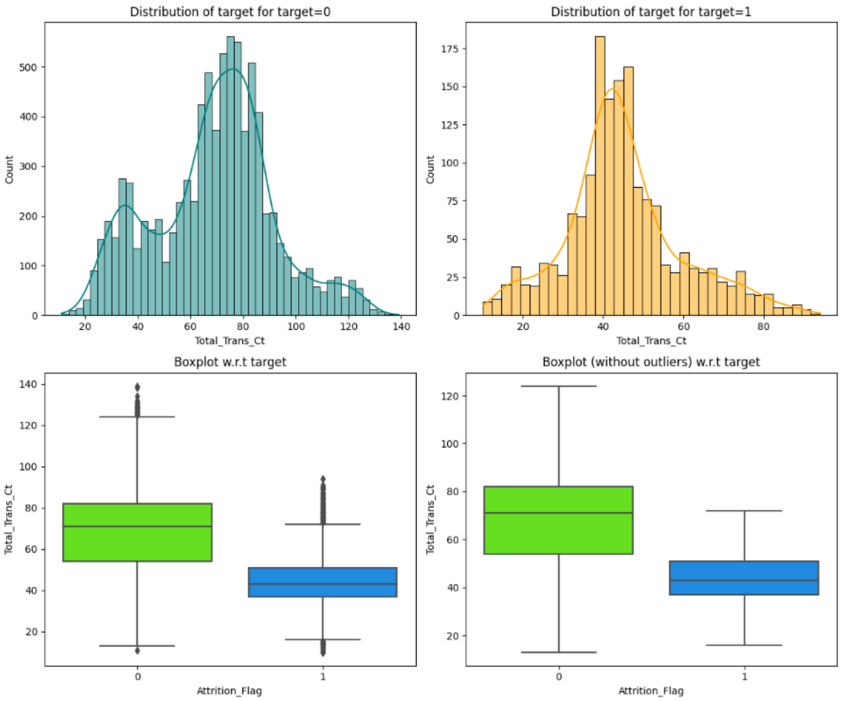
EDA - Bivariate Analysis – Attrition vs Customer Age

- **Attrition_Flag vs Customer_Age:**
  - Minimum: Customers that attrited had a minimum age of 26 years
  - Maximum: Customers that attrited had a maximum age of 68 years
  - Q1: 25% of the customers that attrited were less than 41 years
  - Q3: 75% of the of the customers that attrited were less than 52 years
  - The mean age for attrited customer was 46 years
  - The median age for attrited customer was 47 years



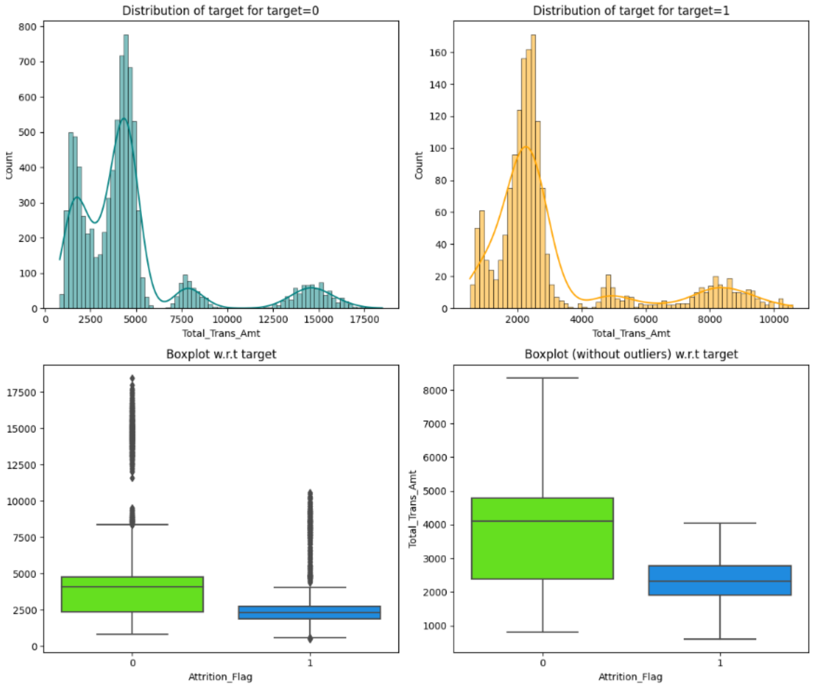
EDA - Bivariate Analysis – Attrition vs Total Transaction Count

