Thera Bank

Contents

- Business Problem Overview and Approach
- Objective
- Data Overview
- Exploratory Data Analysis (EDA) – Univariate Analysis
- EDA – Bivariate Analysis Between Two Attributes
- Model Performance Summary
- Business Insights and Recommendations

Business Problem Overview and Solution Approach

- **Core Business Idea**
  Thera Bank is a company that wants to enable and establish a viable business model to maintain and expand their customer base. Currently, there are 4 types of credit card products that the company offers. Fees are charged to every user irrespective of usage, while others are charged under specified circumstances.

- **Problem to Tackle**
  Customers' leaving credit cards services would lead bank to loss. The company wants to analyze available customer data to identify customers who will leave their credit card services so that bank could improve upon those service areas.

- **Financial Implications**
  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.

- **Machine Learning Model to Solve the Problem**
  In building and applying different models from the dataset provided, Thera Bank will be able to predict if the customer is going to churn or not and the model will help the bank identify and improve its services so that customers do not renounce their credit cards.

Objective

To identify the best possible model that will give the required performance necessary to help the bank improve its services so that customers do not renounce their credit cards

Based upon the provided data set, conduct a statistical analysis to complete the following:

- Explore and visualize the dataset
- Build a classification model to predict if the customer is going to churn or not
- Optimize the model using appropriate techniques
- Generate a set of insights and recommendations that will help the bank improve and retain credit card customers

Data Overview

- Dataset contains information in relation to their personal and customer banking profile

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

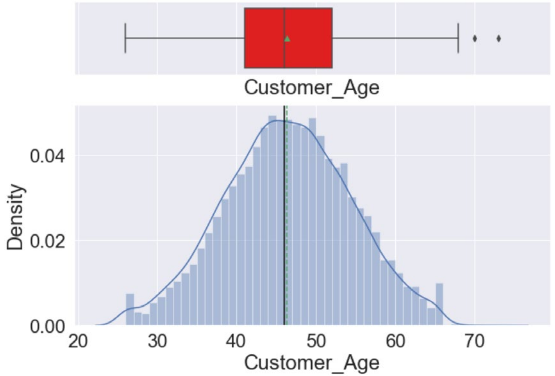
Entries	Variables	Elements of Missing Data
10127	21	3,380

Data Prep Required:

- Identified invalid string entries and processed them as missing data
- Missing values were imputed using KNN methodology
- High Outliers Identified and capped for analysis models

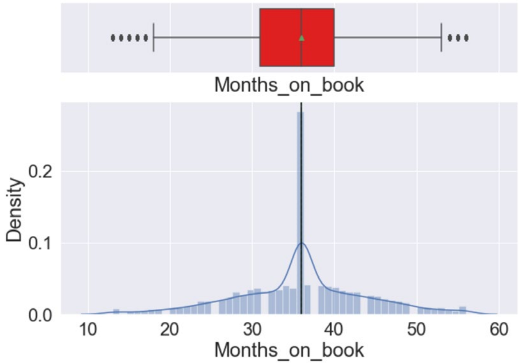
EDA – Univariate Analysis: Customer Age, Months on Book, label, Total Trans Amount

Customer Age



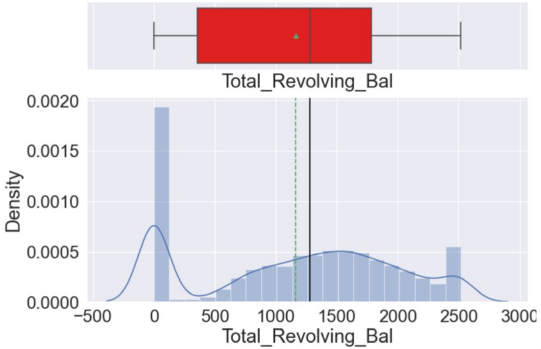
- Average customer is 46 years of age, attribute range is 26 - 73 which appears normal
- Dataset contains 45 unique entries, and a count range of 398 - 500 for the top 10 common entries

