ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

ROS-LightGBM-Based Prediction of Term Deposit Subscriptions in Bank Telemarketing and CRM Strategy Analysis

Yuxuan Chen[a], Exiang Luo[b]

Business School, University of Shanghai for Science and Technology, Shanghai, 200093, China
[a]daniel_chencc@foxmail.com, [b]luoex126@126.com

Abstract: *Deposits are an important source of funding that enables commercial banks to maintain liquidity and operate prudently, while telemarketing is a major channel through which banks promote term deposit products. A key challenge in precision marketing is to identify prospective subscribers while reducing ineffective contacts and to translate predictions into customer relationship management actions. Using the UCI Bank Marketing dataset, this study develops a five-stage framework that integrates model comparison, class imbalance treatment, feature selection, SHAP interpretation, and customer relationship management. By comparing seven types of machine learning models and multiple sampling methods and ratios, the study identifies the optimal Optuna-tuned ROS-LightGBM model, which achieves an 88.78% recall and a 93.17% ROC-AUC on a test set that retains the original class distribution. SHAP analysis shows that marketing interaction features account for 74.31% of the total contribution, with call duration, contact channel, previous campaign outcome, and month exerting the strongest effects. This study further combines predicted probabilities with customer balances to classify customers into four groups: priority conversion, potential cultivation, relationship maintenance, and low-cost maintenance. It then proposes differentiated contact and maintenance strategies, establishing a practical pathway from prospective customer identification and prediction interpretation to customer relationship management, and providing a reference for the refined management of bank term deposit telemarketing.*

Keywords: *Bank Telemarketing; Class Imbalance; LightGBM; SHAP; CRM*

1. Introduction

Deposits are an important source of funding that enables commercial banks to maintain liquidity, support lending, and operate prudently[1,2]. Adequate and stable deposits can alleviate depositors' concerns about bank solvency and liquidity while strengthening banks' ability to withstand financial shocks[3]. As market competition and capital constraints intensify, attracting customers to subscribe to deposit products at a lower cost has become an important task in bank marketing management[4,5]. Compared with undifferentiated mass promotion, direct marketing can offer specific products based on customer information, reduce transaction costs, and obtain timely feedback[6]. Telemarketing, characterized by remote contact, direct communication, and relatively low organizational costs, has therefore become an important channel through which banks sell term deposits and other financial products[7].

However, the low cost of telemarketing does not justify contacting customers indiscriminately and repeatedly. Because customers differ in their needs, financial circumstances, and willingness to respond, unscreened calls not only consume agent time and marketing resources but may also inadvertently cause customers to feel disturbed, thereby damaging customer experience and long-term relationships[8,9]. The central issue in bank telemarketing is therefore not simply to expand call volume, but to use existing customer profiles and historical campaign information to identify customers who are more likely to subscribe[10]. Banks can then determine whom to contact, through which channel, at what time, and how frequently, thereby coordinating short-term product conversion with long-term customer relationship maintenance[11,12]. Customer relationship management research also shows that firms need to adjust communication and resource allocation according to customer value and relationship stage, and convert customer information into sustainable managerial actions[13,14].

As banks accumulate increasing volumes of customer attributes, account information, and campaign

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

records, data mining offers new methods for telemarketing decisions. Moro et al. compared LR, DT, neural networks, and SVM, demonstrating that data-driven models can predict bank telemarketing outcomes and support customer prioritization[7]. Subsequent studies have improved predictive capability through model ensembles, heterogeneous data processing, and algorithmic optimization. Feng et al. proposed a dynamic ensemble selection method that incorporated predictive accuracy and average marketing profit into a single optimization framework, achieving an AUC of 89.44%[8]. Tekouabou et al. proposed a classification framework for heterogeneous features and emphasized prospective customer identification before a campaign begins[15]. Guo et al. optimized a neural network using a metaheuristic algorithm and achieved an accuracy of 88.76%[16]. Xie et al. further combined ensemble learning with interpretable analysis to identify key determinants and achieved an accuracy of 94.80% [9]. Dong et al. combined resampling with cost-sensitive learning to improve minority-class identification and model transparency on imbalanced data, achieving an AUC of 90.80%[17]. Nasir et al. developed a two-stage evaluation framework that achieved an AUC of 91%[18]. Together, these studies have shifted bank telemarketing from experience-based screening toward customer-data-driven predictive decision-making.

Nevertheless, three interrelated problems remain in the literature. First, subscribers usually represent only a small minority in bank telemarketing data, resulting in severe class imbalance. Even when a model attains high overall accuracy, it may rely primarily on correctly classifying nonsubscribers and consequently miss genuinely valuable prospective subscribers[17]. Existing studies have focused on selecting among resampling methods[19], but have paid insufficient attention to the effects of different sampling ratios. Second, ensemble models can capture complex nonlinear relationships, but they also make it more difficult for managers to understand the basis of predictions. Third, most studies stop at model performance or feature rankings and do not fully explain how predictions can support actual banking decisions.

In response to these issues, this study develops a five-stage framework integrating class imbalance treatment, feature selection, model optimization, interpretability analysis, and CRM strategy. Using the UCI Bank Marketing dataset, the study develops decision tree, KNN, random forest, AdaBoost, XGBoost, LightGBM, and stacking models as seven candidate models. It then focuses on minority-class customer identification by systematically comparing no sampling, random oversampling, SMOTE, and SMOTETomek under different sampling ratios, and performs feature selection based on gain importance. Finally, SHAP is used to explain global feature contributions, dependence relationships among key features, and individual predictions for representative customers, after which feasible customer relationship management strategies are proposed. This study extends bank telemarketing research from predicting whether customers subscribe to explaining why they subscribe and supporting subsequent management through a concrete analytical process. Meanwhile, the optimal model achieves strong performance on the original, unsampled test set, with a recall of 88.78% and an AUC of 93.17%. The framework can help banks reduce the missed identification of prospective subscribers, improve marketing strategies, and support differentiated customer maintenance.

2. Research Design and Methods

2.1 Research Framework

This study develops a five-stage research framework around the question of how bank telemarketing predictions can be translated into customer relationship management strategies. The first stage covers data cleaning, feature encoding, and modeling preparation. The second stage compares candidate machine learning models and identifies the model to be optimized. The third stage sequentially compares class imbalance treatments and feature schemes. The fourth stage evaluates the final model and uses SHAP to explain overall model patterns and individual predictions. The fifth stage applies the interpretation results to customer identification, customer differentiation, interaction management, and customized marketing. Figure 1 illustrates the relationships among these stages.

