WEP Warwick Evans Publishing

International Journal of Global Economics and Management

ISSN: 3005-9690 (Print), ISSN: 3005-8090 (Online) | Volume 10, Number 4, Year 2026
DOI: https://doi.org/10.62051/ijgem.v10n4.06
Journal homepage: https://ijgem.org

Machine Learning Methods for Bank Term Deposit Subscription Prediction

Xiaoqian Wu

Chongqing Jiaotong University, Chongqing, China

ABSTRACT

To address the problem of customer purchase behavior prediction in bank term-deposit services, this paper develops a machine-learning-based prediction framework using the bank marketing dataset. First, the raw data are processed through missing-value inspection, categorical variable encoding, relevant feature selection, imbalance handling, and data standardization. Then, three models, namely Random Forest, Logistic Regression, and LightGBM, are constructed to analyze customer attributes, marketing-related attributes, and economic background attributes. Experimental results show that, among the three models, LightGBM achieves the best overall performance, with an accuracy of 91.95% and a test F1-score of 60.22%, outperforming both Random Forest and Logistic Regression. Further analysis indicates that features such as call duration and campaign frequency have strong influence on customer subscription decisions. The results demonstrate that machine learning methods can effectively improve the accuracy of identifying potential term-deposit customers, and provide useful data support for precision marketing and resource optimization in banking services.

KEYWORDS

Bank marketing; Term-deposit Subscription Prediction; Machine Learning; LightGBM; Feature Engineering

1. INTRODUCTION

With the development of digital finance and precision marketing, banking services are gradually shifting from experience-driven practices to data-driven decision-making. For term-deposit services, identifying target customers who are more likely to subscribe directly affects marketing costs and conversion performance. Therefore, customer purchase behavior prediction has become an important research issue in intelligent bank marketing [1, 2].

Existing studies have shown that machine learning methods can effectively improve the accuracy of bank telemarketing prediction. Compared with traditional linear models, ensemble methods such as Random Forest, XGBoost, and LightGBM usually have stronger capabilities in handling nonlinear relationships and feature interactions [3, 6, 7].

However, this task still faces two practical challenges. First, the data contain both numerical and categorical variables, making preprocessing relatively complex. Second, the number of "subscribed" samples is significantly smaller than that of "non-subscribed" samples, and this class imbalance reduces the model's ability to identify target customers [5].

To address these issues, this paper develops a unified analytical framework for bank term-deposit customer prediction, including categorical variable encoding, correlation analysis, imbalance handling, and standardization, and compares Logistic Regression, Random Forest, and LightGBM

under the same experimental setting. Experimental results show that, on an imbalanced dataset with an approximate ratio of 8:1, the accuracies of all three models exceed 90%. Among them, LightGBM achieves the best overall performance, with an accuracy of 91.95% and a test F1-score of 60.22%, outperforming the other two models. Further analysis indicates that variables such as duration and campaign are strongly associated with customer subscription behavior. These results demonstrate that a standardized data-processing pipeline combined with appropriate machine learning models can effectively support bank customer screening and precision marketing.

