Predicting

CUSTOMER CHURN

BY GUARDIAN TRI ANGGORO

introduce

CUSTOMER CHURN

What

Customer churn refers to the natural business cycle of losing and acquiring customers. Every company — no matter the quality of its products or customer service — experiences churn. Generally speaking, the less churn you have, the more customers you keep.

Why

Understanding your customer churn is essential to evaluating the effectiveness of your marketing efforts and the overall satisfaction of your customers. It's also easier and cheaper to keep customers you already have versus acquiring new ones. Due to the popularity of subscription business models, it's critical for many businesses to understand where, how, and why their customers may be churning.

How to reduce Customer Churn

1. Target the right audience
2. Offer incentives
3. Analyze why churn occurs
4. Know Your Customer
5. Pay attention to complaints

OBJECTIVE

PREDICTING CUSTOMER CHURN WITH 2 MODELS

FIND THE IMPORTANT FEATURE THAT CAN MITIGATE CUSTOMER CHURN

GIVE RECOMMENDED ACTION TO BANK

ABOUT THE DATASET

The data used in this repository is an Credit Card Customer dataset available at Kaggle through this https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers?datasetId=982921&sortBy=voteCount

Features	Definition
Customer Age	Customer's Age in Years
Gender	M=Male, F=Female
Education Level	Educational Qualification of the account holder
Marital Status	Married, Single, Divorced, Unknown
Income Category	Annual Income Category of the account holder
Card Category	Type of Card (Blue, Silver, Gold, Platinum)
Total Relationship Count	Total no. of products held by the customer
Months Inactive 12 mon	No. of months inactive in the last 12 months
Contacts Count 12 mon	No. of Contacts in the last 12 months
Credit Limit	Credit Limit on the Credit Card
Total Revolving Bal	Total Revolving Balance on the Credit Card
Avg Open To Buy	Open to Buy Credit Line
Total Trans Amt	Total Transaction Amount (Last 12 months)
Total Trans Ct	Total Transaction Count (Last 12 months)
Avg Utilization_Ratio	Average Card Utilization Ratio

Data Information

```
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Attrition_Flag          10127 non-null  object
 1   Customer_Age            10127 non-null  int64
 2   Gender                  10127 non-null  object
 3   Education_Level         10127 non-null  object
 4   Marital_Status          10127 non-null  object
 5   Income_Category         10127 non-null  object
 6   Card_Category           10127 non-null  object
 7   Months_on_book          10127 non-null  int64
 8   Total_Relationship_Count  10127 non-null  int64
 9   Months_Inactive_12_mon  10127 non-null  int64
 10  Contacts_Count_12_mon   10127 non-null  int64
 11  Credit_Limit            10127 non-null  float64
 12  Total_Revolving_Bal     10127 non-null  int64
 13  Avg_Open_To_Buy         10127 non-null  float64
 14  Total_Trans_Amt         10127 non-null  int64
 15  Total_Trans_Ct          10127 non-null  int64
 16  Avg_Utilization_Ratio   10127 non-null  float64
dtypes: float64(3), int64(8), object(6)
memory usage: 1.3+ MB
```

The dataset contains 17 columns and 10127 entries. The target column is a `Attrition_Flag`

Data Observation :

- There are 17 columns and 10127 entries
- There are no missing values
- There are no duplicted data

Data Cleaning

MISSING VALUE AND DUPLICATED DATA

- MISSING VALUE

```
df.isna().sum()
```

```
Attrition_Flag              0
Customer_Age                0
Gender                      0
Education_Level             0
Marital_Status              0
Income_Category             0
Card_Category               0
Months_on_book              0
Total_Relationship_Count    0
Months_Inactive_12_mon      0
Contacts_Count_12_mon       0
Credit_Limit                0
Total_Revolving_Bal         0
Avg_Open_To_Buy             0
Total_Trans_Amt             0
Total_Trans_Ct              0
Avg_Utilization_Ratio       0
dtype: int64
```

There is no missing value in dataset

- DUPLICATED DATA

```
[ ] df.duplicated().sum()
```

```
0
```

There is no duplicated data in dataset

Standard EDA

DIVIDES INTO 2 TYPES OF DATA CATEGORIES: NUMERICAL AND CATEGORICALS

- **NUMERICALS**

	Customer_Age	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Trans_Amt	Total_Trans_Ct	Avg_Utilization_Ratio
count	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000	10127.000000
mean	46.325960	35.928409	3.812580	2.341167	2.455317	8631.953698	1162.814061	7469.139637	4404.086304	64.858695	0.274894
std	8.016814	7.986416	1.554408	1.010622	1.106225	9088.776650	814.987335	9090.685324	3397.129254	23.472570	0.275691
min	26.000000	13.000000	1.000000	0.000000	0.000000	1438.300000	0.000000	3.000000	510.000000	10.000000	0.000000
25%	41.000000	31.000000	3.000000	2.000000	2.000000	2555.000000	359.000000	1324.500000	2155.500000	45.000000	0.023000
50%	46.000000	36.000000	4.000000	2.000000	2.000000	4549.000000	1276.000000	3474.000000	3899.000000	67.000000	0.176000
75%	52.000000	40.000000	5.000000	3.000000	3.000000	11067.500000	1784.000000	9859.000000	4741.000000	81.000000	0.503000
max	73.000000	56.000000	6.000000	6.000000	6.000000	34516.000000	2517.000000	34516.000000	18484.000000	139.000000	0.999000

- Minimum and maximum values for all columns seemed reasonable
- Customer_Age, Months_Inactive_12_mon, Contacts_Count_12_mon, Credit_Limit, Avg_Open_To_Buy, Total_Trans_Ct, Avg_Utilization_Ratio has skewed distribution (mean > median)

