Credit Card User

Nibu Kuriakose

Jun 2021

Content

1. Introduction and Proposed Approach

Background

- Thera bank recently saw a decline in the number of users of its credit card service. Bank charges different types of fees on credit cards. Decrease in number of credit card users will lead to a loss for the bank.
- Bank wants to identify the credit card customers who might leave and the reason for the same.

Purpose and Benefits

- Perform exploratory data analysis to understand about the customers, and potential reasons for customers leaving the credit card service.
- Create a model to predict potential customers who might leave the credit card service. This will allow for an effective way for targeting these customers to avoid them leaving
- Provide insights and recommendation based on the data.

Proposed Approach

1. Perform exploratory data analysis.
2. Build prediction models
2. Find the appropriate performance measure to evaluate the models and perform model improvement to get the best model.

2. Dataset Information and Feature Engineering

- There are 10,127 samples. Each sample has 21 attributes
- It is an imbalanced dataset as only 16% of customers have left which is the target dependent variable for building the model. Hence, sampling techniques might improve model performance.
- There are no null values in the dataset, but there are 'unknown' values in 'educational_level', 'Martital_Status' and 'Income_category' fields. KNN imputation will be used to fill these columns to see whether the model performance can be improved.
- Feature Engineering- Following new fields were created using the existing data:
    - Average transaction amount (Avg_Trans_Amt)= Total Transaction Amount/Total Transaction Count
    - Ratio of average transaction amount in Q4 and Q1 (Avg_Trans_Amt_Chng_Q4_Q1)= Total_Amt_Chng_Q4_Q1/Total_Ct_Chng_Q4_Q1

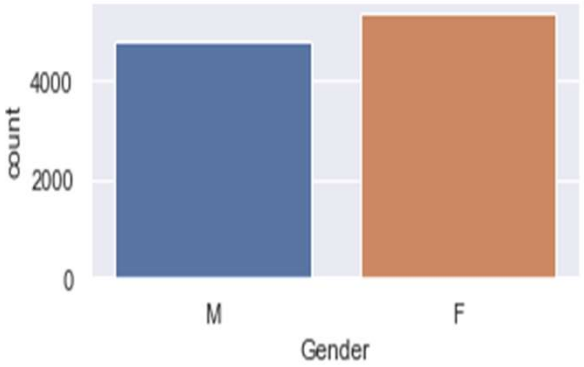
3.1 Univariate Analysis

Attribute	Description
Customer Age	• Slightly right skewed with 50% of customers between 41 and 52 years.
Dependent Count	• Customers with 3 dependents is the highest followed by 2 and 1 dependents.
Months on book	• Customers with 3 years is the highest, with significantly lower customers either more than or less than 3 years.
Total_Relationship_Count	• Customers with 3 products is the highest. The next highest is customers with 4,5 or 6 products.
Months_Inactive_12_mon	• Right skewed, with customers inactive for more than 3 months is much smaller than customers who are inactive for 3 months or less.
Contacts_Count_12_mon	• Right skewed, with highest number of customers are ones who made 3 contacts in the last 12 months followed by 2 and 1 contact.
Credit_Limit	• Right skewed data, with 50% of customers having limit of $4,549 or lesser. There are around 5% of customers with around $34k as credit limit.
Total_Revolving_Bal	• Left skewed with 50% of customers having revolving balance of $1276.00 • Also, the revolving balance of customers with higher credit limit (>$30k) is low.
Avg_Open_To_Buy	• Left skewed with 50% of customers having open to buy as $3,474. • Also, the open to buy for customers with higher credit limit (>$30k) is high showing that they are spending less on their credit card.
Total_Amt_Chng_Q4_Q1, Total_Ct_Chng_Q4_Q1, Avg_Trans_Amt_Chng_Q4_Q1 (newly created field)	• For majority of the customers the spend, transaction count and average transaction amount has gone down in Q4 compared to Q1.
Total_Trans_Amt, Total_Trans_Ct, Avg_Trans_Amt ((newly created field)	• 50% of customers have yearly transaction amount of $3,899, 67 transactions and average transaction amount of circa. $57.
Avg_Utilization_Ratio	• 50% of customers spent only 18% of their credit limit and 75% spent only 50% of their limit.

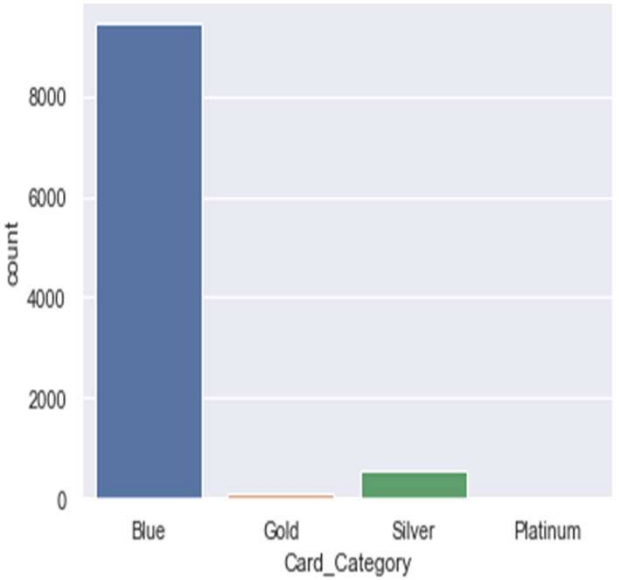
3.1 Univariate Analysis

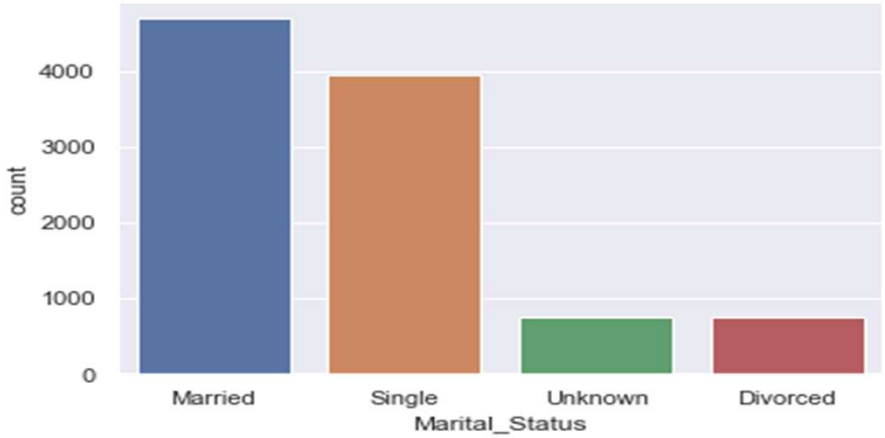
	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
count	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00	10127.00
mean	46.33	2.35	35.93	3.81	2.34	2.46	8631.95	1162.81	7469.14	0.76	4404.09	64.86	0.71	0.27
std	8.02	1.30	7.99	1.55	1.01	1.11	9088.78	814.99	9090.69	0.22	3397.13	23.47	0.24	0.28
min	26.00	0.00	13.00	1.00	0.00	0.00	1438.30	0.00	3.00	0.00	510.00	10.00	0.00	0.00
25%	41.00	1.00	31.00	3.00	2.00	2.00	2555.00	359.00	1324.50	0.63	2155.50	45.00	0.58	0.02
50%	46.00	2.00	36.00	4.00	2.00	2.00	4549.00	1276.00	3474.00	0.74	3899.00	67.00	0.70	0.18
75%	52.00	3.00	40.00	5.00	3.00	3.00	11067.50	1784.00	9859.00	0.86	4741.00	81.00	0.82	0.50
max	73.00	5.00	56.00	6.00	6.00	6.00	34516.00	2517.00	34516.00	3.40	18484.00	139.00	3.71	1.00

3.1 Univariate Analysis



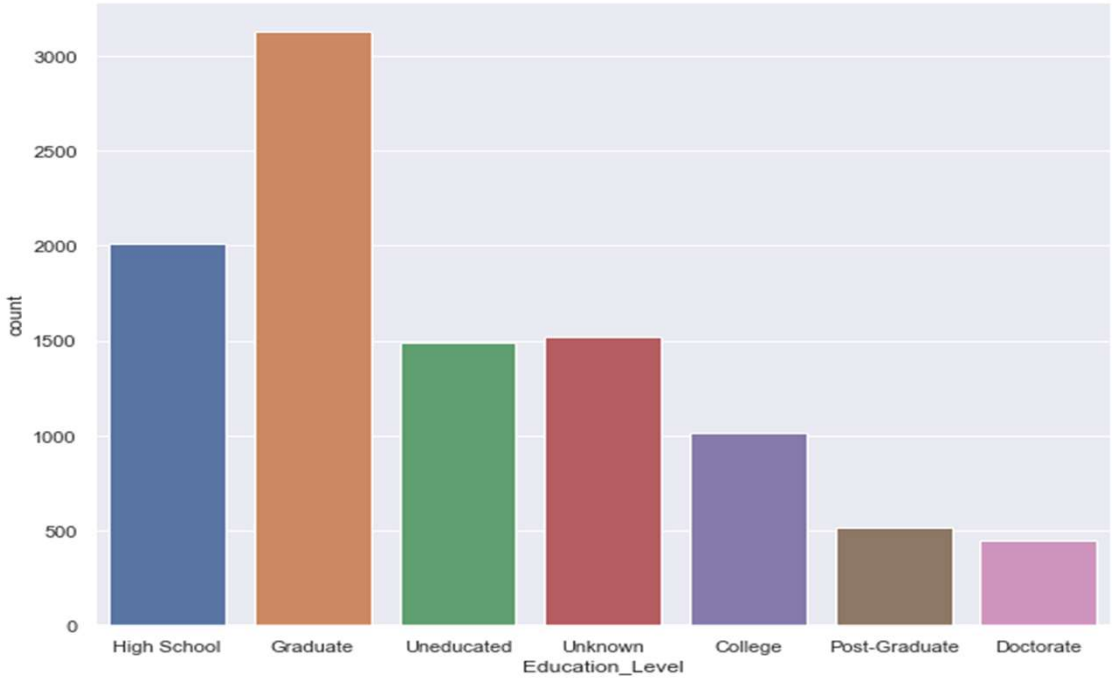
- Majority of customers are females
- Married customers are the highest followed by singles
- Majority of customers have blue card with low numbers for all other categories. Only 0.1% have platinum card.

