Machine Learning Based Customer Churn Prediction in Banking Sector

Shangrong Han

Department of Statistical Science, University College London, London, UK

zcaksh4@ucl.ac.uk

Abstract. Customers in 21st century have access to a wide range of ways to deposit money, both online and offline, which leads to constant customer churn for the whole banking industry. In order to retain existing customers, the bank sector has been prioritizing building models which aim to predict clients who may exit in the future. In this paper, based on machine learning techniques, different models such as XGboost, Catboost and LightGBM are fitted to the churn modelling dataset from Kaggle, contributing to the prediction of potential bank customer churn. In addition, some methods of feature selection and hyperparameter tunning are used to enhance the performance of final prediction results. The results generated by different models are compared in terms of accuracy, precision, recall, etc. Age and the number of products purchased from the bank are suggested to be 2 factors that greatly influence the prediction results. LightGBM model shows the best general performance and therefore is recommended for future prediction.

Keywords: Bank Churn, Customer Churn Prediction, Machine Learning.

1. Introduction

There has been increasingly growing turmoil and competition between enterprises in virtually every industry, due to the availability of a huge number of service providers and difficulties in satisfying ever-changing consumer demands. Thus, it is a significant issue for companies that companies should conceive innovative and entrenched strategy to stabilize customer basis. Depending on a large customer basis, companies can promote further expansion and gain higher profits, which implies two prerequisites for success: attracting new customers and maintaining existing ones [1]. Liu and Shih state that, to achieve these two goals, personalized marketing campaigns could be utilized to attain high-level consumer satisfaction [2]. However, this method could be fairly time and budget consuming, predominantly for the sake of expanding new customer basis. Rather than raising new customers' attraction by promoting advertisement, other studies hold the belief that improving existing customers' loyalty is much more vital, as it is 25 times more economically efficient [3, 4]. Moreover, the omission of churn customers may contribute to disorientation of operations and loss of funding for companies [5]. Thus, the major focus of customer-oriented retention for a company should be identifying customers who have shown churn tendency.

Customer churn, or customer attrition, is related to those who tend to departure the current company or industry where they have received benefits from their purchasing behavior [6, 7]. The possibility of customer churn is linked to a variety of aspects, which have complex dependence relationships. Consequently, it is tough to identify potential customer churn directly by filtering specified customer features, and a fully trained model is required to meet the demand of targeting customers who tend to leave the company.

The field of building models based on machine learning techniques to predict customer churn has been fully and extensively exploited by previous researchers due to their excellent prediction ability in many other tasks [8, 9], who make use of many universal and well-developed algorithms such as K-Nearest Neighbors (KNN), Decision Tree (DT), Logistic Regression (LR), Support Vector Classifier (SVC), etc. [10]. These models are configured with specified parameters based on empiricism, which may cause discrepancy with the optimal figures. Therefore, the authors in [11] scrutinize the importance of altering different hyperparameter values when using a Deep Neural Network (DNN) algorithm to better predict the customer churn in banking sector. Also, the

experimentation reveals a huge gap in performance between DNN models with distinct monotonic activation functions, emphasizing the significance of hyperparameter tunning.

Another research chiefly concentrates on the impact of the prediction accuracy of different models when highly imbalanced dataset (with approximately 83.9% of non-churners and 16.1% of churners) is applied to the fitting progress [12]. The research compares the performance of 3 models, whose training sets undertake Synthetic Minority Oversampling Techniques (SMOTE), to the performance of another 6 models proposed by other researchers, and successfully finds that after SMOTE preprocessing stage, the RF model predicts customer churn with the highest accuracy rate 88.7% among all 9 models.

In this paper, different models of machine learning, including Extreme Gradient Boost (XGBoost), CatBoost and Light Gradient-Boosting Machine (LightGBM), are applied to the Bank Customer Churn Prediction dataset from Kaggle for future churn prediction for the banking sector. In terms of performance as accuracy, precision, recall, etc., these models are compared to find the model of best fit, which is LightGBM. In addition, the feature importance rankings are also investigated to find the dominant factors that lead to customer churn.

2. Methods

2.1. Dataset Preparation

The dataset used in this study is taken from Kaggle [13]. The dataset contains information of 10,000 bank clients in total and is composed of 14 features (attributes), including "Credit Score", "Tenure", "Balance", etc. All the binary feature values in the dataset are transformed to numbers, where "1" means "True/Yes" and "0" means "False/No". The target feature for prediction results is "Exited". Among all the bank clients, 2037 people have already left the bank, accounting for 20.37% of the sample space, while 7963 people, who account for 79.63% of the sample space, are still retained in the bank, showing a highly skewed characteristic of the sample data. In addition, the type of this work is binary classification.

