

ENSEMBLE TECHNIQUES AND MODEL TUNING

Project 3 Churn Prediction

AIML UT Austin McCombs School of Business

AUGUST 18, 2023

Presented by Greg Wenzel

Executive
Summary

Business Problem
Overview and
Solution
Approach

Exploratory
Data Analysis

Data
Preprocessing

Performance
Summary
Hyperparameter
Tuning.

Appendix

TABLE OF CONTENTS

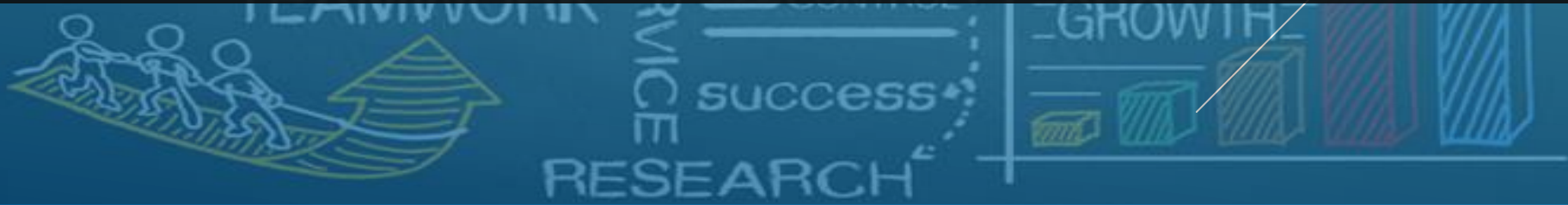


EXECUTIVE SUMMARY



In today's competitive financial landscape, retaining customers is paramount for the sustained growth of any bank. Thera Bank has encountered a concerning decline in credit card users, leading to potential revenue loss. This decline could be attributed to factors such as inadequate service quality or unaddressed customer concerns.

As a solution, we propose the implementation of a classification model to predict customer attrition, thus enabling Thera Bank to proactively address customer needs, enhance service offerings, and ultimately prevent credit card churn.




BUSINESS PROBLEM

- Thera Bank is facing a significant challenge due to the declining number of credit card users. This *decline* not only *affects revenue* but also raises concerns about customer satisfaction and loyalty. Identifying the reasons behind customer attrition is critical for devising targeted strategies to retain existing customers.
- From a data science perspective, the approach involves building a classification model to predict whether a customer will contribute to 'attrition' or continue using the credit card services as an 'existing customer'. By analyzing historical customer data, including demographic information, transaction behavior, credit card usage patterns, and other relevant attributes, we can uncover patterns and insights that contribute to customer 'churn', or attrition.

SOLUTION APPROACH

The proposed solution involves the following steps

- Data Preparation and Exploration
 - Feature Engineering
 - Model Selection
 - Hyperparameter Tuning
 - Model Evaluation
 - Interpretability and Insights
 - Deployment and Monitoring
- 
- A series of white diagonal lines of varying lengths and thicknesses, located in the bottom right corner of the slide, creating a modern, abstract graphic element.

SOLUTION APPROACH – CONTINUED

Data Preparation and Exploration

The provided data will be preprocessed, handling missing values and ensuring consistency. Exploratory data analysis (EDA) will be conducted to gain insights into customer attributes and their potential impact on attrition.

Feature Engineering

We will engineer new features if necessary and transform existing ones to improve the model's predictive power. For instance, we will calculate the credit utilization ratio and utilize it as an informative feature.

Model Selection

Several classification algorithms will be considered, including the AdaBoost classifier mentioned in the problem statement. We will evaluate different algorithms' performance using appropriate metrics and cross-validation techniques.

SOLUTION APPROACH – CONTINUED

Hyperparameter Tuning

We will perform hyperparameter tuning using techniques like RandomizedSearchCV to find the best set of hyperparameters for the chosen model. This step aims to enhance the model's generalization performance.

Model Evaluation

The model's performance will be evaluated using various metrics such as recall, precision, and F1-score, with a focus on recall due to the nature of the problem. The goal is to identify potential churners with high accuracy.

Interpretability and Insights

The model's performance will be evaluated using various metrics such as recall, precision, and F1-score, with a focus on recall due to the nature of the problem. The goal is to identify potential churners with high accuracy.


Deployment and Monitoring

Once the model is finalized, it can be deployed to predict customer attrition in real-time. Regular monitoring and periodic updates will be necessary to ensure the model's continued accuracy as customer behaviors evolve.

SOLUTION APPROACH

Conclusion

The proposed classification model aims to provide Thera Bank with a powerful tool to predict customer attrition, enabling the bank to make informed decisions and take proactive measures to retain customers. By addressing the root causes of attrition and improving service quality, Thera Bank can foster customer loyalty, enhance its financial performance, and maintain a competitive edge in the market.

Three white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, extending from the right edge towards the center.