Months on Book



- Average customer has been on the books 36 months, 13 - 56 month attribute range
- 44 unique entries for this attribute, with 36 months the highest count of 2463 (~24%). Second highest count being 358 for 37 months

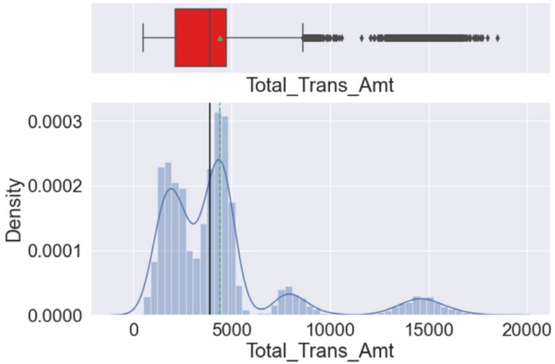
Total Revolving Balance



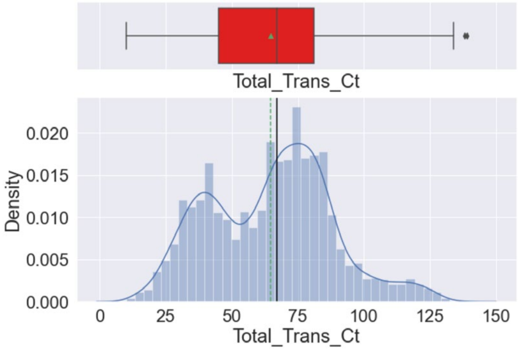
- Average customer total revolving balance is 1162.81, with 359 - 2517 being typical based on IRQ
- Over 1,970 unique entries, with 24% of the customer base having a zero balance, and over 500 (~5%) customers with a balance of 2,517, leaving ~70% of customers having a balance of less than 2000

EDA – Univariate Analysis: Total Trans Amount, Total Trans Amount

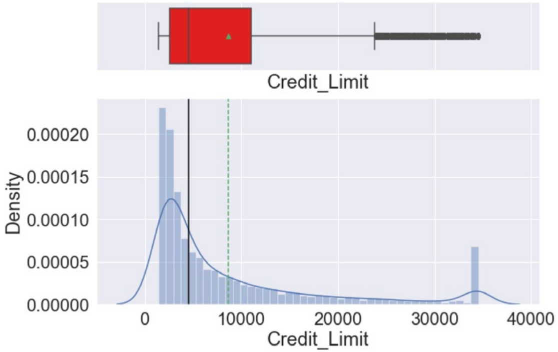
Total Trans Amount



Total Trans Count



Credit Limit



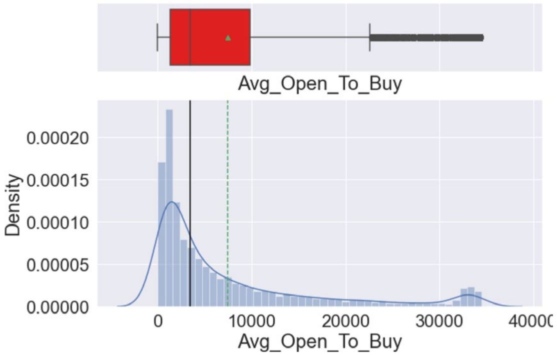
- Average customer total transaction amount has been 4,404, with 510 – 4,741 being a typical range based on IRQ values
- The most common transaction amounts being over 4000

- Average customer has had almost 65 transactions in the last 12 months, with 45 - 81 being a typical range based on IRQ values
- Over 125 unique entries, with the most common transaction counts being in the range of high 60's to the low 80's

- Average customer has a 8,631 credit limit, while the attribute range is 1,438 to 3,4516
- Two credit limits of 34,516 and 1,438 being the most common in having over 500 counts (~10%), leaving ~90% of the customers in the other balances

