PREDICTING CUSTOMER SUBSCRIPTION TO TERM DEPOSITS USING MACHINE LEARNING: A RANDOM FOREST APPROACH

Dr.Naresh Dembla, Ravindra Yadav
Assistant Professor, IIPS, DAVV
nareshdembla@gmail.com, rydav@ietdavv.edu.in

Abstract

Customer retention and engagement are critical for banks offering term deposits. This study develops a machine learning-based predictive model to identify the likelihood of customer subscription to term deposits based on demographic, financial, and campaign data. Using a Random Forest Classifier, the model achieved an accuracy of 91% and a ROC-AUC score of 0.94. The methodology incorporates comprehensive data preprocessing, feature engineering, and model evaluation to ensure robust predictions. The insights derived can assist banks in optimizing marketing strategies and resource allocation.

Keywords customer retention, customer engagement, term deposits, machine learning

Introduction

The banking industry relies on personalized marketing to drive customer engagement and subscription to financial products such as term deposits.Understnding customer behavior and predicting subscription like hood is vital for effective campion management .traditional methods often fail to leverage the wealth of the data available ,limiting the accuracy of the predictions .Machine learning (ML) techniques offer a powerful alternative ,enabling the analysis of large dataset and uncovering hidden patterns .n this study we aim to develop a predictive model for term deposited subscription using a demographic ,financial and camping related features .The proposed model identifies influent factors contributing to customer decision, assisting bank in targeting high potential customers decisions ,assisting bank in targeting high potential customers and improving market marketing efficiency.

Related study Estimating the explainable machine learning model using actual data and testing numerous machine learning model using test data are the objective of the study. Based on the findings the Boost model performed better than other machine learning techniques in identifying churn clients and offering guidance on what aspects should be considered while managing existing ones

This study uses Shapely Additive explanations (SHAP) values to support the evaluation and interpretability of machine learning models for customer churn analysis, especially explainable models, even though the literature uses a variety of models for this purpose [1]. Consider Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Linear Discriminant Analysis (LDA) as examples of classification models, and compare their accuracy and performance. The model with the highest accuracy is the best classifier. Random Forest is used practically in this work to choose the key classification feature. The comparative analysis of the RF method from many viewpoints is evident from our trials [2]. The experiment's findings for the selected datasets demonstrate that decision tree-based models perform better. To evaluate the performance of the aforementioned classifiers in the particular context of credit rating, we add a measure of accuracy based on notches termed "Notch Distance" to the standard accuracy metric of classifiers. This metric indicates the degree to which the forecasts deviate from the actual ratings[3].This study is significant for a number of reasons, including: (i) the potential of Type-2 Fuzzy Logic as a highly flexible, effective, and explicable AI technique is not well understood; (ii) there is a lack of cross-disciplinary knowledge between financial services and AI expertise, and this work attempts to close that gap; (iii) regulatory thinking is changing with little guidance globally, and this work aims to support that thinking; and (iv) it is crucial that banks maintain market stability and customer trust as the use of AI grows[4].This study looked at how consumer loyalty is impacted by how consumers view banks' marketing communication methods. A survey questionnaire was used to collect 313 valid responses from bank clients in Nigeria. The measurement model and study hypotheses were Examined using the partial least squares structural equation modelling (PLS-SEM) process. The findings show that while public relations, sales promotion, advertising, and personal selling are important components of bank marketing communication that predict client loyalty, direct marketing is not[5]