EXPLORATORY DATA ANALYSIS

Data Summary

Rows : 10127

Columns/Features (Beginning) : 21

Columns/Features (Final) : 29

Duplicate Values : 0

Null Values : 'Education' : 1519 , 'Marital_Status' : 749

Mean Customer Age : 46

Average Number of Dependents : 2.4

Average Number of Products Held by Customer : 3.8

Education List : Doctorate, College, Graduate, High School, Post-Graduate, Uneducated

Income Categories : < \$40K, \$40K-\$60K, \$60K-\$80K, \$80K-\$120K, \$120K+

Mean Age of Account : 36 Months

Mean Credit Limit : \$8600

Mean Revolving Balance : \$1161

[Link to Appendix slide on data background check](#)

EXPLORATORY DATA ANALYSIS

Data Summary - Continued

Gender : Female = 5358(53%) , Male = 4769(47%)

Marital Status : Null = 7.4%, Divorced = 7.4%, Married = 46%, Single = 39%

Total Attrition (Non-Card Holders): 1627 = 19%

Attrition by Gender : Female = 57.1% , Male = 42.8%

Attrition/Marital Status : Married = 47% , Single = 45% , Divorced = 8%

Attrition/Income*: <\$40K=37%, \$40-60K=17%, \$60-80K=12%, \$80-120K=14%, \$120<8%, Null=12%

Attrition/Edu: Doc=7% , Post=7%, No Edu= 17%, Grad=36% , College=11% , High School=22%

[Link to Appendix slide on data background check](#)

DATA PREPROCESSING

Duplicate value check: No duplicate values present

Missing value treatment: Three features require attention. Education_Level, Marital_Status, and 'Income_Category' features have string values that were converted to 'nan'.

Outlier check: The feature with the highest percentage of outliers is 'Attrition_Flag' which is not a concern as it will be the 'target' variable. There are other outliers within the data set however, after reviewing and evaluating the data no further process will be implemented.

Feature engineering – The following categories were assigned to the related features:

Education: Graduate, HighSchool, Uneducated, College, Post-Graduate, Doctorate

Gender: F – Female , M - Male

Marital Status: Single, Divorced, Married

Card_Category: Blue, Silver, Gold, Platinum

Income_Category: <\$40k, \$40-60K, \$60-80K, \$80-120K, \$120K<

Data preparation for modeling – Categorical values were encoded resulting in 29 columns. Missing values were imputed. Training, Validation, and Testing sets were created.

MODEL PERFORMANCE SUMMARY

Initial Training and Validation

Original Data

Training Performance:

Bagging: 0.9845679012345679
Random forest: 1.0
Adaptive Boost: 0.8526234567901234
Gradient Boost: 0.8904320987654321
XGBoost: 1.0

Validation Performance:

Bagging: 0.8444444444444444
Random forest: 0.8469135802469135
Adaptive Boost: 0.8222222222222222
Gradient Boost: 0.8493827160493828
XGBoost: 0.9185185185185185

Undersampled Data

Training Performance:

Bagging: 0.9938271604938271
Random forest: 1.0
Adaptive Boost: 0.9567901234567902
Gradient Boost: 0.9807098765432098
XGBoost: 1.0

Validation Performance:

Bagging: 0.9407407407407408
Random forest: 0.9679012345679012
Adaptive Boost: 0.9555555555555556
Gradient Boost: 0.9679012345679012
XGBoost: 0.9703703703703703

Oversampled Data

Training Performance:

Bagging: 0.9977957384276267
Random forest: 1.0
Adaptive Boost: 0.97060984570169
Gradient Boost: 0.9822189566495224
XGBoost: 1.0

Validation Performance:

Bagging: 0.8641975308641975
Random forest: 0.8938271604938272
Adaptive Boost: 0.8641975308641975
Gradient Boost: 0.9135802469135802
XGBoost: 0.928395061728395

[Link to Summary of Training and Validation Results](#)

MODEL PERFORMANCE SUMMARY

Adaptive Boosting Oversampled: Tuning, Training, Validating & Testing

Training performance comparison:

	Adaptive boosting with Undersampled Training	Adaptive boosting with Oversampled Training	Adaptive boosting with Original Training	Gradient boosting with Undersampled Training	Gradient boosting with Original Training
Accuracy	0.990	0.991	0.960	0.970	0.989
Recall	0.956	0.962	1.000	0.978	0.951
Precision	0.981	0.979	0.798	0.963	0.976
F1	0.968	0.970	0.888	0.971	0.964

Validation performance comparison:

	Adaptive boosting Undersampled Validated	Adaptive boosting Oversampled Validated	Adaptive boosting Original Validated	Gradient boosting Undersampled Validated	Gradient boosting Original Validated
Accuracy	0.975	0.974	0.938	0.938	0.972
Recall	0.899	0.886	0.951	0.963	0.889
Precision	0.941	0.950	0.739	0.734	0.935
F1	0.919	0.917	0.832	0.833	0.911

```
ada_test_over = model_performance_classification_sklearn(tuned_ada_over , X_test , y_test)
ada_test_over
```

	Accuracy	Recall	Precision	F1
0	0.990	0.962	0.978	0.969



Test Data
Adaptive Boost
Oversampled

MODEL PERFORMANCE SUMMARY

Supplement: Testing Gradient Boosting Undersampled

Training performance comparison:

	Adaptive boosting with Undersampled Training	Adaptive boosting with Oversampled Training	Adaptive boosting with Original Training	Gradient boosting with Undersampled Training	Gradient boosting with Original Training
Accuracy	0.990	0.991	0.960	0.970	0.989
Recall	0.956	0.962	1.000	0.978	0.951
Precision	0.981	0.979	0.798	0.963	0.976
F1	0.968	0.970	0.888	0.971	0.964

