Long Term Deposit Subscription Prediction using Machine Learning

Nikhil Kumar Reddy Buddula
11545974
University of North Texas
NikhilKumarReddyBuddula@my.unt.edu

Vijay Raghav Reddy Chirra Reddigari
11434971
University of North Texas
vijayraghavreddychirrareddigari@my.unt.edu

Ragha Harshita Namana
11515169
University of North Texas
RaghaHarshitaNamana@my.unt.edu

Preethi Chowdary Bollipalli
11446366
University of North Texas
PreethiChowdaryBollipalli@my.unt.edu

I. Introduction

The banking sector employs two primary approaches for marketing initiatives: mass marketing and focused marketing. Mass marketing involves large-scale public marketing efforts, such as TV and print advertisements, that aim to reach a broad audience. However, this approach typically has a lower rate of favorable reactions from customers to buy a product or sign up for a service, which can be attributed to its lack of personalization and relevance to the individual customer [1].

On the other hand, focused marketing involves specialized promotions that are targeted at specific customer segments, such as individuals who have shown an interest in a particular financial product or service. This approach typically yields higher rates of success as the messages are tailored to the individual customer's needs and preferences.

Although mass marketing campaigns can be expensive, they can still contribute to the sale of the product. Direct marketing, which involves reaching out to customers directly, can also be effective in generating sales. For example, a salesperson may contact a potential customer via indirect telemarketing on their landline or mobile phone. However, finding prospective customers who fall within a specific category can be challenging, which is where statistical methods and tools come into play [2].

Marketing managers have started using data-driven decisions to identify potential customers for financial products and services. By analyzing customer data, such as purchase history and demographics, companies can determine who is most likely to invest in banks and avoid potential pitfalls. This approach enables marketers to create personalized messages that resonate with the customer, increasing the likelihood of generating a sale. In recent years, data-driven decisions have become increasingly popular as companies strive to stay ahead of the competition and provide the best possible customer experience [3].

Machine learning has been widely adopted by the banking industry in recent years to support their marketing efforts. Machine learning is an AI technique that allows machines to acquire knowledge and proficiency through experience rather than manual instruction. Machine learning algorithms can predict future customer behavior by examining massive data sets for patterns.

II. Problem Statement

The ongoing economic instability in a number of countries has compelled banks to sell more term deposits in order to replenish their currency reserves [4]. This has put tremendous pressure on marketing professionals to convince the general public to purchase term deposits. Marketing managers require an effective method for anticipating whether a consumer will sign up for a term deposit in order to optimize resources and increase sales. Therefore, we propose a project that employs supervised machine-learning algorithms to predict bank term deposit consumer sign-ups in order to address this issue.

III. PROJECT SCOPE AND OBJECTIVES

The goal of this project is to use customer profile data to train machine learning models that can reliably predict which customers are more likely to be interested in signing up for a term deposit than others. The purpose of this project is to compare the effectiveness of the following ensemble machine learning algorithms in classifying consumers according to the attributes provided in the dataset.

- Bagging Classifier
- AdaBoost Classifier
- Gradient Boosting Classifier
- Random Forest Classifier

The objective of this study will be achieved through three stages: customer profile research, model creation, and model assessment.

IV. Dataset

The data used in this analysis came from the UCI Machine Learning Repository, and it was all about the marketing campaigns run by different banks.To access the dataset publicly. (https://archive.ics.uci.edu/ml/datasets/bank+marketing).
This project makes use of data from a Portuguese financial

institution as its data source. The bank used its call center to conduct direct marketing campaigns for its term deposit offering and to reach out to potential new customers. This dataset offers a wealth of information on customer behavior and preferences, providing valuable insights into the bank's marketing strategy. By analyzing this dataset in detail, we can better understand the factors that drive customer behavior and use this knowledge to improve our marketing efforts. The insights gained from this analysis will enable us to develop effective predictive models and tailor our marketing campaigns to maximize their impact.