Standard EDA

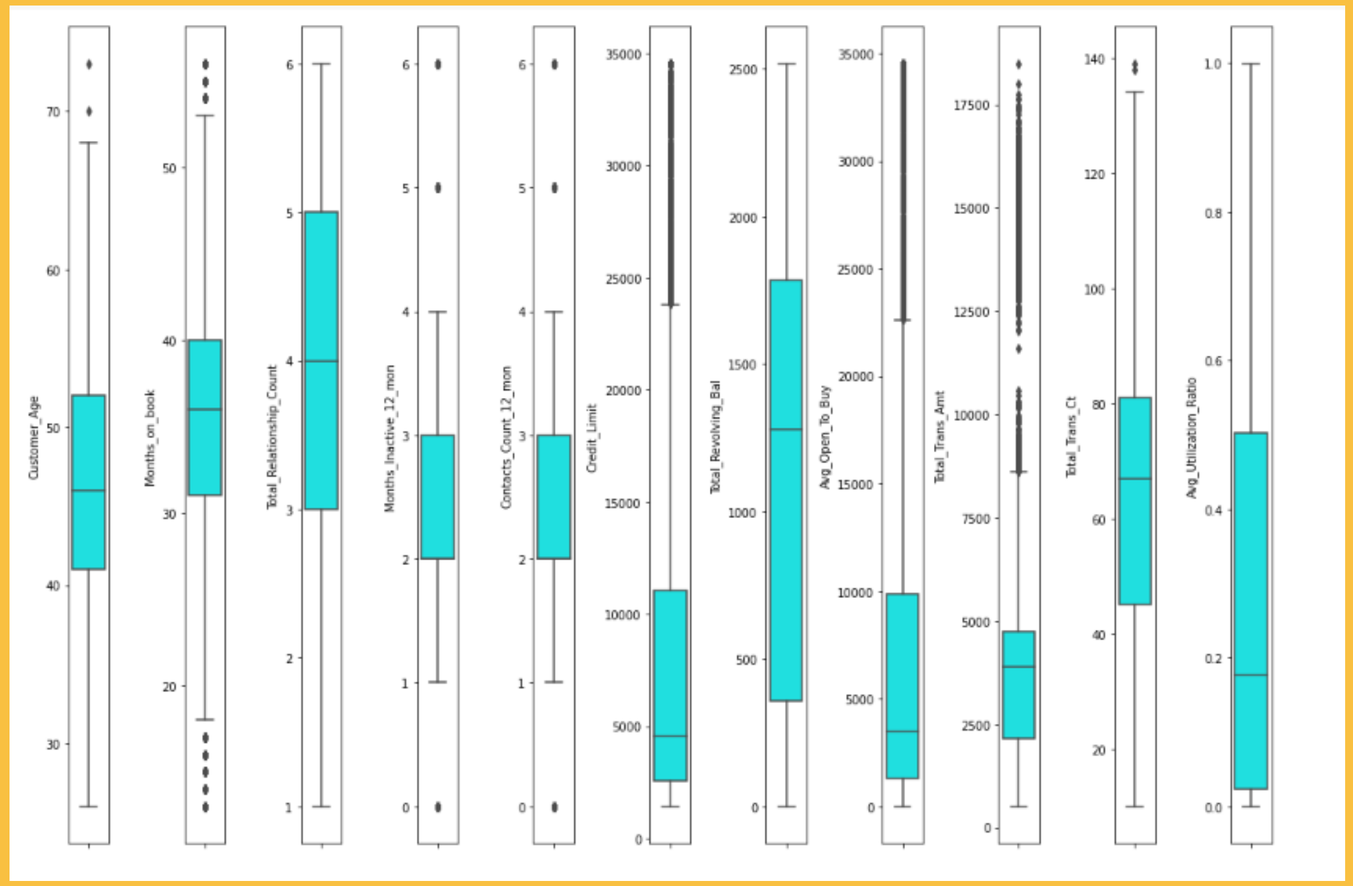
DIVIDES INTO 2 TYPES OF DATA CATEGORIES: NUMERICAL AND CATEGORICALS

- CATEGORICALS

	Attrition_Flag	Gender	Education_Level	Marital_Status	Income_Category	Card_Category
count	10127	10127	10127	10127	10127	10127
unique	2	2	7	4	6	4
top	Existing Customer	F	Graduate	Married	Less than $40K	Blue
freq	8500	5358	3128	4687	3561	9436

Standard EDA

BOXPLOT



There are many outlier for some features

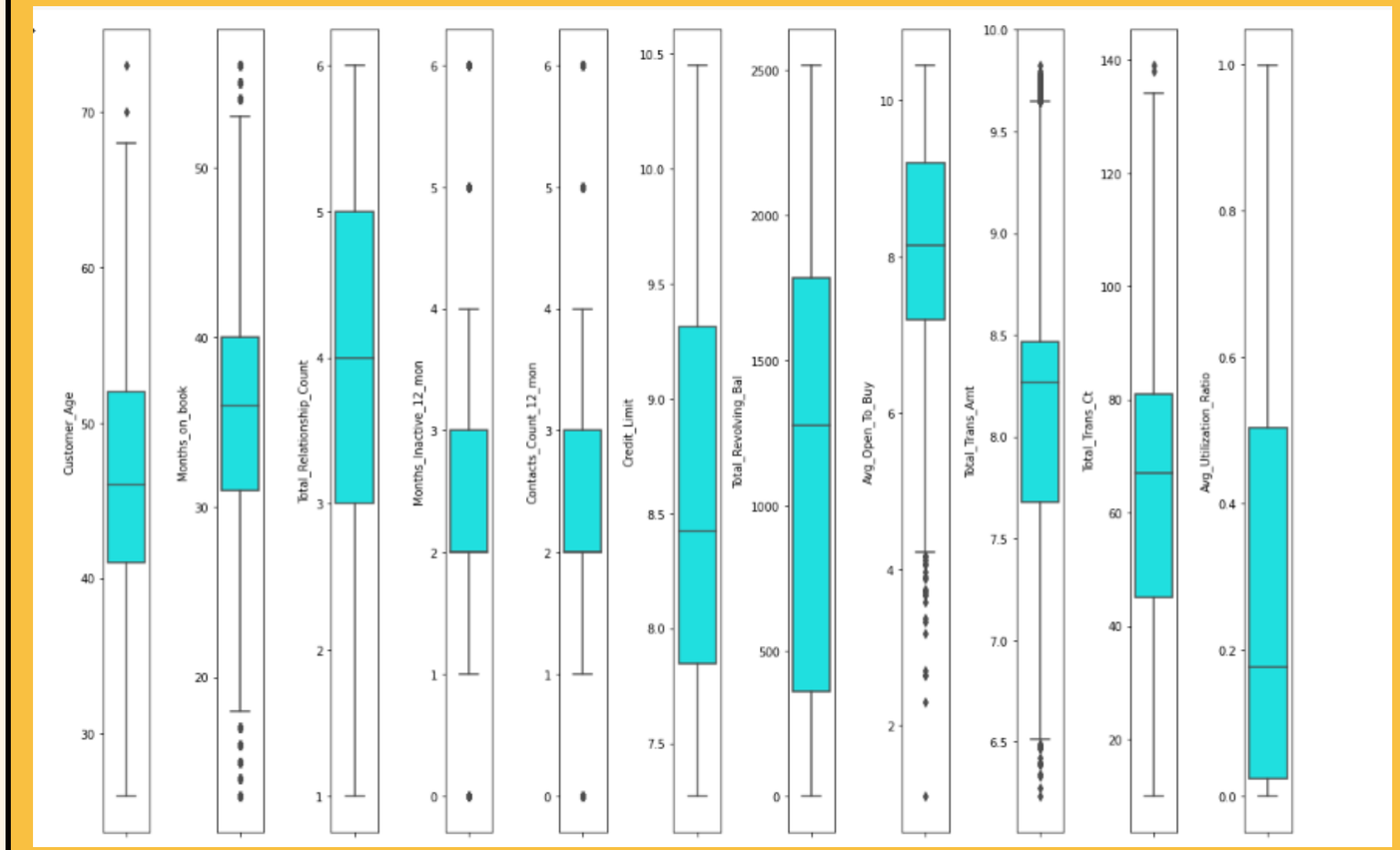
- Credit_Limit
- Avg_Open_To_Buy
- Total_Trans_Amount

(Handling with Log Transformation)

- Customer_Age,Months_on_book,Months_Inactive_12_mon,Contacts_Count_12_mon have outliers but still normal. no need to drop it

