2025-03

Predicting customer subscription in bank telemarketing campaigns using ensemble learning models

Peter, Michael

Elsevier

https://doi.org/10.1016/j.mlwa.2025.100618

Machine Learning with Applications 19 (2025) 100618

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Predicting customer subscription in bank telemarketing campaigns using ensemble learning models

Michael Peter [a,b,*], Hawa Mofi [a], Said Likoko [a], Julius Sabas [a], Ramadhani Mbura [a], Neema Mduma [a]

[a] Department of Applied Mathematics, The Nelson Mandela African Institution of Science and Technology (NM-AIST), P.O. Box 447, Arusha, Tanzania
[b] Department of Mathematics and ICT, The Open University of Tanzania (OUT), P.O Box 23409, Dar es Salaam, Tanzania

ARTICLE INFO

Keywords:
Ensemble learning
Subscription prediction
Telemarketing campaign
Feature importance
Predictive modeling

ABSTRACT

This study investigates the use of ensemble learning models bagging, boosting, and stacking to enhance the accuracy and reliability of predicting customer subscriptions in bank telemarketing campaigns. Recognizing the challenges posed by class imbalance and complex customer behaviors, we employ multiple ensemble techniques to build a robust predictive framework. Our analysis demonstrates that stacking models achieve the best overall performance, with an accuracy of 91.88% and an Receiver Operating Characteristic Area Under the Curve (ROC-AUC) score of 0.9491, indicating a strong capability to differentiate between subscribers and non-subscribers. Additionally, feature importance analysis reveals that contact duration, economic indicators like the Euro interbank offered (Euribor) rate, and customer age are the most influential factors in predicting subscription likelihood. These findings suggest that by focusing on customer engagement and economic trends, banks can improve telemarketing campaign effectiveness. We recommend the integration of advanced balancing techniques and real-time prediction systems to further enhance model performance and adaptability. Future work could explore deep learning models and interpretability techniques to gain deeper insights into customer behavior patterns. Overall, this study highlights the potential of ensemble models in predictive modeling for telemarketing, providing a data-driven foundation for more targeted and efficient customer acquisition strategies.

1. Introduction

In today's competitive banking sector, telemarketing campaigns remain a crucial channel for promoting financial products and services, particularly term deposits. However, these campaigns often face significant challenges in terms of resource allocation and customer targeting efficiency. The ability to accurately predict customer subscription likelihood has become increasingly vital for optimizing marketing strategies and improving return on investment. As financial institutions continue to generate vast amounts of customer interaction data, there is a growing opportunity to leverage advanced machine learning techniques for more precise prediction of customer responses.

Studies have demonstrated the significant potential of machine learning in banking marketing optimization, particularly through ensemble learning methods. Moro et al. (2014) conducted pioneering work using data mining approaches to predict term deposit subscriptions, achieving notable success with neural networks and support vector machines. Their study established the importance of feature selection and data preprocessing in improving prediction accuracy.

Building on this foundation, Zhang et al. (2022) introduced a hybrid approach combining random forests with gradient boosting, demonstrating superior performance in handling imbalanced banking datasets and achieving a 15% improvement in prediction accuracy compared to traditional single-model approaches. Additionally, further work has shown that ensemble techniques can provide robust predictive capabilities under challenging data conditions. For example, Thabet et al. (2024) highlighted the efficacy of bagging in reducing variance and increasing stability for banking classification tasks. Meanwhile, Sikri et al. (2024) explored boosting-based models, revealing improved accuracy for customer churn predictions. Talukder et al. (2024) investigated hybrid stacking frameworks tailored to imbalanced data distributions in finance, emphasizing versatility in model adaptation.