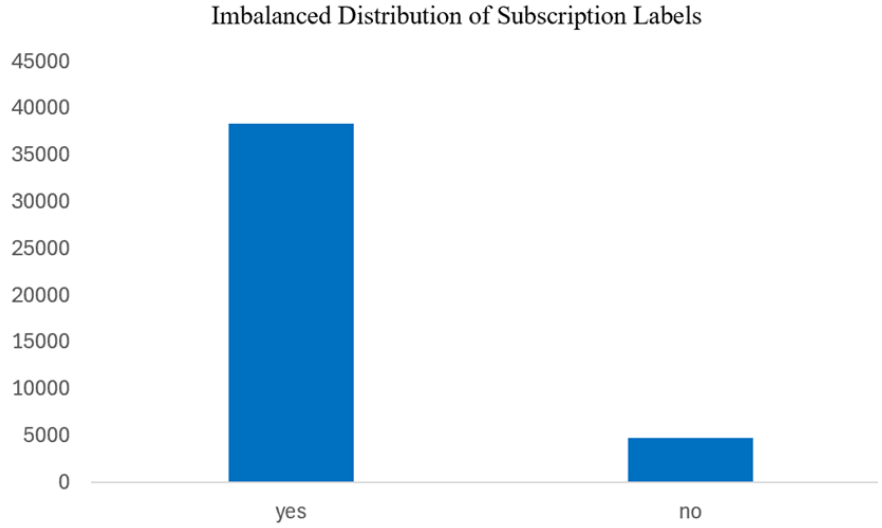

Figure 1. Imbalanced distribution of subscription labels in the bank marketing dataset

2. RELATED WORK

Bank telemarketing prediction is a typical binary classification task in financial marketing. Moro et al. conducted early research based on real bank marketing data and established this problem as a classic benchmark for subsequent studies [1].

On this basis, related research has gradually expanded from traditional classification models to ensemble learning and interpretability analysis frameworks. Existing studies indicate that models such as Logistic Regression have good interpretability, while Random Forest and boosting tree models usually achieve better predictive performance [2, 3, 4].

3. METHOD

We improve the conventional bank term-deposit subscription prediction framework by integrating categorical encoding, feature screening, imbalance handling, and data standardization into a unified pipeline [10]. This framework is designed to provide a more suitable input representation for bank marketing data and to enable a fair comparison among different machine learning models. The main idea is to reduce noise and class bias in the original data, so that the models can better identify potential customers for term-deposit subscription.

3.1. Baseline Prediction Framework

The problem studied in this paper is the prediction of customer purchase behavior for bank term deposits, which is essentially a binary classification task. Given customer-related features, the model is required to determine whether a customer will subscribe to a bank term-deposit product.

A conventional prediction framework usually consists of three basic steps: input feature construction, data preprocessing, and classifier training. Specifically, the raw marketing data are first transformed

into a feature representation suitable for modeling, and the processed features are then fed into a classifier to generate the final prediction of customer purchase behavior [4]. For this type of tabular bank marketing data, Logistic Regression, Decision Trees, and their ensemble variants are the most common baseline modeling approaches.

In the baseline prediction framework adopted in this study, the input variables are mainly composed of three parts. The first part includes customer profile attributes, such as age, job, marital status, education level, credit status, housing loan, and personal loan information [3]. The second part consists of marketing-related attributes, such as contact communication type, contact month, day of the week, call duration, number of contacts, and previous campaign outcomes [2].

The third part contains economic background attributes, such as employment variation rate, consumer price index, consumer confidence index, three-month interest rate, and number of employees [6]. The target variable is whether the customer subscribes to a term deposit, with labels represented as "yes" or "no". Therefore, the core objective of the baseline prediction framework is to characterize customer status using these multi-source features and establish the mapping relationship between input features and purchase outcomes, thereby providing a unified baseline for subsequent improved data processing and model comparison.

3.2. Categorical Encoding and Feature Screening

Since the bank marketing data contain both numerical and categorical variables, the raw features cannot be directly used as inputs to classification models. Therefore, this study first performs numerical transformation on categorical variables and removes irrelevant and redundant features to construct an input representation suitable for subsequent modeling. Let x_j denote an original categorical variable. Its encoding process can be expressed as

$$x_j^{(\mathrm{enc})} = g_j(x_j),\ j \in \mathcal{C}$$

Where $g_j(\cdot)$ denotes the label encoding mapping function corresponding to the j-th categorical variable, $\mathcal{C}$ represents the set of all categorical variables, and $x_j^{(\mathrm{enc})}$ denotes the encoded numerical feature. Through this mapping, the original categorical variables are transformed into numerical representations that can be used as inputs to classification models.

