Customer Churn Analysis

Brown University, Data Science Institute
Zhirui Li

1. Introduction

As the banking sector faces an increasing trend of customers abandoning their credit card services, managing customer churn has become a paramount concern for bank management. This paper delves into the nuanced challenge of predicting customer churn. This endeavor is critical without direct customer feedback and essential for enabling proactive customer retention strategies. Our analysis demonstrates how predictive modeling can empower Customer Relationship Management (CRM) and customer experience teams, potentially curtailing customer attrition by up to 11% with timely interventions [1].

Leveraging a comprehensive dataset from LEAPS [2], which includes 21 features across 10,127 data samples, we focus on the ‘Attrition_Flag’ variable to identify 'Attrited Customers.' Our approach combines advanced analytical techniques with practical insights, offering a robust framework for churn prediction. By integrating machine learning algorithms and exploring feature interrelationships, we aim to provide bank managers with actionable intelligence to enhance customer engagement and loyalty.

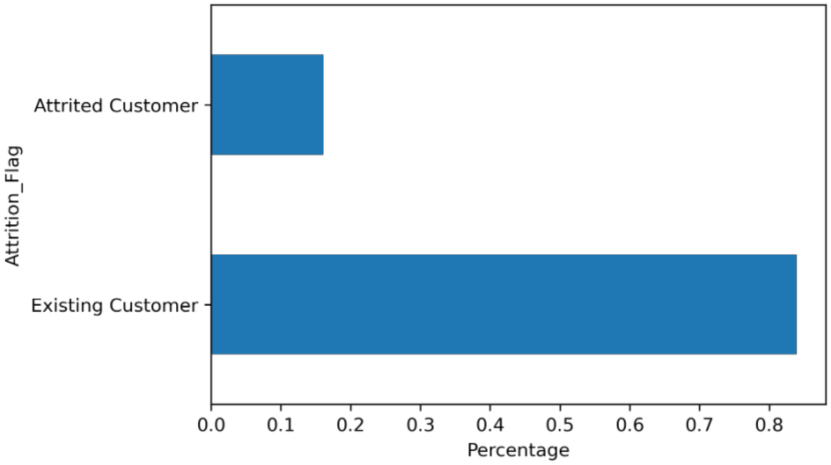
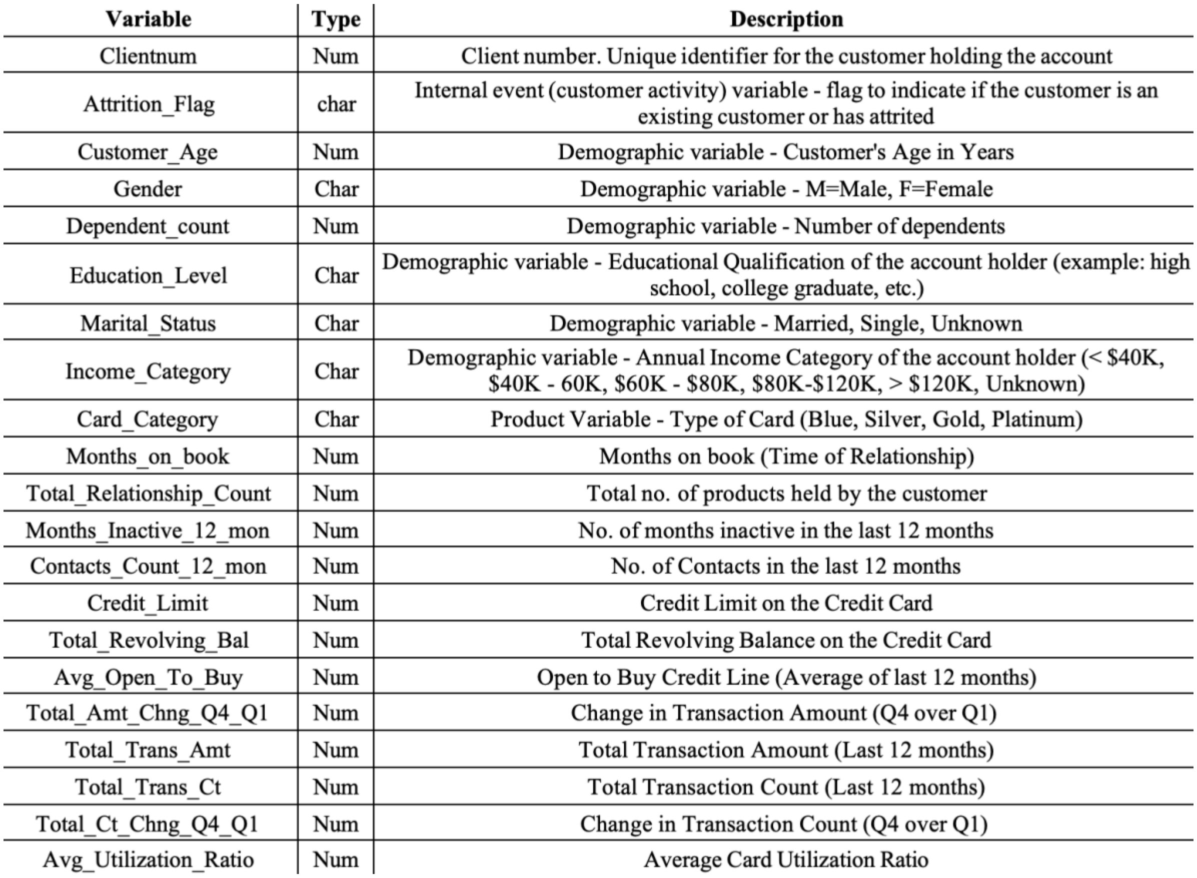


