Performance of Machine Learning Techniques and Applied Statistics for Predicting in Financial Organizations

Research extracted from a Master thesis of Statistics

By

Walaa Ibrahim El-shahat Mohamed

A researcher in the Statistics, Mathematics, and Insurance department, Faculty of Commerce-Benha University

walaaibrahim362@gmail.com

Dr. Mohamed Goda Khalil

Assistant Professor of Statistics, Mathematics, and Insurance Department

Faculty of Commerce-Benha University

Dr. Rehab Shehata Mahmoud

Lecturer of Statistics, Mathematics, and Insurance Department

Faculty of Commerce-Benha University

Scientific Journal for Financial and Commercial Studies and Research (SJFCSR)

Faculty of Commerce – Damietta University

Vol.6, No.1, Part 1., January 2025

APA Citation

Mohamed, W. I. E.; **Khalil**, M. G. K. and **Mahmoud**, R. S. (2025). Performance of Machine Learning Techniques and Applied Statistics for Predicting in Financial Organizations, *Scientific Journal for Financial and Commercial Studies and Research*, Faculty of Commerce, Damietta University, 6(1)1, 1-40.

Website: https://cfdj.journals.ekb.eg/

Performance of Machine Learning Techniques and Applied Statistics for Predicting in Financial Organizations

Walaa Ibrahim; Dr. Mohamed Goda and Dr. Rehab Shehata

Abstract

This study scrutinizes the efficacy of established statistical methodologies alongside cutting-edge machine learning processes for projecting trends in bank deposit dynamics, utilizing data from a series of direct marketing initiatives by a bank in Portugal. Initially, the analysis employs both descriptive and inferential statistical approaches to delineate consumer behavior and evolving market scenarios, paving the way for sophisticated predictive analytics.

As the examination unfolds, it incorporates an array of machine learning techniques, encompassing logistic regression, decision trees, and support vector machines, enriched with complex ensemble frameworks such as random forests and gradient boosting. The research further harnesses the capabilities of neural networks to delve into deeper behavioral analytics of consumers.

The application of the Synthetic Minority Over-sampling Technique (SMOTE) significantly refines the accuracy of the predictive models by addressing imbalances within the class distributions, thereby enhancing overall model performance. For instance, post-SMOTE application, the Decision Tree model's accuracy escalated from 87.22% to 89.76%, with a concurrent rise in its ROC AUC from 70.19% to 89.74%. The Random Forest model similarly benefited, with its accuracy jumping from 90.06% to 93.68% and ROC AUC increasing from 92.18% to 96.05%.

Further advancements in model reliability are achieved through strategic applications of bagging and stacking techniques, which fortify the models' accuracy and stability. Through bagging, accuracy was heightened to 94% and ROC to 98.6%, while stacking brought about an accuracy of 92% and a ROC of 97%. These enhancements not only bolstered the predictive precision but also provided richer insights into the variables influencing commitments to bank deposits, underscoring the vital importance of hybrid modeling strategies in the optimization of financial decision-making.

Key Words: Bank deposit, Machine Learning, Machine Learning Algorithms, SMOTE.

Introduction:

The banking sector is a cornerstone of economic stability and growth, playing a critical role in the financial well-being of individuals, businesses, and nations. Banks operate by pooling customer deposits and utilizing these funds to extend loans and invest in profitable ventures. This dual role of banks as both financial intermediaries and economic stimulators makes understanding and predicting bank deposit trends crucial for economic planning and stability. The ability to accurately forecast bank deposits is not just a tactical advantage but a strategic necessity in today's volatile financial environment (Koroniotis, 2020).

Historically, the focus of financial predictive modeling has largely been on securities like bonds and stocks. However, deposits represent a significant portion of the resources raised by banks, accounting for roughly two-thirds of total funds. Despite their importance, predictive modeling for deposit trends has not been as extensively explored as other financial metrics. This is partly due to the complex and dynamic nature of the factors that influence deposit levels, such as economic policies, interest rates, consumer confidence, and broader market conditions (Heider and Leonello, 2021).

The determinants of bank deposits can be broadly categorized into macroeconomic and microeconomic factors. Macroeconomic factors include elements like interest rates, inflation, and overall economic growth, which influence the general economic environment and affect consumer behavior towards savings and investments (Wilmarth, 2020). Microeconomic factors pertain more directly to individual banks and include their policy decisions, product offerings, marketing strategies, and customer service quality, which can attract or deter depositors (Liu et al., 2021).

In the realm of predictive modeling, the challenge lies in selecting and effectively utilizing the right set of variables that can capture the essence of these complex dynamics. Variables such as, economic growth, disposable personal income, education and employment levels are essential as they directly influence consumers' liquidity preferences and investment decisions relative to alternative financial instruments. Their effectiveness in predictive

models depends significantly on how well they orientate these consumer behaviors and market trends (Tsui et al.2023).

This research focuses on enhancing the predictive accuracy of bank deposit trends by employing advanced statistical and machine learning techniques. Traditional approaches have often fallen short due to their inability to adapt quickly to changing market conditions or to handle large and diverse datasets that reflect the multifaceted nature of economic behaviors.

This study examines the application of some sophisticated supervised learning techniques—and neural networks utilized across distinct dataset derived from the direct marketing efforts of a Portuguese bank, also archived in the UCI machine learning repository. These techniques were chosen for their proven efficacy in handling complex, nonlinear relationships and their capacity for feature importance evaluation, which is crucial for understanding the impact of various predictors in the models.

This methodological framework employs a comprehensive data preprocessing strategy to ensure the robustness and generalizability of the algorithms across various data subsets. This meticulous approach facilitates the evaluation of each model's effectiveness in isolation while simultaneously allowing for a comparative analysis of their performance. This rigorous process underscores the commitment to achieving precise and dependable predictive outcomes.

The transition to machine learning is further exemplified by the employment of the Synthetic Minority Over-sampling Technique (SMOTE), which significantly enhances model performance by addressing class imbalances prevalent in the training datasets. The efficacy of machine learning is demonstrated through notable improvements in predictive accuracies and ROC AUC values, as evidenced by the enhanced performance of models post-SMOTE application.

Moreover, the deployment of ensemble methods such as bagging and stacking introduces an additional layer of sophistication, substantially elevating the reliability and accuracy of the predictive models. The application of neural networks, tailored with multiple layers and specific

activation functions, further underscores the depth of analysis possible with these techniques, achieving impressive accuracies and ROC scores.

The integration of these advanced statistical and machine learning techniques not only refines the accuracy of predictive models but also deepens our understanding of the variables influencing bank deposit subscriptions. By highlighting the benefits of a hybrid approach that combines traditional statistical methods with modern machine learning algorithms, this research contributes significantly to the banking sector's ability to make informed decisions and tailor marketing strategies effectively. This study not only charts a path for future research but also acts as a beacon for financial institutions aiming to harness the power of advanced analytics to better understand and serve their customers.

Conceptual framework of the study variables:

- **Bank Deposit Definition**

Bank deposits constitute a core component of the financial intermediation process, serving as the primary mechanism through which individuals and entities transfer monetary resources into banking institutions for safeguarding, management, and earning potential interest. These deposits are essential to the liquidity and capital reserves of banks, enabling them to finance lending and investment activities critical for economic development and stability (MOHD-KARIM, 2010).

- **Machine Learning Definition**

Machine Learning (ML) stands as a vibrant and swiftly advancing area, covering a broad array of methods and strategies. It's crucial to understand that machine learning can be defined in numerous ways, each reflecting different elements of its comprehensive nature. Moreover, the diverse branches of machine learning highlight the extensive variety of methods used by researchers and practitioners to tackle intricate challenges. Acknowledging the breadth of machine learning definitions enables us to recognize its flexibility and capacity for adaptation, promoting ongoing innovation and progress within this thrilling field.

ML emerges as an innovative discipline at the intersection of statistics, applied mathematics, and computer science. It processes data, detects patterns, and operates with minimal human intervention. These algorithms are capable of executing nearly any task that follows a data defined pattern or ruleset (Dutta, 2021).

- **Machine Learning Algorithms**

In the realm of machine learning, there are several distinct types or categories (Supervised learning, unsupervised learning, reinforcement learning, and deep learning) that capture different learning approaches and techniques. Understanding these types is crucial for selecting the appropriate methodology for a given problem or task.

Here is an outline of the algorithms used to obtain the results:

- **Logistic Regression**

Logistic regression is a supervised ML that works by extracting some set of weighted features from the input, taking logs and combining them linearly, which means that each feature is multiplied by a weight and then added up (Nasteski, 2017).

It is a type of regression that predicts the probability of occurrence of an event by fitting data to a logistic function. Just as many forms of regression analysis, logistic regression makes use of several predictor variables that may be numerical or categorical.