- **Attrition_Flag vs Total_Trans_Ct:**
  - Minimum: The minimum Total_Trans_Ct for attrited customers was 10
  - Maximum: The upper fence and maximum Total_Trans_Ct for attrited customers was 72 and 94 respectively
  - Q1: 25% of Total_Trans_Ct for attrited customers was less than 37
  - Q3: 75% of the Total_Trans_Ct for attrited customers was less than 51
  - The mean Total_Trans_Ct for attrited customers was 44.9
  - The median Total_Trans_Ct for attrited customers was 43



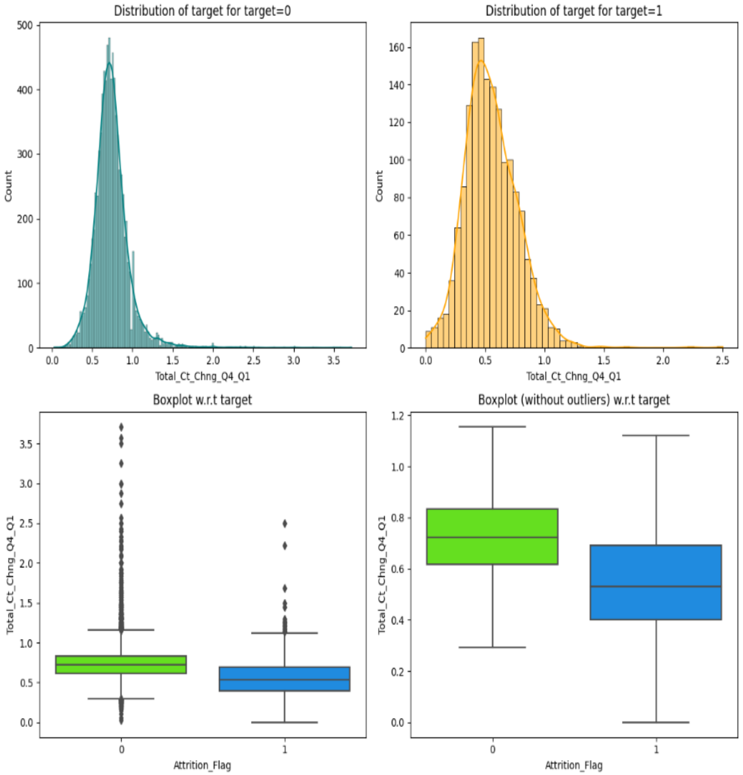
EDA - Bivariate Analysis – Attrition vs Total Transaction Amount

- **Attrition_Flag vs Total_Trans_Amt:**
  - Minimum: The minimum Total_Trans_Amt for attrited customers was $510
  - Maximum: The upper fence and maximum Total_Trans_Amt for attrited customers was $4,041 and 10.58K respectively
  - Q1: 25% of Total_Trans_Amt for attrited customers was less than $1,903.5
  - Q3: 75% of the Total_Trans_Amt for attrited customers was less than $2,772
  - The mean Total_Trans_Amt for attrited customers was $3,095
  - The median Total_Trans_Amt for attrited customers was $2,329