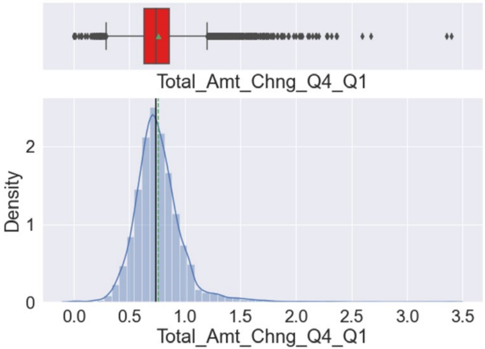
EDA – Univariate Analysis: Ave. Open to Buy, Total Amt Change Q4/Q1, Total Amt Change Q4/Q1

Ave. Open to Buy



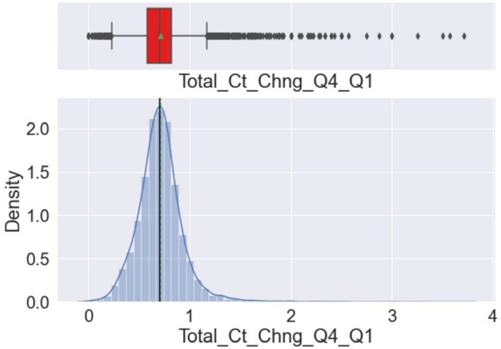
- Average customer has over 7,469 available, with 1,324 – 9,859 being a typical range based on IRQ values
- The most common entry being 1,438.3, and two high values of 34,516 (98 counts) and 31,999 (26 counts) which could possibly be outliers or skew analysis

Total Amt Change Q4/Q1



- Average customer ratio is approximately 0.76,with a typical range being 0.63 - 0.86 based on IRQ values. Attribute displays a possible outlier max value of 3.40 to be investigated further

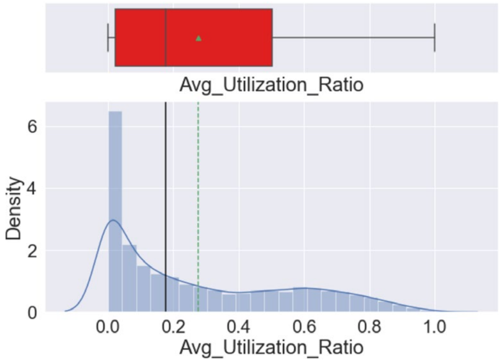
Total Ct Change Q4/Q1



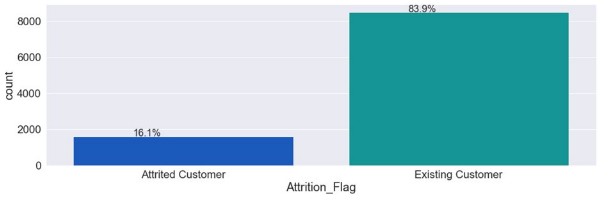
- Average customer ratio is approximately 0.71, with a typical range being 0.58 - 0.82 based on IRQ values
- 166 entries displaying a 1:1 change ratio, other more common ratios range from ~.55 to ~85

EDA – Univariate Analysis: Avg Utilization Ratio, Attrition Flag, Gender

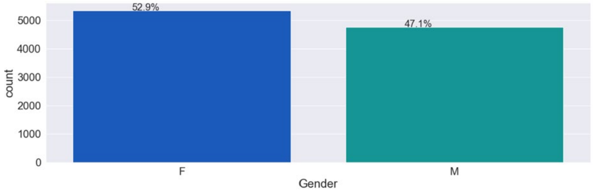
Avg Utilization Ratio



Attrition Flag



Gender



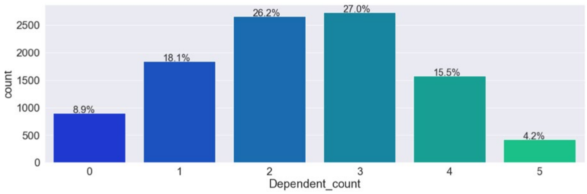
- Average customer ratio is 0.27, with a typical range being 0.02 - 0.50 based on IRQ values
- Over 24% (2470) of the customer base not spending their available balance, other more common ratios are less than .08

- The dependent binary attribute variable shows that the overall customer base consists of almost 84% (8500) existing customers, to 16% (1627) attrite customers