Validation performance comparison:

	Adaptive boosting Undersampled Validated	Adaptive boosting Oversampled Validated	Adaptive boosting Original Validated	Gradient boosting Undersampled Validated	Gradient boosting Original Validated
Accuracy	0.975	0.974	0.938	0.938	0.972
Recall	0.899	0.886	0.951	0.963	0.889
Precision	0.941	0.950	0.739	0.734	0.935
F1	0.919	0.917	0.832	0.833	0.911

```
gbm1_test = model_performance_classification_sklern(tuned_gbm1 , X_test , y_test)
gbm1_test
```

	Accuracy	Recall	Precision	F1
0	0.944	0.978	0.750	0.849

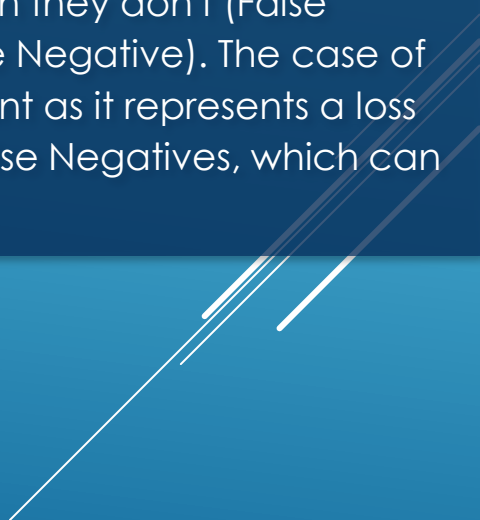


Test Data
Gradient Boosting
Undersampled

MODEL PERFORMANCE SUMMARY

Conclusion:

Considering data science, the primary focus is to develop a model that effectively predicts customer attrition (churn) for the bank. The key consideration in evaluating the model is the trade-off between making two types of wrong predictions: predicting a customer will attrite when they don't (False Positive) and predicting a customer won't attrite when they actually do (False Negative). The case of predicting that a customer won't attrite but they do is deemed more important as it represents a loss of valuable customers or assets. Therefore, the bank's priority is to minimize False Negatives, which can be achieved by maximizing the recall.

Three white lines of varying lengths and slopes are positioned in the bottom right corner of the slide, extending from the right edge towards the center.

MODEL PERFORMANCE SUMMARY

Conclusion Continued:

The given output presents a comparison of different model performances based on various evaluation metrics for both training and validation datasets. The models used include Adaptive Boosting with Undersampled Training, Adaptive Boosting with Oversampled Training, Adaptive Boosting with Original Training, Gradient Boosting with Undersampled Training, Gradient Boosting with Original Training, XGBoost with Original Training, and XGBoost with Undersampled Training.

From the output, it's evident that both Adaptive Boosting with Undersampled Training and Gradient Boosting with Oversampled Training have strong performance in terms of recall on the test data. Recall is a crucial metric for the bank's needs, as it measures the ability to correctly identify customers who are at risk of attrition (True Positives) out of all the actual attrition cases (True Positives + False Negatives).

MODEL PERFORMANCE SUMMARY

Conclusion Continued - Based on the recall scores for the test data:

Adaptive Boosting Undersampled Test Data:

Accuracy: 0.990
Recall: 0.956
Precision: 0.980
F1: 0.968

Gradient Boosting Oversampled Test Data:

Accuracy: 0.944
Recall: 0.978
Precision: 0.750
F1: 0.849

Both models achieve high recall scores, indicating a strong ability to identify customers at risk of attrition. However, the Adaptive Boosting model with undersampled training data performs slightly better in terms of precision, accuracy, and F1-score compared to the Gradient Boosting model with oversampled training data.

Considering the bank's objective to reduce the loss associated with False Negatives (missed attrition cases), the Adaptive Boosting model with undersampled training data seems to be the better choice. It strikes a balance between recall, precision, and overall accuracy while effectively identifying customers at risk of attrition. This would allow the bank to take proactive measures to retain valuable customers who might otherwise be lost if not detected.

MODEL PERFORMANCE SUMMARY

Original, Undersampled, and Oversampled Results

Based on the provided performance metrics for Gradient Boosting Oversampled and Adaptive Boosting Undersampled models, it's evident that both of these approaches are effective for improving recall on the given dataset. Let's explore why these models could be chosen based on the data characteristics and performance results:

Original	Undersampled	Oversampled
Training Performance: Bagging: 0.9845679012345679 Random forest: 1.0 Adaptive Boost: 0.8526234567901234 Gradient Boost: 0.8904320987654321 XGBoost: 1.0	Training Performance: Bagging: 0.9938271604938271 Random forest: 1.0 Adaptive Boost: 0.9567901234567902 Gradient Boost: 0.9807098765432098 XGBoost: 1.0	Training Performance: Bagging: 0.9977957384276267 Random forest: 1.0 Adaptive Boost: 0.97060984570169 Gradient Boost: 0.9822189566495224 XGBoost: 1.0
Validation Performance: Bagging: 0.8444444444444444 Random forest: 0.8469135802469135 Adaptive Boost: 0.8222222222222222 Gradient Boost: 0.8493827160493828 XGBoost: 0.9185185185185185	Validation Performance: Bagging: 0.9407407407407408 Random forest: 0.9679012345679012 Adaptive Boost: 0.9555555555555556 Gradient Boost: 0.9679012345679012 XGBoost: 0.9703703703703703	Validation Performance: Bagging: 0.8641975308641975 Random forest: 0.8938271604938272 Adaptive Boost: 0.8641975308641975 Gradient Boost: 0.9135802469135802 XGBoost: 0.928395061728395