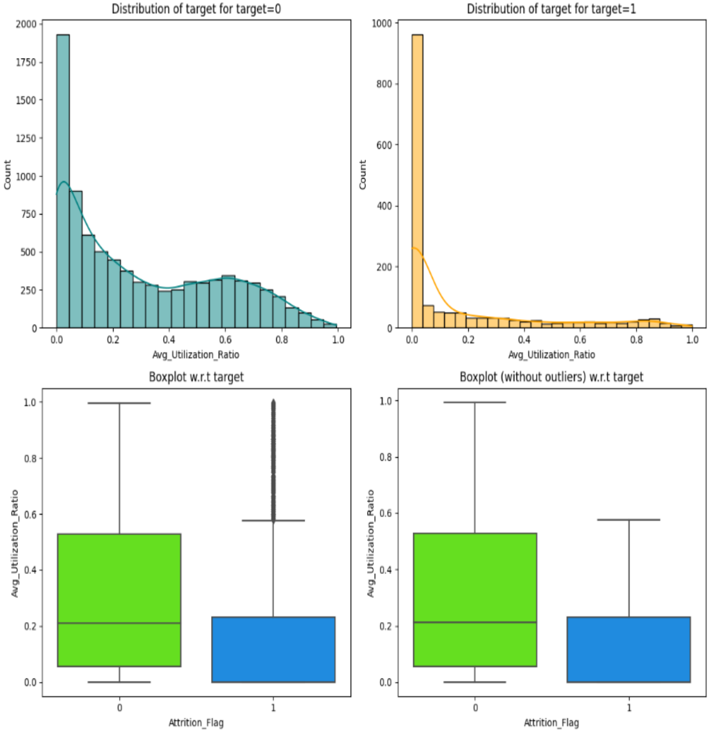
EDA - Bivariate Analysis – Attrition vs Total CT Change Q4-Q1

- **Attrition_Flag vs Total_Ct_Chng_Q4_Q1:**
  - Minimum: The minimum Total_Ct_Chng_Q4_Q1 for attrited customers was 0
  - Maximum: The upper fence and maximum Total_Ct_Chng_Q4_Q1 for attrited customers was 1.22 and 2.5 respectively
  - Q1: 25% of Total_Ct_Chng_Q4_Q1 for attrited customers was less than 0.4
  - Q3: 75% of the Total_Ct_Chng_Q4_Q1 for attrited customers was less than 0.69
  - The mean Total_Ct_Chng_Q4_Q1 for attrited customers was 0.55
  - The median Total_Ct_Chng_Q4_Q1 for attrited customers was 0.53



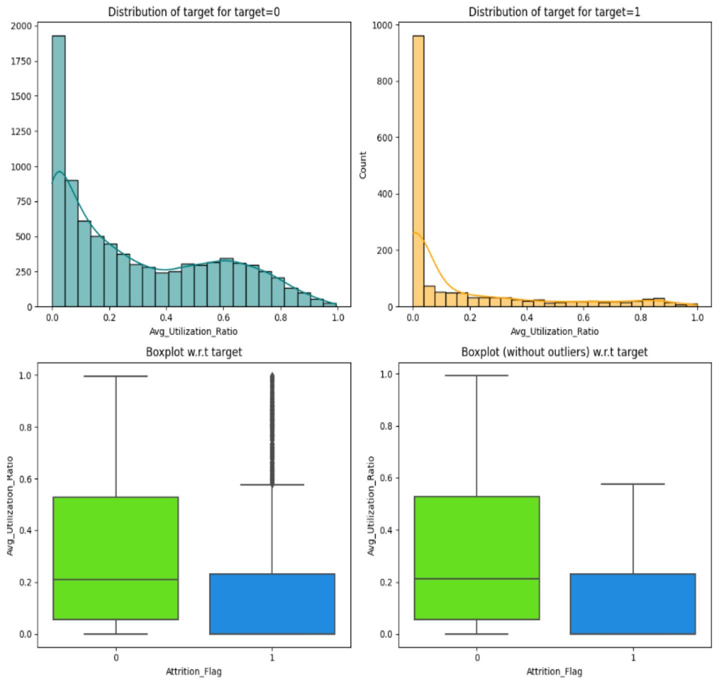
EDA - Bivariate Analysis – Attrition vs Avg Utilisation Ratio

- **Attrition_Flag vs Avg_Utilization_Ratio:**
  - Minimum: The minimum Avg_Utilization_Ratio for attrited customers was 0
  - Maximum: The upper fence and maximum Avg_Utilization_Ratio for attrited customers was 0.57 and 0.99 respectively
  - Q1: 25% of Avg_Utilization_Ratio for attrited customers was less than 0
  - Q3: 75% of the Avg_Utilization_Ratio for attrited customers was less than 0.23
  - The mean Avg_Utilization_Ratio for attrited customers was 0.16
  - The median Avg_Utilization_Ratio for attrited customers was 0



