Business Presentation

Credit Card Users Churn Prediction Case Study

Background & Context

- **Thera Bank** saw a **steep decline** in the **number of users** of their **credit card**.
- **Credit cards** are a **good source of income** for banks because of different kinds of fees charged by the banks **like annual fees, balance transfer fees, and cash advance fees, late payment fees, foreign transaction fees, and others**. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances.
- **Customers leaving credit cards** services **would lead bank to loss**, so the **bank wants to analyze the data of customers** and **identify** the **customers who will leave their credit card services** and **reason for same** – so that bank could improve upon those areas.

Objective

- **Explore and visualize data**
- **Build a model**

  Model should be able to

  - To predict whether a Credit card customer will leave the bank or not.
  - Which variables are most significant in customer attrition.
  - Which customers should be focused more for retention.
- **Draw Conclusions and Business Recommendations**

Data Overview

- Data contains **10127 (rows)** customer data with **21 (columns)** characteristics.

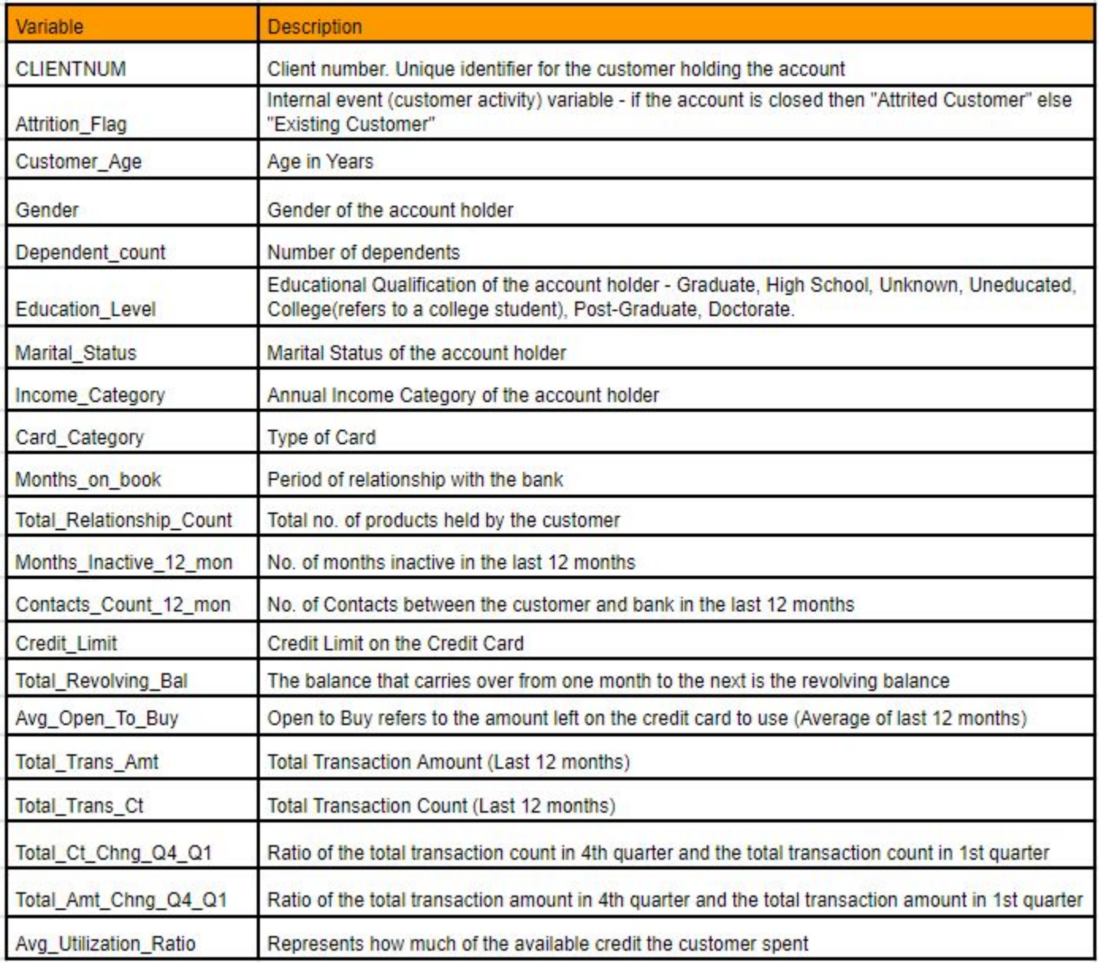
- There are **missing values** in **Education_Level** and **Marital_Status columns.**

- **No duplicates** in dataset.

- The **CLIENTNUM - Customer ID** column is of no significance and will be **removed** for analysis.

- The **Income Category** has some **'abc' values** which will be **imputed**.

- Customer_Age, Months_on_book, Credit_Limit, Total_Revolving_Bal, Avg_Open_To_Buy,Total_Trans_Amt,Total_Trans_Ct,Total_Ct_Chng_Q4_Q1,Total_Amt_Chng_Q4_Q1,Avg_Utilization_Ratio will be binned and mapped to new range columns respectively for EDA purpose and will be dropped before modelling.