* Corresponding author at: Department of Applied Mathematics, The Nelson Mandela African Institution of Science and Technology (NM-AIST), P.O. Box 447, Arusha, Tanzania.
E-mail addresses: peterm@nm-aist.ac.tz (M. Peter), mofih@nm-aist.ac.tz (H. Mofi), likokos@nm-aist.ac.tz (S. Likoko), massawej@nm-aist.ac.tz (J. Sabas), mburar@nm-aist.ac.tz (R. Mbura), neema.mduma@nm-aist.ac.tz (N. Mduma).

https://doi.org/10.1016/j.mlwa.2025.100618
Received 22 November 2024; Received in revised form 27 December 2024; Accepted 1 January 2025
Available online 7 January 2025

The complexity of customer decision-making in banking products, coupled with the inherent imbalance in subscription datasets where successful subscriptions typically represent a minority class, presents substantial challenges for traditional predictive models. These challenges are further compounded by the multifaceted nature of features influencing customer decisions, including demographic factors, economic indicators, and previous campaign outcomes (Sharma et al., 2024; Zhang et al., 2024). Various techniques such as cost-sensitive learning and Synthetic Minority Over-sampling Technique (SMOTE) have been developed to address these imbalance issues, with studies showing significant improvements in model performance through resampling methods (Bansal et al., 2022).

Ensemble learning methods have emerged as particularly effective solutions in this domain. Milli et al. (2024) developed an advanced ensemble framework that effectively handled class imbalance in banking data, achieving a balanced accuracy of 82% and demonstrating the superiority of ensemble methods over individual classifiers. The effectiveness of ensemble learning in financial applications has been further validated by Wang and Chi (2024), who implemented a stacking-based approach that combined multiple base learners to predict customer churn in retail banking. Their research showed that ensemble methods could capture complex patterns in customer behavior more effectively than traditional approaches, leading to a 20% improvement in prediction precision. Recent research has also highlighted the success of hybrid ensemble models, achieving up to 95.6% accuracy in predicting customer responses to telemarketing campaigns (Saeed et al., 2022).

This study aims to enhance the accuracy and reliability of subscription predictions in bank telemarketing campaigns through the application of ensemble learning models. Specifically, we investigate the effectiveness of three prominent ensemble techniques: Bagging, Boosting, and Stacking. Our approach focuses on developing a comprehensive predictive framework that not only addresses the challenge of class imbalance but also capitalizes on the complementary strengths of different base learners to improve overall prediction performance.

The primary contributions of this research are three fold:

- We propose a systematic approach to preprocessing and feature engineering specifically tailored for bank marketing data, addressing common challenges such as missing values, categorical variable encoding, and feature scaling.
- We implement and compare three distinct ensemble learning strategies Bagging (using Random Forest), Boosting (using Gradient Boosting and AdaBoost), and Stacking with optimized hyperparameters through extensive cross-validation.
- We provide a comprehensive evaluation framework that goes beyond traditional accuracy metrics to include balanced accuracy, Matthews Correlation Coefficient, and Cohen's Kappa, offering deeper insights into model performance across different aspects of prediction quality.

This research extends the current literature by offering a detailed comparative analysis of ensemble learning approaches in the specific context of bank telemarketing success prediction. While previous studies have typically focused on individual techniques or limited combinations, our work provides a comprehensive evaluation of multiple ensemble approaches and their synergistic effects. Our findings contribute to both the theoretical understanding of ensemble methods' effectiveness in handling imbalanced financial datasets and the practical application of these techniques in real-world marketing scenarios. The results of this study provide valuable insights for banking institutions seeking to optimize their telemarketing campaigns through data-driven decision-making, while acknowledging the ongoing challenges of model selection and computational resource requirements in practical applications.

The organization of the paper is as follows: Section 2 describes the data description and methodology used for experimentation and evaluation. In Section 3, results and discussion are presented. Finally, the paper concludes along with the future studies and recommendation scope in Section 4.

Table 1
Feature categories and variable description.

