Project 3

Credit Card Users Churn Prediction
PGP AI/ML October 23

Saurabh Mittal

3/1/24

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model performance summary for hyperparameter tuning.
- Appendix

Executive Summary

- The objective of this study is to identify various factors that would help the Thera Bank address attrition and target the existing customer base to prevent attrition
- The data contains 10127 records: 21 attributes, with missing values for 3 of the attributes, which after analysis were found to be less impactful. The data was of good quality.
- The data from the past 12 months was analyzed using Machine Learning methodology that revealed the following insights
    - Total transaction count, total transaction amounts, revolving balances and holders of various Thera products are the top criteria influencing attrition rate
    - About 38% of customer were inactive for 3 months and 33% for 2 months, making 71% inactive for at least 2 months
    - About 80% of customer base holds more than 3 Thera products
    - About 66% of the customers were contacted at least 2 times per year.
    - The attrition rate is not that high. It is under 18% of the current customer base
- Recommendations
    - Customers who are not utilizing their current Thera products in an active manner are most likely to attrite, which is an intuitive result. The Bank must increase efforts for upselling various products to customers who are not utilizing their existing products fully
    - Bank should incentivize customers who log onto to the website or utilize services actively

Business Problem Overview and Solution Approach

- Problem
  - 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.
  - Develop a classification model that will help the bank improve its services so that customers do not renounce their credit cards, and identify the best possible model that will give the required performance
- Solution approach / methodology
  - Bank would want Recall to be maximized, greater the Recall higher the chances of minimizing false negatives. Hence, the focus should be on increasing Recall or minimizing the false negatives or in other words identifying the true positives(i.e. Class 1) so that the bank can retain their valuable customers by identifying the customers who are at risk of attrition.
  - Build a suite of models: Bagging, Random Forrest, Gradient Boosting, Ada Boosting, and Decision Tree
  - Utilize Hyperparameter Tuning by undersampling and oversampling data to tune the models
  - Select the optimally tuned model and verify with test set

EDA Results

EDA Summary

- The dataset has 10127 rows and 21 columns
- Only 6 variables are object types, rest all are numerical types.
- 2 columns have less than 10127 non-null values i.e. columns have missing values.

Data Cleaning

- Drop ”CLIENTNUM" as ”CLIENTNUM" is unique for each candidate and might not add value to modeling.
- Missing value imputation will be done after splitting the data.

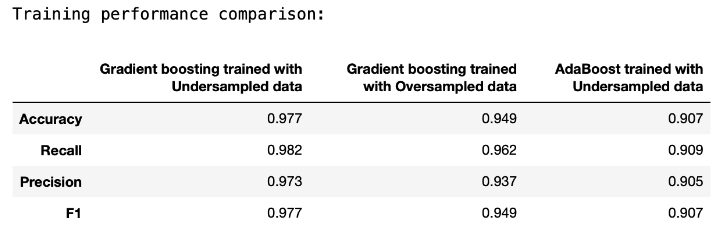
Link to Appendix slide on data background check

Data Preprocessing

- Duplicate value check: No duplicates
- Missing value treatment: The below columns were imputed
  - Education_Level has 15% missing data: 1519 records
  - Marital_Status has 7.4% missing data: 749 records
  - Income_Category has 11% missing data: 1112 records
- Outlier check
  - Few of the variables have Outlier data: Credit_Limit (10%), Avg_Open_To_Buy(10%), Total_Transac_Amt (9%). The other columns with less than 6% outliers are: Months_on_book (4%), Months_Inactive_12_mon (3%), Contacts_Count_12_mon(6%), Total_Amt_Chng_Q4_Q1(4%), and Total_Ct_Chng_Q4_Q1(4%)
- Data preparation for modeling
  - Training set has 8101 records, Validation set has 507 records, Test set has 1519 records
  - Total columns: 19
  - After encoding, total columns = 29

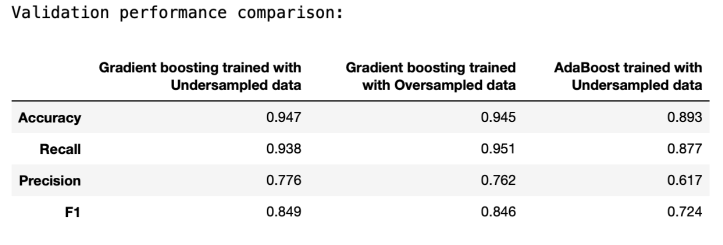
Model Performance Summary

- Model Selection after Hyperparameter Tuning



```
Training performance comparison:
```