Standard EDA

BOXPLOT AFTER LOG TRANSFORMATION



There are still many outlier for some features :

- Avg_Open_To_Buy
- Total_Trans_Amount

Standard EDA

VARIANCE INFLATION FACTOR (VIF)

	feature	vif_score
1	Customer_Age	2.697315
2	Months_on_book	2.701003
3	Total_Relationship_Count	1.106288
4	Months_Inactive_12_mon	1.012022
5	Contacts_Count_12_mon	1.033422
6	Credit_Limit	24.283942
7	Total_Revolving_Bal	3.161376
8	Avg_Open_To_Buy	49.608277
9	Total_Trans_Amt	5.960724
10	Total_Trans_Ct	5.816538
11	Avg_Utilization_Ratio	15.407085

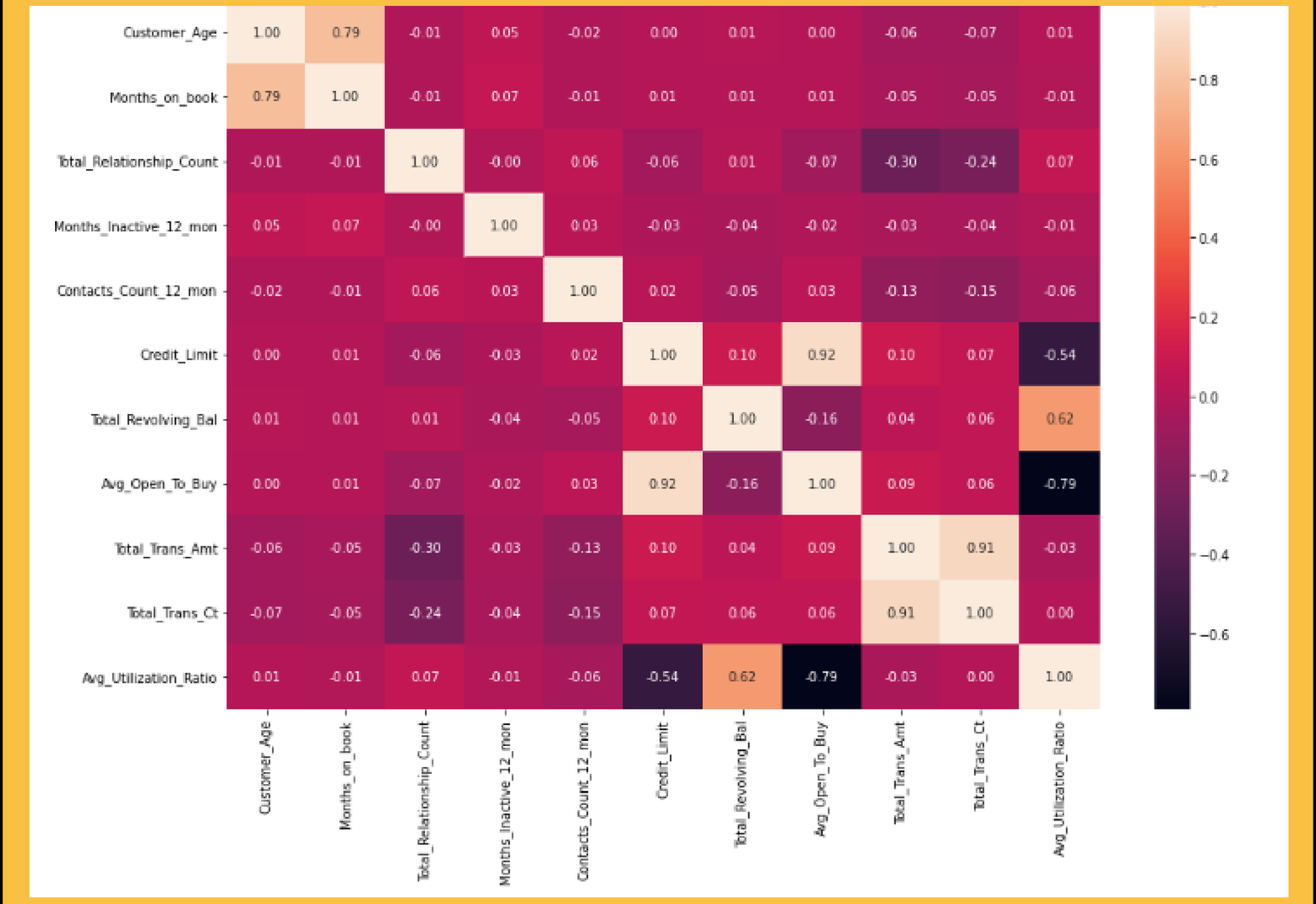
Interpretation :

There are some feature that have VIF score > 4

- Credit_Limit
- Avg_Open_To_Buy
- Total_Trans_Amt
- Total_Trans_Ct
- Avg_Utilization_Ratio

Standard EDA

CORRELATION HEATMAP



INTREPRETATION

IVariables inside cream rectangle are highly correlated each other

- Months_on_books have a high correlation score 0.79 with Customer_Age
- Credit_Limit have a high correlation score 0.92 with Avg_Open_To_Buy
- Total_Trans_Amt have a high correlation score 0.92 with Total_Trans_Ct

This means they contain redundant information. We can choose only 1 of them to modelling process.

According VIF Score and Correlation Heatmap. I will drop some feature below here

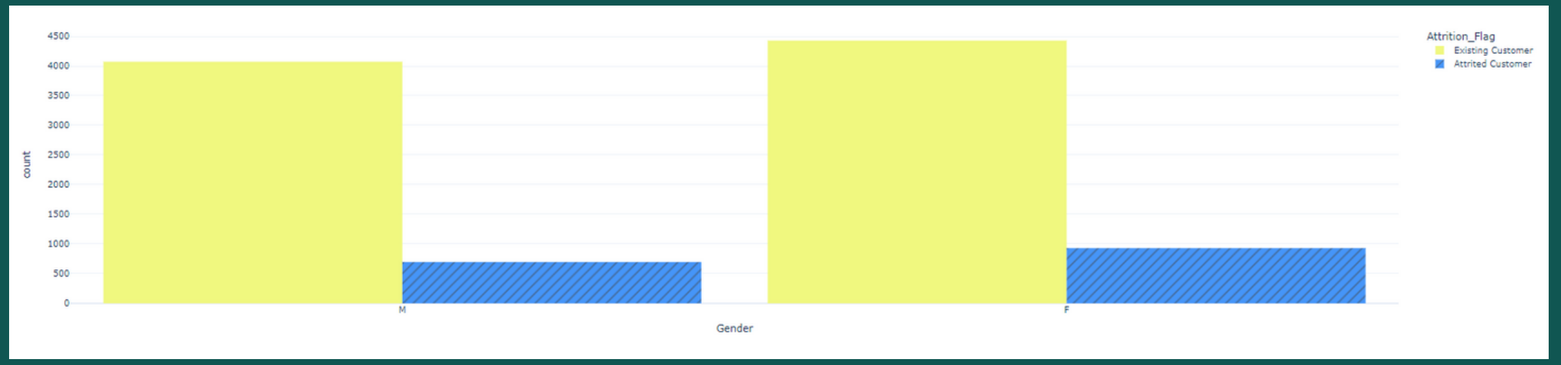
- Avg Open to Buy
- Months of Books
- Avg Ultization Ratio
- Total Trans Ct

