The Thera Bank

Case Study

Sven Meydell

Areas of Focus

- **Core Business Idea:**
  - To predict which customers are most likely to renounce their credit cards and to help improve services so as to prevent the issue from prevailing in the future

- **Financial Implications:**
  - The classification model should be able to **accurately predict which customers are at the highest risk of closing out their Credit Card Account**
  - It is critical that the bank **accurately target <u>all</u> customers at risk** of closing their accounts, to **prevent further business loss**
    - The model therefore needs to be as accurate as possible in predicting all customers at risks who we could lose business from
    - Specifically, **Recall testing should be the focus** – we want all customers predicted to close out their accounts addressed, while minimizing the potential of missing any customers
      - Any customers that incorrectly categorized that end up closing their accounts will cost the bank lost revenue and should be minimized as best as possible

Solving Problems with ML

- **Problem:**
  - Saving the most possible revenue through correctly identifying and catering to those customers most likely to leave the bank or close out their Credit Card services
    - There are opportunities for the company to grow prior relationships and prevent further brand loyalty diminishment through **targeting specific customers based on profiling their, and other similar, customer profiles**
  - Additional cost is a secondary concern and not to be curbed for this campaign as maintaining current revenues is crucial for business/department survival
- **Solution:**
  - Machine Learning can help reduce the uncertainty by factoring in each of the many variables in the current customer dataset and **numerically assigning values collectively to predict one final result**, the likelihood of a customer closing their Credit Card accounts with the bank
  - The **final prediction is accurate within a very high, statistically significant, acceptance level** and the automated model can be easily replicated and tuned as new data becomes available

Objectives

- To predict whether/not a customer is currently, or in the near future, at risk of closing out their credit account/s
- Identify the most significant variables in the sample dataset affecting the target outcome and suggest ways to mitigate the risks of attrition
- Determine which segments of customers should be most/least targeted for best results of retaining at-risk customers

Which Scoring Metric to Use

All Ensemble Models (Decision Trees, Bagging, and Boosting Techniques) will use: **Recall as the Scoring Metric**

- **Precision:** Out of all the customers predicted to close out their credit cards, how many did?
  - True Positives / (True Positive + False Positives)

- **Recall (Sensitivity):** Of all the customers that did actually cancel their credit cards, how many did the model predict would?
  - True Positive / (True Positives + False Negatives)

- **F1-Score (Combo of Precision & Recall/Sensitivity):** What is the Harmonic Mean split between the Precision and Recall results?
  - 2 * (Recall * Precision) / (Recall + Precision)

- **Specificity:** Of all the customers that did not close out their credit cards, how many did the model predict wouldn't?
  - True Negative / True Negatives + True Positives

