Intelligent System for Predicting Bank Policy Acceptance by Ensemble Machine Learning and Model Explanation

Remigio Hurtado[1] and Eduardo Ayora[1]

Universidad Politécnica Salesiana, Cuenca, Ecuador
aayorao@est.ups.edu.ec, rhurtadoo@ups.edu.ec

Abstract. Efficient management of financial resources is crucial for the sustainability and competitiveness of banks, particularly in optimizing term deposit subscriptions to maintain liquidity. This paper introduces an advanced intelligent system for predicting term deposit acceptance using ensemble machine learning techniques. Our approach combines Random Forest and K-Nearest Neighbors (KNN) models to enhance prediction accuracy while providing clear explanations. The system follows the CRISP-DM methodology, which includes detailed phases of data preparation, modeling, fine-tuning, and model explanation. We utilize Random Forest for its feature importance metrics and KNN for assessing feature relevance through nearest neighbor analysis. The integration of these methods allows us to generate comprehensive explanations of prediction outcomes by identifying and interpreting key features influencing decision-making. By applying this method to the Bank Marketing Data Set, we demonstrate improved performance across standard metrics such as accuracy, precision, recall, and F1-score. The detailed explanation phase helps understand the model's decision process, providing actionable insights for refining telemarketing strategies. This research presents a robust framework for implementing explainable machine learning in financial marketing, enhancing both predictive accuracy and interpretability for better-informed decision-making.

Keywords: Machine Learning · Intelligent System · Ensemble Learning · Model Explanation · Data Science · Bank Policy Acceptance

1 Introduction

Efficient management of financial resources is vital for the sustainable and competitive operation of the banking industry. Optimizing fund procurement through policies is crucial for providing liquidity. An intelligent decision support system can predict policy acceptance, optimize offers, reduce acquisition costs, and increase customer retention, facilitating better financial planning and ensuring a stable source of funds for lending to SMEs. This promotes economic development, job creation, and sustainable growth [1]. Leveraging AI and machine learning, particularly Ensemble Methods, can improve predictive performance and reliability by combining multiple models to mitigate biases and errors [2]. Ensemble

methods involve combining multiple machine learning models to enhance predictive performance, robustness, and generalization capabilities. Ensemble Methods offer versatility by accommodating various types of base learners and ensemble techniques, such as bagging, boosting, and stacking, each providing unique advantages in different contexts.

Explaining model predictions is crucial for understanding and trust, especially in banking. Three key approaches are saliency-based explanations, transparent model logic-based explanations, and exemplar-based explanations [3] [4]. Transparent models like decision trees and K-Nearest Neighbors (KNN) provide clear, intuitive explanations by showing decision paths and similarities to training data. In our approach, we have selected Random Forest and KNN as part of our ensemble method to achieve both high classification performance and robust explainability. Random Forest provides an ensemble of decision trees, which collectively offer insights into feature importance and decision rules, while KNN facilitates intuitive explanations through its neighborhood-based approach. This combination ensures our model delivers high performance with clear, justifiable predictions, crucial for transparency in banking decisions. Following the CRISP-DM methodology [5], data science processes [6] and processes adjusted from the architectures of our previous researches [7], [8] and [9], our approach covers data collection, preparation, modeling, optimization, and explanation phases. Demonstrated with the Bank Marketing Data Set, the system learns from historical data, makes accurate predictions, and adapts to changes while providing transparent explanations. This approach enhances workflow, decision-making, efficiency, reliability, and responsiveness in the banking sector. Therefore, the main contributions of this research are:

- **Intelligent Decision Support System:** Development of an advanced predictive system to optimize term deposit subscriptions, enhancing financial planning and resource management in banks.
- **Ensemble Predictive Models:** Implementation of a combined K-Nearest Neighbors (KNN) and Random Forest approach, improving prediction accuracy and providing robust explanations.
- **CRISP-DM Framework:** Integration of the CRISP-DM methodology, ensuring systematic data collection, data preparation, modeling, fine-tuning, and explanation of predictions.
- **Explainable AI in Banking:** Emphasis on transparency and interpretability, providing clear and justifiable explanations of predictions for decision-making in banking services.
- **Adaptive Response:** Design of the system to continuously adapt to new data, ensuring an effective and reliable approach to decision support in dynamic financial environments.

The remainder of the document is organized as follows: II. Related work, III. Intelligent System Proposed, IV. Design of Experiments, V. Results and Discussion, and VI. Conclusions.

