Dean&Francis
ISSN 2959-6130

# Logistic Regression for Customer Churn Analysis: The Role of Data Proportion and Class Balancing in Model Performance

**Huimin Pan**

Department of Art and Science, Lehigh University, Bethlehem, United State of America
hup229@lehigh.edu

**Abstract:**

Customer churn prediction has become a critical challenge for the banking industry, as retaining existing clients is often more cost-effective than acquiring new ones. This study applies logistic regression to the Bank Customer Churn dataset (Kaggle, 2017), which contains 10,000 records and 13 features. After preprocessing, including removal of irrelevant identifiers, categorical encoding, and feature selection, seven key variables were retained. The research explores the impact of training set proportions (50%, 60%, 70%, 80%) and class balancing techniques (oversampling vs. no balancing) on model performance. Results show that model accuracy remains stable at approximately 0.81 across different training ratios, suggesting that increasing training size does not yield significant gains. In contrast, balancing the dataset reduced overall accuracy (0.72 vs. 0.81), reflecting the trade-off between accuracy and minority class recall in imbalanced classification. Logistic regression coefficients further revealed interpretable patterns: customers in Germany had higher churn odds, while active membership, longer tenure, and multiple product ownership reduced churn likelihood. These findings contribute to understanding how data preprocessing choices affect churn modeling outcomes and provide actionable insights for banking institutions seeking to strengthen retention strategies.

**Keywords:** Customer churn prediction; logistic regression; machine learning.

## 1. Introduction

In today's digital age, banks aren't just competing with each other—fintechs and new financial services are everywhere. This means customers have more choices than ever, making it easier to leave. The impact of churn is huge. Directly, it cuts revenue and raises operational costs. But there's hidden damage too: high churn signals poor service or irrelevant products, eroding trust and hurting the brand. There-

ISSN 2959-6130

fore, study churn is very important.

Recent studies have emphasized the significance of data preprocessing strategies such as the size of the training set, the distribution of categories, and feature selection in improving the predictive performance of customer churn models. For instance, Provost and Weiss pointed out that although an imbalanced training set may lead the model to higher accuracy, in cases where the Area Under the Curve (AUC) is used as the evaluation metric, a balanced category distribution (such as 50:50) is always better than other scenarios [1]. This finding was later confirmed by Albisua and Provost, who made further analysis and discovered that a category ratio of 30% to 70% typically yields the best AUC results in cost-sensitive learning environments [2]. Furthermore, Tantithamthavorn et al. studied the impact of re-balancing techniques such as oversampling, undersampling, and Synthetic Minority Oversampling Technique (SMOTE). The results showed that although these methods can increase the recall rate of minority classes, they may have adverse effects on the interpretability and accuracy of the model [3]. In addition to category balance, effective feature selection is also crucial. Raja et al. demonstrated that applying feature selection methods based on wrappers and filters can significantly improve the accuracy of telecommunications customer churn prediction models [4].

Although there has been an increasing amount of research on customer churn prediction in recent years, many scholars have used machine learning models to model churn. However, studies on the combined influence of training set proportion, class balance, and variable selection are still insufficient, especially in bank customer churn prediction models based on logistic regression. Existing literature often analyzes these factors separately or uses them with default parameters, lacking systematic evaluation of their combined application, especially in comparisons across multiple metrics such as accuracy, recall rate, and AUC. At the same time, although oversampling and SMOTE and other class imbalance re-balancing techniques have been widely adopted [5-7], few studies have explored their effects when combined with different training set proportions. Therefore, this paper attempts to fill this research gap, from an experimental perspective, systematically analyzing the impact of training proportion (50%, 60%, 70%, 80%) and class balance techniques on model performance, and further examining the improvement effect of feature selection on model accuracy and interpretability, thereby providing practical data processing suggestions for customer churn prediction modeling in the financial industry.

# 2. Method

## 2.1 Dataset Preparation

The source of the dataset is from Kaggle's Bank Customer Churn Prediction dataset and contains 10,000 customer records with 13 features [8]. These features include demographic information such as Geography (France, Spain, Germany) and Gender (Male/Female), numerical attributes such as CreditScore, Age, Tenure (years as a customer), Balance, NumOfProducts (number of bank products), HasCrCard (credit card ownership), IsActiveMember (activity status), and EstimatedSalary, along with three identifier fields (RowNumber, CustomerId, Surname). The target variable is Exited, indicating whether a customer churned (binary classification: 0 for stayed, 1 for churned). Preprocessing steps included removing irrelevant identifier columns and the variable that just have a small impacts such as RowNumber, CustomerId, Surname, CreditScore, HasCrCard and EstimatedSalary, converting categorical variables into numerical factors, and splitting the dataset into an 80% training set and a 20% test set. Feature selection retained 7 relevant features for modeling, and no normalization was applied since logistic regression in R handles mixed data types without prior scaling [8].