- Attribute displays that Females account for almost 53% of the customer base, as compared to 47% (4769) of the customers being male

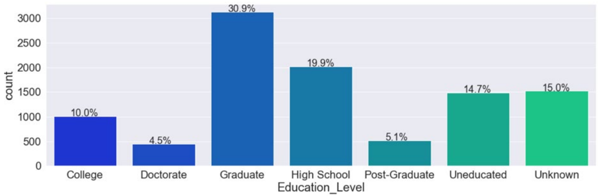
EDA – Univariate Analysis: Dependent Count, Education Level, Marital Status

Dependent Count



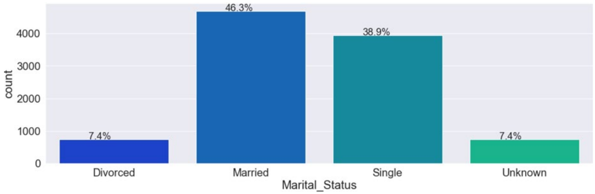
- Category range is from 0 - 5, with most customers having 2 & 3 dependents containing the highest counts greater than 2600

Education Level



- Graduates account for the most common education level of the customer base
- Attribute contained 1519 (~15%) missing/unknown entries

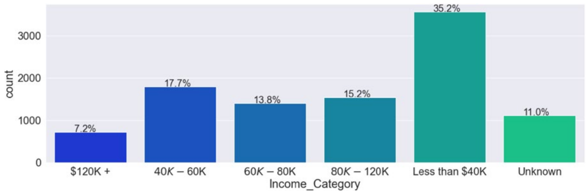
Marital Status



- Married customers account for almost 50% of the 3 -categories within the customer base
- Attribute contained 749 (~15%) missing/unknown entries

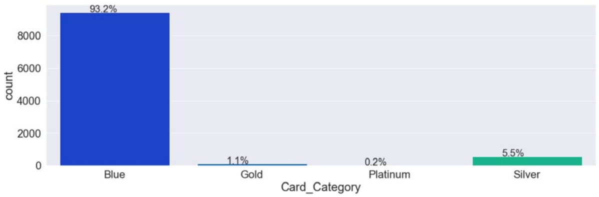
EDA – Univariate Analysis: Income Category, Card Category, Total Relationship Count

Income Category



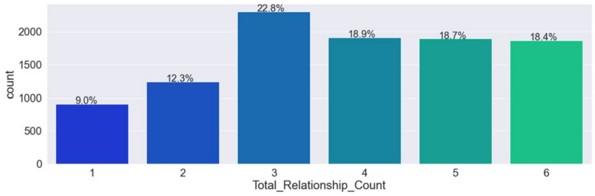
- Customers earning less than 40K account for the most common (~35%) income category within the customer base
- Attribute contains 1112 (~11%) invalid string entries, that were changed to “Unknown” during the EDA portion of the analysis

Card Category



- The most common card carried by customers is the Blue card from the 4 cards available, which accounts for ~93% of the customer base

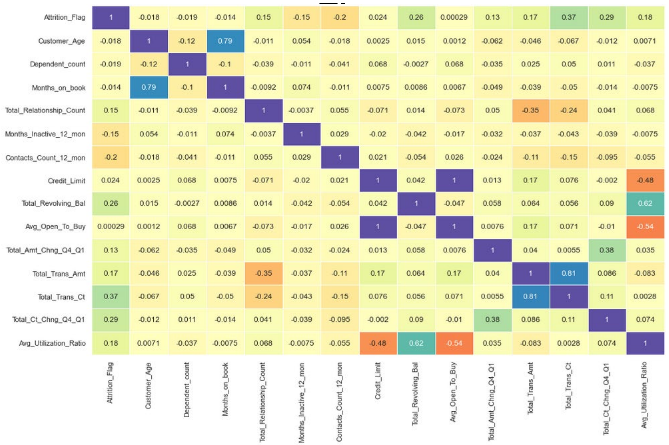
Total Relationship Count



- Average customer has almost 4 bank products
- 3 being the highest and 4, 5, 6, having extremely close counts
- 1 - 6 category type attribute range