2 Related Work

In the field of prediction and segmentation in banking marketing, various approaches have been explored to enhance the accuracy and efficiency of models. The study in [10] proposes a data analytics and machine learning-based method for predicting credit risk in financial applications, showcasing the application of advanced techniques in banking risk assessment. The research in [11] provides a comparative analysis of credit rating predictions using neural networks, support vector machines, and decision trees, highlighting the effectiveness of each technique in credit risk forecasting. In the area of banking risk classification, [12] integrates data preprocessing techniques and machine learning, using LightGBM and SMOTE techniques to improve banking risk classification, while [13] presents a hybrid machine learning approach to forecast banking performance, combining multiple techniques to enhance predictions.

For comparing our model, we have selected three prominent baselines. **NN** (Neural Network) uses a neural network as an initial reference for evaluating performance in telemarketing classification [14]. **SMOTEENN-XGBoost** combines oversampling and undersampling techniques with the XGBoost boosting algorithm, providing a robust solution to improve accuracy in predicting deposit subscriptions through advanced ensemble techniques [15]. Finally, **NN-KMeans** integrates a neural network with the K-Means clustering method to segment users, thereby optimizing marketing strategies by combining classification with data clustering [16]. Our proposed method stands out for its integrated and multifaceted approach. We combine a Random Forest model, which provides transparent explanations by identifying the most relevant variables through feature importance, with K-Nearest Neighbors (KNN), which allows precise identification of relevant variables based on similarities among users. This dual approach not only enhances segmentation by identifying user groups with similar characteristics but also facilitates the detection of the most influential variables in predictions. Additionally, we follow a structured process based on CRISP-DM, covering everything from data loading and preparation to detailed model explanation. This comprehensive methodology ensures a deep understanding of the system and offers an innovative solution for prediction and optimization in banking marketing, integrating advanced techniques for greater accuracy and efficiency in decision-making.

3 Intelligent System Proposed

In this section, we present the developed intelligent system. This system enables learning from historical data and predicting new clients to determine whether a client will subscribe to a policy. Figure 1 provides a visual representation of the methodology guiding our process. The general algorithm formalizes the phases of the methodology applied to our specific problem. The proposed methodology follows the CRISP-DM approach, which includes the following detailed steps. First,

data loading is performed, where the relevant dataset for analysis is imported. Second, data preparation is carried out, which includes data cleaning, transformation of categorical variables, and correlation analysis to identify the most relevant variables before learning, thus formulating initial hypotheses. Third, modeling is done using ensemble machine learning techniques, utilizing both voting and stacking methods, combining K-Nearest Neighbors (KNN) and Random Forest. In the stacking method, the final model used to make predictions is Random Forest. Fourth, fine-tuning is performed to identify the best combination of hyperparameters in the ensemble method. Finally, in the fifth step, the Model Explanation phase is conducted, where the model is explained using information obtained from the tuning of KNN and Random Forest to determine the most relevant variables for the predictions.

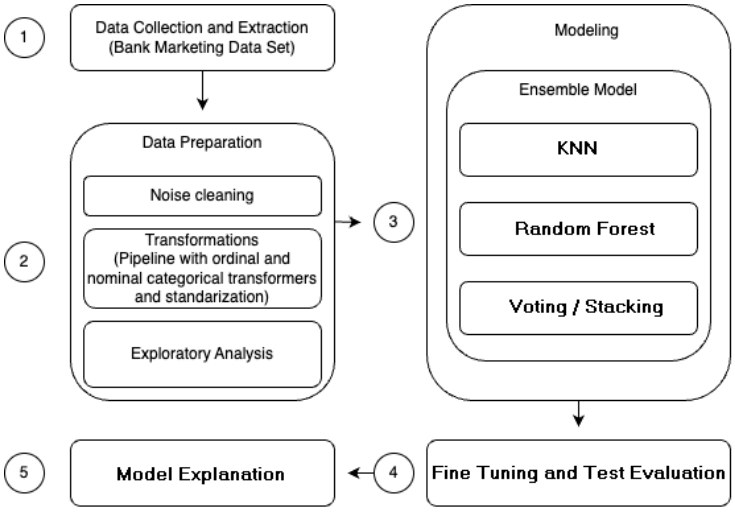

Fig. 1. Proposed method following CRISP-DM process

The methods are combined to obtain a more robust ensemble model, capable of explaining the most relevant variables for the decision-making process in the model, according to the individual KNN and Random Forest models. This supports the understanding of the decision and helps bank executives or the system itself to make the final decision. The voting method allows using the predictions of each base model as a vote, thus contributing to the creation of a more robust and balanced system in predicting policy acceptance. Stacking, on the other hand, combines the predictions of multiple base models through a higher-level model, improving the system's ability to capture complex relationships in the data and providing greater accuracy in the final predictions.