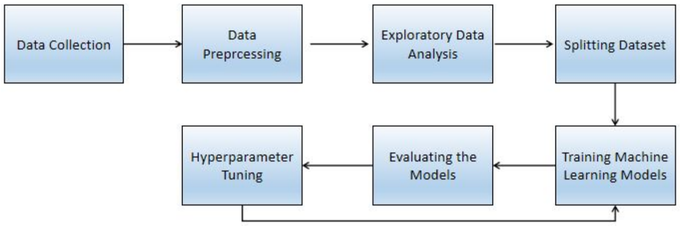

Fig. 1. Working flow of this project.

A. *Features in Dataset*

1) Age-numeric form
2) Job Description (A categorical factor)
3) Relationship Status (categorical)
4) Education (general)
5) Do you have a mortgage? (categorical)
6) Do you have a personal loan? (categorical)
7) Kind of contact communication (categorical)
8) The year's final contact month (categorical)
9) Final day of the week for contact (categorical)
10) Duration of the last touch in seconds-numeric
11) Promotion
12) how many days have passed since the client was last contacted for a marketing campaign?
13) Number of contacts made before this campaign
14) Previous marketing campaign outcome (category)
15) Rate of fluctuation in employment (numeric)
16) Consumer Price Index (numeric)
17) Index of consumer confidence (numeric)
18) 3-month Euribor rate (numeric)
19) Total number of workers (numeric)
20) Variable output (desired outcome)

V. RELATED WORK

In this section, we will mostly discuss prior studies that have dealt with issues similar to our own. The work of S. Palaniappan, A. Mustapha, and coworkers [5] is particularly noteworthy since it uses a number of algorithms, including Decision Tree and Rough Set Theory, to develop principles that may be used to predict whether or not a consumer will sign up for a bank term deposit. In a related study [6], M. A. T. Rony, M. M. Hassan, E. Ahmed, A. Karim, S. Azam, and D. A. Rez used Machine Learning models like Logistic Regression, Random Forest, Support Vector Machine, and K-nearest neighbors to foretell whether a customer would choose to make a bank term deposit.

In their analysis, the logistic regression model yielded the best results (90.64 percent accuracy). The research conducted by S.Hou, Z. Cai, J. Wu, H. Du, and P. Xie [7] is yet another significant advancement in this field. These studies show that using Machine Learning models to predict consumer behavior and pinpoint areas where marketing strategies and customer engagement may be strengthened is a sound business practice. Our studies hope to expand upon these findings by making use of new methodological advances and assessing alternative models for higher precision and more trustworthy outcomes. We believe this will be a major step forward in our ability to develop marketing campaigns that better address the demands of our target demographic and anticipate their decisions to make term deposits at our bank.

VI. HYPOTHESIS

A. *Hypothesis 1:*

A customer's age and occupation are the most significant factors in predicting whether they will subscribe to a term deposit. This is because older customers with stable occupations may have more savings and be more interested in long-term investments.

B. *Hypothesis 2:*

Predicting whether a customer would sign up for a term deposit is most difficult without knowing how many times they have been contacted during a marketing campaign and how they have responded in the past. This is because customers who have been contacted multiple times and have shown interest in previous campaigns may be more likely to subscribe in the future.

VII. RESEARCH DESIGN AND METHODOLOGY

These are the steps that we followed to train the machine learning models:

1) Exploratory Data Analysis(EDA)
2) Pre-processing of data
3) Splitting the data
4) Training and testing the proposed models.

A. *Exploratory Data Analysis(EDA)*

In our project, we conducted an extensive exploratory data analysis to gain a comprehensive understanding of the dataset we were working with. The dataset contained a total of 41,188 examples and 20 features. To gain more insights into the dataset, we performed various types of analyses, including univariate and bivariate analyses.

We looked at how the various categories were spread throughout the dataset. We identified two groups of customers: those who would not sign up for the term deposit scheme and those who may. Figure 2 displays the computed percentages of each category, showing that 88.7% of the cases fell into the former category and 11.3% fell into the latter. Since there are more customers who are unlikely to be enrolled to the term deposit program than others who could be interested, this suggests that the dataset is unbalanced.