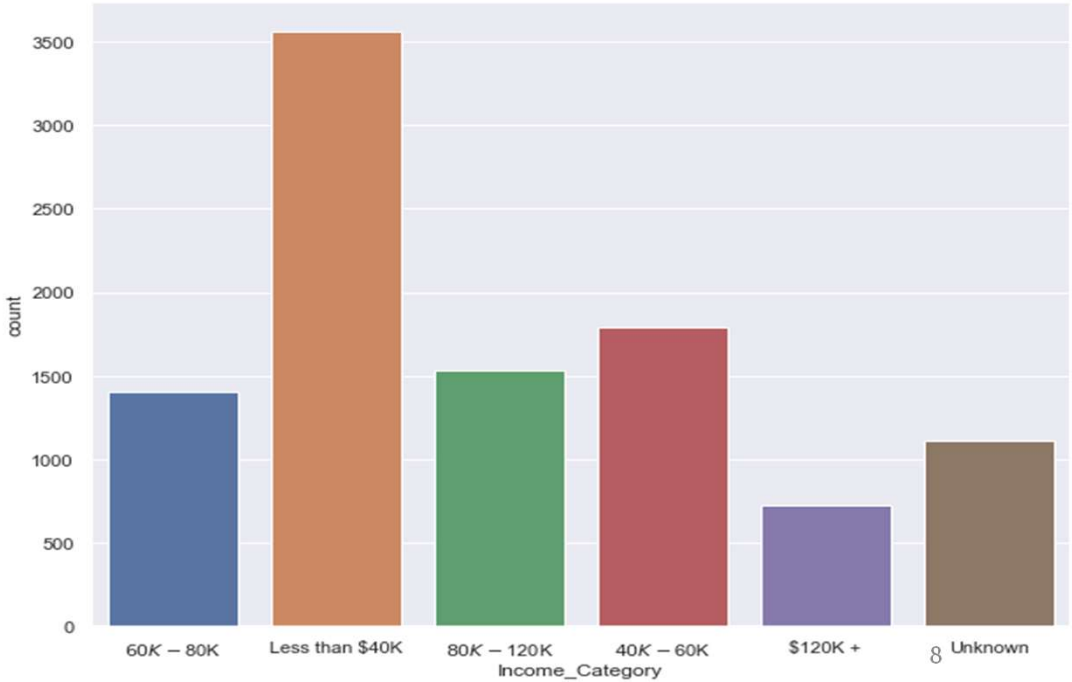




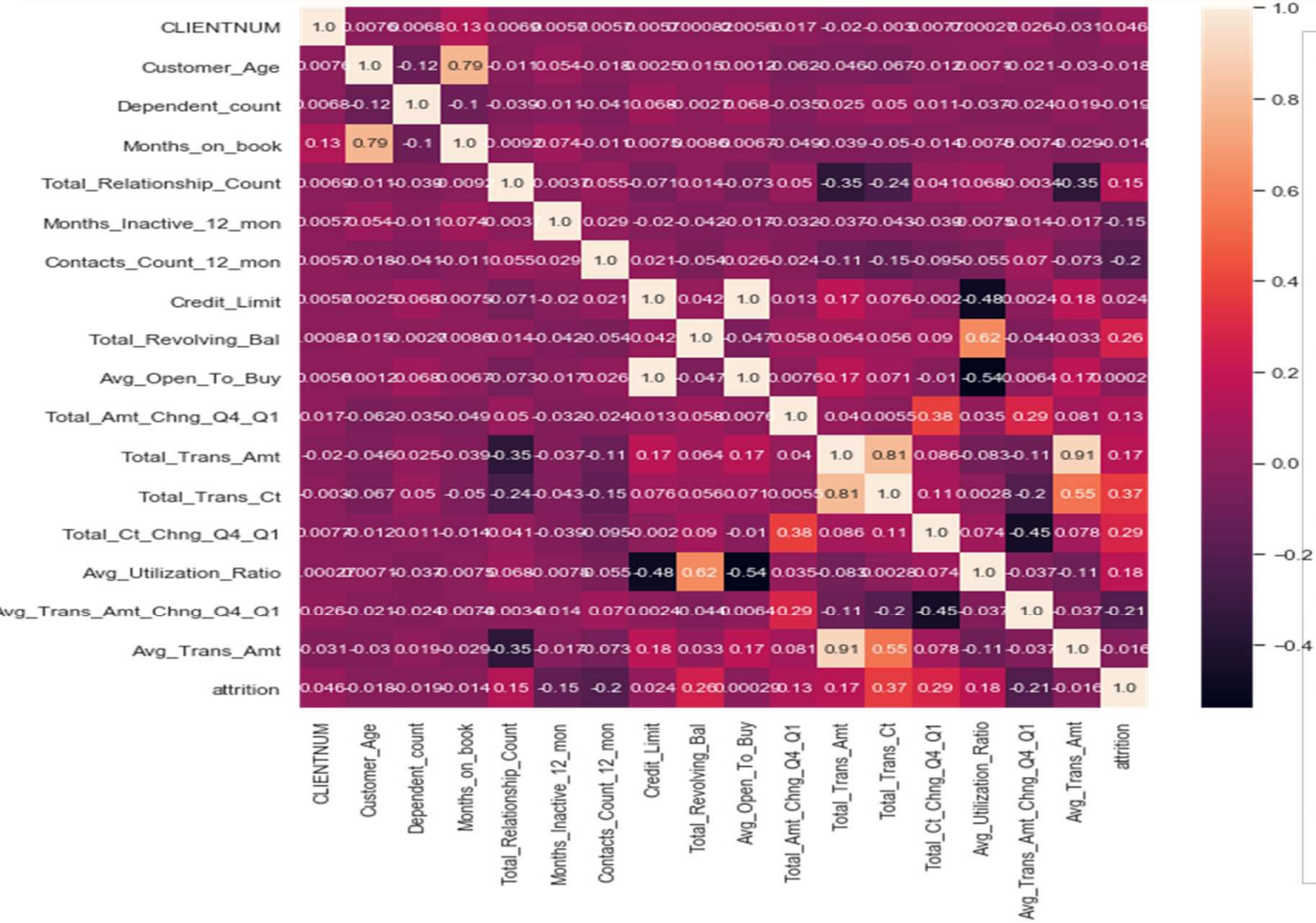
3.1 Univariate Analysis

- Customers who have a graduate degree is the biggest group followed by customers who have attended high school. Low number of post graduate and doctorate customers.
- Majority of customers have income less than $40k.



