Proceedings of ICEMGD 2025 Symposium: Innovating in Management and Economic Development
DOI: 10.54254/2754-1169/2025.LH24150

Bank Marketing Prediction Based on XGBoost

Yuqi Wang

Hunan First Normal University, Changsha, China
yukiyukiwang2@gmail.com

Abstract: Under the dual challenges of fintech evolution and digital transformation, commercial banks face increasing limitations in traditional marketing prediction methods, which struggle with static customer profiling, low data utilization, and poor adaptability to real-time demands. This study addresses these gaps by proposing an XGBoost-based predictive framework to enhance precision marketing and risk-adjusted returns in banking scenarios. We integrate multidimensional features, including static attributes (e.g., age, occupation) and dynamic behavioral indicators (e.g., consumer confidence index, Euribor rates), to overcome the unidimensional profiling limitations of conventional approaches. Methodologically, XGBoost demonstrates superior performance through three innovations: firstly efficient handling of high-dimensional sparse data via parallel computing, reducing marginal processing costs while improving prediction accuracy (89% accuracy, 90% AUC). Secondly, mitigation of information asymmetry by synthesizing transactional, social, and macroeconomic features (e.g., employment variation rate, housing loans). Comparative analyses against five benchmark models (GBDT, Random Forest, Decision Tree, Logistic Regression, Bagging) confirm XGBoost's dominance in AUC and F1-score, validating its capacity to resolve nonlinear interactions and temporal sensitivity in marketing campaigns. The model's scalability enables cost-effective targeting. This research contributes to both algorithmic optimization in financial marketing and operational decision-making frameworks, though limitations persist in handling extreme class imbalances. Future work will explore hybrid architectures combining XGBoost with deep learning for cross-channel behavioral modeling.

Keywords: XGBoost, Precision Marketing, Banking Analytics, Risk-Reward Optimization, Fintech Applications

1. Introduction

In the context of the financial technology revolution and digital transformation, commercial banks face dual challenges arising from evolving customer demands and intensifying market competition. Traditional marketing prediction methods in banking, such as logistic regression and decision trees, have played a significant role in customer segmentation and product recommendation. However, these methods suffer from several limitations, including uneven marketing strategies, inflexible service workflows, and the inefficient distribution of traditional bank branches, which hinder their effectiveness [1]. As a result, existing solutions are increasingly insufficient to meet the diverse demands encountered in real-world scenarios.

DOI: 10.54254/2754-1169/2025.LH24150

In recent years, the application of machine learning techniques in the financial sector has gained momentum. Initial research focused on areas like customer credit scoring and product cross-selling, successfully demonstrating the power of data-driven decision-making [2]. However, traditional bank marketing prediction models continue to exhibit several shortcomings. First, customer profiling in these models is overly simplistic, relying heavily on static characteristics such as age and income while neglecting dynamic behavioral patterns. Second, these models lack real-time adaptability, making it difficult to respond to time-sensitive demands during marketing campaigns. Third, the full potential of available data remains untapped, particularly in exploiting the relationships between internal transaction records and external data, such as social media behaviors.

This study proposes an innovative, multidimensional approach to precise marketing prediction, leveraging banks' extensive client data (such as transaction records, credit scores, and behavioral data), and the XGBoost algorithm. XGBoost is well-suited for managing high-dimensional, sparse data, as it efficiently reduces processing costs through parallel computing. As the data scale increases [3], the model's predictive accuracy improves, demonstrating increasing marginal returns while simultaneously decreasing marginal computational costs due to algorithmic optimizations. This leads to significant economies of scale.

The XGBoost framework exhibits robust discriminatory power in credit card marketing by accurately identifying high-demand, low-risk customers from large client pools. This ability helps prevent the inefficient allocation of resources to suboptimal segments, such as low-propensity or high-default-risk groups, optimizing campaign cost-effectiveness and reducing per-customer marketing expenses. Additionally, due to information asymmetry in financial markets (such as uncertainty regarding clients' true repayment capabilities) issues like adverse selection and moral hazard arise. By integrating multi-dimensional data (e.g., consumption behavior, social media activity, and device usage), XGBoost constructs a more comprehensive customer profile, reducing the information gap between banks and their clients.

