







Data Insights into Credit Card Service Retention: Thera Bank's Proactive Model

Project : Credit Card Users Churn Prediction
Course: AIML
Date: 05/18/2024

Prosjoma Vivore
mot on firrennn keemae
woianer

Contents

	01	Executive Summary
	02	Business Problem Overview and Solution Approach
	03	EDA Results
	04	Data Preprocessing
	05	Model performance summary for hyperparameter tuning.
	06	Appendix

Executive Summary

Enhancing Customer Retention through Predictive Modeling



Introduction:

Thera Bank has launched a strategic initiative to leverage advanced data analytics for enhancing customer retention amid a notable decline in credit card utilization. This executive summary outlines the methodology behind developing a predictive model to adeptly identify potential customer churn, providing actionable insights to effectively address this challenge. Through comprehensive analysis of customer data and rigorous evaluation of model efficacy, the initiative aims to not only reverse the downturn in credit card usage but also to strengthen customer relationships and stabilize revenue streams derived from credit-related fees.

Executive Summary Business Implications

Key Findings

Customer Engagement and Retention: Retention is strongly linked to engagement metrics such as transaction count and revolving balance. Thera Bank should enhance interactions via credit card services to reduce churn.

High-Impact Features: Critical features like transaction amount, transaction count, revolving balance, relationship count, and transaction amount are key to predicting customer loyalty and can help pinpoint at-risk customers for timely interventions.

Predictive Model Utilization: Leveraging predictive models known for its high recall, is essential for effectively identifying potential churn, supporting proactive retention efforts.

Executive Summary Key Recommendations

Enhance Customer Engagement

1. **Transactional Bonuses:** Implement rewards programs that incentivize frequent transactions or higher spending, thus encouraging continual engagement with the card
2. **Engagement Programs:** Develop targeted engagement initiatives for customers showing a decline in transaction activity or revolving balance, as these are key indicators of potential churn

Product Cross-Selling

1. **Personalized Offers:** Utilize insights from the total relationship count to offer personalized product bundles or services, increasing the number of bank touchpoints per customer
2. **Loyalty Programs:** Enhance loyalty programs to reward customers with multiple product holdings, which could solidify their relationship with the bank

Financial Health Monitoring

1. **Credit Management Tools:** Provide tools and advisories that help customers manage their credit effectively, especially those with high revolving balances, to prevent financial stress that could lead to account closure.
2. **Early Warning Systems:** Develop predictive models to identify early signs of financial distress based on changes in transaction behavior and revolving balances

Executive Summary Key Recommendations

Customer Retention Strategies

1. **Proactive Retention Campaigns:** Implement retention campaigns that proactively address the needs of customers predicted to churn, offering tailored solutions based on their transaction behaviors and product usage
2. **Customer Satisfaction Surveys:** Regularly engage with customers through satisfaction surveys, especially those who show signs of decreased engagement, to identify and address grievances promptly

Risk Management

1. **Adjust Credit Limits:** Use insights from revolving balances and credit utilization to adjust credit limits responsibly, ensuring customers have enough credit for their needs without increasing risk excessively.
2. **Segment-Based Risk Analysis:** Analyze transaction and revolving balance data to segment customers by risk level, tailoring credit offers and terms to balance customer needs with bank risk policies

Executive Summary

Conclusion :

The implementation of the AdaBoost model trained on original data, coupled with strategic initiatives targeting identified risk factors for customer attrition, is recommended to enhance customer retention and satisfaction. This approach balances the need to detect potential attritors with the operational efficiency of managing false positives, ensuring that resources are utilized effectively to maintain and grow the customer base



Business Problem Overview and Solution Approach

Problem Statement

The core challenge is identifying the key factors contributing to customer churn related to credit card services and developing an effective strategy to predict and mitigate this churn. Addressing this issue is crucial for maintaining the bank's profitability and improving customer satisfaction



