CREDIT CARD CUSTOMER CHURN PREDICTION

COURSE PROJECT (SI- 515)

DATA MINING

SHIV YADAV 22N0064 | MSC ASI

UNDER THE GUIDANCE OF PROF. SANJEEV V. SABNIS & DR. RADHENUSHKA SRIVASTAV (IIT- BOMBAY)

Background and Context:

A manager at the bank is disturbed with more and more customers leaving their credit card services. They would really appreciate if one could predict for them who is going to get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction.
we need to come up with a classification model that will help the bank improve its services so that customers do not renounce their credit cards.

Objective:

- **Explore and visualize the dataset (EDA).**
- **Build different classification model (logistics, LDA, QDA, KNN, Decision Tree and Random Forest) to predict if the customer is going to churn or not**
- **Optimize the model using appropriate techniques**
- **Generate a set of insights and recommendations that will help the bank**

Data Description: https://leaps.analyttica.com/home

I got this dataset from a website with the URL as https://leaps.analyttica.com/home.

Also, the dataset is present on Kaggle:

https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

1. **CLIENTNUM**: Client number. Unique identifier for the customer holding the account

2. **Attrition_Flag:** Internal event (customer activity) variable - if the account is closed then "Attired Customer" else "Existing Customer"

3. **Customer_Age:** Age in Years

4. **Gender:** Gender of the account holder

5. **Dependent_count:** Number of dependents

6. **Education_Level:** Educational Qualification of the account holder - Graduate, High School, Unknown, Uneducated, College (refers to a college student), Post-Graduate, Doctorate.

7. **Marital_Status:** Marital Status of the account holder

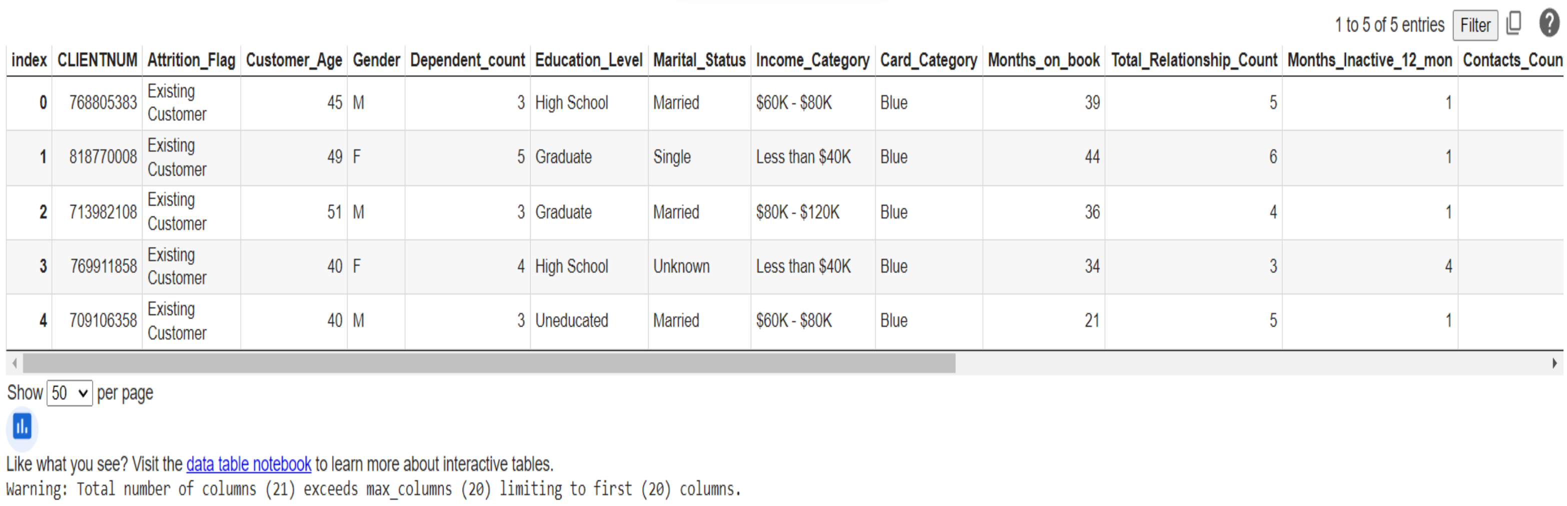
8. **Income_Category:** Annual Income Category of the account holder

9. **Card_Category:** Type of Card
10. **Months_on_book:** Period of relationship with the bank
11. **Total_Relationship_Count:** Total no. of products held by the customer
12. **Months_Inactive_12_mon**: No. of months inactive in the last 12 months
13. **Contacts_Count_12_mon:** No. of Contacts between the customer and bank in the last 12 month
14. **Credit_Limit:** Credit Limit on the Credit Card
15. **Total_Revolving_Bal:** The balance that carries over from one month to the next is the revolving balance
16. **Avg_Open_To_Buy:** Open to Buy refers to the amount left on the credit card to use (Average of last 12 months)
17. **Total_Trans_Amt:** Total Transaction Amount (Last 12 months)
18. **Total_Trans_Ct:** Total Transaction Count (Last 12 months)
19. **Total_Ct_Chng_Q4_Q1:** Ratio of the total transaction count in 4th quarter and the total transaction count in 1st
20. **Total_Amt_Chng_Q4_Q1:** Ratio of the total transaction amount in 4th quarter and the total transaction amount in 1st quarter
21. **Avg_Utilization_Ratio:** Represents how much of the available credit the customer spent

this dataset consists of 10,000 customers mentioning their age, salary, marital status, credit card limit, credit card category, etc. There are nearly 20 features.

We have only 16.07% of customers who have churned. Thus, it's a bit difficult to train our model to predict churning customers.

Glimpse of Customer Churn Data set:



1 to 5 of 5 entries Filter

index	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Coun
0	768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue	39	5	1	
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	6	1	
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue	36	4	1	
3	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	3	4	
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K - $80K	Blue	21	5	1	

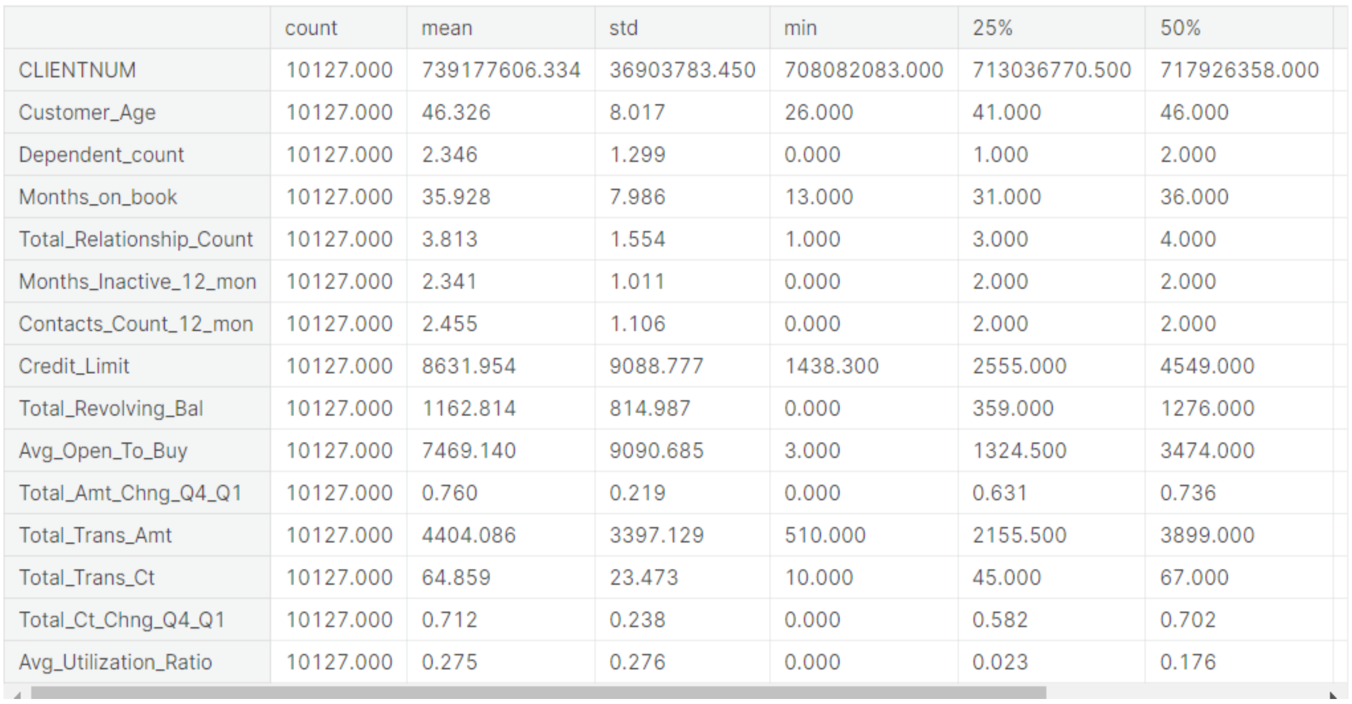
Show 50 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

```
Warning: Total number of columns (21) exceeds max_columns (20) limiting to first (20) columns.
```

Shape of Data: (10127, 23)

statistical summary of the numerical columns in the data:



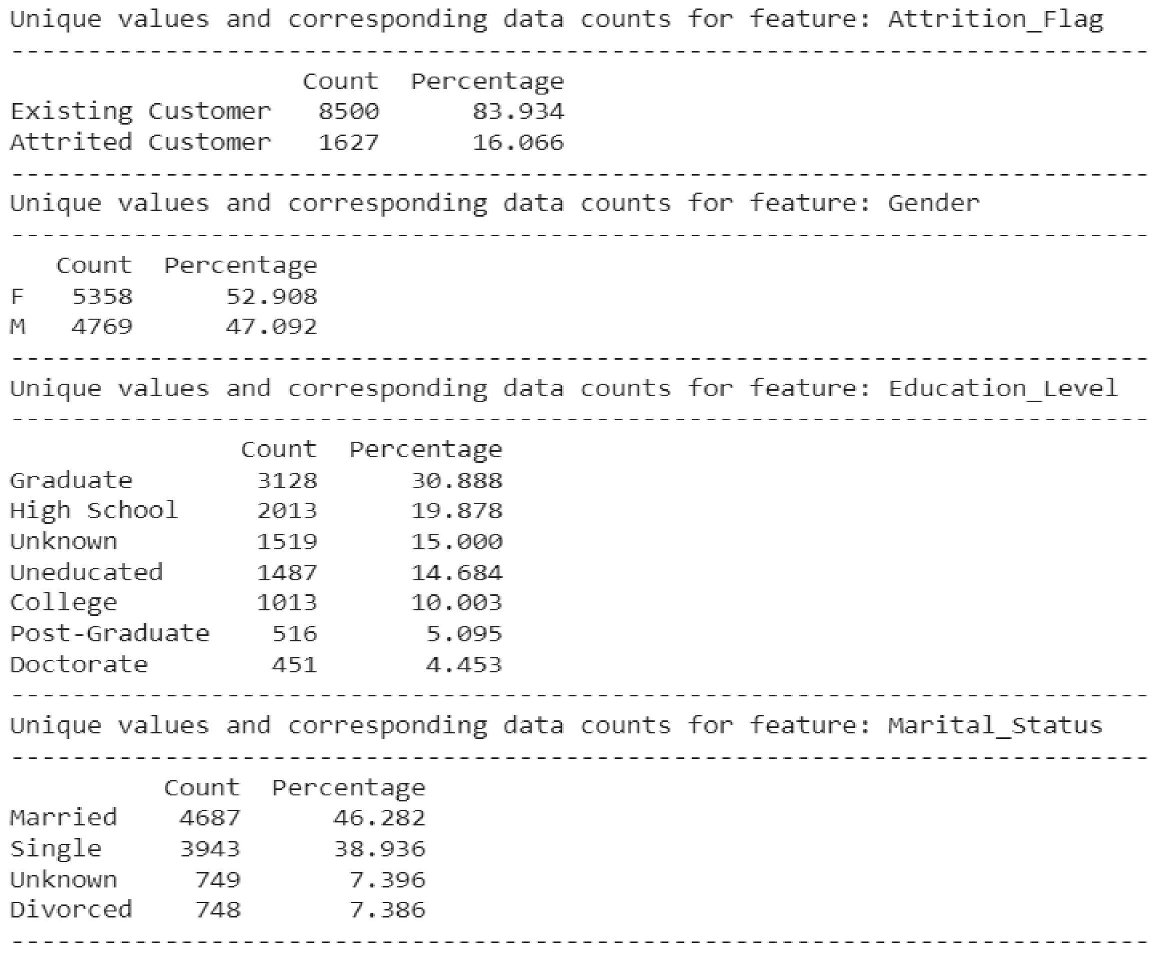
	count	mean	std	min	25%	50%
CLIENTNUM	10127.000	739177606.334	36903783.450	708082083.000	713036770.500	717926358.000
Customer_Age	10127.000	46.326	8.017	26.000	41.000	46.000
Dependent_count	10127.000	2.346	1.299	0.000	1.000	2.000
Months_on_book	10127.000	35.928	7.986	13.000	31.000	36.000
Total_Relationship_Count	10127.000	3.813	1.554	1.000	3.000	4.000
Months_Inactive_12_mon	10127.000	2.341	1.011	0.000	2.000	2.000
Contacts_Count_12_mon	10127.000	2.455	1.106	0.000	2.000	2.000
Credit_Limit	10127.000	8631.954	9088.777	1438.300	2555.000	4549.000
Total_Revolving_Bal	10127.000	1162.814	814.987	0.000	359.000	1276.000
Avg_Open_To_Buy	10127.000	7469.140	9090.685	3.000	1324.500	3474.000
Total_Amt_Chng_Q4_Q1	10127.000	0.760	0.219	0.000	0.631	0.736
Total_Trans_Amt	10127.000	4404.086	3397.129	510.000	2155.500	3899.000
Total_Trans_Ct	10127.000	64.859	23.473	10.000	45.000	67.000
Total_Ct_Chng_Q4_Q1	10127.000	0.712	0.238	0.000	0.582	0.702
Avg_Utilization_Ratio	10127.000	0.275	0.276	0.000	0.023	0.176

Remark:

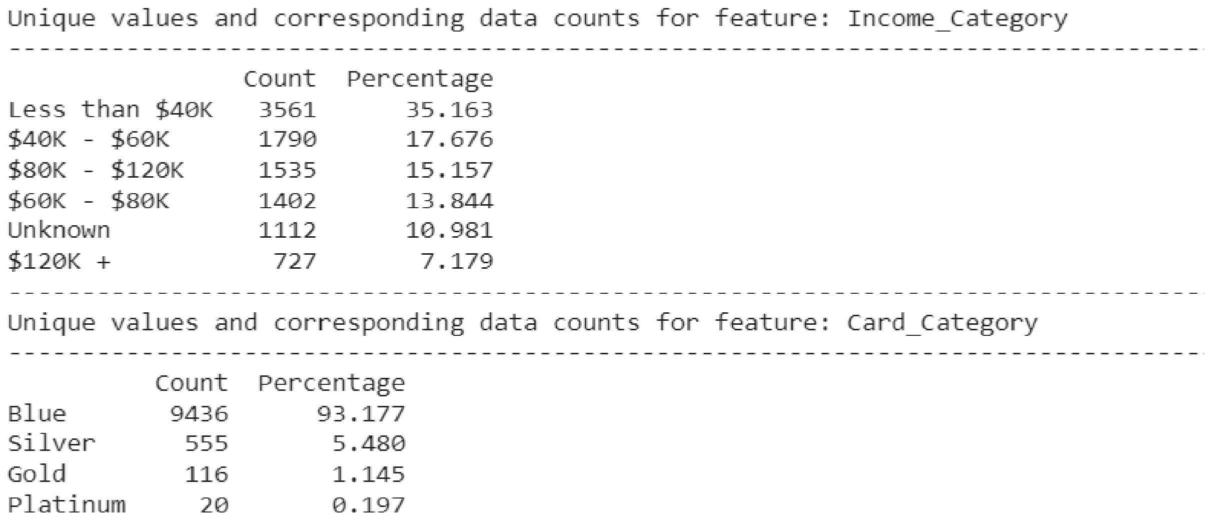
- Mean value for the Customer Age column is approx. 46 and the median is also 46. This shows that majority of the customers are under 46 years of age.