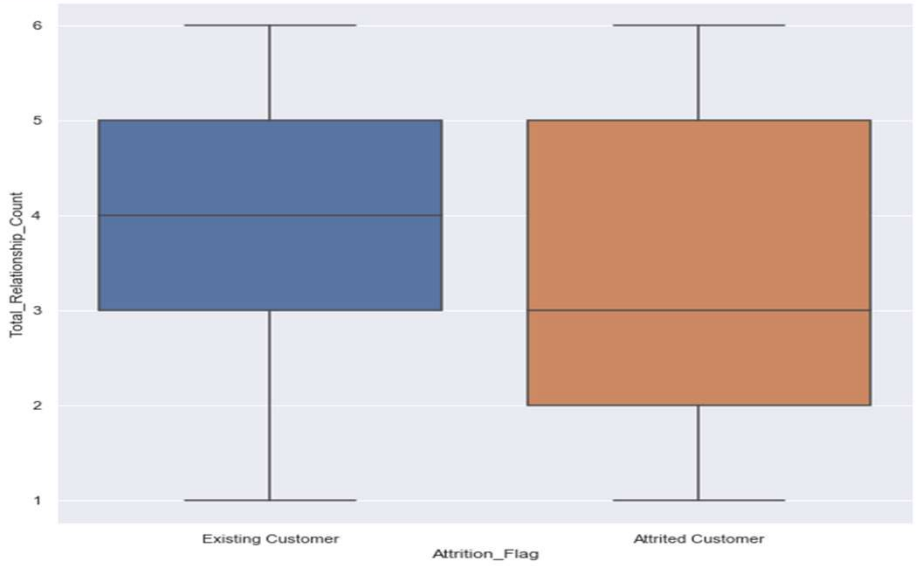


3.2 Multivariate Analysis

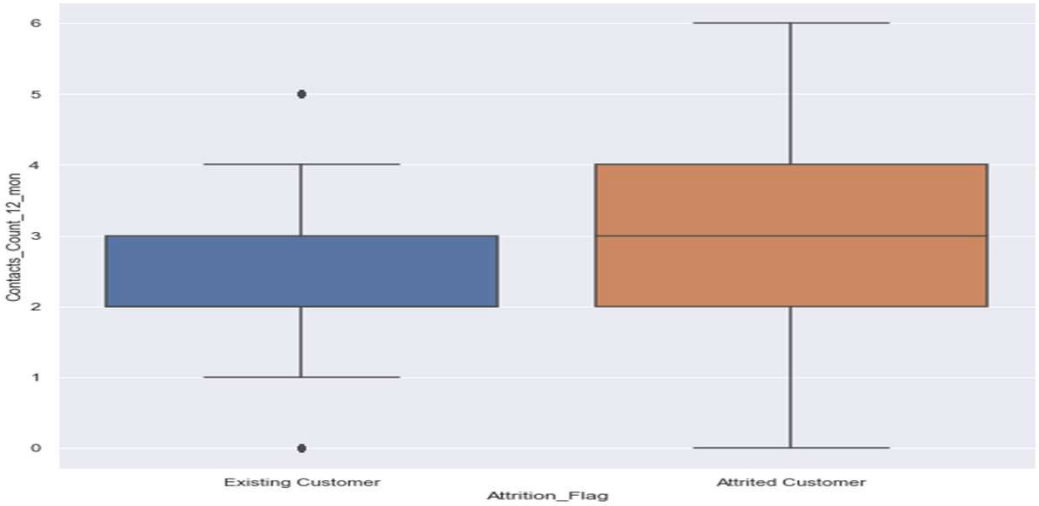


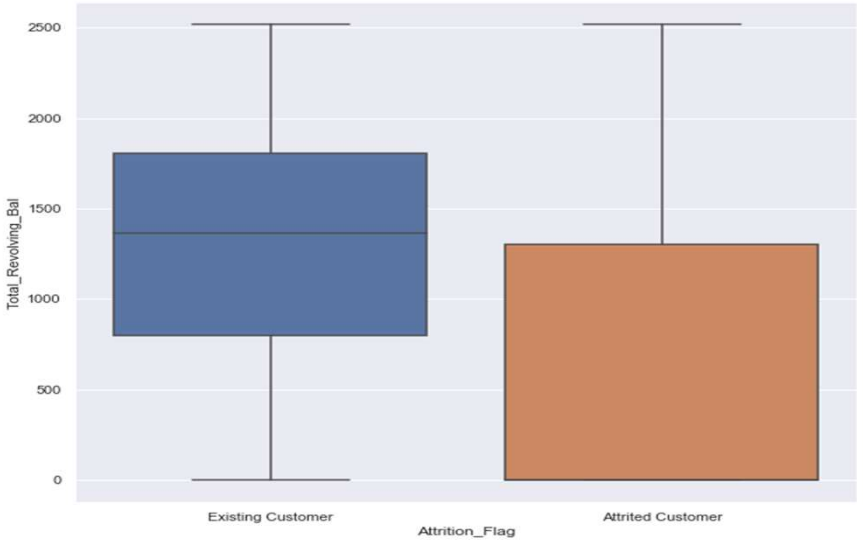
- Attrition has a small positive correlation with total revolving balance, total transaction count, transaction count ratio between Q4 and Q1, and number of products held by customer.
- Months on book and customer age are highly correlated.
- Total transaction amount and number of transactions are negatively correlated to the number of products held by customer.
- Credit limit is perfectly correlated to average open to buy.
- Utilisation ratio is negatively correlated to the credit limit.
- Avg. utilisation ratio is positively correlated to total revolving balance and negatively with average open to buy.
- Total transaction count and amount are positively correlated.

4. Customer Churn Key Factors (I/IV)



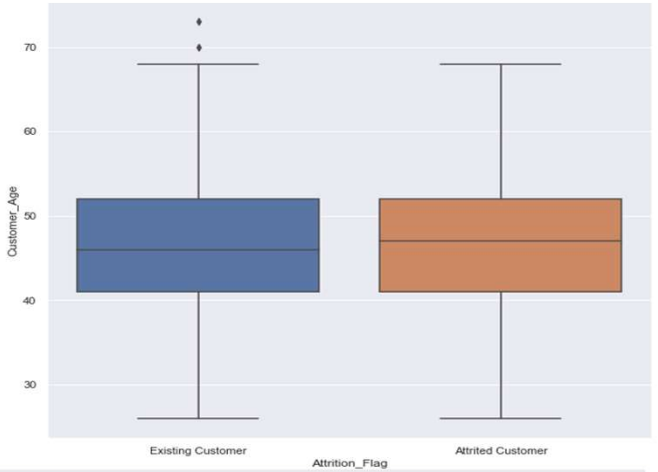
- Average number of products held by customers who are leaving are lesser than one who stay.
- More customer contacts are made by customers who are leaving.
- Balance carried over from one month to next is lower for customers who are leaving compared to the ones who are staying.

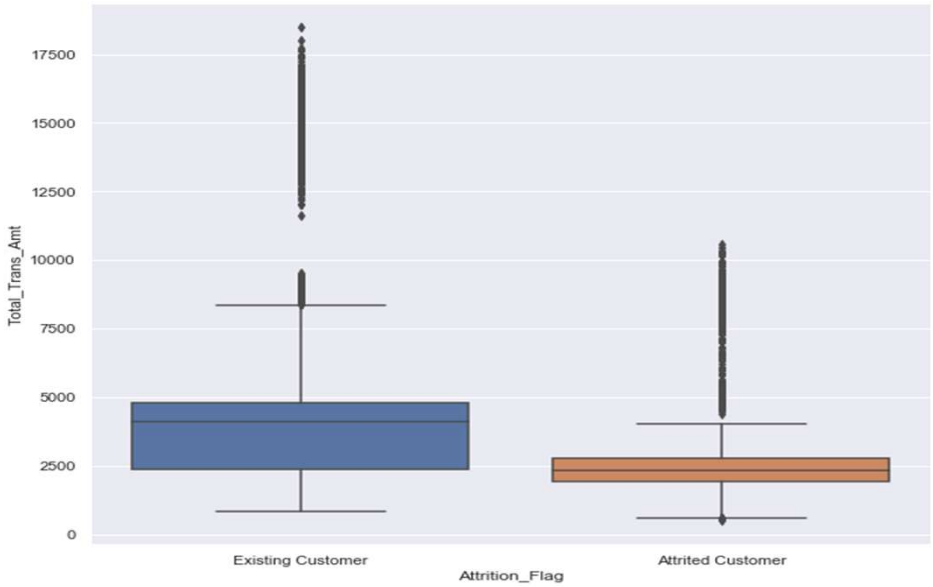


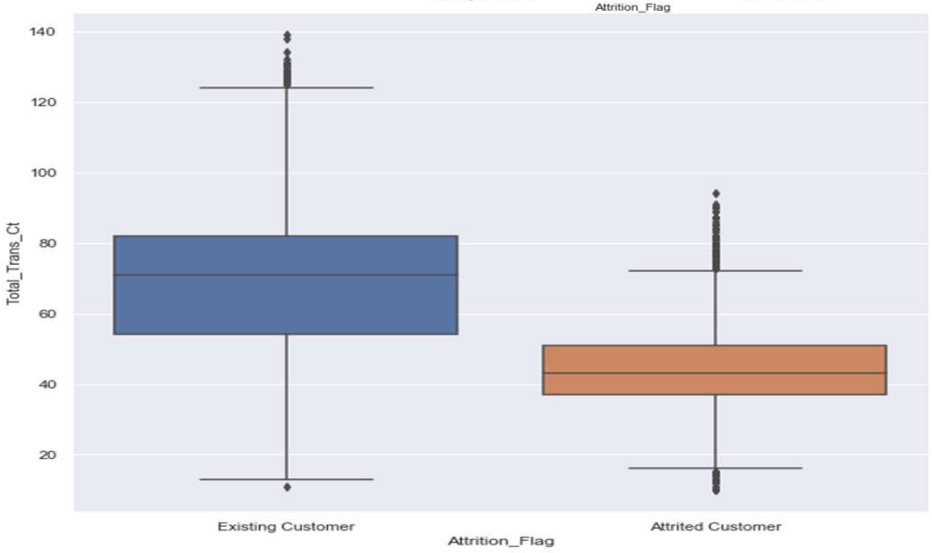


4. Customer Churn Key Factors (II/IV)

- Total transactions made and total amount of transaction are lower for customer who leave.
- Average age of customers who leave is higher than who stay.

