

Optimizing Bank Customer Churn Prediction with LightGBM and Data-Driven Strategies

Weixuan Wu

Happy Valley International School, Guiyang, Guizhou Province, 550001, China

1710011048@stu.sqxy.edu.cn

Abstract. This study utilizes the LightGBM model to enhance the prediction of bank customer churn. By utilizing Kaggle's comprehensive datasets and feature engineering, to solve the missing value problem, and eliminate the useless data, so that the data becomes unified, centralized, and easy to identify the advantages. Used data visualization for in-depth analysis, an intervention to further narrow down the content and characteristics of the data by manually identifying correlations and characteristics of various aspects of the data and then conducting more precise checks. The use of LightGBM takes full advantage of handling massive datasets, outperforming traditional algorithms such as Random Forest and XGBoost in terms of efficiency and speed. The combination of new features such as age category and account balance further improves the prediction accuracy of the model, and more deeply complete, In conclusion, this study takes an important step in applying machine learning to improve bank customer churn prediction by proposing a model that balances complexity and practicality.

Keywords: Bank Customer Churn, Prediction, LightGBM.

1. Introduction

In the fierce competition of the bank, the decrease in clients has constituted a big challenge for the bank market [1]. The success of this company is that it can keep a good mood and satisfy customer demand, nowadays, In the bank market, too much competition, basically every bank provides the same service, and the ability to retain customers depends on how well the bank provides refined and generalized services, but most of the time customer will leave is because of some certain service detail is not so good as others, Because of this small detail, customers may not use the products of this bank, because in this industry, there are many competitors, and customers will choose the best [2]. Similarly, if the service of this bank is too bad, a large number of dissatisfied customers will migrate, which will cause the bank to go bankrupt or suffer major losses. To retain these old customers and attract new customers, it is necessary to analyze the details of each customer's age preference, divide customers, and analyze the reasons for their departure. So Machine learning prediction technology is needed by the market [3]. Use deep learning technology to predict bank customer churn, and use it to provide an efficient retain customer relationships strategy so that bank can boost customer satisfaction and their compactness [4]. The key to the business's success is the customer. Customer leave not only will cause an income decrease but also will increase the cost of new customers that the bank acquired, thus influencing bank profit and business development. So, predicting customers accurately and then taking some prevention measures is important for banks [5]. In recent years, deep learning technology has shown great potential in dealing with large-scale data and digging complex patterns, this discovery provides several methods for predicting customer churn. Although there have been studies on the prediction of bank customer churn, most of them focus on traditional statistical methods and relatively low applications of deep learning. In addition, many studies did not apply complex relationships and implicit patterns in customer data [6].

This study takes the LightGBM algorithm, based on decision trees and a deep learning model [7]. To analyze bank customers' data and predict customer churn. The content of this study includes data preprocessing, features engineering, model training, and judgments. These results to judge the accuracy and effect of this model, the goal of this study is aim to explore the potential power of machine learning. This study aims to provide a useful, effective deep learning technology, for

applying this prediction model, banks can understand the customers more deeply, so that can make a more effective strategy, and increase the advance about them self. according to calculus, AUC scores can measure the ability of this model, this is an important indicator, except AUC scores, which also judge the accuracy, recall rate, and F1 scores. these indexes provide a view of different aspects and performances, to know the better predictive power of this model. according to calculus, AUC scores can measure the ability of this model, this is an important indicator, except AUC scores, I also judge the accuracy, recall rate, and F1 scores [8]. these indexes provide a view of different aspects and performance, to know the better predictive power of this model. The accuracy rate is used to measure the sample proportion of prediction, recall rate expression the true ability of the scan sample, F1 is the average value of the recall rate and the accuracy rate which is used to consider the balance of these two things comprehensively [9]. In addition, the growth of digital banking has intensified the need for banks to understand and predict customer behavior [10]. A Bain & Company report highlights that digital channels are increasingly becoming the preferred way for customers to conduct banking interactions, making it critical for banks to leverage advanced analytics to personalize service and retain customers. Additionally, a study published in the Journal of Business Research highlights the role of customer experience in digital banking, noting that a smooth and engaging digital experience can significantly reduce churn. By integrating these insights into the current study, this research aims to contribute to the development of more effective customer retention strategies in the digital banking era.

2. Data Part

2.1. Data Source

The data source is from Kaggle, Kaggle is an online platform specifically designed for data scientists and machine learning engineers, it provides datasets and other resources, and it contains a series of details containing customer Credit, Age, and Balance, this dataset was related to original, but not the same, its generated from a deep learning model trained on the Bank Customer Churn Prediction datasets. These details can help to get a more accurate result.