The logistic regression hypothesis is explained by (Nasteski, 2017) as:

$$\boldsymbol{h_\theta(x) = g(\theta^T x)} \quad \textbf{(1)}$$

Where the function g is sigmoid function defined as:

$$\boldsymbol{g(z) = \frac{1}{1+e^{-z}}} \quad \textbf{(2)}$$

- **Support Vector Machine**

Support Vector Machines (SVMs) are supervised learning models used for classification and regression, particularly effective for binary classification problems. The core concept is to find a hyperplane that best separates data into classes while maximizing the margin between them. The

margin is the distance between the hyperplane and the nearest data points of each class, known as support vectors (Wu & Zhou, 2006).

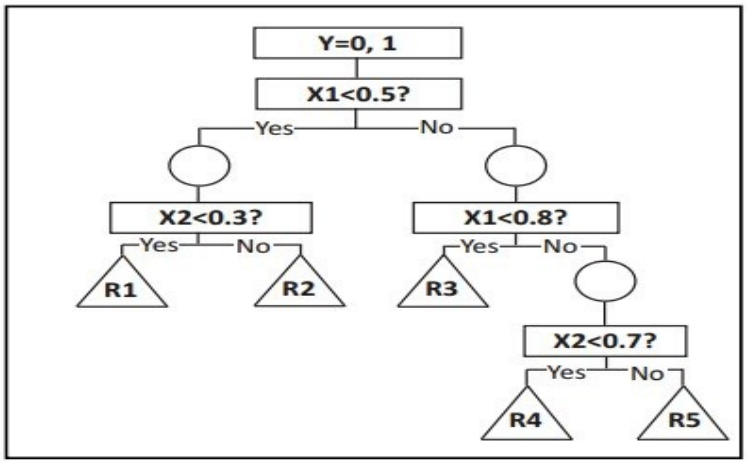
In a binary classification problem, the SVM algorithm seeks to find a hyperplane that separates the data points of two classes. The hyperplane equation is defined as:

$$\boldsymbol{wx + b = 0} \qquad \textbf{(3)}$$

Where:

- w is the weight vector perpendicular to the hyperplane.
- x is the input feature vector.
- b is the bias term (intercept).

- **Decision Tree:**



Decision trees are non-parametric models that make predictions by partitioning the feature space based on input variable values. They classify data by asking a series of hierarchical questions, each represented by a node. Each internal node directs to child nodes based on answers, forming a tree structure. Items are sorted into classes by following paths from the root to leaf nodes. Some variations include probability distributions at leaves to estimate class probabilities, although unbiased probability estimation can be challenging (Kingsford & Salzberg, 2008).

Figure (1) Decision Tree (Song & Ying, 2015)

- **Random forest**

Random Forest is a robust algorithm used in supervised learning for both classification and regression tasks. It is an ensemble method that combines multiple decision trees to enhance accuracy and robustness. By aggregating predictions from multiple models, it often outperforms a single model. In Random Forests, the selection of predictor variables for splitting nodes is based on predetermined criteria, such as entropy for classification, formulated as an optimization problem (Schonlau & Zou, 2020).

$$\boldsymbol{E = -\sum_{i=1}^{c} p_i \times \log(p_i)} \quad \textbf{(4)}$$

where $\boldsymbol{c}$ is the number of unique classes and $\boldsymbol{p_i}$ is the prior probability of each given class. This value is maximized to gain the most information at every split of the decision tree. For regression problems, a commonly used splitting criterion is the mean squared error at each internal node.

- **Gradient Boosting**

Gradient Boosting is an ensemble learning technique for regression and classification that sequentially builds models, each correcting errors from the previous ones. It minimizes a specified loss function, by iteratively adding models (usually decision trees) to correct residual errors. This process resembles gradient descent optimization in function space, combining multiple weak learners into a strong learner. The method effectively reduces bias and variance, leading to robust and accurate predictions (Sun et al., 2020).

The loss function used can vary depending on the task. For regression, a common choice is the Mean Squared Error (MSE):

$$\boldsymbol{L\big(y, F(x)\big) = \frac{1}{2}\sum_{i=1}^{N}(y_i - F(x_i))^2} \quad \textbf{(5)}$$

- **K-nearest neighbors**

K-nearest neighbors (KNN) is a non-parametric algorithm used for classification and regression tasks in machine learning. It predicts the target value of a new data point based on the majority class or average value of its k nearest neighbors in the training dataset. The algorithm uses distance metrics, commonly Euclidean distance, to measure similarity between data points. For classification, KNN assigns the most frequent class among the k-nearest neighbors to the query point (Guo et al., 2003).

The algorithm calculates the distance between data points using a distance metric, most

commonly the Euclidean distance, though other metrics like Manhattan distance can also be used. The distance between two points is a measure of their similarity. The Euclidean distance between two points, 'p' and 'q', in a d-dimensional space is calculated as: (Zhang et al., 2017)

$$\boldsymbol{\sqrt{\sum_{i=1}^{d}(p_i - q_i)^2}} \quad \textbf{(6)}$$

where '$\boldsymbol{p_i}$' and '$\boldsymbol{q_i}$' are the respective coordinates of the two points in the '$\boldsymbol{i}$' dimension. For classification, the k-NN algorithm identifies the k training instances that are closest to the query point based on the chosen distance metric.

- **Neural Network**

A neural network is a computational model inspired by the human brain, consisting of interconnected layers of nodes (neurons). It includes an input layer, one or more hidden layers, and an output layer. Nodes process data through weighted connections and activation functions, introducing non-linearity. Neural networks are trained on large datasets using optimization algorithms to adjust weights, minimizing a loss function to improve prediction accuracy, and are used for tasks like classification, regression, and pattern recognition (Ibnu et al.2020).

When setting up a neural network, the configuration of various parameters is essential for tuning the model's performance (Shibuya & Hotta, 2022). Below is a brief overview of key parameters:

Table (1)
Brief Explanation of Neural Network Configuration Parameters.

Parameter	Description
Batch Size	Defines the number of samples processed before the model updates its parameters. Larger batches offer more precise gradient estimates but use more memory; smaller batches lead to quicker but less stable updates.
Epochs	Represents one complete cycle through the training data, with more epochs potentially improving model accuracy at the risk of overfitting.
Optimizer Type	Algorithms that adjust network attributes (like weights) to minimize losses; common types include SGD, Adam, and RMSprop.
Learning Rate	Controls the adjustment magnitude of model weights during training, critical for efficient convergence.
Number of Layers	Indicates the depth of the network, affecting the model's ability to learn complex patterns but increasing training difficulty.
Activation Function	Functions that define the output of network nodes, crucial for learning non-linear data patterns. Examples include ReLU (hidden layers) and Sigmoid or Softmax (output layers).

- **Multi-Layer Perceptron (MLP) Classifier**

It is a type of artificial neural network used for classification tasks. It consists of an input layer, one or more hidden layers, and an output layer, each composed of neurons. These neurons are fully connected, meaning each neuron in one layer is connected to every neuron in the next layer. The MLP utilizes non-linear activation functions, such as the sigmoid function or Rectified Linear Unit (ReLU), to capture complex patterns in data (Mohammed & Al-Bazi, 2022).

The ReLU function can be expressed as:

$$\boldsymbol{ReLU(z) = max\ (0, z)} \qquad \textbf{(7)}$$

- **Bagging**

The Bagging Algorithm, or Bootstrap Aggregating, enhances the stability and accuracy of machine learning models by training multiple versions on various subsets of the dataset, using random sampling with replacement. This method effectively mitigates overfitting and reduces variance by leveraging the strengths of each model subset. In classification tasks, it employs majority voting from all models for final predictions, whereas in regression, it averages all outputs (Ganaie et al.2022).

The technique utilizes the central limit theorem, which suggests that the distribution of sample means approaches a normal distribution as the sample size increases, thus improving predictive performance with minimal increase in bias.

- **Stacking**

The Stacking Algorithm is a sophisticated ensemble machine learning technique aimed at enhancing prediction accuracy by strategically combining multiple predictive models to form a superior meta-model.

It involves constructing a new model, known as a meta-learner or blender, that learns to optimally integrate the predictions of several base models. The process starts by training various models on the same dataset. These models, typically diverse, generate predictions used as inputs for the meta-learner. The final prediction is derived from the meta-learner, ensuring it is informed by a comprehensive blend of insights from the initial models (Cui et al., 2021).

Stacking is based on the premise that the aggregation of multiple predictions leads to more accurate and robust outcomes than any single model alone. This concept is supported by statistical theory, which suggests that combining various estimates tends to converge towards the true underlying value, effectively reducing variance.

- **Synthetic Minority Over-sampling Technique (SMOTE)**

SMOTE is an advanced over-sampling approach that aims to balance class distribution through the creation of synthetic samples rather than by simply duplicating minority class instances. This technique is particularly useful in

scenarios where the minority class is under-represented, and traditional over-sampling methods might lead to overfitting (Khushi et al.2021).