Figure 1 The distribution of ‘Attrition_Flag’. It is an imbalanced dataset since the proportion of the ‘Attrited customer’ is about 16.07%.



Variable	Type	Description
Clientnum	Num	Client number. Unique identifier for the customer holding the account
Attrition_Flag	char	Internal event (customer activity) variable - flag to indicate if the customer is an existing customer or has attrited
Customer_Age	Num	Demographic variable - Customer's Age in Years
Gender	Char	Demographic variable - M=Male, F=Female
Dependent_count	Num	Demographic variable - Number of dependents
Education_Level	Char	Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.)
Marital_Status	Char	Demographic variable - Married, Single, Unknown
Income_Category	Char	Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, > $120K, Unknown)
Card_Category	Char	Product Variable - Type of Card (Blue, Silver, Gold, Platinum)
Months_on_book	Num	Months on book (Time of Relationship)
Total_Relationship_Count	Num	Total no. of products held by the customer
Months_Inactive_12_mon	Num	No. of months inactive in the last 12 months
Contacts_Count_12_mon	Num	No. of Contacts in the last 12 months
Credit_Limit	Num	Credit Limit on the Credit Card
Total_Revolving_Bal	Num	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Num	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Num	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Num	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Num	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Num	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Num	Average Card Utilization Ratio

Table 1 Description of the rest 20 features

Our primary objective with this dataset is twofold: firstly, to enhance the prediction accuracy of churned customers, and secondly, to identify critical factors influencing customer churn. Given their interrelated nature, these tasks are effectively combined into one. Previous studies in this domain have yielded notable results. Thomas [3] applied the SMOTE technique to balance the dataset, significantly boosting the F1 score from 0.6 to 0.9. Andi's [4] exploratory data analysis revealed correlations between customer churn and variables such as annual expenditure, months of inactivity, and credit limits. Joseph [5] achieved a recall of 97% and 95% accuracy using Random Forest and LightGBM, highlighting transaction features as particularly influential. These insights emphasize the importance of a thorough transactional data analysis in my approach.

2. EDA

I've analyzed the correlation between each feature and the target variable, uncovering several noteworthy relationships.

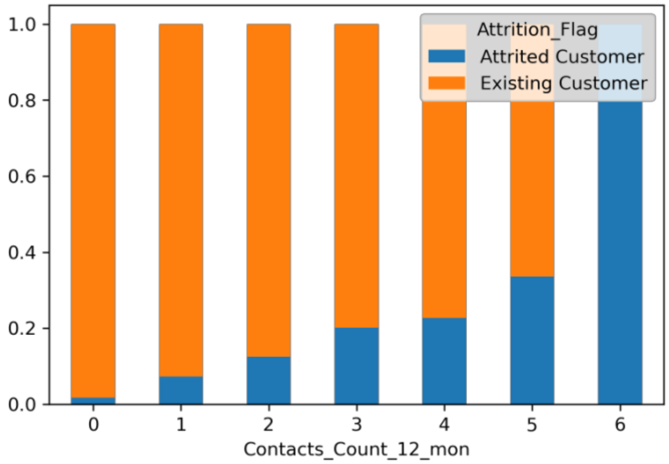


Figure 2 This graph illustrates the distribution of contacts between two customer groups. Customers who churned had more frequent interactions with bank managers over the past 12 months.

