Predict Churning Customers of Bank Credit Card

Helen Zheng   Shuran Wen

BLUF

This business report presents the findings and analysis of our project, which aims to investigate the factors contributing to customer churn in the credit card system utilizing machine learning models. We used recall metric as our primary metrics. Among the 10 examined classifiers, we discovered that the Light GBM model exhibited exceptional performance, achieving the highest recall score of 0.893 and the highest F1 score of 0.888. Building upon this success, we conducted a grid search on Light GBM model, the new parameter resulted in a more robust performance. Continually, both feature importance method and Sharpley value revealed 6 key features that have a significant impact on customer churn within the credit card system: 'Total_Trans_Ct,' 'Total_Trans_Amt,' 'Total_Revolving_Bal,' 'Total_Ct_Chng_Q4_Q1,' 'Avg_Utilization_Ratio,' and 'Total_Relationship_Count.' After cutting off other insignificant features, the recall score became 0.889 and the learning curve revealed a noticeable reduce in difference between training set (Recall: 0.96) and test set (Recall: 1).

Analysis

-Business Problem

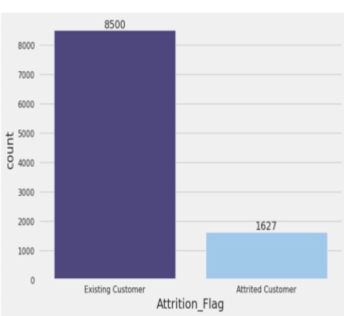
Customer churn poses a significant risk for financial institutions, resulting in revenue loss. This project is of utmost importance for the banking institution, as it has the potential to enhance customer retention and drive profitability. By creating a reliable churn classification model, we can help institutions proactively address customer concerns and foster lasting relationships. The model enables targeted retention strategies, boosting customer satisfaction and loyalty.

-Metric Selection

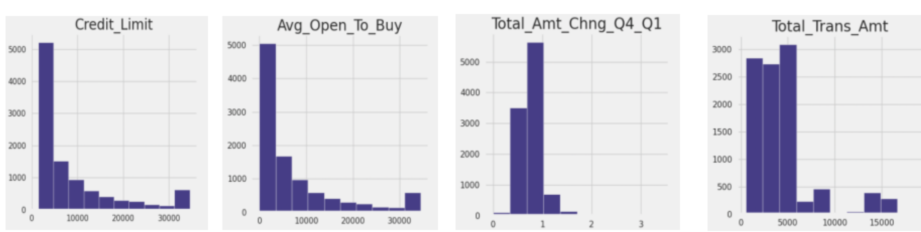
In our classification model, we deemed “Recall” as the principal evaluation metric. Recall measures the proportion of true positive predictions among all actual positive instances, as well as the proportion of true negative predictions among all actual negative instances. When using recall, we aim to accurately assess the model's performance in correctly identifying the false negatives, that is to what extent the " attrite customers" is identified as "existing customers". Additionally, when presenting the leaderboard and learning curve, F1 is introduced as supplementary references as it maintains a balance between the precision and recall.