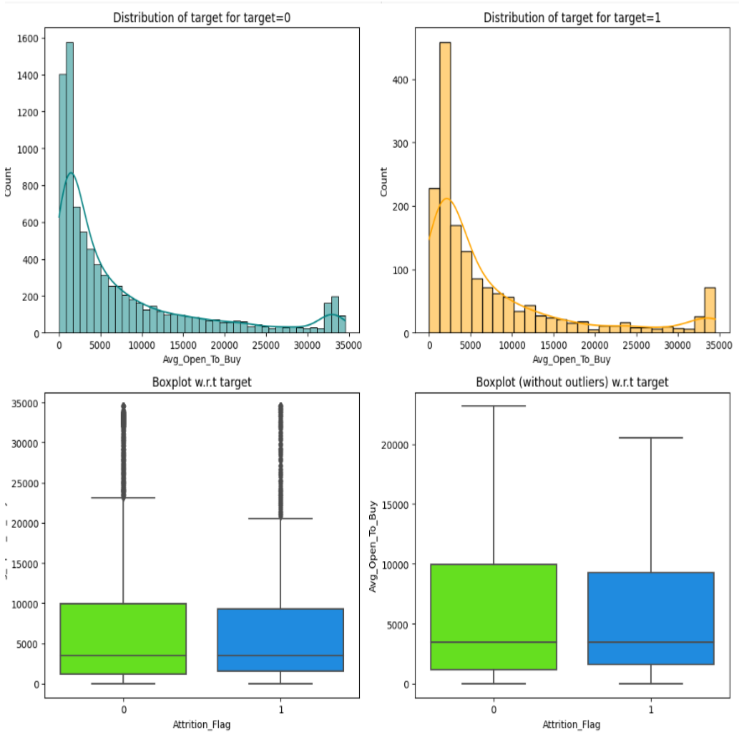
EDA - Bivariate Analysis – Attrition vs Months On Book

- **Attrition_Flag vs Months_on_book:**
  - Minimum: The minimum Months_on_book for attrited customers was 13 months and the lower fence was 20 months
  - Maximum: The upper fence and maximum Months_on_book for attrited customers was 52 months and 56 months respectively
  - Q1: 25% of Months_on_book for attrited customers was less than 32 months
  - Q3: 75% of the Months_on_book for attrited customers was less than 40 months
  - The mean Months_on_book for attrited customers was 36.17 months
  - The median Months_on_book for attrited customers was 36 months



EDA - Bivariate Analysis – Attrition vs Avg Open To Buy

- **Attrition_Flag vs Avg_Open_To_Buy:**
  - Minimum: The minimum Avg_Open_To_Buy for attrited customers was $3
  - Maximum: The upper fence and maximum Avg_Open_To_Buy for attrited customers was $20.5K and $34.5K respectively
  - Q1: 25% of Avg_Open_To_Buy for attrited customers was less than $1,587
  - Q3: 75% of the Avg_Open_To_Buy for attrited customers was less than $9,257
  - The mean Avg_Open_To_Buy for attrited customers was $7,463
  - The median Avg_Open_To_Buy for attrited customers was $3,488



EDA - Bivariate Analysis – Key Observations & Insights

- **Key Observations :**
  - There is a significant positive correlation of 0.79 between Customer_Age and Months_On_Book, which implies that customers have been with the bank for a signifcant period of time.
  - There is a significant positive correlation of 0.81 between Total_Trans_Amt and Total_Trans_Ct, which implies that higher the number of transactions greater is the generated revenue
  - There is a strong positive correlation of 0.62 between Total_Revolving_Bal and Avg_Utilisation_Ratio
  - Customers with an income above $120K and those with an income below $40K have higher rates of attrition than compared to customers who have an income within the range of $40K-$120K
  - Customers having 1 or 2 relationship counts / products have a higher rate of attrition that compared to those who have a relationship count of 3 and above
  - Nearly half of the customers that have been fully active over a 12-month period have attrited than compared to those customers who have been inactive for longer periods

EDA - Bivariate Analysis – Key Observations & Insights (Cont'd)