MODEL PERFORMANCE SUMMARY

Original, Undersampled, and Oversampled Results - Continued

1. Gradient Boosting Oversampled:

Recall Improvement: The validation recall for the Gradient Boosting model with oversampled data is consistently higher compared to the original dataset. This indicates that the model is better at identifying the positive class (attrited customers) which is the primary focus of the problem.

Balancing Class Distribution: Oversampling addresses the class imbalance issue by creating duplicate instances of the minority class. This allows the model to learn from a more balanced dataset and capture patterns in the minority class more effectively.

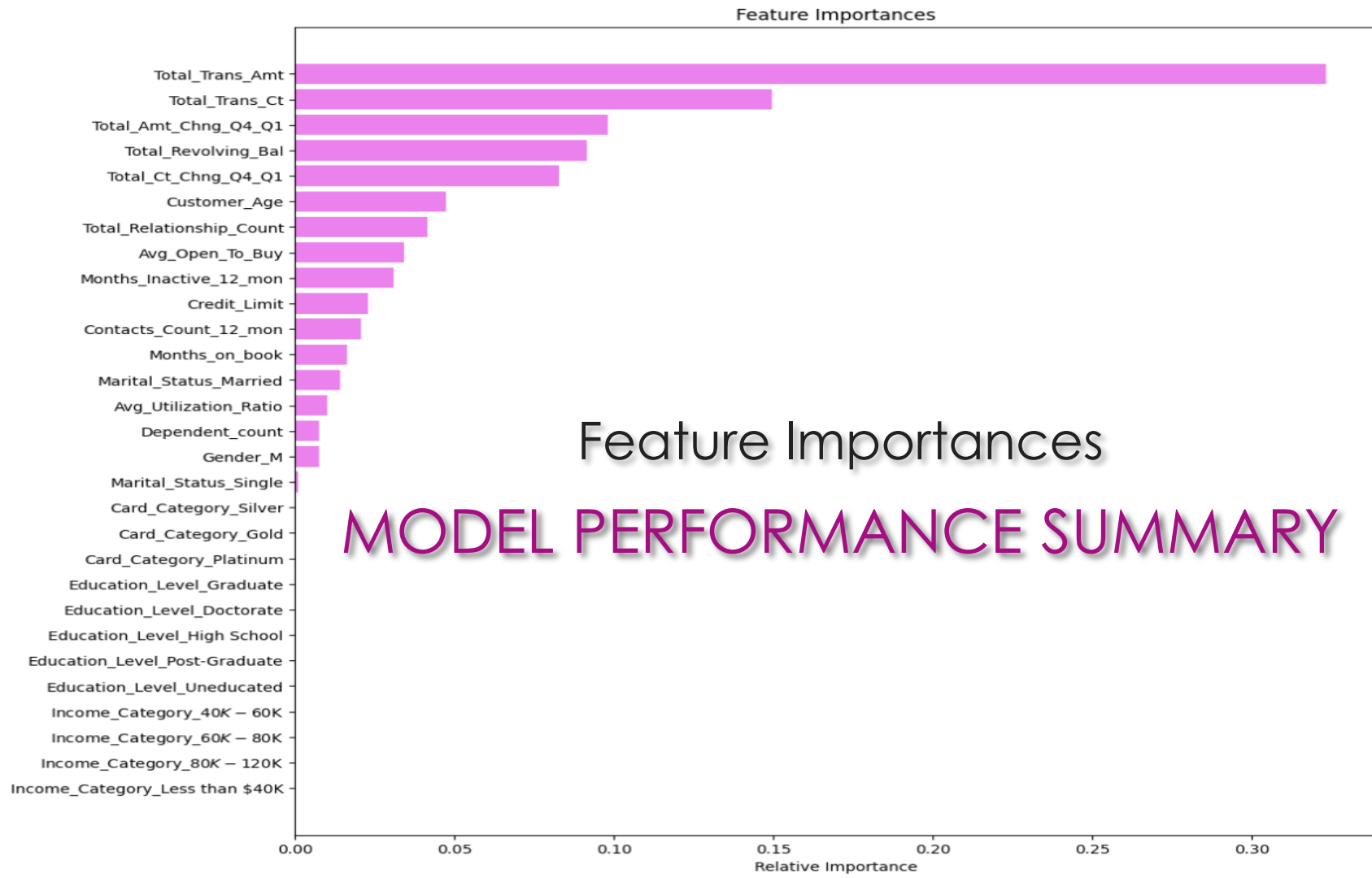
High Training Performance: The model achieves high recall in both training and validation, which suggests it's not heavily overfitting and is generalizing well to new data.

2. Adaptive Boosting Undersampled:

Recall Improvement: The validation recall for the Adaptive Boosting model with undersampled data is also notably higher than the original dataset. This means that the model is better at correctly identifying the attrited customers.