It operates by selecting instances that are close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line.

Mathematically, a synthetic sample S is generated as follows:

$$\mathbf{S = x_i + \lambda(x_{zi} - x_i)} \quad \textbf{(8)}$$

Where:

$\mathbf{x_i}$ **:** is a randomly chosen minority class sample.

$\mathbf{x_{zi}}$ **:** is one of its k-nearest neighbors.

$\boldsymbol{\lambda}$ **:** is a random number between 0 and 1.

Related Work

This section explores the integration of machine learning algorithms in banking, discussing their impacts and outlining current trends, challenges, and future prospects in the banking sector.

Here is an overview of scholarly perspectives on banking systems:

Ruangthong and Jaiyen (2015) presented a compelling solution to the challenge of predicting customer responses in bank direct marketing campaigns by effectively handling imbalanced data. The proposed method not only improves prediction accuracy but also offers insights into the application of machine learning techniques in direct marketing strategies within the banking sector.

The study presented in the paper discovered that combining the Synthetic Minority Over-Sampling Technique (SMOTE) with the Rotation Forest Principal Component Analysis (PCA)-J48 algorithm effectively addressed the challenge of imbalanced data in predicting customer responses to bank direct marketing campaigns. This method surpassed several other machine learning algorithms in terms of accuracy, sensitivity, and specificity.

Lu et al. (2016) presented a study on the application of Artificial Immune Systems (AIS) for data analysis and pattern recognition in bank term

deposit recommendations. Also, introduced an Artificial Immune Network (AIN) model for collaborative filtering as a classification tool after

applying feature selection to identify key characteristics for classification purposes.

The study successfully applied an AIN combined with feature selection techniques to improve the accuracy of bank term deposit recommendations. The AIN model's ability to handle imbalanced data and generate numerous classification rules highlights its potential for accurate customer prediction in the banking sector.

Ramesh (2017) presented a study on developing a machine learning model for predictive analytics on banking dataset. This dataset, which comes from the University of California Irvine Machine Learning Repository, includes customer details and their responses to a bank's marketing campaigns, specifically, whether they would subscribe to a bank term deposit.

The study employed a dataset related to direct marketing campaigns of a Portuguese banking institution, focusing on phone calls made to clients. Data preprocessing involved cleaning and converting the dataset's binary outcome (yes/no) into numerical values (1/0) for model training and testing. Logistic regression is used for the binary classification model, leveraging Amazon Web Services (AWS)'s capabilities for easy deployment and prediction. It concluded that AWS Machine Learning offers a robust platform for developing and deploying machine learning models for predictive analytics in banking. The model shows high accuracy and can be used to predict customer behavior in response to marketing campaigns effectively.

Le and Viviani (2018) evaluated the performance of two traditional statistical approaches.

(Discriminant Analysis and Logistic Regression) against three machine learning approaches (ANN, SVM, and K-Nearest Neighbors). The study considered 31 financial ratios covering aspects such as loan quality, capital quality, operational efficiency, profitability, and liquidity, collected over a five-year period before the banks became inactive.

ANNs and K-Nearest Neighbors (k-NN) methods outperformed traditional statistical methods in predicting bank failures, with ANNs showing the highest accuracy. It illustrated the potential of machine learning

techniques in enhancing the predictive accuracy of bank failure models beyond traditional statistical methods, leveraging a broad set of financial ratios to capture various aspects of bank operations and health.

Ilham et al. (2019) explored various machine learning algorithms to classify potential bank customers based on their likelihood to be interested in long-term deposit products through telemarketing strategies. The main objective was to maximize customer value and increase corporate earnings by accurately targeting potential customers.

The study compares seven different classification algorithms: DT, Naïve Bayes (NB), RF, K-NN, SVM, Neural Network (NN), and Logistic Regression (LR), using a dataset from the Protestal Bank of the UCI Machine Learning repository. The performance of these algorithms was evaluated based on two metrics: AUC and Accuracy. The results indicated that the SVM algorithm performed the best in terms of both Accuracy (91.07%) and AUC (92.5%), making it the most suitable choice for classifying prospective customers interested in time deposit products offered via telephone.

Machine (2020) focused on utilizing Bank Marketing dataset, traditionally used for predicting long-term deposit subscriptions, to predict loan approval decisions for bank clients. This innovative approach aims to support bank decision-makers by not only achieving high model prediction accuracy but also providing interpretable predictions.

Utilizing a range of ensemble machine learning methods, such as Bagging and Boosting, the study developed a predictive model with an impressive 83.97% accuracy, markedly surpassing the performance of most current loan prediction models by around 25%. The research explored various ensemble techniques, including AdaBoost, LogitBoost, Bagging, and RF, finding that the LogitBoost model delivered superior results when all dataset features were employed.

Borugadda et al. (2021) investigated the efficacy of telemarketing strategies in promoting long term deposits within the banking sector, analyzing data from direct marketing campaigns of a Portuguese bank. The research offered valuable insights for banks to improve their marketing strategies and laid a foundation for future studies on optimizing telemarketing efforts through machine learning techniques.

Utilizing a range of machine learning models RF, SVM, Gaussian Naive Bayes (GNB), DT, and LR, the study found that the LR model was the most effective, boasting a 92.48% accuracy rate in identifying potential customers for long-term deposits through telemarketing.

Tuan (2022) explored the use of deep neural network models to predict customer decisions on term deposits in a banking context. The study aimed to utilize information such as age, job, marital status, education, and other factors through deep learning models to identify potential depositors.

The research applied several deep learning models, including Longshot Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (BiLSTM), Bidirectional GRU (BiGRU), and Simple Recurrent Neuron Network (SimpleRNN). The findings revealed that the GRU model achieved the highest accuracy of 90.08% at the 50th epoch, closely followed by the BiLSTM model with 90.05% accuracy.

Hayder et al. (2023) focused on utilizing machine learning algorithms to predict customers' reactions to bank marketing efforts, specifically regarding fixed-term deposit offers. The data used in the study was sourced from the UCI machine learning repository, related to the Portuguese banking institution's marketing campaigns, and contained 20 input variables alongside one output variable.

The study employed four machine learning classifiers: k-NN, DT, NB, and Support SVM. The findings indicated that the Decision Tree classifier yielded the highest accuracy of 91%, followed closely by SVM with an accuracy of 89%. The study also involved data balancing techniques like SMOTE and feature selection methodologies to optimize the performance of the machine learning models.

Zaki et al. (2024) investigated the use of predictive analytics and machine learning to enhance direct marketing strategies for bank term deposits. Using datasets from Kaggle, the study examined data through various visualization techniques and applied several machine learning models. Among these, the Random Forest Classifier showed the highest effectiveness, achieving notable metrics: an accuracy of 87.5%, a negative predictive value (NPV) of 92.99%, and a positive predictive value (PPV) of

87.83%. These results highlighted the potential of machine learning to enhance predictive capabilities in banking marketing strategies.

Problem Statement:

The financial sector is fiercely competitive, prompting banks to refine their marketing strategies to attract and keep customers. A major challenge for banks is identifying potential clients likely to subscribe to term deposit products, as traditional marketing methods are often costly and inefficient.

Using personalized, data-driven marketing can greatly enhance the effectiveness of direct marketing campaigns. However, the vast and complex nature of customer data requires sophisticated analytical tools and robust predictive models to derive actionable insights and predict customer behavior.

A dataset from a Portuguese bank, including various customer and campaign details, allows for the exploration of these challenges through machine learning. This method improves understanding of customer behavior and optimizes marketing strategies, leading to better resource use and higher conversion rates.

The growing availability of customer data necessitates advanced models capable of making accurate predictions and considering diverse data types and factors. These models revolutionize bank marketing by enabling more targeted, cost-effective, and customer-focused campaigns.

Furthermore, applying statistical analysis to this data helps uncover crucial relationships between customer attributes and their likelihood of subscribing to a term deposit. This analysis also assesses the effectiveness of marketing strategies and identifies key factors influencing customer decisions.

Aim of the work:

- Develop a predictive model to determine the likelihood of a customer subscribing to term deposit based on their attributes and past interactions.
- Perform customer segmentation analysis to identify distinct groups with varying propensities to subscribe, thereby enabling tailored marketing strategies.
- Optimize marketing campaign strategies by analyzing the effectiveness of different communication methods, contact frequencies, and previous campaign outcomes.

- Develop a generalizable framework that other financial institutions can adopt enhance their marketing strategies through similar analytical and predictive methodologies.

Data Description:

The dataset used in this study is sourced from the UCI Machine Learning Repository and specifically pertains to direct marketing campaigns of a Portuguese banking institution. The data was collected from 2008 to 2013 and aims to predict whether a client will subscribe to a term deposit. This makes it a valuable resource for developing targeted marketing strategies.

