A Comparative Analysis of Machine Learning Algorithms for Predicting the Telemarketing Campaigns of Portuguese Banking Institutions

Shiyue Huang[1,a,*]

[1]School of Computer Science, South China Normal University, Guangzhou, 510631, China
a. 20223802008@m.scnu.edu.cn
*corresponding author

Abstract: A financial institution is an entity of business that handles deposits, money circulation, saving, and other corresponding operations. Banks supply a broad spectrum of services. This differs depending on the capacity of each bank. The bank is going to provide further amenities, which is one of the many industries that frequently engage in direct product launches. Thus, banks might employ technological innovations to do market investigation before directly releasing products, which may assist with decision-making. Marketing initiatives that offer prospective consumers the option of embracing or rejecting an offered product can be carried out using direct email, phone calls, and email campaigns. The volume of incoming data continues to expand as time goes by. One of the bank institutions in Portugal faced difficulties judging whether or not its customer base would choose to sign up for a term deposit because of the growing amount of data. Consequently, to ascertain if the consumer will subscribe to a term deposit, the data mining processes utilized here will incorporate Ensemble Learning (Gradient Boosting (GB), Xg Boost (XGB), Random Forest (RF), Linear Models (Logistic Forest (RF), Linear Models (Logistic Regression (LR), Non-Linear Models (K-Near Neighbors (KNN), Support Vector Machine (SVM), Decision Tree Classifier (DT)), and Probabilistic Models (Naive Bayes (NB))included in classification techniques. Among these, Xg Boost yields the most impressive results, with the best performance in terms of AUC, model accuracy, and cross-validation accuracy in a minimum amount of computation time.

Keywords: Direct Marketing, Telemarketing Campaigns, Machine Learning

1. Introduction

A mathematical expression for data representation in the context of a business problem is machine learning. Management is now able to act swiftly on big data-driven choices to boost promotional efforts, thanks in substantial part to machine learning strategies. Given that data continually shifts, adopting the machine learning technique consumes time and occasionally yields inaccurate findings. The objective of this academic endeavor is to deploy machine learning techniques to project the effectiveness of marketing initiatives for a Portuguese banking business.

Financial institutions have to optimize the greatest use of their financial assets, and their efforts to promote should center on ways to enhance the clientele's experience. Strategies for marketing ought

to focus on luring in prospective clients in order to motivate them to accept the deals that are being presented. To efficiently sell the products and services it produces to prominent clients, the Portuguese banking system additionally utilizes a range of marketing strategies. Certain financial institutions have also implemented alternative machine learning-based marketing initiatives to take advantage of their resources and engage with prospective clients. There have been a lot of marketing campaigns run by banks, but most of these have lost their effect on most of the general public [1]. This is one of the main drivers of firms and financial institutions altering their tactics; the majority of top marketers have been reported to devote more funds to specialized campaigns, which involve determining meticulously which clients to reach out to and how to accomplish so.

The main focus of these studies has been to strengthen the direct campaign process and machine learning. That is, applying data to mitigate financial losses, which has been put forth as an approach. Banks may utilize machine learning to figure out which clients they ought to focus on with the aim of increasing their loan volume. Multiple investigators have employed diverse machine learning methods to perform this process. Real-world data from a Portuguese marketing campaign for bank deposit subscriptions was used in almost all of this research. For data collection, banks used their contact centers and organized directed marketing campaigns [2].

In the present investigation, the researcher will scrutinize the data set and implement machine learning techniques to foresee the success of banks telemarketing business operations. The data set will be correlated with a Portuguese banking institution's direct marketing initiatives. On phone calls, the marketing strategies were predetermined. From May 2008 to November 2010, 41188 entries with 20 attributes were incorporated into the assortment.

The project's researcher will endeavor to figure out the primary determinants of customers' decisions regarding signing up for a term deposit with the bank. With the objective of enhancing campaign efficiency, this investigation project will look into the customer behavior that occurs while founding a term deposit in a bank. Customer behavior is an essential element in decision-making. Recognizing the components of the banking system that substantially shape consumers' decision-making and induce them to sign up for distinct bank services is crucial.