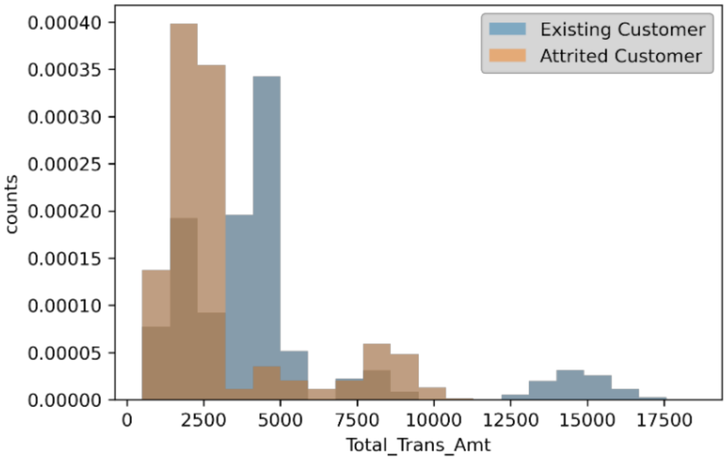


Figure 3 This graph demonstrates that attrited customers generally have a lower total transaction amount than existing customers, accounting for this feature's high ranking in both models used in Joseph's studies.

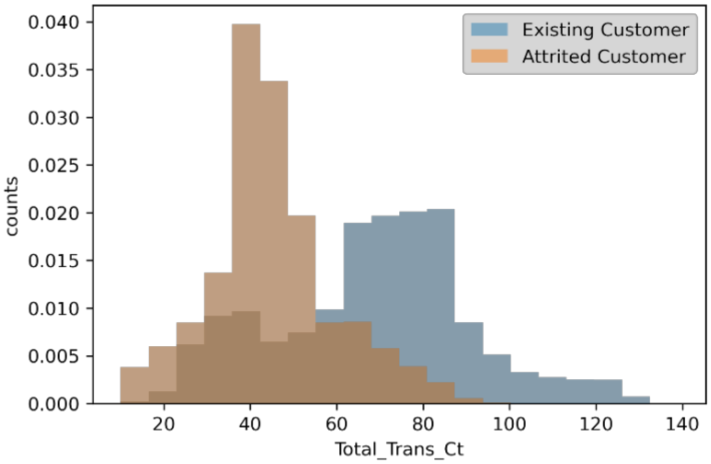


Figure 4 This graph exhibits a pattern akin to Figure 3, indicating that total transaction count is a significant feature in distinguishing attrited customers from existing ones.

3. Method

3.1 Preprocessing

This dataset is not Independent and Identically Distributed (IID) as the features show varying distributions, though it lacks group structures or time-series elements. Due to its imbalance, a stratified split approach is used, with 60% of data allocated for training and 20% each for validation and testing — a standard split for large datasets.

Discrete numerical features (dis_fea) are treated as ordinal and used directly in the model. Categorical features (cat_fea) undergo One Hot Encoding, as ordering them (like gender or marital status) isn't logical. Ordinal features (ord_fea) such as educational level, income category, and card category are processed with Ordinal Encoder to preserve their inherent order. For continuous features (con_fea), Standard Scaler is chosen, considering the near-normal distribution of 'Customer_Age,' 'Months_on_book,' and 'Total_Trans_Ct,' despite some skewness and long-tailed distributions in other features.