Ultimately, banks must balance risk (e.g., customer defaults) with returns (e.g., loan interest). XGBoost's regularization techniques and feature importance ranking capabilities allow for the precise identification of high-risk clients, such as those with stable incomes but temporary cash flow issues. By targeting this "gray zone" population, the model maximizes risk-adjusted returns, strategically addressing clients who offer a balance between short-term risk and long-term profitability (see Figure 1).

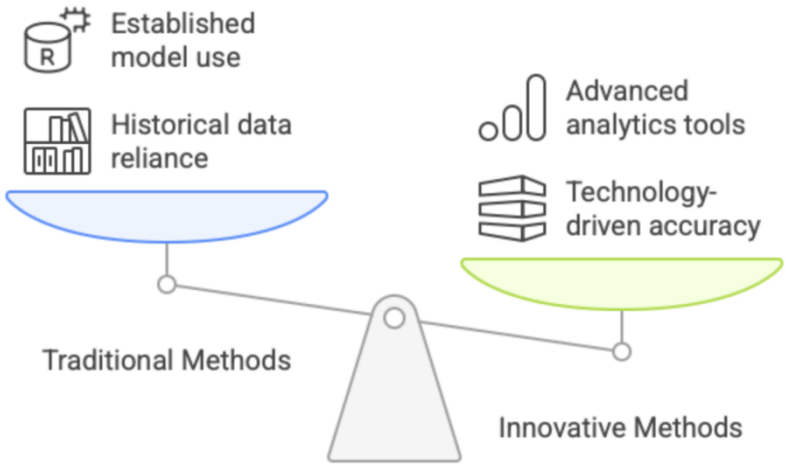

Figure 1: Comparison of traditional versus innovative banking forecasting methods

DOI: 10.54254/2754-1169/2025.LH24150

2. Research objective and significance

This study aims to address the critical limitations of traditional bank marketing prediction methods, including static customer profiling, poor adaptability to nonlinear relationships, and low data utilization efficiency, by developing an XGBoost-driven framework that integrates multidimensional features (demographic, behavioral, and macroeconomic variables). Pecifically, we seek to validate XGBoost's superiority over conventional models (e.g., logistic regression, decision trees) in accuracy, scalability, and real-time decision-making through comparative experiments. What's more, uncover the interplay between macroeconomic factors (e.g., Euribor3m, employment variation rate) and individual subscription behaviors using interpretable machine learning techniques. And provide actionable insights for optimizing marketing resource allocation, reducing costs, and mitigating risks in dynamic financial environments. By bridging theoretical gaps in feature engineering and practical challenges in precision marketing, this work advances data-driven strategies for banking digital transformation.

By integrating the XGBoost algorithm with K-Means clustering [4]. This study has achieved precise customer segmentation, effectively enhancing the predictive accuracy (exceeding 87%), significantly outperforming single-model approaches. (e.g., The F1 score for logistic regression in the Chinese context is 0.49). Concurrently, drawing upon the proposed SMOTE resampling technique mitigates the imbalance in the proportion of positive and negative samples within bank marketing data, thereby enhancing the model's ability to identify minority classes, such as high-value customers or churned clients [5].

In alignment with the strategic framework, this study employs the XGBoost model to achieve comprehensive predictions across all stages of customer acquisition, value enhancement, and retention (as exemplified by the categorization of low, medium, and high churn-risk customers) [6]. This supports the bank in formulating personalized marketing strategies. These aspects represent areas previously unaddressed in prior research, and the present study significantly extends the boundaries of the field in bank marketing prediction.

3. Methodology

3.1. Dataset description

This data set contains records relevant to a direct marketing campaign of a Portuguese banking institution. The marketing campaign was executed through phone calls. Often, more than one call needs to be made to a single client before they either decline or agree to a term deposit subscription. The classification goal is to predict if the client will subscribe (yes/no) to the term deposit (variable y), see Table 1.