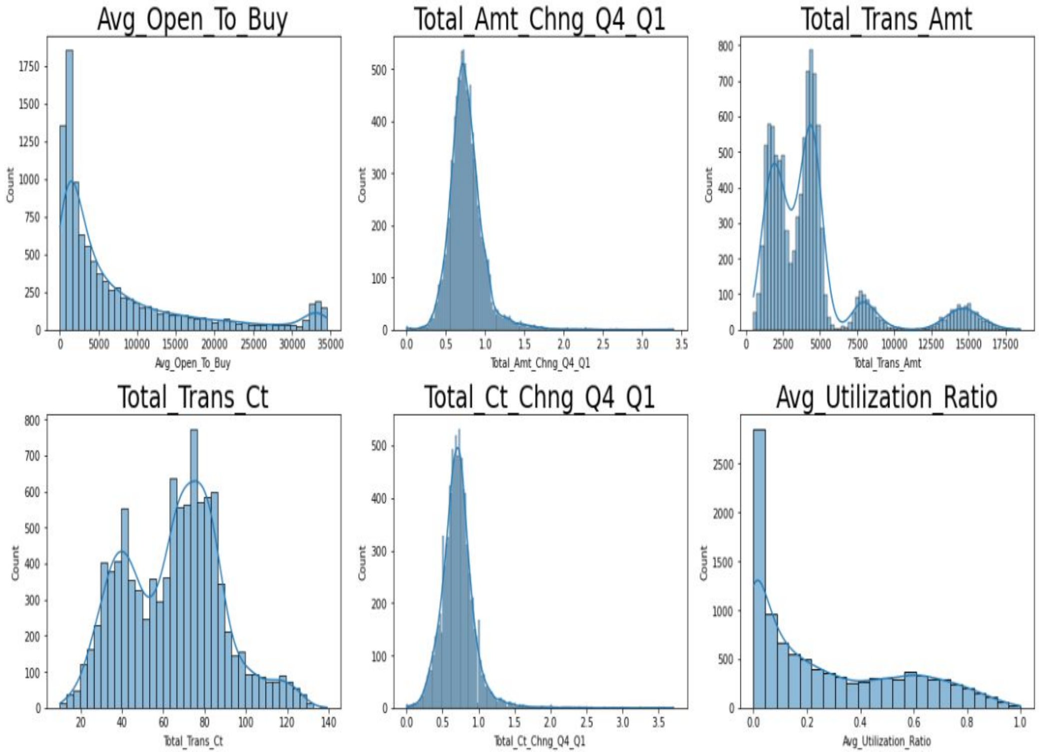


Variable	Description
CLIENTNUM	Client number. Unique identifier for the customer holding the account
Attrition_Flag	Internal event (customer activity) variable - if the account is closed then "Attrited Customer" else "Existing Customer"
Customer_Age	Age in Years
Gender	Gender of the account holder
Dependent_count	Number of dependents
Education_Level	Educational Qualification of the account holder - Graduate, High School, Unknown, Uneducated, College(refers to a college student), Post-Graduate, Doctorate.
Marital_Status	Marital Status of the account holder
Income_Category	Annual Income Category of the account holder
Card_Category	Type of Card
Months_on_book	Period of relationship with the bank
Total_Relationship_Count	Total no. of products held by the customer
Months_Inactive_12_mon	No. of months inactive in the last 12 months
Contacts_Count_12_mon	No. of Contacts between the customer and bank in the last 12 months
Credit_Limit	Credit Limit on the Credit Card
Total_Revolving_Bal	The balance that carries over from one month to the next is the revolving balance
Avg_Open_To_Buy	Open to Buy refers to the amount left on the credit card to use (Average of last 12 months)
Total_Trans_Amt	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Ratio of the total transaction count in 4th quarter and the total transaction count in 1st quarter
Total_Amt_Chng_Q4_Q1	Ratio of the total transaction amount in 4th quarter and the total transaction amount in 1st quarter
Avg_Utilization_Ratio	Represents how much of the available credit the customer spent

Exploratory Data Analysis - Univariate

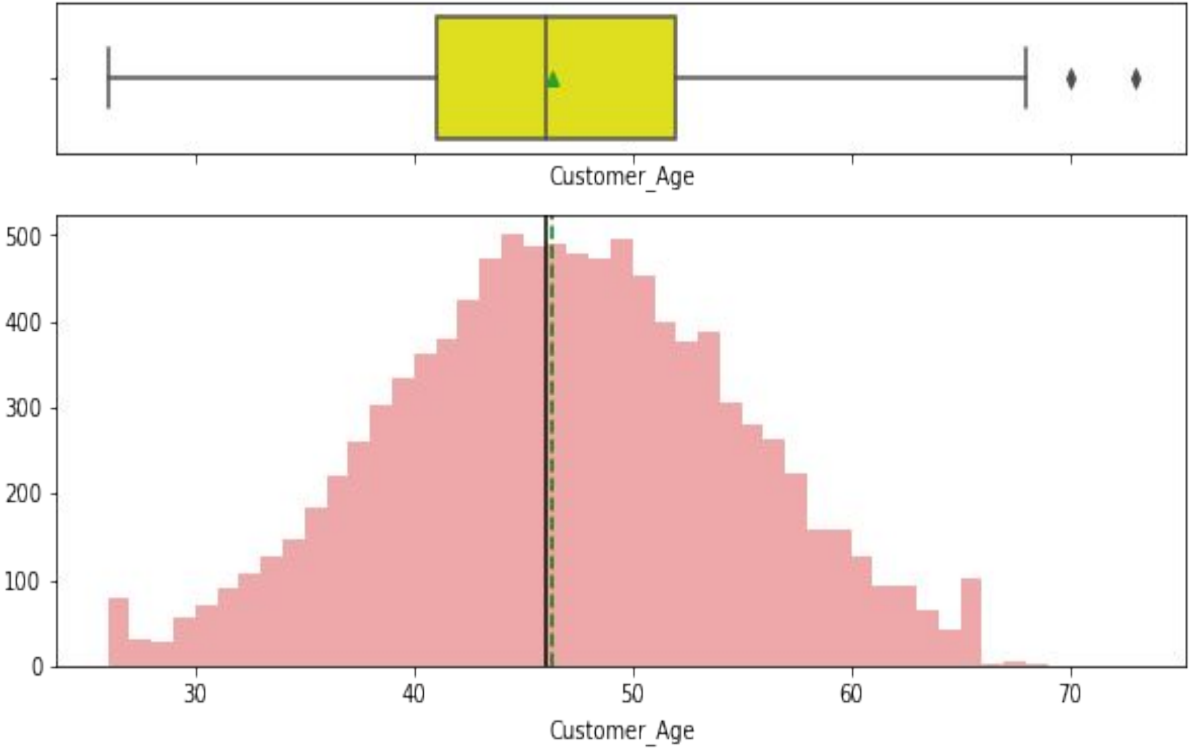
Observations:

- **Avg_Open_To_Buy:** Data is heavily right skewed with lot of outliers on the right tail.
- **Total_Amt_Chng_Q4_Q1:** Data has a normal distribution with a lot of outliers on the right end.
- **Total_Trans_Amt:** This distribution is multimodal with lots of peaks.
- **Total_Trans_Ct:** This distribution is bimodal with two peaks.
- **Total_Ct_Chng_Q4_Q1:** Data has a normal distribution with a lot of outliers on the right end.
- **Avg_Utilization_Ratio:** Data is heavily right skewed with lot of outliers.