Goal: Transforming Customer Interaction to Enhance Loyalty and Revenue



Solution Approach



Data Analysis



Model Development



Implement Solutions



Strategic Execution

Solution Approach

Data Analysis

- Comprehensive Data Collection
- Pattern Recognition

Model Development

- Predictive Modeling
- Model Refinement and Validation

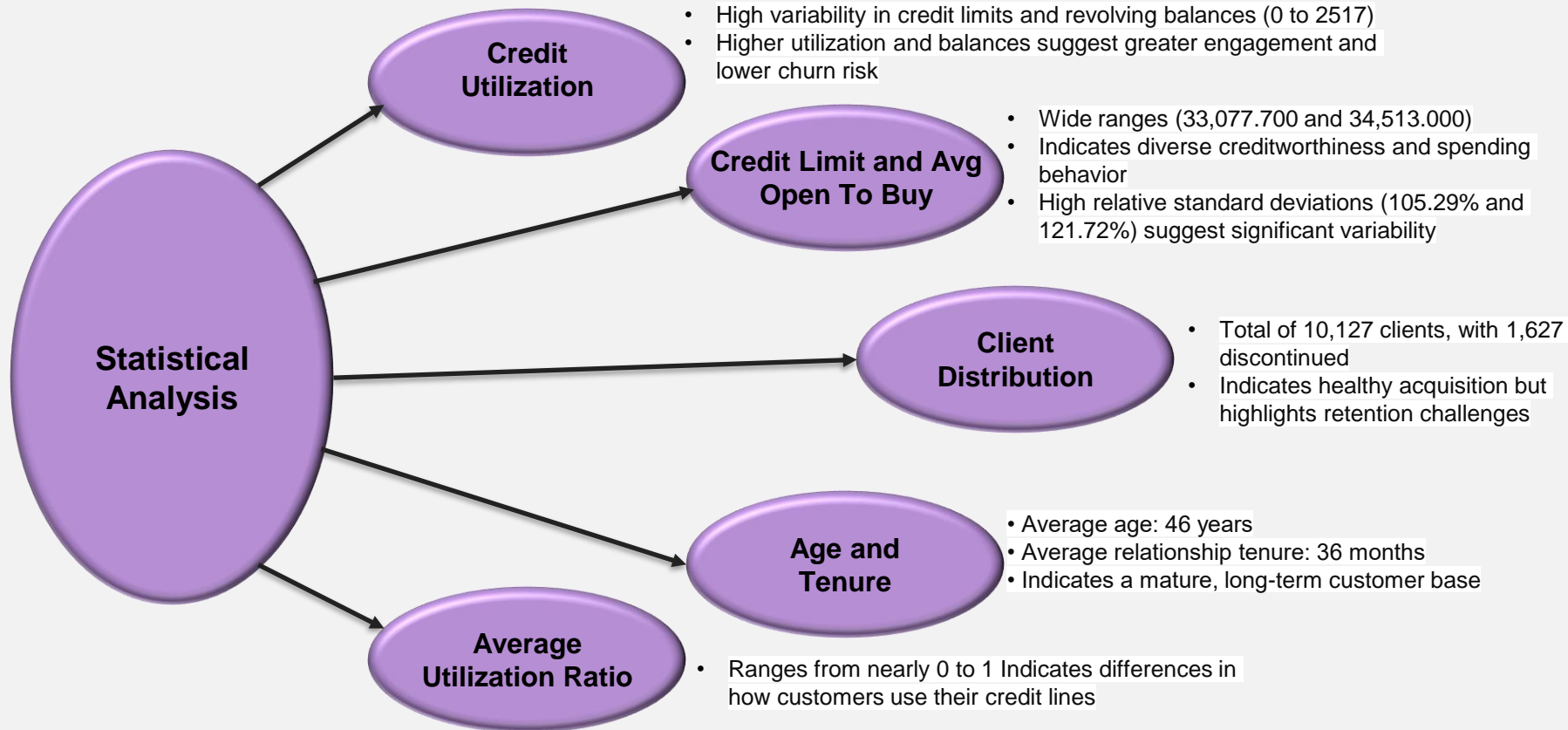
Implement Solutions

- Targeted Interventions
- Operational Integration

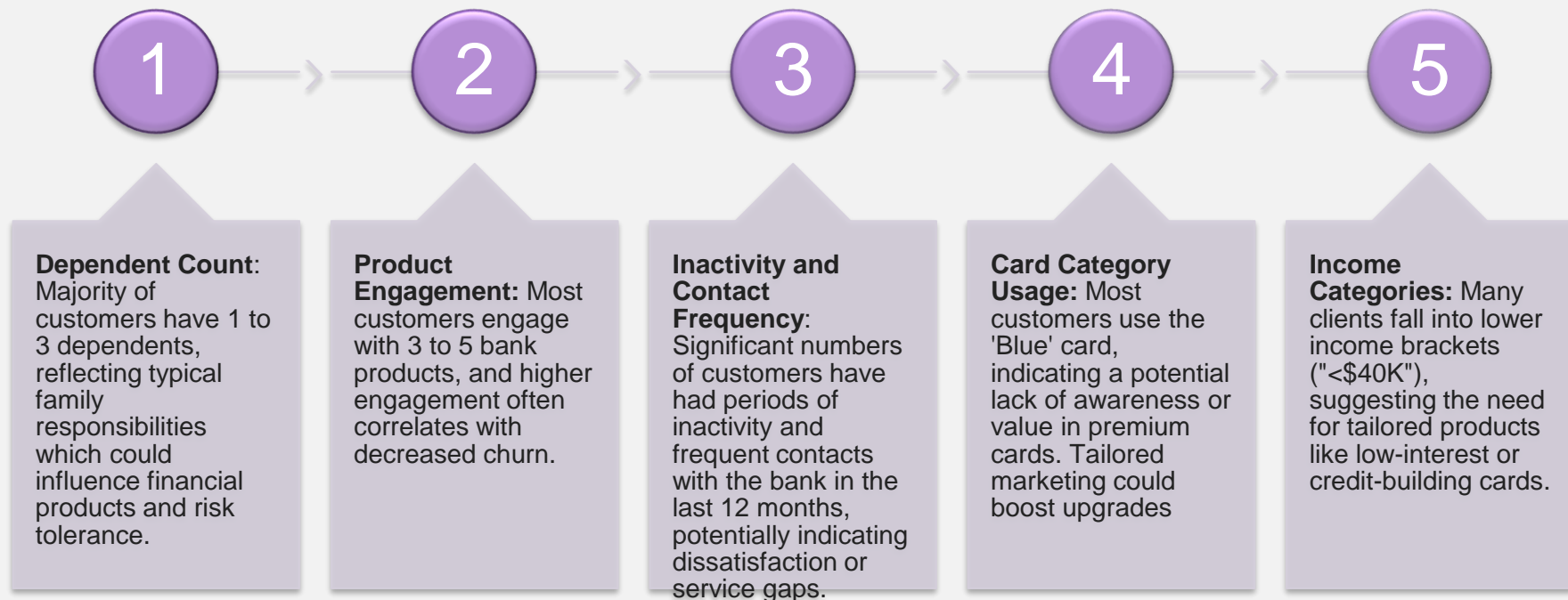
Strategic Execution

- Phased Roll-Out
- Feedback Loop

EDA Results Statistical Analysis



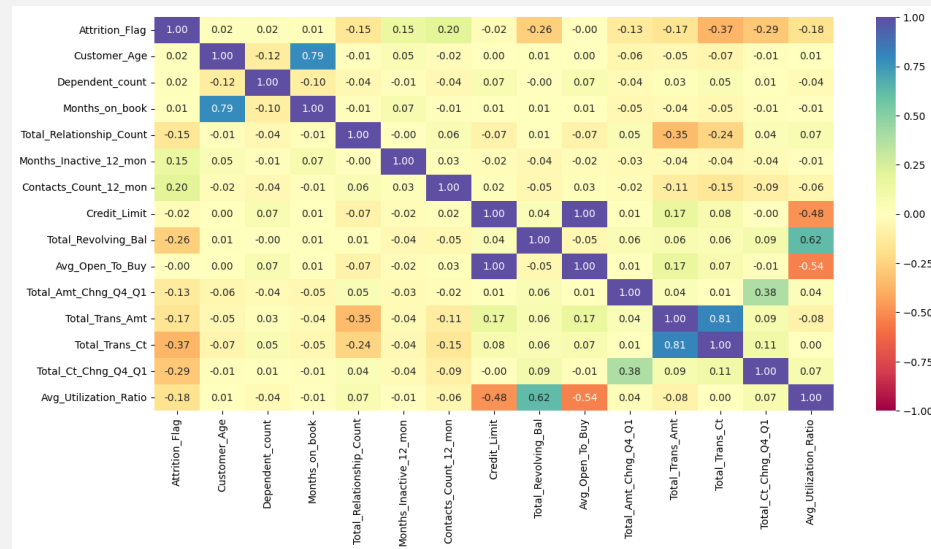
EDA Results Univariate Analysis



*Refer slides 21- 23 in
Appendix for the graphs*

EDA Results of Bivariate Analysis

- Attrition and Transactions:** There is a strong negative correlation of -0.37 between attrition and the total transaction count and amount. Lower engagement in these areas is a strong indicator of potential churn.
- Credit Management:** A positive correlation of 0.62 between total revolving balance and average utilization ratio suggests that higher credit engagement reduces the likelihood of account closure.
- Customer Contacts and Satisfaction:** A positive correlation of 0.2 between the number of contacts and attrition highlights that increased customer service interactions often relate to issues or dissatisfaction, leading to higher churn rates.
- Demographic Influence:** No significant differences in attrition rates among genders, but marital status and education level show varying effects on loyalty, with married and highly educated customers exhibiting lower attrition rates.



[Refer slides 24- 28 in Appendix for the graphs](#)

Data Preprocessing

Data Overview

- Columns (10127)
- Rows(21)



- Dropped CLIENTNUM column as non-informative
- Detected and retained outliers indicating natural variation

- Identified missing entries in
- Education Level (1519)
- Marital Status (749)

- Encoded categorical variables for model compatibility

- Training set: (8101, 19)
- Validation set: (507, 19)
- Test set: (1519, 19)

Model Performance Summary

Model	Data Type	Training Recall	Validation Recall	Model Observation