Proposed Methodology

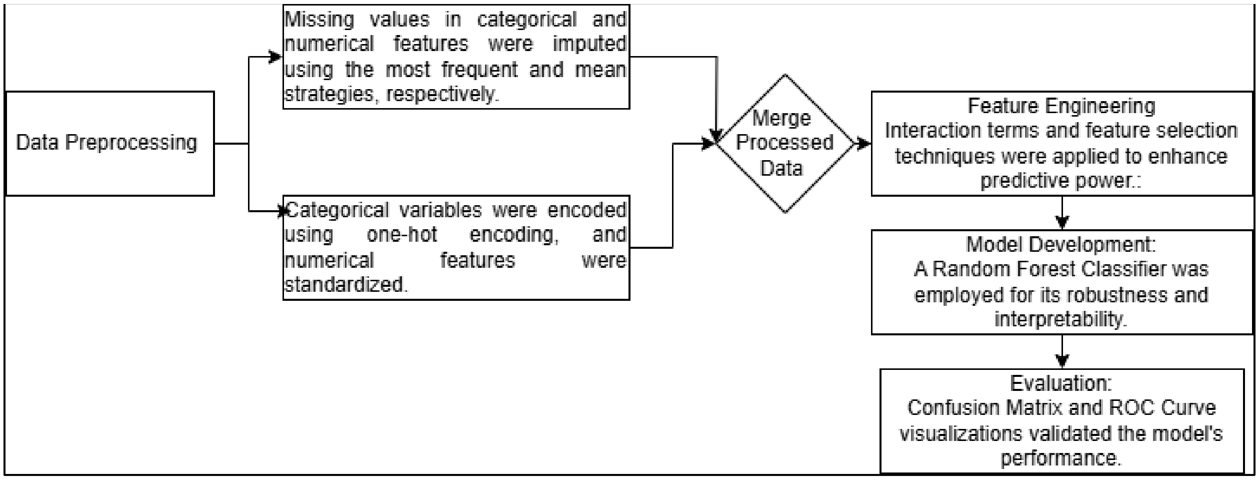
Data Preprocessing
Missing values in categorical and numerical features were imputed using the most frequent and mean strategies, respectively.
Categorical variables were encoded using one-hot encoding, and numerical features were standardized.
Merge Processed Data
Feature Engineering
Interaction terms and feature selection techniques were applied to enhance predictive power.:
Model Development:
A Random Forest Classifier was employed for its robustness and interpretability.
Evaluation:
Confusion Matrix and ROC Curve visualizations validated the model's performance.

Figure 1: Proposed Methodology

Dataset

A financial institution must examine the trends of its past marketing campaign in order to increase the efficacy of its next initiatives. This will allow us to determine the most effective tactics to use in order to increase the success of subsequent campaigns. There are 39188 samples and 21 characteristics in this dataset.

One Hot Encoding

In order to complete the incomplete dataset, missing values are first addressed using the suggested missing-data preprocessing procedure. The finished dataset is then subjected to the classification and regression tree (CART) model in order to assess the effectiveness of various preprocessing techniques. According to the experimental findings, a high missing rate is optimal for the suggested one-hot encoding technique. The random sample (RS) imputation approach outperforms the other imputation methods in this study when the missing rate is low, but it comes at a higher computing cost[6]

Using one-hot encoding, binary matrix representations are created from category information. For a specific category characteristic. With k distinct categories, x, one-hot encoding generates k variables that are binary.

Let x be the categorical variable with the following categories: {C 1,C 2,…,C k }, xi be the i th observation of x .The definition of the one-hot encoding $O(xi)$ of xi is as follows: $O(xi)$=[oi 1, oi 2,…, oik], where:oij = {1 if xi = Cj, 0else.o ij = 0.

Decision Tree

A decision tree uses feature-value-based decision rules to divide data into subgroups. The structure is depicted as a tree where: Every leaf denotes a class label or regression value, every internal node denotes a decision rule, and every branch denotes the result of a rule[7].

The splitting criterion, one of the fundamental components, establishes how to divide a node. Measures such as entropy or Gini impurity are used for classification. Variance reduction or mean squared error (MSE) are utilized for regression.

$$G=1-\sum_{i=1}^{k} p_i^2 \qquad (1)$$

where the percentage of class I in the subset is denoted by p i.A Decision Rule Represented Mathematically: Every node: The split rules are as follows: x j ≤t, where: The threshold value is denoted by t, while the j-th feature is represented by x j.The D dataset is separated into:

$$D\ \text{right} = \{x \in D | xj > t\}. \qquad (2)$$

$$D\ \text{right} = \{x \in D | xj < t\}. \qquad (3)$$

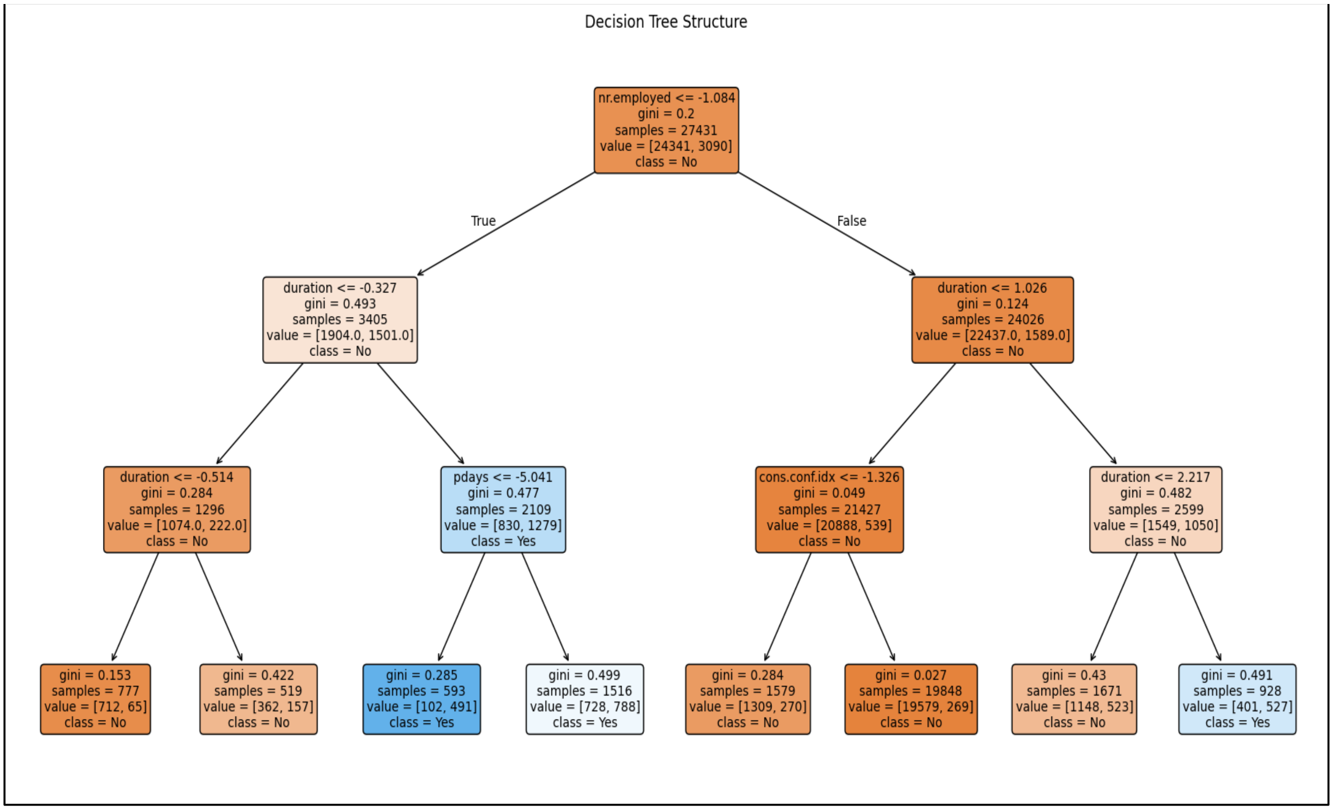

Figure :Decision tree

Results

The model's prediction accuracy for each class is shown by the confusion matrix (0: No subscription, 1: Subscription).The most significant features in forecasting client subscription are highlighted in the Feature

Importance Plot. With a high AUC value of 0.94, the ROC curve illustrates the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate).

Class 1 (subscribed) precision: 67%
Class 1 recall: 46%
91% overall accuracy
F1-score weighted: 91%
Strong discriminative ability is indicated by the ROC-AUC score of 0.94.

