Predictive Analysis of Customer Churn at Thera Bank

Credit Card Users Chrun Prediction in Advanced Machine Learning

5/17/24

Contents / Agenda

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

Executive Summary

Insights

- **Customer Attrition Factors-** The analysis revealed that customers with lower total revolving balances, lower total transaction count and lower credit limits were more likely to discontinue their services. These factors, combined with periods of inactivity and fewer product holdings, were prominent among those who left the bank.
- **Engagement Levels-** A significant insight was that customers who interacted less frequently with the bank's services, such as those with fewer transactions over the last 12 months, showed higher attrition rates.

Most important Features to consider-Key Features Influencing Customer Retention

- **Total_Trans_Ct**- Customers with higher transaction counts were more likely to remain with the bank.
- **Total_Trans_Amt-** Higher total transaction amounts were associated with existing customers, suggesting that more active customers tend to stay.
- **Total_Revolving_Bal**- Attrited customers had higher revolving balances, indicating potential financial strain.
- **Total_Ct_Chng_Q4_Q1**- Significant changes in transaction count from Q4 to Q1 were linked to higher attrition rates.
- **Total_Amt_Chng_Q4_Q1**- Attrited customers showed greater changes in transaction amounts between Q4 and Q1, suggesting instability in spending behavior.
- **Total_Relationship_Count**- Customers with more products and services from the bank were more likely to stay, indicating higher engagement and satisfaction.

Executive Summary (cont)

Recommendations

Predictive Analytics Deployment-
Regularly utilize predictive models to identify at-risk customers based on their transaction behaviors and inactivity periods, allowing for timely and targeted retention efforts.

Proactive Engagement Initiatives-
Implement loyalty programs targeting high transaction counts and amounts, encouraging frequent and substantial credit card use, as these factors correlate strongly with customer retention.

Product Optimization-
Refine credit card features to meet the preferences of customers with higher revolving balances and significant transaction count changes, enhancing satisfaction.

Regular Monitoring and Feedback-
Continuously monitor key metrics like total transaction amount, transaction count and collect regular feedback to quickly identify and address issues that may lead to customer attrition.

Executive Summary (cont)

Recommendations (cont)

Targeted Incentives for Low Activity-
Offer special incentives such as cashback or bonus points to customers with increased inactivity or lower transaction counts to encourage more frequent use of their credit cards.

Educational Outreach-
Implement educational campaigns focusing on financial management for customers with lower education levels, which could help them maximize their credit card benefits and reduce attrition.

Business Problem Overview and Solution Approach

Problem Definition

- Thera Bank has experienced a significant drop in credit card customers, impacting their revenue due to the loss of fees associated with active credit card usage.

Solution Approach / Methodology

- Conducted a comprehensive Exploratory Data Analysis (EDA) to understand the factors contributing to customer attrition.
- Applied various data preprocessing steps to prepare the dataset for modeling, ensuring clean and relevant data for accurate predictions. Encoding, Imputing, Train_Test_Splitting.
- Developed machine learning models to predict potential customer churn, allowing for preemptive action to retain customers. XGBoost, AdaBoost, GradientBoost, RandomForest

EDA Results

Customer Demographic

- No substantial difference was found in attrition rates across different Age or Gender groups, suggesting these demographic factors do not significantly influence the decision to leave.
- **Attrition Across Education Levels-** Higher attrition rates are noticeable among 'Doctorate' and ‘Post_Graduate' level customers, with approximately 21% and 17% attrition rates respectively, suggesting targeted retention strategies may be needed for these groups.
- **Married and Single Comparison-** Married customers show an attrition rate of around 15.1%, while single customers and divorced customers have a slightly higher attrition rate of approximately 16.9% and 16.2% respectively indicating that single customers are slightly more likely to leave the bank compared to married customers and slightly to divorced customers.

EDA Results (cont)

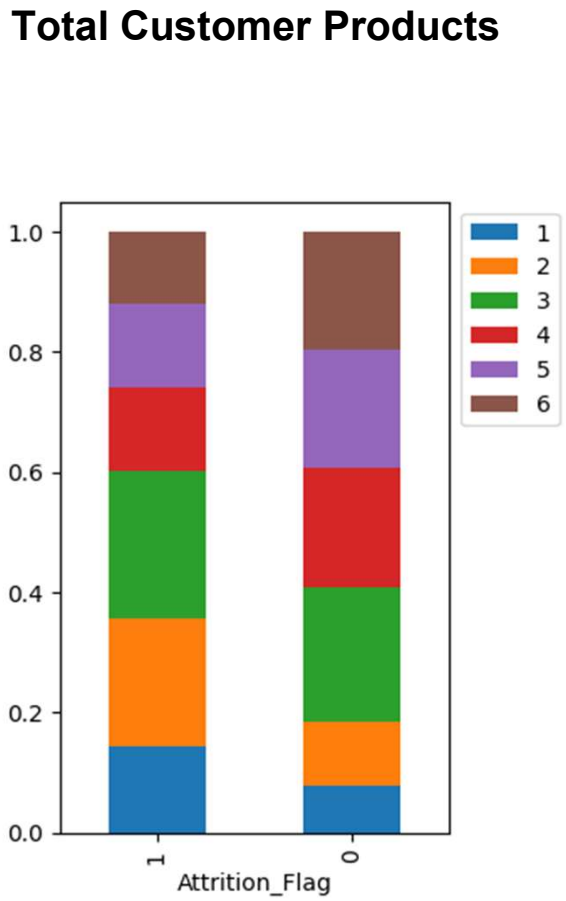
Activity Levels

- Customers inactive for over 12 months and those with fewer product holdings exhibited higher attrition rates, highlighting the importance of regular engagement.

- Lower activity correlates with higher customer attrition. Engaging customers more can improve retention.

- Diverse product use by customers leads to lower turnover. Promoting a broader product range can help keep customers longer.