-EDA

Graph1. Imbalance of the Data



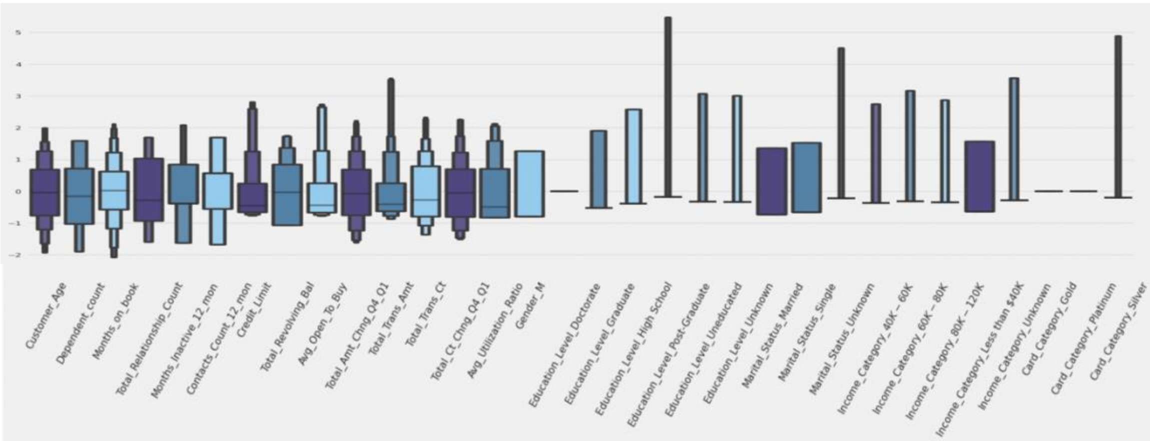
Graph2. Distribution of features with skewness



The dataset from Kaggle: ‘Credit Card Customers’ has 10127 instances and 23 features. The last two columns and the client number column are dropped, as they do not provide significant importance for modeling purpose. In EDA process reveals that the dataset has no duplicates and missing values, but it does have skewness that are greater than 1 for several features. Additionally, the targeting variable “Attrition flag” has serious imbalance problem: with 8500 in “Existing Customer” and 1627 in “Attrited Customer”. This severe class imbalance poses a challenge for building an accurate predictive model, as it may result in biased predictions favoring the majority class. Addressing this class imbalance will be a critical step in developing a reliable churn prediction model.

-Preprocessing

Graph3. Box Plot of features after processing



In this process, the target "Attrition Flag" is encoded as 1 for "Attrited Customer", and 0 for "Existing Customer". We then conducted skewness correction to problematic features and limited their skewness within a safe range of [-1, 1]. To mitigate the influence of the extreme values, we use Turkey's method to identify outliers and winsorized X to eliminate outliers. After these steps, we split the dataset into training and testing data, and utilized SMOTE in training dataset and applied Scaling to both training and testing dataset of feature variables.

-Selecting Classifiers

Starting with forming a classifier horse race of 10 classifiers to through 5-fold cross validation: "Logistic Regression", "Perceptron", "Logistic Regression", "SVM (RBF kernel)", "Decision Tree", "Naive Bayes", "k Nearest Neighbors", "MLP", "Random Forest", "XG Boost", "Light GBM", the Light GBM classifier emerged as the ideal choice based on its superior performance. The leaderboard presented below highlights the performance ranked by Recall:

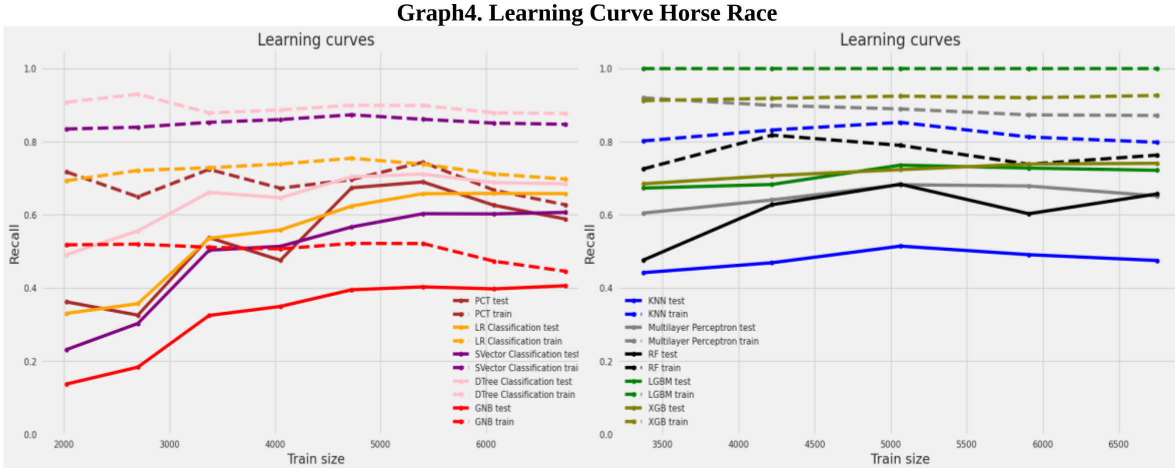
As we can see from the leaderboard: for Recall metrics, Light GBM outperforms all other models in terms of the Recall metric, achieving a Recall score of 0.883. Following closely behind is XG Boost (0.880). Notably, Light GBM also demonstrates superior performance in F1 scores (0.883). This is consistent with the result of low code obtained with PyCaret code (Recall: 0.88, F1:0.87).

