Thera Bank

Credit Card Users Churn Prediction

August 2021

Brian Hall – DSBA – Project 6

August 2021

Problem

The Thera Bank recently saw a steep decline in the number of users of their credit card, which is a good source of income due to the different kinds of fees charged by the bank. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances. Customers leaving credit card services would lead the bank to a loss, so Thera Bank would like to predict and understand customers that leave so it can improve upon those areas.

Objective

By analyzing customer data provided by Thera Bank we will develop a classification model to identify and understand customers that are likely to churn from credit card services and make recommendations on ways to retain those customers.

Recommendations

- Account activity seems to play a critical role in identifying customers who are at risk of Attiring, finding ways to incite customers to use their card more often could reduce churn.
- Customers who have higher contact rates are more likely to attire, it is possible that the bank is identifying at risk customers earlier but are not successful in their outreach. Consider revising outreach protocols once churn customers are identified.
- The Silver card category has the lowest attrition rate, consider offering low activity Blue card customers an upgrade with incentives to use, such as rewards or waiving certain fees.
- There is no attrition for customers with total transaction amounts > $11,000 and few between $2700 and $5200, consider researching these customers further to determine the reason.
- Total_Revolving_Bal & Total_Trans_Ct are the top 2 important features in the selected model. Nearly all customers with a revolving balance < $500 attired, most of those have a Total transaction amount of < $3000
- Providing additional data such as purchases, transaction types, type of other bank products, customer location, etc. could lead to better understanding why customers are leaving and developing strategies to target specific types of customers.

EDA Overview

10,127 Observations and 21 Features

There are 2 variables with missing values
- Education_Level - 1519
- Marital_Status - 749

There are no Duplicate Rows

Target variable Attrition_Flag is imbalanced with 'Attrited Customer' representing 16% of the data

All missing values were imputed; no observations deleted

Education_Level & Marital_Status missing values were imputed to Unknown

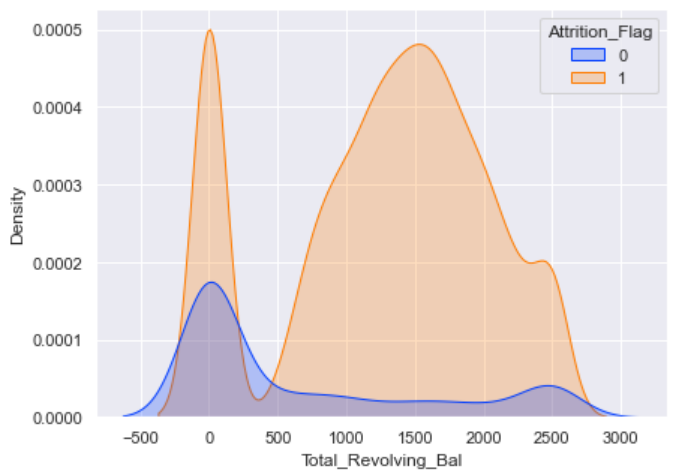
Income_Category has a category of 'abc' and accounts for 11% of the data. No patterns were found to determine where this category belongs. It was not changed.

