Cuans Bank
Customer Attrition Prediction Model

By Riza Arfiqih

Project Overview

01
Business & Data Understanding

02
Data Wrangling

03
Data Pre-processing

04
Exploratory Data Analysis

05
Model Building & Interpretation

06
Conclusion

Business Context

Customer attrition, also known as customer churn, is a significant challenge for banks and financial institutions. It refers to the phenomenon where customers stop using a bank's services, leading to a loss of revenue and potentially harming the bank's reputation. Cuans Bank faces a similar challenge and aims to proactively address customer attrition to enhance customer retention and overall business performance.

Objectives

Identify Key Drivers of Attrition

Determine the features that have the most significant impact on customer attrition, providing insights into the factors leading to attrition.

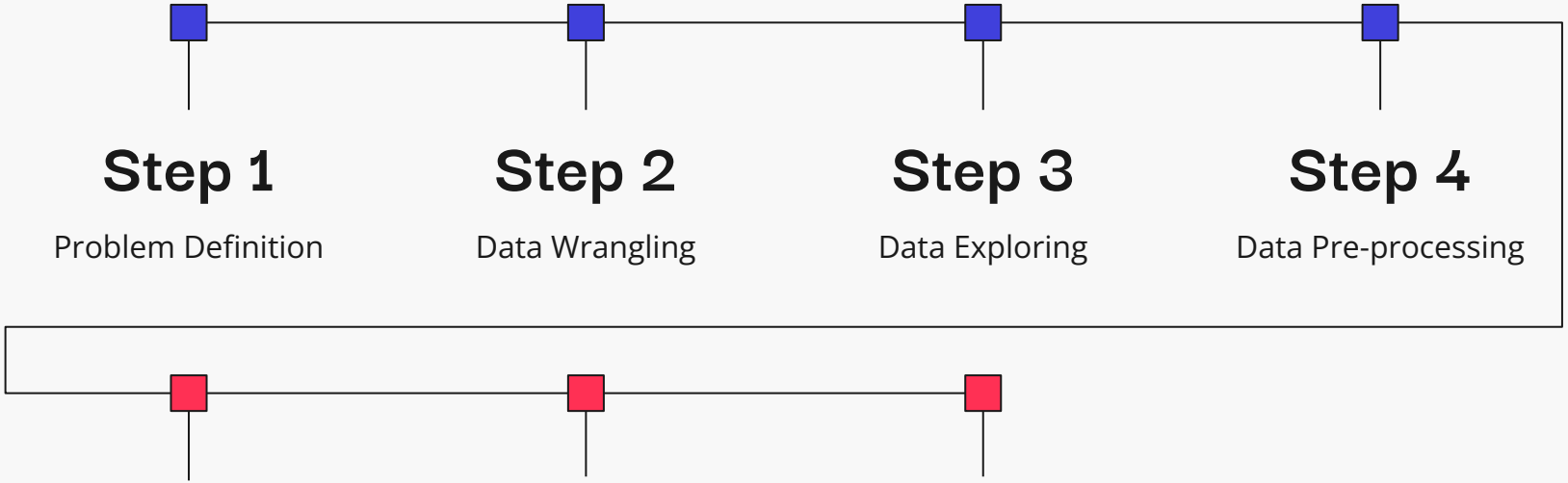
Develop a Predictive Model

Build a robust machine learning model that predicts customers at high risk of attrition. The model should prioritize high recall, F1-score, and a low False Negative Rate to minimize the chances of missing potential attrition cases.

Gain Actionable Insights

Understand the underlying patterns and relationships between customer behavior and attrition to inform targeted retention strategies. And develop specific recommendations on how to mitigate the issues found.

Machine Learning Workflow



Data Overview

The project will utilize a dataset named "*bank_churn_data.csv*". This dataset includes 21 columns capturing customer profiles (age, gender, education, marital status, income, etc.), account details (credit limit, card category, months on book, etc.), and customer activity (transactions, inactivity periods, contacts with the bank, etc.).

#	Column	Non-Null Count	Dtype
0	user_id	10127 non-null	int64
1	attrition_flag	10127 non-null	object
2	customer_age	10127 non-null	int64
3	gender	10127 non-null	object
4	dependent_count	10127 non-null	int64
5	education_level	10127 non-null	object
6	marital_status	10127 non-null	object
7	income_category	10127 non-null	object
8	card_category	10127 non-null	object
9	months_on_book	10127 non-null	int64
10	total_relationship_count	10127 non-null	int64
11	months_inactive_12_mon	10127 non-null	int64
12	contacts_count_12_mon	10127 non-null	int64
13	credit_limit	10127 non-null	float64
14	total_revolving_bal	10127 non-null	int64
15	avg_open_to_buy	10127 non-null	float64
16	total_amt_chng_q4_q1	10127 non-null	float64
17	total_trans_amt	10127 non-null	int64
18	total_trans_ct	10127 non-null	int64
19	total_ct_chng_q4_q1	10127 non-null	float64
20	avg_utilization_ratio	10127 non-null	float64

Customer Profile & Account

- **user_id**: customer account number.
- **attrition_flag**: customer status (Existing and Attrited).
- **customer_age**: age of the customer.
- **gender**: gender of customer (M for male and F for female).
- **dependent_count**: number of dependents of customers.
- **education_level**: customer education level (Uneducated, High School, Graduate, College, Post-Graduate, Doctorate, and Unknown).
- **marital_status**: customer's marital status (Single, Married, Divorced, and Unknown).
- **income_category**: customer income interval category (Less than 40k, 40k- 60k, 60k- 80k, 80k-120k, 120k +, and Unknown).
- **card_category**: type of card used (Blue, Silver, Gold, and Platinum). months_on_book: period of being a customer (in months).