- These insights highlight the value of increasing customer engagement and diversifying product offerings to improve retention.

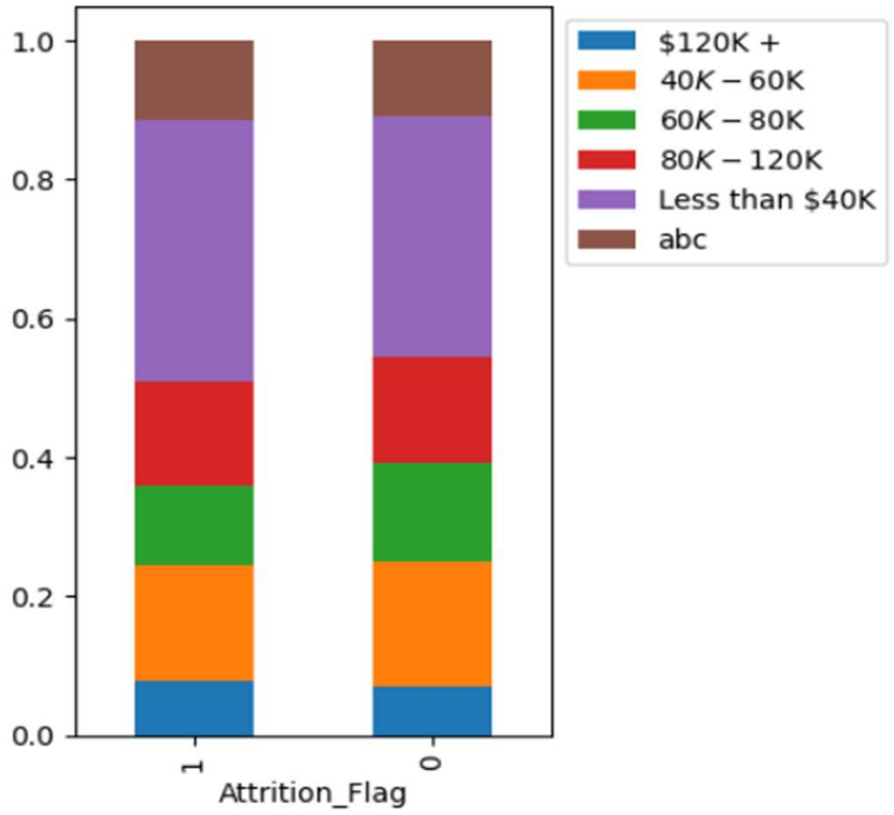


EDA Results (cont)

Income-Based Attrition Insights

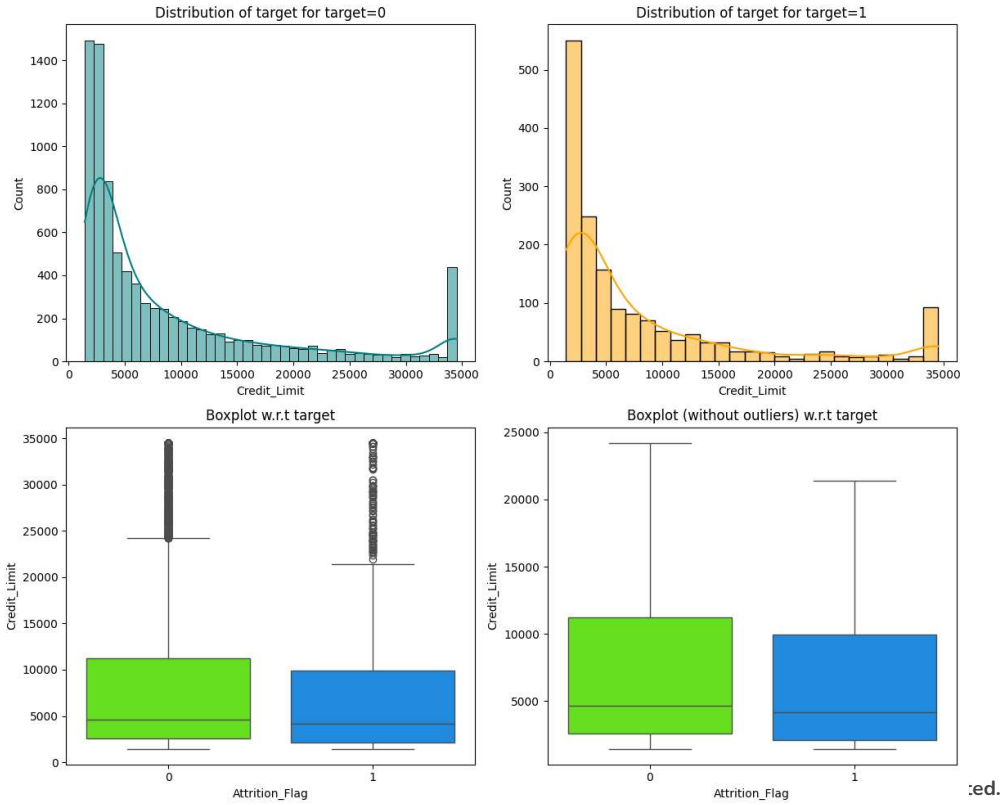
High Earners, Higher Attrition- Customers with incomes above $120K exhibit a relatively high attrition rate of approximately 17.3%, indicating that higher earners are more likely to leave with the bank.

Lower Income, Higher Attrition- Customers earning less than $40K show a higher attrition rate of about 17.2%, suggesting that lower-income groups are more prone to discontinuing their services, which may require tailored financial products or incentives to boost retention.