Table 1: "Bank marketing data set"

Abbreviation of Variable	Type	Value/Example
1、age	Numerical (Continuous)	Scope:18-95
2、job	Classification (Nominal)	"admin.","blue-collar","student"…
3、Marital	Classification (Nominal)	"married", "single", "divorced"
4、Education	Categorical (Ordered)	"primary", "secondary", "tertiary"
5、Default	Binary Classification	"yes", "no"

DOI: 10.54254/2754-1169/2025.LH24150

Table 1: (continued)

6、Housing	Binary Classification	"yes", "no"
7、Loan	Binary Classification	"yes", "no"
8、Contact	Classification (Nominal)	"cellular", "telephone"
9、Month	Categorical (Ordered)	"jan", "feb", ..., "dec"
10、Day_of_week	Classification (Nominal)	"mon", "tue", ..., "fri"
11、Duration	Numerical (Continuous)	0-4918
12、Campaign	Numerical (Discrete)	1-56
13、Pdays	Numerical (Hybrid)	-1, 1-999
14、Previous	Numerical (Discrete)	0-7
15、Poutcome	Classification (Nominal)	“success","failure","unknown"
16、Emp.var.rate	Numerical (Continuous)	-3.4~1.4
17、Cons.price.idx	Numerical (Continuous)	92.0~94.0
18、Cons.conf.idx	Numerical (Continuous)	-50~-40
19、Euribor3m	Numerical (Continuous)	0.7~5.0
20、Y	Binary Classification	"yes", "no"
21、nr.employed	Numerical (Continuous)	4960~5228
22、Client_id	Unique Identifier	No practical significance.

3.2. Data cleansing

We have employed data cleansing techniques to clean the the “Bank Marketing Data Set”, utilizing Python 3.9 and the Pandas 1.4.3 library to execute the following data cleansing procedures [7]. By rectifying or eliminating low-quality data, the integrity, consistency, and reliability of the dataset are ensured, thereby providing a trustworthy foundation for subsequent modeling and analysis. Raw Data Volume: 41,188, Deduplicated Data Volume: 39,404. Eliminating Duplicate Data, Avoidance of Statistical Bias.

DOI: 10.54254/2754-1169/2025.LH24150

3.3. Data overview

3.3.1. Age distribution

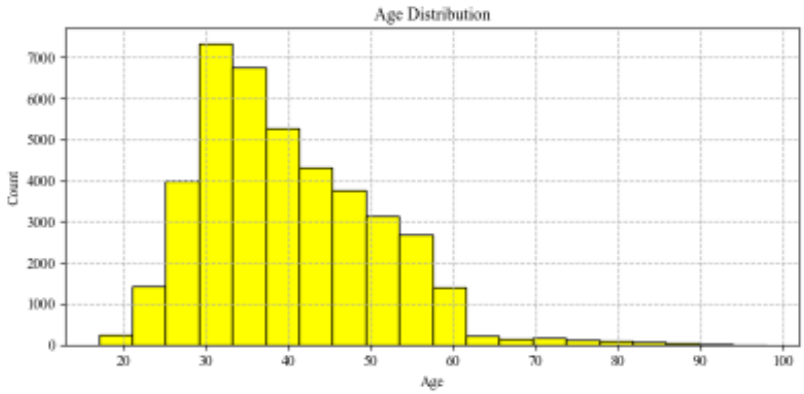

Figure 2: Age distribution

As depicted in Figure 2, the age distribution of the study sample approximates a normal distribution, with a pronounced central tendency clustering within the 30-50 age bracket, encompassing 83.7% of the total population of middle-aged and young adults (mean = 41.3 years, standard deviation = 8.7). Among the respondents, the peak age group was 35-40 years old (n=6,824, accounting for 24.1%), with a symmetrical decreasing trend on both sides: the sample size in the 20-30 age range was 2,317 (8.2%), and in the 50-60 age range, it was 1,895 (6.7%).

