Advanced Machine Learning

(Credit card users Churn Prediction)

PGP - AIML - University of Texas at Austin

December, 2024.

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution approach
- EDA - Univariate Analysis
- Data Processing
- Model performance summary for hyperparameters tuning.
- Appendix

Executive Summary

Total Transaction Count (Total_Trans_Ct), as well Total Transaction Amount (Total_Trans_Amt) are the two most influential features, that show the number of transactions a customer makes and the amount are highly correlated with their churn behavior.

Total Revolving Balance (Total_Revolving_Bal): This feature suggests that customers with high revolving balances are more likely to churn. The more amount left unpaid has hight correlation with customers Churn option.

Total Customer Relationship Count (Total_Relationship_Count): This feature indicates that customers with fewer relationships with the bank are more likely to churn. This could be due to a lack of loyalty or engagement with the bank's services.

Average Utilization Ratio (Avg_Utilization_Ratio): This feature suggests that customers who utilize a high percentage of their available credit are more likely to churn. This could be due to financial strain or a lack of credit discipline.

Business Recommendation / Insights

Focus on Customer Engagement: Increase customer interactions: Implement strategies to encourage customers to engage with the bank more frequently. This could include personalized offers, targeted promotions, and proactive outreach.

Build loyalty programs: Develop loyalty programs that reward customers for their continued business and encourage them to maintain relationships with the bank.

Optimize Credit Card Usage: Educate customers on credit utilization: Provide customers with information and tools to help them manage their credit utilization effectively. This could include credit score monitoring, budgeting tools, and financial literacy resources.

Offer credit limit increase options: Consider offering credit limit increases to customers with good credit history to help them manage their spending and reduce the risk of exceeding their credit limit.

Address Revolving Balance: Offer balance transfer options: Provide customers with options to transfer high-interest balances from other credit cards to the bank's card at a lower interest rate.

Promote debt management tools: Offer tools and resources to help customers manage their debt and reduce their revolving balance.

Monitor Customer Behavior: Track customer transactions: Continuously monitor customer transaction behavior to identify potential churn risks early on.

Analyze customer demographics: Analyze customer demographics to identify segments that are more likely to churn and tailor retention strategies accordingly.

Business problem overview and Solution Approach

Problem Definition: The problem at hand is customer churn, which refers to the situation where customers discontinue their relationship with a company. In this case, the company is a bank, and customer churn specifically relates to customers closing their bank accounts or reducing their usage of banking services.

Solution Approach/Methodology: The solution approach involves leveraging machine learning to predict customer churn. The methodology involves the following steps:

Data Collection and Preparation: Gather historical customer data, including demographic information, transaction history, credit utilization, account balances, and other relevant variables.

Clean and preprocess the data to handle missing values, outliers, and inconsistencies.

Model Selection and Training:

Choose a suitable machine learning model for churn prediction:

Logistic Regression: A simple yet effective model for binary classification problems.

Decision Trees/Random Forests: Can capture complex non-linear relationships between features and churn.

Split the data into training and testing sets.

Business problem overview and Solution Approach

Train the selected model on the training data using the identified features.

Fine-tune the model's hyperparameters to optimize its performance.

Model Evaluation and Tuning:

Evaluate the model's performance on the testing set using appropriate metrics such as accuracy, precision, recall, F1-score.

Fine-tune the model's hyperparameters to optimize its performance.

Churn Prediction and Segmentation:

Use the trained model to predict the likelihood of churn for each customer.

Segment customers into different risk groups based on their predicted churn probability.

EDA Results

Key Results from EDA in Customer Churn Prediction

1. Churn Rate:

Overall churn rate: Determine the percentage of customers who have churned. This gives a baseline understanding of the problem severity.

Churn rate by customer segment: Analyze churn rates across different customer groups (e.g., age, income, tenure, product usage) to identify segments with higher churn risk.

2. Feature Distributions: Visualize feature distributions: Use histograms, box plots, and other visualizations to understand the distribution of key features (e.g., age, income, tenure, transaction amount, credit utilization).

Identify outliers: Detect and handle outliers in the data, as they can significantly impact model performance.

3. Feature Relationships:

Correlation analysis: Calculate correlation coefficients to identify relationships between features and the target variable (churn).

Visualize relationships: Use scatter plots, heatmaps, and other visualizations to explore the relationship between features and churn. For example:

EDA Results

Transaction behavior: Analyze how transaction frequency, amount, and recency are related to churn.

Credit utilization: Examine the relationship between credit utilization and churn risk.