General Algorithm

Input: Dataset
Phase 1: Data Collection and Extraction
1. Load the dataset for analysis.
Phase 2: Data Preparation
2. Data cleaning: removal of null and duplicate values.
3. Transformation of categorical variables into numerical variables.
4. Correlation analysis to identify the most relevant variables and initial hypotheses.
Phase 3: Modeling with Machine Learning Ensemble
5. Training base models:
   a. K-Nearest Neighbors (KNN).
   b. Random Forest.
6. Combining base models using Voting and Stacking techniques.
Phase 4: Fine Tuning and Test Evaluation
7. Identification of the best combination of hyperparameters.
Phase 5: Model Explanation
   a. **KNN for identifying Relevant Variables:**
8. Select the best KNN to predict the labels of the test set.
9. Calculate the absolute differences between the test instances and their nearest neighbors in the training set:
   $d_{ij} = |x_i - x_j|$, where d_{ij} is the absolute difference between instance i and its neighbor j.
10. Calculate the mean of these absolute differences for each feature:
   $\bar{d}_k = \frac{1}{n}\sum_{i=1}^{n} d_{ik}$, where $\bar{d}_k$ is the mean of the absolute differences for feature k.
11. Identify the features with the highest $\bar{d}_k$ as the most relevant.
12. Generate a figure showing these features and their mean absolute differences.
   b. **Random Forest for identifying Relevant Variables:**
13. Select the best Random Forest model with the scaled training set.
14. Calculate the importance of each feature using the *feature_ importances_* metric:
   $FI_k = \sum_{t=1}^{T} \left(\frac{I_t \cdot \Delta f_t}{\sum_{k=1}^{K} I_t \cdot \Delta f_t} \right)$, where FI_k is the importance of feature k, I_t is the importance of tree t, and Δf_t is the decrease in impurity caused by feature k.
15. Sort the features by their importance and select the most relevant ones.
16. Generate a figure showing these features.

Output: Predictions, Relevant Variables and Quality Results.

The **Model Explanation phase** helps us understand which features are most influential in making predictions, providing valuable insights for decision-making in a banking institution and contributing to the system's autonomy. This process enables the automatic execution of the contact or telemarketing process to communicate with clients who are estimated to be interested in a bank policy. In the Model Explanation phase, we identify the most important features using two different methods: K-Nearest Neighbors (KNN) and Random Forest. The feature importance metric in **Random Forest** allows identifying the most relevant variables by analyzing the structure of the tree and how each feature affects the purity of the splits within the model. In a Random Forest, each decision tree partitions the data based on the available features. The importance of a feature

is evaluated by looking at how much the purity of splits in the tree improves when that feature is used. Specifically, we measure how the inclusion of a feature reduces the impurity in the leaves of the tree, i.e., how it helps to make the data sets more homogeneous in terms of the target variable. The features that contribute most to reducing impurity in the tree partitions are considered the most important, as they have a significant impact on the quality of the model and its ability to make accurate predictions. This evaluation helps to identify the key variables that are crucial for decision making in the model. For **KNN** we measure how different each test instance is from its closest neighbors in the training data. We calculate the average of these differences for each feature. The features with the highest average differences are considered the most important. Features with the highest average differences in a K-Nearest Neighbors (KNN) model are considered the most important because they have the greatest impact on the model's predictions. This significant variation indicates that these features are key factors in the decision-making process, providing the model with essential information to make accurate predictions. Understanding these important features helps us gain valuable insights into what drives the predictions, enabling better-informed strategies and actions. For instance, in predicting whether customers will subscribe to a policy using a K-Nearest Neighbors (KNN) model with three customers and two variables (age and income), the standardized data is shown in the Table 1. By calculating the absolute differences in age between Customer A and its nearest neighbors (Customers B and C), we obtain differences of 0.70 and 2.82, with a mean absolute difference of 1.76. Similarly, the differences in income between Customer A and its neighbors are 0.95 and 2.44, yielding a mean absolute difference of 1.69. Since age has the highest average difference, it is identified as the most significant variable in predicting policy subscription, indicating its critical role in the KNN model's decision-making process.