Exploratory Data Analysis - Univariate

Customer_Age

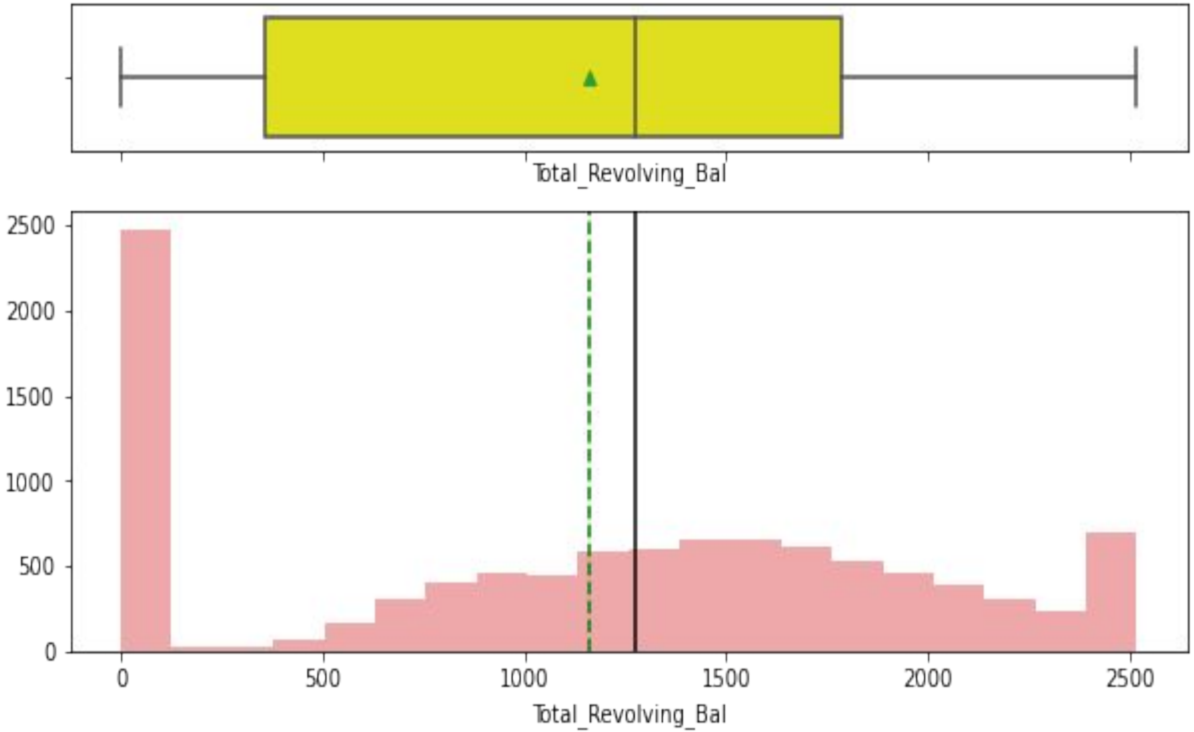


Observations:

- Average age of people in the dataset is 46 years, age has a wide range from 26 to 73 years.
- There are outliers beyond 66 years.

Exploratory Data Analysis - Univariate

Total Revolving Balance



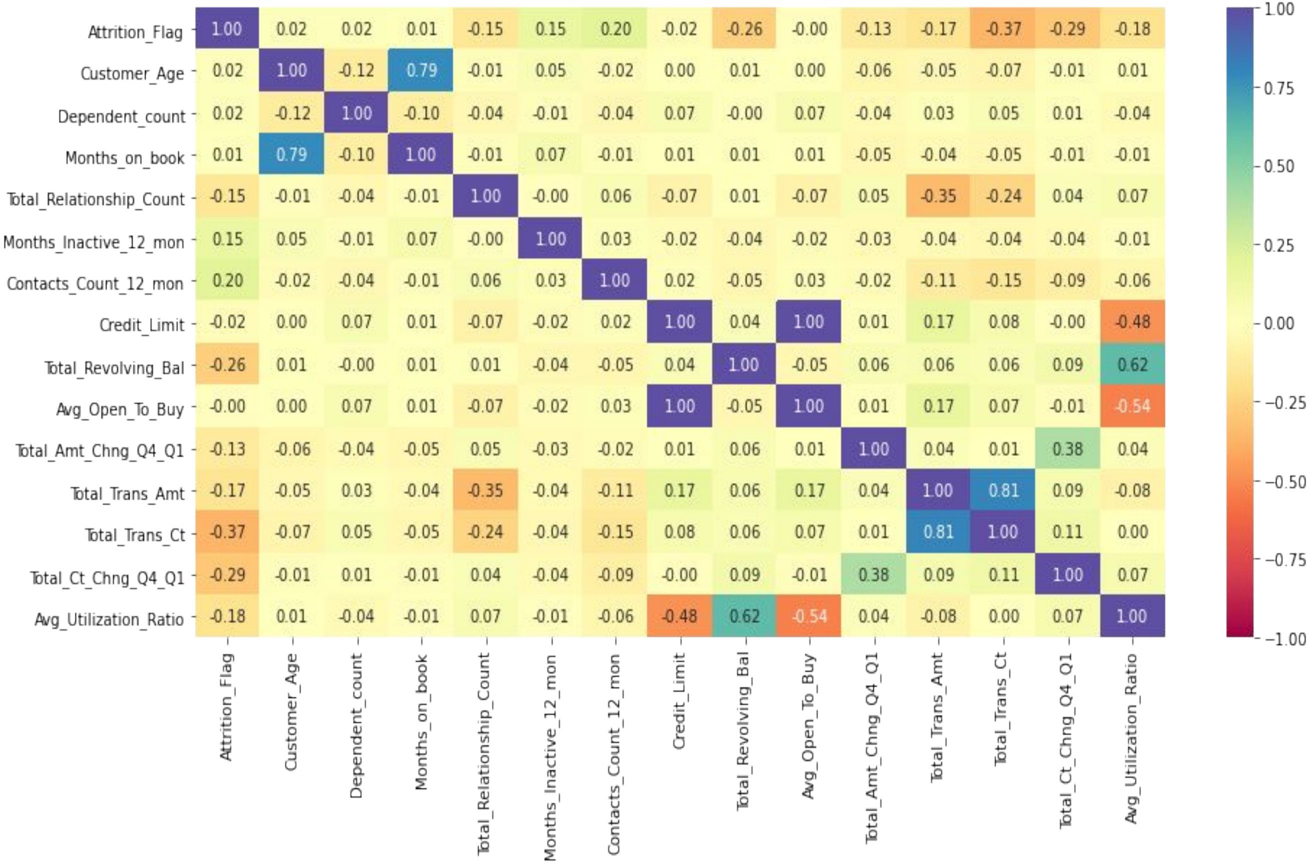
Observations:

- 25% of the customers do not use the card actively and has a revolving balance of 0 dollars.
- 8% of the customers heavily use the card with a balance of 2500 dollars.
- We can see higher counts at the max and minimum values.

Exploratory Data Analysis - Bivariate Heat Map

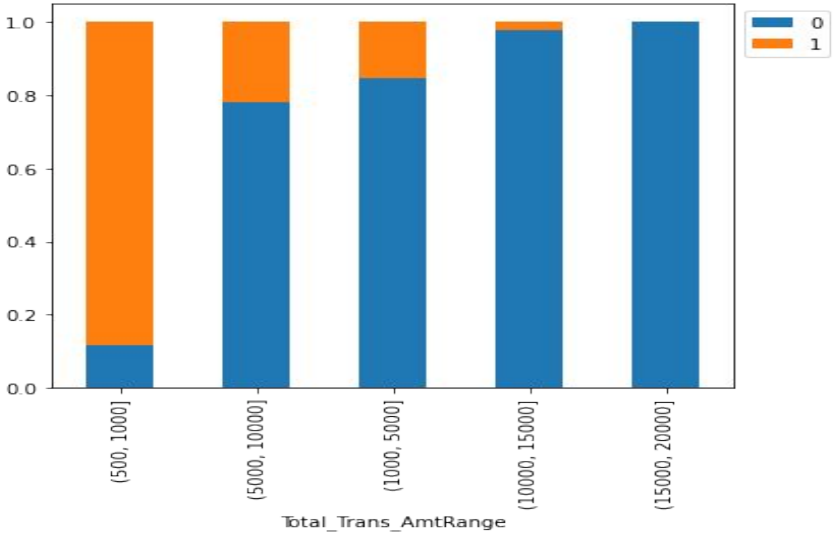
Observations:

- **Credit_Limit and Avg_Open_To_Buy are strongly correlated with a 1 value and we can drop one of these columns during modelling.**
- CCAvg and Income is also 0.8 correlated.
- Months_on_book and Customer Age have a high correlation of 0.79
- Avg utilization ratio is derived using credit limit , Avg_Open_to_buy and Total Revolving Bal, so we can see some weaker negative and positive correlation for those three variables.
- We could not see any strong correlation between our dependent variable Attrition_Flag and others.