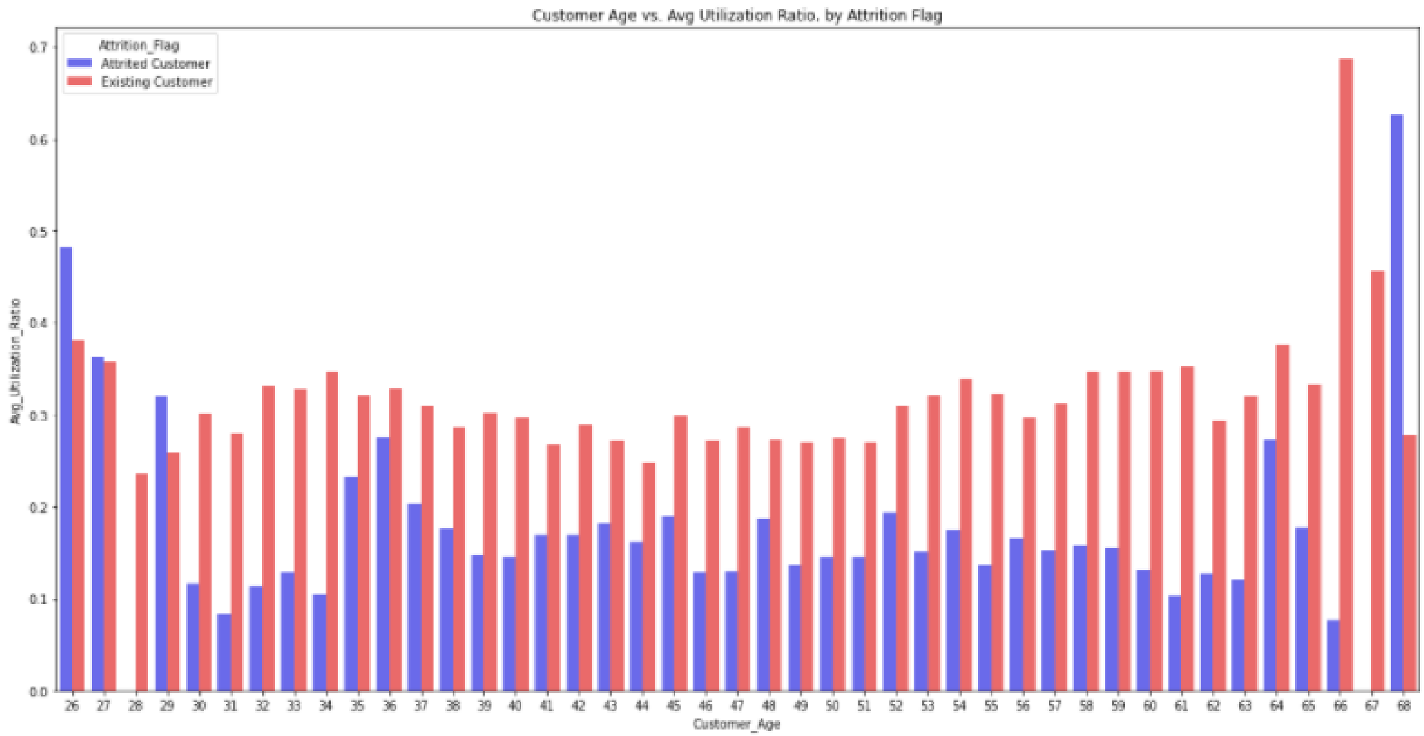
Data Provided:
Customer Details

- **CLIENTNUM:** Client number for the customer holding the account (Unique)
- **Attrition Flag:** Indicates if the customer account is closed ('Attrited' Customer) or active (Existing Customer)
- **Customer Age:** Age in Years
- **Gender:** Gender of the Account Holder
- **Dependent Count:** Number of Dependents
- **Education Level:** Education of Account holder (Ordinal): Graduate, High School, Unknown, Uneducated, College (Still a 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 Utilized by Customer
- **Months on Book:** Length of Relationship with the Bank
- **Total Relationship Count:** Total (Bank) Products Held by the Customer
- **Months Inactive 12 Months:** Count of Months Account Inactive in the last Year (12 Months)
- **Contacts Count** 12 Months: Count of Customer Contacts in the last Year (12 Months)
- **Credit Limit:** Credit Limit on the Credit Card
- **Total Revolving Bal:** The balance Carrying Over Month-to-Month (Revolving)
- **Avg. Open To Buy:** Balance Left on Credit Card Still Available for Use (Average of Prior 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 4th Quarter Total Transactions to 1st Quarter Total Transactions
- **Total Amt Chng Q4 Q1:** Ratio of 4th Quarter Total Amount to 1st Quarter Total Amount
- **Avg. Utilization Ratio:** Represents Amount of Available Credit Used/Spent by Customer

Manipulations of Raw Data

- Removal of **CLIENTNUM** variable as it offered little to no value
- **Renaming** fields or types:
  - 'Uneducated' education level to 'High School', as it is assumed that each Customer has at least some level of High School education
- Converting object datatypes to **Categorical** first then encoding to numeric through **One-Hot-Encoding** process
  - Gender
  - Education Level
  - Marital Status
  - Income Category
  - Card Category
- Subgrouping numeric values (addressing Outliers) into smaller groups and converting to **Categorical**
  - Months on Book
  - Total Transaction Amount
- Capping Outliers to their max values for select columns
  - Customer Age
  - Total Transaction Count
- Missing values **Imputed with Median (**in Training Data only):
  - Education Level: 1519
  - Marital Status: 749

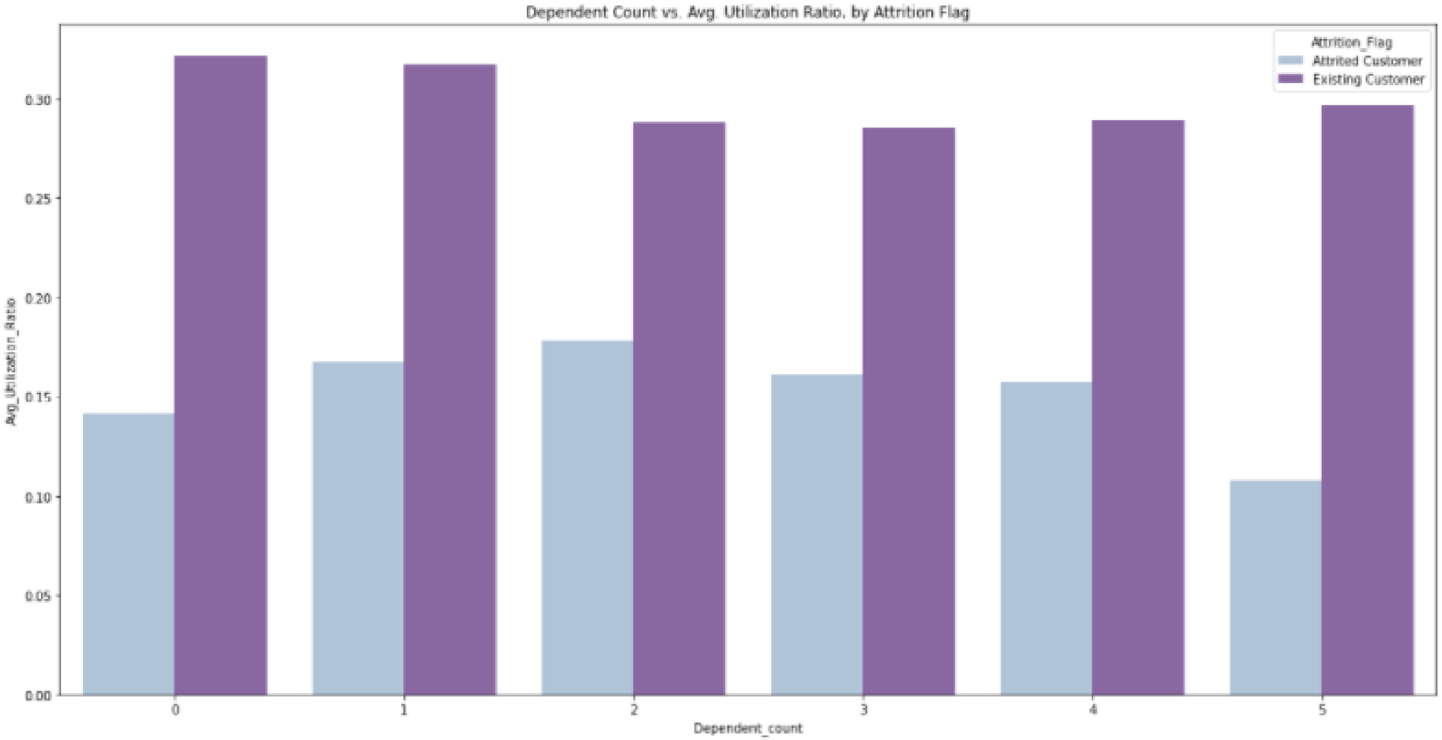
Attrition Flag:
Customer Age & Avg. Utilization Ratio