Table 1. Data example for Model Explanation using K-Nearest Neighbors (KNN)

Customer	Age	Income
A	-1.41	-1.22
B	-0.71	-0.27
C	1.41	1.22

4 Design of Experiments

This section presents the characteristics of the dataset, the baselines, the parameters in the fine tuning and the quality measures. Table 2 presents the general description of dataset and their most relevant characteristics. As explained in the related work section, the baselines selected to evaluate the model performance in this study are **NN** (Neural Network) [14], **SMOTEENN-XGBoost** which is a baseline of ensemble learning [15], and **NN-KMeans** [16]. Table 3 presents the parameters to be experimented to train and optimize the models.

For the evaluation of the methods, the Cross-Validation K-Folds technique (with K=3) has been used in order to obtain an adequate generalization of the results and mitigating the risk of overfitting. The quality measures used are: Precision, Recall, F1-Score and Accuracy, particularly tailored for classification tasks. The average of K experiments with the best parameters of each method is presented in the results section.

Table 2. Description of dataset

Dataset	#Customers	#Features	Output
Bank Marketing [17]	45,211	17	y: subscribed or not

Table 3. Optimization parameters of predictive models

Method	Parameters
K-Nearest Neighbors (KNN)	K (Number of neighbors): 3, 5, 7, 10, 15, 20
Random Forest (RF)	n_estimators (Number of trees): 50, 100, 150 max_depth (Depth that trees can reach): 10, 20, 30, 50
Ensemble Model - Voting	Combination of KNN and Random Forest voting: hard (majority of votes) voting: soft (average of the probabilities) * voting is the mechanism for combining predictions
Ensemble Model - Stacking	Combination of KNN and Random Forest

5 Results and Discussion

We began our method by data collection and data preparation. **Phase: Data preparation** involved cleaning the dataset, transforming categorical variables into numerical ones, and performing correlation analysis to identify the most relevant variables and initial hypotheses as shown in Fig. 2. In this study, the categorical variables include the type of job of the client (job), marital status (marital), educational level (education), credit default status (default), housing loan status (housing), personal loan status (loan), type of contact communication (contact), the last month of contact (month), and the outcome of the previous marketing campaign (poutcome). All other variables are numerical. Since the objective of this study is to predict the success of bank telemarketing campaigns, we formulate the following initial hypotheses. Variables such as the duration of the last contact (duration), the successful outcome of the previous campaign (poutcome_success), and contact via cellular phone (contact_cellular) are expected to have a significant influence on the probability that a customer will subscribe to a term deposit. In addition, specific months of contact (month_oct, month_mar, month_sep), the number of previous contacts (previous), and the absence of

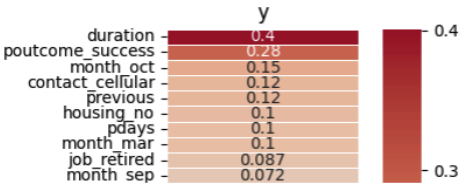

Fig. 2. Correlation analysis to identify the initial relevant variables

mortgage loans (housing_no) could also be determining factors. These hypotheses are based on the observed correlation between these variables and the desired outcome, suggesting that these characteristics could be key indicators in the effectiveness of telephone marketing strategies. In **Phase: Modeling** we trained our base models, which included K-Nearest Neighbors (KNN) and Random Forest, and combined them using Voting and Stacking techniques. In **Phase: Fine Tuning**, the best combination of hyperparameters was identified, as depicted in Fig. 3. The best values for the parameters are as follows: the number of neighbors (K) is 5, the number of trees (n_estimators) is 50, the maximum depth that trees can reach (max_depth) is 50, and the voting mechanism for the ensemble model is soft. Thus leveraging the strengths of both the K-Nearest Neighbors and Random Forest models to enhance overall predictive performance. **Phase: Model Explanation** involved KNN and Random Forest for identifying relevant variables. We selected the best KNN model to predict the labels of the test set. The absolute differences between the test instances and their nearest neighbors in the training set were calculated. The mean of these absolute differences for each feature was then computed. The features with the highest mean absolute differences were identified as the most relevant. We selected the best Random Forest model with the scaled training set. The importance of each feature was calculated, and the features were sorted by their importance to identify the most relevant ones. The results are shown in Panel (a) and Panel (b) of Fig. 4. In Panel (c) of Fig. 4, an example is provided to illustrate how visualizing the structure of a decision tree within a Random Forest can help identify the most relevant variables and criteria for classifying clients. To achieve this, a decision tree was trained with a maximum depth of three. This depth constraint limits the tree's complexity, making it easier to interpret while still capturing the key decision points used in classification. By examining the tree's structure, including the nodes and branches, one can observe how different variables are used to split the data and how they contribute to the classification decisions. The tree's nodes indicate the variables and the corresponding thresholds that are most influential in distinguishing between different classes, thereby highlighting the criteria that have the greatest impact on the model's predictions.