Addressing Class Imbalance: Undersampling reduces the instances of the majority class, allowing the model to focus more on the minority class. This can help prevent the model from being biased towards the majority class.

High Training Performance: Similar to Gradient Boosting, Adaptive Boosting achieves high recall in both training and validation, indicating good generalization.



MODEL PERFORMANCE SUMMARY

Feature Importances: Comparing Final Models

These summaries provide insight into how different features impact the performance of the two boosting algorithms under different sampling techniques. Features like Total_Trans_Amt, Total_Trans_Ct, and Total_Amt_Chng_Q4_Q1 consistently show high importance across both models, while other features may vary in their impact.

Adaptive Boosting (Undersampled)		Gradient Boosting (Oversampled)	
Total_Trans_Amt: 32.32%	Dependent_count: 0.76%	Total_Trans_Amt: 32.56%	Avg_Utilization_Ratio: 1.12%
Total_Trans_Ct: 14.96%	Gender_M: 0.74%	Total_Trans_Ct: 16.28%	Gender_M: 1.00%
Total_Amt_Chng_Q4_Q1: 9.80%	Marital_Status_Single: 0.10%	Total_Revolving_Bal: 10.42%	Months_on_book: 0.84%
Total_Revolving_Bal: 9.14%	Card_Category_Silver: 0.04%	Total_Amt_Chng_Q4_Q1: 8.76%	Card_Category_Silver: 0.52%
Total_Ct_Chng_Q4_Q1: 8.28%	Card_Category_Gold: 0.03%	Total_Ct_Chng_Q4_Q1: 7.92%	Education_Level_Graduate: 0.40%
Customer_Age: 4.74%	Card_Category_Platinum: 0.02%	Customer_Age: 5.14%	Income_Category_\$40K - \$60K: 0.24%
Total_Relationship_Count: 4.16%	Education_Level_Graduate: 0.02%	Total_Relationship_Count: 3.96%	Income_Category_\$80K - \$120K: 0.18%
Avg_Open_To_Buy: 3.41%	Education_Level_Doctorate: 0.00%	Months_Inactive_12_mon: 3.02%	Income_Category_\$60K - \$80K: 0.18%
Months_Inactive_12_mon: 3.09%	Education_Level_High School: 0.00%	Credit_Limit: 2.68%	Card_Category_Gold: 0.16%
Credit_Limit: 2.28%	Education_Level_Post-Graduate: 0.00%	Contacts_Count_12_mon: 2.32%	Card_Category_Platinum: 0.12%
Contacts_Count_12_mon: 2.08%	Education_Level_Uneducated: 0.00%	Avg_Open_To_Buy: 2.08%	Education_Level_Doctorate: 0.10%
Months_on_book: 1.62%	Income_Category_\$40K - \$60K: 0.00%	Marital_Status_Married: 1.64%	Education_Level_High School: 0.08%
Marital_Status_Married: 1.41%	Income_Category_\$60K - \$80K: 0.00%	Dependent_count: 1.46%	Education_Level_Post-Graduate: 0.06%
Avg_Utilization_Ratio: 1.01%	Income_Category_\$80K - \$120K: 0.00%	Marital_Status_Single: 1.34%	Education_Level_Uneducated: 0.04%

APPENDIX

- General Insights and Conclusions

Exploratory Data Analysis

- Headmap
- Revolving Balance Vs. Attrition
- Credit Limit Vs. Attrition
- Transaction Amount Vs. Attrition
- Transaction Count Vs. Attrition
- Transaction Amount Change Vs. Attrition

APPENDIX - I

General Insights and Conclusions

It would be helpful to have further clarification regarding certain features:

Months_Inactive - Is months inactive related to how long they have stopped using the credit card or is how long they didn't use the credit card prior to canceling... or is it how long they have been a customer and now applied for a credit card?

Contacts_Count - Are the same contacts / touches with the 'attrite' customer going to all customers? Did they receive contacts before they had a credit card? Did they receive contact regardless of a credit card... do current customers receive touches... do the non-card customers receive targeted information when being contacted or is it the same contact cycle as card holders?

Card_Category - What are the specifics of each type of card... Income, Credit limit, interest rate, What is the difference between the cards? Also, Has the customer ever had a credit card at the bank? Do they have a credit card with another bank? Are they homeowners? Do they have car loans, boat loans, mortgages, student loans, etc? What about credit scores... it would be helpful to know credit history, or some indicator of creditworthiness so specific customers can be targeted based on their soundness for a credit card.

Total_Amt_Chng_Q4_Q1 - Does this feature represent the total dollar amount change from quarter to quarter, the total change from the beginning number to the ending number, is it evaluated on a monthly basis and then reported quarterly, etc.?