Customers, of all ages, who **Utilize a higher portion of their Credit Limit are more likely to stay active** with their Credit Card service

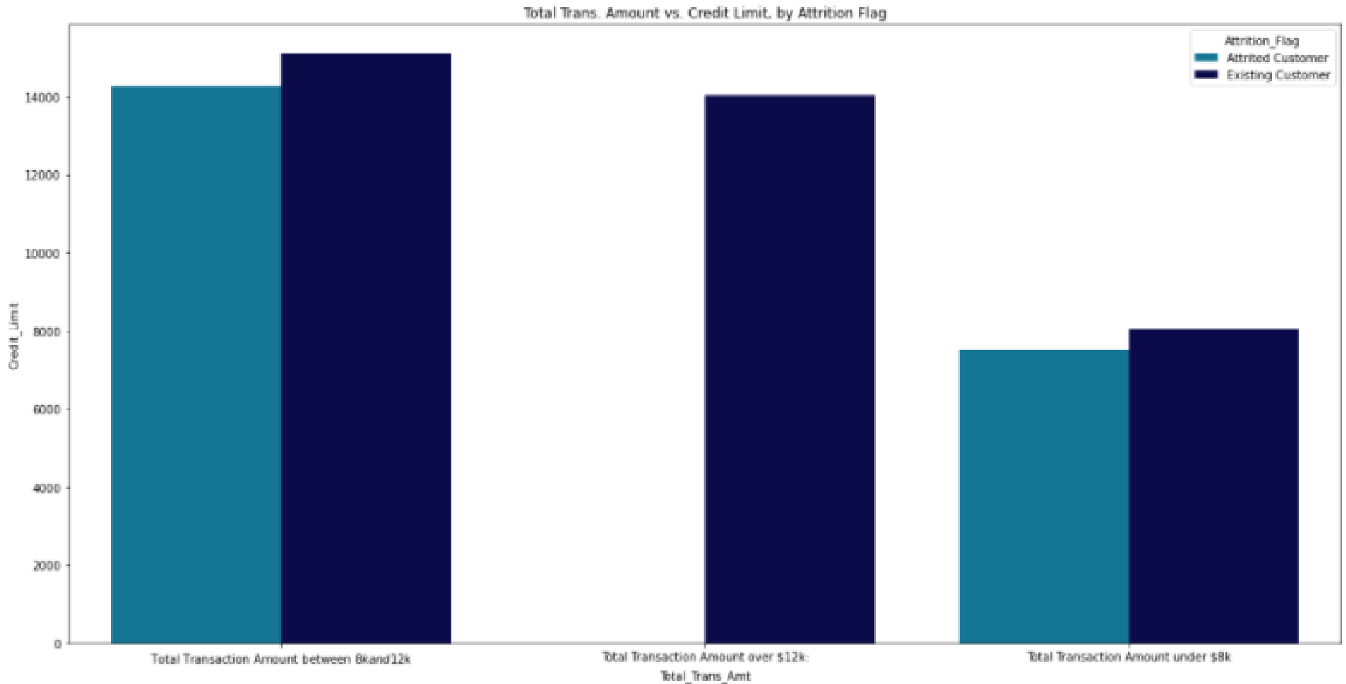
- There are anomalies for ages 26 and 68 who, for whatever reasons, spend a large portion of their available Credit Limits and still close out their accounts
  - This may be due to **Balance Transfer Promotions** and other incentives from other competitors or possibly the downsizing and closure of Credit facilities by **older customers transitioning to a retired lifestyle**