Table1. Leader board

	Recall	F1		Recall	F1
Light GBM	0.883	0.883	Support Vector	0.690	0.728
XG Boost	0.880	0.855	Logistic Regression	0.650	0.643
Decision Tree	0.864	0.772	Perceptron	0.582	0.558
Multilayer Perceptron	0.782	0.781	K Neighbors	0.502	0.520
Random Forest	0.715	0.620	Gaussian NB	0.415	0.397

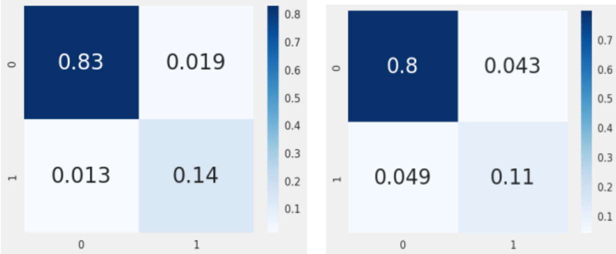
For better accuracy, we also graphed the classifiers' learning curves to see which one performs well. It indicates that the Light GBM's learning curve also perform well and has no sign of significant overfitting. The initial difference between training set (Recall: 0.07) and test set (Recall:1.0) is roughly 0.3.

Graph4. Learning Curve Horse Race



Considering its consistent top performance across multiple evaluation metrics, we can conclude that Light GBM is the most suitable model for studying the credit card classification problem.

Graph5. Confusion Matrix (left: before PCA, right: after PCA)



Following the selection of the Light GBM classifier, we performed PCA (Principal Component Analysis) to reduce the dimensionality of the dataset and improve computational efficiency. The percentage of instances with propensities is computed in [0.7,0.3]. Graph5 shows an improved performance in the confusion matrix. The increasing proportion of false negatives and false positives indicating a reduced number of these 2 terms (of only 37 and 38), indicating a high a recall of 0.867 and a F1 score of 0.831.

To tune the hyper parameters of LightGBM, we use cross validation of 5 folds. The grid search process identified the following optimal hyperparameters: {'bagging_fraction': 0.7, 'feature_fraction': 0.7, 'learning_rate': 0.15, 'num_leaves': 70, 'reg_lambda': 0.1}. The best score achieved through grid search of 0.679 showcasing the robustness and effectiveness of the tuned LightGBM model.

-XAI

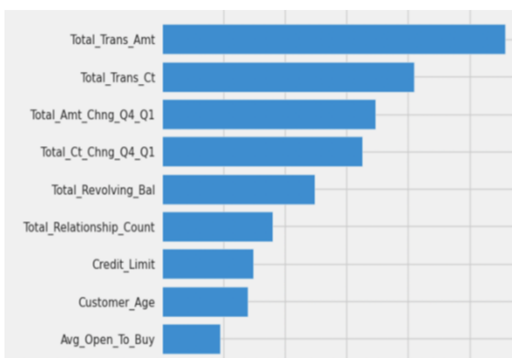
As shown in Table2, in the beginning of feature importance, we utilized RFE (Recursive Feature Elimination) to select 9 features that best fit in the model prediction in accuracy of 1.