- Dependent Count column has mean and median of ~2
- Months on Book column has mean and median of 36 months. Minimum value is 13 months, showing that the dataset captures data for customers with the bank at least 1 whole years
- Total Relationship Count has mean and median of ~4
- Credit Limit has a wide range of 1.4K to 34.5K, the median being 4.5K, way less than the mean 8.6K
- Total Transaction Count has mean of ~65 and median of 67

Similarly for categorical Column:



```
Unique values and corresponding data counts for feature: Attrition_Flag
--------------------------------------------------------------------------
                    Count  Percentage
Existing Customer    8500      83.934
Attrited Customer    1627      16.066
--------------------------------------------------------------------------
Unique values and corresponding data counts for feature: Gender
--------------------------------------------------------------------------
   Count  Percentage
F   5358      52.908
M   4769      47.092
--------------------------------------------------------------------------
Unique values and corresponding data counts for feature: Education_Level
--------------------------------------------------------------------------
               Count  Percentage
Graduate        3128      30.888
High School     2013      19.878
Unknown         1519      15.000
Uneducated      1487      14.684
College         1013      10.003
Post-Graduate    516       5.095
Doctorate        451       4.453
--------------------------------------------------------------------------
Unique values and corresponding data counts for feature: Marital_Status
--------------------------------------------------------------------------
          Count  Percentage
Married    4687      46.282
Single     3943      38.936
Unknown     749       7.396
Divorced    748       7.386
--------------------------------------------------------------------------
```



```
Unique values and corresponding data counts for feature: Income_Category
--------------------------------------------------------------------------------
                Count  Percentage
Less than $40K   3561      35.163
$40K - $60K      1790      17.676
$80K - $120K     1535      15.157
$60K - $80K      1402      13.844
Unknown          1112      10.981
$120K +           727       7.179
--------------------------------------------------------------------------------
Unique values and corresponding data counts for feature: Card_Category
--------------------------------------------------------------------------------
          Count  Percentage
Blue       9436      93.177
Silver      555       5.480
Gold        116       1.145
Platinum     20       0.197
```

Remark:

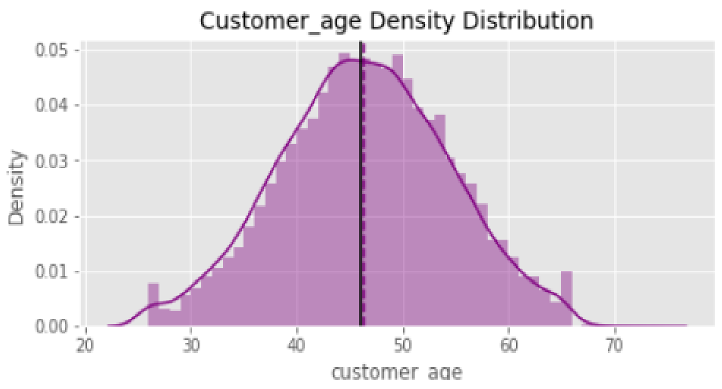
- The target variable Attrition Flag has Existing to Attrited ratio of 83.9 : 16.1. There is imbalance in the dataset
- 93% customers are having Blue Card
- Income Category has a value abc for 10% records, which we'll change to Unknown

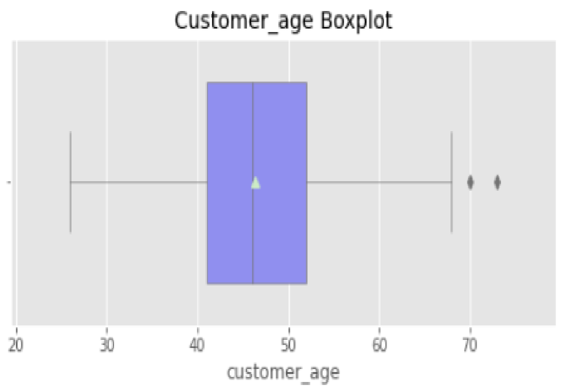
Exploratory Data Analysis

Univariate Analysis of Numerical feature:

1. Customer Age

```
+-------+-------+------+------+------+-------+
|       |   Min |   Q1 |   Q2 |   Q3 |   Max |
|-------+-------+------+------+------+-------|
| Value |    26 |   41 |   46 |   52 |    73 |
+-------+-------+------+------+------+-------+
```

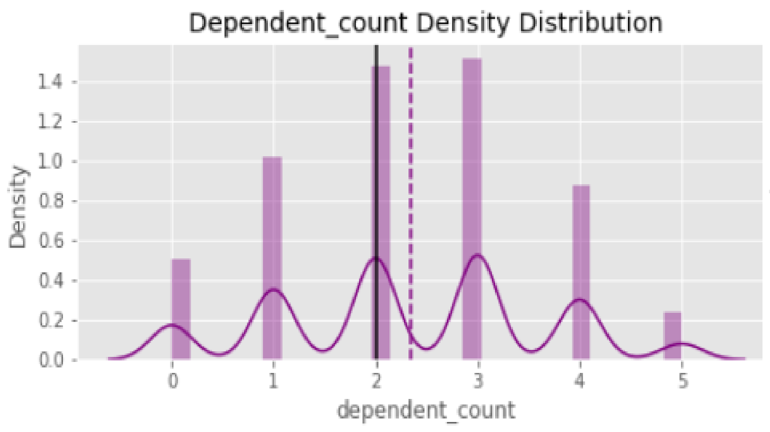


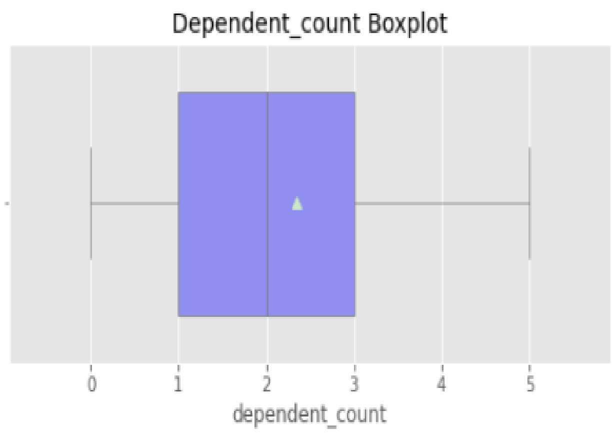


Note: The data is normally distributed, with only 2 outliers on the right side (higher end)

2. Dependent Count

```
5 Point Summary of Dependent_count Attribute:


+-------+-------+------+------+------+-------+
|       |   Min |   Q1 |   Q2 |   Q3 |   Max |
|-------+-------+------+------+------+-------|
| Value |     0 |    1 |    2 |    3 |     5 |
+-------+-------+------+------+------+-------+
```

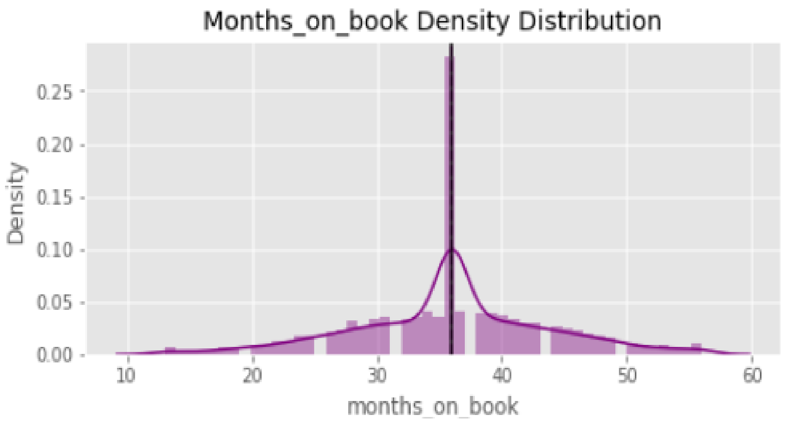


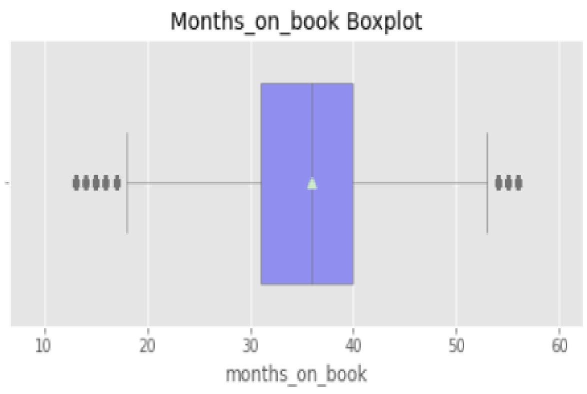


Note: Dependent Count is mostly 2 or 3

3. **Months on Book:**

```
5 Point Summary of Months_on_book Attribute:


+-------+-------+------+------+------+-------+
|       |   Min |   Q1 |   Q2 |   Q3 |   Max |
|-------+-------+------+------+------+-------|
| Value |    13 |   31 |   36 |   40 |    56 |
+-------+-------+------+------+------+-------+
```

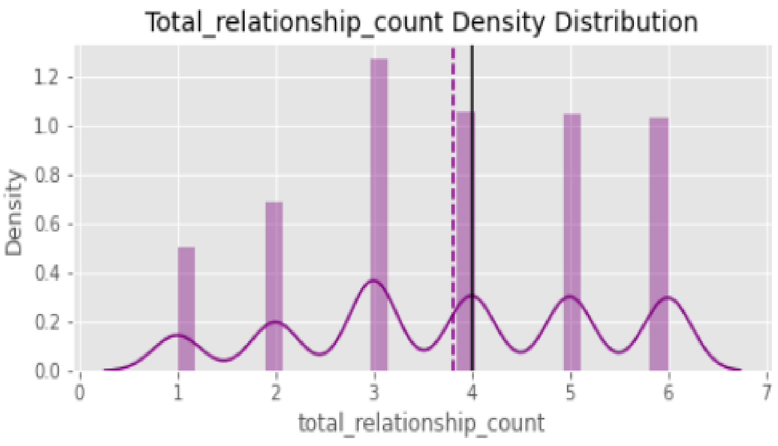


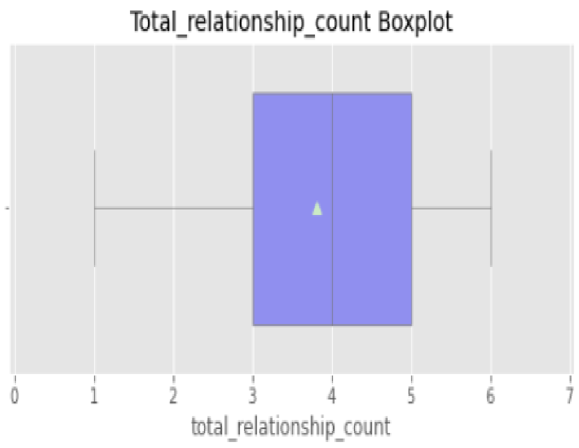


Note: Most customers are on the books for 3 years. There are outliers on both lower and higher end

4. **Total Relationship Count**

```
5 Point Summary of Total_relationship_count Attribute:

+-------+-------+------+------+------+-------+
|       |   Min |   Q1 |   Q2 |   Q3 |   Max |
|-------+-------+------+------+------+-------|
| Value |     1 |    3 |    4 |    5 |     6 |
+-------+-------+------+------+------+-------+
```



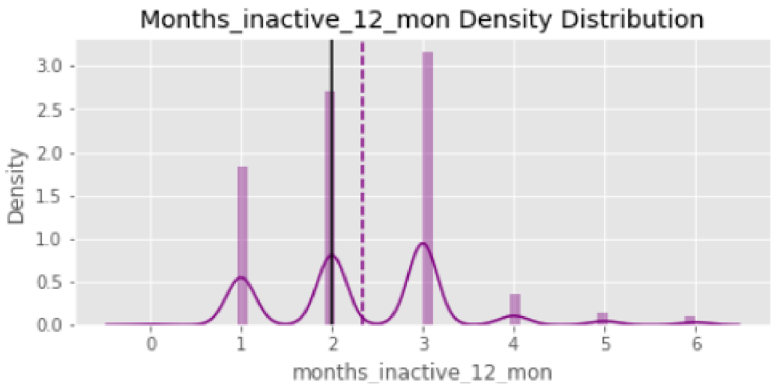


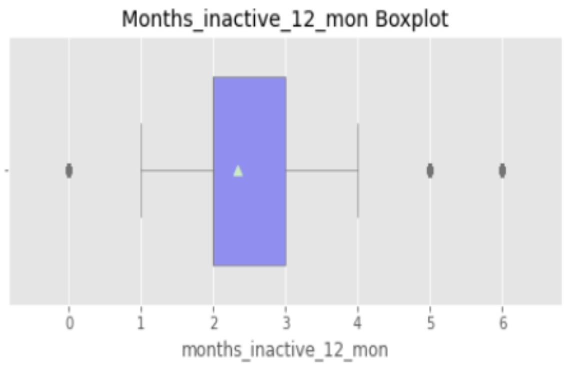
Note: Most of the customers have 4 or more relations with the bank

5. Months Inactive (12 months)

5 Point Summary of Months_inactive_12_mon Attribute:

```
+-------+-------+------+------+------+-------+
|       |   Min |   Q1 |   Q2 |   Q3 |   Max |
|-------+-------+------+------+------+-------|
| Value |     0 |    2 |    2 |    3 |     6 |
+-------+-------+------+------+------+-------+
```



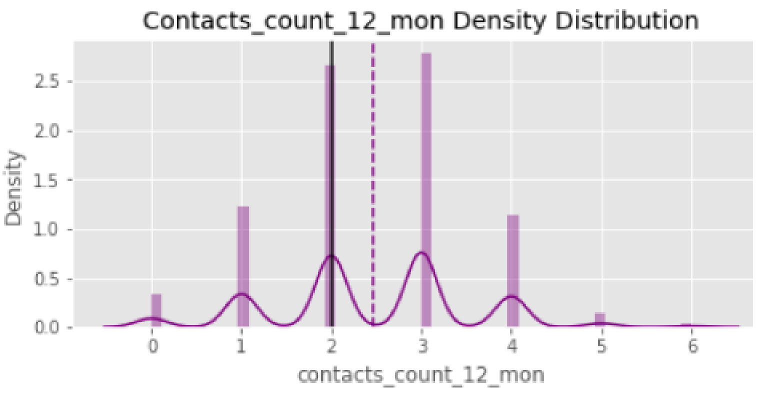


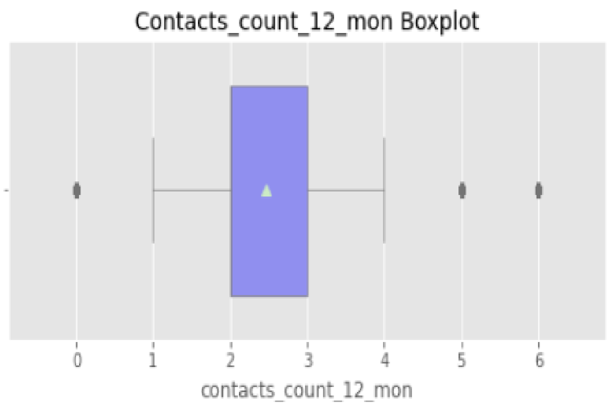
Note: There are lower and higher end outliers for Months inactive in last 12 months.
Lower end outliers are not concerning since 0 value means the customer is always active.
The customers who are inactive for 5 or more months are to be concerned about.