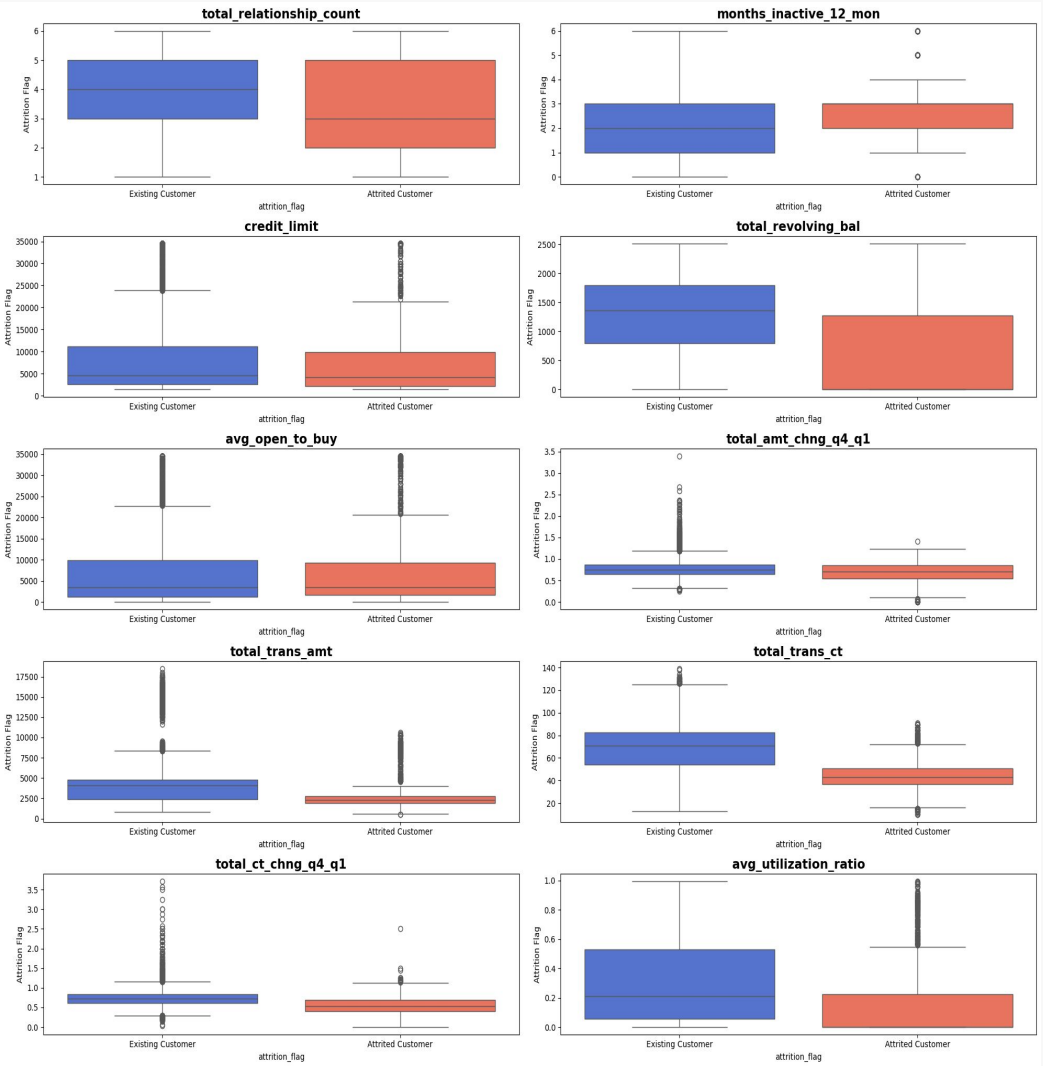
Customer Activity

- **months_on_book**: months they've been a customer.
- **total_relationship_count**: the number of products used by customers in the bank.
- **months_inactive_12_mon**: period of inactivity for the last 12 months.
- **contacts_count_12_mon**: the number of interactions between the bank and the customer in the last 12 months.
- **credit_limit**: credit card transaction nominal limit in one period.
- **total_revolving_bal**: total funds used in period.
- **avg_open_to_buy**: the difference between the credit limit set for the cardholder's account and the current balance.
- **total_amt_chng_q4_q1**: increase in customer transaction nominal between quarter 4 to 1.
- **total_trans_amt**: total nominal transaction in the last 12 months.
- **total_trans_ct**: the number of transactions in the last 12 months.
- **total_ct_chng_q4_q1**: the number of customer transactions increased between quarter 4 and quarter 1.
- **avg_utilization_ratio**: percentage of credit card usage.