The notion of data preprocessing is important in data mining. The research conducted in [14] indicates that preprocessing perturbations could markedly impact the performance of models of machine learning. As it is difficult to apply the "Gender" variables to the models directly, some methods should be used to transform these categorical variables to numerical ones. One feasible way is to utilize Label Encoder, which turns "Male" into "1" and "Female" into "0".

According to [15], after selecting features, the performance of classifiers employed can show significant enhancement, which reveals the effectiveness of feature selection in data mining. Features which have barely relevancy with target feature are deemed redundant and may have a negative influence on the performance of models. Thus, the features "RowNumber", "Surname" and "Geography" are abandoned. In various previous studies, the dataset is utilized to develop and assess the performance of different models, but one issue is that the potential dependence between variables of the dataset has not been paid enough attention to. This work also finds the possible relationships between existing features of the dataset and creates some new features as are shown in Table 1, which can be applied to the development of models to increase the prediction accuracy. The ratio of the number of data contained in the training set to the test set is 3:1. To acquire the training, test and validation sets, the library scikit-learn is imported and utilized.

Table 1. Additional Feature Description

Feature Name	Feature Description
TenurePerProduct	The amount of tenure divided by the number of products purchased by the customer
IsActiveCrCard	Whether the customer has a credit card and is an active member. Calculated by multiplying "HasCrCard" and "IsActiveMember"
AgeGroup	Categorizing customers into 3 groups (young, middle and old), based on their ages
BalanceGroup	Categorizing customers into 4 groups (low, medium, high and very high), based on their balance

2.2. Modelling

In this section, 3 different classifiers, including XGboost, Catboost and LightGBM, are used. Before fitting training set to the models, hyperparameter tunning technique is applied to the models to obtain the optimal parameters for the arguments of these classifiers, greatly enhancing the performance of model prediction. All the modules used for modelling are from scikit-learn library. The eval metric, which is used to determine when to stop developing models, is Area Under the Receiver Operating Characteristic Curve (AUC).

2.2.1 XGboost

The primary principle of Extreme Gradient Boost (XGboost) is to integrate many less potent classifiers (Decision Trees, Random Forests, Gradient Boost, etc.), followed by a sequential training process, and create a robust classifier [16]. Initially, a weak prediction is produced. A series of residuals computed based on observed data and predicted data are then added to the initial prediction value. As a result, the eventual prediction could have much higher accuracy than that produced by only one classifier from the combinations.

2.2.2 Catboost

Catboost is conducted by merging Gradient-Boosted Decision Trees (GBDT) with categorical features [17], which greatly enhances the capability of predicting customer churn based on a dataset with numerous categorical features. Similar to Gradient Boost algorithm, Catboost works by iteratively construct decision trees to gradually improves the accuracy of prediction.

2.2.3 LightGBM

Compared to other algorithms growing trees horizontally, Light Gradient-Boost Machine (LightGBM) works in a vertical way, which enables the algorithm users to attain highest splitting-gain while minimizing training losses [18]. The leaf-wise running principle of LightGBM reduces memory when handling a large size of data, hence high speed to deliver prediction results.

2.2.4 Randomized Search CV

Hyperparameters are specific values which govern how an algorithm is processed. To obtain an optimal prediction result, hyperparameter tunning/optimization is required. One way to conduct hyperparameter tunning, Randomized Search CV, is used in this section, which chooses hyperparameters randomly from the given parameter grid and reports the combination of parameters with which the model can outperform others.