6. Contact Counts (12 months)

5 Point Summary of Contacts_count_12_mon Attribute:

```
+-------+-------+------+------+------+-------+
|       |   Min |   Q1 |   Q2 |   Q3 |   Max |
|-------+-------+------+------+------+-------|
| Value |     0 |    2 |    2 |    3 |     6 |
+-------+-------+------+------+------+-------+
```

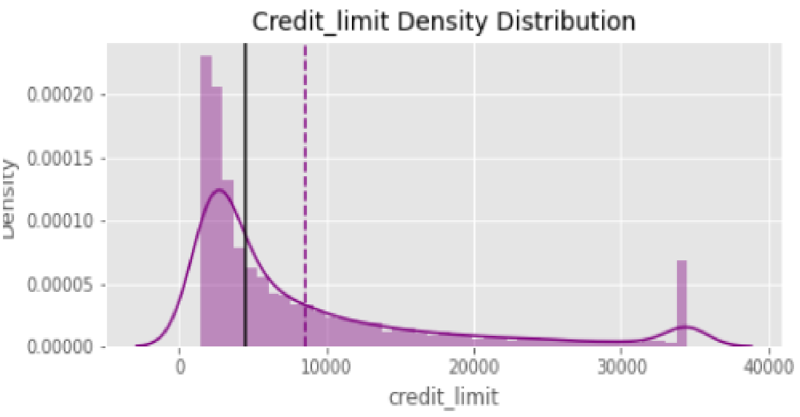


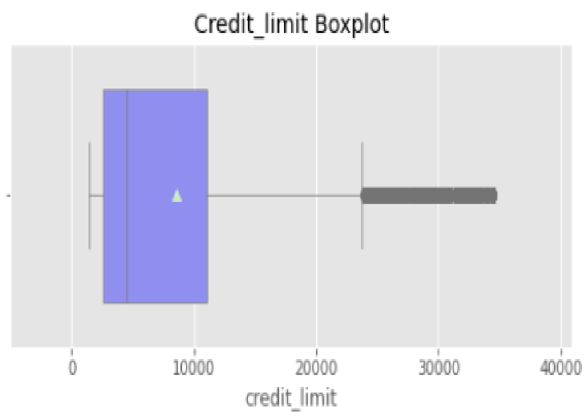


Note: Again lower and higher end outliers are noticed. Here less number of contacts between the bank and the customer should be interesting to be checked

7. Credit Limit

```
5 Point Summary of Credit_limit Attribute:

+-------+--------+------+------+---------+-------+
|       |    Min |   Q1 |   Q2 |      Q3 |   Max |
|-------+--------+------+------+---------+-------|
| Value | 1438.3 | 2555 | 4549 | 11067.5 | 34516 |
+-------+--------+------+------+---------+-------+
```

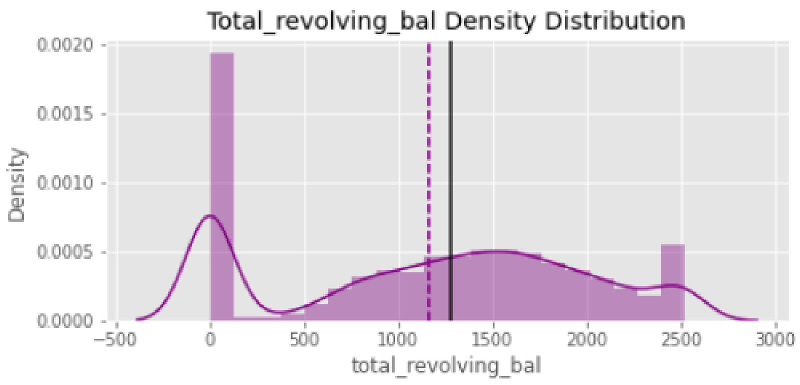


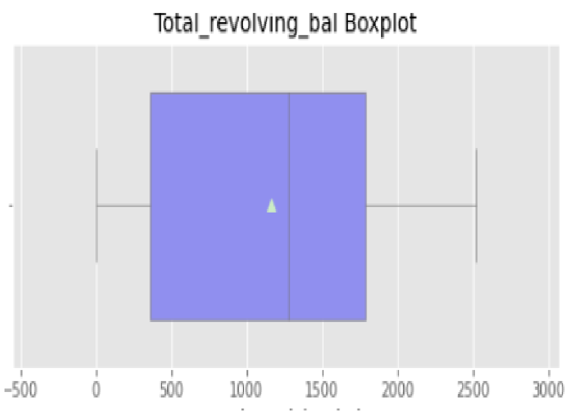


Note: There are higher end outliers in Credit Limit. This might be because the customers are high end.

8. Total revolving balance:

```
5 Point Summary of Total_revolving_bal Attribute:

+-------+-------+------+------+------+-------+
|       |   Min |   Q1 |   Q2 |   Q3 |   Max |
|-------+-------+------+------+------+-------|
| Value |     0 |  359 | 1276 | 1784 |  2517 |
+-------+-------+------+------+------+-------+
```

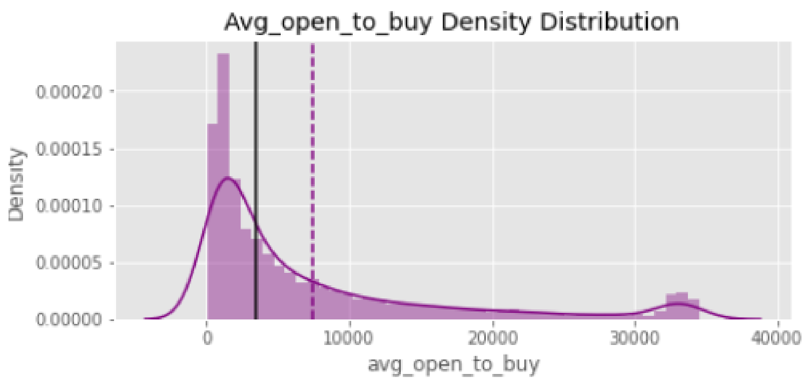


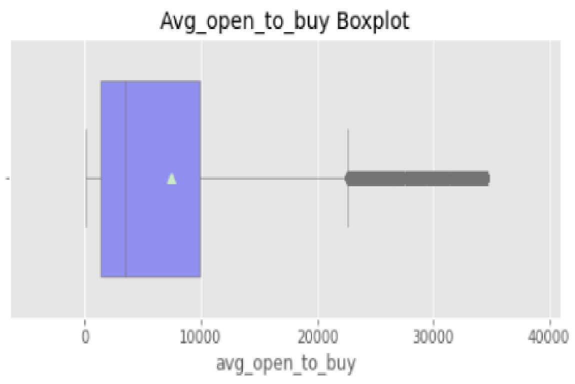


Note: Total revolving balance of 0 would mean the customer never uses the credit card

9. **Average open to buy**

```
5 Point Summary of Avg_open_to_buy Attribute:


+-------+-------+--------+------+------+-------+
|       |   Min |     Q1 |   Q2 |   Q3 |   Max |
|-------+-------+--------+------+------+-------|
| Value |     3 | 1324.5 | 3474 | 9859 | 34516 |
+-------+-------+--------+------+------+-------+
```

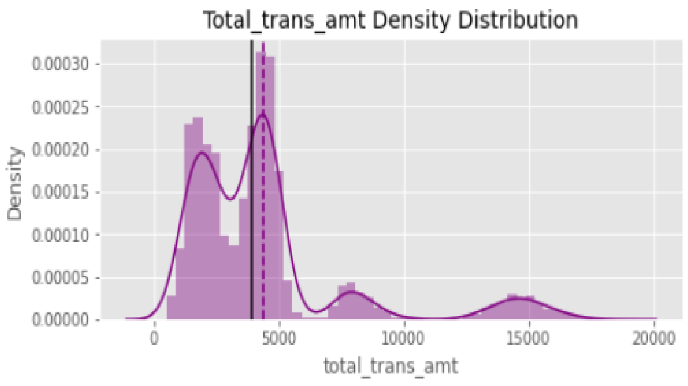


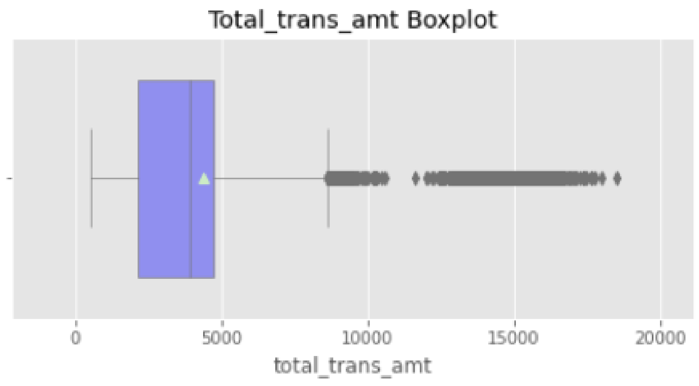


Note : Average Open to Buy has lots of higher end outliers, which means there are customers who uses only very small amount of their credit limit

10. Total transaction amount:

```
5 Point Summary of Total_trans_amt Attribute:


+-------+-------+--------+------+------+-------+
|       |   Min |     Q1 |   Q2 |   Q3 |   Max |
|-------+-------+--------+------+------+-------|
| Value |   510 | 2155.5 | 3899 | 4741 | 18484 |
+-------+-------+--------+------+------+-------+
```

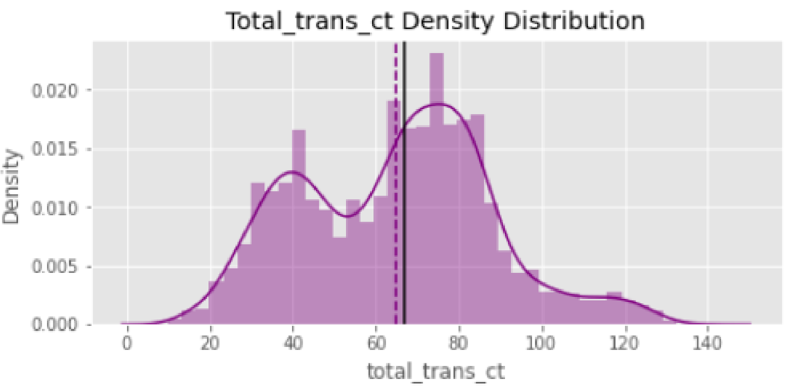


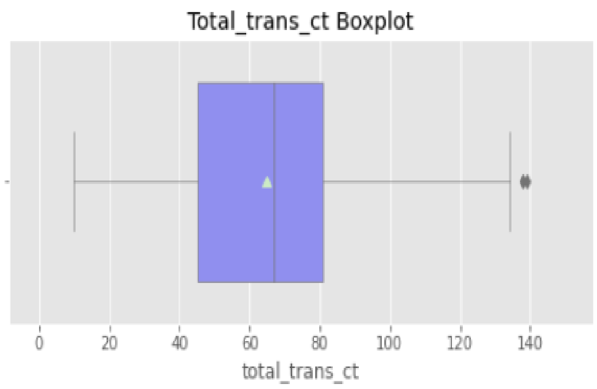


Note : Total Transaction Amount has lots of higher end outliers

11. Total transaction count:

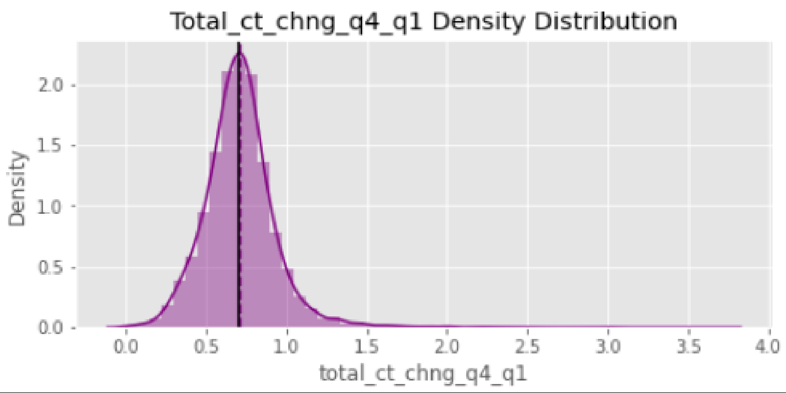
```
5 Point Summary of Total_trans_ct Attribute:

+-------+-------+------+------+------+-------+
|       |   Min |   Q1 |   Q2 |   Q3 |   Max |
|-------+-------+------+------+------+-------|
| Value |    10 |   45 |   67 |   81 |   139 |
+-------+-------+------+------+------+-------+
```