The Bank Marketing Dataset comprises approximately 45,211 entries and 17 features, each entry representing an individual client. This comprehensive dataset covers demographic, financial, contact, and campaign-related information, suggesting its use in marketing or CRM analytics, potentially to predict customer behavior or the success of marketing campaigns. Categorical and Numerical Balance: With 10 categorical and 7 numerical features, this dataset offers a balanced mix for applying various data analysis techniques, including classification algorithms that can handle both data types effectively.

The primary objective of using this dataset is to predict the binary outcome 'y', which indicates whether the client subscribed to a term deposit. This outcome helps in understanding the effectiveness of the bank's direct marketing campaigns and in optimizing future marketing strategies to increase conversion rates.

Statistical Analysis for Numerical Features

Table (1.2)
Descriptive Statistical Metrics for Numerical Features.

features	mean	std	min	2.5%	25%	50%	75%	97.5%	max
age	40.936210	10.618762	18	25	33	39	48	61	95
balance	1362.27	3044.77	-8019	-375	72	448	1428	8411	102127
day	15.81	8.32	1	2	8	16	21	30	31
duration	258.16	257.53	0	19	103	180	319	974	4918
campaign	2.76	3.1	1	1	1	2	3	11	63
pdays	40.2	100.13	-1	-1	-1	-1	-1	354.75	871
previous	0.58	2.3	0	0	0	0	0	5	275

From the percentiles, we can see that 50% of clients are 39 or younger, and 25% are 48 or older, suggesting a relatively young client base. The account balance is quite skewed, with 25% of clients having a balance of 72 or less, while a small percentage (2.5%) have a negative balance, which could be a concern for the financial institution. The contact durations are also skewed; the majority of contacts are relatively short, but there are some very long ones.

Most clients are not repeatedly contacted (50% have only been contacted twice or less), but a small percentage have been contacted 11 times or more (2.5% of clients). These statistics help the bank understand the demographic and behavioral patterns of their clients, which can inform strategies for future marketing campaigns.

Statistical Analysis for Categorical Features.

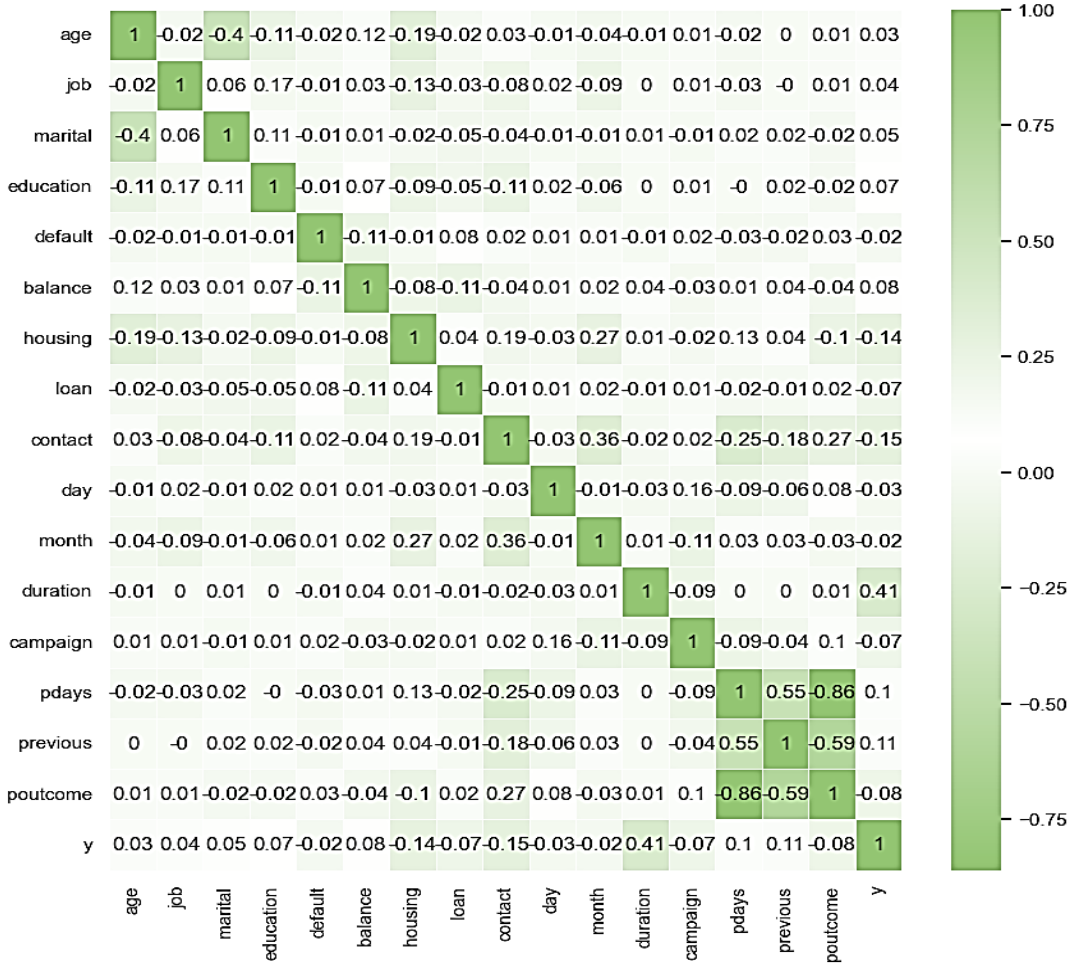
Table (1.3)
Descriptive Statistical Metrics for Categorical Features

features	count	unique	top	frequency
job	45211	12	blue-collar	9732
marital	45211	3	married	27214
education	45211	4	secondary	23202
default	45211	2	no	44396
housing	45211	2	yes	25130
loan	45211	2	no	37967
contact	45211	3	cellular	29285
month	45211	12	may	13766
poutcome	45211	4	unknown	36959
y	45211	2	no	39922

From the analysis of the categorical features, we observe several key characteristics about the client base. Most clients are employed in blue-collar jobs, are married, and have attained a secondary education level. Regarding financial commitments, the vast majority have not defaulted on any loans. It is more common for clients to have housing loans than not, while personal loans are less prevalent.

In terms of communication, cellular phones emerge as the most favored method. May stands out as the busiest month for client contact. When looking at the outcomes of previous campaigns, it is noted that for most clients, these outcomes remain unknown. Ultimately, the majority of clients did not

subscribe to a term deposit, despite this being the primary goal of the campaign.



Correlation

Figure (2) Correlation Matrix for the Data

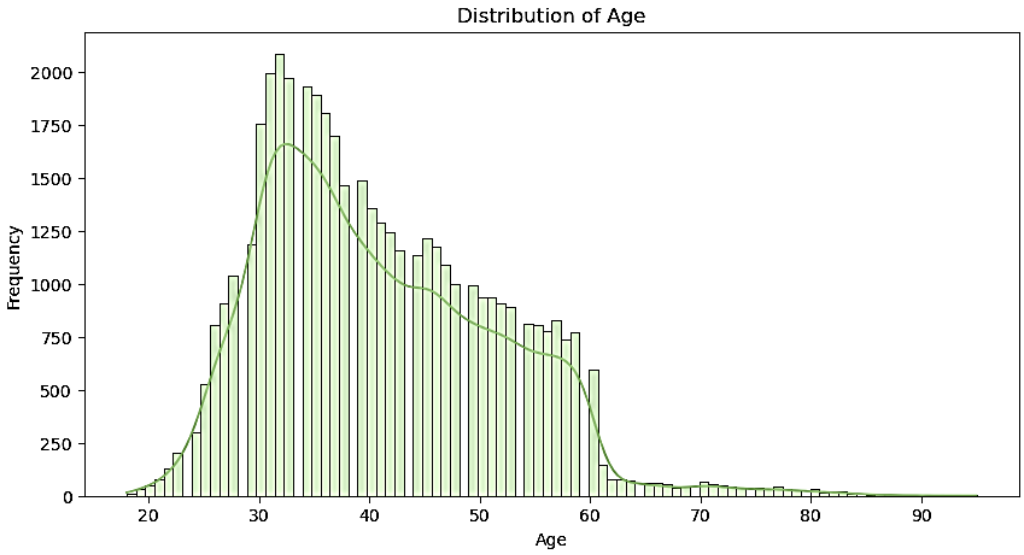
The correlation matrix illustrates the pairwise correlation coefficients between various variables in the dataset, which is crucial for understanding linear relationships and potential multicollinearity among the variables. Notably, a strong correlation exists between poutcome and previous (0.86), suggesting a significant association between the outcome of previous marketing campaigns and the number of prior contacts.

Additionally, the duration of the call shows a moderate correlation with the target variable y (0.41), implying that longer call durations could be

predictive of successful outcomes. Other correlations, such as the slight inverse relationship between housing and loan (-0.14), provide insights that may impact the development of predictive models.