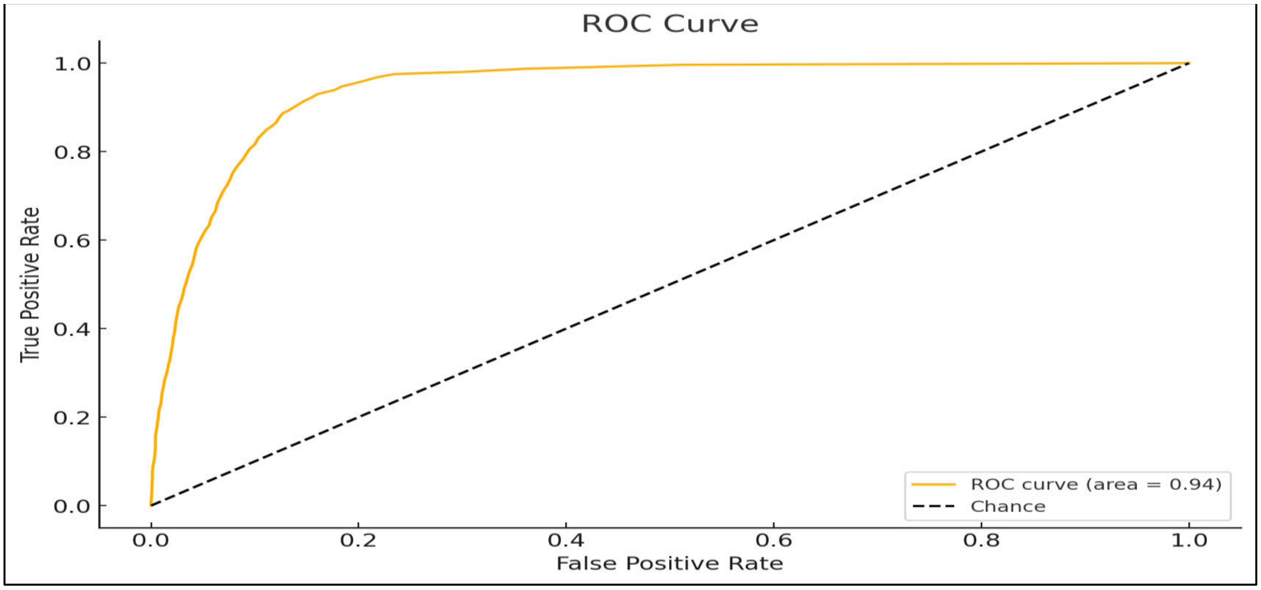

Figure 2: ROC curve area

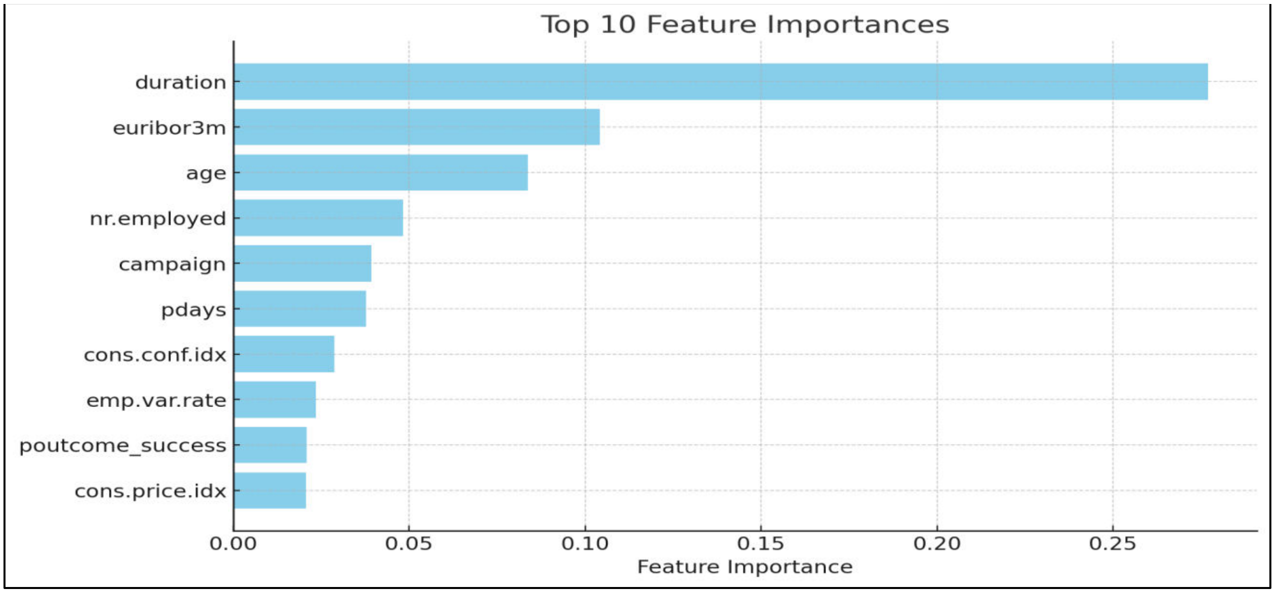

Figure 3: Feature Importance

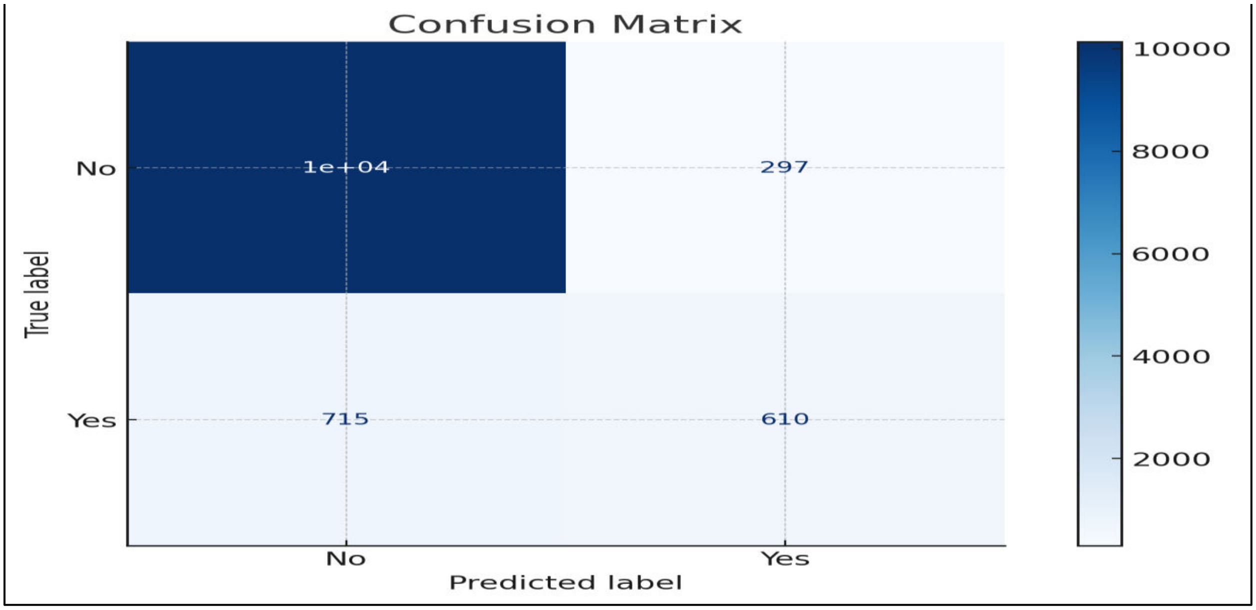

Figure 4: Confusion matrix

Conclusion

This study demonstrate how term deposited customer subscription may be predicted using machine learning ,more especially the random forest classifier .The models capacity to effectively differentiate between likely subscribers and non-subscriber is demonstrated by its high accuracy of 91% and solid ROC-AUC score of 0.94 ,which is achieved by using demographic ,financial and marketing data >by using the data from this model to better manage resources ,optimize marketing efforts and target the appropriate customer s at the right time bank can increase customer engagement.

Future work

Future research could examine the addition of more comprehensive or external data sources such as social media activity, customer transaction history, or macroeconomics factors, in orders to increase the predicted accuracy of the model.

Model comparison and optimization: Despite the Random forest classifier's promising result more investigation into other machine learning algorithm such as XGBOOST, LIGHTGBM or Neural Networking produce superior results

References

[1] H. Guliyev And F. Yerdelen Tatoğlu, "Customer churn analysis in banking sector: Evidence from explainable machine learning model," Journal of Applied Microeconometrics , vol.1, no.2, pp.85-99, 2021.