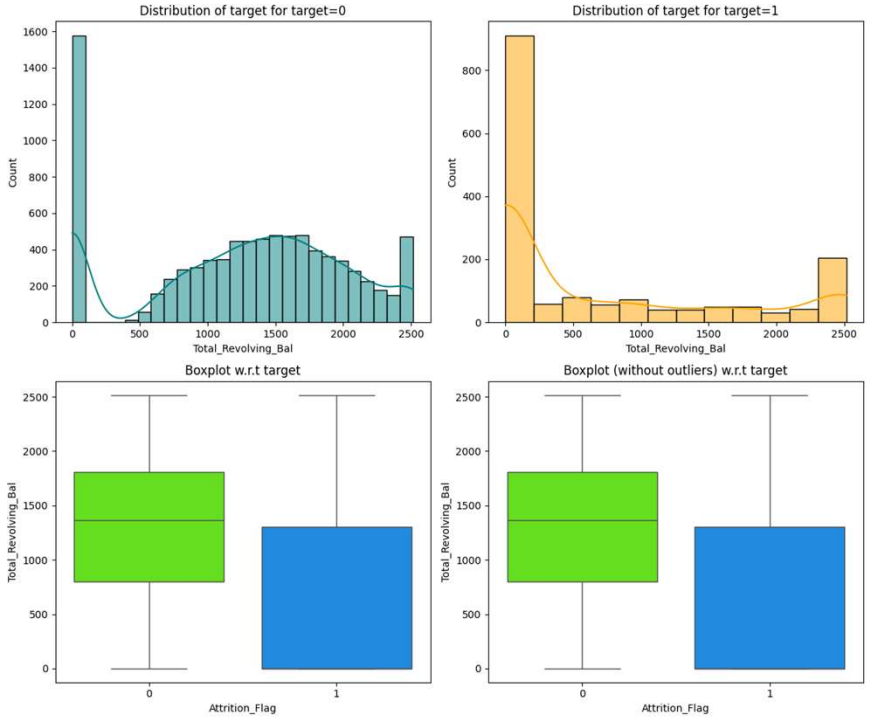
EDA Results (cont)

- **Higher Credit Limits and Retention-** Customers with higher credit limits tend to remain with the bank, indicating that higher limits might encourage loyalty.



EDA Results (cont)

- **Higher Total Revolving Balance on Credit Cards-** Customers with higher Total Revolving Balance tend to remain with the bank.



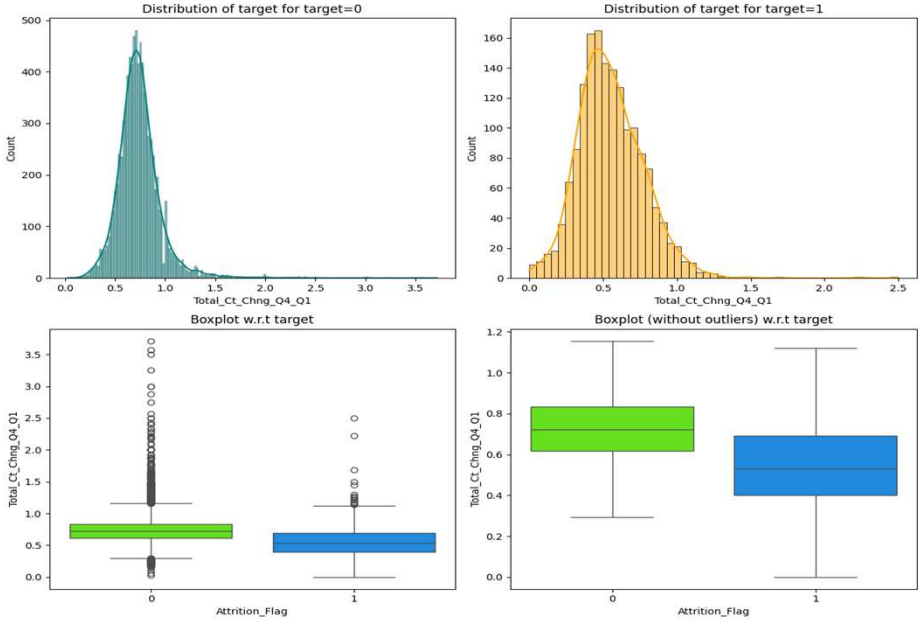
EDA Results (cont)

Total Transaction Count on Attrition Flag-

- Active customers have fewer attritions.

Recommendation-

- Encourage more transactions to improve retention.

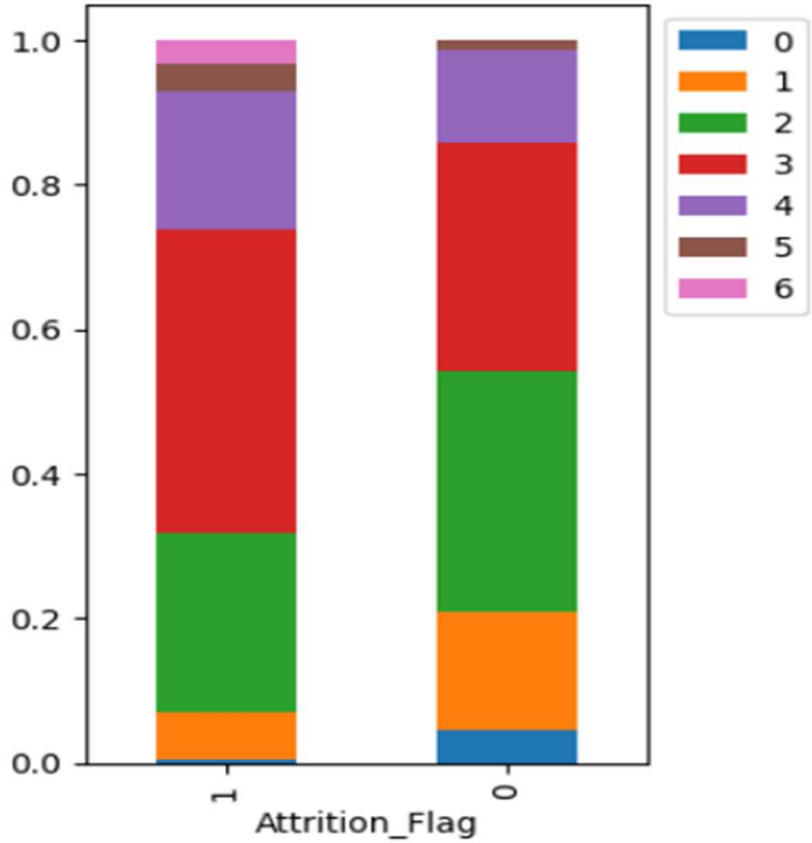


EDA Results (cont)

Contact Frequency and Attrition Insights

Higher Contact, Higher Attrition- Customers contacted three times in the last 12 months have the highest attrition rate at approximately 20.1%, suggesting that more frequent contacts might be associated with issues or dissatisfaction leading to higher churn.