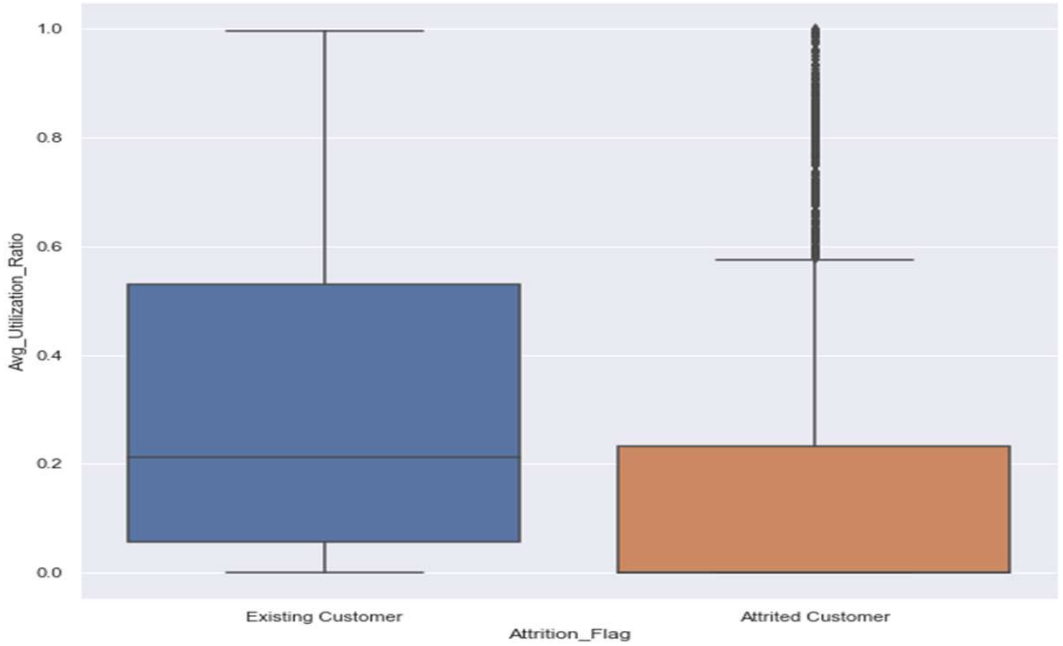


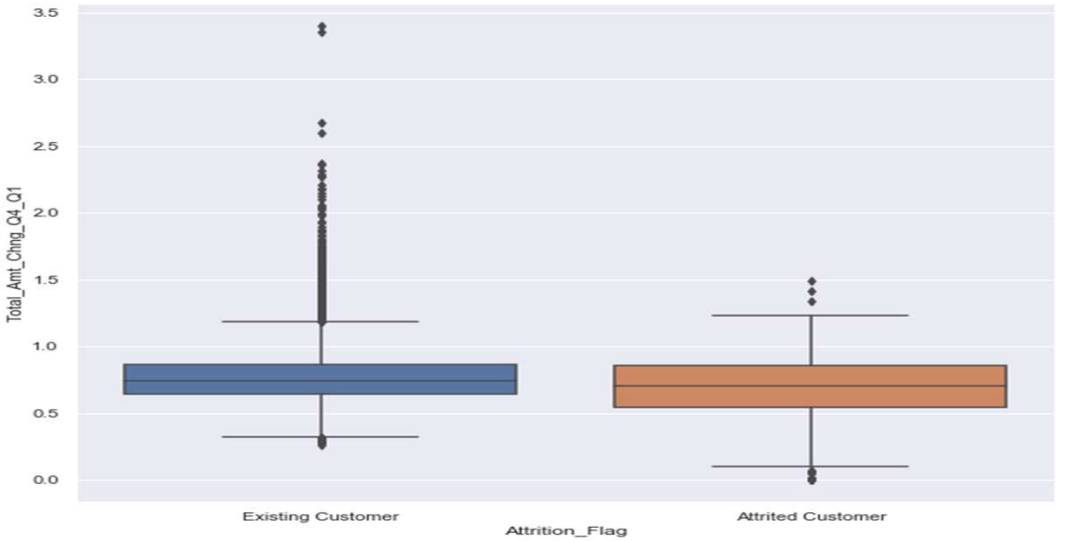


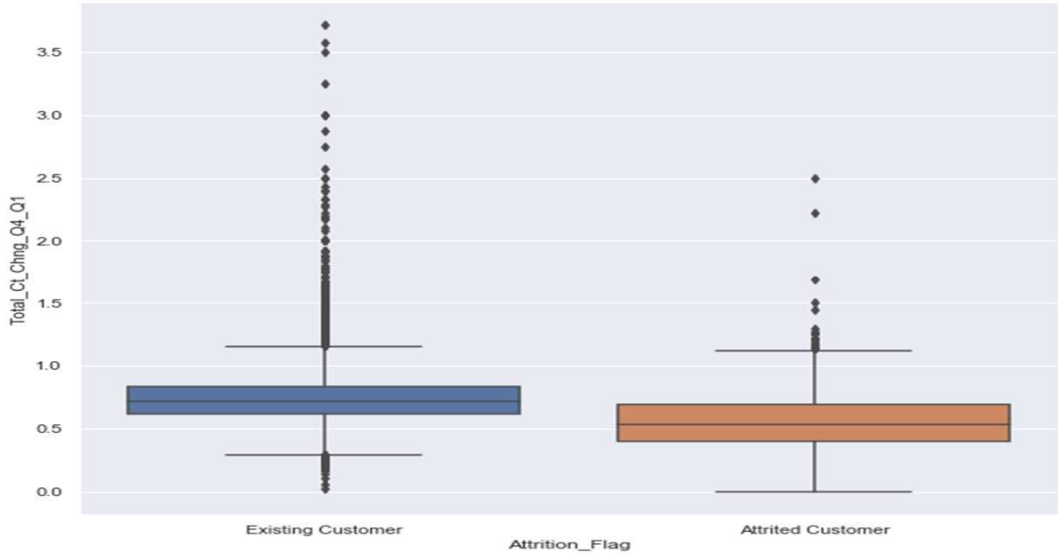


4. Customer Churn Key Factors (III/IV)

- The fall in transaction count and amount between Q1 and Q4 is higher for customers who leave.
- Average utilisation of the credit limit is lower for customers who are leaving.

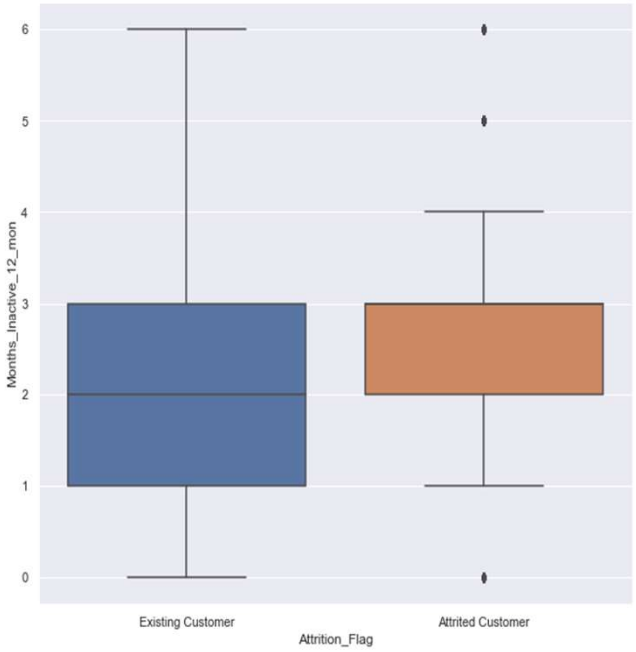


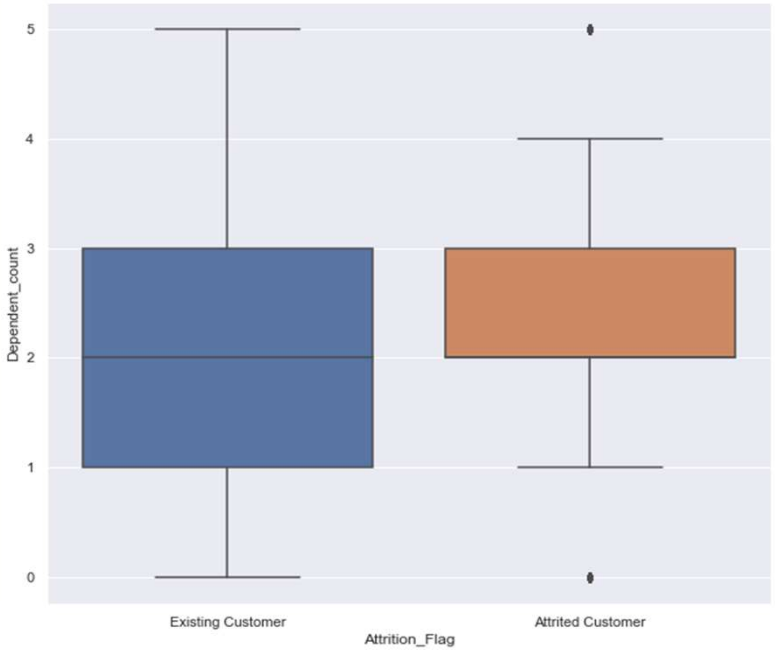


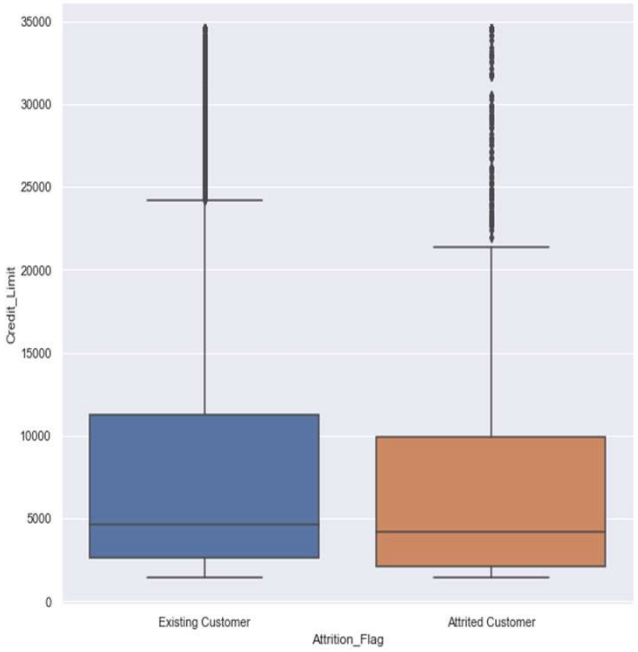


4. Customer Churn Key Factors (IV/IV)

- Higher months of inactive in the last 12 month is shown by customers who are leaving.
- Customers with more dependents leave than customers with lower number of dependents.
- Credit limit of customers who leave is slightly lower compared to customers who stay.