21 Columns, 10127 Rows

20 Features, 1 Target

0 Missing Values, 0 Duplicated



Exploratory Data Analysis

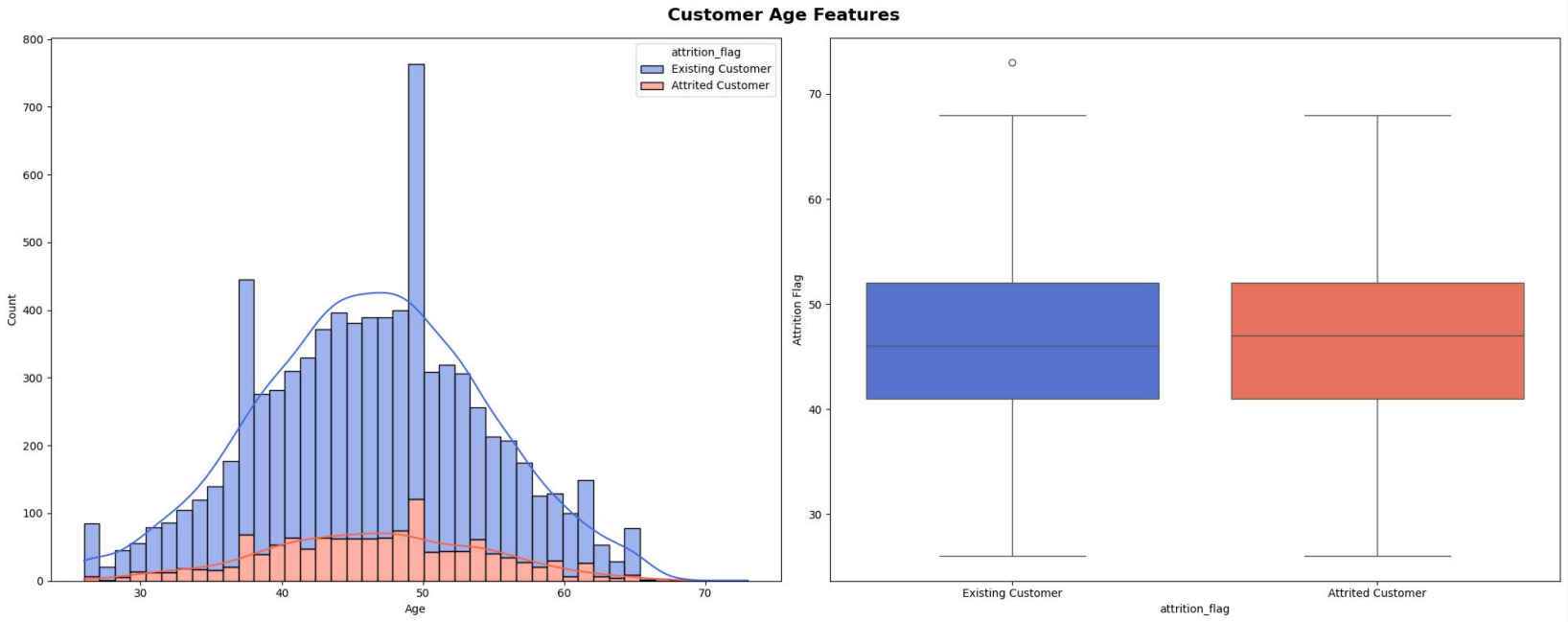
The distribution of numerical features reveal several key differences between Attrited and Existing Customers:

- **Inactivity:** Attrited Customers tend to be more inactive, with a higher median for **months_inactive_12_mon**.
- **Credit Utilization:** Attrited Customers have a higher median **avg_utilization_ratio**, suggesting they are more likely to max out their credit limits.
- **Financial Activity:** Existing Customers generally exhibit higher levels of financial activity, with higher median values for **total_trans_amt**, **total_trans_ct**, and **avg_open_to_buy**.

These findings suggest that inactivity and high credit utilization are strong indicators of customer attrition.

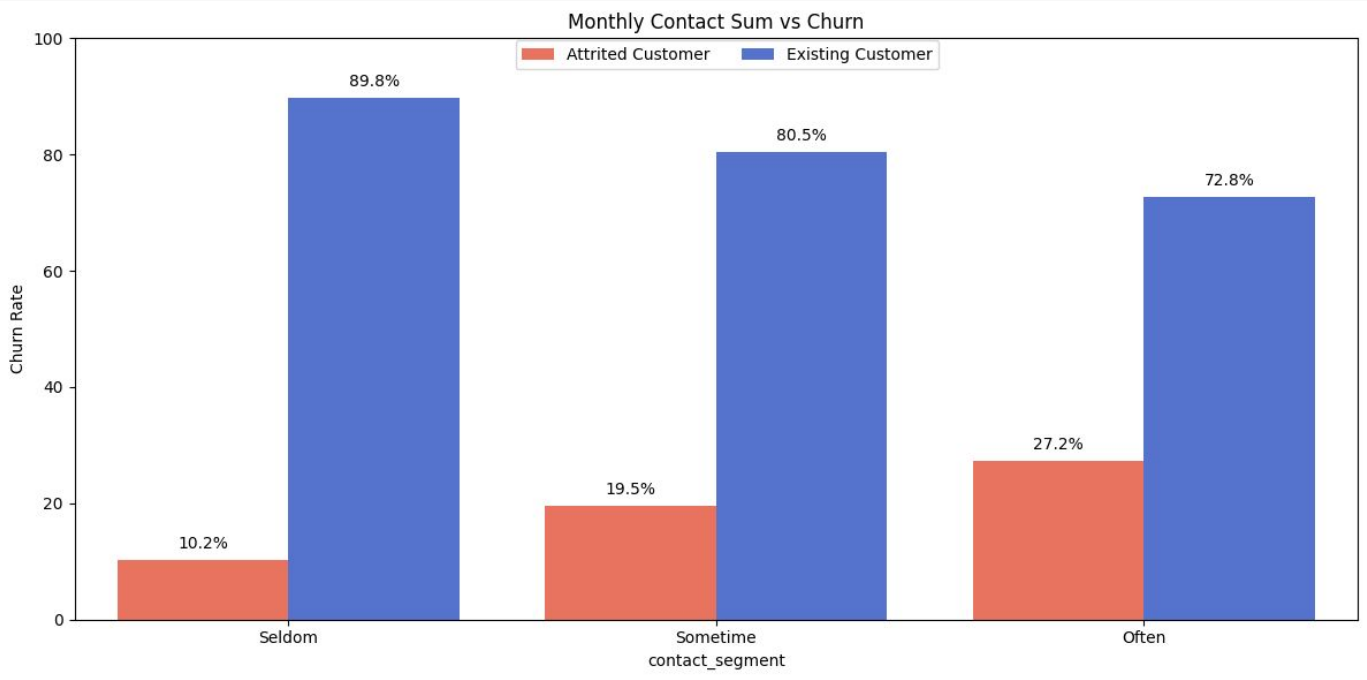
Exploratory Data Analysis

customer_age is normally distributed. And the age range of 40-55 years tends to be high for attrition.