This research that has been carried out regarding a Portuguese financial company is exceptionally noteworthy. This research project will examine the adoption of machine learning approaches to predict marketing campaign effectiveness. A detailed examination of the correlation between numerous factors and a marketing campaign's success will also be undertaken. This research project will effectively contribute to existing studies related to machine learning techniques and will also provide valuable information to key stakeholders in the banking industry [1].

Ensemble learning (GB, XGB, RF), Linear non-linear models (KNN, SVM, DT) and probabilistic models (NB) are included in classification techniques. Figure 1 below displays the set of algorithms employed in this work to estimate term deposit subscriptions.
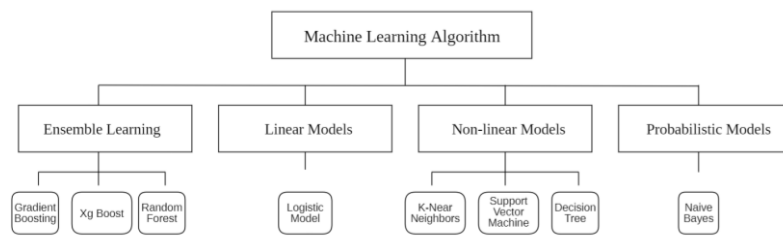

Figure 1: Algorithms used in this research (Photo/Picture credit: Original).

The section that follows functions as a reminder regarding this paper. The applicable work is put forward in Section 2. The methodologies are covered in Part 3, the outcomes of the experiment are demonstrated in Section 4, and the conclusion will be provided in Section 5.

2.   Literature Review

Portugal's banking system is highly advanced, with all commercial banks connected to the national multibanco system and offering numerous services directly from user accounts. The banking sector employs various marketing strategies to effectively promote products and services to key customers. In the banking sector, telemarketing is a key tool for promoting financial products and services. The banks are using telemarketing services to provide information to customers regarding new services and offerings [3]. Factors including interest rates, maintenance schedules, risk management, fees for service, and account duration have an impact on customers' decisions regarding opening term deposits. Practically, these investigations utilized certifiable information gathered from a Portuguese showcasing effort identified with bank store memberships [4]. The practice of telemarketing has gained sufficient traction, and banks are wisely using this tool for advertising their products and offerings.

In Portugal, the banks also encourage customers to control their spending and investment in term deposits to improve their overall earnings [4]. In telemarketing data mining, issues like unbalanced class distribution can affect learning algorithms' performance. The reliability of predictions may also be challenging to determine.

Another crucial consideration when distributing marketing tactics is subscription-based services. In Mari's study, the intent, methodology, and advantages associated with machine learning for marketing are outlined by the author. He also addresses a marketing approach powered by artificial intelligence [5]. The paper found that Random Forest (RF) outperforms other classifiers in forecasting. In another study, a multi-layer perceptron (MLP) classifier integrated with a deep learning neural network was used to optimize predictions and mitigate the impact of unbalanced data [5]. This approach achieved an accuracy of 89.98%, demonstrating the effectiveness of advanced machine learning techniques in telemarketing analytics [5].

Arushi mentions the juxtaposition of random forest and neural network algorithms using R to evaluate a client's willingness [6]. Zaki et al.'s study individually implements models such as SGD Classifier, k-nearest neighbor Classifier, and Random Forest Classifier [7]. The results indicate that the best performance among the evaluated models was exhibited by the Random Forest Classifier, achieving an accuracy of 87.5%, a negative predictive value (NPV) of 92.9972%, and a positive predictive value (PPV) of 87.8307% [7]. Li et al.'s research is designed to compare the five metrics (Type I sample f1-score, Type II sample f1-score, accuracy, Area Under Curve, Kolmogorov-Smirnov) to distinguish the feasibility of different algorithms [8]. The higher index represents the better algorithm. As the results, the neural network has the highest AUC (0.85) and highest Type I sample f1-score(0.50), while the logistic regression has the highest accuracy (0.90), KS (0.64), and Type II sample f1-score (0.50). (0.94) [8]. According to reality, the neural network is suggested to be the optimal algorithm that needs to be adopted by bank managers for client prediction [8]. Malyadri's empirical results reveal that the accuracy of random forest model reaches 92%, while the accuracy of SVM model reaches 87%. This indicates that the ensemble learning model has higher accuracy and forecasting ability than the single model [9].