Exploratory Data Analysis – Bivariate Analysis

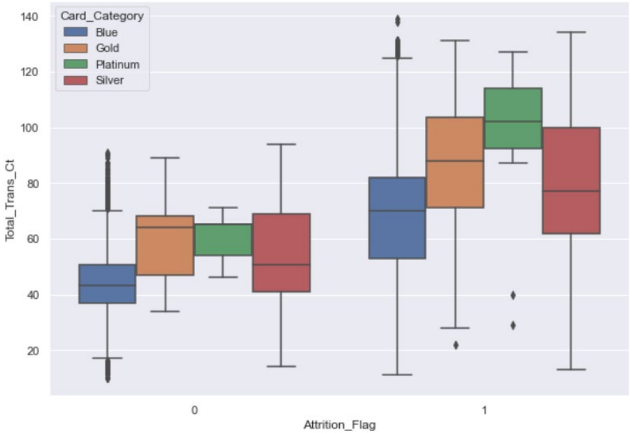
Correlation in Relation to Attrition



	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	-0.018	-0.019	-0.014	0.15	-0.15	-0.2	0.024	0.26	0.00029	0.13	0.17	0.37	0.29	0.18
Customer_Age	-0.018	1	-0.12	0.79	-0.011	0.054	-0.018	0.0025	0.015	0.0012	-0.062	-0.046	-0.067	-0.012	0.0071
Dependent_count	-0.019	-0.12	1	-0.1	-0.039	-0.011	-0.041	0.068	-0.0027	0.068	-0.035	0.025	0.05	0.011	-0.037
Months_on_book	-0.014	0.79	-0.1	1	-0.0092	0.074	-0.011	0.0075	0.0086	0.0067	-0.049	-0.039	-0.05	-0.014	-0.0075
Total_Relationship_Count	0.15	-0.011	-0.039	-0.0092	1	-0.0037	0.055	-0.071	0.014	-0.073	0.05	-0.35	-0.24	0.041	0.068
Months_Inactive_12_mon	-0.15	0.054	-0.011	0.074	-0.0037	1	0.029	-0.02	-0.042	-0.017	-0.032	-0.037	-0.043	-0.039	-0.0075
Contacts_Count_12_mon	-0.2	-0.018	-0.041	-0.011	0.055	0.029	1	0.021	-0.054	0.026	-0.024	-0.11	-0.15	-0.095	-0.055
Credit_Limit	0.024	0.0025	0.068	0.0075	-0.071	-0.02	0.021	1	0.042	1	0.013	0.17	0.076	-0.002	-0.48
Total_Revolving_Bal	0.26	0.015	-0.0027	0.0086	0.014	-0.042	-0.054	0.042	1	-0.047	0.058	0.064	0.056	0.09	0.62
Avg_Open_To_Buy	0.00029	0.0012	0.068	0.0067	-0.073	-0.017	0.026	1	-0.047	1	0.0076	0.17	0.071	-0.01	-0.54
Total_Amt_Chng_Q4_Q1	0.13	-0.062	-0.035	-0.049	0.05	-0.032	-0.024	0.013	0.058	0.0076	1	0.04	0.0055	0.38	0.035
Total_Trans_Amt	0.17	-0.046	0.025	-0.039	-0.35	-0.037	-0.11	0.17	0.064	0.17	0.04	1	0.81	0.086	-0.083
Total_Trans_Ct	0.37	-0.067	0.05	-0.05	-0.24	-0.043	-0.15	0.076	0.056	0.071	0.0055	0.81	1	0.11	0.0028
Total_Ct_Chng_Q4_Q1	0.29	-0.012	0.011	-0.014	0.041	-0.039	-0.095	-0.002	0.09	-0.01	0.38	0.086	0.11	1	0.074
Avg_Utilization_Ratio	0.18	0.0071	-0.037	-0.0075	0.068	-0.0075	-0.055	-0.48	0.62	-0.54	0.035	-0.083	0.0028	0.074	1