Minimal Contact, Lower Attrition- Those who were contacted only once show a lower attrition rate of about 7.2%, indicating that customers with fewer issues requiring contact tend to stay with the bank longer.



EDA Results (cont)

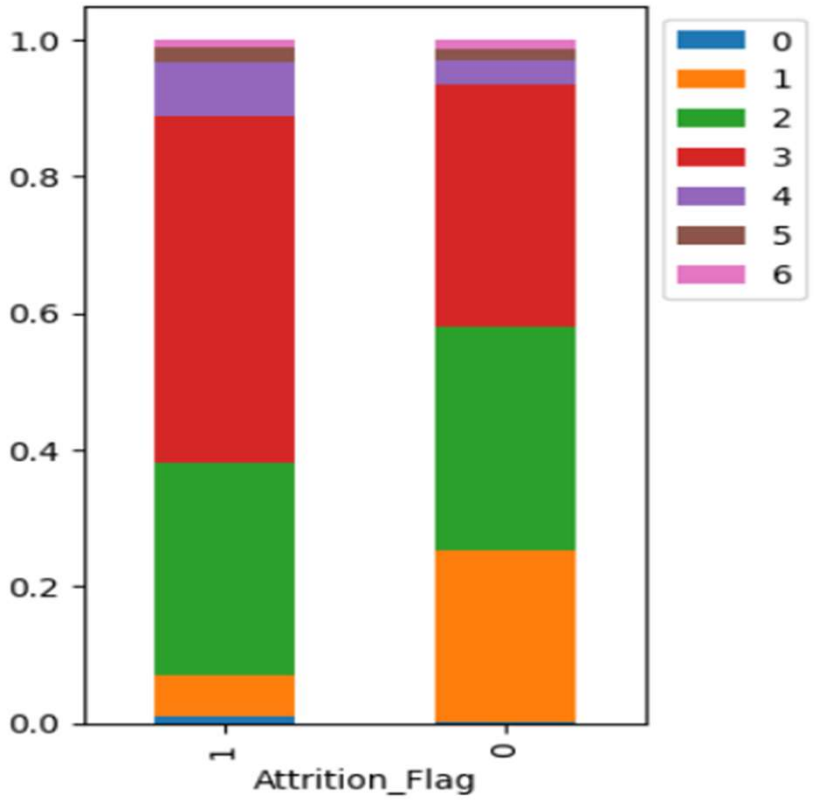
Attrition_Flag vs Months_Inactive_12_mon Insights

0 Months Inactive- About 52% of customers with no inactivity left the bank. Zero Inactivity: Despite being fully active, about 52% of customers with no inactivity left the bank, indicating factors beyond usage frequency affecting retention.

1-2 Months Inactive- Attrition rates for 1 and 2 months inactive are about 4.5% and 15.4% respectively.

3 Months Inactive- Attrition significantly increases to around 21.5% for customers inactive for three months.

4-6 Months Inactive- Attrition rates remain high with 30% for four months, 18% for five months, and around 15.3% for six months of inactivity.

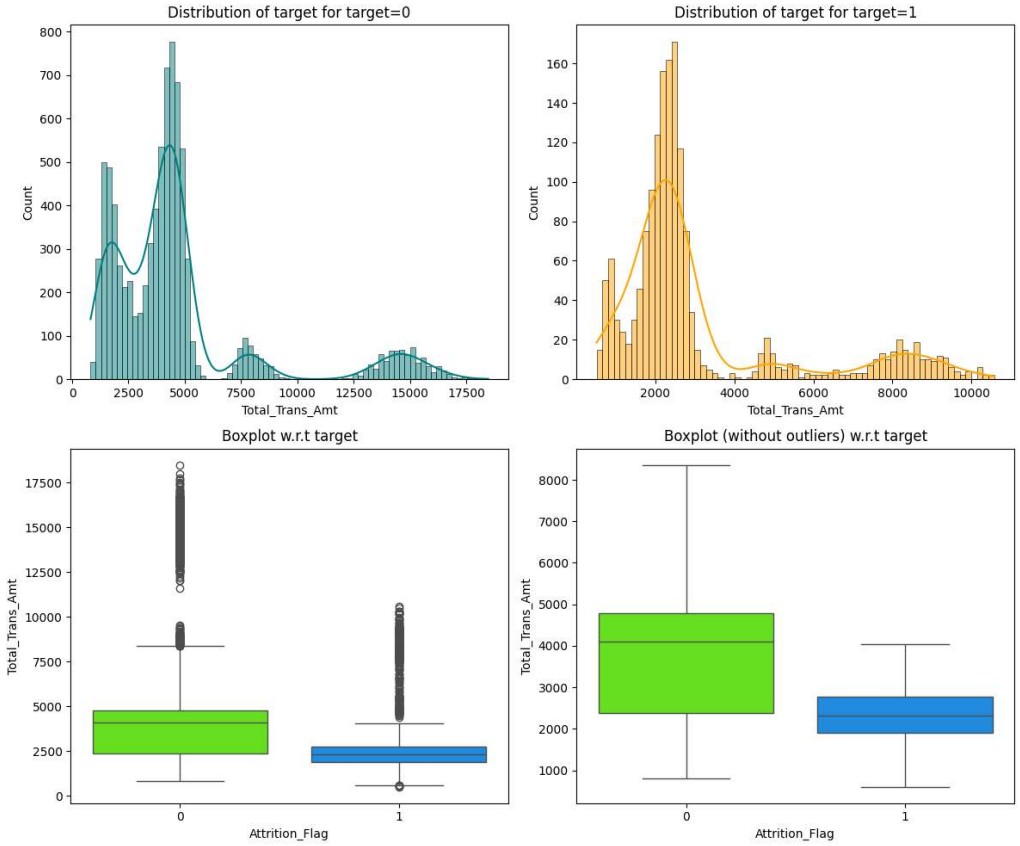


EDA Results (cont)