## 2.2 Logistic Regression

Logistic regression is a statistical method that is used for binary classification problems [9, 10]. The gold of logistic regression is to predict a data point that belonging to one of two classes. Logistic regression is widely used in binary classification problems due to its simplicity, high computational efficiency, and ease of interpretation. It can produce probability outputs, which are helpful for adjusting the decision threshold and making decisions. The coefficients can clearly reveal the degree and direction of the influence of features on the log-odds of the outcome. However, this method inherently assumes a linear relationship between the predictor variables and the log-odds, is sensitive to multicollinearity among predictor variables, and can only model linear decision boundaries. Therefore, it is not very suitable for datasets with complex nonlinear patterns, unless feature engineering or transformation is performed. Additionally, logistic regression may be overfit on high-dimensional datasets without regularization, and is susceptible to the influence of outliers, which may distort coefficient estimates and model predictions

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_2 + \ldots + \beta_n x_n}} \quad (1)$$

$\beta_0$ is the intercept, $\beta_i$ are the coefficients, $x_i$ are the input features. The coefficients ( $\beta_i$ ) indicate the direction and

strength of the relationship between each feature and the log-odds of the outcome. By applying a threshold, logistic regression classifies an observation into one of the two categories.

## 3. Result and Discussion

Fig. 1 illustrates the relationship between the proportion of data allocated to the training set and the classification accuracy of the logistic regression model used to predict customer churn. This stability indicates that within the tested proportion range, increasing the size of the training set does not significantly improve the model's predictive performance. One possible explanation is that even with a 50% data set for training, the model has already been able to obtain a sufficient number of representative samples to learn the intrinsic patterns of customer churn behavior. As a relatively simple linear classifier, the logistic regression model may have reached its optimal state before its performance reaches its limit, and the additional training data brings negligible gains. Furthermore, the results also indicate that the model's accuracy is less sensitive to changes in the proportion of the training set. This might be due to the balanced distribution of the main features in the dataset and the lack of highly complex nonlinear relationships.

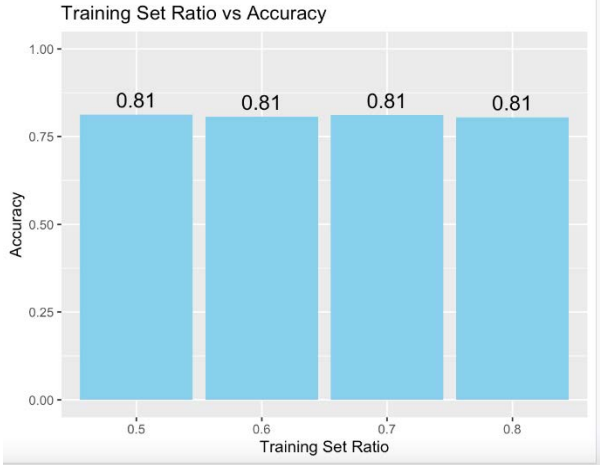


**Fig. 1 Accuracy result (Picture credit : Original)**

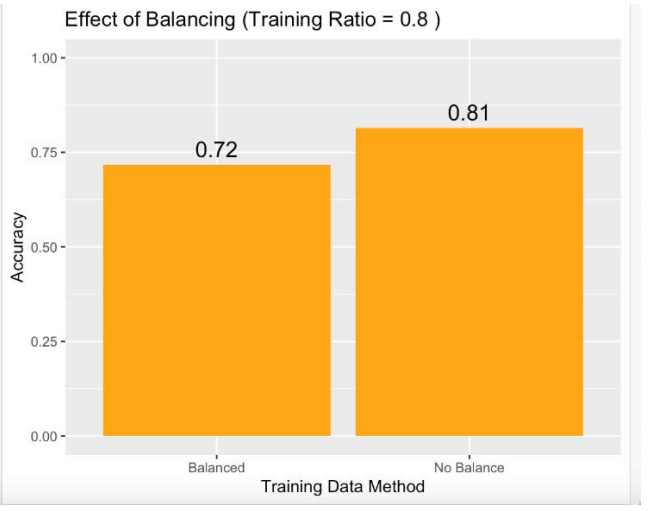


**Fig. 2 Effect of Balancing (Picture credit : Original)**

This bar chart shown in Fig. 2 compares the prediction accuracy of the logistic regression model under two training data configurations (balanced and unbalanced). The fixed training set proportions for both configurations are 80%. In the "balanced" configuration, oversampling is applied to the data to make the number of churned and non-churned customers equal; while in the "unbalanced" configuration, the original class distribution is used, which is biased towards non-churned customers. The results show that the accuracy rate of the unbalanced processing model is 0.81, which is higher than that of the balanced processing model (0.72). This seemingly abnormal phenomenon can be explained by the dependence of the accuracy rate on the majority class. In the unbalanced dataset, even if the model's prediction for the minority class (lost customers) is poor, as long as it can mostly correctly predict the majority class (non-lost customers), the accuracy rate will be "raised". However, in the balanced dataset, the model is forced to treat both types of samples equally. Although it may increase the recall rate of the minority class, the prediction accuracy rate of the majority class may decrease, thereby leading to a decrease in the overall accuracy rate. Therefore, although the "unbalanced" configuration seems to perform better in terms of accuracy, it may not necessarily have an advantage in identifying churned customers. Other evaluation metrics, such as precision, Recall, F1 score, or AUC, should be considered to conduct a fairer assessment of the model's performance in the context of imbalanced classification problems.