3.   Methodology

3.1.   Data Selection

The Portuguese Bank dataset includes personal data from 41,188 participants, collected from May 2008 to November 2010. It contains seventeen features: six numerical (age, average annual balance, last contact day, contact time, current campaign contacts, and prior campaign contacts) and ten

categorical (job type, marital status, education level, credit default status, personal loans, contact type, housing loans, last contact month, previous campaign outcomes, and term deposit subscription status).

3.2. Data Processing

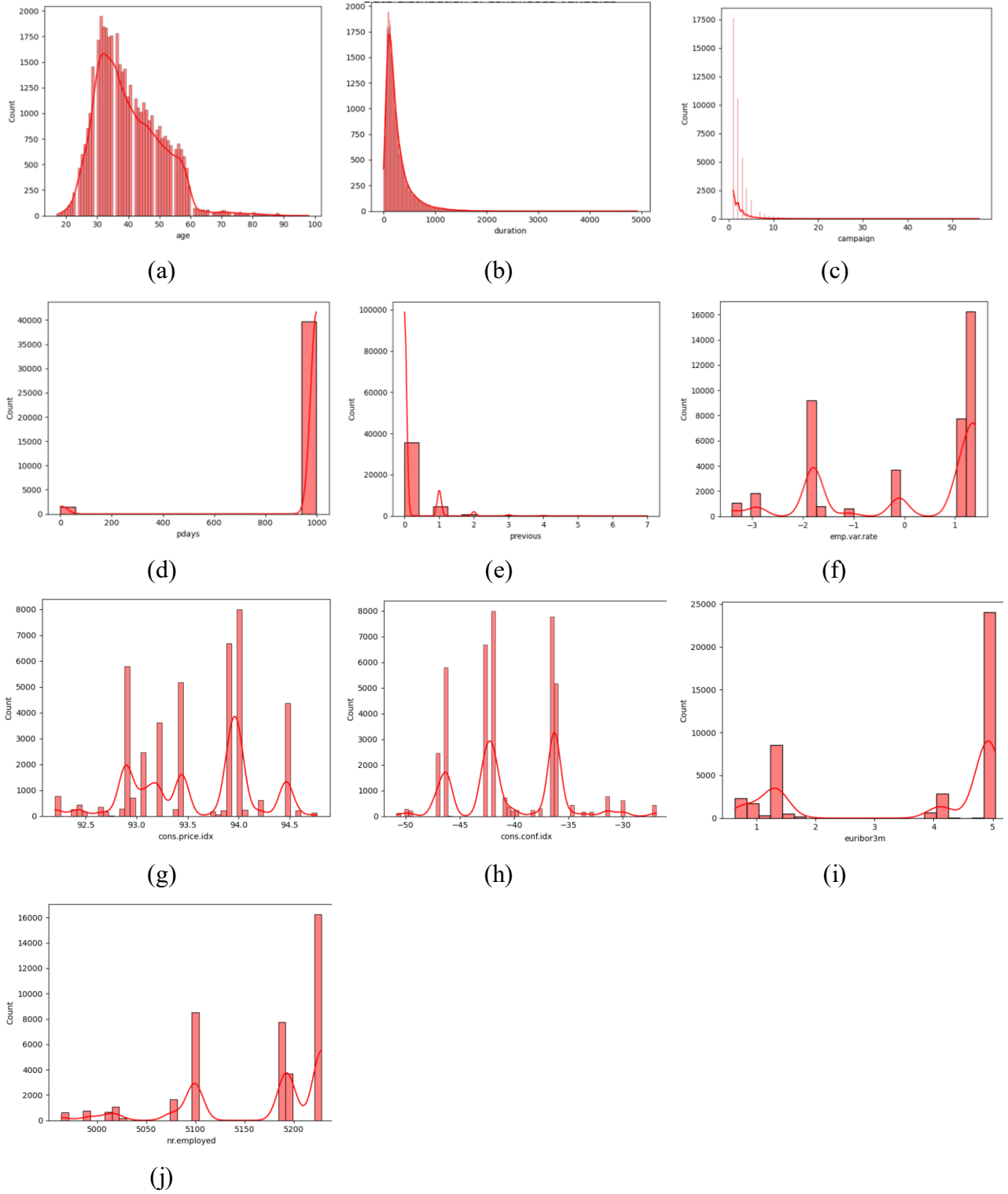(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Figure 2: Data distribution of continuous variables (Photo/Picture credit: Original).