High correlations like those observed could indicate multicollinearity, especially problematic in predictive modeling as they can inflate the variance of coefficient estimates and destabilize the model. To mitigate such issues, techniques like Principal Component Analysis (PCA) or regularization methods like LASSO might be necessary.

The color gradient from green to white in the matrix effectively highlights these correlations, with darker shades indicating stronger



correlations, aiding in the quick identification of significant relationships. This analysis is imperative for selecting appropriate features in model building, aiming to enhance model reliability and predictive performance.

Application of Shapiro-Wilk Test on Age Feature

Figure (3) Application of Shapiro-Wilk Test

Results

- Shapiro-Wilk Test Statistic: 0.9605457782745361, p-value: 0.0
- Age distribution is likely not normal (reject H_0)
- Skewness of age distribution: 0.6847952047866451

- Kurtosis of age distribution: 0.31940232676995794
- Age distribution is positively skewed.
- Age distribution has a sharper peak than a normal distribution (Leptokurtic).

The results strongly indicate that the age distribution is not normal.

Application of Chi-square Test

Table (4)
Chi-square Test Results for Association with Target Variable 'y'.

feature	Chi-square	P-value	Degree of freedom
job	836.105488	-162.93	11
marital	196.495946	-37.17	2
education	238.923506	-46.58	3
default	22.202250	0.67	1
housing	874.822449	-184.07	1
loan	209.616980	-42.47	1
contact	1035.714225	-221.60	2
month	3061.838938	0	11
poutcome	4391.506589	0	3

The Chi-square test results indicate significant associations between all the categorical features and the target variable 'y', as all p-values are significantly below the standard threshold of 0.05. This implies that each categorical feature has a strong relationship with the target variable 'y', which could be crucial for predictive modeling and analysis.

Data Preprocessing:

Here are the steps used to make preprocessing for the data:

- **Data Cleaning:**

The dataset is free from missing values and duplicates, ensuring it is primed for analysis. This clean data allows for direct application in analytical

models and offers ample opportunities for insightful feature engineering to enhance predictive accuracy.

- **Handling outliers**

In the pursuit of data normalization, particularly in the presence of outliers, traditional scaling methods, such as standard Z-score normalization, can be significantly influenced by the deviant points. To address this challenge, is employed due to its effectiveness in managing outliers.

Robust Scaler is particularly effective in datasets where outliers are present. Unlike traditional methods, it uses the median and the interquartile range (IQR) for scaling. Specifically, it subtracts the median from each data point and then divides by the IQR.

$$\mathbf{IQR = Q_3 - Q_1} \quad \textbf{(9)}$$

- **Label Encoding**

In the preprocessing stage, label encoding is applied to the categorical features using a dedicated encoder for each column. This approach ensures that the data for each feature was independently transformed, maintaining the integrity and uniqueness of the information for the machine learning models.

- **Data Splitting**

The dataset is strategically divided into two subsets: 80% for training and 20% for testing. This splitting ratio is chosen to ensure a robust training process while still providing a substantial test set to evaluate the model's performance accurately.

Model Performance and Evaluation

Following meticulous preprocessing and strategic partitioning of the dataset into training and test sets, this section is dedicated to the application and detailed evaluation of various machine learning algorithms. The performance of each algorithm is rigorously analyzed using a diverse array of metrics: accuracy, precision, recall, F1 score, Area Under the Curve (AUC), and Receiver Operating Characteristic (ROC) curves.

Results Before using SMOTE

- **Logistic Regression Performance**

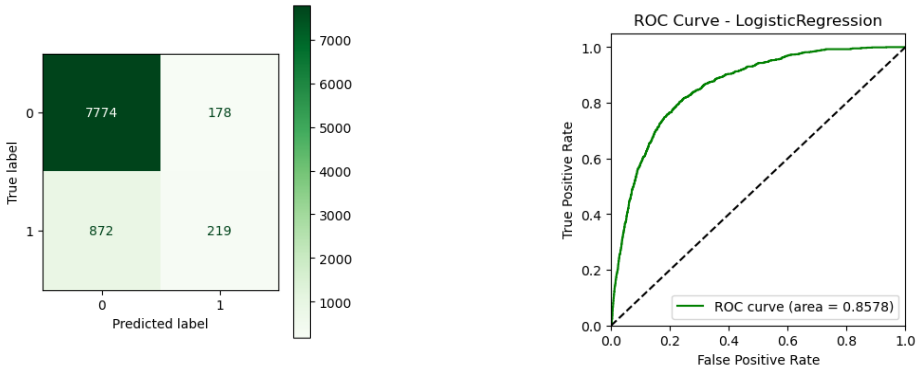


Figure (4) Confusion Matrix and ROC of LR

Table (5)
Performance of LR

Metric Name	Percentage
Accuracy	0.8839
Precision	0.8572
Recall	0.8839
F1 Score	0.8592
ROC	0.8578

The Logistic Regression model demonstrates strong performance with an Accuracy of 88.39%, Precision of 85.72%, and an F1 Score of 85.92%, indicating effective balance in classifying the outcomes accurately. The ROC curve further supports its robustness with an AUC of 85.78%, showcasing good discriminative ability between classes.

- **Support Vector Machine Performance**

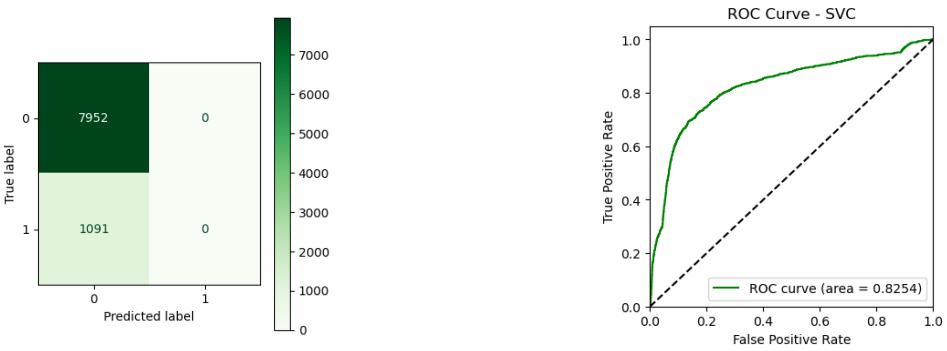


Figure (5) Confusion Matrix and ROC of SVM

Table (6)
Performance of SVM

Metric Name	Percentage
Accuracy	0.8794
Precision	0.7733
Recall	0.8794
F1 Score	0.8229
ROC	0.8253

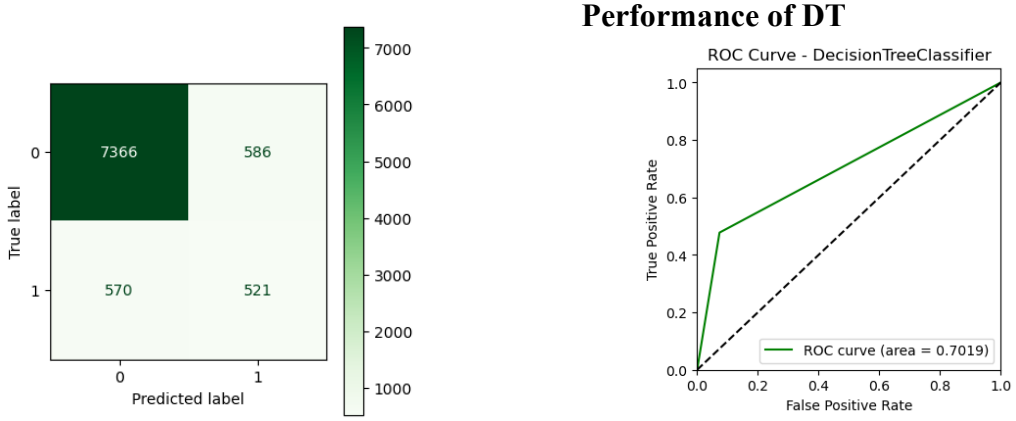
The SVM model displayed in the results demonstrates a solid predictive performance with an Accuracy of 87.94% and ROC AUC of 82.53%. Comparatively, this model exhibits slightly lower accuracy and ROC AUC LR model.

- **Decision Tree Performance**

Figure (6) Confusion Matrix and ROC of DT

Table (7)

Performance of DT



Metric Name	Percentage
Accuracy	0.8722
Precision	0.8730
Recall	0.8722
F1 Score	0.8726
ROC	0.7019

The Decision Tree model yields an Accuracy of 87.22% and ROC AUC of 70.19%. When compared to the Logistic Regression and Support Vector Machine models, the Decision Tree exhibits a slightly lower performance. Specifically, its accuracy is lower than the Logistic Regression's and marginally lower than the SVM's The ROC AUC also demonstrates a considerable decline, where Logistic Regression achieved and SVM scored.