Attrition Flag:
Dependent Count & Avg. Utilization Ratio



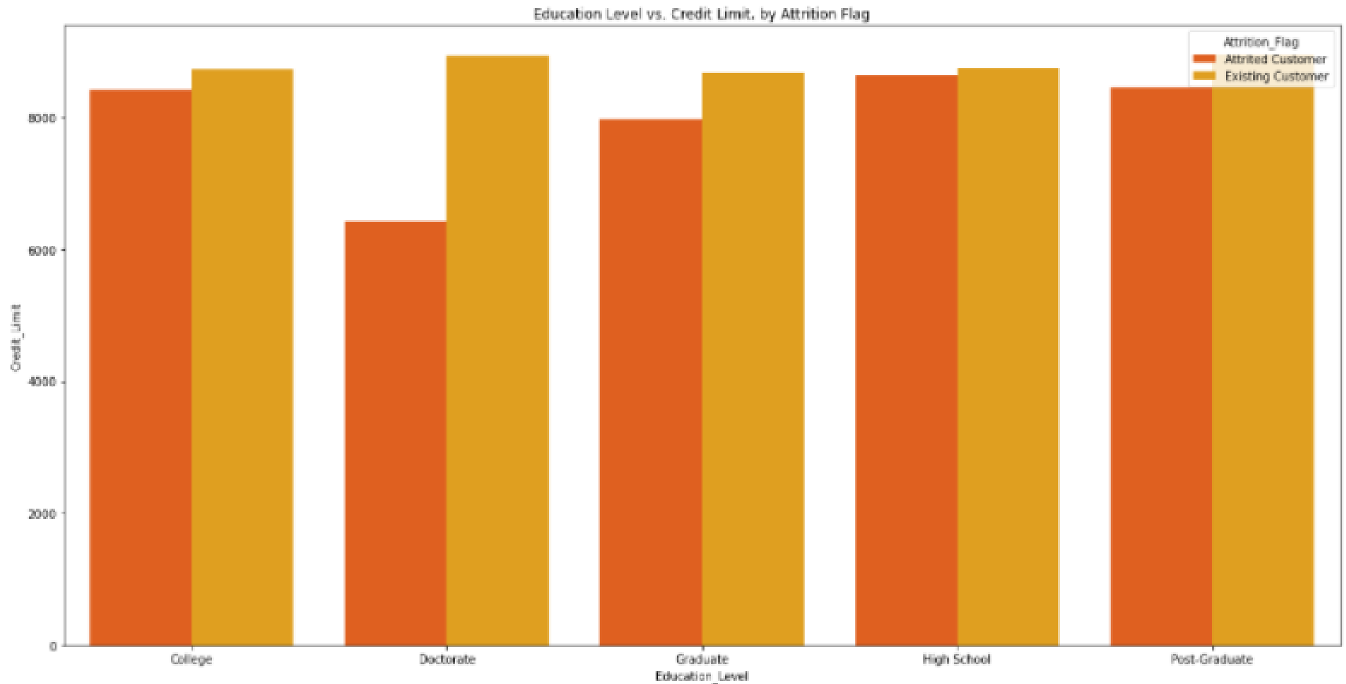
- Dependent counts have little effect on determining customer attrition as it relates to Credit Cards
- **Average Utilization Ratios**, however, show a strong increase in likelihood of staying active as a customer, the higher they increase **-** particularly when **surpassing 30%**

