Bank Customer Churn Prediction

Norin Hossam
Computer Engineering
AAST
Alexandria, Egypt

Zyad Shrin
Computer Engineering
AAST
Alexandria, Egypt

I. Abstract

Nowadays, there are many choices available in the banking world prompting a rise in the competition globally. Therefore, nearly all banks face the hurdle of client churn. This paper centers on the classification problem known as customer churn prediction, which examines client's behavior to study the probability of customers leaving the bank in the near future. With the aim of building a reliable predictive system, feature engineering techniques were conducted. Several machine learning algorithms were explored such as LightGBM, XGBoost, and Random Forest. The results of the models were analyzed and compared, concluding that LightGBM was the leading model in terms of AUC.

Keywords— Classification, LightGBM, XGBoost, Random Forest, Oversampling.

II. Introduction

In today's extremely competitive financial industry, maintaining existing customers is just as important, if not more than obtaining new ones. Customer churn poses a serious risk to banks' profitability and long-term growth. Understanding and anticipating customer churn has become a strategic goal for financial institutions due to the significantly higher cost of gaining a new customer vs retaining an existing one.

Customer churn can be caused by a variety of circumstances, such as poor service quality, better offers from competitors, financial changes, or unfavorable client experiences. Accurately identifying clients who are likely to churn enables banks to execute tailored retention initiatives, reducing churn and increasing customer loyalty.

III. Related Work

The initial study [1] focused on predicting customer churn in a bank, the dataset used was obtained from Kaggle which contained 10000 rows of data, 7693 were positive class samples and 2037 were negative class samples. Data preprocessing techniques were carried out such as oversampling by resampling the negative class which was the minority class due to dataset imbalance and feature selection using MRMR and Relief methods to reduce dimensionality. The data split was 70% for training, and 30% for testing. The machine learning techniques experimented with were KNN, SVM, Decision Tree, and Random Forest. After hyper-tuning models' parameters, a 10-fold cross validation was performed to prevent overfitting. Overall, the top model was Random Forest achieving an accuracy of 95.74%.

The researchers concentrated on predicting customer churn in Chinese commercial banks [2] to assist in maximizing revenues and bypass the loss of clients. The dataset used contained 50000 rows of data, but due to preprocessing, 46404 data entries were present. Two types of SVM models were investigated: SVM with radial basis and linear SVM. One preprocessing method, which greatly helped improve the model's performance and increased its forecasting capability, was under-sampling. Ultimately, the model achieved an accuracy of 80.84%.

Abdelrahim, Assef, and Kadan's research [3] aimed at predicting client churns for telecommunication companies. The dataset was obtained from the Syriatel company. Four machine learning algorithms were explored: Gradient Boosted Machine Tree (GBM), Decision Tree, XGBoost, and Random Forest. K-fold cross validation helped hyper parameter tune, and data was balanced using an under-sampling method. The evaluation metric utilized was AUC. The GBM model outperformed Random Forest and Decision Tree models. Overall, the results indicate that XGBoost is the optimal model with 180 trees securing an AUC of 93.301%.

IV. Proposed Model

Three powerful machine learning algorithms were utilized to forecast customer churn: Random Forest Classifier, XGBoost, and LightGBM. To determine the ideal model, each algorithm was examined in three distinct manners:

1) **Baseline model with default parameters:** Initially, each model was trained with the dataset's default setting. This served as a benchmark to assess future performance improvements.
2) **Hyperparameter tuning:** The second strategy involved optimizing each model's performance by modifying its hyperparameters. Grid search was used to determine the optimal combination of parameters for maximizing accuracy and other key metrics.
3) **Feature Engineering and Data up-sampling:** The third solution combined feature engineering and data up-sampling techniques. We balanced the dataset using the Synthetic Minority Over-sampling Technique (SMOTE), ensuring that the minority class was appropriately represented. Moreover, additional features were developed that helped capture the underlying patterns associated with customer churn. This method likewise employed the hyper-tuned parameters identified in the preceding step.