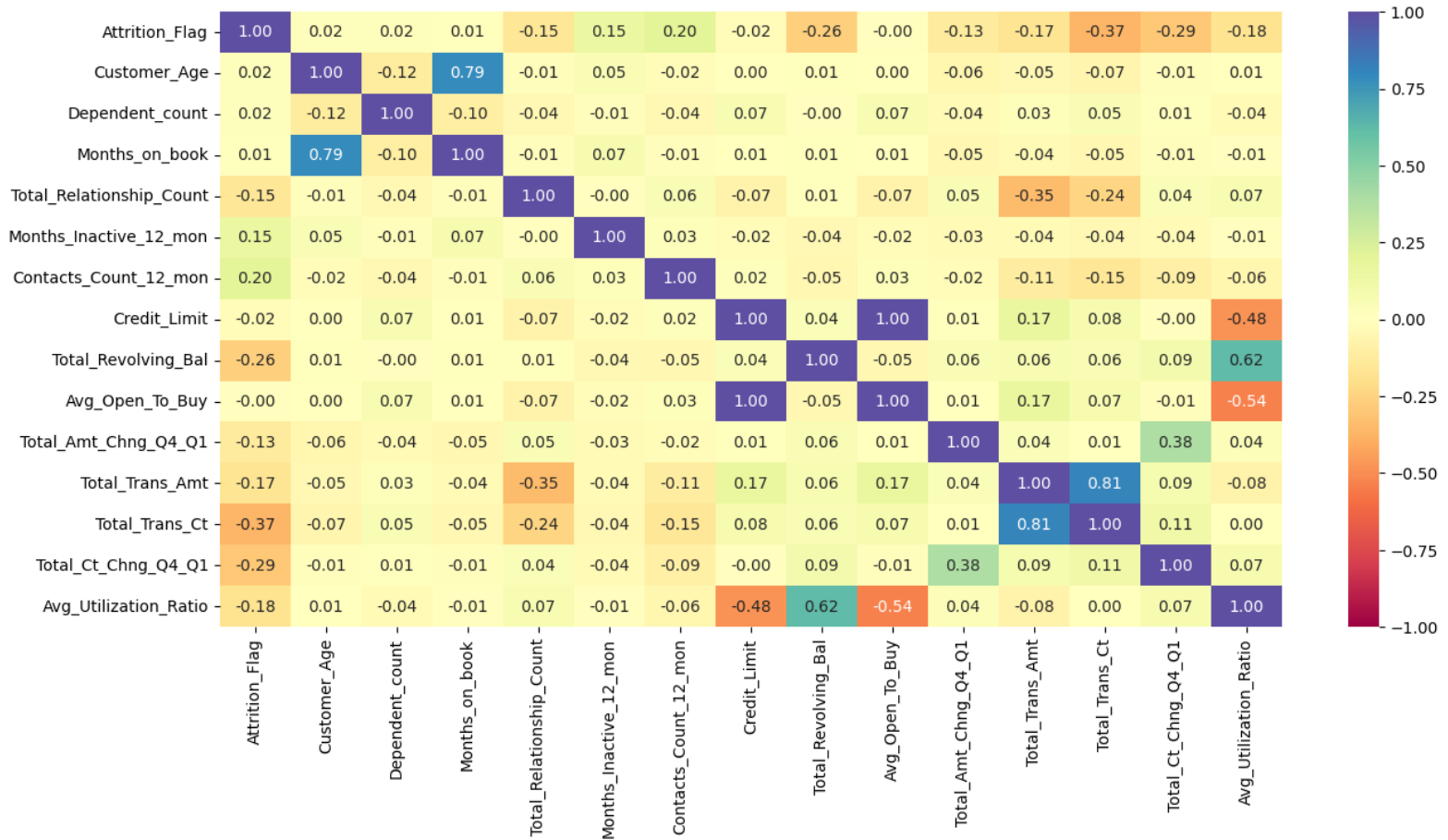
Adaptive Boosting Undersampled and Gradient Boosting Oversampled indicate to be good fits for the data set. It is important to note that there may be slight overfitting using either model. It is important to continue to monitor the ongoing results with future data to avoid data leakage and maintain accurate results.

Outliers were random and worth monitoring for certain features. Using the chosen models helps to minimize outlier influence and it is important to continue to monitor and or retrain future data sets related to different models.