- The Correlation in Relation to Attrition Flag displayed the following:
- Attributes such as Total Trans Ct, Total_Ct_Chng_Q4_Q1, and Total Revolving Bal have the highest positive correlation amounts (0.2<) respectively when compared to the Attrition Flag attribute
- Contacts_Count_12_mon and Months_Inactive_12_mon attributes display largest negative correlation amounts, yet by only a -0.2 amount

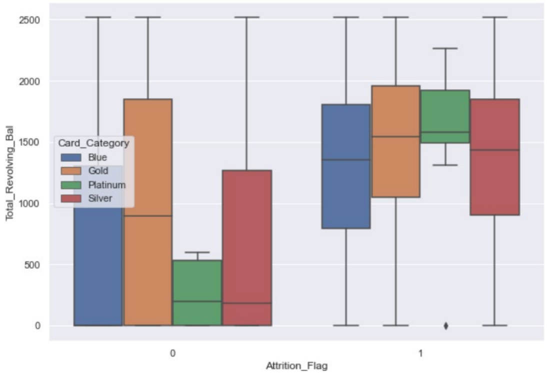
Attrition vs. Total Transaction Count



- This plot displaying Total Transaction Counts, broken down by Credit Card Category and the Attrition Flag provides some strong indicators of customers that will be likely looking to leave Thera Bank
- Other indicators that a customer may possibly leave Thera Bank's Credit Card Service included ratio attributes, Utilization balance, and Transaction counts

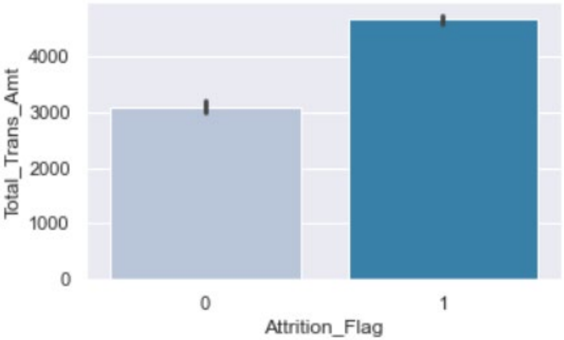
Exploratory Data Analysis – Customer Attrition Indicators

Attrition Flag vs. Total Revolving Balance



- This plot displaying Total Revolving Balances, broken down by Credit Card Category and the Attrition Flag provides some strong indicators of customers that will be likely looking to leave Thera Bank
- Except for the customers having a Gold Card, customers having a Total Revolving Balance greater than 13,000 are more likely to leave Thera Bank

Attrition Flag vs. Total Transaction Amounts



- The above two plots show that card customers with Total Transaction Amounts greater that 3,000, and transaction amounts much greater than 11,000 are likely to leave Thera Bank's credit card services

Model Overview

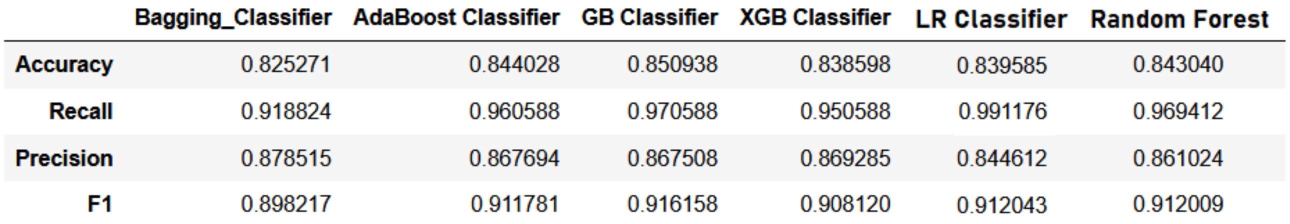
3 sets, each containing 6 models were built using the following models to measure False Negatives:

  - Logistic Regression
  - Random Forest Classifier
  - Bagging Classifier
  - AdaBoost Classifier
  - Gradient Boosting Classifier
  - XGBoost Classifier Model