Feature category	Variables
Demographic Information	age, job, marital status, and education
Financial Indicators	housing loan, personal loan, and default status
Campaign Data	number of contacts in the campaign, type of last contact, and duration of the last contact
Past Campaign Outcomes	previous campaign contacts and results
Economic Context	Macro-economic indicators, including the employment variation rate, consumer price index, consumer confidence index, Euribor (3-month rate), and the number of employees

2. Methodology

2.1. Data description

The dataset used in this study was obtained from the UCI Machine Learning Repository, originally collected from a Portuguese banking institution's direct marketing campaigns conducted between May 2008 and November 2010 (Moro et al., 2014). The primary goal of the campaign was to promote term deposit subscriptions to potential customers. The dataset contains both demographic and behavioral features, as well as socio-economic indicators, which collectively provide a comprehensive view of each customer. The target variable, labeled *subscription status*, indicates whether a customer subscribed to the term deposit. The data includes 41,188 entries as described in Table 1.

2.2. Data preprocessing

Given the complexity and high dimensionality of the dataset, several steps were performed to prepare the data for analysis as presented in Fig. 1, this include, data Cleaning, encoding Categorical Variables, feature scaling, handling class imbalance,final to best prediction model.

2.3. Ensemble models

This study explores three distinct ensemble learning approaches—bagging, boosting, and stacking to predict the likelihood of subscription with high accuracy and reliability.

Bagging was implemented using the random forest model. By training multiple decision trees on bootstrapped samples and aggregating their predictions, the model leverages diversity in individual tree outputs to improve accuracy and reduce overfitting. To optimize the model's performance, a grid search was performed to fine-tune hyperparameters such as the number of trees, maximum depth, and minimum samples per split.

Boosting was applied using gradient boosting and Adaboost. These models were chosen for their ability to improve predictive performance through sequential learning, where each new tree corrects errors from the previous ones. The gradient boosting model builds trees sequentially to minimize residual errors. Its performance was optimized by adjusting parameters such as learning rate, the number of estimators, and depth. Adaboost was selected for its efficiency and flexibility in handling missing values. Cross-validation was used to tune hyperparameters like learning rate, maximum depth, and gamma.

Stacking was implemented as a combination of base learners. The stacked ensemble model included random forest, bagging classifier, gradient boosting, and Adaboost as base models. Each base model was trained independently, and a portion of the training set was reserved to generate predictions. These predictions were then used as inputs for a logistic regression meta-model, which combined the strengths of individual models to maximize predictive accuracy.