APPENDIX - II

Exploratory Data Analysis

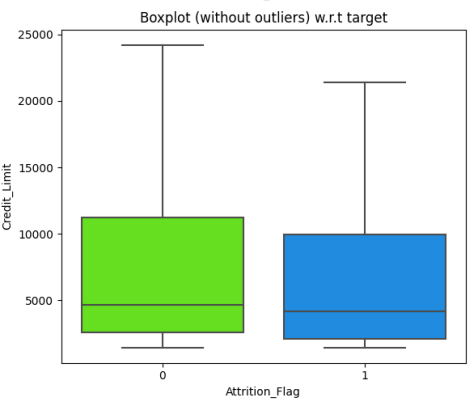
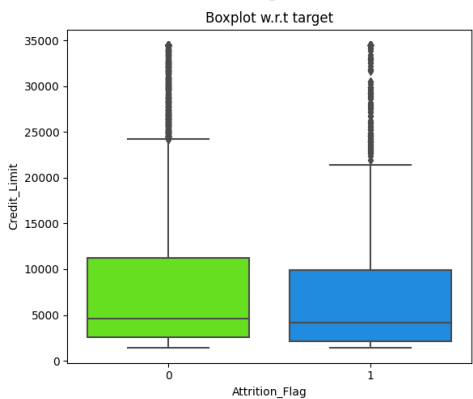
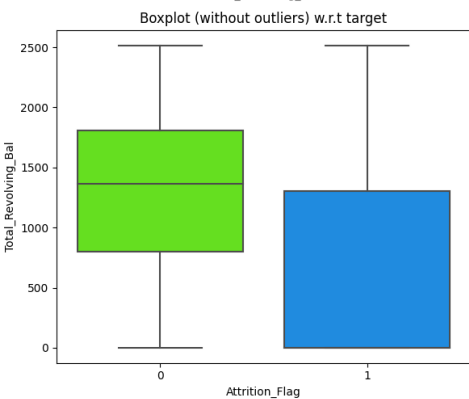
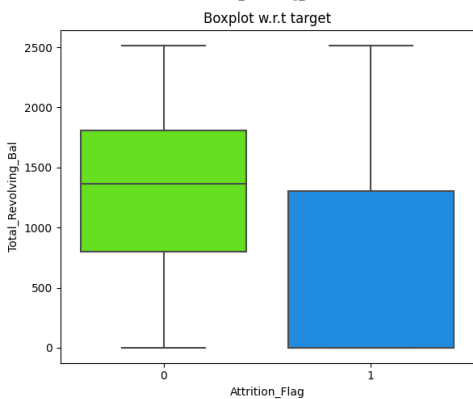
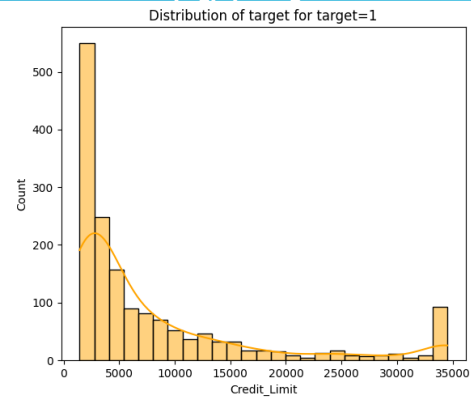
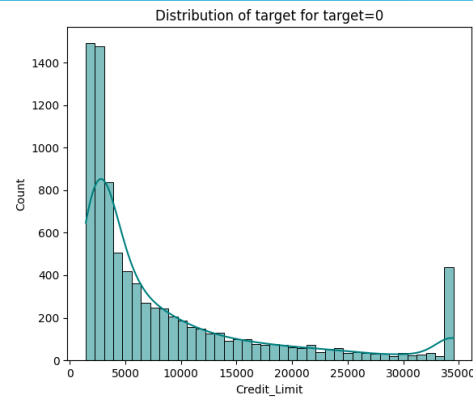
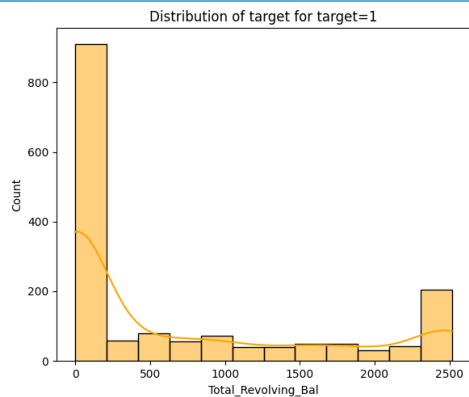
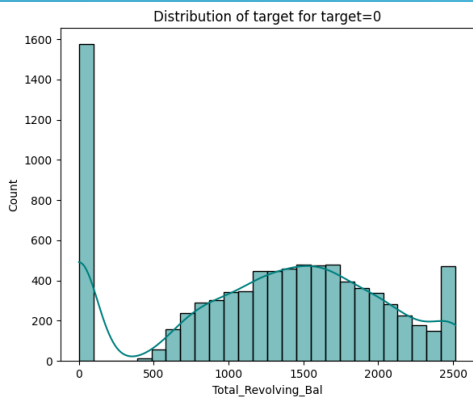
HeatMap



