Credit Card Users Churn Prediction

Ensemble Techniques and Model Tuning

Contents

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

- Attrition is highest among the customers who are using 1 or 2 products offered by the bank - together they constitute ~55% of the attrition - Bank should investigate here to find the problems customers are facing with these products, customer support, or more transparency can help in retaining customers.

- Female customers should be the target customers for any kind of marketing campaign as they are the ones who utilize their credits, make more and higher amount transactions. But their credit limit is less so increasing the credit limit for such customers can profit the bank.

- As inactivity increases the attrition also increases, 2-4 months of inactivity are the biggest contributors of attrition -Bank can send automated messages to engage customers, these messages can be about their monthly activity, new offers or services, etc.

- Highest attrition is among the customers who interacted the most with the bank, this indicates that the bank is not able to resolve the problems faced by customers leading to attrition - a feedback collection system can be set-up to check if the customers are satisfied with the resolution provided, if not, the bank should act upon it accordingly.

Business Problem Overview and Solution Approach

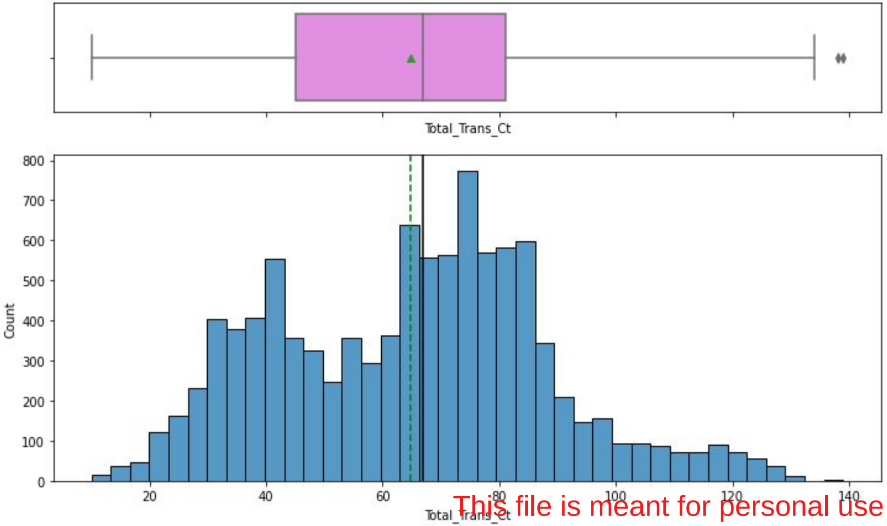
- The Thera bank recently saw a steep decline in the number of users of their credit card, credit cards are a good source of income for banks because of different kinds of fees charged by the banks like annual fees, balance transfer fees, and cash advance fees, late payment fees, foreign transaction fees, and others. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances.

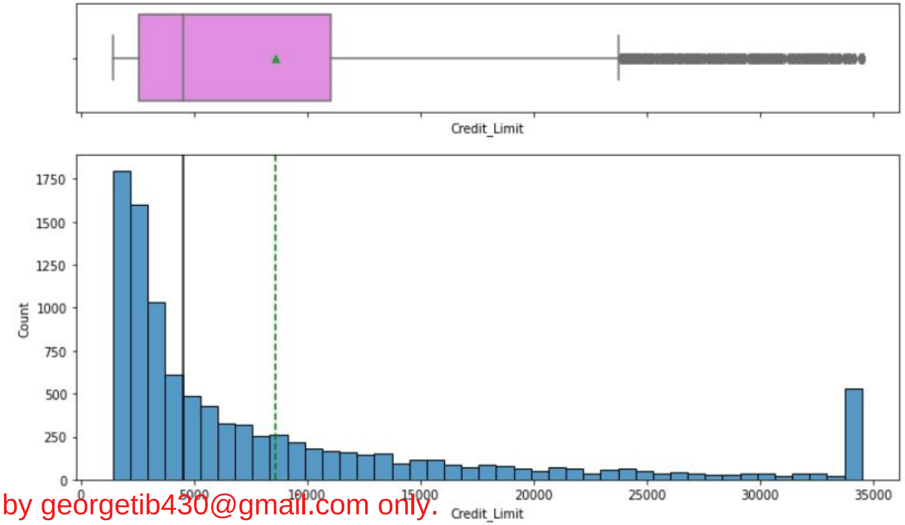
- Customers' leaving credit cards services would lead bank to loss, so the bank wants to analyze the data of customers and identify the customers who will leave their credit card services and reason for same – so that bank could improve upon those areas