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

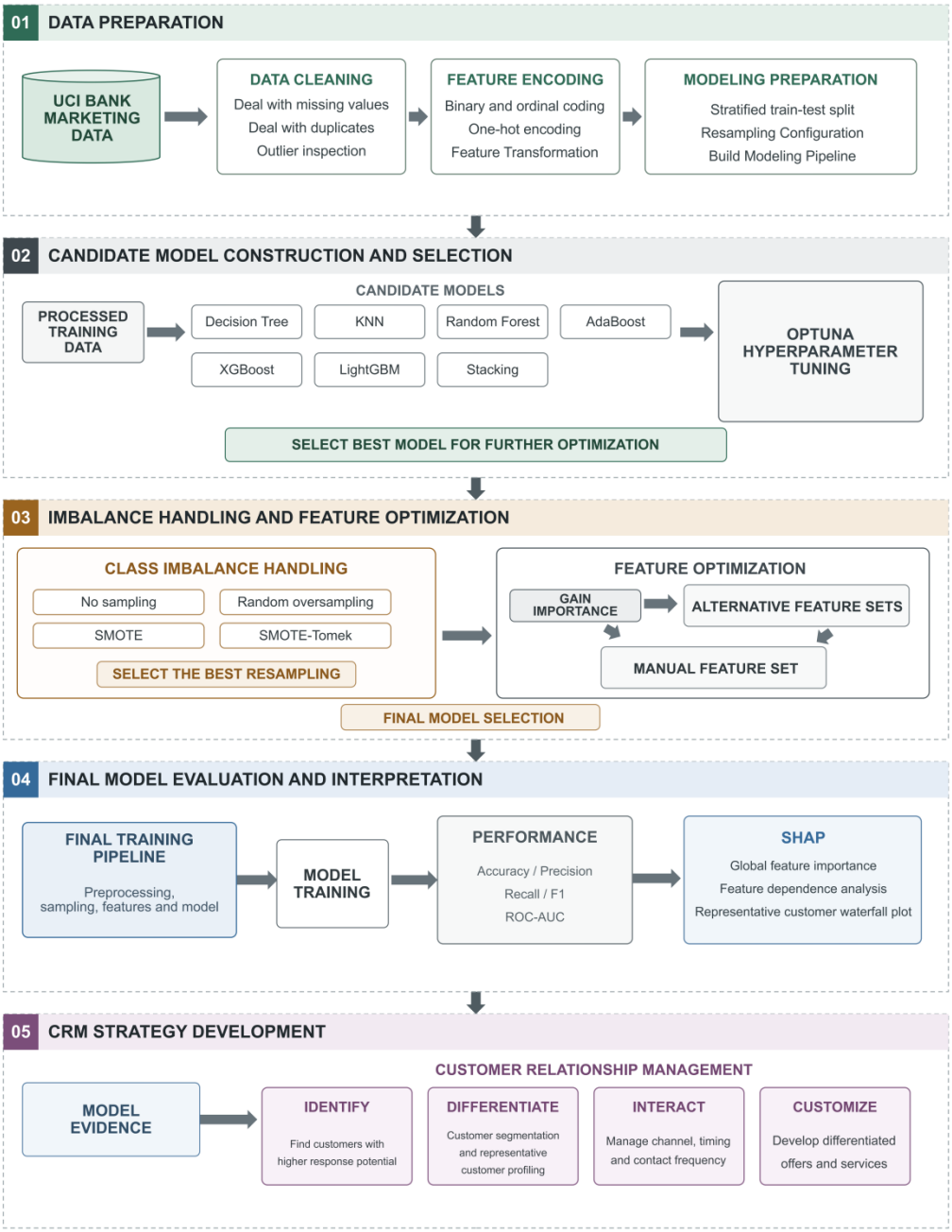

Figure 1. Research framework

2.2 Data Source and Feature Description

The data used in this study are drawn from the publicly available UCI Bank Marketing dataset[7]. The dataset records customer information and campaign outcomes from a Portuguese bank's telemarketing campaign. It contains 45,211 observations, 16 input variables, and one target variable, y. The target variable takes the value yes or no, indicating whether the customer subscribed to a term deposit.

To align the variable structure with the bank telemarketing process, this study reorganizes the 16 input variables into four groups based on their meanings and roles in the campaign: customer attributes, financial attributes, marketing interactions, and previous campaign information. This grouping provides the basis for the subsequent model interpretation and customer relationship management analysis. Table 1 presents the definitions and groupings of all variables. It should be noted that duration is available only after a marketing contact occurs and is therefore commonly removed[15]. In this study, however, duration is retained for call-process analysis, campaign review, and subsequent customer

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

management, rather than being used as screening information before the first call.

Table 1. Variable definitions and business meanings

Variable	Type	Group	Business meaning
age	Numeric	Customer attributes	Customer age
job	Nominal		Customer occupation
marital	Nominal		Customer marital status
education	Ordinal		Customer education level
default	Binary	Financial attributes	Whether the customer has a credit default record
balance	Numeric		Average annual account balance, in euros
housing	Binary		Whether the customer has a housing loan
loan	Binary		Whether the customer has a personal loan
contact	Nominal	Marketing interactions	Contact channel used in the current campaign
day	Integer		Day of the month on which the most recent contact occurred
month	Nominal		Month in which the most recent contact occurred
duration	Numeric		Duration of the most recent call (seconds)
campaign	Integer		Number of contacts with the customer during the current campaign
pdays	Ordinal	Previous campaigns	Days since the previous campaign contact; -1 indicates no previous contact
previous	Integer		Number of previous contacts before the current campaign
poutcome	Nominal		Outcome of the previous marketing campaign
y	Binary	Target variable	Whether the customer subscribed to a term deposit: yes = 1, no = 0

2.3 Data Preprocessing

First, duplicate values, missing values, and outliers were processed. No NA values, blank values, or fully duplicated records were found in the raw data. However, job, education, contact, and poutcome contained categories recorded as the string unknown. Because unknown indicates that the relevant information could not be confirmed and is not equivalent to an NA or blank value, it was retained as an explicit category in subsequent processing. The outlier check showed that the maximum value of previous was 275 and occurred only once, whereas the second-largest value was 58; the former was 4.74 times the latter. Given this degree of isolation, the observation with previous = 275 was removed, leaving 45,210 modeling samples. After removal, the data contained 5,289 positive samples and 39,921 negative samples, with a positive-class proportion of 11.70%.

Second, features were encoded and transformed. The binary variables default, housing, loan, and y were encoded as 1/0 for yes/no. Education was encoded from 1 to 4 in the order unknown, primary, secondary, and tertiary. One-hot encoding was applied to job, marital, contact, month, and poutcome, with the first reference category in each group removed. After processing, the dataset contained a total of 40 variables. The duration feature was markedly right-skewed and included zero values; therefore, a log1p transformation was applied to reduce the influence of extremely long calls. Large values of campaign and balance may reflect genuine customer differences, so these variables were not winsorized; instead, RobustScaler was applied within the pipeline to reduce the influence of extreme values on scale. A value of -1 for pdays has the business meaning of no previous contact and cannot be treated as an ordinary negative value. Pdays was divided into seven intervals: -1 for no previous contact, 0-7 for within one week, 8-30 for within one month, 31-90 for within three months, 91-180 for within six months, 181-365 for within one year, and above 365 for more than one year. These intervals were

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

then ordinally encoded from 0 to 6.

The cleaned data were divided using stratified random sampling at a 7:3 ratio with a random seed of 42. The training set contained 31,647 observations, and the test set contained 13,563 observations. Age, balance, day, duration, campaign, previous, education, and pdays were transformed with RobustScaler within the pipeline to prevent data leakage from inflating model evaluation results, while one-hot encoded features retained their 0/1 values.

2.4 Model Development and Staged Optimization

All 40 encoded features were used in the candidate models, and RandomOverSampler with a sampling ratio of 0.5 was consistently applied to oversample each training fold. Decision Tree, KNN, Random Forest, AdaBoost, XGBoost, and LightGBM were selected[20] to compare the performance of different machine learning models in bank telemarketing prediction. A stacking model was also constructed to examine the effect of model combination on subscriber identification.

For each candidate model, 50 hyperparameter trials were conducted using Optuna's TPE sampler. Each trial used five-fold StratifiedKFold cross-validation with shuffling, a random seed of 42, and mean recall as the optimization objective. For Decision Tree, the main search dimensions were max_depth ranged from 2-15, min_samples_split from 8-50, min_samples_leaf from 4-30, ccp_alpha from 0-0.05, and max_features from 0.2-0.6, with the splitting criterion selected between gini and entropy. For KNN, the search compared odd numbers of neighbors from 3-35, uniform and distance weighting, Euclidean and Manhattan distances, and leaf_size values from 20-50. For Random Forest, the search covered 50-300 trees, tree depths of 2-20, minimum split sizes of 2-50, minimum leaf sizes of 1-30, feature proportions of 0-0.3, maximum leaf-node counts of 10-100, and pruning coefficients of 0-0.1. For AdaBoost, the search covered base-estimator depths of 1-4, minimum split sizes of 2-20, minimum leaf sizes of 1-10, 50-400 base estimators, and learning rates of 0.01-0.4. For XGBoost, the search covered tree depths of 1-8, learning rates of 10^-7-0.3, 50-400 trees, sample and feature subsampling ratios of 0.5-1.0, and regularization parameters including gamma, reg_alpha, and reg_lambda. For LightGBM, the search covered 10-60 leaves, tree depths of 5-20, learning rates of 0.001-0.2, 100-600 trees, minimum leaf sizes of 1-80, and sample and feature subsampling ratios of 0.1-1.0. The stacking model used the tuned models as base learners and Logistic Regression with balanced class weights as the meta-learner.

After the model for subsequent analysis was identified, the class imbalance experiment compared No Sampling, RandomOverSampler, SMOTE[19], and SMOTETomek. For each of the latter three methods, sampling_strategy was set to 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0, yielding 18 sampling settings. For each setting, 50 Optuna trials were conducted anew within the training set, and the sampling rule was ultimately selected according to five-fold cross-validated recall.