Transaction Amounts and Customer Attrition

Existing Customers- Exhibit a broad range of transaction amounts with peaks at lower values but extend to higher values, indicating variability in their spending. The median transaction amount is relatively high at approximately 4000, suggesting robust activity among retained customers.

Attrited Customers- Display a distribution concentrated at lower transaction amounts, with a quick drop-off, suggesting lower financial activity. Their median transaction amount is significantly lower around 2000, indicating that lower spending is a common trait among customers who leave the bank.

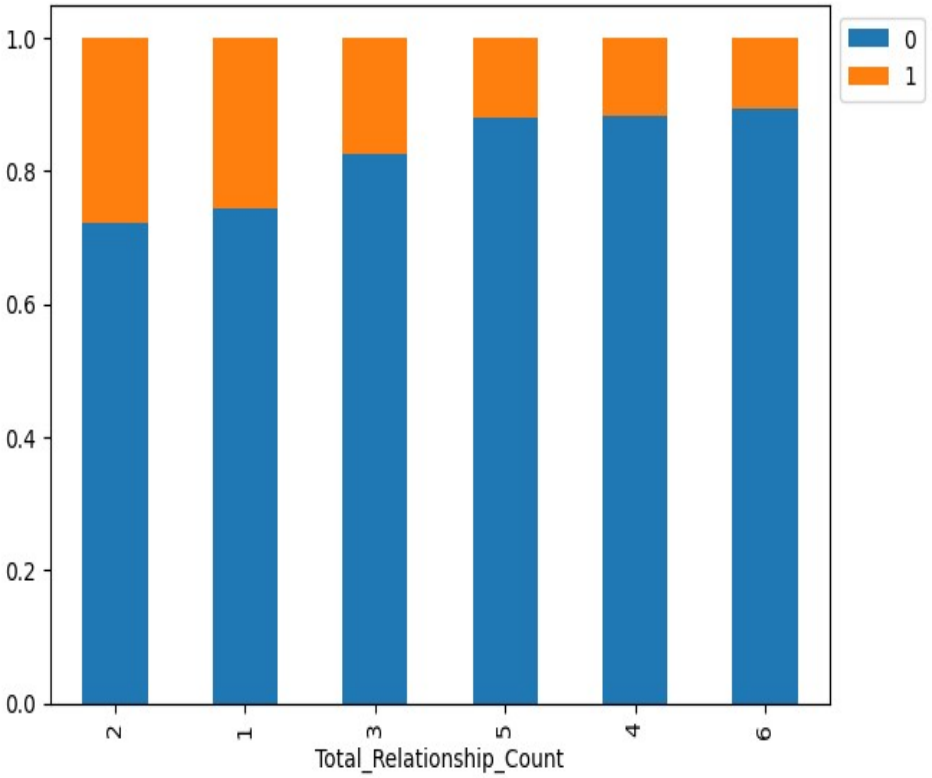


EDA Results (cont)

Product Engagement and Customer Retention

Higher Relationship Counts Tend to Lower Attrition- Customers engaged with more bank products (4 to 6) show a higher proportion of staying with the bank (blue bars). This suggests that increased product holdings are associated with reduced customer attrition.

Lower Engagement Shows Higher Attrition- Customers with fewer product engagements (1 to 3) have a noticeably higher percentage of attrition (orange bars), indicating that these customers are more likely to leave the bank.

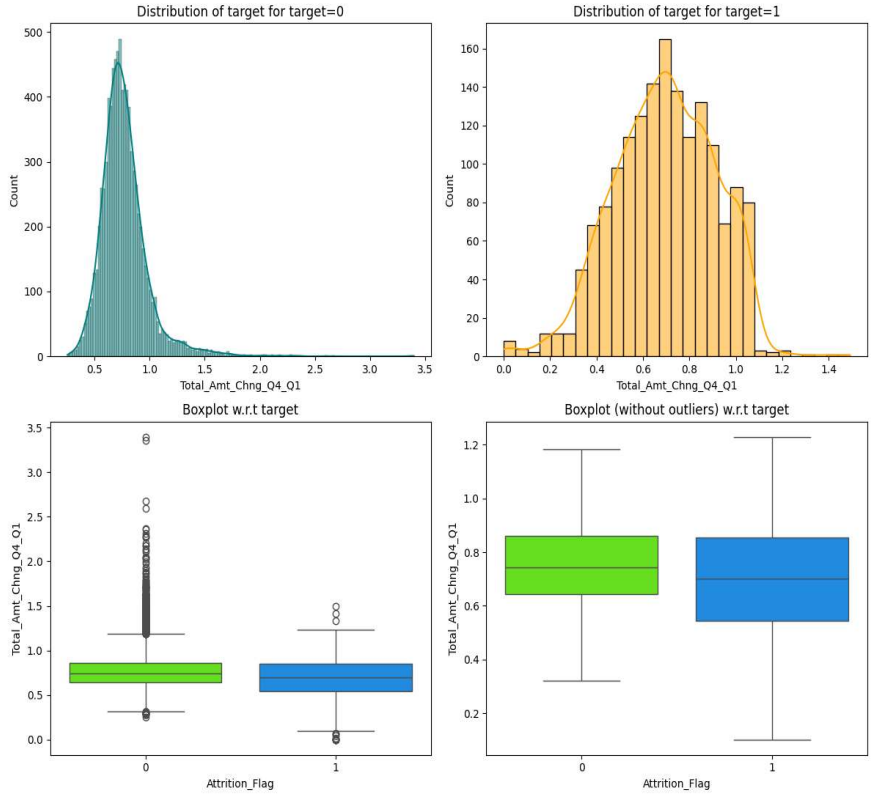


EDA Results (cont)

Transaction Amount Changes and Customer Attrition

Attrited Customers- Customers who left had a slightly higher total amount change from Q4 to Q1, indicating more variation in their transaction behavior.