- You as a Data scientist at Thera bank 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

EDA Results

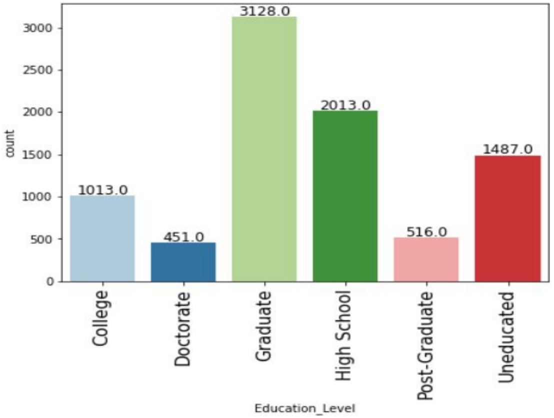
- The distribution of Total_Trans_Ct shows two peaks on 40 and 80 transactions in a year which indicates that customers used credit cards 3 to 6 times a month to make transactions.
- The distribution of the Credit_Limit is skewed to the right.
- There are quite a few customers with a maximum Credit Limit of 35000.
- 50% of the customers of the bank have a credit limit of less than <5000.



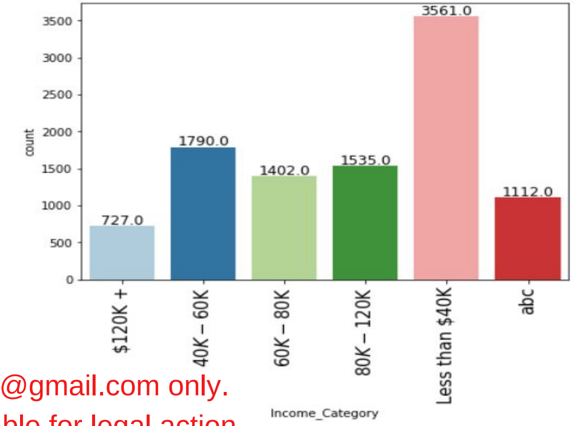


EDA Results

- 35.2% of the customers lie in the Less than 40k income category group, followed by 17.7% of the customers in the 40k-60k income group.
- Percentage of missing value in Income_Category column - 11%.

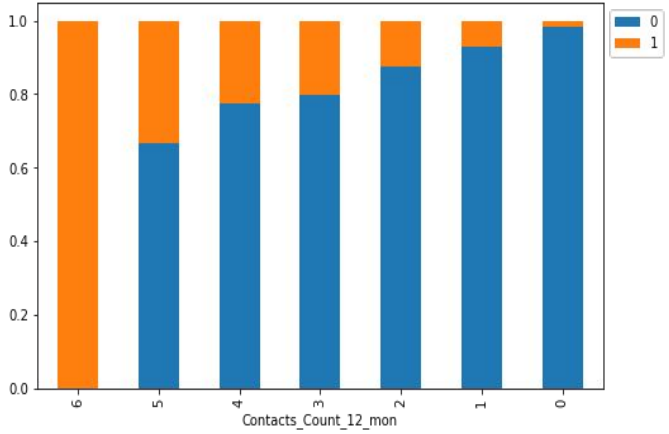


- 30.9% of the customers are graduates, followed by 19.9% of the customers who completed high school.
- Percentage of missing value in Education_Level column - 15%.