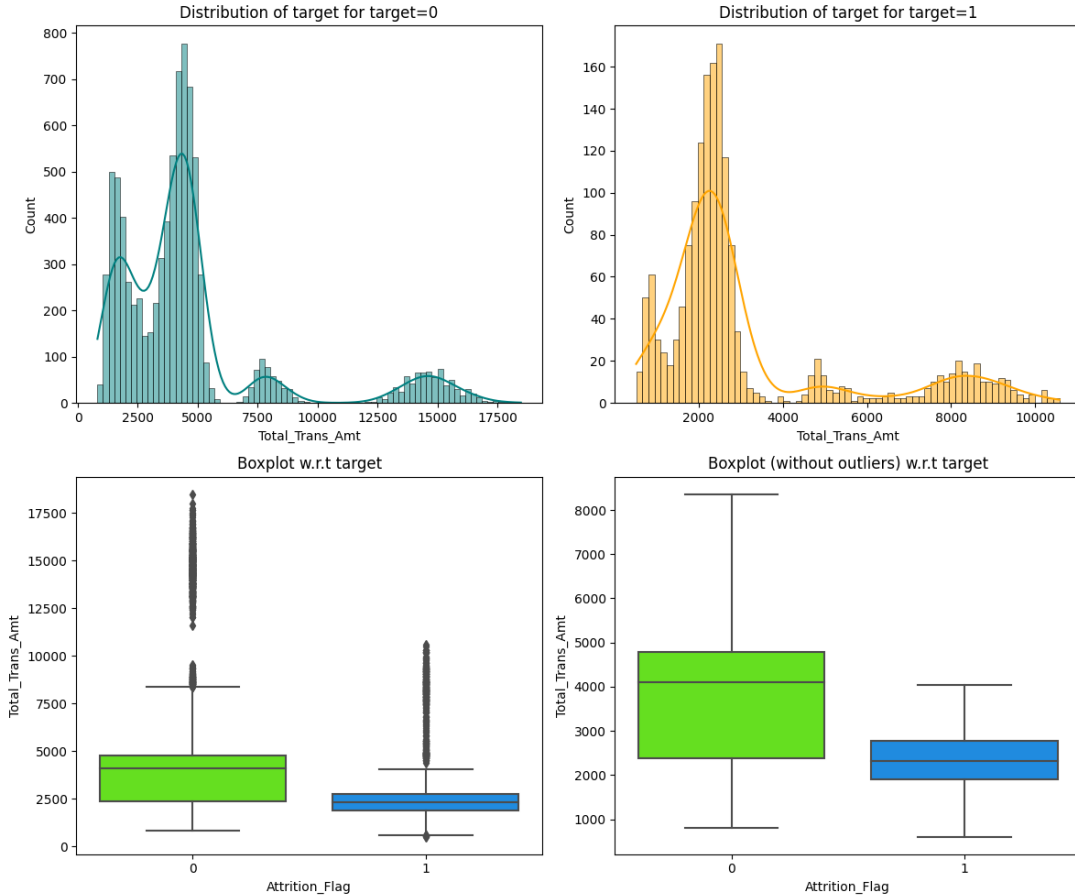
APPENDIX - III Exploratory Data Analysis

Revolving Balance Vs. Attrition

Credit Limit Vs. Attrition



APPENDIX - IV Exploratory Data Analysis

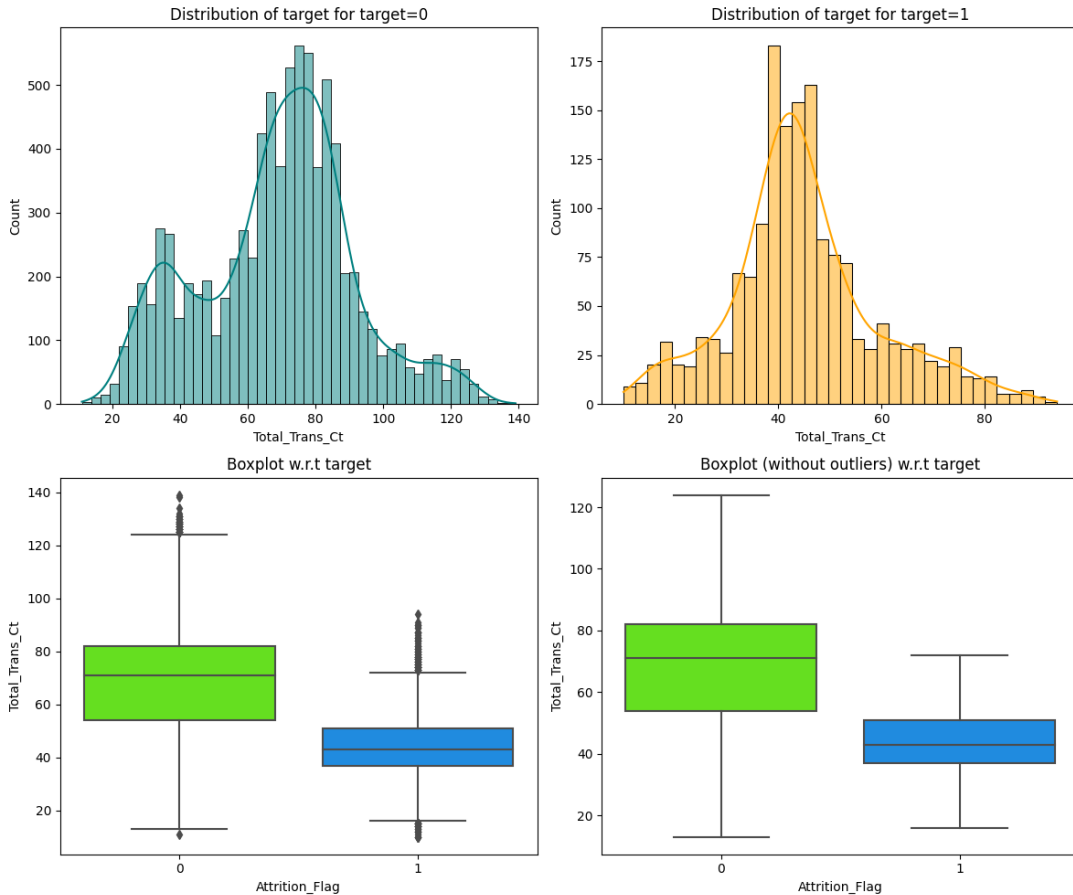


Transaction Amount
Vs. Attrition

Total_Trans_Amount was the most
influential feature of the model.

APPENDIX - V

Exploratory Data Analysis



Transaction Counts
Vs. Attrition

Total_Trans_Ct was the second most
influential feature of the model.

APPENDIX - VI Exploratory Data Analysis