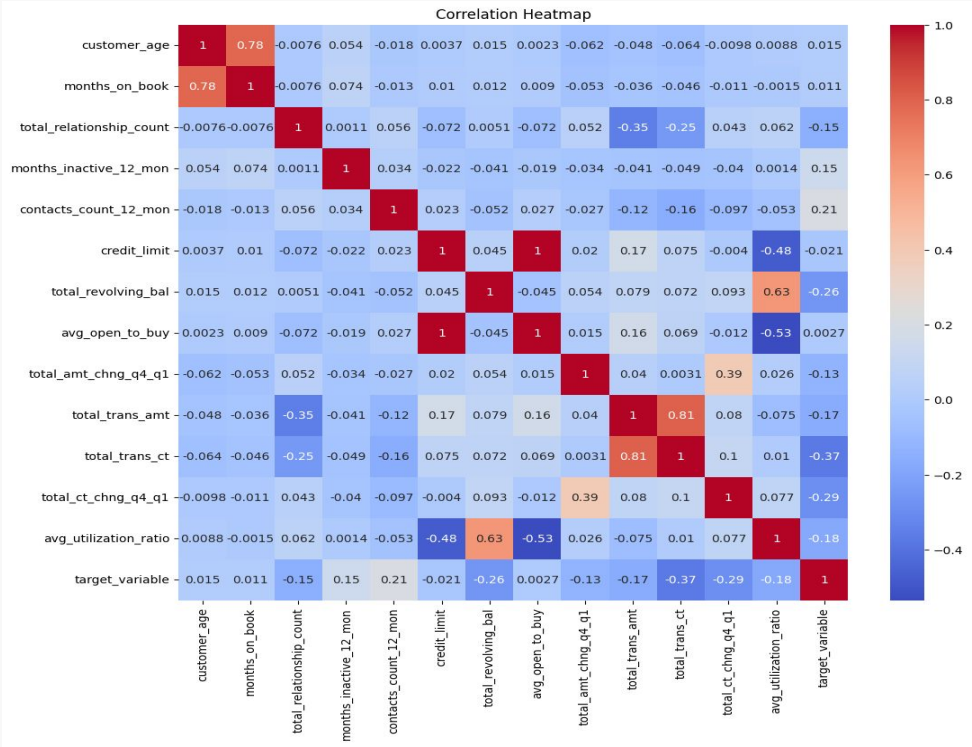
Exploratory Data Analysis

The more often (4-6 times) the customers interacted with the bank in one year, the more likely the customers were to attrition at a rate of 28% compared to those who did so seldom or only sometime.



Data Pre-Processing

Feature Selection



In the heatmap, we can see (threshold > 0.8) strong correlations between:

- **total_trans_amt** and **total_trans_ct** (0.82)
- **avg_open_to_buy** and **credit_limit** (1)
- **customer_age** and **months_on_book** (0.78)

Based on those correlation with the target_variable (attrition_flag). We'll drop **avg_open_to_buy, total_trans_ct features**, and **months_on_book.**

And I'll drop features that don't significantly impact the target **attrition_flag** based on EDA. The aims is to avoid complexity, simplify the model, making it easier to interpret and potentially improving performance by reducing noise. Then we'll be comparing the models with and without these features using metrics.

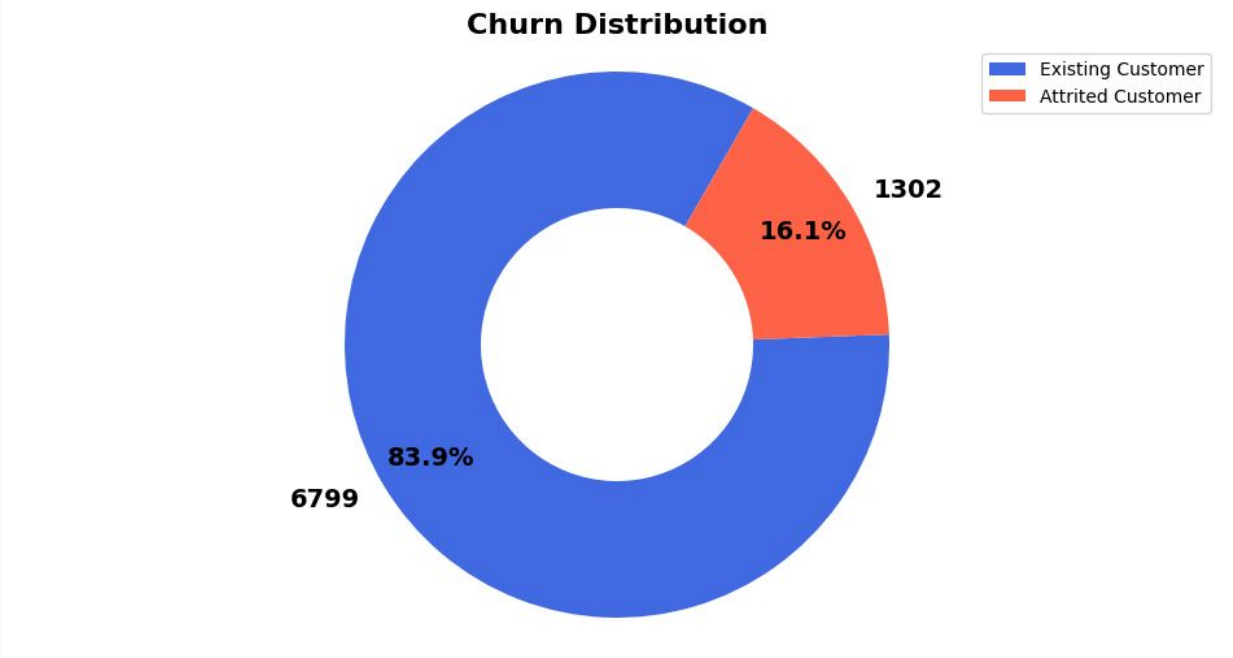
Encoding Categorical Data

Encoding Target **attrition_flag** into:
'Existing Customer': 0, **'Attrited Customer**': 1

keep = [
'customer_age',
'months_inactive_12_mon',
'contacts_count_12_mon',
'total_revolving_bal',
'total_amt_chng_q4_q1',
'total_trans_amt',
'total_relationship_count',
'total_ct_chng_q4_q1',
'credit_limit']