- **Key Insights :**
  - The attrition levels of females is slightly higher than males.
  - The attrition levels is high for single and divorced than compared to married customers
  - The attrition levels for advanced degrees such as Doctorate and Postgraduates is much higher than compared to other degree / non-degree educated customers
  - Customers having 1 or 2 relationship counts / products have a higher rate of attrition that compared to those who have a relationship count of 3 and above
  - Customers having 3 or more dependents have a higher rate of attrition that compared to those who having less than 3 dependents
  - High concentration of customers who attrited were observed with:
    - A lower total transaction amount
    - A lower total transaction count
    - A lower utilization ratio
    - A lower transaction count change Q4 to Q1
    - A significantly higher contacts with or by the bank

Data Preprocessing

Data Preprocessing – Key Observations & Insights

- **Outlier Check & Treatment :** There are several outliers. Credit_Limit, Avg_Open_To_Buy and Total_Trans_Amt have approx. 9% outliers, whereas Months_on_book, Months_Inactive_12_mon, Total_Amt_Chng_Q4_Q1 and Total_Ct_Chng_Q4_Q1 have approx. 3% outliers. Contacts_Count_12_mon has 6.2% of outliers, whereas Customer_Age, Dependent_count, Total_Relationship_Count, Total_Revolving_Bal, Total_Trans_Ct and Avg_Utilization_Ratio has almost neglible outliers

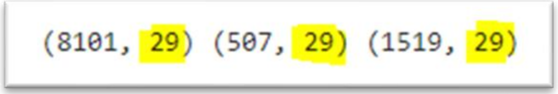
- **Feature Engineering:**
  - CLIENTNUM variable has been dropped since it does not add value to the dataset
  - Attrition_Flag values such “Existing Customer” and “Attrited Customer” have been repalced with 0 and 1 respectively
  - Income_Category has 1,112 anomalous values such as “abc” . These have been initially replaced with NAN values. and then have been imputed with the “most frequent” occurring value using the simple imputer function

- **Dataset Segregation:** The data has been split into Training, Validation & Testing datasets.

- **Missing Value Treatment:** Education_Level has 1,519 N/A values, Marital_Status has 749 NAN values, and Income_Category now has 1,112 NAN values. All these values have been imputed in the segregated Training, Validation & Testing datasets with the “most frequent” occurring values for their respective columns using the simple imputer function

- **Categorical Variables:** Categorical variables within the Training, Validation & Testing have been encoded, which has added 8 new columns (29 total columns) to each dataset.

- **Dataset Shape:** Training, Validation & Testing dataset snapshot respectively



```
(8101, 29) (507, 29) (1519, 29)
```

Model Building

Model Evaluation Criteria & Approach

- **Model Evaluation Criteria :** The primary objective for building the model is to predict whether an existing customer will attrit or not and the key reasons for leaving the Credit Card Services. Using the confusion matrix as guiding principle, it is imperative to focus on reducing the False Negatives (FN) i.e., predicting that a customer will not attrite, but eventually attrites the Credit Card Services. Losing an existing customer would be a significant loss of revenue to the Bank. So, if FN is high, that means the attrition will be high. This implies that **reducing False Negatives** should be of utmost importance to the business
  - Key Criteria – Recall: The bank should therefore use Recall as the key model evaluation criteria – higher the Recall, greater are the chances of minimising False Negatives

- **Model Building Approach:** We have split the data into Training, Validation and Testing datasets, and built 5 base models.
  - Model 1 - Bagging
  - Model 2 - Random Forest
  - Model 3 - Gradient Boosting (GBM)
  - Model 4 - AdaBoost
  - Model 5 - Decision Tree (DTree).

  Using the original data, and Over-Sampling & Under-Sampling techniques, we have evaluated the performance (Recall) of the above models on Training & Validation datasets. Based on their **Recall** scores, we have **shortlisted** the **3 best performing** models (**1-GBM using Original dataset, 2-GBM using Under-Sampled dataset and 3-AdaBoost using Original dataset**) and tuned their hyper-parameters to achieve the best performance. We have then selected the **best model (GBM using Under-Sampled dataset)** and evaluated its performance on the Testing dataset

Model Building – Original, Oversampled & Undersampled Data

- **Model Performance – Original Dataset:** Following is the model performance on the original dataset. GBM has the best performance score followed by Adaboost as per the Validation performance