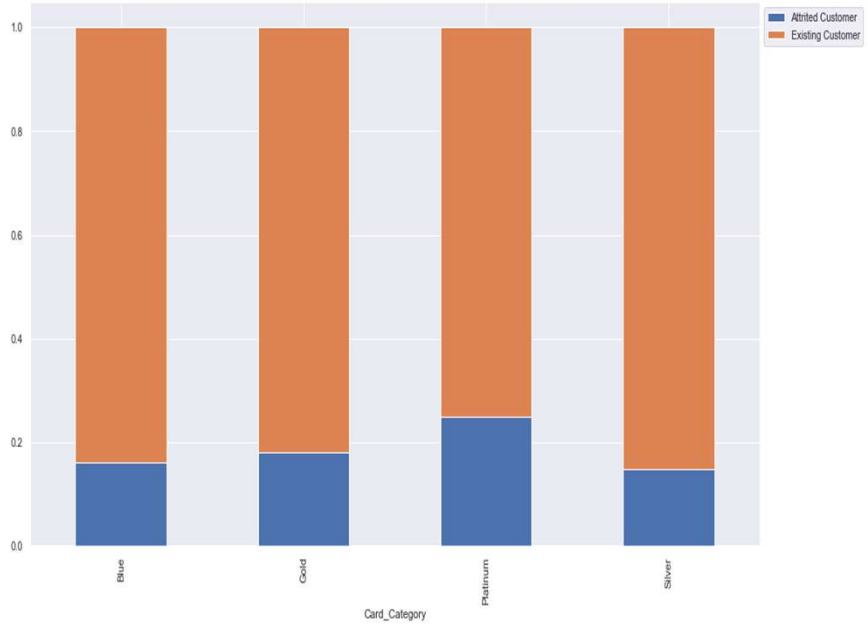


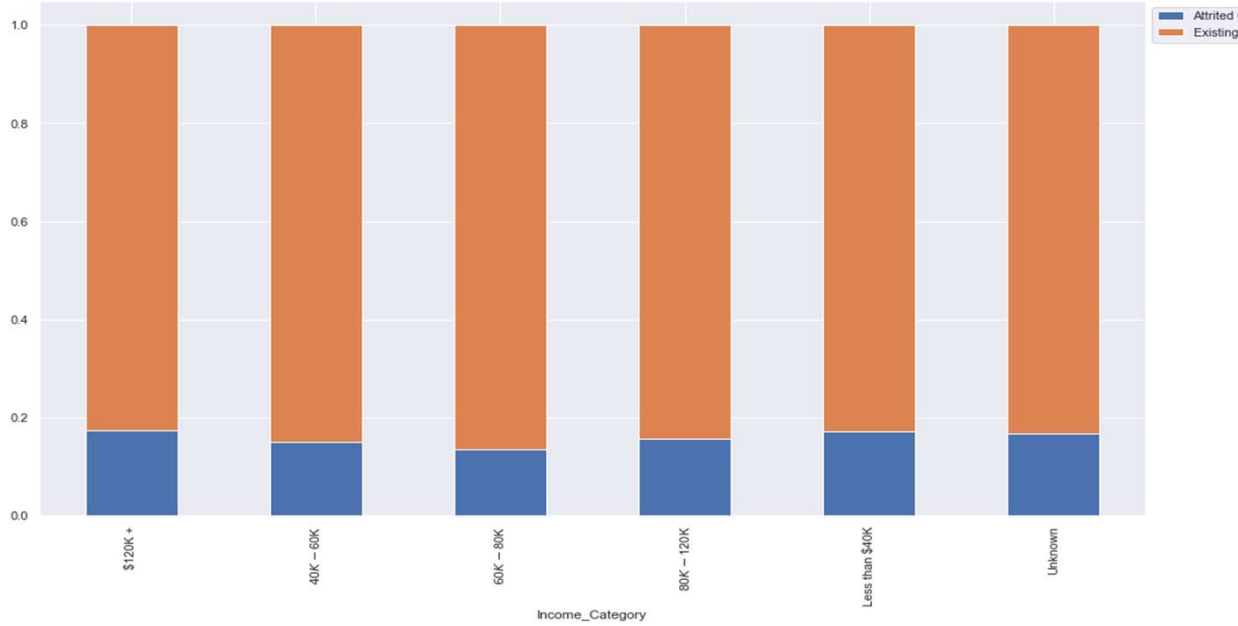




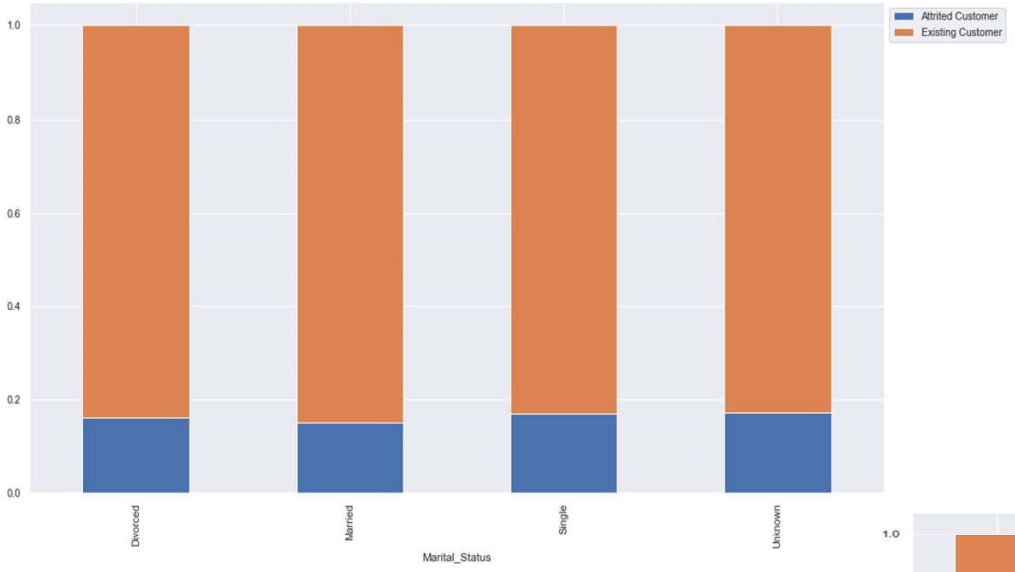
4. Customer Churn Non-Key Factors

- Attrition across customers within various income categories are similar.
- Attrition among platinum customers is slightly higher than customers with other type of cards.

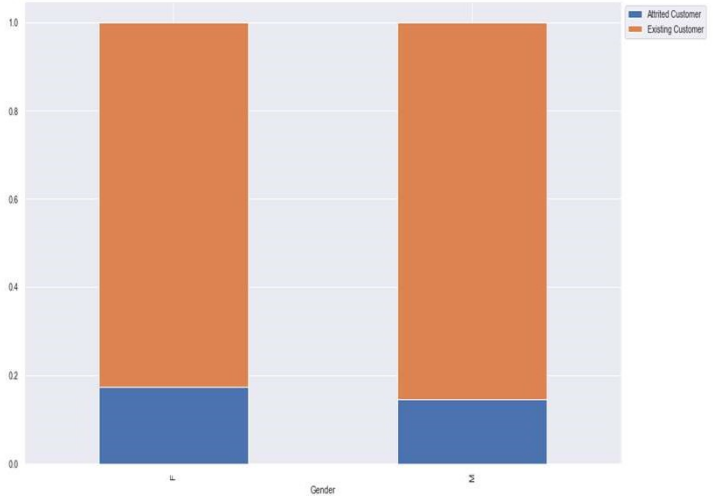


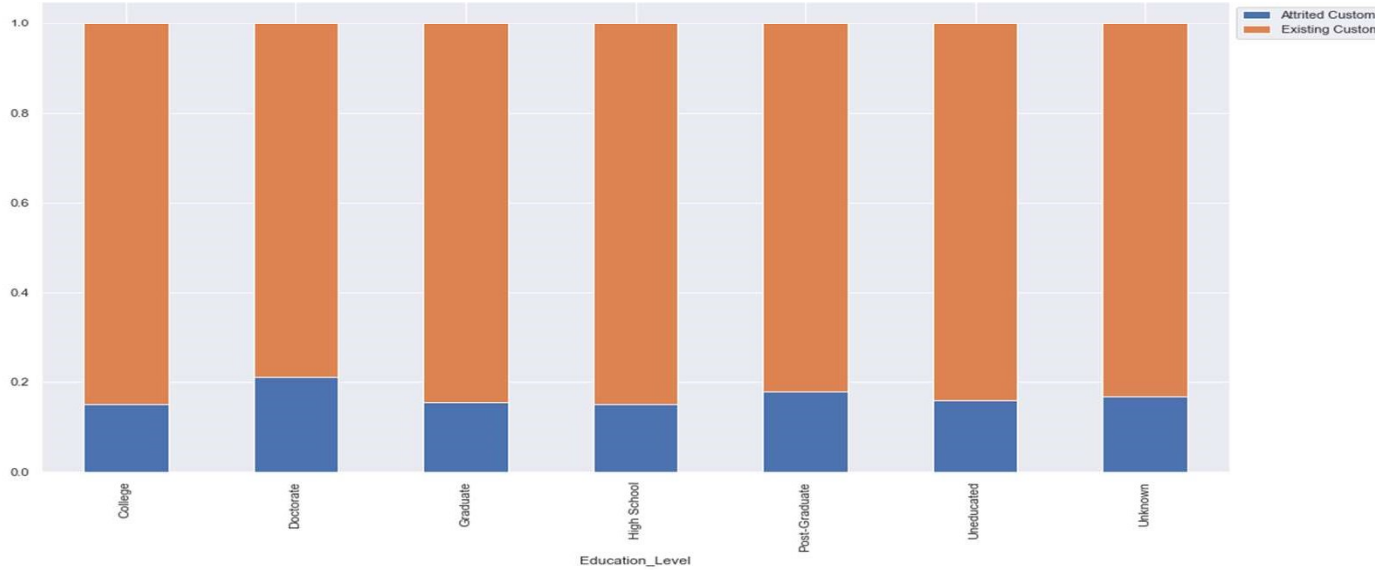


4. Customer Churn Non-Key Factors

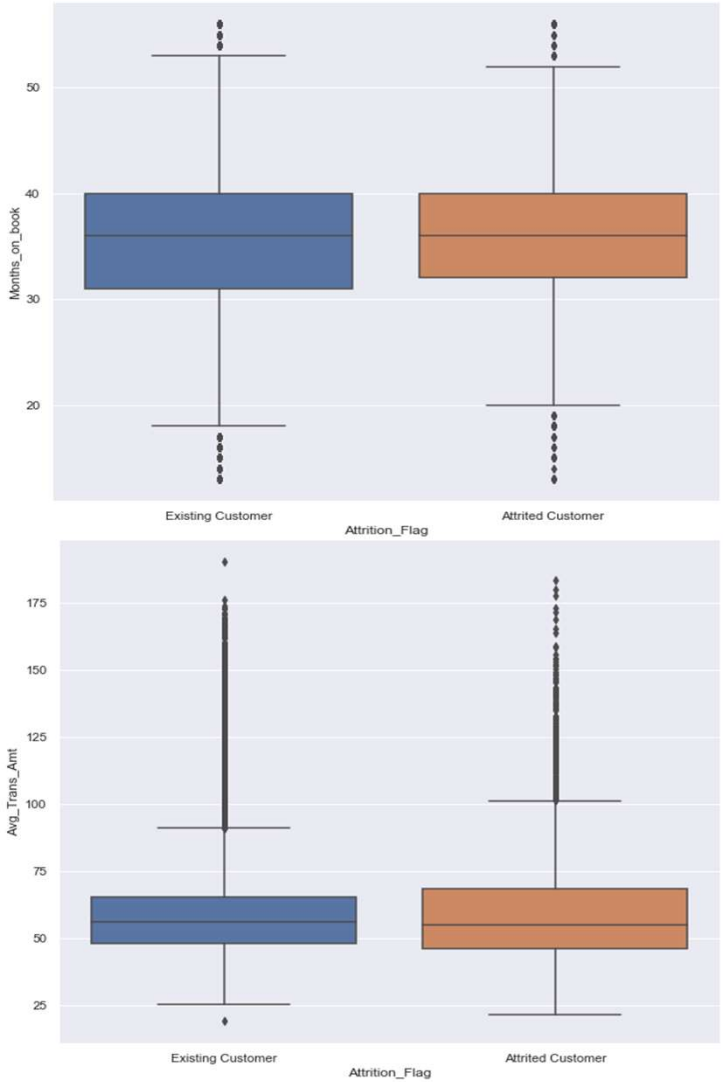


- Attrition across females, males and various educational levels are similar.
- Attrition is similar irrespective of the customer is married, single or divorced.