Machine Learning Modeling & Evaluation

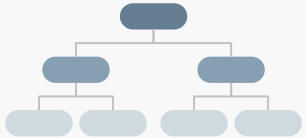
Due to imbalance of target feature **attrition_flag** where class 0 ("Existing Customer" 83.9 %) has significantly more samples than the class 1 ("Attrited Customer" 16.1 %). In this case, I think if using metric accuracy score can be misleading. And metric **recall score** becomes a more important metric because it focuses on correctly identifying the minority class (attrited customers). By minimizing false negatives (FN) is crucial because we want to avoid missing potential attrition cases, which is the primary business concern here.



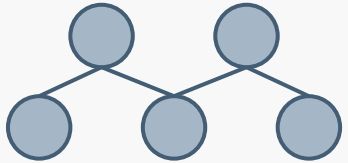
Machine Learning Modeling & Evaluation

3 Tree-Based models below to compare and choose the best performance model.

Decision Tree



Random Forest



XGboost

Model	F1-Score Train	F1-Score Test	Recall Train	Recall Test
XGBoost	1.00	**0.81**	1.00	**0.85**
Random Forest	1.00	0.80	1.00	0.80
Decision Tree	1.00	0.68	1.00	0.77

The **XGBoost model is the top performer**, excelling in identifying true attrition cases. To further evaluate its performance, we'll use a classification report, confusion matrix, and ROC-AUC curve.

Machine Learning Modeling & Evaluation

We **escalated the performance** after tuning the XGBoost model with the best parameters from RandomizedSearchCV method when the Training Metrics no longer showed a perfect score and the Testing Metrics improved.

Best Parameters: {
'**colsample_bytree**': 0.9147897164123967,
'**gamma**': 0.19448538911715274,
'**learning_rate**': 0.28414499077183736,
'**max_depth**': 14, 'n_estimators': 93,
'**reg_alpha**': 2.743245814889095,
'**reg_lambda**': 97.902347572739,
'**subsample**': 0.5924332538341432
}

Score	Training	Testing
F1-score	0.86	**0.84**
Recall	0.89	**0.82**
Precision	0.93	**0.85**

And to gain a more comprehensive understanding, we will use a classification report score, confusion matrix to visualize true positives, true negatives, false positives, and false negatives. Then ROC-AUC visualization will help assess the model's ability to distinguish between the two classes.

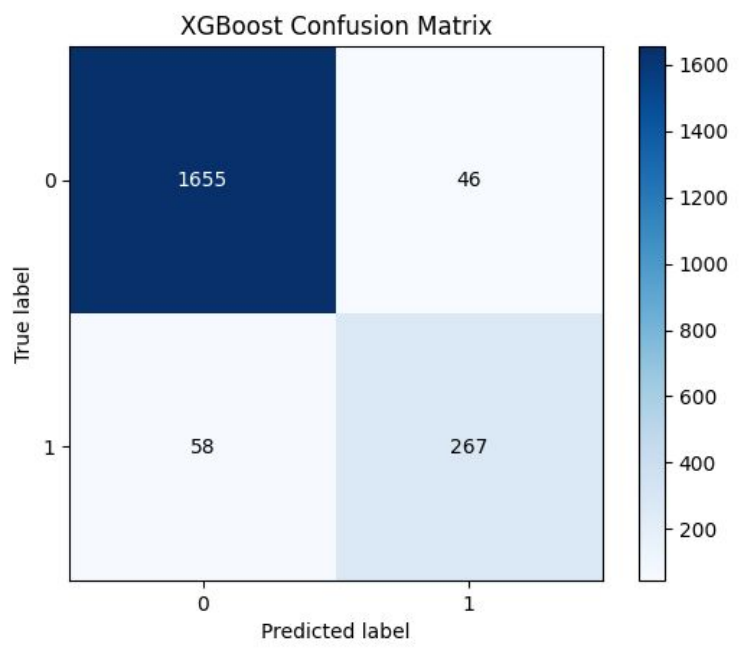
Machine Learning Modeling & Evaluation

The **XGBoost** model demonstrates strong performance in predicting customer attrition. It has a **high recall score of 0.82 for the "Attrited Customer" class**, means **it successfully identifies a large proportion of customers who are likely to** attrition. Despite the dataset imbalance, the XGBoost model shows promising results in identifying the minority class. The high recall score is particularly valuable in this scenario, as it minimizes the risk of missing potential attrition cases.

	Precision	Recall	F1-Score	Support
0	0.97	0.97	0.97	1701
1	0.85	**0.82**	**0.84**	325
Accuracy			0.95	2026
Macro Avg	0.91	0.90	0.90	2026
Weighted Avg	0.95	0.95	0.95	2026

- **Precision**: *Out of all the customers predicted as "Attrited" (churn), what proportion was actually "Attrited"?* Precision for class 1 (Attrited Customer) is 0.86, means **86% of the customers predicted to churn actually did churn.**
- **Recall**: *Out of all the actual "Attrited" customers, what proportion did the model correctly identify?* Recall for class 1 is 0.82, means **the model correctly identified 82% of the customers who actually churned.**
- **F1-score**: The harmonic mean of precision and recall, provides a balanced measure of the model's performance. F1-score for class 1 is 0.84, representing a good balance between precision and recall.
- **Accuracy**: The overall proportion of correctly classified instances. The model achieved an accuracy of 0.95, meaning it correctly classified 95% of the customers in the test set. However, accuracy can be misleading in imbalanced datasets.