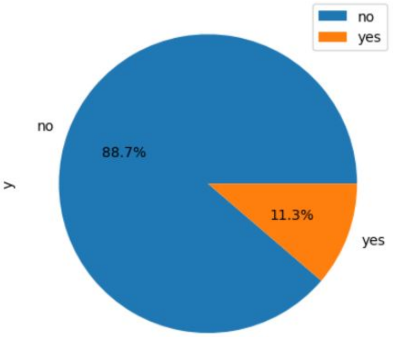

Fig. 2. Percentage of Each Class in the Dataset.

B. Research Questions

1) How do customers' backgrounds and levels of education affect how well marketing efforts work?: Figure 3 shows the Customers Subscription rate with different jobs. From the figure, students have the highest rate of subscription rate of 31.43%. Retired people have also a high subscription rate of 25.23%. Blue Collar people have the lowest subscription rate of 6.89%. Figure 4 shows the rate of subscription based on education. As we can see illiterate people have the highest rate of subscription deposits. It means people with no educational background are more likely to sign a term deposit.

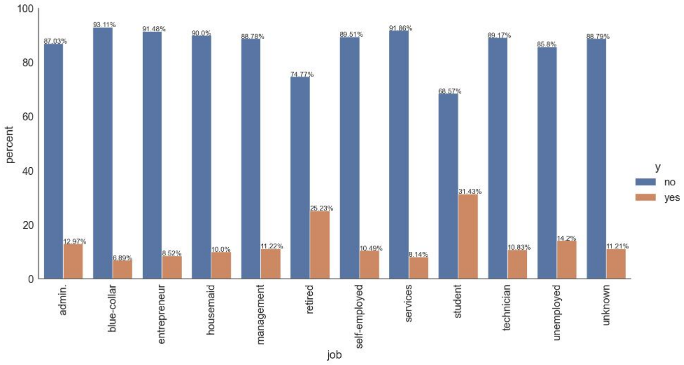

Fig. 3. Customer Subscription rate with different background

2) Do marketing efforts that run at different times of the year or in different seasons have different success rates?: Figure 5 shows the Customer's Subscription rate in a different month of the year. From the figure, in March, December, September, and October rate of a subscription a term deposit is high as compared to other months. May and November have the lowest subscription rate.

3) Is there a correlation between marital status and the performance of marketing campaigns?: Figure 6 shows the Customer's Subscription rate based on their marital status. From the figure, we can see that there is not a big difference in customer subscription rates based on marital status.