The **final results** of our proposed method and selected methods from related work, including accuracy, precision, recall, and F1-score, are summarized in Table 4. In comparing our model with prominent baselines, we observe dis-

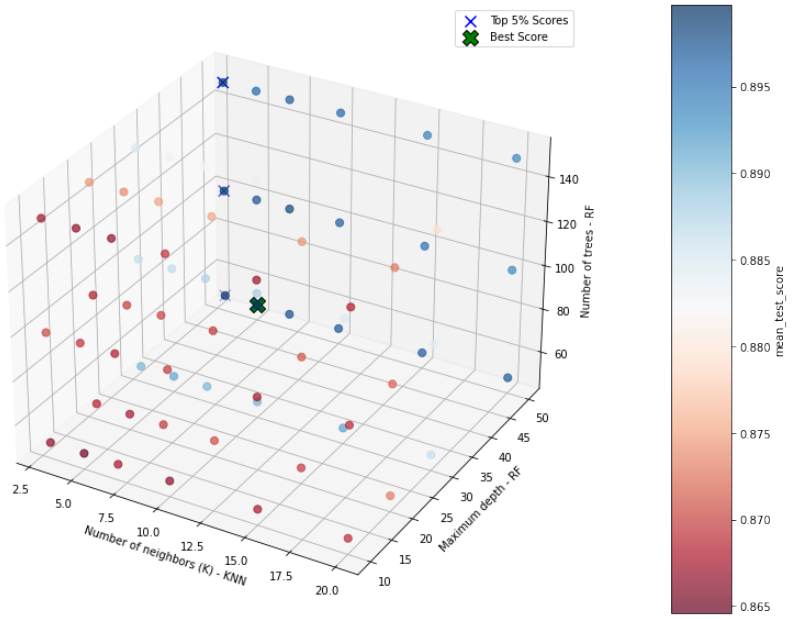

Fig. 3. Hyperparameter tuning

tinct advantages. Neural Network (NN), a widely used reference for telemarketing classification, demonstrates competitive performance, particularly in recall, though its precision and F1-Score are lower. SMOTEENN-XGBoost excels in precision and F1-Score due to its robust ensemble techniques combining oversampling and undersampling with XGBoost, achieving the highest values among the baselines. NN-KMeans provides a strong precision score but falls short in accuracy compared to our method. Our proposed method, integrating Random Forest and K-Nearest Neighbors (KNN), offers significant improvements. The Ensemble Stacking approach achieves the highest accuracy, precision, recall, and F1-Score, indicating a robust and comprehensive model. This success is attributed to our combination of variable importance analysis with Random Forest and similarity-based variable identification with KNN, enhancing both segmentation and prediction accuracy. This dual approach, supported by a structured CRISP-DM methodology, ensures a deep and accurate understanding of customer behaviors and prediction outcomes, thus providing a superior solution in banking marketing. Despite the impressive performance of our proposed method, several areas could benefit from further refinement. While our method excels in accuracy and F1-Score, the recall for some models indicates room for improvement in identifying positive instances. Exploring alternative models or hybrid approaches that enhance recall without compromising other metrics could be advantageous.