After selecting the sampling rule, Gain Importance was extracted from the full-feature LightGBM model. Gain Importance represents the cumulative gain contributed by a feature in tree splits[20]. Several Top-N feature sets were constructed from the gain ranking, with the full feature set used as the benchmark. After reviewing the gain ranking and the performance of the existing feature schemes, the researchers manually defined an additional candidate feature set and included it as an exploratory scheme.

2.5 Model Evaluation, Interpretation, and Development of Customer Relationship Management Strategies

Customers who subscribed are defined as the positive class. TP denotes subscribers correctly identified as subscribers, TN denotes nonsubscribers correctly identified as nonsubscribers, FP denotes nonsubscribers incorrectly classified as subscribers, and FN denotes subscribers who were missed. Accuracy measures the overall proportion of correct classifications; precision represents the proportion of valid customers in the marketing list; recall measures the extent to which actual subscribers are identified; F1 balances precision and recall; and ROC-AUC measures overall discrimination across different thresholds.

Under class imbalance, high accuracy may result from conservative predictions for the majority class[19]; therefore, model performance cannot be assessed independently of the confusion matrix. Xie et al. noted that in bank telemarketing, the loss from missing prospective subscribers deserves greater priority than the loss from contacting uninterested customers, making recall more important than

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

reducing precision[9]. So this study uses cross-validated recall as the primary search metric to reduce missed identification of prospective subscribers. ROC-AUC is used to evaluate overall discriminatory ability. Precision and F1 are also reported to determine whether improvements in recall are accompanied by excessive ineffective contacts.

TreeExplainer is used to calculate SHAP values for the final model[21]. For global interpretation, mean absolute SHAP values and a beeswarm plot identify the primary bases of prediction; dependence plots describe the relationships between key feature values and model output; and waterfall plots explain individual predictions for high-score, low-score, and boundary customers.

Following the IDIC framework, this study translates model predictions and SHAP interpretations into four categories of CRM recommendations: customer identification, customer differentiation, customer interaction, and customized management[22,23]. In customer identification, model prediction scores determine follow-up priority. In customer differentiation, SHAP contributions from customer attributes, financial attributes, marketing interactions, and previous campaigns are combined to segment customer groups and construct representative customer profiles. In customer interaction, the contributions of marketing interaction features inform recommendations on subsequent contact channels, timing, and frequency. In customized management, combinations of positive and negative contributions in individual predictions support differentiated communication and service recommendations.

SHAP explains the basis on which the model forms predictions in the observed sample; it does not identify causal effects of variables on subscription behavior[21]. Accordingly, the CRM measures described above are managerial recommendations supported by model evidence. They should not be interpreted as showing that changing a feature will necessarily increase the probability of subscription, nor do they indicate that the proposed contact strategies have been validated through actual marketing interventions. Their effectiveness must be tested in subsequent marketing experiments or causal analyses.

3. Model Results and Performance Comparison

3.1 Performance Comparison of Baseline Models

The seven candidate model types were compared using the same train-test split, encoded features, and sampling rule (ROS 0.5). All models were optimized for five-fold cross-validated recall on the training set over 50 Optuna trials. Table 2 reports the test-set performance of the default and tuned models, distinguishing changes attributable to the algorithms themselves from those produced by hyperparameter optimization.

Table 2. Test-set performance of candidate models using all features

Model	Accuracy	Precision	Recall	F1	ROC-AUC
DecisionTree Default	0.8730	0.4559	0.4430	0.4493	0.6865
DecisionTree Tuned	0.8477	0.4189	0.7795	0.5449	0.8927
KNN Default	0.8437	0.3868	0.5740	0.4622	0.7944
KNN Tuned	0.8722	0.4650	0.6150	0.5296	0.8875
RandomForest Default	**0.9035**	**0.6073**	0.4959	0.5460	0.9278
RandomForest Tuned	0.8731	0.4741	0.7738	0.5880	0.9206
AdaBoost Default	0.8714	0.4655	0.6679	0.5487	0.9007
AdaBoost Tuned	0.8857	0.5080	0.7448	0.6040	0.9237
XGBoost Default	0.8938	0.5349	0.7051	0.6083	0.9250
XGBoost Tuned	0.8851	0.5057	0.7851	0.6152	0.9313
LightGBM Default	0.8843	0.5036	0.7965	**0.6170**	0.9317
LightGBM Tuned	0.8823	0.4981	0.8059	0.6156	**0.9323**
Stacking	0.8459	0.4219	**0.8557**	0.5651	0.9232

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

After Optuna tuning, test-set recall increased for all six individual models, although the magnitude of improvement varied. Recall increased from 0.4430 to 0.7795 for Decision Tree and from 0.4959 to 0.7738 for Random Forest, substantially improving their coverage of actual subscribers. KNN recall increased only from 0.5740 to 0.6150, representing a relatively small improvement. The tuned AdaBoost model achieved a recall of 0.7448, an F1 of 0.6040, and a ROC-AUC of 0.9237. XGBoost recall increased from 0.7051 to 0.7851; after tuning, its F1 reached 0.6152 and its ROC-AUC reached 0.9313, indicating strong performance among the boosting models. LightGBM recall increased slightly from 0.7965 to 0.8059, while its tuned ROC-AUC further increased to 0.9323. By comparison, stacking achieved the highest recall among the candidate schemes (0.8557), but its precision was only 0.4219 and its F1 was 0.5651.

Considering all metrics rather than recall alone, this study selects LightGBM over stacking for subsequent optimization. LightGBM achieved a recall of 0.8059, which was lower than that of stacking, but its precision and F1 reached 0.4981 and 0.6156, respectively, both substantially higher than those of stacking. Its ROC-AUC of 0.9323 was also the highest among the candidate schemes. LightGBM performed similarly to the tuned XGBoost model, but exceeded it by 0.0208 in recall, 0.0004 in F1, and 0.0010 in ROC-AUC, making it more consistent with the objective of prioritizing fewer missed prospective subscribers. In addition, LightGBM directly outputs Gain Importance and supports subsequent TreeSHAP analysis, thereby connecting model selection, feature optimization, and model interpretation. LightGBM is therefore used in subsequent comparisons of sampling ratios and feature schemes. The ROC curves of the tuned models are shown in Figure 2.

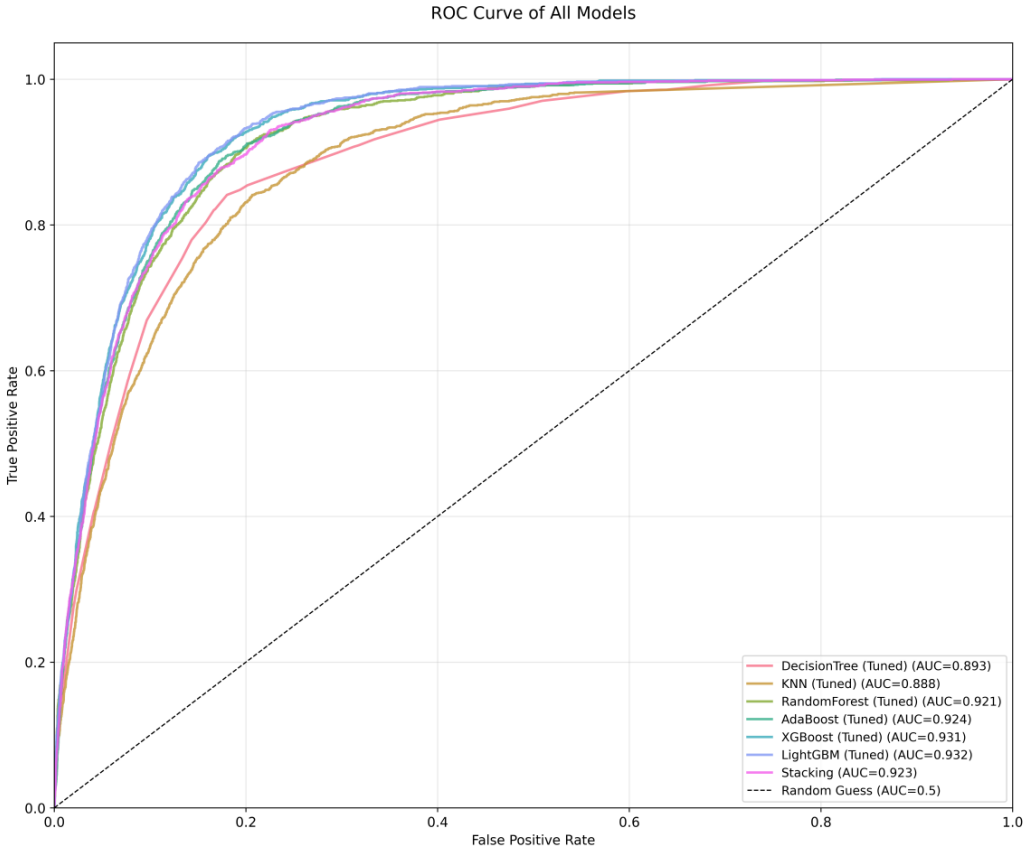

Figure 2. ROC curves of candidate models after tuning

3.2 Comparison of Sampling Methods and Sampling Ratios

The sampling experiment fixed all 40 features and the LightGBM search space, and compared No Sampling with RandomOverSampler, SMOTE, and SMOTETomek at sampling ratios from 0.5 to 1.0. Each sampling rule underwent 50 new Optuna trials. Table 3 reports the corresponding model results on the independent test set.