- **Random Forest Performance**

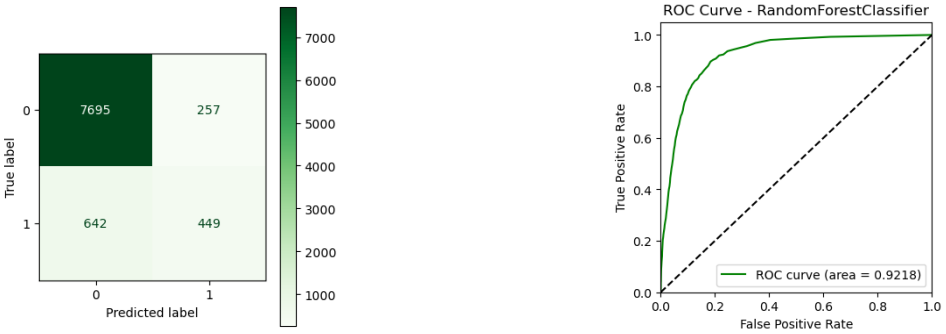


Figure (7) Confusion Matrix and ROC of RF

Table (8)
Performance of RF

Metric Name	Percentage
Accuracy	0.9006
Precision	0.8884
Recall	0.9006
F1 Score	0.8911
ROC	0.9218

The Random Forest (RF) model exhibits outstanding performance with an Accuracy of 90.06% and a ROC AUC of 92.18%, surpassing the Logistic Regression, Support Vector Machine, and Decision Tree models in both metrics.

- **Gradient Boosting Performance**

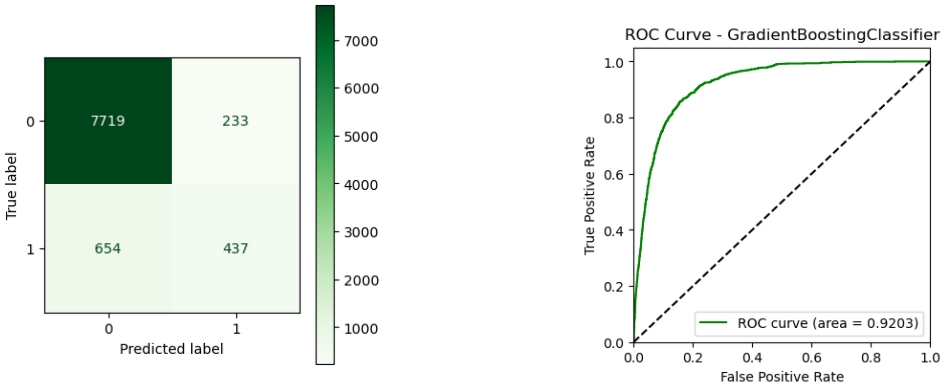


Figure (8) Confusion Matrix and ROC of GB

Table (9)
Performance of GB

Metric Name	Percentage
Accuracy	0.9019
Precision	0.8894
Recall	0.9019
F1 Score	0.8915
ROC	0.9203

The Gradient Boosting Classifier (GB) demonstrates impressive performance with an Accuracy of 90.19% and a ROC AUC of 92.03%, closely competing with the Random Forest model and surpassing the Logistic Regression, Support Vector Machine, and Decision Tree models.

- **K-Neighbors Classifier**

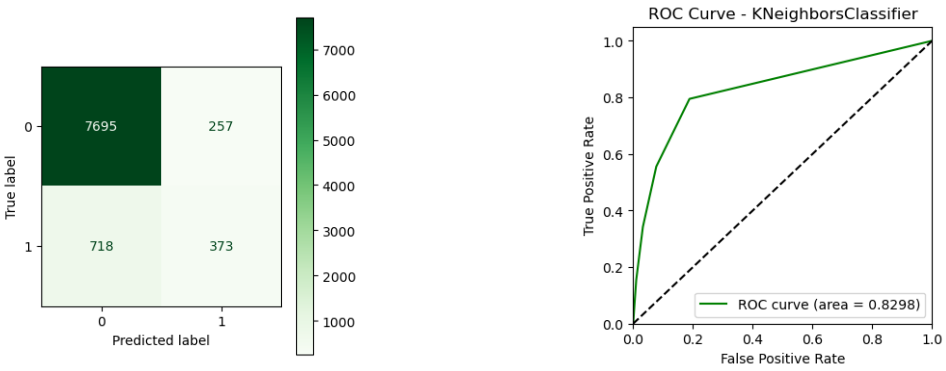


Figure (9) Confusion Matrix and ROC of KNN

Table (10)
Performance of KNN

Metric Name	Percentage
Accuracy	0.8922
Precision	0.8757
Recall	0.8922
F1 Score	0.8793
ROC	0.8298

The KNN model displays a solid performance with an accuracy of 89.22% and a ROC AUC of 82.98%. It achieves a commendable balance in its classification metrics, with a precision of 87.57%, recall of 89.22%, and an F1 score of 87.93%. These figures underscore its consistent classification performance across different scenarios.

- **MLP Classifier**

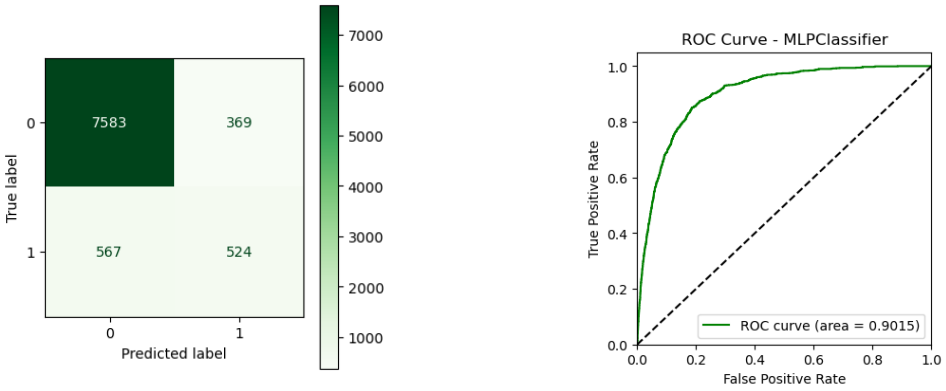
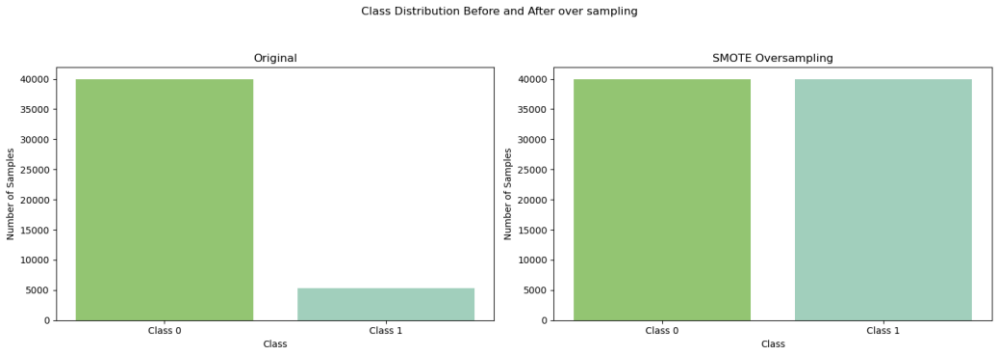


Figure (10) Confusion Matrix and ROC of MLP

Table (11)
Performance of MLP

Metric Name	Percentage
Accuracy	0.8922
Precision	0.8757
Recall	0.8922
F1 Score	0.8793
ROC	0.8298

The Multi-Layer Perceptron (MLP) classifier exhibits robust performance, achieving an accuracy of 89.22% and a ROC AUC of 90.15%. It demonstrates well-balanced classification capabilities, as evidenced by its commendable precision of 87.57%, recall of 89.22%, and an F1 score of 87.93%. These metrics highlight the model's consistent ability to accurately identify and classify cases across both categories.



Enhancing Results using SMOTE

Figure (11) Impact of SMOTE on Class Balance

This figure illustrates the effectiveness of SMOTE in addressing class imbalance. The left graph displays the original class distribution with a substantial disparity: 39,922 instances in Class 0 compared to only 5,289 in Class 1. After implementing SMOTE both classes are perfectly balanced, each with 39,922 instances.