Machine Learning Modeling & Evaluation



- **True Negatives (TN)**: 1655 customers were correctly predicted as not churning.
- **False Positives (FP)**: 46 customers were incorrectly predicted as churning when they actually did not.
- **False Negatives (FN)**: 58 customers were incorrectly predicted as not churning when they actually did.
- **True Positives (TP)**: 267 customers were correctly predicted as churning.

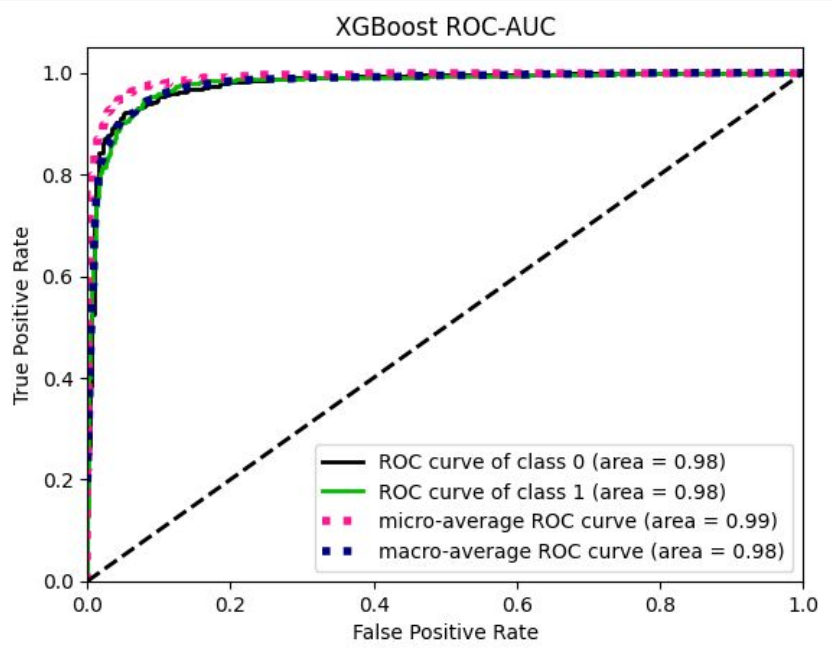
From 2026 rows of test data, the FN number seems lower when it has fewer of the number predicted class 1 (58 of total class 1 = 325) than the TP. When the aim is to minimize the FN, I think this XGBoost model is very well to represent it.

False Negative Rate: 0.1785

FNR = FN / (FN + TP)

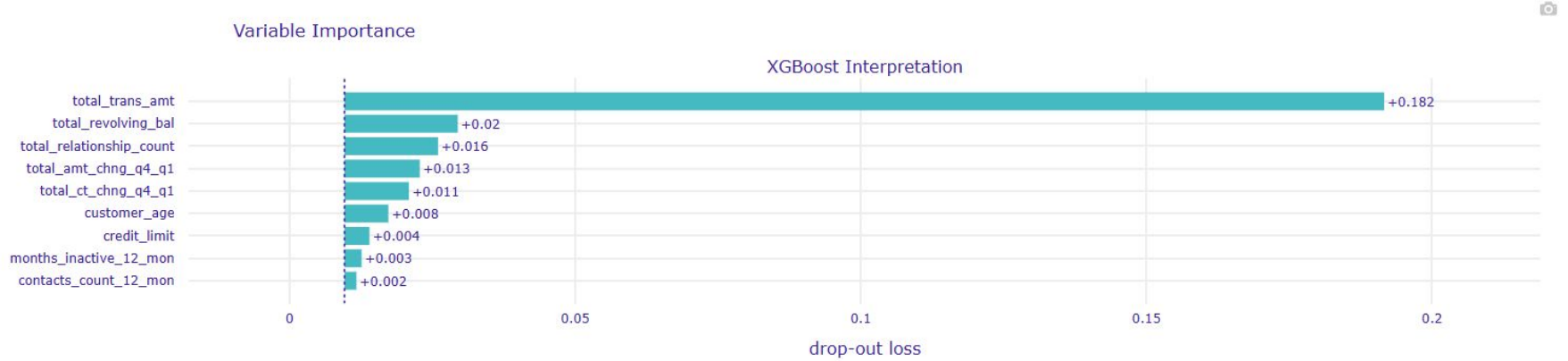
Then an FNR of 0.1785 means that 17.85% of customers who actually churned were incorrectly predicted by the model as not churning. In other words, for every 325 customers in the test data who churned, the model missed about 56-58 of them.

Machine Learning Model Interpretation



From the ROC-AUC plot, **it shows a curve that hugs the top-left corner, indicating high TPR and low FPR and the auc values for both classes are high (0.98), further confirming their strong performance**. It means that the XGBoost model demonstrates exceptional performance in predicting both classes. The high auc values for both individual classes and the overall model suggest that it is highly accurate and reliable in distinguishing between positive and negative cases.