Following this step, we bring out Global methods of feature importance. LGBM's feature importance (Graph 6) and Sharpley value (Graph 7) tell similar story, in which we can see some features with high importance in common. Among all the 23 features, Total Transaction Amount and Total Transaction Count are 2 most significant driver of credit card churn. As can be shown in Graph 8, Change in Transaction Count (Q4 over Q1), Change in Transaction Amount (Q4 over Q1), Total Revolving Balance on the Credit Card, Total Number of Products Customer Held are also important, with churn rate decreasing, respectively increasing with these features.

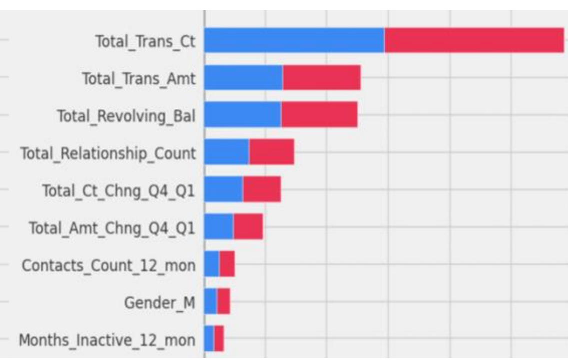
Table2.Important Features based on RFE

Customer Age
Total No. of Products Customer Held
Credit Limit on Card
Total Revolving Balance on the Card
Open to Buy Credit Line (Avg of Last 12 months)
Total Transaction Amount (Last 12 months)
Total Transaction Count (Last 12 months)
Change in Transaction Count (Q4 over Q1)
Change in Transaction Amount (Q4 over Q1)
Avg Card Utilization Ratio

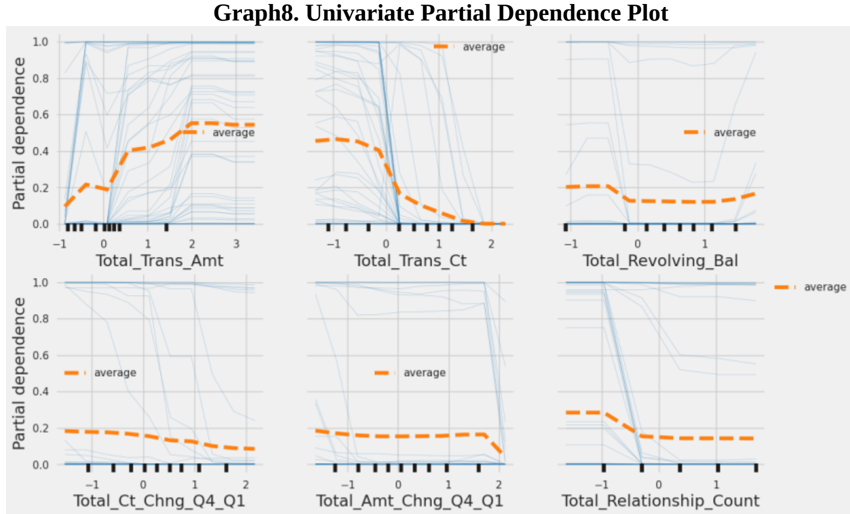
Graph6. LGBM Feature Importance



Graph7. Sharpley Values



Graph8. Univariate Partial Dependence Plot



-Rerun the model

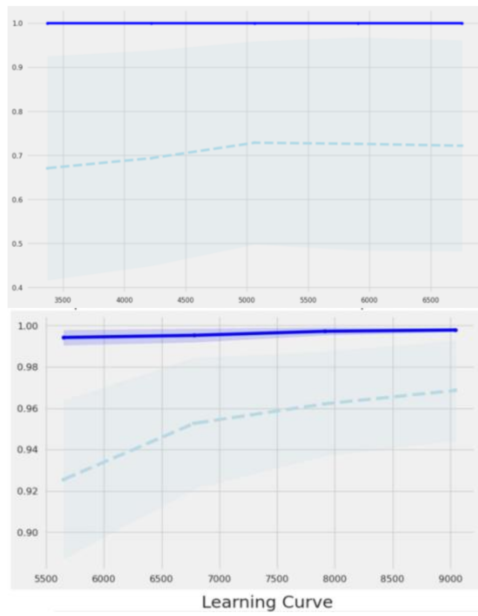
In the light of feature importance, we tested two sample size: 4 features and 6 features by rerunning the LGBM model using 5-fold cross-validation. We dropped the 4 features model as its recall and F1 is relevantly worse than the original model's.