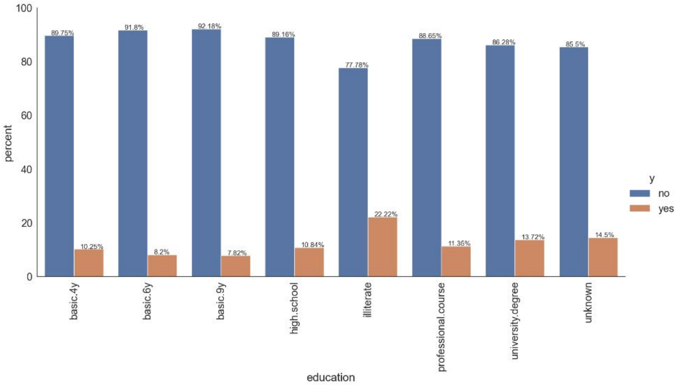

Fig. 4. Customer Subscription rate with different Education

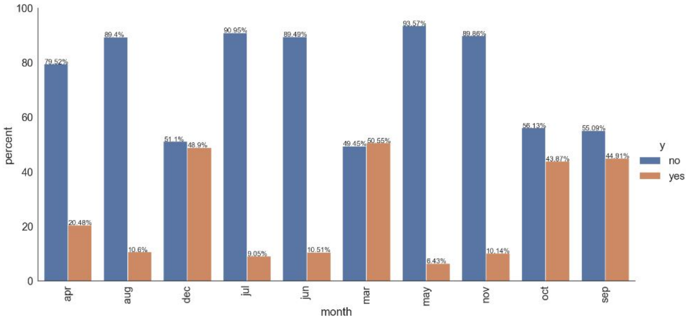

Fig. 5. Customer Subscription Rate in Different Months

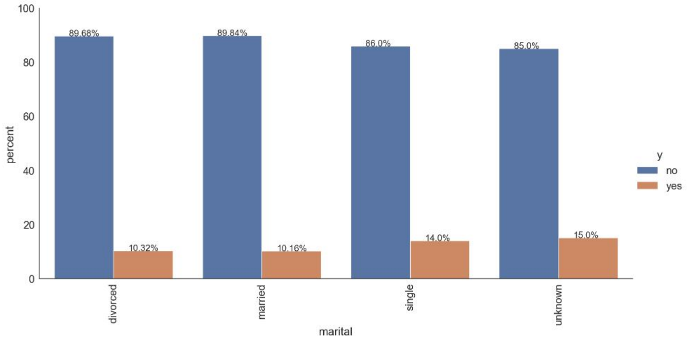

Fig. 6. Customer Subscription Rate based on their marital status

VIII. Data Preprocessing

Pre-processing is a crucial stage in preparing a dataset for machine/deep learning. Before training any models, it is necessary to ensure that the data is clean, consistent, and normalized. In this stage, we perform various operations to eliminate NAN values, check for missing values, and normalize our features.

Firstly, we eliminate NAN values, which are essentially missing values that could negatively impact the performance of our models. We either drop these values or impute them using various techniques such as mean, median, or mode imputation.

Secondly, we check for missing values and replace them with appropriate values based on the context of the data. We may also choose to drop certain features if they contain too many missing values or if they are deemed irrelevant to the problem at hand.

Finally, we normalize our features to make sure they are all of similar relevance in the model and have the same scale. To avoid making inaccurate forecasts, normalization is crucial since some features may have a wider range of values than others. Minmax scaling and z-score scaling are two frequent normalizing methods.

By performing these pre-processing steps, we ensure that the data is suitable for training and testing our machine/deep learning models, leading to more accurate and reliable predictions.

A. *Features Extraction and Selection*

When it comes to boosting a model's efficiency in machine learning, feature extraction and selection are crucial processes. To make it possible for the model to produce predictions, it must first take into account the features of the data. We used the **Chi-Square Test** to determine which of our dataset's 20 attributes were most significant for our research. When working with categorical data, the chi-square test is an invaluable tool. It determines if two categorical variables are significantly related to one another.

For each of the two variables being compared, this test computes the actual count **O** and anticipated count **E**. The chi-square measure, which exhibits a particular distribution, is then used to determine the difference between the expected count **E** and the real count **O**. Therefore, it is feasible to determine which features have a meaningful relationship with the goal variable by using the chi-square test on the remaining features after removing overfitted ones. Then, these features can be used to create a more precise forecast model while any unnecessary features can be eliminated to save time and money.

$$X_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

where

- c degree of freedom
- O is observed values
- E is the expected values

Based on the importance of each feature in the dataset, this test gave us a p-value for that feature. Commonly, a p-value threshold of less than 0.05 is used to rule out the null hypothesis, which suggests that there is a connection between the feature and the objective variable.

1) Null Hypothesis: There is no significant relationship between the customer's attributes and their decision to subscribe to a term deposit if p_value is greater than 0.05.

2) Alternative Hypothesis: There is a significant relationship between the customer's attributes and their decision to subscribe to a term deposit if p_value is less than 0.05.

p_value for **month, age, day_of_week** and **loan** is greater than 0.05 as shown in table I so we will remove these features

TABLE I
Chi-Square test result.

Feature	p-value
age	0.13686
job	0.00401
marital	0.00231
education	0.00000
default	0.00000
housing	0.11277
loan	0.37187
contact	0.00000
month	0.64436
day_of_week	0.10971
duration	0.00000
campaign	0.00006
pdays	0.00000
previous	0.00000
poutcome	0.00000
emp.var.rate	0.00000
cons.price.idx	0.00000
cons.conf.idx	0.00104
euribor3m	0.00000
nr.employed	0.00000

IX. Training and Evaluation of Machine Learning models

In this project, we hope to find the best machine learning model for predicting whether or not a consumer would sign up for a term deposit by training it on our dataset. In order to choose the most effective model, we want to compare their performance measures.