Machine Learning Model Interpretation



Here are the features importance from XGBoost model's prediction:

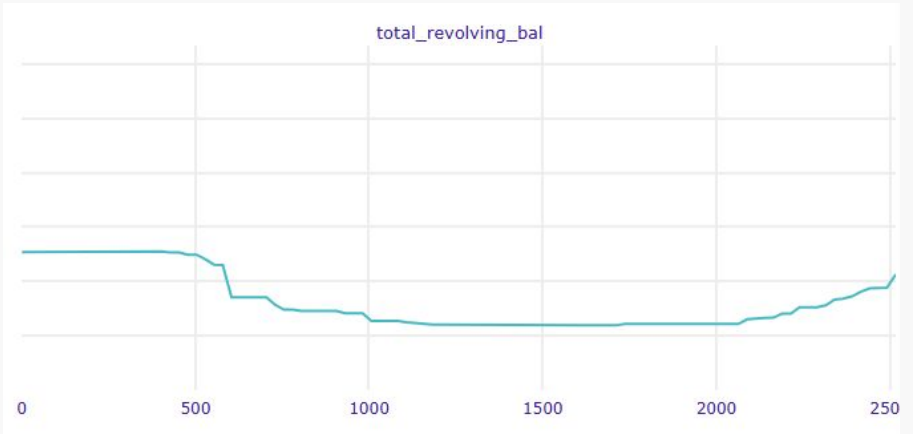
- **total_trans_amt**: The highest importance, suggesting that total nominal transaction in the last 12 months is a crucial factor in predicting the target variable.
- **total_revolving_bal**: Indicates that changes in the total funds which is used in a period are important predictors.
- **total_relationship_count**: Indicates that more or less the number of products used by customers in the bank are important predictors.
- **total_amt_chng_q4_q1 & total_ct_chng_q4_q1**: Also have significant importance, indicating that changes in transaction amounts are important predictors.
- **customer_age**: Have relatively low importance, suggesting that the age impact on the model's predictions is limited.

And the rest is relatively low importance too on model's predictions.

Machine Learning Model Interpretation

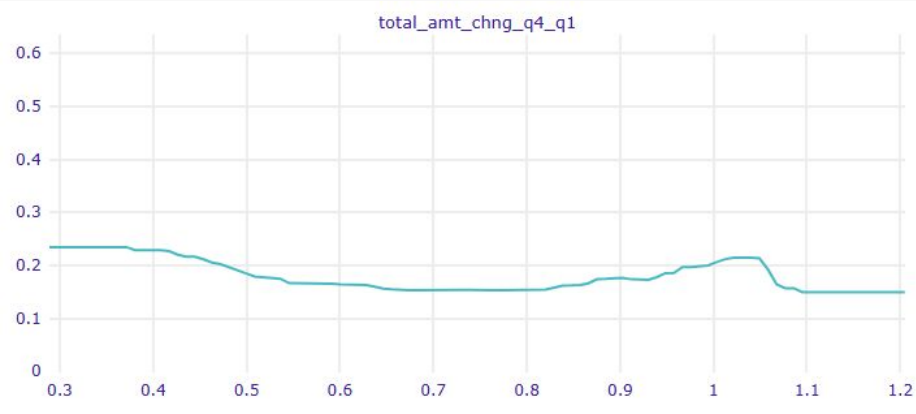
`total_trans_amt` (Total Transaction Amount): Higher amounts are generally associated with lower predicted values. Customers exhibiting a decreasing trend in transaction amounts pose a higher risk of attrition. This could signal a shift in customer spending habits and indicate potential dissatisfaction.

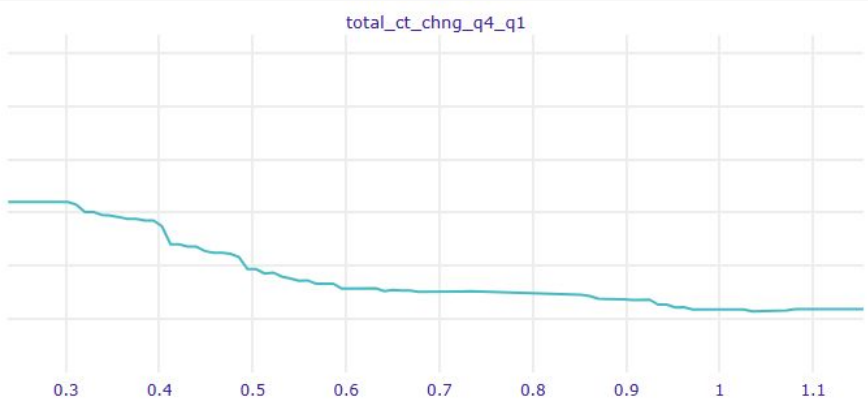
`total_revolving_bal` (Total Revolving Balance): Higher balances are associated with higher predicted values. Customers with higher balances are more likely to attrition. This could indicate a greater reliance on credit, which can be a risk factor.



Machine Learning Model Interpretation

`total_amt_chng_q4_q1` & **`total_ct_chng_q4_q1`** **(**Total Amount and Credit Change Q4-Q1): A decrease in spending from Q4 to Q1 is also associated with an increased attrition probability. This emphasizes the need to address any negative change in customer spending patterns.





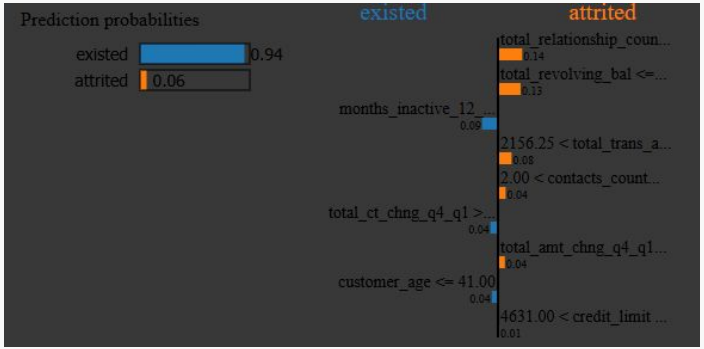
`total_relationship_count`: Indicates that more or less the number of products used by customers in the bank is an important predictor. Customers who had less number of products (1-3) are associated with an increased attrition probability while the more products customers took, the more they still loyal.

Machine Learning Model Interpretation

User ID : **713426358**

The Shapley value of **-2.76** suggests this user has a **low attrition risk**. The model attributes this to high total transaction and activity, indicating greater customer engagement. The user's high total transaction and activity is the main reason the model predicts a low attrition risk.