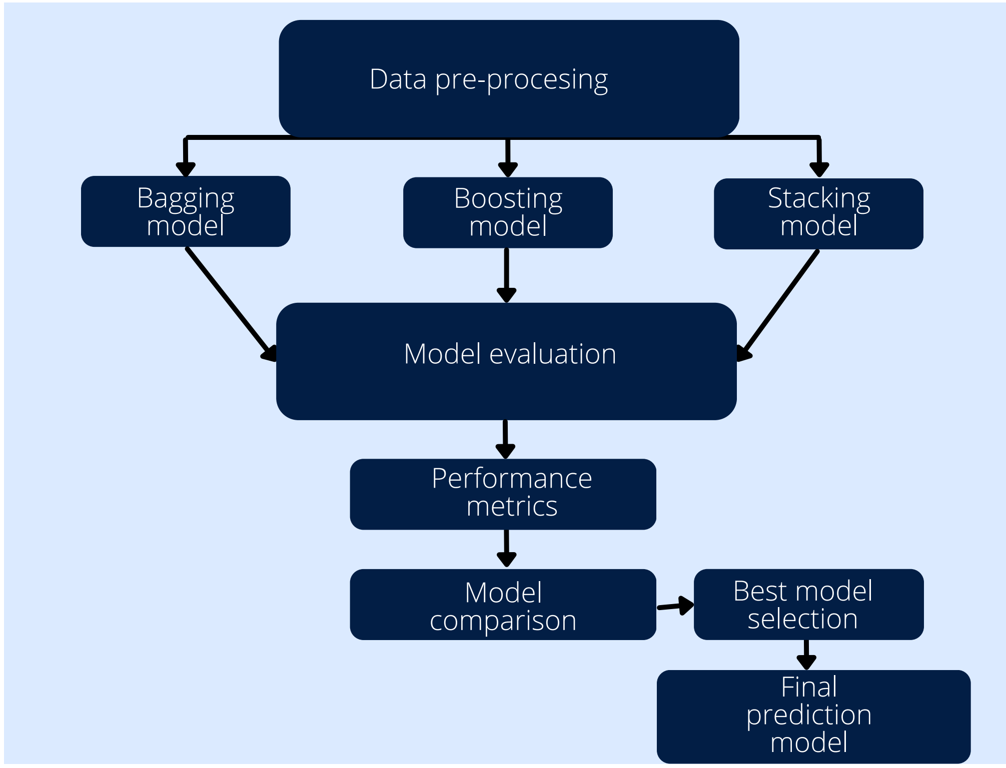

Fig. 1. Framework for Building an Ensemble-Based Predictive Model: This flowchart outlines the process beginning with data preprocessing, followed by training ensemble models (bagging, boosting, and stacking). The models are evaluated using performance metrics, compared, and the best-performing model is selected for the final prediction.

2.4. Model evaluation

To assess model performance, we used Stratified K-Fold Cross-Validation, which ensures that each fold maintains the class distribution of the target variable. This approach allows for robust performance estimates across all data subsets.

2.5. Evaluation metrics

Given the imbalanced nature of the data set, the following metrics were chosen to evaluate model effectiveness, accuracy, while commonly used, accuracy alone may be misleading in imbalanced datasets, it is reported alongside other metrics for completeness. Precision, this measures the proportion of true positive subscriptions among all predicted positives, indicating the model's ability to avoid false positives. Recall, also known as sensitivity, this metric indicates the proportion of true positives detected among actual positives, reflecting the model's ability to capture true subscribers. F1-Score, the harmonic mean of precision and recall, which balances both metrics to provide a single performance measure. ROC-AUC, the area under the receiver operating characteristic curve provides a threshold-independent metric, evaluating the model's ability to distinguish between classes. Lastly, Matthews correlation coefficient (MCC) and Cohen's Kappa offers deeper insights into model reliability are used.

2.6. Feature importance and interpretability

To understand which features contributed most to model predictions, feature importance was examined for each ensemble model. Additionally, shapley additive explanations values were used to interpret predictions on an individual level, clarifying how specific features influenced each customer's likelihood of subscription.

2.7. Implementation tools

The implementation was performed using Python libraries such as scikit-learn for data preprocessing, and ensemble models such as random forest, gradient boosting, Adaboost, and a stacking of model which involved combining the three ensemble models. Imbalanced learn for resampling techniques in interpretability of model predictions were employed. This methodology establishes a rigorous approach to predictive modeling, using ensemble techniques to capture complex patterns within the data and ensuring robustness against the challenges posed by imbalanced classes and high-dimensional data.

3. Results and discussion

3.1. Model performance comparison

Based on the results presented in Table 2, this section provides an analysis of the performance of various ensemble models specifically bagging, boosting, and stacking models in predicting customer subscription to bank services. Each model's performance is evaluated in terms of accuracy, precision, recall, F1-score, ROC-AUC, MCC, and Cohen's kappa. All collectively, offer insights into the models' strengths and weaknesses in addressing the class imbalance challenge and enhancing prediction reliability. The experimental results reveal that the stacking model outperforms other ensemble methods in predicting customer subscription status in bank telemarketing campaigns. With an accuracy of 91.88% and a ROC-AUC of 0.9491, the stacking model achieves the highest scores across all evaluated models. These metrics highlight its ability to reliably differentiate between customers likely to subscribe and those who are not, making it a robust choice for bank marketing teams aiming to prioritize high-potential leads. Additionally, the stacking model displays a balanced precision of 0.6832 and recall of 0.5305, indicating its ability to accurately identify subscribed customers without excessively misclassifying non-subscribed customers as