The framework setup contains numerous phases, namely data collection, pre-planning, training, testing, and algorithm implementation. The intended outcome is the last phase. The unprocessed information that was collected may be insufficient or controversial [10]. Before employing the data for learning, it should go through the preparation stage to clean the data. Highlight extraction is a further stage in the training model's evolution. Additionally, the figures that follow could potentially be of assistance for both model selection and data processing. According to the result of the isnull() method, there are no missing values in the dataset for each attribute. In addition, Figure 2 visualizes the distribution of continuous variables, excluding column y. Moreover, Figure 3 provides a heatmap to visualize the correlations between various variables, which facilitates the discovery of hidden linkages and the understanding of correlations between variables.
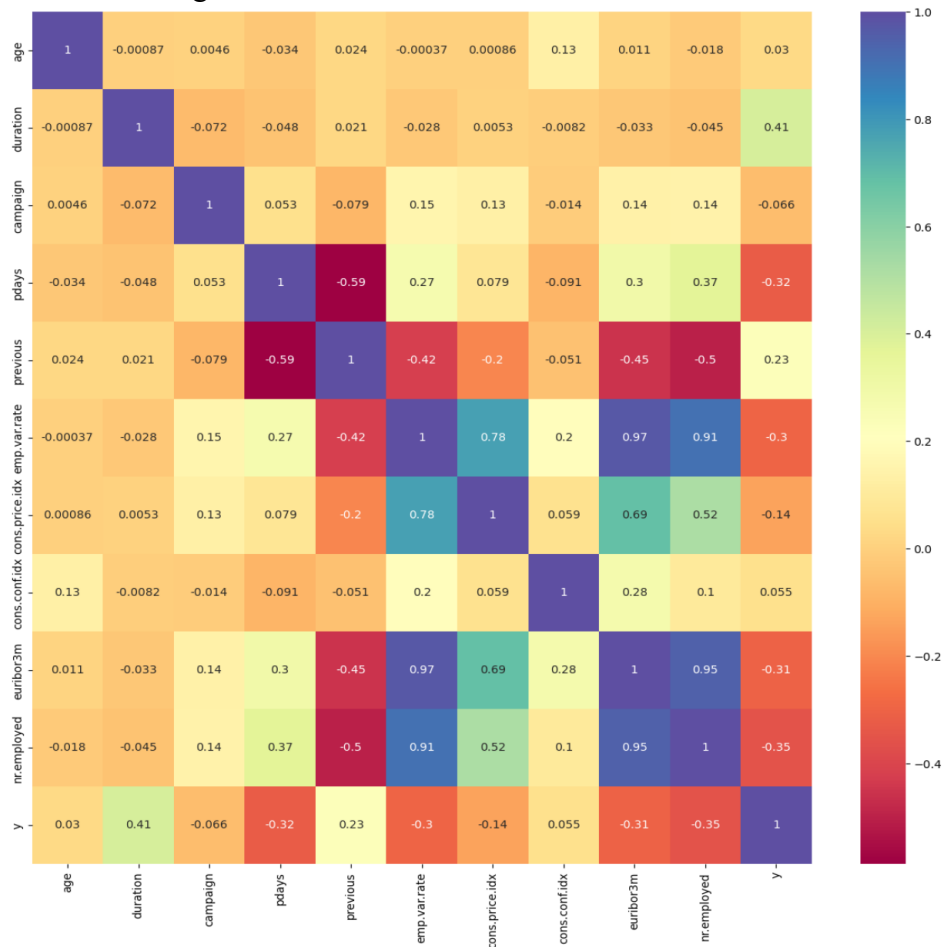

Figure 3: Correlation matrix (Photo/Picture credit: Original).