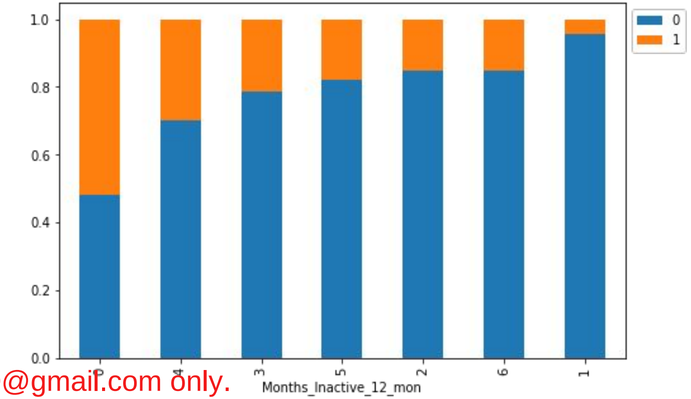


EDA Results

- As inactivity increases attrition also increases (2-4 months)
- The interpretation from here for 0 months and 6 months is difficult as customers who recently used the card attrited the most while those who were inactive for 6 months attrited less.

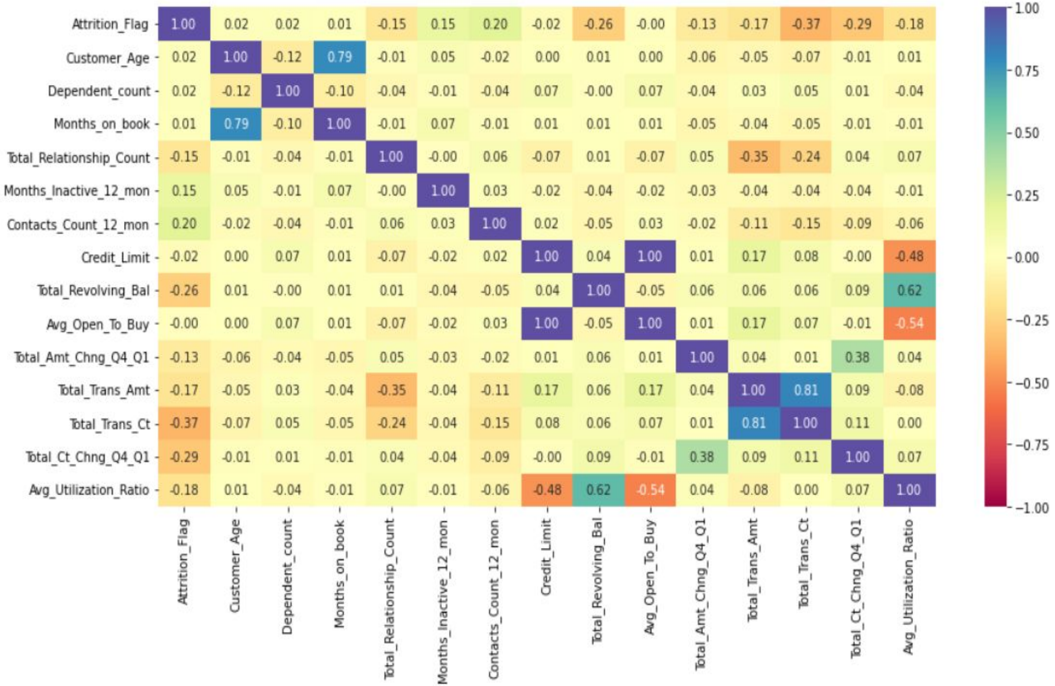


- The highest attrition is among the customers who interacted the most with the bank.
- This signifies that the bank is not able to resolve the problems faced by customers leading to attrition
- A preliminary step to identify attriting customers would be to look out for customers who have reached out to them repeatedly.