We will use **Optuna** for hyperparameter tuning, which is the process of modifying our models' parameters for maximum precision, to guarantee their peak performance. With this method, we can optimize our models for optimal performance.

We plan to use several ensemble models in this study, which include but are not limited to Random Forest, Gradient Boosting, and AdaBoost. Ensemble models are a combination of multiple machine learning models, which are trained together to provide a more robust and accurate prediction. By using these models, we can identify the most important features that affect customer behavior and create effective marketing strategies accordingly.

A. *Bagging Classifier*

A sort of ensemble machine learning technique known as the bagging classifier combines predictions from various separate decision tree models to increase the precision of the total prediction.

Several training datasets are produced using the bagging (short for bootstrap aggregating) technique by randomly selecting sample of the original training data and replacing it with new data. An individual decision tree model is then trained using each training dataset. The final forecast is calculated by averaging the predictions produced by each individual tree (for regression problems) or by a majority vote (for classification problems). Several parameters of scikit-learns' Bagging Classifier can be adjusted to improve the model's predictive accuracy. Some key factors are listed below:

- **base_estimator:** This is the decision tree model used as the base estimator for the bagging classifier. The default is DecisionTreeClassifier, but other models can also be used.
- **n_estimators:** The number of decision trees to use in the ensemble, denoted by n estimators. Adding more estimators to a model can enhance its accuracy, but doing so may increase the time and space needed to train the model.
- **max_samples:** The maximum amount of training samples for each particular decision tree. It might be an integer (like 100 for 100 samples) or a fraction (like 0.5 for 50% of the data). Overfitting can be mitigated by setting this parameter to a value below 1.0.
- **max_features:** In deciding how to divide each node of the decision tree, this is the maximum amount of features to take into account.It can be given as an integer (e.g., 10) or a fraction (e.g., 0.5) of the total number of characteristics. If you set this value lower than the total number of features, you can minimize the effect of irrelevant or noisy characteristics.
- **bootstrap:** This parameter controls whether or not to use bootstrapping when creating the individual training datasets. If set to True (the default), each training dataset will be created by randomly sampling the data with replacement. If set to False, each training dataset will be a subset of the original data without replacement.

B. AdaBoost Classifier

AdaBoost (Adaptive Boosting) Classifier is an ensemble learning approach that combines multiple ”weak” learners into a single ”strong” one. A model is said to be ”weak” if it just marginally improves upon the results of random guessing. The core idea behind AdaBoost is to train many, relatively weak classifiers in parallel on weighted versions of the training data, with each subsequent classifier providing greater weight to the samples that its predecessors mistakenly labeled. During each iteration, the algorithm finds the misclassified samples and gives them more weight. After a certain threshold has been reached, the process is reset and a new, less robust classifier is trained using the updated dataset.

AdaBoost Classifier in scikit-learn:

- **base_estimator**
- **n_estimators**
- **learning_rate**
- **algorithm**
- **random_state**

C. Gradient Boosting Classifier

The Gradient Boosting Classifier is a sort of ensemble learning that takes several ”weak” learners and merges them into a single ”strong” learner. It's a widely used and highly effective algorithm for solving classification and regression issues. Unlike AdaBoost, which assigns higher weights to misclassified samples, Gradient Boosting focuses on minimizing the loss function directly by iteratively adding trees that correct the errors of the previous model. This results in a more complex and powerful model that can capture complex interactions and nonlinearities in the data.

D. Random Forest Classifier

For classification and regression jobs, a well-liked machine learning method called random forest is used. The term ”random forest” refers to an ensemble learning technique that mixes various decision trees to form a ”forest” of trees.