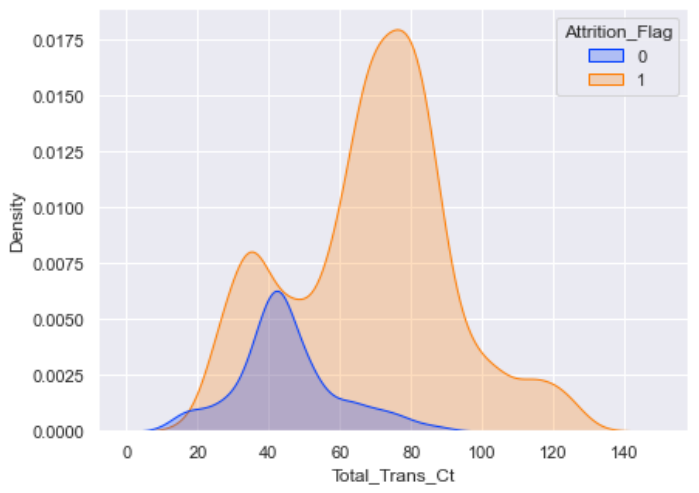
	count	mean	std	min	25%	50%	75%	max
Customer_Age	10127.000	46.326	8.017	26.000	41.000	46.000	52.000	73.000
Dependent_count	10127.000	2.346	1.299	0.000	1.000	2.000	3.000	5.000
Months_on_book	10127.000	35.928	7.986	13.000	31.000	36.000	40.000	56.000
Total_Relationship_Count	10127.000	3.813	1.554	1.000	3.000	4.000	5.000	6.000
Months_Inactive_12_mon	10127.000	2.341	1.011	0.000	2.000	2.000	3.000	6.000
Contacts_Count_12_mon	10127.000	2.455	1.106	0.000	2.000	2.000	3.000	6.000
Credit_Limit	10127.000	8631.954	9088.777	1438.300	2555.000	4549.000	11067.500	34516.000
Total_Revolving_Bal	10127.000	1162.814	814.987	0.000	359.000	1276.000	1784.000	2517.000
Avg_Open_To_Buy	10127.000	7469.140	9090.685	3.000	1324.500	3474.000	9859.000	34516.000
Total_Amt_Chng_Q4_Q1	10127.000	0.760	0.219	0.000	0.631	0.736	0.859	3.397
Total_Trans_Amt	10127.000	4404.086	3397.129	510.000	2155.500	3899.000	4741.000	18484.000
Total_Trans_Ct	10127.000	64.859	23.473	10.000	45.000	67.000	81.000	139.000
Total_Ct_Chng_Q4_Q1	10127.000	0.712	0.238	0.000	0.582	0.702	0.818	3.714
Avg_Utilization_Ratio	10127.000	0.275	0.276	0.000	0.023	0.176	0.503	0.999

	count	unique	top	freq
Attrition_Flag	10127	2	Existing Customer	8500
Gender	10127	2	F	5358
Education_Level	8608	6	Graduate	3128
Marital_Status	9378	3	Married	4687
Income_Category	10127	6	Less than $40K	3561
Card_Category	10127	4	Blue	9436

EDA Overview

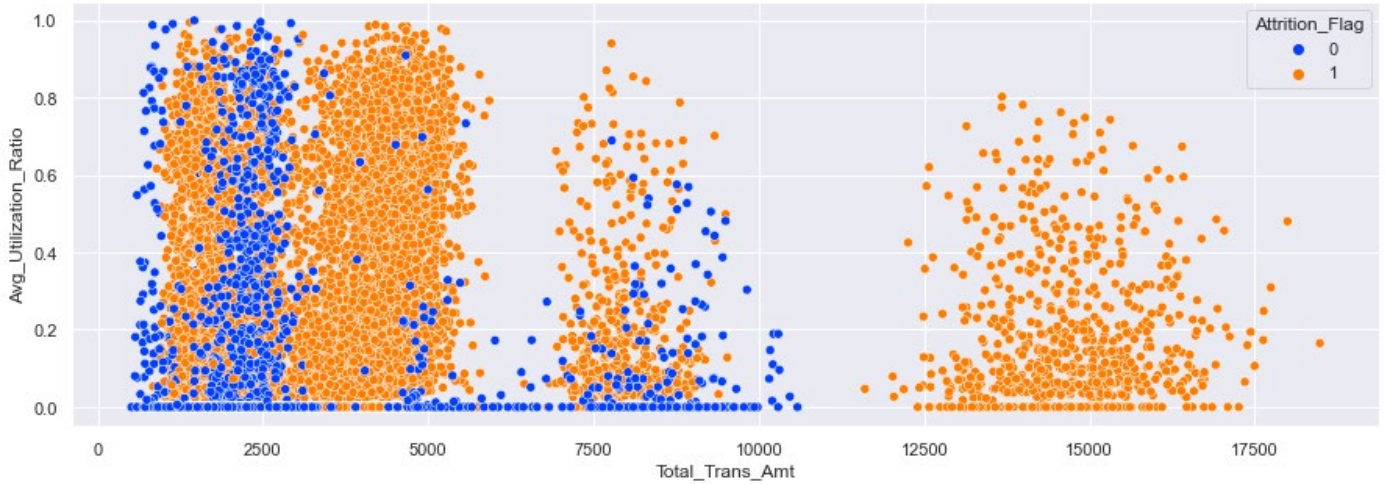
Total_Revolving_Bal & **Total_Trans_Ct** are the top 2 important features in the selected model