EDA Results

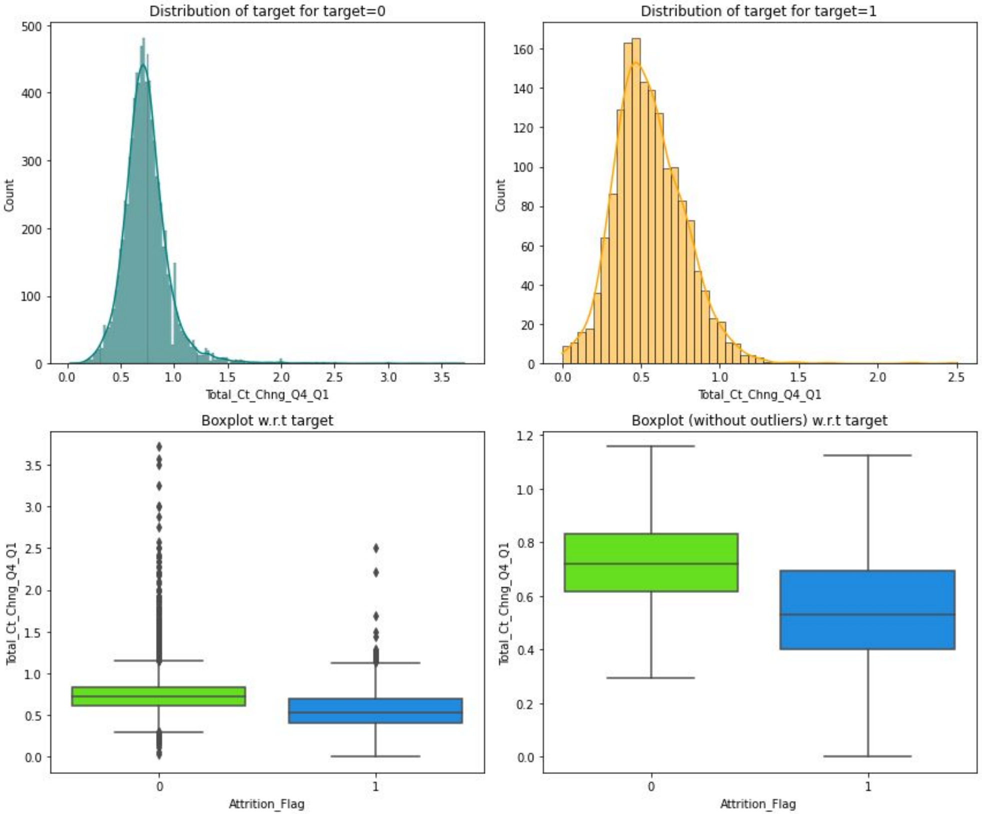
- The attrition flag shows a bit of a negative correlation with total transactions count and total transaction amount
- There's a strong positive correlation between months on book and customer age, total revolving_Bal and Avg utilization ratio, total trans amt and Total trans count
- There's a negative correlation of Total relationship count with the total trans amt and total trans count, avg utilization ratio with the credit limit and avg open to buy.



	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

EDA Results

- Customers who didn't attrite showed less variability across Q4 to Q1 as compared to the ones who attrited.



Data Preprocessing

- The data shared is a ciphered version containing 10127 observations.

- The characteristics include client number, age, gender, count of dependents, education level, marital status, income category, card category, period of relationship with the bank, number of products held, inactive months and number of contacts in past 12 months, credit limit, total revolving balance, average open to buy,change in transaction amount, total transaction amount (Last 12 months), Total Transaction Count (Last 12 months), change in transaction count(Q4 over Q1), avg utilization ratio

- There were few missing values in 2 columns and 1 column have ‘abc’ values- we’ll treated them as missing values. We imputed them using the most_frequent values and to avoid data leakage we imputed missing values after splitting train data into train and validation sets.

- ~84% of the customers not attrited and only ~16% customers are attrited. The dataset is highly imbalanced so we tried undersampling and oversampling techniques to balance the data.

Comparing Model Performance

Training Performance

Metrics	XGBoost trained with Original data	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Original data
Accuracy	0.946	0.996	0.962	0.992
Recall	0.999	0.999	0.999	0.967
Precision	0.749	0.994	0.809	0.982
F1-Score	0.856	0.996	0.894	0.975