After encoding, correlation analysis is further applied for feature screening. Let y denote the target variable. Then, the selected feature set is defined as

$$S = \{x_j \mid |\, corr(x_j, y) \,| > \delta\}, \quad \delta = 0.35$$

Where $\mathrm{corr}(x_j, y)$ denotes the correlation coefficient between feature x_j and the target variable y, δ is the screening threshold, and S represents the final selected feature set. According to the original experimental setting, δ is set to 0.35, and 9 variables with relatively strong correlations to customer subscription behavior are retained [9]. This process provides a more effective feature space for subsequent imbalance handling and model training.

3.3. Change-Point Aware Attention Mechanism

Since the number of "subscribed" samples is significantly smaller than that of "non-subscribed" samples in the bank marketing data, the class distribution is highly imbalanced, which reduces the model's ability to identify minority-class customers [8]. To alleviate this issue, the SMOTE method is applied to the training set for minority-class oversampling. Let x_i denote a minority-class sample and x_{zi} denote one of its nearest minority-class neighbors. Then, a synthetic sample can be generated as

$$x_{\text{new}} = x_i + \lambda(x_{z_i} - x_i), \quad \lambda \in [0,1]$$

Where x_i represents the original minority-class sample, x_{zi} represents a neighboring minority-class sample, and λ is a random coefficient in the interval [0, 1]. This equation shows that the new sample is generated by linear interpolation between two nearby minority-class samples rather than by simply duplicating the original sample. In this way, the distribution of minority-class samples can be expanded and the risk of overfitting can be reduced.

After imbalance handling, feature standardization is further conducted to reduce the influence of different measurement scales on model training. Let x_j denote the original feature value, and let μ_j and σ_j denote the mean and standard deviation of the j-th feature, respectively. Then, the standardized feature can be expressed as