#	Model Type	Original Dataset (Recall Score)		
		Training Dataset	Validation Dataset	Difference
1	Model – 1: Bagging	0.9838	0.8148	0.1691
2	Model – 2: Random Forest	1.0	0.7530	0.2469
3	Model – 3: Gradient Boosting (GBM)	0.8840	0.9012	-0.0172
4	Model – 4: AdaBoost	0.8471	0.8641	-0.0170
5	Model – 5: Decision Tree (DTree)	1.0	0.8024	0.1975

- **Model Performance – Oversampled Dataset:** Following is the model performance on the oversampled dataset. GBM has the best performance score followed by Adaboost as per the Validation performance

#	Model Type	Oversampled Dataset (Recall Score)		
		Training Dataset	Validation Dataset	Difference
1	Model – 1: Bagging	0.9986	0.8765	0.1221
2	Model – 2: Random Forest	1.0	0.8888	0.1111
3	Model – 3: Gradient Boosting (GBM)	0.9810	0.9382	0.0428
4	Model – 4: AdaBoost	0.9670	0.8888	0.0782
5	Model – 5: Decision Tree (DTree)	1.0	0.8024	0.1975

Model Building – Original, Oversampled & Undersampled Data

- **Model Performance – Undersampled Dataset:** GBM has the best performance score followed by Adaboost

#	Model Type	Undersampled Dataset (Recall Score)		
		Training Dataset	Validation Dataset	Difference
1	Model – 1: Bagging	0.9946	0.9382	0.0564
2	Model – 2: Random Forest	1.0	0.9506	0.0494
3	Model – 3: Gradient Boosting (GBM)	0.9823	0.9382	0.0441
4	Model – 4: AdaBoost	0.9516	0.9506	0.0010
5	Model – 5: Decision Tree (DTree)	1.0	0.9135	0.0864

- **Model Performance Comparison:** Following is the performance comparison of all models for their Recall scores on Training & Validation datasets. The model performance was optimised and evaluated using original (no-sampling), oversampled and undersampled data. **Gradient Boosting (GBM) had the best performance followed by AdaBoost**

#	Model Type	Original Dataset		Oversampled Dataset		Undersampled Dataset	
		Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset
1	Model – 1: Bagging	0.9838	0.8148	0.9986	0.8765	0.9946	0.9382
2	Model – 2: Random Forest	1.0	0.7530	1.0	0.8888	1.0	0.9506
3	Model – 3: Gradient Boosting (GBM)	0.8840 ①	0.9012	0.9810	0.9382	0.9823 ②	0.9382
4	Model – 4: AdaBoost	0.8471 ③	0.8641	0.9670	0.8888	0.9516	0.9506
5	Model – 5: Decision Tree (DTree)	1.0	0.8024	1.0	0.8024	1.0	0.9135

Hyper Parameter Tuning 3 Best Models

Based on the Recall score and the performance on the Validation dataset, we have shortisted the following 3 best models. Hyper Parameter Tuning these models would give us the 'best' performing model, which can then be used on the Testing datset.

- **Hyper Tuning – GBM Using Original Data:** Following are the hyper-parameters:
  - n_estimators : 125, max_features : 0.7, learning_rate : 0.2, CV score = 0.8671087533156498, subsample : 0.7
- **Hyper Tuning – GBM Using Undersampled Data:** Following are the hyper-parameters:
  - n_estimators : 125, max_features : 0.7, learning_rate : 0.2, CV score = 0.9523902151488359, subsample : 0.7
- **Hyper Tuning – AdaBoost Using Original Data:** Following are the hyper-parameters:
  - n_estimators : 100, max_depth : 2, learning_rate : 1, CV score = 0.8893427645151781, subsample : 0.7

Following are the scores of the 3 Hyper Tuned models on the Training & Validation dataset.

#	Model Type - Hyper Tuned	Accuracy		Recall		Precision		F1	
		Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset
1	Model – 3: GBM using Original Data	0.988	0.984	0.948	0.926	0.978	0.974	0.963	0.949
2	Model – 3: GBM using Undersampled Data	0.996	0.955	0.998	0.975	0.994	0.790	0.996	0.873
3	Model – 4: AdaBoost using Original Data	0.993	0.982	0.975	0.914	0.982	0.974	0.978	0.943

Model Performance Comparison Summary

- **Model Performance without Hyper Parameter Tuning:** Following is the performance of all models for their Recall scores on Training & Validation datasets. The model performance was optimised and evaluated using original (no-sampling), oversampled and undersampled data. Based on the Recall score, Gradient Boosting (GBM) had the best performance followed by AdaBoost