Comparing Model Performance

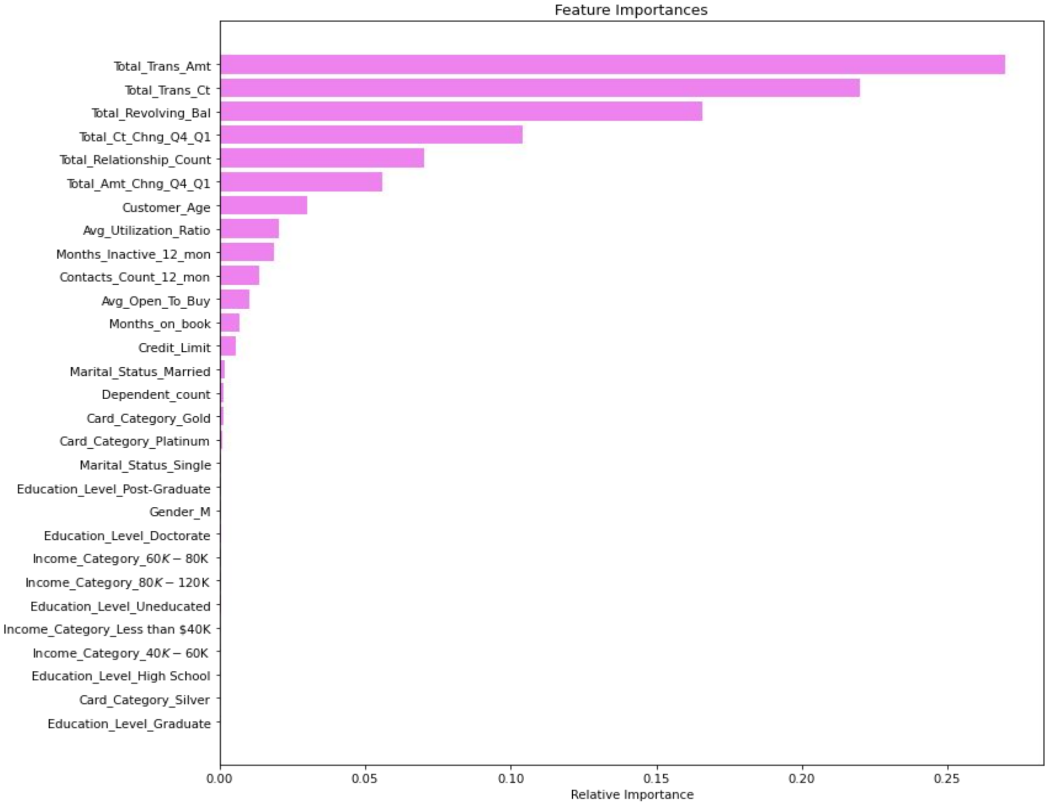
Validation Performance

Metrics	XGBoost trained with Original data	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Original data
Accuracy	0.930	0.944	0.944	0.969
Recall	0.960	0.957	0.957	0.871
Precision	0.710	0.759	0.759	0.934
F1-Score	0.816	0.847	0.847	0.902

- Gradient boosting model trained with original data has generalised performance, so let's consider it as the best model.

Feature Importance

- Total transaction amount is the most important variable in predicting credit card churn followed by total revolving balance, total transaction count (Last 12 months), total no. of products held by the customer and change in transaction Amount (Q4 over Q1)



APPENDIX

Model Performance Summary (original data)

Metrics	Bagging	Random Forest	Gradient Boosting	Adaboost	XGBoost	Decision Tree
Training	0.985	1.0	0.875	0.826	1.0	1.0
Validation	0.812	0.797	0.855	0.852	0.883	0.815

- Xgboost has the best performance on the validation followed by GBM and Adaboost

Model Performance Summary (oversampled data)

Metrics	Bagging	Random Forest	Gradient Boosting	Adaboost	XGBoost	Decision Tree
Training	0.997	1.0	0.980	0.969	1.0	1.0
Validation	0.849	0.868	0.892	0.901	0.898	0.825

- Adaboost has the best performance on validation followed by XGB

Model Performance Summary (undersampled data)

Metrics	Bagging	Random Forest	Gradient Boosting	Adaboost	XGBoost	Decision Tree
Training	0.990	1.0	0.980	0.951	1.0	1.0
Validation	0.920	0.938	0.957	0.960	0.957	0.920

- Adaboost has the best performance followed by Xgboost as per the validation performance

Great Learning

Happy Learning !