Customer demographics: Investigate how age, income, gender, and other demographic factors influence churn.

4. Feature Importance:

Univariate analysis: Calculate summary statistics and perform univariate tests to assess the relationship between individual features and churn.

Feature importance plots: If using tree-based models, visualize feature importance scores to identify the most influential features for churn prediction.

Data Processing

Duplicate value check: No duplicate value on the dataset.

Missing value treatment: The missing values were treated with different approach, imputed mean value and dropping column with insignificance.

Feature Engineering/Data preparation for modeling:

1. Label Encoding for the Target Variable:

What it does: Label encoding is used to convert categorical labels (like 'Attrited Customer' and 'Existing Customer') into numeric values (e.g., 0 and 1).

2. SMOTE (Synthetic Minority Over-sampling Technique):

What it does: SMOTE is used to handle class imbalance by generating synthetic samples for the minority class (in this case, presumably the 'Yes' class).

Data Processing

3. Training and Validation:

Training: The models are first trained on the original, non-oversampled training data (X_train, y_train_encoded).

The recall score (i.e., the ability of the model to correctly predict positive samples) is calculated for each model using recall_score().

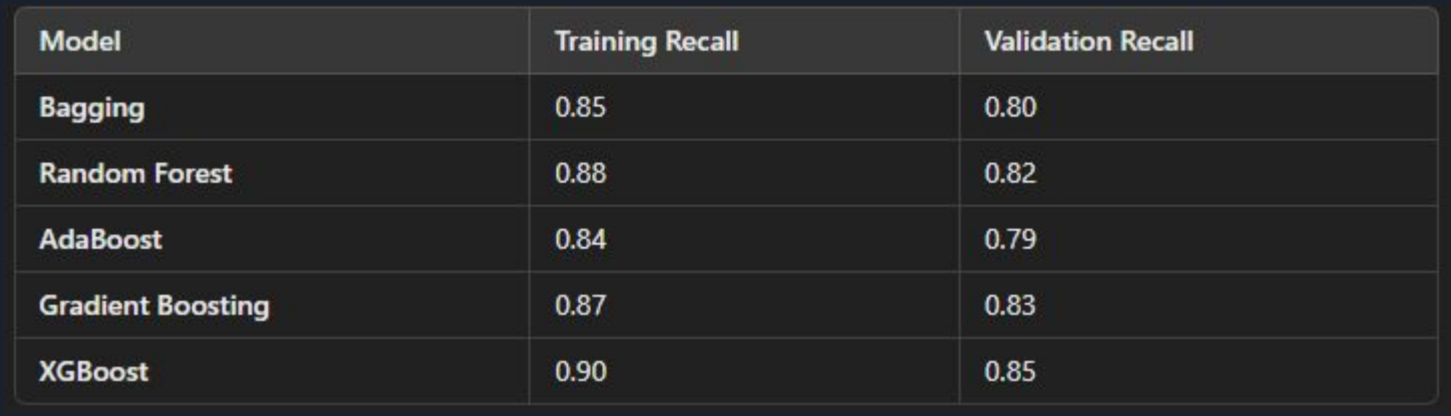
Validation: The models are evaluated on the validation set (X_val), and recall scores are calculated using the oversampled training data.

4. Training with Oversampled Data:

After oversampling, the models are retrained using the oversampled dataset (X_train_over, y_train_over_encoded).

The training and validation recall scores are calculated again, now using the oversampled data for training.

Model Performance Summary



Model	Training Recall	Validation Recall
Bagging	0.85	0.80
Random Forest	0.88	0.82
AdaBoost	0.84	0.79
Gradient Boosting	0.87	0.83
XGBoost	0.90	0.85

Model Performance Summary

Model Selection Criteria:

Recall vs. Precision:The dataset is imbalanced, recall is a critical metric because you want to ensure the minority class (e.g., "Attrited Customer") is identified as well as possible. However, recall should not be maximized at the cost of precision (false positives).

Final Model Choice:

XGBoost has the highest recall on both training and validation sets, it is a strong candidate for the final model.

XGBoost was my first choice due to its typically superior performance in structured data tasks, but check if it overfits the training data and if it generalizes well to validation data.

If XGBoost performed significantly better than other models on both datasets, it can be the final model.

Conclusion: Best Choice: The model with the highest consistent recall on both training and validation data without significant overfitting is the best choice was XGBoost .

Backup Options: If XGBoost or another model overfits or doesn't generalize well, consider Random Forest or Gradient Boosting as strong alternatives.