Results for some Algorithms After using SMOTE

- **Decision Tree after SMOTE**

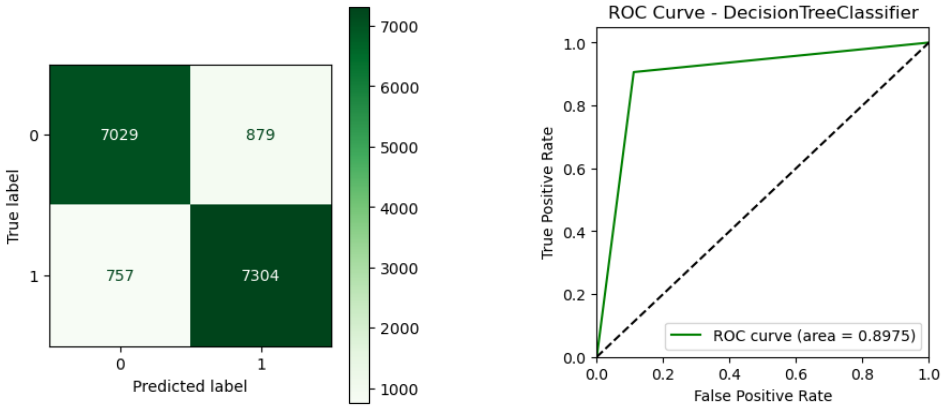


Figure (12) DT after SMOTE

Table (13)
Performance of DT after SMOTE

Metric Name	Percentage
Accuracy	0.8976
Precision	0.8976
Recall	0.8976
F1 Score	0.8975
ROC	0.8974

The DT classifier exhibits substantial improvement following the application of SMOTE, with its performance metrics significantly enhanced. Prior to applying SMOTE, the model achieved an accuracy of 87.22% and a ROC AUC of 70.19%. Post-SMOTE, the accuracy rose to 89.76% and the ROC AUC improved markedly to 89.74%.

- **Random Forest after SMOTE**

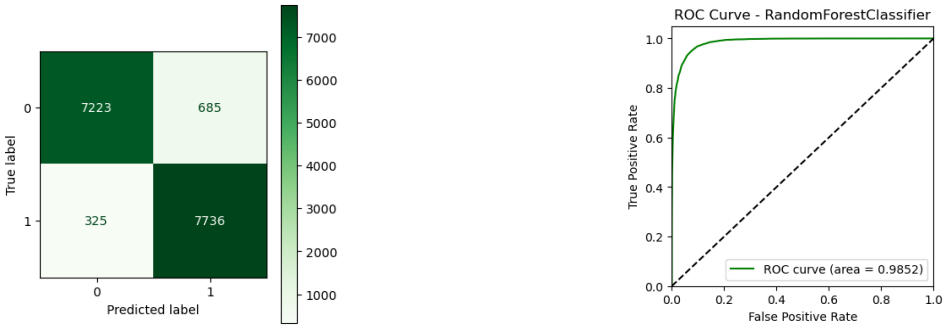


Figure (13) RF after SMOTE

Table (13)
Performance of RF after SMOTE

Metric Name	Percentage
Accuracy	0.9368
Precision	0.9376

Recall	0.9368
F1 Score	0.9367
ROC	0.9852

The RF classifier shows outstanding performance after the application of SMOTE, with a notable increase in its performance metrics. Initially, before SMOTE, the model recorded an accuracy of 90.06% and a ROC AUC of 92.18%. Following the application of SMOTE, accuracy surged to 93.68% and ROC AUC to an impressive 98.52%.

- **K-Neighbors after SMOTE**

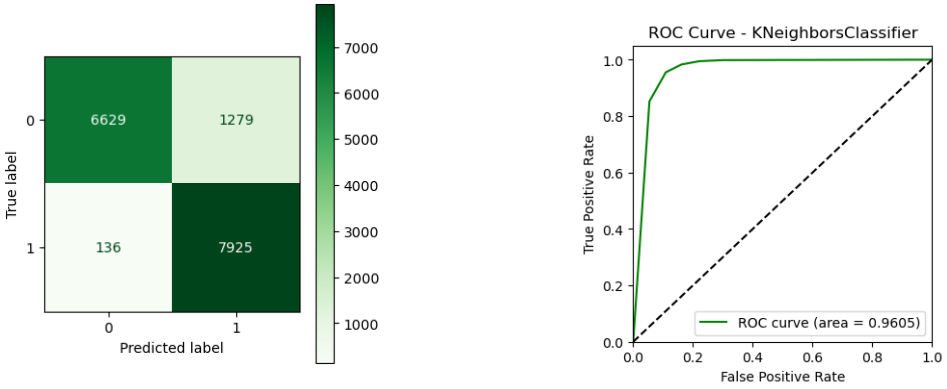


Figure (14) KNN after SMOTE

Table (14)
Performance of KNN after SMOTE

Metric Name	Percentage
Accuracy	0.9114
Precision	0.9199
Recall	0.9114
F1 Score	0.9109
ROC	0.96051

These results display the performance metrics of the KNN classifier after applying SMOTE. The notable improvement in performance metrics can be observed, with the accuracy rising from 89.22% before SMOTE to 91.14%

afterwards. The ROC AUC also shows a significant improvement from 82.98% to 96.05%.

From the results before and after using SMOTE, the improvement highlights the model's enhanced ability to accurately differentiate and classify cases across both categories, greatly benefiting from the strategic use of SMOTE to address class imbalance.

Enhancing Results using Standard Scaler

The Standard Scaler is an essential preprocessing tool in machine learning, particularly useful for datasets with features on varying scales. It standardizes features by removing the mean and scaling to unit variance, transforming each feature to have a standard normal distribution with a mean of zero and a standard deviation of one. The scaler calculates the mean and standard deviation for each feature individually to ensure accurate data transformation.

The standard scaling process can be mathematically expressed as:

$$z = \frac{x-\mu}{\sigma} \quad \textbf{(10)}$$

Where:

- z is the standard value,
- x is the original value,
- μ is the mean of the feature values,
- σ is the standard deviation of the feature values.

Enhancing Results using Neural Network

Table (15) Model Configuration

Parameter	Value
Batch Size	32
Epochs	50
Optimizer Type	Adam
Learning Rate	0.001
Number of Layers	3 (Dense layers)
Activation Function	ReLU (first two layers), Softmax (output layer)

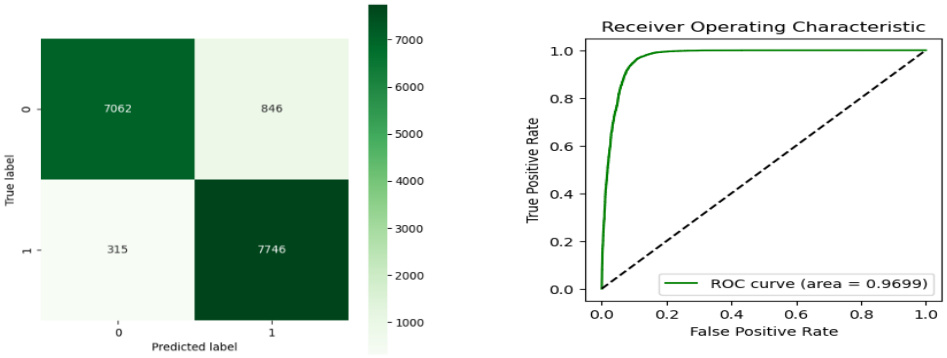


Figure (15) Confusion Matrix and ROC of Enhancing NN

Table (16)
Enhancing Performance of NN

Metric Name	Percentage
Accuracy	0.93
Precision	0.914
Recall	0.95
F1 Score	0.93
ROC	0.972

The results in Table (16) highlight the neural network's effectiveness, showing high accuracy 93%, precision 91.4%, and recall 95%. The F1 score of 93% and a ROC score of 97.2% confirm the model's strong ability to classify accurately and distinguish between classes effectively.

17- Enhancing Performance Using Bagging

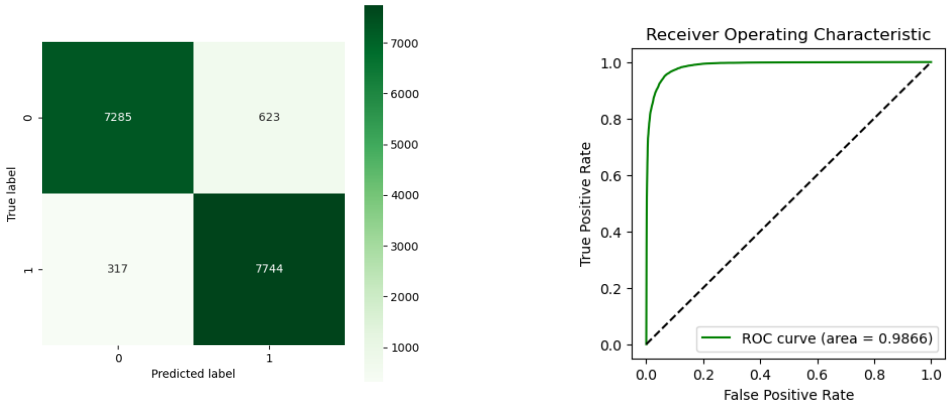


Figure (16) Confusion Matrix and ROC of Bagging

Table (17)
Performance of Bagging

Metric Name	Percentage
Accuracy	0.94
Precision	0.925
Recall	0.96
F1 Score	0.94
ROC	0.986

The results in Table (17) show that bagging significantly improves model performance, achieving an accuracy of 94% and a ROC score of 98.6% With a recall of 96%, precision of 92.5%, and an F1 score of 94%, the model effectively balances true positive identification and false positive minimization, confirming the efficacy of bagging for bank deposit classification.