Gradient Boosting	Original	0.894	0.827	Generalize
Gradient Boosting	Oversampled	0.978	0.864	Overfit
Gradient Boosting	Undersampled	0.978	0.938	Generalize
AdaBoost	Original	0.860	0.815	Generalize
AdaBoost	Oversampled	0.968	0.864	Overfit
AdaBoost	Undersampled	0.963	0.938	Generalize
XGBoost	Original	1.000	0.877	Overfit
XGBoost	Oversampled	1.000	0.901	Overfit
XGBoost	Undersampled	1.000	0.951	Overfit

The models that generalize were chosen for tuning using Randomized Search.

SMOTE technique was used for oversampling
Random Under Sampler was used for under sampling

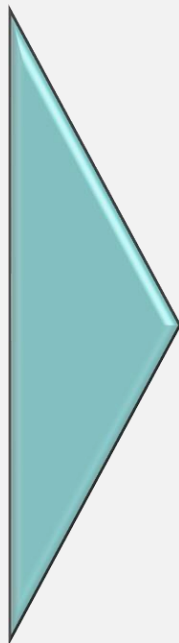
[Refer slides 30-31 in Appendix for more details](#)

Model Performance Post Tuning

Model	Train Accuracy	Train Recall	Train Precision	Train F1	Validation Accuracy	Validation Recall	Validation Precision	Validation F1	Selection
Gradient Boosting (Under)	0.977	0.980	0.974	0.977	0.945	0.938	0.768	0.844	No
Gradient Boosting (Original)	0.975	0.886	0.958	0.921	0.970	0.815	1.000	0.898	No
Gradient Boosting (Over)	0.926	0.860	0.991	0.921	0.970	0.815	1.000	0.898	No
AdaBoost (Original)	0.983	0.925	0.965	0.945	0.982	0.901	0.986	0.942	First Choice
AdaBoost (Under)	0.991	0.991	0.992	0.991	0.941	0.901	0.768	0.830	Second Choice
XGBoost (Original)	0.982	0.995	0.906	0.948	0.947	0.914	0.787	0.846	Third Choice
XGBoost (Over)	0.798	1.000	0.713	0.832	0.631	1.000	0.302	0.464	No
XGBoost (Under)	0.950	0.955	0.945	0.950	0.929	0.926	0.714	0.806	No

Rationale for Model Selection

The final model selection shows AdaBoost (Original Data) is the best option



High Consistency

Achieves strong performance across all metrics in training and validation, indicating good generalization and a reduced risk of overfitting

Balanced Metrics

Maintains a healthy balance between precision and recall, resulting in a high F1 score—essential for minimizing both false positives and false negatives

Strong Validation Precision and F1

Exhibits nearly perfect validation precision and a robust F1 score, ensuring effective and accurate positive case classification

Model Performance on Test Data

Here's the performance of the Selected AdaBoost model (trained on the original training set)