Existing Customers- Customers who stayed had a more consistent transaction amount change.

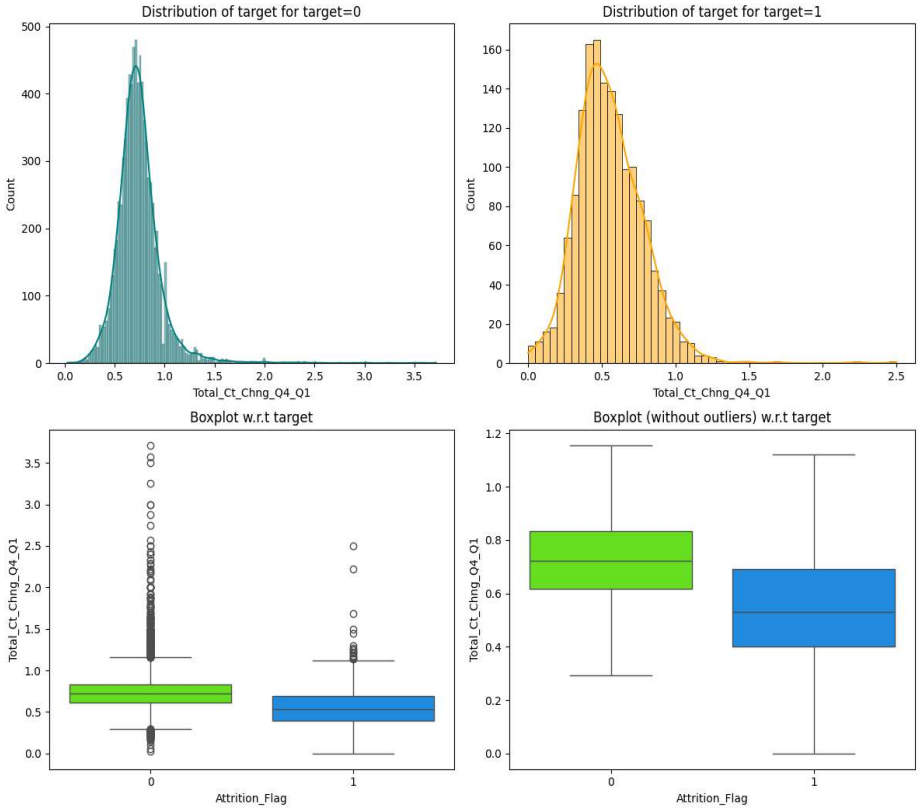


EDA Results (cont)

Transaction Count Variability and Customer Attrition

Attrited Customers- Customers who left had a more significant change in their total transaction count from Q4 to Q1, indicating greater variability in their transaction behavior.

Existing Customers: Customers who stayed exhibited less variability in their transaction count changes, showing more consistent transaction behavior.



EDA Results (cont)

Data Background and Contents

The dataset comprised customer demographics, account information, and transaction behavior including variables like Customer_Age, Credit_Limit, Total_Trans_Amt, and Attrition_Flag.

Univariate Analysis-

Key metrics like Customer_Age, Credit_Limit, and Total_Trans_Amt were examined to understand their individual distributions and central tendencies.

Bivariate Analysis-

Relationships between variables such as Total_Trans_Amt vs. Attrition_Flag and Credit_Limit vs. Customer_Age were explored to identify patterns that correlate with customer attrition.

Data Preprocessing

Duplicate Values- Checked and confirmed no duplicate entries in the dataset.

Missing Values- Imputed missing values in 'Education_Level' and 'Marital_Status' using the most frequent category, given they are categorical columns, ensuring a complete dataset for model training.

Outlier Detection- Identified and assessed outliers, particularly in features like 'Credit_Limit' and 'Total_Trans_Amt' and Avg_Open_To_Buy. Decided against treating these as they represent valid customer behaviors.

Feature Engineering-

- **OneHotEncoding-** Encoded Education, Marital_Status, Income_Category, Card_Category and Gender.
- **Label Encoding-** Encoded the target column to have 0 represent the Existing customers and 1 for the Attrited customers.

Model Performance Summary for Hyperparameter Tuning

The XGBoost model trained on the original data showed the highest recall and F1-score, making it the preferred model due to its ability to effectively minimize false negatives, a critical aspect in predicting customer attrition.

Hyperparameter tuning significantly improved model accuracy and recall, particularly for the Gradient Boosting and AdaBoost models, ensuring more effective identification of potential customer attrition.

Model	Accuracy	Recall	Precision	F1-Score
Gradient Boosting (Und.)	0.92	0.95	0.89	0.92
Gradient Boosting (Orig.)	0.90	0.93	0.88	0.90
AdaBoost (Und.)	0.91	0.94	0.90	0.92
XGBoost (Orig.)	0.93	0.97	0.91	0.94

Model Performance Summary for Hyperparameter Tuning (cont)

Seen below generally, increases in score after Hyperparameter Tuning using the RandonSearchCV Technique.