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

Table 3. Test-set performance of LightGBM under different sampling methods and ratios

Sampling	Ratio	Accuracy	Precision	Recall	F1	ROC-AUC
No Sampling	-	**0.9049**	**0.6204**	0.4820	0.5426	0.9291
ROS	0.5	0.8823	0.4981	0.8059	**0.6156**	0.9323
ROS	0.6	0.8755	0.4813	0.8248	0.6078	0.9321
ROS	0.7	0.8682	0.4654	0.8526	0.6021	**0.9327**
ROS	0.8	0.8593	0.4474	0.8601	0.5886	0.9314
ROS	0.9	0.8556	0.4406	0.8696	0.5849	0.9312
ROS	1.0	0.8504	0.4314	**0.8759**	0.5781	0.9324
SMOTE	0.5	0.8977	0.5508	0.6799	0.6086	0.9255
SMOTE	0.6	0.8898	0.5207	0.7303	0.6079	0.9235
SMOTE	0.7	0.8838	0.5023	0.7606	0.6050	0.9226
SMOTE	0.8	0.8455	0.4141	0.7719	0.5391	0.8996
SMOTE	0.9	0.8643	0.4548	0.8021	0.5805	0.9166
SMOTE	1.0	0.8290	0.3899	0.8166	0.5278	0.9035
SMOTETomek	0.5	0.8974	0.5498	0.6812	0.6085	0.9252
SMOTETomek	0.6	0.8856	0.5080	0.7196	0.5956	0.9196
SMOTETomek	0.7	0.8676	0.4599	0.7561	0.5720	0.9111
SMOTETomek	0.8	0.8390	0.4031	0.7813	0.5318	0.8955
SMOTETomek	0.9	0.8639	0.4536	0.7984	0.5785	0.9165
SMOTETomek	1.0	0.8550	0.4370	0.8286	0.5722	0.9122

The no-sampling model achieved an accuracy of 0.9049 and a precision of 0.6204, but its recall was only 0.4820, indicating that its high overall accuracy was driven primarily by the majority class. When the ROS ratio increased from 0.5 to 1.0, recall increased from 0.8059 to 0.8759, while precision decreased from 0.4981 to 0.4314; accuracy and F1 also decreased from 0.8823 and 0.6156 to 0.8504 and 0.5781, respectively. As the share of positive samples in the training data increased, the model identified more subscribers and FN consequently decreased. However, the model also became more likely to classify nonsubscribers as positive, increasing FP. Because nonsubscribers form the majority in the original test set, the additional FP reduced both accuracy and precision and further affected F1. ROC-AUC remained broadly stable, indicating that these changes arose mainly because the model became more inclined to predict the minority class rather than because its ranking ability declined substantially. By comparison, SMOTE and SMOTETomek produced lower recall and ROC-AUC than ROS at most equivalent ratios, suggesting that synthetic samples did not improve minority-class generalization in the current encoded feature space.

According to the prespecified selection criterion of five-fold cross-validated recall, ROS 1.0 achieved the highest CV recall, at 0.8930; its test-set recall was 0.8759 and its ROC-AUC was 0.9324, representing improvements of 0.3939 and 0.0033, respectively, over the no-sampling model. Xie et al. noted that in bank telemarketing, the loss from missing prospective subscribers exceeds the loss from contacting uninterested customers, so false negatives should receive priority[9]. This study therefore uses ROS 1.0, which achieved the highest five-fold cross-validated recall, in the subsequent experiments. Relative to the no-sampling model, this scheme reduced accuracy by 0.0545; precision and F1 also declined as the sampling ratio increased. If a bank places greater emphasis on accuracy, contact costs, or the risk of disturbing customers, it may use a lower sampling ratio or recalibrate the classification threshold during deployment.

3.3 Feature Selection Based on LightGBM Feature Importance

After determining the optimal sampling scheme, this study calculated Gain Importance using the

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

selected model and performed feature selection accordingly. The results showed that duration, poutcome_success, contact_unknown, day, housing, age, balance, and pdays, among other variables, contributed strongly to tree splits. It should be noted that Gain Importance reflects only the split gain attributable to a feature and does not explain the direction of its effect. Section 4 therefore uses SHAP to analyze the positive and negative contributions of each feature to model output.

Table 4. Cross-validation and test-set performance under different feature schemes

Features	Number	CV Recall	Accuracy	Precision	Recall	F1	ROC-AUC
Top 10	10	0.8552	0.8079	0.3638	0.8570	0.5108	0.8994
Top 15	15	0.8828	0.8242	0.3879	0.8696	0.5364	0.9158
Top 20	20	0.8971	0.8448	0.4222	0.8859	0.5719	0.9309
Top 25	25	0.8952	0.8502	**0.4318**	0.8872	**0.5809**	0.9318
Top 30	30	0.8949	0.8473	0.4264	0.8834	0.5752	0.9309
Top 35	35	0.8936	0.8453	0.4226	0.8790	0.5708	0.9304
All Features	40	0.8930	**0.8504**	0.4314	0.8759	0.5781	**0.9324**
Manual	22	**0.8982**	0.8476	0.4272	**0.8878**	0.5769	0.9317

In Table 4, the manually selected 22-feature scheme achieved the highest CV recall of 0.8982 and a test-set recall of 0.8878. The Top 20 scheme achieved a CV recall of 0.8971, a very small difference, but its test-set accuracy, precision, F1, and ROC-AUC were all slightly lower than those of the manual scheme. For Top 25, all test-set metrics except recall were slightly higher than those of the manual scheme, but the respective gains were only 0.0026, 0.0046, 0.0040, and 0.0001. By contrast, the manual scheme achieved higher CV recall and test-set recall while using three fewer features. Thus, neither scheme dominated the other across all metrics. Because this study places greater emphasis on coverage of prospective subscribers and seeks to reduce the complexity of model inputs and subsequent interpretation, the 22-feature scheme was ultimately selected.

The final 22 features were duration, poutcome_success, contact_unknown, day, age, housing, balance, month_mar, pdays, month_oct, month_jun, month_jul, month_aug, campaign, month_may, month_feb, month_nov, loan, month_jan, previous, month_sep, and education. This set retained information from all four categories: marketing interactions, previous campaigns, customer attributes, and financial attributes, while reducing the input dimension from 40 to 22. Compared with the full feature set, the manual scheme increased recall by 0.0119 and CV recall by 0.0051, while reducing ROC-AUC by only 0.0007, indicating that removing low-contribution features caused only a limited reduction in overall discriminatory ability.

The final model combines LightGBM with ROS 1.0 and the manually selected 22-feature set. On the test set of 13,563 observations, it correctly identified 1,409 of 1,587 subscribers and missed 178, yielding a recall of 88.78%. It also classified 1,889 nonsubscribers as subscribers, resulting in a precision of 42.72% and a ROC-AUC of 93.17%. The main benefit of the final scheme is therefore a reduction in the number of input features and fewer missed subscribers.