2.2. Data Features Engineering

This study used features engineer technical, which means using correlation background knowledge and specialty technical to deal with data, it makes a better process on the machine learning algorithm. Better features mean better flexibility, only the need to use a simple model can get better results, therefore, features engineering occupies a high importance, it is the most crucial and decisive factor in determining the success or failure of the result.

This research according to the definition a function to create new features, and use these new features to apply to the Datasets, for example as shown in the figure, according to divide 10 and multiplying by 10 then, can get an age range, for example, customers between the ages of 30 and 39 are grouped into 30 age group which was created by binning the Age of the customers. Senior_Citizen is created by the Age, define a threshold, if the Age is higher than this threshold, mark it as 1, and mark it as 0 if it is lower than the threshold. Credit_Tier was created based on the “pd. cut” function, it is based on every label, and it is divided into several intervals, like 0 to 4. Active_CreditCard_User was created based on multiplying whether customers hold credit or not and whether they are active members or not these two values, result in a new feature that can reflect whether the customer is actively using the card. Active_CC_User_by_Geo and Active_CC_User_by_Gender are according to the combination of the customer location, Active_CreditCard_User, and gender to create, for instance, if Active_CreditCard_User is 1 and also customer and gender are in a certain place, then provide the service condition based on these two features. Credit_Status is via inspecting the period of the customer relationship with the bank, for example, setting a threshold value, if the relationship between the customer and bank is less than 2 years, it is marked as 0, greater than 2 years, it is marked as 1.

2.3. Missing Value Presessing

Table 1 shows the original data in the data set before missing value detection. It can be seen that these data are not uniform, indicating that there are missing values in some classified data types. Usually, in a dataset, it's hard to avoid missing values, no matter what type of dataset, the first thing is to eliminate the missing value, this will disturb the quality of data analysis and the performance of machine learning.

Table 1. Data set before missing value processing

Column	Non-Null Count	Dtype
id	165034 non-null	int64
CustomerId	165034 non-null	int64
Surname	164000 non-null	object
CreditScore	165000 non-null	int64
Geography	165034 non-null	object
Gender	165020 non-null	object
Age	164500 non-null	float64
Tenure	165034 non-null	int64
Balance	160000 non-null	float64
NumOfProducts	165034 non-null	int64
HasCard	165034 non-null	float64
IsActiveMember	165034 non-null	float64
EstimatedSalary	165030 non-null	float64
Exited	165034 non-null	int64

Table 2 shows the table after processing the missing values, and it can be seen that the values are unified, total has 165034 non-null values, but in the Surname line, it only has 164000 non-null values, which means it has 1034 non-null values. This study used line deletion and column deletion, which bifurcates into two distinct pathways: firstly, delete rows that contain missing values if it is random, and secondly, delete all features if they lost most of the values [11].

Table 2. Data set after missing value processing

Column	Non-Null Count	Dtype
id	165034 non-null	int64
CustomerId	165034 non-null	int64
Surname	165034 non-null	object
CreditScore	165034 non-null	int64
Geography	165034 non-null	object
Gender	165034 non-null	object
Age	165034 non-null	float64
Tenure	165034 non-null	int64
Balance	165034 non-null	float64
NumOfProducts	165034 non-null	int64
HasCard	165034 non-null	float64
IsActiveMember	165034 non-null	float64
EstimatedSalary	165034 non-null	float64
Exited	165034 non-null	int64

After dealing with the missing values, obviously, in Figure 4, every column is 165034, which means it all has non-null values in every column, same as count, it is all non-null, which means there are no missing values. For the Dtype, "int64" intends integer, "float64" intends floating point, and "object" is an object type, this information helps understand dataset structure and other analysis, Knowing that the data set has no missing values can solve many problems, because you do not have to fill padding data. After processing the missing values, then it can visualize the data and deal with some feature selections.

2.4. Data Visualization

Data visualization is a visual data expression that uses computer graphics processing and computer vision technology to display different pattern styles and behaviors on the page. Fig. 1 is a Thermodynamic diagram, showing the correlation of each data, this expression can make observers clearly and intuitively understand the changing trend of data. This technology can be well used in machine learning to facilitate the detection of outliers. Identify those features that have the greatest impact on the target variable for feature selection.