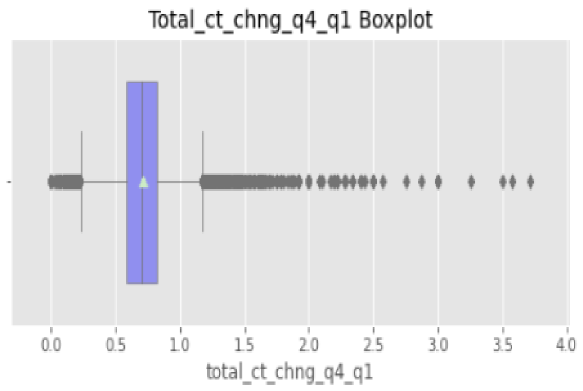




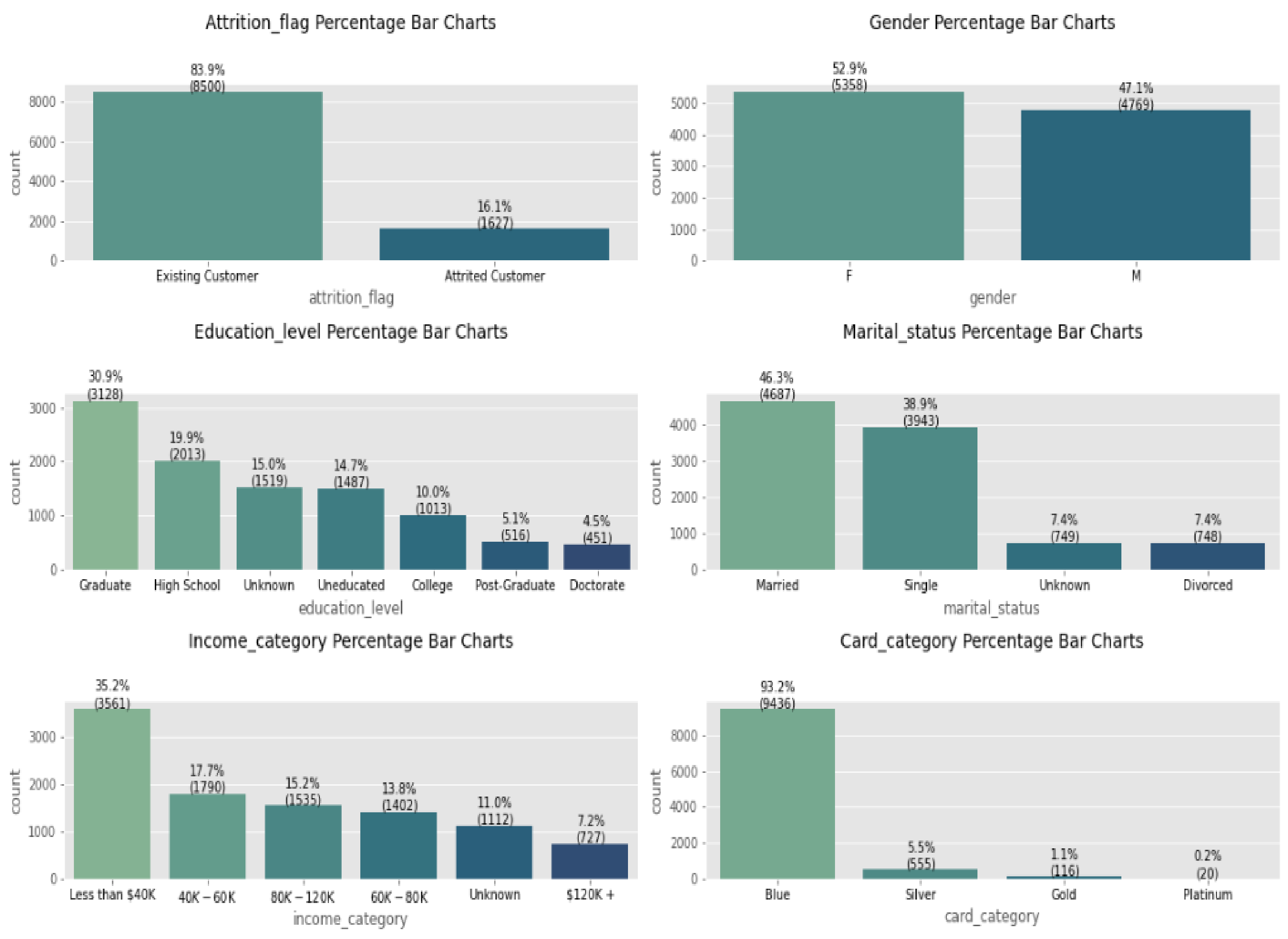
12. Total_ct_chng_q4_q1

```
5 Point Summary of Total_ct_chng_q4_q1 Attribute:

+-------+-------+-------+-------+-------+-------+
|       |   Min |    Q1 |    Q2 |    Q3 |   Max |
|-------+-------+-------+-------+-------+-------|
| Value |     0 | 0.582 | 0.702 | 0.818 | 3.714 |
+-------+-------+-------+-------+-------+-------+
```





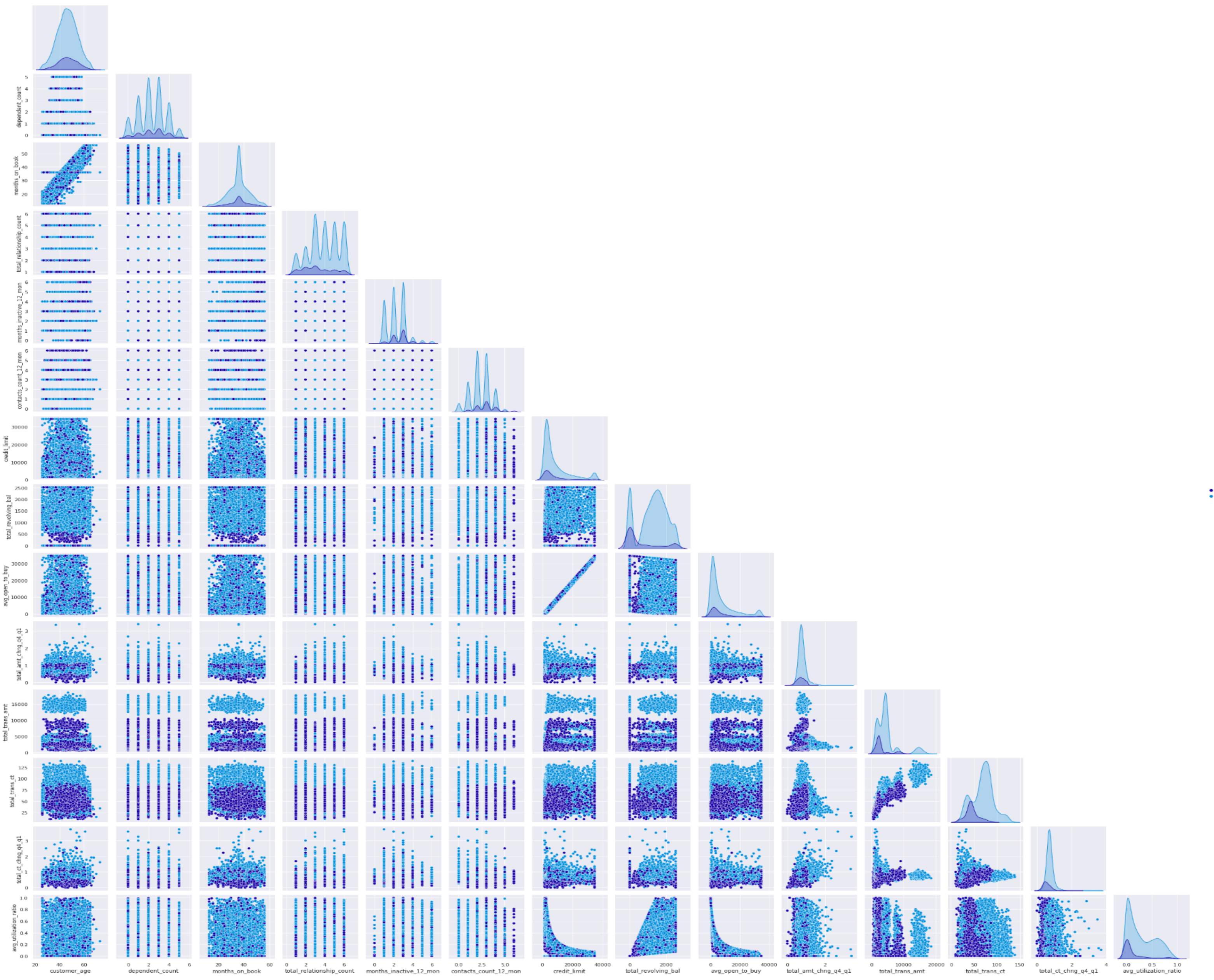
Percentage on bar chart for Categorical Features



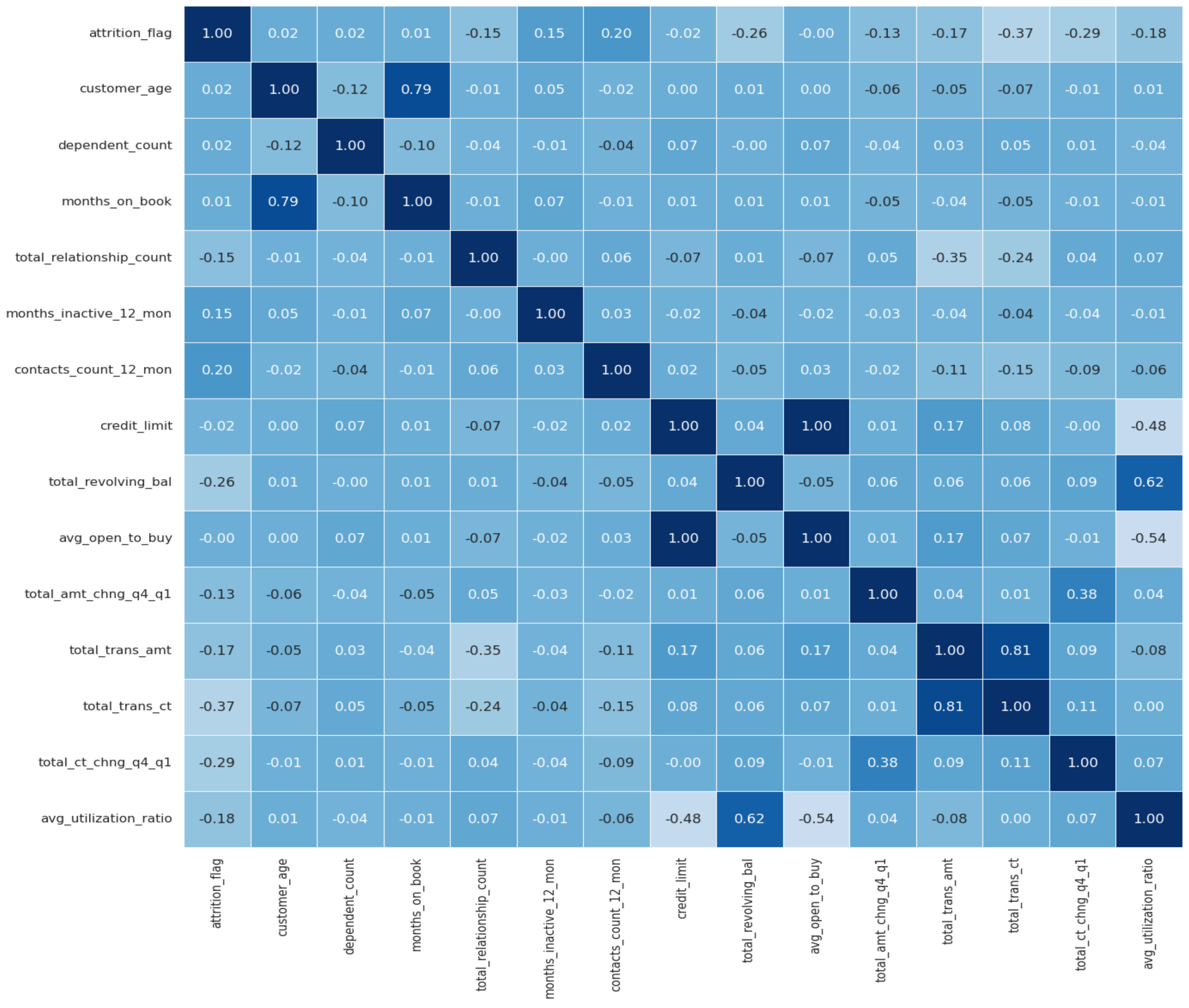
Remark:

- High Imbalance in data since the existing vs. attrited customers ratio is 84:16
- Data is almost equally distributed between Males and Females
- 31% customers are Graduate
- ~85% customers are either Single or Married, where 46.7% of the customers are Married
- 35% customers earn less than $40k and 36% earns $60k or more
- ~93% customers have Blue card

Pair plot of all available numeric columns



Heatmap to understand correlations between independent and dependent variables



	attrition_flag	customer_age	dependent_count	months_on_book	total_relationship_count	months_inactive_12_mon	contacts_count_12_mon	credit_limit	total_revolving_bal	avg_open_to_buy	total_amt_chng_q4_q1	total_trans_amt	total_trans_ct	total_ct_chng_q4_q1	avg_utilization_ratio
attrition_flag	1.00	0.02	0.02	0.01	-0.15	0.15	0.20	-0.02	-0.26	-0.00	-0.13	-0.17	-0.37	-0.29	-0.18
customer_age	0.02	1.00	-0.12	0.79	-0.01	0.05	-0.02	0.00	0.01	0.00	-0.06	-0.05	-0.07	-0.01	0.01
dependent_count	0.02	-0.12	1.00	-0.10	-0.04	-0.01	-0.04	0.07	-0.00	0.07	-0.04	0.03	0.05	0.01	-0.04
months_on_book	0.01	0.79	-0.10	1.00	-0.01	0.07	-0.01	0.01	0.01	0.01	-0.05	-0.04	-0.05	-0.01	-0.01
total_relationship_count	-0.15	-0.01	-0.04	-0.01	1.00	-0.00	0.06	-0.07	0.01	-0.07	0.05	-0.35	-0.24	0.04	0.07
months_inactive_12_mon	0.15	0.05	-0.01	0.07	-0.00	1.00	0.03	-0.02	-0.04	-0.02	-0.03	-0.04	-0.04	-0.04	-0.01
contacts_count_12_mon	0.20	-0.02	-0.04	-0.01	0.06	0.03	1.00	0.02	-0.05	0.03	-0.02	-0.11	-0.15	-0.09	-0.06
credit_limit	-0.02	0.00	0.07	0.01	-0.07	-0.02	0.02	1.00	0.04	1.00	0.01	0.17	0.08	-0.00	-0.48
total_revolving_bal	-0.26	0.01	-0.00	0.01	0.01	-0.04	-0.05	0.04	1.00	-0.05	0.06	0.06	0.06	0.09	0.62
avg_open_to_buy	-0.00	0.00	0.07	0.01	-0.07	-0.02	0.03	1.00	-0.05	1.00	0.01	0.17	0.07	-0.01	-0.54
total_amt_chng_q4_q1	-0.13	-0.06	-0.04	-0.05	0.05	-0.03	-0.02	0.01	0.06	0.01	1.00	0.04	0.01	0.38	0.04
total_trans_amt	-0.17	-0.05	0.03	-0.04	-0.35	-0.04	-0.11	0.17	0.06	0.17	0.04	1.00	0.81	0.09	-0.08
total_trans_ct	-0.37	-0.07	0.05	-0.05	-0.24	-0.04	-0.15	0.08	0.06	0.07	0.01	0.81	1.00	0.11	0.00
total_ct_chng_q4_q1	-0.29	-0.01	0.01	-0.01	0.04	-0.04	-0.09	-0.00	0.09	-0.01	0.38	0.09	0.11	1.00	0.07
avg_utilization_ratio	-0.18	0.01	-0.04	-0.01	0.07	-0.01	-0.06	-0.48	0.62	-0.54	0.04	-0.08	0.00	0.07	1.00

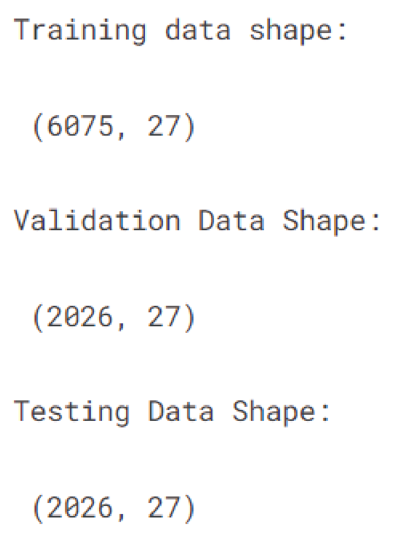
Remark:

- Credit Limit and Average Open to Buy have 100% collinearity
- Months on book and Customer Age have quite strong correlation
- Average Utilization Ration and Total Revolving Balance are also a bit correlated it appears
- Attrition Flag does not have highly strong correlation with any of the numeric variables
- Customer Churn appears to be uncorrelated with Customer Age, Dependent Count, Months on Book, Open to Buy, Credit Limit, we'll remove these from dataset

Data processing

Data pre-processing is one of the most important parts of the job before starting to train the model with the dataset. We need to impute missing values, fix any illogical data value in columns, convert category columns to numeric (either ordinal, or binary using one-hot encoding), scale the data to deal with the distribution skewness and outliers, before feeding the data to a model.