The LIME plot shows that the model predicts this user with a **high probability of existing (94%)** and a low probability of attrition (6%). That indicates that the model predicts a high probability of existence for this customer due to factors like high transaction activity, positive changes in transaction behavior, and customer engagement. The model also suggests that younger customers (below 41) are more likely to exist.

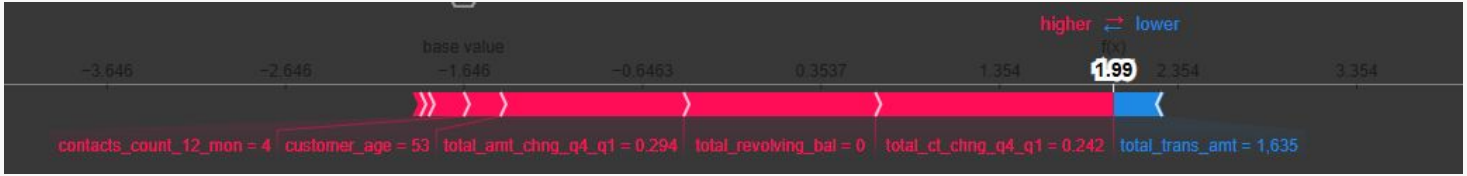


Feature	Value
total_relationship_count	3.00
total_revolving_bal	0.00
months_inactive_12_mon	2.00
total_trans_amt	2395.00
contacts_count_12_mon	3.00
total_ct_chng_q4_q1	1.00
total_amt_chng_q4_q1	1.33
customer_age	28.00
credit_limit	5685.00

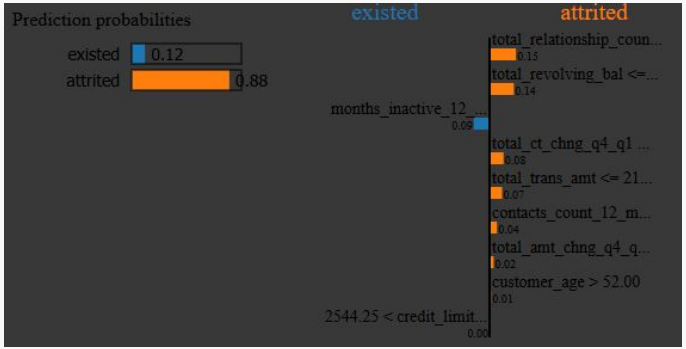
Machine Learning Model Interpretation

User ID : **716564433**

The Shapley value of **1.99** suggests this user has a **high attrition risk**. The model attributes this to low transaction activity, indicating lower customer engagement. The user's low transaction activity is the main reason the model predicts a high attrition risk.



While the LIME plot shows that the model predicts this user with a **high probability of attrition (88%)** and a low probability of existing (12%). The model predicts a low probability of existence for this customer due to factors like low transaction activity and negative changes in transaction behavior. The model also suggests that older customers are more likely to churn.



Feature	Value
total_relationship_count	3.00
total_revolving_bal	0.00
months_inactive_12_mon	2.00
total_ct_chng_q4_q1	0.24
total_trans_amt	1635.00
contacts_count_12_mon	4.00
total_amt_chng_q4_q1	0.29
customer_age	53.00
credit_limit	4287.00

Conclusion

Features driving attrition:

- Higher Impact on attrition probability
  - **Total transaction amount**
  - **Total amount and credit change Q4-Q1**
  - **Total number of products used by customers**
  - **Total revolving balance**

- Smaller Impact on attrition probability
  - **Customer age**
  - **Credit limit**
  - **Customer contacts count on 12 months**
  - **Months of the customer is Inactive in a year**

These factors individually had a smaller impact than the financial parameters mentioned above, they still contribute to attrition predictions. Addressing customer concerns proactively, ensuring ease of use, and providing relevant information can potentially reduce attrition risks.

Conclusion

The **XGBoost model** effectively predicts customer attrition with a **high recall score** of **0.82**, meaning it accurately identifies a large portion of customers likely to churn. This is valuable despite data imbalance and minimizes the risk of missing potential attrition cases.

Therefore, by correctly identifying 82% of churning customers, **the bank can potentially prevent 82% of the total potential loss** due to customer attrition.

Assumptions:

- Total customers: 2026
- Churn rate: 16.1% (0.161)
- CLTV (Customer Lifetime Value): 1,000,000
- Attrition Cost: 1,000,000
- Recall score: 0.82 (82% of churned customers identified)

Calculations:

- Number of churned customers: 2026 * 0.161 = 326.186
- Potential Loss: 326.186 customers * 1,000,000 (Attrition Cost) = 326,186,000
- Customers Saved: 2026 * 0.161 (Churn Rate) * 0.82 (Recall Score) = 267.95 customers
- Loss Prevented: 267.95 customers * 1,000,000 (Attrition Cost) = 267,950,000
- Percentage of Loss Prevented: (267,950,000 / 326,186,000) * 100 = 82%

Recommendations for Bank to Prevent Customers from Attrition:

1. Develop tailored offers and incentives for at-risk customers to address their specific needs and potentially prevent attrition. This could include exclusive promotion per customer, banking product bundles, or loyalty programs.
2. Engage proactively with customers experiencing a decrease in transaction amounts or changes in spending patterns. It can help identify potential issues early on and offer solutions before they lead to attrition.
3. Ensure exceptional customer service and support to address concerns promptly and effectively. It can help prevent customer dissatisfaction and boost retention rates.
4. Implement loyalty programs and rewards for long-term customers. Recognizing and appreciating their continued relationship with the bank can significantly reduce attrition.
5. Utilize digital channels to communicate with customers, offer support, and collect feedback. It can improve customer engagement and satisfaction.

By implementing these recommendations, the bank can make data-driven decisions to improve customer retention and prevent potential losses due to attrition.

Thanks!

Do you have any questions?

Email
LinkedIn
Github

Link to Code