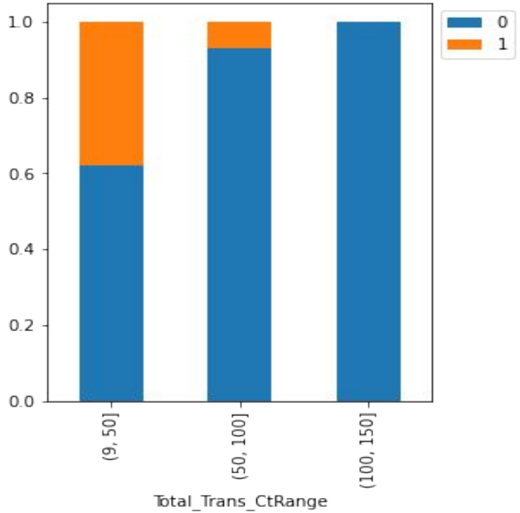
Exploratory Data Analysis - Bivariate

Total_Trans_Amt



- Customers with **less transaction amount** between **500-1000** dollars have **attrited** the most.

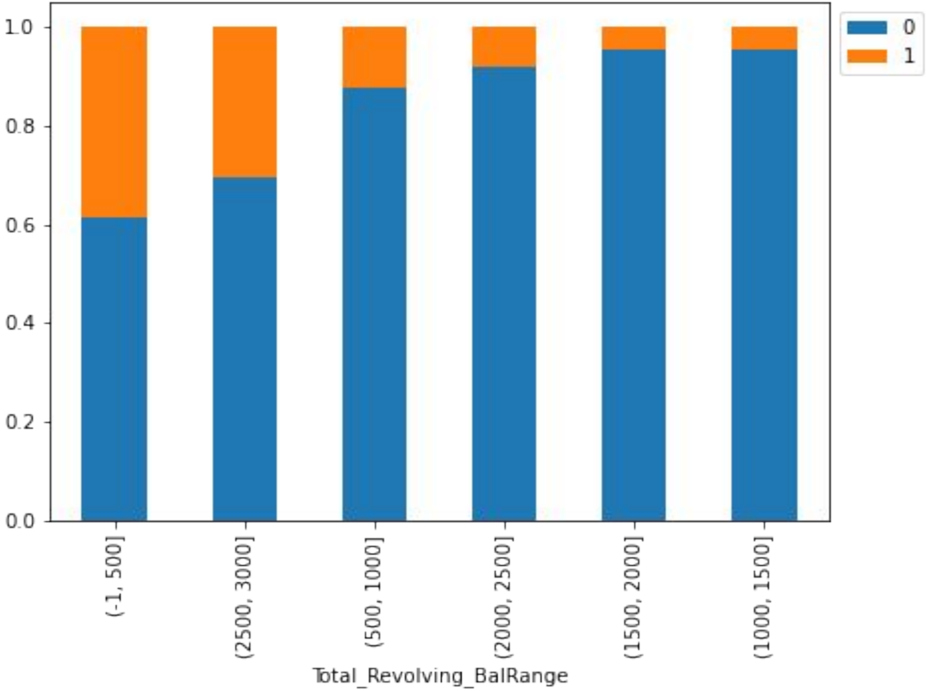
Total_Trans_Ct



- Customer with a **low transaction count** of **9-50** have **attrited** the most.

Exploratory Data Analysis - Bivariate

Total_Revolving_Bal

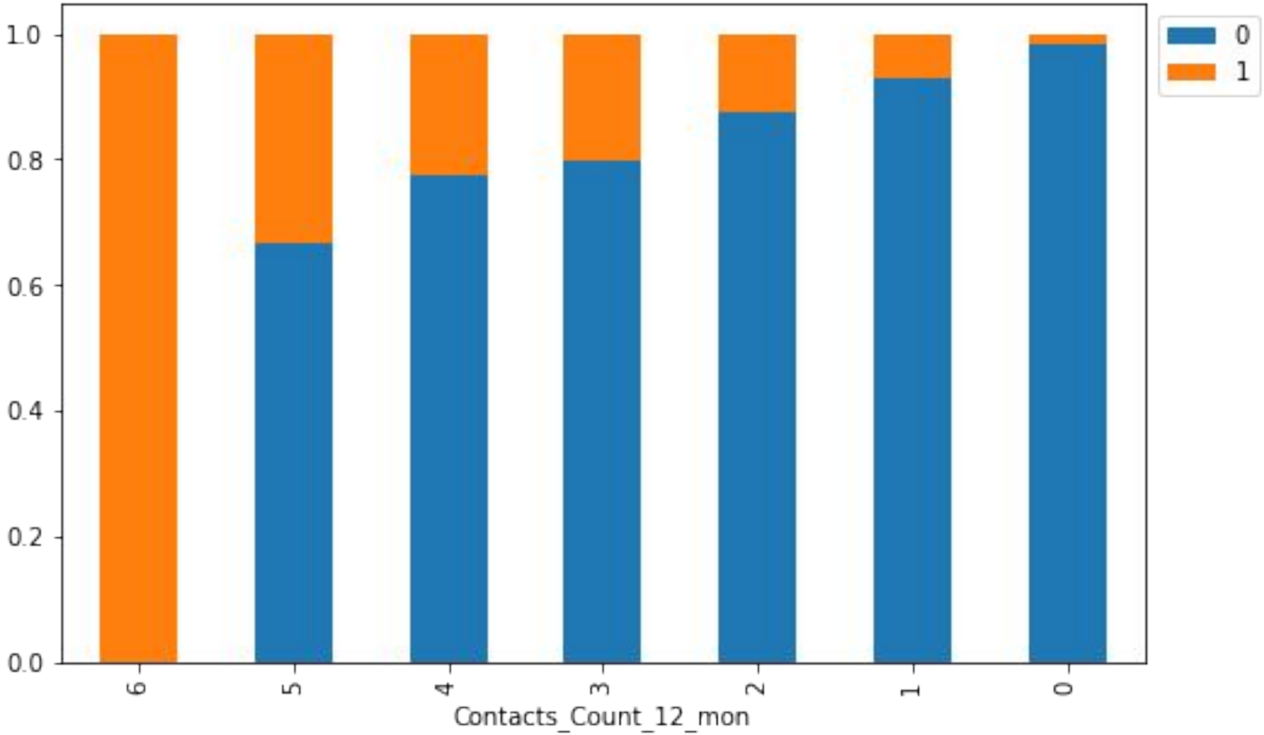


Observations:

- Customer with a **low revolving balance** of **0-500** dollars have **attrited** the most.