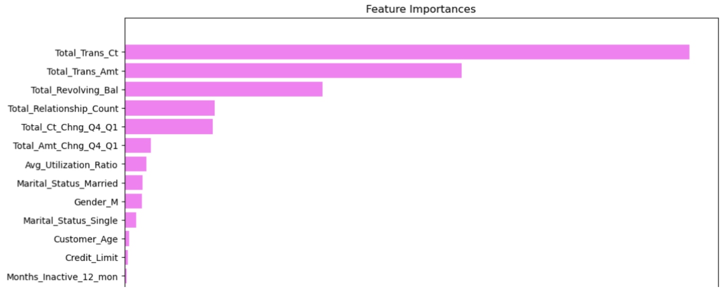
	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.977	0.949	0.907
Recall	0.982	0.962	0.909
Precision	0.973	0.937	0.905
F1	0.977	0.949	0.907



```
Validation performance comparison:
```

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Oversampled data	AdaBoost trained with Undersampled data
Accuracy	0.947	0.945	0.893
Recall	0.938	0.951	0.877
Precision	0.776	0.762	0.617
F1	0.849	0.846	0.724

Gradient Boosting with Oversampled Data has generalized performance and achieves 93% Recall with Test data set



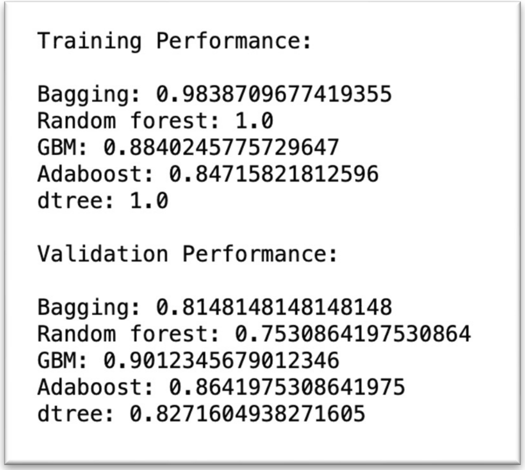
Top 5 Features:

- Total Transaction Count
- Total Transaction Amount
- Total Revolving Balance
- Total Relationship Count
- Total Count for Change Q4 to Q1

APPENDIX

Model Performance Summary (original data)

- Summary of performance metrics for training and validation data in tabular format for comparison for original data.
- GradientBoost (GBM) and Adaboost have the closest numbers in training/validating, i.e., the difference between the corresponding values is minimal.

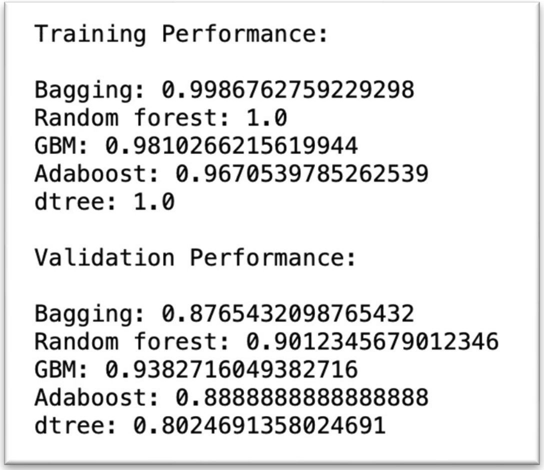


```
Training Performance:

Bagging: 0.9838709677419355
Random forest: 1.0
GBM: 0.8840245775729647
Adaboost: 0.84715821812596
dtree: 1.0

Validation Performance:

Bagging: 0.8148148148148148
Random forest: 0.7530864197530864
GBM: 0.9012345679012346
Adaboost: 0.8641975308641975
dtree: 0.8271604938271605
```

Link to Appendix slide on model assumptions

Model Performance Summary (oversampled data)

- Summary of performance metrics for training and validation data in tabular format for comparison for oversampled data. Oversampling method chosen: SMOTE with 5 K-nearest neighbors
- GradientBoost (GBM) and Adaboost have the closest numbers in training/validating, i.e., the difference between the corresponding values is minimal.