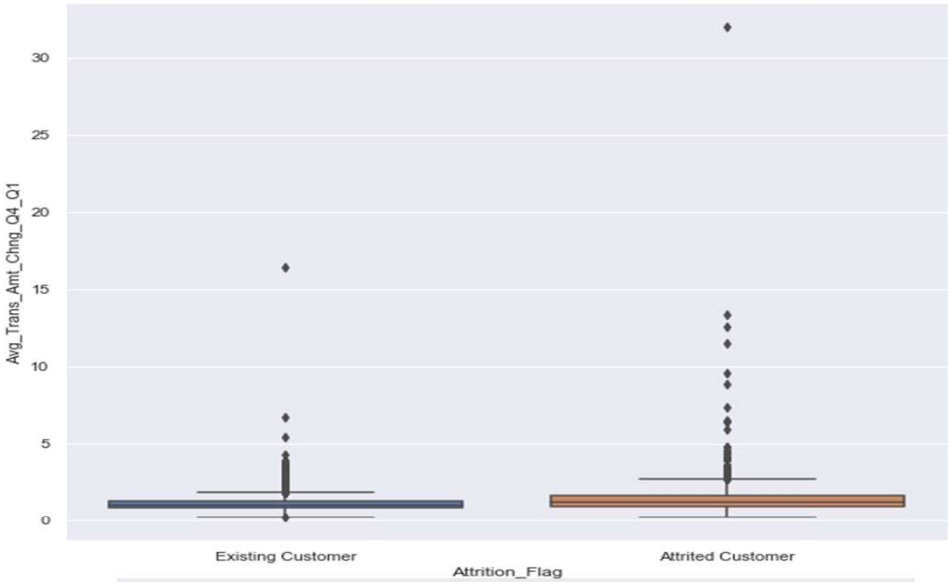


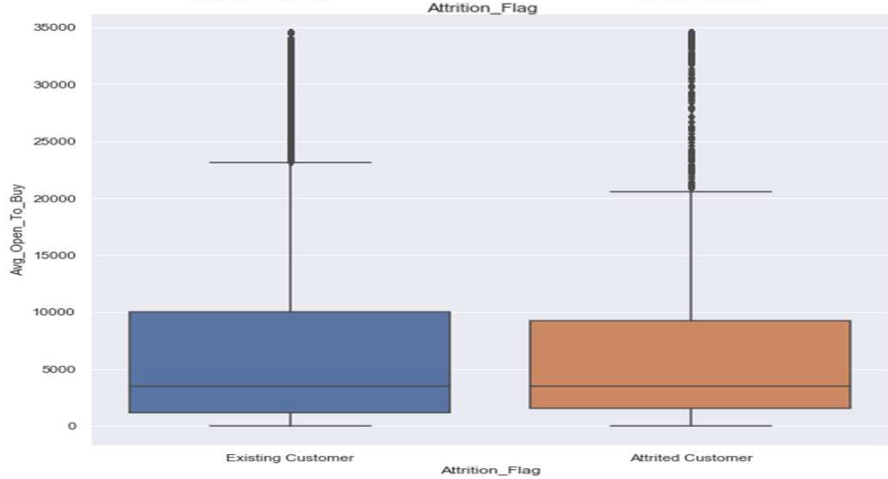


4. Customer Churn Non-Key Factors



Period of relationship with the bank, average transaction amount, ratio of average transaction amount between Q1 and Q4, and average open to buy is similar across customers who are leaving and staying.





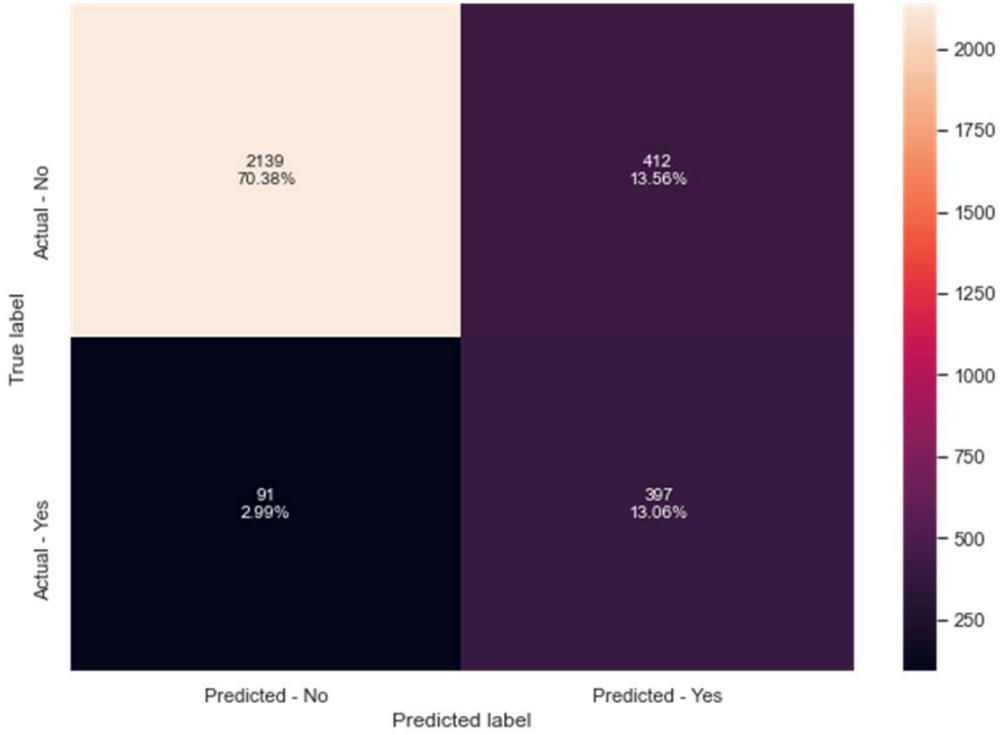
5.1 Model Building- Overview

- Logistic regression, Bagging methods and Boosting methods were used to build models to predict whether a customer is going to leave or not.
- Recall is the performance measure used to compare the performance of models in this scenario rather than accuracy.
  - ✓ Recall measures what proportion of the customers who leave that can be identified.
  - ✓ The cost of not being able to rightly identify a customer who might leave is high which is measured using recall.
  - ✓ Accuracy is less important as it only the proportion of customers which are correctly identified. We will need to keep this high as possible to reduce looking into customers who will not leave.
- Following method was used for the different modelling techniques:
  - ✓ Logistic regression was performed. Over and under sampling was performed to check whether the performance can be improved. Regularisation will also be used to reduce overfitting.
  - ✓ Following models will be created and the best three models will be selected. Performance of the models will be maximised using hyperparameter tuning.
    - Decision tree
    - Random forest
    - Bagging classifier
    - Xgboost
    - AdaBoost
    - Gradient boosting
  - ✓ Models will be selected based on highest recall and lower overfitting.
  - ✓ Random search will be used to get an approximate set of parameters for maximum performance. This will then used for specific search using Gridsearch to find the parameters which give the maximum recall.

5.2 Logistic regression

- **Logistic regression** with under sampling gave the highest recall with minimal overfitting.
- KNN imputation to fill the unknown values have not increased model performance.

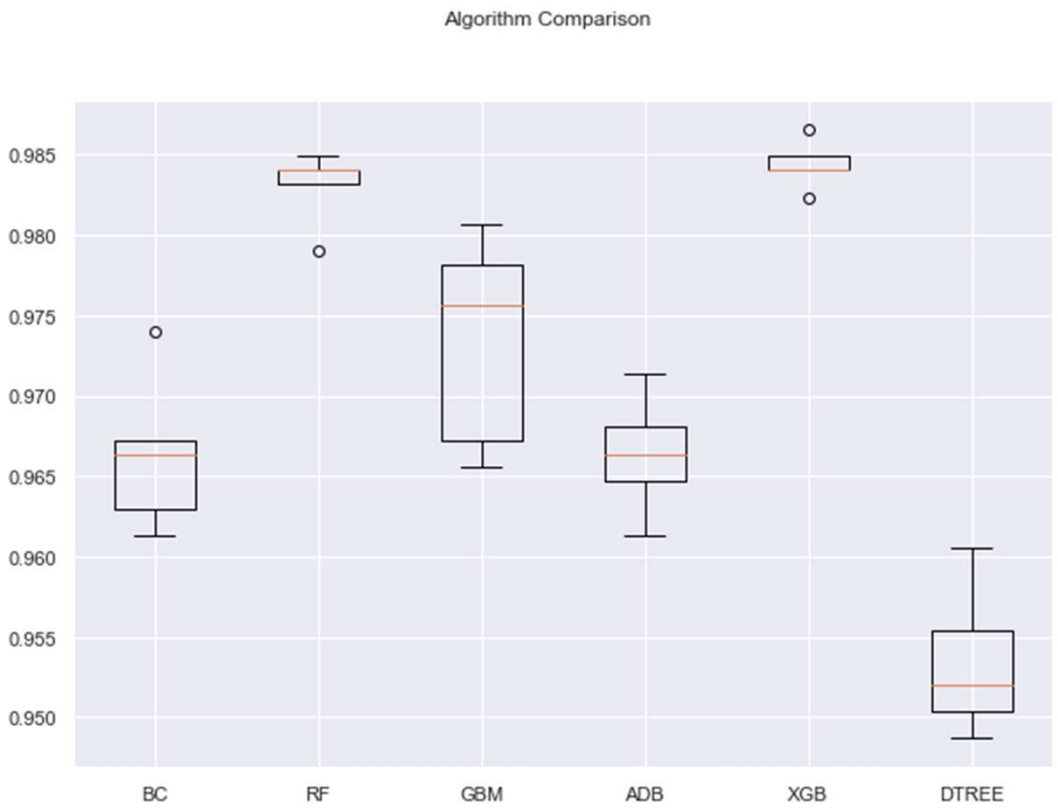
Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
LR- Base model	48.6	50.2	88.8	89.1
LR with oversampling	82.1	78.7	82.5	83.1
LR with oversampling with regularisation	82.1	78.7	82.5	83.1
LR with undersampling	82.6	81.4	81.5	83.4
LR with undersampling With regularisation	36.7	32.9	61.9	78.3
LR with undersampling And filling unknowns with KNN imputations	77.9	78.7	78.7	80.4



5.3 Model Selection for Hyper parameter Tuning

- Models in the table below were assessed to identify the best 3 models for hyperparamter tuning.
- K-fold technique was used to measure the recall and is shown on the right as box plot for each model.
- **XGBoost, RandomForest and Gradient boosting** provided best results using over sampling. Hence, these models were selected for further hyperparamter tuning.