3. Results and Discussion

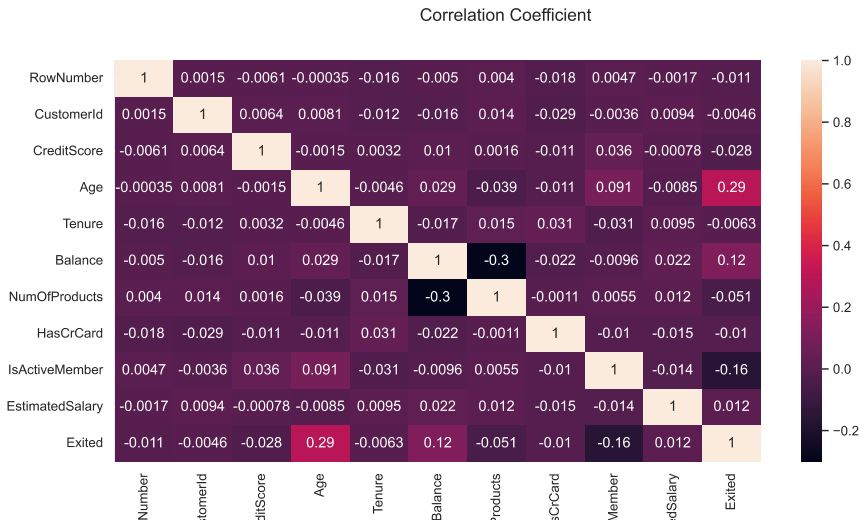

Figure 1. Correlation between variables (Photo/Picture credit: Original)

The bivariate correlation relationships between all numerical data and categorical data that are presented in the form of numbers are calculated and shown in Fig. 1. Compared to other correlation coefficients, the one between "Age" and the target feature, "Exited", is much greater. Therefore, age might play a crucial role in predicting customer churn.

For XGboost, Catboost and LightGBM classifiers, the technique of Randomized Search CV is implemented to find the optimal values for classifier parameters that enable models to increase the probability of making an accurate prediction. The evaluation metric values (AUC) after training set is fitted to XGboost, Catboost and LightGBM classifiers are approximately 0.8648, 0.8659, and 0.8656 respectively. This result suggests that Catboost and LightGBM may be more appropriate to be utilized when predicting customer churn.

After modelling, the testing set is applied to 3 models, generating 3 different prediction results, which can be visualized by confusion matrix shown in Fig. 2. To find the optimal model for prediction, the values of accuracy, precision, recall, false omission rate and F1 score are calculated by the formulae given in (1) – (5) and shown in Table 2. As the major goal of the model is to predict whether a customer will leave the bank, properly predicting the TN cases (Customers who will exit the bank) should be regarded as the priority of the mission. Therefore, XGboost can much properly predict the TN and is the best predicter among 3 models. However, the dataset used for prediction is highly imbalanced, which may be more compatible with XGboost classifier than other 2 classifiers and lead to the situation where XGboost performs better when predicting TN. Moreover, if the general performance of the model is taken into consideration, the LightGBM model, which has the highest values of accuracy, precision and F1 score for churn prediction, may become the optimal choice for churn prediction in the banking sector.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{4}$$

$$False\ Omission\ Rate = \frac{TN}{TN+FN} \tag{5}$$

Table 2. Prediction Performance

Algorithm	Accuracy	Precision	Recall	False Omission Rate	F1 Score
XGboost	82.32%	0.8716	0.9049	0.6253	0.8879
Catboost	85.32%	0.9671	0.8659	0.3870	0.9137
LightGBM	85.64%	0.9716	0.8660	0.3849	0.9158

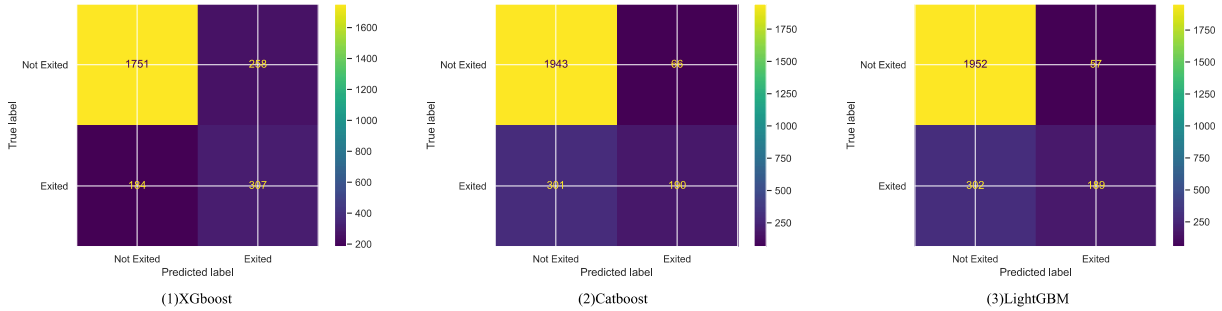

Figure 2. Confusion Matrix for 3 models (Photo/Picture credit: Original)