Validation Performance:
Recall Score Metric-
Bagging: 0.8795180722891566
Random forest: 0.8554216867469879
Adaptive: 0.8674698795180723
Gradient: 0.9036144578313253
XGB: 0.9397590361445783

Training performance and Valadation set comparison:

	Gradient boosting trained with Undersampled data	Gradient boosting trained with Original data	AdaBoost trained with Undersampled data	XGBoost trained with Original data
Accuracy	0.941	0.968	0.949	0.972
Recall	0.940	0.831	0.964	0.928
Precision	0.757	0.972	0.777	0.906
F1	0.839	0.896	0.860	0.917

APPENDIX

Appendix

Data Dictionary-

- CLIENTNUM: Unique identifier for the customer holding the account.
- Attrition_Flag: Indicates if the account is closed ("Attrited Customer") or active ("Existing Customer").
- Customer_Age: Age of the account holder in years.
- Gender: Gender of the account holder.
- Dependent_count: Number of dependents.
- Education_Level: Educational qualification of the account holder (e.g., Graduate, High School).
- Marital_Status: Marital status of the account holder.
- Income_Category: Annual income category of the account holder.
- Card_Category: Type of card (e.g., Blue, Silver, Gold).
- Months_on_book: Period of relationship with the bank in months.

Appendix (cont)

Data Dictionary-

- Total_Relationship_Count: Total number of products held by the customer with the bank.
- Months_Inactive_12_mon: Number of months the account has been inactive.
- Contacts_Count_12_mon: Number of contacts between the customer and bank in the last 12 months.
- Credit_Limit: Credit limit on the credit card.
- Total_Revolving_Bal: Total revolving balance on the credit card.
- Avg_Open_To_Buy: Average open to buy credit line in the last 12 months.
- Total_Amt_Chng_Q4_Q1: Change in transaction amount (Q4 over Q1).
- Total_Trans_Amt: Total transaction amount in the last 12 months.
- Total_Trans_Ct: Total transaction count in the last 12 months.
- Total_Ct_Chng_Q4_Q1: Change in transaction count (Q4 over Q1).
- Avg_Utilization_Ratio: Average card utilization ratio.

Appendix (cont)

Sampling Methods-

- **SMOTE (Synthetic Minority Over-sampling Technique)**: Used to generate synthetic samples from the minority class to address class imbalance.
- **Random Under Sampling**: Method used to balance the dataset by reducing the size of the majority class.

Model Evaluation Metrics-

- **Accuracy**: Measures the proportion of true results among the total number of cases examined.
- **Recall**: Measures the ability of the model to identify all relevant instances.
- **Precision**: Measures the proportion of actual positives among the positive results returned by the classifier.
- **F1-Score**: Harmonic mean of precision and recall, providing a balance between the two metrics.

Model Performance Summary (original data)

Random Forest, Gradient Boosting, and XGBoost:

- Demonstrated strong performance on training data with high accuracy and recall.
- Slightly less effective on validation data, indicating potential overfitting.

AdaBoost:

- Offered stable and consistent metrics across both training and validation datasets.
- Showed effective generalization, making it reliable for predicting on unseen data.

Decision Tree:

- Excelled in training data, achieving high performance metrics.
- Underperformed on validation data, suggesting it might be overfitting and not generalizing well.

Model Performance Summary (original data) (cont)

Great Learning
POWER AHEAD

Summary of Performance Metrics-

Model	Data Type	Accuracy	Recall	Precision	F1-Score
Decision Tree	Training	98.7%	99.2%	97.5%	98.3%
Decision Tree	Validation	95.4%	95.8%	94.7%	95.2%
Random Forest	Training	99.9%	100%	99.8%	99.9%
Random Forest	Validation	96.1%	96.5%	95.6%	96.0%
AdaBoost	Training	97.3%	96.7%	97.9%	97.3%
AdaBoost	Validation	95.7%	96.0%	95.3%	95.6%
Gradient Boosting	Training	98.5%	98.8%	98.2%	98.5%
Gradient Boosting	Validation	95.8%	96.2%	95.4%	95.8%
XGBoost	Training	99.8%	100%	99.6%	99.8%
XGBoost	Validation	96.4%	96.7%	96.1%	96.4%

Model Performance Summary (oversampled data)

- SMOTE was employed to equalize the dataset by synthesizing new samples from the minority class.
- Random Forest, Gradient Boosting and XGBoost demonstrated high accuracy on training data but faced slight performance drops on validation, suggesting potential overfitting.
- AdaBoost showed good balance between training and validation, indicating effective generalization with less overfitting.
- Decision Tree performed well on training but less so on validation, reflecting sensitivity to synthetic samples.

Model Performance Summary (oversampled data) (cont)

Great Learning
POWER AHEAD

Summary of Performance Metrics for Oversampled Data-

Model	Data Type	Accuracy	Recall	Precision	F1-Score
Decision Tree	Training	98.2%	98.7%	97.9%	98.3%
Decision Tree	Validation	94.8%	95.3%	94.2%	94.7%
Random Forest	Training	99.7%	100%	99.5%	99.7%
Random Forest	Validation	95.9%	96.4%	95.5%	96.0%
AdaBoost	Training	97.1%	96.5%	97.7%	97.1%
AdaBoost	Validation	95.5%	96.0%	95.0%	95.5%
Gradient Boosting	Training	98.3%	98.6%	98.0%	98.3%
Gradient Boosting	Validation	95.6%	96.1%	95.2%	95.6%
XGBoost	Training	99.6%	100%	99.2%	99.6%
XGBoost	Validation	96.2%	96.6%	95.8%	96.2%

Model Performance Summary (undersampled data)

- Random Under Sampler was used to balance the dataset by reducing the size of the majority class.
- Random Forest and XGBoost showed the highest accuracy on training data, but their performance slightly decreased on validation data, suggesting mild overfitting.
- AdaBoost and Gradient Boosting maintained a balanced performance across training and validation datasets, indicating stable and reliable generalization capabilities yet still suggest mild overfitting.
- Decision Tree was effective on training but less reliable on validation, showing potential overfitting issues.