Several innovative measures have been employed to enhance machine learning methods' performance and prediction accuracy in the dataset. First of all, in order to avoid high cardinality and correlation issues, the 'duration' attribute has been removed with the highest amount of unique values. Apart from that, 'nr.employed' has been eliminated to avoid data redundancy and multicollinearity. The most highly correlated variables are shown in Figure 4.
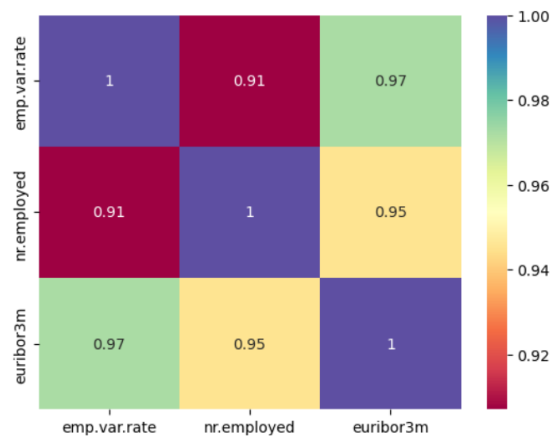
Figure 4: The most highly correlated variables (Photo/Picture credit: Original).

In addition, the interquartile range (IQR) method has been applied to outlier detection. The measures that have been taken are as follows: Firstly, quartiles for various variables and identified outliers have been computed. Secondly, with the purpose of avoiding multicollinearity, enhancing model performance, and improving feature engineering, dummy variables have been created. Additionally, data standardization has been employed to apply mean normalization to center the data around zero. Moreover, features have been scaled to have identical variance and improve algorithm convergence and model interpretability. Figure 5 visualizes the distribution of the target variable y to reveal the target variable y's distribution for aiding with model selection and to detect class imbalance.
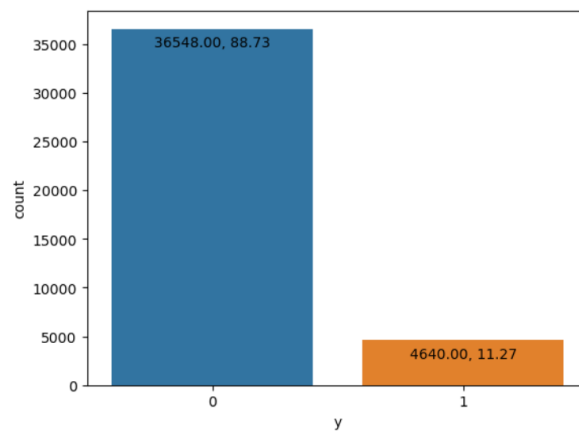

Figure 5: Distribution of the target variable y (Photo/Picture credit: Original).

3.3. Algorithm

In the era of AI, machine learning has significant applications. The paper addresses concerns about individuals signing up for term deposits and discusses binary classification issues. Data mining, using techniques like clustering, classification, association, prediction, and sequencing, extracts and identifies patterns from large datasets to generate valuable information not previously known, effectively drawing conclusions from the gathered data through machine learning techniques. A sophisticated classification technique has been employed with processed datasets in the realm of machine learning. KNN , SVM, DT, RF, XGB, GB, NB, and LR constitute some of the methodologies deployed.

4. Results

4.1. Primary Considerations for A Term Deposit

By generating a table that shows the distribution of categories within each variable's maximum and minimum value discrepancies. Following that, it arranges the table according to these distinctions, as seen in Table 1. For the 'poutcome', a count plot has been made using the biggest difference. Furthermore, by adding color to the target variable y to depict its distribution within each group, it becomes easier to identify the main variables that cause a term deposit.

This makes it possible to comprehend how the target variable, y, is distributed among the many 'poutcome' categories. Engineering and Feature Selection: Well-informed choices about feature selection and engineering can be made on the basis of the analysis. A term deposit in a bank is primarily considered, as shown in Figure 6.