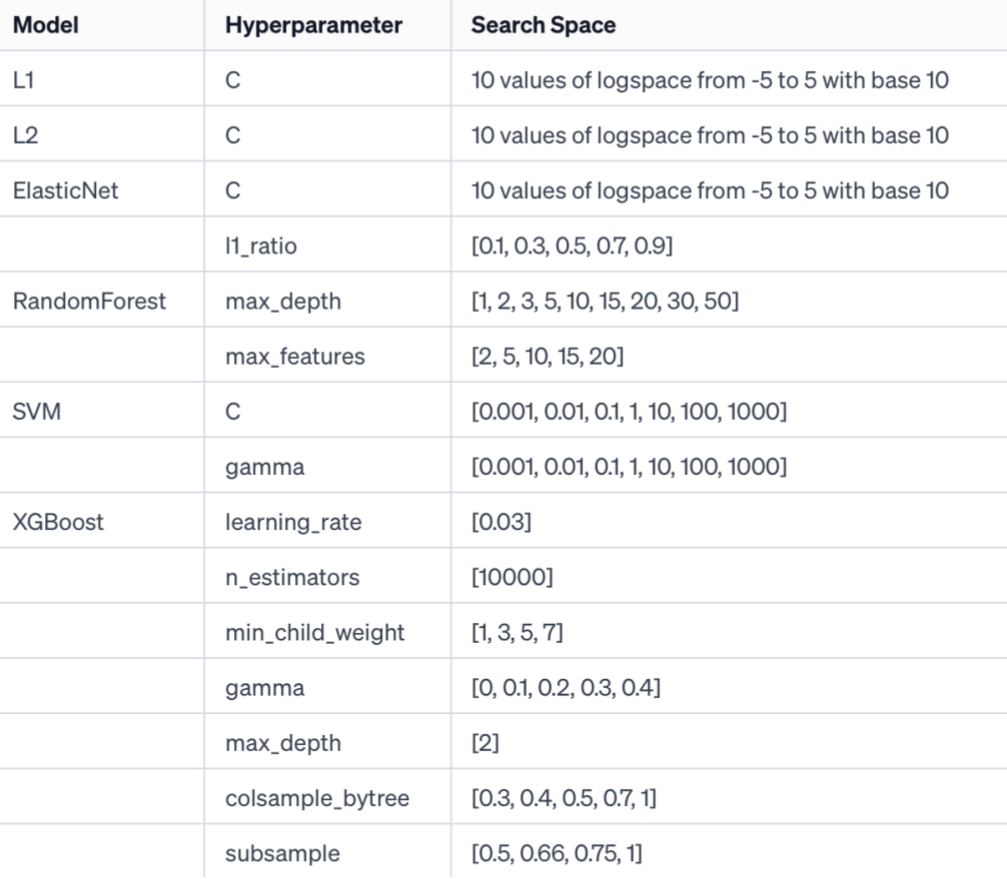
The target variable is transformed for machine comprehension in this binary classification problem: "Existing Customer" to 1 and "Attrited Customer" to 0. Given their categorical nature, missing values in demographic variables like educational level, income category, and marital status are categorized separately.

The dataset comprises 24 features with 6076, 2025, and 2026 data points in the training, validation, and test sets.

3.2 Parameter tuning

Our Machine Learning pipeline is straightforward. For each of the five random states, the data is split and preprocessed as outlined earlier. We then evaluate all parameter combinations on the training and validation datasets to identify the optimal model for each state, subsequently computing the test score for these models. This process yields five test scores per model, one for each random state.

This project explores six distinct models: three logistic regression models with varying regularization methods, along with SVM, Random Forest, and XGBoost. The parameters for each model are detailed in the table below.



Model	Hyperparameter	Search Space
L1	C	10 values of logspace from -5 to 5 with base 10
L2	C	10 values of logspace from -5 to 5 with base 10
ElasticNet	C	10 values of logspace from -5 to 5 with base 10
	l1_ratio	[0.1, 0.3, 0.5, 0.7, 0.9]
RandomForest	max_depth	[1, 2, 3, 5, 10, 15, 20, 30, 50]
	max_features	[2, 5, 10, 15, 20]
SVM	C	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
	gamma	[0.001, 0.01, 0.1, 1, 10, 100, 1000]
XGBoost	learning_rate	[0.03]
	n_estimators	[10000]
	min_child_weight	[1, 3, 5, 7]
	gamma	[0, 0.1, 0.2, 0.3, 0.4]
	max_depth	[2]
	colsample_bytree	[0.3, 0.4, 0.5, 0.7, 1]
	subsample	[0.5, 0.66, 0.75, 1]

Table 2 Parameters used for tuning

We selected the F2 score as our evaluation metric, prioritizing recall due to the significant impact of misclassifying attrited customers as existing ones. This choice reflects the higher cost associated with incorrectly assuming customer retention than wrongly predicting customer attrition. The F2 score, emphasizing recall, effectively addresses this imbalance by placing greater weight on correctly identifying attrited customers.