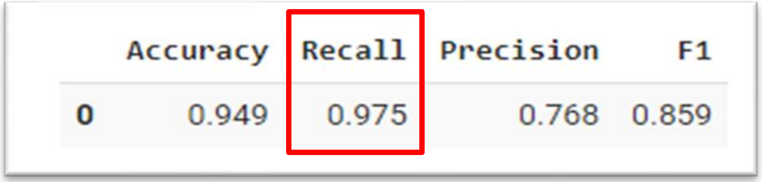
#	Model Type	Original Dataset		Oversampled Dataset		Undersampled Dataset	
		Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset
1	Model – 1: Bagging	0.9838	0.8148	0.9986	0.8765	0.9946	0.9382
2	Model – 2: Random Forest	1.0	0.7530	1.0	0.8888	1.0	0.9506
3	Model – 3: Gradient Boosting (GBM)	0.8840	0.9012	0.9810	0.9382	0.9823	0.9382
4	Model – 4: AdaBoost	0.8471	0.8641	0.9670	0.8888	0.9516	0.9506
5	Model – 5: Decision Tree (DTree)	1.0	0.8024	1.0	0.8024	1.0	0.9135

- **Model Performance - Hyper Parameter Tuned:** Optimised performance of GBM & AdaBoost on Training & Validation Datasets

#	Model Type - Hyper Tuned	Accuracy		Recall		Precision		F1	
		Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset	Training Dataset	Validation Dataset
1	Model – 3: GBM using Original Data	0.988	0.984	0.948	0.926	0.978	0.974	0.963	0.949
2	Model – 3: GBM using Undersampled Data	0.996	0.955	0.998	0.975	0.994	0.790	0.996	0.873
3	Model – 4: AdaBoost using Original Data	0.993	0.982	0.975	0.914	0.982	0.974	0.978	0.943

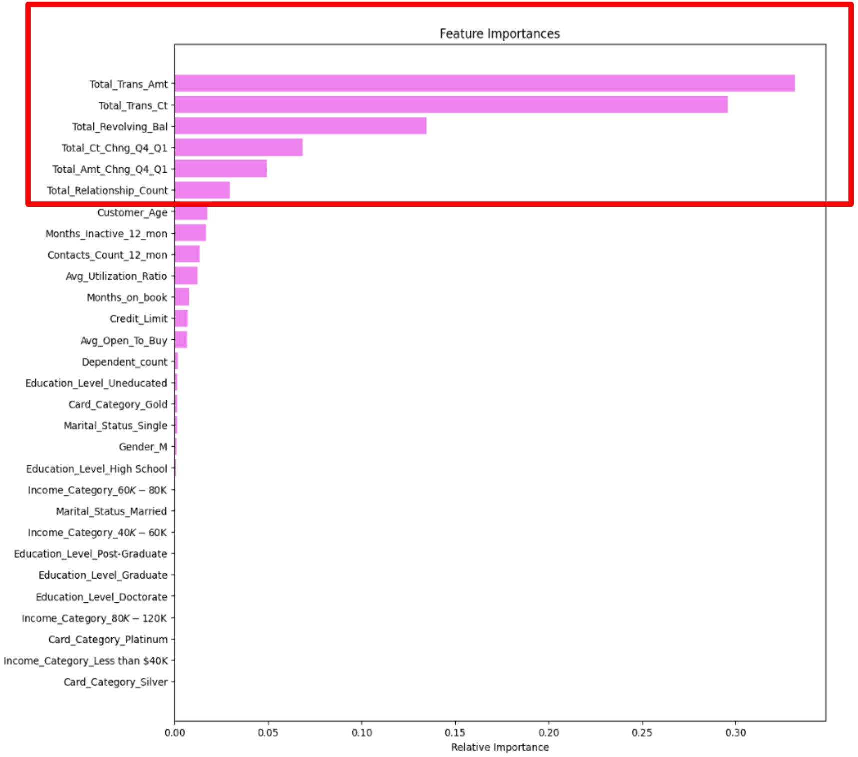
Model Performance Summary – Best Model

- **Best Model - Model – 3: GBM using Undersampled Data:** With a Recall score of 99.8% and 97.5% on Training & Validation datasets the Gradient Boosting Model using Undersampled dataset has generalised its performance and is the best model. **The model has delivered a score of 97.5% on the Testing dataset**



	Accuracy	Recall	Precision	F1
0	0.949	0.975	0.768	0.859

- **Most Important Features:** Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Total_Amt_Chng_Q4_Q1, and Total_Relationship_Count are the most important features



Executive Summary - Conclusion

- Of all the models, Model – 3: GBM using Undersampled Data (post hyper-tuning) seems to be the best fit model for Thera Bank.