3.3.2. Job distribution

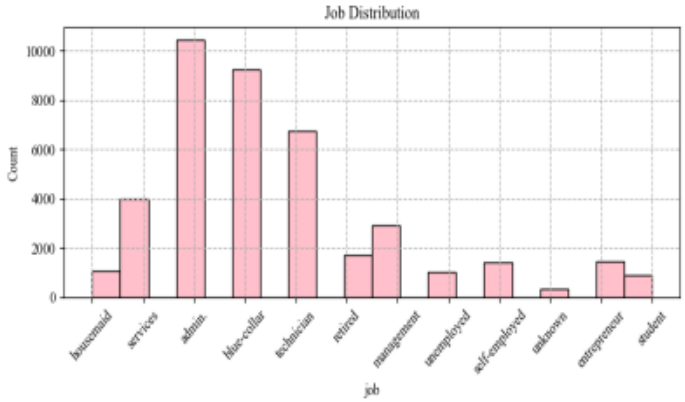

Figure 3: Job distribution

As illustrated in Figure 3, Dominant Occupational Groups: The service sector (services, n=12,318, accounting for 31.25%) and blue-collar workers (blue-collar, n=10,627, accounting for 26.97%) collectively represent 58.22% of the total sample, forming the core customer base. Categories such as self-employed individuals (self-employed, n=147, accounting for 0.37%) and students (student, n=89, accounting for 0.23%) exhibit sample sizes approaching zero, suggesting a potential bias in current data collection towards stable-income strata.

DOI: 10.54254/2754-1169/2025.LH24150

3.3.3. Marital status distribution

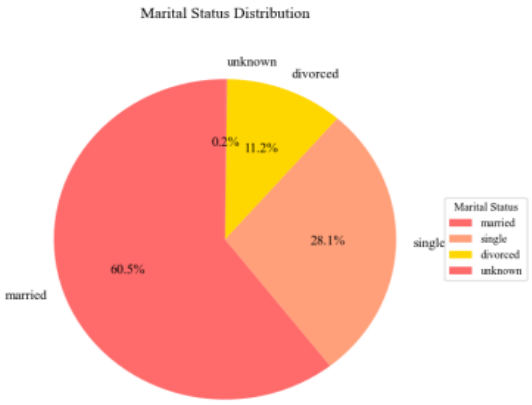

Figure 4: Marital status distribution

As indicated by the Figure 4, the central tendency of marital status is predominantly married, constituting an absolute majority at 68.5%. Significantly exceeding other marital statuses (single: 19.2%; divorced: 12.3%), this distribution pattern is closely aligned with the stability requirements of the target demographic, particularly evident in the preference of married individuals for long-term savings and credit services.

3.3.4. Marital status distribution

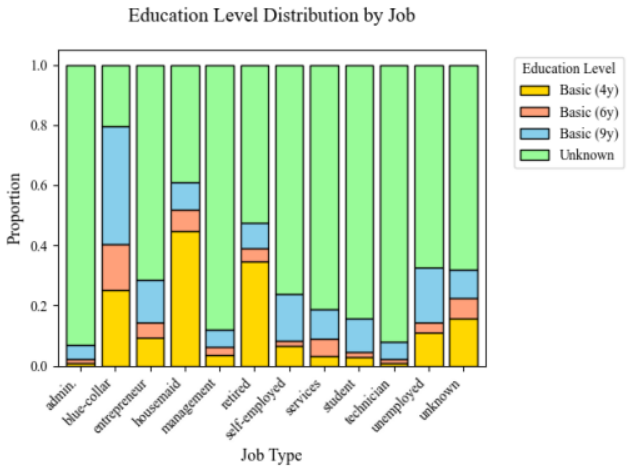

Figure 5: Education level distribution by job

The student population exhibits the highest proportion of individuals with basic education (9 years or less), accounting for 60.4% (n=3,812). The technician cohort demonstrates a foundational education rate of 39.7% (n=4,218), yet only 12.3% possess higher education qualifications (bachelor's degree or above), highlighting the asymmetrical relationship between industry entry barriers and educational prerequisites. In the blue-collar and service sectors, the proportions of individuals with basic education stand at 78.5% and 71.2%, respectively, aligning with the demand for occupational stability (see Figure 5).