We are using the pre-available transformation classes and the custom classes that we created to first fit the training data and then transform the train, validation and test dataset. This is the standard logical practice to keep the influence of test and validation data in the train dataset to prevent/avoid data leakage while training or validating the model.



```
Training data shape:

 (6075, 27)

Validation Data Shape:

 (2026, 27)

Testing Data Shape:

 (2026, 27)
```

Splited data in Train, Validation and Test sets after preprocessing part.

<u>Model Building</u>

Model evaluation criterion:

Model can make wrong predictions as:

1. Predicting a customer will attrite and the customer does not attrite - Loss of resources
2. Predicting a customer will not attrite and the customer attrites - Loss of opportunity for churning the customer

Which case is more important?

- Predicting that customer will not attrite, but actually attrites, would result in loss for the bank since if predicted correctly, marketing/sales team could have contacted the customer to retain them. This would result in losses. So, the false negatives should be minimized.

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

- Company wants Recall to be maximized, greater the Recall lesser the chances of false negatives.

Let's start by building different models using KFold and cross_val_score and tune the best model using Randomized SearchCV

- Stratified K-Folds cross-validation provides dataset indices to split data into train/validation sets. Split dataset into k consecutive folds (without shuffling by default) keeping the distribution of both classes in each fold the same as the target variable. Each fold is then used once as validation while the k - 1 remaining folds form the training set.

I have built 5 models here, Logistic Regression, LDA, QDA, Decision Tree and Random forest.

1. On Original data:

a.) Logistic Regression:

Logistic Regression is a statistical method used for binary classification problems. It models the probability of the outcome belonging to a specific class, given the predictor variables.

Probability Estimation:

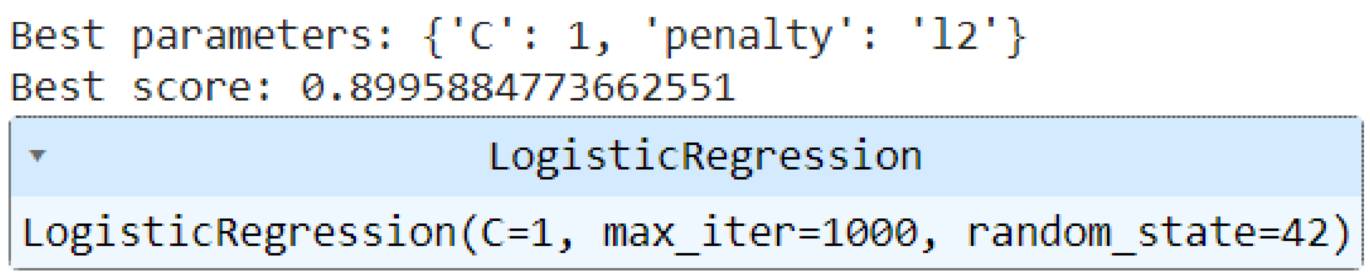
The Logistic Regression model estimates the probability of the binary outcome. It employs the logistic function, also known as the sigmoid function, to map the output to a probability value between 0 and 1.

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\,\dots\dots\dots+\beta_px_p)}}$$

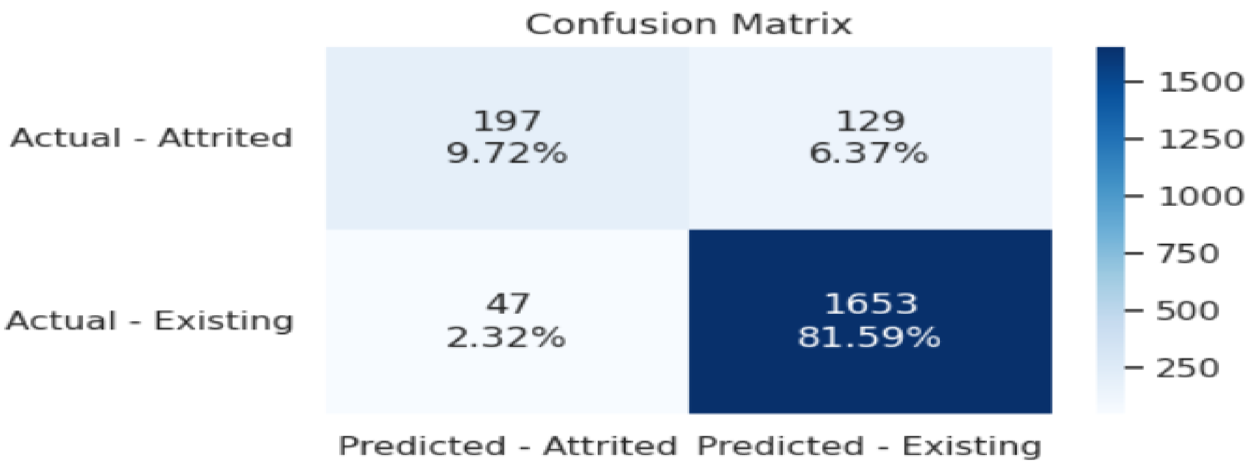
Here, $P(Y=1|X)$ represents the probability of the outcome being class 1 given the predictor variables.

Decision Boundary:

The decision boundary is derived from the logistic function. By setting a threshold (commonly 0.5), predictions are made: if the estimated probability is above the threshold, the outcome is predicted as class 1; otherwise, it's predicted as class 0.


```
Best parameters: {'C': 1, 'penalty': 'l2'}
Best score: 0.8995884773662551
```

```
LogisticRegression(C=1, max_iter=1000, random_state=42)
```



We got TEST Recall = 0.604277

b.) KNN

In K-Nearest Neighbours, a non-parametric algorithm used for classification, the class label of an unclassified sample is predicted based on the classes of its 'k' nearest neighbours.

Distance Metric:

KNN utilizes a distance metric (such as Euclidean, Manhattan, etc.) to measure the proximity between instances in the feature space.

Decision Rule for Binary Classification:

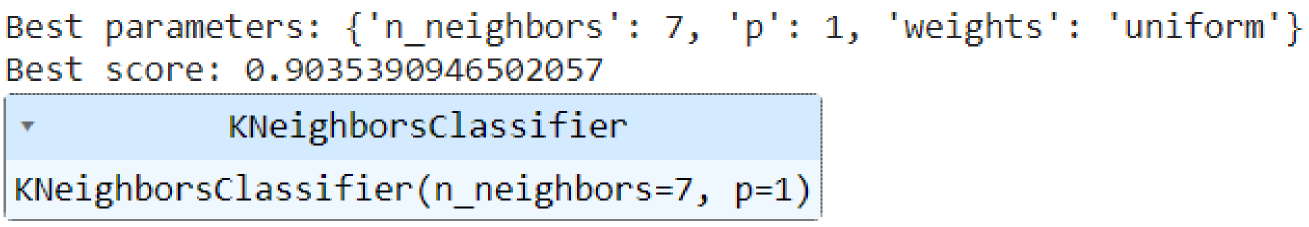
In a binary classification scenario (where the target variable has two classes), the decision rule is to assign a class label to an unclassified instance based on the majority class among its 'k' nearest neighbours.

Voting Mechanism:

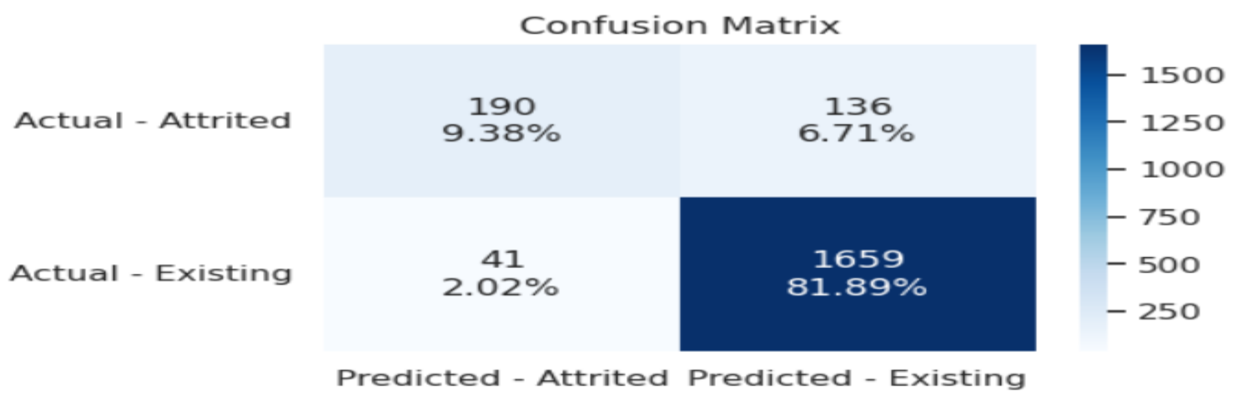
For KNN, the class assignment is determined via a majority vote among the 'k' neighbors. In the binary case, the class with the highest count among the 'k' neighbors is chosen as the predicted class.

Parameter 'k' Selection:

The choice of the 'k' parameter influences the model's performance. A smaller 'k' value tends to capture more localized patterns, possibly leading to overfitting, while a larger 'k' may oversimplify the decision boundary, resulting in underfitting.



```
Best parameters: {'n_neighbors': 7, 'p': 1, 'weights': 'uniform'}
Best score: 0.9035390946502057
KNeighborsClassifier
KNeighborsClassifier(n_neighbors=7, p=1)
```



We got TEST Recall = 0.58282513

c.) LDA (Linear Discriminant Analysis)

Linear Discriminant Analysis (LDA) is a fundamental classification technique predicated on several assumptions. The core assumptions are based on the following principles:

1. **Multivariate Gaussian Distribution Assumption:**
   Let $X = (X1, X2, \ldots, Xp)$ be the predictor vector, and $Y \in \{0,1\}$
   is the response variable. The assumption is that the observations of the predictor variables for a particular class follow a multivariate Gaussian distribution. **T**hus, for class k=1 and 2, the feature vector $\boldsymbol{X(k)}$ **follows a normal distribution** $X(k) \sim N(\mu k, \Sigma)$ where $\boldsymbol{X(k)}$ **denotes the feature vector for class k.**
   Common Variance-Covariance Matrix Assumption:
   Further, the assumption is that the variance-covariance matrix (Σ) across the classes is the same.
   The decision rule utilized in LDA is given by the discriminant function as follows**:**

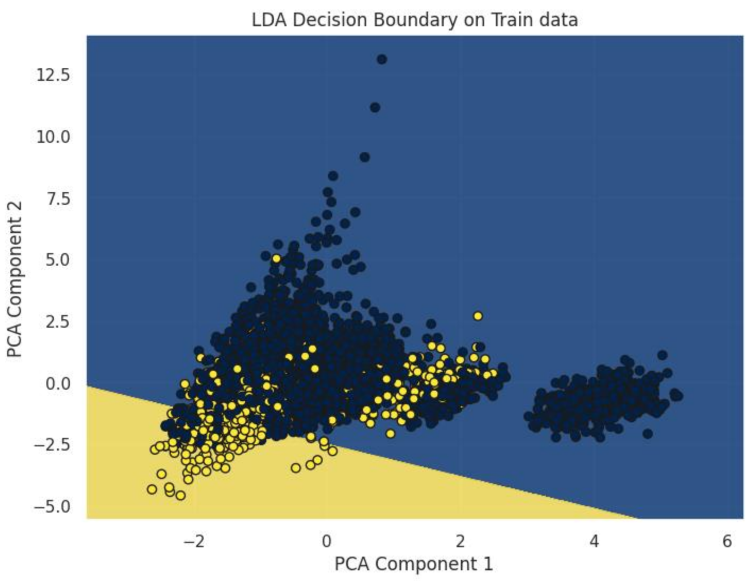
$$\boldsymbol{\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log(\pi_k)}$$

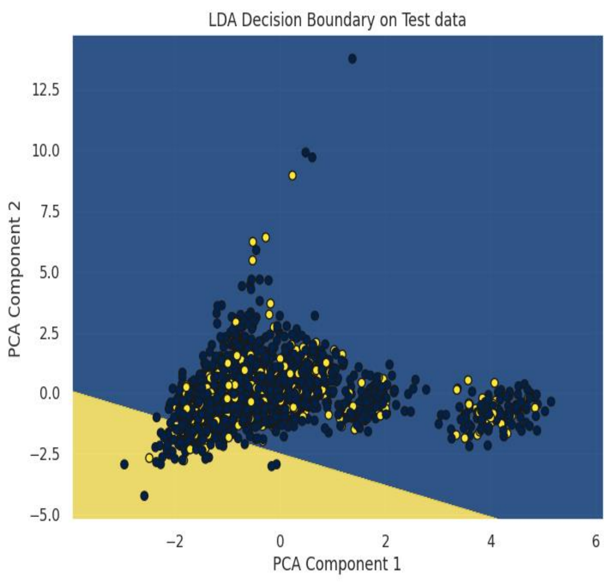
Here, *x* represents an observation of the feature vector *X*, and $\boldsymbol{\pi_k}$ denotes the probability of class *k*. In practice, estimations for the parameters are utilized:

$\boldsymbol{S_{pooled}\ is\ used\ as\ the\ estimate\ of\ \Sigma}$

$\boldsymbol{\overline{x}_k\ is\ used\ as\ the\ estimate\ of\ \mu_k}$

Therefore, the computation of equation (1) is performed for each class, and the observation �*x* is assigned to the class for which $\boldsymbol{\delta_k(x)\ is\ maximal}$





```
•  Training Accuracy Score :   85.7
•  Cross Validation Score : 85.71
❖  Testing Accuracy Score :   79.91
•  Precision Score is : 12.04
•  Recall Score is : 4.0
•  F1-Score Score is : 6.0
-----------------------------------------
```

d.) Quadratic Discriminant Analysis (QDA)

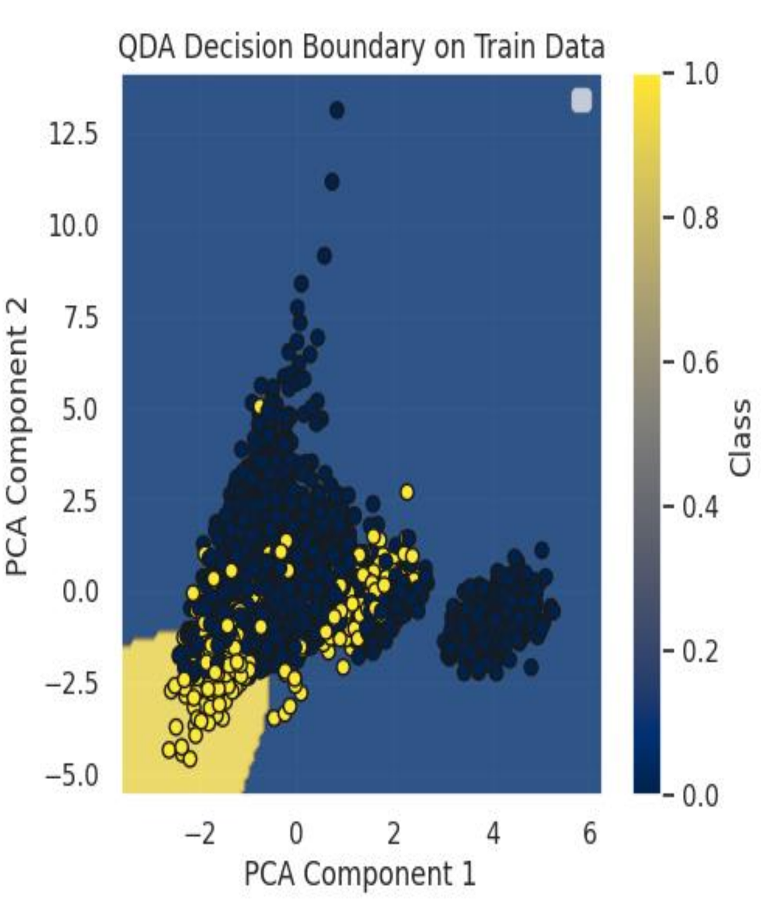
Quadratic Discriminant Analysis (QDA) is an extension of Linear Discriminant Analysis (LDA) that relaxes a critical assumption made by LDA. Unlike LDA, which assumes identical variance-covariance matrices of the feature variables for different classes, QDA allows these matrices to vary between classes.

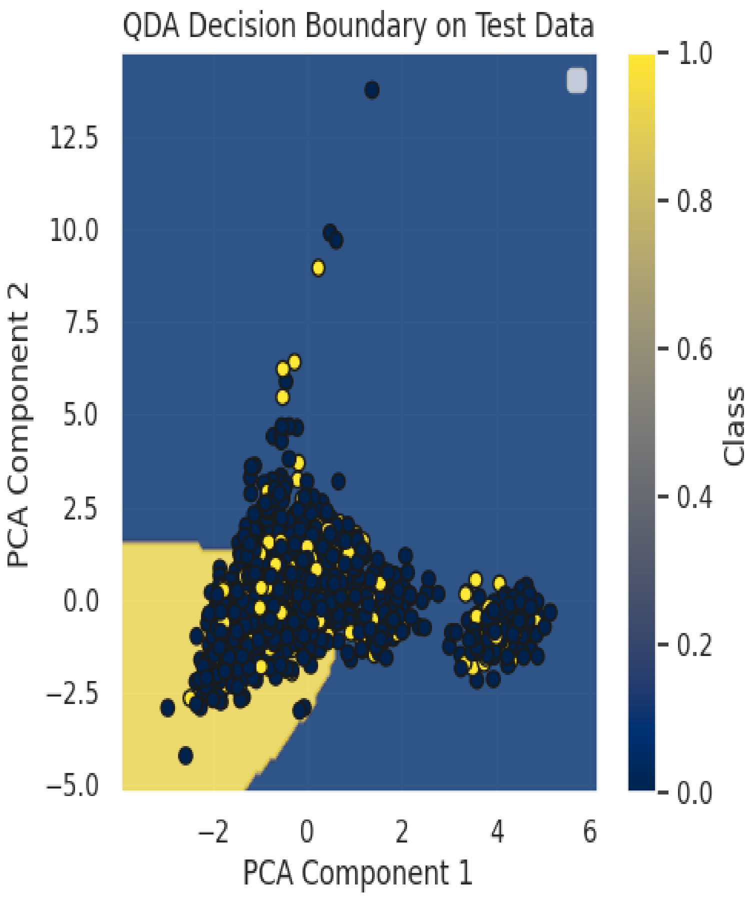
1. **Varying Variance-Covariance Matrices Assumption:**
   Let $X = (X1, X2, \dots, Xp)$ be the predictor vector, and $Y \in \{0,1\}$ is the response variable. The assumption is that the observations of the predictor variables for a particular class follow a multivariate Gaussian distribution. **T**hus, for class k=1 and 2, the feature vector $\boldsymbol{X(k)}$ **follows a normal distribution** $X(k) \sim N(\mu k, \Sigma_k)$ where $\boldsymbol{X(k)}$ **denotes the feature vector for class k.**

   The decision rule in QDA is expressed as follows:

$$\boldsymbol{\delta_k(x) = -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T}\Sigma^{-1}\boldsymbol{\mu_k - \frac{1}{2}log(}\Sigma_k\boldsymbol{) + log(\pi_k)}$$

Here, *x* represents an observation of the feature vector *X*, and $\boldsymbol{\pi_k}$ denotes the probability of class *k*.





```
• Training Accuracy Score :  86.04
• Cross Validation Score : 86.07
❖ Testing Accuracy Score :  77.34
• Precision Score is : 14.36
• Recall Score is : 8.31
• F1-Score Score is : 10.53
-------------------------------------------
```

f.) Decision Tree

In this study, we construct a decision tree classifier utilizing the entropy-based splitting criterion for each split. To mitigate the risk of overfitting, a pruning technique known as Cost-Complexity Pruning (also known as Weakest link pruning) is employed. This method involves a greedy search to determine the optimal alpha parameter, which controls the trade-off between tree complexity and accuracy.

Methodology:

Entropy-based Splitting:

In this study, we employ an entropy-based splitting criterion for each node of the decision tree. Entropy, denoted as *H*(*D*), is calculated as:

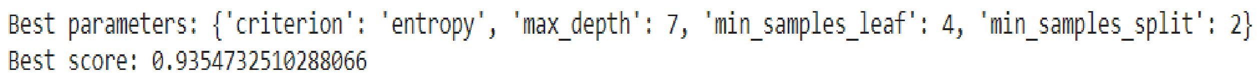
$$\boldsymbol{H(D) = -\sum_{k} p(k).\log\big(p(k)\big)}$$

Here, *H*(*D*) represents the entropy of dataset *D* with *c* classes, and *p*(*i*) signifies the proportion of class *i* within *D*.

Tree Construction:

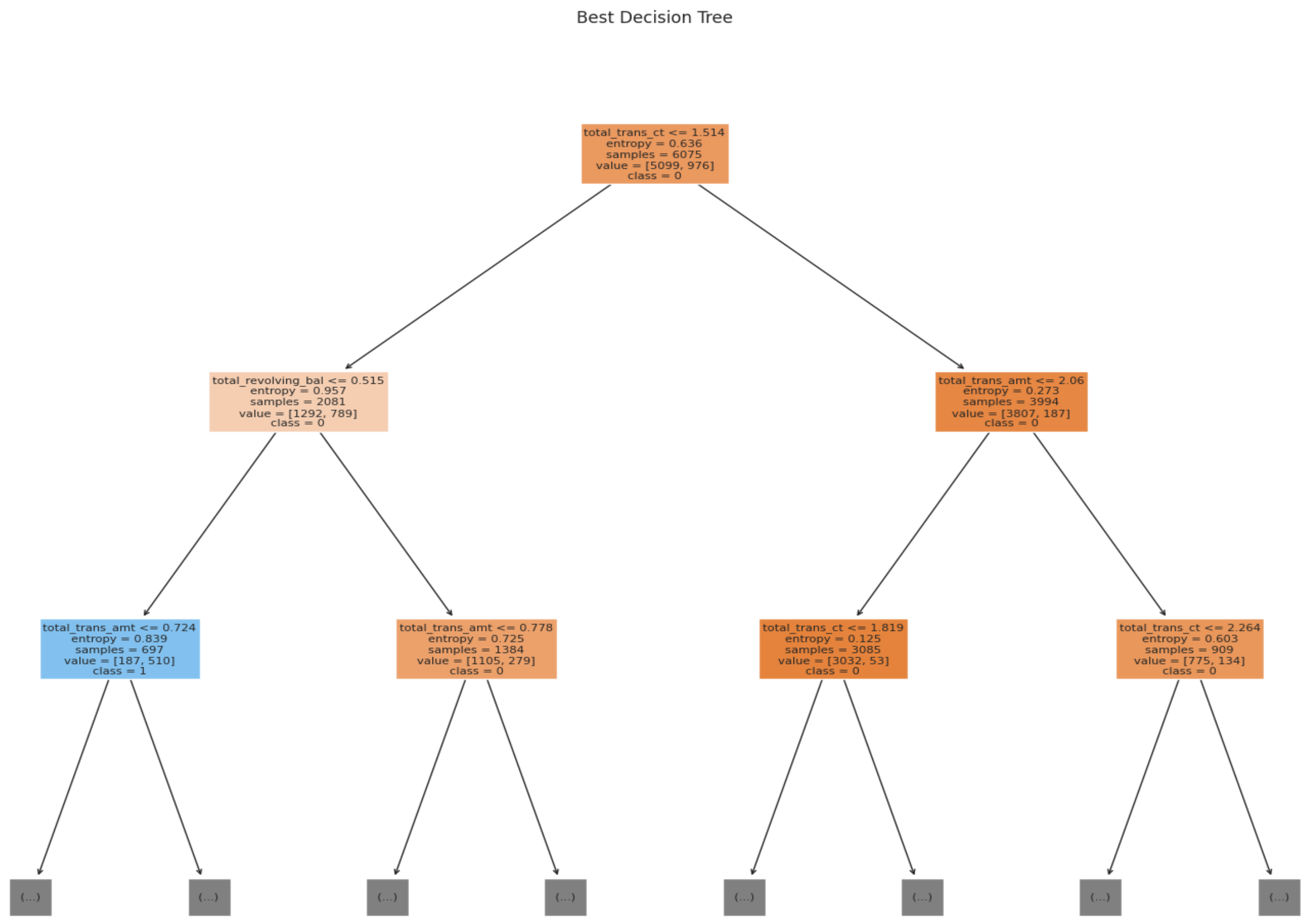
The decision tree is constructed in a recursive manner. At each internal node, the algorithm searches for the optimal feature to split the data. This decision is made by maximizing the information gain, calculated as:

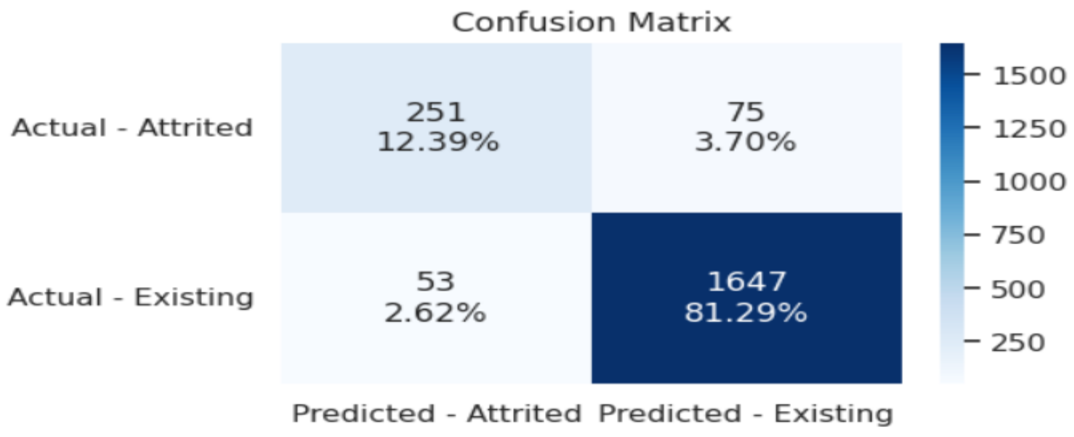
***InformationGain*=Entropy before split−Weighted average of entropy after split**



```
Best parameters: {'criterion': 'entropy', 'max_depth': 7, 'min_samples_leaf': 4, 'min_samples_split': 2}
Best score: 0.9354732510288066
```

Best Decision Tree





We got TEST Recall = 0.7699356

g.) Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It employs the bagging technique to build multiple trees, thereby reducing overfitting and improving accuracy.

Entropy-based Splitting:

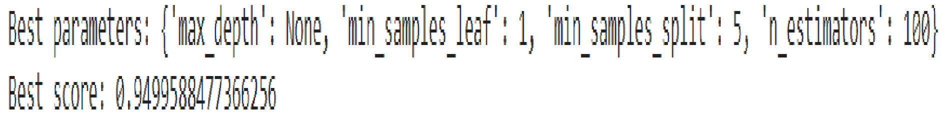
Similar to the decision tree classifier, Random Forest employs entropy as one of the criteria for feature selection during tree construction. Entropy is calculated at each internal node to determine the best feature for splitting the data.