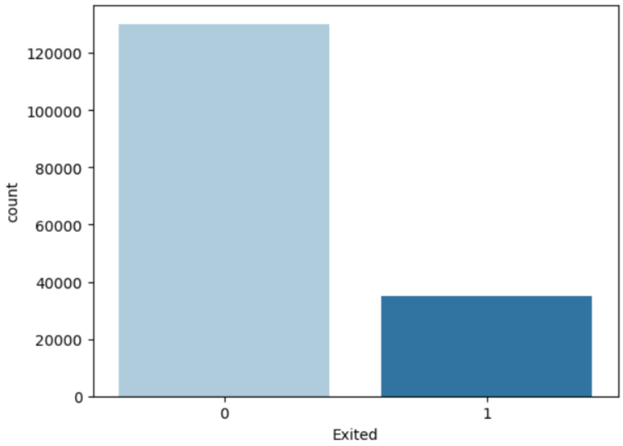
Following these experiments, the third technique provided the best results. Among the three models, the **LightGBM model** stood out. As a result, it was chosen for our customer churn prediction task, which was trained with hyper-tuned parameters, an up-scaled dataset using SMOTE, and feature engineering. This approach produced the most reliable forecasts, allowing the bank to better identify customers at risk of churn and conduct successful retention tactics.

V. Experimental Work

A. Dataset

The bank customer churn dataset is popular for predicting customer turnover in the banking industry (https://www.kaggle.com/datasets/shubhammeshram579/bank-customer-churn-prediction). It contains information about bank customers who have either left the bank or are still customers. The dataset originally had 12 features and a label. To improve our analysis, we created an additional 10 features, for a total of 22 characteristics. Some of the new features are 'IsSenior', 'ZeroBalance', and 'IsActive_by_CreditCard'. The dataset contains 165,034 rows, each representing a client. Of these, 130,113 consumers did not churn, whereas 34,921 did, resulting in a large class imbalance as shown in **Figure 1**. We addressed this imbalance using SMOTE and improving the performance of our prediction models.

Figure 1



B. Evaluation Metrics

The models were assessed using a variety of measures to ensure an accurate evaluation of their effectiveness. We used the accuracy score with the confusion matrix, as well as the AUC score, precision, recall, and F1-score. Accuracy is measured as the ratio of accurately predicted occurrences to total instances. The confusion matrix displays the number of True Positive, True Negative, False Positive, and False Negative predictions. The AUC score summarizes the model's performance with a single value by displaying the true positive rate versus the false positive rate, is important for comparing models. The precision, recall, and F1-score for each class: precision is the ratio of true positives to total predicted positives; recall is the ratio of true positives to actual positives; and F1-score is the harmonic mean of precision and recall. These measures, taken together, provide an overall view of model performance, which is critical for understanding and increasing prediction accuracy for bank customer turnover.

The confusion matrices for the three models tested using the third strategy, which incorporates hyperparameter tuning, feature engineering, and SMOTE-based dataset upsampling, are shown in the figures below. **Figure 2** displays the Random Forest model, whereas **Figure 3** represents the XGBoost and **Figure 4** demonstrates the LightGBM model which yielded the best results. These confusion matrices give a visual depiction of the algorithms' performance in predicting client departures.