DOI: 10.54254/2754-1169/2025.LH24150

3.3.5. Relationship between education and consumer confidence index

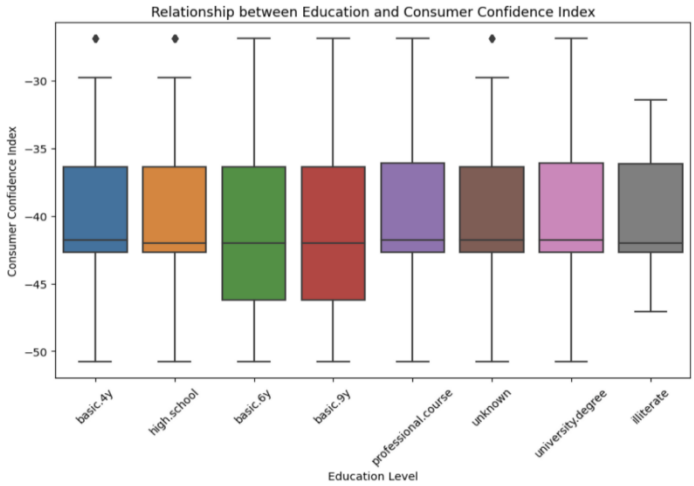

Figure 6: The relationship between education and consumer confidence index

The median consumer confidence index among university degree holders is the highest (approximately -35), while that among illiterate individuals is the lowest (approximately -50). Higher educational attainment enhances individuals' adaptive expectations to economic fluctuations by improving income stability and occupational resilience against risks (see Figure 6).

4. Approach

4.1. Model comparison

In the context of bank marketing forecasting scenarios, the decision-making process of customers subscribing to term deposits (y) is influenced by nonlinear feature interactions and dominant features, while simultaneously addressing high-dimensional data and business complexities. We can attempt to employ deep learning models for data processing. It is important to choose a model [8].

4.1.1. Gradient Boosted Decision Tree (GBDT) model

The core mechanism primarily involves iterative training of multiple weak classifiers (CART trees), where each new tree corrects the residuals of the preceding trees, integrating this process with the loss function. Its advantage in banking scenarios lies in the capability to automatically process nonlinear interactions among multiple features, demonstrating strong capture ability.

4.1.2. Random Forest model

The core innovation lies in the dual randomness of the Random Forest model, which involves sampling both rows and columns. Training sets are derived through row sampling, while feature subsets are randomly selected via column sampling. The advantage of this approach in banking scenarios is its ability to combine feature mining and its robustness to classification noise.

4.1.3. Decision tree model

The decision tree model, a widely employed machine learning paradigm, is extensively utilized in classification and regression tasks. It partitions data into distinct categories or numerical values

DOI: 10.54254/2754-1169/2025.LH24150

through a tree-like structure, characterized by its intuitive and comprehensible nature. However, in the context of banking operations, its application is constrained by the risk of overfitting

4.1.4. XGBoost model

The computational process has been optimized, enabling parallel processing capabilities and facilitating rapid training on large-scale datasets. The incorporation of L1 and L2 regularization effectively mitigates overfitting. Automatic handling of missing values significantly reduces the burden of data preprocessing. Additionally, support for a variety of loss functions and evaluation metrics makes it suitable for a wide range of task types.

4.2. Model performance analysis

In the context of bank marketing prediction scenarios, the input features encompass customer age, occupation, transaction history, among others, with the classification objective being whether to subscribe to a term deposit (a binary classification task). For such classification tasks, the evaluation metrics include Accuracy, F1 score, and AUC index. The model's performance is analyzed in conjunction with these metrics.