3.4 Comparison with Previous Studies

After selecting the final model, this study uses previous research based on the UCI Bank Marketing dataset as a reference. Feng et al.[8] and Xie et al.[9] reported high model performance, but their results are not included in Table 5. Feng et al. first performed outlier processing and undersampling on the complete dataset, balanced the sample to 2,000 observations, and then divided it into training, dynamic selection, and test sets. Xie et al. performed missing-value imputation, Min-Max normalization, and SMOTE augmentation on the complete dataset before conducting 10-fold cross-validation. Research on data leakage indicates that normalization, missing-value imputation, and resampling should be completed within each training fold[24,25]. If these steps are fitted using the complete dataset, information from the validation folds may enter model training, causing performance estimates to deviate from actual performance on new data[24,25]. Because of these differences, the high scores reported in those two studies should not be compared directly with the results obtained here on an independent, nonresampled test set that retains the original class distribution.

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

Table 5. Comparison of model performance between this study and previous research

Research	Model	Accuracy	Precision	Recall	F1	ROC-AUC
Nasir et al.(2026)[18]	XGBoost + B2S	0.8800	**0.5100**	0.5600	0.5300	0.9100
Dong et al.(2026)[17]	CatBoost + SMOTE	-	0.3980	0.8120	0.5400	0.9080
Guo et al.(2023)[16]	EFO-ANN	**0.8876**	-	-	-	0.7931
Our	LightGBM + ROS	0.8476	0.4272	**0.8878**	**0.5769**	**0.9317**

Note: A hyphen indicates that the corresponding metric was not reported in the original study.

Table 5 shows that the recall of the model proposed in this study is 0.8878, exceeding that of XGBoost+B2S reported by Nasir et al. (2026)[18] by 0.3278 and that of CatBoost+SMOTE reported by Dong et al. (2026)[17] by 0.0758. The ROC-AUC of this study is 0.9317, exceeding the two values by 0.0217 and 0.0237, respectively. The precision of 0.4272 obtained here is lower than the 0.5100 reported by Nasir et al. but higher than the 0.3980 reported by Dong et al. The F1 of 0.5769 exceeds the two comparable values by 0.0469 and 0.0369, respectively. Compared with the EFO-ANN of Guo et al. (2023)[16], this study improves ROC-AUC by 0.1386 but reduces accuracy by 0.0400. Its accuracy is also 0.0324 lower than that reported by Nasir et al. Thus, the advantages of this study are concentrated in recall and ROC-AUC, which are particularly important in bank telemarketing, while F1 also exceeds the two comparable results. Overall, the model covers more prospective subscribers and effectively distinguishes customers' subscription propensities.

4. Model Interpretation and Customer Relationship Management Strategy Analysis

4.1 Basis for Customer Identification from Global SHAP Interpretation

This study uses TreeSHAP to interpret the final LightGBM model by decomposing each prediction into feature-level contributions and then aggregating those contributions across samples to measure global feature importance[21].

The result shows that duration has a mean absolute SHAP value of 1.4406, substantially higher than those of the other features. Contact_unknown is 0.5453, month_may is 0.3050, poutcome_success is 0.2643, and housing is 0.2427. The top 10 also include month_jul, age, balance, day, and campaign. These features also rank highly by LightGBM Gain Importance. Gain Importance records the cumulative gain produced by a feature in tree splits, whereas SHAP measures overall influence from contributions to individual predictions. Both interpretation methods show that the final model relies primarily on marketing interaction information to identify customers, while financial attributes, customer attributes, and previous campaign outcomes provide supplementary information.

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

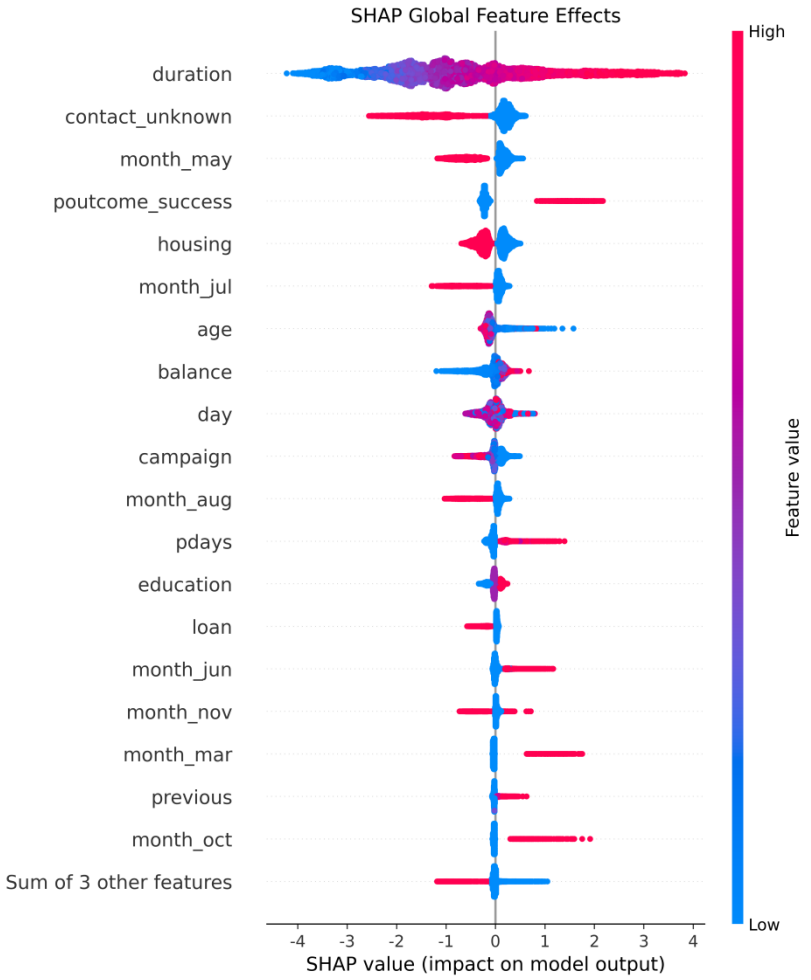

Figure 3. SHAP beeswarm plot for the final LightGBM model

Figure 3 visualizes the distributions of feature values and SHAP values. The vertical axis orders features from highest to lowest mean absolute SHAP value, and each point represents a customer in the test set. The horizontal axis indicates a feature's contribution to model output: a positive SHAP value increases the model-predicted probability of subscription, whereas a negative value decreases it; greater distance from zero indicates a stronger effect. Color indicates the magnitude of the feature value. For binary variables, red and blue correspond to 1 and 0, respectively.

Duration has the widest horizontal spread and is the most influential feature in the final model. Higher values are concentrated mainly in the positive SHAP region, whereas lower values cluster in the negative region, indicating that customers with longer calls are more likely to subscribe and those with shorter calls are more likely not to subscribe. When contact_unknown, month_may, housing, and month_jul equal 1, most points fall in the negative SHAP region, indicating that an unknown contact channel, contact in May or July, and the presence of a housing loan reduce the model-predicted probability of subscription. By contrast, when poutcome_success equals 1, points concentrate in the positive SHAP region, indicating that success in a previous campaign is a relatively stable basis for predicting subscription.

Low balance values generally make negative contributions, whereas higher values occur more often in the positive SHAP region, although their influence is substantially smaller than that of duration. Points with high campaign values are concentrated mainly on the left, indicating that more contacts are generally associated with a lower predicted subscription probability. Loan also contributes negatively in most cases when it equals 1. Higher values of pdays and previous occur mainly in the positive SHAP region. Among the month variables, month_aug and month_nov generally contribute negatively, whereas month_jun, month_mar, and month_oct contribute positively more often. Different values of age and day appear on both sides of zero.

Figure 3 shows that the final model forms predictions primarily from call duration, contact channel,

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

contact month, and previous campaign outcome, with account balance, loan status, and customer attributes providing supplementary information.

Table 6. SHAP contributions of different business feature groups

Business feature group	Mean |SHAP|	Normalized contribution
Marketing Interaction	3.1337	74.31%
Financial Attributes	0.4431	10.51%
Previous Campaign History	0.4074	9.66%
Customer Attributes	0.2331	5.53%

Table 6 groups the final 22 features into customer attributes, financial attributes, marketing interactions, and previous campaigns, and normalizes the sums of mean absolute SHAP values within each group. Marketing interaction features contribute 74.31%, financial attributes 10.51%, previous campaigns 9.66%, and customer attributes 5.53%. The large share of marketing interactions is driven mainly by call-process variables such as duration. However, duration becomes available only after a call has occurred; the model is therefore suitable for call-process analysis, post-call review, and follow-up contacts, but not for customer screening before the first call.