Tree Construction:

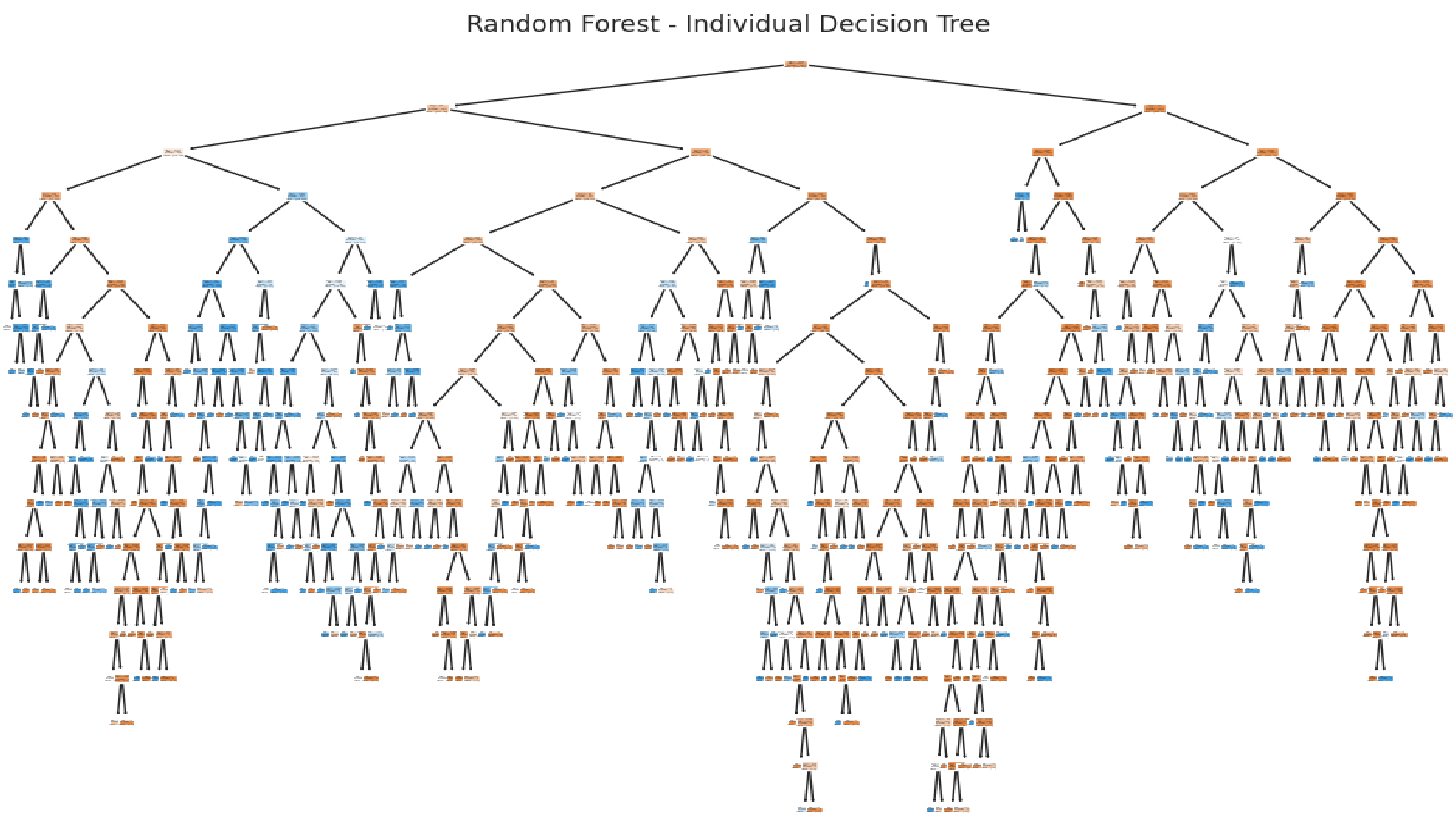
The Random Forest algorithm constructs multiple decision trees using a bagging (bootstrap aggregating) approach. Each tree is built using a random subset of features and data samples to minimize correlation between trees. The decision on splitting nodes in these trees is based on maximizing information gain, following a similar process to the decision tree.

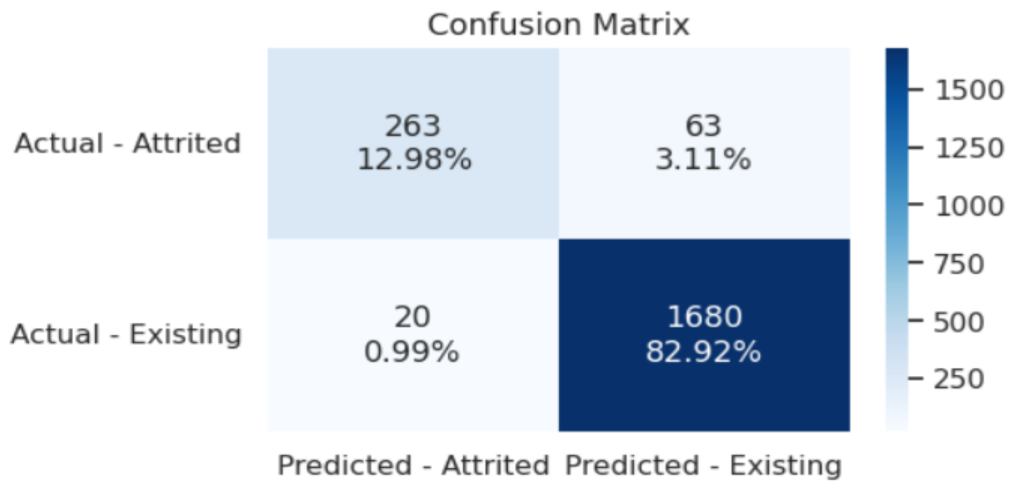
Cost-Complexity Pruning (for individual trees within the forest):

While Random Forests are less prone to overfitting due to the aggregation of multiple trees, individual trees within the forest might still benefit from pruning. A form of Cost-Complexity Pruning may be applied to individual trees, involving the addition of a complexity parameter (α) to the impurity measure. By evaluating the total impurity of each individual tree, the optimal pruning parameter is determined through a greedy search over a range of α values. The goal is to minimize the cost-complexity criterion to strike a balance between the complexity of each tree and its predictive accuracy



```
Best parameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100}
Best score: 0.9499588477366256
```





We got TEST Recall = 0.806729328

Oversampling train data using SMOTE

Our dataset has a huge imbalance in target variable labels. To deal with such datasets, we have a few tricks up our sleeves, which we call Imbalanced Classification.

Imbalanced classification involves developing predictive models on classification datasets that have a severe class imbalance.

The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on the minority class, although typically it is performance on the minority class that is most important, which is the case in our study here.

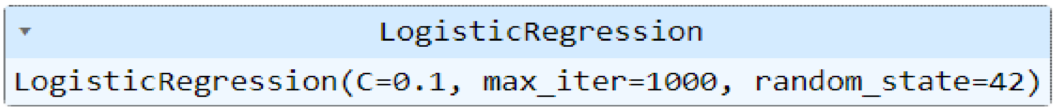
One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don't add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE for short.

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

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

After UpSampling, the shape of train_X: (10198, 27)
After UpSampling, the shape of train_y: (10198,)
```

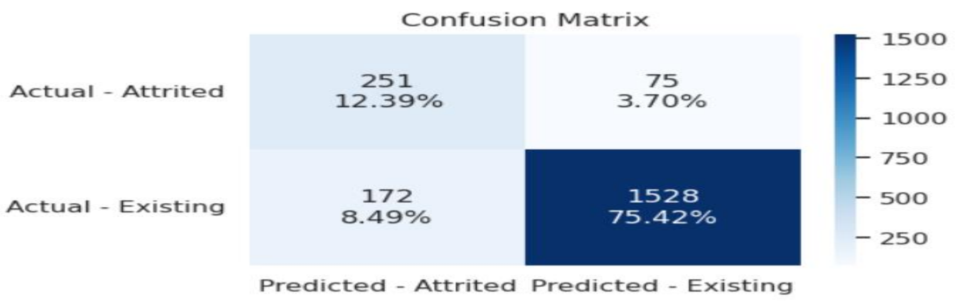
Build Models with Oversampled Data

a.) Logistic Regression:

```
Best parameters: {'C': 0.1, 'penalty': 'l2'}
Best score: 0.8857659944801854
```

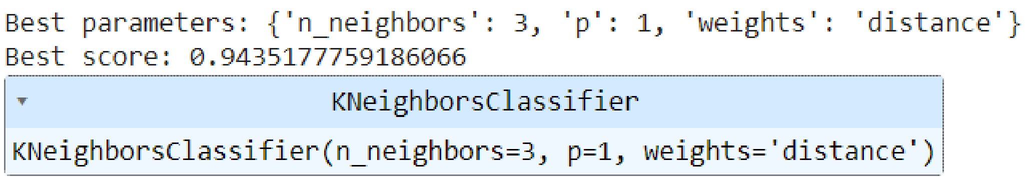


```
LogisticRegression
LogisticRegression(C=0.1, max_iter=1000, random_state=42)
```

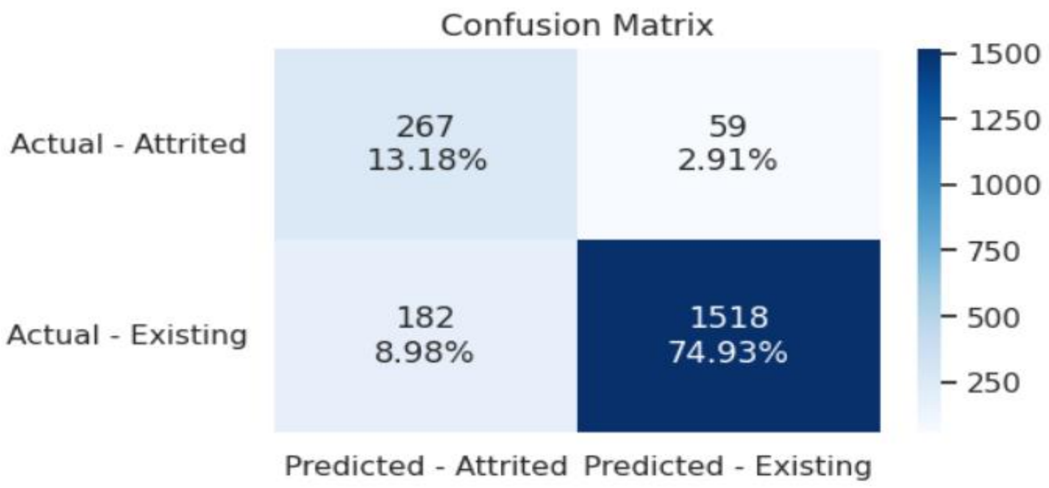


We got TEST Recall = 0.769938

b.) KNN (K- Nearest Neighbour)

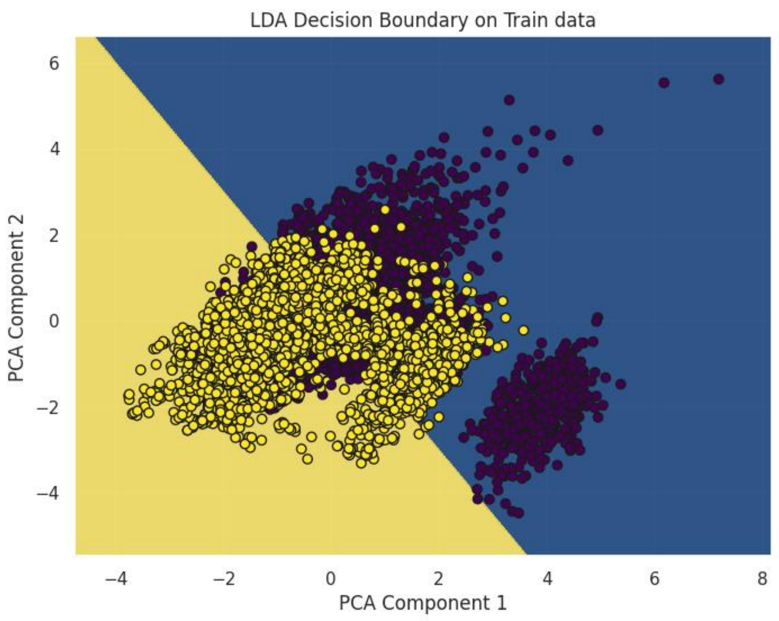


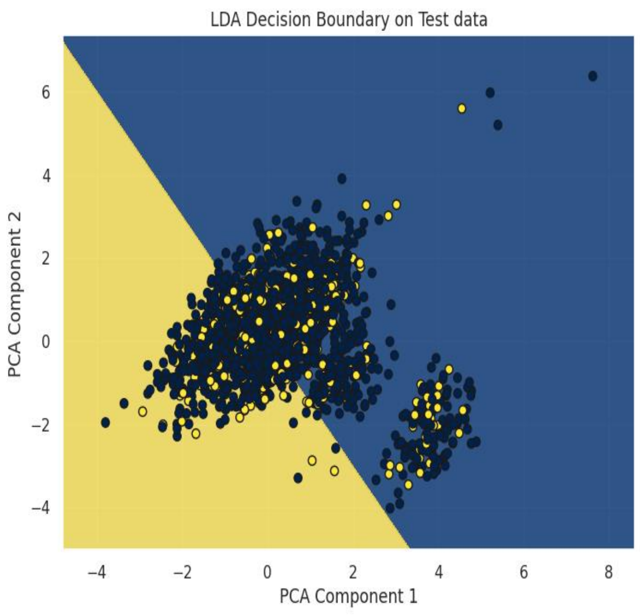
```
Best parameters: {'n_neighbors': 3, 'p': 1, 'weights': 'distance'}
Best score: 0.9435177759186066
▾                        KNeighborsClassifier
KNeighborsClassifier(n_neighbors=3, p=1, weights='distance')
```



We got TEST Recall = 0.81901

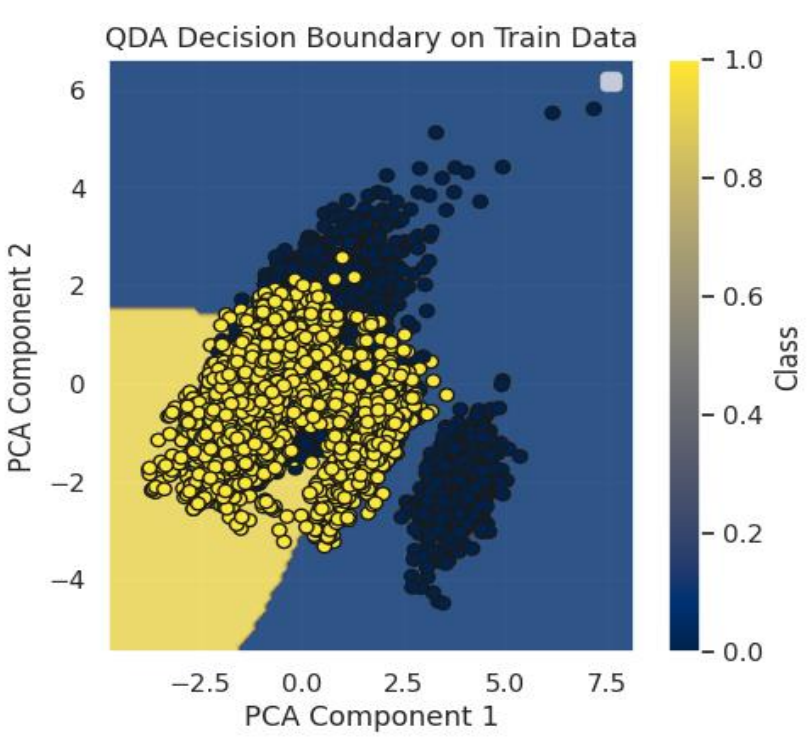
c.) LDA (Linear Discriminant Analysis)

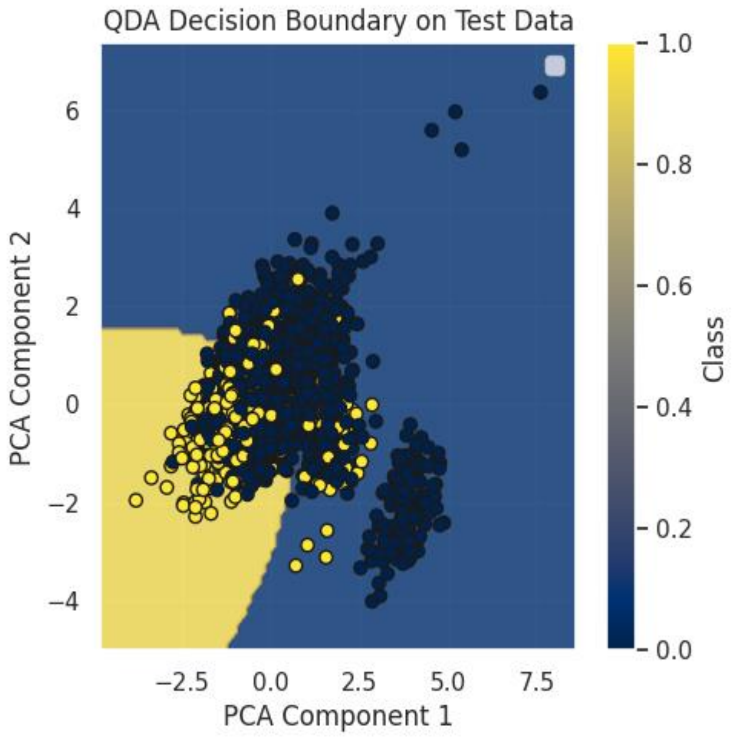