Table 2
Performance evaluation of ensemble learning models.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MCC	Cohen's kappa
BM							
RF	0.9139	0.6619	0.4941	0.5658	0.9413	0.5294	0.5230
BC	0.9133	0.6356	0.5540	0.5920	0.9396	0.5621	0.5604
BOM							
GB	0.9185	0.6828	0.5273	0.5951	0.9476	0.5598	0.5567
ADB	0.9058	0.6916	0.3070	0.4252	0.9361	0.4748	0.4570
STCK							
ST	0.9188	0.6832	0.5305	0.5972	0.9491	0.5588	0.5549

Note: **BM**: Bagging Model, RF: Random Forest, BC: Bagging Classifier, **BOM**: Boosting Model, GB: Gradient Boosting, ADB: AdaBoost, **STCK**: Stacking Model, ST: Stacking.

Table 3
Experimental results of ensemble models.

Feature	Importance
Contact duration	0.298962
Daily three month Euribor rate	0.094858
Age	0.088029
Quarterly average of the total number of employed citizens	0.059697
Employment variation rate, with a quarterly frequency	0.045529
Campaign	0.041979
P.days	0.036019
P.outcome	0.033422
Marital	0.0235337
Cons.conf.index	0.022691

positive cases. This balance of precision and recall is critical in ensuring that marketing efforts are not wasted on uninterested customers, thus optimizing resources and enhancing campaign effectiveness.

Examining Matthews correlation coefficient (MCC) and Cohen's Kappa offers deeper insights into model reliability, particularly for imbalanced data scenarios, where these metrics are considered more robust indicators of performance. The Bagging Classifier achieves the highest MCC of 0.5621 and Kappa of 0.5604, indicating strong predictive consistency and agreement between the predicted and actual classifications. MCC values signify accurate classification across both positive and negative classes, while a high Kappa score shows minimal impact of random chance in the predictions. This demonstrates that the bagging classifier excels at managing imbalanced data, striking a balance between precision, recall, and overall classification accuracy. Although the stacking model performs well across other metrics, its MCC of 0.5549 and Kappa of 0.5588 values are slightly lower, indicating a small gap in handling data imbalance compared to bagging.

Overall, these results emphasize that each ensemble approach offers unique strengths. The stacking model leads in general prediction accuracy and ROC-AUC, making it effective for broader classification tasks. However, the bagging classifier stands out for its exceptional MCC and Kappa scores, suggesting a more balanced and reliable performance for imbalanced data. This distinction highlights the value of using multiple evaluation metrics to comprehensively assess model performance, especially in complex scenarios like predicting customer subscription outcomes, where both high predictive power and balanced classification reliability are crucial.

The significance of these results lies in their potential to transform how banks approach customer outreach in telemarketing campaigns. This ability to accurately target potential subscribers is crucial for banks, as it enables them to optimize marketing resources by focusing efforts on high-probability leads. By improving the efficiency of targeting, banks can reduce the costs associated with reaching uninterested customers, increase conversion rates, and improve the overall return on investment for telemarketing campaigns.

Furthermore, the results highlight the value of using ensemble models to enhance predictive performance, especially in the context of imbalanced data. Bank telemarketing data often contains many more non-subscribers than subscribers, making it challenging to identify the subscribed customers effectively. The stacking model's balanced precision and recall indicate that it not only achieves high accuracy but also performs well in capturing the true positives customers who will actually subscribe. This balance is essential in reducing false negatives, thereby helping banks avoid missed opportunities with potential customers. Ultimately, this predictive framework could serve as a strategic asset for banks, improving customer acquisition processes and driving growth by leveraging advanced ensemble learning techniques.