Metric	Score
Accuracy	0.968
Recall	0.848
Precision	0.945
F1 Score	0.894

Detailed Classification Report:

Class	Precision	Recall	F1-Score	Support
0	0.97	0.99	0.98	1275
1	0.95	0.85	0.89	244

Metric	Score
Macro Avg Precision	0.96
Macro Avg Recall	0.92
Macro Avg F1	0.94
Weighted Avg Precision	0.97
Weighted Avg Recall	0.97
Weighted Avg F1	0.97

Final Model Performance Key Observations

Key Insights:

1. **High Accuracy:** The model achieves a high accuracy of 96.8%, indicating that it correctly predicts the class labels for most of the test data
2. **Balanced Performance:** With a high precision (94.5%) and F1 score (89.4%), the model effectively balances between identifying true positives and minimizing false positives

Class Performance:

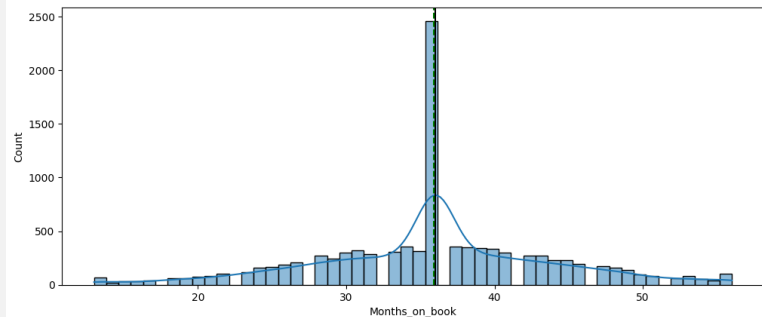
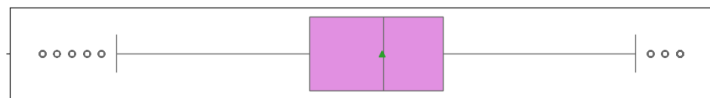
1. **For Class 0**, the model shows excellent precision (97%) and recall (99%), leading to a very high F1 score (98%)
2. **For Class 1**, while the precision remains high at 95%, the recall is slightly lower at 85%, resulting in an F1 score of 89%

Conclusion:

Overall, the AdaBoost model demonstrates robust performance, effectively distinguishing between classes with a strong balance between precision and recall, making it reliable for predicting customer churn.

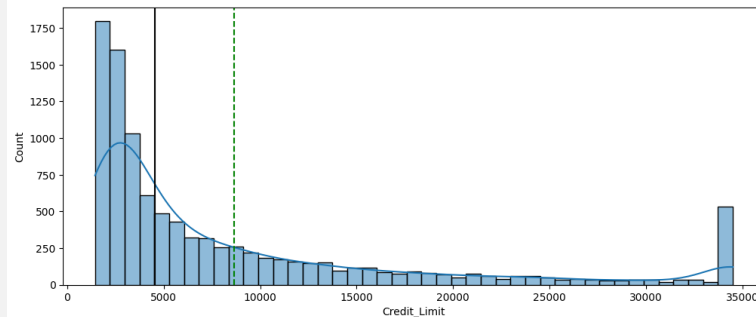
APPENDIX

EDA – Univariate Analysis



Months on Book

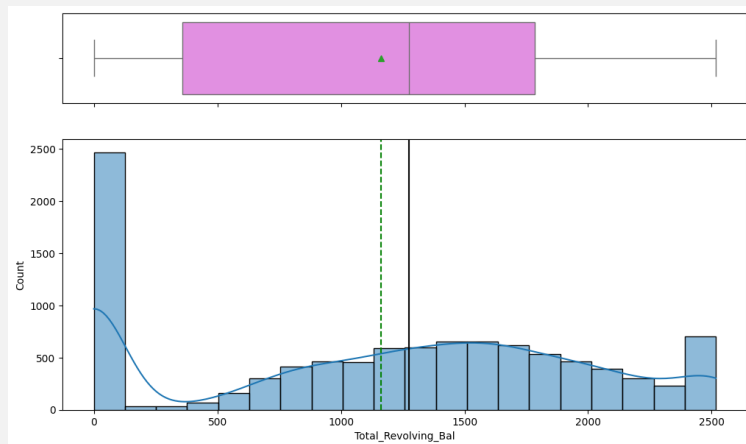
1. Distribution Shape: Peak at 36 months
2. Quartiles and Range: Narrow interquartile range
3. Outliers: Numerous high-end outliers
4. Retention Insights: Critical point at 36 months



Credit Limit:

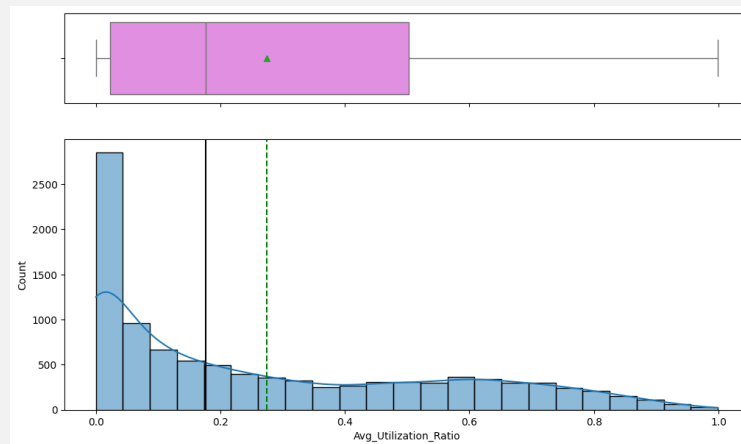
1. Distribution Shape: Highly right-skewed
2. Peaks and Tails: Peaks at lower limits, long tail higher
3. Median and Quartiles: Median below \$10,000, skewed lower range
4. Outliers: Significant high-end outliers
5. Mean: Higher than median, pulled up by outliers
6. Implication: Crucial for understanding behavior, risk, upselling opportunities