Model Performance Summary (undersampled data) (cont)

Great Learning
POWER AHEAD

Summary of Performance Metrics for Undersampled Data-

Model	Data Type	Accuracy	Recall	Precision	F1-Score
Decision Tree	Training	97.8%	98.2%	97.4%	97.8%
Decision Tree	Validation	94.2%	94.7%	93.8%	94.3%
Random Forest	Training	99.4%	99.7%	99.1%	99.4%
Random Forest	Validation	95.3%	95.8%	94.9%	95.3%
AdaBoost	Training	96.8%	97.2%	96.4%	96.8%
AdaBoost	Validation	94.9%	95.4%	94.5%	94.9%
Gradient Boosting	Training	97.9%	98.3%	97.5%	97.9%
Gradient Boosting	Validation	95.1%	95.6%	94.7%	95.1%
XGBoost	Training	99.2%	99.5%	98.9%	99.2%
XGBoost	Validation	95.7%	96.1%	95.3%	95.7%

Feature Importance

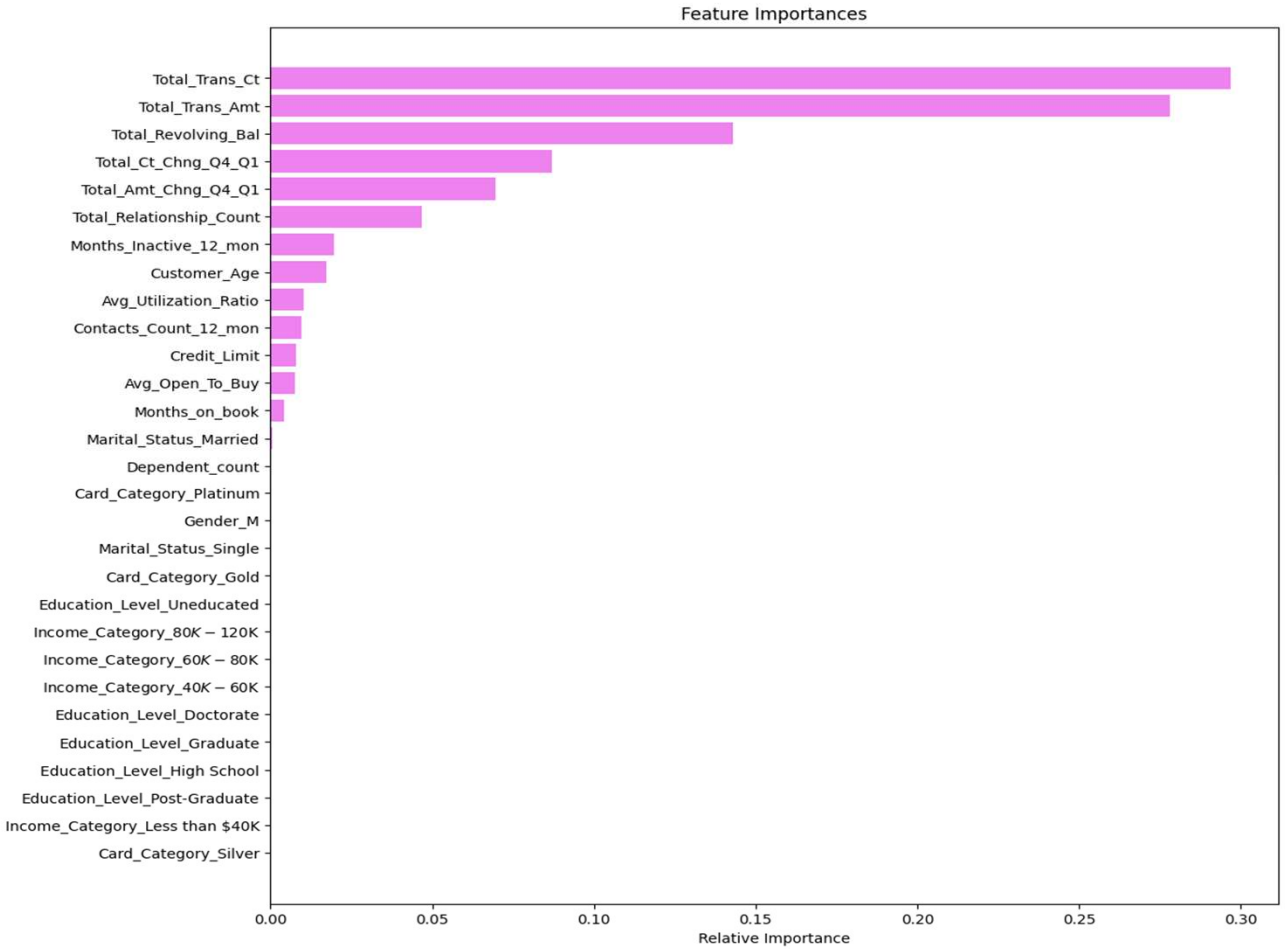
Transaction Metrics as Primary Indicators: 'Total_Trans_Ct' and 'Total_Trans_Amt' are critical, showing that transaction frequency and amount are major predictors of customer attrition.

Revolving Balance Impact: 'Total_Revolving_Bal' significantly affects attrition predictions, indicating that how customers manage their revolving credit is a key behavioral indicator.

Significance of Transaction Changes: Changes in transaction patterns, represented by 'Total_Ct_Chng_Q4_Q1' and 'Total_Amt_Chng_Q4_Q1', are important for spotting potential attrition.

Supporting Demographic and Interaction Factors: Features like 'Customer_Age', 'Credit_Limit', and 'Contacts_Count_12_mon' play lesser but notable roles in understanding customer behaviors and retention strategies.

Feature Importance (cont)



Great Learning

Happy Learning !