4.2 Customer Segmentation Based on Model Results

Given the importance of both customers' subscription probabilities and customer value, this study segments customers using their model-predicted probability of subscription and balance, as shown in Table 7. Let the model-predicted subscription probability be s. A customer is classified as high-score when s >= 0.5 and low-score when s < 0.5. After balance is divided into equal-frequency groups, groups with balances below EUR 300 generally show negative SHAP contributions. Beginning at EUR 300, the median SHAP value generally becomes positive. This study therefore uses EUR 300 as an empirical boundary between high- and low-balance customers. Balance >= 300 denotes high balance, while balance < 300 denotes low balance. It should be emphasized that balance reflects only a customer's average annual account balance; it cannot be treated as the amount subscribed to a deposit or as equivalent to complete customer lifetime value.

Table 7. Criteria for four customer segments

Customer segment	Segmentation criterion	Segment characteristics
Priority conversion customers	s≥0.5, balance≥300	High subscription probability and high customer value
Potential cultivation customers	s≥0.5, balance<300	High subscription probability and low customer value
Relationship maintenance customers	s<0.5, balance≥300	Low subscription probability and high customer value
Low-cost maintenance customers	s<0.5, balance<300	Low subscription probability and low customer value

4.3 Customer Contact Patterns Based on SHAP Dependence Relationships

Among the 22 features retained in the final model, 14 are marketing interaction features, including duration, contact_unknown, day, campaign, and selected encoded month features. Figure 4 presents SHAP dependence relationships for the 10 most important features. Duration, contact_unknown, month_may, month_jul, day, and campaign are marketing interaction features.

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

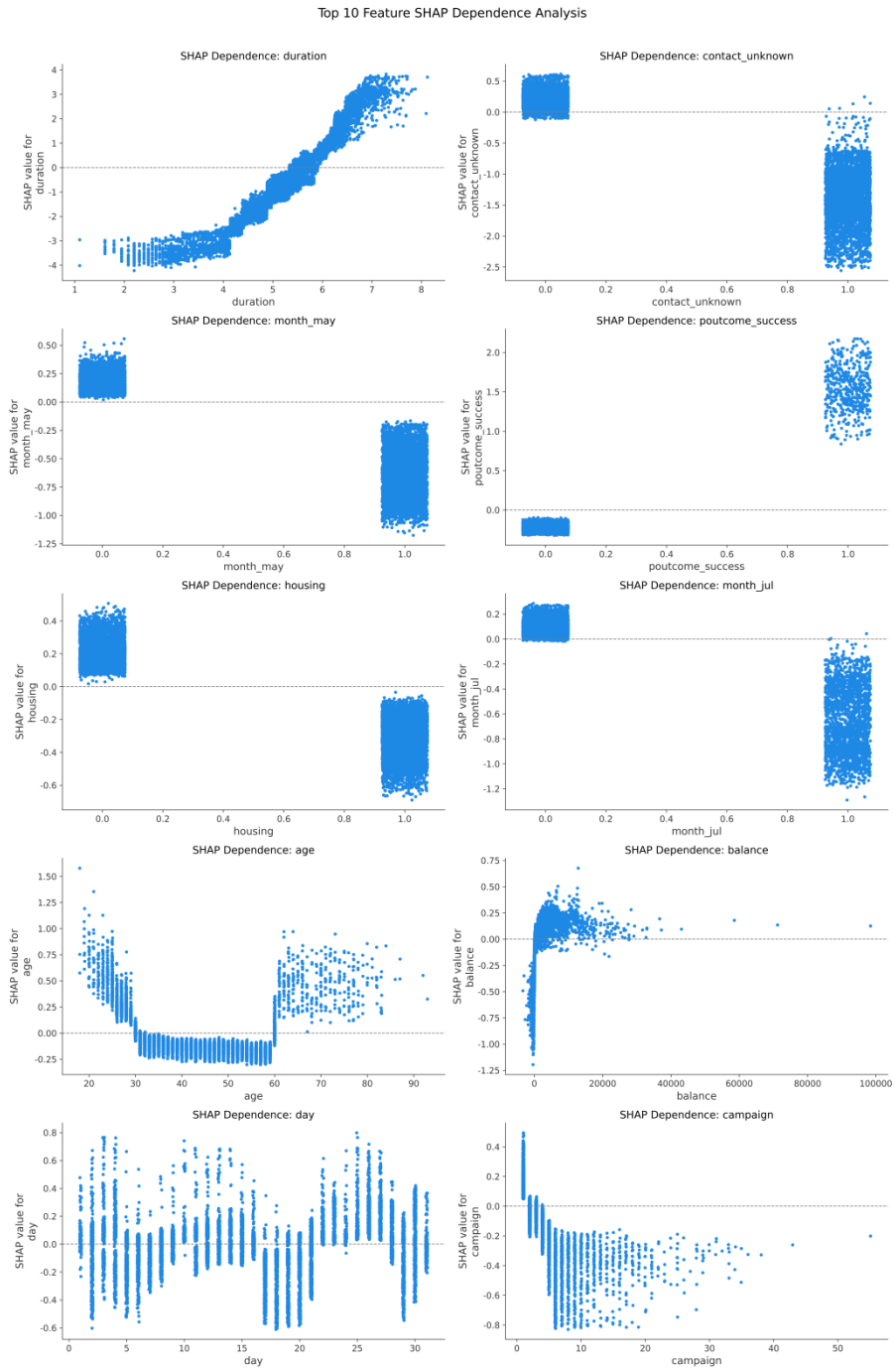

Figure 4. SHAP dependence relationships for key features

As duration increases, its SHAP value generally rises. Shorter calls mostly make negative contributions, while longer calls make positive contributions, indicating that longer call durations generally correspond to higher model-predicted subscription probabilities. When contact_unknown equals 1, its contribution is mainly negative; when it equals 0, its contribution is closer to zero or positive, indicating that an unknown contact channel reduces the model-predicted subscription probability.

When month_may equals 1, SHAP values cluster in the negative region, indicating that contact in May generally corresponds to a lower model-predicted subscription probability. Poutcome_success makes a strong positive contribution when it equals 1 and generally makes a slightly negative contribution when it equals 0, indicating that success in a previous campaign substantially increases the model-predicted subscription probability. Housing contributes mainly negatively when it equals 1 and positively when it equals 0, indicating that customers with housing loans generally have lower model-predicted subscription probabilities. When month_jul equals 1, SHAP values are mostly

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

negative; when it equals 0, values are closer to zero or positive, indicating that contact in July generally corresponds to a lower model-predicted subscription probability.

Age exhibits a nonlinear relationship: samples aged approximately 30-60 generally make slightly negative contributions, whereas younger and older samples make relatively larger contributions. The median SHAP value of balance changes from negative to positive at approximately EUR 300, after which the increase gradually slows. Thus, higher balances generally increase the model-predicted subscription probability, but their marginal influence diminishes. Day has highly variable SHAP values, with days 10-15 and 22-27 generally making positive contributions. Campaign generally makes a positive contribution for the first contact, but becomes negative for most samples as the number of contacts increases, indicating that repeated contact generally corresponds to a lower predicted subscription probability.

Therefore, when contacting an individual customer, the bank should first verify valid contact information and prioritize a channel familiar to the customer, avoiding any reduction in subscription likelihood caused by unknown channel information. Contact can be scheduled outside May and July, which make relatively strong negative contributions in the model; dates between 10-15 and 22-27 are preferable. Because the first contact makes a relatively strong positive contribution, the quality of the initial communication should receive priority. If a customer does not respond after several consecutive attempts, repeated calls should be reduced and a cooling-off period introduced. During a call, provided the customer is willing to engage, the conversation should be extended where appropriate. More substantive communication, a clearly identified contact channel, and fewer repeated contacts are all associated with higher model-predicted subscription probabilities.

4.4 Individual Explanations and Representative Customer Profiles

Global importance indicates only the features on which the model generally relies; customer profiling also requires examining positive and negative contributions for individual samples. Figure 5 shows the three samples which were selected from the test set: the one with the highest subscription probability, the one with the lowest subscription probability, and the one closest to the 0.5 classification boundary. SHAP waterfall plots were used to show how the model formed each predicted subscription probability. Table 8 presents an overview of these representative customers.