18- Enhancing Performance Using Stacking

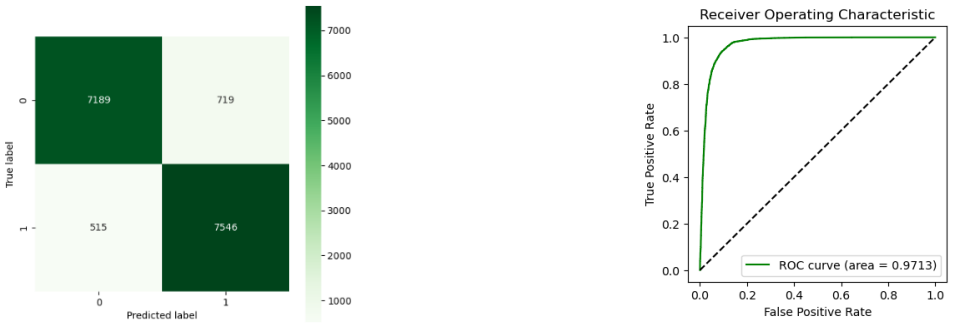


Figure (17) Confusion Matrix and ROC of Stacking

Table (18)
Performance of Stacking

Metric Name	Percentage
Accuracy	0.92
Precision	0.91
Recall	0.93
F1 Score	0.92
ROC	0.97

The results in Table (18) indicate that the stacking method performs exceptionally well, with an accuracy of 92%, precision of 91%, recall of 93%, and an F1 score of 92%. The high ROC score of 97% further confirms the model's strong discriminative capability.

Conclusion

This research has revealed the substantial benefits of integrating advanced statistical methods with machine learning techniques to improve the predictive accuracy of bank deposit trends. By carefully employing both traditional and modern analytical approaches, the study has shown that hybrid models excel in handling the complexities of financial datasets, particularly those from direct marketing efforts in the banking industry.

Using ensemble methods such as bagging and stacking, along with the strategic application of SMOTE, has effectively addressed class imbalances and significantly enhanced the reliability and predictive power of the models. The marked improvements in accuracy and ROC AUC values following these applications highlight the efficacy of these sophisticated techniques in refining predictive analytics.

Additionally, the incorporation of neural networks has deepened the analysis, providing a nuanced understanding of consumer behaviors and the intricate factors influencing bank deposit trends. These insights are crucial for banks aiming to optimize their financial strategies and marketing efforts in an increasingly unpredictable economic environment.

This study advances both academic and practical knowledge of financial predictive modeling, presenting a robust framework that can be adapted and expanded upon in future research. It emphasizes the importance of a balanced approach that combines traditional statistical techniques with the advanced capabilities of machine learning to address real-world financial forecasting challenges.

For banks and financial institutions, these findings offer a strategic guide to leveraging data analytics, enabling more informed decision-making that aligns with consumer needs and market conditions. As the financial landscape continues to evolve, the methodologies validated and refined in this research will be essential in enhancing economic planning, stability, and growth in the banking sector.

Recommendations

To fully capitalize on the insights derived from the statistical and machine learning analyses, the following strategic recommendations are proposed for implementation in banking practices:

- **Continuous Innovation**

Banks should continue to innovate by integrating advanced data analytics into their strategic operations to maintain a competitive edge.

- **Staff Training**

Implement ongoing training programs for bank staff to ensure they are well-versed in the latest analytical tools and methodologies.

- **Enhanced Data Management**

Invest in robust data management systems to ensure high-quality data collection, storage, and analysis capabilities.

- **Client-Centric Strategies**

Utilize insights gained from data analytics to tailor products and services to better meet the needs of individual customers.

- **Collaborative Data Sharing**

Encourage collaborations with other banks and financial institutions to share insights and best practices in data analytics, while ensuring data privacy and security standards are upheld.

- **Risk Management Optimization**

Apply machine learning algorithms to improve risk management strategies, particularly in credit scoring and operational risk assessments.

- **Regulatory Compliance**

Ensure that the use of data analytics complies with all relevant data protection and privacy regulations to maintain customer trust.

References

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.

Cui, S., Yin, Y., Wang, D., Li, Z., & Wang, Y. (2021). A stacking-based ensemble learning method for earthquake casualty prediction. Applied Soft Computing, 101, 107038.

Dutta, P. (2021). A Study On Machine Learning Algorithm For Enhancement Of Loan Prediction. International Research Journal of Modernization in Engineering Technology and Science, 3.

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115, 105151.

Ghasemi, Z., Afshar Kermani, M., & Allahviranloo, T. (2021). Exploring the Main Effect of e-Banking on the Banking Industry Concentration Degree on Predicting the Future of the Banking Industry: A Case Study. Advances in Fuzzy Systems, 2021(1), 8856990.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

Hayder, I. M., Al Ali, G. A. N., & Younis, H. A. (2023). Predicting reaction based on customer's transaction using machine learning approaches. International Journal of Electrical and Computer Engineering, 13(1), 1086.

Heider, F., & Leonello, A. (2021). Monetary policy in a low interest rate environment: Reversal rate and risk-taking.

Ibnu Choldun R, M., Santoso, J., & Surendro, K. (2020). Determining the number of hidden layers in neural network by using principal component analysis. In Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2 (pp. 490-500). Springer International Publishing.

Ilham, A., Khikmah, L., Indra, Ulumuddin, & Bagus Ary Indra Iswara, I. (2019, March). Long-term deposits prediction: a comparative framework of classification model for predict the success of bank telemarketing. In Journal of Physics: Conference Series (Vol. 1175, p. 012035). IOP Publishing.

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., ... & Reyes, M. C. (2021). A comparative performance analysis of data resampling methods on imbalance medical data. IEEE Access, 9, 109960-109975.

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees?. Nature biotechnology, 26(9), 1011-1013.

Koroniotis, N. (2020). Designing an effective network forensic framework for the investigation of botnets in the Internet of Things (Doctoral dissertation, UNSW Sydney).

Le, H. H., & Viviani, J. L. (2018). Predicting bank failure: An improvement by implementing a machine-learning approach to classical financial ratios. Research in international business and finance, 44, 16-25.

Lu, X. Y., Chu, X. Q., Chen, M. H., Chang, P. C., & Chen, S. H. (2016). Artificial immune network with feature selection for bank term deposit recommendation. Journal of Intelligent Information Systems, 47, 267-285.

Machine, M. D. U. E. (2020). Predicting loan approval of bank direct marketing data using ensemble machine learning algorithms.

Mohammed, N. A., & Al-Bazi, A. (2022). An adaptive backpropagation algorithm for long-term electricity load forecasting. Neural Computing and Applications, 34(1), 477-491.

MOHD-KARIM, M. U. H. A. M. M. A. D. (2010). Profit-sharing deposit accounts in Islamic banking: analysing the perceptions and attitudes of the Malaysian depositors (Doctoral dissertation, Durham University).

Nasteski, V. (2017). An overview of the supervised machine learning methods. Horizons. b, 4, 51-62.

Ramesh, R. (2017, February). Predictive analytics for banking user data using AWS machine learning cloud service. In 2017 2nd International Conference on Computing and Communications Technologies (ICCCT) (pp. 210-215). IEEE.

Ruangthong, P., & Jaiyen, S. (2015, July). Bank direct marketing analysis of asymmetric information based on machine learning. In 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 93-96). IEEE.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29.

Shibuya, E., & Hotta, K. (2022). Cell image segmentation by using feedback and convolutional LSTM. The Visual Computer, 38(11), 3791-3801.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130

Sun, R., Wang, G., Zhang, W., Hsu, L. T., & Ochieng, W. Y. (2020). A gradient boosting decision tree based GPS signal reception classification algorithm. Applied Soft Computing, 86, 105942.

Tsui, K. L., Chen, V., Jiang, W., Yang, F., & Kan, C. (2023). Data mining methods and applications. In Springer handbook of engineering statistics (pp. 797-816). London: Springer London.

Tuan, N. M. (2022). Machine Learning Performance on Predicting Banking Term Deposit. In ICEIS (1) (pp. 267-272).

Wilmarth Jr, A. E. (2020). Taming the megabanks: why we need a new Glass-Steagall Act. Oxford University Press, USA.

Wu, Q., & Zhou, D. X. (2006). Analysis of support vector machine classification. Journal of Computational Analysis & Applications, 8(2).

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.

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems, 29(5), 1774-1785.