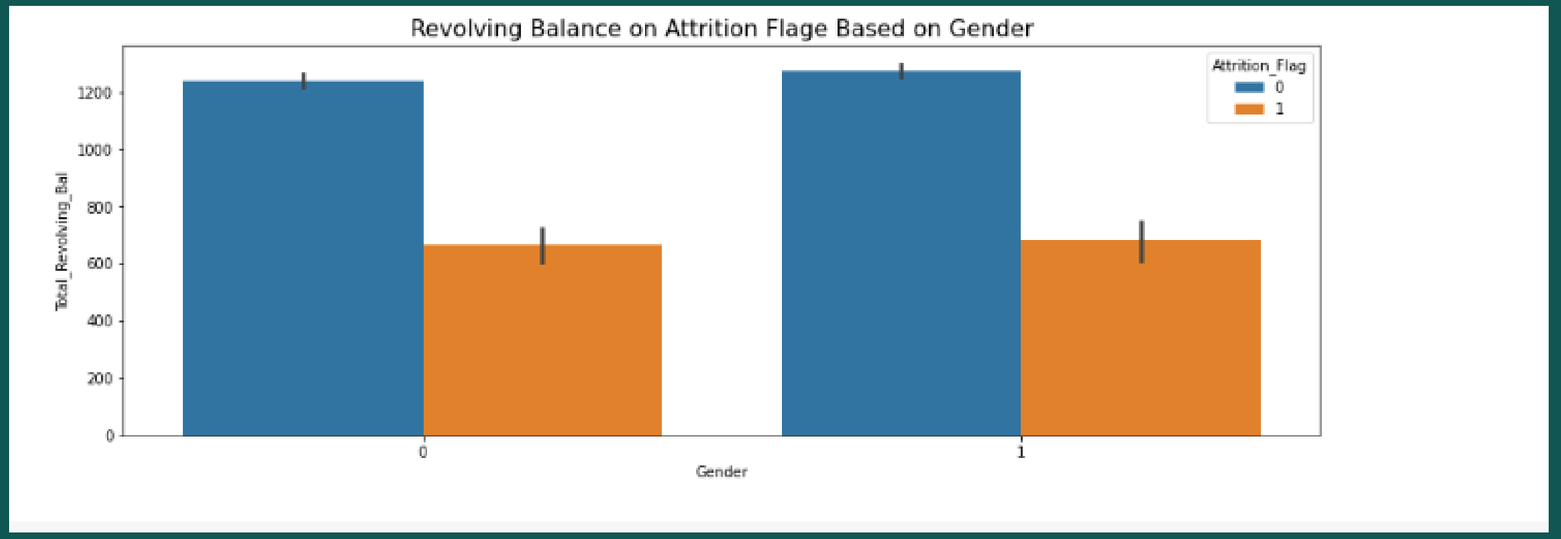
DATA

Visualization

Gender

DATA VISUALIZATION





INTREPRETATION

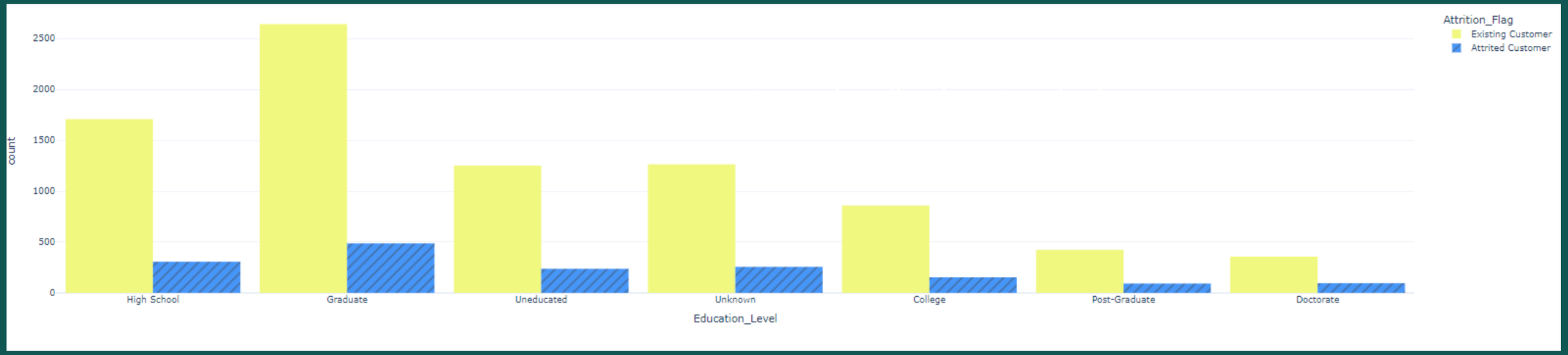
In Data visualization, Gender is divided into female and male. By percentage, female (52.9%) have higher percentage more than men (47.1%).

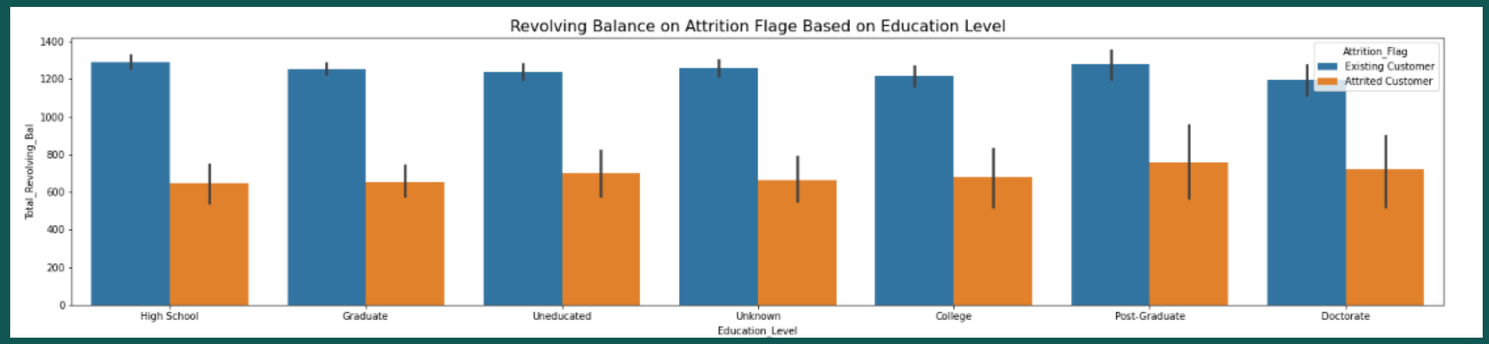
And based on the histogram count, it is said that the female has a higher churn rate than the male

In revolving balance, female has the highest revolving balance (existing and churn)