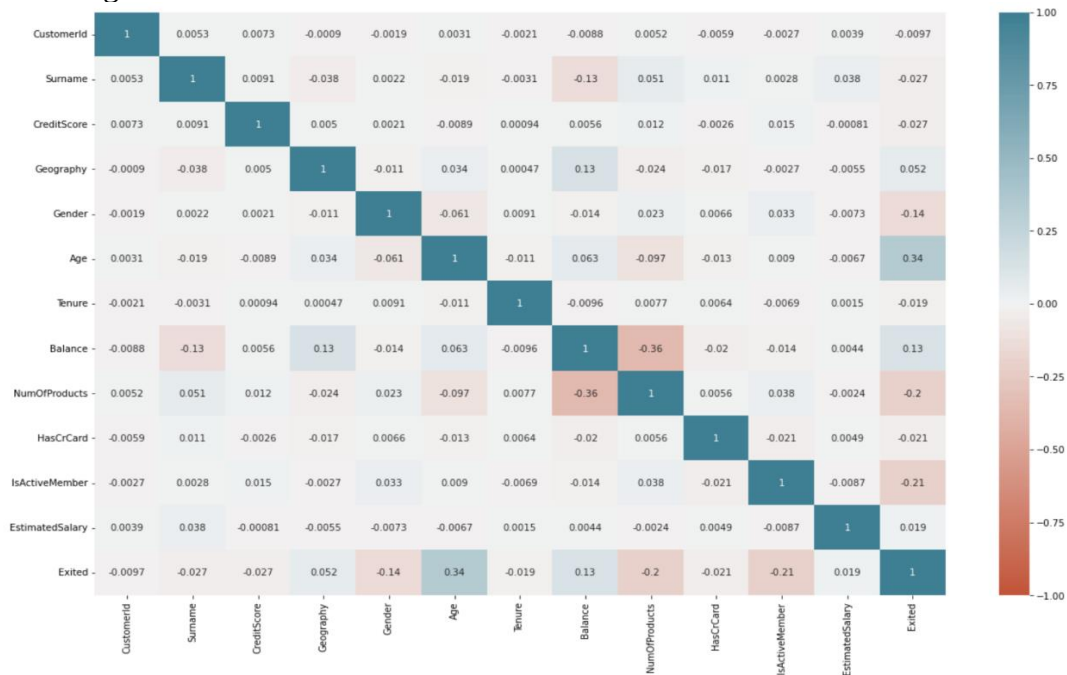


Fig. 1. A Thermodynamic diagram containing all the features in the data set

The thermodynamic diagram is a useful visual tool, it is used to show the correlation between data and features, in this data visual graph, the darkest color is

Closet to 1, the lightest color is closest to 1, and dark blue or other cool colors indicate close to negative 1, every cell represents a correlation coefficient, it can help us to find something that is useless or something helpful, for example, if there is a correlation that has a high positive correlation with the target variable, so this features will be helpful. If there is a high correlation between these features and another, probably only consider one of them. In Fig. 1, the correlation between NumOfProducts and Balance is -0.36, this is one of the darkest color blocks, -0.36 represents it has middle relevance which means that the more customers hold a product, the less account balance is. Age and Exited also had a 0.34 correlation which shows that there will be bank churn in relevant large-age customers and the opposite, the least correlation is Customer ID and Surname, because in the graph, these two things are closest to zero, but the dependency is just a measure, it can't shows every relationship, it needs context to make an accurate measurement, graph5 is specification use for liner measure, for example, it can't measure binary classification, because some variable is not liner, or maybe it has some complex problems [12].

3. Data Standardization

This is the final part of data processing, after this part, this study will move on to the training model part, usually machine learning separated into two steps, first is data processing, second is model training and this step is one of the important parts of machine learning. The main reason is that data standardization can raise algorithm performance, accelerate model convergence, and make the important features easy to explain. Data standardization is now divided into two main types, Standardization, and Normalization.

Standardized formula:

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma} \quad (1)$$

x standardized represents the standardized value, x is the original value, μ is the mean of the values, and σ is the standard deviation of the values.

Normalized formula:

$$x_{\text{normalized}} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (2)$$

x normalized: The normalized value of x , x : The original value, $\min(x)$: The minimum value in the set of values x , $\max(x)$: The maximum value in the set of values x [13].