The results revealed that a model incorporating the following 6 features proved to be an excellent fit: Total Transaction Amount, Total Transaction Count, Change in Transaction Count (Q4 over Q1), Change in Transaction Amount (Q4 over Q1), Total Revolving Balance on the Credit Card, Total Number of Products Customer Held. Notably, this refined model exhibited superior performance compared to the original model, boasting a higher recall value of 0.889 and a higher F1 score of 0.887, which can be also shown in the increased proportion of false negative in confusion matrix. These metrics underscore the model's enhanced ability to accurately identify churned customers. Furthermore, a remarkable improvement was observed in the learning curve, with a noticeable reduce in the discrepancy between the test and train scores, now ranging from 0.96 to 1.0.

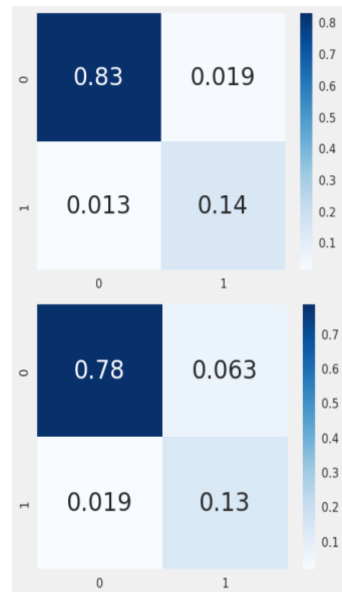
Table3. Metrics Comparison

	Recall	F1
All features	0.883	0.883
6 features	0.889	0.887
4 features	0.873	0.830

Graph9. Changes in Learning Curve (up: all features, down: 6 features)



Graph10. Changes in Confusion Matrix (up: all features, down: 6 features)



Conclusion

-Summary

Our analysis focused on identifying the factors that contribute to customer churn within financial institutions. The comprehensive process involved exploratory data analysis (EDA), data preprocessing, optimal classifier evaluation, feature importance selection, hyperparameter tuning, implementing explainable AI tools, and solving the overfitting and underfitting problem. we identified Light GBM as the top-performing classifier as it has the highest recall and F1 score, and a relatively small subsize in learning curve in the 10 classifiers' horse race.

Additionally, our analysis revealed six influential features that significantly drives customer churn issue within financial institutions. These features, namely 'Total_Trans_Ct,' 'Total_Trans_Amt,' 'Total_Revolving_Bal,' 'Total_Ct_Chng_Q4_Q1,' 'Avg_Utilization_Ratio,' and 'Total_Relationship_Count,' play a crucial role in determining the likelihood of customers churning. By monitoring and strategically addressing these variables, financial institutions can effectively mitigate churn and enhance customer retention rates.

-SWOT

Our analysis showcases several strengths and weaknesses that merit consideration. Firstly, the classifier we selected demonstrates top-performing capabilities across all evaluation metrics and exhibits a strong fit within the learning curve. This achievement underscores the reliability and effectiveness of our chosen model. However, one notable weakness lies in the serious imbalance, which can undermine the effectiveness for model to identify the minority class Although we have taken steps to address this issue by utilizing SMOTE (Synthetic Minority Over-sampling Technique), there are potential weaknesses that need to be acknowledged, including the risk of overfitting and limitations in generalizability. In a real-world scenario, identifying potential customer churn is crucial for the company, while in most cases the customer are not leaving, most potential churn case should be correctly identified.