$$z_j = \frac{x_j - \mu_j}{\sigma_j}$$

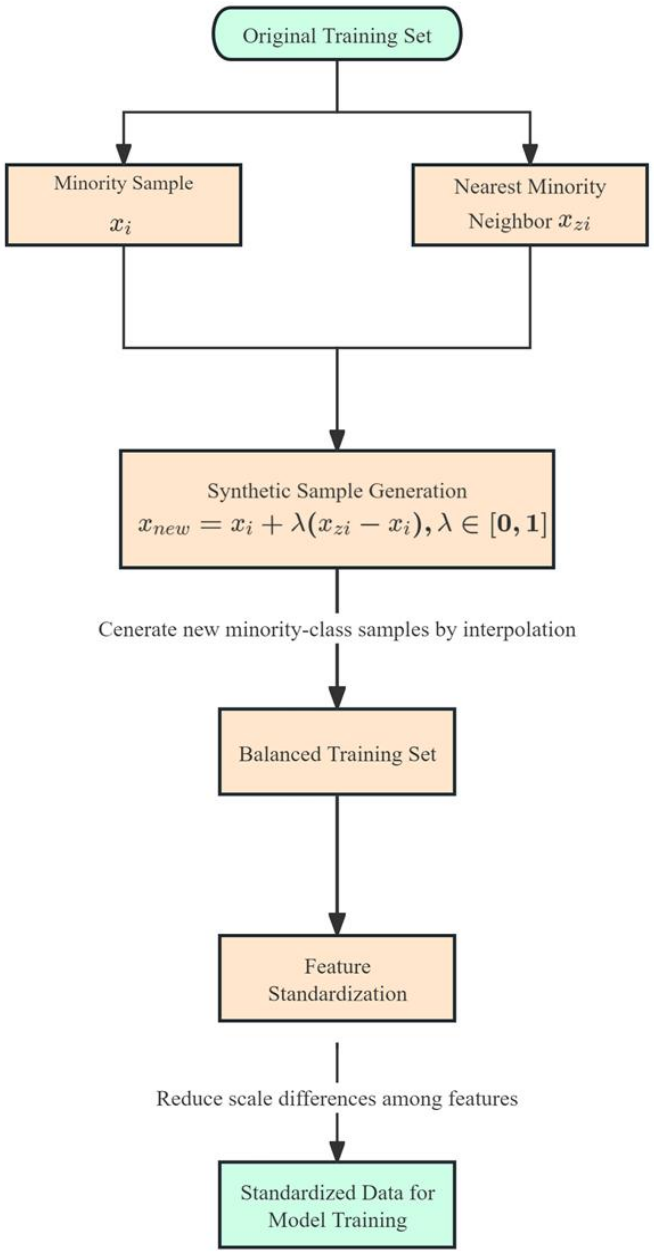

Figure 2. SMOTE-based imbalance handling and data standardization process

3.4. Model Integration and Comparative Framework

After categorical encoding, feature screening, imbalance handling, and data standardization, the processed feature matrix is used as the unified input for three classification models: Logistic Regression, Random Forest, and LightGBM. These three models represent a linear model, a Bagging-based ensemble model, and a Boosting-based ensemble model, respectively, and can therefore reflect the applicability differences of different machine learning methods in the bank marketing prediction task.

To ensure a fair comparison, all three models are trained and tested under the same data partition and preprocessing conditions. Specifically, the training data are first balanced and standardized, and then the same input features are fed into different classifiers to learn the mapping from customer attributes to term-deposit subscription outcomes. Through this unified modeling framework, the performance differences among the three models can be mainly attributed to their modeling mechanisms rather than inconsistent preprocessing conditions.

4. EXPERIMENTAL SETUP

4.1. Dataset

This study conducts experiments on the publicly available bank marketing dataset obtained from the Kaggle platform. The task aims to predict whether a customer will subscribe to a bank term-deposit product based on customer attributes, marketing-related attributes, and economic background attributes. The dataset contains 21 input features, and the target variable is a binary label, i.e., "yes" or "no." According to the nature of the variables, all features can be divided into three categories: customer profile attributes, marketing-related attributes, and economic background attributes.

According to the original statistical results, there are no missing values in any of the features in the dataset. Meanwhile, the label distribution is clearly imbalanced, with the ratio of "no" to "yes"v being approximately 8:1. Based on the missing-value statistics and descriptive results, the dataset is divided into a training set and a test set at a ratio of 7:3 for subsequent model training and performance evaluation.

4.2. Evaluation Metrics

To comprehensively evaluate the classification performance of different models in the bank term-deposit customer prediction task, this study adopts Accuracy, Precision, Recall, and F1-score as evaluation metrics. Accuracy is used to measure the overall proportion of correctly classified samples. Precision reflects the proportion of truly subscribed customers among the samples predicted as subscribed. Recall represents the model's ability to identify actual subscribed customers. F1-score, which jointly considers Precision and Recall, can more effectively reflect the overall performance of a model in imbalanced classification tasks.

Since the number of "yes" samples in this study is significantly smaller than that of "no" samples, using Accuracy alone is not sufficient to fully evaluate the model's ability to identify minority-class customers. Therefore, this study pays particular attention to Precision, Recall, and F1-score, among which F1-score is used as the key metric for comparing the performance differences among different models in the target customer identification task.

4.3. Implementation Details

The experiments in this study are implemented in the Python environment. Following the proposed procedure, categorical variables are first transformed by label encoding, and feature screening is then performed based on correlation analysis. After that, the SMOTE method is applied to the training set

for imbalance handling, and feature standardization is further conducted. The processed data are then uniformly fed into three classification models, namely Logistic Regression, Random Forest, and LightGBM, for training and testing.

To ensure a fair comparison among different models, all experiments are conducted under the same data partition and preprocessing conditions. Specifically, the dataset is divided into the training set and the test set at a ratio of 7:3, and the random seed is set to 0. These three models represent a linear model, a Bagging-based ensemble model, and a Boosting-based ensemble model, respectively, and thus can provide a relatively comprehensive comparison of the applicability of different machine learning methods in the bank marketing prediction task.