```
Accuracy score on the test set: 0.74
Accuracy score on the training set: 0.75
Confusion Matrix for Test Data:
[[1259  441]
 [  88  238]]

Confusion Matrix for Training Data:
[[3818 1281]
 [1231 3868]]
```

d.) QDA (Quadratic Discriminant Analysis)

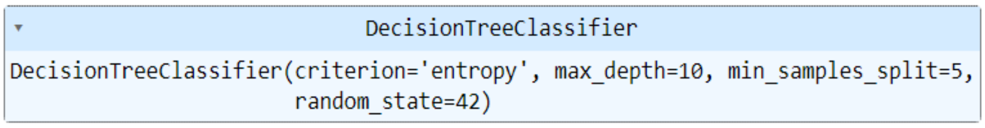




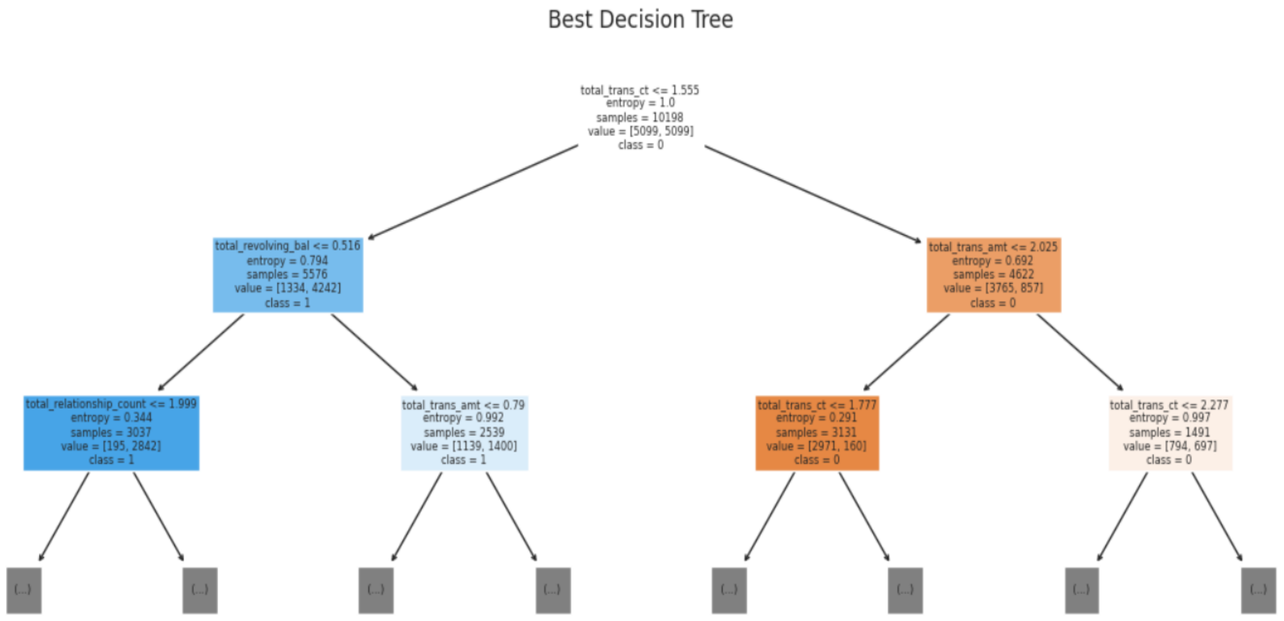
```
• Training Accuracy Score :  72.48
• Cross Validation Score : 72.47
❖ Testing Accuracy Score :  69.15
• Precision Score is : 30.96
• Recall Score is : 74.54
• F1-Score Score is : 43.74
-------------------------------------------
```

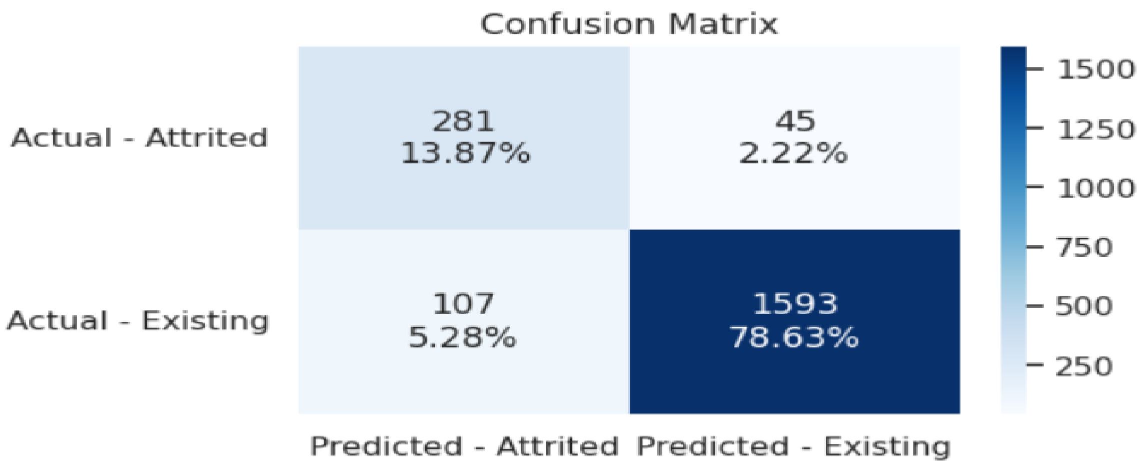
e.)Decision Tree

```
Best parameters: {'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 5}
Best score: 0.9379298771985498
```



```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', max_depth=10, min_samples_split=5,
                       random_state=42)
```

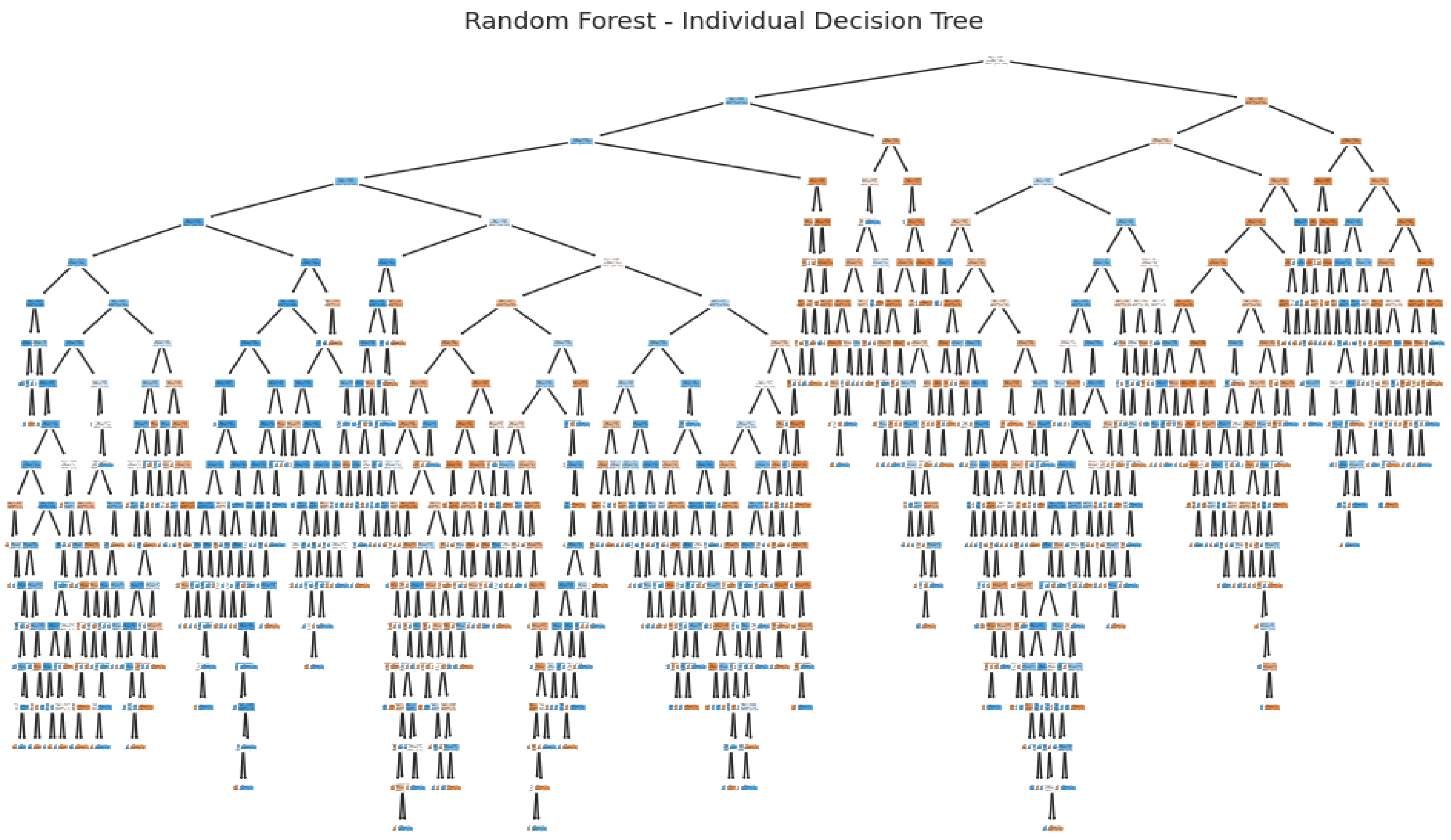


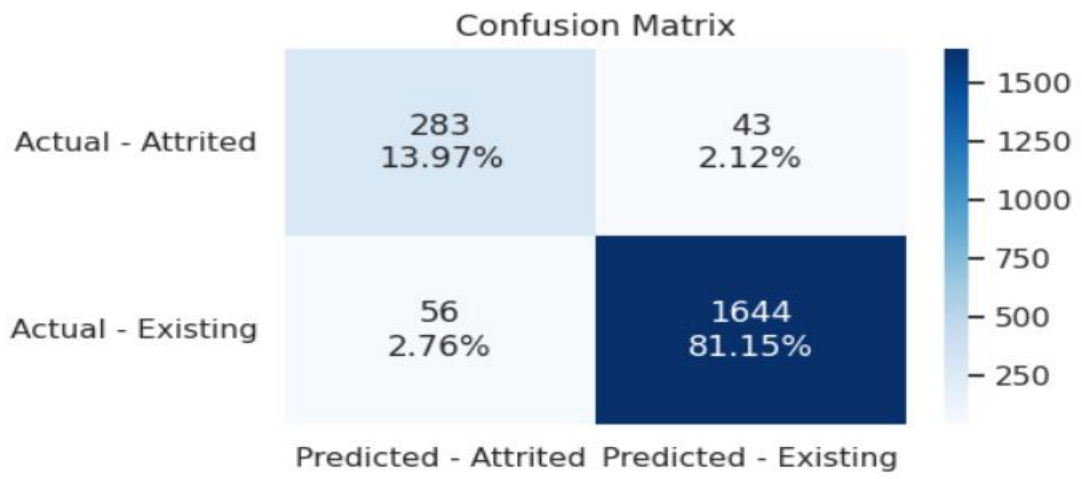


We got TEST Recall = 0.86196319

f.) Random Forest

```
Best parameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Best score: 0.972544163324967
```





We got TEST Recall = 0.8680942

Conclusion:

"After performing an in-depth analysis on credit card customer churn prediction using several models, it was observed that among Logistic Regression, LDA, QDA, Decision Tree, and Random Forest, the **Random Forest model emerged as the most efficient and accurate model for predicting customer churn**. The findings are based on the assessment of accuracy using both original and oversampled data.

Though Decision Trees are prone to overfitting, it offered competitive accuracy rates. It performed well, especially with the original data, but fell short compared to Random Forest, particularly in handling complex interactions within the features.

Therefore, based on the comprehensive evaluation**, the Random Forest model stands out as the most efficient and reliable model for credit card customer churn prediction in our study**, **exhibiting superior accuracy over both the original(recall=0.80672) and oversampled datasets(Recall = 0.868092)**. Its ensemble learning technique and capability to handle complex interactions provide a robust solution for predicting customer churn in this domain.

Key findings:

The most important features to understand customer credit card churn, are

- Total Transaction Count
- Total Transaction Amount
- Total Revolving Balance
- Total Amount Change Q4 to Q1
- Total Count Change Q4 to Q1
- Total Relationship Count

1.) **All of these features are negatively correlated with the Attrition Flag, meaning, the lower the values of these features, the higher the chances of a customer to attrite.**
2.) **Bank should connect with the customer more often to increase the connect, and provide the customer with various offers and schemes to increase relationships of the customer with the bank**
3.) **Bank should also offer credit limit increase for the customers who are regularly using the credit card. This should increase the credit card spends/transaction amounts.**
4.) **With the above model, we can predict which customers are likely to attrite, and according to the predicted probability, at least top 20-30% customers can be reached out to discuss credit card offers, credit limit increase etc, to try retain those customers.**