Exploratory Data Analysis - Bivariate

Contacts_Count_12_mon

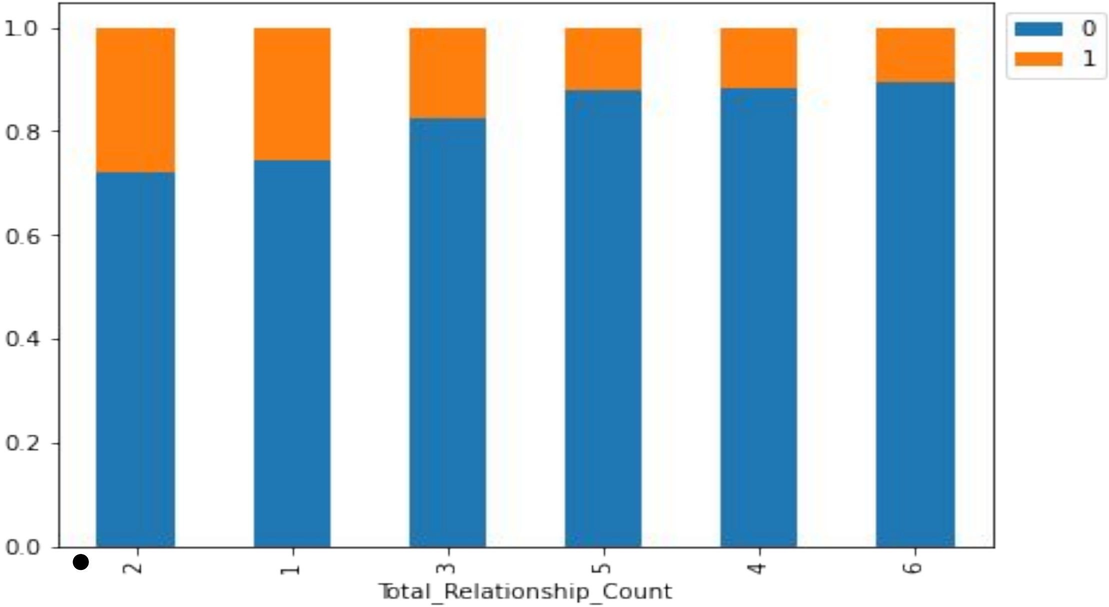


Observations:

- We could see a clear pattern here, the **more number of times** the **customer contacts a bank,** the **more likely** he is going **to attrite**.Looks like the banks Customer service is not that great to make the customer happy.

Exploratory Data Analysis - Bivariate

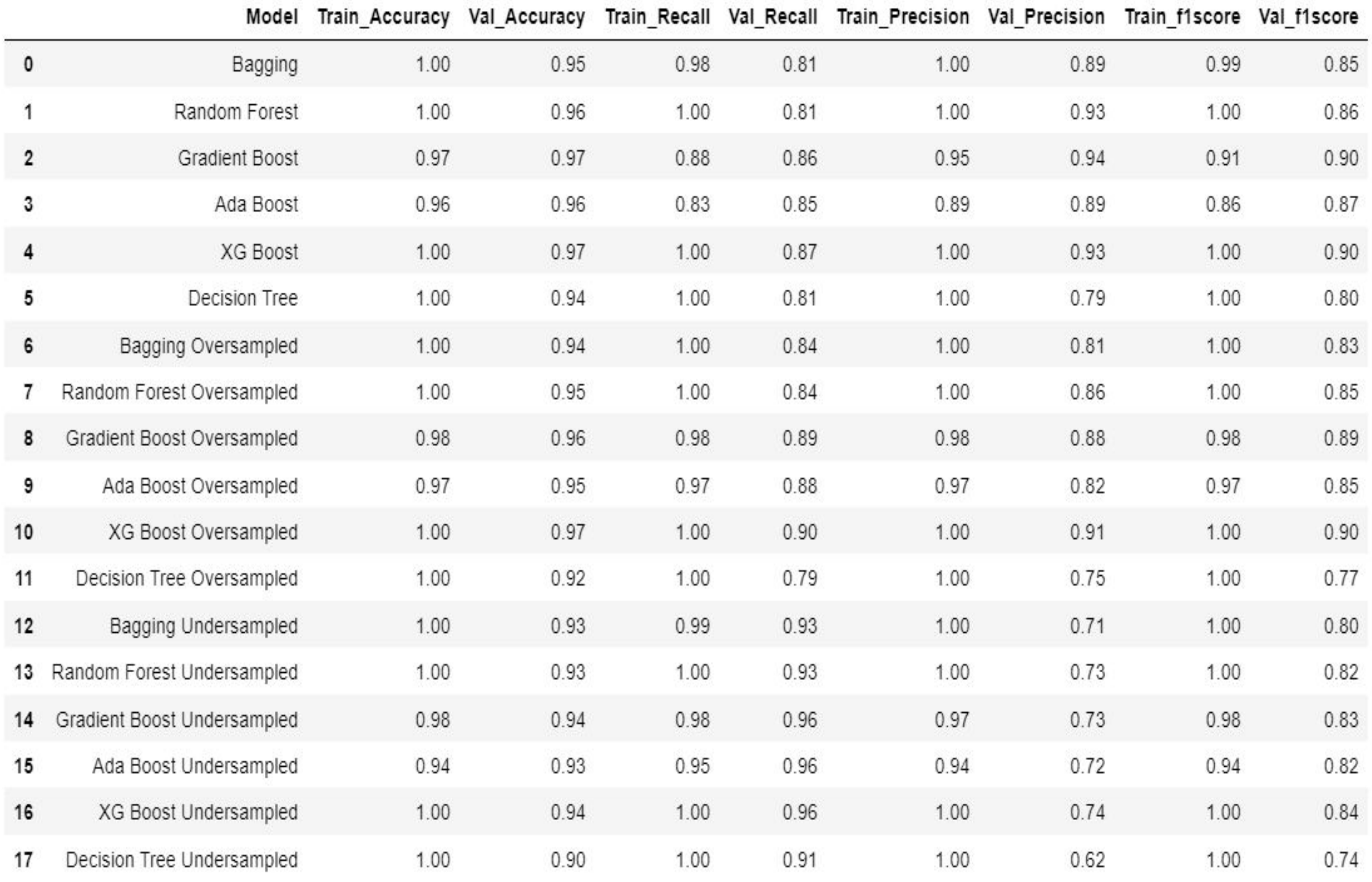
Relationship count



Observations:

- The **more products** customers has from the bank, he/she is **less likely to attrite.**