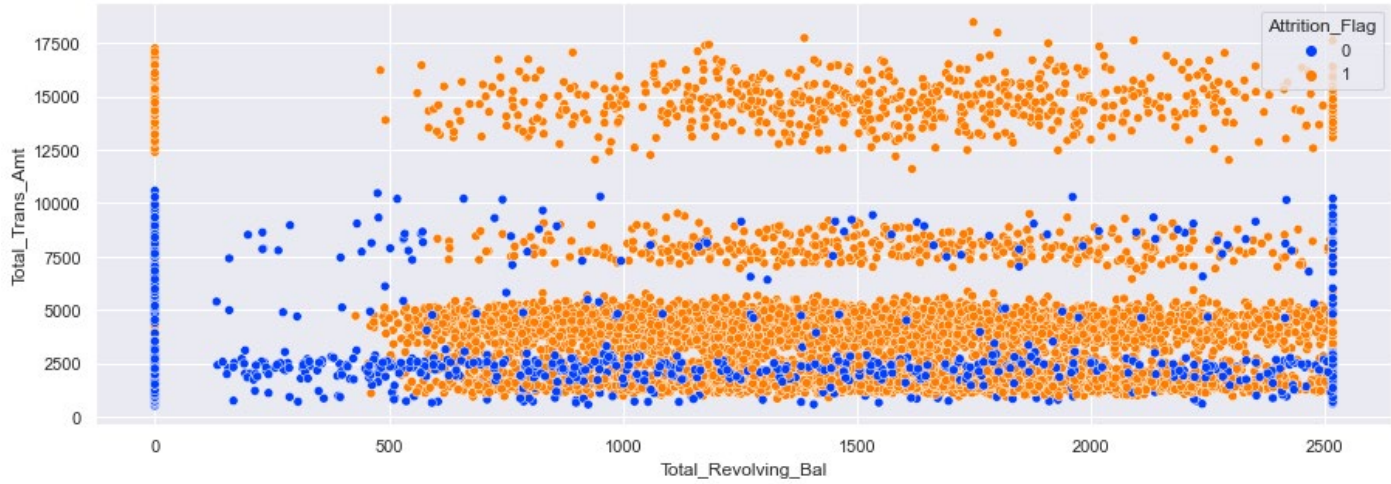


Key Observations

- No attrition for customers with total trans amounts > $11,000 and few between $2700 and $5200
- It is interesting that there are a number of customers with utilization ratios of 0 who have high total transaction amounts. Perhaps total transactions include other types of products