EDA – Univariate Analysis



Total revolving Balance:

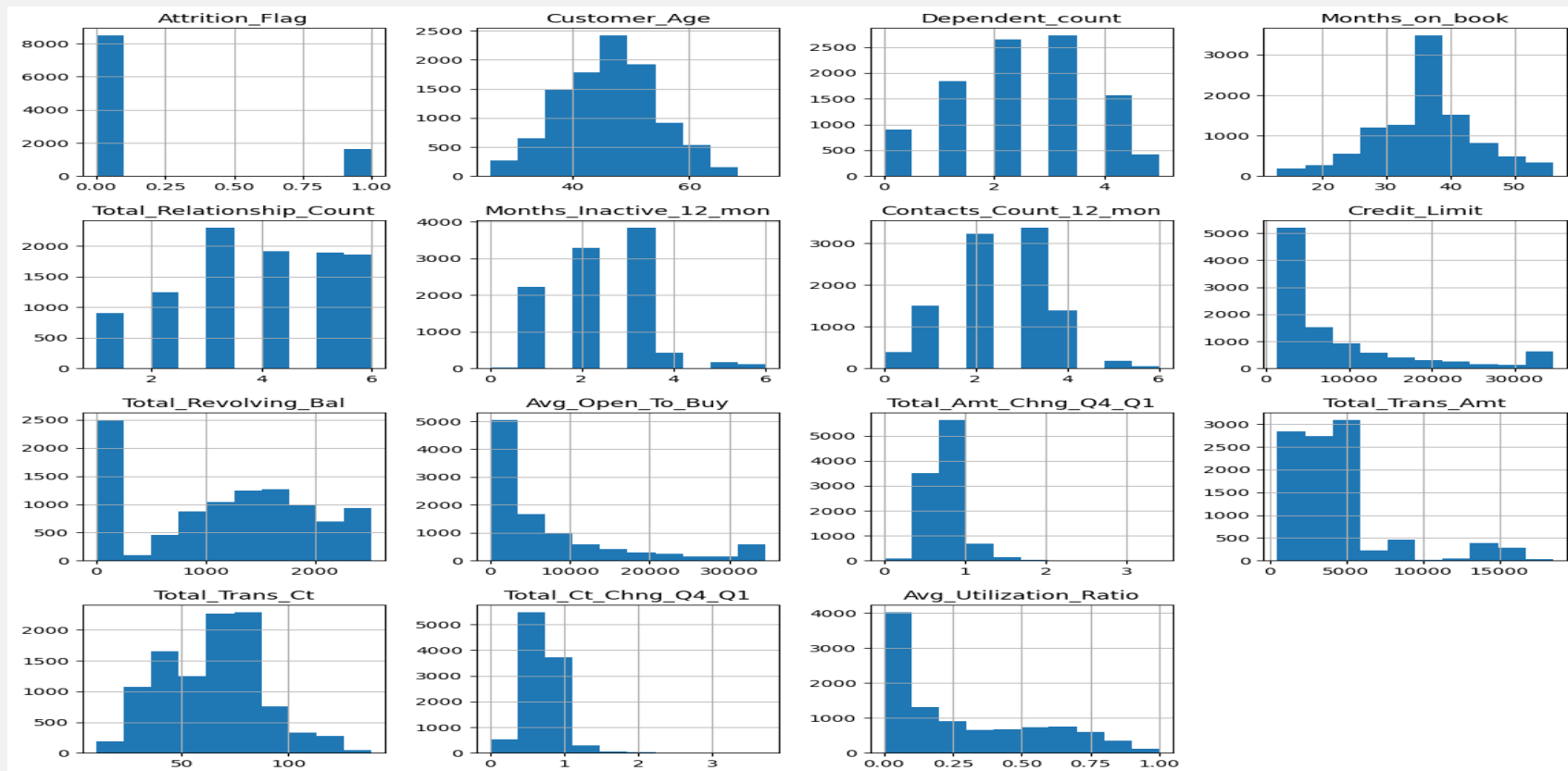
1. Mode at Zero: Significant peak at zero
2. Distribution Shape: Right-skewed distribution
3. Spread and Tail: Long tail towards higher balances
4. Outliers: Few high-end outliers
5. Credit Utilization Patterns: Varying utilization, many zero balances
6. Risk and Revenue Implications: Higher balances imply risk and revenue
7. Customer Financial Behavior: Diverse behaviors, tailored strategies needed



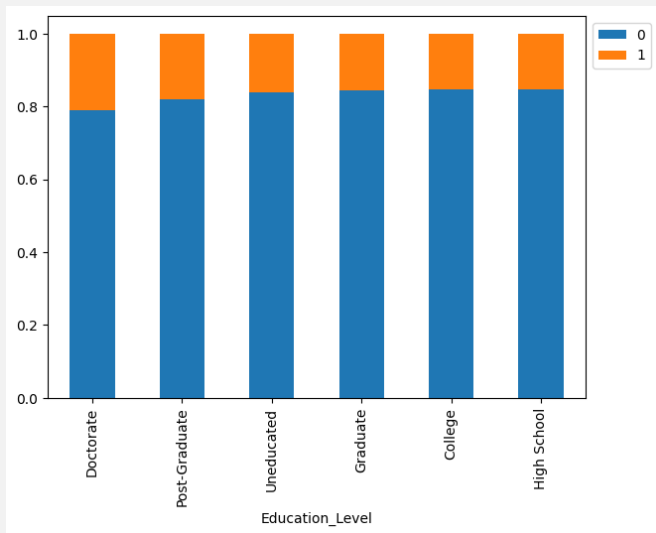
Avg Utilization Ratio Ratio:

1. Distribution Shape: Heavily left-skewed
2. High Frequency at Lower Ratios: Peak close to zero
3. Outliers: Near ratio of 1
4. Financial Health: Low ratios indicate better financial position
5. Marketing Opportunities: Tailor offerings based on utilization patterns
6. Implication: Guides risk management, marketing, and product development

EDA Univariate Analysis

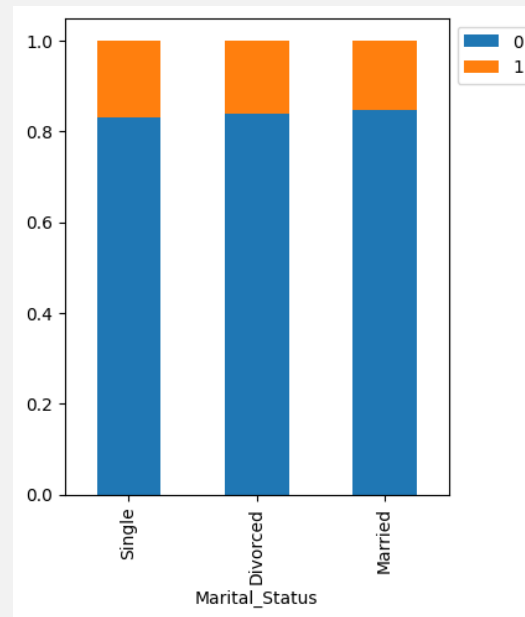


EDA Bivariate Analysis



Marital Status:

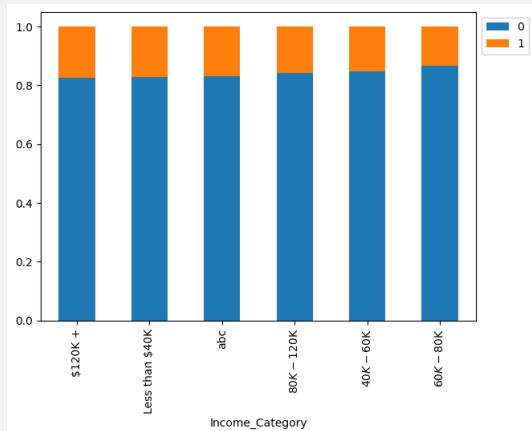
1. Lower attrition among married customers
2. Married individuals may be more financially stable



Education Level:

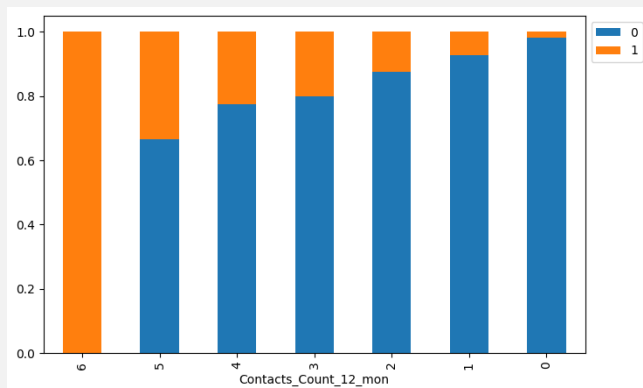
1. Slightly lower attrition for higher education levels (Doctorate and Post-Graduate)
2. Higher financial literacy or satisfaction with services

EDA Bivariate Analysis



Income Category:

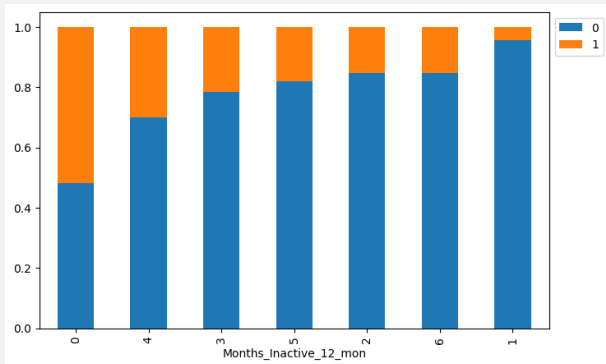
1. Lower attrition in higher income brackets (">\$120K" and "80K - 120K")
2. Higher-income customers may value tailored, premium services



Contacts with Bank (Contacts_Count_12_mon):

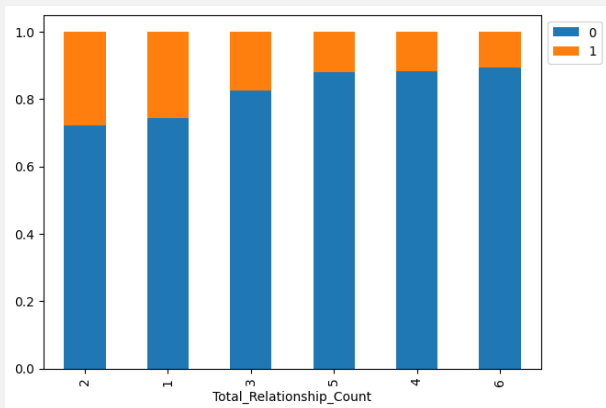
1. Higher contact levels correlate with higher attrition
2. Increased contact may indicate issues or dissatisfaction

EDA Bivariate Analysis



Months Inactive:

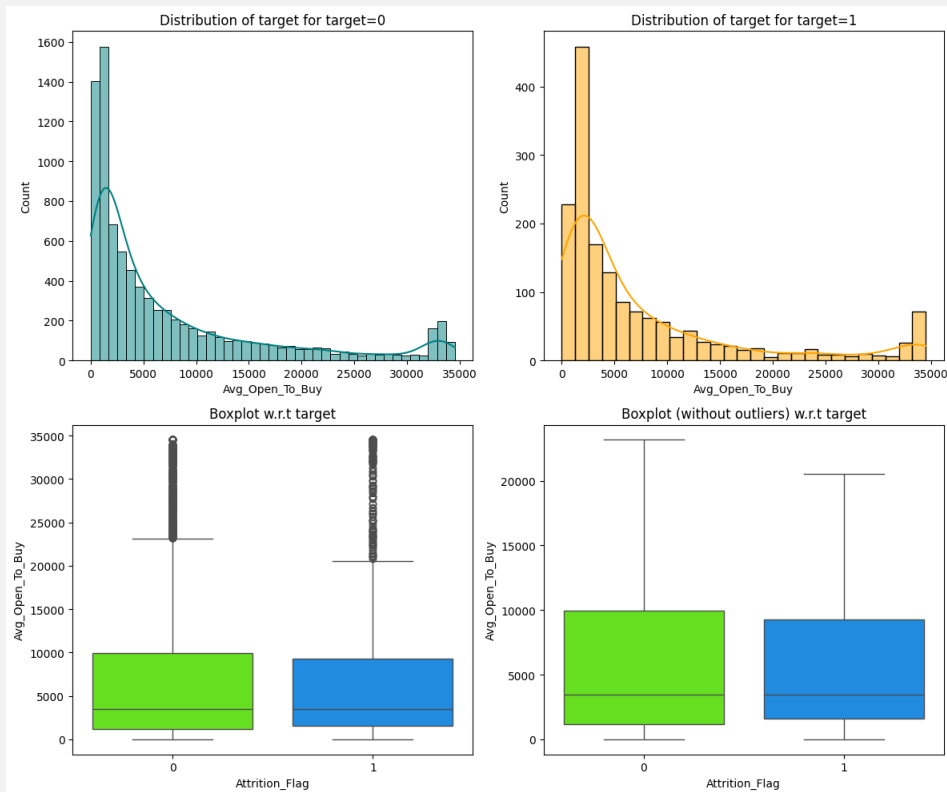
1. Higher inactivity correlates with higher attrition, especially at 3-4 months
2. Disengagement or lack of usage is a strong predictor of attrition



Total Relationship Count:

1. Lower attrition with more products (4-6)
2. Broader relationship with the bank enhances retention

EDA Bivariate Analysis



1. Lower Unused Credit Limits in Customers Who Left:

- Indicates perceived insufficient credit availability or unfavorable terms
- Affects customer satisfaction and loyalty

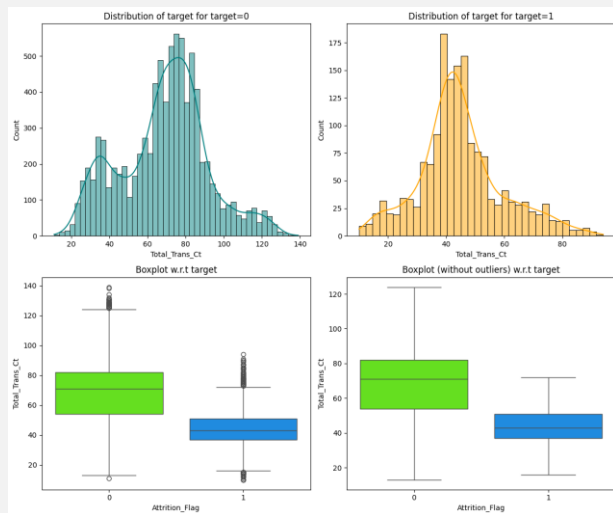
2. Credit Limit Management:

- Reevaluation needed for setting credit limits and offering increases
- Proactively manage limits to retain high creditworthiness customers

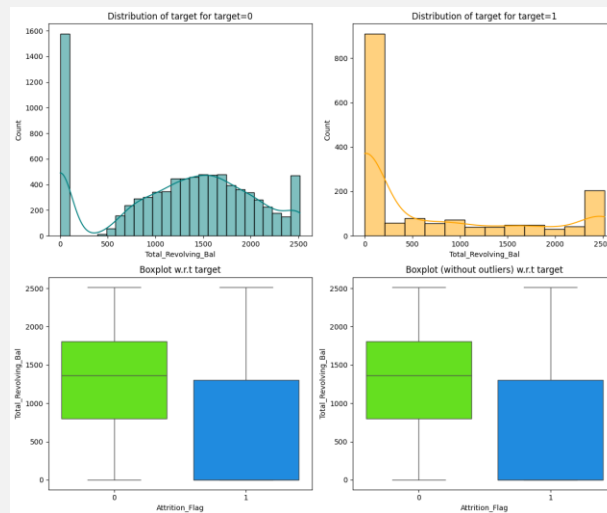
3. Risk Assessment and Customer Engagement:

- Customers with low "Avg_Open_To_Buy" at higher risk of attrition
- Targeted engagement strategies needed to suit their financial needs and improve retention