To merge the prediction outcomes with the studied variables, the feature importance is also illustrated in the form of horizontal bar chart, which is shown in Fig. 3. In XGboost and Catboost algorithms, the number of products is in the dominant position in prediction, which may be due to 2

main reasons. Customers who have purchased multiple products are usually more engaged with the bank and therefore have built a more loyal relationship with the institution. Loyal customers tend to firmly believe the recommendation of upcoming products by the bank and constantly buy these new products, which further deepen the strong bond of bank and them. In addition, once customers have purchased many products from the bank, their costs associated with switching to another bank become inevitably high, which is a great hindrance to departure from the current bank. However, in LightGBM algorithm, age plays the most significant role, which matches the result of correlation plot mentioned above. When customers are in different age stages, the major life goals that they are pursuing are also totally different. Young graduates are prone to seeking higher profits and depositing more money, as they have not yet achieved the economic freedom and need to accumulate enough money in case of any emergencies. As a result, they often change banks to find profit maximization. On the contrary, elderly customers do not have those problems and therefore are in favor of sticking with the bank that they are familiar with. Besides, in LightGBM algorithm, the majority of features are considerably useful, while in other 2 algorithms, many features only have slight influence on the final prediction results. This suggests that the changes in any aspects may lead to the change in prediction results for LightGBM model, which is more sensitive to ever-changing customer conditions, while other 2 models may not immediately react to slight change in customer conditions.

Combining the findings above, LightGBM is regarded to be the optimal algorithm in 3 models, which brings about the best performance when it comes to bank customer churn prediction.

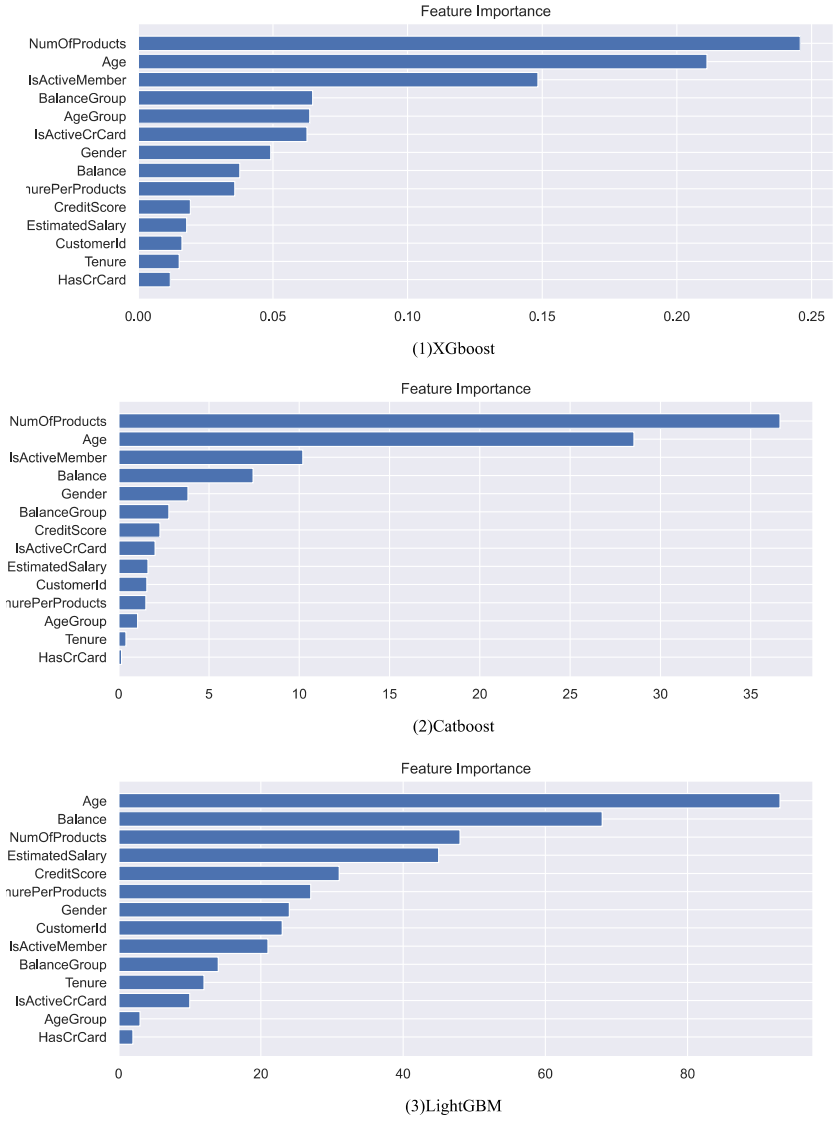

Figure 3. Feature Importance (Photo/Picture credit: Original)