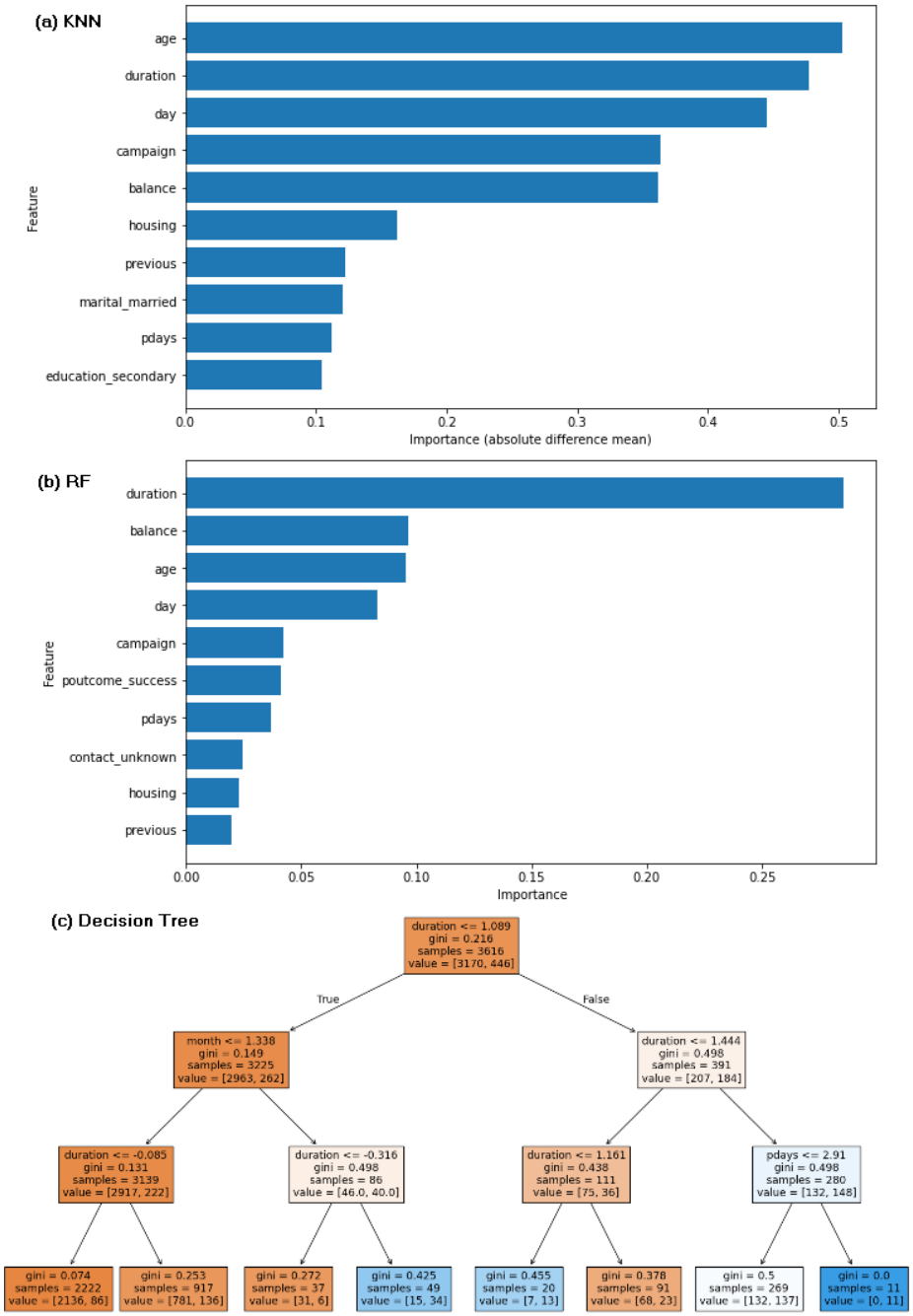

Fig. 4. Panel (a) presents the variables identified as most relevant by the K-Nearest Neighbors (KNN) model, Panel (b) displays the variables deemed most relevant by the Random Forest (RF) model, while Panel (c) presents a Decision Tree from the Random Forest model.

Table 4. Performance metrics for our proposed method and baselines

Predictive Models of Our Method				
Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Ensemble Stacking	**0.9037**	**0.7767**	**0.7131**	**0.7390**
Ensemble Voting	**0.9036**	**0.8074**	0.6498	0.6922
Random Forest (RF)	0.9030	0.7885	0.6707	0.7087
K-Nearest Neighbors (KNN)	0.8947	0.7551	0.6527	0.6856
Baselines				
Method	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Neural Network (NN) [14]	0.8912	0.5112	**0.6723**	0.5850
SMOTEENN-XGBoost [15]	**0.8824**	**0.8650**	**0.9020**	**0.8831**
NN-KMeans [16]	0.8120	0.8230	0.6701	0.7402

6 Conclusions

Our study's initial analysis identified key factors influencing term deposit subscriptions, such as the length of the last contact, the success of prior marketing efforts, and whether the contact was made via a cellular phone. Using Random Forest (RF) and K-Nearest Neighbors (KNN) models, we refined these insights. The RF model highlighted the importance of the number of previous contacts, mortgage status, the time since the last contact, the number of contacts made during the campaign, the success of the previous campaign, and the duration of the last contact. Meanwhile, the KNN model emphasized the significance of the customer's educational background, marital status, and account balance. These findings validate our initial hypotheses, suggesting that telemarketing strategies should focus on these aspects to enhance prediction accuracy and overall effectiveness.