4. Results

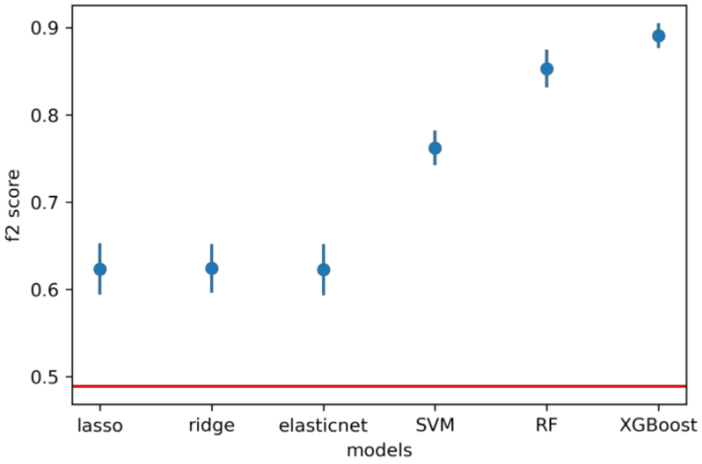


Figure 5 F2-score for the six models over five random states

The error bar plot above reveals that all six models surpass the baseline F2 score, where the baseline model predicts all instances as "Attrited Customer." XGBoost has the highest average test score and most minor variance, performing approximately 29 standard deviations better than the baseline.

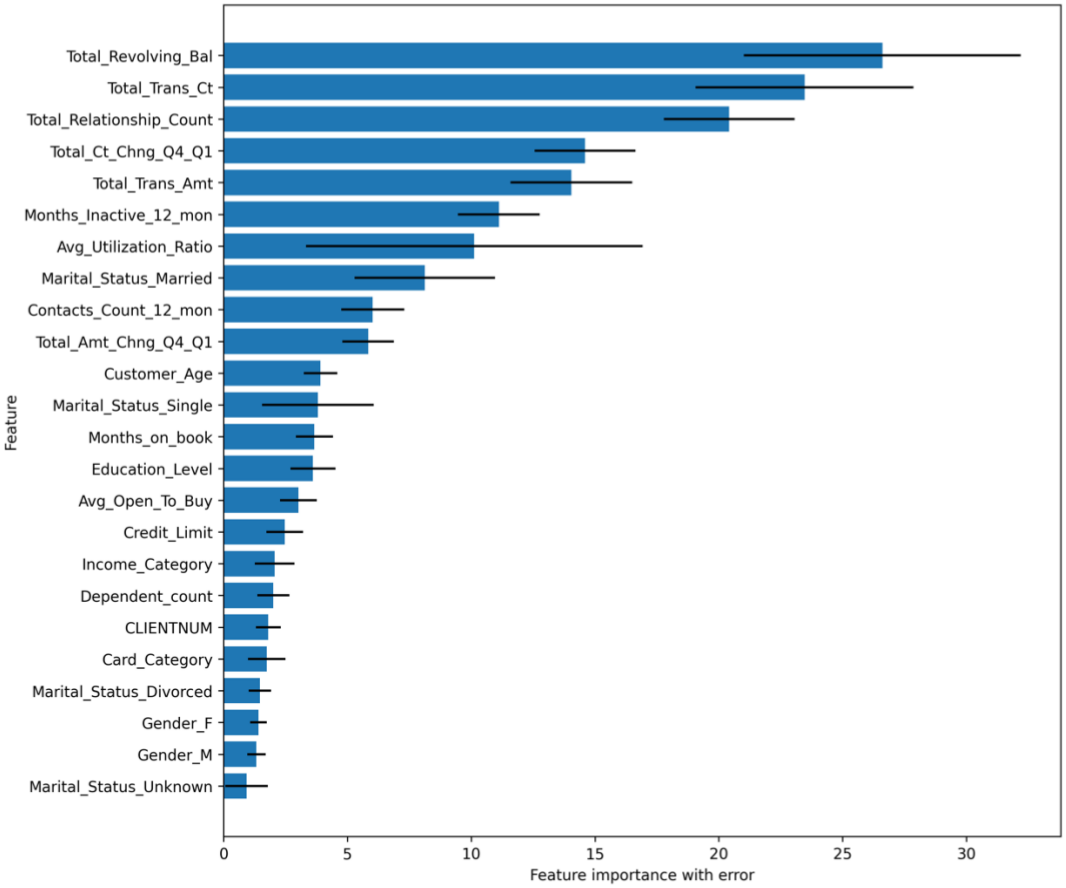


Figure 6 Feature importance for XGBoost models using Gain as the importance metric