Model	K-fold Recall (k=5)		
	No sampling	Over sampling	Under sampling
Decision Tree	77.8	95.3	86.2
Bagging Classifier	81.8	96.6	91.0
Random Forest	76.7	98.3	93.5
Adaptive boosting	83.9	96.6	93.5
Gradient Boosting models	84.1	97.3	94.8
XGBoost	88.8	98.4	94.9



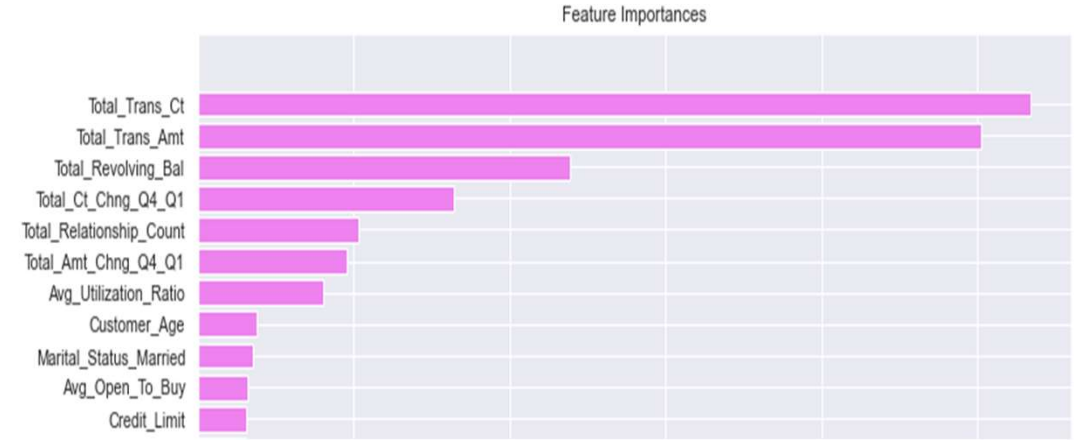
5.4 Random Forest

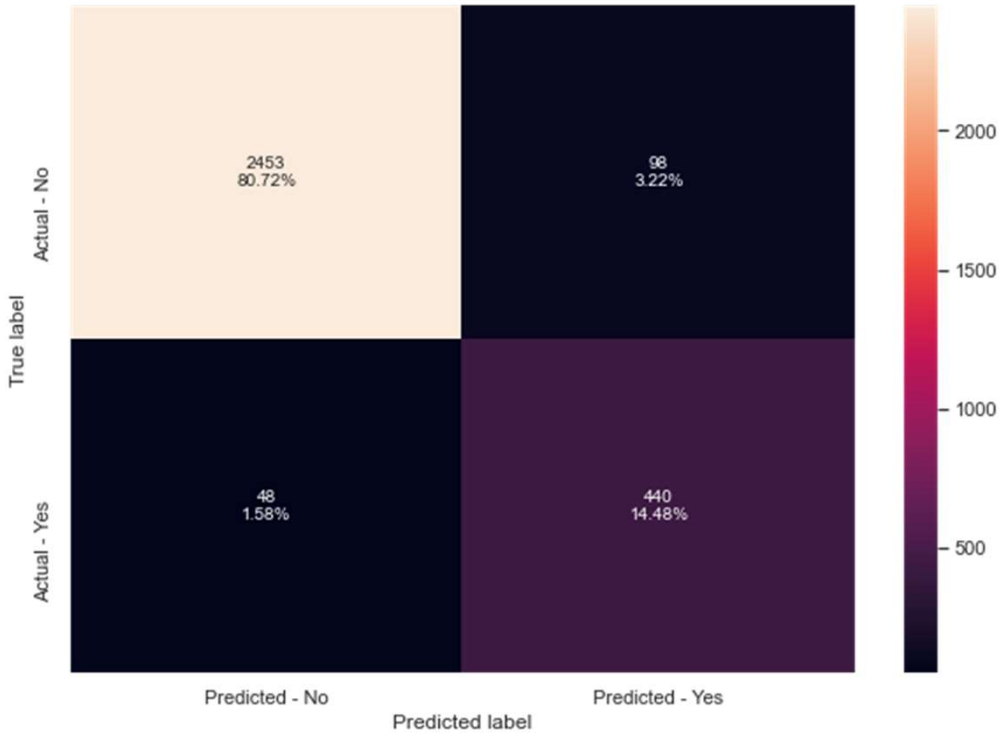
- RandomSearch was used to narrow down the search space for the best hyperparameters.
- The above result was used to do a grid search to obtain the best hyperparameters.
- This method allowed to optimise the time for search.

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
RF-RandomSearch	99.2	89.8	98.4	95.1
RF-GridSearch	99.3	90.2	98.5	95.2

Key Attributes

Following are key attributes by importance for the hypertuned model.





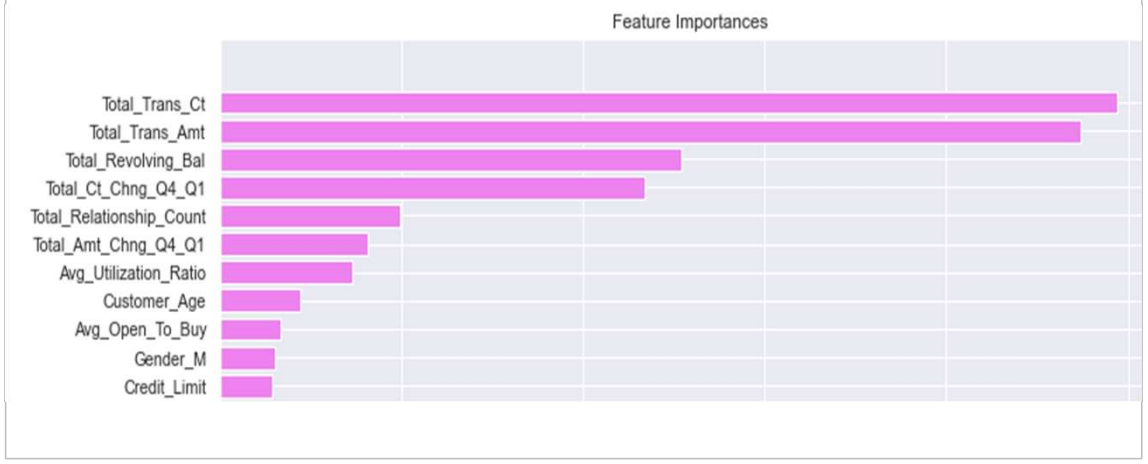
5.5 Gradient Boosting

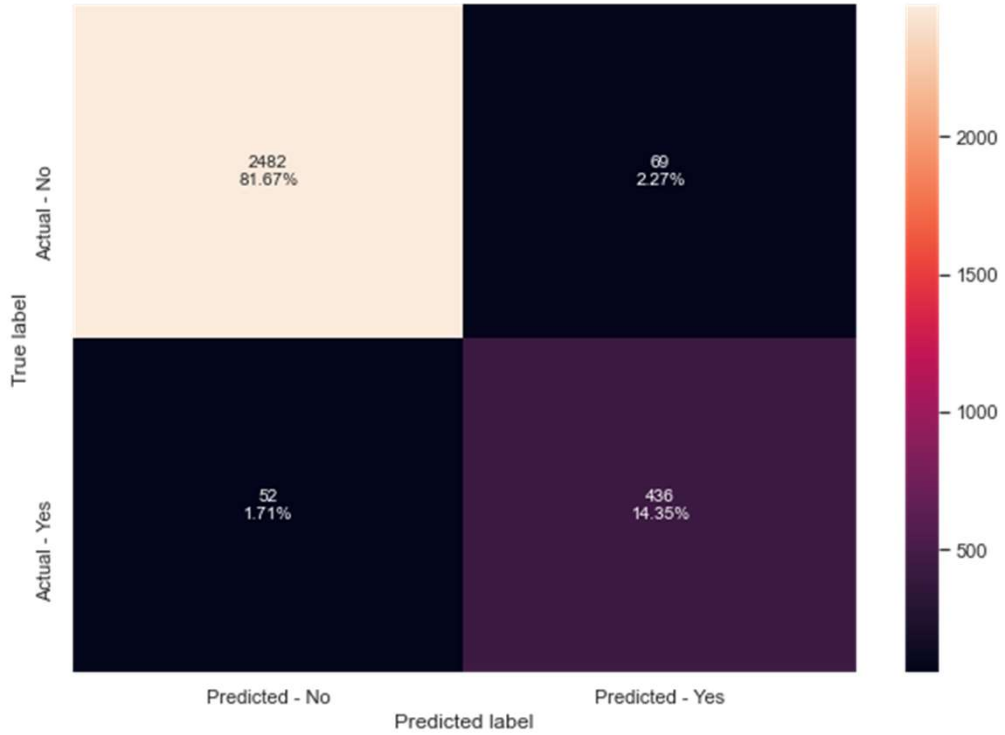
- RandomSearch was used to narrow down the search space for the best hyperparameters.
- The above result was used to do a grid search to obtain the best hyperparameters.
- This method allowed to optimise the time for search.