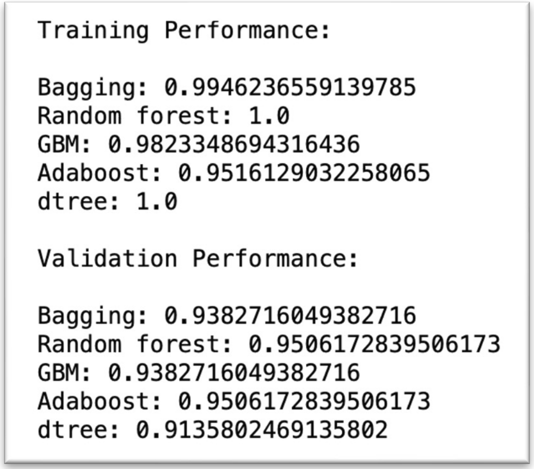


```
Training Performance:

Bagging: 0.9986762759229298
Random forest: 1.0
GBM: 0.9810266215619944
Adaboost: 0.9670539785262539
dtree: 1.0

Validation Performance:

Bagging: 0.8765432098765432
Random forest: 0.9012345679012346
GBM: 0.9382716049382716
Adaboost: 0.8888888888888888
dtree: 0.8024691358024691
```

Link to Appendix slide on model assumptions

Model Performance Summary (undersampled data)

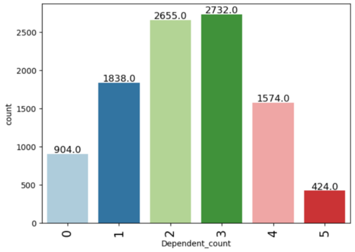
- Summary of performance metrics for training and validation data in tabular format for comparison for undersampled data. Undersampling method chosen: RandomUnderSampler
- GradientBoost (GBM) and Adaboost have the closest numbers in training/validating, i.e., the difference between the corresponding values is minimal.

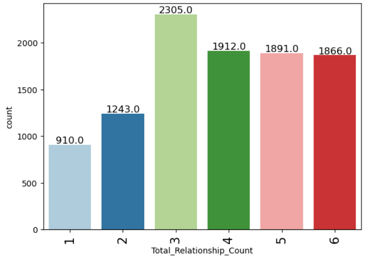


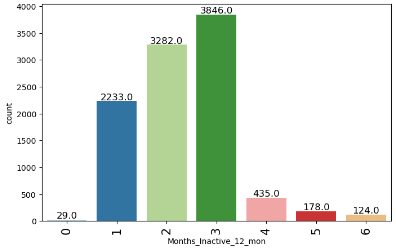
```
Training Performance:

Bagging: 0.9946236559139785
Random forest: 1.0
GBM: 0.9823348694316436
Adaboost: 0.9516129032258065
dtree: 1.0

Validation Performance:

Bagging: 0.9382716049382716
Random forest: 0.9506172839506173
GBM: 0.9382716049382716
Adaboost: 0.9506172839506173
dtree: 0.9135802469135802
```

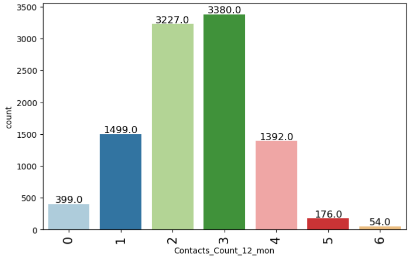
Link to Appendix slide on model assumptions