"Gain" is utilized to assess feature importance, a key metric for interpreting the relative significance of each feature [6]. Our findings show that in XGBoost models, the most impactful features are 'Total_Revolving_Bal,' 'Total_Trans_Ct,' and 'Total_Relationship_Count.' The first two relate to customer banking activities, while the latter is a demographic indicator. In contrast, 'x1_Divorced', 'Income_Category,' and 'Card_Category' are the least influential features.

Figure 7 demonstrates the importance of local features for two specific data points. The base value of -1.8 reflects the average model output across the test dataset, with red features indicating an increase and blue a decrease in the prediction score. For instance, in the upper graph, 'Total_Trans_Ct' (-1.3) increases the likelihood of attrition, while 'Total_Trans_Amt' (-0.9) reduces it. In the lower graph, 'Total_Relationship_Count' (2.0), 'Total_Amt_Chng_Q4_Q1' (-1.8), and 'Total_Trans_Ct' (-1.2) are shown to have the most significant increasing effect on the prediction.

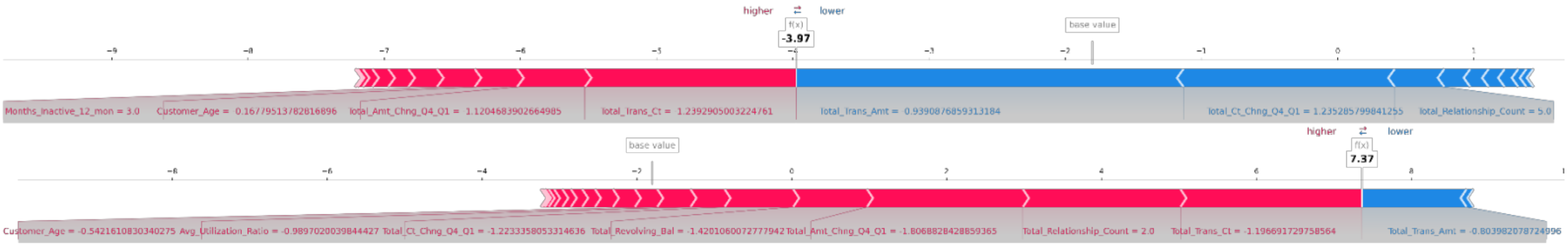


Figure 7 Feature contribution of data points (upper: existing customer; lower: attrited customer)

Additionally, the scatter plot of SHAP and feature values reveals intriguing insights into how feature values influence the model's output.

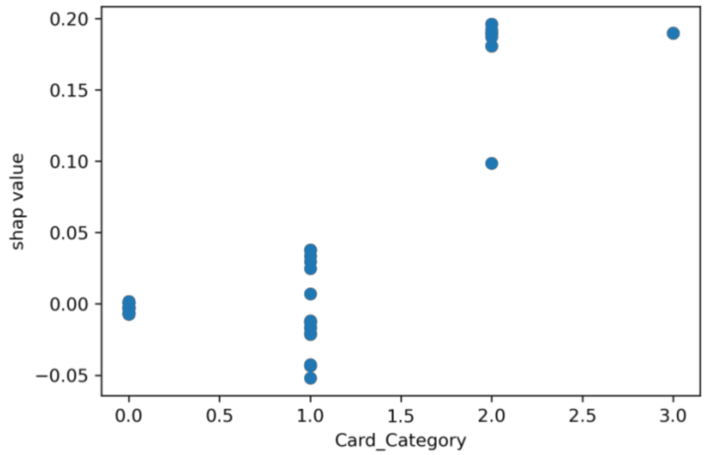


Figure 8 indicates that customers with higher-value card holdings are more likely to be predicted as attrited customers.