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
GB- RandomSearch	99.5	88.9	99.3	96.2
GB- GridSearch	99.6	89.3	99.3	96.0

Key Attributes

Following are key attributes by importance for the hypertuned model.





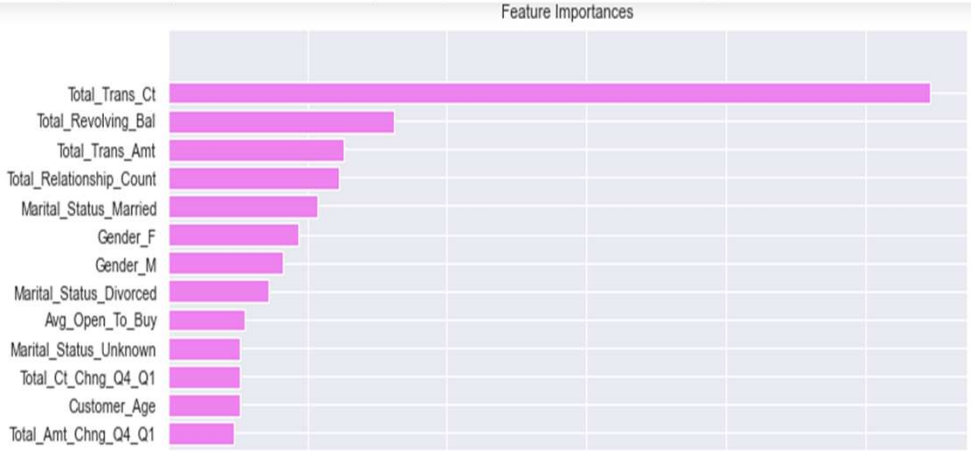
5.6 XGBoost

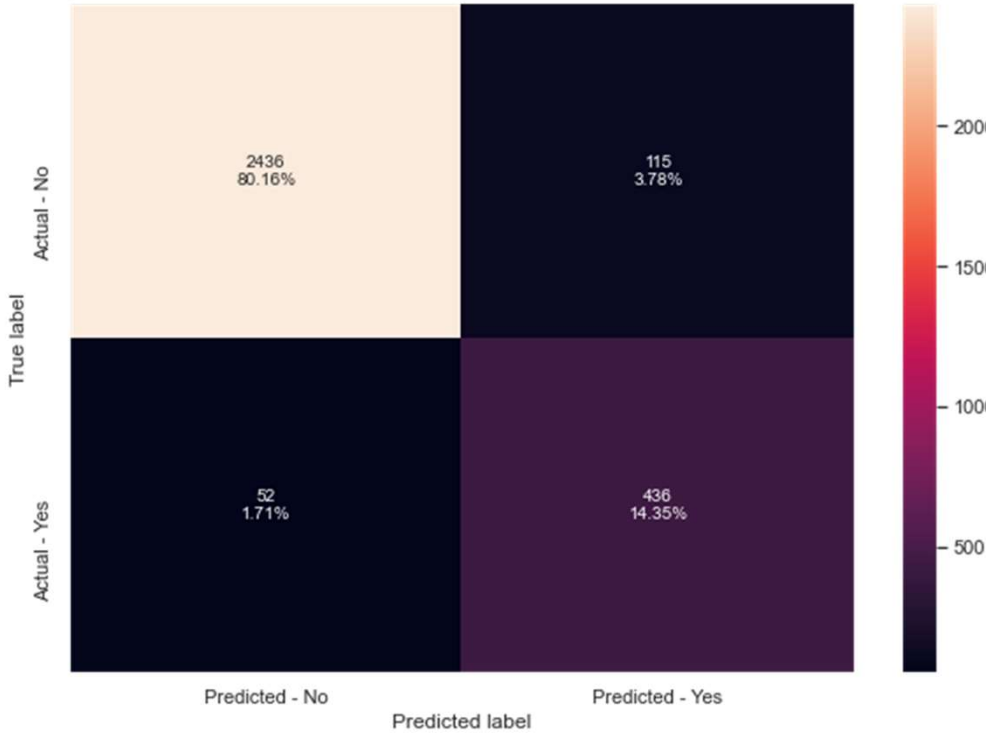
- RandomSearch was used to narrow down the search space for the best hyperparameters.
- The above result was used to do a grid search to obtain the best hyperparameters.
- This method allowed to optimise the time for search.

Model	Recall		Accuracy	
	Training Data	Test Data	Training Data	Test Data
XG-RandomSearch	98.4	89.3	97.6	94.4
XG-GridSearch	98.4	89.3	97.5	94.5

Key Attributes

Following are key attributes by importance for the hypertuned model.





5.7 Model- Selection Summary

- Below is the summary of all the models created.
- RandomForest with gridsearch gave the best test recall. Hence, this model is considered the best model.

Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
RandomForest with GridSearchCV	0.985208	0.951958	0.993108	0.901639	0.977660	0.817844
RandomForest with RandomizedSearchCV	0.984031	0.951300	0.991595	0.897541	0.976817	0.817164
Gradient Boost with GridSearchCV	0.993360	0.960184	0.996806	0.893443	0.989983	0.863366
XGBoost with RandomizedSearchCV	0.975710	0.943731	0.984367	0.893443	0.967614	0.785586
XGBoost with GridSearchCV	0.975458	0.945048	0.984199	0.893443	0.967289	0.791289
Gradient Boost with RandomizedSearchCV	0.993024	0.961500	0.995293	0.889344	0.990797	0.873239

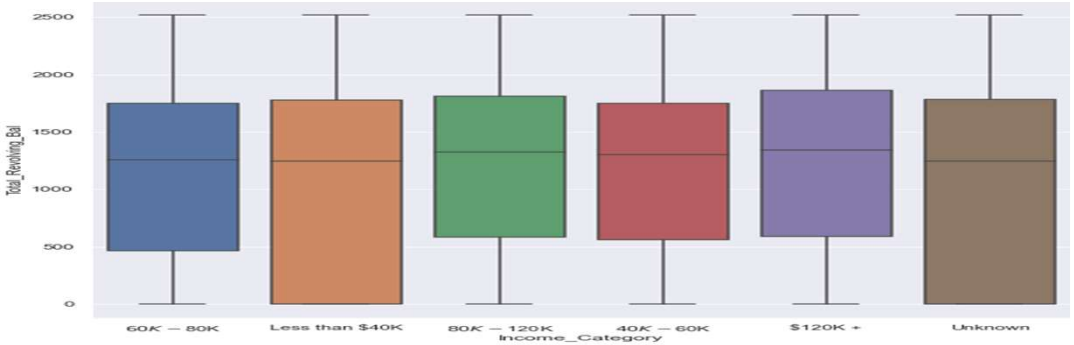
6. Key Insights

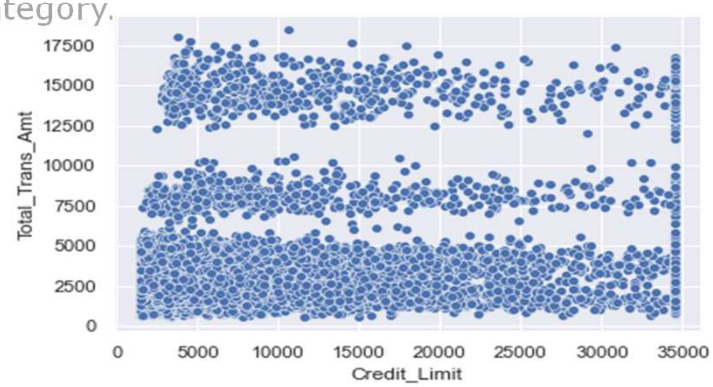
Customer Profile

1. Most of the customers are females
2. Most of the customers are from lower income category of less than $40k, and lower number of customers from higher income category. Similarly, most of the customers are graduates, with lower number of customers with higher qualification.
3. Low number of customers using higher end cards such as silver and platinum.
4. There are low number of customers who are more than three years with the bank.

Credit card finance

1. Credit card limit is low, with the high limit only for a small proportion of customers. This might be due to high number of customers being in the low income category.
2. Customers use only a small part of their credit card limit. As shown below, the total transaction amount does not go up as the credit card limit goes up.
3. The higher income category only spends similar amounts to the lower income category.





6. Key Insights

Key observations for leaving customers

Following customers have a higher chance of leaving the credit card service provided by the bank:

1. Inactive for 2 or more months
2. Lower revolving balance, transaction amount and number of transactions
3. 3 or more contacts were made in the last 12 months
4. Hold 3 or less products provided by the bank
5. Use 20% or lesser of their credit limit
6. Had a significant fall in transaction amount and transaction count in Q4 compared to Q1
7. Have 2 or more dependants

7. Recommendations

Generic

1. Acquire new customers who are from higher income category, and who are more educated by targeting professionals and business owners.
2. Provide incentive for long term customers including providing silver and platinum cards to retain customers beyond 3 years.
3. Increase credit card spend by:
    - reviewing and updating credit limit of lower income customers, if risk is low
    - providing incentives such as gifts and lower interest to make the high credit limit customers to increase their credit card spend
    - partnerships with retailers to provide discounts for credit card use.

Retaining credit card customers

1. Use the model to predict customers who might leave and perform following targeted actions.
    - Proactively reach out to customers who are inactive for more than a month and incentivise them to spend
    - Target discounts and promotions to low credit card spenders
    - Ensure the customers who have contacted more than twice is satisfied with the service and there is no underlying issues in credit card service provided to these customers.
    - Increase incentives for customers with 2 or more dependents by introducing discounts on family holidays, restaurants etc.
    - Increase number of bank products taken up by credit card customers which will increase strength of relationship with the bank.