The logistic regression model developed in this study can predict the logarithmic probability of customer churn based on several demographic and behavioral variables. The estimation formula is as follows:

ISSN 2959-6130

$$logit(p) = -3.559 + 0.8024xGeographyGermany + 0.05274xGeographySpain - 0.4898xGenderMale + 0.06978xAge - 0.03131xTenure + 0.000002516xBalance - 0.1556xNumOfProducts - 1.047xIsActiveMember \quad (2)$$

In this model, positive coefficients indicate that the corresponding variable increases the log odds and the probability of churn, while negative coefficients indicate a protective effect against churn. For example, customers located in Germany (0.8024) have substantially higher odds of churn compared to the reference category, France, whereas male customers (-0.4898) are less likely to churn than females. Age is positively associated with churn, with each additional year increasing the odds by a small margin (0.06978). Longer tenure (-0.03131) and a greater number of products owned (-0.1556) both reduce churn risk. The coefficient for balance is nearly zero, indicating minimal impact on churn probability. The most influential protective factor is active membership (-1.047), showing that active customers are much less likely to leave the bank.

## 4. Conclusion

This study demonstrates the effectiveness and limitations of logistic regression in predicting bank customer churn. The experiments confirm that training set proportion has minimal influence on overall accuracy, as performance stabilizes once sufficient representative samples are included. However, data balancing strategies significantly affect outcomes: while unbalanced datasets achieve higher accuracy due to dominance of the majority class, balanced datasets improve fairness in learning but at the cost of reduced accuracy. The interpretability of logistic regression provides valuable business insights, highlighting that factors such as geography, gender, tenure, and activity level play significant roles in churn behavior. Nevertheless, the model's reliance on linear assumptions and sensitivity to class imbalance limit its predictive power. Future research should extend this work by incorporating advanced methods such as ensemble learning or deep learning, along with evaluation metrics beyond accuracy, to achieve more robust predictions. Overall, this study underscores the importance of carefully considering training data composition and feature selection in churn modeling, offering both methodological guidance and practical implications for customer retention in the financial sector.

## References

[1] Weiss GM, Provost F. Learning when training data are costly: The effect of class distribution on tree induction. J Artif Intell Res. 2003 Oct 1;19:315-54. Fangfang. Research on power load forecasting based on improved BP neural network [dissertation]. Harbin: Harbin Institute of Technology; 2011.

[2] Albisua I, Arbelaitz O, Gurrutxaga I, Lasarguren A, Muguerza J, Pérez JM. The quest for the optimal class distribution: an approach for enhancing the effectiveness of learning via resampling methods for imbalanced data sets. Prog Artif Intell. 2013 Mar;2(1):45-63.

[3] Tantithamthavorn C, Hassan AE, Matsumoto K. The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. IEEE Trans Softw Eng. 2018 Oct 17;46(11):1200-19.

[4] Raja JB, Sandhya G, Peter SS, Karthik R, Femila F. Exploring effective feature selection methods for telecom churn prediction. Int J Innov Technol Explor Eng. 2020;9(3).

[5] Wu X, Meng S. E-commerce customer churn prediction based on improved SMOTE and AdaBoost. In: 2016 13th International Conference on Service Systems and Service Management (ICSSSM). IEEE; 2016. p. 1-5.

[6] Amin A, Rahim F, Ali I, Khan C, Anwar S. A comparison of two oversampling techniques (SMOTE vs MTDF) for handling class imbalance problem: A case study of customer churn prediction. In: New Contributions in Information Systems and Technologies. Vol 1. Cham: Springer International Publishing; 2015. p. 215-25.

[7] Harshini A, Nallagorla NSR, Bathula ST, Chandra S, Syed S. Improving customer churn prediction accuracy: A SMOTE-based approach. In: 2024 8th International Conference on Inventive Systems and Control (ICISC). IEEE; 2024. p. 215-22.

[8] KartikSaini18. Churn Bank Customer [Internet]. 2025. Available from: https://www.kaggle.com/datasets/kartiksaini18/churn-bank-customer

[9] LaValley MP. Logistic regression. Circulation. 2008;117(18):2395-9.

[10] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken: Wiley; 2013.