Model Performance Summary - Comparison

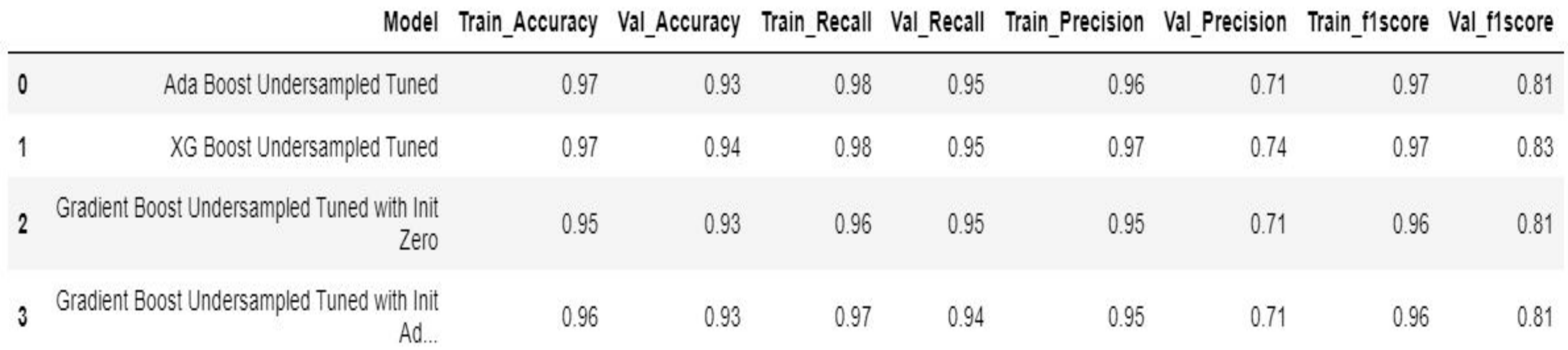


	Model	Train_Accuracy	Val_Accuracy	Train_Recall	Val_Recall	Train_Precision	Val_Precision	Train_f1score	Val_f1score
0	Bagging	1.00	0.95	0.98	0.81	1.00	0.89	0.99	0.85
1	Random Forest	1.00	0.96	1.00	0.81	1.00	0.93	1.00	0.86
2	Gradient Boost	0.97	0.97	0.88	0.86	0.95	0.94	0.91	0.90
3	Ada Boost	0.96	0.96	0.83	0.85	0.89	0.89	0.86	0.87
4	XG Boost	1.00	0.97	1.00	0.87	1.00	0.93	1.00	0.90
5	Decision Tree	1.00	0.94	1.00	0.81	1.00	0.79	1.00	0.80
6	Bagging Oversampled	1.00	0.94	1.00	0.84	1.00	0.81	1.00	0.83
7	Random Forest Oversampled	1.00	0.95	1.00	0.84	1.00	0.86	1.00	0.85
8	Gradient Boost Oversampled	0.98	0.96	0.98	0.89	0.98	0.88	0.98	0.89
9	Ada Boost Oversampled	0.97	0.95	0.97	0.88	0.97	0.82	0.97	0.85
10	XG Boost Oversampled	1.00	0.97	1.00	0.90	1.00	0.91	1.00	0.90
11	Decision Tree Oversampled	1.00	0.92	1.00	0.79	1.00	0.75	1.00	0.77
12	Bagging Undersampled	1.00	0.93	0.99	0.93	1.00	0.71	1.00	0.80
13	Random Forest Undersampled	1.00	0.93	1.00	0.93	1.00	0.73	1.00	0.82
14	Gradient Boost Undersampled	0.98	0.94	0.98	0.96	0.97	0.73	0.98	0.83
15	Ada Boost Undersampled	0.94	0.93	0.95	0.96	0.94	0.72	0.94	0.82
16	XG Boost Undersampled	1.00	0.94	1.00	0.96	1.00	0.74	1.00	0.84
17	Decision Tree Undersampled	1.00	0.90	1.00	0.91	1.00	0.62	1.00	0.74

Observations:

- **Top three models** are **Adaboost Undersampled, XG Boost Undersampled and Gradient Boost Undersampled** with validation **recall** of **96**.
- We could see some **overfitting** in **XGBoost** training vs test recall.
- **Adaboost** has performed very well in **validation** data **better than training**.
- **Gradient boost** has also **generalized well** in **validation** data.

Model Performance Summary - Tuning Comparison(Top 3 Model)



	Model	Train_Accuracy	Val_Accuracy	Train_Recall	Val_Recall	Train_Precision	Val_Precision	Train_f1score	Val_f1score
0	Ada Boost Undersampled Tuned	0.97	0.93	0.98	0.95	0.96	0.71	0.97	0.81
1	XG Boost Undersampled Tuned	0.97	0.94	0.98	0.95	0.97	0.74	0.97	0.83
2	Gradient Boost Undersampled Tuned with Init Zero	0.95	0.93	0.96	0.95	0.95	0.71	0.96	0.81
3	Gradient Boost Undersampled Tuned with Init Ad...	0.96	0.93	0.97	0.94	0.95	0.71	0.96	0.81

What are the Top 3 Models and why we need to tune them?

- **Ada boost, XGBoost, Gradient Boost** all of **Undersampled** data gave the **best recall scores(our metric of interest)** and were selected as **top 3 models** for **hyperparameter tuning.**
- We have to tune them **to reduce overfitting** and to create a **best performing generalized model** with **best parameters** that will **work well on unseen real world data.**

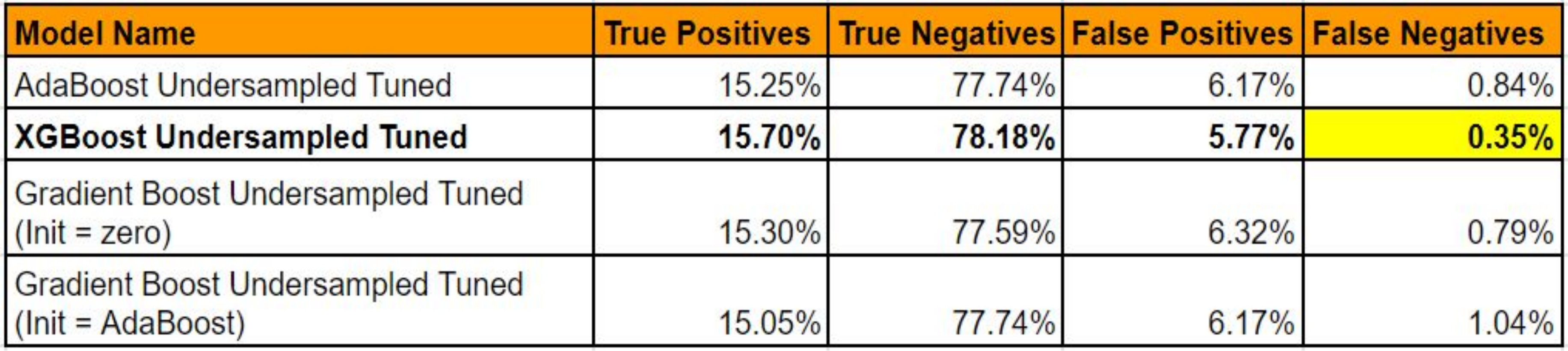
Observations:

- **Out of** the **top three models** that were hypertuned, **XGBoost gave the best results for all the four metrics in comparison.**

Note:

Gradient Boost model was tuned for both Init as zero and Init as AdaBoost model being one of the top models.