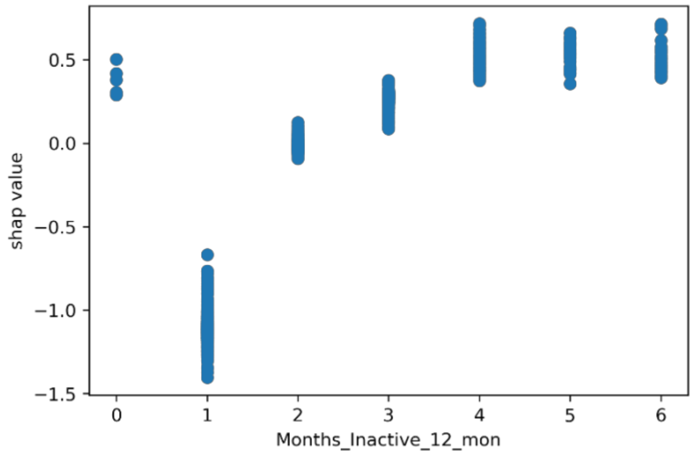


Figure 9 illustrates that more inactive months increase the likelihood of customers being predicted as attrited.

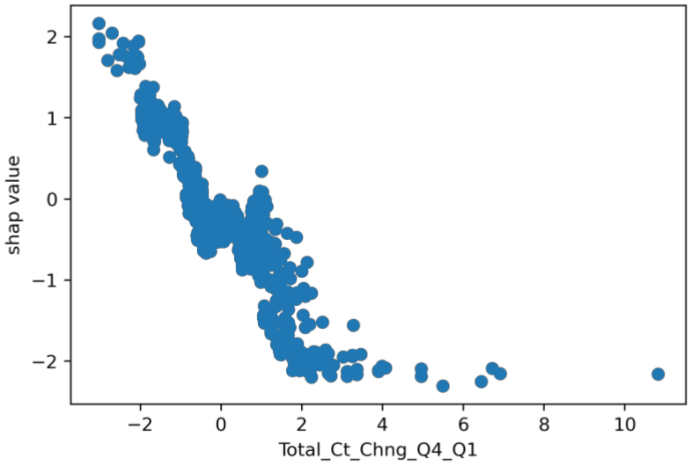


Figure 10 demonstrates that an increase in total transaction count from Q1 to Q4 correlates with a lower probability of a customer being predicted as attrited.

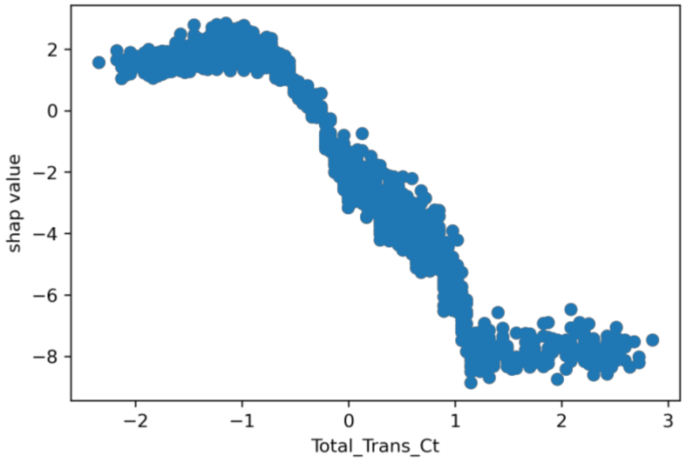


Figure 11 indicates that customers with a higher total transaction count are less likely to be predicted as attrited.

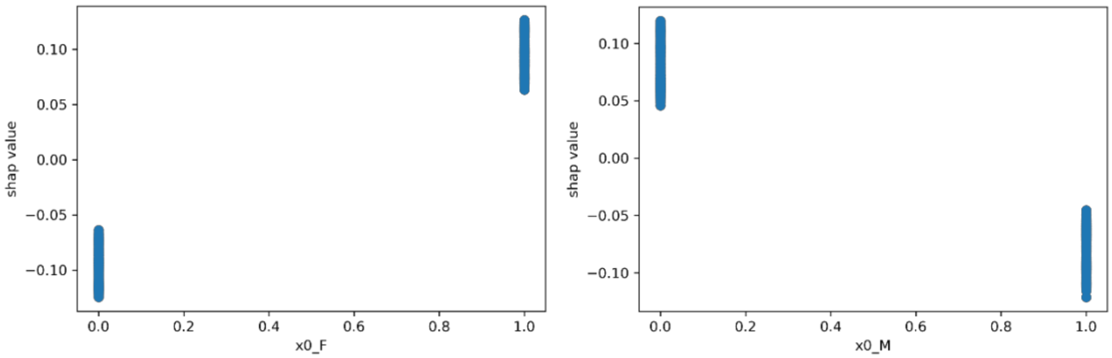


Figure 12 presents scatter plots of SHAP values against the gender feature, revealing a trend where female customers are more likely to be predicted as attrited than male customers.