3.2. Feature importance

The feature importance results in Table 3 highlight the relative influence of different variables on subscription predictions in the bank telemarketing campaign. Contact duration emerges as the most significant predictor, with an importance score of 0.2989. This value suggests that the length of each customer interaction has a considerable impact on whether a customer will subscribe, indicating that the longer a bank representative engages with a potential customer, the more likely they are to secure a positive outcome. Following contact duration, the Daily three-month Euribor rate and Age also show notable importance, with scores of 0.0949 and 0.0880, respectively. The Euribor rate reflects broader economic conditions, implying that economic sentiment and interest rates influence customer financial decisions. Age also provides insights into the demographic profile of likely subscribers, helping banks understand which age groups are more responsive to telemarketing.

The significance of these findings lies in their practical application for optimizing telemarketing strategies. By identifying and focusing on key factors like contact duration, economic indicators, and demographic characteristics, banks can tailor their campaigns more effectively. For instance, customer service teams can be trained to prioritize longer, more engaging calls, especially with demographics and economic conditions that have shown higher subscription rates. Additionally, economic variables such as the Euribor rate and employment variation rate allow the bank to adjust their outreach based on market conditions, potentially scaling efforts during favorable economic times.

4. Conclusion

This study aimed to improve the accuracy and reliability of subscription predictions in a bank telemarketing campaign by leveraging ensemble learning techniques specifically bagging, boosting, and stacking. The results in Table 2 demonstrate that ensemble methods can effectively capture the complexities of customer subscription behavior, yielding competitive performance across key metrics like accuracy, precision, recall, F1-score, and ROC-AUC. Among the tested models, the stacking ensemble achieved the best overall performance, with a high ROC-AUC score of 0.9491, indicating a strong ability to distinguish between subscribers and non-subscribers. Gradient boosting and stacking models outperformed bagging techniques, suggesting that boosting and stacking approaches, which better handle imbalanced data and emphasize challenging cases, are particularly effective in this domain.

Table 3's feature importance analysis further highlights the critical factors influencing subscription predictions. Contact duration emerged as the most influential feature, underscoring the role of customer engagement in telemarketing success; longer call durations may reflect higher interest or engagement, increasing the likelihood of subscription. Economic indicators, such as the daily three-month Euribor rate and employment variation rate, also ranked high in importance, suggesting that broader economic factors impact customers' propensity to subscribe. Demographic factors like age and campaign-specific variables, including the number of contacts made, provide additional insights into customer behavior and preferences. Together, these findings underscore the value of ensemble learning and feature analysis for building predictive models that not only achieve high accuracy but also provide actionable insights for campaign optimization. By focusing on these critical features and refining the choice of ensemble techniques, banks can enhance their targeting strategies, ultimately improving subscription rates and customer satisfaction in telemarketing campaigns.

Based on our findings, we recommend that telemarketing strategies prioritize customer engagement, leveraging factors like contact duration to enhance subscription likelihood, and optimize campaign timing by monitoring economic indicators such as the Euribor rate. Targeting should also be refined to focus on customer demographics and behaviors associated with higher subscription probabilities.

Future research should prioritize the development and implementation of real-time predictive systems to enable immediate, data-driven adjustments during customer interactions, enhancing campaign responsiveness and effectiveness. Exploring the feasibility of such systems involves addressing several technical and operational challenges, including ensuring data quality, minimizing latency in real-time streams, managing the computational demands of ensemble models, and integrating these systems with existing telemarketing infrastructure. Leveraging technologies like Apache Kafka or Spark Streaming for real-time data ingestion and processing, alongside employing lightweight ensemble techniques or model optimization methods (e.g., quantization or distillation), could provide practical solutions to these challenges. Investigating these aspects will pave the way for the deployment of effective real-time systems in telemarketing campaigns.