When using random subsets of the training data to create numerous decision trees, random forest aggregates each tree's forecasts to produce a final estimate. At each split, attributes from a separate randomly selected portion of the training data are considered as potential split candidates, and each decision tree is formed using a new subset of the data. By lowering the likelihood of overfitting, this randomization strengthens the model's accuracy and stability. The most important parameters for a random forest model are:

- **n_estimators**: The number of trees in the forest.
- **max_depth**: The maximum depth of each tree.
- **max_features**: The maximum number of features to consider at each split.
- **min_samples_split**: The minimum number of samples required to split an internal node.
- **min_samples_leaf**: The minimum number of samples required to be at a leaf node.
- **random_state**: Controls the random number generator used for fitting the model.

E. Hyperparameter Tuning

To discover the best hyperparameters for a specific machine learning model, use the well-known open-source **Optuna** hyperparameter optimization tool. To quickly explore the hyperparameter space and identify the ideal set of hyperparameters. Figures 7, 8, 9 10 shows hyperparameter tuning results of each ensemble model.

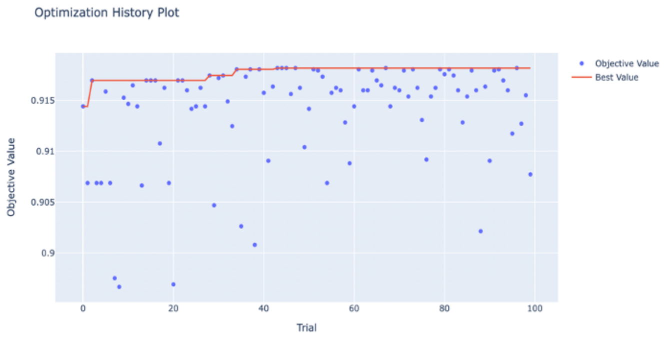

Fig. 7. Hyperparameter Tuning of Bagging Classifier

F. Oversampling

In order to provide our models a comprehensive grasp of each class in our dataset, we oversampled to ensure that there were roughly equal numbers of examples in each class.