Table 1: The maximum and minimum value differences within each variable.

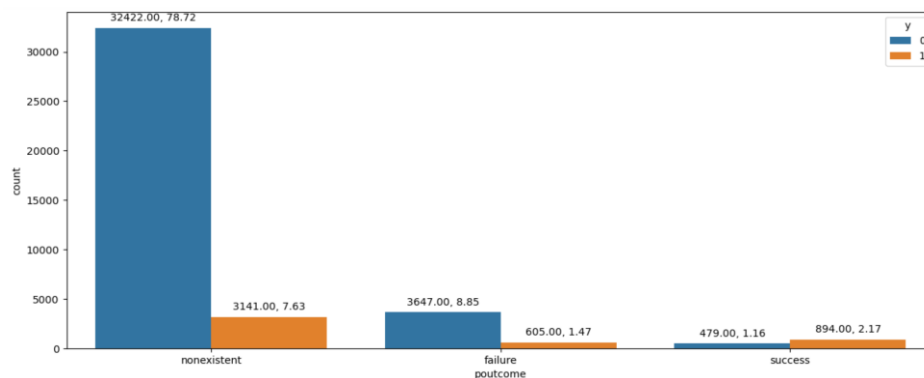| Variable	Min-Max Difference
Poutcome	34190
Loan	32960
Default	32585
Y	31908
Marital	24848
Housing	20586
Month	13587
Education	12150
Contact	11100
Job	10092
Day of week	796

Figure 6: Distribution of y within each category of 'poutcome' (Photo/Picture credit: Original).

4.2. Model Performance

Superior performance has been demonstrated by XGB, with a high F1 score and accuracy. Strong generalization capacity is indicated by cross-validation findings, which coincide with training results. In contrast to most other models, it ranks better in terms of prognosis for Class 1. Table 2 depicts the model performance for this particular model. In addition, Figure 7 displays the AUC for all the models above except the SVM. Table 1 creates a Pandas DataFrame containing multiple models and their corresponding scores, then sorts the DataFrame by the scores in descending order. As is shown in

Table 2, Pandas DataFrame contains model comparison information, sorts it by the Model F1-Score column in descending order, and then formats the DataFrame to display values as percentages highlighted with the maximum value in orange and the minimum value in green in each column separately, sets a caption, and formats the values as percentages.

As it can be seen in Table 3, XGB and GB perform best when proactively taking into consideration all aspects of the performance of all models, particularly accuracy, F1 score, cross-validation results, and generalization features. Both models perform well in cross-validation and have superb precision and F1 scores, implying strong generalization ability.

XGB fares marginally better overall and has a slight edge in the F1 score for Class 1 (0.51 vs. 0.49) when determining the most advantageous model. For such a reason, XGB is the most efficient selection with its superior accuracy, F1 score, and cross-validation results, together with its outstanding generalization capacity utilization and marginal advantage in predicting the minority class (Class 1).

Table 2: The performance of XGBoost.

	Precision	Recall	F1-score	Support
0	0.96	0.93	0.95	7332
1	0.43	0.60	0.51	668
Accuracy			0.90	8000
Macro avg.	0.70	0.77	0.73	8000
Weighted avg.	0.92	0.90	0.91	8000

(a)

(b)

(c)

(d)
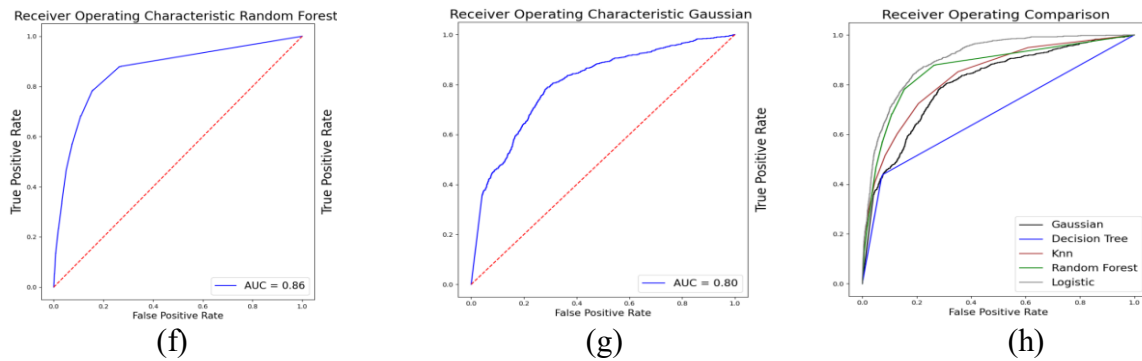

(e)