Figure 2

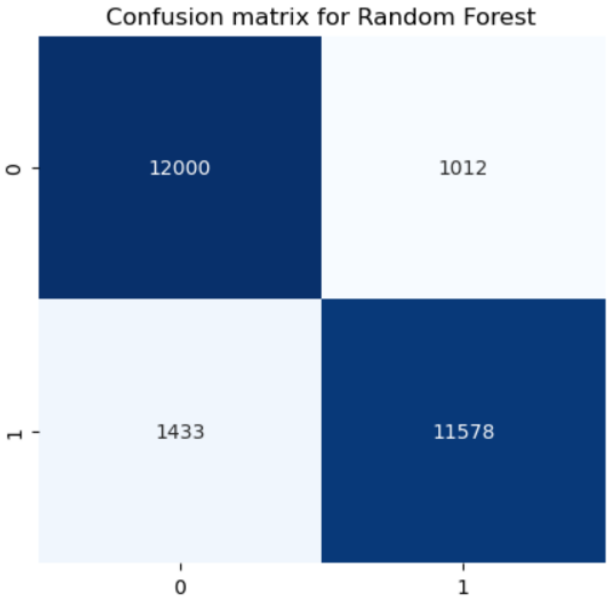


Figure 3

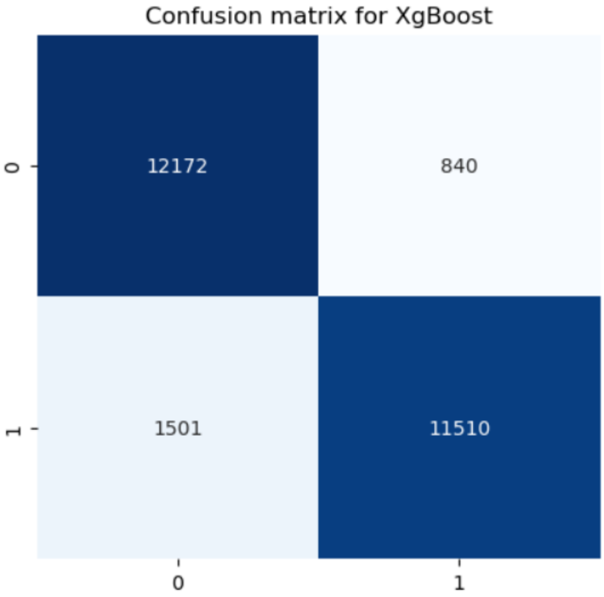
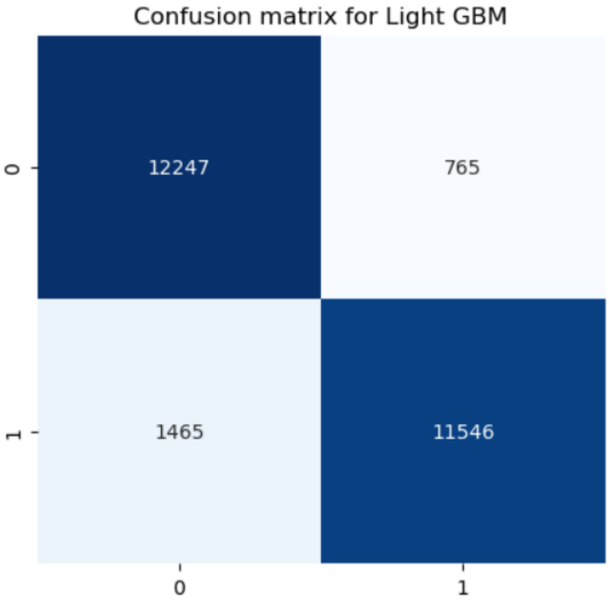


Figure 4



C. Results

The three tables below show the outcomes of each model in terms of precision, recall, F1-score, accuracy, and AUC. Each table provides a thorough comparison of the three approaches used on each model. **Table 1** shows the results for the Random Forest model. **Table 2** displays the assessment metrics for the XGBoost model. **Table 3** shows the performance of the LightGBM model which had the best results. These tables provide a detailed summary of each model's effectiveness under various settings and preprocessing methods.

Table 1 Random Forest

	Precision	Recall	F1	Accuracy	AUC
1st	0.85	0.86	0.85	85.9%	74.3%
2nd	0.85	0.86	0.85	86.3%	74.62%
3rd	0.91	0.91	0.91	90.6%	90.6%

Table 2 XGBoost

	Precision	Recall	F1	Accuracy	AUC
1st	0.86	0.86	0.86	86.43%	75.12%
2nd	0.86	0.87	0.86	86.54%	75.31%
3rd	0.91	0.91	0.91	91%	91%

Table 3 LightGBM

	Precision	Recall	F1	Accuracy	AUC
1st	0.86	0.87	0.86	86.62%	75.39%
2nd	0.86	0.87	0.86	86.68%	75.61%
3rd	0.92	0.91	0.91	91.43%	91.43%

VI. Conclusion and Future Work

The goal of the “Bank Churn Prediction” is to predict the possibility of customers leaving the bank soon, this helps the banks to take immediate action in order to preserve their clients. This paper utilizes three machine learning techniques: LightGBM, XGBoost, and Random Forest. The main limitation was the highly imbalanced dataset, which was addressed by oversampling the minority class. The findings indicate that LightGBM was the standout model achieving an AUC of 88.88%. For future work, the target is to further enhance the model by implementing K-fold cross validation and to experiment with ensemble methods to boost the model’s accuracy.

VII. References

[1] M. Rahman, and V. Kumar, “Machine Learning Based Customer Churn Prediction In Banking”, 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020.

[2] B. He, Y. Shi, Q. Wan, and X. Zhao, “Prediction of customer attrition of commercial banks based on SVM model,” Procedia Computer Science, vol. 31, pp. 423–430, 2014.

[3] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” Journal of Big Data, vol. 6, no. 1, p. 28, 2019.