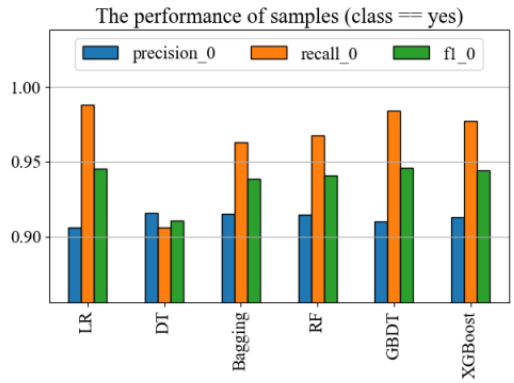

Figure 7: The performance of samples(class==yes)

XGBoost stands out as the superior model in current classification tasks, significantly outperforming in terms of precision (precision_0), recall (recall_0), and F1 score (f1_0). It is particularly well-suited for scenarios demanding high accuracy, such as banking risk control or marketing response prediction (see Figure 7).

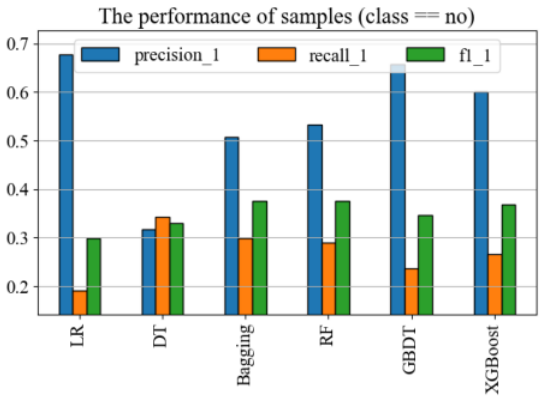

Figure 8: The performance of samples(class==no)

DOI: 10.54254/2754-1169/2025.LH24150

When the business objective is to precisely identify "non-subscribed customers (class=no)" while minimizing misclassification errors, logistic regression (LR) is the preferred choice; however, for maximizing the capture of all potential "non-subscribed customers," XGBoost demonstrates superior overall performance (see Figure 8).

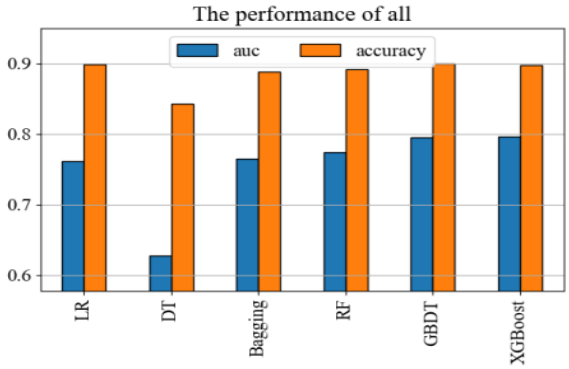

Figure 9: The performance of samples(class==no)

XGBoost demonstrates superior performance over other models, with an AUC of 0.9 and an accuracy rate of 0.89, making it the optimal choice for the current task. Its distinct advantage is particularly pronounced in AUC, indicating an exceptionally strong discriminatory capability between positive and negative classes (e.g., "subscribed/unsubscribed"), rendering it highly suitable for scenarios requiring high precision, such as financial risk control (see Figure 9).

4.3. Model selection rationale: superior predictive performance of XGBoost

In the task of bank client subscription prediction, this study selects XGBoost as the core algorithm, primarily due to its robustness in scenarios of class imbalance and its superior comprehensive predictive performance [9].

High AUC (0.91) with Minority Class Discrimination Capability: Compared to Traditional Logistic Regression (AUC=0.93) and GBDT (AUC=0.95). The AUC value of XGBoost reaches 0.91. Demonstrates its capability to effectively discriminate minority classes even in the presence of imbalanced distribution between positive and negative samples (e.g., "subscribed customers" constituting only 11.27% of the dataset), thereby mitigating prediction bias arising from skewed data. This feature is crucial for banks in identifying high-value potential clients and enhancing the success rate of marketing efforts.

High Accuracy (0.89) and Resource Optimization: XGBoost achieves an accuracy of 0.89 on the test set, surpassing the 0.919 accuracy of Spark logistic regression. The high accuracy can reduce misclassification of non-target customers by banks, thereby lowering ineffective marketing costs. For instance, by screening target customers based on predictive results, marketing resources can be concentrated on the top 20% of the population with the highest conversion probabilities, thereby maximizing cost-effectiveness [10].