(f)   (g)   (h)
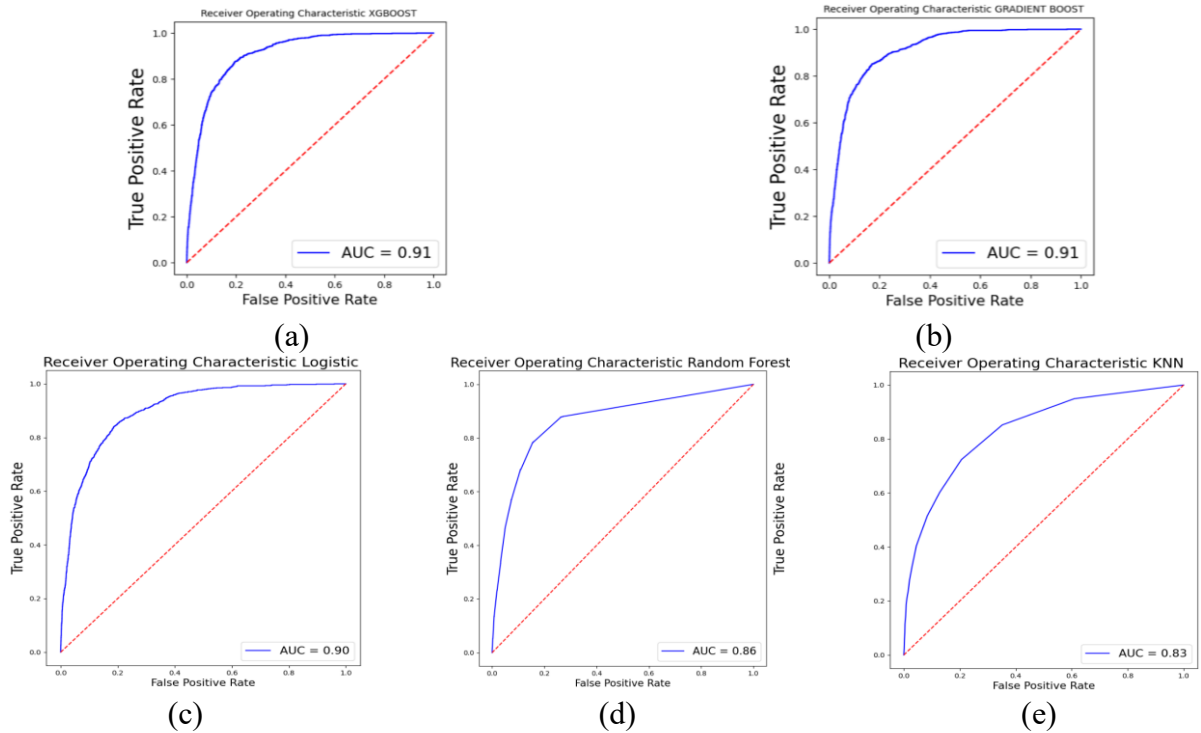
Figure 7: The AUC for all the models above except the SVM (Photo/Picture credit: Original).

Table 3: Sorted model comparison information.

	Model Accuracy	Model F1-score	CV Accuracy	CV std
KNN	89.86%	92.41%	89.82%	0.24%
Logistic Regression	90.18%	91.58%	34.53%	2.32%
XGBoost	90.14%	90.85%	90.56%	0.29%
Random Forest	89.45%	90.62%	89.85%	0.20%
Gradient Boosting	90.41%	89.44%	90.85%	0.40%
SVM	87.76%	88.18%	87.49%	0.37%
Decision Tree	87.08%	87.14%	44.38%	2.02%
Naive Bayes	66.50%	60.47%	66.12%	1.22%

5. Conclusion

Eight algorithms were primarily utilized in this study: SVM, NB, RF, KNN, LR, DT, XGB, and GB. XGB turned out to have the most outstanding performance amongst these eight algorithms, highlighting a high F1 score as well as acceptable accuracy. Cross-validation findings reveal strong generalization power and match training outcomes. It presents a better prognosis for Class 1 than the majority of other models.