Table 1. Key Experimental Settings

Experimental Setting	Value
Number of Features	21
Selected Features	9
Class Ratio (No: Yes)	8:1
Train/Test Split	7:3
Correlation Threshold	0.35
Random Seed	0

5. RESULTS AND DISCUSSION

Comparative experiments were conducted on the test set using Logistic Regression, Random Forest, and LightGBM to evaluate the performance of different methods in the bank term-deposit customer prediction task.

5.1. Quantitative Comparison

Comparative experiments are conducted on the test set to evaluate the performance of Logistic Regression, Random Forest, and LightGBM for bank term-deposit subscription prediction. The results show that all three models achieve satisfactory classification performance, with accuracies above 90%. Among them, LightGBM obtains the best overall performance, indicating its stronger applicability to the imbalanced bank marketing dataset.

5.2. Analysis and Discussion

Table II presents the detailed classification results of the three models on the test set. As shown in the table, LightGBM achieves the highest accuracy of 91.95% and the highest test F1-score of 60.22%, outperforming Random Forest (90.48%, 56.28%) and Logistic Regression (91.01%, 49.93%). These results indicate that LightGBM provides the most competitive overall performance in the bank term-deposit customer prediction task.

Although the differences in accuracy among the three models are relatively small, the differences in F1-score are more obvious. Since the dataset is highly imbalanced, accuracy alone is insufficient to fully reflect the model's ability to identify subscribed customers. In this case, F1-score is more suitable for evaluating the effectiveness of different models on minority-class recognition. The higher F1-score of LightGBM suggests that it achieves a better balance between precision and recall under the imbalanced data condition.

Figure. 3 further shows the confusion matrix of the best-performing LightGBM model on the test set. It can be observed that the model correctly identifies a large number of both subscribed and non-subscribed customers, demonstrating good classification stability. At the same time, some misclassified samples still exist, which indicates that customer identification remains challenging

under the class-imbalanced bank marketing scenario. The confusion matrix results are consistent with the quantitative comparison in Table II and further confirm the advantage of LightGBM in this study.

In addition, the original experimental analysis shows that duration and campaign are strongly related to customer subscription behavior. This suggests that customer decisions are influenced not only by personal profile attributes but also by the interaction process in the marketing campaign. Therefore, under

the unified preprocessing and modeling framework proposed in this study, LightGBM not only achieves the best classification performance but also provides more practical value for customer screening and precision marketing in banking services.

Table 2. Classification Results of the Three Models on the Test Set

Model	Accuracy	Train F1-score	Test F1-score
Random Forest	90.48%	98.15%	56.28%
Logistic Regression	91.01%	47.78%	49.93%
LightGBM	91.95%	65.42%	60.22%