Attrition Flag: Total Transaction Amount & Credit Limit



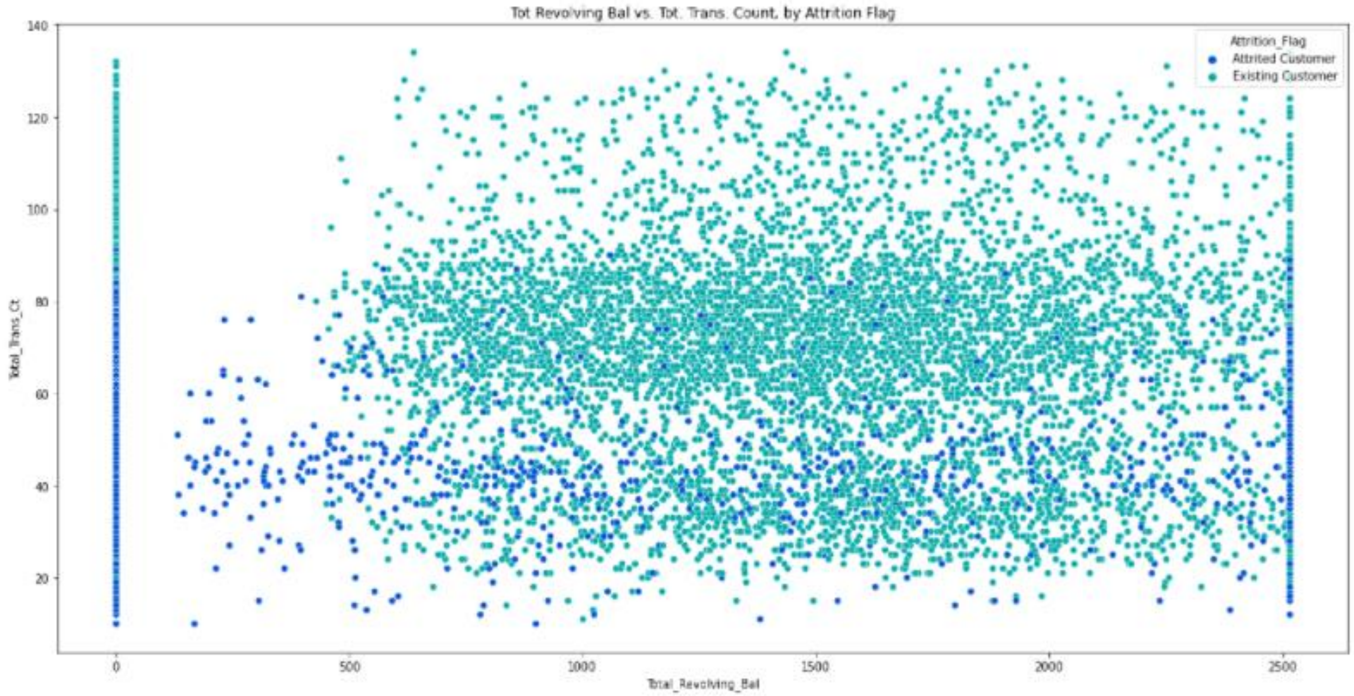
- **Customers spending over $12k a year show no likelihood of attrition**
- The majority of customers sampled spend **between $8k and $12k**
  - A large portion of these customers are shown to cancel their Credit Cards and appear to be grossly underserved by the Bank in terms of customer loyalty, etc.
  - Similarly, to a lesser extent, customers spending **under $8k a year** have higher attrition levels and appear underserved by the bank and a large focal area for the existing customers left

Attrition Flag: Education Level & Credit Limit



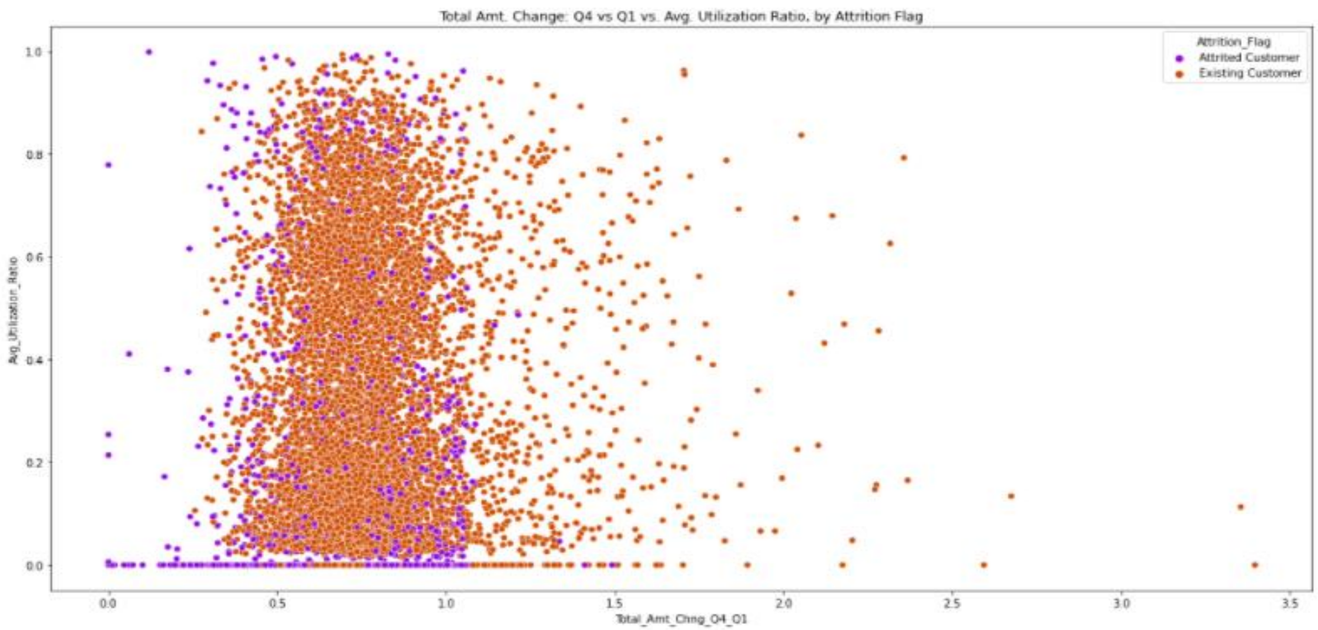
- **Doctorate and Graduate** educated customers have a lower chance of attrition
    - Credit Limits are relatively even across all Education Levels, with high level educated customers having slightly higher limits on average

Attrition Flag:
Total Revolving Bal. & Total Transaction Ct.