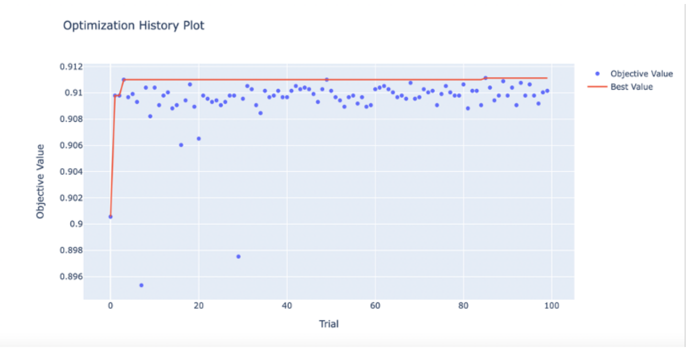

Fig. 8. Hyperparameter Tuning of AdaBoost Classifier

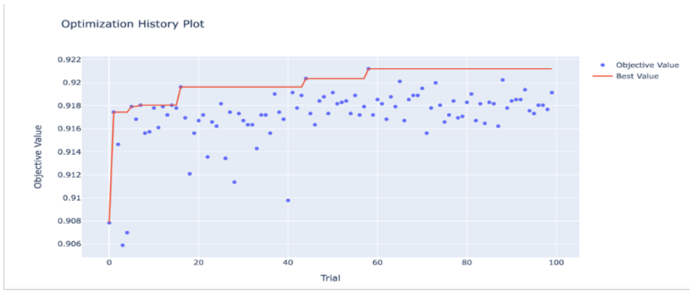

Fig. 9. Hyperparameter Tuning of Gradient Boosting Classifier

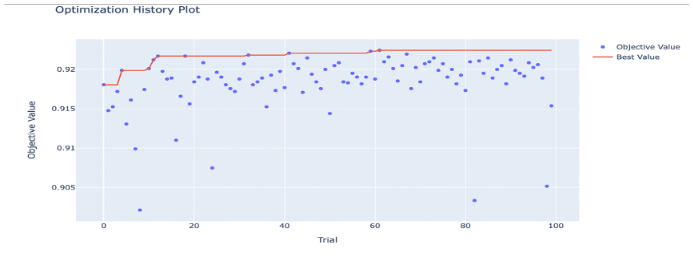

Fig. 10. Hyperparameter Tuning of Random Forest Classifier

G. Model Evaluation

To test how well our model performs, we apply a classification on the remaining 20% of data that serves as validation. The models' classification report is displayed in Table II. Performance in classification problems at different thresholds can also be measured with the use of the AUC - ROC curve. The area under the ROC curve (AUC) is a measure of how well the two categories can be separated. It demonstrates the model's ability to discriminate between groups. The AUC measures how well a model can separate subscribers from non-subscribers, hence a higher AUC indicates better accuracy. Figures 11,12,13 and 14 shows the ROC curve of the model. From the classification report, we can see that all the models performed very well, with AdaBoost Classifier giving the least accuracy of 91% and 0.97 AUC. Bagging Classifier outperformed the other models by a slight margin as it gives AUC=0.99 and accuracy of 94% while Gradient Boosting and Random Forest Classifier give 0.98 AUC and accuracy of 93%.

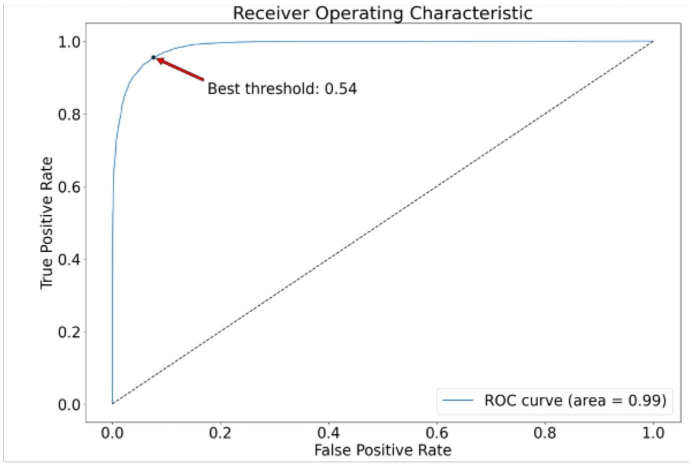

Fig. 11. ROC-AUC Curve of Bagging Classifier

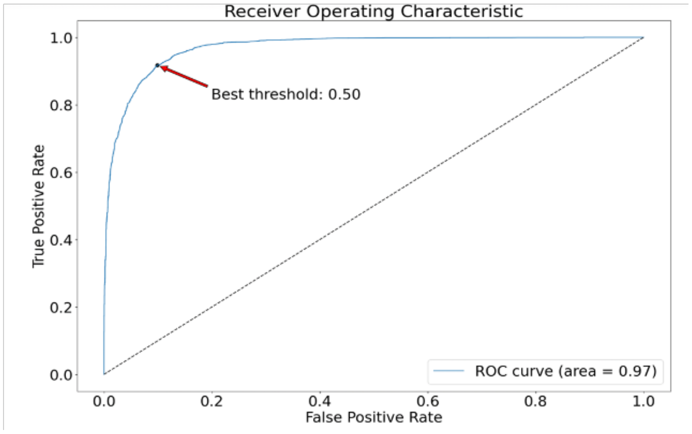

Fig. 12. ROC-AUC Curve of AdaBoost Classifier

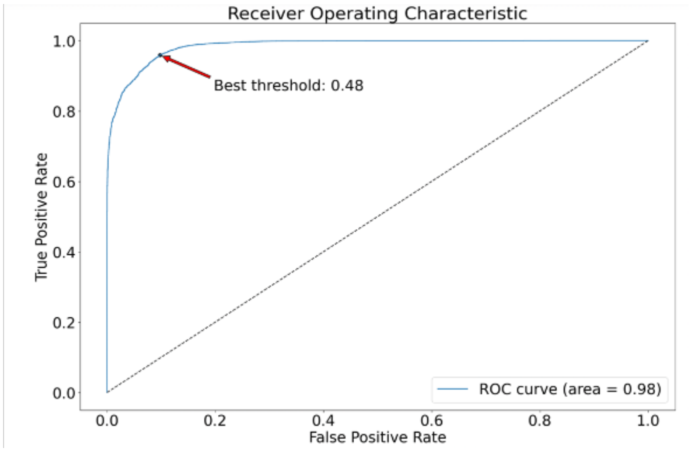

Fig. 13. ROC-AUC Curve of Gradient Boosting Classifier

X. Conclusion

In conclusion, this project has provided valuable insights into the process of exploratory data analysis, data cleaning, and hyperparameter tuning for machine learning models. By accurately predicting customer behavior, marketing strategies