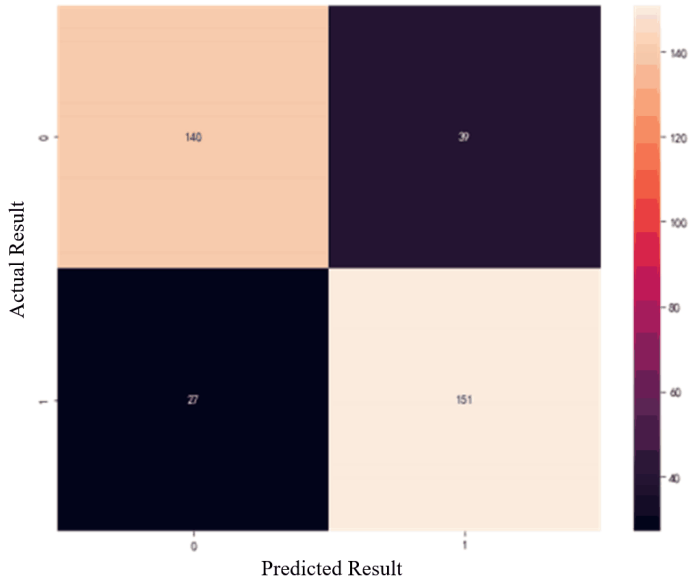

Figure 3. Confusion Matrix of the Best-Performing LightGBM Model on the Test Set

6. CONCLUSION

This paper focuses on the prediction of customer purchase behavior for bank term deposits and develops a unified machine-learning-based analytical framework that includes categorical variable encoding, feature screening, imbalance handling, and data standardization. Under the same experimental setting, three models, namely Logistic Regression, Random Forest, and LightGBM, are compared. Experimental results show that all three models can effectively accomplish the prediction task, with classification accuracies above 90%. Among them, LightGBM achieves the best overall performance, with an accuracy of 91.95% and a test F1-score of 60.22%, outperforming Random Forest (90.48% / 56.28%) and Logistic Regression (91.01% / 49.93%). These results indicate that, on the current imbalanced bank marketing dataset, a proper data-processing pipeline combined with an appropriate machine learning model can effectively improve the identification of target customers.

This study suggests that the superior performance of LightGBM mainly comes from its stronger ability to capture nonlinear relationships and feature interactions in tabular data. Compared with linear models such as Logistic Regression, LightGBM can better learn the complex associations

among customer attributes, marketing-related attributes, and economic background attributes. Compared with Random Forest, LightGBM demonstrates better overall generalization ability in the current task. Meanwhile, the categorical encoding, correlation-based feature screening, SMOTE-based imbalance handling, and standardization adopted in this study also provide a cleaner and more effective input space for model training.

The experimental results further show that marketing-related variables such as duration and campaign are strongly associated with customer subscription behavior, indicating that customer decisions are influenced not only by personal attributes but also by the interaction process in the marketing campaign.

Although the proposed method achieves promising results, several limitations should be acknowledged. First, the study is validated only on a single bank marketing dataset, and its generalization ability under other banking scenarios or different data distributions still requires further investigation. Second, the selected feature-screening threshold and imbalance-handling strategy may affect the final model performance, and these settings may need to be re-adjusted for different datasets.

Finally, this study mainly compares three classical machine learning models and has not yet incorporated more advanced ensemble strategies or interpretability analysis methods. Future work can be extended to multi-dataset validation, feature engineering optimization, automatic parameter tuning, and model interpretability enhancement, so as to further improve the practical value of the proposed framework in real banking applications.

REFERENCES

[1] Moro, S., Cortez, P., and Rita, P. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 2014, 62: 22–31.

[2] Yu, J.-M., and Cho, S.-B. Prediction of Bank Telemarketing with Co-training of Mixture-of-Experts and MLP. In Neural Information Processing: 23rd International Conference, ICONIP 2016, Proceedings, Part IV, 2016, pp. 52–59.

[3] Tékouabou, S. C. K., Gherghina, Ş. C., Toulni, H., Neves Mata, P., Mata, M. N., and Martins, J. M. A Machine Learning Framework towards Bank Telemarketing Prediction. Journal of Risk and Financial Management, 2022, 15(6): 269.

[4] Xie, C., Zhang, J.-L., Zhu, Y., Xiong, B., and Wang, G.-J. How to improve the success of bank telemarketing? Prediction and interpretability analysis based on machine learning. Computers & Industrial Engineering, 2023, 175: 108874.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357.

[6] Chen, T., and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems 30, 2017, pp. 3146–3154.

[8] Khan, M. Z., Munquad, S., and Rao, T. S. M. A Study on Improving Banking Process for Predicting Prospective Customers of Term Deposits using Explainable Machine Learning Models. In Proceeding of International Conference on Computational Science and Applications, 2022, pp. 93–103.

[9] Safarkhani, F., and Moro, S. Improving the Accuracy of Predicting Bank Depositor's Behavior Using a Decision Tree. Applied Sciences, 2021, 11(19): 9016.

[10] Feng, Y., Yin, Y., Wang, D., and Dhamotharan, L. A dynamic ensemble selection method for bank telemarketing sales prediction. Journal of Business Research, 2022, 139: 368–382.