Table 8. Overview of representative customer samples

Sample type	Sample index	Actual label	Predicted label	Predicted probability
High-score customer	44,042	1	1	0.977156
Low-score customer	18,738	0	0	0.004690
Boundary customer	29,757	0	1	0.500310

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

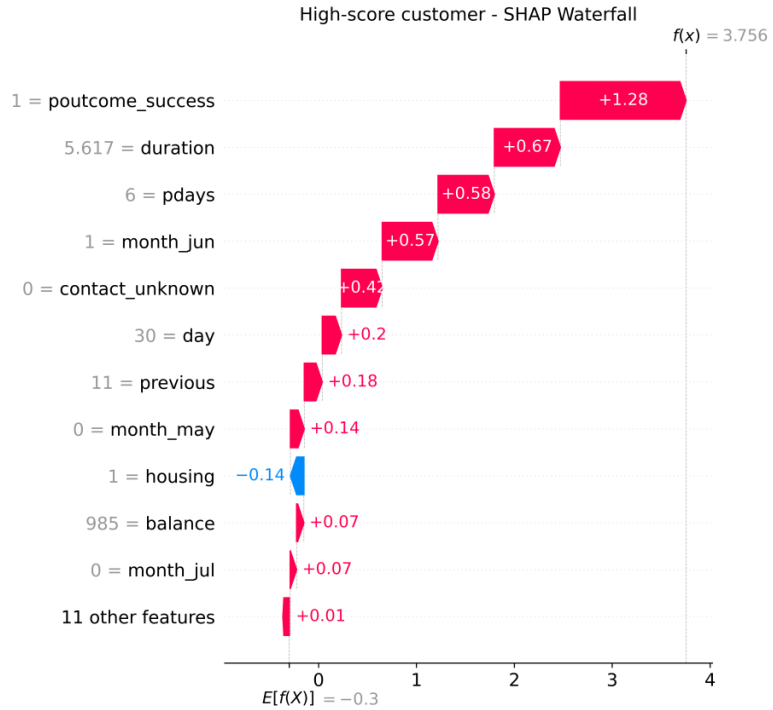

Figure 5(a). Local SHAP explanation for the high-score customer

The model output for the high-score customer is f(x) = 3.756. A successful previous campaign, poutcome_success = 1, contributes +1.28 and is the strongest positive factor. Duration = 5.617 corresponds to an original call duration of approximately 274 seconds. Pdays = 6 indicates that more than 365 days have elapsed since the previous marketing contact. Duration, pdays, month_jun, and a known contact channel also make substantial positive contributions to model output. Housing = 1 makes a small negative contribution. The customer's model score is 0.977156 and balance is 985; according to Table 8, the customer is assigned to the priority conversion segment.

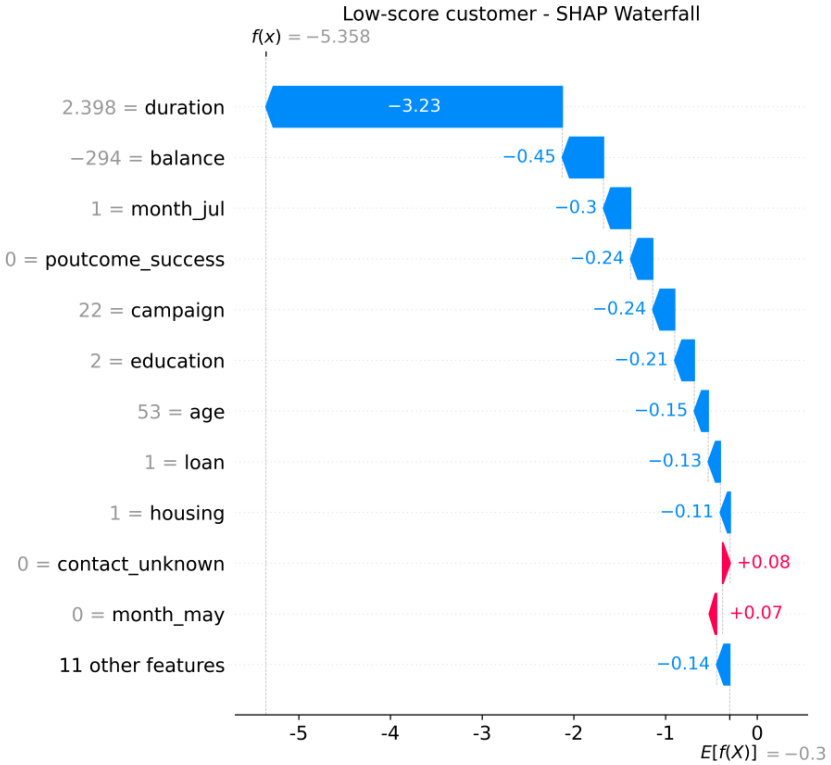

Figure 5(b). Local SHAP explanation for the low-score customer

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

The model output for the low-score customer is f(x) = -5.358. Duration = 2.398 corresponds to an original call duration of approximately 10 seconds, and its SHAP contribution of -3.23 is the principal negative factor. Average annual account balance, contact month, number of contacts, personal loan status, and housing loan status also reduce model output. The customer's model score is 0.004690; according to Table 8, the customer is assigned to the low-cost maintenance segment.

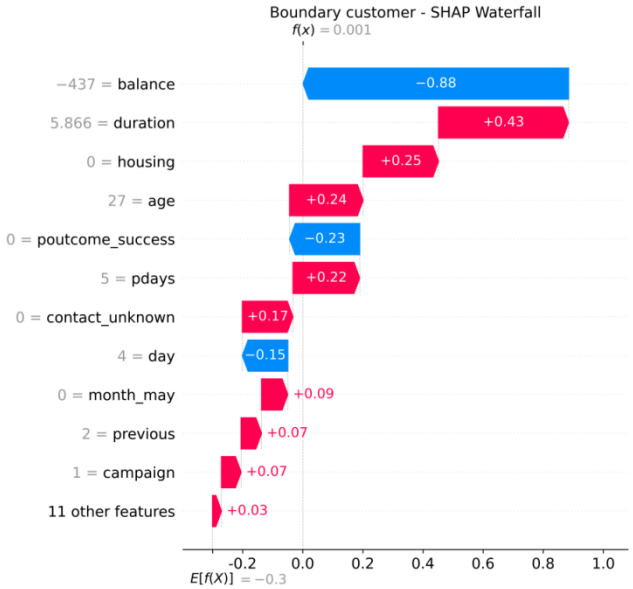

Figure 5(c). Local SHAP explanation for the boundary customer

The model output for the boundary customer is f(x) = 0.001, and the score shown in Table 8 is 0.500310, only 0.000310 above the 0.5 threshold. Balance is -437, corresponding to a SHAP contribution of -0.88. Duration = 5.866 corresponds to an original call duration of approximately 352 seconds; having no housing loan and being 27 years old make positive contributions. Poutcome_success = 0 indicates no record of success in a previous campaign and makes a negative contribution. Under the fixed criteria, this customer belongs to the potential cultivation segment.

4.5 Customized Strategies for Customer Segments

The core of customization is to manage each customer segment according to its specific managerial objective[22]. Both priority conversion and potential cultivation customers have high subscription propensities; the former require an emphasis on long-term relationship maintenance, while the latter require an emphasis on increasing customer value. Relationship maintenance customers have a relatively strong financial base, so management should stabilize the existing relationship and increase their propensity to subscribe. Low-cost maintenance customers should retain basic contact through low-frequency, low-cost methods. Table 9 presents the specific strategies.

Table 9. Customized strategies for four customer segments

Customer segment	Management objective	Customized strategy
Priority conversion customers	Facilitate subscription and strengthen long-term relationship maintenance	Assign a consistent relationship manager to understand the customer's available funds, maturity preferences, and liquidity needs; match products with different maturities; and follow up through channels to which the customer has responded well. After subscription, provide maturity reminders and discuss funding arrangements to sustain the customer relationship.
Potential cultivation	Encourage term deposit subscription and gradually	Match an appropriate minimum deposit and product maturity to the customer's available

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

customers	increase customer value	funds. During subsequent contacts, consider changes in balance and funding arrangements and offer options to increase the subscription amount or adjust the amount and maturity, thereby gradually increasing the subscribed amount and customer value.
Relationship maintenance customers	Stabilize existing customer relationships and increase customers' propensity to subscribe	Verify valid contact information, adjust contact intervals according to customer responses, and reduce repeated contacts. Use communication records to understand product needs and subscription concerns, and adjust product presentations and follow-up timing to gradually increase subscription propensity.
Low-cost maintenance customers	Control maintenance costs and retain basic contact	Provide concise product information infrequently through low-cost channels such as text messaging, thereby reducing reliance on manual calls.