- First, the models were used as they were for a baseline comparison
- Second, Model Building was conducted using oversampled data using SMOTE
- Next, Model Building was conducted using undersampled data using Random Under Sample
- The 3 models were chosen to be tune to see if performance could be improved
- Followed by Hyperparameter Tuning for each model
- Hyperparameter tuning was applied using random search
- A final model using pipelines was then built and tested
- Comparison the models was performed of the models for their performance
- Most important factors used by the ML model for prediction:
  - Data weight balancing was required for a more accurate ML model
  - Applied visualization of Confusion Matrix and Importance Feature
  - Applied Accuracy, Recall, and Importance Feature measures to measure performance
  - Applied Random Search for Hyperparameter Tuning
  - Use of GridSearchCV for Computing Importance Feature

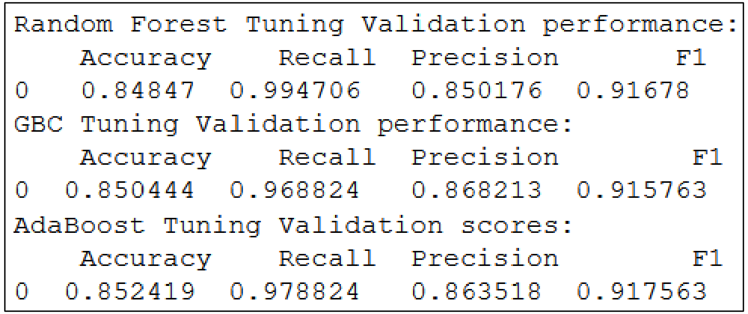
Model Performance Summary

Initial Validation Training Performance Comparison



	Bagging_Classifier	AdaBoost Classifier	GB Classifier	XGB Classifier	LR Classifier	Random Forest
Accuracy	0.825271	0.844028	0.850938	0.838598	0.839585	0.843040
Recall	0.918824	0.960588	0.970588	0.950588	0.991176	0.969412
Precision	0.878515	0.867694	0.867508	0.869285	0.844612	0.861024
F1	0.898217	0.911781	0.916158	0.908120	0.912043	0.912009

- Based up the results above, the three models with the highest Recall values are from AdaBoost, GBC, and the Logistic Regression model respectively
- Random forest displayed the best performance of the three tuned models to the right
- Limitations of hardware prevented successful model building of all hyper tuning boosting models



```
Random Forest Tuning Validation performance:
    Accuracy    Recall  Precision       F1
0    0.84847  0.994706   0.850176  0.91678
GBC Tuning Validation performance:
    Accuracy    Recall  Precision        F1
0   0.850444  0.968824   0.868213  0.915763
AdaBoost Tuning Validation scores:
    Accuracy    Recall  Precision        F1
0   0.852419  0.978824   0.863518  0.917563
```

Business Insights and Recommendations

- Recommendations based on EDA and interpretation of the model input variables
  - Thera Bank should make a strong effort to expand marketing toward the proper income level of their credit card customer to increase Silver, Gold, and Platinum credit cards, especially to customers in higher income categories
  - Based upon the Months Inactive attribute, marketing should conduct more periodic contact to customers that tend to have inactive banking patterns
  - Dependent count did not appear to have much influence within this analysis
- Insight to Profile of Customers' Leaving Thera Bank's Credit Cards Services
  - 75% of customers looking to leave the bank, will likely have only 2-3 contacts with the bank during a 12-month period as compared to customers
  - Existing customers that remain typically have 2-4 contacts or more during the same 12-month time period
  - Except for the customers having a Gold Card, customers having a Total Revolving Balance greater than 13,000 are more likely to leave Thera Bank
  - Except for customers with a zero Total Revolving Balance and a Platinum Card, Platinum Card holders with Total Revolving Balances above 10,000 are extremely likely to leave Thera Bank
  - Card customers with higher ratios regarding Average Utilization Ratio, Total Count Change, Total Amount Changes are more likely to leave Thera Bank

greatlearning

Power Ahead

Happy Learning !