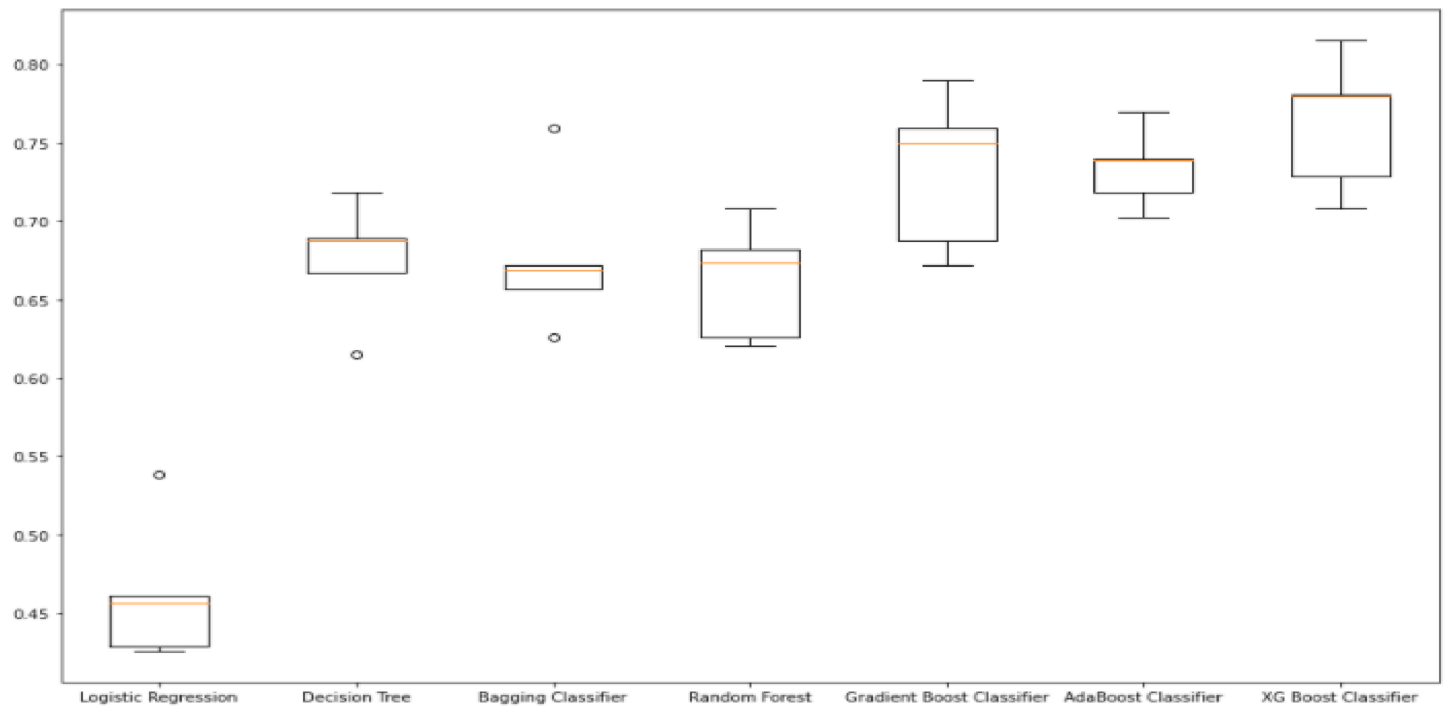
- In general, the **higher the Total Transaction Count**, the greater the chance of customers staying active with the company
- Customers with a \$**0 balance show full payment each month**
  - A lot of those have closed their Credit Cards
- There are also a lot of customers **carrying a promotional balance of $2.5k**, of which around 70% have stayed current with the bank while the remaining **30% have transferred their balances elsewhere**

Attrition Flag:
Total Amt. Change Q4/Q1 & Avg. Util. Ratio



- Any spending in Q4 that is in excess of 1x Q1 **spending is historically done by active customers**
  - **Increased Q4/Holiday promotional activity** could boost and maintain active Credit Card usage by customers
  - The Average Utilization Ratio can be seen to reach **nearly 100% of available balance during this timeframe**
    - This potentially **boosts transactional fees and revolving balance charges (interest)** while simultaneously keeping customers active and less likely to move their Credit Card business elsewhere

Initial Model Results



Observations

The **XG Boost Classifier** is giving the **highest Cross-Validated Recall score**, closely followed by the **AdaBoost and Gradient Boost Classifiers, which scored almost identically**

On the BoxPlot, the **XB Boost Classifier** shows a **larger range in CV Results**, ranging from around 0.72 to 0.82

The **AdaBoost Classifier**, on the other hand, had a **smaller range of CV Results**, ranging from around 0.71 to 0.78