TABLE II
CLASSIFICATION REPORT OF MODELS

	Precision	**recall**	**F1-score**	**support**		**Precision**	**recall**	**F1-score**	**support**
Bagging Classifier					**AdaBoost Classifier**				
No	96%	91%	94%	7264	**No**	91%	90%	91%	7264
Yes	92%	96%	94%	7351	**Yes**	91%	91%	91%	7351
accuracy			94%	14615	accuracy			91%	14615
macro avg	94%	94%	94%	14615	macro avg	91%	91%	91%	14615
weighted avg	94%	94%	94%	14615	weighted avg	91%	91%	91%	14615
Gradient Boosting Classifier					**Random Forest Classifier**				
No	95%	91%	93%	7264	**No**	97%	89%	93%	7264
Yes	91%	96%	93%	7351	**Yes**	90%	97%	93%	7351
accuracy			93%	14615	accuracy			93%	14615
macro avg	93%	93%	93%	14615	macro avg	93%	93%	93%	14615
weighted avg	93%	93%	93%	14615	weighted avg	93%	93%	93%	14615

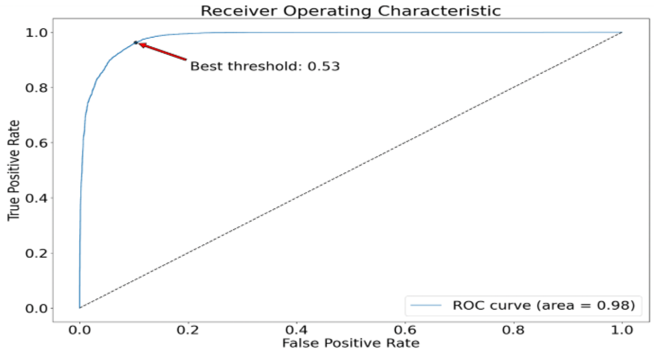

Fig. 14. ROC-AUC Curve of Random Forest Classifier

can be tailored to maximize their impact and improve the bank's bottom line. The best-performing model in this study was the Bagging Classifier, which achieved an accuracy of 94% in classifying potential customers who are more likely to subscribe to a bank term deposit. In this study, all the models have outperformed the previous benchmark accuracy of 90.64% and demonstrated the effectiveness of employing machine learning models in predicting customers' behavior.

REFERENCES

[1] S. Moro, R. Laureano, and P. Cortez, "Using data mining for bank direct marketing: An application of the crisp-dm methodology," 2011.

[2] C. S. T. Koumétio, W. Cherif, and S. Hassan, "Optimizing the prediction of telemarketing target calls by a classification technique," in *2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM)*. IEEE, 2018, pp. 1–6.

[3] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," in *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*. IEEE, 2017, pp. 1–4.

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

[5] S. Palaniappan, A. Mustapha, C. F. M. Foozy, and R. Atan, "Customer profiling using classification approach for bank telemarketing," *JOIV: International Journal on Informatics Visualization*, vol. 1, no. 4-2, pp. 214–217, 2017.

[6] M. A. T. Rony, M. M. Hassan, E. Ahmed, A. Karim, S. Azam, and D. A. Reza, "Identifying long-term deposit customers: A machine learning approach," in *2021 2nd International Informatics and Software Engineering Conference (IISEC)*. IEEE, 2021, pp. 1–6.

[7] S. Hou, Z. Cai, J. Wu, H. Du, and P. Xie, "Applying machine learning to the development of prediction models for bank deposit subscription," *International Journal of Business Analytics (IJBAN)*, vol. 9, no. 1, pp. 1–14, 2022.