4. Model Part

This study uses the LightGBM model, this model is based on GBDT which is called Gradient Boosting Decision Tree, which means using a Gradient Boosting structure to train the DT model, Gradient Boosting is used for regression and classification problems, it combines a series of weak model to build a strongest predict model, the main ideology reduces the mistake step by step so that enhance the power of model predictability. the basic theory of this algorithm is first, to use a relatively basic model to be a basic prediction, usually, the model used at the beginning is very simple, it just provides a fundamental framework. Secondly, iterative build is divided into three steps, the first step is to use every training data that it provides to calculate the mistake and rate of closure, then compare the difference between the true value and predicted value, then use the data that just get in last step as a target to training a new weak learner, after that, multiply the weak learner in Step 2 by a learning rate and then add it to the existing model, the purpose of this is upgrade models prediction. Third, repeat the first step and second step until it reaches the ideal number of iterations or this model reaches a satisfactory level of performance. Finally, combine these models, and take shape into the final model. The advantages of this type of algorithm are that it can optimize step by step, every step can reduce the mistakes of this model and every step can optimize the model. Second is because this algorithm does not have something special rules for the decision trees and regression models, so it's easy to start, and the method that it trains, this pattern of gradient lifting, can disengage the learner rate and iterations, avoiding matching effectively. Meanwhile, in this study, this algorithm is also because it main used mainly to deal with classification problems, so it just fits in this study. After selecting the model, need to adjust the model parameter, but in this study, the default hyper-parameter Settings are applied, but it can adjust according to self-demand. After completing this step, need to use the method of fit to embed into training data. After that assess this model, and do a probability prediction.

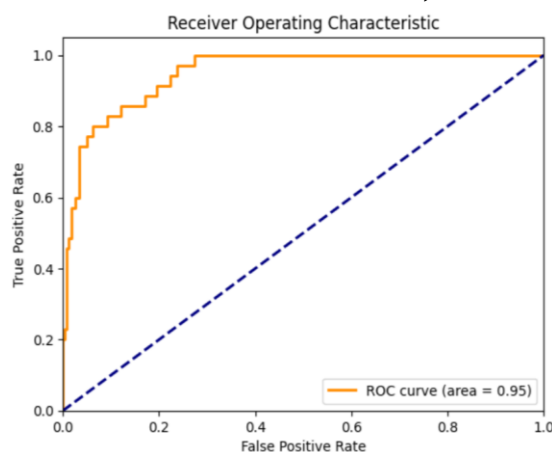


Fig. 2. Model roc score evaluation figure

Figure 2 shows the roc values of the model in this study. Generally speaking, the ROC-AUC score is used to evaluate a model, mainly to evaluate the performance of the model. After evaluating the

performance of this model, use the confusion matrix to move forward, after that, it is better to perform a hyper-parameter adjustment and cross-validation of the model, of course, this is not required, but cross-validation is still more important to evaluate the stability and generalization ability of the model [14].