4. Conclusion

The purpose of this study is to develop a model using machine learning techniques to assist banking sectors in predicting customer churn. The study utilizes 3 different algorithms, including XGboost, Catboost and LightGBM, and finds that LightGBM is the optimal choice, with 0.8656 roc auc, 85.64% accuracy, 0.9716 precision and 0.9158 F1 score for prediction. One drawback of this study is that the dataset used contains only 10,000 highly imbalanced data, and some of the variables are limited in a small range. For instance, the range of estimated salaries is from 11.58 to 199992.48, and 75% of the sample customers have ages between 18 to 44. The prediction model may come to better performance when the sample size and variable value ranges are much larger and the method of SMOTE can be applied in the data preprocessing stage. An innovative point of this study is that the potential dependence relationships between variables of dataset are investigated, and additional features based on arithmetic or grouping are added to the dataset, enabling the prediction results to be better. In future work, ensemble methods like Voting Classifier could be used to bring together the advantages of different models. In addition, some research on geography and surname could be conducted to check whether there is any possible means that enables these two features to be used for customer churn prediction.

References

[1] Singh PP, Anik FI, Senapati R, Sinha A, Sakib N, Hossain E. Investigating customer churn in banking: A machine learning approach and visualization app for data science and management. Data Science and Management, 2024, 7 (1): 7-16.

[2] Liu DR, Shih YY. Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management, 2005, 42 (3): 387-400.

[3] Amoako GK, Arthur E, Bandoh C, Katah RK. The impact of effective customer relationship management (CRM) on repurchase: A case study of (GOLDEN TULIP) hotel (ACCRA-GHANA). African Journal of Marketing Management, 2012, 4 (1): 17-29.

[4] Gallo A. The value of keeping the right customers. Harvard Business Review, 2014, 29 (10): 304-309.

[5] De Caigny A, Coussement K, De Bock KW. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 2018, 269 (2): 760-772.

[6] Kaur I, Kaur J. Customer churn analysis and prediction in banking industry using machine learning. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, 2020 November: 434-437.

[7] Qiu Y, Chen P, Lin Z, et al. Clustering Analysis for Silent Telecom Customers Based on K-means++. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). IEEE, 2020, 1: 1023-1027.

[8] Gu Z, Lv J, Wu B, et al. Credit risk assessment of small and micro enterprise based on machine learning. Heliyon, 2024, 10 (5).

[9] Qiu Y, Hui Y, Zhao P, et al. A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy, 2024: 130866.

[10] Bharathi SV, Pramod D, Raman R. An ensemble model for predicting retail banking churn in the youth segment of customers. Data, 2022, 7 (5): 61.

[11] Domingos E, Ojeme B, Daramola O. Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. Computation, 2021, 9 (3): 34.

[12] Muneer A, Ali RF, Alghamdi A, Taib SM, Almaghthawi A, Ghaleb EA. Predicting customers churning in banking industry: A machine learning approach. Indonesian Journal of Electrical Engineering and Computer Science, 2022, 26 (1): 539-549.

[13] Meshram S. Bank customer churn prediction [Data set]. https://www.kaggle.com/datasets/shubhammeshram579/bank-customer-churn-prediction, 2023.

[14] Zelaya CVG. Towards explaining the effects of data preprocessing on machine learning. In 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019 April: 2086-2090.

[15] Mishra K, Rani R. Churn prediction in telecommunication using machine learning. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), IEEE, 2017 August: 2252-2257.

[16] Mahapatra S, Gupta VR, Sahu SS, Panda G. Deep neural network and extreme gradient boosting based Hybrid classifier for improved prediction of Protein-Protein interaction. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 19 (1): 155-165.

[17] Zhou F, Pan H, Gao Z, Huang X, Qian G, Zhu Y, Xiao F. Fire prediction based on catboost algorithm. Mathematical Problems in Engineering, 2021, 2021: 1-9.

[18] Yu J, Lu Q, Qin Z, Yu J, Li Y, Qin Y. A Multi-Stage Ensembled-Learning Approach for Signal Classification Based on Deep CNN and LGBM Models. J. Commun., 2022, 17 (1): 30-38.