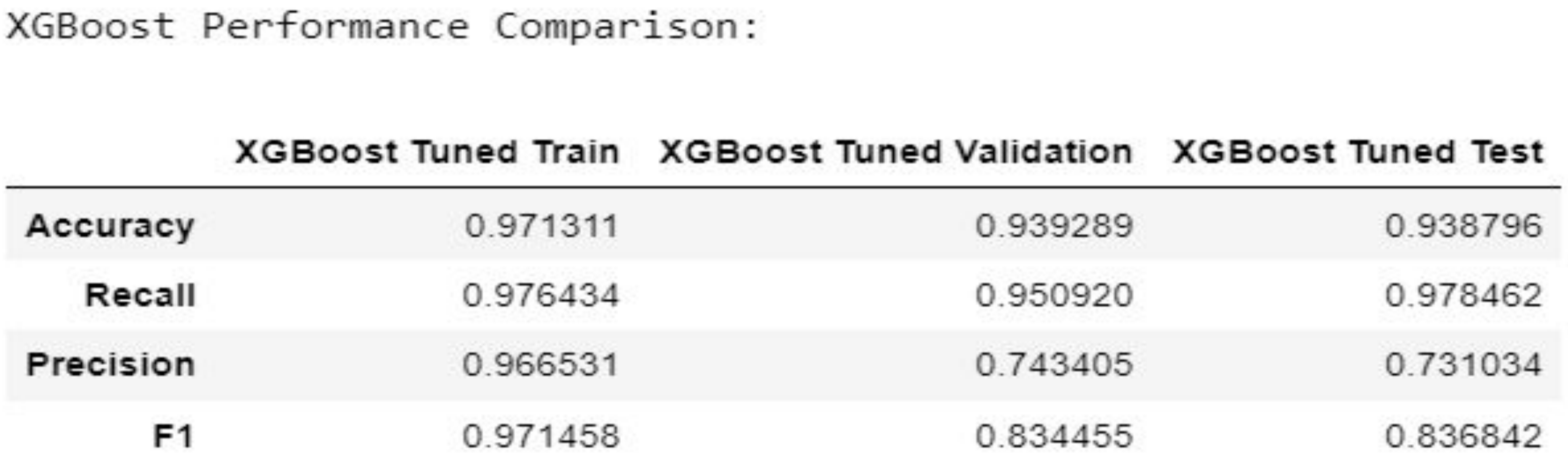
Model Performance - Confusion Matrix Comparison(Top 3 Models)



Model Name	True Positives	True Negatives	False Positives	False Negatives
AdaBoost Undersampled Tuned	15.25%	77.74%	6.17%	0.84%
XGBoost Undersampled Tuned	**15.70%**	**78.18%**	**5.77%**	**0.35%**
Gradient Boost Undersampled Tuned (Init = zero)	15.30%	77.59%	6.32%	0.79%
Gradient Boost Undersampled Tuned (Init = AdaBoost)	15.05%	77.74%	6.17%	1.04%

- In our case of **predicting Attriting Customers,not being able to identify a potential customer who will attrite is the biggest loss Bank can face.**
- **Recall** is the **right metric** to check the performance of the model.
- We should have the **lowest false negative rates** to **avoid the potential customer retention loss.**
- In our **top three models XGBoost Tuned on Undersampled data** gave the **best recall 97.8%** and **lowest False negatives of 0.35%**

Model Performance Summary - Best Model Test Performance



XGBoost Performance Comparison:

	XGBoost Tuned Train	XGBoost Tuned Validation	XGBoost Tuned Test
Accuracy	0.971311	0.939289	0.938796
Recall	0.976434	0.950920	0.978462
Precision	0.966531	0.743405	0.731034
F1	0.971458	0.834455	0.836842

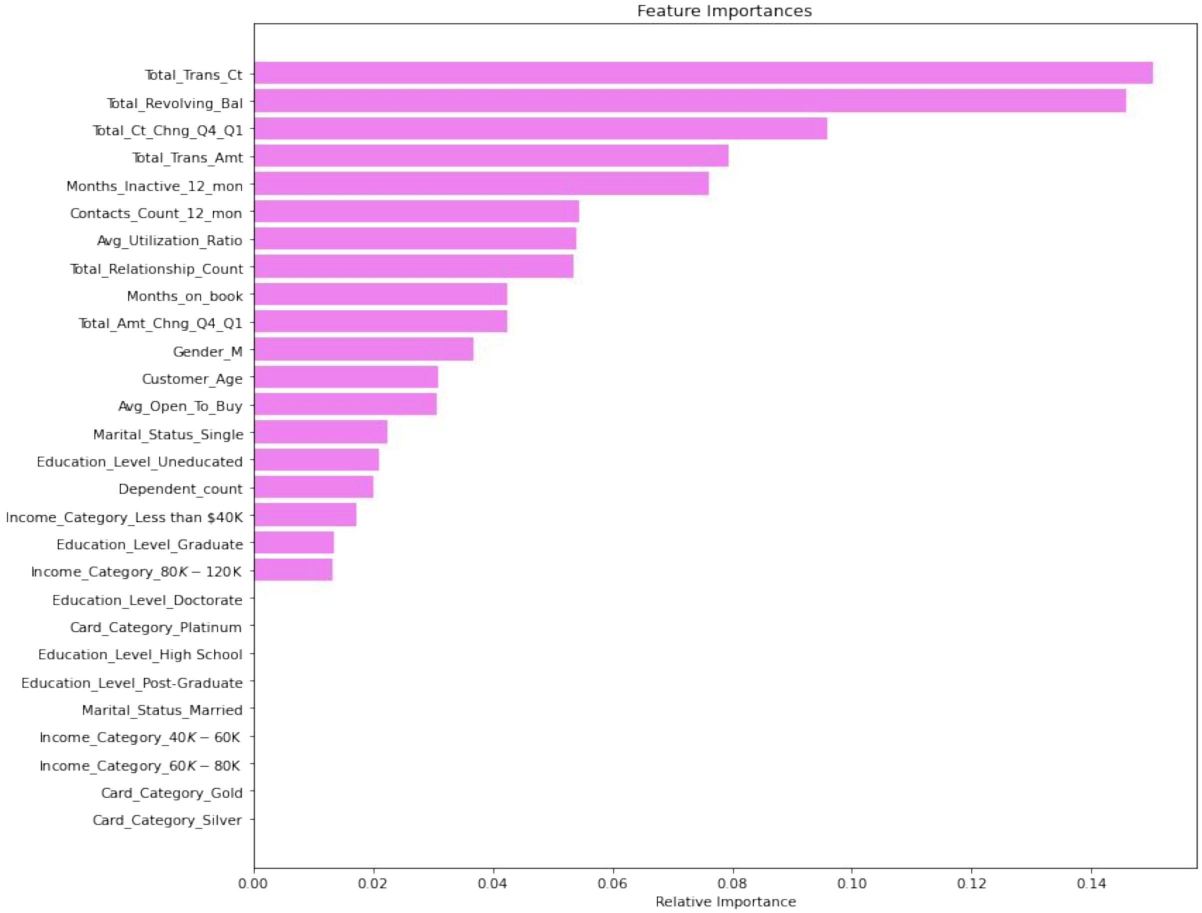
- **XGBoost Undersampled has given the best recall score of 97.8% in test and 97.6% in training indicating a very good performance without any overfitting.**