Thus, with astute insight into speeding up the telemarketing campaign, the likelihood has surfaced to maintain a competitive edge over contemporary and potential suppliers of services in the banking sector. These outcomes might be worthwhile to service providers when researching telemarketing campaign techniques that will boost banks' earnings potential.

As a whole, the good F1 score and respectable accuracy of XGB could help service providers expedite telemarketing operations. A thorough comparison of machine learning models and analysis tailored to each class was carried out in order to fill in the gaps in the literature. In addition to offering fresh viewpoints and insights for managing bank marketing data, the study's methodological contribution and model selection optimization add to the overall caliber of research in this area.

As a guide for other researchers utilizing machine learning models in a variety of datasets and activities, the study offers empirical support for further investigation into how well XGB handles bank marketing data. Additionally, it provides useful advice on how to improve marketing plans and customer service in the banking industry. Nevertheless, the study has many shortcomings, including the need for more research on feature engineering and data preprocessing, the problem of class imbalance, and the absence of validation in a real-world setting. Prospective investigations have to

concentrate on refining feature engineering methodologies, streamlining data preprocessing, and improving model integration approaches.

To make sure that study findings are applicable to a larger population and are effective, real-world testing is also required. By streamlining model exploration and optimization, automated machine learning (AutoML) can increase research yield and performance while lowering the need for human intervention. By addressing these developments in bank marketing prediction research, business operations can benefit from more significant outcomes and useful solutions.

References

[1] Hagen, L., Uetake, K., Yang, N., Bollinger, B., Chaney, A. J. B., Dzyabura, D., Etkin, J., Goldfarb, A., Liu, L., Sudhir, K., Wang, Y., Wright, J. R., & Zhu, Y. (2020). How can machine learning aid behavioral marketing research? Marketing Letters, 31(4), 361–370.

[2] Mitić, V. (2019). Benefits of artificial intelligence and machine learning in marketing. In Sinteza 2019 - International Scientific Conference on Information Technology and Data Related Research (pp. 472-477).

[3] Borugadda, P., Nandru, P., & Madhavaiah, C. (2021). Predicting the success of bank telemarketing for selling long-term deposits: An application of machine learning algorithms. St. Theresa Journal of Humanities and Social Sciences, 7(1), 91-108.

[4] Umam, F. S. A., & R. A. (2021). Determinants of Mudharabah term deposit: A case of Indonesia Islamic banks. Journal of Economics Research and Social Sciences, 5(2), 167-180.

[5] Mari, A. (2019). The rise of machine learning in marketing: Goal, process, and benefit of AI-driven marketing. Swiss Cognitive.

[6] Gupta, A., & Gupta, G. (2018). Comparative study of random forest and neural network for prediction in direct marketing. In Advances in Intelligent Systems and Computing (Vol. 697, pp. 401-410). Springer.

[7] Zaki, A. M., Khodadadi, N., Lim, W. H., & Towfek, S. K. (2024). Predictive analytics and machine learning in direct marketing for anticipating bank term deposit subscriptions. American Journal of Business and Operations Research, 11(1), 79-88.

[8] Li, Z., Xu, Z., & Zhou, Y. (2024). Application of machine learning on client prediction in bank marketing. In Economic Management and Big Data Application: Proceedings of the 3rd International Conference (pp. 1064-1076).

[9] Tang, X., & Zhu, Y. (2024). Enhancing bank marketing strategies with ensemble learning: Empirical analysis. PLOS ONE, 19(1), e0294759.

[10] Koripi, M. (2021). A review on secure communications and wireless personal area networks (WPAN). Wutan Huatan Jisuan Jishu, XVII(VII), 168-174.