Education

DATA VISUALIZATION



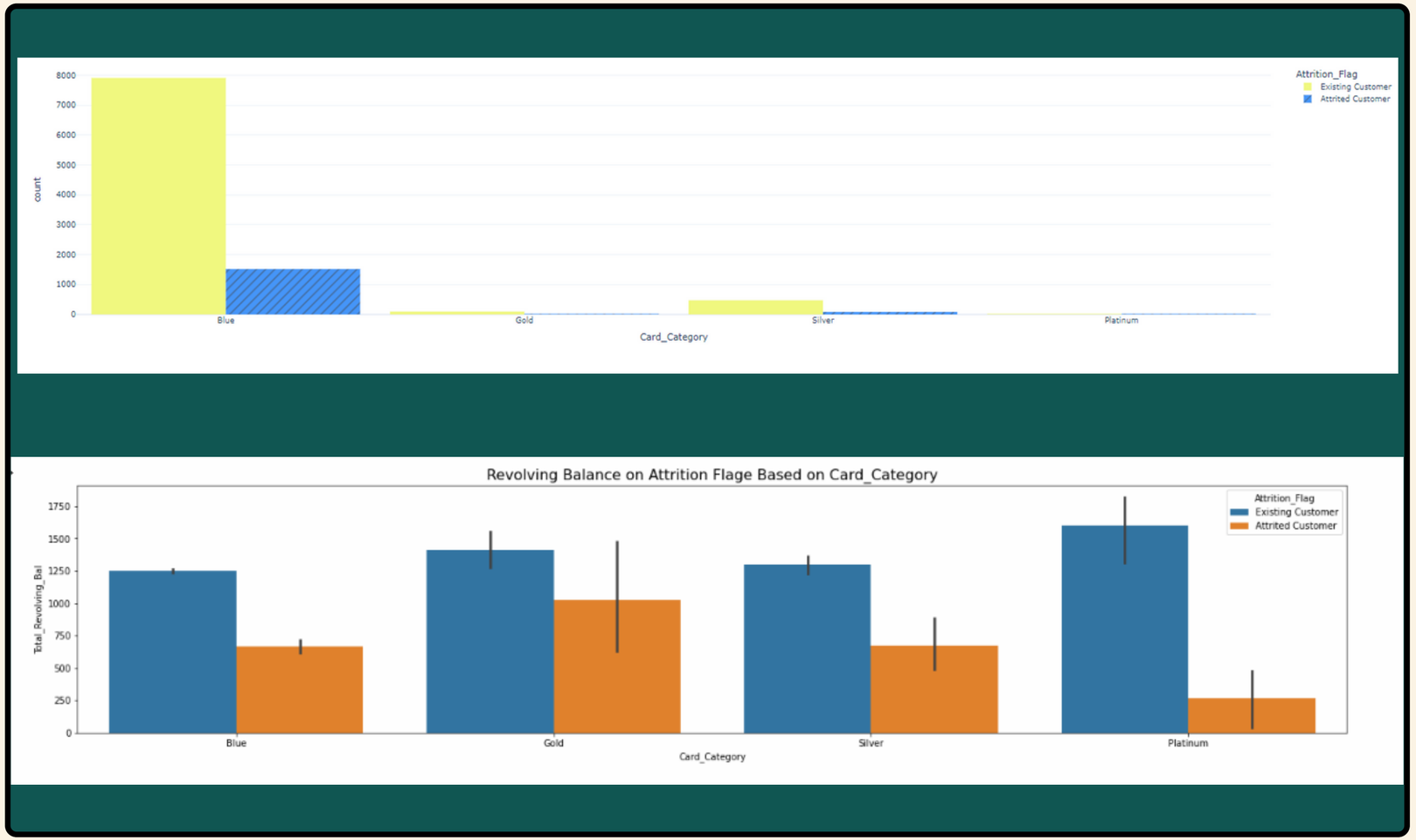


INTREPRETATION

IN DATA VISUALIZATION, EDUCATION LEVEL GRADUATE IS THE HIGHEST PERCENTAGE WITH (30.9%), FOLLOWED BY HIGH SCHOOL (19.9%), UKNOWN (15%), AND UNEDUCATED (14.7%). AND IN HISTOGRAM, IT CAN BE SAID THAT GRADUATE HAS A HIGHER CHURN.

CARD CATEGORIES

DATA VISUALIZATION



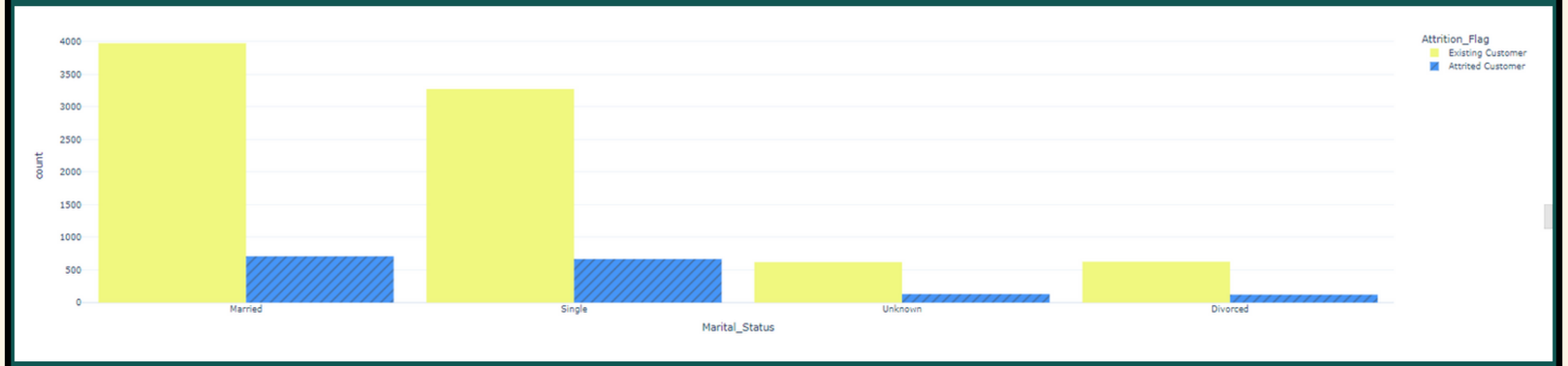
INTREPRETATION

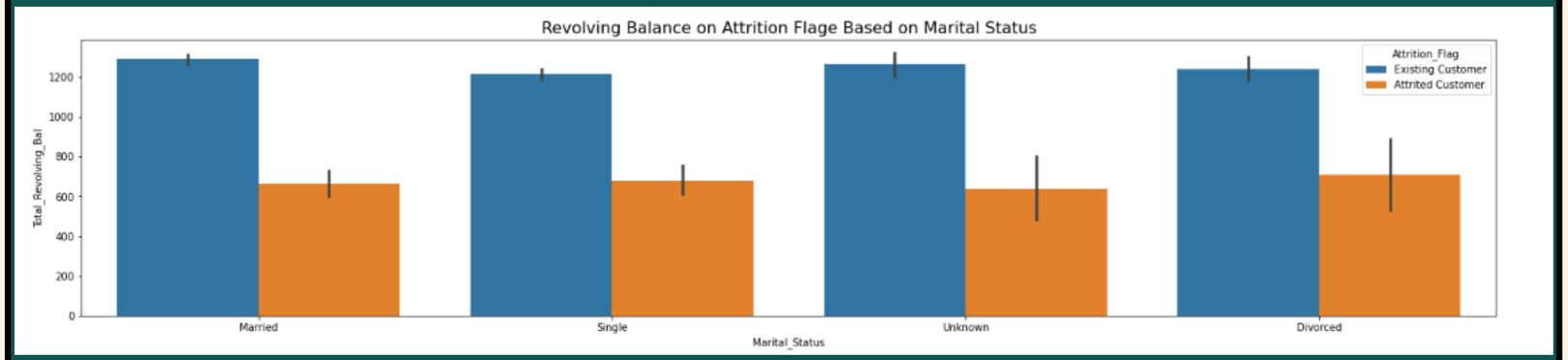
In Data visualization, Card Category Blue Card blue card is the most dominant card used by the customer with percentage score (93.2%). And of course on the histogram, blue card has the highest customer churn