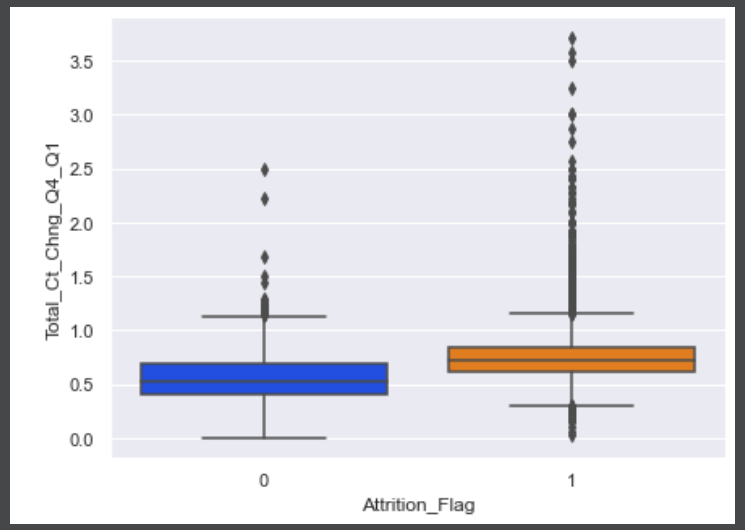


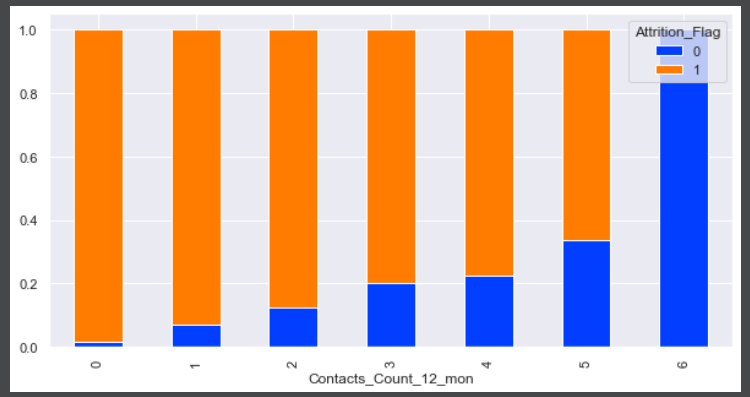
- Nearly all customers with a revolving balance < $500 attired, most of those have a Total transaction amount of < $3000
- Nearly all customers with a 0 revolving balance attired



EDA Overview

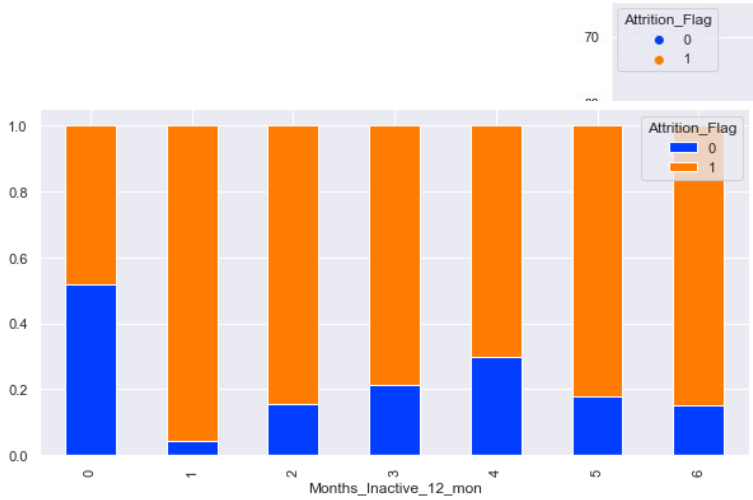
Total_Ct_Chng_Q4_Q1 & **Contacts_Count_12_mon** are the 3nd and 4th most important features in the selected model