Table 9 translates the customer segmentation results into specific customer management plans. In practice, relationship managers can determine follow-up priorities and objectives according to each customer's segment. When customer information changes, segment membership and management strategies can be updated accordingly, thereby improving the precision of bank telemarketing, encouraging term deposit subscriptions, and sustaining long-term customer relationships.

5. Conclusion

This study develops a five-stage bank term deposit telemarketing strategy framework that integrates class imbalance treatment, feature selection, model optimization, interpretability analysis, and CRM strategy. It focuses on three problems: prospective customers are easily missed under class imbalance; the effects of different sampling ratios have not been systematically compared; and model results are difficult to translate into managerial actions.

Using the five-stage framework, this study reaches the following principal conclusions. First, Optuna hyperparameter optimization substantially improves recall across the individual models. The gains for Decision Tree and Random Forest are close to 30%, while the other models also improve markedly, confirming the effectiveness of automated hyperparameter optimization. Second, class imbalance treatment is the main stage for increasing the identification rate of subscribers. The staged experiment shows that the largest increase in recall comes from adjusting the sampling ratio: recall is only 48.2% without sampling but reaches 88.78% in the optimal model. Sampling method and sampling ratio therefore require particular attention when the data are imbalanced. Feature selection can also improve or stabilize model performance while reducing the number of features. Compared with the full-feature model, the final model removes 18 input features, increases recall by 0.0119, and reduces ROC-AUC by only 0.0007. This result indicates that some low-contribution features can be removed while the model retains strong predictive and discriminatory ability with fewer inputs. Finally, SHAP analysis shows that marketing interaction features account for 74.31% of the total contribution, with call duration, contact channel, previous campaign outcome, and month exerting the strongest effects. Combining subscription probability with customer balance divides customers into four segments: priority conversion, potential cultivation, relationship maintenance, and low-cost maintenance, for which corresponding management strategies are proposed. Individual SHAP explanations also reveal how predictions are formed for different customers, providing a basis for banks to prioritize customers and formulate differentiated maintenance plans, and thereby completing the transition from subscription prediction to customer relationship management strategy.

This study constructs a complete chain connecting class imbalance treatment, feature selection, model optimization, SHAP interpretation, and CRM strategy, extending bank telemarketing research from subscription prediction to practical management strategy. By systematically comparing sampling methods and ratios, the study addresses the limited attention paid to sampling ratios in previous research. Compared with studies using the same dataset, the ROS-LightGBM model achieves higher recall and ROC-AUC on the original test set, reaching 88.78% and 93.17%, respectively, and demonstrates strong coverage of prospective customers and overall discriminatory ability. It can

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

effectively help banks determine customer priorities, contact frequencies, and maintenance methods, thereby supporting differentiated contact and the allocation of marketing resources.

The data used in this study come from historical marketing activities at a single bank, and the model's applicability across regions and institutions has not yet been validated. Customer segmentation relies on account balance and assumes that a high balance indicates a high deposit amount, but actual subscription amounts are unavailable. Future research can incorporate data from additional marketing campaigns to examine model stability and scope of applicability, while further optimizing customer segmentation and the allocation of marketing resources.

References

[1] Gatev E, Schuermann T, Strahan P E. Managing bank liquidity risk: How deposit-loan synergies vary with market conditions[J]. The Review of Financial Studies, 2009, 22(3): 995-1020.
[2] Cornett M M, McNutt J J, Strahan P E, et al. Liquidity risk management and credit supply in the financial crisis[J]. Journal of Financial Economics, 2011, 101(2): 297-312.
[3] Ivashina V, Scharfstein D. Bank Lending During the Financial Crisis of 2008[J]. Journal of Financial Economics, 2010, 97(3): 319-338.
[4] Egan M, Hortaçsu A, Matvos G. Deposit competition and financial fragility: Evidence from the US banking sector[J]. American Economic Review, 2017, 107(1): 169-216.
[5] Clemes M D, Gan C, Zhang D. Customer switching behaviour in the Chinese retail banking industry[J]. International Journal of Bank Marketing, 2010, 28(7): 519-546.
[6] Liao S H, Chen Y J, Hsieh H H. Mining Customer Knowledge for Direct Selling and Marketing[J]. Expert Systems with Applications, 2011, 38(5): 6059-6069.
[7] Moro S, Cortez P, Rita P. A Data-Driven Approach to Predict the Success of Bank Telemarketing[J]. Decision Support Systems, 2014, 62: 22-31.
[8] Feng Y, Yin Y, Wang D, et al. A Dynamic Ensemble Selection Method for Bank Telemarketing Sales Prediction[J]. Journal of Business Research, 2022, 139: 368-382.
[9] Xie C, Zhang J L, Zhu Y, et al. How to Improve the Success of Bank Telemarketing? Prediction and Interpretability Analysis Based on Machine Learning[J]. Computers & Industrial Engineering, 2023, 175: 108874.
[10] Moro S, Cortez P, Rita P. Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns[J]. Neural Computing and Applications, 2015, 26(1): 131-139.
[11] Li S, Sun B, Montgomery A L. Cross-selling the right product to the right customer at the right time[J]. Journal of Marketing Research, 2011, 48(4): 683-700.
[12] Godfrey A, Seiders K, Voss G B. Enough is enough! The fine line in executing multichannel relational communication[J]. Journal of Marketing, 2011, 75(4): 94-109.
[13] Ngai E W T, Xiu L, Chau D C K. Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification[J]. Expert Systems with Applications, 2009, 36(2): 2592-2602.
[14] Payne A, Frow P. A Strategic Framework for Customer Relationship Management[J]. Journal of Marketing, 2005, 69(4): 167-176.
[15] Koumétio Tékouabou S C, Gherghina S C, Toulni H, et al. A Machine Learning Framework towards Bank Telemarketing Prediction[J]. Journal of Risk and Financial Management, 2022, 15(6): 269.
[16] Guo W, Yao Y, Liu L, et al. A Novel Ensemble Approach for Estimating the Competency of Bank Telemarketing[J]. Scientific Reports, 2023, 13: 20819.
[17] Dong B, Zhang X, Liu Y, et al. Risk-Sensitive Machine Learning for Financial Decision Modeling Under Imbalanced Data: Evidence from Bank Telemarketing[J]. Entropy, 2026, 28(3): 354.
[18] Nasir F, Ahmed A A, Yevseyeva I, et al. Marketing Analytics in Banking 4.0: A Two-Stage Explainable AI Framework for High-Accuracy and Well-Calibrated Predictions[J]. PLOS One, 2026, 21(5): e0348767.
[19] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-Sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.
[20] Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree[C]. Advances in Neural Information Processing Systems, 2017, 30: 3146-3154.
[21] Lundberg S M, Erion G, Chen H, et al. From Local Explanations to Global Understanding with Explainable AI for Trees[J]. Nature Machine Intelligence, 2020, 2: 56-67.
[22] Peppers D, Rogers M. Managing Customer Relationships: A Strategic Framework[M]. 3rd

ISSN 2616-5902 Vol. 8, Issue 7: 99-117, DOI: 10.25236/AJBM.2026.080714

ed. Hoboken, NJ: John Wiley & Sons, 2016.
[23] Kumar S, Kumar A B, Ravi V. Explainable Artificial Intelligence for Analytical Customer Relationship Management in Banking and Finance[C]. In: Intelligent Systems Design and Applications. Springer, 2024: 101-112.
[24] Apicella A, Isgrò F, Prevete R. Don't Push the Button! Exploring Data Leakage Risks in Machine Learning and Transfer Learning[J]. Artificial Intelligence Review, 2025, 58: 339.
[25] Sasse L, Nicolaisen-Sobesky E, Dukart J, et al. Overview of Leakage Scenarios in Supervised Machine Learning[J]. Journal of Big Data, 2025, 12: 135.