In revolving balance, Platinum card (existing cust) and Gold card (Churn Cust) has the highest revolving balance

Martial Status

DATA VISUALIZATION





INTREPRETATION

In Data visualization, Martial Status Married is the highest percentage with (46.3%), Followed by Single (38.9%), Uknown (7.4%), and Divorced (7.39%).

And in histogram, it can be said that married has a higher churn.

In revolving balance, Married (existing) and divorce (churn) has the highest revolving balance

DATA PREPROCESSING

- ## LABEL ENCODING

```
[ ] df['Gender'] = df['Gender'].replace("F", 0).replace("M", 1)
    df['Attrition_Flag'] = df['Attrition_Flag'].replace("Existing Customer", 0).replace("Attrited Customer", 1)
    df.head()
```

	Attrition_Flag	Customer_Age	Gender	Education_Level	Marital_Status	Income_Category	Card_Category	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Total_Trans_Amt
0	0	45	1	High School	Married	$60K - $80K	Blue	5	1	3	9.448648	777	7.042286
1	0	49	0	Graduate	Single	Less than $40K	Blue	6	1	2	9.018695	864	7.163172
2	0	51	1	Graduate	Married	$80K - $120K	Blue	4	1	0	8.136811	0	7.542744
3	0	40	0	High School	Unknown	Less than $40K	Blue	3	4	1	8.105609	2517	7.065613
4	0	40	1	Uneducated	Married	$60K - $80K	Blue	5	1	0	8.458716	0	6.704414

- ## ONE HOT ENCODING

```
df =pd.get_dummies(data=df,columns=['Education_Level','Marital_Status',
                                    'Income_Category', 'Card_Category'],drop_first=True)
df.head()
```

	Attrition_Flag	Customer_Age	Gender	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Total_Trans_Amt	Education_Level_Doctorate	...	Marital_Status_Single	Marital_Status_Unknown	Income_Category_$40K - $60K	Income_Categ
0	0	45	1	5	1	3	9.448648	777	7.042286	0	...	0	0	0	
1	0	49	0	6	1	2	9.018695	864	7.163172	0	...	1	0	0	
2	0	51	1	4	1	0	8.136811	0	7.542744	0	...	0	0	0	
3	0	40	0	3	4	1	8.105609	2517	7.065613	0	...	0	1	0	
4	0	40	1	5	1	0	8.458716	0	6.704414	0	...	0	0	0	

5 rows × 26 columns

VARIANCE INFLATION FACTOR (VIF)

Interpretation :

There are some feature that have VIF score > 4

- Income_Category_Less than $40K

So i will drop it, and check VIF again

	feature	vif_score
1	Customer_Age	1.017954
2	Gender	3.468891
3	Total_Relationship_Count	1.102118
4	Months_Inactive_12_mon	1.008035
5	Contacts_Count_12_mon	1.029405
6	Credit_Limit	1.991243
7	Total_Revolving_Bal	1.028240
8	Total_Trans_Amt	1.169013
9	Education_Level_Doctorate	1.375805
10	Education_Level_Graduate	2.826335
11	Education_Level_High School	2.393928
12	Education_Level_Post-Graduate	1.427517
13	Education_Level_Uneducated	2.089267
14	Education_Level_Unknown	2.113287
15	Marital_Status_Married	3.871954
16	Marital_Status_Single	3.779709
17	Marital_Status_Unknown	1.814950
18	Income_Category_$40K - $60K	3.930180
19	Income_Category_$60K - $80K	2.631485
20	Income_Category_$80K - $120K	2.660017
21	Income_Category_Less than $40K	7.296238
22	Income_Category_Unknown	3.559322
23	Card_Category_Gold	1.068662
24	Card_Category_Platinum	1.015521
25	Card_Category_Silver	1.234557

	feature	vif_score
1	Customer_Age	1.014012
2	Gender	2.348720
3	Total_Relationship_Count	1.102094
4	Months_Inactive_12_mon	1.007978
5	Contacts_Count_12_mon	1.028979
6	Credit_Limit	1.766849
7	Total_Revolving_Bal	1.027213
8	Total_Trans_Amt	1.167905
9	Education_Level_Doctorate	1.375801
10	Education_Level_Graduate	2.825934
11	Education_Level_High School	2.393871
12	Education_Level_Post-Graduate	1.427478
13	Education_Level_Uneducated	2.089265
14	Education_Level_Unknown	2.113264
15	Marital_Status_Married	3.871372
16	Marital_Status_Single	3.779464
17	Marital_Status_Unknown	1.814915
18	Income_Category_$40K - $60K	1.224137
19	Income_Category_$60K - $80K	1.709898
20	Income_Category_$80K - $120K	1.798538
21	Income_Category_Unknown	1.226389
22	Card_Category_Gold	1.067715
23	Card_Category_Platinum	1.014464
24	Card_Category_Silver	1.217585

Interpretation :

There is no feature that have VIF score > 4

Random Forest VS KNN

WITH OVERSAMPLING SMOTE

RF-SMOTE

```
sm = over_sampling.SMOTE(sampling_strategy=0.5,random_state=42)

X = df.drop(['Attrition_Flag'],axis = 1)
Y = df['Attrition_Flag']
X_sm, Y_sm = sm.fit_resample(X, Y)

print(f'''Shape of X before SMOTE: {X.shape}
Shape of X after SMOTE: {X_sm.shape}''')

print('\nBalance of positive and negative classes (%):')
Y_sm.value_counts(normalize=True) * 100
```

```
Shape of X before SMOTE: (10127, 25)
Shape of X after SMOTE: (12750, 25)

Balance of positive and negative classes (%):
0    66.666667
1    33.333333
Name: Attrition_Flag, dtype: float64
```

f1 Score 0.8909626719056976

KNN-SMOTE

```
sm = over_sampling.SMOTE(sampling_strategy=0.5,random_state=42)

X = df.drop(['Attrition_Flag'],axis = 1)
Y = df['Attrition_Flag']
X_sm, Y_sm = sm.fit_resample(X, Y)

print(f'''Shape of X before SMOTE: {X.shape}
Shape of X after SMOTE: {X_sm.shape}''')

print('\nBalance of positive and negative classes (%):')
Y_sm.value_counts(normalize=True) * 100
```

```
Shape of X before SMOTE: (10127, 25)
Shape of X after SMOTE: (12750, 25)

Balance of positive and negative classes (%):
0    66.666667
1    33.333333
Name: Attrition_Flag, dtype: float64
```