- **This is a generalized model that can be used for predicting the Attriting customers by the bank.**

Feature Importance - XG Boost Undersampled Tuned

The model indicates that the most significant predictors of Potential Attriting Customers are

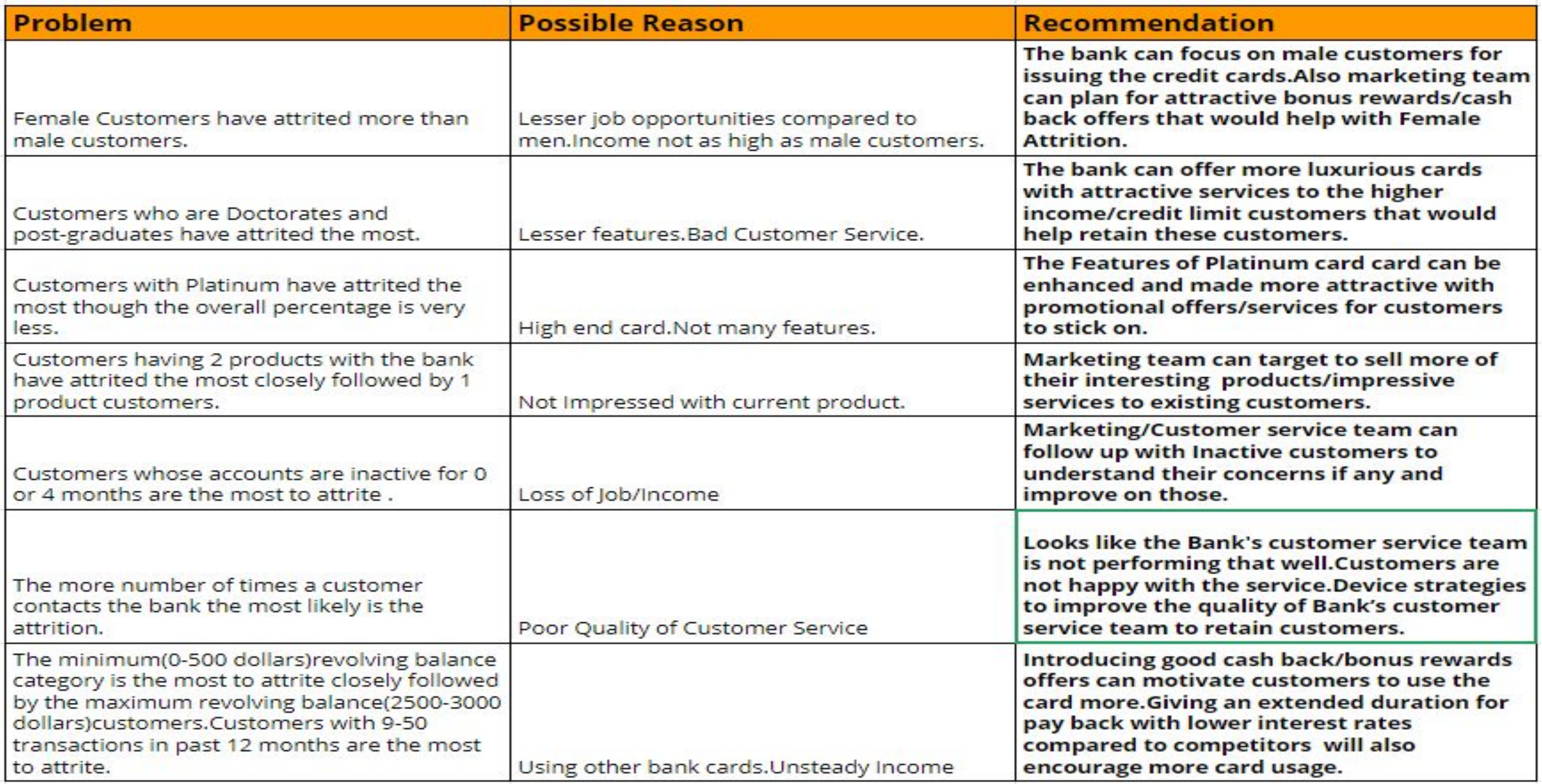
1. **Total_Trans_Ct** is between **9-50 transactions.**
2. **Total_Revolving_Bal** is between **0-500 dollars.**
3. **Total_Trans_Amt** is between **0-500 dollars.**
4. **Total_Ct_Chng_Q4_Q1 ratio** is between **0-0.5 range**.
5. **Months_Inactive_12_mon** value is either **0 or 4.**



Conclusion

- **XGBoost Undersampled Tuned** Model gave us the **highest recall score of 97.6% on train and 97.8% on test data** and the least error of **0.35% False negatives** and 5.77% False positives.
- **Total_Trans_Ct,Total_Revolving_Bal, Total_Trans_Amt, Total_Ct_Chng_Q4_Q1, Months_Inactive_12_mon** are the key variables that has strong relationship with the dependent variable.**Lower the values** for these variables,**higher chances of Customer Attrition.**
- The **bivariate results of EDA clearly matched with the the XGBoost predicted important variables**.

Recommendations



Problem	Possible Reason	Recommendation
Female Customers have attrited more than male customers.	Lesser job opportunities compared to men.Income not as high as male customers.	**The bank can focus on male customers for issuing the credit cards.Also marketing team can plan for attractive bonus rewards/cash back offers that would help with Female Attrition.**
Customers who are Doctorates and post-graduates have attrited the most.	Lesser features.Bad Customer Service.	**The bank can offer more luxurious cards with attractive services to the higher income/credit limit customers that would help retain these customers.**
Customers with Platinum have attrited the most though the overall percentage is very less.	High end card.Not many features.	**The Features of Platinum card card can be enhanced and made more attractive with promotional offers/services for customers to stick on.**
Customers having 2 products with the bank have attrited the most closely followed by 1 product customers.	Not Impressed with current product.	**Marketing team can target to sell more of their interesting products/impressive services to existing customers.**
Customers whose accounts are inactive for 0 or 4 months are the most to attrite .	Loss of Job/Income	**Marketing/Customer service team can follow up with Inactive customers to understand their concerns if any and improve on those.**
The more number of times a customer contacts the bank the most likely is the attrition.	Poor Quality of Customer Service	**Looks like the Bank's customer service team is not performing that well.Customers are not happy with the service.Device strategies to improve the quality of Bank's customer service team to retain customers.**
The minimum(0-500 dollars)revolving balance category is the most to attrite closely followed by the maximum revolving balance(2500-3000 dollars)customers.Customers with 9-50 transactions in past 12 months are the most to attrite.	Using other bank cards.Unsteady Income	**Introducing good cash back/bonus rewards offers can motivate customers to use the card more.Giving an extended duration for pay back with lower interest rates compared to competitors will also encourage more card usage.**