[2] Chen, RC., Dewi, C., Huang, SW. *et al.* Selecting critical features for data classification based on machine learning methods. *J Big Data* **7**, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4.

[3] Parisa Golbayani, Ionuţ Florescu, Rupak Chatterjee,A comparative study of forecasting corporate credit ratings using neural networks, support vector machines, and decision trees,The North American Journal of Economics and Finance,Volume 54,2020,101251,ISSN 1062-9408,https://doi.org/10.1016/j.najef.2020.101251.

[4] J. Adams and H. Hagras, "A Type-2 Fuzzy Logic Approach to Explainable AI for regulatory compliance, fair customer outcomes and market stability in the Global Financial Sector," 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK, 2020, pp. 1-8, doi: 10.1109/FUZZ48607.2020.9177542.

[5] Chinedu Ogbonnaya Zephaniah, Ike-Elechi Ogba, Ernest Emeka Izogo,Examining the effect of customers' perception of bank marketing communication on customer loyalty,Scientific African,Volume 8,2020,e00383,ISSN 2468-2276,https://doi.org/10.1016/j.sciaf.2020.e00383.

[6] Yu, L., Zhou, R., Chen, R., & Lai, K. K. (2020). Missing Data Preprocessing in Credit Classification: One-Hot Encoding or Imputation? *Emerging Markets Finance and Trade*, *58*(2), 472–482. https://doi.org/10.1080/1540496X.2020.1825935

[7] Chen, C., Geng, L. & Zhou, S. RETRACTED ARTICLE: Design and implementation of bank CRM system based on decision tree algorithm. *Neural Comput & Applic* **33**, 8237–8247 (2021). https://doi.org/10.1007/s00521-020-04959-8