This research presents an advanced decision support system for optimizing term deposit subscriptions in banking, integrating RF and KNN models to enhance prediction accuracy and provide transparent explanations. Following the CRISP-DM methodology, the system systematically addresses data preparation, modeling, fine-tuning, and explanation of predictions. Key insights confirm the importance of variables such as the duration of the last contact, the success of prior marketing efforts, and customer demographics. Overall, the proposed system improves predictive performance and offers a reliable framework for decision support in dynamic financial environments, contributing to better financial planning and resource management in banks. Future work could explore integrating deep learning models or hybrid approaches to enhance recall and performance, incorporating real-time data processing for adaptability, and improving model interpretability and computational efficiency for large-scale datasets. Expanding the system's scope to include other financial products and services could provide a comprehensive decision support tool for banks, driving innovation and customer satisfaction across various domains.

References

1. Alvarez Alvarez, Begoña, and Rodolfo Vázquez Casielles. "Consumer evaluations of sales promotion: the effect on brand choice." European Journal of Marketing 39.1/2 (2005): 54-70.
2. Dietterich, T. G. Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg, 2000.
3. Chin, E. K. (2019). "The Deep Learning Architect's Handbook". Packt Publishing.
4. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
5. Wirth, Rüdiger, and Jochen Hipp. "CRISP-DM: Towards a standard process model for data mining." Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Vol. 1. 2000.
6. C. Hayashi, What is data science? Fundamental concepts and a heuristic example, Data science, classication, and related methods. Springer, Tokyo, 1998. 40-51.
7. Hurtado, Remigio, Stefanía Guzmán, and Arantxa Muñoz. "An Architecture and a New Deep Learning Method for Head and Neck Cancer Prognosis by Analyzing Serial Positron Emission Tomography Images." Conference on Cloud Computing, Big Data and Emerging Topics. Cham: Springer Nature Switzerland, 2023.
8. Ortiz, S. Guzmán, et al. "Una arquitectura de análisis de imágenes seriadas con la tomografía por emisión de positrones mediante la aplicación de machine learning combinado para la detección del cáncer de pulmón." Revista Española de Medicina Nuclear e Imagen Molecular (2024).
9. Hurtado, Remigio, et al. "Development of an Intent-Based Network Incorporating Machine Learning for Service Assurance of E-Commerce Online Stores." International Conference on Machine Learning for Networking. Cham: Springer Nature Switzerland, 2022.
10. Ortiz, Remigio Hurtado, et al. "A data analytics method based on data science and machine learning for bank risk prediction in credit applications for financial institutions." 2022 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC). Vol. 6. IEEE, 2022.
11. Golbayani, Parisa, Ionu Florescu, and 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 (2020).
12. Muslim, M. A., et al. "Bank predictions for prospective long-term deposit investors using machine learning LightGBM and SMOTE." Journal of Physics: Conference Series. Vol. 1918. No. 4. IOP Publishing, 2021.
13. Islam, Ummana, et al. "Forecasting of Bank Performance using Hybrid Machine Learning Techniques." 2022 International Conference on Innovations in Science, Engineering and Technology (ICISET). IEEE, 2022.
14. Moro, S., Cortez, P., & Rita, P. "A data-driven approach to predict the success of bank telemarketing." Decision Support Systems, 62, 22-31 (2014).
15. Li, Y., & Wu, Z. "Prediction of customers' subscription to time deposits based on SMOTEENN-XGBoost model." (2022).
16. Hematyar, A. "Data-driven Decision Making for Direct marketing of Banking Products with the use of Deep Learning and Random Forests." (2022), 78-89.
17. Moro, S., Rita, P., & Cortez, P. "Bank Marketing." UCI Machine Learning Repository (2012). https://doi.org/10.24432/C5K306.