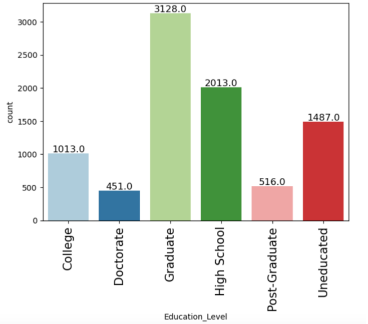
Various other distributions in Data

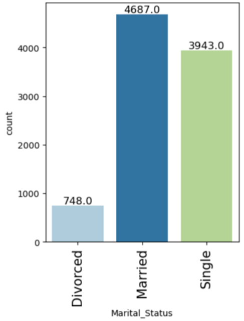


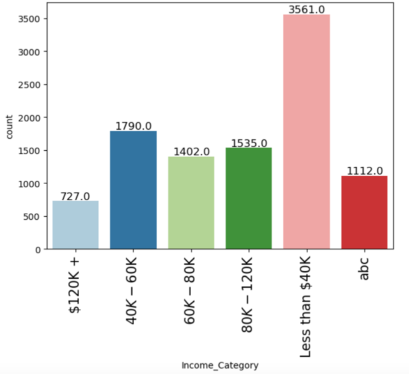


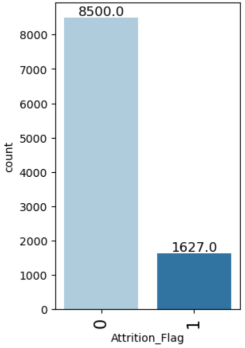




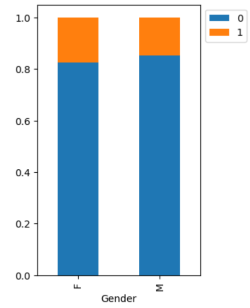




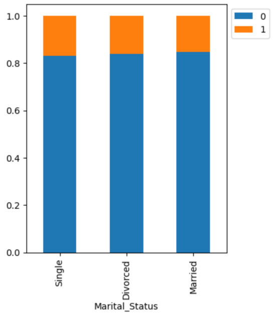




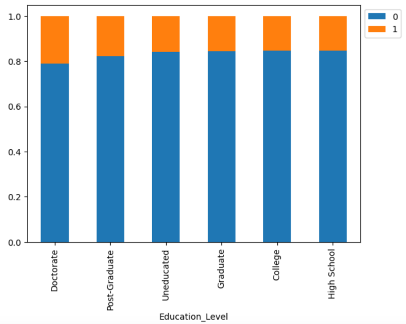
Variables and Distributions for Attrition



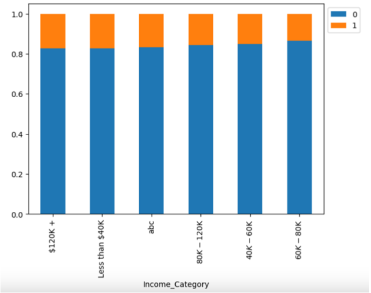
17% of Females and
15% of Males attrited



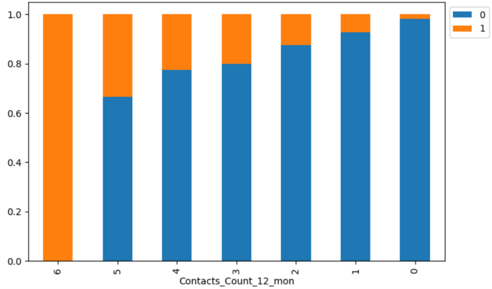
15% of Marital_Status
is attrited



16% of Education_level is
attrited



16% of Income_category is
attrited

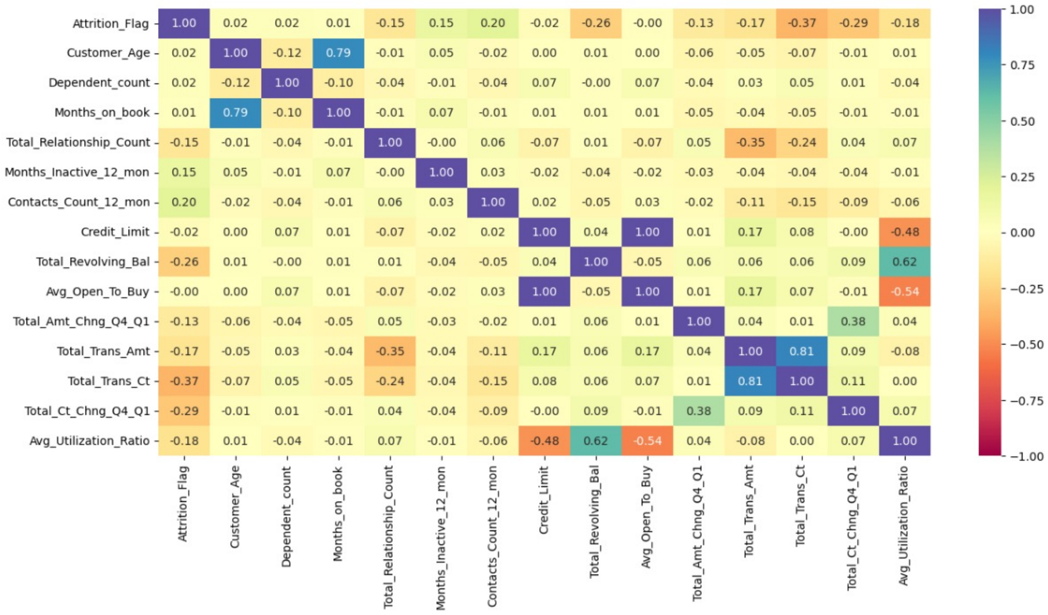


16% of Contacts_Count_12_mon is
attrited. As the contacts are
increased, the attrition increases

For all the specific pairs and their impact on attrition, the dataset is quite uniform in the sense that about 15-18% of customers are attrited.

Correlation

- Positive Correlation exists between four pairs:
    - 100%: Avg_Open_To_Buy and Credit_Limit
    - 79%: Months_on_book and Customer_age
    - 62%: Avg_Utilization_Ratio and Total_Revolving_balance
    - 81%: Total_Trans_Ct and total_Transact_Amount
- No significant negative correlation was found



	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

Data Dictionary

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

Happy Learning !