The **Logistic Regression** model was included for reference against the various Decision Tree and additional Ensemble Models and **scored substantially lower than all other models**

```
Cross-Validation Performance:

Logistic Regression: 46.21245421245422
Decision Tree: 67.51909994767138
Bagging Classifier: 67.62375719518576
Random Forest: 66.18733647305076
Gradient Boost Classifier: 73.15384615384615
AdaBoost Classifier: 73.36002093144951
XG Boost Classifier: 76.2276295133438

Validation Performance:

Logistic Regression: 0.5337423312883436
Decision Tree: 0.6840490797546013
Bagging Classifier: 0.696319018404908
Random Forest: 0.7177914110429447
Gradient Boost Classifier: 0.7760736196319018
AdaBoost Classifier: 0.7760736196319018
XG Boost Classifier: 0.7760736196319018
```

All Model Scores – Training Summary

Training Performance Results - Logistic Regression:

	Logistic Regression	Logistic Regression with OverSampling	Logistic Regression with UnderSampling
Accuracy	0.887	0.830	0.815
Recall	0.468	0.832	0.812
Precision	0.734	0.829	0.816
F1	0.572	0.830	0.814

Training Performance Results - Top 3 Original & Tuned:

	Untuned Gradient Boost Classifier	Untuned Adaptive Boost Classifier	Untuned XG Boost Classifier	Gradient Boost Tuned with Random Search	Adaptive Boost Tuned with Random Search	XG Boost Tuned with Random Search
Accuracy	0.949	0.939	1.000	0.962	0.939	0.954
Recall	0.777	0.766	0.999	0.832	0.766	0.990
Precision	0.895	0.838	0.999	0.927	0.838	0.782
F1	0.832	0.800	0.999	0.877	0.800	0.874

Training Performance Results - Top 3 OverSampled & UnderSampled:

	Gradient Boost Tuned - OverSampled	Adaptive Boost Tuned - OverSampled	XG Boost Tuned - OverSampled	Gradient Boost Tuned - UnderSampled	Adaptive Boost Tuned - UnderSampled	XG Boost Tuned - UnderSampled
Accuracy	0.972	0.946	0.947	0.971	0.920	0.907
Recall	0.969	0.950	0.972	0.969	0.929	0.908
Precision	0.975	0.942	0.926	0.972	0.912	0.906
F1	0.972	0.946	0.949	0.971	0.920	0.907

All Model Scores – Validation Summary

Validation Performance Results - Logistic Regression:

	Logistic Regression	Logistic Regression with OverSampling	Logistic Regression with UnderSampling
Accuracy	0.894	0.822	0.813
Recall	0.534	0.810	0.822
Precision	0.737	0.470	0.455
F1	0.619	0.595	0.586

Validation Performance Results - Top 3 Original & Tuned:

	Untuned Gradient Boost Classifier	Untuned Adaptive Boost Classifier	Untuned XG Boost Classifier	Gradient Boost Tuned with Random Search	Adaptive Boost Tuned with Random Search	XG Boost Tuned with Random Search
Accuracy	0.947	0.937	0.944	0.948	0.948	0.928
Recall	0.776	0.776	0.776	0.791	0.791	0.902
Precision	0.878	0.824	0.863	0.872	0.872	0.722
F1	0.824	0.799	0.817	0.830	0.830	0.802

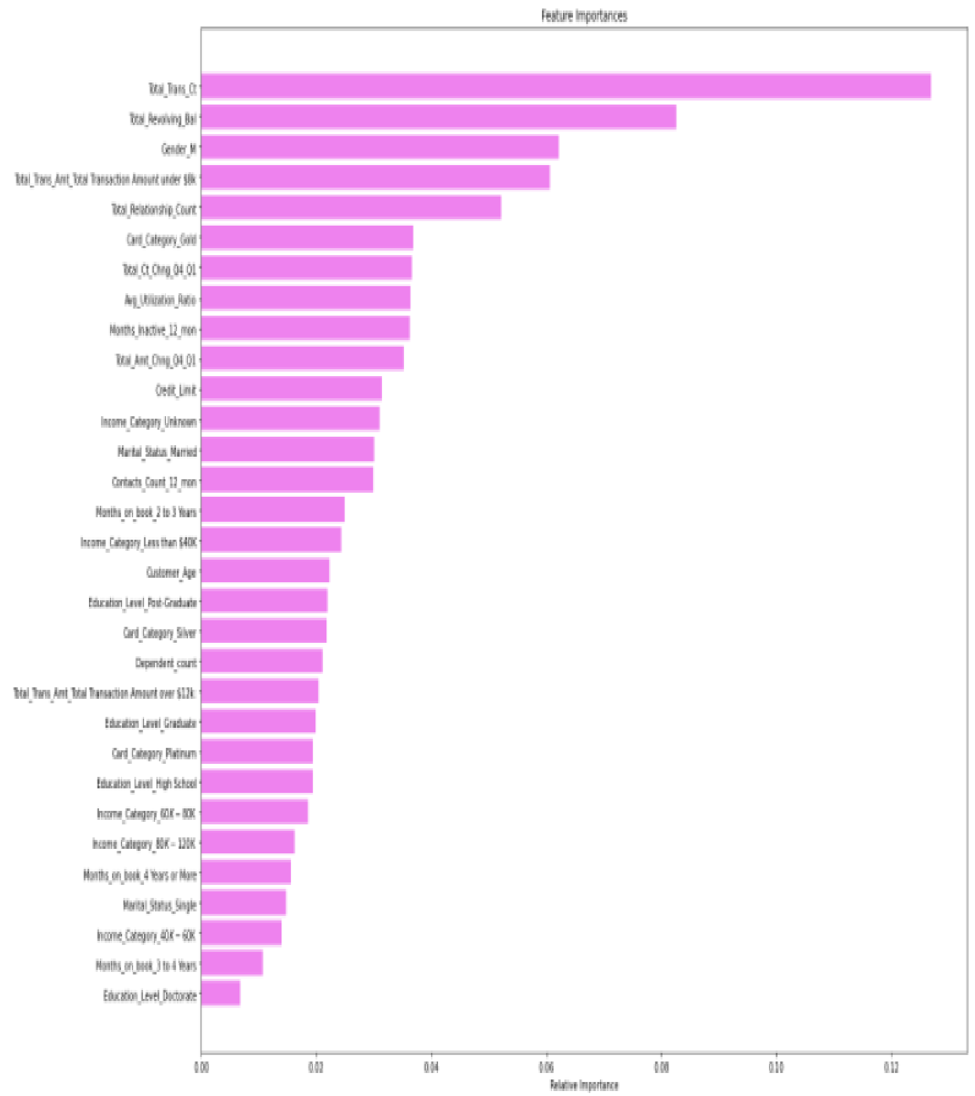
Validation Performance Results - Top 3 OverSampled & UnderSampled:

	Gradient Boost Tuned - OverSampled	Adaptive Boost Tuned - OverSampled	XG Boost Tuned - OverSampled	Gradient Boost Tuned - UnderSampled	Adaptive Boost Tuned - UnderSampled	XG Boost Tuned - UnderSampled
Accuracy	0.942	0.922	0.920	0.917	0.891	0.905
Recall	0.831	0.810	0.871	0.908	0.920	0.911
Precision	0.814	0.733	0.701	0.680	0.606	0.646
F1	0.822	0.770	0.777	0.778	0.731	0.756

Final Model Selected: XG Boost (Tuned)

```
Test Performance:
```

	Accuracy	Recall	Precision	F1
0	0.936821	0.950769	0.733967	0.828418



Observations

- The Model's **results are well Generalized across all tests**, and scored very high in **Recall (0.95)**
- **Total Transaction Count** is by far the most important indicator in determining likelihood of Customer Attrition regarding their Credit Cards
  - This is indicative of customer activity **- the higher the transaction counts, the more often Customers are continuing to use their Credit Cards**
- This is followed by **Total Revolving Balance which shows that customers with higher balances are more committed to the bank** and to paying back the debts owed, while likely also spending more on their Credit Cards

Insights

- The lower one's Credit Limit, etc., the higher the Credit Utilization Ratio could reach each month, in general
- Both the Total Amount and Counts ratios transactions between Q4 and Q1 indicate a strong chance of Customers with Higher Q4 spending (between 1 and 3 times Q1 levels) maintaining their Credit Cards due to **higher seasonal spending**
- Similarly, customers with **Total Transaction Counts greater than around 95 a Year are far more likely to stay active** with their Credit Card service
- Customers with 2 or more dependents have a greater chance of attrition vs. customers with 1 or fewer dependents
- Those customers with **higher months of inactive usage over a 1 year period (2 or more months) have a much higher chance of cancelling their Credit Card** services with the bank
- Customers with higher contacts within a 1 year period are more likely to close their Credit Card with the bank
- The **higher the Revolving Credit Balance, the more likely the customer will maintain their Credit Card service** with the bank
- Customers with **higher Credit Limits are more likely to stay active**
- **The higher the average Credit Utilization Ratio (greater than 20%)**, the more likely customers are to keep their Credit Cards

Recommendations

- In order to target customers most at risk of attrition, the bank should focus on targeting the following individuals:

    - Customers who have been **Inactive** for the last couple months or more
    - Customers with **little to no Revolving Balance** and a large (in relation to Credit Limit) **Average Open to Buy** balance that is largely unused
    - Customers with lower **Transaction Counts under 50 a year** and lower **Average Utilization Ratios under 25%** since they are choosing to not utilize their current credit offerings for some reason or another
    - Customers with **multiple Relationships (Products) with the bank** as those with more Credit options available could easily switch or close their Credit Cards for another option or competitor
    - **Customers Spending under $8k a Year on Average** as they show less likelihood of continuing to use their Credit Cards and remain committed to the bank than those customers spending higher amounts in excess of $12k
    - **Customers with lower Q4 vs. Q1 Spending Ratios** as this indicates that during the holiday season (Q4) when spending increases on average, these customers are choosing other Credit options than what we've provided them