Additionally, future work could explore deep learning models and incorporate of transfer learning techniques and cross domain validation to better capture complex customer behavior patterns while focusing on enhancing model interpretability through techniques like Shapley Additive Explanations (SHAP) or Local Interpretable Model-Agnostic Explanations (LIME). Longitudinal studies across campaigns are also recommended to assess the consistency of model accuracy and adaptability over time. Together, these directions aim to further refine predictive strategies, support robust and adaptive telemarketing frameworks, and maximize the effectiveness of data-driven customer acquisition strategies.

CRediT authorship contribution statement

Michael Peter: Conceptualization, Data curation, Methodology, Software, Investigation, Writing – original draft, Formal analysis, Visualization, Writing – review & editing. **Hawa Mofi:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Writing – review & editing. **Said Likoko:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Writing – review & editing. **Julius Sabas:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Writing – review & editing. **Ramadhani Mbura:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Writing – review & editing. **Neema Mduma:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors extend heartfelt thanks to The Nelson Mandela African Institution of Science and Technology, and The Open University of Tanzania for providing first-rate workspaces and resources. Their assistance has greatly aided the effective completion of this work

Data availability

Data will be made available on request.

References

Bansal, A., Singh, S., Jain, Y., & Verma, A. (2022). Analysis of ensemble classifiers for bank churn prediction. In *2022 international conference on computing, communication, and intelligent systems* (pp. 593–598). http://dx.doi.org/10.1109/ICCCIS56430.2022.10037623.

Milli, M. E. F., Aras, S., & Kocakoç (2024). Investigating the effect of class balancing methods on the performance of machine learning techniques: Credit risk application, 5. (pp. 55–70). http://dx.doi.org/10.56203/iyd.1436742.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. http://dx.doi.org/10.1016/j.dss.2014.03.001.

Saeed, S. E., Hammad, M., & Alqaddoumi, A. (2022). Predicting customer's subscription response to bank telemarketing campaign based on machine learning algorithms. In *2022 international conference on decision aid sciences and applications* (pp. 1474–1478). http://dx.doi.org/10.1109/DASA54658.2022.9765152.

Sharma, V., Khanna, S., Gautam, P., & Kaushik, J. (2024). *Bank customer identification for targeted marketing and revenue optimisation: a comparative analysis of predictive models* (pp. 1–6). http://dx.doi.org/10.1109/ICRITO61523.2024.10522140.

Sikri, A., Jameel, R., Idrees, S. M., & Kaur, H. (2024). Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports*, *14*, 13097. http://dx.doi.org/10.1038/s41598-024-63750-0.

Talukder, M. A., Islam, M. M., Uddin, M. A., Hasan, K. F., Sharmin, S., Alyami, S. A., & Moni, M. A. (2024). Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. *Journal of Big Data*, *11*, 33. http://dx.doi.org/10.1186/s40537-024-00886-w.

Thabet, H. H., Darwish, wish, S. M., & Ali, G. M. (2024). Measuring the efficiency of banks using high-performance ensemble technique. *Neural Computing and Applications*, *36*, 16797–16815. http://dx.doi.org/10.1007/s00521-024-09929-y.

Wang, S., & Chi, G. (2024). Cost-sensitive stacking ensemble learning for company financial distress prediction. *Expert Systems with Applications*, http://dx.doi.org/10.1016/j.eswa.2024.124525.

Zhang, X., Zhou, Y., Lin, Z., & Wang, Y. (2024). Ensemble learning with dynamic weighting for response modeling in direct marketing. *Electronic Commerce Research and Applications*, *64*, Article 101371. http://dx.doi.org/10.1016/j.elerap.2024.101371.

Zhang, Z., Zhu, X., & Liu, D. (2022). Model of gradient boosting random forecast prediction. In *IEEE international conference on networking, sensing and control* (pp. 1–6). http://dx.doi.org/10.1109/ICNSC55942.2022.10004112.