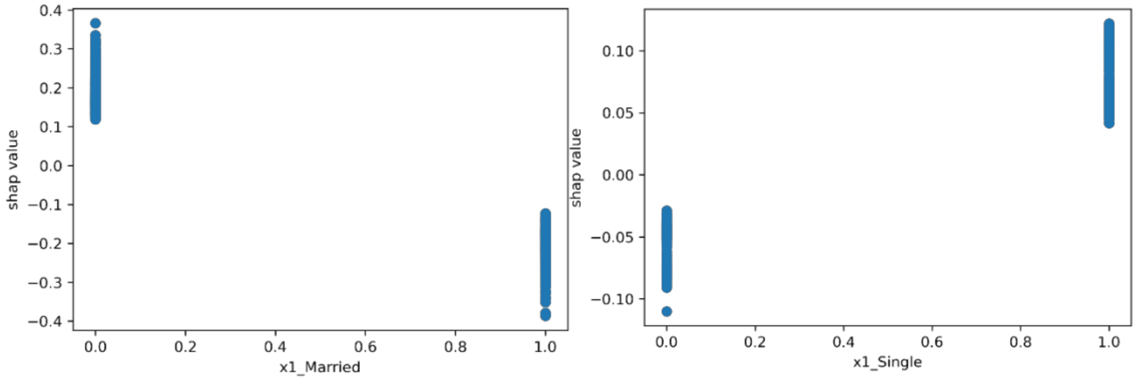


Figure 13 displays scatter plots of SHAP values against marital status, indicating that married customers are less likely to be predicted as attrited than single customers.

5. Outlook

In conclusion, our analysis underscored the importance of examining interrelationships among features to enhance model interpretability in customer churn prediction. Notably, we observed significant patterns in SHAP values, especially between features such as 'Total_Trans_Ct' and 'Total_Ct_Change_Q4_Q1', as well as 'Avg_Open_To_Buy' and 'Credit_Limit.' These insights demonstrate strong associations between transactional and credit-related features and pave the way for innovative feature engineering. The proposed new feature, derived from 'Avg_Open_To_Buy' and 'Credit_Limit,' holds promise for improving predictive accuracy. Furthermore, the adoption of LightGBM and the strategic use of the SMOTE technique for addressing data imbalance show potential for enhancing model performance. Critical to our approach was the reassessment of our evaluation metric to accurately reflect the different impacts of misclassification. This comprehensive analysis provides a robust foundation for future efforts in customer retention strategies, emphasizing the need for continual data enrichment and methodological refinement to predict better and mitigate customer churn.

6. Reference

[1] Why customers leave &amp; what can banks do? Tiger Analytics. (2020, September 16). Retrieved October 12, 2021, from https://www.tigeranalytics.com/blog/addressing-customer-churn-in-banking/.

[2] Predict Customer Attrition Using Naïve Bayes Classification. ATH Leaps. Retrieved October 12, 2021, from https://leapsapp.analyttica.com/cases/11.

[3] Konstantin, T. (2021, May 1). Bank churn data exploration and churn prediction. Kaggle. Retrieved October 12, 2021, from https://www.kaggle.com/thomaskonstantin/bank-churn-data-exploration-and-churn-prediction.

[4] IDW, A. (2021, January 31). Customer churn - EDA, 95% ACC and 85% recall. Kaggle. Retrieved October 12, 2021, from https://www.kaggle.com/paotografi/customer-churn-eda-95-acc-and-85-recall.

[5] Chan, J. (2021, January 13). Bank Churners Classifier (Recall: 97% accuracy: 95%). Kaggle. Retrieved October 12, 2021, from https://www.kaggle.com/josephchan524/bankchurnersclassifier-recall-97-accuracy-95.

[6] Abu-Rmileh, A. (2021, September 2). Be careful when interpreting your features importance in xgboost! Medium. Retrieved December 5, 2021, from https://towardsdatascience.com/be-careful-when-interpreting-your-features-importance-in-xgboost-6e16132588e7.

7. GitHub repository

- https://github.com/ZhiruiLi1/Customer_Churn_Analysis