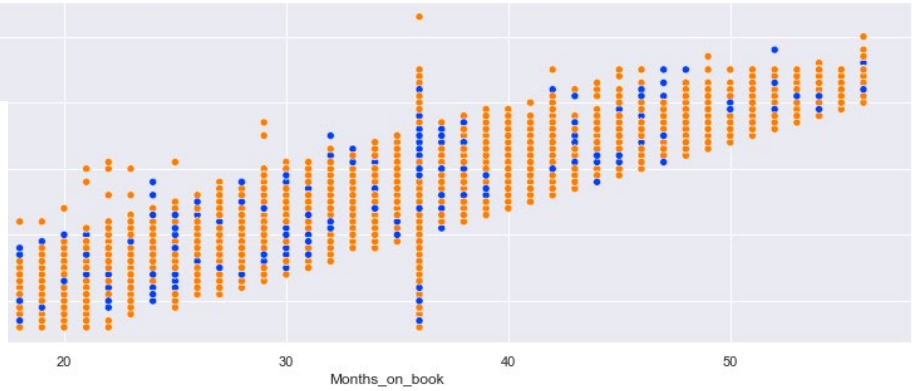




Key Observations

- Surprisingly the more contact the customer had with the bank the more likely they were to attire. Perhaps this reflects the bank increasingly reaching out to customers as time passed.
- All customers, 54, who had 6 contacts are flagged as attrited, with 5 contacts having ~ 33% attrition rate
- The majority of attired customers had a total transaction amount < $3000 and a credit limit < $10000
- There are no attrited customers with more than 100 transactions in the last 12 months
- Customers with an average utilization ratio > 0.25 are far less likely to attire
- Those with 0 inactive months have an attrition rate of ~ 51%, however there are only 30 of these customers
- A surprising number of customers with the max credit limit attired
- For months on book, there is very high spike, nearly 25% of customers, at 36 months (3 years).
- Credit_Limit has many outliers above $25,000 with a spike at the maximum $34,516 (5%). $34516 seems odd as a cap.
- The vast majority of customers have a "Blue" Card at 93%. Platinum card accounts for only 0.2%
- Silver has the lowest rate of attired customers with ~ 14%
- Surprisingly customers with platinum cards have a higher percentage chance of attiring than others. However, there are only 20 total platinum cards.
- Customers with average transactions between $0 - $4000 are more likely to attire
- The highest rate of attrition is for customers with about 38 to 54 transactions in the last 12 months
- It is interesting that there are a number of customers with utilization ratios of 0 who have high total transaction amounts. Perhaps total transactions include other types of products.
- A surprising number of customers with the max credit limit attired





Model Performance

By analyzing customer data provided by Thera Bank we will develop a classification model to identify and understand customers that are likely to churn from credit card services and make recommendations on ways to retain those customers.

12 models trained with Recall as the metric

Conclusion

The best performing model was **XGBoost Tuned** and could predict the likelihood of a customer attiring **98.4 %** of the time on test data.

- Recall on Validation set : 100%
- **Recall on Test set : 98.4%**

	Model	Train Recall	Validation Recall
0	CV Gradient Boosted	97.800	98.410
1	CV XGBoost	97.300	97.880
2	CV Random Forest	97.700	98.230
3	GBM Tuned	98.600	98.400
4	XGB Tuned	100.000	100.000
5	RF Tuned	99.900	99.900
6	GBM Oversample	98.600	98.400
7	XGB Oversample	100.000	100.000
8	RF Oversample	99.900	99.900
9	GBM Undersample	99.200	99.200
10	XGB Undersample	100.000	100.000
11	RF Undersample	100.000	100.000

All Models performed very well on all metrics for train & validation and generalize very well.

We choose XGBoost Tuned model as it performed best across all metrics for cross validation, tuned, oversampling & undersampling.

Test set scores

- Recall - 98.4
- Accuracy - 93.0
- Precision - 93.6
- F1 Score - 95.9

Data Overview

The data is assumed to be a view of a random subset of customer accounts over the last year

Data has been provided via CSV file (BankChurn.csv | 1,176k)

Data Dictionary

- **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

The Data Contains:

- 10,127 rows and 21 columns
- There are 2 variables with missing values
  - Education_Level - 1519
  - Marital_Status - 749
- There are no Duplicate Rows
- Target variable Attrition_Flag is imbalanced with ‘Attrited Customer’ representing 16% of the data

Data Treatment:

- All missing values imputed; no observations deleted
  - Education_Level & Marital_Status missing values were imputed to Unknown
- “Income_Category”, “Card_Category”, “Education_Level” were rank order encoded
- "Customer_Age", "Credit_Limit", "Total_Revolving_Bal", "Total_Trans_Ct“ were scaled with the MinMax algorithm.
- Attrition_Flag was over and under sampled during modeling with minimal impact and was not used in the final model