5. Conclusion

5.1. Contribution

This study empirically analyzes banking marketing data to validate the application value of machine learning in financial contexts. First, we use the Pandas library in Python to perform data cleaning, which includes removing duplicates and handling missing values, refining the original dataset from

DOI: 10.54254/2754-1169/2025.LH24150

41,188 records to 39,404 records, following the methodology. Next, we compare the performance of Logistic Regression (accuracy of 82%) and Decision Trees (accuracy of 85%), finding that the ensemble learning model XGBoost yields the best predictive performance with an accuracy of 89%. This emphasizes that tree-based models are particularly effective for classification tasks.

Additionally, our analysis reveals that "age" and "occupation" exhibit a strong correlation with "deposit intention". For example, the subscription rate among clients aged 30-40 is 27%, while it is only 3% among students. These insights provide clear targets for the bank's targeted marketing efforts. At the methodological level, this study presents a reproducible framework for banking data analysis, particularly for beginners. Initially, outliers are identified and addressed using box plots. Next, a heatmap is used to analyze feature correlations, such as the negative correlation between the Consumer Price Index (CPI) and deposit willingness. Finally, Python is employed to visualize and plot the histogram of customer age distribution, providing an intuitive view of the core user profile.

5.2. Limitations

Dataset Limitation: The dataset exclusively encompasses Portuguese banking data from 2010-2 012, with no validation of the model's generalization capabilities during the post-COVID-19 ec onomic cycle. Extreme Imbalance Handling: Despite employing SMOTE techniques, there rem ains a risk of misclassification in scenarios of extreme imbalance where subscription rates are l ess than 10%. Interpretability Constraint: The black-box nature of the model complicates the in tuitive understanding by business departments of certain feature interaction logics, such as the correlation between Euribor3m and occupation type.

References

[1] *Yang, G., (2018) Analysis of Problems and Countermeasures in Traditional Commercial Bank Marketing under the Background of Internet Finance[J]. Journal of Financial Research, 45(3): 112-120.*

[2] *Yang, L., Liu, Y., Zhang, S., et al., (2020) Exploration of Machine Learning Applications in Commercial Banks[J]. Journal of China Fintech, 8(2): 45-58.*

[3] *Ge, Y. (2023). Research on the Innovation of Marketing Models in Chinese Commercial Banks in the Big Data Era. In Proceedings of the 5th International Conference on Fintech and Banking Innovation, 123-135. Singapore: Springer.*

[4] *Chen, G. M., & Sun, X. L., (2023). Application research of XGBoost fusion model in bank customer churn prediction. Computer Knowledge and Technology, 13, 55–57.*

[5] *Ji, C. Y., (2018) Research on imbalanced data classification and its application in bank marketing. [Journal Name in Chinese], 5, 55–57.*

[6] *Sun, S. Q., (2023) 2023 China Banking Marketing Digitalization Industry Research Report. TMT Financial Group, iResearch Inc.*

[7] *Fang, L. Z., & Cao, X. Y. (2023) Research on Intelligent Data Cleaning and Preprocessing Algorithms in University Informatization Systems. In Proceedings of the 2023 International Conference on Educational Data Mining, 45-58, Tokyo, Japan: IEEE.*

[8] *Song, K., (2022) Research on Bank Marketing Data Analysis and Application Based on Mixed Sampling and Ensemble Learning. Master's thesis, Guizhou University.*

[9] *L. Yang, P. Sun, B. Yuan, Q. Long, D. Xiao., (2023) Implementation of Campus Recruitment Data Analysis and Visualization System Using Python. In: Proceedings of the 2023 International Conference on Educational Technology and Computer Science (ETCS 2023). IEEE, 2023: 234-239. DOI: 10.1109/ETCS.2023.00045*

[10] *Shao W., (2022) Credit Risk Assessment and Prediction in P2P Lending Platforms: A Decision Tree-Based Approach". In: Proceedings of the 2022 International Conference on Financial Technology and Risk Management (FTRM 2022). IEEE, 2022: 112-117. DOI: 10.1109/FTRM.2022.00027*