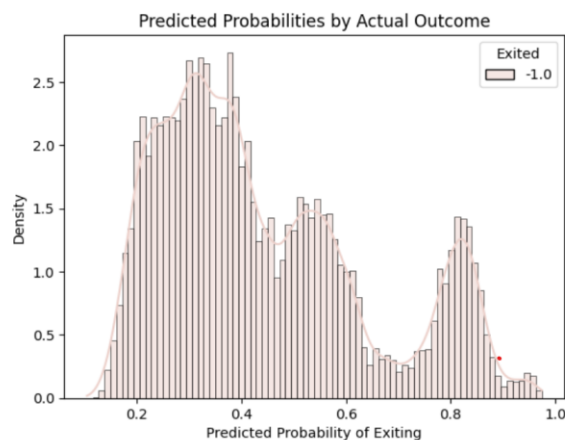


Fig. 3. Prediction result line figure

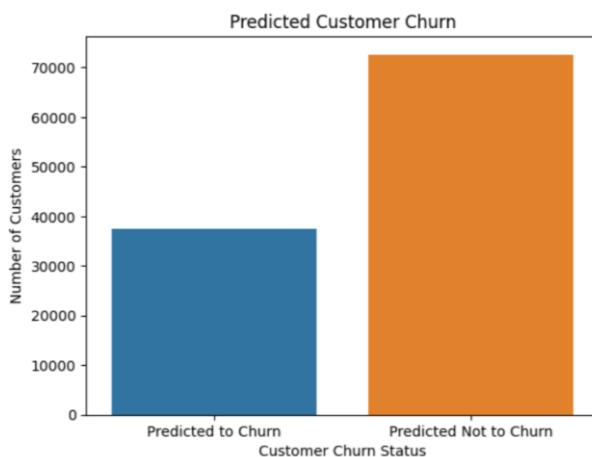


Figure 4. Bar figure of prediction results

Figs 3 and 4 are the prediction probability graphs of this study, it can be seen that in this study, the number of people who are not expected to lose is twice that of those who are expected to lose. LightGBM is mainly used as a major machine learning algorithm to predict the loss of bank customers in this study. Compared with traditional machine learning methods and other popular integrated learning frameworks, such as Random Forest and XGBoost, lightGBM shows significant advantages and uniqueness. Especially when it comes to the efficiency and performance of processing large-scale data. First of all, this model adopts gradient-based unilateral sampling and mutually exclusive feature bunching technology to optimize training speed and memory usage. In the loss prediction with a large number of large data, the model can be trained quickly, while reducing the consumption of computing resources, which brings many advantages compared with other model algorithms. Secondly, data preprocessing, missing value processing, and feature engineering creation in this study further improved the preprocessing ability of the model. This strategy provided more accurate and abundant information for the model, thus improving the accuracy of the prediction. In feature engineering, new features such as age range, Senior_Citizen, and Balance_Salary_Ratio are introduced to enable the model to capture deeper data patterns and associations, which are ignored by traditional models. In summary, by using the advanced LightGBM algorithm, combined with well-designed data preprocessing and feature engineering strategies, this project successfully improves the accuracy and efficiency of bank customer churn prediction. The comprehensive application of these methods makes this project perform well in processing large-scale data sets and improving model performance and also provides a valuable reference for future research on similar problems [15].

5. Conclusion

The findings of this study emphasize the powerful application of LightGBM in predicting bank customer churn. The findings show that strategic feature engineering and skillful handling of large datasets greatly improved the predictive accuracy of the model. The study provides meaningful insights into customer retention strategies for financial institutions, emphasizing the critical role of data-driven decision-making. However, it is important to recognize inherent limitations such as potential biases in the datasets and the generalizability of the model beyond the parameters of the current study. While the findings are promising, they depend on the context and nature of the data used. Future work could extend this study by incorporating more diverse data sources, thereby expanding the applicability and robustness of the model. In summary, the significance of this study lies in its detailed approach to a complex problem that provides a replicated methodology for similar challenges in the banking industry and beyond. By emphasizing the dynamic interplay between advanced algorithmic techniques and thoughtful data organization, this study sets a precedent for subsequent research in the field of predictive analytics.

References

- [1] Jordan, M. I., Mitchell, T. M.: Machine learning: trends, perspectives, and prospects. *Science*, 349 (6245), 255–260 (2020).
- [2] Radhakrishnan, A.: Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science* (2024).
- [3] Bensaoud, A.: A survey of malware detection using deep learning. *Machine Learning with Applications*, 16(1):100546 (2024).
- [4] Wardley, L. J.: A machine learning approach feature to forecast the future performance of the universities in Canada. *Machine Learning with Applications*, 100548 (2024).
- [5] Ahmed, J., Robert G.: Predicting severely imbalanced data disk drive failures with machine learning models. *Machine Learning with Applications*, 100361 (2021).
- [6] Solaiyappan, S., Yuxin W.: Machine learning based medical image deepfake detection: A comparative study. *Machine Learning with Applications*, 100298, (2022).
- [7] Jahangiri, S.: An inpatient fall risk assessment tool: Application of machine learning models on intrinsic and extrinsic risk factors. *Machine Learning with Applications*, 15(1): 100519 (2024).
- [8] Subasi, A.: Practical machine learning for data analysis using Python. Academic Press (2020).
- [9] Emmanuel, T., Maupong, T., Mpoeleng, D. et al.: A survey on missing data in machine learning. *Journal of Big Data*, 8, 1-37 (2021).
- [10] Fernando, M. P., Cèsar, F., David, N., et al.: Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7), 3217-3258 (2021).
- [11] Wang, H., Yuan, Z., Chen, Y., et al.: An industrial missing values processing method based on generating model. *Computer Networks*, 158, 61-68 (2019).
- [12] Sinar, E. F. Data visualization. In *Big Data at Work*, pp. 115-157. Routledge (2015).
- [13] Gal, M. S., Rubinfeld, D. L.: Data standardization. *Nyul Rev.*, 94, 737 (2019).
- [14] Ke, G., Meng, Q., Finley, T., et al.: Lightgbm A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30 (2017).
- [15] Wang, D., Zhang, Y., Zhao, Y.: LightGBM: an effective miRNA classification method in breast cancer patients. In: *Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics*, pp. 7-11 (2017).