F1 Score KNN 0.4302325581395349

Random Forest VS KNN

HYPERTUNNING

RF

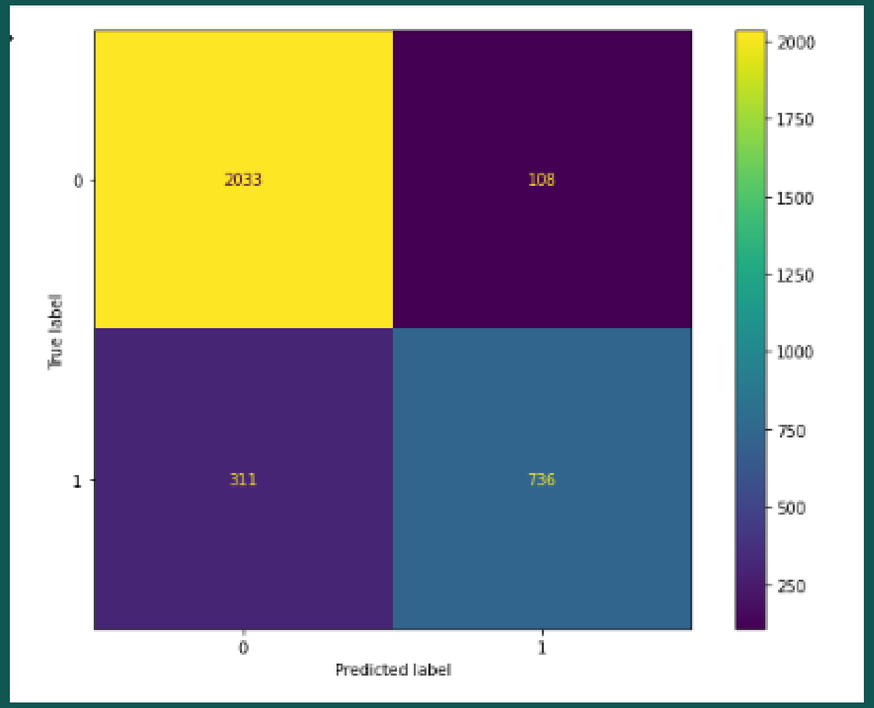
	params	mean_test_score	rank_test_score
20	{'max_depth': 5, 'n_estimators': 10}	0.701219	1
21	{'max_depth': 5, 'n_estimators': 20}	0.697172	2
22	{'max_depth': 5, 'n_estimators': 30}	0.686235	3
23	{'max_depth': 5, 'n_estimators': 40}	0.685608	4
24	{'max_depth': 5, 'n_estimators': 50}	0.682804	5
15	{'max_depth': 4, 'n_estimators': 10}	0.640028	6
16	{'max_depth': 4, 'n_estimators': 20}	0.634719	7
17	{'max_depth': 4, 'n_estimators': 30}	0.626293	8
19	{'max_depth': 4, 'n_estimators': 50}	0.611313	9
18	{'max_depth': 4, 'n_estimators': 40}	0.611311	10
12	{'max_depth': 3, 'n_estimators': 30}	0.552927	11
13	{'max_depth': 3, 'n_estimators': 40}	0.519525	12
11	{'max_depth': 3, 'n_estimators': 20}	0.518895	13
14	{'max_depth': 3, 'n_estimators': 50}	0.518592	14
10	{'max_depth': 3, 'n_estimators': 10}	0.511404	15
5	{'max_depth': 2, 'n_estimators': 10}	0.434608	16
7	{'max_depth': 2, 'n_estimators': 30}	0.376536	17
6	{'max_depth': 2, 'n_estimators': 20}	0.328757	18
8	{'max_depth': 2, 'n_estimators': 40}	0.317209	19
9	{'max_depth': 2, 'n_estimators': 50}	0.289118	20
4	{'max_depth': 1, 'n_estimators': 50}	0.000000	21
3	{'max_depth': 1, 'n_estimators': 40}	0.000000	21
2	{'max_depth': 1, 'n_estimators': 30}	0.000000	21
1	{'max_depth': 1, 'n_estimators': 20}	0.000000	21
0	{'max_depth': 1, 'n_estimators': 10}	0.000000	21

KNN

	params	mean_test_score	rank_test_score
1	{'n_neighbors': 2, 'weights': 'distance'}	0.404982	1
3	{'n_neighbors': 3, 'weights': 'distance'}	0.372270	2
2	{'n_neighbors': 3, 'weights': 'uniform'}	0.368730	3
5	{'n_neighbors': 4, 'weights': 'distance'}	0.366079	4
9	{'n_neighbors': 6, 'weights': 'distance'}	0.355433	5
7	{'n_neighbors': 5, 'weights': 'distance'}	0.354540	6
6	{'n_neighbors': 5, 'weights': 'uniform'}	0.351004	7
11	{'n_neighbors': 7, 'weights': 'distance'}	0.341297	8
10	{'n_neighbors': 7, 'weights': 'uniform'}	0.341297	9
13	{'n_neighbors': 8, 'weights': 'distance'}	0.334213	10
12	{'n_neighbors': 8, 'weights': 'uniform'}	0.271455	11
8	{'n_neighbors': 6, 'weights': 'uniform'}	0.267015	12
4	{'n_neighbors': 4, 'weights': 'uniform'}	0.251129	13
0	{'n_neighbors': 2, 'weights': 'uniform'}	0.230806	14

Random Forest VS KNN

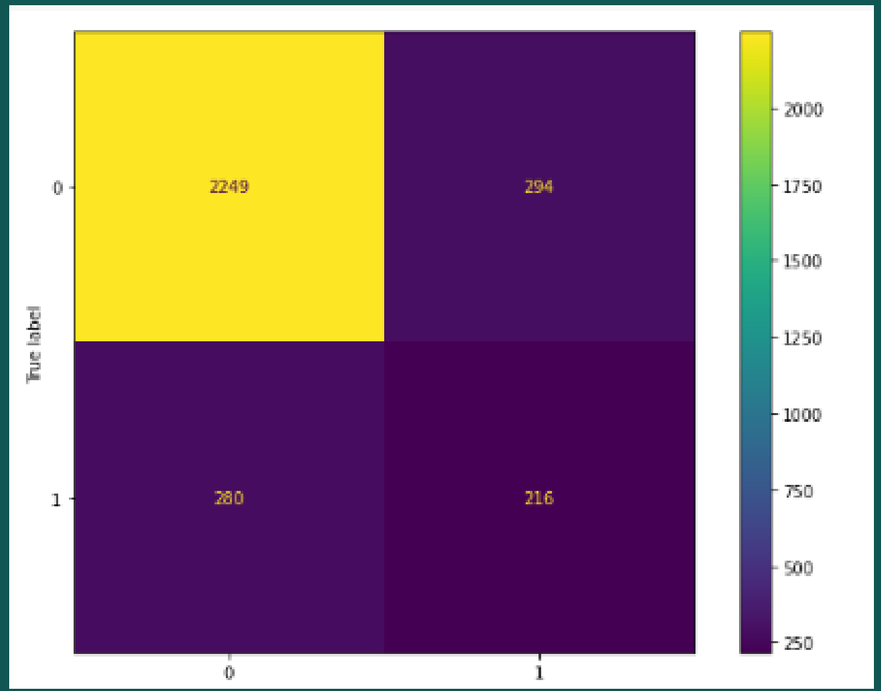
RF-CONFFUSION MATRIX



Interpretation :

From the RF conffusion matrix, it can be said that there are 2033 customers who are predicted to be existing customers, while for churn customers there are 736 customers.