- The model has high Recall score of 0.975 on the Testing dataset. This is in line with the Recall scores of 0.998 and 0.975 achieved on the Training & Validation datasets, respectively. With such high Recall scores, the model will minimise False Negatives, which is of utmost importance to the business.

- The model built can be used to predict whether a customer will attrit or not., which will help the bank to target the potential customers, who have a higher probability of attriting, and proactively incentivise them in order to maximise customer retention

- The Bank should focus on improving key services that are related to Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Total_Amt_Chng_Q4_Q1, and Total_Relationship_Count features

Executive Summary - Key Actionable Business Insights

- **Summarised Key Observations :**
  - There is a significant imbalance in dataset since there are high number of existing customers (~84%) than compared to 16% of attrited customers
  - High concentration of customers who attrited were observed with 1) A lower total transaction amount, 2) A lower total transaction count, 3) A lower utilization ratio, 4)A lower transaction count change Q4 to Q1 and 5) A significantly higher contacts with or by the bank
  - Majority of customers are in Blue card category and very few in Gold & Platinum category, which implies that customers are not using their credit card as much as possible and there are opportunities for service improvement
  - Average Open to Buy has lots of high-end outliers, which implies that there are customers who use only very small amount of their credit limit
  - Approx. 65% of customers have been contacted by the bank very often, which implies that that there have been issues where the customer needed support
  - The attrition levels of females is slightly higher than males
  - The attrition levels is high for single and divorced than compared to married customers
  - The attrition levels for advanced degrees such as Doctorate and Postgraduates is much higher than compared to other degree / non-degree educated customers

Executive Summary - Key Actionable Business Insights

- **Summarised Key Insights :**
  - Customer attributes such as Total_Trans_Amt, Total_Trans_Ct, Total_Revolving_Bal, Total_Ct_Chng_Q4_Q1, Total_Amt_Chng_Q4_Q1, and Total_Relationship_Count are the most important features in predicting potential customers that will attrit. These features are negatively correlated with the Attrition_Flag, which implies that lower the values of these features, the higher the chances of a customer to attrite
  - Bank should increase the frequency of customer contact and provide them with various incentives and schemes to increase relationships of the customer with the bank
  - Bank should incentivise customers with cashback schemes on using credit cards, which might encourage customers on using their credit cards more often
  - Bank should also increase the credit limit for customers who regularly us their credit cards, which will lead to an increase in credit card spend / transaction amounts
  - The banks could potentially introduce an annual 0% interest free offer on large products, which will encourage customers to buy high value items using their credit cards. The payments towards this can be made on a monthly basis using a credit card. This would increase the total transaction amount, transaction counts and the revolving balance.

Executive Summary - Our Recommendation

Based on our key observations and insights, we recommend the following areas of improvement / opportunities that will drive business growth and lead to a better customer experience

- **Implement Customer Incentivisation Scheme:** Incentivising customers by offering them cashback schemes and discounts / vouchers on credit card purchases will encourage frequent spending and will drive customer gowth and increase revenue
- **Implement Tier based Rewards:** The Bank should introduce a Tier based Loyalty & Rewards Scheme for credit card purchases. Cumulative loyalty points above a certain threshold will promote the customer to a new tier, that will offer specific rewards such as First-Class Lounge access at Airports, Spa & Well-Being discounts etc
- **Implement Customer Satisfaction Survey:** The Bank should initiate a targeted Customer Satisfaction Survey to understand customer pain points and implement the findings to improve retention ratio of such customers

APPENDIX

Appendix - Notes

- Further analysis would be required on a comprehensive dataset to provide customer segmentation strategies

Great Learning

Happy Learning !