EDA Bivariate Analysis

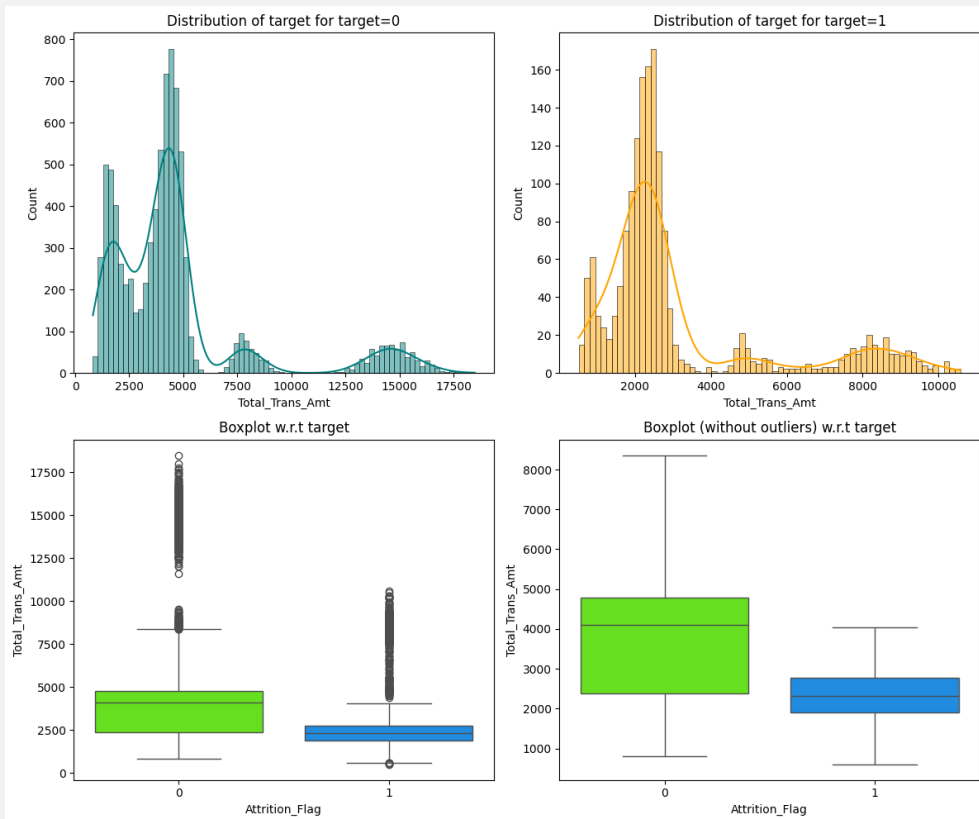


Attrition and Transactions (Total_Trans_Ct): There is a strong negative correlation of -0.37 between customer attrition and the total number of transactions. Higher transaction activity is linked to lower attrition rates, indicating active customers are less likely to leave.



Attrition and Revolving Balances (Total_Revolving_Bal): A negative correlation of -0.26 suggests that customers with higher revolving balances are less likely to close their accounts, possibly due to ongoing credit utilization.

EDA Bivariate Analysis



- 1. Enhanced Customer Engagement:**
Increase transaction volumes with loyalty programs, personalized offers, competitive products
- 2. Targeted Interventions:**
Identify and address needs of customers with decreasing transaction volumes
- 3. Customer Segmentation:**
Tailor marketing and service strategies based on transaction amounts to maximize customer lifetime value

Model Performance Summary (Original Data)

Model	Training Recall	Validation Recall
Bagging	0.975	0.765
Random Forest	1.000	0.802
AdaBoost	0.860	0.815
Logistic Regression	0.512	0.444
Gradient Boosting	0.894	0.827
XGBoost	1.000	0.877

- The Random Forest and XGBoost models show perfect recall in training, suggesting a strong fit to the training data but potential overfitting, as seen by lower recall in the validation set
- Logistic Regression shows notably lower recall scores in both training and validation, indicating difficulties in capturing the complexity of the dataset
- AdaBoost and Gradient Boosting show more balanced recall performance between training and validation, suggesting better generalization compared to the more extreme results of Bagging, Random Forest, and XGBoost.

Model Performance Summary (oversampled data)

Smote method for over sampling

Model	Training Recall	Validation Recall
Bagging	0.998	0.852
Random Forest	1.000	0.877
AdaBoost	0.968	0.864
Logistic Regression	0.822	0.741
Gradient Boosting	0.978	0.864
XGBoost	1.000	0.901

- Both Random Forest and XGBoost display perfect training recall, suggesting they can identify all churned customers in the training data, potentially indicating overfitting.
- Logistic Regression shows the lowest recall performance, reflecting challenges in handling the complexity of the oversampled dataset.
- Gradient Boosting and AdaBoost exhibit strong performance, striking a balance between effective recall and avoiding overfitting

Model Performance Summary (undersampled data)

- Using Random Under Sampler

Model	Training Recall	Validation Recall
Bagging	0.992	0.938
Random Forest	1.000	0.926
AdaBoost	0.963	0.938
Logistic Regression	0.780	0.728
Gradient Boosting	0.978	0.938
XGBoost	1.000	0.951

- Both Random Forest and XGBoost exhibit perfect training recall scores, potentially indicating overfitting to the under sampled data, but still perform very well on validation data.
- Logistic Regression shows the lowest recall scores, suggesting it struggles more with the complexities and reduced sample size of the under sampled dataset.
- Gradient Boosting and AdaBoost provide strong and balanced performance on both training and validation sets, suggesting effective handling of the dataset's constraints.

Model Performance on Test Data

Model	Accuracy	Recall	Precision	F1 Score
XGBoost (Original)	0.958	0.898	0.849	0.873
XGBoost (Oversampled)	0.669	0.996	0.326	0.491
AdaBoost (Original)	0.968	0.848	0.945	0.894

AdaBoost (Original) performs the better overall

1. XGBoost (Original): High accuracy with balanced metrics, excelling in recall and precision, indicating effective performance across the dataset.
2. XGBoost (Oversampled): Extremely high recall with poor precision, likely overfitting the minority class, reducing its general applicability.
3. AdaBoost (Original): Very high accuracy and excellent precision, slightly lower recall than XGBoost but effective in classifying positive cases with few false positives.

Data Insights into Credit Card Service Retention: Thera Bank's Proactive Model

The End