KNN-CONFFUSION MATRIX



Interpretation :

From the KNN conffusion matrix, it can be said that there are 2249 customers who are predicted to be existing customers, while for churn customers there are 216 customers.

Random Forest VS KNN

RF-CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.87	0.95	0.91	2141
1	0.87	0.70	0.78	1047
accuracy			0.87	3188
macro avg	0.87	0.83	0.84	3188
weighted avg	0.87	0.87	0.86	3188

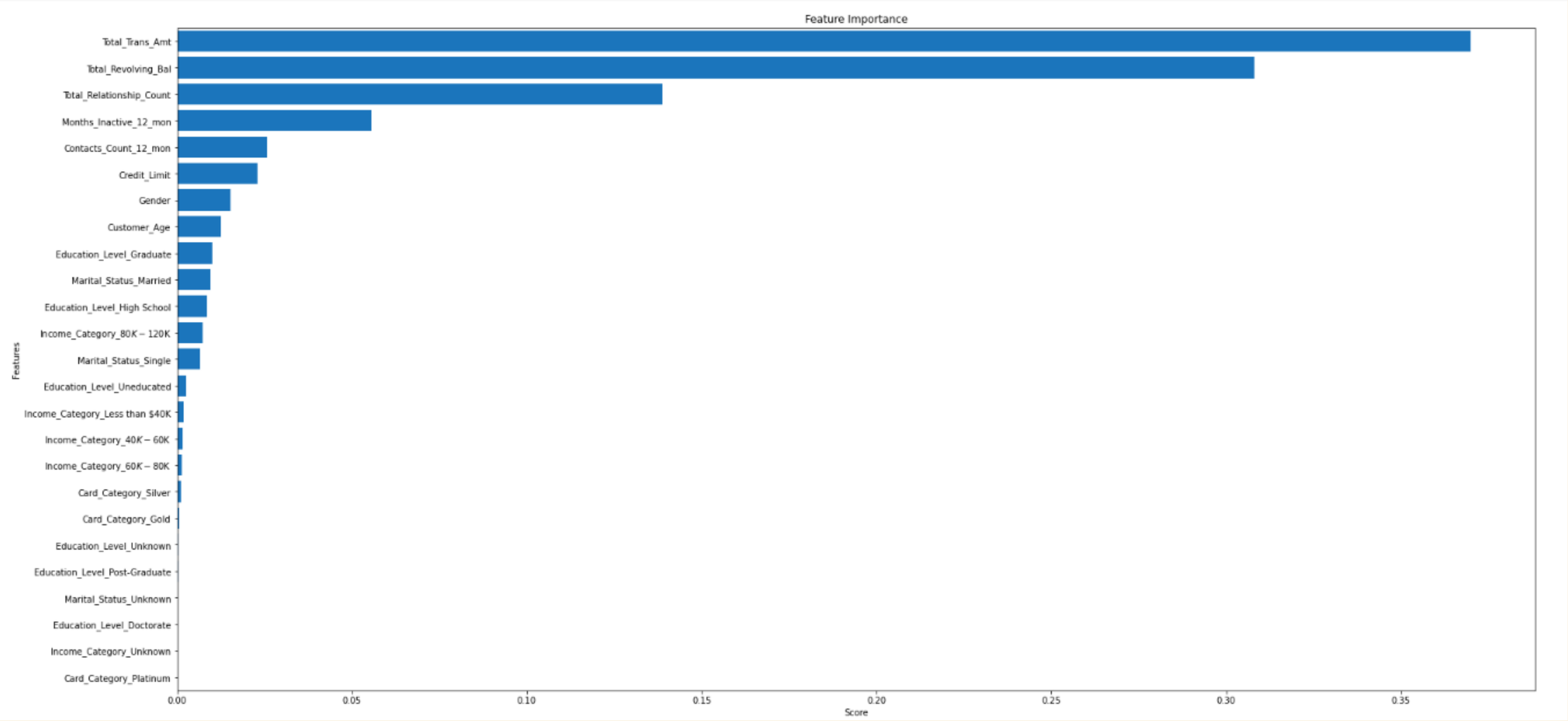
KNN-CLASSIFICATION REPORT

	precision	recall	f1-score	support
0	0.89	0.88	0.89	2543
1	0.42	0.44	0.43	496
accuracy			0.81	3039
macro avg	0.66	0.66	0.66	3039
weighted avg	0.81	0.81	0.81	3039

Interpretation :

From the comparison of the random forest and KNN models, it can be seen that the random forest is the best model by having an f1 score that is better than KNN with a value of 0.91 and 0.78 with a higher accuracy value than the KNN model with a value of 0.87.

FEATURE IMPORTANCE



There are three important feature in predicting the target variable customer churn.

`Total_Trans_Amount, Total_Revolving_Balance, Total_Relantionship_Count

RECOMMENDATIONS

ACCORDING FEATURE IMPORTANCE,

Total_Trans_Amount

- if Total_Trans_Amount is larger, this will increase the bill and interest on credit card payments (if the customer cannot pay on time) this can result in the customer running away from his responsibilty, although there are risks if the customer does not fulfill his responsibilty such as a blacklist. The thing that can be done by the bank is to provide a credit card payment relief program if the customer has entered the fail to pay category, so this will benefit both parties.
- Banks can provide credit card promos if customers make large transactions, such as cash back, discount, earn points to be exchanged for hotel stays or flights or changing customer category to privilage or priority customer if they have big amount.

Total_Revolving_Balance

- Revolving balance is called by If you don't pay the balance on your revolving credit account in full every month, the unpaid portion carries over to the next month. To prevent swelling of revolving credit card balances for customers which will later have an impact on customers leaving is to improve the system on Know Your Customer investigations, Conducting due diligence, Setting accurate credit limits and Using a reputable credit reference agency for more in-depth due diligence can give you invaluable insight and alert you to any potential red flags.

Total_Relantionship_Count

- change or add credit card products according to customer needs, such as if the customer is married, the bank provides health or education insurance benefits according to the credit card limit given or the number of transactions made. or for people who like to travel, banks can make special credit cards for people who have a hobby of traveling with the benefit of collecting points from each traveling transaction. points that have been collected can be exchanged for hotel accommodation or airline tickets. so that if customers have credit cards according to their respective profiles this can reduce the customer churn rate

Thankyou